Final Project Report on :
Comparison of Content across the Streaming Services

Author: Harshit Paliwal
hpaliwal1@hawk.iit.edu
A20449708

Submitted on: 12/06/2020

Course: Applied Statistics(MATH 564)
Instructor: Dr. Lulu Kang

Table of Contents

1. Abstract

In today's modern world where everything has shifted to online service including shopping, food delivery, playing video games, listening to music, banking and many more. So the film and tv show industry have adopted the same with several good services to choose from. Gone are the days of cable subscription and people have adopted the need to change with this and so is the introduction of numerous big platforms for streaming. This project includes analysis of four major streaming services Netflix, Amazon Prime, Hulu and Disney+ in order to determine which among them is the most suitable for long term purchase and what type of users are attracted by different services, I have also analysed I started with analyzing the dataset which I extracted from Kaggle and performed exploratory data analysis using python scripts. I came to analyse several key aspects from the data which can help people identify which service is the best for them as different services are better for each viewer category. Following that I did hypothesis testing in order to determine how new released movies have a lower rating and runtime then older movies and tv-shows. I also applied the clustering technique in order to help viewers determine which movies/tv-shows have higher ratings so they can be better for viewing without wasting time. Lastly, I have developed a recommendation system for Netflix movies and tv-shows which outputs 5 predicted movies/tv-shows which the viewer would like to watch.

2. Introduction

2.1 Problem Description

Generally people tend to choose the best service in every aspect of their lives so why not in streaming services. We have many options to choose from and in this project I have compared four different streaming services and gave the users which one to choose according to their interests. This can save people their time and money in trying each service and then ultimately choosing one of them.

2.2 Problem Solving Strategies

Three different datasets are used in this project in order to give insights on different facets of the streaming services. I have applied different algorithms in order to understand the dataset better and help viewers in finally choosing the ultimate streaming service for them.

Models used-

1. Hypothesis Testing (One tailed and two tailed)
2. Correlation Analysis
3. K-Means clustering
4. Recommendation System algorithm

3. Problem Statement

3.1 Data Sources

In order to properly evaluate the problem I have used three different datasets which I took from Kaggle. These datasets are open source and open to anyone to perform their analysis, below are the links for the datasets-

1. https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney
   This dataset contains Title, Year, Age, Imdb rating, Rotten Tomatoes rating, and service in which the Title is available by showing by treating them as categorical variables. The type is a categorical variable showing 0 for movies and 1 for tv-shows

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

Table 1: Dataset from Kaggle-1

2. https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney
   This dataset contains Title, Year, Age, Imdb rating, Rotten Tomatoes rating, and service in which the Title is available by showing by treating them as categorical variables. The type is a categorical variable showing 0 for movies and 1 for tv-shows. The Genres, Country of the Title and Languages in which it is available is shown in the dataset.

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

Table 2: Dataset from Kaggle-2

3. https://www.kaggle.com/shivamb/netflix-shows
   This dataset contains show_id, type, Title, director, cast, country of origin, date_added, release_year, rating for viewers, duration of the show, Listed_in which category of viewers and finally a small description of the show.

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	81145628	Movie	Norm of the North: King Sized Adventure	Richard Finn, Tim Maltby	Alan Marriott, Andrew Toth, Brian Dobson, Cole...	United States, India, South Korea, China	September 9, 2019	2019	TV-PG	90 min	Children & Family Movies, Comedies	Before planning an awesome wedding for his gra...
1	80117401	Movie	Jandino: Whatever it Takes	NaN	Jandino Asporaat	United Kingdom	September 9, 2016	2016	TV-MA	94 min	Stand-Up Comedy	Jandino Asporaat riffs on the challenges of ra...
2	70234439	TV Show	Transformers Prime	NaN	Peter Cullen, Sumalee Montano, Frank Welker, J...	United States	September 8, 2018	2013	TV-Y7-FV	1 Season	Kids' TV	With the help of three human allies, the Autob...
3	80058654	TV Show	Transformers: Robots in Disguise	NaN	Will Friedle, Darren Criss, Constance Zimmer, ...	United States	September 8, 2018	2016	TV-Y7	1 Season	Kids' TV	When a prison ship crash unleashes hundreds of...
4	80125979	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper, Kate Walsh, John Michael Higgins...	United States	September 8, 2017	2017	TV-14	99 min	Comedies	When nerdy high schooler Dani finally attracts...

Table 3: Dataset from Kaggle-3

3.2 Complications with Dataset:

The raw dataset consisted of data from three different sources which needed to be clubbed together to perform various analysis. Change of variable needed to be made for example - rotten tomatoes needed to be changed from str which used % to float.
Also, IMDB ratings needed to be calculated out of 100 because of uniformity with rotten tomatoes rating in order to perform correlation analysis and plot several charts.
Descriptions from the third dataset needed to be tokenized in order to build a recommendation system using the separate words.
Variable names needed to be made similar to club together the datasets.

