High-Grossing Movie Sentiments

DNSC 6211: Programming for Analytics

Kelly Berdelle
Jason Houghton
Yuebo Li
Qinya Wang
Gaoshuang Zhu

Abstract

Our project focuses on the highest-grossing US films, adjusted to today's dollar. If the performance of a movie could predict future perception, producers could make more informed long-term decisions. We scraped the film data to find what these films were and to collect related data, conducted a sentiment analysis on these movie titles using twitter data, and then analyzed the data to see if there was a relationship between high-grossing films and sentiment. In doing so, we found no relationship between the amount of money a movie earned and its twitter sentiment, but we did uncover a relationship which helps explain how the adjusted gross revenue value was calculated for each movie.

Contents

1 Introduction **3**

2 Background **3**

3 Method **3**

4 Organization **3**
4.1 Workflow 4
4.2 Project structure 4
4.3 Figures and Tables 5

5 Discussion **6**
5.1 Learnings 6
5.2 Challenges 7

6 Conclusion **7**

1 Introduction

The topic of our project is an analysis of the highest-grossing US movies adjusted for inflation to 2016 dollars, as reported by Box Office Mojo. We want to see if there is a relationship between the movies that earned the most money at time of release and current sentiments on twitter. Because we are gathering twitter data, and many of these movies were released before twitter existed, this will help us to analyze if movies that were financially successful at time of release are still viewed favorably today.

2 Background

Our initial project idea was to analyze heavy rail systems in the United States. We wanted to see if there was a relationship between ridership or budget and sentiment of various train systems, such as the DC Metrorail or the New York Subway. However, once we started looking for a data set, we decided that none of the available data sets allowed us to perform the desired analysis.

Next, we decided to see if the financial performance of a movie is an indicator of future sentiment of that movie. If the performance of a movie could predict future perception, producers could make more informed long-term decisions. A common website for movie information is IMDb, and this was the first website we looked to scrape for data. However, IMDb did not provide the adjusted gross revenue of movies, which would not allow us to analyze movies over the years. Instead, IMDb provides a score for each movie based on user preferences.

We finally landed on the Box Office Mojo website, which contains a table of the information we were looking for: the top 200 US movies by adjusted gross. Our initial concern was that Box Office Mojo did not provide the method by which it calculated adjusted gross, but we went ahead with this data set since it provided the data we wanted to analyze.

We originally wanted to include a viewer rating score; however, the rating provided by IMDb is not a moment-in-time sentiment, but rather an average of user preferences over time. Therefore, we limited our analysis to the data found in Box Office Mojo and twitter.

3 Method

Overall, we looked for a relationship between Adjusted Gross and Sentiment for the top 200 highest-grossing movies. Once we scraped the data from Box Office Mojo and calculated a sentiment score for each movie based on tweets, we looked for relationships using single linear regression and multiple linear regression. Once we did this, however, we found that there was no relationship between Adjusted Gross and Sentiment Score, contrary to our hypothesis. Then, we began exploring other potential relationships in our data, but continued to find no relationships of significance. For example, we analyzed our data to see if a relationship existed between Studio and Sentiment, but once again found that no such relationship exists. Finally, we analyzed the relationship between Year, Unadjusted Gross, and Adjusted Gross in order to see if we could uncover Box Office Mojo's method of calculating Adjusted Gross. Lastly, we used an interactive web application so that any user could create a graph of any one or two of our variables.

4 Organization

Jason and Gaoshuang worked on web scraping, and Jason cleaned the data. Yuebo and Jason worked on gathering the tweets and conducting the sentiment analysis. Kelly created some

initial plots and regression models, and then Qinya created the interactive web application, with minor changes from Yuebo. Kelly composed the initial draft of the documentation, and together the group finalized the documentation.

