

Streaming Services Analysis

Team Members:

Salita Santiago
Shayna Chernak
Hima Gharat
Marcellis Valentin



Hulu

30 Million subscribers

Netflix

183 Million subscribers

Netflix's reported
2020 revenue was
5.77 Billion

More than 55% of the
US population uses
streaming services

Disney

60 Million subscribers

Disney launched a
year ago and has
already surpassed
Hulu's subscription
total

Amazon

150 Million subscribers

Biggest viewing
increase is due to
viewing time on
smartphones

Do Ratings Matter?

- Streaming platforms have changed how we consume content.
- Platforms like Rotten Tomatoes and IMDB give viewers the ability to rate this content on a public forum.
- We wanted to explore a possible link between the average ratings of shows and movies on streaming platforms and which type of content streaming platforms are leaning more towards.

Data Extraction

- Data sources- CSV files from Kaggle
- Loaded into Pandas/Jupyter Notebook

```
In [1]: import pandas as pd
        from sqlalchemy import create_engine
```

```
In [2]: movie_file = ("C:\\UPenn\\ETL_Project\\Netflix_Data_Analysis\\movies.csv")
        movie_df = pd.read_csv(movie_file)
        movie_df.head()
```

Out[2]:

	ID	Title	Year	Age	IMDb	Rotten Tomatoes	Netflix	Hulu	Prime Video	Disney+	Type	Directors	Genres	Country
0	1	Inception	2010	13+	8.8	87%	1	0	0	0	0	Christopher Nolan	Action,Adventure,Sci-Fi,Thriller	United States,United Kingdom,English,Japanese
1	2	The Matrix	1999	18+	8.7	87%	1	0	0	0	0	Lana Wachowski,Lilly Wachowski	Action,Sci-Fi	United States
2	3	Avengers: Infinity War	2018	13+	8.5	84%	1	0	0	0	0	Anthony Russo,Joe Russo	Action,Adventure,Sci-Fi	United States
3	4	Back to the Future	1985	7+	8.5	96%	1	0	0	0	0	Robert Zemeckis	Adventure,Comedy,Sci-Fi	United States
4	5	The Good, the Bad and	1966	18+	8.8	97%	1	0	1	0	0	Sergio Leone	Western	Italy,Spain,West Germany

```
In [3]: tv_file = ("C:\\UPenn\\ETL_Project\\Netflix_Data_Analysis\\tv_shows.csv")
        tv_df = pd.read_csv(tv_file)
        tv_df.head()
```

Out[3]:

	ID_Number	Title	Year	Age	IMDb	Rotten Tomatoes	Netflix	Hulu	Prime Video	Disney+
0	0	Breaking Bad	2008	18+	9.5	96%	1	0	0	0
1	1	Stranger Things	2016	16+	8.8	93%	1	0	0	0
2	2	Money Heist	2017	18+	8.4	91%	1	0	0	0
3	3	Sherlock	2010	16+	9.1	78%	1	0	0	0
4	4	Better Call Saul	2015	18+	8.7	97%	1	0	0	0

```
In [4]: #Transform Movie DataFrame
```

Data Transformation

- Cleaned data to include only pertinent information and assigned unique names to each column

```
In [4]: #Transform Movie DataFrame
movie_cols= ["Title", "IMDb", "Netflix", "Hulu", "Prime Video", "Disney+" ]
movie_transformed= movie_df[movie_cols].copy()

#Rename column headers
movie_transformed = movie_transformed.rename(columns={"Title": "movietitle_id",
                                                    "IMDb" : "movieimdb_id",
                                                    "Netflix": "movienetflix_id",
                                                    "Hulu": "moviehulu_id",
                                                    "Prime Video": "movieprime_id",
                                                    "Disney+": "moviedisney_id"})

movie_transformed.head()
```

```
Out[4]:
```

