CAPSTONE PROJECT WORK REPORT

Phase II

NETFLIX DATA ANALYSIS

Submitted work done by

KABIN B

A report submitted in part fulfilment of the degree of

B.Sc. in Computer Science with Data Analytics

Supervisor: Mrs.JAYAPRIYA.P, M.C.A.,M.E.,(Ph.D),Associate professor Dept. of CS with DA

KPRCAS
LEARN BEYOND

Department of Computer Science with Data Analytics

KPR College of Arts Science and Research

(Affiliated to Bharathiar University,
Coimbatore) Avinasi Road, Arasur,
Coimbatore – 641 407

NOVEMBER – 2022

CAPSTONE PROJECT WORK REPORT

Phase II

NETFLIX DATA ANALYSIS

Bonafide Work Done by

KABIN B

REG. NO. 2028B0018

Learn Beyond

KPR College of
Arts Science
and Research

Avinashi Road, Arasur, Coimbatore.

Dissertation submitted in partial fulfillment of the requirements for the award of Bharathiar University, Coimbatore-46.

Signature of the Guide **Signature of the HOD**

[Mrs.JAYAPRIYA. P]

Submitted for the Viva-Voce Examination held on ____________________

Internal Examiner **External Examiner**

ACKNOWLEDGEMENT

In the accomplishment of completion of my Capstone Project Work Phase – II on **Highest Paid Athlete Analysis with Python** I would like to convey my special gratitude to **Dr. S. Balusamy, Principal of KPR College of Arts Science and Research** and **Mrs.JAYAPRIYA. P, Associate Professor , Department of Computer Science with Data Analytics**. Your valuable guidance and suggestions helped me in phase - II of the completion of this project. I will always be thankful to you in this regard. I am ensuring that this project was finished by me and not copied.

Student Signature

Place:

Date:

Content

ORGANIZATION PROFILE

KPR COLLEGE OF ARTS SCIENCE AND RESEARCH

(Affiliated to Bharathiar University, Coimbatore)

Avinashi Road, Arasur, Coimbatore – 641 407

ABOUT THE COLLEGE

KPR College of Arts Science and Research is the latest addition to the KPR fleet. The College is located in a picturesque campus of about 11. Acres. The College is run by KPR charities under the leadership of our Chairman Dr. K.P. Ramasamy. The KPR Group is one of the largest industrial conglomerate in the country with interest in Textiles, Sugar, Wind Turbines, Automobiles and Education. The College was established in the year 2019 with a vision of providing top class education and life skills to students and thereby serve the nation and beyond. KPRCAS today offers 12 UG programmes in Management, Commerce and Computer Science streams. The Students of KPRCAS undergo intense training not only in the syllabus and curriculum of the affiliating University but are also trained in various areas. So that they emerge as industry ready graduates to meet the varying demands of the competing industries. Character building and Leadership qualities are inculcated into the students to make them responsible citizens focusing on the development of society and nation. A plethora of Clubs and Events encouraged the students to take part in sports and other cultural activities. KPRCAS offers three years undergraduate courses, which are exclusively for Business, Commerce and Computer Science Stream. The students are equipped with skills and knowledge needed to take up various leadership positions and to develop the society. Beyond Book Teaching help them to be professionals. KPRCAS emphasis on making the students academically brilliant, and also prepare them for the real corporate world. The learning curve begins here for the students of KPRCAS.

ABOUT THE DEPARTMENT

Bachelor of Computer Science with Data Analytics (B.Sc. (CS with DA)) was established in the year 2020. Data Analytics helps to raise the quality of data in the entire business system. The goal of data analytics is to construct the means for extracting business-focused insights from data This requires an understanding of how value and information flows in a business, and the ability to use that understanding to identify business opportunities. The primary aim of a data analyst is to increase efficiency and improve performance by discovering patterns in data. Data analysts exist at the intersection of information technology, statistics and business. They combine these fields in order to help businesses and organizations succeed.

SYNOPSIS

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, I’m going to introduce you to a data science project on Netflix data analysis with Python.

CHAPTER 1

1.INTRODUCTION

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.I’ll 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
- and sentiment analysis of content available on Netflix.

1.1SYSTEM SPECIFICATION

1.1.1. HARDWARE CONFIGURATION

Processor	intel® core™ i5-8265u cpu @ 1.60GHZ 1.80GHZ
RAM	8.00GB (7.88GB usable)
HD	64-bit operating system, x64-based processor

1.1.2. SOFTWARE SPECIFICATION

Windows Operating System

Intel i5 Processor

Python software used

CHAPTER 2

2. SYSTEM STUDY

2.1. EXISTING SYSTEM

We all are familiar with Netflix services. It handles large categories of movies and television content and users pay the monthly rent to access these contents. Netflix has **180+M** subscribers in **200+** countries. Netflix works on two clouds…**AWS** and **Open Connect**. These two clouds work together as the backbone of Netflix and both are highly responsible for providing the best video to the subscribers.

The application has mainly 3 components

Client

Device (User Interface) which is used to browse and play Netflix videos. TV, XBOX, laptop or mobile phone, etc

OC (Open connect) or Netflix CDN

CDN is the network of distributed servers in different geographical locations and Open Connect is Netflix's own custom global CDN (Content delivery network). It handles everything which involves video streaming. It is distributed in different locations and once you hit the play button the video stream from this component is displayed on your device. So if you're trying to play the video sitting in North America, the video will be served from the nearest open connect (or server) instead of the original server (faster response from the nearest server).

