PARENTAL ADVISORY:

ANALYZING THE EFFECT OF THE MATURE CONTENT LABEL

ON THE MUSIC INDUSTRY

A Thesis

Presented to the faculty of the Department of Economics

California State University, Sacramento

Submitted in partial satisfaction of
the requirements for the degree of

MASTER OF ARTS

in

Economics

by

William Michael Matsuoka

SPRING
2013
Student: William Michael Matsuoka

I certify that this student has met the requirements for format contained in the University format manual, and that this thesis is suitable for shelving in the Library and credit is to be awarded for the thesis.

__________________________, Graduate Coordinator
Kristin Kiesel, Ph.D.        Date

Department of Economics
Abstract

of

PARENTAL ADVISORY:

ANALYZING THE EFFECT OF THE MATURE CONTENT LABEL

ON THE MUSIC INDUSTRY

by

William Michael Matsuoka

Profanity in popular music has been blamed for causing an increase in adolescent violence following recent mass shootings. Using Visual Basic for Application scripts, I compile a dataset that includes profanity counts from 2688 songs (1970-2012) and 2186 movies (1980-2012) in order to analyze the effect of the Parental Advisory Label. This study finds that, using differences-in-differences estimators, there was a decrease in profanity in popular music following the adoption of the label. Furthermore, after evaluating the survival characteristics of number one hits, we find significant reductions in the longevity of songs with profanity; which implies the decrease in profanity following the labeling may be due to an artist response to consumer preferences.
Finally, using socioeconomic data from various government agencies and the aforementioned profanity set, we find no significant relationship between profanity and juvenile crime or teen pregnancy.

________________________, Committee Chair
Jonathan Kaplan, Ph.D.

________________________
Date
ACKNOWLEDGEMENTS

None of this would have been as rewarding as it has been without the discipline, devotion, and passion instilled in me by everyone who has left a mark on my life. To those who taught me to think for myself and to challenge the dogma that discourages freedom and to those who instilled a love for personal growth and learning, I thank you.

To the great Jonathan Kaplan, I cannot begin tell you how much I appreciate all of your comments added to this thesis. Your abundant use of question marks taught me that some of the most meaningful comments contain no words; a punctuation mark can be either extremely informative or completely devastating. Thank you for your haste and insight.

To the hip Terri Sexton, it has been a pleasure working with you this semester, both on my thesis and with teaching advice. After learning about your extracurricular activities and experiences, I can say you deserve nothing less than the title “hip,” especially when considering the context of this thesis. Thank you very much.

Special thanks goes to Kristin Kiesel, for without your food economics course, my GPA would be much higher (last time I mention that, I swear); but I would not have been pushed to grow and adapt to the responsibilities that are a part of being a graduate student. This definitely helped me in the long run. You have my sincere gratitude.

Finally, thank you to my family for support, my friends who had to endure my daily thesis babble, and everyone else who had an impact on this endeavor.
TABLE OF CONTENTS

Acknowledgements ................................................................................................................ vii

List of Tables ............................................................................................................................ x

List of Figures ........................................................................................................................... xi

Chapter

1. INTRODUCTION ............................................................................................................... 1

2. LITERATURE REVIEW .................................................................................................... 5
   2.1 Introduction ........................................................................................................... 5
   2.2 The Economics of Popular Music ................................................................................. 6
   2.3 Music as an Economic Good ......................................................................................... 10
   2.4 Profanity as an Externality ......................................................................................... 12
   2.5 Determinants of Juvenile Violence and Teen Pregnancy ............................................. 13
   2.6 Natural Experiement ................................................................................................. 14
   2.7 Summary .................................................................................................................... 15

3. DATA SEARCH METHODOLOGY ................................................................................ 16
   3.1 Introduction ............................................................................................................... 16
   3.2 Music ......................................................................................................................... 16
   3.3 Movies ........................................................................................................................ 27
   3.4 Summary .................................................................................................................... 31

4. DATA ................................................................................................................................. 32
   4.1 Introduction ............................................................................................................... 32
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Tables</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Subtitle Formats</td>
<td>30</td>
</tr>
<tr>
<td>4.1 Profanity Legend</td>
<td>35</td>
</tr>
<tr>
<td>4.2 Song Data Summary Table</td>
<td>36</td>
</tr>
<tr>
<td>4.3 Song Summary per Decade</td>
<td>37</td>
</tr>
<tr>
<td>4.4 Singles Data Summary</td>
<td>42</td>
</tr>
<tr>
<td>4.5 Singles Summary per Decade</td>
<td>43</td>
</tr>
<tr>
<td>4.6 Movie Data Summary Table</td>
<td>48</td>
</tr>
<tr>
<td>4.7 Movie Data Summary per Decade</td>
<td>49</td>
</tr>
<tr>
<td>4.8 Data Source and Units</td>
<td>52</td>
</tr>
<tr>
<td>4.9 Societal Data Summary</td>
<td>53</td>
</tr>
<tr>
<td>4.10 Societal Data Summary by Decade</td>
<td>53</td>
</tr>
<tr>
<td>5.1 Treatment Effect Statistical Significance</td>
<td>59</td>
</tr>
<tr>
<td>5.2 Interaction Term Sign for Significant Values</td>
<td>60</td>
</tr>
<tr>
<td>5.3 Maximum Likelihood Estimation of Parametric Models</td>
<td>63</td>
</tr>
<tr>
<td>5.4 Weibull Distributed Results by Year</td>
<td>64</td>
</tr>
<tr>
<td>5.5 Dependent Variable: ln(Teen Pregnancy)</td>
<td>68</td>
</tr>
<tr>
<td>5.6 Dependent Variable: ln(Juvenile Crime)</td>
<td>69</td>
</tr>
<tr>
<td>5.7 Teen Pregnancy with Difference Estimators</td>
<td>71</td>
</tr>
<tr>
<td>5.8 Juvenile Crime with Difference Estimators</td>
<td>72</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figures</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Screenshot of Wikipedia Hot 100 Page</td>
<td>17</td>
</tr>
<tr>
<td>3.2</td>
<td>Screenshot of Wikipedia Song Infobox</td>
<td>18</td>
</tr>
<tr>
<td>3.3</td>
<td>Screenshot of Wikipedia Artist Infoboxes</td>
<td>20</td>
</tr>
<tr>
<td>3.4</td>
<td>Wikipedia Excerpt with Highlighted Pronouns</td>
<td>22</td>
</tr>
<tr>
<td>3.5</td>
<td>Screenshot of Musicnotes.com Song Information</td>
<td>23</td>
</tr>
<tr>
<td>3.6</td>
<td>Screenshot of Bing’s Structure for Metrolyrics.com Searches</td>
<td>24</td>
</tr>
<tr>
<td>3.7</td>
<td>Screenshot of Metrolyrics.com Single Artist Pages</td>
<td>25</td>
</tr>
<tr>
<td>3.8</td>
<td>Screenshot of Wikipedia Number One Singles Index</td>
<td>27</td>
</tr>
<tr>
<td>3.9</td>
<td>Screenshot of Movie Data from Boxofficemojo.com</td>
<td>28</td>
</tr>
<tr>
<td>3.10</td>
<td>Screenshot of Single Movie Data</td>
<td>29</td>
</tr>
<tr>
<td>3.11</td>
<td>Screenshot of Movie Search Engine</td>
<td>29</td>
</tr>
<tr>
<td>4.1</td>
<td>Major Scales per Year</td>
<td>38</td>
</tr>
<tr>
<td>4.2</td>
<td>Profanity in Songs per Year</td>
<td>39</td>
</tr>
<tr>
<td>4.3</td>
<td>Average Profanity per Minute per Year</td>
<td>39</td>
</tr>
<tr>
<td>4.4</td>
<td>Song Genre per Year</td>
<td>40</td>
</tr>
</tbody>
</table>
Chapter I

Introduction

The “Parental Advisory: Explicit Content” label (PAL) was introduced in 1995 by the Recording Industry Association of America (RIAA) to help parents make the right decision when deciding what songs are appropriate for their children.¹ According to the RIAA, the label comes with a set of uniform standards that can apply to both single songs as well as albums. These standards are supposed to represent cultural morals and take into account many factors such as artist target audience, words, phrases, sounds, context, and frequency when applied to the song or album. One downside of the current label is that it does not convey information to the consumer besides the fact that there could be offensive material.

Purging the music industry of objectionable content has been a central focus for many agencies and politicians alike since the early 1980s. In 1985, after realizing that profanity exposure in the media was becoming an increasing issue, Tipper Gore, Susan Baker, Pam Hower, and Sally Nevius founded the Parental Music Resource Center (PMRC). The goals of the PMRC were to limit children’s exposure to drug use, sexual content, and profane material by use of a label, which predated the RIAA’s label. However, this faced much opposition by artists in the music industry who believed they were being unjustifiably censored. The purpose of these labels was to provide information to parents, but the effect of these labels on the industry is still unknown.

¹ Additional information on the parental advisory label can be found at www.riaa.com/toolsforparents.php?content_selector=parental_advisory.
This thesis looks at the music labeling system and analyzes the economic and linguistic impact of the PAL. We take existing economic models such as differences-in-differences, accelerated time failure models, and time series models with growth rates and first differences and we modify them by including profanity variables in order to fit within the context of consumer preferences and social benefits. Because we need data on curse words in songs and movies, this thesis also serves as a guide for extracting this data from various websites. Visual Basic for Applications (VBA) scripts are used to automate this process and are provided in Appendix A. Once these data are catalogued, we continue with the development of our economic models by including these newly mined profanity indicators.

The differences-in-differences model looks at profanity in both the music and movie industry and uses the movie industry as a baseline for social profanity prevalence. Because the movie industry was not subject to the same labeling policy as the music industry, while also representing a sample whose players have similar preferences to those of the music industry, we are able to use movies as a control group. We test the change in profanity in the periods before and after the implementation of the PAL using different representations of each curse word of interest. The accelerated time failure model is a parametric model in which we evaluate the time a song stays on the charts, while controlling for attributes of the song, the artist, and the lyrics. This model serves to test the survivability of songs with profanity compared to those without. Our final model is a time series model that evaluates teen pregnancy and juvenile crime against determinants established in Chapter II (Literature Review) as well as three different
indices that represent profanity in songs. We look at both growth rates and first differences when deciding if the index is a significant determinant.

The results of the analysis still have some ambiguity as to whether the profanity label affects society as a whole. Our juvenile crime and teen pregnancy models show that there is no statistical significance for a profanity indicator as a determinant of either problem. More encouraging, the differences-in-differences estimators found a reduction in profanity per year and profanity per minute per year in songs following the introduction of the PAL compared to the movie industry. The consumer preference accelerated time failure model also found that, after 1995, curse words decreased the time a song spent in the number one position on the Billboard *Hot 100* list. Because of limitations in the data, we are unable to evaluate if there was a change in consumer preferences pre- and post-PAL. This raises the question of whether or not producers were responding to the PAL itself, or if they changed composition in their songs due to consumer preferences. It is worth noting that the observed consumer preference suggests PAL was effective in reducing profanity in popular music.

This remainder of this thesis contains five chapters and an appendix that presents the VBA scripts. The next chapter is a review of the literature, which summarizes existing literature on economic studies of popular music, profanity, teen violence and pregnancy, and a discussion of the foundations for the models used in this thesis. Chapter III explains the search methodology for the data and provides the necessary steps for compiling all of the song, artist, movie, and profanity variables in each model. The purpose of Chapter III is to provide some transparency on the extensive steps taken to
provide accurate, unbiased data. Chapter IV (Data) is a summary of the data retrieved through the web-based search algorithms described in Chapter III, and also includes socioeconomic variables gathered from different government agencies. This chapter explores relationships between variables and analyzes trends or changes in the variables of interest over time. Chapter V describes the empirical analysis Chapter VI concludes with a summary of the results, and discussion of inferences drawn from these results and of future research needs.
Chapter II

Literature Review

2.1 Introduction

Economic studies of popular music have been limited by the lack of publicly available data. Sales, revenue, and consumption data are available but require a steep fee to acquire. Because of the scarcity of the data, and perhaps a general disinterest in the formal econometric study of popular music, the literature surrounding media studies by economists is quite limited. A quick database search reveals that commodities such as steel (2,402 results), rubber (5,544 results), and electricity (7,779 results)) are much more widely studied than music (646 results) or movies (402 results). Recently, with tragic events such as the Sandy Hook Elementary Shooting and even earlier Columbine, many people have cited the media as a culprit for these violent attacks by adolescents. While this may seem conducive to the goals of the parental advisory label, little evidence exists in the academic libraries that illustrate a causal link between curbing explicit content and reducing violence. While efforts such as background checks for purchasing firearms and gun control regulations are being implemented to act as a buffer between these socially unstable adolescents and firepower, some individuals are more concerned with addressing what they believe to be the real culprit: violence in music, movies, and video games. An even smaller group, Coyne et al. (2011), believes that this violence can be directly linked to profanity use; ergo, profanity in the media is one of the determinants of

---

2 Niesen SoundScan, RIAA Shipment Data, MPAA Reports & Data

violent crimes. Given the notion that this correlation exists, we are interested in determining what role music plays in evoking violence in teenagers to justify a social need for the parental advisory label.

Because the literature surrounding econometric music analysis is not as extensive as that other goods, the rest of this chapter will be devoted to establishing key assumptions and summarizing existing models in order to build a comprehensive viewpoint of the industry. The first section is a review of the economics of popular music that includes market trends and identifies some key behaviors of buyers and sellers in the market. Similar to the first section, the second part looks at music as an economic good. The third section analyzes the literature regarding profanity and violence in music as externalities. The fourth section looks at the societal determinants of violence and teen pregnancy. Finally, the fifth section is a summarization of existing models in the analysis of popular music that will provide a basis for our empirical analysis.

2.2 The Economics of Popular Music

The North American Music Industry has seen a dramatic change in revenue sources in the past few years. Since 2001, British CD sales fell by 40%, and in 2009 alone, Japan’s CD sales fell by 24% after an 8% decrease in the previous year (The Economist 2010). In the United States, the music industry faced a 13.5% drop in physical album sales from 2011 to 2012, and a 14.1% increase in digital album sales as well as a 5.1% increase in digital track sales (The Nielsen Company and Billboard 2012).
Digital music sales increased over time and overtook physical sales in 2011. Concert ticket revenues also experienced growth from $1.5 billion in 1995 to an estimated $4.6 billion in 2009 or an estimated 141% growth after accounting for inflation (The Economist 2010). The broader music industry’s total worth was estimated to be close to $140 billion in 2009 and grew to an estimated $167.7 billion in 2010 (IFPI 2010).

Turning to the producers in the music industry, concert sales in the United States for the top 3 artists (U2, Taylor Swift, and Kenny Chesney) generated $338 million alone in 2011 (Pollstar 2011). People are generally willing to pay up to $72 dollars on average for these tickets, and this figure does not include any price discrimination or black markets (Connolly and Krueger 2006).

When considering the players in the market for music, we will define the consumers as those who demand and purchase music in any format, and producers as the artists or bands that produce the songs. Recording companies and studios are not considered the producers of music since the consumer’s purchase of music is presumably independent from the record company representing the artist. Every song can be treated as an individual good with many close substitutes or complements which represent all other songs that are similar in composition. I use the qualifier “similar in composition” because popular music of the 2000s and popular music of the 1940s are likely unrelated goods given current consumer preferences for new music. Likewise, a song that shares the same artist is likely to be a complementary good since they are consumed together on a playlist or album, whereas songs from different artists or playlists can be treated as

---

substitutes. Asai (2011) analyzes demand for CD’s in the Japanese popular music industry and finds that album and single sales are statistically dependent on whether the consumer is a fan of the artist. While these results seem like common knowledge, Asai also found that the artist, the gender of the artist, and the price of albums are not statistically significant for album sales. This seems to suggest that consumers prefer characteristics of individual songs rather than album characteristics. This notion of a fan-based economy seems to be supported by the data. Crain and Tollison (2002) study the market distribution of artists in the music industry and the returns to artist quality. They use the Billboard Hot 100 number one singles list to compute the share of the weeks held by the top four and five artists. Their analysis regresses artist concentration per year on earnings, teenage share, interest rates, death rates of active military members, inflation rates, musical variety, and number of FM radio stations to control for economic factors that might influence an artist’s popularity as found in Crain and Tollison (1997). Their results seem to confirm the existence of the “superstar” effect in popular music, where the share of number one hits is disproportionate to the artist’s talents. This implies that an artist’s earnings are not an accurate representation of their ability, but are skewed due to a “superstar” effect.

Over the last decade, digital downloads have been a changing force in both the consumption and production of songs. Stevans and Sessions (2005) study the Recording Industry Association of America’s (RIAA) estimates that CD sales fell by 4.5% in 2001 and 8.1% in 2002 as a direct result of illegal downloading. While they find a significant decrease in sales as a direct result of file sharing websites, ultimately, legal digital
downloads and DVD prices were large contributors to this drop in sales. Because consumers can now download a single song for $1.29 on iTunes instead of having to buy an entire album, producers theoretically have more of an incentive to create higher quality songs to maximize revenue. Let us say an album costs $11.99 on iTunes. If a consumer has a marginal willingness to pay above $1.29 for two songs on an album, and a marginal willingness to pay of $0 for the rest of the songs combined, then the consumer would be better off simply buying the two songs for $2.58 compared to the $11.99 they would have to pay for the entire album. Previously, they would have to value the two songs above $11.99.

Perhaps the main issue that is facing suppliers in the music industry is the prevalence of piracy that has been made possible by torrents and file sharing websites. Piracy is a form of copyright infringement that arises when there is a duplication of a song that is transferred from one user to another. Peitz and Waelbroeck (2005) find that piracy is a factor for the decline in demand for CDs, but much like Stevans and Sessions (2005), digital downloads were found to be a large factor. In regards to file sharing and displacement of CD sales, Waldfogel (2010) reaffirms that file sharing did reduce CD sales; however, Piolatto and Schuett (2012) find that the file sharing prevalence can actually have a positive impact on producers. In their results, the data show that, partially due to the aforementioned “superstar effect,” popular artists actually receive increased revenue from having their music widely disseminated because this effect has the potential to boost the artist’s listener base which has a positive impact on concert and merchandise
sales. Conversely, artists that do not have a large listener base to begin with suffer from file sharing because music sales make up the majority of their income.

2.3 Music as an Economic Good

Like any economic good, songs provide utility when they are “consumed” by an individual; in some cases, they can provide negative utility. The issue then arises as to how the consumer derives their benefit from listening to songs. Each consumer has different tastes and preference regarding song choice. While the determinants of why individuals prefer certain songs will not be examined in this thesis, we can easily break songs down into smaller elements. This hedonic pricing point of view is a way to determine the value, on average, of observable characteristics. Such characteristics might include genre, the artist, the artist’s gender, choice of instruments, length of the song, theme, song key, chord structure, other aspects rooted in musical theory, and the main focus of this analysis: lyrics. The idea of analyzing the individual parts of music is not new. Crain and Tollison (1997) compile songs from the Billboard Hot 100 list and examine the song structure over time, which they refer to as the “music architecture.” This is appropriately named since the structure of a song can be as intricate as blueprints for a building. Unfortunately, due to time and technology constraints, Crain and Tollison only use beats per minute (BPM) and song duration as their structural elements. This is similar to only drafting the top floor of a tall skyscraper: there are many other aspects to the music architecture that they did not consider. Their control variables include artist popularity, weeks on the charts, and measures for demographic and economic trends such
as teenage population share and interest rates. They find, not surprisingly, that music is largely tied to the economic climate. Song meter is statistically influenced by the TV share, or the operating broadcast television stations divided by the total number of operating broadcast radio and TV stations in a year, and the misery index, which is the probability of dying in war. Crain and Tollison argue that TV share is the availability of close substitutes where the BPM of the song is driven by the rate of time preferences and availability of close substitutes. They also explain that when there is a higher probability of death in war, songs tend to be faster on average. For further research, Crain and Tollison (1997) mention using this same methodology and adding control variables for certain words or themes in music to analyze the effect on drug use and teenage pregnancy, which is exactly what this thesis does.

Crain and Tollison (1997) are not the only economists to capitalize on using the Billboard Hot 100 list as a metric. Bradlow and Fader (2001) use a Bayesian model to estimate which factors influence an artist’s ability to keep their number one spot on the chart. They use both internal and external societal factors much like those used in Crain and Tollison (1997). Giles (2007) devises a parametric accelerated time failure model building upon the idea that internal aspects of a song might influence its popularity. He uses variables such as the year the song debuted, dummy variables for the Beatles and Elvis, variables for artist gender, groups, if the song had a “bounce back” on the charts, purely instrumental songs, and a control for the digital age. Giles found that there appears to be a “superstar effect” in the data and that there are statistically significant coefficients for female and purely instrumental songs.
2.4 Profanity as an Externality

Similar to music, profanity can be viewed as an economic good or an economic bad given one’s age, financial situation, gender, religion, or cultural background. Many have questioned whether profanity influences violent behavior, hypersexual behavior, or criminal activities. For example, Coyne et al. (2011) claim a correlation exists between an adolescent’s exposure to profanity in the media and their level of violence or aggression. However, their analysis primarily focuses on which of their models best fits the data. The data are subject to many biases that the authors have identified. First, there exists a response bias because the data comes from a survey they conducted. Second, there is a large threat to the external validity of their model because of their chosen sample: 223 students aged 11-15 from the same middle school in the Midwest. This sample is not likely representative of the entire population of teenagers in the United States and therefore one might see national trends extrapolated from this sample as unreliable. Their model also does not establish the direction of the correlation. However, the media have widely cited this study implying that profanity leads to aggressive behavior when aggressive behavior could actually be affecting profanity use. The authors do not control for this simultaneous causality. Coyne et al. (year) also omits income, education level of parents, mental health, and other key determinants, and further threatening the internal validity of their analysis. This thesis explores these previously excluded factors by regressing violence and teen pregnancy on these determinants as well as a profanity index variable.

2.5 Determinants of Juvenile Violence and Teen Pregnancy

Violence, teen pregnancy, and substance abuse seem to share similar determinants and are even used as controls for one another in separate studies, which taken together suggest some level of simultaneous causality. Molina and Duarte (2006) look at the risk determinants for suicide attempts among adolescents and find statistical significance for substance abuse, gun possession, sexual activity, age, race, and weight obsession. In an earlier study regarding teenage abortion demand, King, Myers, and Byrne (1992) use a Probit Likelihood model and find significance for other factors such as the unemployment rate, income, education, religiosity, and race. In the same year, Hayward, Grady, and Billy (1992) analyze similar characteristics; however, they place more of an emphasis on age, religiosity, and contraception and their effect on teenage pregnancy. While this thesis does not have a measure for religiosity or contraception, the other determinants are taken into account for the societal regression. This is because we are looking at national level data rather than panel state data. Overall, religiosity is not expected to change much over time and national contraception data is limited.

Past research on juvenile crime includes similar independent variables to those used in the teen pregnancy and abortion studies. Mocan and Rees (2005) look at the main determinants: race, age, religious background, unemployment, but they also include determinants such as population density, single parent households, and violent crime. They state that there is a growing literature that shows that crime is correlated with past values due to an increase in criminal human capital. Similarly, Mocan and Tekin (2006) include micro level data such as if the juvenile has tattoos, their school GPA, and
previous fighting experience compared to the likelihood of violence. These studies add
two important elements to the societal regression. The first element is the other key
determinants such as single parent households. The second element is the fact that the
time series data is highly correlated with past values, which suggests that the data might
not be stationary. To correct for unit roots, we use the first difference when evaluating
our societal issues model. We also attempt to control for correlation in the errors by
using a Prais-Winsten transformation.

