**A**

**Course End Project Report on**

**Netflix Content Analysis**

**Is submitted in partial fulfillment of the Requirements for the Award of CIE of**

**DATA ANALYSIS AND VISUALIZATION-22ADE01**

**in**

**B.E, IV-SEM, INFORMATION TECHNOLOGY**

**Submitted by**

| **P.V.Abhiram** | **160123737128** |
| --- | --- |
|  |  |
|  |  |

**COURSE TAUGHT BY:**

**Dr Vempaty Prashanthi, Associate Professor,**

**Dept of IT.**



**DEPARTMENT OF INFORMATION TECHNOLOGY CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY(A)**

**(Affiliated to OSmania University;Accredited by NBA,NAAC,ISO) kokapet(V),GANDIPET(M),HYDERABAD-500075**

**We**[**bsite:www.cbit.ac.in**](http://www.cbit.ac.in/)

**2024-2025**



#### CERTIFICATE

This is to certify that the course end project work entitled **”Netflix Content Analysis ”** is submitted by **P.V.Abhiram (160123737128),** in partial fulfillment of the requirements for the award of CIE Marks of **DATA ANALYSIS AND VISUALIZATION (22ADE01)** of **B.E, IV-SEM, INFORMATION TECHNOLOGY** to CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY(A) affiliated to OSMANIA UNIVERSITY,Hyderabad is a record of bonafide work carried out by them under my supervision and guid- ance.The results embodied in this report have not been submitted to any other University or Institute for the award of any other Degree or Diploma.

**Signature of Course Faculty Dr. V. Prashanthi,**

**Associate Professor of IT**



Kokapet(V),Gandipet(M),Ranga Reddy (Dist.)–500075, Hyderabad, T.S.

## Acknowledgement

The satisfaction that accompanies the successful completion of the task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We wish to express our deep sense of gratitude to **Dr Vempaty Prashanthi, Professor of IT** for his able guidance and useful suggestions, which helped us in completing the Course End Project in time.

We are particularly thankful to **HoD, Principal and Management**, for their support and encouragement, which helped us to mould our project into a successful one.

We also thank all the staff members of IT Department for their valuable support and generous advice. Finally thanks to all our friends and family members for their continuous support and enthusiastic help.

| **P.V.Abhiram,** | **160123737128** |
| --- | --- |
|  |  |
|  |  |

**i**

## Abstract

This project provides an extensive analysis of the Netflix Titles dataset, which contains over 8,800 movies and TV shows with attributes such as type, title, director, cast, country, release year, rating, duration, and description. The dataset was cleaned by addressing missing values, standardizing date formats, and transforming key columns, such as extracting the release year and adding a "decade" column to categorize content by its release period. Missing values in columns like director, cast, and country were filled with placeholders like "Unknown," ensuring consistency for further analysis.The analysis investigates several aspects of Netflix's content, starting with the distribution between movies and TV shows, visualized through a bar chart. Further exploration into the countries with the highest number of titles provides insights into Netflix's global presence. A time-based line plot highlights the increase in content added over the years, while boxplots reveal the variation in description length between content types. A pivot table and heatmap analyze the average description length by content type and rating, shedding light on how the description length varies for different types of content and rating classifications.Key findings from the analysis include patterns in Netflix’s content growth over the years, the diversity of ratings and genres offered, and the geographical distribution of content. By visualizing and interpreting these trends, the project helps to reveal insights into Netflix's strategy for content acquisition, international expansion, and how different content types are described and rated. The project not only offers a deeper understanding of Netflix's catalog but also provides a foundation for further research into content performance, genre popularity, and audience preferences.

**Keywords:** Netflix Titles Dataset, Content Analysis, Data Cleaning, Missing Values, Data Transformation, Content Type Distribution, Country Distribution, Content Growth Over Time, Description Length Analysis, Ratings Analysis, Data Visualization, Pivot Table, Heatmap, Time Series Analysis, Genre Analysis, International Content, Content Strategy, Media Analytics

