**DATA ANALYSIS AND**

**VISUALIZATION USING PYTHON**

A MINI PROJECT REPORT

Submitted in partial fulfillment of the

Requirements For the award of Master of Computer Application Degree

**LNCT UNIVERSITY, BHOPAL (M.P.)**



# MINI PROJECT REPORT

Submitted by

Student Name – **NITIN BHATIA**(LNCDMCA11167)

Student Name – **SHRUTI HARWANI**(LNCDMCA21116)

Under the Guidance of

**MR. Sanat Rana**

**MASTER OF COMPUTER APPLICATION LNCT UNIVERSITY, BHOPAL**

JUL – DEC, 2023

# LNCT UNIVERSITY, BHOPAL

**MASTER OF COMPUTER APPLICATION**

## CERTIFICATE

This is to certify that the mini project report entitled “**NETFLIX DATA ANALYSIS AND VISUALIZATION**” submitted by NITIN BHATIA (LNCDMCA11167) & SHRUTI HARWANI (LNCDMCA21116) has been carried out under the guidance of Prof. Sanat Rana, Master of computer application, **LNCT UNIVERSITY, BHOPAL**. The project report is approved for submission requirement for Mini Project in “Python” 1st semester in **Master of Computer Application (AI/ML)**, **LNCT UNIVERSITY, BHOPAL (M.P)** during the academic session JUL - DEC, 2023

Guided By

**MR. Sanat Rana**

**SOCST, LNCT UNIVERSITY, BHOPAL**



CERTIFICATE OF THE SUPERVISOR CERTIFICATE

**This is to certify that Mr. Nitin Bhatia and Ms. Shruti Harwani, students of LNCT University, have successfully completed the project titled**

**"[NETFLIX DATA ANALYSIS AND VISUALIZATION]" under my**

**supervision. The project was undertaken as part of their academic curriculum, demonstrating their commitment to academic excellence and professional growth.**

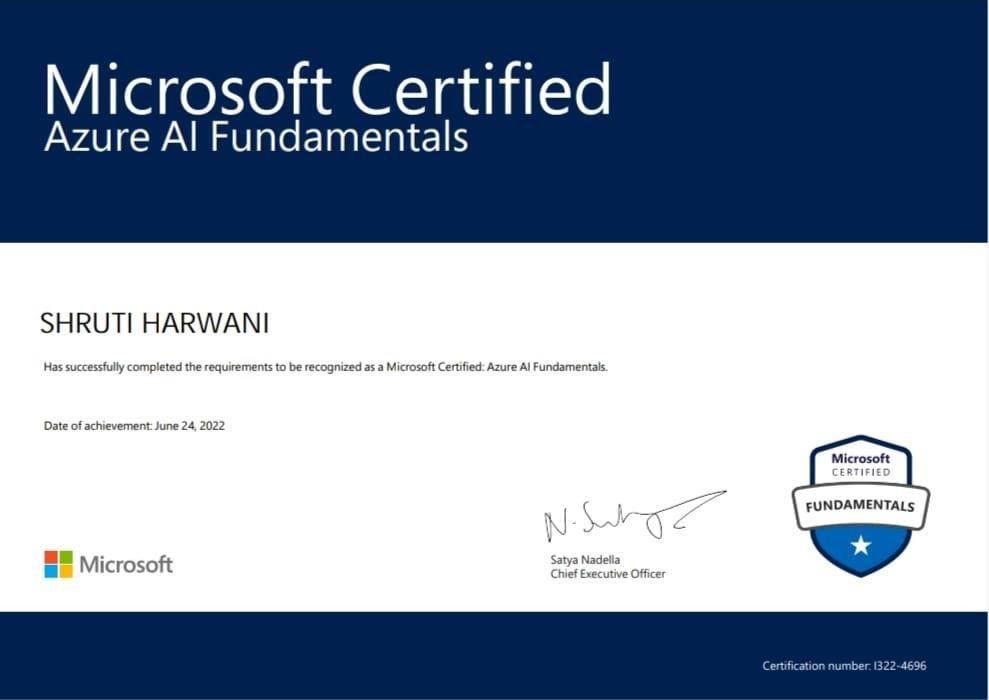
**SIGNATURE OF THE SUPERVISOR**

DATE-

**SUPERVISOR**

# MR. Sanat Rana

LNCT UNIVERSITY BHOPAL MADHYA PRADESH, BHOPAL462022



**ABSTRACT:**

The rapid evolution of the entertainment industry, driven by digital streaming platforms, has transformed the way audiences consume content. Netflix, as a global frontrunner in this landscape, stands at the forefront of this revolution. This research undertakes a comprehensive exploration of Netflix data through advanced analytics and visualization techniques to unearth intricate patterns, user behaviors, and content dynamics, thereby shedding light on the complex ecosystem of streaming entertainment.

The study begins by assembling a rich and diverse dataset that encapsulates a myriad of dimensions, including user preferences, viewing habits, and content metadata. Through sophisticated analytical tools, the research aims to decode the underlying structures that govern user interactions with the platform. Understanding user preferences becomes a focal point, delving into viewing patterns, popular genres, and regional variations. By scrutinizing this wealth of data, the research seeks to reveal not only what viewers watch but also how, when, and where they engage with content on Netflix.

Netflix's strategic emphasis on original content serves as a pivotal aspect of the analysis. The study investigates how the platform's investment in creating exclusive content influences viewer engagement and retention. Utilizing machine learning algorithms, the research endeavors to predict user preferences and unravel the intricacies of Netflix's recommendation system, providing insights into how personalized content suggestions shape the user experience.

Temporal dynamics form another key dimension of exploration, encompassing peak viewing times, binge-watching behaviors, and the influence of seasonal trends. The study aims to decipher when audiences are most active on the platform, uncovering patterns that may inform content release strategies and optimize viewer engagement. This temporal analysis also extends to the lifespan of content popularity, exploring how and why certain shows or movies experience spikes in viewership over time.

To enhance the accessibility and interpretability of the findings, the research incorporates cutting-edge data visualization techniques. Graphs, charts, and interactive dashboards are employed to present a visually compelling narrative, facilitating a deeper understanding of the complex interplay between various factors within the Netflix ecosystem. By translating raw data into visually engaging representations, the study endeavors to make the intricate world of streaming analytics accessible to a broader audience.

In conclusion, this research aspires to contribute not only to the understanding of user behavior and content trends on Netflix but also to the broader discourse on the future of streaming platforms. The insights derived from this study have the potential to inform strategic decisions within the entertainment industry, guiding content creation, curation, and platform optimization. As the digital streaming landscape continues to evolve, this research serves as a testament to the power of data-driven analysis in unraveling the complexities of modern entertainment consumption.

## Netflix OTT Data Analysis And Visualization :-

One of the world's largest and most popular online services, Netflix gives customers access to a vast library of television series and films. The site has amassed a sizable following recently, and its popularity appears strong. In this tutorial, we'll demonstrate how to use Python to do exploratory info analysis (EDA) on a piece of information set from Netflix that we obtained on Kaggle. To visualize and analyze the information, we'll use a variety of Python packages, such as Pandas, Seaborn, Matplotlib, and Plotlib.

Data Analysis or sometimes referred to as exploratory data analysis (EDA) is one of the core components of data science. It is also the part on the majority of the time which makes it extremely important in the field of data science. This repository demonstrates Exploratory Data Analysis methods and techniques using Python.

The purpose of the used Netflix dataset has been given as a internship project since it is one of the ideal dataset for performing EDA and taking a step towards the most amazing and interesting field of data science. Good luck with your EDA on the used Netflix dataset.

