***Final Project Report***

***Extract***

Our original data sources were CSV files predominantly from Kaggle. We were fortunate enough to have many options on movie-related data.

***Transform***

First, we downloaded are datasets into Pandas and then cleaned the data. This included checking variable types, checking for NaNs, and checking for consistency.

As many of our datasets contained many unnecessary fields, we deleted many columns for the sake of simplicity.

For the sake of our analysis and filtering, we opted to filter for only movies that were produced in the U.S between 2000 and 2017.

***Load***

As a final step, we loaded our various datasets into SQL / Postgres. We also found it useful to create the tables first in postgres, stating the variable type, and then working to push the pandas dataframe to Postgres. This helped us further clean our data. Once our data was in Postgres, we performed a join, as if we were to do further analysis, we would likely have done many joins on the movie title key.