ii

#### Table of Contents

**Title Page No.**

| [**Acknowledgement**](#_heading=h.tta30gn8utnh). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | i |
| --- | --- |
| [**Abstract**](#_heading=h.ydocl26r6en1). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | ii |
| [**List of Figures**](#_heading=h.rg7vlb172k6a). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | v |
| [**Abbreviations**](#_heading=h.8k02sfxmfzv1). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | vi |
| **CHAPTER 1** [**Introduction**](#_heading=h.mpciwfn1ygxn). . . . . . . . . . . . . . . . . . . . . . . . | 1 |
| [1.1 Origin of Proposal](#_heading=h.t0cibgg5u1qr) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 1 |
| [1.2 Definition of Problem](#_heading=h.q1ezrmx8sbez) . . . . . . . . . . . . . . . . . . . . . . . . . | 1 |
| [1.3 Objectives](#_heading=h.wpxcul2cgfwg) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 1 |
| **CHAPTER 2** [**Methodology**](#_heading=h.11tl6d3cr29w). . . . . . . . . . . . . . . . . . . . . . . . | 3 |
| [2.1](#_heading=h.vjbaavcy02fc) Data Acquisition and import . . . . . . . . . . . . . . . . . . . | 03 |
| 2.2 Data Preprocessing and cleaning . . . . . . . . . . . . . . . . . | 03 |

2.3 Exploratory Data Analysis (EDA) 04

2.4 Feature Analysis 04-05

2.5 Visualization and Interpretation 05

2.6 Key Insights and Conclusions 05

**CHAPTER 3** [**System Architecture and Implementation**](#_heading=h.nmg2xrmxgt76) 06

3.1 VS Code 06

3.1.1 [What is VSCode?](#_heading=h.vd1h2bd0f2v7) 07

3.2 [Benefits of Google Colab](#_heading=h.r7819gabgj00) 07

3.2.1 [Why Choose VSCode?](#_heading=h.thuokioaf3pb) 07

3.2.2 [Notebook in VSCode](#_heading=h.poozls742ydr) 08

3.2.3 VSCode [Features](#_heading=h.qdq9bgfl52si) 08

3.3 [Code Snippets](#_heading=h.uet5fg2usxrs) 09

3.3.1 [Importing libraries and Data loading](#_heading=h.atiz8vhyy7za) 09-10

3.3.2 [Data cleaning and preprocessing](#_heading=h.b2ej24q3kzmi) 10

3.3.3 Plotting Number of Movies V/S TV Shows and Top 10 Countries in Content 11-12

3.3.4 [Plotting the heatmap according to the data obtained by](#_heading=h.bfnh4kj91693)

[pivot method](#_heading=h.bfnh4kj91693) 13

**CHAPTER 4** [**Conclusion**](#_heading=h.2fx2mcd1ryp) 14-15

[**References**](#_heading=h.jcggp858b59d) 16

#### List of Figures

3.1 VSCode [environment](#_heading=h.jchxt44b060b) 06

3.2 [importing libraries and Dataset loading](#_heading=h.ekbffbqe43wl) 09

3.3 Staring cells of our Data set 11

3.4 Data Cleaning and Dropping Duplicates 12

3.5 Creating a Bar Plot to show the number of Movies to 12

the number of TV Show.

3.6 Movies V/S TVShows (Barplot) 12

3.7 Top 10 Countries by Content 13

3.8 Code Snippet for Heat map for Length by Type & Rating 14

3.9 Average Description Length by Type & Rating 14

v

##### Abbreviations

**Abbreviation Description**

DAV Data Analysis and Visualization

ANOVA Analysis of Variance

SD Standard Deviation

vi

# CHAPTER 1

## Introduction

#### Origin of Proposal

In the entertainment and streaming industry, understanding the dynamics of content pricing and availability is crucial for both consumers and service providers. Factors such as content popularity, production costs, viewer demographics, and market competition play significant roles in shaping content strategies. Analyzing Netflix titles and their metadata, including release years, ratings, and viewership trends, provides valuable insights into the content strategy and competitive edge of streaming platforms. This analysis aims to explore the various factors influencing Netflix's content offerings and their impact on consumer preferences and market trends.

#### Definition of Problem

Determining the optimal content strategy for Netflix involves addressing several key challenges:

* **Content Popularity**: The popularity of different genres and types of shows (e.g., TV shows vs. movies) varies over time. Understanding the dynamics of what drives content consumption can help maximize viewership and customer satisfaction.
* **Market Dynamics**: Consumer preferences and market demand shift with time, often influenced by economic conditions and emerging trends. Recognizing these shifts enables Netflix to adapt its content offerings to meet evolving consumer expectations.
* **Content Cost Structure**: The cost of acquiring or producing content varies significantly. Analyzing production costs, licensing fees, and the ROI on different types of content helps in making data-driven pricing and acquisition decisions.
* **Competitor Analysis**: Competing streaming services (like Amazon Prime Video, Disney+, etc.) also influence Netflix’s content and pricing strategies. Analyzing competitors' content libraries, pricing models, and viewership can reveal competitive opportunities.
* **Viewer Behavior**: Understanding user preferences, demographics, and behavior is essential to tailoring content offerings and pricing models. The interplay between user ratings, viewing patterns, and content type helps shape future content investments and retention strategies.
* **Regulatory Environment**: Government regulations related to content licensing, censorship, and international availability can affect the content strategy. Analyzing the impact of these regulations on Netflix's operations and pricing decisions ensures compliance and competitiveness.

