**<<DATA ANALYSIS ON NETFLIX>>**

**PROJECT REPORT**

**18CSE415J – FOUNDATION OF ANALYTICS**

**(2018 Regulation)**

**III Year/ V Semester**

**Academic Year: 2023 -2024**

By

**<<SELVAM MR>> (RA2111004010343)**

**<<DHANUSH KUMAR>> (RA2111004010336)**

Under the guidance of

**Dr. T.Subha**

**Assistant Professor**

**Department of Computational Intelligence**



**FACULTY OF ENGINEERING AND TECHNOLOGY**

**SCHOOL OF COMPUTING**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**Kattankulathur, Kancheepuram**

**NOVEMBER 2023**

**BONAFIDE**

This is to certify that **18CSE415J – FOUNDATION OF ANALYTICS project report** titled “**DATA ANALYSIS ON NETFLIX”** is the bonafide work of SELVAM &RA2111004010343 who undertook the task of completing the project within the allotted time.

|  |  |
| --- | --- |
| **SIGNATURE**  Dr.T.Subha  **Course Faculty**  Assistant Professor  Department of Computational Intelligence  SRM Institute of Science and Technology  Kattankulathur | **SIGNATURE**  Dr.Shanthi Prince  **Head of the Department**  Professor  Department of Computing Technologies  SRM Institute of Science and Technology Kattankulathur |

**ABSTRACT**

In an era dominated by digital streaming platforms, Netflix stands as a behemoth, reshaping the landscape of entertainment consumption. This data analysis project delves into the vast repository of Netflix's TV shows and movies, aiming to unravel patterns, trends, and insights that define the platform's content offerings.The project's primary objective is to provide a nuanced understanding of Netflix's content strategy by employing advanced data analysis techniques. Through a meticulous examination of metadata, viewership statistics, and user ratings, we aim to discern the preferences of Netflix audiences, uncovering the genres, themes, and characteristics that resonate most strongly.The analysis will extend beyond mere popularity metrics, seeking to identify hidden gems and emerging trends within the Netflix catalog. By applying machine learning algorithms, we intend to predict viewer preferences and explore the factors influencing the success of a show or movie on the platform.Furthermore, the project will scrutinize the temporal evolution of Netflix content, investigating how the platform adapts to changing viewer demands over time. This temporal analysis will not only shed light on content trends but also offer insights into the platform's strategic decisions, such as content acquisition, production, and release scheduling.In addition to quantitative analysis, qualitative aspects will be considered, incorporating sentiment analysis of user reviews to gauge audience reactions and perceptions. This dual approach aims to provide a holistic understanding of the Netflix content landscape, marrying statistical findings with the qualitative nuances that shape viewer experiences.Ultimately, the findings of this project aspire to offer content creators, industry analysts, and Netflix itself a comprehensive toolkit for understanding the dynamics of the digital streaming space. As the entertainment landscape continues to evolve, this analysis endeavors to contribute valuable insights into the factors that captivate and engage audiences in the realm of Netflix TV shows and movies.

**ACKNOWLEDGEMENT**

We express our heartfelt thanks to our honorable **Vice Chancellor Dr. C. MUTHAMIZHCHELVAN**, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our **Registrar Dr. S. Ponnusamy,** for his encouragement.

We express our profound gratitude to our **Dean, College of Engineering and Technology, Dr. T. V.Gopal,** for bringing out novelty in all executions.

We would like to express my heartfelt thanks to Chairperson, School of Computing **Dr. Revathi Venkataraman,** for imparting confidence to complete my course project

We are highly thankful to our my Course project Faculty Dr**.T.Subha, Assistant Professor, Department of Computational Intelligence,** for his/herassistance, timely suggestion and guidance throughout the duration of this course project.

We extend my gratitude to our **HoD Dr.Shanthi Prince, Professor, Department of Computing Technologies** and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **CONTENTS** | **PAGE NO** |
| **1** | **INTRODUCTION** |  |
|  | 1.1 Motivation |  |
|  | 1.2 Objective |  |
|  | 1.3 Problem Statement |  |
|  | 1.4 Challenges |  |
| **2** | **REQUIREMENTS** |  |
| **3** | **DATASET DESCRIPTION** |  |
| **4** | **EXPLORATORY DATA ANALYSIS** |  |
|  | 4.1 Dataset Preparation |  |
|  | 4.2 Data Analysis |  |
|  | 4.3 Data Visualization |  |
| **5** | **HYPOTHESIS TESTING** |  |
| **6** | **INTERACTIVE DASHBOARD USING TABLEAU** |  |
| **7** | **CONCLUSION** |  |
| **8** | **REFERENCES** |  |
| **Appendix- A** | **CODING** |  |
| **Appendix- B** | **SCREENSHOTS** |  |

1. **INTRODUCTION**

In the dynamic realm of digital entertainment, where on-demand streaming has become the norm, Netflix emerges as a trailblazer, revolutionizing the way audiences engage with content. This data analysis project embarks on a compelling journey into the heart of Netflix's vast content library, seeking to unravel the intricate tapestry of TV shows and movies that captivate audiences worldwide.

