

Ramrao Adik Institute of Technology (Affiliated to the University of Mumbai) Dr D. Y. Patil Vidyanagar, Sector 7, Nerul, Navi Mumbai 400706.

**A Mini Project Report**

On

**“Netflix Big Data Analysis”**

Submitted by

Shreyas Churi 19CE7040

Divyansh Mishra 19CE8004

**Signature of Faculty**

**Contents:**

|  |  |  |
| --- | --- | --- |
| **Sr.No** | **Contents** | **Page No.** |
|  | Introduction | 3-4 |
|  | Project Description | 5 |
|  | Implementation Details | 6-10 |
|  | Result Analysis | 11 |
|  | Conclusion | 12 |

1. **Introduction:**

Today, Netflix is one of the most loved streaming apps in the market. With the number of users increasing every second from 115 million users, there is no doubt that this streaming channel has won the hearts of millions of people becoming the kind of streaming. Today, Netflix is one of the most loved streaming apps in the market. With the number of users increasing every second from 115 million users, there is no doubt that this streaming channel has won the hearts of millions of people becoming the kind of streaming world today world today.

Netflix’s current company valuation is $234 billion. It is currently renowned as the most valued company/media company in the world and transcends even Disney. The success lies in a secret term that is no secret (but the way it’s used in a certain way is a secret) — customer retention.

Data analysis is the process of analyzing data in various formats. Even though data is abundant nowadays, it’s available in different forms and scattered over various sources. Data analysis helps to clean and transform all this data into a consistent form so it can be effectively studied.

Once the data is cleaned, transformed, and ready to use, it can do wonders. Not only does it contain a variety of useful information, studying the data collectively results in uncovering very minor patterns and details that would otherwise have been ignored

**2. Project Description:**

* How Netflix uses data and big data analytics?
  + The question looks so simple and straightforward, but only people having the background or experience of working, studying, or playing with data can understand the depth of this question.
  + For any company or organization, data collection is essential. Imagine Netflix with its 203 million subscribers. Studying the traits of the data of this many customers would be a tremendous task. Netflix uses the collected information, converts them into insights, results, or visualizations, and recommends TV shows and movies as per customers’ preferences and interests. Just read this line again — it almost feels like a supernatural talent or power.
  + You should be able to relate if you’re a Netflix user. According to Netflix’s study, viewer activity depends on personalized recommendations and the results are true for over 75% of subscribers. Diving deeper into it, several data points have been collected and a detailed profile of each subscriber has been generated. It's hard to believe but, the profile of a subscriber created by Netflix is much more detailed than the information or preferences provided by the subscriber at the beginning of their Netflix usage
  + If I want to generalize this, data collected by Netflix is mostly about customer interaction on the application or webpage and responsiveness to shows or movies. To put it simply, if you’re watching any TV show or movie on Netflix, it knows the date, location, and device being used to watch, as well as the time of your watching. On top of that, Netflix also knows about how and when you pause and resume your shows and movies. They also take into consideration if you are completing the show or not, how many hours, days, or weeks to complete the episode or a season or a movie.
  + Netflix’s ability to collect and use the data is the reason behind its success. It results in better customer retention per year. The study says the rate of customer retention is increasing on Netflix because 80% of users follow the recommendation, and the recommended show or movie is streamed

**3. Implementation details:**

* **DATASET**
  + There are so many datasets under the topic netflix analysis. From that, we have selected a dataset "Netflix Movies and TV shows" based on the fact that more the rows the more data we can infer. This dataset consists of tv shows and movies available on Netflix as of 2020.This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Fixable which is a third-party Netflix search engine.
* Using both the classification algorithms we predicted the ratings of TV show and movies. The accuracy score is calculated for both the algorithms and as you can see based on the k value the accuracy score changes in KNN algorithm and it covered all the records within seconds which we have evaluated using confusion matrix. We also have a procedure to calculate the accurate k value which is a additional step through which we can get the k value for the dataset and calculate its accuracy score.
* But in Naive Bayes algorithm, we can predict the exact accuracy score in the first time of execution itself. Comparing the accuracy score from KNN and Naive Bayes, I conclude that the coverage of records in a lesser time with the accuracy score of 64.007 KNN algorithm predicts better than Naive Bayes algorithm for the selected dataset.
* We used the Apriori and GSP Algorithms to assess country involvement in the production process (for example, which countries produce the most movies/TV shows with the United States). According to the implementation results, 7% of all films/shows produced in the United States were co-produced with the United Kingdom. Canada holds a 6% share, while France holds a 3% share.
* International Movies appears to be the most popular genre, but this makes no sense in terms of the genre. As a result, let's take a look at the most common genre pairings with International Movies. We evaluated the most popular genre pairings with International Movies using the Apriori and GSP Algorithms. According to the implementation results, dramas account for 53% of all international films (2437). Comedy films account for 30% of all films, while action and adventure films account for 15%.
* The most popular genre of international television shows is drama (40 percent). Nonetheless, romantic and crime dramas are ranked second and third in the TV Shows category, respectively (people choose to watch detective and love dramas on TV).

