University of Denver
Data Analytics Bootcamp
Project 2: ETL

Technical Report:
ETL Analysis on Show Streaming Data

By Holly Shirmenzagas, Daniel Jackson, Daniel Carrillo

In this project, we were able to execute each step in the ETL process (extract, transform, and load), which was the priority. Secondary to that priority, we wanted a topic that would be both practical as well as feasible within our given timeline. On our previous project, we took a hard focus on finance as all team-members have an interest or are prospecting in that area of data science; however, this time around, we were leaning a little bit more towards what could be a more fun topic. We ended up being more into the idea of having fun while practicing our ETL skills on real data. We finally settled on the topic of movies and shows streaming on major platforms when we all couldn't stop smiling while talking about it. We couldn't help it, we had to ditch our original fascination with financial data to run with this fun alternative.

Though we had discussed movies, we decided to focus solely on show streaming data (so no movies are included in this analysis, only shows). We also decided to keep our queries and representations fairly simple for the sake of practice, meeting the deadline, and fulfilling the criteria.

Extraction:

The first step was to extract the data we wanted from reliable sources. We originally toyed with the idea of web-scraping all of our data, but found that this kind of data is not nearly as accessible from a webpage because much of is housed in high-level databases like IMDB—otherwise, we likely would have had to have scraped our data one web-page at a time from different sites and sort of "Frankenstein" it all together. So in the end, we thought would be best to start with a simple API and dataset (MovieDB and Kaggle). Please see Jupyter Notebook titled "data_extraction.ipynb" to observe our methodology.

Transformation:

The primary transformations of the data necessary for our purposes were joining, amalgamating, and prettifying, with the goal being a useable master dataset. Much of that can also be found in "data_extraction.ipynb".

Load:

The last steps were that of loading this data into SQLite, running queries, and generating some simple visualizations. We had some technical difficulty in loading our master CSV into an SQLite database file, including failing to run SQL Studio, attempting to start with pgAdmin, and not quite being able to convert CSV to SQLite. We finally had success in loading SQLite. Please review our Jupyter Notebook "sql.ipynb" and our SQLite file "tv_data.sqlite".

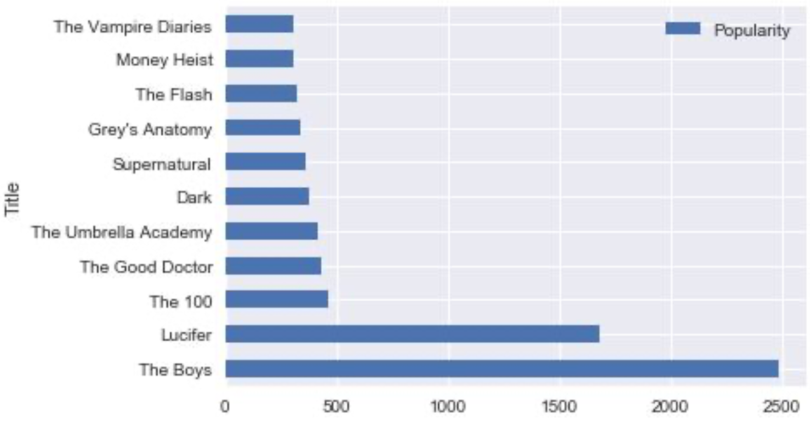
Analysis – The “Fun” Part:

We ran some simple queries, although we were excited at the prospect of everything we could do with this data and with our SQLite. However, for the sake of executing the project, our simple queries include:

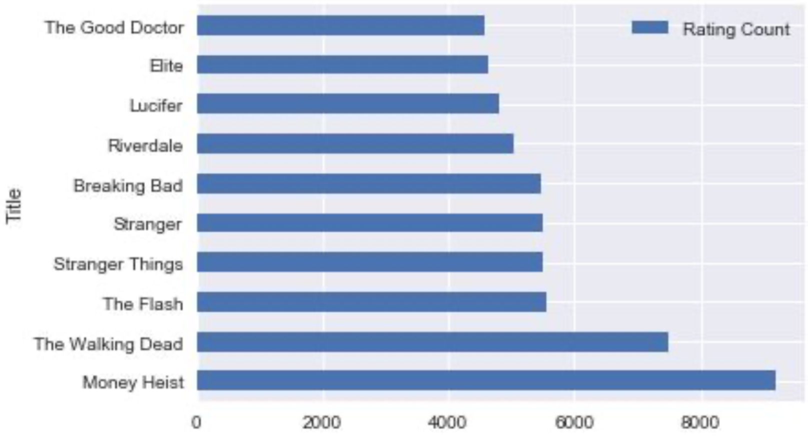
- What were historically the most popular shows across all streaming platforms?
- What were historically the most rated shows across all streaming platforms?
- On average, which country’s shows are most popular?

These were our results (please see Jupyter Notebooks “visualizations1.ipynb” and “visualizations2.ipynb”, as well as the “Graphs.pdf” file):

Most Popular Titles Across all Streaming Services:



Most Rated Titles Across all Streaming Services:



We were essentially able to leverage the power of SQL to generate valid results, which we were very pleased about. Some of us had never even heard of “The Boys”, and yet it scored the highest popularity. One might not necessarily stop to think about how rating count can also be an entirely different metric for what shows people are watching and reviewing—“Money Heist” had actually scored the highest frequency of ratings. We were surprised that Columbia actually has the highest rank of popular shows—looks like we have many foreign films to queue on our watchlist! Furthermore, we know that we are only beginning to scratch the surface of what’s possible with this powerful technology and excited at the possibilities!