The Team

Betsy Shane

Adrienne Lewis

Zachy Girmay

Problem Statement a.k.a Decompose The Ask

We are trying to ascertain if there is a relationship between the most anticipated movies of a year prior to release, and the movies that are ultimately rated as the best movies of that given year.

Identify Data Sources

*Describe the data source(s) you will use. Make sure you have access to the data you want to use \*in the quantity and quality you need\*. Ideally you should get all the data in your hands before starting your project.*

*Describe how you plan to obtain the data, or how you got it if you already have it.*

We plan on using a most anticipated movies of the year list from Screenrant and a compilation of top 10 movie rankings from a variety of websites and film critics. The 2017 compilation movie data will be pulled from a Kaggle dataset that collected various top 10 rankings of 2017 movies from different online sources and critics. That data will then be compared to a list of the most anticipated movies of 2017 as published by Screenrant.

Define Strategy and Metrics

*List some goals of your analysis, ideally in the form of testable hypothesis, or via well-defined success metrics. These can be tentative, and you don’t need to stick to them throughout your project. Again, since you haven’t done any exploratory analysis yet, you might assume that the data has structure that it doesn’t, and you might not have seen other interesting patterns in the data. But you should always approach the data with some expectations so that your efforts are focused.*

We’ll try to discover if there is a positive correlation between a movie being highly anticipated prior to release, and critically acclaimed at the end of the year by comparing the film’s anticipating ranking to an aggregated ranking score based on the compiled top 10 lists for the best movies of 2017.

Description of Data Analysis Tools You Plan to Use

*Describe the tools you plan to use throughout the project. As you might expect, there will be several stages in the project. Make sure to use* ***at minimum*** *the required tools (see Project1 power point).*

We plan on web scraping the ScreenRant list of the most anticipated movies of 2017 to gather the titles and ranks of the films on that list. That will then be compared to a downloaded dataset of aggregated top 10 film rankings which will imported into pandas for cleaning. The cleaned data will then be stored in a SQL database.

The webpage that we plan on scraping data from: <http://www.screenrant.com/most-anticipated-movies-2017/>

Describe the Data Products Your Project Will Produce

*Data products include results of statistical tests, visualizations of the data, websites/apps, or model parameters. What will be the tangible evidence of your discovered insights?*

We’re providing a dataframe that will include the most anticipated films of 2017 and a weighted ranking score for the aggregated top 10 ranking data, which will allow us to see how the most anticipated movies of 2017 performed critically.