Backend (Database):

This part handles everything that doesn't involve video streaming (before you hit the play button) such as onboarding new content, processing videos, distributing them on servers located in different parts of the world, and managing the network traffic. Most of the processes are take care of by Amazon Web Services.

How Netflix Onboard a Movie/Video

Netflix receives very high-quality videos and content from the production houses so before serving the videos to the users it does some preprocessing. Netflix supports more than 2200 devices and each one of them requires different resolutions and formats. To make the videos viewable on different devices Netflix performs transcoding or encoding which involves finding errors and converting the original video into different formats and resolutions.

Drawbacks

- Limited Regional Selections.
- An Outdated Library
- Internet Requirements
- Data Cap Consumption
- No Ownership of Media
- Subscription Value
- Loss of Channel Surfing

Limited Regional Selections

If you live in the United States, there will inevitably be times when you want to watch something that's only available on Netflix Canada or Netflix UK. This happens very rarely, but when it does, it's annoying. We can't imagine how frustrating it is for people outside of the US wanting to watch Netflix content available exclusively in the United States.

An Outdated Library

The other big complaint about Netflix---which has been one of its sore spots ever since the streaming service went live---is that its library is really be up to-date. These days, only Netflix originals can really be considered timely and trendy.

Internet Requirements

The thing about Netflix (along with any other streaming app) is that the entire service is contingent upon your internet connection quality. Whether you're watching YouTube, Twitch, or Netflix, your ISP could be the difference between watching in 240p, 720p, or 4K video.If your internet goes down then there's no Netflix. If people on your network are watching YouTube or playing games, and consequently hogging up your bandwidth, Netflix will stutter. And if your internet speed is bad, video quality will suffer

Data Cap Consumption

While we're on the topic of internet connections, let's not forget that data caps are a very real nuisance to consider when streaming media---especially for videos, which can eat up more than 1GB/hour depending on how much quality you demand when watching movies and TV shows.

No Ownership of Media

Of all the reasons not to sell your CDs and DVDs, this one is the most relevant: even though you pay for Netflix, you don't own anything on it. If you buy a DVD, it's yours. With Netflix, your payments disappear into thin air. This means that after one year you will have paid anywhere from $108 to $192 depending on which Netflix plan you choose. However, you’ll have nothing   to show for it except the memories of whatever TV shows and films you watched during that time.

Subscription Value

just because Netflix makes it really easy to move from one episode to the next, although that does play a big part. It's because Netflix is a subscription service. There's no free Netflix trial and you pay the same no matter how much you watch, so watching more in a month means wringing more value out of your subscription.On the other hand, if you don't watch much at all, then Netflix may not be worth the price tag. If you go a month without watching anything, then you've basically thrown away your money.

Loss of Channel Surfing

This last point is minor in the bigger picture, but still worth considering if you haven't cut the cord yet: you can't surf channels and just watch whatever's playing. You always have to pick something, and sometimes this isn't that easy.Some workarounds to simulate channel surfing based on certain genres exist, but even those tend to be riddled with bugs and/or veer too far from the real thing. There's a charm to knowing that a show is playing live, and Netflix doesn't have that.

2.2 PROPOSED SYSTEM

Netflix is a popular platform for online streaming videos that makes the user to view and save the tv shows, medias, video streaming, movies and other popular series which was uploaded in the Netflix. This data provides a unique insight into the user’s personality. Of particular interest to our research work, are the user’s interests and eager on the shows or movies available there and analysing it and gives the proper result to users which the user is particularly interested on or in the state of mind. The dataset that we using for the task of Omicron sentiment analysis is collected from tweet download, which was initially collected from Twitter when people were sharing their opinions about the Omicron variant.

2.2.1 FEATURES

To classify the user needs and analysing it to what the user is actually looking for to stream or viewing any videos to deliver the right content to the user and make the application more reliable and useful

Performing sentiment analysis on Netflix data

Make the user to ownership of a channel.The user can able to create a channel or create their own playlist to stream or view any file content as per their wish. the user don’t need to keep scrolling for their favorite movies.Use channel surfing for the user to look for his/her requirement the user is able to search for any content streaming in online as per their wish. This is very useful to search any content through online without waiting for any streaming videos to arrive it makes the user convenient to search any ott files in any application .Use less data for streaming videos.There is a availability of choosing the video quality which uses less data data for streaming videos. which means less data is observed for low quality videos and more data is observed for high quality videos .Users are allowed to request for an streaming videos to upload

