Movie Review Metrics vs Box Office Gross Values

DNSC 6211: Programming for Analytics

Abhinav Chandel
Tingting Ju
Xinyi Wang
Yunning Zhu
Daniel Chen

Abstract

For graduate students, going to the movies is perhaps the most popular pastime. The goal of this project is to examine the relationship between movie reviews and box office success through a data-driven methodology, which involves scrapping websites to collect data via Python, and running regressions on the collected data and displaying the results graphically via R. The population includes the top 100 grossing movies worldwide from 2015 and their corresponding revenues (dependent variables), review ratings, sentiment scores, and production budgets (independent variables). The resulting regressions not only answered our question but also revealed an insight that we were not expecting. There is indeed a positive linear relationship between revenues and review ratings plus sentiment scores. However, the strength of the relationship is extremely low, and there is a confounding effect in play that is difficult to interpret. We then included production budgets as an additional independent variable. This course of action led to the surprising discovery of production budget being a powerful predictor of box office revenues. The regression result between revenues and review ratings plus production budget shows not only global statistical significance, but also high strength in relationship. Based on our analysis, to generate box office successes, movie studios should track trending comments and review scores when they plan for productions but most importantly, they should not be frugal with their budgets.

Contents

1 Introduction

For full-time graduate students, going to the movies is perhaps the most available and low-cost pastime option. We often check movie ratings on popular review sites such as Rotten Tomato, IMDB, Metacritic, etc. to help us to decide whether or not we should go to see a particular movie. Therefore, it is natural to expect that higher-rated movies generate more moviegoers, and consequently, more box office successes. But is that truly the case? We decided to conduct our project by examining the relationship between movie review metrics and box office gross values. More specifically, we decided to focus on the top 100 grossing movies in 2015 and their corresponding movie reviews. The problem we identified is as follows: How much explanatory power does movie review metrics have over box office success, in terms of worldwide revenues?

2 Background

Our storyline was expected to be: The combination of positive Twitter Sentiment Scores, IMDB Ratings, Rotten Tomato Ratings, and Metacritic Ratings has a positive effect on Worldwide Revenues. Our storyline was driven by the following data sources:

1. Revenue-Related Data:
   (a) Box Office Mojo (programming file name is BudgetRevenue.ipynb)
2. Budget-Related Data:
   (a) The Numbers Movie Budget (programming file name is BudgetRevenue.ipynb)
3. Review-Related Data:
   (a) Rotten Tomato: ratings by audience (programming file name is RottenTomatoes.ipynb)
   (b) IMDB: ratings by audience (programming file name is IMDBMetacritic.ipynb)
   (c) Metacritic: ratings by critics (programming file name is IMDBMetacritic.ipynb)
   (d) Twitter: commentaries from general populace (programming file name is TwitterSentiment.ipynb

We scrapped 2015s top 100 grossing movies and their associated revenues off Box Office Mojo utilizing the Python package beautifulsoup and pandas. Similar programming was done to extract data from IMDB and Metacritic. We obtained Twitter commentaries via API and the Python package tweepy, and translated the commentaries into sentiment scores using the Python package textblob. With Rotten Tomato, we succeeded in taking a different approach, using the Python package selenium to automatically select movies and extract scores. Some of the same movie names have different spellings across different data sources. So, we used the Python package difflib to do a similarity match with a high cutoff.

During our research, we encountered a website called The Number Movie Budget that publishes data related to movie budgets. We decided to write another program to extract the data and add it to our regression as an additional independent variable. We had to abandon only one data source, Facebook, due to time constraint and lack of experience with Facebooks API.

Moreover, part of our original plan was to build a database to consolidate all data points. We saw little relevance of doing so after deciding to do regression and visualization in R. It was much easier process-wise to output data, transfer data, and retrieve data through intermediate csv files.

Finally, we employed the min max methodology to standardize all of our variables to a 0-to-1 range. The normalization was done in R with the formula: Zi = [Xi min(X)] / [max(X) min(X)].

3 Method

The resulting multiple regression between the dependent variable worldwide revenues and the independent variables Twitter sentiment scores, IMDB ratings, Rotten Tomato ratings, and Metacritic ratings turned out to be statistically significant, but weak in terms of explanatory power. Yes, we succeeded in answering our question that there is indeed a positive linear relationship between movie ratings and box office success. But the answer was semi-satisfactory.