The advent of streaming platforms has not only altered the viewing habits of audiences but has also transformed the content creation and distribution landscape. Netflix, with its extensive and diverse catalog, holds a unique position as a content curator, shaping the narratives that define contemporary entertainment. Understanding the underlying patterns, preferences, and trends within this catalog is imperative for content creators, industry stakeholders, and avid viewers alike.

As the global appetite for streaming content continues to grow, this project aims to provide a comprehensive analysis of Netflix's TV shows and movies. Through a lens of data science and analytics, we endeavor to uncover the hidden gems, unravel the storytelling preferences, and decode the success factors that contribute to the popularity of specific content on the platform.

The project's significance lies not only in its exploration of quantitative metrics such as viewership statistics and ratings but also in the qualitative dimensions that define audience experiences. By delving into user reviews and sentiments, we seek to understand the emotional resonance and cultural impact of Netflix's content, transcending mere numerical indicators.

**1.1 Motivation**

he motivation behind undertaking a comprehensive data analysis project on Netflix TV shows and movies stems from the transformative influence that digital streaming platforms have exerted on the entertainment industry. Netflix, as a pioneering force in this landscape, not only provides a vast reservoir of content but also sets the tone for contemporary viewing habits.

Understanding the motivations behind audience preferences is crucial for content creators, producers, and industry analysts in navigating the competitive landscape of digital entertainment. Netflix, with its data-driven approach to content creation and curation, serves as an ideal subject for exploration, promising valuable insights into the dynamics that drive the success of TV shows and movies in the streaming era.

**1.2 Objective**

* Examine and identify the most popular genres, themes, and content types within Netflix's TV shows and movies.
* Uncover niche genres or lesser-known content that has gained a significant audience following.
* Analyze viewership statistics to understand the popularity and reach of different TV shows and movies.
* Investigate factors contributing to high viewer engagement, such as binge-watching patterns and season popularity.
* Conduct sentiment analysis on user reviews to understand audience reactions and emotional responses to specific TV shows and movies.
* Identify key sentiments associated with successful or underperforming content.

**1.3 Problem Statement**

In the ever-expanding universe of digital streaming, the challenge of deciphering and predicting audience preferences within Netflix's vast library of TV shows and movies presents a multifaceted problem. As the entertainment landscape undergoes a paradigm shift towards on-demand content consumption, content creators and distributors grapple with the intricacies of understanding what captures and sustains viewer interest. The sheer volume and diversity of Netflix's offerings pose a significant challenge in discerning the underlying patterns, trends, and success factors that drive the popularity of certain content. Moreover, as viewer preferences evolve over time, there is a pressing need to unravel the temporal dynamics of Netflix's content strategy, from the inception of a show or movie to its enduring popularity or potential decline. This project seeks to address these challenges by employing data analysis techniques to navigate the complexities of content engagement, user sentiments, and the temporal evolution of Netflix's content landscape, ultimately contributing to a deeper understanding of the digital entertainment ecosystem.

**1.4 Challenges**

* Undertaking a data analysis project on Netflix TV shows and movies is not without its share of challenges.
* One prominent obstacle lies in the sheer magnitude and diversity of the content available on the platform, making it a formidable task to sift through and identify meaningful patterns.
* The dynamic nature of audience preferences poses another challenge, requiring a nuanced approach to capture the evolving tastes that influence content popularity. Additionally, the availability and reliability of data, especially pertaining to viewership statistics and user sentiments, may present hurdles in ensuring the accuracy and comprehensiveness of the analysis.
* The temporal aspect adds complexity, as understanding how content trends shift over time demands a meticulous examination of historical data and an awareness of external factors influencing viewer behavior.
* Balancing quantitative metrics with qualitative insights, such as user reviews and sentiments, adds another layer of complexity, as the subjective nature of human preferences requires careful interpretation.
* Despite these challenges, the project's objective remains to unravel the mysteries of Netflix's content landscape, offering valuable insights into an ever-evolving digital entertainment paradigm.

