A picture containing clock, knife

Description automatically generated

Netflix ETL Data Analysis Report

Gredel, CJ, and Angie

The purpose of this report is to use data analysis to allow our business to make the best decisions about which movie offerings to invest in. Our profit depends on our subscription numbers. One vital component to increase subscriptions is to provide a strong reason for the consumer to pay a monthly fee to have access to our offerings. More information about the movies that we currently offer will help us make better predictions about potential movies and choices for which movies to acquire in the future. This initial data analysis will focus on movies produced in the United States.

Extract, Transform, Load

The current Netflix database (source: csv file) was filtered to movies produced in the United States. Two popular databases; OMDb API and TMDB were used to extract movie data. We wrote an API call in Jupyter Notebook for both data bases and used the titles of each movie in our database as the search criteria. This ensured that we only get data for movies that Netflix currently offers to our subscribers.

After the data was obtained from the API calls we prepared a database in PgAdmin called movie\_db to store the extracted data. Three tables were created for the Netflix, OMDb, and TMDB data using SQL in PgAdmin. After careful consideration, only columns of interest were created and datatypes were chosen based on data information. From Jupyter Notebook the data was cleaned further before loading to the SQL database. Columns for each dataframe were renamed appropriately for the tables in the database. The box office revenue column in the OMDb dataframe was converted from a string into an decimal which was required for use in data analysis. After initial cleaning of the dataframes in Jupyter Notebook they were loaded into the SQL database.

Data Analysis

We would like to know if there are any correlations between the rating systems of each movie database. The OMDb also provided box office revenues for many movies which could provide useful information. Using python in jupyter notebook we transformed the data such that scatter plots were produced and subsequent linear regression analysis and r-squared values could indicate correlation.

Sources:

<https://www.kaggle.com/shivamb/netflix-shows> (csv download)

<http://www.omdbapi.com/>

<https://www.themoviedb.org>