



Project ID: 2020CSEPID10

**Project Report
on
"A Data Analysis on Netflix Using Python"**

**Submitted in Partial Fulfilment of the Requirement
For the Degree of
Bachelor of Technology
In
Computer Science and Engineering**

**By
Abhinav Gupta (2002900100003)**

**Under the Supervision
of
Dr. Shipra Saraswat**

**ABES INSTITUTE OF TECHNOLOGY, GHAZIABAD
AFFILIATED TO**

**Dr. A.P.J. ABDUL KALAM TECHNICAL
UNIVERSITY, UTTAR PRADESH,
LUCKNOW (ODD SEM, 2022-23)**

DECLARATION

I hereby declare that the work presented in this report entitled “**A Data Analysis on Netflix Using Python**”, was carried out by me. I have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute. I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable.

Name: Abhinav Gupta

Roll. No: 2002900100003

Branch: Computer Science and Engineering

(Candidate Signature)

CERTIFICATE

Certified that **Abhinav Gupta** (Roll no. 2002900100003) have carried out the research work presented in this thesis entitled “**A Data Analysis on Netflix Using Python**” for the award of **Bachelor of Technology** from Dr. APJ Abdul Kalam Technical University, Lucknow under my/our (print only that is applicable) supervision. The report embodies results of original work, and studies are carried out by the student himself/herself (print only that is applicable) and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Supervisor Signature

Dr. Shipra Saraswat
Associate Professor
ABES Institute of Technology, Ghaziabad

HOD Signature

Dr. Rizwan Khan
Head of Department (CSE)
ABES Institute of Technology

ABSTRACT

As we all know in today's world with day-to-day advancement in technologies people are searching for more and more easy and convenient ways to operate, So, we decided to work on making 'Data Analysis on Netflix' process simpler and more convenient. By creating this we are also using and improving our knowledge to do some real-world application. We can say that data visualization is basically a graphical representation of data and information. It is mainly used for data cleaning, exploratory data analysis, and proper effective communication with business stakeholders. Right now, the demand for data scientists is on the rise. Day by day we are shifting towards a data-driven world. It is highly beneficial to be able to make decisions from data and use the skill of visualization to tell stories about what, when, where, and how data might lead us to a fruitful outcome.

Data Analysis and visualization is going to change the way our analysts work with data. They're going to be expected to respond to issues more rapidly. And they'll need to be able to dig for more insights – look at data differently, more imaginatively.

ACKNOWLEDGEMENT

With deep gratitude I express my earnest thanks to my esteemed supervisor Dr. Shipra Saraswat, Associate Professor, Department of Computer Science & Engineering for his constant involvement, energetic efforts and proficient guidance, which gave me direction and body to work, respond here. Without his counsel and encouragement, it would have been impossible to complete the thesis work in this manner.

I wish to express my gratitude to Dr. Rizwan Khan (Head of Department), and Dr. Manish Kumar Jha (Director), for their support, guidance and advice throughout this work. I am thankful to all the faculty members of the Computer Science and Engineering Department especially for their intellectual support during my research work. I also want to thank my friends for their valuable support whenever I needed. I would like to thank all those people who have helped me some way or the other in my thesis work.

Lastly, and most importantly, I thank my parents for their moral support and encouragement towards completing my thesis successfully. In the last, I want to thank Almighty God.

Abhinav Gupta
B.Tech (Third Year)
Roll No: 2002900100003
Computer Science & Engineering
ABES Institute of Technology, Ghaziabad

TABLE OF CONTENT

Description

- Declaration
- Certificate
- Abstract
- Acknowledgment
- Table Of Content
- Introduction
- Purpose of Project
- Methodology
- Result
- Conclusion
- References

Introduction

Analytics is all about solving problems and Data analytics is the soul of the internet of things (IoT) technology. Analytics is everywhere, this could be working in a variety of different industries such as aviation, industries or government. With so many organizations looking to capitalize on data to improve their processes, it's a hugely exciting time to start a career in analytics. Data Analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

We first wanted to get an overview of the dataset that we were dealing with. First, we loaded up tidy verse for a simple data analysis purpose. We got the dataset from Kaggle and we are going to utilize data that the Kaggle website provides to understand the trend of movies and TV shows released on the platform.