2.6 Natural Experiment

While popular opinion seems to suggest that profanity has increased over time, it
is impossible to tell if the growth rate of profanity decreased after the implementation of
the PAL. This is why we use movies in the same time period to act as a control group to
look at trends in profanity over time in two similar but independent groups. We employ a
differences-in-differences model to capture the effectiveness of the label relative to the
movie industry. Differences-in-differences estimators were first used by Card and
Krueger (1995) whose results in in evaluating minimum wage laws in New Jersey were
seen as controversial. Looking at New Jersey alone, it appeared that the minimum wage
laws decreased (full time equivalent) FTE employment in the fast food sector. However,
when controlling for changes in all omitted variables that do not change over time by
using the neighboring state of Pennsylvania, it is found that FTE employment actually
increased in New Jersey. We replicate these results in the analysis by using movies as
our control group, and music as our treatment group.
2.7 Summary

To understand the change in profanity in music following the label, we use differences-in-differences estimators, which treat the movie industry as a control group since it was not subject to the restrictions of the parental advisory label. In order to obtain the information we need, we use the idea of music architecture to extract profanity from music and treat them as an input to look at changing patterns over time. We also build upon Giles (2007) by replicating his framework and variables selected to see how consumers react to certain characteristics. However, we build upon the model by adding variables such as song duration as presented in Crain and Tollison (1997), song key, as well as profanity counts to see if they increase or reduce the risk of being knocked out of the number one position. Our societal regression takes key determinants from studies such as Molina and Duarte (2006); King, Myers, and Byrne (1992); and Mocan and Rees (2005) to build a model that controls for certain characteristics. The limitation of this latter model is that it is used to understand national-level time series data.
Chapter III

Data Search Methodology

3.1. Introduction

This section is divided into two main parts to explain the process behind building each dataset. The first part relates to music and the process undertaken in order to organize data in a representative manner of popular songs. This part is further broken down into two sections. The first section talks about gathering data from the *Hot 100* list while the second section deals with the *Hot 100* Year-End number one singles list.\(^6\) Contrary to how similar they sound, both sets require very different processes for data extraction. The second part of the methodology explains what steps need to be taken to gather a comprehensive movie dataset.\(^7\)

3.2 Music

First and foremost, we use Wikipedia for a majority of this section. Although many do not cite Wikipedia as a credible source, it provides an excellent indexing system that makes it easier to merge datasets with one another. This process will be further discussed later in this chapter.

---

\(^6\) The *Hot 100* year-end singles list is the top 100 hits per year as opposed to the year-end number one singles list, which is a representation of the songs which stayed in the number one position for each week of the year.

\(^7\) The scripts used to extract the data are contained in Appendix A.
3.2.1 The Hot 100 List

The Hot 100 Year-End Single list, more commonly referred to as the Hot 100 list, is available at http://www.billboard.com/charts/year-end. Wikipedia’s version can be found at http://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles_of_2010, and the single factor that makes Wikipedia much easier to extract data from is its URL. Notice that the last four characters of the Wikipedia version is the year. A simple loop from 1970 to 2012 allows us to gather information from each page. Figure 3.1 is a screenshot of the list in 2010.

![Figure 3.1. Screenshot of Wikipedia Hot 100 Webpage](image)

<table>
<thead>
<tr>
<th>No</th>
<th>Title</th>
<th>Artist(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&quot;Tik Tok&quot;</td>
<td>Kesha</td>
</tr>
<tr>
<td>2</td>
<td>&quot;Need You Now&quot;</td>
<td>Lady Antebellum</td>
</tr>
<tr>
<td>3</td>
<td>&quot;Hey, Soul Sister&quot;</td>
<td>Train</td>
</tr>
<tr>
<td>4</td>
<td>&quot;California Gurls&quot;</td>
<td>Katy Perry featuring Snoop Dogg</td>
</tr>
<tr>
<td>5</td>
<td>&quot;OMG&quot;</td>
<td>Usher featuring will.i.am</td>
</tr>
</tbody>
</table>

Each row of the table in the figure contains five key points of information. We are able to get the rank of the song, the name of the song, and the name of the artist(s). We also obtain the links of both the song’s Wikipedia page and the artist’s Wikipedia page. These factors serve to provide a simple way to merge datasets even when names are incorrectly spelled or altered and to give us a link to more information. The visual basic for applications (VBA) script is available in Appendix A.1.
3.2.2 Song Information

When we follow the link to the song summary page, we see a brief description of the song in the Wikipedia “Infobox.” This is the box on the right side of the webpage that contains a synopsis the year the song was released, the album it came from, and other information pertaining to the song.

**Figure 3.2. Screenshot of Wikipedia Song Infobox**

<table>
<thead>
<tr>
<th>&quot;Need You Now&quot;</th>
<th>LADY ANTEBELLUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single by Lady Antebellum</td>
<td>from the album Need You Now</td>
</tr>
<tr>
<td>Released</td>
<td>August 24, 2009&lt;sup&gt;31&lt;/sup&gt;</td>
</tr>
<tr>
<td>Format</td>
<td>CD single, Music download</td>
</tr>
<tr>
<td>Genre</td>
<td>Country pop, soft rock</td>
</tr>
<tr>
<td>Length</td>
<td>3:56 (single/main version), 4:38 (video version)</td>
</tr>
<tr>
<td>Label</td>
<td>Capitol Nashville, Parlophone (UK release)</td>
</tr>
</tbody>
</table>
We can see that there are two areas of interest: genre and length. In this example, there are two genres and two lengths. Both genres will be captured; however, only the first length will be recorded in the dataset as it is typically the standard version of the song.₈ (Appendix A.3.)

3.2.3 Artist Information

Following the link to the artist’s Wikipedia page will allow us to construct a profile of each artist, which can be merged into other sets. Once again, the Infobox is an important tool in this step and is shown in Figure 3.3.

Besides the artist’s name, the Infobox provides us with information about the artist’s age and/or years active. Most importantly, the color matters. Golden boxes signify an individual artist, blue boxes represent groups, and other colors represent orchestras, conductors, or non-vocal instrumentalists. After adding these data to our set, we gather the single artist’s gender. The method for capturing this information involves finding key gender specific related words.

₈ https://en.wikipedia.org/wiki/Template:Infobox_album: “…only list the most common length or that of the standard edition…”
Figure 3.3. Screenshots of Wikipedia Artist Infoboxes

<table>
<thead>
<tr>
<th>Taylor Swift</th>
<th>Zac Brown Band</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Taylor Swift" /></td>
<td><img src="image2" alt="Zac Brown Band" /></td>
</tr>
<tr>
<td><strong>Swift performing in St. Louis as part of her 2013 Red Tour</strong></td>
<td><strong>Zac Brown Band at Walmart Soundcheck L-R: Coy Bowles, Clay Cook, Daniel Reyes, Zac Brown (Front), Jimmy Martini, John Hopkins, Chris Fryar</strong></td>
</tr>
<tr>
<td><strong>Background information</strong></td>
<td><strong>Background information</strong></td>
</tr>
<tr>
<td><strong>Birth name</strong></td>
<td>Zac Brown Band</td>
</tr>
<tr>
<td>Taylor Alison Swift</td>
<td><strong>Origin</strong></td>
</tr>
<tr>
<td><strong>Born</strong></td>
<td><strong>Genres</strong></td>
</tr>
<tr>
<td>December 13, 1989 (age 23)</td>
<td></td>
</tr>
<tr>
<td>Reading, Pennsylvania, United States</td>
<td></td>
</tr>
<tr>
<td><strong>Genres</strong></td>
<td><strong>Years active</strong></td>
</tr>
<tr>
<td>Country, country pop, pop, pop rock</td>
<td><strong>Labels</strong></td>
</tr>
<tr>
<td><strong>Occupations</strong></td>
<td></td>
</tr>
<tr>
<td>Singer-songwriter, musician, actress</td>
<td></td>
</tr>
<tr>
<td><strong>Instruments</strong></td>
<td><strong>Associated acts</strong></td>
</tr>
<tr>
<td>Vocals, acoustic guitar, electric guitar, banjo guitar, ukulele, piano</td>
<td><strong>Website</strong></td>
</tr>
<tr>
<td><strong>Years active</strong></td>
<td><strong>Members</strong></td>
</tr>
<tr>
<td>2006–present</td>
<td>Taylor Swift</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.4 is a quote from an actual Wikipedia page. Notice the highlighted words. A script is created to take the entire article and search for subjective pronouns (he, she, they), objective pronouns (him, her, them), possessive pronouns (his, her, their), and absolute possessive pronouns (his, hers, theirs). The script (Appendix A.2) catalogues each reference to a female, male, or plural pronoun and the total count is added to the dataset. As illustrated in Figure 3.4, we can see that many articles may contain more than one gender specific pronoun. The green highlighted words represent masculine pronouns and the yellow highlighted words represent feminine pronouns. Because each article contains a mixture, a simple majority is taken to find the most prevalent gender specification. Since the color of the box is a definitive way to establish groups, we use the plural pronouns as a check to test the accuracy of this method.

Approximately 90% of the artists are correctly identified as single artists or groups. Turning to single artists, when we restrict the sample to only single artists and look at the difference between the count of feminine and masculine pronouns, we find that only 87 artists have an ambiguously close difference in the two. Selecting a small sample of 100 individual artists at random, we find that each artist had the correct gender assigned using this technique. A caveat should be added that this technique is not completely robust to errors-in-variables bias; however, in our sample I was unable to find any errors having to do with the gender of a single artist being calculated incorrectly.
Taylor Alison Swift was born on December 13, 1989 in Reading, Pennsylvania. Her father, Scott Kingsley Swift, is a Merrill Lynch financial adviser. He was raised in Pennsylvania, and is the descendant of three generations of bank presidents. Her mother, Andrea (née Finlay), is a homemaker who previously worked as a mutual fund marketing executive. Andrea spent the first ten years of her life in Singapore, before settling in Texas; her father was an oil rig engineer who worked throughout Southeast Asia. Swift was given a gender-neutral name because her mother believed it would help her forge a successful business career. She has a younger brother, Austin, who attends the University of Notre Dame.

She spent the early years of her life on an eleven-acre Christmas tree farm in Montgomery County, Pennsylvania. She attended preschool and kindergarten at the Alvernia Montessori School, run by Franciscan nuns, and was later educated at the fee-paying Wyndcroft School. When Swift was nine years old, the family moved to Wyomissing, Pennsylvania, where she attended West Reading Elementary Center and Wyomissing Area Junior/Senior High School. She summered at her parents' vacation home in Stone Harbor, New Jersey and has described it as the place "where most of my childhood memories were formed."

Swift's family owned several Quarter horses and a Shetland pony and her first hobby was English horse riding. Her mother first put her in a saddle when she was nine months old and she later competed in horse shows. At the age of nine, Swift became interested in musical theatre. She performed in many Berks Youth Theatre Academy productions and traveled regularly to Broadway, New York for vocal and acting lessons. Swift then turned her attention to country music and spent her weekends performing at local festivals, fairs, coffeehouses, karaoke contests, garden clubs, Boy Scout meetings and sporting events. At the age of eleven, after many attempts, Swift won a local talent competition by singing a rendition of LeAnn Rimes's "Big Deal", and was given the opportunity to appear as the opening act for Charlie Daniels at a Strausstown amphitheater. This growing ambition began to isolate Swift from her middle school peers.

### 3.2.3 Song Key

The key of the song is of interest because it can roughly describe if the song is “happy” or “sad” in tone (Halpern, Martin, and Reed 2008). The key of the song is found by looking at the sheet music associated with the song from musicnotes.com. The search engine, Bing, is used to query “site:musicnotes.com Title Artist.” This provides us with a list of each song and their respective sheet music. The corresponding script is under Appendix A.4. An error correction is added that will be explained further in section 3.2.4. Following the links takes us to the sheet music page where we are primarily interested in information shown in Figure 3.5.
23

Figure 3.5. Screenshot of Musicnotes.com Song Information

We will take the “Original Published Key” and add it to the dataset.

3.2.4 Song Lyrics

Retreiving song lyrics is, by far, the most difficult part of this process. After reviewing nine prominent lyric websites, the one that provided the most extensive data and comprehensive index system is metrolyrics.com. In what follows, the commands are relevant to this site alone, but the general steps can be applied to other lyric databases. The first step is to use the search engine Bing to locate artists. A screenshot of the format is shown in Figure 3.6, and the script is in Appendix A.5.
Reviewing Figure 3.6, we have to control for the difference between song and artist webpages. As we can see, there are patterns that arise from these search results. All song pages contain the artist name followed by a “dash.” Notice, the actual artist page is labeled “Coldplay Lyrics,” without any dashes. Once we tell the script to gather these data, we can filter through our list of artists. After creating a master list, we are able to follow the links to the next page which is illustrated in Figure 3.7.
At this point, we have to employ the error correction previously mentioned. We use a Levenshtein distance calculator in order to find the closest matching string to the title of the song as explained in Ristad and Yianilos (1998). The equation is as follows:

\[
lev_{a,b}(i,j) = \begin{cases} 
\max(i,j) \\
\min\left(lev_{a,b}(i - 1,j) + 1, lev_{a,b}(i,j - 1) + 1, lev_{a,b}(i - 1,j - 1) + [a_i \neq b_j]\right)
\end{cases}, \quad \text{if } i, j > 0
\]

This equation gives the minimum number of steps it takes to transform one string into another. The subscripts \(a\) and \(b\) are two different strings. For instance, \(a\) could equal “Hello World” and \(b\) could equal “Hallo Wald.” The minimum edit distance from “Hello World” to “Hallo Wald” would be 3. The first step would be to change “Hello” to “Hallo.” Step two changes “World” to “Warld.” Finally, step three deletes the “r” and we are left with the string “Hallo Wald.” The advantage of the Levenshtein distance is
that it accounts for changes, deletion, insertion, and substitution. A Levenshtein distance of zero means that the two strings are an exact match; therefore, it took zero edits to change one string to the other. Likewise, if we obtain a minimum edit difference of four, which would imply that it took four different edits to change one string to the other. In this case, it was common to find that the only edit was removing the word “the” from an artist’s name. All of the Levenshtein distances are added to the dataset and any Levenshtein distance greater than five is manually reviewed. From this point, we are able to find the closest matching string for each song given the artist’s metrolyrics.com page. Once we have a list of song pages, it is easy to extract the lyrics given the script in Appendix A.7. As soon as the lyrics are added to our set, we are able to use the script in Appendix A.8. to conduct the search for profanity.

3.2.5 Singles

The Hot 100 Year-End number one singles charts (see Figure 3.8) are structured a little differently than the Hot 100 Year-End charts. Two things to keep in mind about Figure 3.8 are that each date row represents a week and that some songs span more than one week. If a song reappears on the chart, it no longer contains a link to the webpage. We can take advantage of this feature to calculate which songs had a reoccurrence in the same year after being knocked out of its position using the script in Appendix A.9. From this chart, we are also interested in the song link as well as the artist link so that we can merge the sets later.

---

3.3 Movies

Movie data searches are similar to song data searches in the steps we need to catalogue the data. We need information about the movie, and information about the words in the movie. A ranking is compiled from boxofficemojo.com whereas words are taken from movie subtitles available at subscene.org.
3.3.1 Movie Information

Each movie is indexed by boxofficemojo.com, a subsidiary of IMDb.com. The script available in Appendix A.10 extracts this data. We are interested in each row from Figure 3.9.

**Figure 3.9. Screenshot of Movie Data from Boxofficemojo.com**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Movie Title (click to view)</th>
<th>Studio</th>
<th>Total Gross / Theaters</th>
<th>Opening / Theaters</th>
<th>Open</th>
<th>Close</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Toy Story 3</td>
<td>BV</td>
<td>$415,004,880</td>
<td>4,028</td>
<td>6/18</td>
<td>12/2</td>
</tr>
<tr>
<td>3</td>
<td>Iron Man 2</td>
<td>Par.</td>
<td>$312,433,331</td>
<td>4,390</td>
<td>5/7</td>
<td>8/15</td>
</tr>
<tr>
<td>4</td>
<td>The Twilight Saga: Eclipse</td>
<td>Sum.</td>
<td>$300,531,751</td>
<td>4,468</td>
<td>6/30</td>
<td>10/23</td>
</tr>
</tbody>
</table>

The links to each movie take us to the respective webpage of each movie (Figure 3.10). These pages include the movie rating, genre, and length. (Appendix A.11.)
3.3.2 Subtitles

The key to downloading subtitles is to make sure that the information from the movie matches up with the search. For this, we need to know the title of the movie, but more important, the year of the movie. The search is modifiable from subscene.org’s URL, so the movie title and the year can be added to the end of the search query.
For example, there are many versions of *The Lord of the Rings: The Fellowship of the Rings*, so we must limit our search to only including titles with 2001. If this gives us an exact match, we capture the web address. However, if there still exists more than one title, we are able to deploy the Levenshtein distance calculator to choose the closest matching string. Appendix A.12 contains the script to measure Levenshtein distance and resulting extraction process. At this point, we download each of the subtitles based on three criteria: (1) the subtitles are rated “good,” (2) the subtitles are not “hearing impaired” subtitles – which means they include sound effects, and (3) there exists only one file – some subtitles are broken into two parts. The script to conduct this procedure is in Appendix A.13.

There are four main types of subtitles that we need to be concerned about. These are SubRip text files (.srt), MicroDVD subtitle files (.sub), Synchronized Accessible Media Interchange files (.smi), or SubStation Alpha (.ssa or .ass) files. Each of these files is unique in that they contain a different type of format to tell the media player when to display the subtitles on the screen. Below is a table of each format.

<table>
<thead>
<tr>
<th></th>
<th>SRT</th>
<th>SUB</th>
<th>SMI</th>
<th>SSA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>00:00:01 --&gt; 00:00:02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Hello World</em></td>
<td>{1} {144} <em>Hello World</em></td>
<td>HTML Formatting</td>
<td>No Example Available</td>
<td></td>
</tr>
<tr>
<td><strong>Subtitle 1</strong></td>
<td>{s1} {e1} <em>Subtitle 1</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Subtitle 2</strong></td>
<td>{s2} {e2} <em>Subtitle 2</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Start Time --> End Time*  
*Subtitles*
SRT files can be cleaned by finding each “-->” in the time string and deleting the line above and the current line containing the arrow.\textsuperscript{10} SUB files are cleaned by replacing any string containing brackets with blanks. SMI files do not need any cleaning if you open them in Microsoft Word, since Word automatically compiles the html script. SSA files are ignored which is why no example is given. After moving all the lyrics to a spreadsheet, the same search created for finding words in song lyrics can be applied to movie subtitles (See Appendices A.16 and A.17).

3.4 Summary

This process serves as a guideline for extracting data from different sources using automated scripts and element recognition in Visual Basic for Applications. The key advantage of using VBA scripts rather than using faster scripts such as Python and Perl is that VBA scripts require less scripting knowledge and can be easily manipulated and run using Microsoft Excel. Additionally, instead of using invasive harvesting techniques of entire webpages, this method takes far less bandwith since we are only extracting the information specify in advanced. This way, the script acts in a similar fashion to the way a human would manually copy this data. Additional error correction methods can be added and an increased number of databases can be used to create a more robust method that gathers data from every single song. The next chapter provides a detailed summary of the data gathered using these scripts as well as a summary of socioeconomic data used in the analysis.

\textsuperscript{10} By cleaning, we mean removing uncessary data such as the time queues and extra numbers so that only the subtitles will remain for our search.
Chapter IV

Data

4.1. Introduction

Interestingly, the dataset has the ability to tell a story about our preferences, cultures, and virtues throughout the last few decades. We can easily look at the data and answer questions such as: Which songs contain the most profanity? What movie genre was most popular in the 1990s? What is the average artist’s age in popular music? The data also capture well known societal trends from the rise of hip hop in the late 1980s to the large spike in disco songs in the late 1970s. From the data, we are able to see that, after the 1980s, disco’s popularity fell faster than housing prices in 2007. In this thesis an extensive data search of the internet has been conducted to analyze the relationship between music and profanity. These data are structured and collected as an economic input that can be used to make meaningful comparisons between society and producers of music. The data discussed in this chapter and used in the analysis in the next chapter were gathered using internet search scripts written in Visual Basic for Applications (VBA). These scripts are contained in Appendix A and are explained in Chapter 3. The data described below have been grouped into three different sets: music data, movie data, and socioeconomic data.

4.2 Music Data

Music can be characterized as a non-rivalous good, since ones person’s consumption of the good does not affect another person’s ability to consume the same
good and as an almost non-exclusive good as well. I use the qualifier *almost* because music can be freely listened to on the radio, streamed over the internet, or (in some cases) obtained through downloading. If we are interested in the lyrics, then there are even more freely accessible sites to find this information. I exploit this nature of music to create a dataset from the rankings provided by Billboard Magazine’s *Hot 100* list on the top 100 songs of the year, artist data, and the number one hits per week.

4.2.1 Song Data

The Billboard *Hot 100* list is a ranking of the 100 most popular songs per year. The list itself gives us information on the year, rank, title, and artist of the song. In order to dig deeper into the question at hand, we need two key categories: information about the song and the song lyrics. In other words, we need facts about the musical side and the lyrical side, which happen to be slightly intertwined. After all, if we were to take hardcore rap lyrics and place them over a traditional religious folk beat, it might be difficult to classify in which category it belongs.

For music data, I focus on the traditional metrics as well as some simple theory. **Year** represents the year that the song first appeared on the *Hot 100* list while **Rank** is what it was ranked in that given year. **Length** is the length of the song in minutes and, given the choice, is the album length rather than the radio edited version because the lyrics are typically uncensored. **Major** is a dummy variable that is equal to 1 if the song is in a major key and 0 if the song is in a minor key. The original published key of the music, which could potentially affect the length of a song’s stay in the number one spot,
is omitted and an explanation of which is available in Appendix B Table B.10. Many songs span multiple genres, which makes it difficult to create dummy variables to control for genre. For instance, Taylor Swift’s song “Eye’s Open” from “The Hunger Games” soundtrack is classified as both alternative rock and country rock. Therefore, under the category of genre, it would have values of 1 for both Rock and Country. These variables are included in the regression because there is a likely correlation between genre and survivability, and genre and profanity preferences. The other genre variables chosen were Pop for popular, RB for rhythm and blues, HipHop for hip hop, Funk for funk, Disco for disco, Folk for folk, Blues for blues, and Other for any other genre that did not have a large enough sample size.

Profanity extracted from the lyrics of the songs is uniform for each dataset. The words chosen are a select sample to represent words that are commonly censored by the FCC, those not suitable for young children, and other epithets that represent racial or sexual oriented slurs. The bundle of words chosen is as follows and the uncensored key is provided in Table 4.1: as, btch, cnt, cck, dk, fg, fck,ngg, pss, sht, and tts.

---

11 http://www.fcc.gov/guides/obscenity-indecency-and-profanity
Table 4.1: Profanity Legend

<table>
<thead>
<tr>
<th>Censored</th>
<th>Uncensored</th>
</tr>
</thead>
<tbody>
<tr>
<td>as</td>
<td>Ass</td>
</tr>
<tr>
<td>bitch</td>
<td>Bitch</td>
</tr>
<tr>
<td>cnt</td>
<td>Cunt</td>
</tr>
<tr>
<td>cck</td>
<td>Cock/Cocksucker</td>
</tr>
<tr>
<td>dk</td>
<td>Dyke</td>
</tr>
<tr>
<td>fg</td>
<td>Fag/Faggot</td>
</tr>
<tr>
<td>fck</td>
<td>Fuck/Motherfucker(a)</td>
</tr>
<tr>
<td>ngg</td>
<td>Nigger(a)</td>
</tr>
<tr>
<td>pss</td>
<td>Pussy</td>
</tr>
<tr>
<td>sht</td>
<td>Shit</td>
</tr>
<tr>
<td>slt</td>
<td>Slut</td>
</tr>
<tr>
<td>tts</td>
<td>Tit</td>
</tr>
</tbody>
</table>

Each number represents a count of the curse words per song. I must make the caveat that these numbers may not represent the true count of the frequency of the word. Some lyrics tend to have a command similar to “x2” or “repeat” which means that the line should be repeated. Therefore, we also create a dummy variable that represents if the song contains at least one occurrence of the corresponding curse word which is used later in the analysis. The results are provided in Table 4.2.