**Code:**

# Visualization

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from wordcloud import WordCloud, STOPWORDS

import warnings

warnings.filterwarnings("ignore")

sns.set\_style("white")

## Loading data

df = pd.read\_csv("netflix\_titles.csv")

df.head()

df.info()

## Data Visualization

### Movies vs TV Shows

As we can see below, movies have a more expressive amount of attractions in Netflix.

types\_df = df.groupby(["type"]).size().reset\_index(name="counts")

types\_df["percent"] = round(types\_df["counts"] / sum(types\_df["counts"]), 2) \* 100

plt.figure(figsize=(12, 6))

g = sns.barplot(x="percent", y="type", palette="rocket\_r", data=types\_df, orient="h")

g.text(

0,

-0.7,

"Attraction type in Netflix by percent (%)",

fontsize=14,

fontweight="bold",

color="black",

)

col = "percent"

for i in range(2):

g.annotate(

f"{int(types\_df[col][i])}%",

xy=(types\_df[col][i] / 2, i),

ha="center",

va="center",

fontsize=50,

fontweight="bold",

color="white",

)

g.annotate(

"Movie" if i == 0 else "TV Show",

xy=(types\_df[col][i] / 2, i + 0.2),

ha="center",

va="center",

fontsize=12,

fontweight="bold",

color="white",

)

for i in ["top", "left", "right", "bottom"]:

g.spines[i].set\_visible(False)

g.set\_xlim(0, 100)

g.set(xticklabels=[], yticklabels=[])

plt.ylabel("")

plt.xlabel("")

### Attractions ratings

We can also see the difference between the ratings of each one!

plt.figure(figsize=(14, 7))

g = sns.countplot(df["rating"], hue=df["type"], palette="rocket\_r")

g.text(

0,

2200,

"Rating by attraction type in Netflix",

fontsize=18,

fontweight="bold",

color="black",

)

for i in ["top", "left", "right"]:

g.spines[i].set\_visible(False)

plt.legend(loc="upper right", frameon=False, prop={"size": 15})

plt.xlabel("Rating", fontfamily="Fira Sans")

plt.ylabel("")

### Movies and TV Shows production

movies\_df = df[df["type"] == "Movie"]

top5countries\_movies = (

movies\_df.groupby(["country"]).size().sort\_values(ascending=False)[0:5]

)

plt.figure(figsize=(12, 6))

g = sns.barplot(x=top5countries\_movies.index, y=top5countries\_movies, palette="rocket")

g.text(

0,

2200,

"Top 5 countries in Netflix movie production",

fontsize=18,

fontweight="bold",

color="black",

)

for i in ["top", "left", "right"]:

g.spines[i].set\_visible(False)

for i in g.patches:

g.text(

i.get\_x() + i.get\_width() / 2.5,

i.get\_height() + 60,

round(i.get\_height()),

fontsize="18",

)

g.set(yticklabels=[])

plt.xlabel("")

plt.ylabel("")

Looking at the top 5 countries again, but for TV Shows, in the chart below we can notice USA still leads the productions!

tvshows\_df = df[df["type"] == "TV Show"]

top5countries\_tvshows = (

tvshows\_df.groupby(["country"]).size().sort\_values(ascending=False)[0:5]

)

plt.figure(figsize=(12, 6))

g = sns.barplot(

x=top5countries\_tvshows.index, y=top5countries\_tvshows, palette="rocket"

)

g.text(

0,

900,

"Top 5 countries in netflix TV Show production",

fontsize=18,

fontweight="bold",

color="black",

)

for i in ["top", "left", "right"]:

g.spines[i].set\_visible(False)

for i in g.patches:

g.text(

i.get\_x() + i.get\_width() / 2.5,

i.get\_height() + 60,

round(i.get\_height()),

fontsize="18",

)

g.set(yticklabels=[])

plt.xlabel("")

plt.ylabel("")

