Extract, Transform, Load

# Project 2 by Kayla Kidwell, Rick Springer, Adam Alvarado, Vanessa Gonzales

Through Kaggle.com, we were able to pull datasets that were in CSV files that contained data for Hulu, Netflix, Disney+, and Amazon Prime. These were the streaming services that we chose to work with. We used Jupyter Notebook to run Python to read the CSV files. The data was then copied to each subsection of information that we would be analyzing to protect the original data. When we analyzed the data, specifically “genres”, we had to split the columns into multiple columns using a comma delimitator. While each streaming service had 3 genres each, Amazon prime had 5 genres. The next step was to remove every unnecessary column that was irrelevant to the specific information we were looking at for our analysis. This where having the original data came into play. Once the tables were using clean data, we could then use Matplotlib to produce graphs that would display the correct, clean data. We made sure that the scale stayed the same for each graph. From there, we saved the new tables that were created into a new CSV file. We wanted to combine the data but to do so, we needed to create identifiers that were specific to each dataset to keep it organized. To achieve this, we added a column to each with an identifying separator for each dataset. Finally, all datasets were able to be merged into one combined database and loaded into MongoDB.