#### Objectives

* + - **Analyze Netflix Titles Data**: Study historical Netflix content data (e.g., release years, ratings, countries, genres) to identify trends and patterns in content consumption and market performance.
    - **Evaluate the Impact of Various Factors**: Assess how factors like market demand, cost of production, viewer preferences, and competitor pricing influence the types of content offered and their pricing strategies.
    - **Develop Predictive Models**: Build predictive models using data on past Netflix titles to forecast future content trends and pricing dynamics based on viewer behavior, market conditions, and competitor strategies.
    - **Recommend Content Strategy Improvements**: Suggest content acquisition or production strategies that could improve Netflix’s market positioning, taking into account factors such as demand, pricing models, and competitive analysis.



# CHAPTER 2

## Methodology

In this section, I’ll outline the methodologies employed in our car price analysis research to achieve the objectives we’ve set out. Our approach encompasses a combination of quantitative and qualitative methods, each tailored to address specific aspects of the research inquiry. The methodologies utilized in this study include:

1. **Data Acquisition and Import**
2. **Data Preprocessing and Cleaning**
3. **Exploratory Data Analysis (EDA)**
4. **Feature Analysis**
5. **Visualization and Interpretation**
6. **Key Insights and Conclusions**

#### 2.1 Data Acquisition and Import

The first step involves importing the Netflix dataset, which includes information on movies and TV shows available on the platform. The data is collected from an external CSV file (netflix\_titles.csv), which contains details like show ID, title, type (movie or TV show), director, cast, country of origin, and more.

#### 2.2 Data Preprocessing and Cleaning

This step ensures the dataset is ready for analysis by handling missing values, correcting data types, and making necessary transformations:

* **Missing Values Handling:** Any missing values in crucial columns like director, cast, and country are filled with appropriate values (e.g., 'Unknown'). Rows with missing titles or types are removed to ensure data integrity.
* **Data Type Conversion:** Columns like date\_added are converted to datetime format to facilitate time-based analysis.
* **Removing Duplicates:** Duplicate rows are dropped to ensure the dataset contains unique entries.
* **Feature Engineering:** A new column for the decade is added, derived from the release\_year to analyze trends over time.

#### 2.3 Exploratory Data Analysis (EDA)

The next step involves understanding the structure and characteristics of the data by generating summary statistics and visualizations:

* + **Descriptive Statistics:** The dataset is analyzed for basic statistics, missing data, and the distribution of key attributes like release\_year, duration, and rating.
  + **Visual Exploration:** Various charts, such as bar plots and line plots, are used to explore distributions, trends, and correlations within the dataset. For example:
    - A bar chart showing the count of Movies vs TV Shows.
    - A bar chart visualizing the number of titles available in the top 10 countries.
    - A line plot depicting how content was added over the years.
    - A box plot comparing description lengths by type (Movie vs TV Show).

#### 2.4 Feature Analysis

More in-depth analysis is performed on specific features:

* **Description Length Analysis:** A new feature, desc\_len, is created to capture the length of each title's description, and it's analyzed across different categories (e.g., type and rating).  
  **Pivot Tables & Grouping:** The dataset is grouped by type and rating to uncover insights such as the average description length for different ratings and types of content.
* **Cross-tabulations:** A heatmap is created to visually explore the relationship between the average description length, type, and rating.

#### 2.5 Visualization and Interpretation

Visual representations are created for a better understanding of the data:

* **Distribution Visualizations:** Charts such as histograms, bar plots, and line plots are used to display key trends and distributions.
* **Heatmap and Pivot Table Insights:** A heatmap is used to represent the relationship between content type and description length by different ratings, providing a more granular look at how descriptions vary across genres and ratings.