3.3 Problem Questions-

I have tried to answer several questions which the viewers would have regarding the streaming services-

1. Which is the best streaming service to pay a premium for ?
2. Which streaming service has the most varied content available ?
3. Which streaming service has the highest rated shows ?
4. Top words used in a title(Helpful for writers to make a new movie name)

5. What is the age distribution of the shows available ?
6. How does the Netflix movie recommendation algorithm work ?

4. Proposed Methodology

In order to solve the problem questions we need to first analyse our dataset which was done by making several plots and using **sweetviz** and **seaborn** library from python which outputs a high-density visualization of our dataset.

After analysis of the dataset using multiple different techniques, I will move on ahead to verify the findings with the help of some models. The idea to do hypothesis testing in order to verify that data provided is bent towards which direction.

Correlation between the year of shows and their ratings could provide us with an appropriate idea how the newly made shows will be appropriate to which kind of audience.

Clustering is important in order to verify how many good rated shows are entered in the streaming services every year and at what percent.

Lastly, I will try to make my own recommendation system which will help me to understand how Netflix does the work in the backend to recommend some great movies similar to our interests.

Hypothesis Testing:

A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.

A two-tailed test is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater or less than a range of values. If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.

Cluster analysis:

KMeans is one of the simple but popular unsupervised learning algorithms. Here K indicates the number of clusters or classes the algorithm has to divide the data into. The algorithm starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster. It performs repetitive calculations to optimize the positions of the centroids