This dataset consists of from the code, we could see the column names that the CSV file contains.

Purpose of The Project

The purpose of any project is to benefit the society in an effective way. The purpose of building this project is to solve real world problem and expand our knowledge. This project will help to do an extensive analysis of the data, to have a better recommendation for the subscribers and to add more content in a way so that more subscribers are added.

Aims

1. Expand our knowledge and contribute to society.
2. Build a real-world project.

Objective

Some of the most important tasks that we can analyze from Netflix data are:

1. To learn the common trends of users in a region.
2. To understand what content is available.
3. To understand the similarities between the content.
4. To understand what exactly Netflix is focusing on.
5. To do Sentiment Analysis.

Research Approach/Methodology

In B-tech Third year we decided to make a project through which we can do some real-world task.

After some brainstorming, we landed up on making a Data Analysis on Netflix Using Python. Our work distinguishes the type of content that is available on Netflix, the similarities between the content and what the ultimate goal of Netflix is and could be planning. Not only that, but also gives new content creators and filmmakers the opportunity to experiment with the users to give them a better experience altogether. Sentiment analysis, also referred to as opinion mining. Through this organizations can determine and categorize opinions about a product, service, or idea. In turn helping the end users get a better watching experience. This gives a clear idea on how data analysis can aid in the prediction and development of various technology and industries of new era.

There are mainly 5 modules in this project they are as follows:

Numpy: Used for making Linear Algebraic Calculation

Pandas: Used to read and prepare Data

Plotly: Used for Data Visualization

Kaggle: To Obtain latest Data Netflix

Data set: To read the available Data

Project Background

The report presents a data analysis research and coordination activity one in the Netflix dataset found on Kaggle to establish a new data exploration on the trend of movies on the platform. When Netflix was first launched, it did not have any data analysis to understand the trend of audience/users using the platform as previously mentioned. As time passes by, the importance of utilizing data analysis started to emerge, and the biggest hit shows were released on the platform after scrutinizing the data the company collected from the users. In the project, the team is working on a same procedure to understand the trend of the platform and attempt to understand the average time the platform takes to upload the contents, the top ten film directors in top ten countries where the platform is streaming, and the top genres that the top actors/actresses are in. The team will, then, attempt to understand the overall trend of the Netflix platform according to the dataset the team got from Kaggle.

Methodology

Before initiating the data analysis, the dataset has been separated into two different sets: one is for the movies only and another dataset is TV shows only. Every date on the dataset is transformed to day and year format.

```
#To read the data set
dff=pd.read_csv('netflix_titles.csv')
dff.columns
print(dff.columns)

#We use pandas here. dff gives the dataframe
#we use dff columns here to get access the columns

#To get distribution of content type

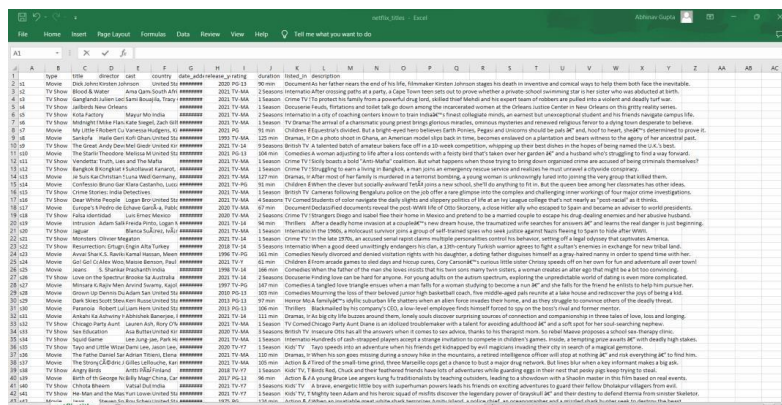
z = dff.groupby(['rating']).size().reset_index(name='counts')

#groupby by splitting used to split obj and combine res
#resindex used to treat index as columns

pieChart = px.pie(z, values='counts', names='rating',
                  title='Content Ratings on Netflix', labels='Deon',
                  color_discrete_sequence=px.colors.qualitative.Set3)
pieChart.show()
```