The time Netflix invests more in movies or TV shows

df["date\_added"] = pd.to\_datetime(df["date\_added"])

df["year\_added"] = df["date\_added"].dt.year

index = [2014, 2015, 2016, 2017, 2018, 2019, 2020]

ts\_df = df[df["year\_added"] > 2013]

ts\_df = (

ts\_df[ts\_df["year\_added"] < 2021]

.groupby("type")["year\_added"]

.value\_counts()

.unstack()

.T

)

fig, ax = plt.subplots(1, 1, figsize=(12, 6))

fig.text(

0.15,

1,

"Amount of Movies and TV Shows added over years",

fontsize=18,

fontweight="bold",

color="black",

)

g1 = sns.lineplot(x=ts\_df.index, y=ts\_df["Movie"], color="#593262", label="Movie")

g2 = sns.lineplot(x=ts\_df.index, y=ts\_df["TV Show"], color="#db6e59", label="TV Show")

for i in ["top", "left", "right"]:

ax.spines[i].set\_visible(False)

ax.legend(loc="upper left", frameon=False, prop={"size": 15})

plt.xlabel("")

plt.ylabel("")

def quarter\_expression(i):

if i == 1:

return "1st Quarter"

elif i == 2:

return "2nd Quarter"

elif i == 3:

return "3rd Quarter"

else:

return "4th Quarter"

df["quarter\_added"] = df["date\_added"].dt.quarter

qrt\_df = df["quarter\_added"].value\_counts()

prct\_qrt = pd.DataFrame(round(qrt\_df / sum(qrt\_df), 2)).T

fig, ax = plt.subplots(1, 1, figsize=(12, 4))

ax.barh(prct\_qrt.index, prct\_qrt[1.0], color="#593262", alpha=0.9)

ax.barh(prct\_qrt.index, prct\_qrt[2.0], color="#772b58", alpha=0.9, left=prct\_qrt[1.0])

ax.barh(

prct\_qrt.index,

prct\_qrt[3.0],

color="#b53158",

alpha=0.9,

left=prct\_qrt[1.0] + prct\_qrt[2.0],

)

ax.barh(

prct\_qrt.index,

prct\_qrt[4.0],

color="#db6e59",

alpha=0.9,

left=prct\_qrt[1.0] + prct\_qrt[2.0] + prct\_qrt[3.0],

)

fig.text(

0.15,

1,

"Percentage of attractions added per quarter",

fontsize=18,

fontweight="bold",

color="black",

)

sum\_list = []

for i in range(1, 5):

ax.annotate(

f"{int((prct\_qrt[i][prct\_qrt.index])\*100)}%",

xy=(sum(sum\_list) + prct\_qrt[i][prct\_qrt.index] / 2, prct\_qrt.index),

ha="center",

va="center",

fontsize=40,

fontweight="bold",

color="white",

)

ax.annotate(

quarter\_expression(i),

xy=(sum(sum\_list) + prct\_qrt[i][prct\_qrt.index] / 2, -0.25),

ha="center",

va="center",

fontsize=15,

fontweight="bold",

color="white",

)

sum\_list.append(prct\_qrt[i][prct\_qrt.index])

for i in ["top", "left", "right", "bottom"]:

ax.spines[i].set\_visible(False)

ax.set(yticklabels=[], xticklabels=[])

ax.set\_xlim(0, 1)

# Genres

df["genre"] = df["listed\_in"].apply(

lambda x: x.replace(" TV", "")

.replace("TV ", "")

.replace(" ,", ",")

.replace(", ", ",")

.split(",")

)

stopwords = set(STOPWORDS)

stopwords.update(["Shows", "Movies", "British", "International"])

all\_summary = " ".join(" ".join(s) for s in df["genre"])

wordcloud = WordCloud(

stopwords=stopwords,

background\_color="white",

colormap="rocket",

width=1600,

height=800,

).generate(all\_summary)

fig, ax = plt.subplots(figsize=(10, 6))

fig.text(

0.15,

1,

"Word cloud of Netflix attractions genres",

fontsize=18,

fontweight="bold",

color="black",

)

ax.imshow(wordcloud, interpolation="bilinear")

ax.set\_axis\_off()

plt.imshow(wordcloud)