5. Analysis and Results

5.1 Exploratory Data Analysis

5.1.1 Service with the most Content available

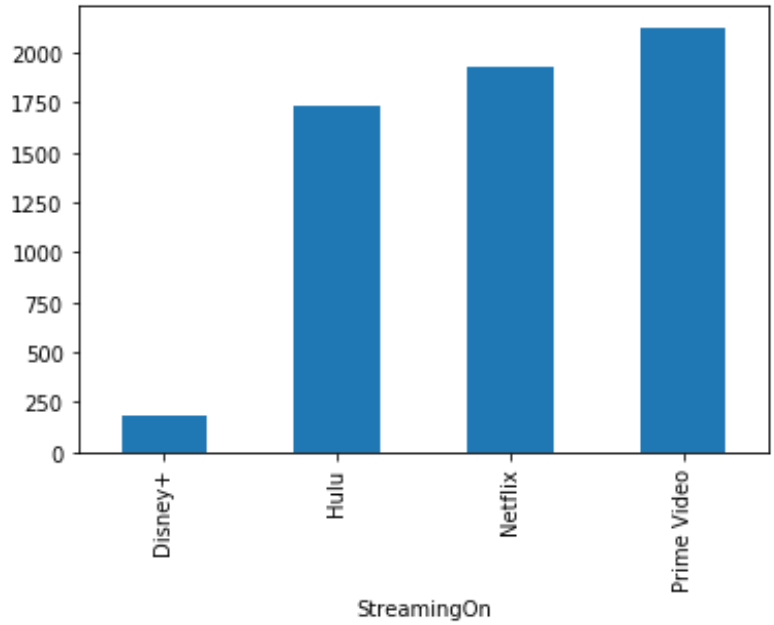

Fig 1: Bar-chart of shows in each streaming service

5.1.2 Scatterplot between IMDB and Rotten Tomato ratings

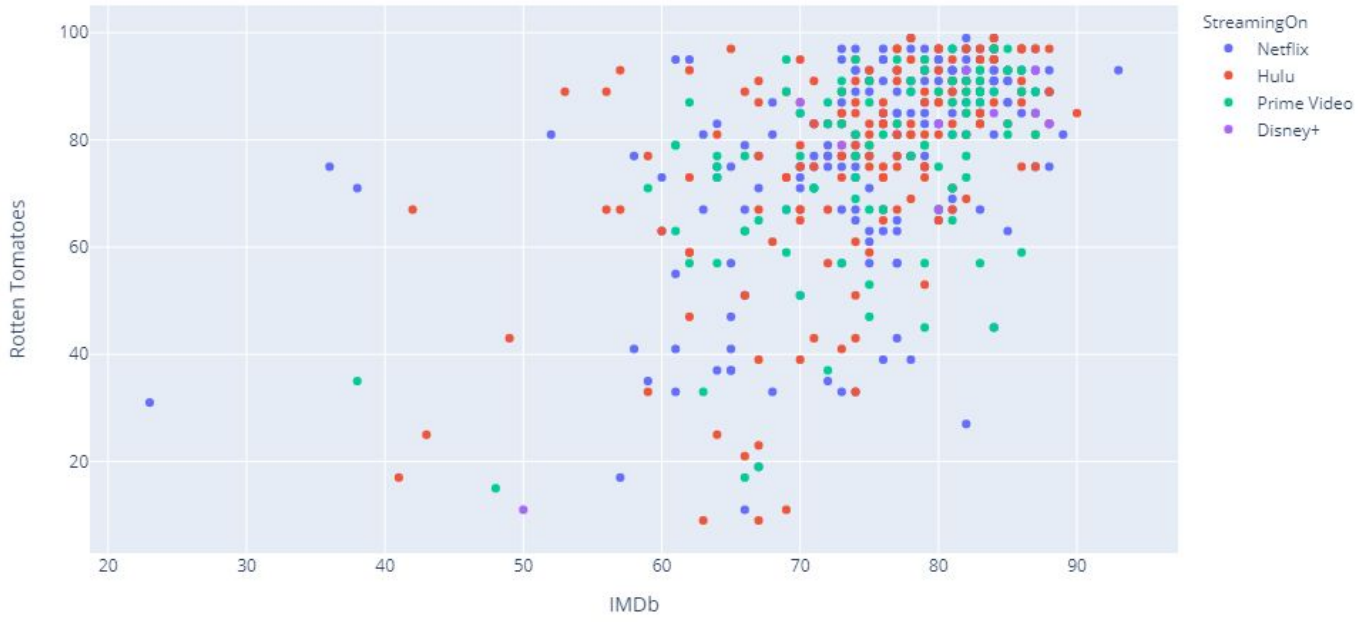

Fig 2: Scatterplot of rotten tomatoes ratings vs IMDB ratings in each streaming service

Fig 1:
Bar Plot Amazon Prime wins this race in this one. So looking at all three we can conclude Amazon Prime is both about quality and quantity.
Fig 2:
Scatter Plot With this another view, it is quite evident, Amazon Prime performs very well in the fourth quadrant. Which verifies our first inference