4.1 Workflow

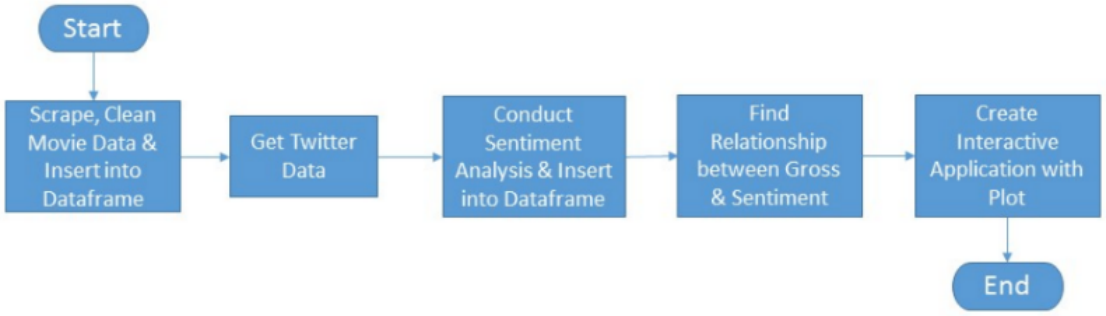

Figure 1: Final project workflow

To complete this project, first we scraped data from Box Office Mojo. While scraping, we also cleaned the data by removing special characters and making a few adjustments for movies from 2016. Simultaneously, we inserted the data into a pandas dataframe. Then, we gathered tweets, conducted a twitter sentiment analysis, and put the sentiment score for each movie into our dataframe. We exported our data into a CSV for analysis in R. Then, we looked for single and multiple linear regression relationships in the data and disproved our hypothesis. Additionally, we plotted the data to visually represent the relationships between our variables. Lastly, we created an interactive web application so that any user could create plots to analyze one-way distributions and multi-way relationships in our data.

4.2 Project structure

We had two very different data sources. The first, a website called Box Office Mojo had data listed in a table format. The values listed for each movie were Ranking, Title, Studio, Adjusted Gross, Unadjusted Gross, and Year. In order to collect this data, we used web scraping. We also had to clean this data due to several issues. First, many of the year values had a caret, indicating that it had multiple theatrical releases. We removed this caret. Then, movies that were released in 2016 had several issues. First, the year was incorrect; it was listed as the year of the movie directly preceding it in ranking. Additionally, the Unadjusted Gross was listed as "2016," when in reality the Unadjusted Gross should have been equivalent to the Adjusted Gross. We replaced the Unadjusted Gross value with the Adjusted Gross value for these rows, and changed the Year value to 2016. Finally, for all values listed in dollar amounts, we removed dollar signs and commas.

Our second data source was twitter, which we used to gather tweets in order to conduct a sentiment analysis. Together, these data sources were put into a single dataframe. Once we had a single dataframe, we were able to compare all of our variables to see if there was a relationship between sentiment score and revenue.