The results show that the average length of any song is about 4.11 minutes from 1970-2012. About three quarters of the songs are in a major scale and the genres seem to follow intuition: pop should be prevalent in the popular music chart. The profanity seems to suggest a distribution with large outliers for many variables. For example, while sht only occurs around .44 times per song, it can range from 0 sht’s per song to 76 sht’s per song. The next most prevalent word is ngg which caps out at 62 ngg’s per song followed by fck which happens around .26 times per song. The lower maximum for fcks suggests that it occurs more often in all songs than as because as has a higher maximum.
Table 4.2: Song Data Summary Table

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Info</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>1991</td>
<td>1970</td>
<td>2012</td>
<td>4300</td>
</tr>
<tr>
<td>Rank</td>
<td>51</td>
<td>1</td>
<td>100</td>
<td>4300</td>
</tr>
<tr>
<td>Length</td>
<td>4.11</td>
<td>1.28</td>
<td>12.30</td>
<td>3497</td>
</tr>
<tr>
<td>Major</td>
<td>0.76</td>
<td>0.00</td>
<td>1.00</td>
<td>3083</td>
</tr>
<tr>
<td><strong>Genre</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop</td>
<td>0.34</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>Rock</td>
<td>0.29</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>OtherMiss</td>
<td>0.23</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>RB</td>
<td>0.20</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>HipHop</td>
<td>0.12</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>Country</td>
<td>0.06</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>Funk</td>
<td>0.04</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>Disco</td>
<td>0.04</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>Folk</td>
<td>0.02</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td>Blues</td>
<td>0.01</td>
<td>0.00</td>
<td>1.00</td>
<td>4300</td>
</tr>
<tr>
<td><strong>Profanity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sht</td>
<td>0.44</td>
<td>0.00</td>
<td>76.00</td>
<td>2688</td>
</tr>
<tr>
<td>ngg</td>
<td>0.36</td>
<td>0.00</td>
<td>62.00</td>
<td>2688</td>
</tr>
<tr>
<td>fck</td>
<td>0.26</td>
<td>0.00</td>
<td>21.00</td>
<td>2688</td>
</tr>
<tr>
<td>btch</td>
<td>0.18</td>
<td>0.00</td>
<td>39.00</td>
<td>2688</td>
</tr>
<tr>
<td>as</td>
<td>0.16</td>
<td>0.00</td>
<td>44.00</td>
<td>2688</td>
</tr>
<tr>
<td>pss</td>
<td>0.03</td>
<td>0.00</td>
<td>20.00</td>
<td>2688</td>
</tr>
<tr>
<td>tts</td>
<td>0.01</td>
<td>0.00</td>
<td>13.00</td>
<td>2688</td>
</tr>
<tr>
<td>cck</td>
<td>0.01</td>
<td>0.00</td>
<td>8.00</td>
<td>2688</td>
</tr>
<tr>
<td>fg</td>
<td>0.00</td>
<td>0.00</td>
<td>3.00</td>
<td>2688</td>
</tr>
<tr>
<td>slit</td>
<td>0.00</td>
<td>0.00</td>
<td>4.00</td>
<td>2688</td>
</tr>
<tr>
<td>cnt</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>2688</td>
</tr>
<tr>
<td>dk</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>2688</td>
</tr>
</tbody>
</table>

Given that the nature of music in the United States tends to go through different trends per era, a more meaningful way of looking at the data is to break it apart by decades: 1970s, 1980s, 1990s, and 2000s. The years 2010, 2011, and 2012 will also be included in the 2000s bracket. The results are provided in Table 4.3.
Table 4.3: Song Summary per Decade

<table>
<thead>
<tr>
<th></th>
<th>1970s</th>
<th>1980s</th>
<th>1990s</th>
<th>2000s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>3.81</td>
<td>4.24</td>
<td>4.43</td>
<td>3.99</td>
</tr>
<tr>
<td>Major</td>
<td>0.82</td>
<td>0.81</td>
<td>0.79</td>
<td>0.65</td>
</tr>
<tr>
<td>as</td>
<td>0.01</td>
<td>0</td>
<td>0.12</td>
<td>0.33</td>
</tr>
<tr>
<td>bitch</td>
<td>0.04</td>
<td>0</td>
<td>0.10</td>
<td>0.38</td>
</tr>
<tr>
<td>cnt</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cck</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td>dk</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fg</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>fck</td>
<td>0.03</td>
<td>0</td>
<td>0.24</td>
<td>0.52</td>
</tr>
<tr>
<td>ngg</td>
<td>0.06</td>
<td>0</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td>pss</td>
<td>0.00</td>
<td>0</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>sht</td>
<td>0.05</td>
<td>0</td>
<td>0.53</td>
<td>0.81</td>
</tr>
<tr>
<td>slt</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>tts</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Rock</td>
<td>0.33</td>
<td>0.44</td>
<td>0.21</td>
<td>0.20</td>
</tr>
<tr>
<td>Pop</td>
<td>0.23</td>
<td>0.46</td>
<td>0.30</td>
<td>0.37</td>
</tr>
<tr>
<td>Country</td>
<td>0.08</td>
<td>0.04</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>RB</td>
<td>0.08</td>
<td>0.14</td>
<td>0.27</td>
<td>0.28</td>
</tr>
<tr>
<td>HipHop</td>
<td>0.00</td>
<td>0.01</td>
<td>0.15</td>
<td>0.26</td>
</tr>
<tr>
<td>Folk</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Blues</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>Disco</td>
<td>0.10</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Funk</td>
<td>0.07</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>OtherMiss</td>
<td>0.34</td>
<td>0.18</td>
<td>0.27</td>
<td>0.17</td>
</tr>
</tbody>
</table>

From this new partitioned view of the data, we are able to see trends starting to emerge with changes over time. First, lengths of popular songs were slightly longer in the 1980s and 1990s than they were in either the 1970s or 2000s. As found by Schellenberg and Scheve (2012) in their research of the top 40 songs, minor keys have
become much more prevalent in the 2000s than they were in any of the previous decades.

Figure 4.1 shows this trend for the full dataset as well as the results when we restrict the data to the top 40 hits.

**Figure 4.1: Major Scales per Year**

![Major Scale Percentage per Year](image)

We also can see the changing prevalence of average profanity occurrence per song for each decade. The 1970s had on average less than 0.10 occurrences of each profane word per song. Contrasted with the 1970s, the 2000s contained many more curse words per song. Figures 4.2 through 4.4 show the trends in profanity as well as changing genres over time.
Figure 4.2: Profanity in Songs per Year

Figure 4.3: Average Profanity per Minute per Year
4.2.2 Artist Data

Artists can be popular for a variety of reasons. They might be extremely talented individuals and are endowed with great singing voices, they can be talented songwriters that know how to evoke emotions, they can be popular because people say that they are popular, and they can be revered for their aesthetics. Understanding these characteristics is beyond the scope of this thesis. Nonetheless, several artist factors are considered such as whether the artist performing the song is a **Group** or single individual, then whether or not that single individual is a **Male** or **Female**, as well as his or her date of birth. Given the year that a song debuted, we can easily deduce the artist’s **Age** and the **YearsActive**, which represents how long they have been performing. This data directly feeds into the
third set and the results are available in Table 4.4 and Table 4.5. These variables are paramount to the accelerated time failure model presented by Giles (2007), which we will be replicating in Chapter 5 (Empirical Analysis).

4.2.3 Singles Data

The singles data set is derived from the Billboard Hot 100 Year-End number one singles chart. This chart differs from the previous list in that it only contains songs that stayed in the number one position and how many weeks it remained in that spot. It provides data on the weeks the song spent in the number one position, which is labeled WeeksNo1 for the non-consecutive weeks the song spends in the number one position. This variable will act as our dependent variable for the survival analysis, which will be used to determine the effect of profanity on song longevity. A second variable NonCon is derived from this chart which is a dummy variable equal to one if the song was overtaken by another song, but retook its top-seat in any following week, and zero otherwise. Because songs that “bounce-back” to the number one position have some extra boost in popularity, it is important to control for this potential anomaly. Finally, the new variable Collaboration is added to test whether a song “featuring” another popular artist will promote profanity or if it will boost weeks spent in the number one spot ceterus paribus. The other variables are merged from existing song and artist data sets such as profanity, genre, length, major, and other music and artist qualities. These are important for the survival analysis to control for factors that might also be related to profanity. The summary of the data are found in Table 4.4 and the decade breakdown in Table 4.5.
## Table 4.4: Singles Data Summary

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Song</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>1991</td>
<td>1970</td>
<td>2012</td>
<td>2230</td>
</tr>
<tr>
<td>Week</td>
<td>26.5</td>
<td>1</td>
<td>53</td>
<td>2230</td>
</tr>
<tr>
<td>WeeksNo1</td>
<td>4.98</td>
<td>1</td>
<td>16</td>
<td>2230</td>
</tr>
<tr>
<td>NonCon</td>
<td>0.04</td>
<td>0</td>
<td>1</td>
<td>2230</td>
</tr>
<tr>
<td>Collaboration</td>
<td>0.19</td>
<td>0</td>
<td>1</td>
<td>2230</td>
</tr>
<tr>
<td>Length</td>
<td>4.16</td>
<td>1.93</td>
<td>12.02</td>
<td>2141</td>
</tr>
<tr>
<td>Major</td>
<td>0.70</td>
<td>0</td>
<td>1</td>
<td>1882</td>
</tr>
<tr>
<td><strong>Profanity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>as</td>
<td>0.09</td>
<td>0</td>
<td>7</td>
<td>1640</td>
</tr>
<tr>
<td>btch</td>
<td>0.02</td>
<td>0</td>
<td>6</td>
<td>1640</td>
</tr>
<tr>
<td>cmt</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
<td>1640</td>
</tr>
<tr>
<td>cck</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1640</td>
</tr>
<tr>
<td>dk</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1640</td>
</tr>
<tr>
<td>fg</td>
<td>0.01</td>
<td>0</td>
<td>3</td>
<td>1640</td>
</tr>
<tr>
<td>fck</td>
<td>0.11</td>
<td>0</td>
<td>6</td>
<td>1640</td>
</tr>
<tr>
<td>ngg</td>
<td>0.23</td>
<td>0</td>
<td>11</td>
<td>1640</td>
</tr>
<tr>
<td>pss</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
<td>1640</td>
</tr>
<tr>
<td>sht</td>
<td>0.40</td>
<td>0</td>
<td>76</td>
<td>1640</td>
</tr>
<tr>
<td>slt</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1640</td>
</tr>
<tr>
<td>tts</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
<td>1640</td>
</tr>
<tr>
<td><strong>Genre</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rock</td>
<td>0.28</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>Pop</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>Country</td>
<td>0.03</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>RB</td>
<td>0.30</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>HipHop</td>
<td>0.14</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>Folk</td>
<td>0.02</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>Blues</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>Disco</td>
<td>0.09</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>Funk</td>
<td>0.07</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td>OtherMiss</td>
<td>0.07</td>
<td>0</td>
<td>1</td>
<td>2209</td>
</tr>
<tr>
<td><strong>Artist</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.34</td>
<td>0</td>
<td>1</td>
<td>2104</td>
</tr>
<tr>
<td>Female</td>
<td>0.33</td>
<td>0</td>
<td>1</td>
<td>2103</td>
</tr>
<tr>
<td>Group</td>
<td>0.32</td>
<td>0</td>
<td>1</td>
<td>2082</td>
</tr>
<tr>
<td>Started</td>
<td>1980.74</td>
<td>1948</td>
<td>2008</td>
<td>2089</td>
</tr>
<tr>
<td>YearsActive</td>
<td>10.30</td>
<td>0</td>
<td>36</td>
<td>2062</td>
</tr>
<tr>
<td>Age</td>
<td>28.92</td>
<td>14</td>
<td>53</td>
<td>1413</td>
</tr>
</tbody>
</table>
Table 4.5: Singles Summary per Decade

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean 1970s</th>
<th>Mean 1980s</th>
<th>Mean 1990s</th>
<th>Mean 2000s</th>
</tr>
</thead>
<tbody>
<tr>
<td>WeeksNo1</td>
<td>3.13</td>
<td>3.41</td>
<td>6.62</td>
<td>6.38</td>
</tr>
<tr>
<td>NonCon</td>
<td>0.04</td>
<td>0.01</td>
<td>0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>Collaboration</td>
<td>0.05</td>
<td>0.10</td>
<td>0.16</td>
<td>0.40</td>
</tr>
<tr>
<td>Length</td>
<td>3.99</td>
<td>4.24</td>
<td>4.40</td>
<td>4.02</td>
</tr>
<tr>
<td>Major</td>
<td>0.78</td>
<td>0.71</td>
<td>0.74</td>
<td>0.61</td>
</tr>
<tr>
<td>as</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>0.21</td>
</tr>
<tr>
<td>btch</td>
<td>0</td>
<td>0.00</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>cnt</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>cck</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dk</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fg</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td>fck</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.29</td>
</tr>
<tr>
<td>ngg</td>
<td>0</td>
<td>0</td>
<td>0.24</td>
<td>0.47</td>
</tr>
<tr>
<td>pss</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>sht</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
<td>1.05</td>
</tr>
<tr>
<td>slt</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tts</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>Rock</td>
<td>0.36</td>
<td>0.53</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td>Pop</td>
<td>0.35</td>
<td>0.57</td>
<td>0.48</td>
<td>0.47</td>
</tr>
<tr>
<td>Country</td>
<td>0.08</td>
<td>0.01</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>RB</td>
<td>0.07</td>
<td>0.18</td>
<td>0.47</td>
<td>0.42</td>
</tr>
<tr>
<td>HipHop</td>
<td>0</td>
<td>0.01</td>
<td>0.16</td>
<td>0.34</td>
</tr>
<tr>
<td>Folk</td>
<td>0.09</td>
<td>0</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Blues</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Disco</td>
<td>0.25</td>
<td>0.06</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Funk</td>
<td>0.13</td>
<td>0.09</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>OtherMiss</td>
<td>0.12</td>
<td>0.05</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>Group</td>
<td>0.44</td>
<td>0.40</td>
<td>0.32</td>
<td>0.17</td>
</tr>
<tr>
<td>Male</td>
<td>0.38</td>
<td>0.32</td>
<td>0.24</td>
<td>0.41</td>
</tr>
<tr>
<td>Female</td>
<td>0.18</td>
<td>0.27</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>Age</td>
<td>29.88</td>
<td>31.98</td>
<td>28.72</td>
<td>26.78</td>
</tr>
<tr>
<td>YearsActive</td>
<td>9.13</td>
<td>12.43</td>
<td>10.13</td>
<td>9.59</td>
</tr>
</tbody>
</table>
Overall, the average time spent in the number one spot is close to 5 weeks. If we look at the decade breakdown, we can see that songs in the 1970s and 1980s tended to hold their spots around half the weeks than they did in the 1990s and 2000s. Groups tend to dominate the first three decades and sharply fell in the 2000s, which could be a result of collaboration between artists increasing almost three-fold from the 1990s. The average age of the artist at the time of their hit seems to be decreasing slightly with each decade and the experience (YearsActive) seems to be more prevalent in the 1980s along with the age of the artist. Although we see country have a resurgence in the song dataset after 2000, in the singles set the genre went from a highly of 8% of the singles to none in 2000. Once again, we see that there is a change in the key of songs from predominantly major to an increasing minor minority.

Although each set contains missing data points, we still see that we have around 62% of the Hot 100 data with most of the missing points occurring in the earlier decades and 71% of the data for number one chart position. Excluding profanity gives us a much larger dataset; however, the aim of this thesis is to analyze changes in producer behavior. This leaves us with 672 songs per decade on average if we were to break the analysis into four different time periods or 1620 songs pre-Parental Advisory and 1068 songs post-Parental Advisory. This will be sufficient to conduct least squared regressions and survival model analysis on both song and single datasets. The next task is to discuss the collected movie data, which acts as the control in the differences–in-differences estimation discussed in the next chapter.
4.3 Movie Data

Lyrics are to songs as subtitles are to movies. There are vast databases on the internet that contain movie subtitles. From these subtitles, we are able to gather data we need to analyze with even greater accuracy than song lyrics due to that aforementioned repeating factor. Song lyrics may contain the “repeat” commands; however, movie subtitles are much more thorough. Additional factors need to be considered when characterizing profanity in movies: (1) movies range in length and production cost, (2) they are tailored to different audiences, and (3) they have the capability to contain many more curse words than songs. When constructing the dataset, we must be mindful of these factors so that we can create a standard metric, such as profanity per minute, which likens the two sets. Luckily, we are able to compile these elements from websites with the same process for song data (see Appendix A for scripts for extracting movie data).

The ranking of movies differ from the ranking of songs in that movies are ranked based off of their total gross revenue received or TotalGross. After identifying the title of the movie and the Year the movie came out (1980-2012), we are able to get data on the Length of the movie as well as the genres. Genres for movies differ from song genres in a few ways. First, movie genres are limited to one classification. For instance, movies that may contain elements from romance, comedy, action, and science fiction may only be classified as a Romantic Comedy. Second, the lines between genres can be undefined. The difference between action, thriller, war, and adventure may be minimal. Third, there is a plethora of genres to choose from. A simple tabulation would result in over 100 different classifications. We account for these issues by creating dummy
variables for genre categories like that in songs. Movie genres are spliced into the two
different elements and then categorized based off the following key genres: Comedy,
Drama, Action, Thriller, Horror, Romance, Adventure, Family, Scifi, Animation, War, Western, Musical, Documentary, and Other. Each movie is limited to a
maximum of two genres given this breakdown. Finally, the last variable of interest for
movies is the rating of the movie itself. Movie ratings are determined by the Motion
Picture Association of America based on the sexual, violent, graphic, and profane nature
of a movie and therefore should be highly correlated with profanity and perhaps genre. G
is the lowest rating that is suitable for all audiences. PG requires some parental
guidance. PG13 suggests that the material may not be suitable for children under the age
of 13. R is restricted to mature audiences. NC17 means that no persons under the age of
17 may be admitted. An Unrated rating encompasses all other movies, since there are
observations that contain no rating. These are omitted in the final regressions.

Profanity in the film industry should be treated differently than profanity in the
music industry given the length of most movies. In the analysis, we will control for the
difference in movie length by dividing each profanity variable by the length of the movie
in minutes. This will allow us to establish a baseline profanity average or “acceptability”
rate. The same curse words are collected for movies as described above for music lyrics.
These include as, b'tch, cnt, cek, dk, fg, fck, ngg, pss, sht, slt, and tts. The count of
each of these variables is more precise than for music given that subtitles are not

\[^{12}\text{http://www.mpaa.org/ratings/what-each-rating-means}\]
censored. The summary of the movie dataset are available in Table 4.6 and the decade approach is shown in Table 4.7.
### Table 4.6: Movie Data Summary Table

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>$n$</th>
<th>Variable</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Info</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>1996</td>
<td>1980</td>
<td>2012</td>
<td>3300</td>
<td>TotalGross</td>
<td>5.73E+07</td>
<td>1169875</td>
<td>7.50E+08</td>
<td>3300</td>
</tr>
<tr>
<td>Rank</td>
<td>50.5</td>
<td>1</td>
<td>100</td>
<td>3300</td>
<td>Length</td>
<td>107.991</td>
<td>38</td>
<td>227</td>
<td>3299</td>
</tr>
<tr>
<td><strong>Genre</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comedy</td>
<td>0.34</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>Family</td>
<td>0.06</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>Drama</td>
<td>0.17</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>Sci-fi</td>
<td>0.06</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>Action</td>
<td>0.15</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>Animation</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>Thriller</td>
<td>0.11</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>War</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>Other</td>
<td>0.10</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>Western</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>Horror</td>
<td>0.09</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>Musical</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>Romance</td>
<td>0.09</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>Documentary</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>Adventure</td>
<td>0.07</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Profanity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fck</td>
<td>13.03</td>
<td>0</td>
<td>425</td>
<td>2186</td>
<td>cck</td>
<td>0.23</td>
<td>0</td>
<td>21</td>
<td>2186</td>
</tr>
<tr>
<td>sht</td>
<td>9.34</td>
<td>0</td>
<td>132</td>
<td>2186</td>
<td>tts</td>
<td>0.16</td>
<td>0</td>
<td>10</td>
<td>2186</td>
</tr>
<tr>
<td>as</td>
<td>4.07</td>
<td>0</td>
<td>83</td>
<td>2186</td>
<td>fg</td>
<td>0.14</td>
<td>0</td>
<td>11</td>
<td>2186</td>
</tr>
<tr>
<td>btch</td>
<td>2.41</td>
<td>0</td>
<td>42</td>
<td>2186</td>
<td>slit</td>
<td>0.11</td>
<td>0</td>
<td>11</td>
<td>2186</td>
</tr>
<tr>
<td>nrgg</td>
<td>0.65</td>
<td>0</td>
<td>114</td>
<td>2186</td>
<td>cnt</td>
<td>0.06</td>
<td>0</td>
<td>21</td>
<td>2186</td>
</tr>
<tr>
<td>pss</td>
<td>0.41</td>
<td>0</td>
<td>36</td>
<td>2186</td>
<td>dk</td>
<td>0.02</td>
<td>0</td>
<td>3</td>
<td>2186</td>
</tr>
<tr>
<td><strong>Rating</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>R</td>
<td>0.40</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>PG</td>
<td>0.24</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>NC17</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
<tr>
<td>PG13</td>
<td>0.31</td>
<td>0</td>
<td>1</td>
<td>3300</td>
<td>Unrated</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
<td>3300</td>
</tr>
</tbody>
</table>
Table 4.7: Movie Data Summary per Decade

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean 1980s</th>
<th>Max 1980s</th>
<th>Mean 1990s</th>
<th>Max 1990s</th>
<th>Mean 2000s</th>
<th>Max 2000s</th>
</tr>
</thead>
<tbody>
<tr>
<td>TotalGross</td>
<td>2.94E+07</td>
<td>3.59E+08</td>
<td>5.01E+07</td>
<td>6.01E+08</td>
<td>8.43E+07</td>
<td>7.50E+08</td>
</tr>
<tr>
<td>Length</td>
<td>105.62</td>
<td>219</td>
<td>109.69</td>
<td>227</td>
<td>108.50</td>
<td>200</td>
</tr>
<tr>
<td>Comedy</td>
<td>0.31</td>
<td>1</td>
<td>0.39</td>
<td>1</td>
<td>0.34</td>
<td>1</td>
</tr>
<tr>
<td>Action</td>
<td>0.11</td>
<td>1</td>
<td>0.17</td>
<td>1</td>
<td>0.17</td>
<td>1</td>
</tr>
<tr>
<td>Scifi</td>
<td>0.07</td>
<td>1</td>
<td>0.06</td>
<td>1</td>
<td>0.06</td>
<td>1</td>
</tr>
<tr>
<td>Romance</td>
<td>0.07</td>
<td>1</td>
<td>0.11</td>
<td>1</td>
<td>0.08</td>
<td>1</td>
</tr>
<tr>
<td>Drama</td>
<td>0.13</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.17</td>
<td>1</td>
</tr>
<tr>
<td>Horror</td>
<td>0.09</td>
<td>1</td>
<td>0.08</td>
<td>1</td>
<td>0.10</td>
<td>1</td>
</tr>
<tr>
<td>Thriller</td>
<td>0.07</td>
<td>1</td>
<td>0.14</td>
<td>1</td>
<td>0.11</td>
<td>1</td>
</tr>
<tr>
<td>Animation</td>
<td>0.04</td>
<td>1</td>
<td>0.04</td>
<td>1</td>
<td>0.08</td>
<td>1</td>
</tr>
<tr>
<td>Western</td>
<td>0.01</td>
<td>1</td>
<td>0.01</td>
<td>1</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>Musical</td>
<td>0.01</td>
<td>1</td>
<td>0.00</td>
<td>1</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>Adventure</td>
<td>0.06</td>
<td>1</td>
<td>0.06</td>
<td>1</td>
<td>0.09</td>
<td>1</td>
</tr>
<tr>
<td>Family</td>
<td>0.03</td>
<td>1</td>
<td>0.09</td>
<td>1</td>
<td>0.07</td>
<td>1</td>
</tr>
<tr>
<td>Documentary</td>
<td>0.00</td>
<td>1</td>
<td>0.00</td>
<td>1</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>War</td>
<td>0.01</td>
<td>1</td>
<td>0.01</td>
<td>1</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>0.25</td>
<td>1</td>
<td>0.02</td>
<td>1</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td>as</td>
<td>3.53</td>
<td>47</td>
<td>4.32</td>
<td>61</td>
<td>4.20</td>
<td>83</td>
</tr>
<tr>
<td>btch</td>
<td>2.16</td>
<td>23</td>
<td>2.79</td>
<td>37</td>
<td>2.28</td>
<td>42</td>
</tr>
<tr>
<td>cnt</td>
<td>0.04</td>
<td>4</td>
<td>0.09</td>
<td>21</td>
<td>0.05</td>
<td>5</td>
</tr>
<tr>
<td>cck</td>
<td>0.15</td>
<td>9</td>
<td>0.28</td>
<td>21</td>
<td>0.23</td>
<td>10</td>
</tr>
<tr>
<td>dk</td>
<td>0.03</td>
<td>3</td>
<td>0.03</td>
<td>3</td>
<td>0.02</td>
<td>3</td>
</tr>
<tr>
<td>fg</td>
<td>0.15</td>
<td>8</td>
<td>0.17</td>
<td>10</td>
<td>0.12</td>
<td>11</td>
</tr>
<tr>
<td>fck</td>
<td>7.56</td>
<td>211</td>
<td>16.30</td>
<td>425</td>
<td>13.80</td>
<td>379</td>
</tr>
<tr>
<td>ngg</td>
<td>0.26</td>
<td>15</td>
<td>0.89</td>
<td>85</td>
<td>0.70</td>
<td>114</td>
</tr>
<tr>
<td>pss</td>
<td>0.34</td>
<td>28</td>
<td>0.41</td>
<td>36</td>
<td>0.44</td>
<td>28</td>
</tr>
<tr>
<td>sht</td>
<td>7.95</td>
<td>102</td>
<td>10.43</td>
<td>132</td>
<td>9.37</td>
<td>129</td>
</tr>
<tr>
<td>slt</td>
<td>0.11</td>
<td>11</td>
<td>0.10</td>
<td>9</td>
<td>0.13</td>
<td>10</td>
</tr>
<tr>
<td>tts</td>
<td>0.17</td>
<td>4</td>
<td>0.14</td>
<td>6</td>
<td>0.16</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>0.05</td>
<td>1</td>
<td>0.05</td>
<td>1</td>
<td>0.05</td>
<td>1</td>
</tr>
<tr>
<td>NC17</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>PG</td>
<td>0.36</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.19</td>
<td>1</td>
</tr>
<tr>
<td>PG13</td>
<td>0.12</td>
<td>1</td>
<td>0.29</td>
<td>1</td>
<td>0.47</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>0.47</td>
<td>1</td>
<td>0.46</td>
<td>1</td>
<td>0.30</td>
<td>1</td>
</tr>
<tr>
<td>Unrated</td>
<td>0.00</td>
<td>1</td>
<td>0.00</td>
<td>1</td>
<td>0.00</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.6 shows that the average non-adjusted gross for movies is around $57 million. The average length of a movie from 1980 to 2012 is approximately 108 minutes (1.8 hours) with the shortest movie being 38 minutes and the longest being 227 minutes.
Comedies seem to be the most saturated genre with 34% of movies, followed by dramas with 17%, action movies with 15%, and thrillers with 11%. In regards to ratings, “R” movies have been the most prevalent type of movies, followed by “PG-13” and “PG.” “G” movies surprisingly made up only 5% of the entire top grossing movies during this time span. Turning to profanity, *fck* has the highest average count with 13 *fcks* per movie, which is heavily skewed by the movies that contain over 300 occurrences. *Sht* comes in second with 9 *shts* per movie and is not as heavily weighted by extreme values. Although *as* follows *sht* for the next position, it is important to note that while *ngg* is in fifth for the average count, it is heavily skewed compared to the other lesser-used words.