5.1.3 Violin Chart to gauge the content rating across all the streaming services

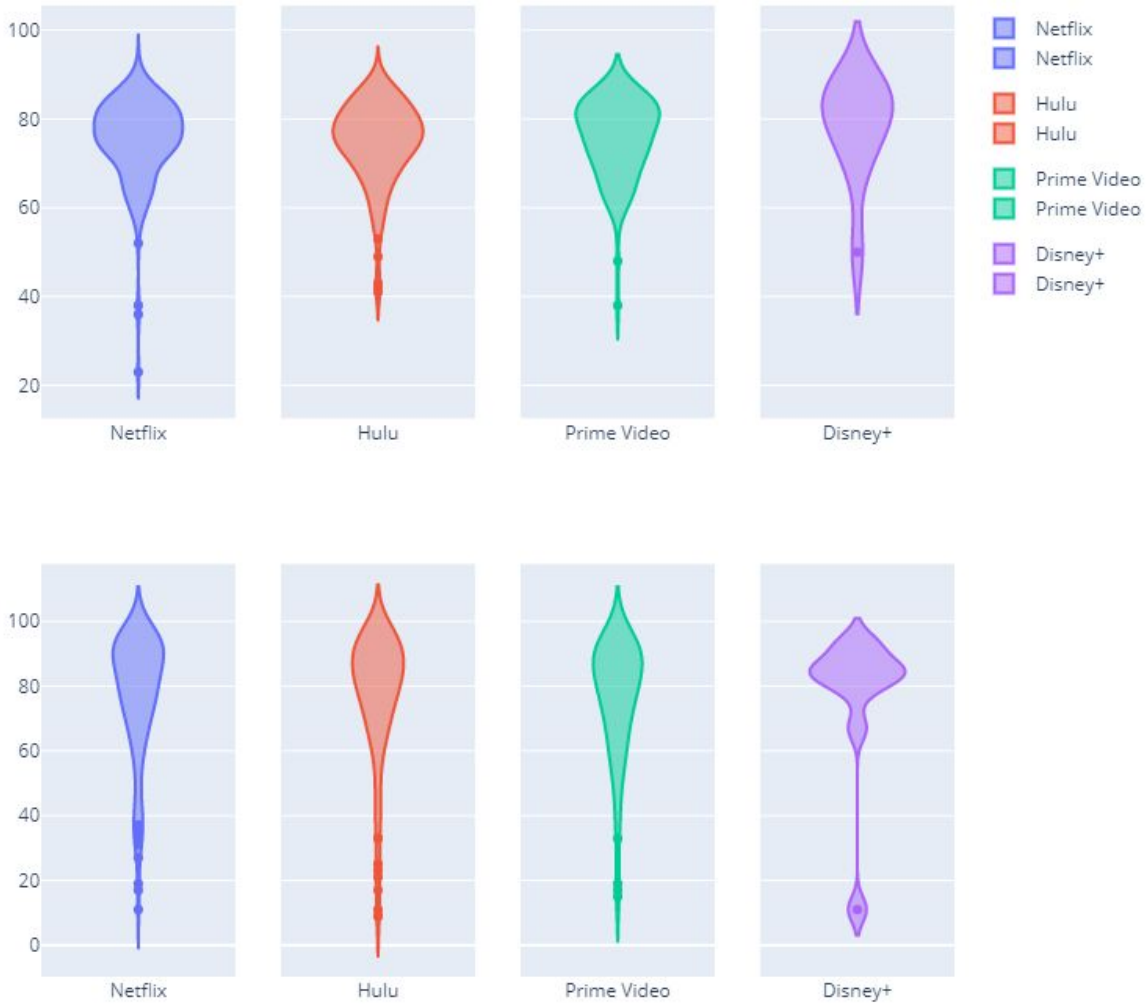

Fig 3: Violin Chart to measure content rating

Fig 3:
Hulu Netflix and Amazon Videos all three have got substantial data in lower end of the ratings.As the content increases so the quality decreases for all three.

