A PROJECT REPORT

on

"BEST STREAMING SERVICE ANALYSIS"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN

INFORMATION TECHNOLOGY

BY

Rakesh Kumar	**2129144**
Harshvardhan Jha	**2129142**
Antaryami Sing	**2129146**
Basudev Mallick	**2129147**
Supreeti Singh	**2129139**

UNDER THE GUIDANCE OF

Suchismita Das

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

April 2024

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

“BEST STREAMING SERVICE ANALYSIS“

submitted by

Rakesh Kumar	2129144
Harshvardhan Jha	2129142
Antaryami Sing	2129146
Basudev Mallick	2129147
Supreeti Singh	2129139

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Sci-ence & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2024-2025, under our guidance.

Date: 04/02/2024

Suchismita Das
Project Guide

Acknowledgements

We are profoundly grateful to **Suchismita Das** of **Affiliation** for her expert guidance and continuous encouragement throughout to see that this project meets its target since its commencement to its completion.

Rakesh Kumar

Harshvardhan Jha

Antaryami Sing

Basudev Mallick

Supreeti Singh

ABSTRACT

In the competitive landscape of streaming services, discerning the best platform is a compelling challenge for data scientists. This article presents a data science project utilizing Python to analyze and compare major streaming services such as Netflix, Prime Video, Hulu, and Disney+. Leveraging ratings of TV shows across these platforms, the study employs visualization techniques to discern trends and draw conclusions regarding the optimal streaming service choice.

The analysis begins with data preparation steps, including handling duplicates and null values, to ensure the dataset's integrity. Visualizations such as violin charts and scatter plots are then utilized to assess the content ratings and platform performance across user-rated platforms like IMDb and Rotten Tomatoes. Through these visualizations, patterns emerge, highlighting strengths and weaknesses of each platform. Ultimately, the study concludes that Amazon Prime emerges as the top choice, excelling in both the quality and quantity of content offered. This project not only showcases the capabilities of Python in data analysis but also provides valuable insights for consumers navigating the plethora of streaming options available today.

Keywords: Streaming services, data analysis, Python, ratings, consumer choice.

Dataset Description

The dataset was obtained from kaggle.com. From this dataset we present a comparison of various streaming platforms - Netflix, PrimeVideo, Disney+ and Hulu. The dataset used for the task of Best Streaming service analysis contains a comprehensive list of all the TV shows which are available on the 4 platforms that we are comparing in this task.

Following are the columns contained in the dataset:

- ID: The associated ID with a row of the table.
- Title: Title of the show.
- Year: Year of release of the show.
- Age: Age suitability factor.
- IMDb: IMDb rating on a scale of 0 to 10.
- Rotten Tomatoes: Rating on a scale of 0 to 100.
- Netflix, Hulu, Prime Video, Disney+: The four streaming platforms to be analyzed here.

Individual Contributions

~ Report Compilation ~

Supreeti Singh (2129139)
Basudev Mallick (2129147)

~ Coding ~

Harshvardhan Jha (2129142)
Rakesh Kumar (2129144)
Antaryami Sing (2129146)