Next, we ran all scenarios of simple linear regression and multiple regression (see our script Regression.R). Viewed side-by-side, the bivariate and multivariate regressions revealed a confounding effect in play. For example, the beta coefficients of the independent variables IMDB ratings and Rotten Tomato ratings at the multivariate level appeared much lower than the ones at the bivariate level. The interplay between the bivariate and multivariate regressions requires further interpretation and a deeper statistical knowledge that we do not possess at this point of our study.

Next, we introduced the independent variable production budget into the mix. The resulting multiple regression showed not only global statistical significance but also strength in relationship. At the individual P-value level, however, only the independent variables Rotten Tomato scores and production budgets exceeded the 95

Therefore, we conducted the multiple regression between worldwide revenues and Rotten Tomato scores plus production budgets. The results showed global-level statistical significance, individual-level statistical significance, and strength in relationship. We found the regression model with the best goodness-of-fit.

4 Organization

Our division of labor was fluid we shifted the workloads around the team to fit our individual schedules as we went through the project. We all contributed to the project idea formulation and data gathering. Dan extracted the revenue-related data; Xinyi, Yunning, and TingTing created the sentiment scores; Abhinav obtained the Rotten Tomato, IMDB, and Metacritic scores; Abhinav and Tingting managed the overall data consolidation. Then Abhinav and Dan built regressions; TingTing, Xinyi, and Yunning produced the visual presentations and shiny applications; and Dan wrote the report. Overall, we all contributed a great deal to the project.

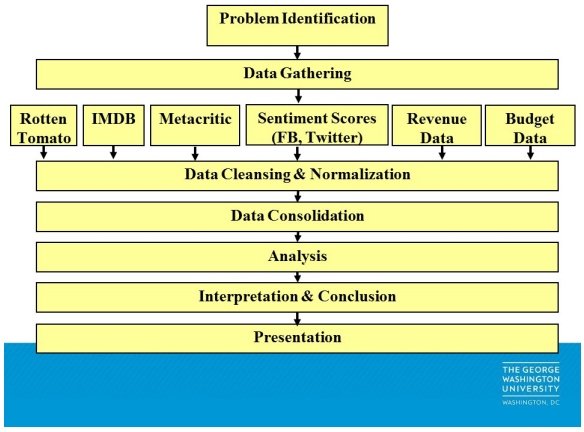

Figure 1: Our projected workflow

4.1 Workflow

1. Data Gathering: Scrap revenue and review data off web in Python
2. Data Transformation: Create sentiment scores in Python
3. Data Cleansing: Ensured clean DataFrames that can easily be output into intermediate csv files
4. Data Normalization: Applied min max method to standardize data
5. Analysis: Build simple linear regressions / multiple regressions in R
6. Presentation: Displayed results via visualization (ggplot and matplotlib) and interactive web applications (shiny)

4.2 Project structure

1. Revenue-Related Data:
   (a) Box Office Mojo (programming file name = BudgetRevenue.ipynb)
2. Budget-Related Data:
   (a) The Numbers Movie Budget (programming file name = BudgetRevenue.ipynb)
3. Review-Related Data:
   (a) Rotten Tomato: ratings by audience (programming file name = RottenTomatoes.ipynb)
   (b) IMDB: ratings by audience (programming file name = IMDBMetacritic.ipynb)
   (c) Metacritic: ratings by critics (programming file name = IMDBMetacritic.ipynb)
   (d) Twitter: commentaries from general populace converted into sentiment scores (programming file name = TwitterSentiment.ipynb)

Again, we scrapped 2015s top 100 grossing movies and their associated revenues off Box Office Mojo utilizing the Python package beautifulsoup and pandas. Similar programming was done to extract data from IMDB and Metacritic. We obtained Twitter commentaries via API and the Python package tweepy, and translated the commentaries into sentiment scores

using the Python package textblob. With Rotten Tomato, we succeeded in taking a different approach, using the Python package selenium to automatically select movies and extract scores. Some of the same movie names have different spellings across different data sources. So, we used the Python package difflib to do a similarity match with a high cutoff.

During our research, we encountered a website called The Number Movie Budget that publishes data related to movie budgets. We decided to write another program to extract the data and add it to our regression as an additional independent variable.

4.3 Figures and Tables

See below.

5 Discussion

We often hear people asking is this a good movie? and people answering what does Rotten Tomato say? The best selling point of our project is that we proved the relevance of movie ratings (particularly Rotten Tomato scores!) in terms of aiding box office success. With a population size of only 100 observations, we showed that there is indeed a positive linear relationship between movie ratings (particularly Rotten Tomato scores) and box office success. Movie production companies should pay attention to movie rating sites.