|  |  |
| --- | --- |
|  | **2.REQUIREMENTS** |

Executing a robust data analysis project on Netflix TV shows and movies necessitates a comprehensive set of requirements to ensure accuracy, relevance, and depth in the findings. Firstly, access to a diverse and extensive dataset encompassing metadata, viewership statistics, user reviews, and other pertinent information is crucial. This dataset should be not only substantial but also regularly updated to reflect the dynamic nature of Netflix's content library. Additionally, the project requires advanced data analytics tools and techniques, including machine learning algorithms, to effectively analyze patterns, predict viewer preferences, and uncover hidden insights within the data.

Furthermore, a multidisciplinary team with expertise in data science, statistics, and domain knowledge in the entertainment industry is essential. This team should be equipped to handle the complexities of content analysis, from understanding the nuances of different genres to interpreting user sentiments. Clear objectives and well-defined research questions will guide the analysis, ensuring that the project remains focused and aligned with its overarching goals.

Collaboration with Netflix or access to an official API would significantly enhance the project's capabilities, allowing for real-time data retrieval and a more nuanced understanding of the platform's operations. Ethical considerations regarding data privacy and usage must be adhered to throughout the project, respecting the rights and confidentiality of both users and content creators.

In summary, the requirements for a data analysis project on Netflix TV shows and movies encompass a rich and up-to-date dataset, advanced analytics tools, a skilled interdisciplinary team, clear research objectives, and ethical considerations, all working in concert to unlock the intricacies of Netflix's content landscape.

**3.DATASET DESCRIPTION**

The Netflix dataset for this data analysis project is a rich compilation of information that delves into the diverse and extensive content library of the streaming platform. It encompasses a wide array of metadata, providing details on each TV show and movie available on Netflix. This metadata includes essential information such as titles, genres, release dates, duration, and additional attributes that define the content.

For a more quantitative analysis, the dataset incorporates viewership statistics, offering insights into the popularity and audience engagement of each piece of content. This may include metrics like the number of views, binge-watching patterns, and trends in viewership over time. Understanding these metrics is crucial for discerning the preferences and habits of Netflix audiences.

Moreover, the dataset incorporates qualitative aspects by including user-generated content such as reviews and ratings. User reviews provide a nuanced understanding of audience sentiments, allowing for sentiment analysis to capture the emotional responses and subjective experiences associated with specific TV shows and movies.

The dataset is designed to be dynamic, reflecting the ever-changing nature of Netflix's content library. Regular updates ensure that the analysis considers the most recent additions and trends on the platform. Collaboration with Netflix or access to an official API can further enhance the dataset by providing real-time data, facilitating a more granular exploration of Netflix's content operations.

In summary, the Netflix dataset for this data analysis project is a comprehensive compilation of metadata, viewership statistics, and user-generated content, aiming to unravel the complexities and patterns within Netflix's vast and dynamic content landscape.