	movietitle_id	movieimdb_id	movienetflix_id	moviehulu_id	movieprime_id	moviedisney_id
0	Inception	8.8	1	0	0	0
1	The Matrix	8.7	1	0	0	0
2	Avengers: Infinity War	8.5	1	0	0	0
3	Back to the Future	8.5	1	0	0	0
4	The Good, the Bad and the Ugly	8.8	1	0	1	0

```
In [5]: #Transform TV DataFrame
tv_cols= ["Title", "IMDb", "Netflix", "Hulu", "Prime Video", "Disney+" ]
tv_transformed= tv_df[tv_cols].copy()

#Rename column headers
tv_transformed = tv_transformed.rename(columns={"Title": "tvttitle_id",
                                                "IMDb" : "tvimdb_id",
                                                "Netflix": "tvnetflix_id",
                                                "Hulu": "tvhulu_id",
                                                "Prime Video": "tvprime_id",
                                                "Disney+": "tvdisey_id"})

tv_transformed.head()
```

```
Out[5]:
```

	tvttitle_id	tvimdb_id	tvnetflix_id	tvhulu_id	tvprime_id	tvdisey_id
0	Breaking Bad	9.5	1	0	0	0
1	Stranger Things	8.8	1	0	0	0
2	Money Heist	8.4	1	0	0	0
3	Sherlock	9.1	1	0	0	0

Data Loading

- Once data was transformed, we created a database connection to PostgreSQL

```
#Rename column headers
tv_transformed = tv_transformed.rename(columns={"Title": "tvtitle_id",
                                              "IMDb": "tvimdb_id",
                                              "Netflix": "tvnetflix_id",
                                              "Hulu": "tvhulu_id",
                                              "Prime Video": "tvprime_id",
                                              "Disney+": "tvdisney_id"})

tv_transformed.head()
```

```
Out[5]:
```

	tvtitle_id	tvimdb_id	tvnetflix_id	tvhulu_id	tvprime_id	tvdisney_id
0	Breaking Bad	9.5	1	0	0	0
1	Stranger Things	8.8	1	0	0	0
2	Money Heist	8.4	1	0	0	0
3	Sherlock	9.1	1	0	0	0
4	Better Call Saul	8.7	1	0	0	0

```
In [6]: #create database connection
connection_string = "postgres:rycbar706@localhost:5432/Streaming_Service_Analysis"
engine= create_engine(f'postgresql://{connection_string}')
```

```
In [7]: #confirm tables
engine.table_names()
```

```
Out[7]: []
```

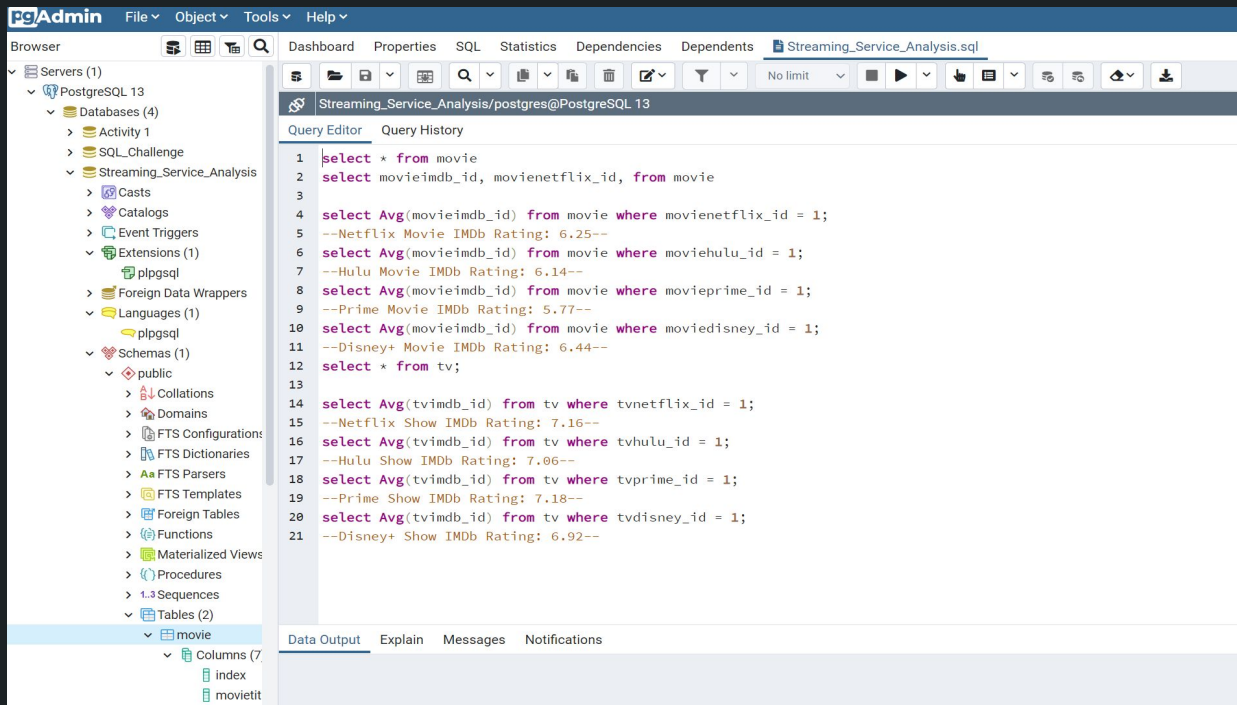
```
In [8]: #Load dataframes into database
movie_transformed.to_sql(name= 'movie', con=engine, if_exists='append', index= True)
```

```
In [9]: tv_transformed.to_sql(name= "tv", con=engine, if_exists= 'append', index= True)
```

```
In [ ]: #
```

