A PROJECT REPORT

on

"Netflix Data Analysis"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE & ENGINEERING

BY

ADYASHA SOUMYA ROUTRAY	2105943
ABHISHEK ANAND	21052216
SONALIKA SAHOO	21052534
SUBHASHREE PANDA	21052539

UNDER THE GUIDANCE OF

Prof. Partha Sarathi Paul

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

A PROJECT REPORT

on

"Netflix Data Analysis"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE
COMPUTER SCIENCE AND ENGINEERING

BY

ADYASHA SOUMYA ROUTRAY (2105943)
ABHISHEK ANAND (21052216)
SONALIKA SAHOO (21052534)
SUBHASHREE PANDA (21052539)

UNDER THE GUIDANCE OF

Prof. Partha Sarathi Paul

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAE, ODISHA -751024

April 2024

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

"Netflix Data Analysis"

submitted by

ADYASHA SOUMYA ROUTRAY	2105943
ABHISHEK ANAND	21052216
SONALIKA SAHOO	21052534
SUBHASHREE PANDA	21052539

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

Date: 13/04/2024

(Prof. Partha Sarathi Paul)
Project Guide

Acknowledgements

We are profoundly grateful to **Prof. Partha Sarathi Paul** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. .....

ADYASHA SAUMYA ROUTRAY 2105943
ABHISHEK ANAND 21052216
SONALIKA SAHOO 21052534
SUBHASHREE PANDA 21052539

ABSTRACT

This project is dedicated to revolutionizing the Netflix user experience through the strategic use of advanced data analysis techniques, particularly harnessing the capabilities of NumPy arrays and DataFrame concepts. The primary objective is to delve deep into Netflix's datasets to uncover insights that can significantly enhance the platform's usability and content delivery.

By meticulously scrutinizing the datasets, we aim to identify and mitigate redundancy, a common challenge that can clutter the user experience and diminish satisfaction. Through the utilization of NumPy arrays, we can efficiently manipulate large volumes of data, facilitating swift analysis and extraction of meaningful patterns. Concurrently, employing DataFrame concepts enables us to organize and structure the data in a manner conducive to insightful exploration.

Central to our approach is a keen focus on understanding user preferences and behaviors. By analyzing viewing patterns, genre preferences, and historical interactions, we can tailor recommendations with precision, ensuring that users are presented with content that aligns closely with their interests. This personalized approach not only increases user engagement but also fosters a sense of connection and satisfaction with the platform.

Moreover, by optimizing content curation based on user feedback and engagement metrics, we can further refine the recommendation algorithms, continually improving their accuracy and relevance. This iterative process of data-driven decision-making forms the backbone of our strategy, allowing us to adapt and evolve in response to changing user needs and preferences.

Ultimately, this project represents a convergence of innovation, efficiency, and personalization. By harnessing the power of data analysis and leveraging insights gleaned from Netflix's datasets, we aim to create a more intuitive and enjoyable streaming experience for users. Through our efforts, we seek to propel Netflix towards the forefront of user-centric content delivery, setting new standards for excellence in the entertainment industry

.

Contents

Chapter 1

Introduction

Netflix, the premier streaming platform, offers diverse content searchable by genre and language. With personalized recommendations and advanced data analysis, users enjoy tailored selections and top picks, enriching their viewing experience with entertainment suited to their preferences.

The significance of this project lies in its utilization of NumPy arrays and Pandas DataFrames to enhance the user experience on Netflix. By analyzing data, reducing redundancy, and offering personalized recommendations, it effectively addresses evolving viewer demands. The report structure encompasses objectives, methodology, findings, implementation insights, and reflections, driving innovation in streaming content delivery.

"In today's digital landscape, streaming platforms like Netflix play a pivotal role in entertainment consumption. However, with vast content libraries, users face challenges in discovering relevant choices. This project aims to address this issue by employing advanced data analysis techniques, including NumPy arrays and Pandas DataFrames, to optimize content recommendations and enhance user satisfaction."

.

Chapter 2

Basic Concepts/ Literature Review

NumPy, with its N-dimensional arrays, powers most data science libraries, offering speed advantages over Python lists. Pandas, an open-source library, streamlines data manipulation with less coding. Pandas Data Frame, a two-dimensional structure, facilitates data cleaning and operations on rows and columns.