Table 4.7 shows the changing dynamic of movie profiles in each decade. Somewhat surprising, there has been little change over time. Movie length averages and maximums hardly fluctuate across decades. The genres categories hardly change either except for a small rise in adventure and animation movies, which could be a result of increasing computer animation technology. On the other hand, the rating averages seem to suggest that there has been a slow convergence towards PG-13 movies. R-rated movies fell from 47% in the 1980s to only 30% in the 2000s while PG rated movies fell from 36% to 19% and PG-13 movies grew from 12% to 47%. Another key difference about this set deals with the profanity profiles. Movies contained more profanity on average in the 1990s than they did in the 2000s. Both the 1990s and 2000s were also much more profane than the 1980s movies. This is a noticeable difference from the profanity profiles of songs during these periods. The final task is to compile national
yearly averages for societal measures to gain a good understanding of the crime, poverty, and other factors that could be contributing to the changing culture.

4.4 Socioeconomic Data

Socioeconomic data were collected to analyze the effect of profanity on juvenile violence and teenage pregnancy, two issues facing the young population. Pop is the entire United States population estimates from 1970 to 2012, which ranges from 205 million to 313 million people. White is the count of non-Hispanic white persons living in the United States in millions and spans from 1970 to 2011. Teen is the population between the ages of fifteen and nineteen. The variable Bachelors is the number of people that hold a bachelor’s degree. rGDP is real US GDP respectively. ChildPov is the poverty rate of children, AdultPov is the poverty rates of adults, and SeniorPov is the poverty rate of senior citizens. Emp is the unemployment rate. SingleMom is the proportion of single parent families in the United States. JuvCrime is the rate of crime for juveniles per 1000, JuvDrug is the rate of drug abuse for juveniles per 1000, JuvAssalt is the rate of aggravated assault for juveniles per 1000, and JuvRape is the rate of rape for juveniles per 1000. All juvenile crime data only span from 1980-2010. The next set of crime data is per 100,000 people and includes ViolentCrime, Murder, Rape, Robbery, and Assault. Finally, TeenPreg is the rate of 15 to 19 year old childbirths per 1000. The sources for the data are available in Table 4.7 and the summaries are available in Tables 4.9 and 4.10
<table>
<thead>
<tr>
<th>Variable</th>
<th>Source</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop</td>
<td>World Bank</td>
<td>Millions</td>
</tr>
<tr>
<td>rGDP</td>
<td>World Bank</td>
<td>Billions</td>
</tr>
<tr>
<td>White</td>
<td>Bureau of Labor Statistics</td>
<td>Percentage</td>
</tr>
<tr>
<td>Bachelors</td>
<td>Census</td>
<td>Millions</td>
</tr>
<tr>
<td>Teen</td>
<td>Census</td>
<td>Millions</td>
</tr>
<tr>
<td>Emp</td>
<td>Bureau of Labor Statistics</td>
<td>Percentage</td>
</tr>
<tr>
<td>Single</td>
<td>Census</td>
<td>Percentage</td>
</tr>
<tr>
<td>ChildPov</td>
<td>Census</td>
<td>Percentage</td>
</tr>
<tr>
<td>AdultPov</td>
<td>Census</td>
<td>Percentage</td>
</tr>
<tr>
<td>SeniorPov</td>
<td>Census</td>
<td>Percentage</td>
</tr>
<tr>
<td>TeenPreg</td>
<td>Centers for Disease Control</td>
<td>Rate per 1000</td>
</tr>
<tr>
<td></td>
<td>Federal Bureau of Investigations</td>
<td></td>
</tr>
<tr>
<td>ViolentCrime</td>
<td>Investigations</td>
<td>Rate per 1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rate per 100,000 ages</td>
</tr>
<tr>
<td>JuvCrime</td>
<td>Office of Justice Programs</td>
<td>10-17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rate per 100,000 ages</td>
</tr>
<tr>
<td>JuvDrug</td>
<td>Office of Justice Programs</td>
<td>10-17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rate per 100,000 ages</td>
</tr>
<tr>
<td>JuvAssault</td>
<td>Office of Justice Programs</td>
<td>10-17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rate per 100,000 ages</td>
</tr>
<tr>
<td>JuvRape</td>
<td>Office of Justice Programs</td>
<td>10-17</td>
</tr>
</tbody>
</table>
Table 4.9: Societal Data Summary

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Max</th>
<th>Min</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>1991</td>
<td>12.56</td>
<td>1970</td>
<td>2012</td>
<td>43</td>
</tr>
<tr>
<td>Pop</td>
<td>255.47</td>
<td>32.78</td>
<td>205.05</td>
<td>311.59</td>
<td>42</td>
</tr>
<tr>
<td>rGDP</td>
<td>6.68</td>
<td>4.43</td>
<td>1.02</td>
<td>14.99</td>
<td>42</td>
</tr>
<tr>
<td>White</td>
<td>0.85</td>
<td>0.03</td>
<td>0.79</td>
<td>0.89</td>
<td>43</td>
</tr>
<tr>
<td>Bachelors</td>
<td>35.22</td>
<td>15.31</td>
<td>12.06</td>
<td>63.29</td>
<td>43</td>
</tr>
<tr>
<td>Teen</td>
<td>20.13</td>
<td>1.40</td>
<td>17.18</td>
<td>22.21</td>
<td>41</td>
</tr>
<tr>
<td>Emp</td>
<td>6.39</td>
<td>1.55</td>
<td>3.97</td>
<td>9.71</td>
<td>43</td>
</tr>
<tr>
<td>Single</td>
<td>0.24</td>
<td>0.07</td>
<td>0.11</td>
<td>0.32</td>
<td>43</td>
</tr>
<tr>
<td>ChildPov</td>
<td>0.19</td>
<td>0.02</td>
<td>0.14</td>
<td>0.23</td>
<td>43</td>
</tr>
<tr>
<td>AdultPov</td>
<td>0.11</td>
<td>0.01</td>
<td>0.08</td>
<td>0.14</td>
<td>43</td>
</tr>
<tr>
<td>SeniorPov</td>
<td>0.13</td>
<td>0.03</td>
<td>0.09</td>
<td>0.25</td>
<td>43</td>
</tr>
<tr>
<td>TeenPreg</td>
<td>51.43</td>
<td>7.79</td>
<td>34.20</td>
<td>68.30</td>
<td>41</td>
</tr>
<tr>
<td>ViolentCrime</td>
<td>544.64</td>
<td>106.93</td>
<td>363.50</td>
<td>758.20</td>
<td>41</td>
</tr>
<tr>
<td>JuvCrime</td>
<td>337.25</td>
<td>73.64</td>
<td>261.60</td>
<td>497.44</td>
<td>30</td>
</tr>
<tr>
<td>JuvDrug</td>
<td>468.39</td>
<td>143.67</td>
<td>268.75</td>
<td>684.50</td>
<td>30</td>
</tr>
<tr>
<td>JuvAssault</td>
<td>192.55</td>
<td>46.37</td>
<td>127.96</td>
<td>282.02</td>
<td>30</td>
</tr>
<tr>
<td>JuvRape</td>
<td>16.39</td>
<td>3.93</td>
<td>9.49</td>
<td>22.38</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 4.10: Societal Data Summary by Decade

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Mean</th>
<th>Mean</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop</td>
<td>215.03</td>
<td>236.96</td>
<td>264.54</td>
<td>297.05</td>
</tr>
<tr>
<td>rGDP</td>
<td>1.65</td>
<td>4.03</td>
<td>7.29</td>
<td>12.58</td>
</tr>
<tr>
<td>White</td>
<td>0.88</td>
<td>0.86</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>Bachelors</td>
<td>15.98</td>
<td>27.10</td>
<td>37.82</td>
<td>54.28</td>
</tr>
<tr>
<td>Teen</td>
<td>20.84</td>
<td>19.92</td>
<td>18.26</td>
<td>21.37</td>
</tr>
<tr>
<td>Emp</td>
<td>6.22</td>
<td>7.27</td>
<td>5.76</td>
<td>6.32</td>
</tr>
<tr>
<td>Single</td>
<td>0.14</td>
<td>0.20</td>
<td>0.29</td>
<td>0.30</td>
</tr>
<tr>
<td>ChildPov</td>
<td>0.16</td>
<td>0.20</td>
<td>0.21</td>
<td>0.18</td>
</tr>
<tr>
<td>AdultPov</td>
<td>0.09</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>SeniorPov</td>
<td>0.16</td>
<td>0.14</td>
<td>0.11</td>
<td>0.10</td>
</tr>
<tr>
<td>TeenPreg</td>
<td>57.63</td>
<td>52.17</td>
<td>55.91</td>
<td>41.05</td>
</tr>
<tr>
<td>ViolentCrime</td>
<td>451.72</td>
<td>593.71</td>
<td>672.89</td>
<td>467.90</td>
</tr>
<tr>
<td>JuvCrime</td>
<td>-----</td>
<td>305.57</td>
<td>424.51</td>
<td>281.67</td>
</tr>
<tr>
<td>JuvDrug</td>
<td>-----</td>
<td>326.67</td>
<td>504.90</td>
<td>573.59</td>
</tr>
<tr>
<td>JuvAssault</td>
<td>-----</td>
<td>152.06</td>
<td>247.67</td>
<td>177.93</td>
</tr>
<tr>
<td>JuvRape</td>
<td>-----</td>
<td>18.36</td>
<td>19.00</td>
<td>11.81</td>
</tr>
</tbody>
</table>

National Statistics by Decade
As we can see by the decadal breakdown, there has been a decrease in teen pregnancy from 1970 to 2012 and an increasing trend in juvenile violence from 1980 to 2000, and a decline from 673 crimes per 100,000 adolescents to 468 per 100,000 adolescents in the 2000s. Juvenile drug use has actually seen an increase with each subsequent decade. We also see a large increase in number of bachelor degrees from 16 million to 54 million people over the time period and a slight decline in the percentage of the population that is white (0.88 to 0.81 percent). Besides GDP, all other variables remain fairly constant.

4.5 Summary

From the data, we are able to see trends in both music and movies from 1970 to 2012. First, we are able to reproduce the results from Schellenberg and Scheve (2012) and find that the prevalence of minor keys is much higher in the last decade than any of the previous decades, which could have an impact on the longevity of songs if consumers prefer these keys. Second, songs tend to stay on charts longer in the period from 1990 to 2012 by almost twice the amount they did in 1970 to 1980. Third, while groups mainly dominated from the 1970s to the 1990s, the 2000s saw a huge rise in collaboration between artists for number one hits. Finally, profanity counts in songs seem to increase with every subsequent decade and have become much higher than they were in 1970.

Movies saw an increasing trend in their ratings. Both the proportion of PG movies and R-rated movies fell which led to an increase in PG-13 rated movies. This could be one factor that caused movies to have the most profanity in the 1990s compared
with the music industry, which saw profanity peak in the 2000s. The average length of movies remained relatively constant over time, as did the genre shares.

The next Chapter focuses on analyzing the data we have gathered using the automated internet scripts. The profanity variables for both movies and music are compared using a differences-in-differences estimator. We also use the song profanity set in the accelerated time failure model to analyze the effect of profanity on a number one hit’s longevity on the Billboard Hot 100 list along with the artist variables and song variables. Finally, the socioeconomic data are used in two different models using determinates for teenage pregnancy and juvenile crime along with three metrics for profanity in society.
Chapter V
Empirical Analysis

5.1. Introduction

This chapter is structured sequentially to explore the effect of the parental advisory label on profanity in music, the effect of profanity on longevity following implementation of PAL, and the effect of profanity on social issues such as teenage pregnancy and juvenile violence. The differences-in-differences estimator analyzes whether the parental advisory label had an effect on the music industry while controlling for the social acceptability of profanity by using the profanity in the movie industry as the control. Each curse word is individually examined with five different measurements that capture profanity saturation, total curse words, average curse words per song, and per song length. The second model modifies an accelerated time failure model developed by Giles (2007). Our adapted model tests the consumer response to profanity by evaluating the popularity of number-one hits of songs that contain profanity compared with those that do not. The final model uses an ordinary least squares regression to analyze whether or not profanity in popular music is a determinant of teen pregnancy or juvenile crime on an aggregate level.
5.2 Estimating Effect of Parental Advisory Label on Profanity in *Hot 100*

5.2.1 Purpose

The current black and white “Parental Advisory: Explicit Content” label has been in place since 1995, which has given us an *ex-post* period of observation of 17 years. A 2005 Gallop poll of teenagers in the US ages 19 to 17 suggests that the labels are ineffective because teenage consumers will either ignore the labels; or in some cases, the labels will act as an incentive for rebellious teens to purchase the album. This section evaluates whether profanity in popular music changed following the implementation of the labeling standard. To evaluate the effect of the label we need profanity and song data in the periods before and after 1995. A differences-in-differences estimator is used to evaluate the change in profanity in popular music.

5.2.2 Methodology

A differences-in-differences (DD) regression compares a group that received a treatment to a control in order to control for pre-treatment differences between the treatment and control groups in the use of profanity. This is conducted by choosing a control group that is related to the treatment group but is not subject to the same policy that caused the effect in the treatment group. We will use songs as our treatment group, as they were subject to the parental advisory label in 1995. Our control group will consist of movies, which are similar to songs in that they provide entertainment, can contain fowl language, are information goods, and are privately regulated by recording
and motion picture associations. Movies act as a control because their rating system has
not changed significantly over this period and the RIAA’s label had no regulatory impact
on the movie industry. The equation of the DD regression is as follows:

\[ Y_{ipt} = \beta_0 + \beta_1 Media_{it} + \beta_2 Y1995_{it} + \beta_3 (Media_{it} \cdot Y1995_{it}) + \varepsilon_{it} \]  \hspace{1cm} (5.1)

\( Y \) is our dependent variable and represents profanity per observation \( i = 1, ..., N \)
over time \( t = 1980, ..., 2012 \). Our profanity matrix \( p \) contains nine different curse word
sets and are measured in five different metrics. The individual words included as
dependent variables include \( btch, fck, sht, \) and \( ngg \) while the bundles being evaluated for
an effect are Bundle 1 (\( fck, cnt, sht, btch \)), Bundle 2 (\( fck, cnt, sht, btch, ngg \)), Bundle 3
(\( as, dk, fg, pss, slt \)), Bundle 4 (\( as, dk, fg, pss, slt, ngg \)), and SumCurse which contains
every curse word in the dataset. The dependent variables are counted in five different
ways including an average saturation of profanity (number of songs with swear words) in
the top 100 hits per year, a total summation of curse words in the top 100 hits per year, a
total summation of curse words in the top 100 hits per year per minute to control for
length, the average profanity per song per year, and the average profanity per song per
year per minute. \( Media \) is a dummy variable that takes a value of 0 if the observation is
from a movie or 1 if from a song. \( Y1995 \) is also a dummy variable that is mean to
represent the period after the treatment parental advisor label was implemented. Finally,
the coefficient from our interaction term (\( Media \cdot Y1995 \)) is our DD estimator which
measures the impact of the parental advisor label on music content.
5.2.3 Results

Relative to the motion picture industry, profanity in the music industry significantly decreased following the parental advisory labeling for profanity per minute and profanity counts and produce slightly less significant results for average profanity per song both in the count and per minute measures. The individual results are available in Appendix B Tables B1 – B9, and the summary is provided in the following Tables 5.1 and 5.2.

<table>
<thead>
<tr>
<th>Table 5.1: Treatment Effect Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturation</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>btch</td>
</tr>
<tr>
<td>fck</td>
</tr>
<tr>
<td>sht</td>
</tr>
<tr>
<td>ngg</td>
</tr>
<tr>
<td>B1</td>
</tr>
<tr>
<td>B2</td>
</tr>
<tr>
<td>B3</td>
</tr>
<tr>
<td>B4</td>
</tr>
<tr>
<td>Sum</td>
</tr>
</tbody>
</table>

* 10% sig   ** 5% sig   *** 1% sig
Table 5.2: Sign for DD for Significant Values

<table>
<thead>
<tr>
<th></th>
<th>Saturation</th>
<th>Count</th>
<th>Per Min</th>
<th>Average</th>
<th>AvgPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>btch</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fck</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sht</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ngg</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B3</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B4</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sum</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- (Negative) + (Positive)

The results are a representation of the four most prevalent significant curse words: btch, fck, sht, and ngg. Each of the first two bundles (B1, B2) are the more offensive words (cnt, fck, sht, btch) excluding and including ngg respectively. Ngg is omitted from certain bundles because, on average, it tends to have a different sign than the rest of the curse words implying that there was an increase in its frequency per year following the parental advisory label. The second two bundles (B3, B4) are, in general, slightly less offensive (as, dk, fg, pss, slt) excluding and including ngg respectively. Sum is the entire aggregation of all curse words in the dataset.

The models using counts and counts per minute as dependent variables appear to provide the most significant results and show a decrease in profanity usage in the music industry with the exception of ngg. When we look at the average profanity per song, we can see btch is statistically significant at the 5% level, and fck and ngg are statistically significant at the 10% level, yet the rest of the variables showed little significance. Finally, only the third bundle (the more offensive words without ngg), saw a significant decrease in saturation per year.
5.3 The Effect of Profanity on Survival on Hot 100 List

5.3.1 Purpose

The count model used in the previous section shows a decrease in profanity in the period after the parental advisory label was adopted. While these results appear to be favorable for the parental advisory label’s intended purpose, we now wish to consider the effect profanity has on the popularity of a song.

5.3.2 Methodology

Giles (2007) serves as the foundation for this methodology. We embellish the original model by including variables to account for other song characteristics such as key, artist attributes not included in the base model, and our profanity set. This information feeds directly into an accelerated time failure model in which four different distributions are tested: exponential, Weibull, log-normal, and log-logistic. For our dependent variable, we use total weeks spent in the number one position. The failure event happens when the song no longer reappears on the chart rather than being “kicked” off the chart. Therefore, we model our survival time as total weeks as opposed to consecutive weeks due to the “bounce back” effect which is captured in our variable NonCon.

The period of observation spans from 1970 to September 2012 as opposed to Giles (2007) which looks at data from 1955 to 2003. After running each model with the song, artist, and profanity data, the distribution with the lowest Akaike Information
Criterion (AIC) is chosen to represent the best fitting distribution. We also evaluate the Bayesian Information Criterion (BIC) in selecting which distribution is used in the final analysis. The model is then parsed into two sections, pre- and post-1995 to test for changes in preferences.

5.3.3 Results

Unlike Giles (2007), we find that our model is best fitted by a Weibull distribution as opposed to the log-logistic distribution given that the Weibull distribution model has both the lowest AIC and BIC as seen in Table 5.3. Given that the Weibull distribution is the best fitting distribution it is no surprise to see that it contains highly statistically significant hazard ratios. Additionally, our log-normal and log-logistic models still provide highly significant coefficients.