CHAPTER 3

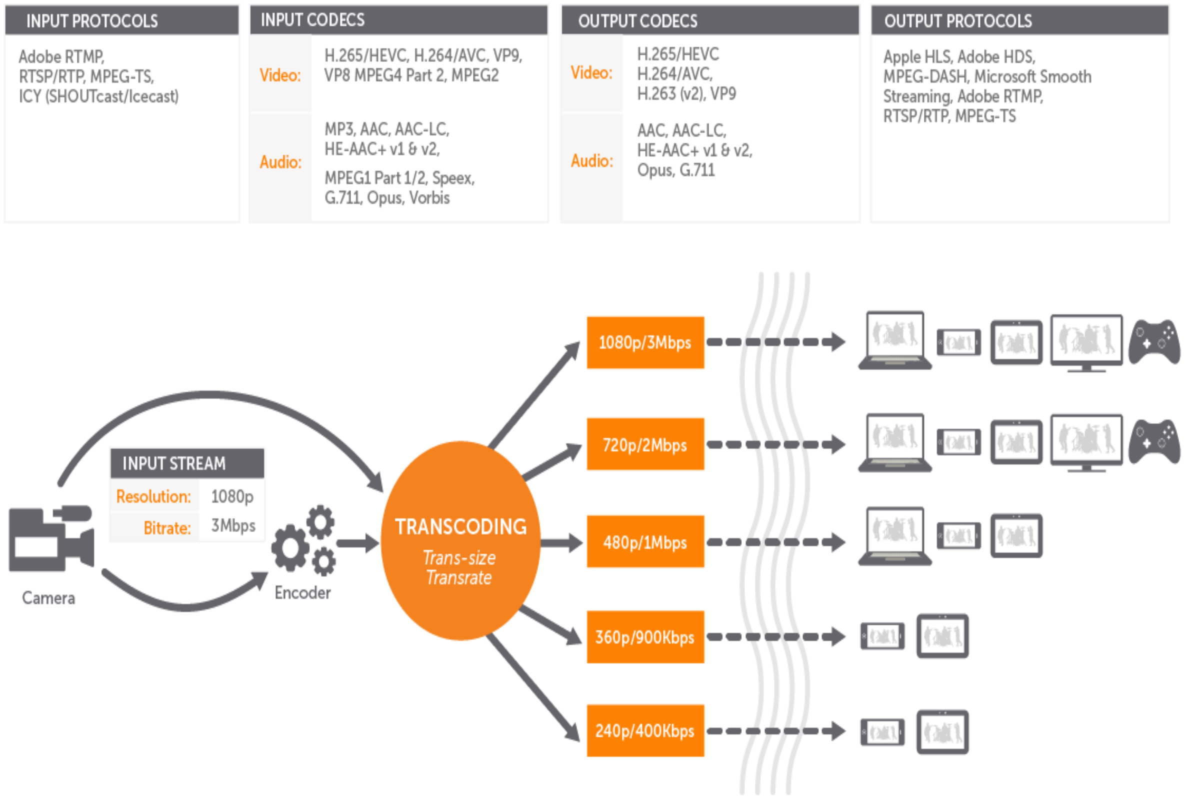
3. SYSTEM DESIGN

Before this movie is made available to users, Netflix must convert the video into a format that works best for your device. This process is called transcoding or encoding.Transcoding is the process that converts a video file from one format to another, to make videos viewable across different platforms and devices.

Whys do we need to do it? why can't we just play the source video?

The original movie/video comes in a high definition format that's many terabytes in size. Also, Netflix supports 2200 different devices. Each device has a video format that looks best on that particular device. If you're watching Netflix on an iPhone, you'll see a video that gives you the best viewing experience on the iPhone.

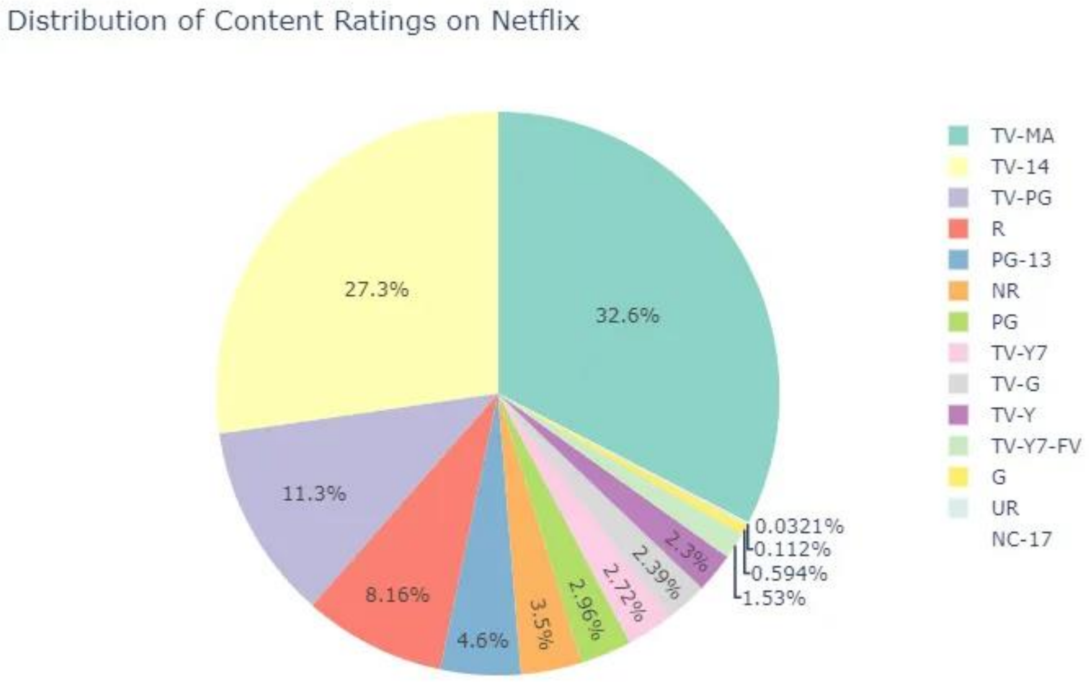
Netflix also creates files optimized for different network speeds. If you're watching on a fast network, you'll see the higher quality video than you would if you're watching over a slow network. And also depends on your Netflix plan. that said Netflix does create approx 1,200 files for every movie