NumPy, integral to data science, boasts N-dimensional arrays and underpins numerous libraries. It offers significant speed enhancements compared to Python lists, accelerating computation-intensive tasks by up to 50 times.

Pandas, a Python library, simplifies data manipulation and analysis, minimizing coding effort. Its flexibility and customization options make it a preferred choice for data scientists seeking efficiency in handling large datasets.

Pandas DataFrame, a two-dimensional data structure, enables seamless operations on rows and columns. It excels in cleaning and organizing data, facilitating error correction and duplicate removal to ensure data integrity.

Chapter 3

Problem Statement / Requirement Specifications

In the era of digital streaming, users of platforms like Netflix encounter challenges in efficiently discovering content aligned with their preferences. Despite the platform's vast library, the current recommendation system often fails to provide personalized suggestions, leading to user frustration and dissatisfaction. This project aims to address this issue by leveraging advanced data analysis techniques, including NumPy arrays and Pandas DataFrames, to enhance Netflix's recommendation algorithm. By optimizing content delivery based on user behavior and preferences, the goal is to improve user satisfaction and engagement on the platform.

3.1 Project Planning

- Gather user feedback and requirements.
- Collect and preprocess Netflix datasets.
- Develop algorithms for content analysis.
- Design intuitive user interfaces.
- Test and validate system performance.
- Deploy enhancements and monitor user feedback.
- Continuously optimize based on insights.

3.2 Project Analysis

After the requirements are collected or the problem statements is conceptualized, this needs to be analyzed for finding any short of ambiguity, mistake, etc.

3.3 System Design

Design a system integrating data preprocessing, advanced analysis algorithms, and user interface enhancements to optimize Netflix's recommendation system for personalized content delivery and user satisfaction.

.

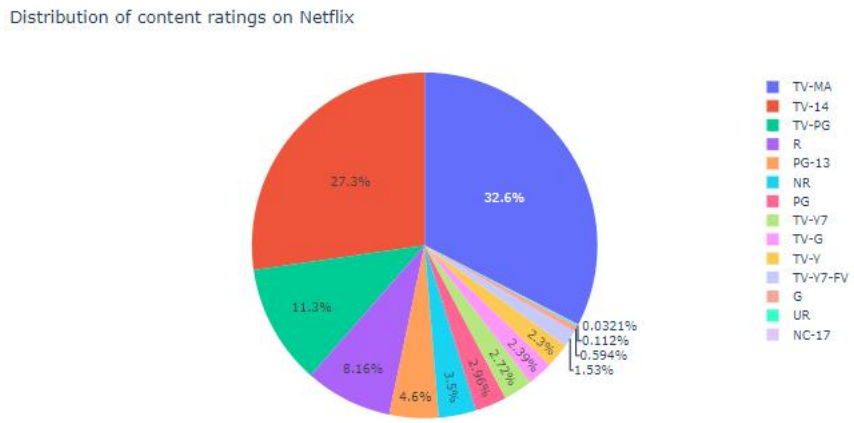
CODE BREAKDOWN

1)Creating a Pie Chart based on Content Rating:

```
pieChart = px.pie(x, values='counts', names='rating', title='Distribution of content ratings on Netflix')
pieChart.show()
```

Explanation:

The code uses the Plotly Express library (px) to create a pie chart (px.pie).
The x variable likely represents a DataFrame containing the data to be visualized.
values='counts' indicates the numerical values to be represented in the pie chart.
names='rating' specifies the category labels for each slice of the pie chart.
title='Distribution of content ratings on Netflix' sets the title for the chart.
Finally, pieChart.show() displays the pie chart.

