ETL Project Summary
Team Name: Data Miners (Seobin, Vaughn, Johnny, William, Habiba)
Project Title: Netflix Trends by Google

Project Description:
Utilizing two datasets: “Netflix Movies and TV Shows” by Shivam Bansal and “Google Trends Dataset” by Dhruvil Dave, we are going to explore the popular movies that are accessible on Netflix. The former dataset is updated monthly while the latter contains data from 2001 to 2020. Both of the datasets selected are sourced from Kaggle and have a usability rating of 10.0. These datasets source global data from their respective companies.

Extract:
As mentioned above, we collected ‘Netflix Movies and Shows’ and ‘Google Trends Datasets’ from Kaggle. We used the datasets available in CSV format so that we can easily import the datasets to clean and merge.

Transform:
For our workspace, we used a combination VS Code and Jupyter Notebooks with Pandas to create the working data frames. After extracting the CSV files, we combined the Google Trends and the Netflix data on an inner join. With the combined data we then got rid of null and extraneous data from the google trends that did not relate to watching habits of consumers, while the Google Trends dataset includes the data such as music, sports, or foods. The Google Trends data only provided the top 5 search rank for each category so it ended up limiting how much of the Netflix data that would appear.

Load:
After transforming the datasets by merging and filtering, we created the datasets of movies, available in Netflix, which was ranked top 5 in annual google search ranking when the movies were initially released. With this dataset, we created the database using the PGAdmin PostgreSQL. This allows us to access the information in the database much faster and easier without having the CSV file. We then connected the data to our Jupyter notebook using SQLAlchemy to run queries. To add on, creating a database allows us to develop multiple tables using the queries in the future

Additional Comments:
From the database that we cleaned, we can observe that ‘Black Panther’ ranked the most frequent top 5 searches around the world among the Netflix movies, and ‘Avengers, Infinity Wars’ and ‘The Conjurings’ are placed the next most frequent top 5 searches around the world.

While transforming the original datasets, we initially wanted to focus on the google searches in the ‘United States’ and the ‘Global’ locations. However, there were a limited number of movies that are both available on Netflix and ranked top 5 in the Google search, so we had to change our plan to every country. Also, we had to make sure the keywords that are ranked in the top 5 Google searches refer to the exact movies by adding the condition regarding the released year.