The other selling point of our project is that we also uncovered the strong positive linear relationship between movie production budget and box office success. Costly films tend to become blockbuster movies.

The final selling point of our project is that we demonstrated the ability to consolidate and standardize numerous data sources. The data community could see how we successfully used different techniques to gather and cleanse data from various sources.

5.1 Learnings

The better moments came at the end of the successful data gathering, cleansing, and consolidation phase. The team was overjoyed to finally be able to build regressions, see the results, and plot them graphically.

5.2 Challenges

Most of the difficult moments happened during the data gathering, cleansing, and consolidation phase. The work was tedious and time consuming. During the Twitter sentiment score development process, we kept getting the timeout status code of 429. To get around the error, we had to extract the scores in pieces.

6 Conclusion

With a global P-value of 0 and coefficient of determination of 0.60, there is a statistically significant and strong linear relationship between movie ratings, production budgets, and box office revenues. More specifically, Rotten Tomato scores and production budgets have the best explanatory powers over box office revenues. The multiple regression model among the 3 variables yielded a global P-value of 0, coefficient of determination of 0.60, and individual P-values that are all under 0.05. It is statistically significant at both the multivariate and single-variable levels, and it explains about 60 percent of the variation in box office revenues. Based on our analysis, to generate box office successes, movie studios should track trending comments and review scores when they plan for productions but most importantly, they should not be frugal with their budgets. With more time allowed, we would like to conduct the same analysis for each movie genre and/or movie production company and explore further insights.