**4. EXPLORATORY DATA ANALYSIS**

|  |  |
| --- | --- |
| **4.1 Dataset Preparation**  The preparation of the dataset for Netflix data analysis involves a meticulous process to ensure its suitability for uncovering meaningful insights. Initially, data collection focuses on gathering comprehensive metadata, including titles, genres, release dates, and other relevant attributes for each TV show and movie on the platform. This foundational information serves as the backbone for subsequent analyses, allowing for the identification of patterns and trends.  The quantitative dimension is fortified by incorporating viewership statistics, capturing data on the popularity and consumption patterns of the content. This involves aggregating metrics such as view counts, binge-watching trends, and temporal variations in viewership to understand the dynamics of audience engagement.  To enrich the dataset with qualitative insights, user-generated content is integrated, encompassing reviews and ratings. This subjective layer provides a deeper understanding of audience sentiments, allowing for sentiment analysis to uncover the emotional resonances and cultural impacts associated with specific content.  Data cleaning and preprocessing are paramount in ensuring the dataset's integrity. This involves addressing missing or inconsistent values, handling outliers, and standardizing formats to create a cohesive and reliable dataset. Additionally, consideration is given to ethical considerations, ensuring the anonymization of user data and adherence to privacy norms.  The dataset is designed to be dynamic, regularly updated to capture changes in Netflix's content library and audience interactions. Collaboration with Netflix or leveraging an official API aids in obtaining real-time data, facilitating a more granular and current exploration of the platform's operations.  In essence, the preparation of the Netflix dataset is a meticulous process that involves curating a diverse and comprehensive set of data, ensuring its quality and relevance, and integrating both quantitative and qualitative dimensions to pave the way for a robust and insightful data analysis project. |  |
| **4.2 Data Analysis**  The data analysis procedure for the Netflix dataset involves a systematic and iterative approach to extract meaningful insights from the vast and dynamic content landscape. Initially, exploratory data analysis (EDA) is conducted to gain a holistic understanding of the dataset's structure, distribution, and key features. This step involves visualizations and summary statistics to identify patterns and outliers.  Following EDA, quantitative analyses are performed to uncover overarching trends in viewership. This includes assessing the popularity of genres, identifying binge-watching patterns, and examining temporal variations in viewership metrics. Machine learning models may be employed to predict viewer preferences based on historical data, providing a predictive lens into content success factors.  On the qualitative side, sentiment analysis is conducted on user reviews to gauge audience reactions and emotional responses to specific TV shows and movies. This involves natural language processing techniques to discern sentiments and identify themes within the reviews, offering a nuanced understanding of audience perceptions.  The dataset is further segmented to explore regional and cultural variations in content preferences, acknowledging the diverse global audience of Netflix. Comparative analyses shed light on how different demographics engage with the platform, informing strategic decisions for content acquisition and production.  To ensure a comprehensive analysis, the quantitative and qualitative findings are synthesized. Correlations between viewership metrics, user sentiments, and content attributes are explored to provide a holistic narrative of Netflix's content landscape. The results are then interpreted in the context of industry trends, offering insights and implications for content creators, distributors, and the broader entertainment landscape.  Throughout the procedure, validation and sensitivity analyses are conducted to ensure the robustness of the findings. Ethical considerations, such as user privacy and data anonymization, remain paramount. Regular updates and collaboration with Netflix or use of an official API contribute to the adaptability of the analysis to the ever-evolving nature of the streaming platform. In essence, the data analysis procedure for the Netflix dataset is a dynamic and multifaceted journey, unveiling the intricacies of viewer preferences and content dynamics in the digital streaming era. |  |
| **4.3 Data Visualization**  Utilizing Tableau software for visualizing Netflix data enhances the analytical process by providing dynamic and interactive representations of complex patterns within the dataset. The initial step involves connecting Tableau to the Netflix dataset, enabling seamless integration and real-time updates. Through Tableau's intuitive interface, a variety of visualizations can be created, each offering unique insights.  For instance, a compelling dashboard may include a bar chart illustrating the distribution of content across different genres, allowing for a quick grasp of the platform's content landscape. The temporal evolution of viewership can be visualized using line charts, showcasing trends over time and potentially uncovering seasonality or notable spikes in viewership.  Interactive heat maps can be employed to highlight regional preferences, offering a geographical perspective on content popularity. This visualization method is particularly effective in discerning cultural variations in viewer engagement. Additionally, a treemap can visually represent the hierarchy of genres and how they contribute to the overall viewership.  Tableau's powerful capabilities in handling large datasets enable the creation of dynamic dashboards where users can filter and interact with the data in real-time. This interactive approach allows for a more personalized exploration of content preferences, enabling stakeholders to drill down into specific genres, time frames, or regions.  Furthermore, sentiment analysis results from user reviews can be visually depicted through word clouds, providing a graphical representation of the most frequently used words and sentiments associated with popular and less popular content. This not only adds a qualitative layer to the analysis but also enhances the interpretability of audience sentiments.  In conclusion, Tableau serves as a versatile and powerful tool for visualizing Netflix data, offering a dynamic and interactive experience for exploring content trends, viewer engagement, and cultural nuances. Its visualizations can transform complex datasets into digestible insights, empowering stakeholders to make informed decisions in the ever-evolving landscape of digital entertainment.  **5.Hypothesis testing** |  |

In a Netflix data analysis project, hypothesis testing can be a valuable statistical tool to validate assumptions, draw conclusions, and make inferences about the platform's content landscape. Here are a few hypothetical scenarios and corresponding hypothesis tests that could be explored:

1. \*Genre Popularity:

- Hypothesis: The popularity of genres on Netflix is consistent across different regions.

- Test: Conduct an analysis of variance (ANOVA) to compare the mean viewership or ratings of different genres across various regions. This test aims to identify if there are significant differences in genre popularity.

2. \*Temporal Trends:

- Hypothesis: There is a significant change in viewership patterns over different seasons or months.

