**Technical Write-up report for ETL-Project 2**

**The sources of data:**

Web Scrape from Rotten Tomatoes ( <https://www.rottentomatoes.com/> )

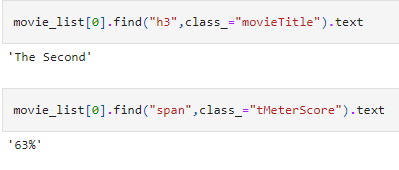
Netflix Movies & TV Shows (<https://www.kaggle.com/datasets/shivamb/netflix-shows>)

**Transformation of Data:**

At first, we attempted to load the data from Rotten Tomatoes using Jupyter Notebook, however we ran into an issue of only being able to extract 32 movies at a time. We were able to code the computer to clicking the “Load More” button on the website to extract the entire list. Totaling 601 pages.

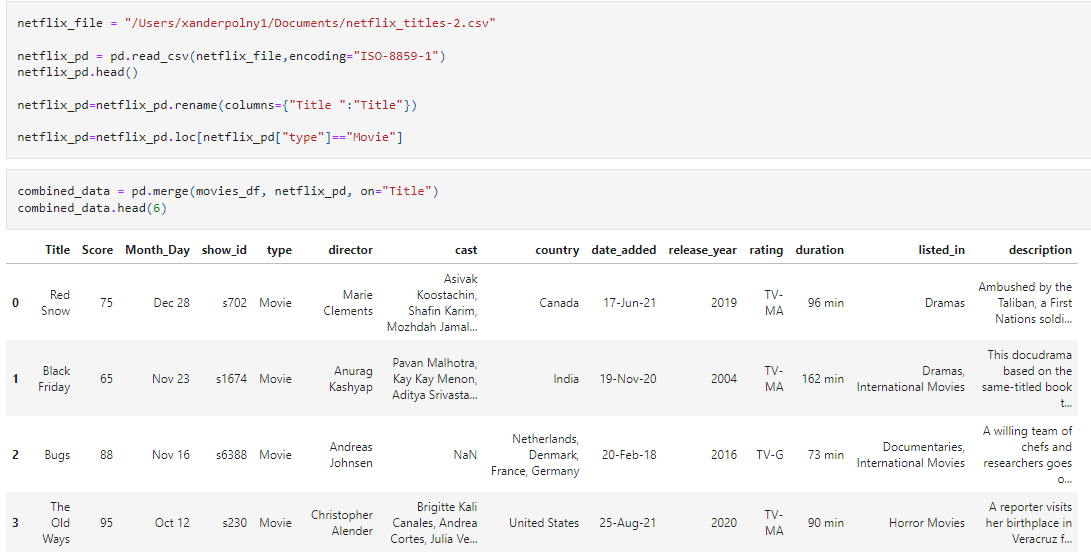


We then tested the headers for the data frames. Assuring we retrieved the data we were looking for. Movie Title, Score and Release date. We were then able to print movies\_df with over 9000 lines of data.