INPUT PROTOCOLS
Adobe RTMP, RTSP/RTP, MPEG-TS, ICY (SHOUTcast/Icecast)
INPUT CODECS
Video:
H.265/HEVC, H.264/AVC, VP9, VP8 MPEG4 Part 2, MPEG2
Audio:
MP3, AAC, AAC-LC, HE-AAC+ v1 & v2,
MPEG1 Part 1/2, Speex, G.711, Opus, Vorbis
OUTPUT CODECS
Video:
H.265/HEVC H.264/AVC, H.263 (v2), VP9
Audio:
AAC, AAC-LC, HE-AAC+ v1 & v2, Opus, G.711
OUTPUT PROTOCOLS
Apple HLS, Adobe HDS, MPEG-DASH, Microsoft Smooth Streaming, Adobe RTMP, RTSP/RTP, MPEG-TS
INPUT STREAM
Resolution: 1080p
Bitrate: 3Mbps
Camera
Encoder
TRANSCODING
Trans-size
Transrate
1080p/3Mbps
720p/2Mbps
480p/1Mbps
360p/900Kbps
240p/400Kbps

3.1. INPUT DESIGN

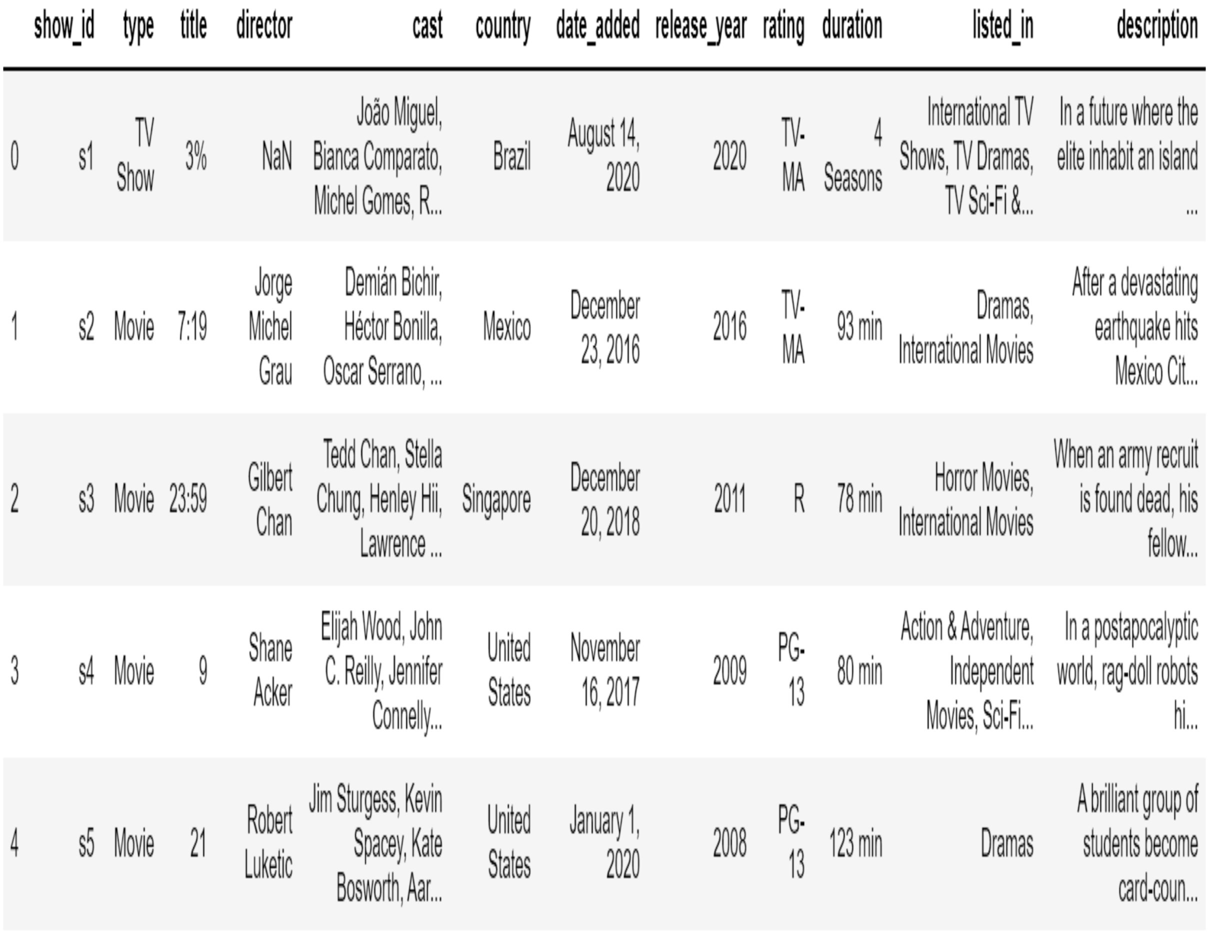
```
z = dff.groupby(['rating']).size().reset_index(name='counts')

pieChart = px.pie(z, values='counts', names='rating',

          title='Distribution of Content Ratings on Netflix',

          color_discrete_sequence=px.colors.qualitative.Set3)

pieChart.show()
```