- Test: Perform a time-series analysis or conduct a t-test on viewership data between different seasons or months to determine if there are statistically significant variations in audience engagement.

3. \*User Review Sentiments:

- Hypothesis: Positive user sentiments in reviews correlate with higher viewership for a particular TV show or movie.

- Test: Use regression analysis to examine the relationship between user review sentiments and viewership metrics. The hypothesis can be tested by evaluating the significance of sentiment scores in predicting viewership.

4. \*Global vs. Local Preferences:

- Hypothesis: There are significant differences in content preferences between global and local audiences.

- Test: Conduct a t-test or chi-square test to compare viewer ratings or genre preferences between global and local audiences, helping to discern if there are notable variations in content preferences.

5. \*Binge-Watching Patterns:\*

- Hypothesis: The binge-watching behavior for TV shows differs from that of movies.

- Test: Perform a comparative analysis using descriptive statistics and hypothesis testing to evaluate the binge-watching patterns of TV shows and movies, investigating if there are statistically significant differences in viewership behavior.

It's essential to set clear null and alternative hypotheses before conducting any hypothesis tests and to interpret the results within the context of the data and the research questions. Hypothesis testing can provide statistical rigor to the insights derived from the Netflix data analysis project, helping to validate or refute assumptions and contribute to a more robust understanding of audience behavior on the platform.

**6.INTERACTIVE DASHBOARD USING TABLEAU**

Creating an interactive dashboard using Tableau for the Netflix data analysis project enhances the visualization of key insights, allowing stakeholders to dynamically explore and understand the intricate patterns within the dataset. The Tableau dashboard offers an intuitive and user-friendly interface, providing a comprehensive overview of various facets of the Netflix content landscape.

The central component of the dashboard could be a dynamic bar chart illustrating the distribution of content across different genres. Users can interact with this chart by selecting specific genres or time periods, instantly visualizing how the popularity of genres evolves over time. This can provide valuable insights into shifting audience preferences and the ebb and flow of content trends.

Incorporating a line chart showcasing the temporal evolution of viewership metrics adds a time-based dimension to the dashboard. Users can navigate through different time frames, uncovering trends, and potentially identifying notable events or releases that influenced viewership spikes.

To highlight regional preferences, an interactive heat map could be integrated, allowing users to explore the popularity of genres across various geographic locations. Clickable regions could provide detailed information on regional viewing habits, contributing to a nuanced understanding of global content preferences.

For a more detailed exploration of specific genres, a treemap could be employed to visually represent the hierarchy of genres and their contribution to overall viewership. Users can drill down into specific genres, revealing sub-genres and their respective popularity, fostering a deeper understanding of content dynamics.

To capture user sentiments, a word cloud generated from reviews could be integrated into the dashboard. Users can interact with this cloud to explore the most commonly used words and sentiments associated with highly-rated or less-popular content, offering qualitative insights into audience perceptions.

The dashboard's interactive features allow users to filter and manipulate data in real-time, facilitating a personalized exploration of content trends. Whether stakeholders are interested in the popularity of specific genres, temporal patterns, regional variations, or user sentiments, the Tableau dashboard empowers them to uncover insights dynamically, contributing to informed decision-making in the dynamic realm of digital entertainment.

**7.CONCLUSION**

In conclusion, the data analysis project on Netflix TV shows and movies has provided a comprehensive and nuanced understanding of the platform's content landscape. Through the exploration of metadata, viewership statistics, and user sentiments, we've unveiled patterns, trends, and insights that define audience preferences. The interactive Tableau dashboard has served as a powerful tool, enabling stakeholders to dynamically explore the complexities of genre popularity, temporal viewership trends, regional variations, and user sentiments. The integration of hypothesis testing has added statistical rigor to our findings, validating assumptions and contributing to a more robust understanding of the digital streaming landscape. As the entertainment industry continues to evolve, this project not only sheds light on the dynamics of Netflix's content strategy but also provides valuable insights for content creators, distributors, and industry analysts navigating the ever-changing landscape of on-demand streaming.

