



VNR Vignana Jyothi Institute of Engineering and Technology

(Affiliated to J.N.T.U, Hyderabad)

Bachupally(v), Hyderabad, Telangana, India.

NETFLIX DATASET ANALYSIS

**A course based project submitted in partial fulfilment of the requirements
for the award of the degree of
BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND BUSINESS SYSTEMS**

Submitted by:

Tejaswi Gaddam(21071A3219)

Pranav Guddeti(21071A3224)

Sravanth Mirtipati(21071A3245)

Mohammad Abdul Sami (21071A3247)

P.Manhith Sai(21071A3253)

P.Deepika Lalitha Sai (21071A3255)

Under the guidance of:

**Mr. P. Venkateswara Rao -Assistant Professor Dept. of Computer Science and
Engineering**



**VNR Vignana Jyothi Institute of Engineering and Technology (Affiliated
to J.N.T.U, Hyderabad)**

Bachupally(v), Hyderabad, Telangana, India.

CERTIFICATE

This is to certify that Tejaswi Gaddam(21071A3219)
Pranav Guddeti(21071A3224) ,Sravanth Mirtipati(21071A3245),
Mohammad Abdul Sami (21071A3247) , P. Manhith Sai(21071A3253),
P.Deepika Lalitha Sai (21071A3255) have completed their course based
project work (Computational Statistics) at CSE Department of VNR
VJIET, Hyderabad entitled:

" **NETFLIX** DATA ANALYSIS " in partial fulfilment of the requirements
for the award of B.Tech degree during the academic year 2021-2022.
This work is carried out under my supervision and has not been
submitted to any other University/Institute for award of any
degree/diploma.

Mr. P. Venkateswara Rao

Assistant Professor

CSE Department

VNRVJIET

Dr.S Nagini

Professor and HOD

CSE Department

VNRVJIET

DECLARATION

This is to certify that our project report titled “NETFLIX DATA ANALYSIS” submitted to Vallurupalli Nageswara Rao Institute of Engineering and Technology in complete fulfilment of requirement for the award of Bachelor of Technology in Computer Science and Engineering is a bona fide report to the work carried out by us under the guidance and supervision of Mr.P.Venkateswara Rao, Assistant Professor, Department of Computer Science and Engineering, Vallurupalli Nageswara Rao Institute of Engineering and Technology. To the best of our knowledge, this has not been submitted in any form to other university or institution for the award of any degree or diploma.

Tejaswi Gaddam(21071A3219),Pranav Guddeti(21071A3224) ,Sravanth Mirtipati(21071A3245), Mohammad Abdul Sami (21071A3247) , P. Manhith Sai(21071A3253), P.Deepika Lalitha Sai (21071A3255)

ACKNOWLEDGEMENT

Behind every achievement lies an unfathomable sea of gratitude to those who activated it, without which it would ever never have come into existence. To them I lay the words of gratitude imprinting within us.

VNRVJIET has helped us transform ourselves from mere amateurs in the field of Computer Science into skilled engineers capable of handling any given situation in real time. I am are highly indebted to the institute for everything that it has given us.

We would like to express our gratitude towards the principal of our institute, Dr. Challa Dhanunjaya Naidu and the Head of the Computer Science & Engineering Department, Dr.S Nagini for their kind co- operation and encouragement which helped us complete the project in the stipulated time.

Although we had spent a lot of time and put in a lot effort into this project, it would not have been possible without the motivating support and help of our project guide Mr.P.Venkateswara Rao We thank him for his guidance, constant supervision and for providing necessary information to complete this project. Our thanks and appreciations also go to all the faculty members, staff members of VNRVJIET, and all our friends who have helped me put this project together.

Scheme of Course Based Project

Name of the course: Course Based project

Year / Semester : II-B.Tech I-semester

Project Title : NETFLIX DATA analysis

Done by :

Tejaswi Gaddam(21071A3219)

Pranav Guddeti(21071A3224)

Sravanth Mirtipati(21071A3245)

Mohammad Abdul Sami (21071A3247)

P.Manhith Sai(21071A3253)

P.Deepika Lalitha Sai (21071A3255)

Project Objectives :

The main objective of Netflix dataset analysis is getting know about the people around the world about their usage of Netflix ,which genre of ott content is popular, top actors ,top directors,etc..