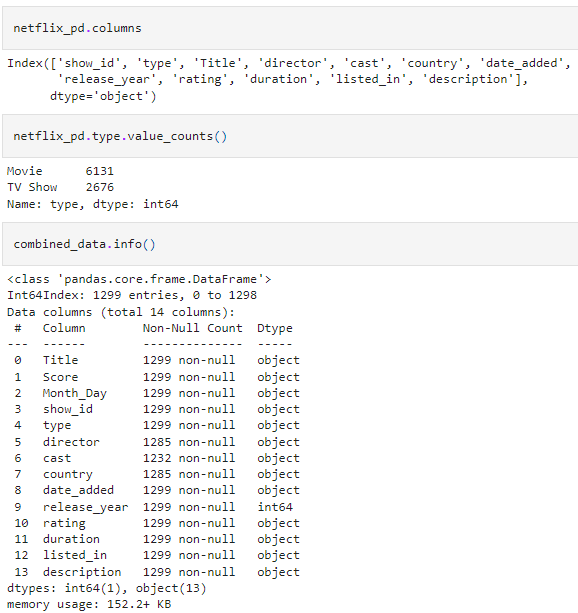




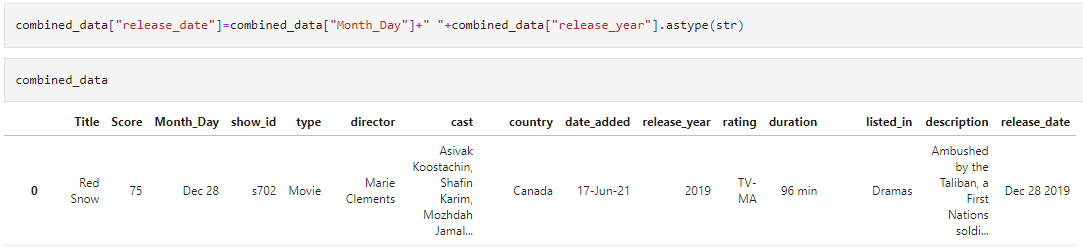
The next steps involved importing data from a CSV file we downloaded from Kaggle, Netflix Titles. In the same lines we changed the title on the CSV file. We also merged both tables into one data frame.

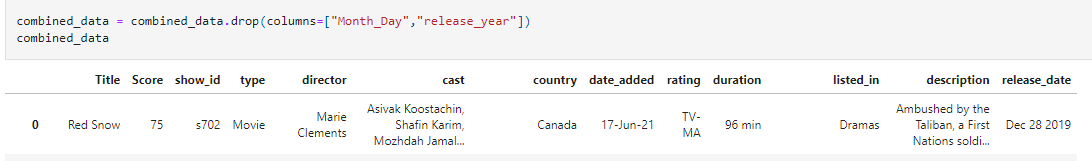


Next, columns were created using PostgreSQL. We were able to see how may data points were in the columns. We were also able to print the data frame.



Nex we combited columns, release\_date from Rotten Tomatoes and Month\_Day for the CSV file. We also concatenated the year from release\_year. After that, we dropped columns Month\_Day and release\_year.

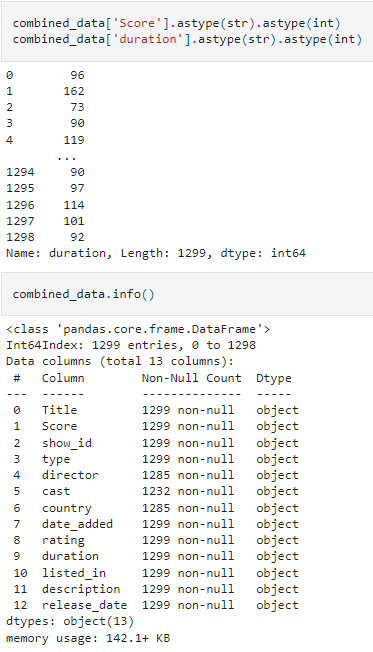




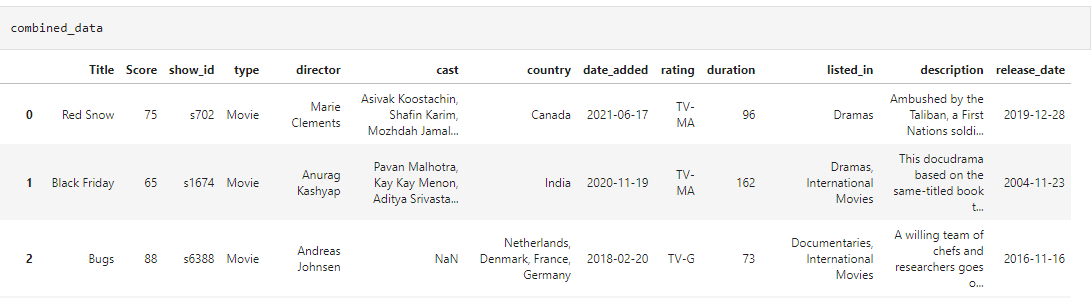
Next, we used str.replace for the duration column, to show in minutes.



Next, we changed the data type of Score and duration from str to int. Then we displayed the data frame info.



Next, we displayed the data frame for the final verification of all data we wanted to display.



The final steps were loading the data back into PostgreSQL.