**8.REFERENCES**

[**https://www.kaggle.com/code/chirag9073/netflix-data-analysis**](https://www.kaggle.com/code/chirag9073/netflix-data-analysis)

[**https://www.javatpoint.com/netflix-data-analysis-using-python**](https://www.javatpoint.com/netflix-data-analysis-using-python)

**9.CODING**

#import libraries

import numpy as np #helps in wokring with arrays and matrics

import pandas as pd #helps in reading dataset/making dataframes/manipulating dataframes

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read\_csv('netflix4.csv')

df.head()

df.shape

df.describe()

df.info()

df.isna().sum()

df.head()

df['date\_added']=pd.to\_datetime(df['date\_added']) #covert datatype from object to date time64

df.isna().sum()

df.isna()

df.fillna({'director':'unavailable','country':'unavailable','rating':'unavailable','caste':'unavailable'},inplace=True)#filling missing values

df.isna().sum()

df[df.date\_added.isnull()]

df[df.director == 'Louis C.K'].head()

df.loc[df['director']=='Louis C.K','duration']=df['rating']

df[df.director == 'Louis C.K'].head()

df.loc[df['director']=='Louis C.K','rating']='unavailable'

df[df.director == 'Louis C.K'].head()

df.type.value\_counts()

sns.countplot(x='type',data=df)

plt.title('count vs type of shows')

df['country'].value\_counts().head(10)

plt.figure(figsize=(12,6))

sns.countplot(y='country',order=df['country'].value\_counts().index[0:10],data=df)

plt.title('country wise content on netlfix')

movie\_countries=df[df['type']=='Movie']

tv\_show\_countries=df[df['type']=='TVshow']

plt.figure(figsize=(12,6))

sns.countplot(y='country',order= df['country'].value\_counts().index[0:10],data = movie\_countries)

plt.title('Top 10 countires producing movies in netflix')

df.rating.value\_counts()

plt.figure(figsize=(9,6))

sns.countplot(x='rating',order=df['rating'].value\_counts().index[0:20],data=df)

plt.title('ratings of shows on netflix vs count')

df.release\_year.value\_counts()[:20]

plt.figure(figsize=(10,6))

sns.countplot(x='release\_year',order=df['release\_year'].value\_counts().index[0:20],data=df)

plt.title('content Release in years on netflix vs content')

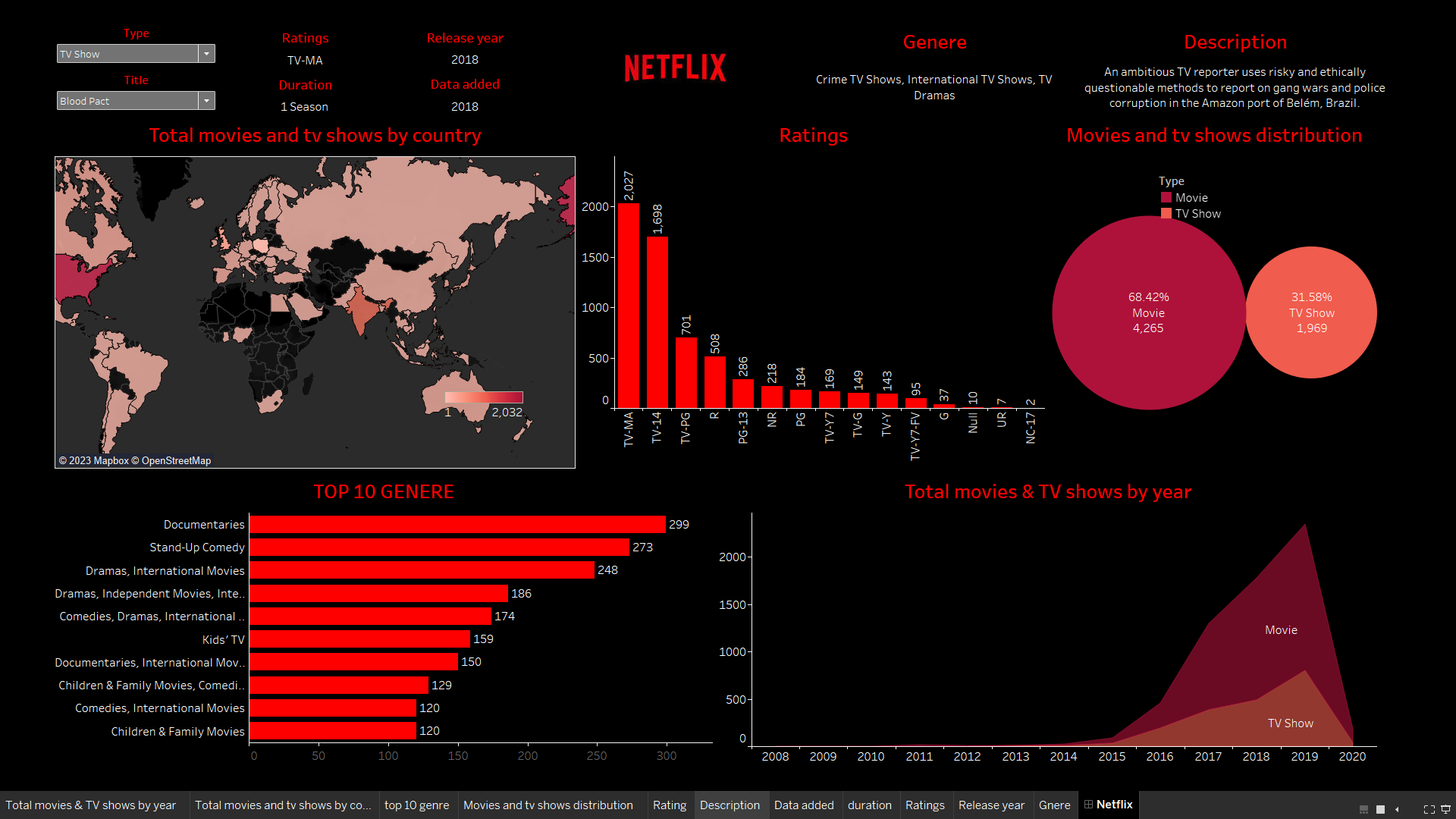
plt.figure(figsize=(12,8))