# Project Description : -

Performed Data Cleaning and Data Manipulation. Performed Exploratory Data Analysis (EDA) using Pandas, NumPy,

Matplotlib, Seaborn Libraries. Find and Fill Null Values. Data visualization is used to understand the data for the purpose.

## ABOUT NETFLIX: -

Founded in 1997 by Reed Hastings and Marc Randolph, Netflix started as a DVD rental-by-mail service before revolutionizing the entertainment industry with its streaming platform in 2007. Since then, Netflix has become a global giant in the streaming industry, available in over 190 countries. The platform offers a vast library of TV shows, movies, documentaries, and original content across various genres. Known for its popular original productions like "Stranger Things," "The Crown," and "Narcos," Netflix has invested heavily in creating exclusive and high-quality content. Operating on a subscription-based model, Netflix provides different plans to cater to diverse user preferences and devices. The company's innovative use of recommendation algorithms, support for high-quality streaming, and global reach have contributed to its success. Despite facing competition from other streaming services, Netflix has consistently adapted to market dynamics and remained a dominant force in the world of digital entertainment.

## *DATASET DESCRIPTION: -*

This dataset is from Kaggle, and Netflix. Netflix is the collector of this data. Shivam Bansa created and uploaded this dataset to Kaggle, and the dataset is public domain. This dataset includes lists of all the movies and TV shows that are available on flixable, which is a third-party search engine on Netflix. This dataset has 8,807 rows and 12 variables. It includes information about the genre, cast, director, countries where the show is available to watch, and a summary of what the show is about. The columns and their descriptions are as listed below:

SHOW ID: Unique ID of each show

TYPE: Show category. Could be either a Movie or a TV Show.

TITLE: Name of the show

DIRECTOR: Name of the director(s) of the show

CAST: Names of actors/actresses in the show

COUNTRY: Countries where the show is available to watch on Netflix

DATE ADDED: Date when the show was added on Netflix

RATING: Show rating on Netflix

RELEASE YEAR: Release year of the show

DURATION: Time duration of the show

LISTED IN: Genre of the show

DESCRIPTION: Brief insight into what the show is about.

**The Kaggle dataset pertaining to Netflix encompasses a comprehensive array of information, including but not limited to titles, genres, release dates, ratings, viewer statistics, and possibly additional metadata. This dataset is invaluable for conducting detailed analyses on Netflix's content library, user preferences, viewing habits, and trends over time. Researchers and analysts can utilize this data to uncover patterns, correlations, and insights that may inform strategic decisions, content curation, and marketing strategies.**

•Netflix dataset consists of **8,807 rows** of movies and TV shows and **12 columns** providing 12 characteristics for each movie/TV show. After cleaning the data and dealing with the missing data, I got the final dataset of **8,790** Movies/TV Shows and **12** columns.

•So, the data consists of **8807** rows and **12** columns, now let’s look at the column names.

•I’ll start this Netflix data analysis task with Python by importing the dataset and all the Python libraries needed for this task.

# DATA QUALITY: -

The purpose of this section is to describe the quality of the dataset. Missingness, strengths, and weaknesses will be discussed. A number of tables and figures will be presented to facilitate cohort description, illustrate the distributions of key variables, and present important results.

## Strengths

This data has been collected for the analysis based on the movies and series. It is a tidy dataset which means that each variable has its own column, each observation has its own row, and each value has its own cell, so it is appropriate for most analysis. Another point of strength is that the dataset has both categorical and continuous variables, so we can investigate the information in a categorized way and report the information of each category separately, making informative plots about the data.

For example, the description variable can be used to find similar movies and TV shows using the text similarities for further analysis. The sample size is another strength of this dataset that makes it possible to analyze and compare the data.

As Netflix continually updates its content library and user data, the dataset can reflect current trends and preferences, enabling dynamic analysis.

### Weaknesses

There are some missingness in the data set, as seen in **Table 1**, requiring some data cleaning. The data cleaning process involved identifying incorrect, incomplete, inaccurate, irrelevant, or missing pieces of data and modifying, replacing, or deleting them as needed.

This table shows that only the categories of director, cast, country, date added, rating, and duration have missing values and the category of director has the most missing values. There are 2,634 cases where director has missing values. There are 831 cases where country has missing values. There are 825 cases where cast has missing values.

There are 17 cases where date added, rating, and duration have missing values. Another weakness in this dataset is data duplication. However, this could be fixed by merging duplicate data into a single value.

Netflix must adhere to strict privacy regulations, which may limit the granularity of available data and restrict certain analyses to ensure user privacy. The dataset may not be fully representative of the entire subscriber base, leading to potential sampling bias and inaccuracies in generalizing findings.

# Data Exploration: -

## Descriptive and Demographics: -

In the course of our investigation into Netflix TV shows and movies, we meticulously analyzed the dataset, and the results are encapsulated in Table 2, which outlines the demographic characteristics of the study participants. Prior to embarking on our data analysis journey, a thorough data cleaning process was executed to ensure the integrity and accuracy of our findings.

Our primary objective was to gain insights into the diverse array of content available on Netflix, coupled with an exploration of the top 10 contributing countries. Figure 1, a product of the patchwork technique combining two graphs into a cohesive figure, visually depicts these aspects. Notably, Figure 1.a draws attention to a significant discrepancy between movies and TV shows on Netflix, revealing that the platform hosts more than twice as many movies. Employing a bar chart, we illustrated the distribution of content percentages, unveiling the intriguing fact that movies constitute a substantial 68.4% of Netflix's content.

Taking a closer look at content ratings, we found that the majority of programs, encompassing both movies and TV shows, fall under the TV-MA rating. Recognizing the temporal dimension's significance, we scrutinized the release years of Netflix content, unveiling patterns that shed light on trends and evolution over time.