3.2. OUTPUT DESIGN



Distribution of Content Ratings on Netflix

3.3. DATABASE DESIGN



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

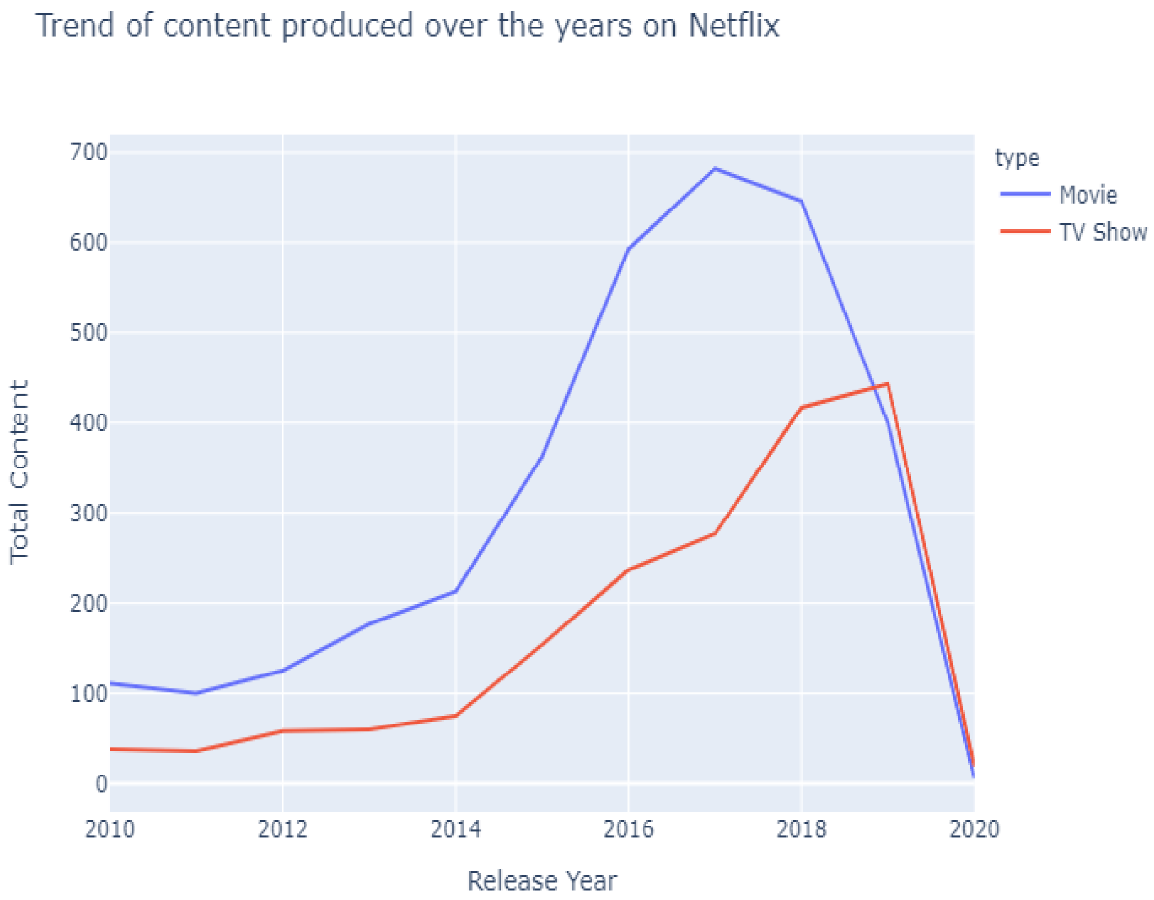
3.4 PERFORMANCE

Analyzing Content on Netflix

```
df1=dff[['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()
```

The next thing to analyze from this data is the trend of production over the years on Netflix:



The above line graph shows that there has been a decline in the production of the content for both movies and other shows since 2018.

CHAPTER 4

4.CONCLUSION

data science project on Netflix Data Analysis with Python programming language. Sometimes data visualization should be captivating and attention-grabbing which I think we have achieved here even if it isn't precise. So by customizing our visualization like what we did here reader's eye is drawn exactly where we want.

4.1 BIBLIOGRAPHY

[1] Data science https://en.wikipedia.org/wiki/Data_science Accessed on27-06-2020.

[2] A book on data science by Dr. Ossama Embarak, https://www.academia.edu/37886932/Data_Analysis_and_Visualization_Using_Python_-_Dr._Ossama_Embarak.pdf Accessed on 27-06-2020.

[3] A blog on quorahttps://www.quora.com Accessed on 27-06-2020.

[4] Smart data collective sitehttps://www.smartdatacollective.com Accessed on 28-06-2020

[5] A blog on by Grenoble School of businesshttps://www.stoodnt.com/index.php/blog/ Accessed on 28-06-2020.

[6] Python website https://www.python.org/doc/essays/blurb/ Accessed on29-06-2020.

[7] Data camp tutorialhttps://www.datacamp.com/community/tutorials/data-structures-python#adt Accessed on 29-06-2020

[8] AutomAte the Boring Stuff with Python Practical Programming for total Beginners (Au-thor AL SWEIGART) Accessed on 29-06-2020

