Taylor Imhof

Bellevue University | DSC 540

Project Milestone #1

1/30/2021

**Project Subject Area**

For my project, I plan to analyze movie data to gain insights into what factors contribute most to the movie’s success.

**Identify Datasets**

1. **Flat File**: tab-separated values (.tsv) data files provided from IMDB’s

Link: [IMDb interfaces](https://datasets.imdbws.com/)

* title.basics.tsv
  + provides a lot of general information about movies, TV shows, video games, etc.
* title.ratings.tsv
  + provides average IMDb rating as well as total votes casted for each title
  + can be merged with title.basics on the unique ID value (‘tconst’)

1. **Website:** Movie Production Budgets

Link: [Movie Budget – The Numbers](https://www.the-numbers.com/movie/budgets/all)

* provides a lot of useful information regarding movie finances (production budget, domestic and worldwide gross)
* also provides release date and movie title data
* should be able to join with flat file data frame on unique movie title column values
* will need to find way to scrape multiple pages (different links) to build a more robust data frame
  + provides thousands of observations, but only 500 per page

1. **API:** IMDb’s API

Link: [IMDb Developer Portal](https://developer.imdb.com/)

* provided via AWS Data Exchange
* put in request for API product access
* plan to use extra information to create additional columns in final data frame for more robust analysis

**Relationships**

I should be able to connect/merge each of the resulting data frames via their respective movie titles. The .tsv files from IMDb Interface provide a unique ID column, but that will only be useful for combining the two flat files I plan to use for this project.

**Project Overview**

I am a huge movie nerd and can remember enjoying films since I was a wee lad. Over the last several years, I have grown interested in the box office returns for movies and TV shows and what factors contribute most to their overall success (or failure). While there are general factors that lead to success (e.g., Marvel movies will almost always be very profitable), I thought it would be fun and a good test of my newfound data analytics skills to see if I can identify other contributing factors.

***What makes a movie successful?***

The best metric for determining whether a movie is successful will be the total box office returns. However, I also felt it would be important to consider the movie budget, so I plan on creating a ratio (or some weighted metric) of the budget vs. box-office revenue data from the website above.

***Plan of attack!***

The first step will be extracting and cleaning the data from the three different data sources. The flat files (.tsv) should be straightforward to load and begin cleaning. I took a quick peek at the data types from the first .tsv file, and they are all encoded as string values, so some cleaning efforts will be required. I also plan on merging the reviews file, so I will have to make sure the column IDs match. Then, I will perform some preliminary exploratory data analysis to get a feel for how the data is distributed.

For the website data (*The Numbers*), I plan to use the BeatifulSoup library to scrape the data from the HTML tables. The trickiest part will be finding a way to scrape multiple pages, as the website only provides 500 observations per page. I will likely have some data cleaning/formatting work after getting the bs4 objects converted into a Pandas data frame. Then, I plan on merging it with the data frame created from the flat file data sources.

As I am still waiting to be approved access to IMDb’s API via AWS Data Exchange, I am not exactly sure what type of data will be available for each request. Therefore, I might have to find a different API endpoint that is easier to access for this project. However, I ultimately want additional columns to add to a final data frame to find more insights potentially.

Ultimately, I hope to refine further my data preparation and analytical skills. I am becoming more comfortable with the tasks required for modern data science endeavors with each successive course in this program.