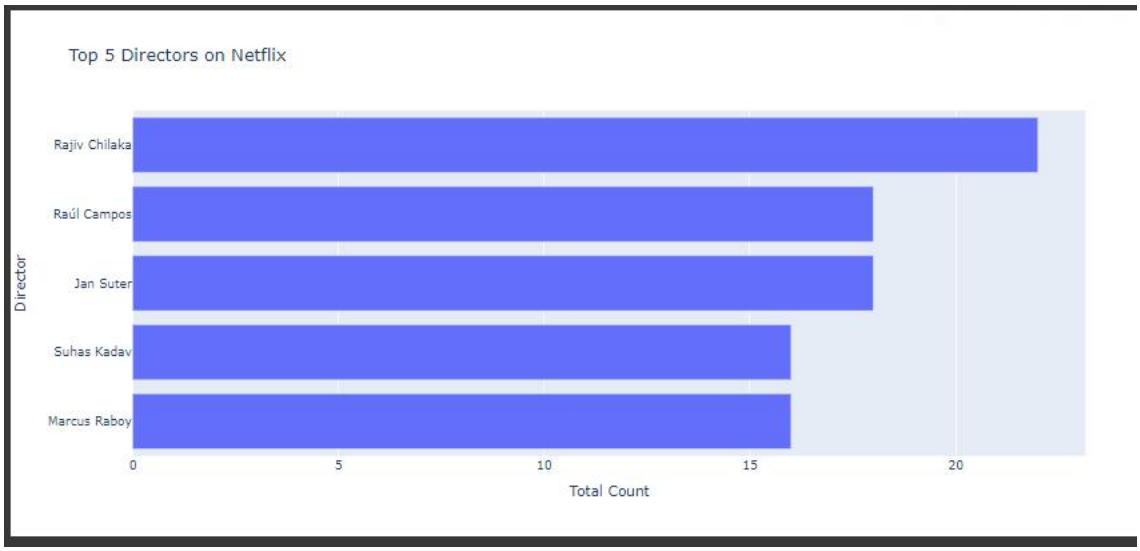


2)Analyzing the Top 5 Directors on Netflix:

```
df['director'] = df['director'].fillna('Director not specified')
df.head()
```

Explanation:

This code snippet fills missing values in the 'director' column with the string 'Director not specified'.
df.head() is then used to display the first few rows of the DataFrame (df).

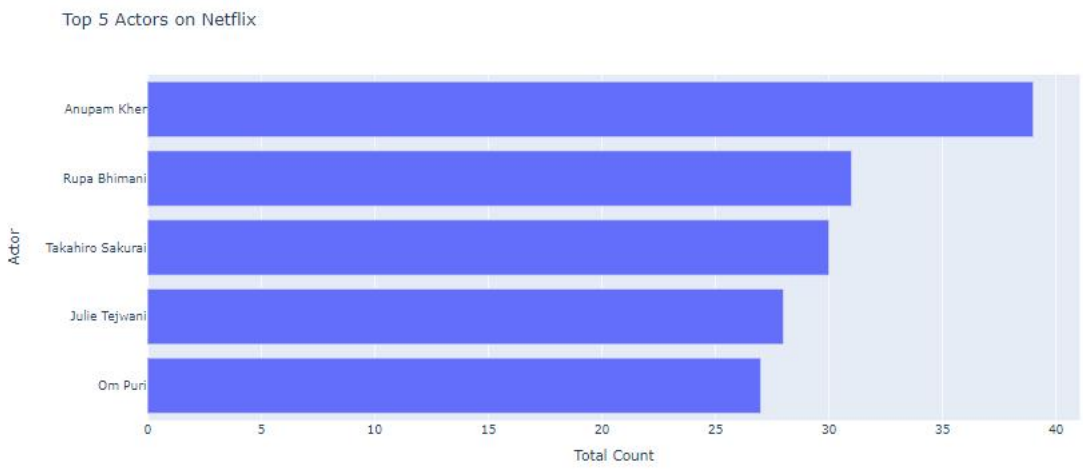


3)Analyzing the Top 5 Actors on Netflix:

```
df['cast'] = df['cast'].fillna('No cast specified')
cast_df = pd.DataFrame()
cast_df = df['cast'].str.split(',', expand=True).stack()
cast_df = cast_df.to_frame()
cast_df.columns = ['Actor']
actors = cast_df.groupby(['Actor']).size().reset_index(name='Total Count')
actors = actors[actors.Actor != 'No cast specified']
actors = actors.sort_values(by=['Total Count'], ascending=False)
top5Actors = actors.head()
top5Actors = top5Actors.sort_values(by=['Total Count'])
barChart2 = px.bar(top5Actors, x='Total Count', y='Actor', title='Top 5 Actors on
Netflix')
barChart2.show()
```

Explanation:

This code snippet processes the 'cast' column to extract individual actors' names.
Missing values in the 'cast' column are filled with the string 'No cast specified'.
The data is split by commas and stacked to create a DataFrame (cast_df) containing actor names.
Grouping by actor and counting occurrences yields the total count of each actor's appearances.
The top 5 actors are selected, and a horizontal bar chart (px.bar) is created to visualize their counts.
The chart is then displayed using barChart2.show().