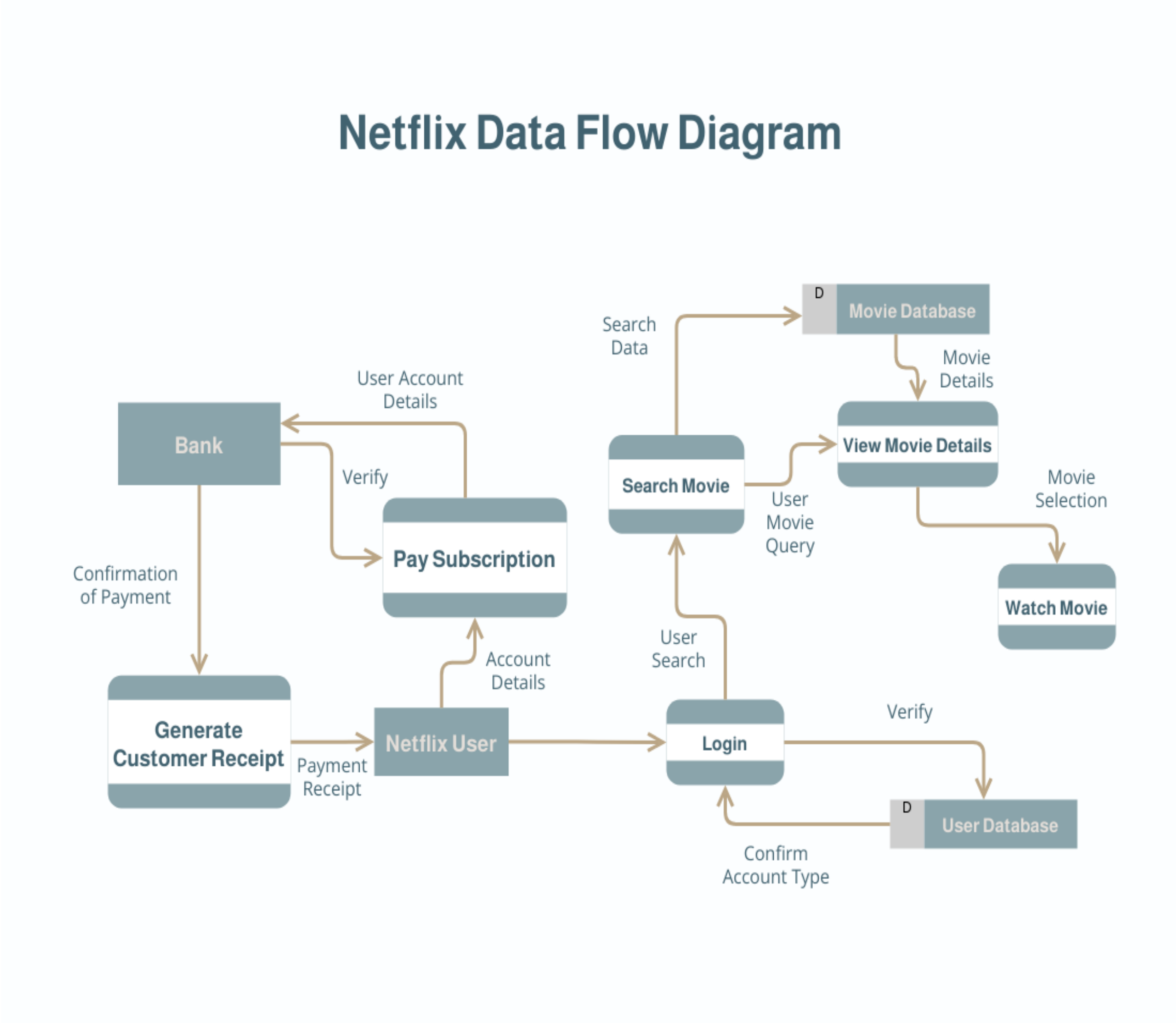
[9] https://wiki.python.org/ Accessed on 03-07-2020