**4. Result and Analysis:**

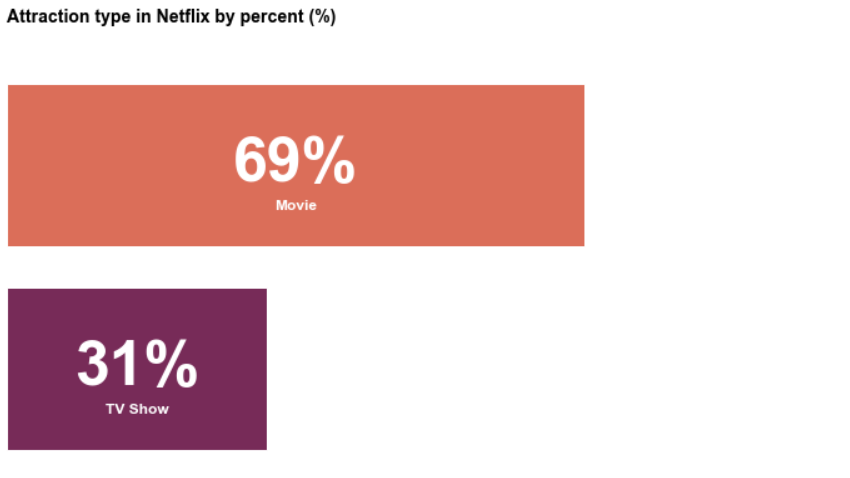


Fig 1: Attraction type in Netflix by percent

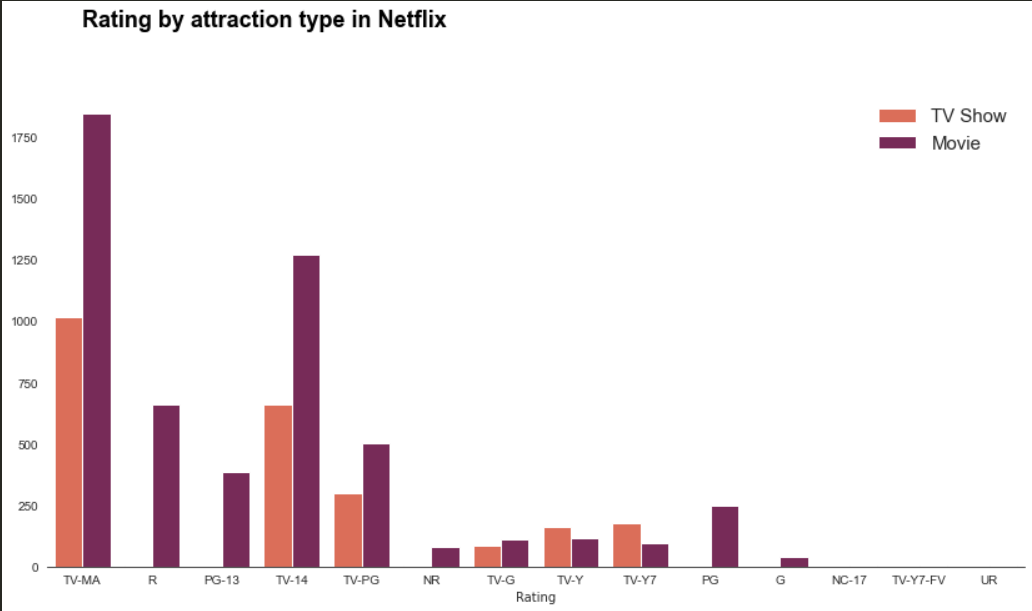


Fig 2: Rating by attraction type in Netflix

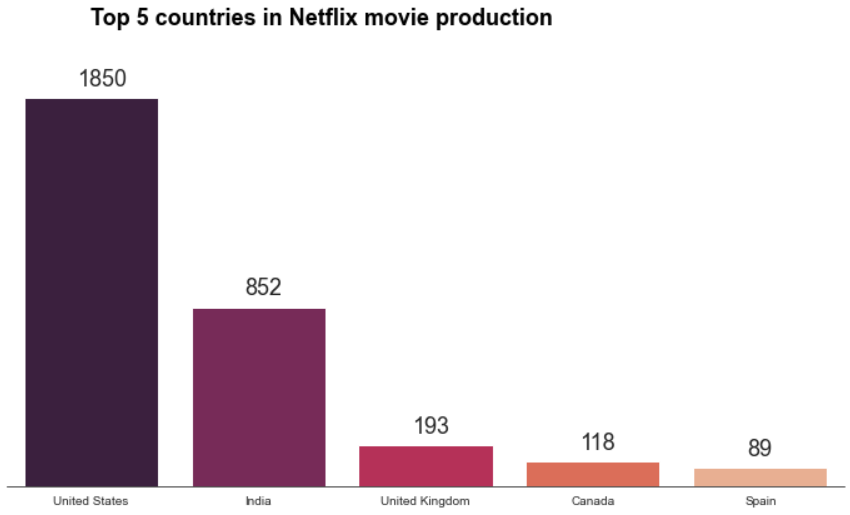


Fig 3: Top 5 countries in Netflix movie production

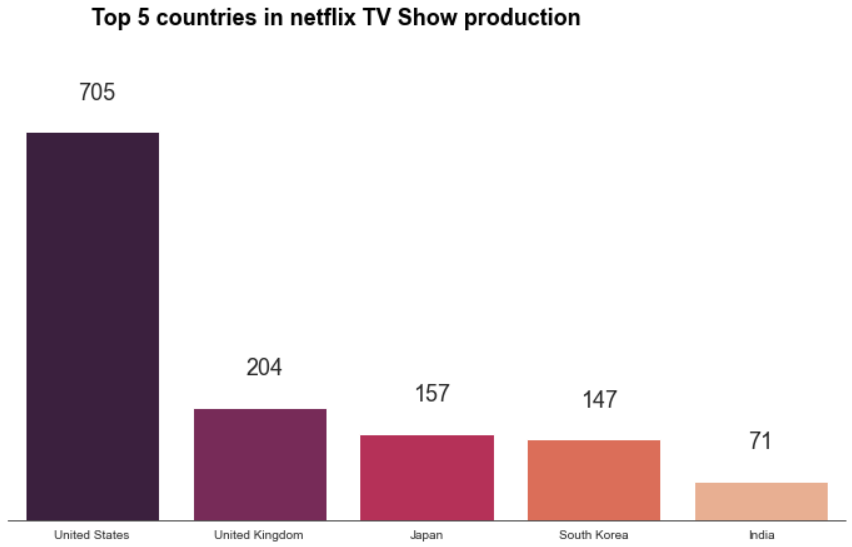


Fig 4: Top 5 countries in Netflix TV show production

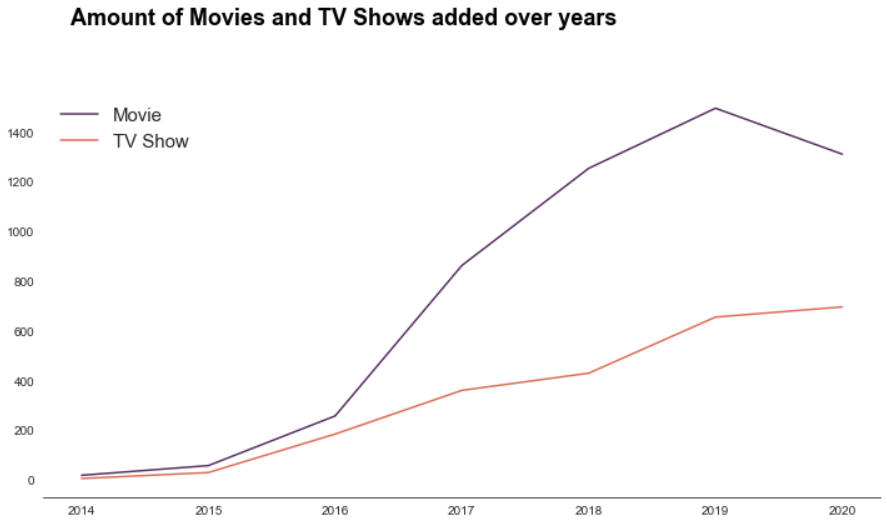


Fig 5: Amount of movies and TV Shows added over years



Fig 6: Word Cloud of Netflix attractions genres

**5. Conclusion:**

* This is nothing but tracking the actions of subscribers and collecting their data based on this. One technique that is very traditional and Netflix uses that too is to take feedbacks from subscribers. The feedback is then converted into a rating and then the team works on system improvement or recommendations.
* In future, the Results can be used for creating a recommendation system