Code Contribution Git Log

Unset

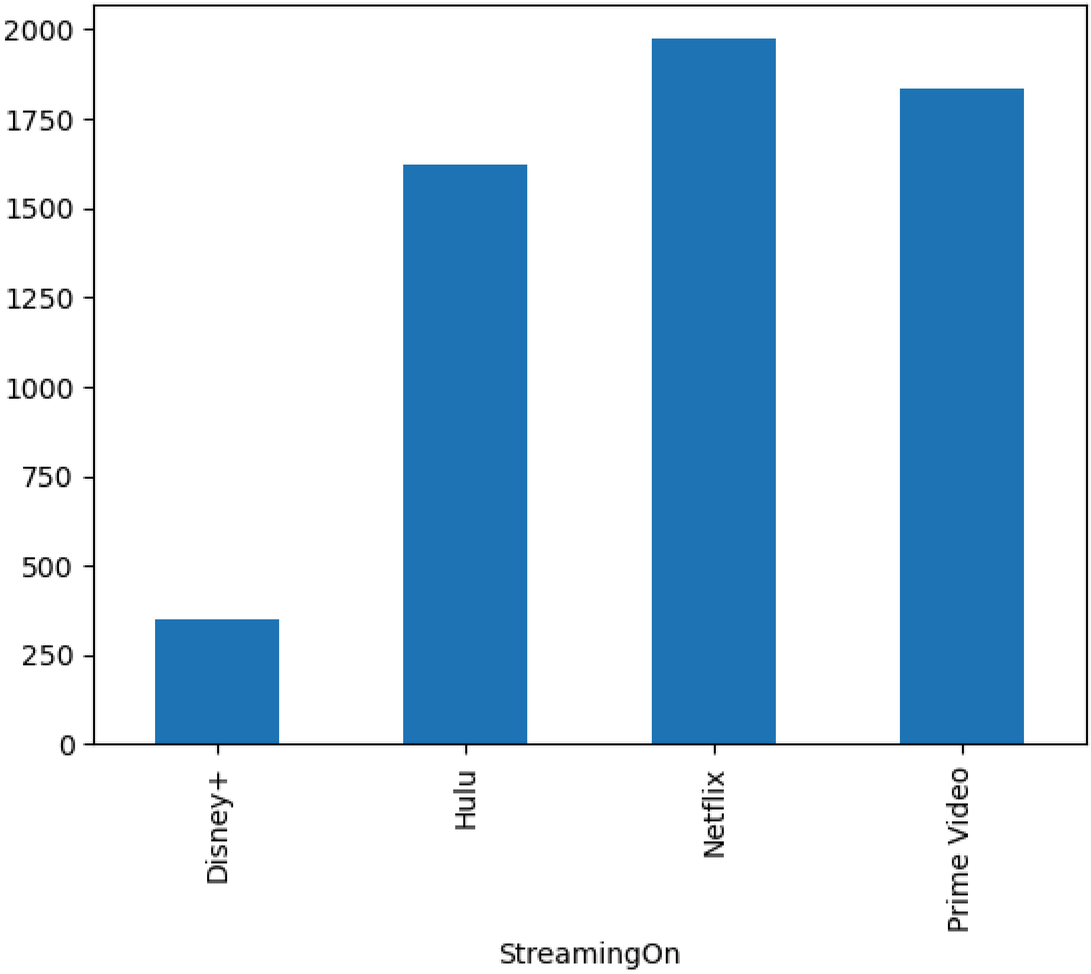
```
commit 9939fc552391683e3cf438894ca5bf1a44e81ffd (HEAD -> main, origin/main)
Author: imraklr <44721620+imraklr@users.noreply.github.com>
Date:   Thu Apr 4 01:17:37 2024 +0530

    add titles for sections in markdown cells

commit ef07b3804780bd60d49c9d52d4163f4c9447f652
Author: imraklr <44721620+imraklr@users.noreply.github.com>
Date:   Thu Apr 4 00:51:06 2024 +0530

    add scatter plot for IMDb vs. Rotten Tomatoes

commit a7f7ac6e5f5fd4adbf0930acd341a8eb4c06b6f1
Merge: a6711e2 d6a25fd
Author: Rakesh Kumar <44721620+imraklr@users.noreply.github.com>
Date:   Wed Apr 3 16:12:52 2024 +0530

    Merge pull request #2 from yami-antar/main

    adds violin plot

commit d6a25fd2539806bd8b3ceed68d699565db56862b
Author: yami-antar <2129146@kiit.ac.in>
Date:   Wed Apr 3 16:12:02 2024 +0530

    adds violin plot

commit a6711e2dc7f0eafcaa2b8ca39b9e839e1fd6956f
Merge: aa410f8 a998ada
Author: Rakesh Kumar <44721620+imraklr@users.noreply.github.com>
Date:   Wed Apr 3 02:51:03 2024 +0530

    Merge pull request #1 from 62026/main

    adds code for EDA and Data preparation

commit a998adad13455a930f8e34ac207bc535d8b3b169
Author: 62026 <jhavardhan4579@gmail.com>
Date:   Wed Apr 3 02:47:10 2024 +0530

    adds code for EDA and Data preparation

commit aa410f8af7181b9e23e4faf69393cbca3062d4cb
Author: imraklr <44721620+imraklr@users.noreply.github.com>
Date:   Wed Apr 3 02:21:36 2024 +0530

    Initial Commit
```