Description :

Netflix is the one of the most popular OTT platform around the world. It was launched in the United States way back in 2005, but in India it was launched in 2017. But within our country too ,the number of users of Netflix have increased rapidly and people started to rejuvenate themselves through Netflix. Especially, during the covid and few months of post-covid era, people often preferred ott platforms over large screens. So, it becomes quite important to study and analyse the dataset of Netflix ,about what people browse, like and their favourite movies and T.V shows.

ABSTRACT

Netflix, Inc. is an American [subscription video on-demand over-the-top streaming](#) service and [production company](#) based in [Los Gatos, California](#). Founded in 1997 by [Reed Hastings](#) and [Marc Randolph](#) in [Scotts Valley, California](#), it offers a film and television series library through [distribution deals](#) as well as its own productions, known as [Netflix Originals](#).

Netflix is one of the largest providers of online streaming services. It collects a huge amount of data because it has a very large subscriber base. In this article, We are going to work on a project on Netflix data analysis with Python.

We can analyze a lot of data and models from Netflix because this platform has consistently focused on changing business needs by shifting its business model from on-demand DVD movie rental and now focusing a lot about the production of their original shows.

In this article, We shall take a look at some very important models of Netflix data to understand what's best for their business. Some of the most important tasks that we can analyze from Netflix data are:

- ★ understand what content is available
- ★ understand the similarities between the content
- ★ understand the network between actors and directors what exactly Netflix is focusing on
- ★ sentiment analysis of content available on Netflix



INDEX

Contents

- 1. Abstract**
- 2. Data set description**
- 3. Libraries used**
- 4. Concepts used**
- 5. Coding snippets**
- 6. Conclusion**
- 7. References**

PANDAS

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Why Use Pandas?

- Pandas allows us to analyze big data and make conclusions based on statistical theories.
- Pandas can clean messy data sets, and make them readable and relevant.
- Relevant data is very important in data science.

MATPLOTTING

A Python matplotlib script is structured so that a few lines of code are all that is required in most instances to generate a visual data plot. The matplotlib scripting layer

overlays two APIs:

- The pyplot API is a hierarchy of Python code objects topped by *matplotlib.pyplot*
-

An OO (Object-Oriented) API collection of objects that can be assembled with greater flexibility than pyplot. This API provides direct access to Matplotlib's backend layers.

The pyplot API has a convenient MATLAB-style stateful interface. In fact, matplotlib was originally written as an open-source alternative for MATLAB. The OO API and its interface is more customizable and powerful than pyplot, but considered more difficult to use. As a result, the pyplot interface is more commonly used, and is referred to by default in this article.

NUMPY

NumPy is a Python library used for working with arrays.

Why Use NumPy?

In Python we have lists that serve the purpose of arrays, but they are slow to process.

NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.

The array object in NumPy is called **ndarray**, it provides a lot of supporting functions that make working with **ndarray** very easy.

SEABORN

Seaborn is a Python data visualization library based on [matplotlib](#). It provides a high-level interface for drawing attractive and informative statistical graphics. Seaborn helps you explore and understand your data.

Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

NLTK

[Natural language processing](#) (NLP) is a field that focuses on making natural human language usable by

computer programs. **NLTK**, or [Natural Language Toolkit](#), is a Python package that you can use for NLP.

A lot of the data that you could be analyzing is [unstructured data](#) and contains human-readable text.

Before you can analyze that data programmatically, you first need to preprocess it.

TEXT BLOB

Text Blob is a Python library for Natural Language Processing. Using Text Blob for sentiment analysis is quite simple. It takes text as an input and can return **polarity** and **subjectivity** as outputs.

Polarity determines the sentiment of the text. Its values lie in $[-1, 1]$ where -1 denotes a highly negative

sentiment and 1 denotes a highly positive sentiment.

Subjectivity determines whether a text input is factual information or a personal opinion. Its value lies between $[0, 1]$ where a value closer to 0 denotes a piece of factual information and a value closer to 1 denotes a personal opinion.

LEMMATIZATION

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meanings to one word.

Text preprocessing includes both [Stemming](#) as well as Lemmatization. Many times people find these two terms confusing. Some treat these two as the same. Actually, lemmatization is preferred over stemming because lemmatization does morphological analysis of the words.

Stop words are **a set of commonly used words in any language**. For example, in English, “the”, “is” and “and”, would easily qualify as stop words. In NLP and text mining applications, stop words are used to eliminate unimportant words, allowing applications to focus on the important words instead.