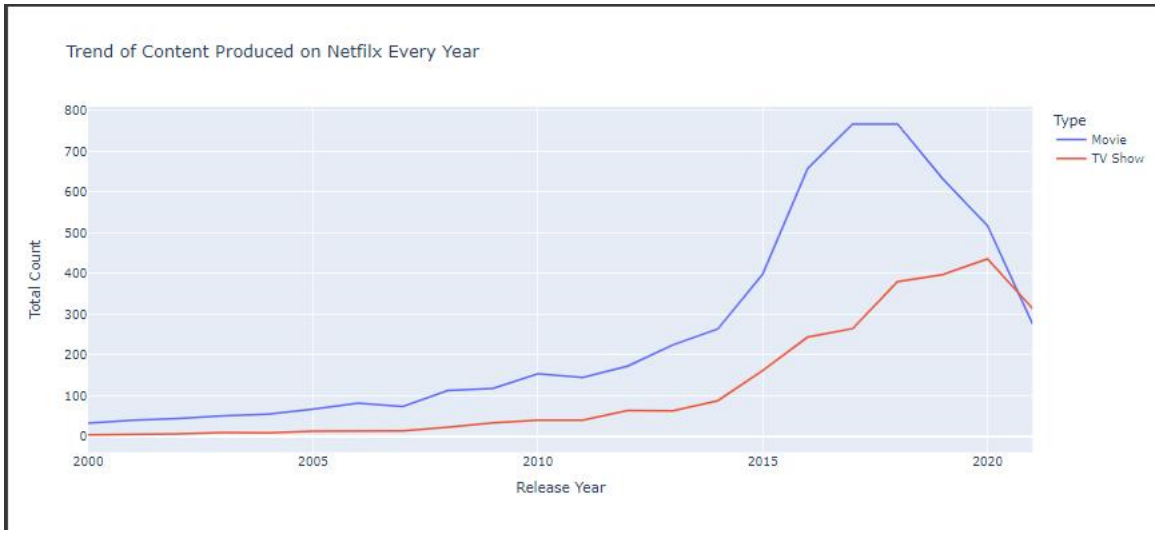


5)Analyzing the Content Produced on Netflix based on Years:

```
df1 = df[['type', 'release_year']]
df1 = df1.rename(columns={"release_year": "Release Year", "type": "Type"})
df2 = df1.groupby(['Release Year', 'Type']).size().reset_index(name='Total Count')
```

Explanation:

This code snippet selects the 'type' (e.g., movie, TV show) and 'release_year' columns from the DataFrame (df).
The column names are renamed for clarity.
The data is grouped by release year and content type, and the count of each combination is calculated.
The results are stored in a new DataFrame (df2), which represents the count of each type of content released in each year.