5.1.4 Age Distribution of the Tv-shows

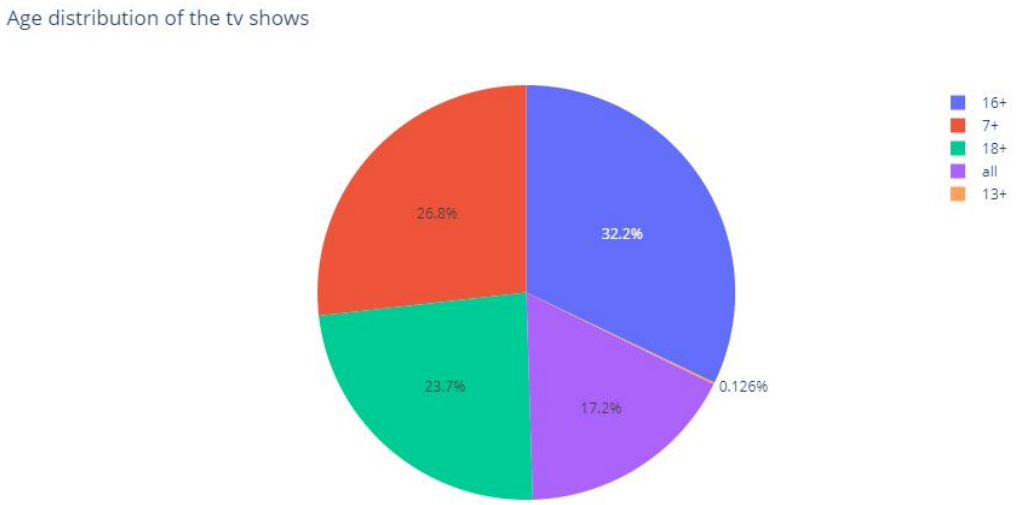

Fig 4: Pie-chart showing the Age distribution of tv-shows

5.1.5 Top words used in Titles

Fig 5: Top words used in Titles

5.1.6 Child Suitable streaming service

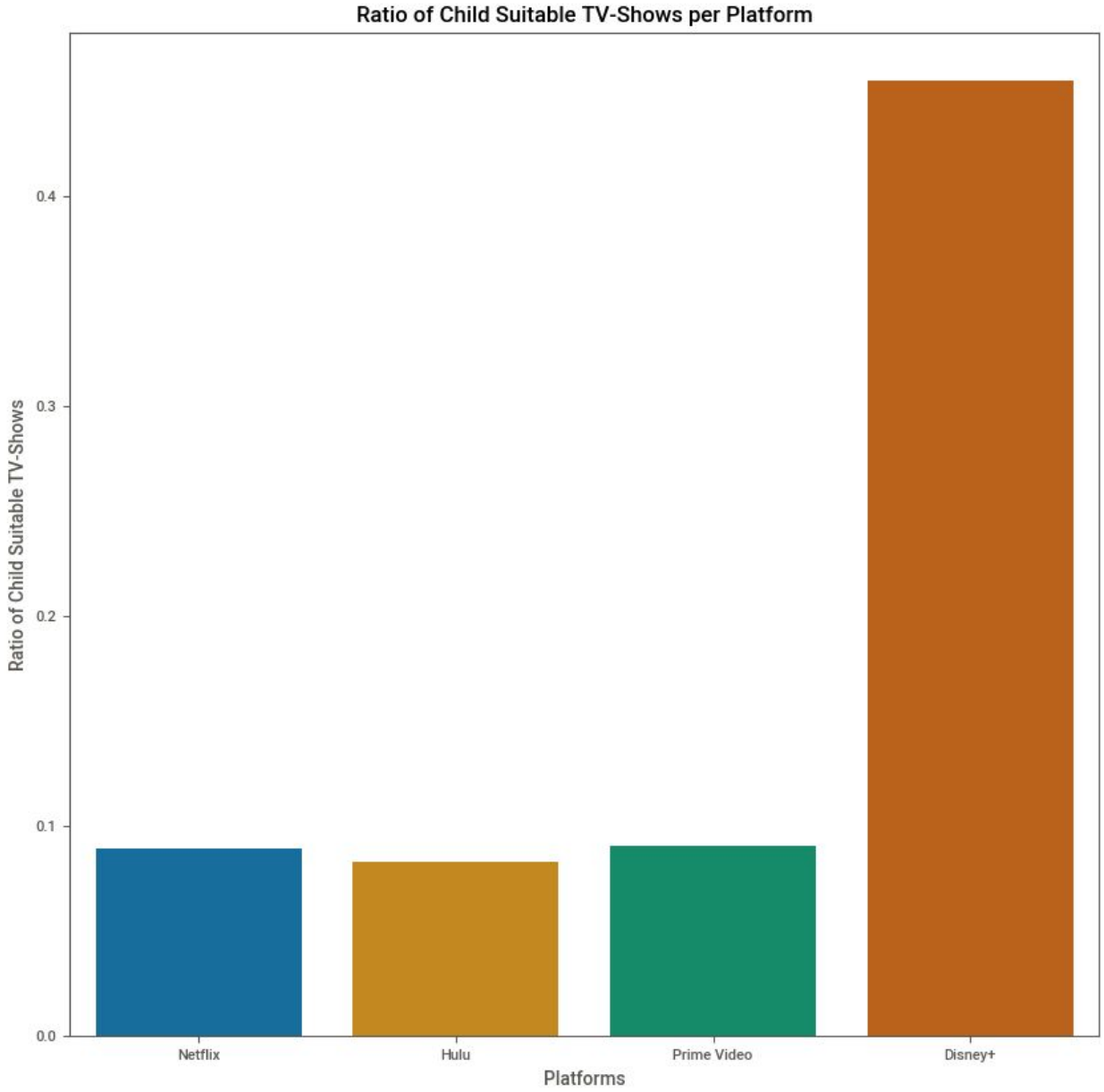

Fig 6: Bar-chart depicting the ratio of child suitable tv-shows

Fig 4 and 6:
They depict that the majority of the shows are for older audiences and only Disney+ has about 45% of the shows for children.

Fig 5:
The top words used in titles of shows are Man, Life, Love and Story.

(For more data visualization refer to Appendix)