WORD CLOUD

Word Cloud is a data visualization technique used for representing text data in which the size of each

word indicates its frequency or importance. Significant textual data points can be highlighted using a

word cloud.

The WordCloud function provides a lot of parameters that we can tweak according to our desire. Let us

understand a few of them.

- `width/height` : To adjust the height and width of the wordcloud.
- `random_state` : To recreate the same plot every time we run the function.
The `random_state` parameter has to be an integer value.
- `background_color` : To set a `background_color`. The default value for this parameter is 'black'. [This](#) page displays a list of colours that can be used.
- `colormap` : To set up the color theme for the words. [This](#) link provides a list of colormaps that can be used. The default value is 'viridis'
- `collocations` : To include bigrams of two words when set to True. The default value is True
- `stopwords` : To set the list of words that needs to be eliminated. This list can include trivial words like this, that, is, was, the, etc. If this parameter is set to None, then function will consider a built-in list of STOPWORDS

CODING SNIPPETS

```
[5] import numpy as np # linear algebra
import io
from google.colab import files
import pandas as pd # for data preparation
import plotly.express as px # for data visualization
from textblob import TextBlob # for sentiment analysis
uploaded = files.upload()
dff=pd.read_csv(io.BytesIO(uploaded['netflix.csv']))
dff.shape
```

Choose Files netflix.csv

- **netflix.csv**(text/csv) - 3408472 bytes, last modified: 1/31/2023 - 100% done
Saving netflix.csv to netflix.csv
(8807, 12)



dff.dtypes

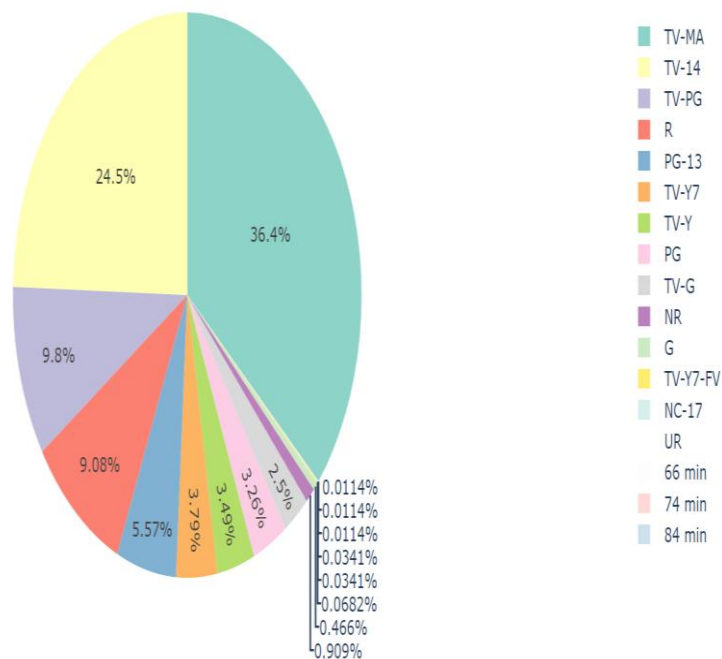
show_id	object
type	object
title	object
director	object
cast	object
country	object
date_added	object
release_year	int64
rating	object
duration	object
listed_in	object
description	object
dtype:	object

```
[6] dff.columns
```