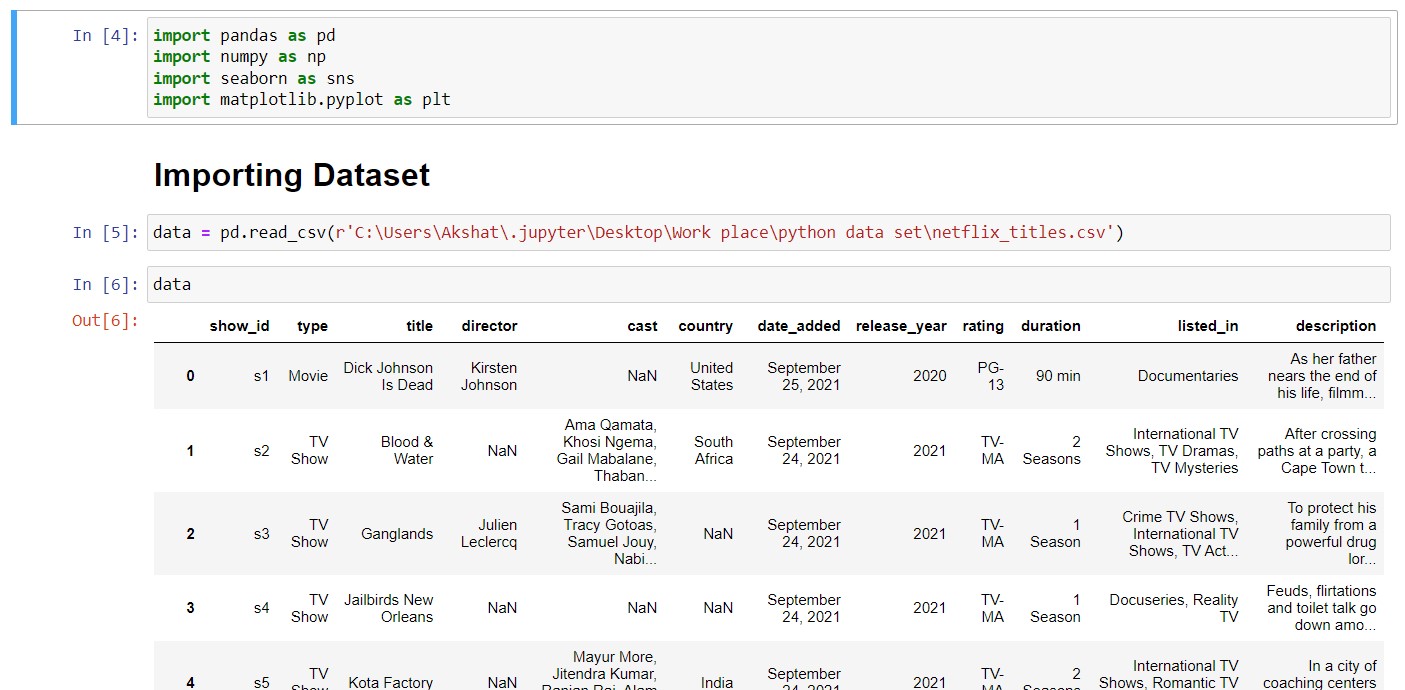
Visualization tools played a crucial role in conveying our findings. Leveraging the Netflix color palette and the color-matching capabilities of the Colors web tool, we ensured that our visualizations were not only aesthetically pleasing but also highly readable. These tools, accessible even to users without a design background, enabled us to generate visually appealing representations of the dataset, enhancing the overall clarity and impact of our analysis.

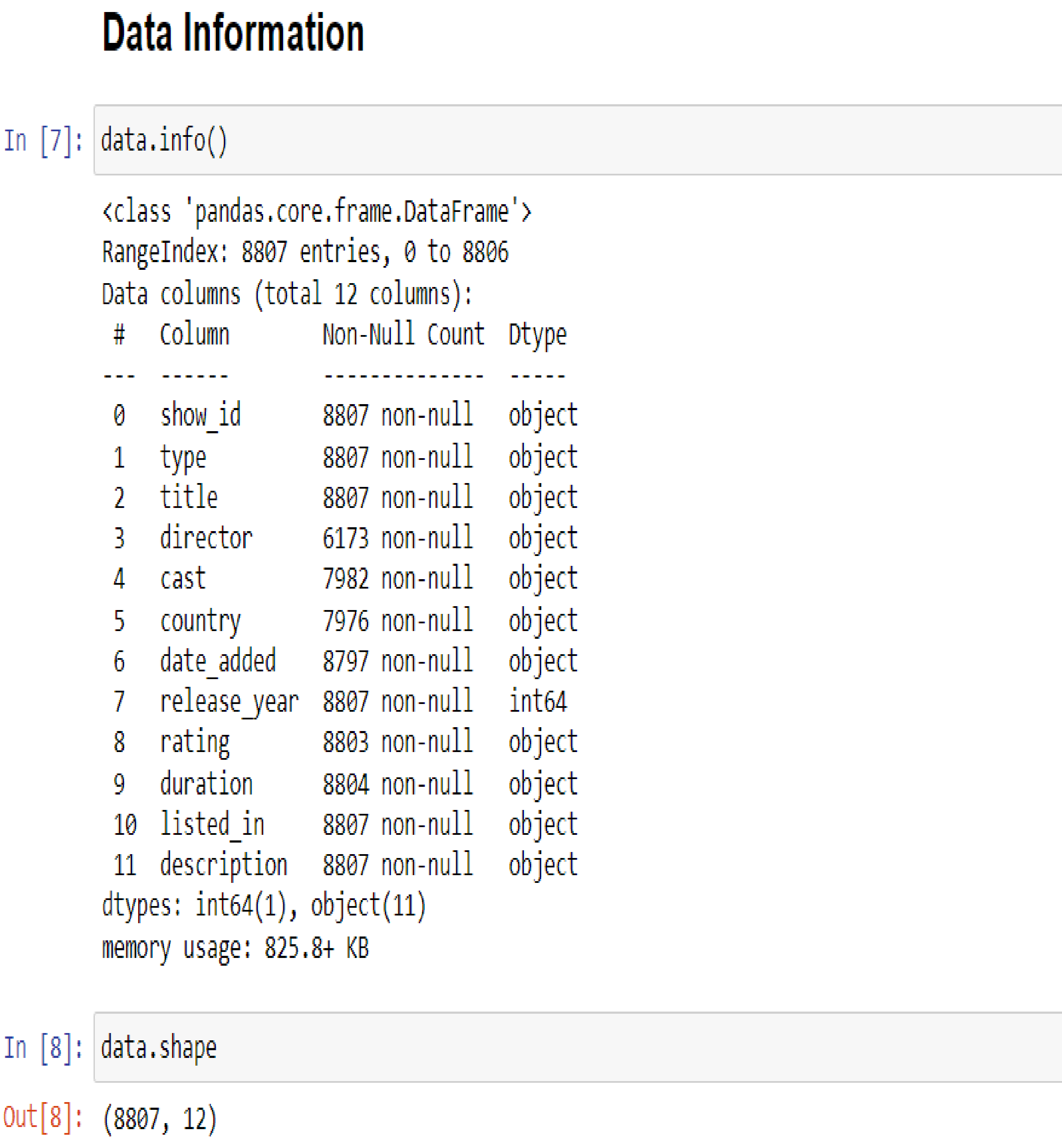
In summary, our comprehensive analysis of Netflix TV shows and movies provides valuable insights into the platform's content landscape, demographics, and temporal patterns. The utilization of advanced visualization tools not only made the findings accessible but also contributed to the overall effectiveness of communicating complex data to a wider audience.

