**ETL Project Report:**

**Netflix Revenue and Subscribers data**

Group 5

Jessica Layne, Nagender Reddy Panyala, Nidhi Sethi and Ryan Youngblood

**Extract Data**

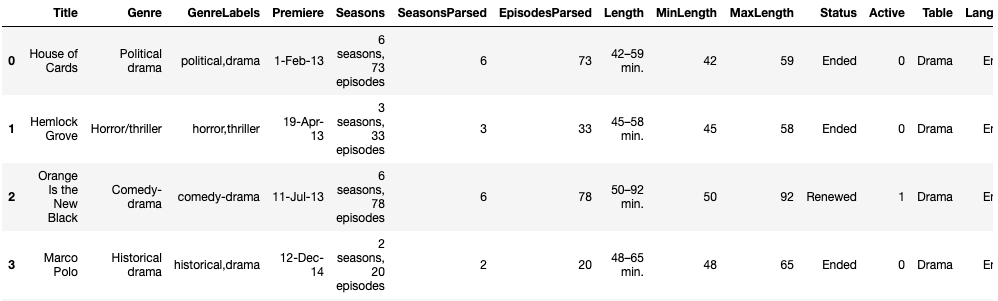
**Data Sources:(Kaggle)**

· Netflix Originals CSV- (https://www.kaggle.com/pariaagharabi/netflix2020)

· Netflix Subscribers and Revenue CSVs- (https://www.kaggle.com/pariaagharabi/netflix2020)

**Transform Data**

The first step in transforming this data was to clean the Netflix Originals CSV to only show the relevant columns. Using pandas, we filtered the table to only reflect Title, Genre and Premiere columns. This is shown in Figures 1 & 2:



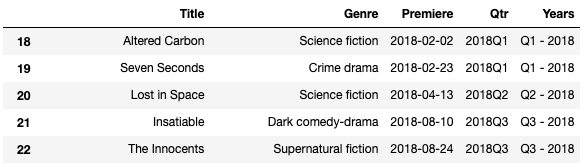
**(Figure 1:** Netflix Originals Original**)**

**↓**



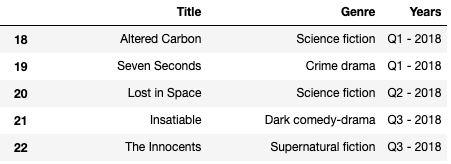
**(Figure 2:** Filtered Netflix Originals**)**

Next, using the DateTime variable, we adjusted the Netflix Originals, “Premiere” column to the same title of “Years” and Qtr - YYY format of the Subscriber and Revenue CSVs. This is reflected in Figures 3 & 4.



**(Figure 3:** Added Years column in Netflix Originals**)**

**↓**

****

**(Figure 4:** Netflix Originals without unneeded columns, Premiere and Qtr)

Then we filtered the Subscriber and Revenue Data. To do this, we filtered the area to only reflect The United States and Canada. You can see this shown in Figures 5 & 6:



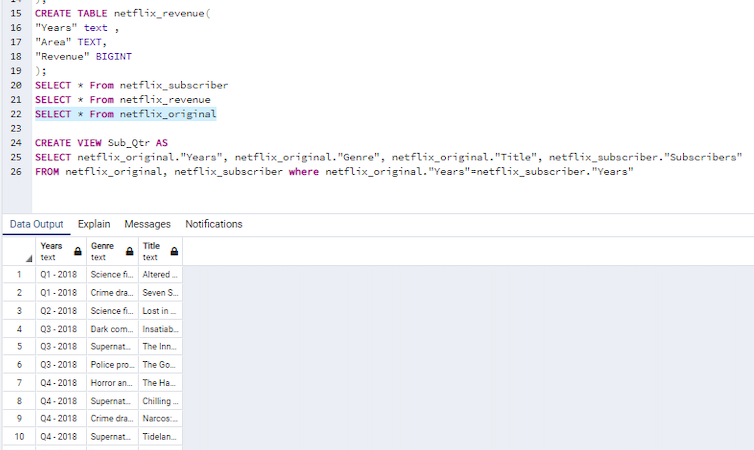
**(Figure 5:** Filtered Netflix Subscriber Data**)**