4.3 Figures and Tables

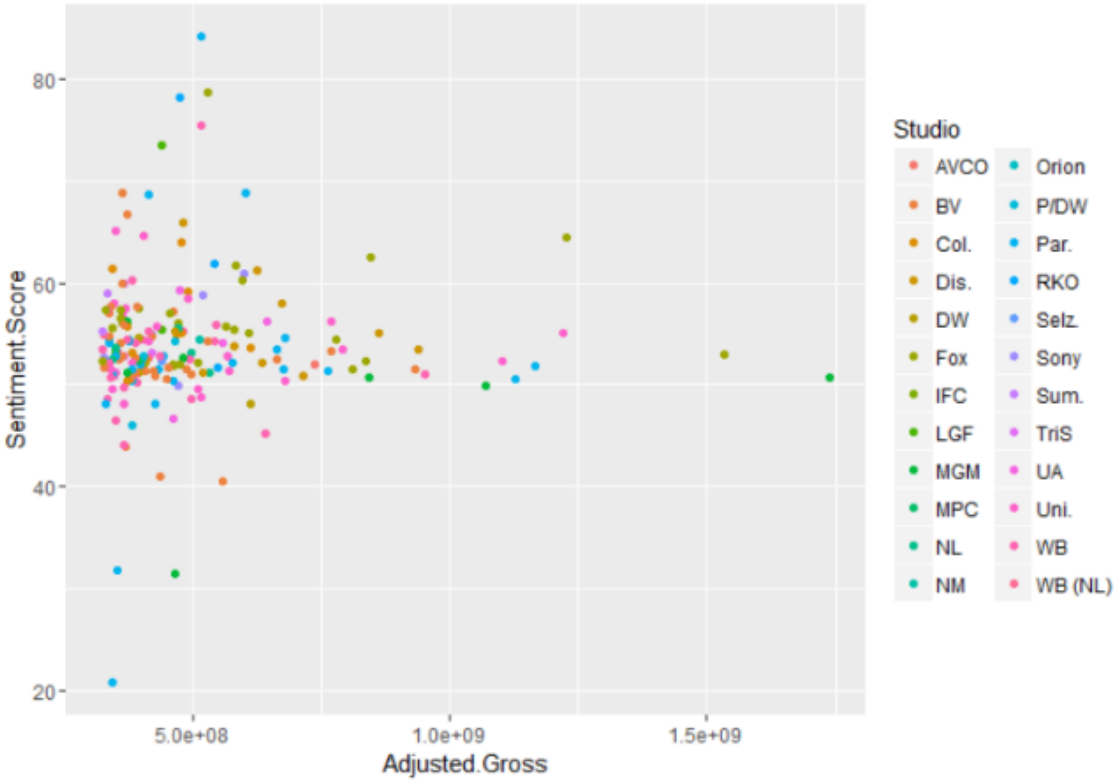

Figure 2: Sentiment Score vs Adjusted Gross

Our graph of Adjusted Gross versus Sentiment Score (figure 2) shows a lack of a linear relationship. Additionally, we can see that Studio has no effect on the distribution of Adjusted Gross or Sentiment Score. This figure allows us to visually see the relationship between Adjusted Gross, Sentiment Score, and Studio in a manner that is easily digested.

Box Office Mojo's calculation of Adjusted Gross is related to Year and Unadjusted Gross, as shown by the existence of a linear relationship between the variables (figure 3), but must also contain additional parameters in order to fully explain the variability in the Adjusted Gross values. These four plots provide an understanding of the multiple linear regression relationship. Specifically, the Residuals vs Fitted graph allows us to see that a linear relationship exists, but shows evidence of heteroskedasticity.

The number of movies per decade in our data set increases as time progresses (figure 4). However, Sentiment Score and Adjusted Gross do not increase with Year. Instead, there is a slight negative linear relationship between Adjusted Gross and Year (figure 5).

Additionally, a studio that has more movies in our data set does not indicate higher Sentiment Scores for each movie (figure 6). Lions Gate Entertainment and PKO seem to have produced movies with higher Sentiment Scores. However, the number of movies produced by

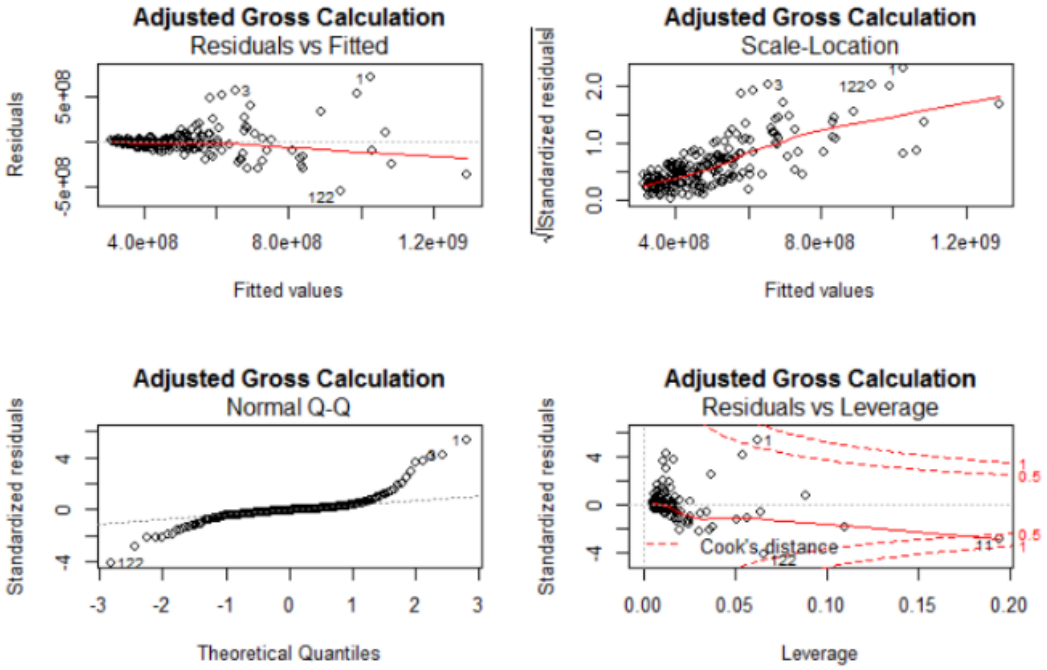

Figure 3: Calculation of Adjusted Gross

these studios are relatively low compared to other studios, thus we cannot conclude that a studio that has more films on the list produces films with higher Sentiment Scores.