Table 5.4 represents the isolated results of the Weibull distribution and the results from the truncation of the model. We also include \( \ln \ p \), which is our logged shaped parameter. The value of 1.899 implies that the hazard functions increase monotonically, which is exactly what we would expect for the life of songs when consumers get tired of their popularity.
Table 5.3: Maximum Likelihood Estimation of Failure Models

<table>
<thead>
<tr>
<th>Variable</th>
<th>Exponential</th>
<th>Weibull</th>
<th>Log-Normal</th>
<th>Log-Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Haz</td>
<td>Haz</td>
<td>Haz</td>
<td>Haz</td>
</tr>
<tr>
<td>Year</td>
<td>0.975 ***</td>
<td>0.955</td>
<td>1.025 ***</td>
<td>1.026 ***</td>
</tr>
<tr>
<td>Length</td>
<td>0.98</td>
<td>0.944 *</td>
<td>1.017</td>
<td>1.009</td>
</tr>
<tr>
<td>NonCon</td>
<td>0.915</td>
<td>0.864</td>
<td>1.153 *</td>
<td>1.077</td>
</tr>
<tr>
<td>Collaboration</td>
<td>0.786 ***</td>
<td>0.656</td>
<td>1.289 ***</td>
<td>1.286 ***</td>
</tr>
<tr>
<td>as</td>
<td>1.106 *</td>
<td>1.251 ***</td>
<td>0.934 *</td>
<td>0.916 ***</td>
</tr>
<tr>
<td>bitch</td>
<td>1.418 *</td>
<td>1.988 ***</td>
<td>0.728 ***</td>
<td>0.712 ***</td>
</tr>
<tr>
<td>fg</td>
<td>1.046</td>
<td>1.252</td>
<td>1.053</td>
<td>1.005</td>
</tr>
<tr>
<td>fck</td>
<td>1.015</td>
<td>1.033</td>
<td>0.98</td>
<td>0.966</td>
</tr>
<tr>
<td>nng</td>
<td>0.921</td>
<td>0.846 ***</td>
<td>1.073 **</td>
<td>1.097 ***</td>
</tr>
<tr>
<td>pss</td>
<td>1.519</td>
<td>2.374</td>
<td>0.69</td>
<td>0.646</td>
</tr>
<tr>
<td>sht</td>
<td>1.006</td>
<td>1.014 *</td>
<td>0.996</td>
<td>0.995</td>
</tr>
<tr>
<td>tts</td>
<td>3.263 *</td>
<td>11.364 ***</td>
<td>0.367 **</td>
<td>0.298 ***</td>
</tr>
<tr>
<td>Major</td>
<td>1.086</td>
<td>1.147 **</td>
<td>0.916 **</td>
<td>0.904 **</td>
</tr>
<tr>
<td>Group</td>
<td>0.995</td>
<td>0.962</td>
<td>0.974</td>
<td>0.965</td>
</tr>
<tr>
<td>Male</td>
<td>1.131 *</td>
<td>1.277 ***</td>
<td>0.89 **</td>
<td>0.899 **</td>
</tr>
<tr>
<td>YearsActive</td>
<td>0.986 ***</td>
<td>0.971 ***</td>
<td>1.012 ***</td>
<td>1.013 ***</td>
</tr>
<tr>
<td>Rock</td>
<td>1.166</td>
<td>1.43 ***</td>
<td>0.926</td>
<td>0.914</td>
</tr>
<tr>
<td>Pop</td>
<td>1.139 *</td>
<td>1.273 ***</td>
<td>0.887 ***</td>
<td>0.876 ***</td>
</tr>
<tr>
<td>Country</td>
<td>0.818</td>
<td>0.591 **</td>
<td>1.044</td>
<td>0.977</td>
</tr>
<tr>
<td>RB</td>
<td>0.866 *</td>
<td>0.752 ***</td>
<td>1.14 ***</td>
<td>1.138 ***</td>
</tr>
<tr>
<td>HipHop</td>
<td>0.824 *</td>
<td>0.721 ***</td>
<td>1.267 ***</td>
<td>1.303 ***</td>
</tr>
<tr>
<td>Folk</td>
<td>0.774</td>
<td>0.671 **</td>
<td>1.282 *</td>
<td>1.336 **</td>
</tr>
<tr>
<td>Blues</td>
<td>0.994</td>
<td>1.12</td>
<td>1.081</td>
<td>1.11</td>
</tr>
<tr>
<td>Disco</td>
<td>1.148</td>
<td>1.412 ***</td>
<td>0.917</td>
<td>0.931</td>
</tr>
<tr>
<td>Funk</td>
<td>1.494 ***</td>
<td>2.22 ***</td>
<td>0.701 ***</td>
<td>0.669 ***</td>
</tr>
<tr>
<td>OtherMiss</td>
<td>1.455 **</td>
<td>2.235 ***</td>
<td>0.761 **</td>
<td>0.734 ***</td>
</tr>
<tr>
<td>D1998</td>
<td>1.367 ***</td>
<td>1.749 ***</td>
<td>0.742 ***</td>
<td>0.746 ***</td>
</tr>
<tr>
<td>constant</td>
<td>2.13E+21 ***</td>
<td>1.71E+38 ***</td>
<td>0 ***</td>
<td>0 ***</td>
</tr>
<tr>
<td>AIC</td>
<td>3327.388</td>
<td>2669.599</td>
<td>2841.324</td>
<td>2879.142</td>
</tr>
<tr>
<td>BIC</td>
<td>3473.701</td>
<td>2821.138</td>
<td>2992.863</td>
<td>3030.681</td>
</tr>
</tbody>
</table>

† Lowest AIC or BIC
### Table 5.4: Weibull Distributed Results by Year

<table>
<thead>
<tr>
<th>Variables</th>
<th>All Years</th>
<th>Pre-1995</th>
<th>Post-1995</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Haz</td>
<td>SE</td>
<td>Haz</td>
</tr>
<tr>
<td>Year</td>
<td>0.955 ***</td>
<td>0</td>
<td>0.954 ***</td>
</tr>
<tr>
<td>Length</td>
<td>0.944 *</td>
<td>0.03</td>
<td>1.017</td>
</tr>
<tr>
<td>NonCon</td>
<td>0.864</td>
<td>0.11</td>
<td>0.624 **</td>
</tr>
<tr>
<td>Collaboration</td>
<td>0.656 ***</td>
<td>0.05</td>
<td>0.562 ***</td>
</tr>
<tr>
<td>as</td>
<td>1.251 ***</td>
<td>0.07</td>
<td>10.621 **</td>
</tr>
<tr>
<td>btch</td>
<td>1.988 ***</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>fg</td>
<td>1.252</td>
<td>0.28</td>
<td>1.199</td>
</tr>
<tr>
<td>fck</td>
<td>1.033</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>ngg</td>
<td>0.846 ***</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>pss</td>
<td>2.374</td>
<td>2.29</td>
<td></td>
</tr>
<tr>
<td>sht</td>
<td>1.014 *</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>tts</td>
<td>11.364 ***</td>
<td>8.55</td>
<td></td>
</tr>
<tr>
<td>Major</td>
<td>1.147 **</td>
<td>0.08</td>
<td>1.138</td>
</tr>
<tr>
<td>Group</td>
<td>0.962</td>
<td>0.07</td>
<td>1.096</td>
</tr>
<tr>
<td>Male</td>
<td>1.277 ***</td>
<td>0.1</td>
<td>1.192</td>
</tr>
<tr>
<td>YrActive</td>
<td>0.971 ***</td>
<td>0</td>
<td>1.009</td>
</tr>
<tr>
<td>Rock</td>
<td>1.43 ***</td>
<td>0.12</td>
<td>1.272 **</td>
</tr>
<tr>
<td>Pop</td>
<td>1.273 ***</td>
<td>0.09</td>
<td>1.115</td>
</tr>
<tr>
<td>Country</td>
<td>0.591 **</td>
<td>0.12</td>
<td>0.381 ***</td>
</tr>
<tr>
<td>RB</td>
<td>0.752 ***</td>
<td>0.06</td>
<td>0.539 ***</td>
</tr>
<tr>
<td>HipHop</td>
<td>0.721 ***</td>
<td>0.08</td>
<td>1.815 **</td>
</tr>
<tr>
<td>Folk</td>
<td>0.671 **</td>
<td>0.13</td>
<td>0.498 ***</td>
</tr>
<tr>
<td>Blues</td>
<td>1.12</td>
<td>0.24</td>
<td>1.014</td>
</tr>
<tr>
<td>Disco</td>
<td>1.412 ***</td>
<td>0.18</td>
<td>1.213</td>
</tr>
<tr>
<td>Funk</td>
<td>2.22 ***</td>
<td>0.36</td>
<td>1.913 ***</td>
</tr>
<tr>
<td>OtherMisc</td>
<td>2.235 ***</td>
<td>0.39</td>
<td>1.795 **</td>
</tr>
<tr>
<td>D1998</td>
<td>1.749 ***</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>ln_p</td>
<td>1.899 ***</td>
<td>0.04</td>
<td>1.904 ***</td>
</tr>
<tr>
<td>AIC</td>
<td>2669.599</td>
<td></td>
<td>1421.763</td>
</tr>
</tbody>
</table>
Unfortunately, given the scarcity of curse words in the number one songs pre-1995, most of our curse words are omitted and therefore we are not able to evaluate if there was a change in the preference for profanity after the implementation of the label. However, we can still make meaningful inferences from the data. First, we see that after 1995 longer songs had less of a risk of being kicked off the list. We see little overall significance for non-consecutive hits except in values pre-1995. There is a large significance to collaborative songs (songs with two popular artists working together), which are a little over half as likely to lose their number one spot as songs that feature a single artist or band. We also see that after 1995, songs in a major key are less popular than songs in a minor key, which could explain the 30% increase in the last decade.

While group suggests that there is no significant advantage for being a group as opposed to a single artist, like Giles (2007), we find that gender matters especially in recent years. YearsActive is relatively close to 1 in value implying that it has little effect. In regards to genres, disco and funk genres seem to have an extremely brief period in the number one spot as opposed to other significant genres. Pre-1995, country songs did extremely well at holding their positions, but after 1995, they suffer greatly which seems to be the case for folk songs as well. HipHop is the only significant genre after 1995 that seems to enjoy more time at the top of the charts.

Turning to our profanity variables, even though there is not enough data in the period prior to the label, we are able to make inferences about the nature of profanity after the labeling. While we are not able to look as if there was a change in preferences, we see that songs which contain repeated profanity experience shorter longevity in the
number one spot with the exception of *ngg*, which statistically increases the likelihood of songs remaining in the number one spot at the 1% significance level. The results show that consumers do not prefer songs with profanity as much as they do songs without after 1995.

### 5.4 Profanity as a Determinant of Juvenile Crime and Teenage Pregnancy

#### 5.4.1 Purpose

This section differs from the previous two sections, by evaluating whether profanity in popular music is a determinant of juvenile crime or teenage pregnancy. While many may believe that profanity is harmful for children, the exact effect of profanity is still ambiguous. If profanity leads to increased violence, we would expect to see an increase in juvenile violence when more profanity is widely circulated in public. Songs which contain sexually explicit profanity may also increase a teenager’s risk of sexual activity which, if not properly discussed, may lead to teenage pregnancies. We wish to test if an increase in profanity exposure in songs results in criminal behavior or increased sexual activity in the teenage population.
5.4.2 Methodology

We evaluate the growth and the first difference of dependent variables (juvenile crime and teen pregnancy rates) in this section. Using an Ordinary Least Squares (OLS), we regress juvenile crime and teen pregnancy separately on the determinants described previously in Chapter II (Literature Review) and include variables that represent yearly profanity in the *Hot 100* list. Both models contain four different variations of the base model. The second model omits variables to evaluate the robustness of the coefficients. Models three through five each include a profanity variable. The first variable represent the saturation of profanity in the *Hot 100* list, the second variable measures the average profanity per song, and the third variable measures the average profanity per year. The following models are estimated using a Prais-Winsten transformation to adjust for autocorrelation.

5.4.3 Results

Regression results are available in Tables 5.5 through 5.8. While we were unable to find any significance for profanity as a determinant of either teen pregnancy or juvenile crime, this could be due to misspecifications in the model, as the research for aggregate determinants of these dependent variables are quite limited. Although the adjusted $R^2$ of the regression is close to one, we usually expect these results with time-series data. The first regressions are represented in the log-log format to represent the data in changes rather than levels and are adjusted for autocorrelated errors. Finally, our omitted model helps to show that our model is robust to truncation parameters.
Table 5.5: Dependent Variable: ln(Teen Pregnancy)

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop</td>
<td>1.867 **</td>
<td>1.093 *</td>
<td>0.705</td>
<td>1.299 **</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.64</td>
<td>1.6</td>
<td>0.61</td>
<td>1.61</td>
</tr>
<tr>
<td>rGDP</td>
<td>-0.527 **</td>
<td>-0.531 **</td>
<td>-0.751 ***</td>
<td>-0.579 **</td>
<td>-0.734 ***</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>0.21</td>
<td>0.22</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>White</td>
<td>0.682</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelors</td>
<td>0.298</td>
<td>0.317</td>
<td>1.17 **</td>
<td>0.328</td>
<td>1.146 **</td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.31</td>
<td>0.51</td>
<td>0.32</td>
<td>0.5</td>
</tr>
<tr>
<td>Teen</td>
<td>-1.113 ***</td>
<td>-0.635 ***</td>
<td>-1.34 ***</td>
<td>-0.695 ***</td>
<td>-1.347 ***</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.23</td>
<td>0.29</td>
<td>0.22</td>
<td>0.29</td>
</tr>
<tr>
<td>Emp</td>
<td>-0.166 ***</td>
<td>-0.138 ***</td>
<td>-0.312 ***</td>
<td>-0.139 ***</td>
<td>-0.31 ***</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Single</td>
<td>-0.214 **</td>
<td>0.039</td>
<td>-0.582 ***</td>
<td>0.043</td>
<td>-0.582 ***</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.11</td>
<td>0.15</td>
<td>0.11</td>
<td>0.15</td>
</tr>
<tr>
<td>CPov</td>
<td>0.024</td>
<td>0.033</td>
<td>0.192 **</td>
<td>0.026</td>
<td>0.188 **</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.08</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>VC</td>
<td>0.354 ***</td>
<td>0.421 ***</td>
<td>0.334 **</td>
<td>0.428 ***</td>
<td>0.328 **</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.1</td>
<td>0.11</td>
<td>0.1</td>
<td>0.11</td>
</tr>
<tr>
<td>SCpm</td>
<td></td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DASC</td>
<td></td>
<td>-0.026</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASCpm</td>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>cons</td>
<td>-1.69</td>
<td>0.926</td>
<td>4.349</td>
<td>0.271</td>
<td>4.375</td>
</tr>
<tr>
<td></td>
<td>4.18</td>
<td>3.32</td>
<td>7.33</td>
<td>3.22</td>
<td>7.36</td>
</tr>
<tr>
<td>R²</td>
<td>0.981</td>
<td>0.923</td>
<td>0.999</td>
<td>0.937</td>
<td>0.999</td>
</tr>
<tr>
<td>AIC</td>
<td>-178.79</td>
<td>-179.03</td>
<td>-99.74</td>
<td>-177.81</td>
<td>-99.72</td>
</tr>
</tbody>
</table>

Note: standard errors are shown under each coefficient
Models 2 - 5 use HAC standard errors
<table>
<thead>
<tr>
<th>Model</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop</td>
<td>2.938</td>
<td>2.43</td>
<td>-10.383</td>
<td>1.669</td>
<td>-9.814</td>
</tr>
<tr>
<td></td>
<td>1.96</td>
<td>1.69</td>
<td>8.82</td>
<td>2.07</td>
<td>9.55</td>
</tr>
<tr>
<td>rGDP</td>
<td>-0.834</td>
<td>-0.533</td>
<td>* 2.787</td>
<td>-0.489</td>
<td>* 2.669</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.26</td>
<td>2.18</td>
<td>0.28</td>
<td>2.37</td>
</tr>
<tr>
<td>White</td>
<td>-5.842</td>
<td>-6.947</td>
<td>* 11.156</td>
<td>-7.902</td>
<td>* 10.794</td>
</tr>
<tr>
<td></td>
<td>3.62</td>
<td>3.8</td>
<td>9.28</td>
<td>4.25</td>
<td>9.81</td>
</tr>
<tr>
<td>Bachelors</td>
<td>0.431</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teen</td>
<td>-0.876</td>
<td>* -0.997</td>
<td>** -1.99</td>
<td>-0.933</td>
<td>** -1.96</td>
</tr>
<tr>
<td></td>
<td>0.42</td>
<td>0.4</td>
<td>1.3</td>
<td>0.41</td>
<td>1.42</td>
</tr>
<tr>
<td>Emp</td>
<td>-0.162</td>
<td>-0.117</td>
<td>0.148</td>
<td>-0.077</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td>0.11</td>
<td>0.07</td>
<td>0.2</td>
<td>0.09</td>
<td>0.21</td>
</tr>
<tr>
<td>Single</td>
<td>-0.49</td>
<td>** -0.54</td>
<td>** -1.152</td>
<td>** -0.409</td>
<td>-1.106</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.19</td>
<td>0.45</td>
<td>0.27</td>
<td>0.49</td>
</tr>
<tr>
<td>APov</td>
<td>0.551 ***</td>
<td>0.455 ***</td>
<td>0.959 ***</td>
<td>0.37 *</td>
<td>0.942 ***</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>0.13</td>
<td>0.21</td>
<td>0.18</td>
<td>0.21</td>
</tr>
<tr>
<td>VC</td>
<td>1.296 ***</td>
<td>1.239 ***</td>
<td>0.243</td>
<td>1.251 ***</td>
<td>0.278</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.14</td>
<td>0.63</td>
<td>0.14</td>
<td>0.68</td>
</tr>
<tr>
<td>JuvDrug</td>
<td>0.258 ***</td>
<td>0.299 ***</td>
<td>0.445 **</td>
<td>0.312 ***</td>
<td>0.435 **</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.08</td>
<td>0.15</td>
<td>0.08</td>
<td>0.16</td>
</tr>
<tr>
<td>SCpm</td>
<td></td>
<td></td>
<td></td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>DASC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.049</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>AvgSCpm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>cons</td>
<td>-12.112</td>
<td>-10.457</td>
<td>42.984</td>
<td>-7.247</td>
<td>40.608</td>
</tr>
<tr>
<td></td>
<td>8.31</td>
<td>7.62</td>
<td>36.84</td>
<td>9.22</td>
<td>40.03</td>
</tr>
<tr>
<td>R²</td>
<td>0.985</td>
<td>0.986</td>
<td>0.999</td>
<td>0.985</td>
<td>0.999</td>
</tr>
<tr>
<td>AIC</td>
<td>-116.30</td>
<td>-112.55</td>
<td>-83.02</td>
<td>-111.23</td>
<td>-82.69</td>
</tr>
</tbody>
</table>

Note: standard errors are shown under each coefficient
Models 2 - 5 use HAC standard errors
For each of the logged models, model 1 is an OLS regression; model 2 demonstrates the robustness of the regression after omitting *White* for teenage pregnancy and *Bachelors* for juvenile crime as well as introducing HAC standard errors; model 3 adds the variable for all profanity per minute in a given year, model 4 replaces the profanity variable added in model 3 with the profanity saturation variable, and model 5 replaces the saturation variable with an average curse word per song variable. Each of these variables is found to be statistically insignificant determinants of teenage pregnancy and juvenile crime. Furthermore, the coefficients on our socioeconomic variables tend to change greatly with the sum of curse words and average curse words while having overinflated standard errors, indicating problems with multicollinearity.

For teenage pregnancy, model 2 shows that teenage pregnancy rates decrease by 0.531% with every 1% increase in real GDP and is significant at the 5% significance level. Teenage share of the population and unemployment are found to be negatively correlated with teenage pregnancy rates at the 1% significant level, and a 1% increase in violent crime rates is correlated with a 0.42% increase in teen pregnancy rates at the 1% significance level. Turning to the juvenile crime model, we see that adult poverty rates, violent crime, and juvenile drug abuse are the most statistically significant determinants in model 2. Each of these determinants is positively correlated with juvenile crime rates. Once again, we see that there is no statistical significance for our profanity variables.

Since our logged variables provided no statistical significance for the profanity variables, we then evaluate the same models using difference estimators.
Table 5.7: Teen Pregnancy with Difference Estimators

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔPop</td>
<td>0.206</td>
<td>-0.293</td>
<td>-0.478</td>
<td>-0.283</td>
<td>-0.375</td>
</tr>
<tr>
<td></td>
<td>0.59</td>
<td>1.01</td>
<td>1.04</td>
<td>1.03</td>
<td>1.03</td>
</tr>
<tr>
<td>ΔrGDP</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ΔWhite</td>
<td>167.995</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>221.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔBachelors</td>
<td>0.587</td>
<td>0.425</td>
<td>0.348</td>
<td>0.447</td>
<td>0.389</td>
</tr>
<tr>
<td></td>
<td>0.56</td>
<td>0.41</td>
<td>0.41</td>
<td>0.42</td>
<td>0.41</td>
</tr>
<tr>
<td>ΔTeen</td>
<td>-2.213</td>
<td>***</td>
<td>0</td>
<td>0.165</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>0.64</td>
<td>0.67</td>
<td>0.67</td>
<td>0.68</td>
<td>0.67</td>
</tr>
<tr>
<td>ΔEmp</td>
<td>-0.193</td>
<td>0.164</td>
<td>0.195</td>
<td>0.177</td>
<td>0.172</td>
</tr>
<tr>
<td></td>
<td>0.27</td>
<td>0.29</td>
<td>0.29</td>
<td>0.3</td>
<td>0.29</td>
</tr>
<tr>
<td>ΔSingle</td>
<td>-0.07</td>
<td>-5.482</td>
<td>-11.976</td>
<td>-6.133</td>
<td>-11.685</td>
</tr>
<tr>
<td></td>
<td>29.29</td>
<td>27.01</td>
<td>26.75</td>
<td>27.4</td>
<td>27.11</td>
</tr>
<tr>
<td>ΔC pov</td>
<td>-6.525</td>
<td>-3.975</td>
<td>-4.548</td>
<td>-4.782</td>
<td>-4.348</td>
</tr>
<tr>
<td></td>
<td>13.49</td>
<td>9.43</td>
<td>9.18</td>
<td>9.79</td>
<td>9.29</td>
</tr>
<tr>
<td>ΔVC</td>
<td>0.024</td>
<td>**</td>
<td>0.014</td>
<td>0.015</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>ΔSC pm</td>
<td>-0.008</td>
<td></td>
<td>-0.864</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td></td>
<td>2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔASC</td>
<td></td>
<td></td>
<td></td>
<td>-0.521</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>cons</td>
<td>-2.349</td>
<td>-0.748</td>
<td>-0.079</td>
<td>-0.783</td>
<td>-0.368</td>
</tr>
<tr>
<td></td>
<td>1.76</td>
<td>2.73</td>
<td>2.85</td>
<td>2.79</td>
<td>2.81</td>
</tr>
<tr>
<td>R²</td>
<td>0.518</td>
<td>0.132</td>
<td>0.176</td>
<td>0.135</td>
<td>0.166</td>
</tr>
<tr>
<td>AIC</td>
<td>147.76</td>
<td>138.34</td>
<td>138.20</td>
<td>140.17</td>
<td>138.70</td>
</tr>
</tbody>
</table>

Note: standard errors are shown under each coefficient
Models 2 - 5 use HAC standard errors
<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔPop</td>
<td>1.486</td>
<td>-4.882</td>
<td>-4.774</td>
<td>-7.732</td>
<td>-4.649</td>
</tr>
<tr>
<td></td>
<td>6.02</td>
<td>8.41</td>
<td>8.54</td>
<td>10.49</td>
<td>8.5</td>
</tr>
<tr>
<td>ΔrGDP</td>
<td>0.034</td>
<td>0.044</td>
<td>*</td>
<td>0.044</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>ΔWhite</td>
<td>1979.397</td>
<td>2781.474</td>
<td>2705.796</td>
<td>3963.353</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>2052</td>
<td>1929.31</td>
<td>2051.21</td>
<td>1963.53</td>
<td>2075.55</td>
</tr>
<tr>
<td>ΔBachelors</td>
<td>8.367</td>
<td></td>
<td></td>
<td></td>
<td>5.71</td>
</tr>
<tr>
<td>ΔTeen</td>
<td>-14.192</td>
<td>-14.332</td>
<td>-14.092</td>
<td>-17.311</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>8.97</td>
<td>9.19</td>
<td>9.75</td>
<td>9.7</td>
<td>9.85</td>
</tr>
<tr>
<td>ΔEmp</td>
<td>2.132</td>
<td>2.964</td>
<td>2.947</td>
<td>2.536</td>
<td>2.959</td>
</tr>
<tr>
<td></td>
<td>3.96</td>
<td>4.01</td>
<td>4.14</td>
<td>4.1</td>
<td>4.17</td>
</tr>
<tr>
<td>ΔSingle</td>
<td>194.563</td>
<td>404.959</td>
<td>394.746</td>
<td>496.591</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>295.23</td>
<td>286.95</td>
<td>296.01</td>
<td>271.68</td>
<td>297.74</td>
</tr>
<tr>
<td>ΔAPov</td>
<td>1509.216</td>
<td>*</td>
<td>1090.675</td>
<td>*</td>
<td>1545.012</td>
</tr>
<tr>
<td></td>
<td>2052</td>
<td>1929.31</td>
<td>2051.21</td>
<td>1963.53</td>
<td>2075.55</td>
</tr>
<tr>
<td>ΔVC</td>
<td>0.767</td>
<td>***</td>
<td>0.613</td>
<td>***</td>
<td>0.617</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.11</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>ΔJuvDrug</td>
<td>0.264</td>
<td>***</td>
<td>0.255</td>
<td>***</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>ΔSCpm</td>
<td></td>
<td>-0.012</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>ΔDASC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>46.578</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31.17</td>
</tr>
<tr>
<td>ΔAvgSCpm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.716</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.4</td>
</tr>
<tr>
<td>cons</td>
<td>-23.329</td>
<td>1.286</td>
<td>1.182</td>
<td>5.361</td>
<td>0.901</td>
</tr>
<tr>
<td></td>
<td>18.3</td>
<td>22.65</td>
<td>23</td>
<td>28.35</td>
<td>22.81</td>
</tr>
<tr>
<td>R^2</td>
<td>0.856</td>
<td>0.755</td>
<td>0.758</td>
<td>0.748</td>
<td>0.759</td>
</tr>
<tr>
<td>AIC</td>
<td>232.35</td>
<td>225.31</td>
<td>227.27</td>
<td>224.42</td>
<td>227.28</td>
</tr>
</tbody>
</table>

Note: standard errors are shown under each coefficient
Models 2 - 5 use HAC standard errors
After differencing our variables, we find no statistical significance for our profanity variables chosen to represent profanity in the music industry. The differenced teenage pregnancy model only finds two variables, violent crime and teen share, statistically significant at the 5% and 1% significance levels respectively. Once adjusting for autocorrelation, we find no significant variables. When we look at the juvenile crime model, we see results that are slightly more significant. Juvenile drug and violent crime rates are statistically significant at the 1% level in determining the change in juvenile crime rates. Adult poverty and real GDP are significant at the 10% significance level after adjusting for autocorrelation.