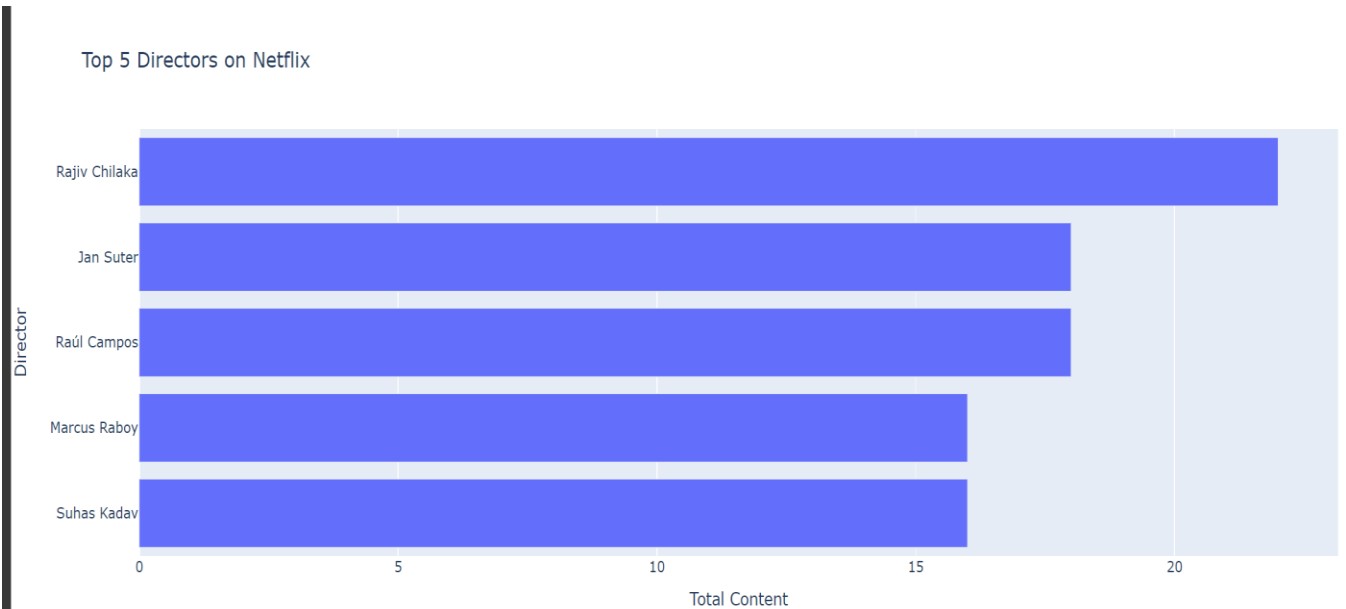
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',  
      'release_year', 'rating', 'duration', 'listed_in', 'description'],  
      dtype='object')
```

```
[7] z = dff.groupby(['rating']).size().reset_index(name='counts')  
    pieChart = px.pie(z, values='counts', names='rating',  
                      title='Distribution of Content Ratings on Netflix',  
                      color_discrete_sequence=px.colors.qualitative.Set3)  
    pieChart.show()
```

	rating	title
	66 min	1
	74 min	1
	84 min	1
	G	41
	NC-17	3
	NR	80
	PG	287
	PG-13	490
	R	799
	TV-14	2160
	TV-G	220
	TV-MA	3207
	TV-PG	863
	TV-Y	307
	TV-Y7	334
	TV-Y7-FV	6
	UR	3

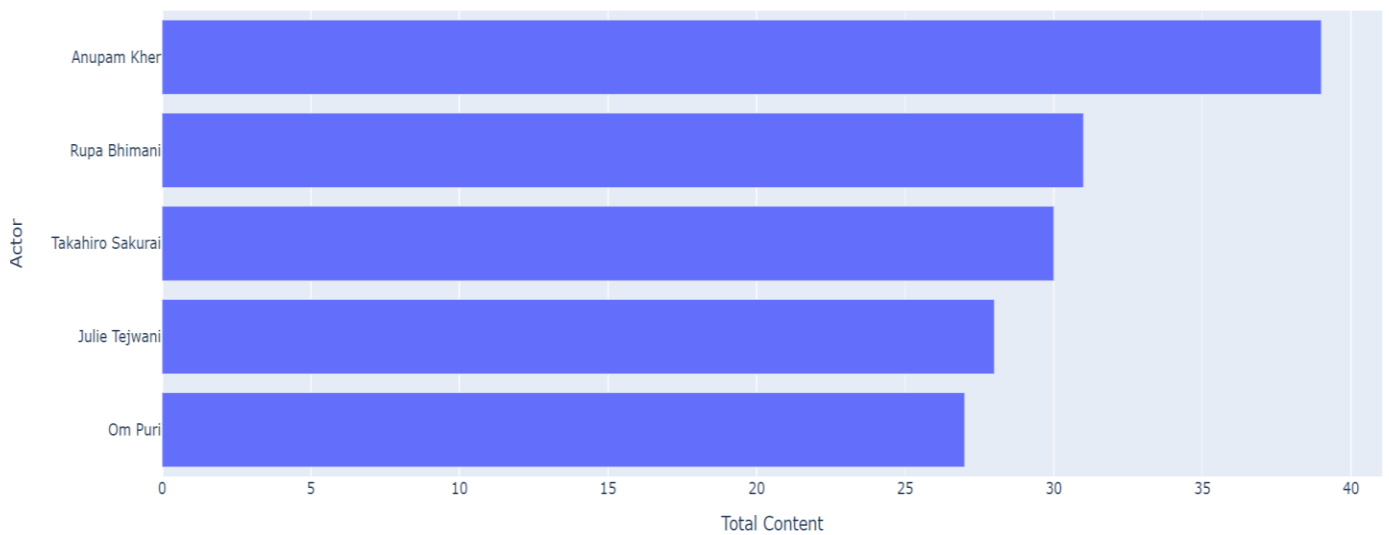