5.2 Models Implemented

5.2.1 Correlation Analysis

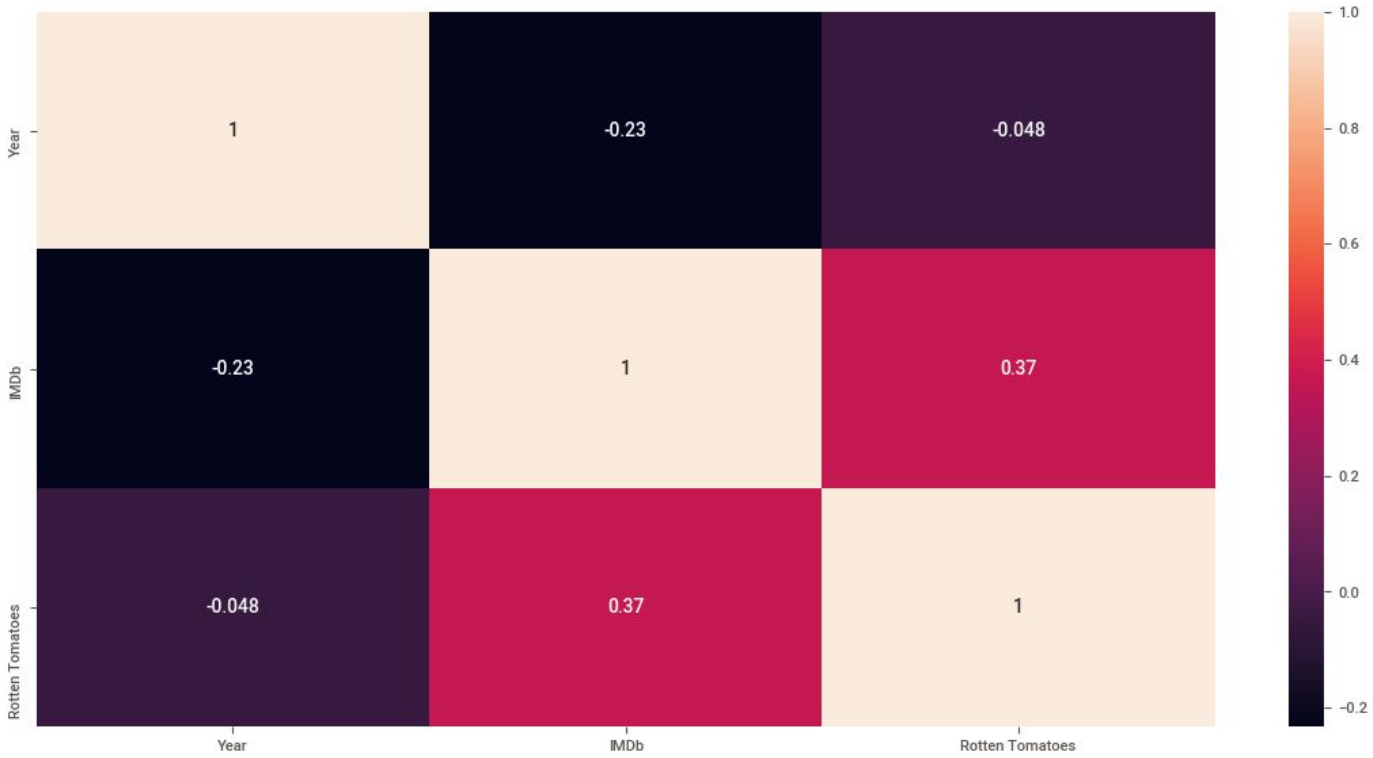

Fig 7: Correlation plot between Rotten Tomatoes, IMDB ratings and Year

By looking at the correlation plot we can clearly infer two things-

1. The IMDB and rotten tomato ratings have declined with the years
2. IMDB and Rotten tomato ratings are positively correlated i.e. a show which is well rated on one platform is well rated on the other two and any one of the ratings is reliable in determining how good the show will be.

5.2.2 Hypothesis Testing

5.2.2.1 One-Tailed Z test

Null hypothesis: The movie runtime has not increased from XX and XXI century
H0 : μx - μy <= 0

Alternative hypothesis: There's significant increase in the movie runtime from XX and XXI century
H1 : μx - μy > 0

In accordance with our result employing a one tailed Z-test does suggest that there's an increase in the movie runtime from XX and XXI century.

5.2.2.2 Two-tailed Z test

Null hypothesis: There's no significant difference in the proportion of well rated movies between movies from the XX and XXI century
H0: : μx - μy = 0

Alternative hypothesis: There's significant difference in the proportion of well rated movies between movies from the XX and XXI century
HA: μx - μy != 0

In accordance with our result employing a two tailed Z-test does suggest that there's significant difference in the proportion of well rated movies between movies from the XX and XXI century.

(For more information refer to Appendix)

5.2.3 K-means Clustering

(For more information on results of WCSS refer to Appendix)

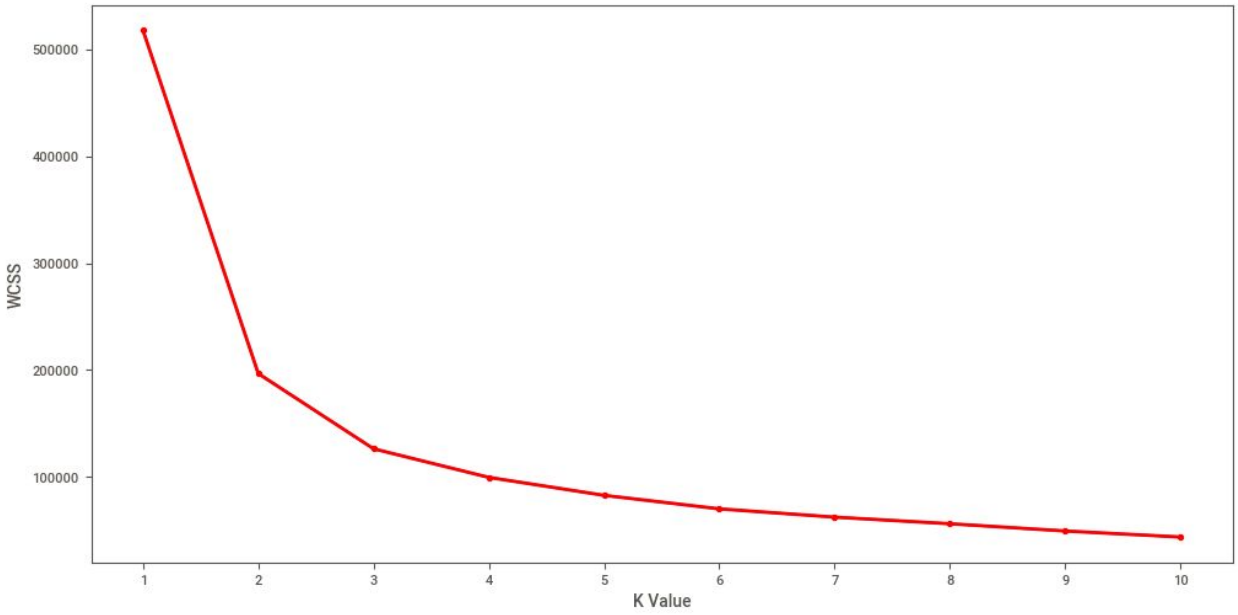

Fig 8: Elbow curve to determine optimum value of K

Elbow curve depicts optimal K value should be around 4, which I have used in my model