5.5 Summary

Each model adds its own contribution to the overall analysis of the profanity labels in the music industry. Although we are able to see a steady increase in profanity over time from the data, our differences-in-differences model has found that, relative to the movie industry, profanity in the music industry has decreased in the period following the implementation of the parental advisory sticker for almost all profane word counts per year. However, we did not find any significant change in the number of songs per year that contained profanity (saturation) or the average profanity count per song on the Hot 100 list.

Next, we use the accelerated time failure model to evaluate the effect profanity has on the longevity of a song on the Hot 100 list. This list can be characterized as a representation of average consumer preferences for songs in the United States each year.
We find that, after 1995, words such as *btch, tts, and sht* decrease a song’s longevity at the 1% significance level, *fck* decreases a song’s longevity at the 10% significance level, and *ngg* actually increases a song’s longevity at the 5% significance level. Seeing that the aggregated consumer preferences tend to favor lower counts of profanity, this decrease in profanity in songs following the label may be due to producers changing the composition of the songs to respond to a change in consumer preferences. Due to the lack of profanity data pre-1995, we cannot say definitively that this drop in profanity was caused by changing consumer preference because our model dropped these observations.

Finally, our societal analysis shows that after accounting for growth rates and differences, our three indexes for aggregate profanity in songs show no statistically significant effect on teen pregnancy and juvenile crime rates. It must be noted that our difference model for teenage pregnancy did not seem to fit the data as closely as our juvenile crime model, especially after adjusting for autocorrelation. However, given our sample size and the availability of data, these models take advantage of the existing dataset to derive inferences that are otherwise unexplained. We have shown that profanity has dropped after the implementation of the parental advisory label; however, this did not have much of an effect on social issues such as teenage pregnancy and juvenile crime.
Chapter VI

Conclusion

In the current political climate, there has been a rush to pinpoint the cause of violent outbursts in people that results in tragic casualties. Some attribute violent behavior to a child’s exposure to profanity. With efforts such as Tipper Gore’s push to clean up the music industry facing such a large backlash from artists, recording companies resorted to informing parents of explicit content using labels. Unlike the 1980s, parents today have better access to information about the music their children listen by accessing watchdog websites and lyric databases. While these resources are also available for movies, we see that the movie industry has a vastly more informative rating system than the music industry. Therefore, understanding the effect of the PAL is important when asking questions about profanity usage and the violence that is claimed to arise from increased profanity exposure.

This study was designed to look at the three segments the label is meant to impact: producers, consumers, and society as a whole. To evaluate if there was a decrease in profanity as a result of the label, we used a differences-in-differences model and compared profanity in the music industry to profanity in movies as a control for other factors such as social and moral responsibility. Next, we used an accelerated time failure model to estimates song longevity and the effect of profanity on the hazard of reducing the longevity of the song. This methodology allowed us to evaluate the popularity of hit songs in the United States and determine if increasing profanity usage leads to consumers substituting away from the good. Our final model is a time series model that evaluates
the effect of profanity on teenage pregnancy rates and juvenile crime rates by using
growth rates and difference estimators on socioeconomic variables as well as a profanity
index.

In order to extract the necessary data, we gathered data using Visual Basic for
Applications (VBA) scripts and compiled information from various websites. After
indexing a list of all songs on the Billboard Hot 100 list from 1970 to 2012, we gathered
information about each song as well as the artist using scripts provided in Appendix A.
We then collected the lyrics of the music and the subtitles of each movie and searched for
each usage of profanity, which are then recorded in our set.

The analysis of Parental Advisory label (PAL) suggests that there was a reduction
in profanity in popular recordings in the period following the implementation of the label.
The results from the differences-in-differences (DD) estimators show that there was a
decrease in total profanity per year (and per minute) for less offensive profanity (dk, fg,
slt, and tts) as well as the more offensive profanity (fck, cnt, sht, and btch) and for the
individual words fck and sht. We find that the DD sign for the word ngg was positive
when using count data, and was the only word among those selected as profanity words
with a positive sign.

The hazard model, built off Giles (2007), contained many more significant
variables than the original model. After truncating the model to values before and after
1995, we lose observations. Most curse variables had to be omitted from the regression
because there are too few observations and STATA automatically dropped these
variables. Instead, an examination of the entire dataset suggests that ngg decreases the
hazard of a song getting kicked off the number one position, while as and bitch increase that likelihood. Looking at values strictly post-1995, sht and fck increase the hazard of losing the number one spot. Other factors such as genre, gender, years of experience, key, and collaborative songs also have significant hazard ratios. We can conclude that as songs contain more profanity, the risk of losing their number one spot increases which shows that consumers actually respond negatively to profanity on the Hot 100 list. This effect is also prevalent in the data since the number one singles already contain less profanity on average than songs in the entire set.

Our final analysis of socioeconomic data including the profanity indicators shows that there is no statistical significance between profanity in music and national juvenile crime or teen pregnancy rates. After implementing five models for each social issue as well as using both logged and differenced variables, we find no significance for each of the three chosen profanity indexes. Even after adjusting for autocorrelation in the time series data and testing the robustness of our coefficients, there seems to be no effect. However, these results are subject to the fact that the model’s standard errors suggest some multicollinearity in the regressions. Even with this issue, our profanity saturation index model does not show the same indicators for multicollinearity, and that model also found profanity to be insignificant in determining teen pregnancy or crime.

The thesis shows that profanity rates in the top 100 songs and top 100 movies have been increasing over time. Even with this increasing trend in profanity, we see that the labeling was able to reduce profanity counts in songs relative to movies. On average, songs in the number one spot on the Billboard’s Hot 100 list enjoyed more time in that
position if they contained less curse words, had a minor key, were recorded by female artists, or was a collaborative effort between two artists. Finally, we see that our profanity indexes were not statistically significant in determining teenage pregnancy when using growth rates or first differences.

Our results seem to suggest that there might be an issue with profanity variables and genres being covariates. We would expect omitted variable bias because sexual content and violence are both correlated with ratings (as given by MPAA standard), as well as profanity and genre. These two variables are likely explaining the same variance that has to do with sexual content. Future studies may consider applying a similar process presented in this thesis to extract data from adult entertainment websites that catalogue sexual content by levels of graphic content.

One potential concern with the analysis in this thesis is that the dataset is missing nearly 1,000 observations. Obtaining this data would require merging multiple lyric databases and expanding the subtitles search to include poorly rated subtitles or perhaps even movie transcripts. The other, more costly method would be to catalogue each observation manually. Unfortunately, the time cost was too great for the purposes of this thesis. Additionally, the dataset used in the analysis contains over 2,000 observations - close to two-thirds of the overall population of songs and their lyrics matching the criteria established in the data chapter. This sample provides a good representation of the overall population of songs, allowing us to make inference about the effects of profanity and the parental advisory labeling.
The purpose of this thesis is to provide a foundation for data search methodology and analysis of the music industry using these search techniques that are not widely implemented for this type of analysis. We demonstrate the benefits of automation and indexing by providing a low cost way of obtaining data, freeing up more time to focus on analysis.

It is important to consider the effect of profanity on society when we add labels to songs and albums. Additionally, with the digital age came a change in the methods people go about acquiring these goods. Recently, there has been a declining physical compact disc market compared to a rising digital market given the price of songs, prevalence of MP3 players, and other factors that lead to an advantage in digital downloads as explained in Chapter II (Literature Review). This change in the way we purchase music has made it much easier for children to buy songs. To download a song, all they need is their parent’s credit card or account information on a popular music database. They are also able to stream songs online and, if they are technically inclined, they are easily able to download all of the songs they need (otherwise known as pirating). Given that parents do not have to physically accompany their children when purchasing songs, perhaps a new rating system for all songs is necessary given that they are available at a click of a button. While the thesis research suggests that the PAL was effective in reducing profanity, it was created before the prevalence of digital downloads back when CDs were the only means to buying songs. If the purpose of the PAL is to inform parents of profanity, and their children are easily able to download songs without parental supervision, then it is possible that we could see the effect of the label become less
significant over time. Therefore, further research is necessary to prove the effect of profanity in popular music on the teenage population in the United States of America by evaluating individual juvenile preferences for music, characteristics, and violent or immoral behavior.
Appendix A

Visual Basic for Application Scripts

The following scripts correspond are described in Chapter III (Data Search Methodology) and are sorganized in the order in which they should be executed.

**A1. Hot 100 List Indexer**

```vbnet
Sub LyricComp()

    Dim ie As Object
    Set ie = CreateObject("internetexplorer.application")

    Range("A1").Select
    ActiveCell = "year"
    ActiveCell.Offset(0, 1) = "rank"
    ActiveCell.Offset(0, 2) = "title"
    ActiveCell.Offset(0, 3) = "titlelink"
    ActiveCell.Offset(0, 4) = "artist"
    ActiveCell.Offset(0, 5) = "artistlink"
    ActiveCell.Offset(0, 6) = "artistindex"
    ActiveCell.Offset(0, 7) = "other artists"

    For iYear = 1970 To 2012
        Range("A" & (iYear - 1970) * 100 + 2).Select

        With ie
            .Visible = True
            .navigate "http://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles_of_" & iYear
            Do While .readystate <> 4 Or .busy: DoEvents: Loop
            Set objHTML = .document.body.getelementsbytagname("table")
        End With

        Set objTR = objHTML.Item(0).getelementsbytagname("tr")

        For j = 1 To objTR.Length - 1
            Set objTD = objTR.Item(j).getelementsbytagname("td")

            If objTD.Length = 3 Then
                iRank = objTD.Item(0).innerText

                If InStr(LCase(objTD.Item(1).innerText), "<a") Then
                    strTitleLink = objTD.Item(1).getelementsbytagname("a").Item(0).getAttribute("href")
                    strTitle = objTD.Item(1).innerText
                Else
                    strTitle = objTD.Item(1).innerText
                    strTitleLink = ""
                End If
```
If InStr(LCase(objTD.Item(2).innerHTML), "<a") > 0 Then
    Set ObjA = objTD.Item(2).getelementsbytagname("a")
    strArtistLink = ObjA.Item(0).getAttribute("href")
    strArtist = ObjA.Item(0).innerText
    strArtistIndex = objTD.Item(2).innerText
    If ObjA.Length > 1 Then
        ActiveCell.Offset(0, 7).Select
        ActiveCell = 1
        For m = 0 To ObjA.Length - 1
            ActiveCell.Offset(0, m * 2 + 1) = ObjA.Item(m).innerText
            ActiveCell.Offset(0, (m * 2) + 2) = ObjA.Item(m).getAttribute("href")
        Next m
        ActiveCell.Offset(0, -7).Select
    End If
    Else
        strArtist = objTD.Item(2).innerText
        strArtistLink = 
        strArtistIndex = strArtist
    End If
Else
    iRank =
    objTR.Item(j).getelementsbytagname("th").Item(0).innerText
    If InStr(LCase(objTD.Item(0).innerHTML), "<a") Then
        strTitleLink =
        objTD.Item(0).getelementsbytagname("a").Item(0).getAttribute("href")
    Else
        strTitle = objTD.Item(0).innerText
        strTitleLink = ""
    End If
    End If
    If InStr(LCase(objTD.Item(1).innerHTML), "<a") > 0 Then
        Set ObjA1 = objTD.Item(1).getelementsbytagname("a")
        strArtistLink = ObjA1.Item(0).getAttribute("href")
        strArtist = ObjA1.Item(0).innerText
        strArtistIndex = objTD.Item(1).innerText
        If ObjA1.Length > 1 Then
            ActiveCell.Offset(0, 7).Select
            ActiveCell = 1
            For m = 0 To ObjA1.Length - 1
                ActiveCell.Offset(0, m * 2 + 1) = ObjA1.Item(m).innerText
                ActiveCell.Offset(0, (m * 2) + 2) = ObjA1.Item(m).getAttribute("href")
            Next m
            ActiveCell.Offset(0, -7).Select
        End If
        Else
            strArtist = objTD.Item(1).innerText
        End If
strArtistLink = ""
strArtistIndex = strArtist
End If
End If

ActiveCell = iYear
ActiveCell.Offsets(0, 1) = iRank
ActiveCell.Offsets(0, 2) = strTitle
ActiveCell.Offsets(0, 3) = strTitleLink
ActiveCell.Offsets(0, 4) = strArtist
ActiveCell.Offsets(0, 5) = strArtistLink
ActiveCell.Offsets(0, 6) = strArtistIndex
ActiveCell.Offsets(1, 0).Select

Next j
Next iYear
End Sub

A2. Artist Information and Gender
Sub ArtistInfo()

Dim ie, objTable As Object
Dim strURL, strURL2, Identifier, strReleased, strBday As String
Dim y, iRe, iReStart, iReEnd As Integer

'************************** START **************************
Set ie = CreateObject("internetexplorer.application")
For i = 1 To 1729
Range("A" & i + 1).Select
strURL = "http://en.wikipedia.org"
strURL2 = ActiveCell

With ie
.Visible = False
.navigate strURL & strURL2
Do While .busy Or .ReadyState <> 4: DoEvents: Loop
Set objTable = ie.document.getElementsByTagName("table")
End With

On Error Resume Next
For y = 0 To objTable.Length
Identifier = objTable.Item(y).getAttribute("class")
If Identifier = "infobox vcard" Then
Exit For
End If
Next y

'************************** INFORMATION **************************
'Debug.Print InStr(ie.Document.body.innerHTML, "infobox vcard")
If InStr(ie.document.body.innerHTML, "infobox vcard") > 0 Then
    If InStr(objTable.Item(y).innerHTML, "rgb(240, 230, 140)") > 0 Then
        'Years active
        If InStr(objTable.Item(y).innerHTML, "Years active") > 0 Then
            'start
            iRe = InStr(objTable.Item(y).innerHTML, "Years active")
            iReStart = InStr(iRe, objTable.Item(y).innerHTML, "<td>")
            'end
            iReEnd = InStr(iReStart, objTable.Item(y).innerHTML, "</td>")
            strReleased = Mid(objTable.Item(y).innerHTML, iReStart + 4, iReEnd - iReStart - 4)
            ActiveCell.Offset(0, 3) = strReleased
        End If
    End If
    'Birthdate
    If InStr(objTable.Item(y).innerHTML, "bday") > 0 Then
        'start
        iRe = InStr(objTable.Item(y).innerHTML, "bday")
        iReStart = InStr(iRe, objTable.Item(y).innerHTML, ">")
        'end
        iReEnd = InStr(iReStart, objTable.Item(y).innerHTML, "</")
        strBday = Mid(objTable.Item(y).innerHTML, iReStart + 1, iReEnd - iReStart - 1)
        ActiveCell.Offset(0, 4) = strBday
    End If
    ActiveCell.Offset(0, 2) = 0
End If
Else
    If InStr(objTable.Item(y).innerHTML, "rgb(176, 196, 222)") > 0 Then
        ActiveCell.Offset(0, 2) = 1
        'group stuff
        'Years active
        If InStr(objTable.Item(y).innerHTML, "Years active") > 0 Then
            'start
            iRe = InStr(objTable.Item(y).innerHTML, "Years active")
            iReStart = InStr(iRe, objTable.Item(y).innerHTML, "<td>")
            'end
            iReEnd = InStr(iReStart, objTable.Item(y).innerHTML, "</td>")
            strReleased = Mid(objTable.Item(y).innerHTML, iReStart + 4, iReEnd - iReStart - 4)
            ActiveCell.Offset(0, 3) = strReleased
        End If
Else

ActiveCell.Offset(0, 2) = x
If InStr(objTable.Item(y).innerHTML, "Years active") > 0 Then

' start
iRe = InStr(objTable.Item(y).innerHTML, "Years active")
'iReStart = InStr(iRe, objTable.Item(y).innerHTML, 
"<td>"
)' end
iReEnd = InStr(iReStart, objTable.Item(y).innerHTML, ",</td>"
strReleased = Mid(objTable.Item(y).innerHTML, 
iReStart + 4, iReEnd - iReStart - 4)
ActiveCell.Offset(0, 3) = strReleased
End If
End If
Else
'skip
End If

'******************************************************************************* Gender *******************************************************************************
strall = ie.document.body.innerText

'MASCULINE
strWord = " he ", iCharCount = Len(Replace(strall, strWord, ""))
iHe = (Len(strall) - iCharCount) / Len(strWord)

strWord = " him ", iCharCount = Len(Replace(strall, strWord, ""))
iHim = (Len(strall) - iCharCount) / Len(strWord)

strWord = " his ", iCharCount = Len(Replace(strall, strWord, ""))
iHis = (Len(strall) - iCharCount) / Len(strWord)

'FEMININE
strWord = " she ", iCharCount = Len(Replace(strall, strWord, ""))
iShe = (Len(strall) - iCharCount) / Len(strWord)

strWord = " her ", iCharCount = Len(Replace(strall, strWord, ""))
iHer = (Len(strall) - iCharCount) / Len(strWord)

strWord = " hers "
iCharCount = Len(Replace(strall, strWord, ""))
iHers = (Len(strall) - iCharCount) / Len(strWord)

'PLURAL
strWord = " they "
iCharCount = Len(Replace(strall, strWord, ""))
iThey = (Len(strall) - iCharCount) / Len(strWord)

strWord = " their "
iCharCount = Len(Replace(strall, strWord, ""))
 iT heir = (Len(strall) - iCharCount) / Len(strWord)

strWord = " theirs "
iCharCount = Len(Replace(strall, strWord, ""))
i Theirs = (Len(strall) - iCharCount) / Len(strWord)

'Record
ActiveCell..Offset(0, 5) = iHe + iHim + iHis
ActiveCell. Offset(0, 6) = iShe + iHer + iHers
ActiveCell. Offset(0, 7) = iT hey + iT heir + iT heir s
Next i

ie.Quit

End Sub

A3 Song Information
Sub SongInfo()
Dim ie As Object

For i = 1 To 4300
Range("A" & i + 1).Select
If IsEmpty(ActiveCell) Then GoTo skippy
strURL = "wikipedia.org" & ActiveCell
Set ie = CreateObject("internetexplorer.application")
With ie
 .Visible = False
 .navigate strURL
 Do While .busy Or .readystate <> 4: DoEvents: Loop
 Set objHTML = .document.getelementsbytagname("table")
End With

For j = 0 To objHTML.Length - 1
 If objHTML.Item(j).classname = "infobox vevent" Then Exit For
 End If
Next j
If j = objHTML.Length Then GoTo skippy:

Set objTR = objHTML.Item(j).getelementsbytagname("tr")
For k = 0 To objTR.Length - 1
If InStr(LCase(objTR.Item(k).innerHTML), "length") > 0 Then
  Exit For
End If
Next k
If k = objTR.Length Then GoTo skippyl

Set objTD = objTR.Item(k).getelementsbytagname("td")
If InStr(objTD.Item(0).innerHTML, "span") > 0 Then
  strLength = objTD.Item(0).getelementsbytagname("span").Item(0).innerText
Else
  strLength = objTD.Item(0).innerText
  If Len(strLength) > 5 Then
    strLength = Left(strLength, 5)
  End If
End If
End If

ActiveCell.Offset(0, 2) = strLength
skippyl:
'***********************Genre
For m = 0 To objTR.Length - 1
  If InStr(LCase(objTR.Item(m).innerHTML), "genre") > 0 Then
    Exit For
  End If
Next m
If m = objTR.Length Then GoTo skippy

Set objA = objTR.Item(m).getelementsbytagname("td").Item(0).getelementsbytagname("a")
ReDim arrA(objA.Length)
For n = 0 To objA.Length - 1
  arrA(n + 1) = objA.Item(n).innerText
Next n
For o = 1 To UBound(arrA)
  ActiveCell.Offset(0, 2 + o) = arrA(o)
Next o

ie.Quit

skippy:
strLength = ""
Next i

End Sub

A4. Song Key
Sub BingSearch()

Dim ie As Object
Set ie = CreateObject("internetexplorer.application")
'i = 1
For i = 1 To 4300
    Range("C" & i + 1).Select
    strTitle = ActiveCell
    strArtist = ActiveCell.Offset(0, 1)
    With ie
        .Visible = False
        .navigate "http://www.bing.com/search?q=site%3Amusicnotes.com+" & strTitle & "+" & strArtist & "+sheet+music"
        Do While .readystate <> 4 Or .busy: DoEvents: Loop
        Set objUl = .document.body.getelementsbytagname("ul")
        End With
        For j = 0 To objUl.Length - 1
            If objUl.Item(j).ClassName = "sb_results" Then
                x = j
            End If
        Next j
        Set objA = objUl.Item(x).getelementsbytagname("a")
        If objA.Length = 0 Then
            GoTo nexter
        End If
        ActiveCell.Offset(0, 2) = objA.Item(0).getAttribute("href")
    Next i
End Sub

Sub Cleaner()
    Range("f2").Select
    For i = 1 To 4300
        If InStr(ActiveCell.Offset(0, -1), 
            "http://www.musicnotes.com/sheetmusic/mtd.asp") > 0 Then
            ActiveCell = ActiveCell.Offset(0, -1)
            ActiveCell.Offset(1, 0).Select
        Else
            ActiveCell.Offset(1, 0).Select
        End If
    Next i
End Sub

Sub KeyGet()
    Dim ie As Object
    Set ie = CreateObject("internetexplorer.application")
    For i = 1 To 4300
        Range("G" & i + 1).Select
If Not IsEmpty(ActiveCell.Offset(0, -1)) Then
    With ie
        .Visible = False
        .navigate ActiveCell.Offset(0, -1)
        Do While .readystate <> 4 Or .busy: DoEvents: Loop
        Set objHTML = .document.body.getelementsbytagname("li")
    End With

'Second Source
Set objTD = ie.document.body.getelementsbytagname("td")
For k = 0 To objTD.Length - 1
    If InStr(LCase(objTD.Item(k).innerText), "original published") > 0 Then
        y = k
    End If
Next k
ActiveCell.Offset(0, 1) = objTD.Item(y + 1).innerText
Else
    ActiveCell.Offset(0, 1) = ""
End If
Next i
ie.Quit
End Sub
Sub second()
    For j = 0 To objHTML.Length - 1
        If objHTML.Item(j).ClassName = "active" Then
            x = j
        End If
    Next j
    If Len(objHTML.Item(x).innerText) > 5 Then
        ActiveCell = objHTML.Item(x).innerText
    Else
        'Do nothing
    End If
End Sub
Sub spliter()
    For i = 1 To 4300
        Range("i" & i + 1).Select
        If IsEmpty(ActiveCell.Offset(0, -1)) Then GoTo nexter
        strKey = ActiveCell.Offset(0, -1)
        Splitter = Split(strKey, " ")
        ActiveCell = Splitter(0)
ActiveCell.Offset(0, 1) = Splitter(1)
nexter:
Next i
End Sub

A5. Artist Indexer
Sub metrolyrics()
Dim ie As Object

Set ie = CreateObject("internetexplorer.application")

For i = 1 To 4300
x = 0
y = 0
Range("E" & i + 1).Select
strA = ActiveCell.Offset(0, -1)
strT = ActiveCell.Offset(0, -2)

With ie
  .Visible = True
  Do While .readystate <> 4 Or .busy: DoEvents: Loop
  Set objHTML = .document.body.getelementsbytagname("ul")
End With

If InStr(ie.document.body.innertext, "No results found for") > 0 Then GoTo skippy

For j = 0 To objHTML.Length - 1
  If objHTML.Item(j).ClassName = "sb_results" Then
    x = j
  End If
Next j

If objHTML.Item(x).getelementsbytagname("h3").Length = 1 Then
  y = 0
Else
  For k = 0 To objHTML.Item(x).getelementsbytagname("h3").Length - 1
    If InStr(objHTML.Item(x).getelementsbytagname("h3").Item(k).getelementsbytagname("a").Item(0).getattribute("href"), "metrolyrics") > 0 And
       InStr(objHTML.Item(x).getelementsbytagname("h3").Item(k).getelementsbytagname("a").Item(0).innertext, "-") = 0 Then
      y = k
      Exit For
    End If
Next k
End If
strAlink =
objHTML.Item(x).getelementsbytagname("h3").Item(y).getelementsbytagname("a").Item(0).getAttribute("href")
    If InStr(strAlink, "metrolyrics.com") > 0 Then
        ActiveCell = strAlink
    Else
        End If
    