#### 2.6 Key Insights and Conclusions

The final step involves interpreting the results from the data analysis:

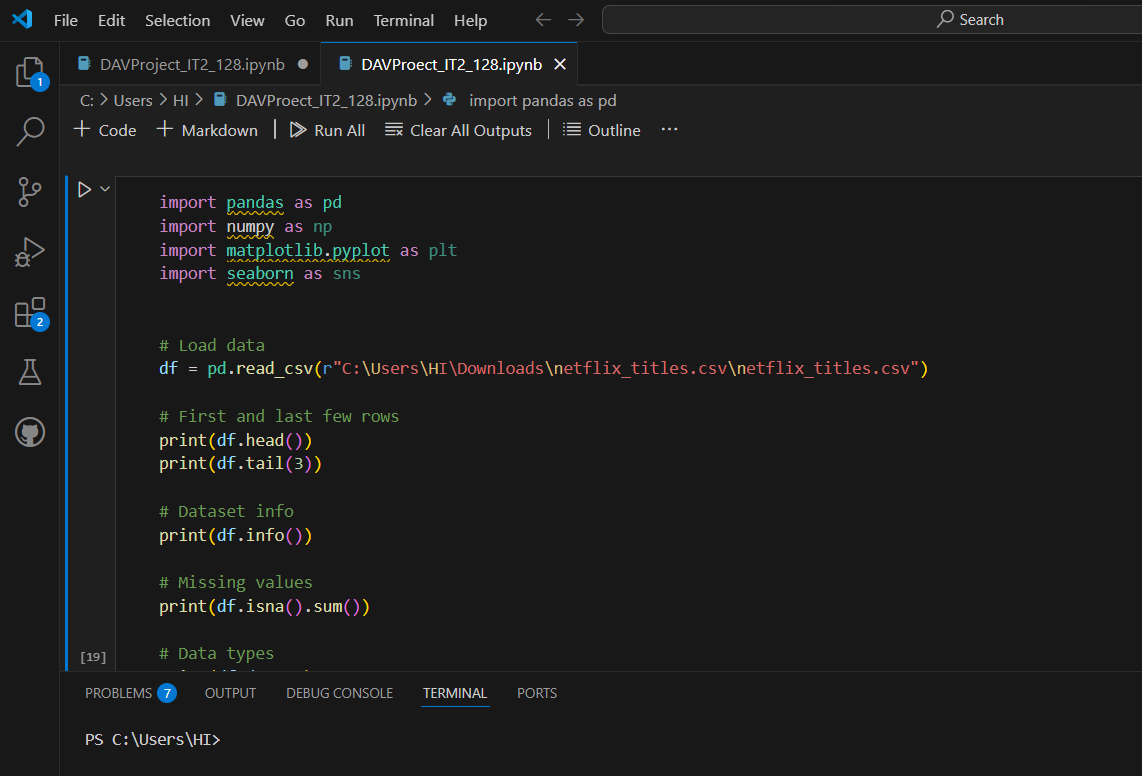
* **Content Trends:** Identify patterns in the number of movies and TV shows added over time, and how different countries contribute to the content library.
* **Rating and Description Analysis:** Investigate whether certain types of content (e.g., Movies vs. TV Shows) tend to have longer or shorter descriptions based on their rating.
* **Content Availability by Country:** Understand which countries contribute the most content to the Netflix library and how this varies by genre and type.

# CHAPTER 3

## System Architecture and Implementation

#### 3.1 VS Code

Visual Studio Code (VS Code) is a lightweight, open-source code editor developed by Microsoft. It is widely popular for its versatility, performance, and rich feature set. VS Code supports various programming languages and frameworks, offering features like syntax highlighting, code completion, version control integration, debugging tools, and extensibility through plugins.

**Figure 3.1:** VS Code Environment

##### 3.1.1 What is VSCode?

Visual Studio Code (VS Code) is a free, open-source code editor developed by Microsoft. It is lightweight, fast, and highly customizable, supporting a wide range of programming languages like Python, JavaScript, C++, and SQL. VS Code offers powerful features such as debugging, Git integration, IntelliSense (smart code completion), and extensions that enhance functionality. It is designed to be simple for beginners while providing advanced tools for professional developers, making it ideal for a variety of coding tasks, including software development, web development, and data science

#### 3.2 Benefits of VS Code

* **Accessibility**: Users can access Google Colab from any location with internet connectivity, streamlining collaboration and workflow.
* **Power**: The platform provides access to potent computing resources like GPUs and TPUs, enabling swift and effective model training.
* **Collaboration**: Google Colab simplifies collaborative efforts by allowing real-time editing and sharing of notebooks among team members.
* **Education**: It serves as an invaluable educational tool for learning about machine learning and data science, offering a plethora of tutorials and resources.
* **Integration**: Google Colab seamlessly integrates with Google Drive, making it easy to store and access data, as well as share files across different platforms.