[10] https://www.python-course.eu/python3_packages.php Accessed on03-07-2020

[11] Matplotlib https://en.wikipedia.org/wiki/Matplotlib Accessed on 04-07-2020

[12] Numpy online https://en.wikipedia.org/wiki/NumPy Accessed on 07-07-202056

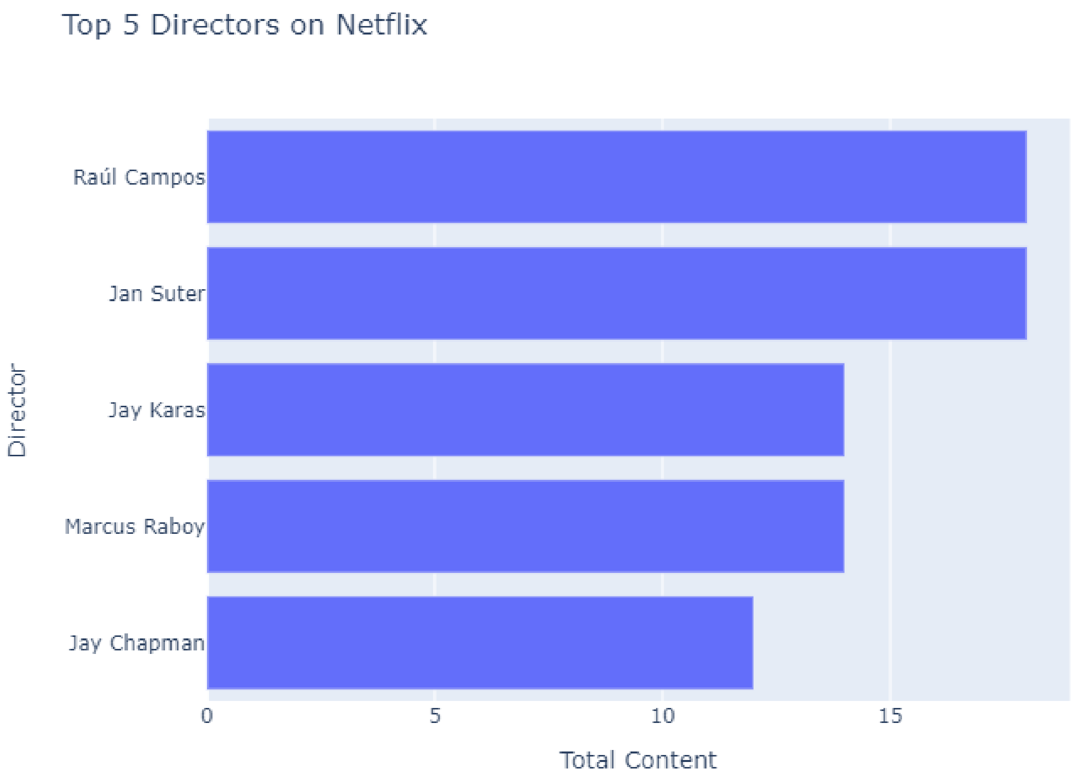
A. Data Flow Diagram



B.Source Code

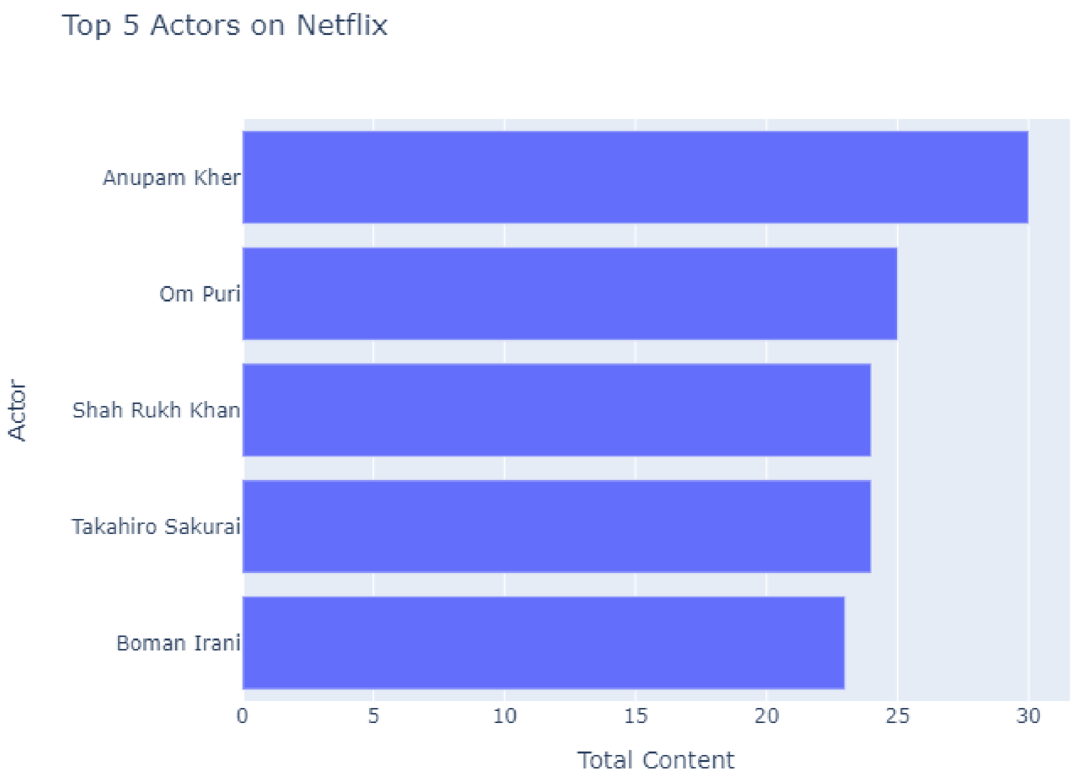
```
dff['director']=dff['director'].fillna('No Director Specified')
filtered_directors=pd.DataFrame()
filtered_directors=dff['director'].str.split(',',expand=True).stack()
filtered_directors=filtered_directors.to_frame()
filtered_directors.columns=['Director']
directors=filtered_directors.groupby(['Director']).size().reset_index(name='Total 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()
```