    If InStr(objHTML.Item(x).getelementsbytagname("h3").Item(y).getelementsbytagname("a").Item(0).innerText, "-") > 0 Then
        ActiveCell = "*"
        ActiveCell.Offset(0, 4) = strAlink
    End If

    If Len(strAlink) - InStr(12, strAlink, "/") - 12 < 0 Then GoTo skippy
    strA2 = Mid(strAlink, InStr(12, strAlink, "/") + 1, Len(strAlink) - InStr(12, strAlink, "/") - 12)
    'Debug.Print strA2
    strA2 = LCase(Replace(strA2, "-", ""))
    strA3 = LCase(Replace(strA, " ", ""))
    strA3 = LCase(Replace(strA3, ".", ""))
    
    strA4 = LCase(Replace(strA3, "the", ""))
    'Debug.Print strA2 & " " & strA3
    ActiveCell.Offset(0, 1) = LevenshteinDistance(strA2, strA3)
    ActiveCell.Offset(0, 2) = LevenshteinDistance(strA2, strA4)
    If LevenshteinDistance(strA2, strA3) >= LevenshteinDistance(strA2, strA4) Then
        iMin = LevenshteinDistance(strA2, strA4)
    Else
        iMin = LevenshteinDistance(strA2, strA3)
    End If
    ActiveCell.Offset(0, 3) = iMin
    iMin = 0
    strAlink = ""
    'Debug.Print
objHTML.Item(x).getelementsbytagname("h3").Item(y).getelementsbytagname("a").Item(0).innerText
    'Debug.Print
InStr(objHTML.Item(x).getelementsbytagname("h3").Item(y).getelementsbytagname("a").Item(0).innerText, "-")
    skippy:
ActiveWorkbook.Save
Next i
End Sub

A6. Lyric Indexer
Sub bArtistPage()
Dim ie As Object
Dim arrT
For i = 1 To 4300
On Error GoTo errHandler
label1:
Set ie = CreateObject("internetexplorer.application")
Range("G" & i + 1).Select
If IsEmpty(ActiveCell.Offset(0, -1)) _
Or ActiveCell.Offset(0, -1) = "+" _
Or ActiveCell.Offset(0, 1) > 4 Then GoTo skippy

strT = ActiveCell.Offset(0, -3)
strT2 = ActiveCell.Offset(0, -2)
strURL = ActiveCell.Offset(0, -1)

With ie
  .Visible = False
  .navigate strURL
  Do While .busy: DoEvents: Loop
  Set objHTML = .document.body.getelementsbytagname("table")
End With

For k = 0 To objHTML.Length - 1
  If InStr(objHTML.Item(k).ClassName, "deftbl") > 0 Then
    x = k
  End If
Next k

Set objA = objHTML.Item(x).getelementsbytagname("A")
iLength = objA.Length - 1
ReDim iLev(objA.Length - 1)

For j = 0 To objA.Length - 1
  iLev(j) = LevenshteinDistance(LCase(Left(objA.Item(j).innerText, Len(objA.Item(j).innerText) - 7)), LCase(strT))
Next j

iCount = iLev(0)
iCount2 = 0

For n = 1 To objA.Length - 5
  If iLev(n) < iCount Then
    iCount = iLev(n)
    iCount2 = n
  End If
Next n

ActiveCell = objA.Item(iCount2).getattribute("href")

skippy:
x = 0: y = 0: j = 0: k = 0: n = 0: iCount = 0

Set objA = Nothing
Set objHTML = Nothing

ie.Quit

errHandler:
   If Err.Number > 0 Then
      ie.Quit
      Set ie = Nothing
      Application.Wait Now + TimeValue("00:00:1")
      Resume label1:
   End If

Next i
End Sub
A7. Lyric Extraction

Sub cLyrics()
Dim ie As Object

For i = 1 To 4300
On Error GoTo errHandler
label:
Range("I" & i + 1).Select

If IsEmpty(ActiveCell.Offset(0, -2)) Then GoTo skippy
strURL = ActiveCell.Offset(0, -2)

If Left(strURL, 1) = "/" Then
strURL = "http://www.metrolyrics.com" & strURL
End If

Set ie = CreateObject("internetexplorer.application")

With ie
.Visible = False
.navigate strURL
    Do While .busy: DoEvents: Loop
    Set objHTML = .document.getElementById("lyrics-body").getElementsByTagName("span")
    "Do While .busy: DoEvents: Loop
End With

strLYRICS = ""

For x = 0 To objHTML.Length - 1
strLYRICS = strLYRICS & " <br> " & objHTML.Item(x).innerText
Next x

ActiveCell = strLYRICS

ie.Quit

errHandler:
    If Err.Number > 0 Then
        ie.Quit
        Set ie = Nothing
        Application.Wait Now + TimeValue("00:00:1")
        Resume label1:
    End If

skippy:
If i Mod 50 = 0 Then
    ActiveWorkbook.Save
End If

Next i
End Sub
A8. Lyric Profanity Search

Sub censored()
For i = 1 To 4300
Range("J" & i + 1).Select
If IsEmpty(ActiveCell.Offset(0, -1)) Then GoTo skippy
If InStr(ActiveCell.Offset, "*") > 0 Then ActiveCell = "*

skippy:
Next i
End Sub

Sub cleaner()
Dim arrPunct(14)
arrPunct(1) = ","
arrPunct(2) = "."
arrPunct(3) = "/"
arrPunct(4) = "?"
arrPunct(5) = ":"
arrPunct(6) = ">
arrPunct(7) = ";"
arrPunct(8) = ":"
arrPunct(9) = ":"
arrPunct(10) = ":"
arrPunct(11) = ":"
arrPunct(12) = ":"
arrPunct(13) = ":"
arrPunct(14) = ":"

For i = 1 To 4300
Range("J" & i + 1).Select
If IsEmpty(ActiveCell.Offset(0, -1)) Then GoTo skippy
ActiveCell = LCase(ActiveCell.Offset(0, -1))
For j = 1 To 14
ActiveCell = Replace(ActiveCell, arrPunct(j), " ")
Next j

skippy:
Next i
End Sub

Sub profanity()
' ass bitch cunt cock dyke faggot fuck nigger/a pussy shit slut tits

For i = 1 To 4300
Range("K" & i + 1).Select
If IsEmpty(ActiveCell.Offset(0, -1)) Then GoTo skippy

strLyrics = ActiveCell.Offset(0, -1)

'Ass ******************************************************* K
strAss = Replace(strLyrics, " ass ", "")
strAss2 = Replace(strLyrics, "a$$", "")
strAsshole = Replace(strLyrics, " asshole ", "")
iAss = ((Len(strLyrics) - Len(strAss)) / Len(" ass ")) +
(((Len(strLyrics) - Len(strAss2)) / Len("a$$")) +
(((Len(strLyrics) - Len(strAsshole)) / Len(" asshole "))) +
(((Len(strLyrics) - Len(strAssC)) / Len(" *ss")))
Range("K" & i + 1).Value = iAss

'Bitch ******************************************************* L
strBitch = Replace(strLyrics, "bitch", "")
strBitchC = Replace(strLyrics, "b*tch", "")
iBitch = ((Len(strLyrics) - Len(strBitch)) / Len("bitch")) +
(((Len(strLyrics) - Len(strBitchC)) / Len("b*tch")))
Range("L" & i + 1).Value = iBitch

'Cunt ******************************************************* M
strCunt = Replace(strLyrics, " cunt ", "")
strCuntC = Replace(strLyrics, "c*nt", "")
iCunt = ((Len(strLyrics) - Len(strCunt)) / Len("cunt")) +
(((Len(strLyrics) - Len(strCuntC)) / Len("c*nt")))
Range("M" & i + 1).Value = iCunt

'Cock ******************************************************* N
strCock = Replace(strLyrics, " cock ", "")
strCockC = Replace(strLyrics, "c*ck", "")
strCocks = Replace(strLyrics, " corks", "")
iCock = ((Len(strLyrics) - Len(strCock)) / Len(" cock ")) +
(((Len(strLyrics) - Len(strCockC)) / Len("c*ck")) +
(((Len(strLyrics) - Len(strCocks)) / Len(" corks")))
Range("N" & i + 1).Value = iCock

'Dyke ******************************************************* O
strDyke = Replace(strLyrics, "dyke", "")
iDyke = ((Len(strLyrics) - Len(strDyke)) / Len("dyke"))
Range("O" & i + 1).Value = iDyke

'Fag ******************************************************* P
strFag = Replace(strLyrics, " fag", "")
iFag = ((Len(strLyrics) - Len(strFag)) / Len(" fag"))
Range("F" & i + 1).Value = iFag

'Fag ******************************************************* Q
strFuck = Replace(strLyrics, "fuck", "")
strFuckC = Replace(strLyrics, " f*ck", "")
iFuck = ((Len(strLyrics) - Len(strFuck)) / Len("fuck")) +
(((Len(strLyrics) - Len(strFuckC)) / Len(" f*ck")))
Range("Q" & i + 1).Value = iFuck

'Nigg ******************************************************* R
strNigg = Replace(strLyrics, "nigg", "")
strNiggC = Replace(strLyrics, "n*gg", "")
iNigg = ((Len(strLyrics) - Len(strNigg)) / Len("nigg")) +
(((Len(strLyrics) - Len(strNiggC)) / Len("n*gg")))
Range("R" & i + 1).Value = iNigg

'Pussy ******************************************************* S
strPussy = Replace(strLyrics, "pussy", "")
strPussyC = Replace(strLyrics, "p*ssy", "")
iPussy = ((Len(strLyrics) - Len(strPussy)) / Len("pussy")) +
((Len(strLyrics) - Len(strPussyC)) / Len("p*ssy"))
Range("S" & i + 1).Value = iPussy
'Shit ******************************************************* T
strShit = Replace(strLyrics, "shit", "")
strShitC = Replace(strLyrics, "sh*t", "")
iShit = ((Len(strLyrics) - Len(strShit)) / Len("shit")) +
((Len(strLyrics) - Len(strShitC)) / Len("sh*t"))
Range("T" & i + 1).Value = iShit
'Slut ******************************************************* U
strSlut = Replace(strLyrics, " slut", "")
strSlutC = Replace(strLyrics, " sl*t", "")
iSlut = ((Len(strLyrics) - Len(strSlut)) / Len(" slut")) +
((Len(strLyrics) - Len(strSlutC)) / Len(" sl*t"))
Range("U" & i + 1).Value = iSlut
'tit ******************************************************* V
strTit = Replace(strLyrics, " tit", "")
strTitC = Replace(strLyrics, " t*t", "")
iTit = ((Len(strLyrics) - Len(strTit)) / Len(" tit")) +
((Len(strLyrics) - Len(strTitC)) / Len(" t*t"))
Range("V" & i + 1).Value = iTit

skippy:
Next i
End Sub

A9. Number-One Singles Indexer
Sub Singles()
Dim ie As Object
Range("G2").Select
Set ie = CreateObject("internetexplorer.application")

For iYear = 1970 To 2012
With ie
.Visible = False
Do While .readyState <> 4 Or .busy: DoEvents: Loop
Set objHTML = .document.getelementsbytagname("table")
End With

For j = 0 To objHTML.Length - 1
 If objHTML.Item(j).classname = "wikitable" Then Exit For
Next j

Set objTR = objHTML.Item(j).getelementsbytagname("tr")
iCount = 1
For i2 = 1 To objTR.Length - 1
Set objTD = objTR.Item(i2).getelementsbytagname("td")
For k = 0 To objTD.Length - 2
    RowSpan = objTD.Item(k).getAttribute("rowspan")
    Range(ActiveCell, ActiveCell.Offset(RowSpan - 1, 0)) = objTD.Item(k).innerText
    ActiveCell.Offset(0, 1).Select
Next k
ActiveCell.Offset(1, (objTD.Length - 1) * (-1)).Select
ActiveCell.Offset(-1, 4) = iYear
ActiveCell.Offset(-1, 5) = iCount
iCount = iCount + 1
Next i2
Next iYear
ie.Quit
End Sub

Sub weeksnum()
    Range("C2").Select
    Do Until IsEmpty(ActiveCell)
        If Not IsEmpty(ActiveCell.Offset(0, 4)) Then GoTo Skipp:
        iCount = 0
        strYes = ActiveCell
        For i = 1 To 52
            If ActiveCell = strYes Then iCount = iCount + 1
            ActiveCell.Offset(1, 0).Select
        Next i
        ActiveCell.Offset(-52, 0).Select

        For i = 1 To 52
            If ActiveCell = strYes Then ActiveCell.Offset(0, 4) = iCount
            ActiveCell.Offset(1, 0).Select
        Next i
        ActiveCell.Offset(-52, 0).Select
    Skipp:
    ActiveCell.Offset(1, 0).Select
    Loop
End Sub

A10. Movie Indexer
Sub ListMovies()
    'An excel web query table is created in a datasheet and the information
    'is dumped into a data sheet.

    For i = 1980 To 2012
Sheets("query").Select
Range("A1").Select
With Selection.QueryTable
 .Connection = "URL:http://boxofficemojo.com/yearly/chart/?yr=" & i
 .WebSelectionType = xlSpecifiedTables
 .WebFormatting = xlWebFormattingNone
 .WebTables = "7"
 .WebPreFormattedTextToColumns = True
 .WebConsecutiveDelimitersAsOne = True
 .WebSingleBlockTextImport = False
 .WebDisableDateRecognition = False
 .WebDisableRedirections = False
 .Refresh BackgroundQuery:=False
End With

Set rngA = Range("A3:H102")

Sheets("data").Select
rngA.Copy
ActiveSheet.Paste
Selection.End(xlDown).Select
ActiveCell.Offset(1, 0).Select
Sheets("query").Select

Next i

End Sub

Sub BOM_ExtractList()
 'Individual movie links are scraped using the movie title

Dim ie As Object
Dim strURL As String
Dim iYear As Integer

strURL = "http://boxofficemojo.com/yearly/chart/?yr="

iYear = 1980
For iYear = 1980 To 2012
    Set ie = CreateObject("internetexplorer.application")
    With ie
        .Visible = True
        .navigate strURL & iYear
        Do While .busy Or .ReadyState <> 4: DoEvents: Loop
        objHTML = ie.Document.body.innerHTML
    End With
    For iRank = 1 To 100
        Range("D" & (iYear - 1980) * 100 + iRank + 1).Select
        strSearch = ">" & ActiveCell.Offset(0, -1) & ">"
iSearch = InStr(objHTML, strSearch) - 75
If iSearch = -75 Then
    ActiveCell = "Error"
    GoTo error1
End If
iStart = InStr(iSearch, objHTML, "href") + 5
iEnd = InStr(iStart, objHTML, ">")
strHTML = Mid(objHTML, iStart, iEnd - iStart)
ActiveCell = strHTML

error1:
    Next iRank
ie.Quit
Next iYear
End Sub

A11. Movie Information
Sub IndMovieData()
    'Each individual movie is queried and information is taken from the site
    Dim ie As Object
    Dim strURL As String
    Dim iYear As Integer
    strURL = "http://boxofficemojo.com"
    For iYear = 1980 To 2012
        For iRank = 1 To 100
            Range("L" & (iYear - 1980) * 100 + iRank + 1).Select
            strSearch = ActiveCell
            Set ie = CreateObject("internetexplorer.application")
            With ie
                .Visible = False
                .navigate strURL & strSearch
                Do While .busy Or .ReadyState <> 4: DoEvents: Loop
                objHTML = ie.Document.body.innerHTML
            End With
            'Runtime
            iSearch = InStr(objHTML, "<!--------------------------Site")
            If iSearch = 0 Then
                ActiveCell = "Error"
                GoTo error1
            End If
    Next iRank
    ie.Quit
Next iYear
End Sub
End If
iStart = InStr(iSearch, LCase(objHTML), ">runtime") + 9
iEnd = InStr(iStart, LCase(objHTML), "</td>")
strHTML1 = Mid(objHTML, iStart, iEnd - iStart)
' Debug.Print strHTML

error1:

' MPAA
iSearch = InStr(objHTML, "<!--------------------------Site")
If iSearch = 0 Then
ActiveCell = "Error"
GoTo error2
End If
iStart = InStr(iSearch, LCase(objHTML), ">mpaa") + 14
iEnd = InStr(iStart, LCase(objHTML), "</td>")
strHTML2 = Mid(objHTML, iStart, iEnd - iStart)
' Debug.Print strHTML

error2:

' Genre
iSearch = InStr(objHTML, "<!--------------------------Site")
If iSearch = 0 Then
ActiveCell = "Error"
GoTo error3
End If
iStart = InStr(iSearch, LCase(objHTML), ">genre:") + 8
iEnd = InStr(iStart, LCase(objHTML), "</td>")
strHTML3 = Mid(objHTML, iStart, iEnd - iStart)
' Debug.Print strHTML

error3:

ActiveCell.Offset(0, 2) = strHTML1
ActiveCell.Offset(0, 3) = strHTML2
ActiveCell.Offset(0, 4) = strHTML3

ie.Quit

Next iRank

Next iYear

End Sub

A12. Subtitle Search
Sub gatherlinksSubs()
Dim ie As Object

For i = 1 To 3300
    Range("E" & i + 1).Select
    strTitle = ActiveCell.Offset(0, -2)
iYear = ActiveCell.Offset(0, -3)
If Not IsEmpty(ActiveCell.Offset(0, -1)) Then iYear = 
ActiveCell.Offset(0, -1)
If InStr(strTitle, "(") > 3 Then
strSearch = Left(strTitle, InStr(strTitle, "(") - 1)
Else
strSearch = strTitle
End If
Set ie = CreateObject("internetexplorer.application")
With ie
.Visible = False
.navigate strURL & strSearch
Do While .readystate <> 4 Or .busy: DoEvents: Loop
Set ObjHTML = .document.getelementsbytagname("div")
End With
For j = 0 To ObjHTML.Length - 1
If LCase(ObjHTML.Item(j).ClassName) = "bytitle" Then
x = j
Exit For
End If
Next j
Set objH2 = ObjHTML.Item(x).getelementsbytagname("h2")
For k = 0 To objH2.Length - 1
If LCase(objH2.Item(k).ClassName) = "exact" Then
strHREF = 
ObjHTML.Item(x).getelementsbytagname("ul").Item(0).getelementsbytagname
("a").Item(0).getAttribute("href")
strText = 
ObjHTML.Item(x).getelementsbytagname("ul").Item(0).getelementsbytagname
("a").Item(0).innerText
If InStr(strText, iYear) > 1 Then
GoTo skippy
Else
Exit For
End If
End If
Next k
'write the next step here
Set objDiv = ObjHTML.Item(x).getelementsbytagname("a")
iCount = 0
ReDim arrA(50)
For b = 0 To objDiv.Length - 1
If InStr(objDiv.Item(b).innerText, iYear) > 0 Then
iCount = iCount + 1
arrA(iCount) = b
End If
Next b
iLev = 100
For c = 1 To iCount
  Debug.Print LCase(objDiv.Item(0).innerText)
  If LD(LCase(objDiv.Item(arrA(c)).innerText), LCase(strSearch)) < iLev Then
    iLev = arrA(c)
  End If
Next c

If iLev = 100 Then GoTo skippy2
ActiveCell = objDiv.Item(iLev).innerText
ActiveCell.Offset(0, 1) = objDiv.Item(iLev).getAttribute("href")
If InStr(objDiv.Item(iLev).innerText, strSearch) > 0 Then
  ActiveCell.Offset(0, 2) = 0
Else
  ActiveCell.Offset(0, 2) = LD(strText, strSearch)
End If
GoTo skippy2

skippy:
ActiveCell.Offset(0, 1) = strHREF
ActiveCell = strText
If InStr(strText, strSearch) > 0 Then
  ActiveCell.Offset(0, 2) = 0
Else
  ActiveCell.Offset(0, 2) = LD(strText, strSearch)
End If

skippy2:

ie.Quit
Next i

End Sub

A13. Optimal Subtitle
Sub downloader()
Dim ie As Object
For i = 1 To 3300
  Range("D" & i + 1).Select
  If IsEmpty(ActiveCell.Offset(0, -1)) Then GoTo skippy2
  strURL = ActiveCell.Offset(0, -1)

  Set ie = CreateObject("internetexplorer.application")
  With ie
    .Visible = False
    .navigate "subscene.com" & strURL
    Do While .readystate <> 4 Or .busy: DoEvents: Loop
    Set objTable = .document.getElementsByTagName("table")
  End With

  Set ie = Nothing
  DoEvents
  If iLev = 100 Then GoTo skippy2

  Call downloader()
Next i
End Sub
Set objTd = objTable.Item(0).getelementsbytagname("td")

ReDim arrTd(50)
iCount = 0
For iTd = 0 To objTd.Length - 1
    If objTd.Item(iTd).ClassName = "a3" And InStr(objTd.Item(iTd).innerText, "1") > 0 Then
        iCount = iCount + 1
        If iCount = 50 Then Exit For
        arrTd(iCount) = iTd
    End If
Next iTd

For iDec = 1 To iCount
    If InStr(objTd.Item(arrTd(iDec) - 1).innerHTML, "positive-icon") > 0 Then
        strHREF = objTd.Item(arrTd(iDec) - 1).getelementsbytagname("a").Item(0).getAttribute("href")
        GoTo skippy
    End If
Next iDec

'ERRORS
For iEr = 0 To objTd.Length - 1
    If objTd.Item(iEr).ClassName = "a1" Then
        strHREF = objTd.Item(iEr).getelementsbytagname("a").Item(0).getAttribute("href")
        ActiveCell.Offset(0, 1) = "*"
        GoTo skippy
    End If
Next iEr

skippy:
ActiveCell = strHREF
strHREF = ""
ie.Quit

skippy2:
Next i

End Sub

Sub getdownloadlink()
Dim ie As Object
For i = 2950 To 3300
    Range("F" & i + 1).Select
    If IsEmpty(ActiveCell.Offset(0, -2)) Then GoTo skippy
    strURL = "subscene.com" & ActiveCell.Offset(0, -2)
    Set ie = CreateObject("internetexplorer.application")
    With ie

Sub download()
    Dim strURL As String
    Dim strFile As String

    For i = 1 To 3300
        Range("F" & i + 1).Select
        If IsEmpty(ActiveCell) Then GoTo skippy:
        strDir = ThisWorkbook.Path & "\subs2\"

        strURL = "http://subscene.com" & ActiveCell
        strFile = strDir & i & ".zip"

        Debug.Print DownloadFile(strURL, strFile)
        Debug.Print strFile
    Next i
End Sub

Sub Unziper()
    On Error GoTo rarer

    For i = 1 To 3300
        If Len(Dir("*DIRECTORY*\temp2\")) > 0 Then
            Kill "*DIRECTORY*\temp2\"
        End If

        If Len(Dir("*DIRECTORY*\subs" & i & ".zip")) = 0 And Len(Dir("*DIRECTORY*\subs" & i & ".rar")) = 0 Then GoTo skippy
        If Len(Dir("*DIRECTORY*\subs" & i & ".rar")) > 0 Then GoTo rarer
        Call UnZip("*DIRECTORY*\temp2", "*DIRECTORY*\subs" & i & ".zip")
        strOld = Dir("*DIRECTORY*\temp2\" & ")"

    Next i

    rarer:
        If Len(strOld) = 0 Then
            Name "*DIRECTORY*\subs" & i & ".zip" As "*DIRECTORY*\subs" & i & ".rar"
            GoTo skippy
        End If
End Sub
strExt = Right(strOld, 3)
strOld2 = "*DIRECTORY*\temp2\" & strOld
strNew = "*DIRECTORY*\ext2\" & i & "." & strExt
Name strOld2 As strNew

GoTo skippy
errHandler:

Debug.Print i & " " & Err.Number
Exit Sub

skippy:

If Len(Dir("*DIRECTORY*\temp2\*")) > 0 Then
Kill "*DIRECTORY*\temp2\*"
End If

strOld = ""
strNew = ""

Next i

End Sub

Sub UnZip(strTargetPath As String, strName As Variant)
Dim zipp As Object
Dim varFileFold As Variant

