**Movie Review Metrics vs Box Office Gross Values**

**Abhinav Chandel**

**Tingting Ju**

**Xinyi Wang**

**Yunning Zhu**

**Daniel Chen**

Abstract

*Describe your project within 200 words. One way is answer the following questions regarding your project: (a) what did you do? (b) why did you choose do that? (c) how did you go about doing your project (d) what did you ﬁnd out?, and ﬁnally (e) What did you ﬁnd out? The content of your abstract and the outline and contents of your report may vary according to the needs of your speciﬁc research topic.*

Contents

1 Introduction 3

2 Background 3

3 Method 3

4 Organization 3

4.1 Workﬂow . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

4.2 Project structure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

4.3 Figures and Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

5 Discussion 4

5.1 Learnings . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

5.2 Challenges . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

6 Bullets and numbered lists (FYI, delete in your report) 4

6.1 Bulleted and Numbered Lists . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

7 Conclusion

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 are likely to 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 look at all the movie reviews from 2015 and see if the top-rated ones actually ended up being one of the top 100 grossing movies of 2015. The problem we identified is: How much predictive power does movie review metrics have over box office success (in terms of both worldwide and domestic revenues)?

1. **Background**

The expected storyline would be something like: “Rotten Tomato Tomatometer ratings are good predictors of box office grosses.” or “Metacritic ratings are not good predictors of box office grosses.” Or “The combination of Rotten Tomato Tomatometer and Metacritic ratings has the best predictive power of box office grosses.” The underlying data sources that drive our storyline would be as follows:

* Revenue-Related Data:
  + Box Office Mojo
* Budget-Related Data:
  + The Numbers – Movie Budget
* Review-Related Data:
  + Rotten Tomato: ratings by audience
  + IMDB: ratings by audience
  + Metacritic: ratings by critics
  + Twitter: commentaries from general populace
  + Facebook: commentaries from general populace

Most of the data could be obtained via web-scrapping. Data from Twitter and/or Facebook would require the use of API. And they would have to be further cleaned for “quality metrics” that include feelings and satisfaction ratings, which would then be translated into sentiment scores. After cleaning up all the datasets, we would like to load all of them into a central database via SQL.

1. **Method**

Our proposed methodology is as follows:

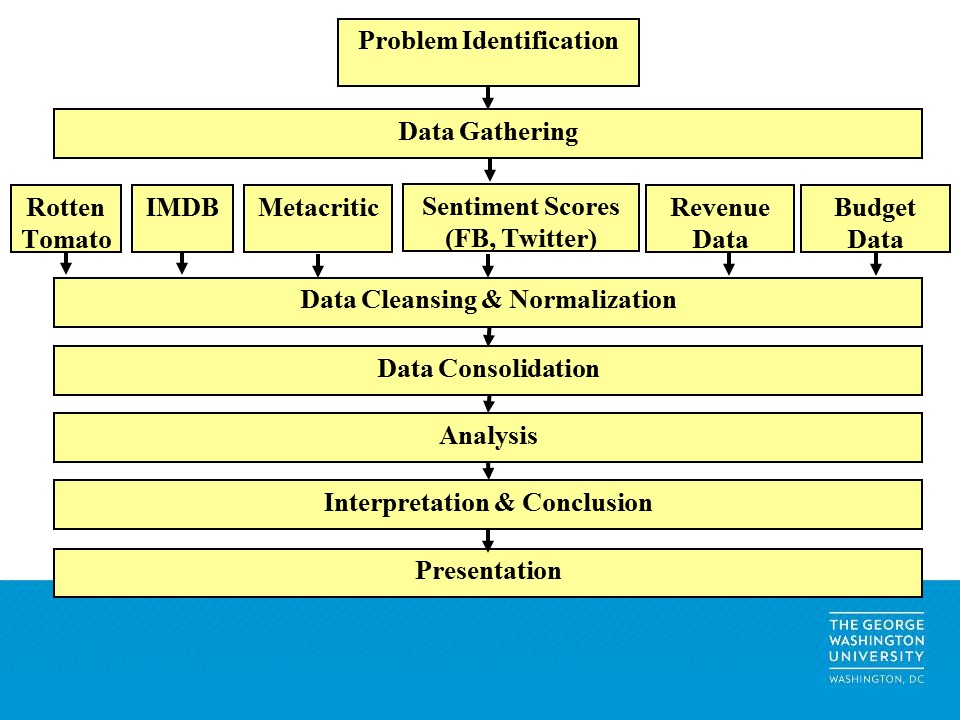
* 1. Data Gathering: Scrap revenue & review data off web in Python
  2. Data Transformation: Create sentiment scores in Python; normalize & standardize rating scores in Python
  3. Data Cleansing: Ensure clean DataFrames that can easily be output into csv files
  4. Data Consolidation: Load all datasets into one database via MySQL
  5. Analysis: Build simple linear regressions / multiple regressions in R
  6. Presentation: Display results via visualization (ggplot & matplotlib) and interactive web applications (shiny)

There are some foreseeable concerns:

1. Sentiment Analysis: None of our group members has any experience with Facebook’s API and cannot guarantee successful data downloads from Facebook.
2. Time Constraint: Our project scope is ambitious with numerous data sources. We fear that we will not have ample time to complete our project.
3. Regression Results: What if none of our regression models turns out to be statistically significant?
4. **Organization**

Our division of labor is fluid – we will shift the workloads around the team to fit our individual schedules as we go through the project. Currently, our division of labor is as follows:

* Abhinav: Project idea formulation, web scrapping (ratings), sentiment scores
* Yunning: Project idea formulation, web scrapping (ratings), sentiment scores
* Xinyi: Project idea formulation, web scrapping (ratings), sentiment scores
* TingTing: Project idea formulation, web scrapping (misc.), initial data consolidation & regression
* Daniel: Project idea formulation, web scrapping (revenues), report/presentation slide write-ups
  1. **Workflow**

****

* 1. **Project Structure**
  2. **Figures and Tables**

1. **Discussion**
   1. **Learnings**
   2. **Challenges**
2. **Bullets and Numbered Lists**
3. **Conclusion**

**Overall, we are comfortable with the data sources and on track to complete the assignment.**