Figure 1: Code for Reading The Dataset

It will read the data from datasets and perform the analysis based on the data given in the figure 2 below:



	title	director	cast	country	date_added	rating	duration	listed_in	description
1	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
2	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
3	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
4	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
5	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
6	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
7	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
8	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
9	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.
10	13	TV Show	David Tennant	United Kingdom	2007-10-12	TV-14	45 min	Doctor Who	David Tennant stars as the 10th Doctor, a time-traveling alien who saves the world from the evil Daleks.

Figure 2: Dataset of Netflix Used in this Project

Tkinter GUI (Graphical User Interface of code)

If want to provide GUI for non-programmers to interact with your code. For creating a GUI in python, Tkinter library is one of the best libraries. So, I made a graphical user interface using Tkinter for code as below (Figure 3).

```
from tkinter import *
from tkinter import messagebox

top = Tk()

#THE MAIN WINDOW

top.title("Netflix Data Analysis")
window_width = 500
window_height = 300

# get the screen dimension
screen_width = top.winfo_screenwidth()
screen_height = top.winfo_screenheight()

# find the center point
center_x = int(screen_width/2 - window_width / 2)
center_y = int(screen_height/2 - window_height / 2)

# set the position of the window to the center of the screen
top.geometry(f'{window_width}x{window_height}+{center_x}+{center_y}')

# NO RESIZE
top.resizable(False, False)

#ICON
top.iconbitmap("./assets/netflix.ico")
```

Figure 3: Code for GUI for non-programmers to interact:

When you run the above code, a GUI pops up as shown below in figure 4

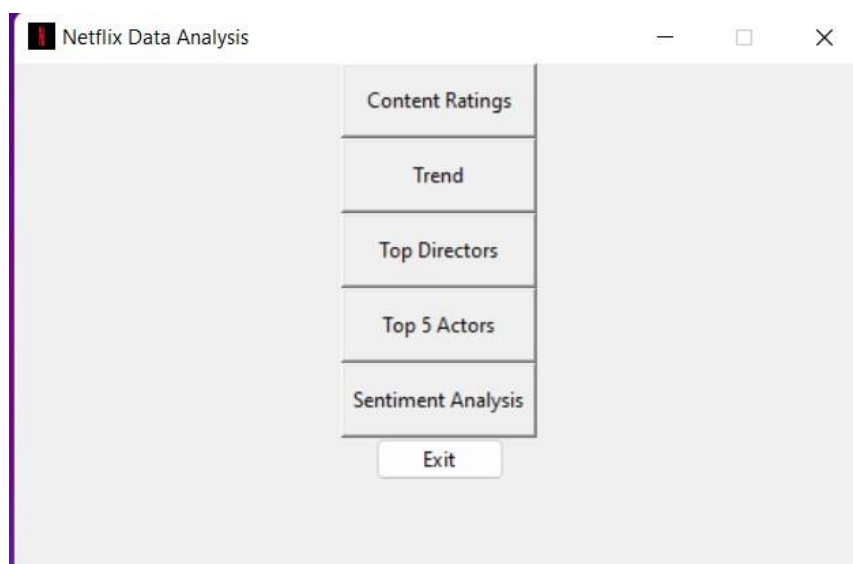


Figure 4: Result of Program

A GUI is like code running in an infinite loop. In the above code, then python program lies in wait for events like click on buttons, menus, mouse hover etc. In ourcode the button named ‘Content Ratings’ is linked to analysis function. In figure 5, it clearly shows when a user clicks on ‘Content Ratings’ button then analysis function is called.