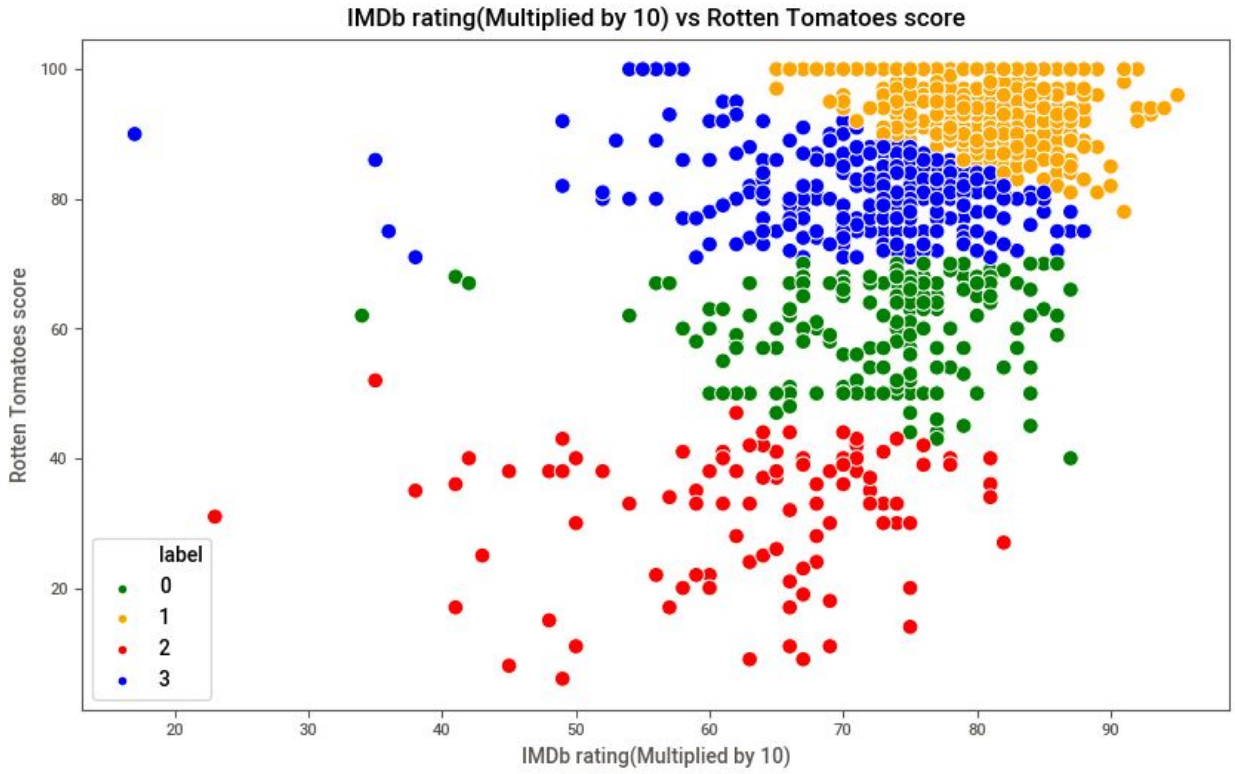

Fig 9: Scatterplot showing clustered objects with labels

Number of Cluster 0 TV Shows are 186

Number of Cluster 1 TV Shows are 423

Number of Cluster 2 TV Shows are 100

Number of Cluster 3 TV Shows are 299

5.2.4 Recommendation System(Only for Netflix shows)

```
To exit Enter "quit"

Enter The Name of a Movie or Tv Show: The Witcher
['Argon', "Marvel's Iron Fist", 'Disappearance', 'The Umbrella Academy', "Oh No! It's an Alien Invasion"]
Enter The Name of a Movie or Tv Show: Poseidon
['Monsters: Dark Continent', 'Highway', 'Faraar', 'The Legacy of a Whitetail Deer Hunter', 'Mohenjo Daro']
Enter The Name of a Movie or Tv Show: Inception
['Apollo 18', 'Forbidden Planet', 'Limitless', 'The Darkest Dawn', 'Spider-Man 3']
Enter The Name of a Movie or Tv Show: Anaconda
The movie or Tv Show does not exist

Enter The Name of a Movie or Tv Show: quit
```

Fig 10: Recommendation system example

The recommendation system here shows 5 predicted movies/tv-shows which will be similar to what the user has just watched.

6. Conclusions

6.1 Inference of Results

Implementation of different models helped to understand which type of movies/tv-shows are better for the general audience. By hypothesis testing we came to the conclusion that the duration of shows has increased from the 20th to 21st century whereas their ratings have decreased, making people less viable to watch them.