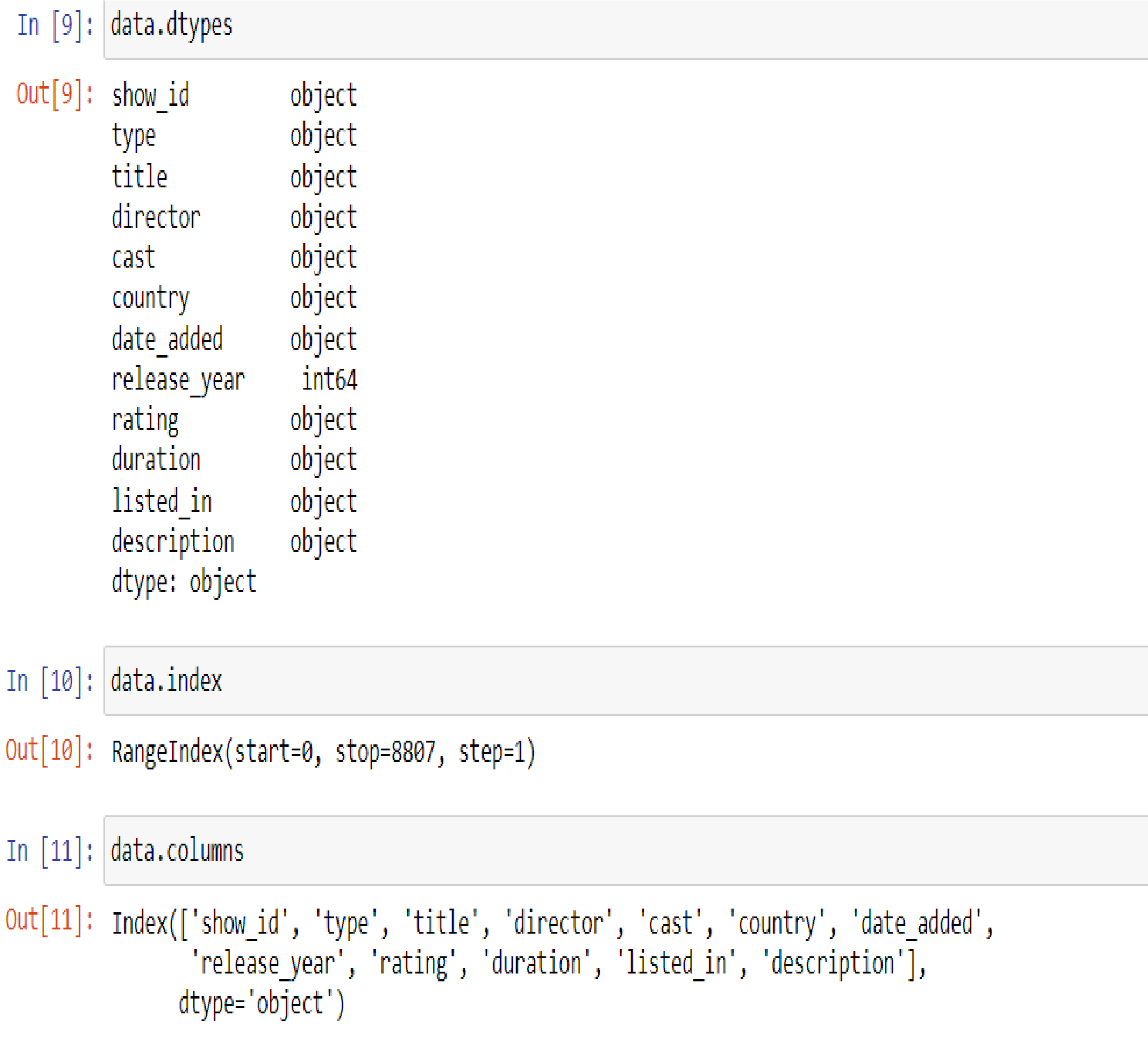
***Pre –***

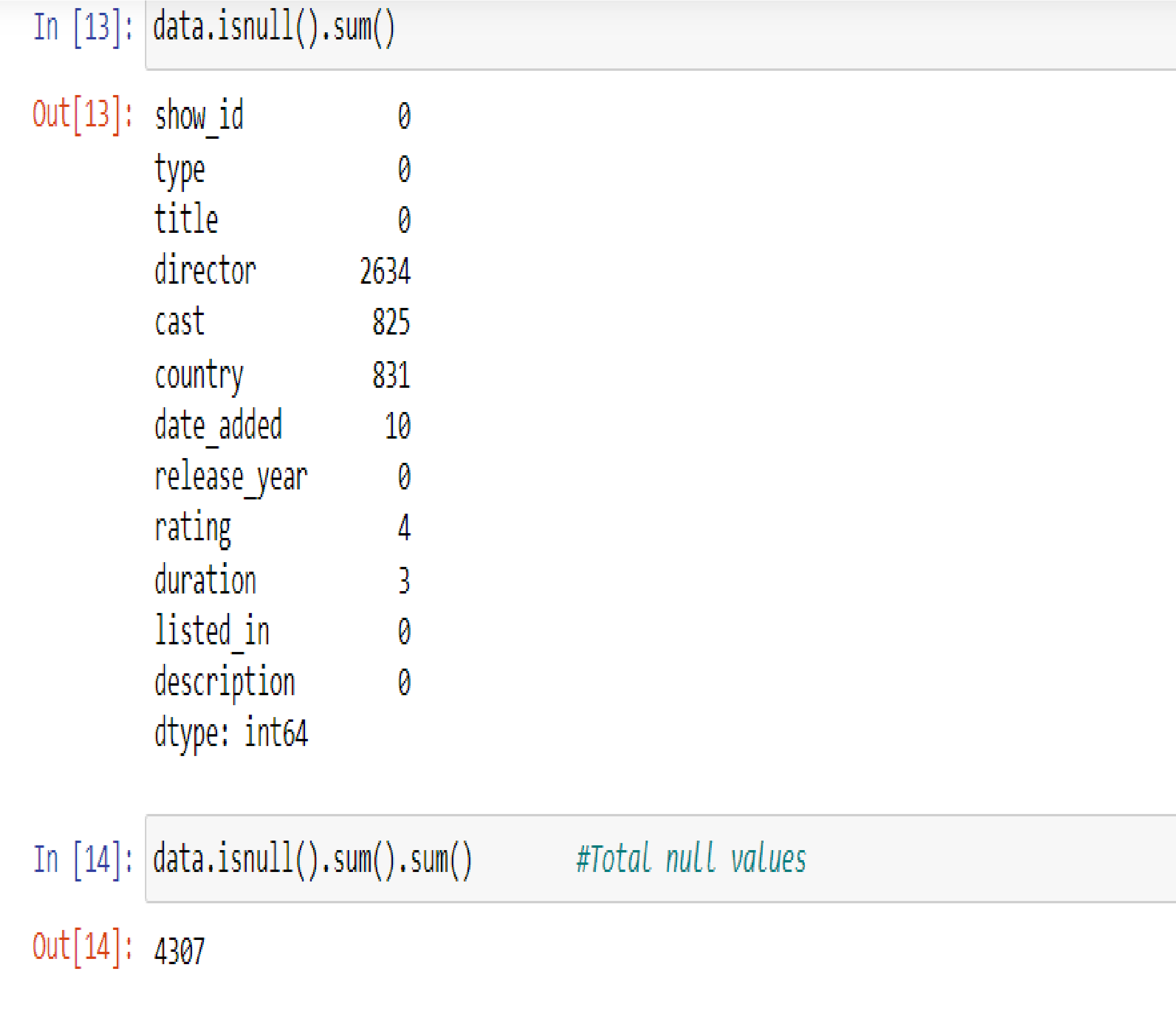
***Processing***

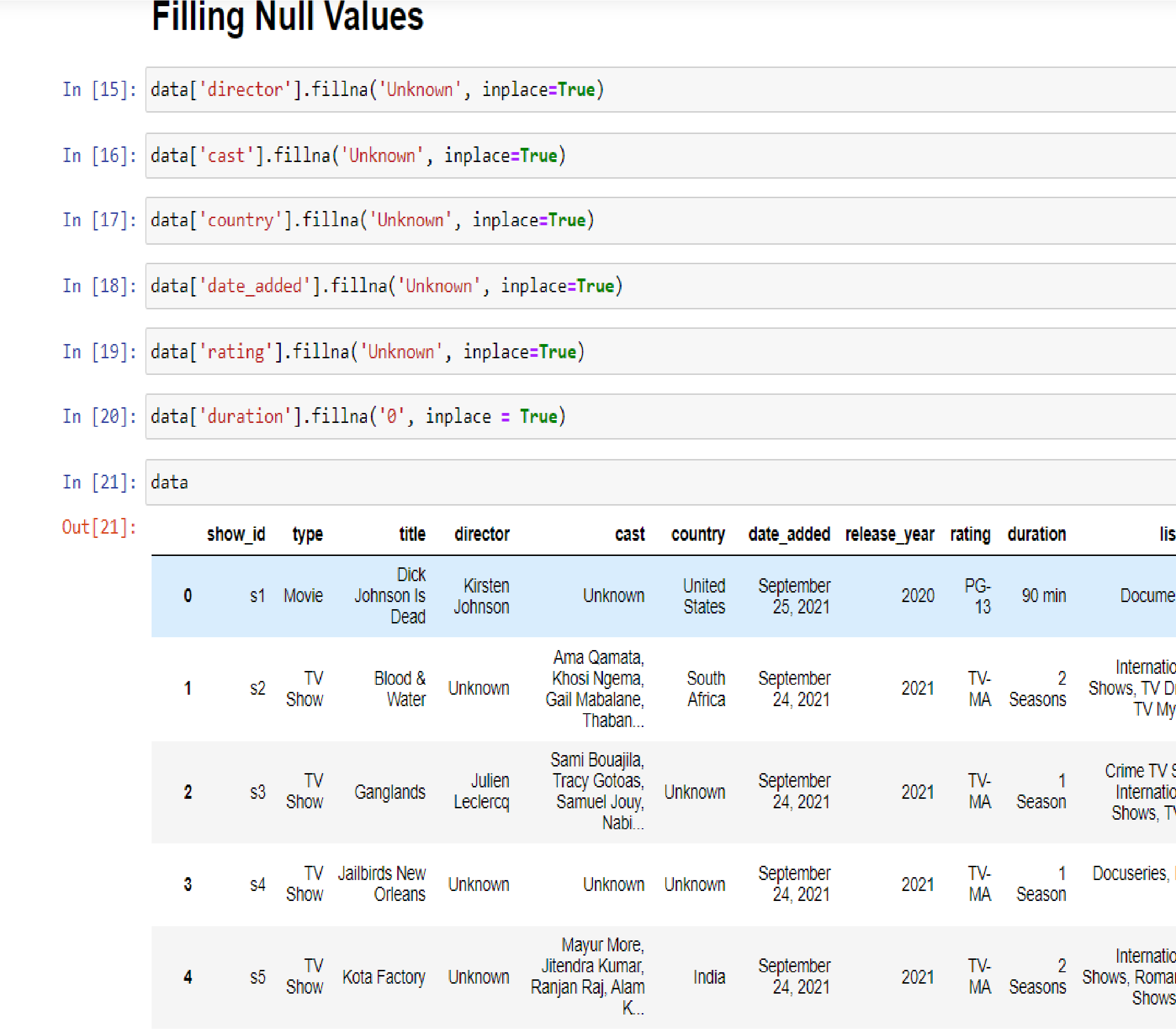
***Steps***

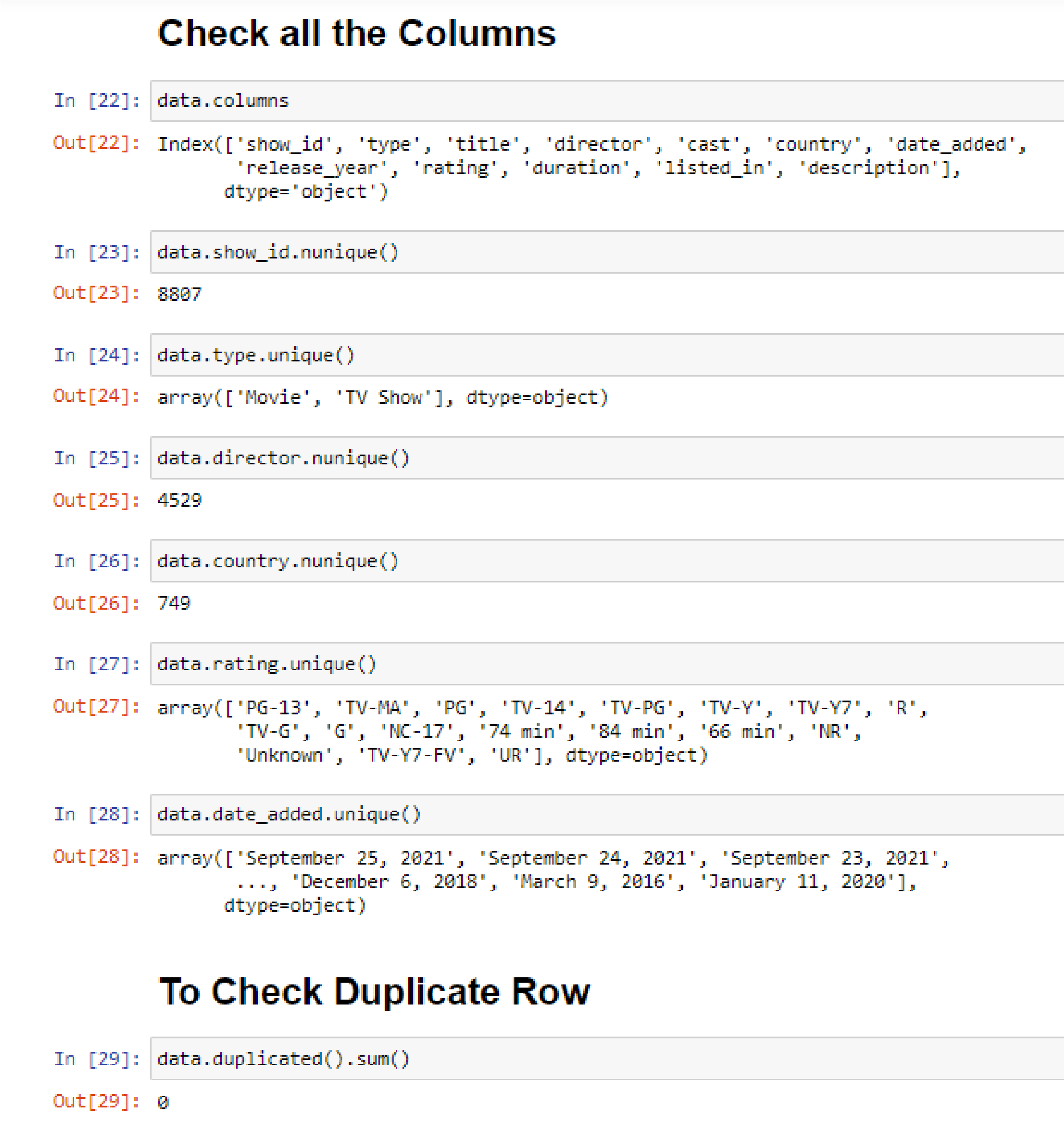




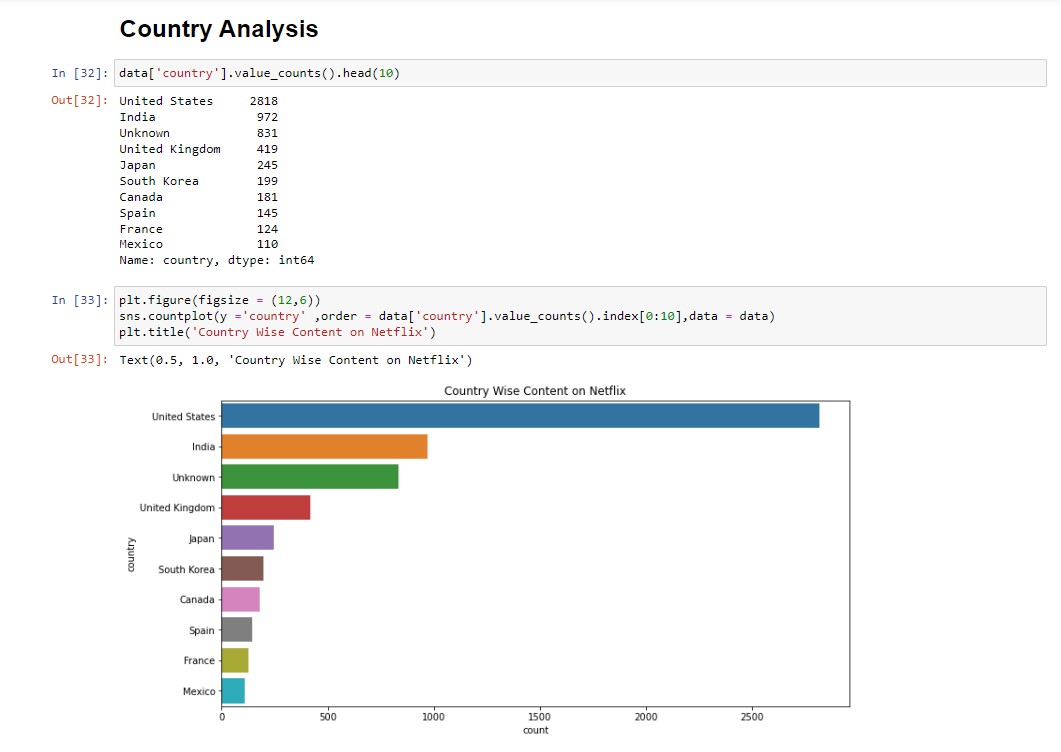


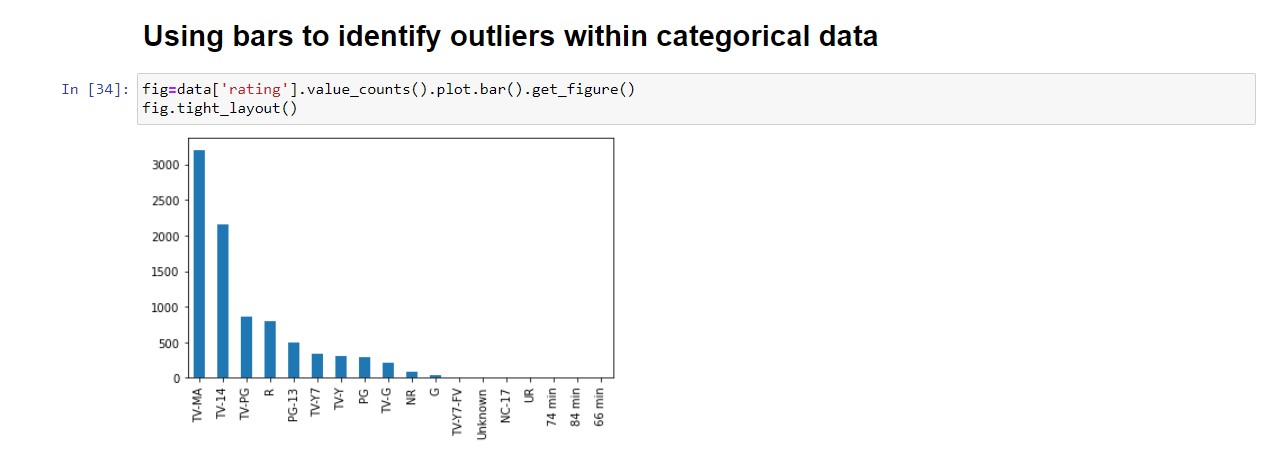


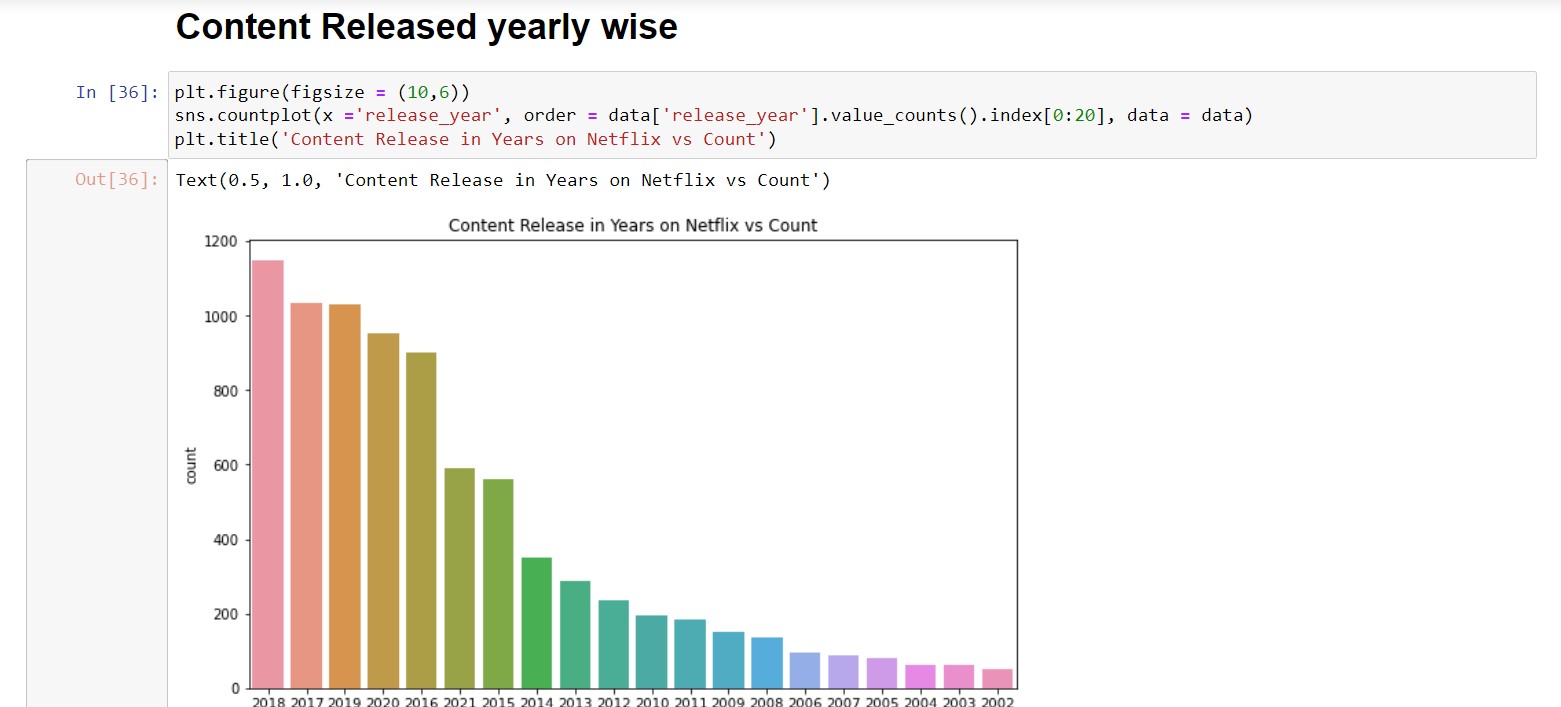


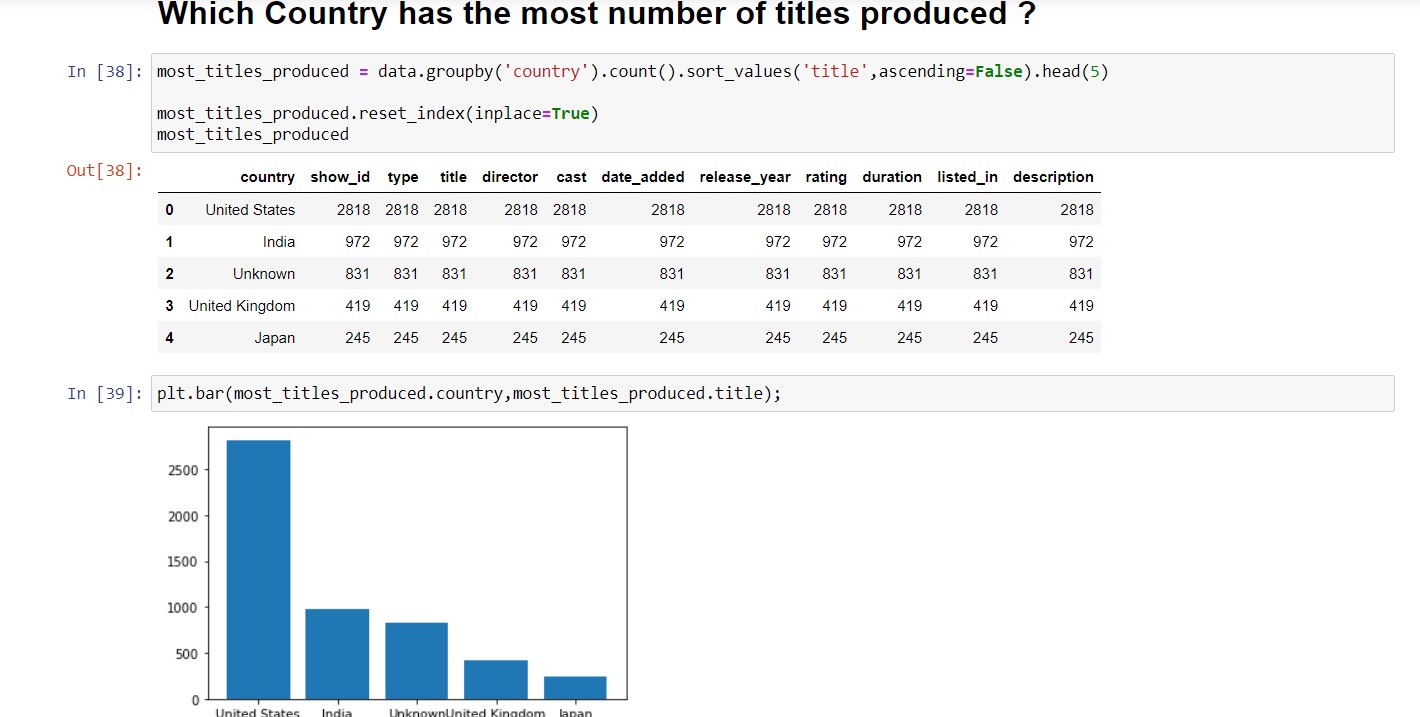


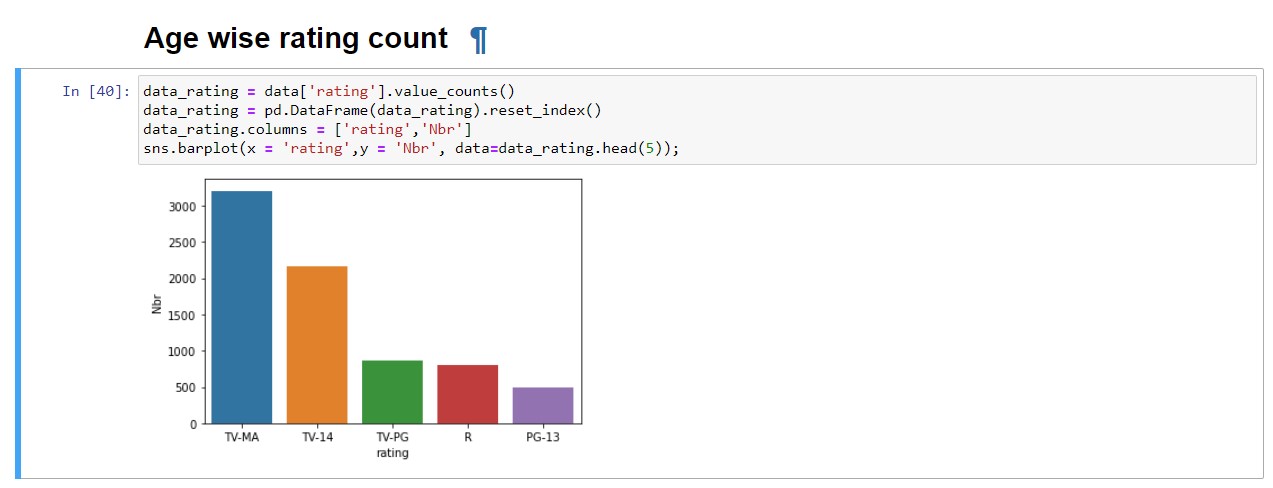


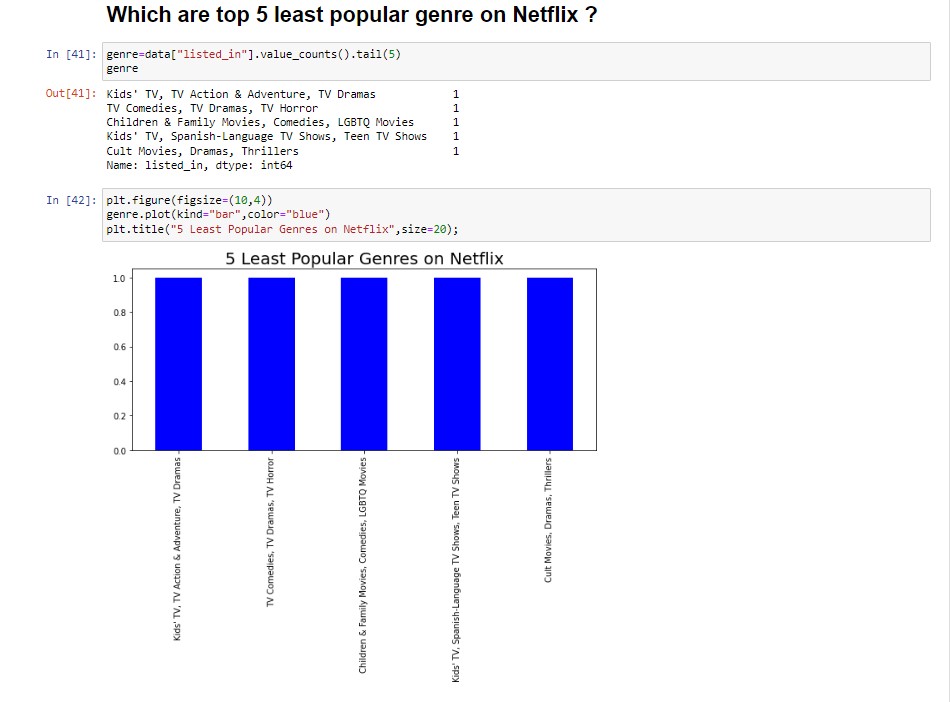


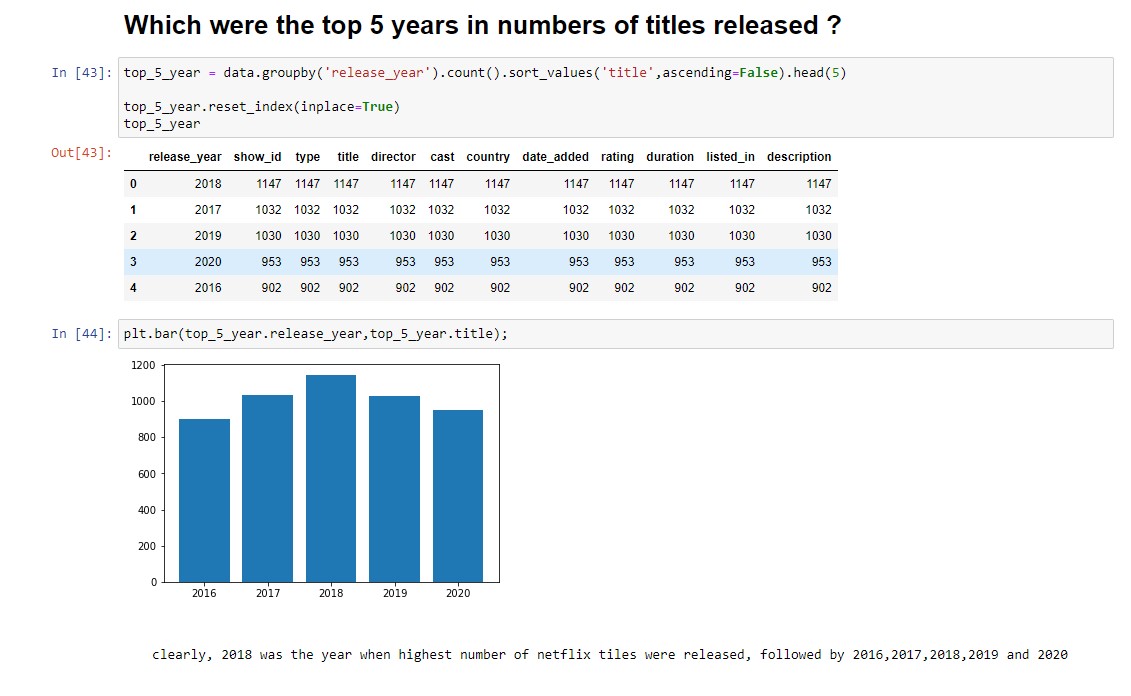


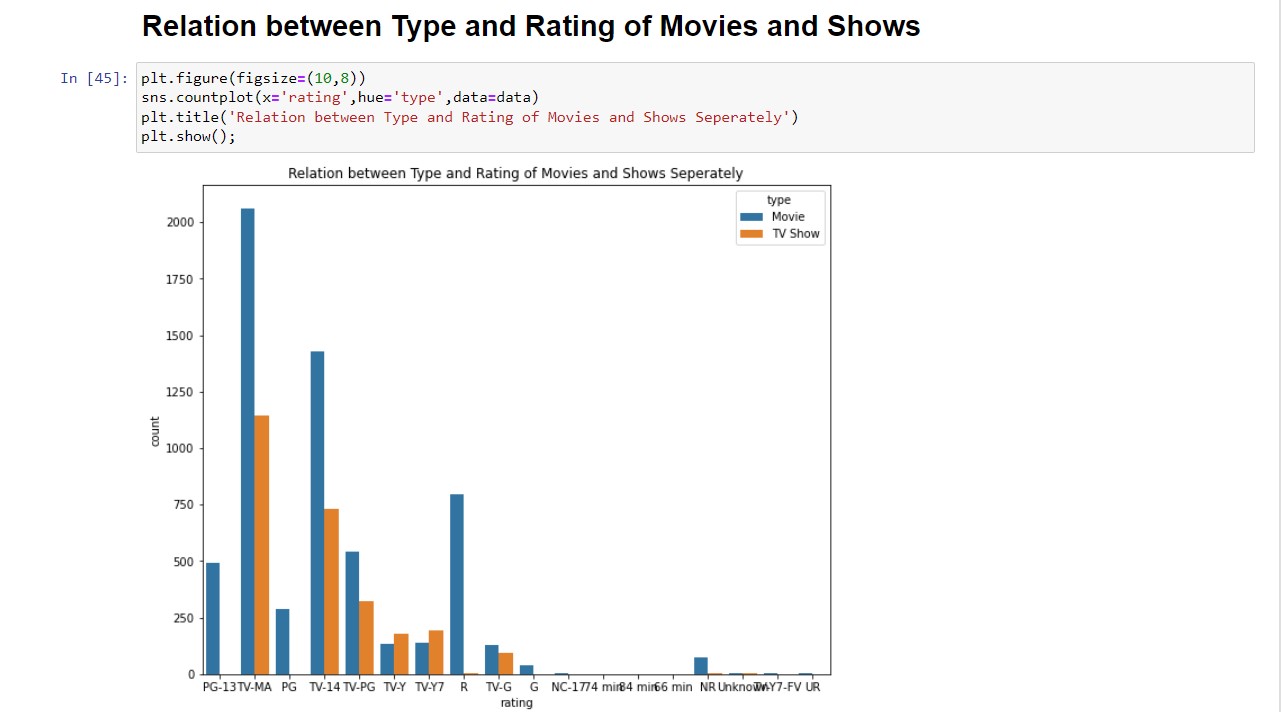