SQL Queries

- Queried averages of IMDb Ratings of shows and movies from each streaming platform to determine which is more popular in each



The screenshot shows the pgAdmin 4 interface with the following components:

- Browser:** A tree view on the left showing the database structure. The 'Streaming_Service_Analysis' database is selected, and the 'movie' table is highlighted under the 'public' schema.
- Query Editor:** The main area on the right displaying a SQL query. The query is titled 'Streaming_Service_Analysis.sql' and is executed against the 'postgres@PostgreSQL 13' connection.
- Query:** The SQL query in the editor is as follows:

```
1 select * from movie
2 select movieimdb_id, movienetflix_id, from movie
3
4 select Avg(movieimdb_id) from movie where movienetflix_id = 1;
5 --Netflix Movie IMDb Rating: 6.25--
6 select Avg(movieimdb_id) from movie where moviehulu_id = 1;
7 --Hulu Movie IMDb Rating: 6.14--
8 select Avg(movieimdb_id) from movie where movieprime_id = 1;
9 --Prime Movie IMDb Rating: 5.77--
10 select Avg(movieimdb_id) from movie where moviedisney_id = 1;
11 --Disney+ Movie IMDb Rating: 6.44--
12 select * from tv;
13
14 select Avg(tvimdb_id) from tv where tvnetflix_id = 1;
15 --Netflix Show IMDb Rating: 7.16--
16 select Avg(tvimdb_id) from tv where tvhulu_id = 1;
17 --Hulu Show IMDb Rating: 7.06--
18 select Avg(tvimdb_id) from tv where tvprime_id = 1;
19 --Prime Show IMDb Rating: 7.18--
20 select Avg(tvimdb_id) from tv where tvdisney_id = 1;
21 --Disney+ Show IMDb Rating: 6.92--
```
- Bottom Bar:** A tabbed interface with 'Data Output', 'Explain', 'Messages', and 'Notifications' tabs. The 'Data Output' tab is currently active.

Conclusion

- Based on the ratings on IMDB, we found that TV shows had higher ratings compared to Movies.
- This leads us to believe that, on average, viewers tend to like TV shows more than movies.

DATA SOURCES

https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney?select=tv_shows.csv

https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney?select=MoviesOnStreamingPlatforms_updated.csv

References:

<https://www.investopedia.com/articles/markets/051215/who-are-netflixs-main-competitors-nflx.asp>

Questions?