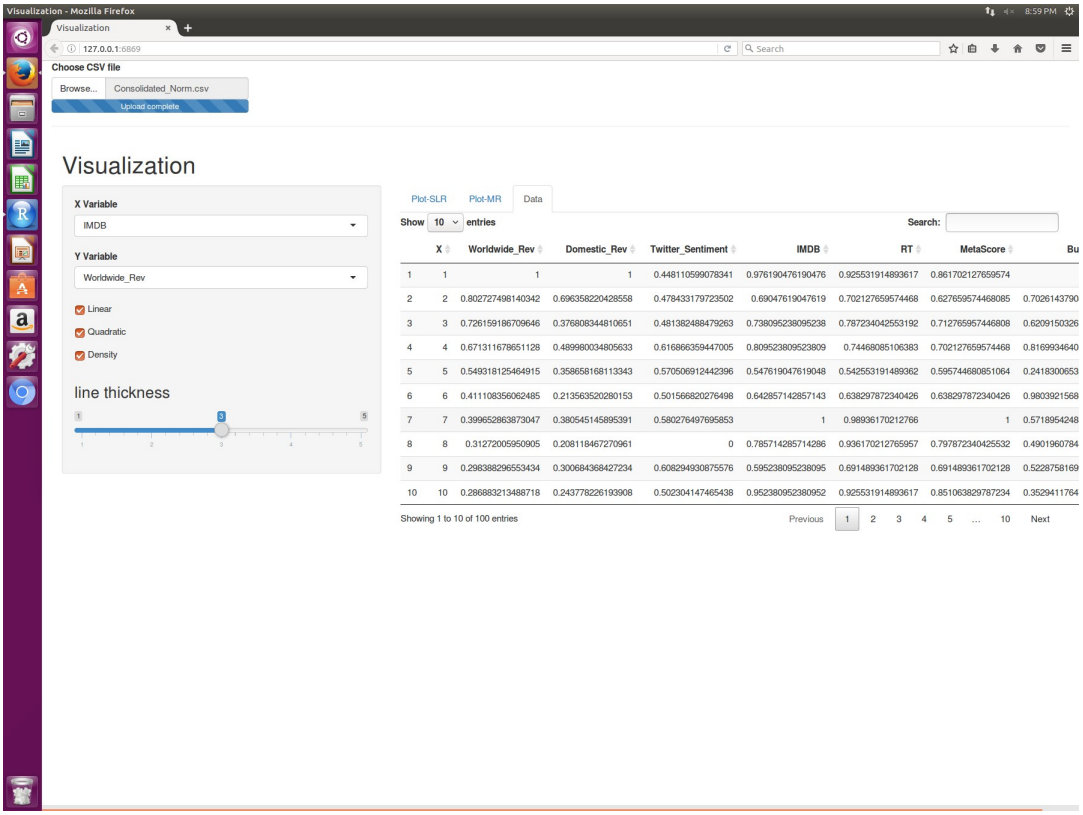

Figure 2: Interactive Data Tab in Shiny

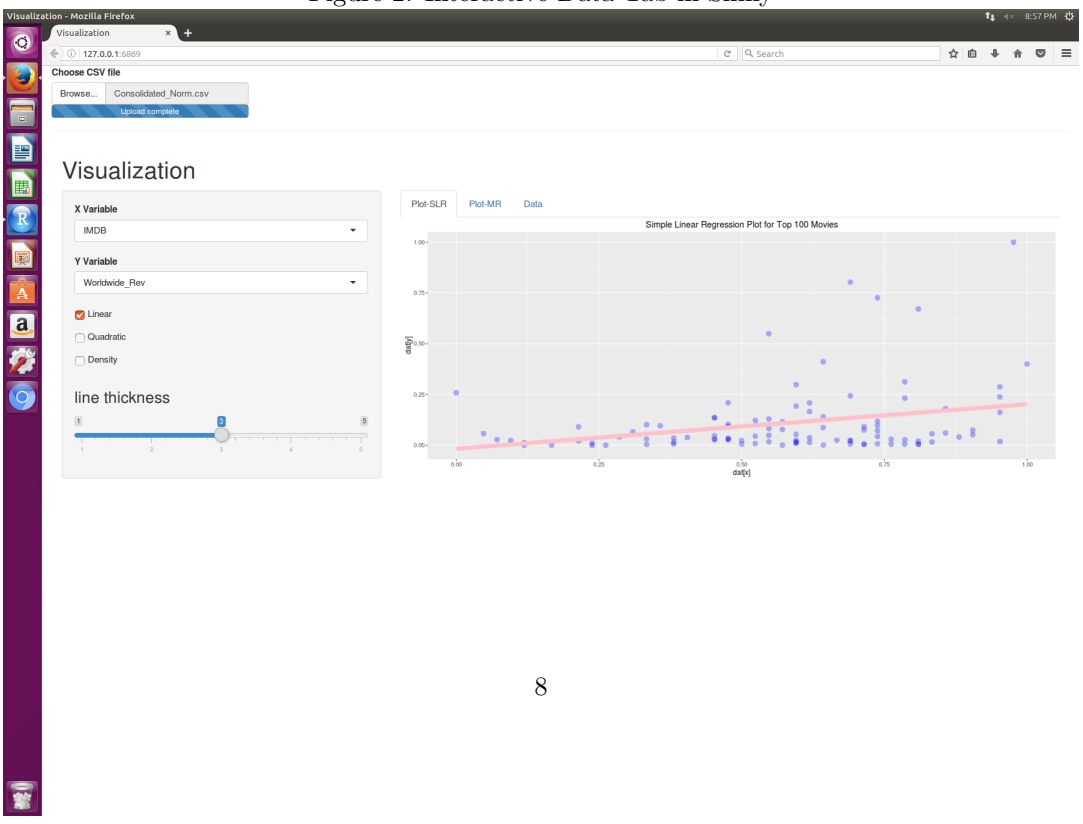

Figure 3: Interactive SLR Tab in Shiny - Linear Line