Output



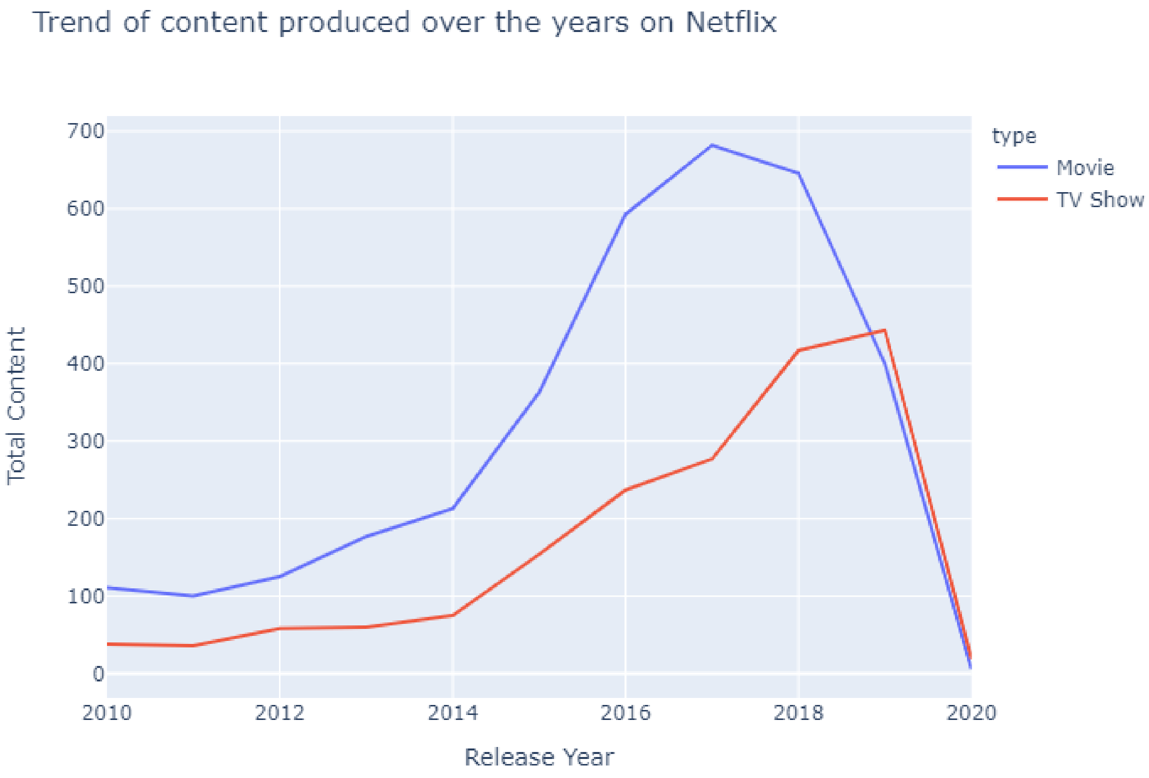
```
dff['cast']=dff['cast'].fillna('No Cast Specified')
filtered_cast=pd.DataFrame()
filtered_cast=dff['cast'].str.split(',',expand=True).stack()
filtered_cast=filtered_cast.to_frame()
filtered_cast.columns=['Actor']
actors=filtered_cast.groupby(['Actor']).size().reset_index(name='Total Content')
actors=actors[actors.Actor !='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()
```

Output



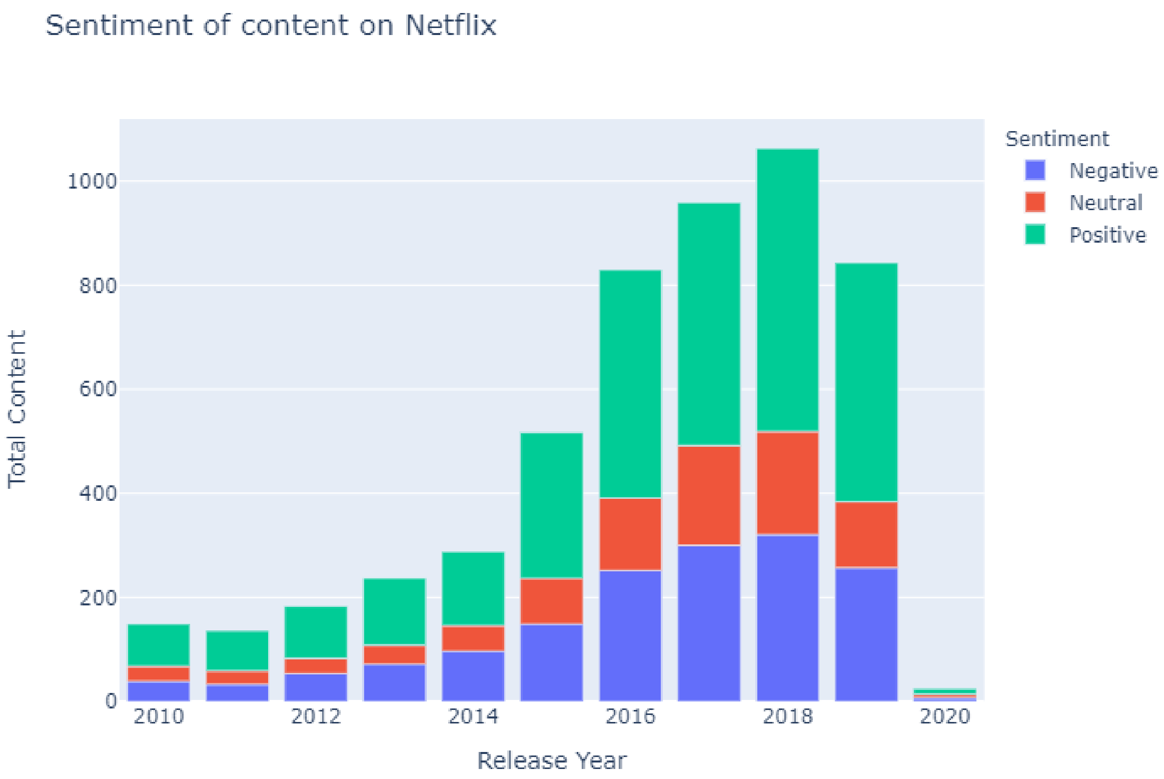
```
df1=dff[['type','release_year']]
df1=df1.rename(columns={"release_year": "Release Year"})
df2=df1.groupby(['Release Year','type']).size().reset_index(name='Total
Content')
df2=df2[df2['Release Year']>=2010]
fig3 = px.line(df2, x="Release Year", y="Total Content",
color='type',title='Trend of content produced over the years on Netflix')
fig3.show()
```

Output



```
dfx=dff[['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]
fig4 = px.bar(dfx, x="Release Year", y="Total Content", color="Sentiment",
title="Sentiment of content on Netflix")
fig4.show()
```

Output


Sentiment of content on Netflix
Total Content
0
200
400
600
800
1000
2010
2012
2014
2016
2018
2020
Release Year
Sentiment
Negative
Neutral
Positive