## Conclusion: -

* Finally, our Python-based exploratory data analysis (EDA) of the Netflix database revealed useful insights into the platform's content landscape and user demographics. We discovered important trends and patterns in Netflix's offers through thorough data cleansing and thoughtful visualizations.
* Our investigation began by looking at the distribution of content genres, which revealed a strong preference for movies over TV series on the site. The bar chart, produced with Python's data visualization packages, presented a clear and visually powerful depiction of Netflix's content distribution, with movies accounting for 68.4% of the total library.
* Looking into content ratings, we noticed that the bulk of programs, including movies and TV episodes, are classed as TV-MA. This research highlights the popularity of mature material on the site and offers insights into the target demographic.
* The time component was also an important part of our investigation, as we looked at the release dates of Netflix material. Unveiling temporal patterns allowed us to identify trends and evolutions across time, offering useful information to content creators, analysts, and enthusiasts alike.
* The strategic use of Python's EDA modules, such as Pandas, Matplotlib, and Seaborn, enabled a smooth analysis process. In addition, using the Netflix color palette and the Colors online tool improved the aesthetics and readability of our visualizations, making the complicated dataset more accessible to a larger audience.
* Ultimately, this EDA of the Netflix database not only deepened our understanding of the platform, but also highlighted the power of

Python to extract meaningful information from massive data sets. As the digital streaming landscape continues to evolve, these analytics will be invaluable to stakeholders looking to adapt and thrive in this dynamic industry.

### BIBLIOGRAPHY

1.Han J. and Kamber M. (2003): “*Data Mining, Concepts and Techniques*”, Academic

Press, 2003.

2.Han J., Pei J., and Yin Y. (2000): “*Mining Frequent Patterns without Candidate Generation*”. In proceedings of International Conference on Management of Data (ACM SIGMOD’00), pages 1-12, ACM Press Dallas, TX, United States, May 2000.

3.Hand D., Mannila H. and Smyth P. (2001): “*Principle of Data Mining*”. MIT Press, Cambridge, Massachusetts, USA, 2001.

4.Hipp J., Guntzer U. and Nakhaeizadeh G. (2000): “*Algorithms for Association Rule Mining: A General Survey and Comparison*”. SIGKDD Explorations, Vol. 2, No. 1, pages 58-64, July 2000.