5 Discussion

This project allowed us to learn two things. First, there is no linear relationship between the adjusted gross revenue of a movie and its sentiment score. Second, the calculation of adjusted gross revenue, according to Box Office Mojo, has more factors than simply the year of a movie's release and its unadjusted gross revenue. Although we were not able to prove our hypothesis, it is useful for movie producers to know that box office performance may not be an accurate predictor of future sentiment.

5.1 Learnings

We enjoyed having the opportunity to work with a topic of interest: movies. While we felt that we had a strong hypothesis, that adjusted gross and sentiment score would be related, we found that such a temporal relationship did not exist. However, it was a good experience to see our hypothesis disproved firsthand by analyzing linear relationships between multiple variables. We found it useful to reapply previous course exercises by modifying the code to fit the needs of this project. In doing so, we better understand the process by which we may repurpose existing code to fit new future challenges. We had also hoped to include mapping in this project through the use of an additional data source containing the geographic location of the

movie's production, but we decided that such an analysis would not be useful to determining any type of conclusion in line with our hypothesis.

5.2 Challenges

Through this project, we learned that websites with simpler HTML are trickier to scrape. Because the HTML is structured differently than Yelp, we needed new methods to scrape the data. Our new method required us to search less structured code. This project was our first experience in this course when the data we had scraped needed to be corrected, so we had to figure out how to handle each unique case.

The way our tweepy code is structured, our sentiment analysis required an extensive amount of time to run. In the future, we would make use of tweepy's streaming functionality for a shorter processing time.

We found working with Shiny tricky because the interactive web application we created was significantly different from example Shiny code available on the web. The syntax proved to be more difficult because its structure was unlike other Python or R code we had previously used.

6 Conclusion

Our project focused on data from the top 200 highest-grossing movies in US history, adjusted for inflation to 2016 dollars. From this project, we have two main conclusions. First, we did not have sufficient evidence to conclude that adjusted gross and sentiment score have a linear relationship; in fact, the data shows that there is truly no linear relationship whatsoever. Therefore, movie producers should not use box office performance as an indicator of future sentiment. Secondly, we concluded that Box Office Mojo's method of calculating adjusted gross revenue certainly included year and unadjusted gross, but it also must have included additional variables. While this project did not uncover any groundbreaking models, it allowed our group to understand how to structure future research using methods learned in this course. We were able to develop a problem statement, find and explore data, apply coding lessons from this semester to real-world data sets, and create meaningful outputs to share our findings.