```
dff['director']=dff['director'].fillna('No Director Specified')  
filtered_directors=pd.DataFrame()  
filtered_directors=dff['director'].str.split(',',expand=True).stack()  
filtered_directors=filtered_directors.to_frame()  
filtered_directors.columns=['Director']  
directors=filtered_directors.groupby(['Director']).size().reset_index(name='Total 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'])  
fig1=px.bar(directorsTop5,x='Total Content',y='Director',title='Top 5 Directors on Netflix')  
fig1.show()
```



```
df['cast']=df['cast'].fillna('No Cast Specified')
filtered_cast=pd.DataFrame()
filtered_cast=df['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'])
fig2=px.bar(actorsTop5,x='Total Content',y='Actor', title='Top 5 Actors on Netflix')
fig2.show()
```

Top 5 Actors on Netflix



```
[10] df1=dfff[['type','release_year']]

df1=df1.rename(columns={"release_year": "Release Year"})

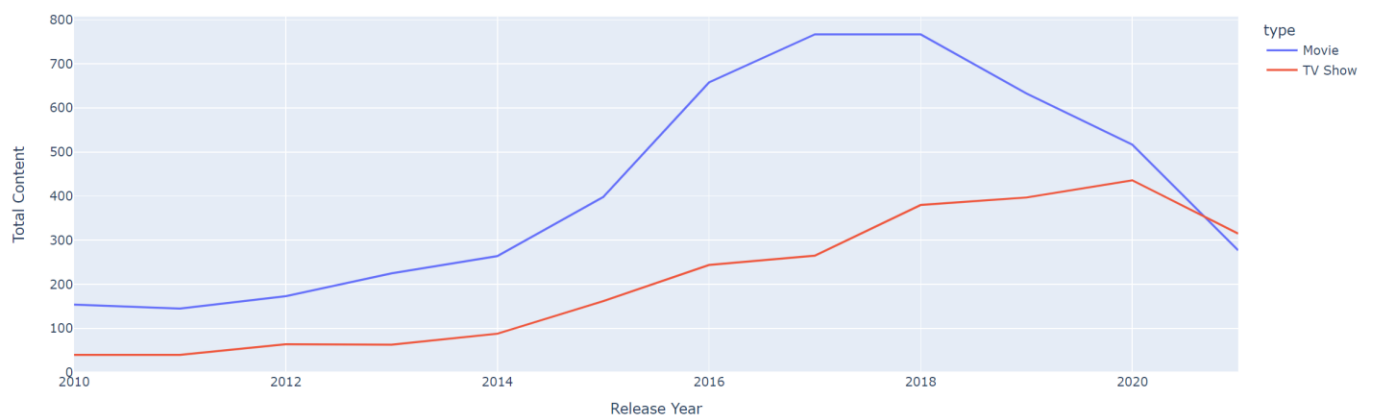
df2=df1.groupby(['Release Year','type']).size().reset_index(name='Total Content')