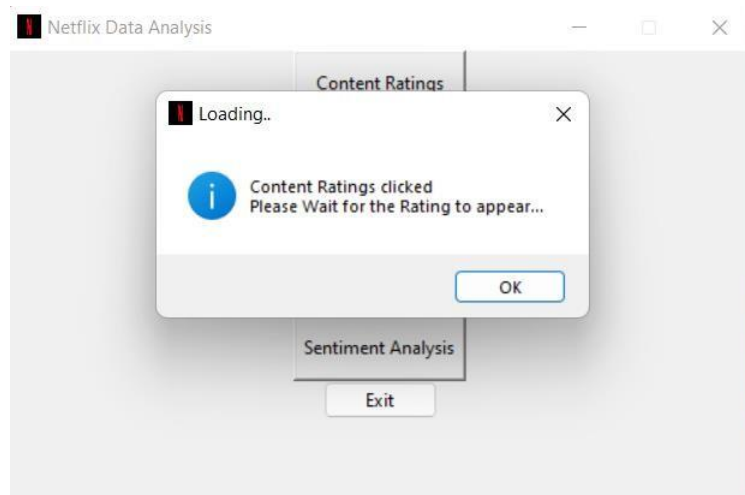


Figure 5: User Clicks on button Content Ratings

Analysis function asks for data file through windows prompt. Once the analysis isdone by fetching the file, the data is processed in the new window. Here are the Results in figure 6;

Content Ratings on Netflix

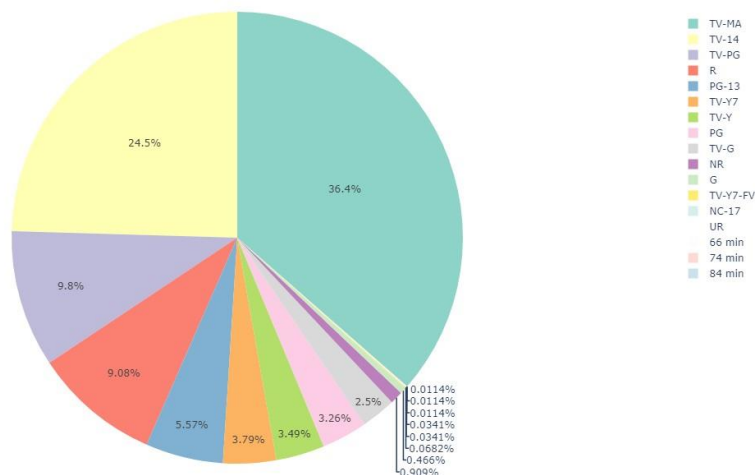


Figure 6: Graph Representation of Netflix Content

This graph (figure 4) shows that most (36.4%) of the content are rated as TV-MA which means the contents are intended for viewing by mature and adult audiences.

the second most (24.5%) number of the contents are rated as TV-14 which means these contents are made for audiences of age 14 or above.

Similarly For other Results, Trend Button is Clicked!

Release year is very important aspect of learning when a content on the platform is released and the average time the platform takes to upload the content to know about its Trend, the same depicts the code in figure 7.

Code:

```
#Analyzing content

df1=dfff[['type','release_year']]
df1=df1.rename(columns={"release_year": "Release Year"})

df2=df1.groupby(['Release Year','type']).size().reset_index(name='Total Content')

df2=df2[df2['Release Year']>=2010]

fig3 = px.line(df2, x="Release Year", y="Total Content", color='type',title='Trend of content produced over the years on Netflix')
fig3.show()
```

Figure 7: Code for Analyzing Content

The plotted graph in figure 8 indicates that most of the movies were released after the year 2000s. From 2010 to 2020, there is an exponential growth in the graph, which indicates that the number of releases that the movies were introduced to the platform started to grow higher.