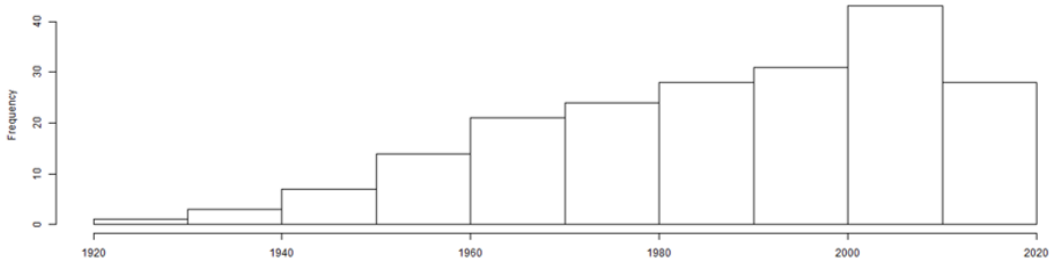

Figure 4: Number of movies in each decade

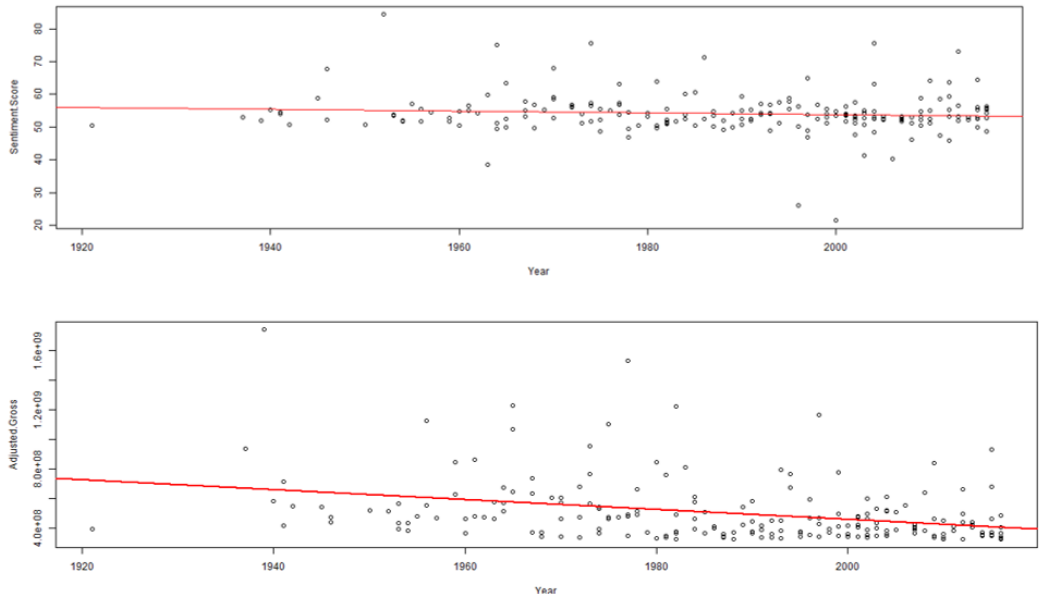

Figure 5: Sentiment Score vs Year, Adjusted Gross vs Year

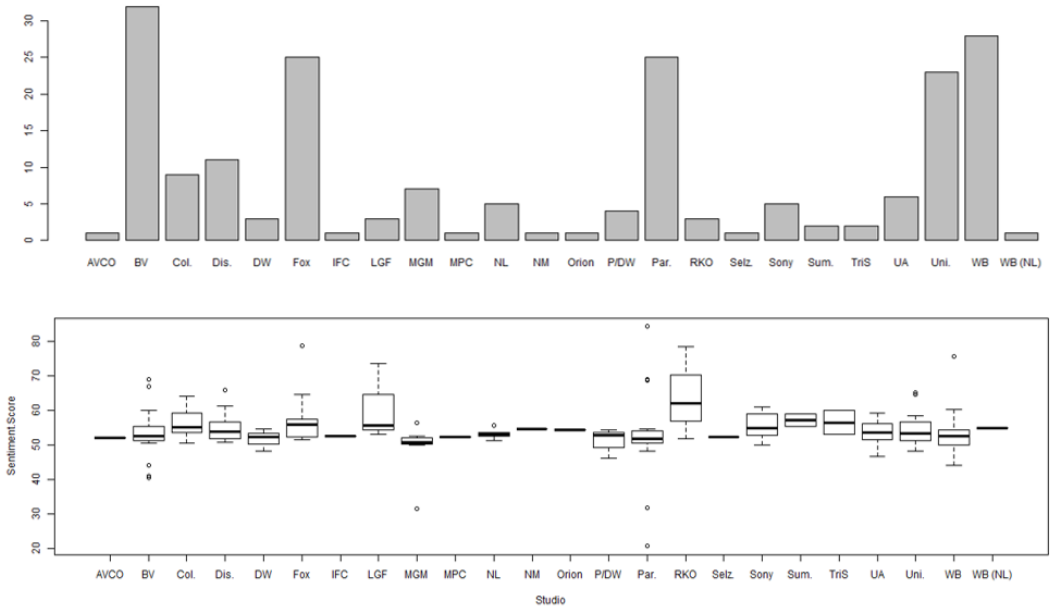

Figure 6: Frequency of Movies by Studio, Sentiment Score by Studio