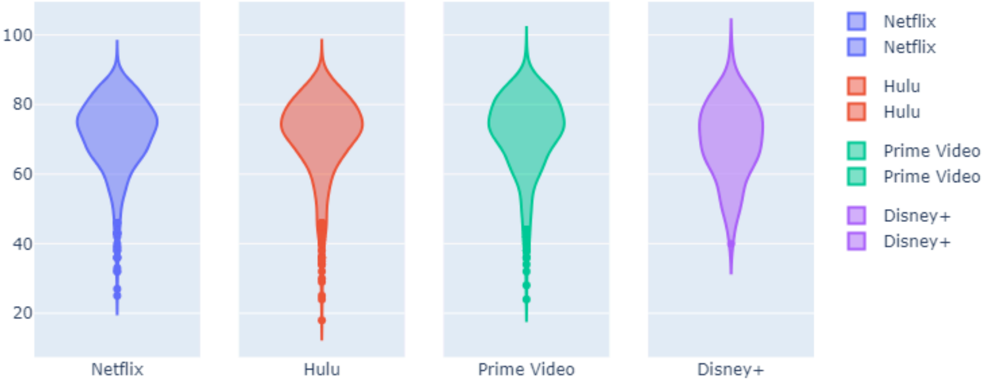
Code

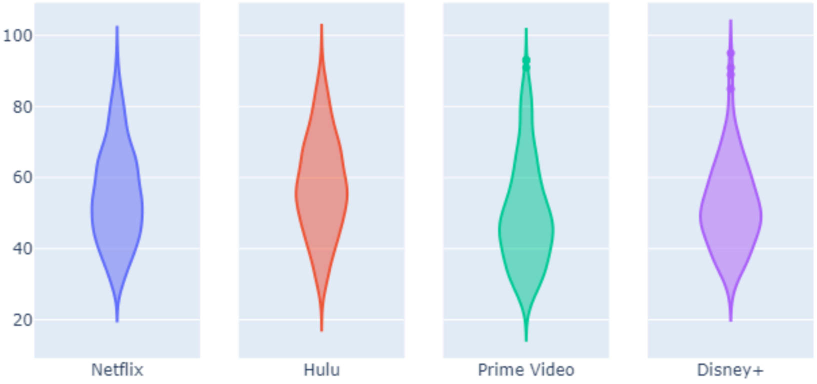
Python

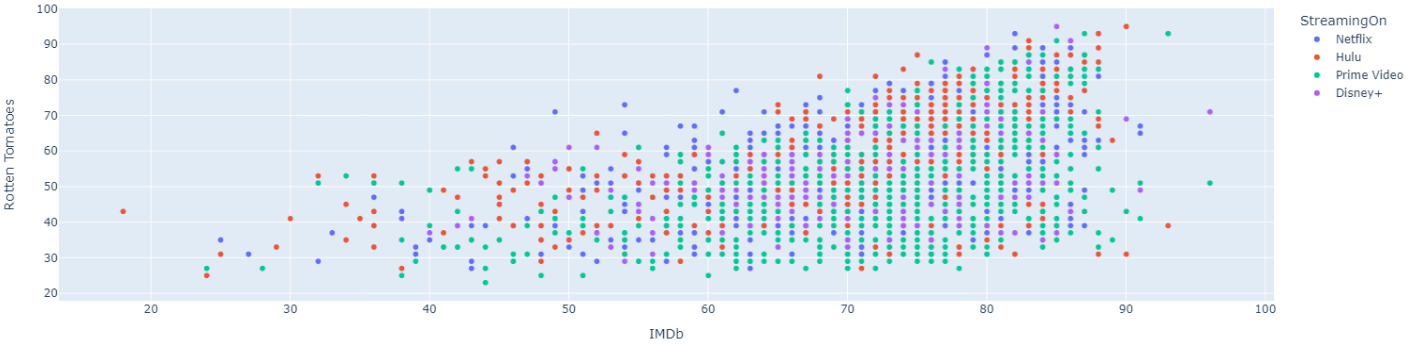
```python
import numpy as np
import pandas as pd

import plotly
import plotly.express as px
from plotly.subplots import make_subplots
import seaborn as sns
import matplotlib.pyplot as plt

tv_shows = pd.read_csv('tv_shows.csv')
tv_shows.head()

tv_shows.drop_duplicates(subset='Title', keep='first',inplace=True)

tv_shows['Rotten Tomatoes'] = tv_shows['Rotten
Tomatoes'].fillna('0.0/10').str.replace('/100', '')
tv_shows['Rotten Tomatoes'] = pd.to_numeric(tv_shows['Rotten Tomatoes'])
tv_shows['IMDb'] = tv_shows['IMDb'].fillna('0.0/10').str.replace('/10', '')
tv_shows['IMDb'] = pd.to_numeric(tv_shows['IMDb'])

tv_shows['IMDb'] = tv_shows['IMDb'].fillna(0)
tv_shows['IMDb'] = tv_shows['IMDb']*10
tv_shows['IMDb'] = tv_shows['IMDb'].astype('int')

tv_shows_long=pd.melt(tv_shows[['Title','Netflix','Hulu','Disney+', 'Prime
Video']],id_vars=['Title'], var_name='StreamingOn', value_name='Present')
tv_shows_long = tv_shows_long[tv_shows_long['Present'] == 1]
tv_shows_long.drop(columns=['Present'],inplace=True)

tv_shows_combined = tv_shows_long.merge(tv_shows, on='Title', how='inner')
tv_shows_combined.drop(columns = ['Unnamed: 0','Netflix', 'Hulu', 'Prime
Video', 'Disney+', 'Type'], inplace=True)

tv_shows_both_ratings = tv_shows_combined[(tv_shows_combined.IMDb > 0) &
tv_shows_combined['Rotten Tomatoes'] > 0]
tv_shows_combined.groupby('StreamingOn').Title.count().plot(kind='bar')

figure = []