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()
```

Trend of content produced over the years on Netflix



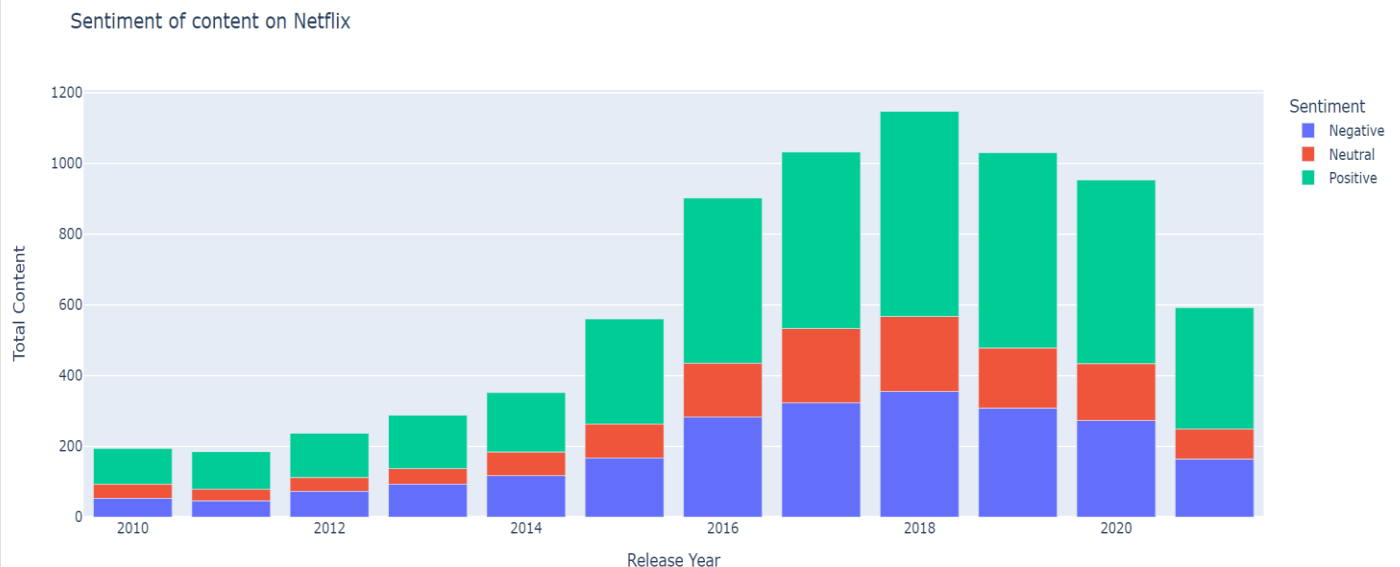

```

import numpy as np # linear algebra
import pandas as pd # for data preparation
import plotly.express as px # for data visualization
from textblob import TextBlob
dfx=dfff[['release_year','review']]
dfx=dfx.rename(columns={'release_year':'Release Year'})
for index,row in dfx.iterrows():
    z=row['review']
    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]
fig4 = px.bar(dfx, x="Release Year", y="Total Content", color="Sentiment", title="Sentiment of content on Netflix")
fig4.show()

```



```

import nltk
nltk.download('all')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize,sent_tokenize

#preprocessing
from nltk.corpus import stopwords #stopwords
from nltk import word_tokenize,sent_tokenize # tokenizing
from nltk.stem import PorterStemmer,LancasterStemmer # using the Porter Stemmer and Lancaster Stemmer and others
from nltk.stem.snowball import SnowballStemmer
from nltk.stem import WordNetLemmatizer # lammatizer from WordNet

#stop-words
stop_words=set(nltk.corpus.stopwords.words('english'))

```

```

new1 = dff[['review']].copy()
def clean_text(review):
    le=WordNetLemmatizer()
    word_tokens=word_tokenize(review)
    tokens=[le.lemmatize(w) for w in word_tokens if w not in stop_words and len(w)>3]
    cleaned_text=" ".join(tokens)
    return cleaned_text

```