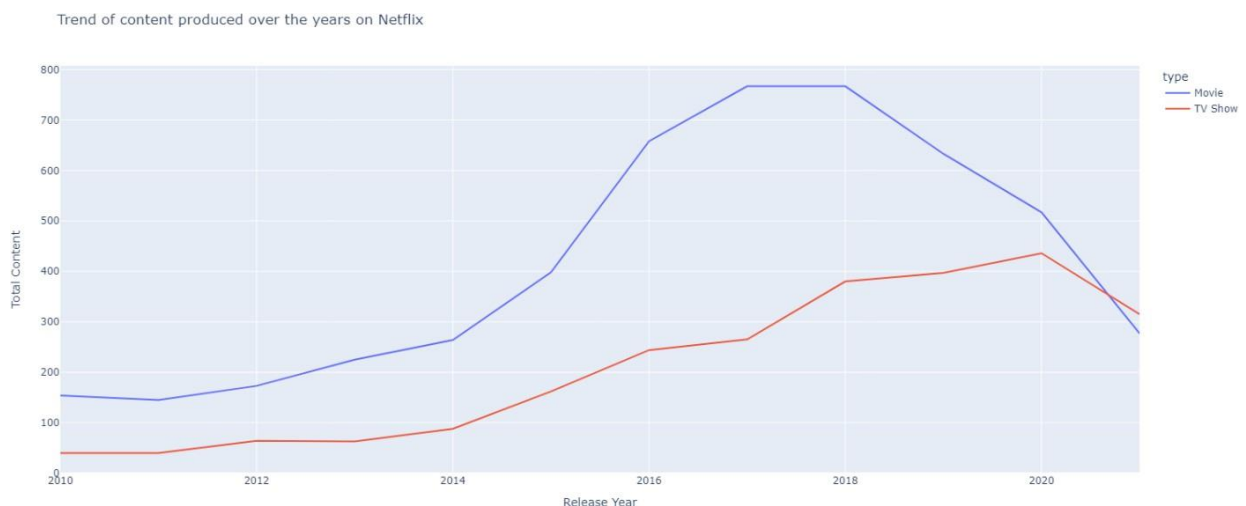


Figure 8: Trend of content produced over the years on Netflix

We can clearly see that from 2018 Netflix faced a heavy decline in movie production, it went further in 2020.

For the TV sections it was increasing until 2020, then again, a huge decline

These are some of the charts which shows the data in the form of charts of last few years, it gives incredible understanding of the raw datasets.

Similarly for Button clicked on Directors, the Analysis is done and code for the same is shown in figure 9.

```
data['director'] = data['director'].fillna('No director specified')
filtered_directors = pd.DataFrame()
filtered_directors = data['director'].str.split(',', expand = True).stack()
filtered_directors = filtered_directors.to_frame()
filtered_directors.columns = ['Director']
directors = filtered_directors.groupby(['Director']).size().reset_index(name = 'Total_Content')
directors = directors[directors.Director != 'No director specified']
directors = directors.sort_values(by = ['Total_Content'], ascending = False)
directorsTop5 = directors.head()
directorsTop5 = directorsTop5.sort_values(by = ['Total_Content'])
fig = px.bar(directorsTop5, x = 'Total_Content', y = 'Director', title = 'Top 5 directors on Netflix', color_discrete_sequence = ['blue'])
fig.show()
```

