ETL Project

By: Michael Munson

**Extraction**

I used two different datasets from the public platform Kaggle to locate game sales data for the 2 platforms Playstation 4 and Xbox One. The data in the two files included the following information:

* Game Name
* Year
* Genre
* Publisher
* North America, Europe, Japan, Rest of the world, and Global Sales totals

The fields of interest include the following:

* Game
* Year
* Platform
* Global Sales

\*Including data from 2014/2020 years.

The following sources for our datasets used:

<https://www.kaggle.com/sidtwr/videogames-sales-dataset?select=XboxOne_GameSales.csv>

<https://www.kaggle.com/sidtwr/videogames-sales-dataset?select=PS4_GamesSales.csv>

**Transformation**

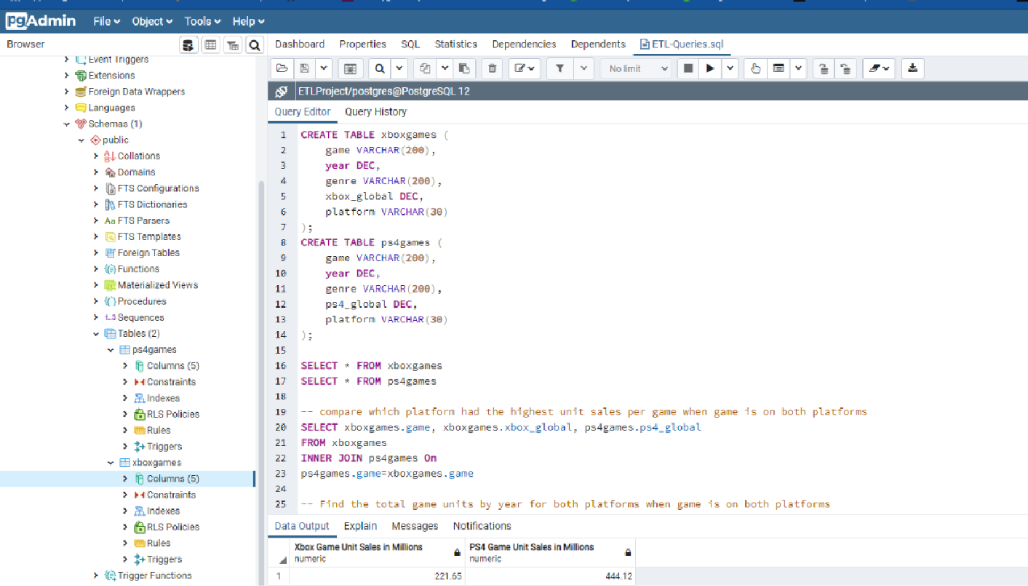
In order to transform the game data and use it in my study I performed the following:

* Used Pandas functions in Jupyter Notebook to load both CSV files.
* Reviewed the files and transformed into data frames
* Removed the unneeded columns from the data frames.
* Deleted any rows that were missing data.

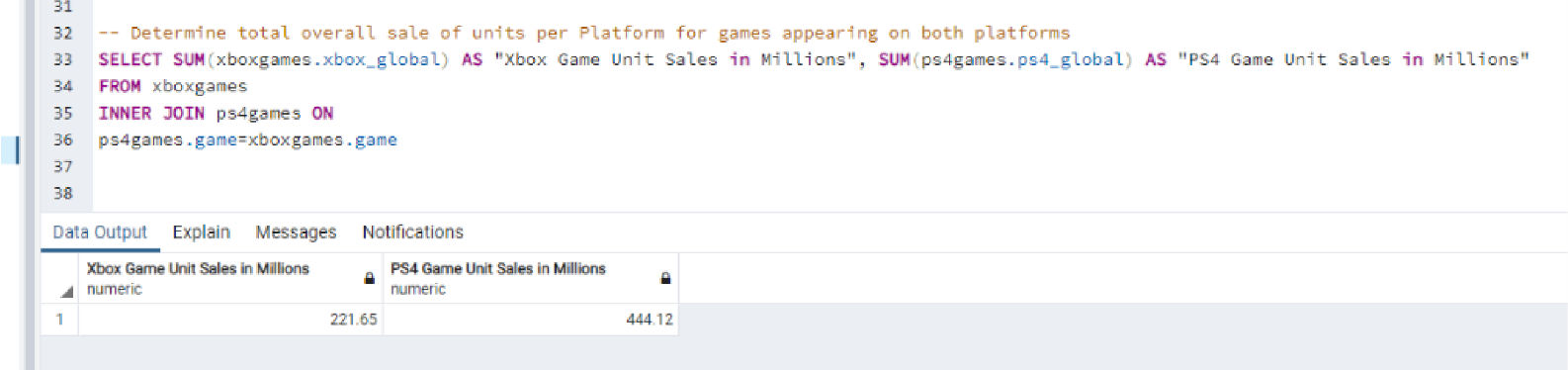
**Load**

After I pulled in the CSV files and loaded them into the data frames, I did an initial connection to the Postgres database using PG admin to store my original clean data sets. I ran my initial queries to create new tables to hold only the relevant information needed from my data sets. I then ran various queries and joins to compare the data of both platforms and determine which had sold the most units.

**Postgres Database:**

****

**Total Unit Sales**



**Summary**

The data sets obtained had different numbers of games due to some titles being exclusive to certain platforms. In order to analyze the data equally I used only games that appeared in both data sets. After reviewing the results of the combined tables, it would appear that Playstation 4 sells units at approximately a 2:1 ration over Xbox One. This would seem to show that PS4 is the more popular platform for the games analyzed.