Project Milestone 1

Harlan Wittlieff

Data Science, Bellevue University

DSC 540: Data Preparation

Catherine Williams

January 9, 2022

Project Subject Area

This project will focus on data surrounding Netflix. Movie titles added will be explored and their impact on subscriber numbers and revenue analyzed.

Data Sources

- Flat File: Netflix Movies and TV Shows
    - This file contains data describing movies in the Netflix library, their director, cast, release date, date added to Netflix, rating, duration, what category their fall into, and a description of the movie.
    - https://www.kaggle.com/shivamb/netflix-shows
- API: IMDB API
    - IMDB's API contains a wide variety of information on all movies. Primarily for this analysis movie titles, rating information, and gross revenue will be used.
    - https://imdb-api.com/api#Title-header
- Website:
    - Wikipedia's Netflix page contains information on subscriber count and total revenue by year.
    - https://en.wikipedia.org/wiki/Netflix

Relationships

Netflix Movies and TV Shows contains data by title in addition to dates that titles were added to Netflix.

IMDB API data contains title data.

Wikipedia's data contains dates.

The Netflix Movies and TV Shows file will be related to IMDB's API via the titles. This data will then be summarized by dates and combined with Wikipedia's data.

Project Summary

For this project, I elected to focus on data surrounding one of my passions. As an avid movie enthusiast and Netflix subscriber, I became interested in understanding how movie performance may motivate Netflix subscriptions. To tackle these questions, the Netflix catalog will be analyzed to discover any impacts it may have on the streaming service's subscriber counts. Information that will be analyzed includes IMDB rating, gross revenue, and release date.

Multiple questions will be explored throughout this analysis. Do higher ranked movies increase subscriber count? Do newer released movies lead to more subscribers? What relationship exists

between gross revenue and subscriber count? Could a combination of factors contribute to the subscriber numbers? These questions should lead to a determination of whether correlation and causation exists relative to subscriber number impacts.

Potential ethical implications of this project must be considered. For example, IMDB ratings could be manipulated if a show is trying to be picked up by Netflix as a result from this study. Additionally, creating an addictive catalog could have a negative impact on Netflix subscribers' quality of life. Finally, movies that do not meet the criteria of high subscriber return could be less likely to be selected for addition to Netflix's catalog.

Various challenges may arise during the analysis of this data. For example, the data may not be clean when attempting to relate data from one database to the next. There could also be data points missing values that would warrant further investigation. Any outliers will need to be thoroughly examined and dealt with accordingly.

The goal of this study is to gain more insight into the relationships between movie performance and Netflix subscriber counts. From this insight, further areas of exploration may be identified.