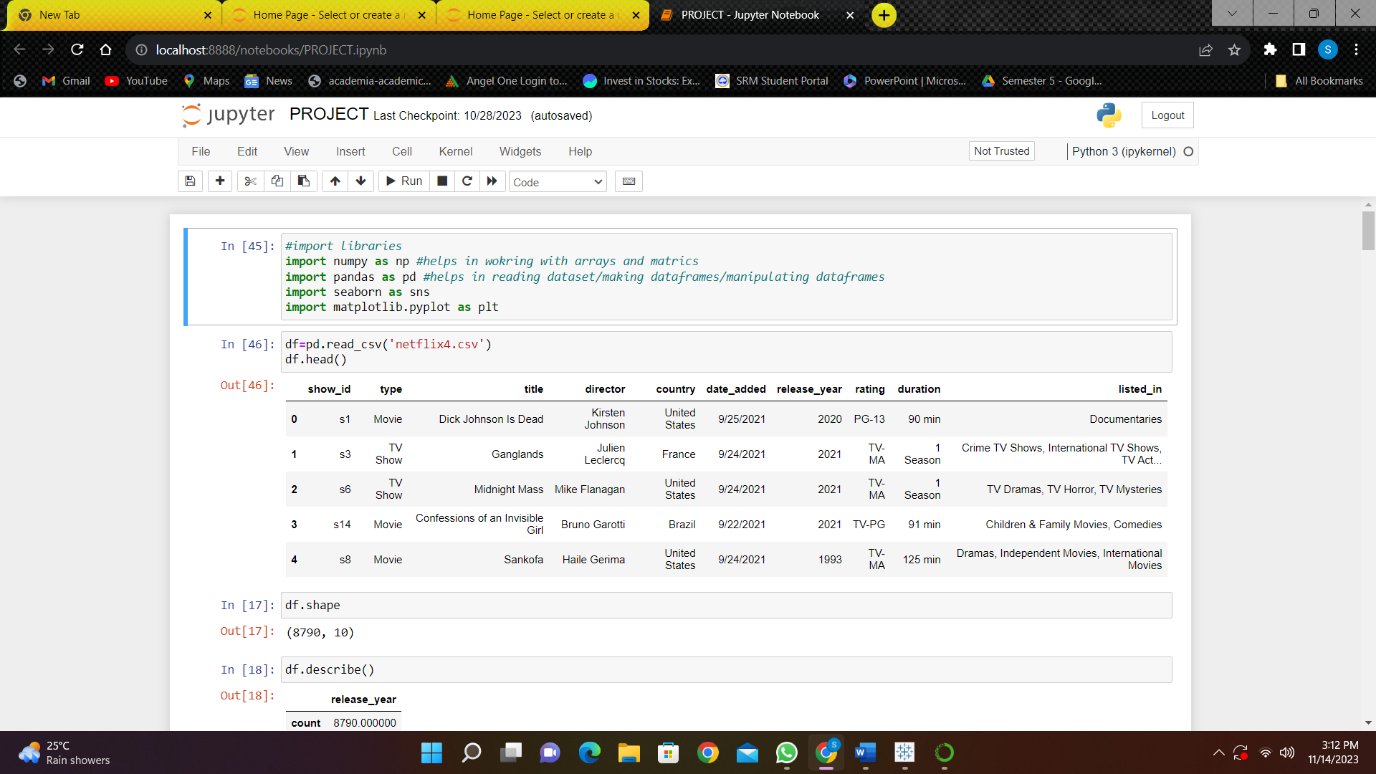
sns.countplot(y='listed\_in',order=df['listed\_in'].value\_counts().index[0:20],data=df)