****

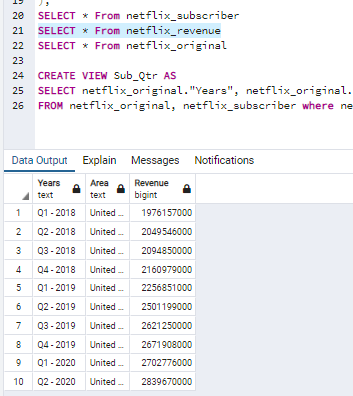
**(Figure 6:** Filtered Netflix Revenue Data**)**

**Load Data**

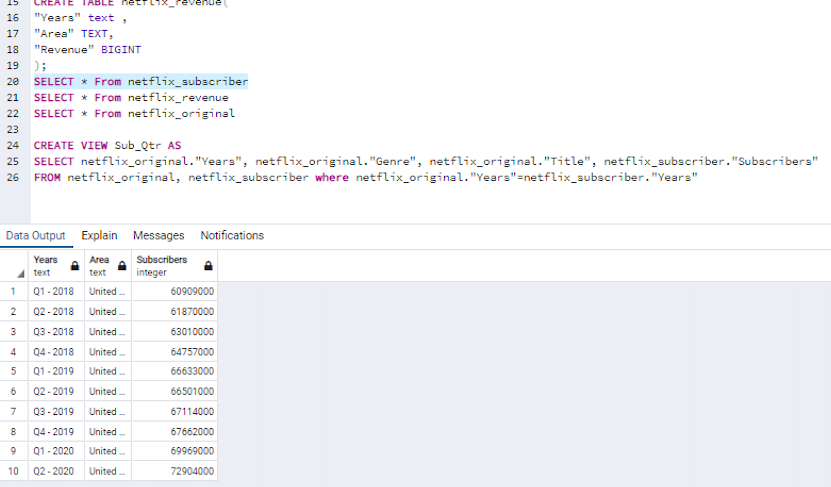
After cleaning the CSVs with Pandas, we used PostgreSQL to read in all of the cleaned CSVs. Shown in Figures 7-9:



**(Figure 7:** Netflix Originals cleaned data**)**

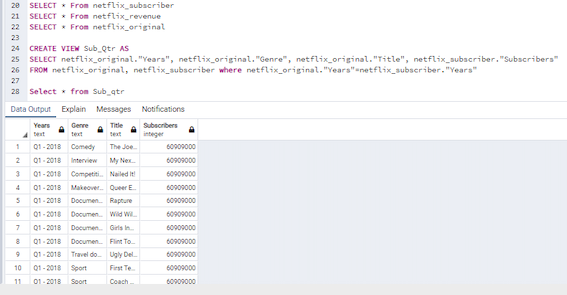
****

**(Figure 8:** Netflix Revenue cleaned data**)**

****

**(Figure 9:** Netflix Subscribers cleaned data**)**

Using the clean tables, we then joined them using the common identifier, Quarter of year. Shown in Figure 10:



**(Figure 10:** Joined clean data**)**

Plotting the Years vs subscribers and Revenue information:

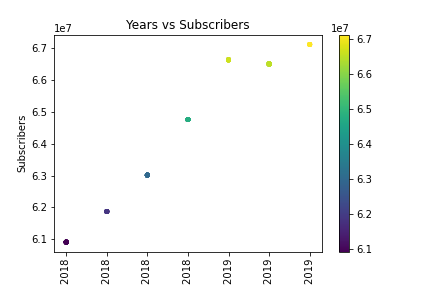


Fig. 11 shows the relation between “Years and Subscribers information”.

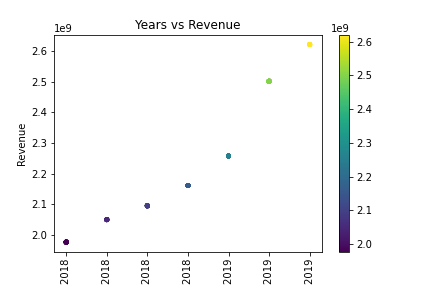


Fig. 12 shows the relation between “Years and Revenue information”.

**Summary of findings**

Through our study, we found that the Netflix Revenue and Subscribers numbers seem to steadily increase with time [**refer to Fig. 11 and 12**]. There doesn’t seem to be a correlation between any Title premieres and Netflix revenue or subscriber numbers.

A challenge that we encountered was that a Title’s premiere date doesn’t reflect the length of the show. Some shows are still running or were out long enough to span multiple quarters. The Revenue and Subscribers data is only reflected for the quarter the Title premiered, not for the duration of the Title. This is something we would look further into if we were to continue our research on the topic.