```
[86] new1['cleaned_text']=new1['review'].apply(clean_text)
```

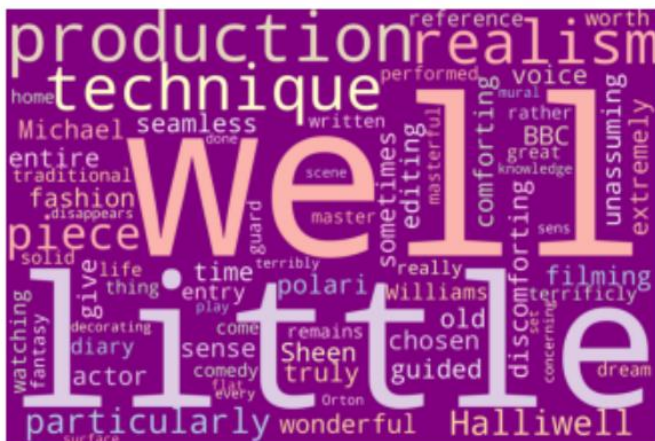
```
[87] new1.head()
```

index	review	cleaned_text
0	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me. The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word. It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentiary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many. Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away. I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a nickel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into prison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable with what is uncomfortable viewing....thats if you can get in touch with your darker side.	reviewer mentioned watching episode hooked They right exactly happened first thing struck brutality unflinching scene violence right word Trust show faint hearted timid This show pull punch regard drug violence hardcore classic word. called nickname given Oswald Maximum Security State Penitentiary focus mainly Emerald City experimental section prison cell glass front face inwards privacy high agenda City home many Aryans Muslims gangsta Latinos Christians Italians Irish scuffle death stare dodgy dealing shady agreement never away. would main appeal show fact go show would dare Forget pretty picture painted mainstream audience forget charm forget romance mess around first episode ever struck nasty surreal could ready watched developed taste accustomed high level graphic violence violence injustice crooked guard sold nickel inmate kill order away well mannered middle class inmate turned prison bitch lack street skill prison experience Watching become comfortable uncomfortable viewing thats touch darker side
1	A wonderful little production. The filming technique is very unassuming- very old-time-BBC fashion and gives a comforting, and sometimes discomforting, sense of realism to the entire piece. The actors are extremely well chosen- Michael Sheen not only "has got all the polari" but he has all the voices down pat too! You can truly see the seamless editing guided by the references to Williams' diary entries, not only is it well worth the watching but it is a terrifically written and performed piece. A masterful production about one of the great master's of comedy and his life. The realism really comes home with the little things: the fantasy of the guard which, rather than use the traditional 'dream' techniques remains solid then disappears. It plays on our knowledge and our senses, particularly with the scenes concerning Orton and Halliwell and the sets (particularly of their flat with Halliwell's murals decorating every surface) are terribly well done.	wonderful little production filming technique unassuming- old-time-BBC fashion give comforting sometimes discomforting sense realism entire piece actor extremely well chosen- Michael Sheen polari voice truly seamless editing guided reference Williams diary entry well worth watching terrifically written performed piece masterful production great master comedy life realism really come home little thing fantasy guard rather traditional 'dream technique remains solid disappears play knowledge sens particularly scene concerning Orton Halliwell set particularly flat Halliwell mural decorating every surface terribly well done
2	I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air conditioned theater and watching a light-hearted comedy. The plot is simplistic, but the dialogue is witty and the characters are likable (even the well bread suspected serial killer). While some may be disappointed when they realize this is not Match Point 2: Risk Addiction, I thought it was proof that Woody Allen is still fully in control of the style many of us have grown to love. This was the most I'd laughed at one of Woody's comedies in years (dare I say a decade?). While I've never been impressed with Scarlet Johanson, in this she managed to	thought wonderful spend time summer weekend sitting conditioned theater watching light-hearted comedy plot simplistic dialogue witty character likable even well bread suspected serial killer While disappointed realize Match Point Risk Addiction thought proof Woody Allen still fully control style many grown love. This laughed Woody

Word Cloud:

```
def cloud(index):  
    tok=new1['cleaned_text'][index].split(" ")  
    text = " ".join(cat for cat in tok)  
    word_cloud = WordCloud(  
        width=3000,  
        height=2000,  
        random_state=1,  
        background_color="purple",  
        colormap="Pastel1",  
        collocations=False,  
        stopwords=STOPWORDS,  
    ).generate(text)  
    plt.imshow(word_cloud)  
    plt.axis("off")  
    plt.show()  
  
cloud(1)
```

Output:



CONCLUSION

Python has been around since 1991. It is one of the best programming languages widely used in data analytics. It is easy to use, fast, and manipulates data seamlessly. It supports various data analytics activities such as data collection, analysis, modelling, and visualization.

The programming language is scalable and flexible. It has a vast collection of libraries for numerical computation and data manipulation. Python provides libraries for graphics and data visualization to build plots. It has broad community support to help solve many kinds of queries.

One of the main reasons why Python is widely used in the scientific and research communities is because of its ease of use and simple syntax which makes it easy to adapt for people who do not have an engineering background. It is also more suited for quick prototyping.

REFERENCES

<https://www.kaggle.com/datasets/shivamb/netflix-shows>

Snippet of Dataset:

[illegible]