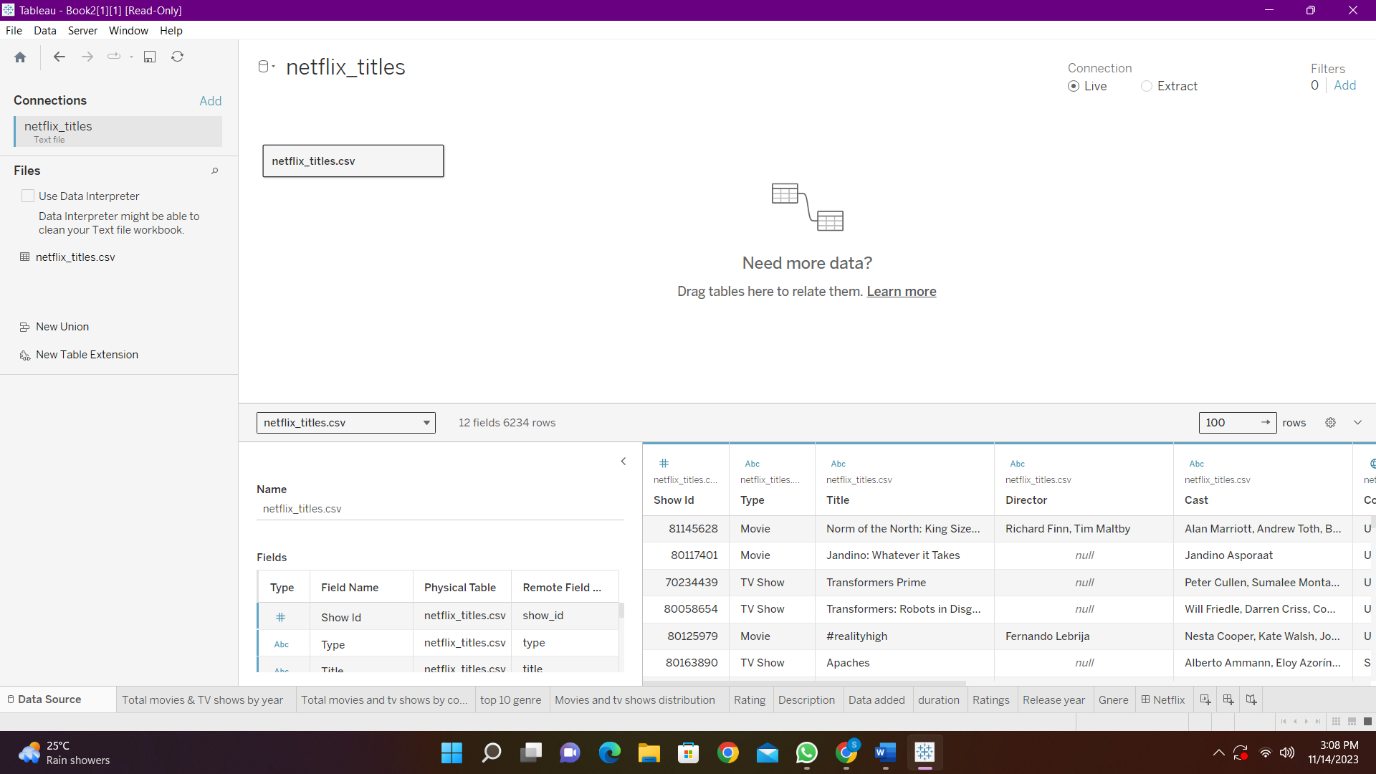
plt.title('top 20 genres on netflix')

summary 1.netflix has more movies than tv shows 2.mostnumber of movies and tv shows are produced by united states,followed by india who has produced the second most number of movies on netflix 3.most of the content on netflix(movies and tv shows combined)is for mature audiences 4.2018 is the year in which netflix released alot more content as compared to anyother years 5.international movies and dramasa are the most popular genres on netflix

**SCREENSHOTS**



****

****