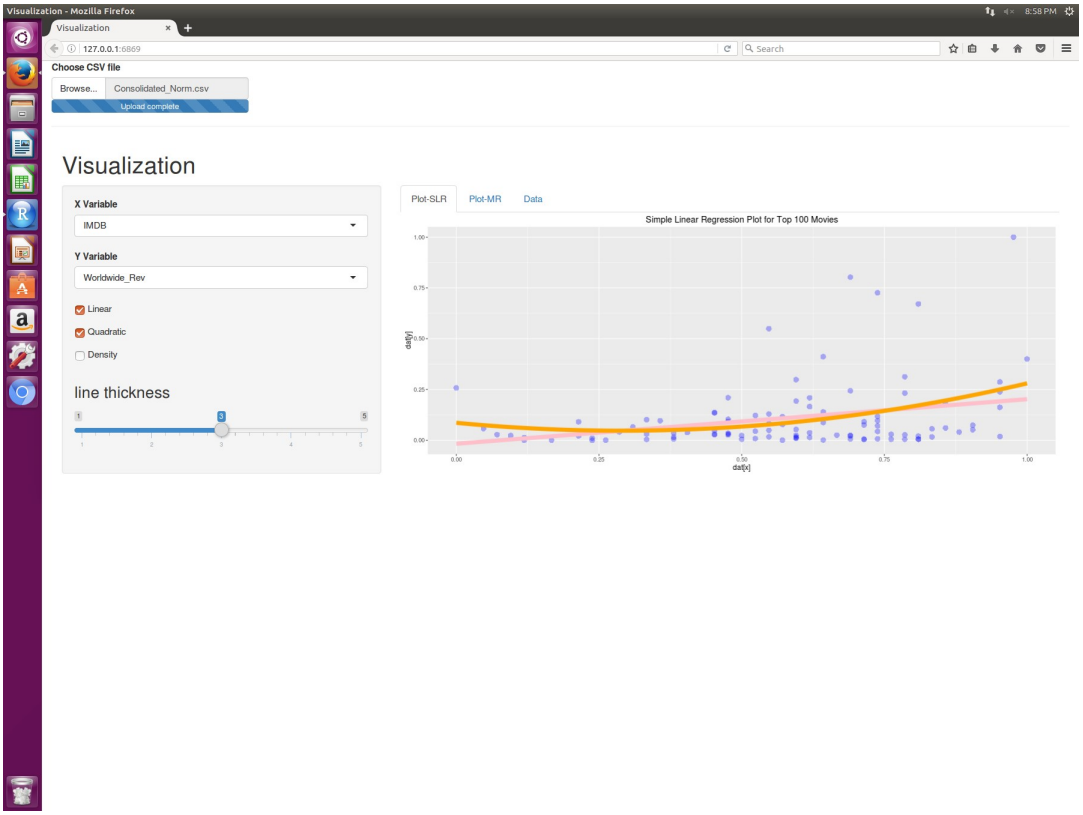

Figure 4: Interactive SLR Tab in Shiny - Quadratic Line

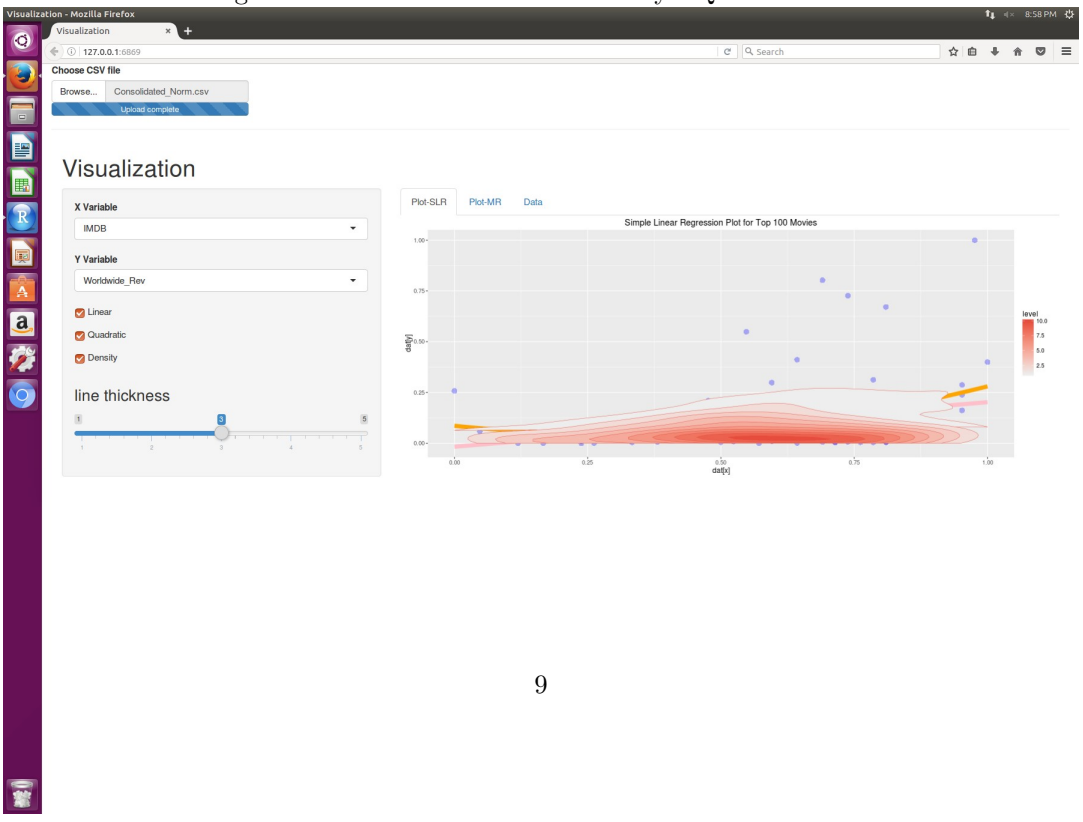

Figure 5: Interactive SLR Tab in Shiny - Density