Sarah Green
Hector Ladero Rivas
UNCC-Data & Analytics

REPORT – Which is the best?

We decided to compare the three main movie entertainment servers (Netflix, Amazon Prime and HBO) to check which server is displaying the highest number of best rating movies.

Data extraction:

- Using Kaggle, we pulled a CSV file with the top 1000 movies according to IMDB.
- Scraping the website: www.reelgood.com, we grabbed the current movies from Netflix, Amazon Prime and HBOMax in a html format, transformed and saved as csv file.
- Extracted HTML code from www.businesssofapps.com/data/Netflix/statisics
  (Was able to obtain statistical information on all of the different streaming services).

Data transformation:

- CSV: Removed duplicated data (from 1,000 rows to 402)
  Dropped unnecessary rows
  Renamed headers
- Netflix, HBOMax and Prime:
  Removed duplicated data.
  Dropped unnecessary rows.
  Renamed headers

- With all the clean data, we joined the IMDB csv with each of the streaming services to see which movies were in both.

- HTML: Pull statistical tables from the website
  Removed duplicated data
  Reset Index
  Dropped unnecessary rows
  Renamed headers
  Change data types to elaborate graphs.

Load Data into SQL:

- We chose to use PgAdmin to create our database. We created a table for each of the data categories we had [Netflix, HBOMax, Prime and IMDB].
- We then linked the database tables to our data in Jupyter Notebook using sqlalchemy.

We did joined Netflix, HBO, Amazon Prime with the CSV File to compare and check which movies were duplicated.