If Right(strTargetPath, 1) <> Application.PathSeparator Then
strTargetPath = strTargetPath & Application.PathSeparator
End If

varFileFold = strTargetPath
Set zipp = CreateObject("Shell.Application")
End Sub

A16. Subtitle Import
Sub subti()
Dim word As Object

Set word = CreateObject("word.application")

For iYear = 1980 To 2012
For iRank = 1 To 100
Sheets("working").Cells.Clear
i = (iYear - 1980) * 100 + iRank + 1
iSub = i - 1
If Len(Dir("*DIRECTORY*\ext2\" & iSub & ".sub")) = 0 Then GoTo skippy
Sheets("working").Select

Next i
Next i
End Sub
Range("A1").Select

With word
  .Visible = True
  .ChangeFileOpenDirectory "*DIRECTORY*\ext2"
  .Documents.Open Filename=iSub & ".sub"
  .Selection.WholeStory
  .Selection.Copy
End With

Sheets(iYear & "]).Select
Range("A1").Select
ActiveCell.Offset(0, iRank - 1).Select
ActiveSheet.Paste

word.ActiveWindow.Close

Application.ScreenUpdating = False
iCount = 1
Do Until IsEmpty(ActiveCell) And IsEmpty(ActiveCell.Offset(1, 0)) And IsEmpty(ActiveCell.Offset(2, 0))
  If IsError(ActiveCell) Then ActiveCell = 
  If IsError(ActiveCell.Offset(1, 0)) Then
  ActiveCell.Offset(1, 0) = 
  If IsError(ActiveCell.Offset(2, 0)) Then
  ActiveCell.Offset(2, 0) = 
  If InStr(ActiveCell, "-->") > 0 Then
    iCount = iCount + 1
    ActiveCell.Offset(1, 0) = LCase(ActiveCell.Offset(1, 0), "," , ","")
    ActiveCell.Offset(1, 0) = replace(ActiveCell.Offset(1, 0), "<i>", ","")
    ActiveCell.Offset(1, 0) = replace(ActiveCell.Offset(1, 0), "</i>", ","")
  If Not IsEmpty(ActiveCell.Offset(2, 0)) Then
  ActiveCell.Offset(2, 0) = LCase(ActiveCell.Offset(2, 0))
  ActiveCell.Offset(2, 0) = replace(ActiveCell.Offset(2, 0), "<i>", ","")
  ActiveCell.Offset(2, 0) = replace(ActiveCell.Offset(2, 0), "</i>", ","")
  End If
ActiveCell.EntireRow.Delete
ActiveCell.Offset(-1, 0).EntireRow.Delete

ActiveCell.Offset(1, 0).Select
  If IsEmpty(ActiveCell) Then ActiveCell.EntireRow.Delete
Else
  ActiveCell.Offset(1, 0).Select
  If IsEmpty(ActiveCell) Then ActiveCell.EntireRow.Delete
End If
If IsEmpty(ActiveCell) Then ActiveCell.EntireRow.Delete
Loop
Application.ScreenUpdating = True
strEnd = ActiveCell.Address(False, False)
Range("A1:" & strEnd).Select
Selection.Copy
Sheets(iYear & ")".Select
Range("A1").Select
ActiveCell.Offset(0, iRank - 1).Select
ActiveSheetPaste
Sheets("working").Select
skippy:
    Next iRank
    ActiveWorkbook.Save
Next iYear

End Sub

A17. Subtitle Profanity Search
Sub cleaner()
    Dim arrPunct(16) as String
    arrPunct(1) = ","
    arrPunct(2) = "."
    arrPunct(3) = "</i>"
    arrPunct(4) = "?"
    arrPunct(5) = "!
    arrPunct(6) = "<i>"
    arrPunct(7) = "</"
    arrPunct(8) = ">"
    arrPunct(9) = ";"
    arrPunct(10) = ":"
    arrPunct(11) = """
    arrPunct(12) = "_"
    arrPunct(13) = "{"
    arrPunct(14) = "}"
    arrPunct(15) = "{"
    arrPunct(16) = "}"
For iYear = 1980 To 2012
    For iRank = 1 To 100
        Sheets("" & iYear).Select
        Range("A1").Select
        ActiveCell.Offset(0, iRank - 1).Select

        If IsEmpty(ActiveCell) And IsEmpty(ActiveCell.Offset(1, 0)) And
        IsEmpty(ActiveCell.Offset(2, 0)) And IsEmpty(ActiveCell.Offset(3, 0))
        And IsEmpty(ActiveCell.Offset(4, 0)) Then GoTo skipRank

        iCount = 0
        Application.ScreenUpdating = False
        Do
            If IsError(ActiveCell) > 0 Then GoTo skippy

        Loop
        Application.ScreenUpdating = True
    Next iRank
Next iYear
End Sub
If InStr(ActiveCell, "-->") > 0 Then GoTo skippy
If IsEmpty(ActiveCell) Then GoTo EmptyCount
If Not IsEmpty(ActiveCell) Then iCount = 0
ActiveCell = LCase(ActiveCell)
For j = 1 To 16
    ActiveCell = Replace(ActiveCell, arrPunct(j), " ")
Next j

GoTo skippy

EmptyCount:
iCount = iCount + 1
skippy:

ActiveCell.Offset(1, 0).Select
If iCount > 5 Then Exit Do
Loop
Application.ScreenUpdating = True

skipRank:
    Next iRank
ActiveWorkbook.Save
skipYear:
    Next iYear
'MsgBox "done"

End Sub

Sub profanity()
'ass bitch cunt cock dyke faggot fuck nigger/a pussy shit slut tits

For iYear = 1980 To 2012
    For iRank = 1 To 100
        i = (iYear - 1980) * 100 + iRank + 1
        Sheets("" & iYear).Select
        Range("A1").Select
        ActiveCell.Offset(0, iRank - 1).Select

        iCount = 0
        iCount2 = 0
        Application.ScreenUpdating = False
        Do
            iCount2 = iCount2 + 1
            If IsEmpty(ActiveCell) Then
                iCount = iCount + 1
            GoTo skippers
        End If
        iCount = 0
        strLyrics = ActiveCell

        'Ass ******************************************************* D
        strAss = Replace(ActiveCell, " ass ", "")
        strAsshole = Replace(ActiveCell, " asshol", "")
        iAss = ((Len(strLyrics) - Len(strAss)) / Len(" ass ") +
            ((Len(strLyrics) - Len(strAsshole)) / Len(" asshol")))

        strLyrics = Replace(strLyrics, strAss,"")
        strLyrics = Replace(strLyrics, strAsshole,"")

        If InStr(ActiveCell, "-->") > 0 Then GoTo skippy
    Next iRank
Next iYear
'MsgBox "done"
End Sub
'Bitch ******************************* E
strBitch = Replace(strLyrics, "bitch", "")
iBitch = (Len(strLyrics) - Len(strBitch)) / Len("bitch")
'Cunt ********************************** M
strCunt = Replace(strLyrics, "cunt", "")
iCunt = (Len(strLyrics) - Len(strCunt)) / Len("cunt")
'Cock ********************************** N
strCock = Replace(strLyrics, "cock", "")
strCocks = Replace(strLyrics, "cocks", "")
iCock = (Len(strLyrics) - Len(strCock)) / Len(" cock ") +
((Len(strLyrics) - Len(strCocks)) / Len(" cocks"))
'Dyke ******************************* O
strDyke = Replace(strLyrics, "dyke", "")
iDyke = (Len(strLyrics) - Len(strDyke)) / Len("dyke")
'Fag ********************************** P
strFag = Replace(strLyrics, " fag", "")
iFag = (Len(strLyrics) - Len(strFag)) / Len(" fag")
'Fuck ********************************** Q
strFuck = Replace(strLyrics, "fuck", "")
iFuck = (Len(strLyrics) - Len(strFuck)) / Len("fuck")
'Nigger/Nigga *************************** R
strNigg = Replace(strLyrics, "nigg", "")
iNigg = (Len(strLyrics) - Len(strNigg)) / Len("nigg")
'Pussy ******************************* S
strPussy = Replace(strLyrics, "pussy", "")
iPussy = (Len(strLyrics) - Len(strPussy)) / Len("pussy")
'Shit ********************************** T
strShit = Replace(strLyrics, "shit", "")
iShit = (Len(strLyrics) - Len(strShit)) / Len("shit")
'Slut ********************************** U
strSlut = Replace(strLyrics, " slut", "")
iSlut = (Len(strLyrics) - Len(strSlut)) / Len(" slut")
'tit ********************************* V
strTit = Replace(strLyrics, " tit ", "")
strTits = Replace(strLyrics, " tits ", "")
iTit = (Len(strLyrics) - Len(strTit)) / Len(" tit ") +
((Len(strLyrics) - Len(strTits)) / Len(" tits "))

iAss2 = iAss2 + iAss
iBitch2 = iBitch2 + iBitch
iCunt2 = iCunt2 + iCunt
iCock2 = iCock2 + iCock
iDyke2 = iDyke2 + iDyke
iFag2 = iFag2 + iFag
iFuck2 = iFuck2 + iFuck
iNigger2 = iNigger2 + iNigger
iPussy2 = iPussy2 + iPussy
iShit2 = iShit2 + iShit
iSlut2 = iSlut2 + iSlut
iTit2 = iTit2 + iTit

skippers:
If iCount > 5 Then Exit Do
ActiveCell.Offset(1, 0).Select
Loop
Application.ScreenUpdating = True

If iCount2 < 10 Then GoTo skipp

skippy:

Sheets("sheet1").Select

Range("D" & i).Value = iAss2
Range("E" & i).Value = iBitch2
Range("F" & i).Value = iCunt2
Range("G" & i).Value = iCock2
Range("H" & i).Value = iDyke2
Range("I" & i).Value = iFag2
Range("J" & i).Value = iFuck2
Range("K" & i).Value = iNigger2
Range("L" & i).Value = iPussy2
Range("M" & i).Value = iShit2
Range("N" & i).Value = iSlut2
Range("O" & i).Value = iTit2

iAss2 = 0: iBitch2 = 0: iCunt2 = 0: iCock2 = 0: iDyke2 = 0: iFag2 = 0: iFuck2 = 0: iNigger2 = 0: iPussy2 = 0: iShit2 = 0: iSlut2 = 0: iTit2 = 0

skipp:

Next iRank
Next iYear
End Sub
Appendix B

Supplemental Tables and Figures

This appendix contains all the supplemental tables and figures not displayed in the thesis. Appendix Tables B.1 – B.9 represent the results from the differences-in-differences model for each of the select curse words and the five different metrics.

<table>
<thead>
<tr>
<th>Table B.1: Btch results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Media</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Y1995</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>MxY1995</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>cons</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>r2</td>
</tr>
<tr>
<td>AIC</td>
</tr>
</tbody>
</table>

As we can see from Table 1, we only find statistical significance in the differences-in-differences estimator for the average profanity, which is significant at the 5% level. While variables such as Y1995, which represents the period before and after the implementation of the label, and Media (takes values of 0 for movies and 1 for songs) are highly significant, our estimators suggest that there was no difference for the word *btch*. 
Table B.2: Fck results

<table>
<thead>
<tr>
<th></th>
<th>Saturation</th>
<th>Count</th>
<th>Per Min</th>
<th>Average</th>
<th>AvgPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>-0.266 ***</td>
<td>-637 ***</td>
<td>-332.31 ***</td>
<td>-10.472 ***</td>
<td>-5.458 ***</td>
</tr>
<tr>
<td></td>
<td>-13.38</td>
<td>-6.49</td>
<td>-6.7</td>
<td>-7.98</td>
<td>-8.27</td>
</tr>
<tr>
<td>Y1995</td>
<td>-0.143 ***</td>
<td>432.82 ***</td>
<td>246.364 ***</td>
<td>3.67 **</td>
<td>2.164 **</td>
</tr>
<tr>
<td></td>
<td>-5.63</td>
<td>2.96</td>
<td>3.21</td>
<td>2.06</td>
<td>2.39</td>
</tr>
<tr>
<td>MxY1995</td>
<td>-0.016</td>
<td>-397.88 ***</td>
<td>-239.54 ***</td>
<td>-3.244 *</td>
<td>-2.069 **</td>
</tr>
<tr>
<td></td>
<td>-0.43</td>
<td>-2.72</td>
<td>-3.12</td>
<td>-1.82</td>
<td>-2.28</td>
</tr>
<tr>
<td>cons</td>
<td>0.717 ***</td>
<td>640.062 ***</td>
<td>332.924 ***</td>
<td>10.532 ***</td>
<td>5.473 ***</td>
</tr>
<tr>
<td></td>
<td>52.2</td>
<td>6.53</td>
<td>6.72</td>
<td>8.03</td>
<td>8.29</td>
</tr>
<tr>
<td>r2</td>
<td>0.818</td>
<td>0.707</td>
<td>0.721</td>
<td>0.758</td>
<td>0.779</td>
</tr>
<tr>
<td>AIC</td>
<td>-148.434</td>
<td>943.092</td>
<td>858.214</td>
<td>361.109</td>
<td>271.664</td>
</tr>
</tbody>
</table>

The differences-in-differences estimators show that *fck* decreased in the music industry following the implementation of the label and it is significant at the 1% significance level for count and per minute count and significant that the 5% level for average count per minute (average profanity per song divided by song length).

Table B.3: Sht results

<table>
<thead>
<tr>
<th></th>
<th>Saturation</th>
<th>Count</th>
<th>Per Min</th>
<th>Average</th>
<th>AvgPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>-0.406 ***</td>
<td>-516.81 ***</td>
<td>-285.95 ***</td>
<td>-8.752 ***</td>
<td>-4.848 ***</td>
</tr>
<tr>
<td></td>
<td>-22.88</td>
<td>-10.81</td>
<td>-11.23</td>
<td>-15.31</td>
<td>-15.88</td>
</tr>
<tr>
<td>Y1995</td>
<td>-0.107 ***</td>
<td>188.364 ***</td>
<td>106.143 ***</td>
<td>0.652</td>
<td>0.374</td>
</tr>
<tr>
<td></td>
<td>-5.19</td>
<td>2.87</td>
<td>2.92</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td>MxY1995</td>
<td>-0.021</td>
<td>-129.83 *</td>
<td>-94.634 **</td>
<td>0.068</td>
<td>-0.208</td>
</tr>
<tr>
<td></td>
<td>-0.61</td>
<td>-1.96</td>
<td>-2.6</td>
<td>0.09</td>
<td>-0.52</td>
</tr>
<tr>
<td>cons</td>
<td>0.858 ***</td>
<td>521.813 ***</td>
<td>286.944 ***</td>
<td>8.852 ***</td>
<td>4.871 ***</td>
</tr>
<tr>
<td></td>
<td>91.75</td>
<td>10.94</td>
<td>11.27</td>
<td>15.58</td>
<td>15.98</td>
</tr>
<tr>
<td>r2</td>
<td>0.911</td>
<td>0.842</td>
<td>0.851</td>
<td>0.901</td>
<td>0.91</td>
</tr>
<tr>
<td>AIC</td>
<td>-160.288</td>
<td>837.836</td>
<td>759.542</td>
<td>244.382</td>
<td>162.589</td>
</tr>
</tbody>
</table>

*Sht* in music is found to have decreased relative to the movie industry at the 10% significance level for count and 5% significance level for per minute count. Other metrics seem to be statistically insignificant.
Unlike the previous results, we find that \textit{ngg} has actually increased following the PAL’s adoption. These values are significant at the 10% significance level.

Sum Curse represents the sum of all curse words in our data set. We see from Table 5 that all profanity variables bundled together has had a significant decrease at 5% for count and 1% for per minute counts. The remaining models are bundles of goods. B1 and B2 represent the combination of \textit{fck}, \textit{cnt}, \textit{sht}, and \textit{btch} excluding and including \textit{ngg} respectively to account for the increase in prevalence. B3 and B4 are the bundle of \textit{as}, \textit{dk}, \textit{fg}, \textit{pss}, and \textit{slt} excluding and including \textit{ngg} respectively.
Table B.6: B1 results

<table>
<thead>
<tr>
<th></th>
<th>Saturation Count Per Min Average AvgPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>-0.433 *** -1299.81 *** -698.84 *** -21.706 *** -11.673 ***</td>
</tr>
<tr>
<td></td>
<td>-24.24 -8.4 -8.79 -11.15 -11.77</td>
</tr>
<tr>
<td>Y1995</td>
<td>-0.096 *** 652.871 *** 370.736 *** 4.238 2.504 *</td>
</tr>
<tr>
<td></td>
<td>-5.05 2.99 3.18 1.67 1.91</td>
</tr>
<tr>
<td>MxY1995</td>
<td>-0.003 ** -347.82 *** -2.787 -2.178</td>
</tr>
<tr>
<td></td>
<td>-0.1 -2.44 -2.98 -1.09 -1.66</td>
</tr>
<tr>
<td>cons</td>
<td>0.888 *** 1309.188 *** 700.72 *** 21.891 *** 11.717 ***</td>
</tr>
<tr>
<td></td>
<td>101.68 8.47 8.81 11.26 11.82</td>
</tr>
<tr>
<td>r2</td>
<td>0.92 0.784 0.798 0.845 0.863</td>
</tr>
<tr>
<td>AIC</td>
<td>-164.532 996.147 913.289 407.78 319.874</td>
</tr>
</tbody>
</table>

Table B.7: B2 results

<table>
<thead>
<tr>
<th></th>
<th>Saturation Count Per Min Average AvgPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>-0.432 *** -1329.63 *** -715.8 *** -22.173 *** -11.94 ***</td>
</tr>
<tr>
<td></td>
<td>-23.84 -8.23 -8.62 -10.9 -11.54</td>
</tr>
<tr>
<td>Y1995</td>
<td>-0.093 *** 672.151 *** 379.202 *** 4.423 * 2.579 *</td>
</tr>
<tr>
<td></td>
<td>-4.92 2.98 3.16 1.68 1.91</td>
</tr>
<tr>
<td>MxY1995</td>
<td>0.009 ** -346.85 *** -2.344 -2.107</td>
</tr>
<tr>
<td></td>
<td>0.28 -2.23 -2.89 -0.88 -1.56</td>
</tr>
<tr>
<td>cons</td>
<td>0.889 *** 1342.438 *** 718.312 *** 22.424 *** 11.999 ***</td>
</tr>
<tr>
<td></td>
<td>102.93 8.32 8.66 11.05 11.6</td>
</tr>
<tr>
<td>r2</td>
<td>0.917 0.776 0.796 0.838 0.86</td>
</tr>
<tr>
<td>AIC</td>
<td>-163.892 1000.808 916.812 412.914 323.928</td>
</tr>
</tbody>
</table>

Table B.8: B3 results

<table>
<thead>
<tr>
<th></th>
<th>Saturation Count Per Min Average AvgPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>-0.365 *** -249.5 *** -139.56 *** -4.273 *** -2.391 ***</td>
</tr>
<tr>
<td></td>
<td>-20.78 -12.78 -12.96 -17.98 -17.76</td>
</tr>
<tr>
<td>Y1995</td>
<td>-0.092 *** 122.221 *** 69.481 *** 0.739 ** 0.429 **</td>
</tr>
<tr>
<td></td>
<td>-4.46 4.52 4.44 2.15 2.14</td>
</tr>
<tr>
<td>MxY1995</td>
<td>-0.081 ** -96.382 *** -63.882 *** -0.433 -0.353 *</td>
</tr>
<tr>
<td></td>
<td>-2.32 -3.47 -4.06 -1.23 -1.75</td>
</tr>
<tr>
<td>cons</td>
<td>0.817 *** 252.25 *** 139.952 *** 4.325 *** 2.4 ***</td>
</tr>
<tr>
<td></td>
<td>79.67 12.95 12.99 18.29 17.83</td>
</tr>
<tr>
<td>r2</td>
<td>0.904 0.89 0.893 0.914 0.914</td>
</tr>
<tr>
<td>AIC</td>
<td>-155.609 723.684 648.658 147.629 73.47</td>
</tr>
</tbody>
</table>
After bundling each of the subgroups together, we find much more significant results in the effect of the label on profanity. All bundles showed a decrease in per minute counts of profanity at the 1% significance level. For counts, B3 shows a significant decrease at the 1% level, B1 and B2 shows a significant decrease at the 5% level, and B4 shows a decrease at the 10% significance level.

<table>
<thead>
<tr>
<th></th>
<th>Saturation</th>
<th>Count</th>
<th>Per Min</th>
<th>Average</th>
<th>AvgPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>-0.364 ***</td>
<td>-279.31 ***</td>
<td>-156.52 ***</td>
<td>-4.74 ***</td>
<td>-2.658 ***</td>
</tr>
<tr>
<td>Y1995</td>
<td>-0.088 ***</td>
<td>141.5 ***</td>
<td>77.947 ***</td>
<td>0.924 **</td>
<td>0.504 **</td>
</tr>
<tr>
<td>MxY1995</td>
<td>-0.035</td>
<td>-66.923 *</td>
<td>-62.916 ***</td>
<td>0.01</td>
<td>-0.282</td>
</tr>
<tr>
<td>cons</td>
<td>0.819 ***</td>
<td>285.5 ***</td>
<td>157.545 ***</td>
<td>4.858 ***</td>
<td>2.682 ***</td>
</tr>
<tr>
<td>r2</td>
<td>0.888</td>
<td>0.836</td>
<td>0.867</td>
<td>0.866</td>
<td>0.89</td>
</tr>
<tr>
<td>AIC</td>
<td>-154.683</td>
<td>763.118</td>
<td>677.648</td>
<td>188.193</td>
<td>103.027</td>
</tr>
</tbody>
</table>
Figure B.1: Major Scale per Year

[Graphs showing the percentage of major scales per year for different keys such as C Major, D Major, E Major, F Major, G Major, and Ab Major.]

Figure B.2: Movie Genre per Year

[Graphs showing the percentage of different movie genres per year such as Comedy, Drama, Action, Thriller, Horror, Romance, Adventure, Family, and Sci-Fi.]
<table>
<thead>
<tr>
<th>Key</th>
<th>Percent</th>
<th>n</th>
<th>Sharps</th>
<th>Flats</th>
<th>Accidentals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Major</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Major</td>
<td>12.83</td>
<td>396</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G Major</td>
<td>11.86</td>
<td>366</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D Major</td>
<td>9.62</td>
<td>297</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>F Major</td>
<td>9.53</td>
<td>294</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A Major</td>
<td>6.9</td>
<td>213</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Eb Major</td>
<td>6.16</td>
<td>190</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>E Major</td>
<td>6.12</td>
<td>189</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Bb Major</td>
<td>5.41</td>
<td>167</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ab Major</td>
<td>2.75</td>
<td>85</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>B Major</td>
<td>1.72</td>
<td>53</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Db Major</td>
<td>1.33</td>
<td>41</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Gb Major</td>
<td>0.65</td>
<td>20</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>F# Major</td>
<td>0.58</td>
<td>18</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td><strong>Minor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Minor</td>
<td>4.67</td>
<td>144</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E Minor</td>
<td>3.56</td>
<td>110</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D Minor</td>
<td>2.69</td>
<td>83</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C Minor</td>
<td>2.66</td>
<td>82</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>G Minor</td>
<td>2.46</td>
<td>76</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>F Minor</td>
<td>2.01</td>
<td>62</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>B Minor</td>
<td>1.91</td>
<td>59</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>C# Minor</td>
<td>1.49</td>
<td>46</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>F# Minor</td>
<td>1.43</td>
<td>44</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Bb Minor</td>
<td>0.68</td>
<td>21</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Eb Minor</td>
<td>0.58</td>
<td>18</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>G# Minor</td>
<td>0.39</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

$n = 3086$

Song key, otherwise more commonly known as pitch, can either be scaled up or down by transposing to different keys. Because I collected the information from the website musicnotes.com, there is no guarantee that the published key is in the same key as the song. Generally, we would expect the key of C Major (or a minor) to be the most prevalent key since it is the easiest to learn on a piano given that there are no sharps or
flats (the black keys). It is also quite possible that performers also perform more songs in these keys because they are much easier to play for all instruments. The minor data is more interesting as it does not follow the same pattern as the major scale. This could be that more songs written in minor keys are denoted by flats rather than sharps when writing the music. Either way, while the pitch might change, the scale will always stay major or minor, as the scale is dependent on the actual notes in the piece. A minor scale depends upon the tonic, the minor third, and the perfect fifth.
References