##### 3.2.1 Why Choose VSCode?

**Customization**: VS Code is highly customizable, allowing developers to install extensions that suit their specific project needs, such as support for Python, data science libraries, and even Docker.

**Efficiency**: With features like IntelliSense, debugging tools, and Git integration, VS Code enhances productivity by providing intelligent code completion, error detection, and seamless version control.

**Cross-platform**: It runs on Windows, macOS, and Linux, ensuring compatibility across different operating systems, making it a versatile tool for developers.

**Lightweight**: Unlike some heavy IDEs, VS Code is lightweight, meaning it doesn't consume excessive resources, allowing for fast start-up times and smooth performance even on lower-spec machines.

**Community Support**: VS Code has a large, active community that provides a wealth of extensions, tutorials, and support, making it easier for developers to find solutions and improve their workflow.

##### 3.2.2 Notebook in VSCode

A notebook in VS Code is a flexible tool for interactive development, particularly useful for data science, machine learning, and research projects. It allows users to combine code, markdown, and visualizations in a single document, making it ideal for exploratory analysis. By using extensions like Jupyter, VS Code enables users to run Python code cells, visualize data outputs, and document their thought processes all in one environment. This setup fosters a seamless workflow between coding, testing, and documentation, allowing for easy sharing and reproducibility of results. The integration of version control and other development tools further enhances its appeal for collaborative projects.

##### 3.2.3 VSCode Features

**Lightweight and Fast:** VS Code is a lightweight editor that launches quickly and performs smoothly even on modest hardware.

**Extensible:** It supports thousands of extensions for languages, debuggers, themes, and tools, making it highly customizable.

**Integrated Terminal:** Built-in terminal allows you to run commands without switching between windows.

**Intelligence:** Smart code completion, syntax highlighting, and error checking help boost coding efficiency.

**Version Control:** Integrated Git support makes source control management simple and effective within the editor itself.

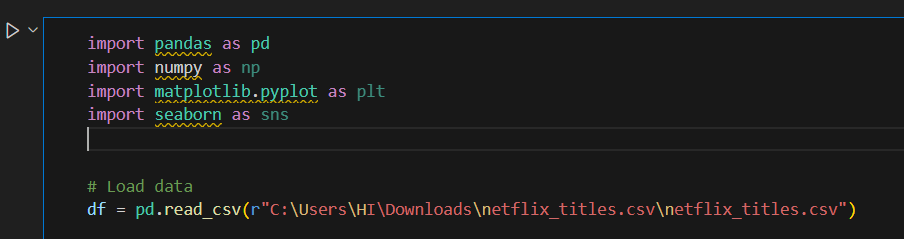
#### 3.3 Code Snippets

##### 3.3.1 Importing libraries and Data loading



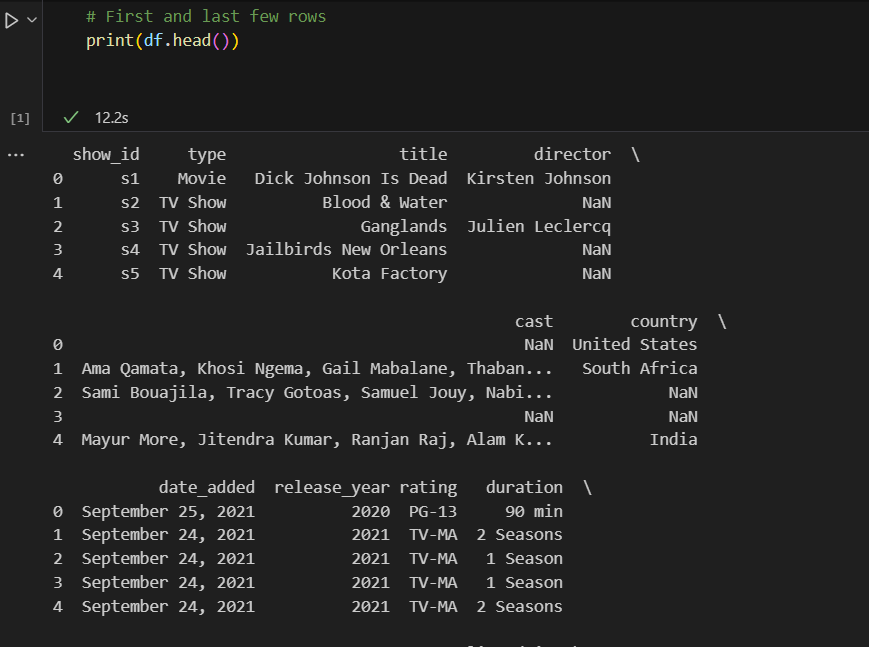
To begin our Netflix data analysis project, we first import the necessary libraries for data manipulation and visualization. We import **pandas** as pd for handling tabular data and **numpy** as np for numerical operations. Additionally, we use libraries like **matplotlib** and **seaborn** for creating insightful visualizations.