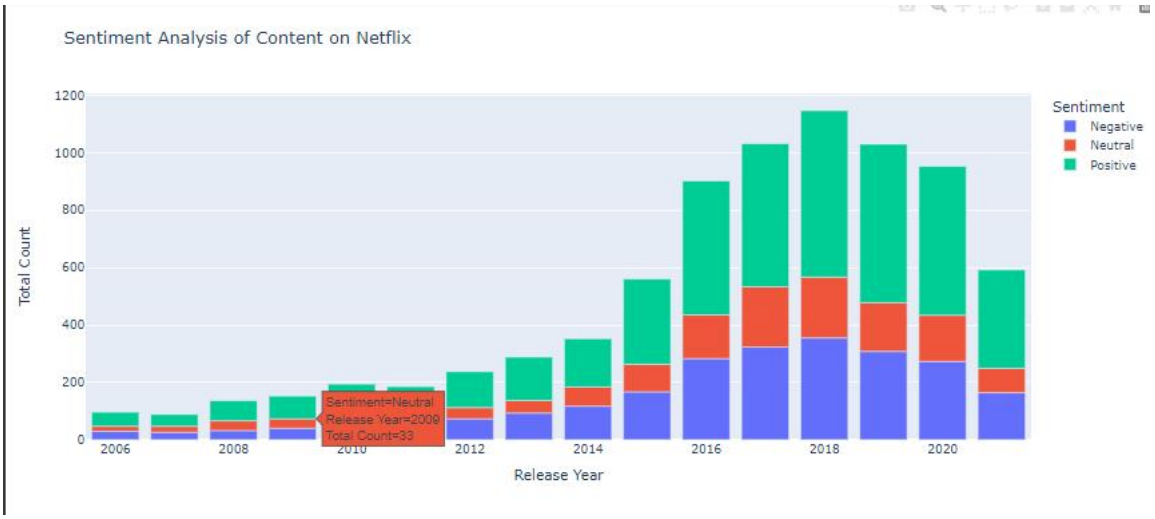
7)Sentiment Analysis of Netflix Content

```
df3 = df[['release_year', 'description']]
df3 = df3.rename(columns={'release_year': 'Release Year', 'description': 'Description'})
for index, row in df3.iterrows():
    d = row['Description']
    testimonial = TextBlob(d)
    p = testimonial.sentiment.polarity
    if p == 0:
        sent = 'Neutral'
    elif p > 0:
        sent = 'Positive'
    else:
        sent = 'Negative'
    df3.loc[[index, 2], 'Sentiment'] = sent

df3 = df3.groupby(['Release Year', 'Sentiment']).size().reset_index(name='Total Count')
df3 = df3[df3['Release Year'] > 2005]
barGraph = px.bar(df3, x="Release Year", y="Total Count", color="Sentiment",
title="Sentiment Analysis of Content on Netflix")
barGraph.show()
```

Explanation:

This code snippet conducts sentiment analysis on the descriptions of Netflix content.
It utilizes the TextBlob library to analyze the sentiment polarity of each description.
Based on the polarity score, each description is classified as 'Positive', 'Negative', or 'Neutral'.
The sentiment analysis results are aggregated by release year and sentiment category.
Only data from the year 2005 onwards is considered.
A grouped bar chart (px.bar) is created to visualize the sentiment analysis results.
The chart is then displayed using barGraph.show().
This breakdown explains the purpose and functionality of each code snippet in the provided code.



Chapter 6

Conclusion and Future Scope

6.1 Conclusion

Utilizing NumPy arrays and Pandas DataFrames, this Netflix project optimizes user experience by minimizing redundancy and tailoring content recommendations. Through advanced data analysis techniques, it enhances data-driven decision-making, fostering a more intuitive and personalized streaming platform. By leveraging these tools, the project aims to streamline content delivery, ultimately revolutionizing the Netflix user experience with innovation and efficiency.

6.2 Future Scope

jd

Future endeavors could expand by incorporating machine learning algorithms to predict user preferences more accurately. Additionally, integrating natural language processing (NLP) techniques could analyze user reviews and feedback, further refining content recommendations. Collaborations with content creators and studios could enhance the diversity and relevance of available content. Continual refinement and adaptation to evolving user behaviors and preferences will ensure sustained improvement in the Netflix user experience.

REFERENCES:

[1] Maddodi, S., 2019. NETFLIX bigdata analytics-the emergence of data driven recommendation. *Srivatsa Maddodi, & Krishna Prasad, K.(2019). Netflix Bigdata Analytics-The Emergence of Data Driven Recommendation. International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, *3*(2), pp.41-51.

[2] Fernández-Manzano, Eva-Patricia, Elena Neira, and Judith Clares-Gavilán. "Data management in audiovisual business: Netflix as a case study." *Profesional de la información/Information Professional* 25, no. 4 (2016): 568-577.

[2] Lad, A., Butala, S. and Bide, P., 2019, November. A comparative analysis of over-the-top platforms: Amazon Prime Video and Netflix. In *International Conference on Communication and Intelligent Systems* (pp. 283-299). Singapore: Springer Singapore.

[4] Wayne, M.L. and Uribe Sandoval, A.C., 2023. Netflix original series, global audiences and discourses of streaming success. *Critical Studies in Television*, *18*(1), pp.81-100.

[5] Wayne, M. L., & Uribe Sandoval, A. C. (2023). Netflix original series, global audiences and discourses of streaming success. *Critical Studies in Television*, *18*(1), 81-100.