**ETL Project**

Jake Vogt, Sarah Miller, Hannah Huntoon

**Proposal:**

We will extract data from two CSV files, netflix\_titles.csv and movie\_metadata.csv. Both data sets come from Kaggle.com. We will clean the data sets, pull out relevant information and structure the data in a relational database (PostgreSQL). Lastly, we will attempt to join the two datasets together on ‘title’ and do analysis on the overlap between IMDB movies and Netflix movies.

To accomplish this project we will use Pandas, Python, PostgreSQL/pgAdmin and Jupyter Notebook.

**Extract:**

We searched Kaggle until we found two semi-related data sets that we were interested in doing analysis on. We found an IMDB movie data set and a Netflix movie data set. Both files had approximately 5,000 rows each. We downloaded the CSV files and then read them in using Jupyter Notebook and Pandas.

**Transform:**

We created data frames from the CSV files and examined the data to familiarize ourselves with the column names and the respective information. Then we cleaned each of the datasets, selecting the columns we wanted to keep for further analysis. The Netflix dataset included a column named ‘type’ which included both movies and tv shows. Since we were only interested in analyzing Netflix movies, we dropped tv shows from the ‘type’ column. Then we renamed the columns of each dataset to match so we could eventually join the data. We realized that both datasets had an existing index, so we reset both of the indexes to the column ‘title’. The cleaned data frames were exported as new CSV files.

**Load:**

We used PgAdmin to create a database called ETLproject in PostgreSQL. From there we created two tables, ‘netflix’ and ‘movies’. After creating the tables, we imported our cleaned CSV files into each respective table. Once the data was imported we decided to join the tables on the column ‘title’. In order to join the two datasets on ‘title’ we had to do additional data cleaning. While merging the data we ran into a few grammatical differences between the two datasets. Once we realized this we further cleaned the datasets to make the data match. We joined the two tables in PostgreSQL and in Jupyter Notebook (using Pandas) so we could easily visualize the newly formed table and perform our analysis.

**Analysis:**

We ran into an issue as we attempted to perform analysis on our combined dataset. We had previously renamed the columns of each dataset to match one another. In order to perform additional analysis we had to drop duplicate column names. We did value colunts on the ‘rating’, ‘release\_year’, ‘director’ and ‘imdb\_score’ columns.

* 492 movie titles from the Netflix dataset match titles from the IMDB movie dataset. Both CSV files had ~5000 rows, so ~10% of movie titles match.
* Of the 492 movie titles from the combined dataset, 72% were either rated ‘R’ or ‘PG-13’.
* 80% of the movie titles from the combined dataset (395 of 492) have a release year of 2000 - 2019.
* Steven Spielberg (8 movies) and Quentin Tarantino (6 movies) were the directors with the most movies listed in the combined dataset.
* 52% (256) of the 492 movies had an imdb score between 6.1-7.6 followed by 26% (130) having a score between 4.6-6.1. Surprisingly, only 17% of the movies had an imdb score above 7.6.