Using K-means(K=4) I clustered the shows according to their ratings and could thereby find most of the shows that come to these streaming services are well rated having rotten tomato ratings above 75 and IMDB rating above 6.5.

I made a replica of the Netflix recommendation system which could recommend 5 similar tv-shows/movies for the particular show entered. The model accuracy is above 85%

Using combination of EDA and models implemented we can finally conclude the following-

1. Best streaming service to pay a premium is Amazon prime for its quantity as well as quality in the shows. Netflix comes as a second winner.
2. The most varied content is available on Amazon Prime.
3. Netflix and Amazon Prime have the highest rated tv-shows in all four but Netflix has higher rated movies.
4. Best streaming service for kids is Disney+
5. The most common genre across all services is family.
6. Netflix has the most added shows from 2018 onwards and could be easily the second choice for viewers.

6.2 Future Scope

More dataset per streaming service could have been used in order to make a generic recommendation model, I could only find Netflix dataset to work on recommendation systems.

Addition of variables in the dataset can be introduced to have deeper insights about the streaming services and better regression analysis can be done on that dataset.

Several missing values can be encountered with their respective data from the internet which could help in not omitting some of the dataset.

7. Appendix

7.1 Data Visualization

Using sweetviz, data is visualized in a simple way. The html file is in the zipped folder for reference.

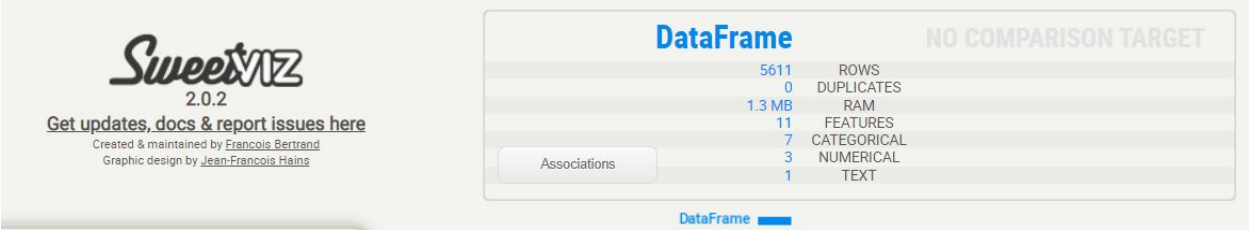

Fig 11: Sweetviz Data Visualization

7.2 Hypothesis Testing

The test statistics for one-tailed and two tailed tests are as follows-

```
H0 : μx - μy <= 0
H1 : μx - μy > 0
alpha value is : 0.05

actual_z: 1.6448536269514729
hypo_z: 2.574539594036462
Reject NULL Hypothesis
```

Fig 12: One-tailed test statistics

```
H0 : μx - μy = 0
H1 : μx - μy != 0
alpha value is : 0.05

actual_z: 1.6448536269514729
hypo_z: 6.11552315156307
Reject NULL Hypothesis
```

Fig 13: Two-tailed test statistics

7.3 K-Means Clustering

I calculated the Within Cluster Sum of Squared Errors (WSS) for different values of k and the results are as follows-

[518145.3650793651, 196771.39860139863, 126340.58764623248, 99566.8877469934, 82736.69423493248, 70099.54419778491, 62767.32824106928, 55163.77042045114, 49711.07639208555, 43718.411010708944]

By looking at the value K=4 was determined to be the optimal K value for clustering.

7.4 Recommendation System

I calculated the cosine similarity using the titles and description of each show in Netflix-

```
[[1.         0.         0.         ... 0.0942809  0.03086067 0.03390318]
 [0.         1.         0.04472136 ... 0.         0.         0.        ]
 [0.         0.04472136 1.         ... 0.         0.10141851 0.07427814]
 ...
 [0.0942809  0.         0.         ... 1.         0.         0.        ]
 [0.03086067 0.         0.10141851 ... 0.         1.         0.21971769]
 [0.03390318 0.         0.07427814 ... 0.         0.21971769 1.        ]]
```

Fig 14: Cosine Similarity Matrix

8. Bibliography and Credits

[1] Tv-shows Dataset
https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney

[2] Movies Dataset
https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

[3] Netflix Movies and Tv-shows
https://www.kaggle.com/shivamb/netflix-shows

[4] Hypothesis Testing
https://mathworld.wolfram.com/HypothesisTesting.html

[5] K-means clustering Documentation
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[6] Recommendation System Python
https://medium.com/@lope.ai/recommendation-systems-from-scratch-in-python-pytholabs-6946491e76c2

[7] Scikit learn: User Guide.
https://scikit-learn.org/stable/user_guide.html

[8] Sweetviz: User Guide.
https://pypi.org/project/sweetviz/

[9] Seaborn: User Guide.
https://seaborn.pydata.org/

[10] Rake: User Guide.
https://pypi.org/project/rake-nltk/