Figure 9: Code for Button clicked on Directors

Top 5 successful directors in this platform

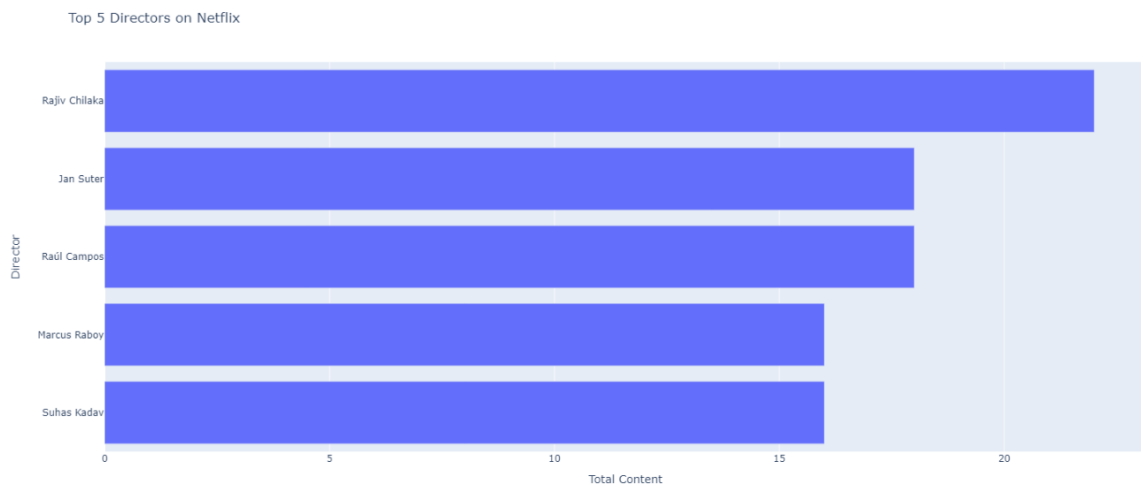


Figure 10: Bar representation of the graph on Directors on Netflix

You can directly understand it by the bar representation of the graph from figure 10. From the above graph it is derived that top 5 directors of Netflix are:

1. Rajiv Chilaka
2. Jan Suter
3. Raul Campos
4. Marcus Raboy
5. Suhas Kadav

Button Clicked on Actor, following code is given and shown in figure 11 and figure 12 displays its results in the form of bar representation:

```
data['cast'] = data['cast'].fillna('No cast specified')
filtered_cast = pd.DataFrame()
filtered_cast = data['cast'].str.split(',', expand = True).stack()
filtered_cast = filtered_cast.to_frame()
filtered_cast.columns = ['Actor']
actors = filtered_cast.groupby(['Actor']).size().reset_index(name = 'Total_Content')
actors = actors[actors.Actor != 'No cast specified']
actors = actors.sort_values(by = ['Total_Content'], ascending = False)
actorsTop5 = actors.head()
actorsTop5 = actorsTop5.sort_values(by = ['Total_Content'])
fig = px.bar(actorsTop5, x = 'Total_Content', y = 'Actor', title = "top 5 actors on Netfli
x", color_discrete_sequence = ['green'])
fig.show()
```

Figure 11: Code for Button Clicked on Actor

Result Obtained:

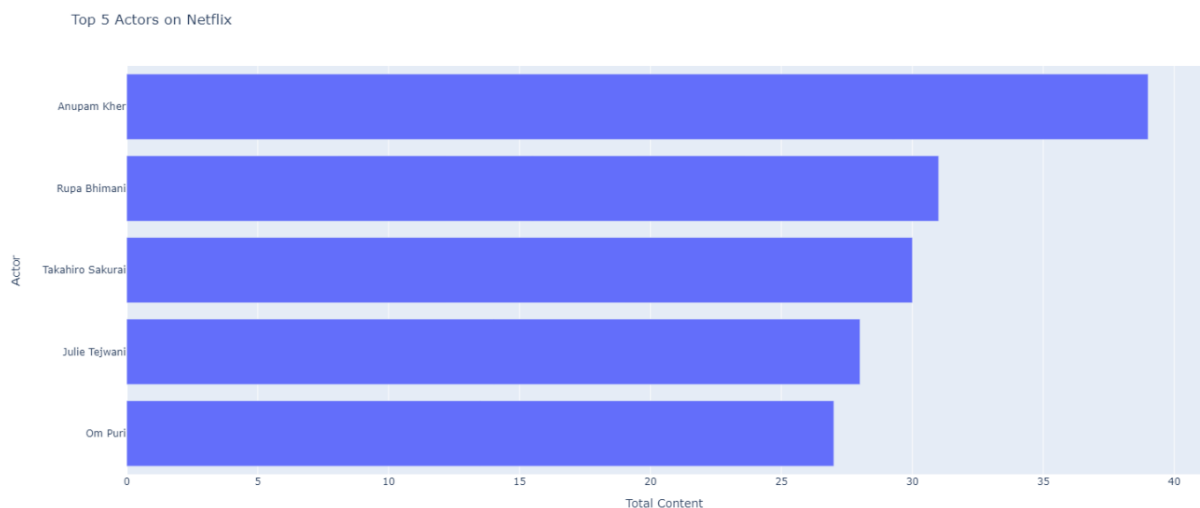


Figure 12: Bar representation of the graph on Actors on Netflix

Sentimental Analysis

In the following, I will show you how to implement a model that can classify *Netflix* reviews as positive or negative. The model will take a whole review as an input (word after word) and provide percentage ratings for checking whether the review conveys a positive or negative sentiment.

Code for Sentimental Analysis in figure 13 shown below:

```
dfx = data[['release_year', 'description']]
dfx = dfx.rename(columns = {'release_year': 'Release year'})
for index, row in dfx.iterrows():
    z = row['description']
    testimonial = TextBlob(z)
    p = testimonial.sentiment.polarity
    if p == 0:
        sent = 'Neutral'
    elif p > 0:
        sent = 'Positive'
    else:
        sent = 'Negative'
    dfx.loc[[index, 2], 'Sentiment'] = sent

dfx = dfx.groupby(['Release year', 'Sentiment']).size().reset_index(name = 'Total_Content')
dfx = dfx[dfx['Release year'] > 2010]
fig = px.bar(dfx, x = 'Release year', y = 'Total_Content', color = 'Sentiment', title = 'Sentiment of content on Netflix')
fig.show()
```