figure.append(px.violin(tv_shows_both_ratings, x = 'StreamingOn', y = 'IMDb',
color='StreamingOn'))
figure.append(px.violin(tv_shows_both_ratings, x = 'StreamingOn', y = 'Rotten
Tomatoes', color='StreamingOn'))
fig = make_subplots(rows=2, cols=4, shared_yaxes=True)

for i in range(2):
    for j in range(4):
        fig.add_trace(figure[i]['data'][j], row=i+1, col=j+1)

fig.update_layout(autosize=False, width=800, height=800)
fig.show()
px.scatter(tv_shows_both_ratings, x='IMDb', y='Rotten
Tomatoes',color='StreamingOn')
```

Graphs


2000
1750
1500
1250
1000
750
500
250
0
Disney+
Hulu
Netflix
Prime Video
StreamingOn


100
80
60
40
20
Netflix
Hulu
Prime Video
Disney+
Netflix
Netflix
Hulu
Hulu
Prime Video
Prime Video
Disney+
Disney+


100
80
60
40
20
Netflix
Hulu
Prime Video
Disney+


Rotten Tomatoes
100
90
80
70
60
50
40
30
20
IMDb
20
30
40
50
60
70
80
90
100
StreamingOn
● Netflix
● Hulu
● Prime Video
● Disney+

Conclusion

By using the violin chart we can observe that:

- Hulu, Netflix, and Amazon Videos all have important data. As content increases, quality decreases for all three.
- Prime Videos has become denser in the top half when looking at IMDB and performs well in cool.
- Disney+ being new, has also been very successful in this area.

Using the scatter plot we can observe that it is quite obvious that Amazon Prime performs very well in the fourth quadrant. Even by using the bar plot, we can observe that Amazon prime had a great quantity of content. So looking at all the streaming platforms we can conclude that Amazon Prime is better in both quality and quantity.