Figure 13: Code for Sentimental Analysis

Result

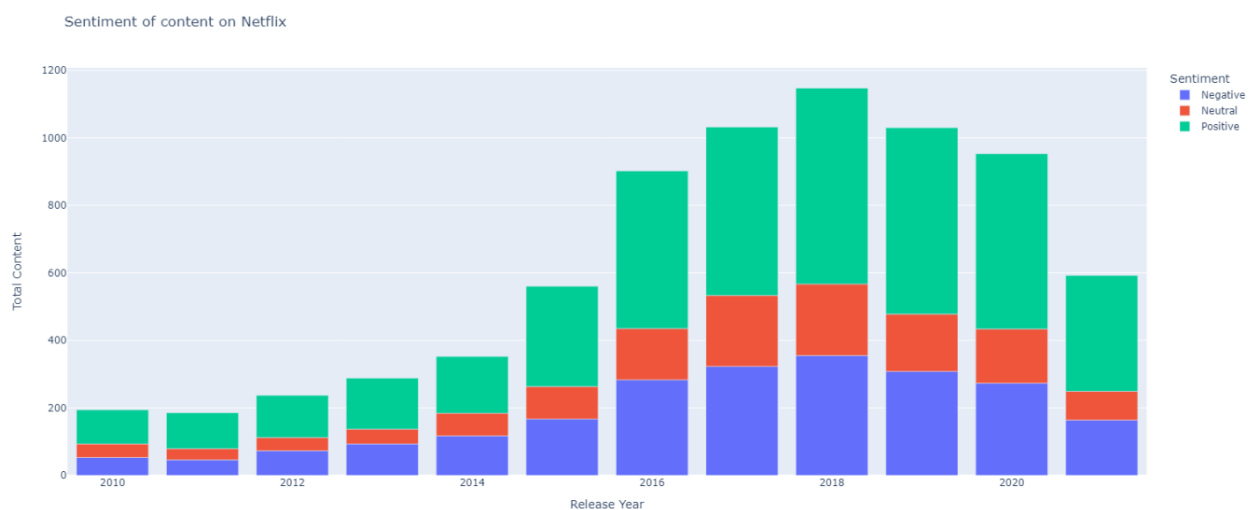


Figure 14: Sentiment of content on Netflix

Sentiment Analysis can help to automatically transform the **unstructured information into structured data** of public opinions about products, services, brands, politics or any other topic that people can express opinions about you can see in figure 14. This data can be very useful for commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, product feedback, and customer service.

Conclusion

To conclude, we have constructed a relatively accurate data analysis to determine the Content Analysis of a given TV show and movie. Processing and narrowing down the features of the Netflix Dataset by identifying the Trends indicates that most of the movies were released after the year and the top directors on Netflix with Actor the audience watch and feeding this data to figure out the trend over the years, we were able to identify which actors/actresses by which directors were popular in country. Steps such as expanding the dataset and feature set and using logistic regression might have the potential to improve upon on results in the future. Our work distinguishes the type of content that is available on Netflix, the similarities between the content and what the ultimate of goal of Netflix is and could be planning. Not only that, but also gives new content creators and filmmakers the opportunity to experiment with the users to give them a better experience altogether. Sentiment analysis, also referred to as opinion mining. Through this organizations can determine and categorize opinions about a product, service, or idea. In turn helping the end users get a better watching experience. This gives a clear idea on how data analysis can aid in the prediction and development of various industries.

References

- [1] Fernández-Manzano, Eva-Patricia, Elena Neira, and Judith Clares-Gavilán. "Data management in audiovisual business: Netflix as a case study." *El profesional de la información (EPI)* 25.4 (2016): 568-576.
- [2] Fernández-Manzano, E. P., Neira, E., & Clares-Gavilán, J. (2016). Data management in audiovisual business: Netflix as a case study. *El profesional de la información (EPI)*, 25(4), 568-576.
- [3] BAGKAR, P., BORUDE, A., & AGA, Z. (2021). Sentiment Analysis on Netflix.
- [4] Fernández-Manzano, Eva-Patricia, Elena Neira, and Judith Clares-Gavilán. "Data management in audiovisual business: Netflix as a case study." *El profesional de la información (EPI)* 25, no. 4 (2016): 568-576.
- [5] BAGKAR, PRITI, AISHWARYA BORUDE, and ZARRIN AGA. "Sentiment Analysis on Netflix." (2021).
- [6] Fernández-Manzano, E.P., Neira, E. and Clares-Gavilán, J., 2016. Data management in audiovisual business: Netflix as a case study. *El profesional de la información (EPI)*, 25(4), pp.568-576.
- [7] Novendri, Risky, et al. "Sentiment analysis of YouTube movie trailer comments using naïve bayes." *Bulletin of Computer Science and Electrical Engineering* 1.1 (2020): 26-32.
- [8] Novendri, R., Callista, A. S., Pratama, D. N., & Puspita, C. E. (2020). Sentiment analysis of YouTube movie trailer comments using naïve bayes. *Bulletin of Computer Science and Electrical Engineering*, 1(1), 26-32.
- [9] Novendri, Risky, Annisa Syafarani Callista, Danny Naufal Pratama, and Chika Enggar Puspita. "Sentiment analysis of YouTube movie trailer comments using naïve bayes." *Bulletin of Computer Science and Electrical Engineering* 1, no. 1 (2020): 26-32.
- [10] Fernández-Manzano EP, Neira E, Clares-Gavilán J. Data management in audiovisual business: Netflix as a case study. *El profesional de la información (EPI)*. 2016;25(4):568-76.