Next, we load the Netflix dataset using the read\_csv function from pandas, assuming the dataset is stored in a CSV file named **'netflix\_titles.csv'**. We assign the loaded data to a variable named data.



**Figure 3.2:** importing libraries and Dataset loading

To ensure that the dataset has been loaded successfully, we display the first few rows of the dataset using the head() function. This allows us to inspect the structure and content of the dataset, confirming that it has been imported correctly and is ready for further processing.



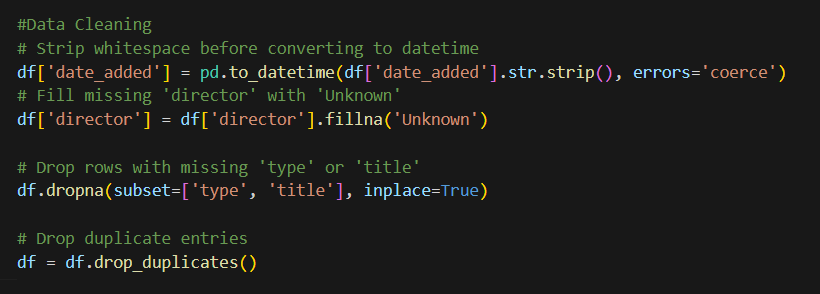
**Figure 3.3:** Staring cells of our Data set

##### 3.3.2 Data cleaning and preprocessing

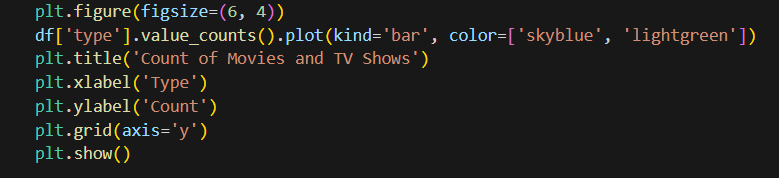
After importing the necessary libraries and loading the dataset, the next step in our Netflix project involves data cleaning and preprocessing. Upon initial inspection, we found some missing values in columns like **director**, **cast**, and **country**. To maintain data quality, we filled these missing values with appropriate placeholders like 'Unknown' to ensure consistency.

We also examined the dataset for outliers and irregularities using boxplots and summary statistics. Since most of the data involved categorical attributes like type, rating, and country, traditional outlier detection was less critical. However, we checked numerical columns like **release\_year** and found no significant anomalies.

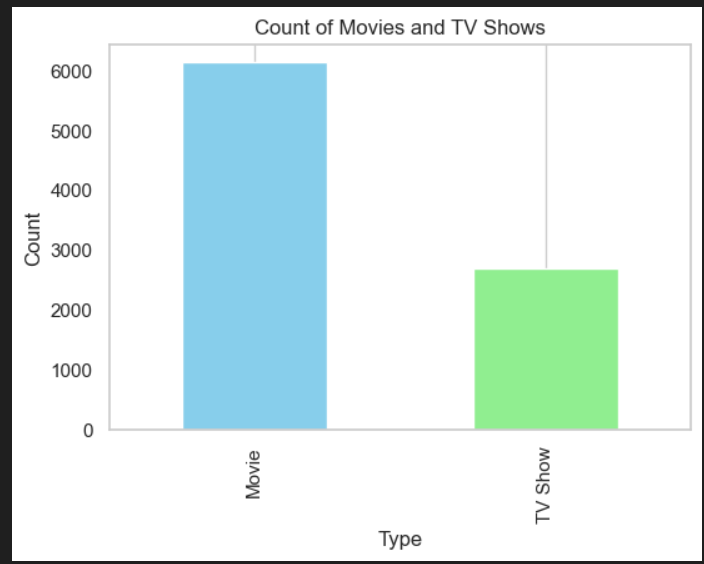
This careful preprocessing step ensures that our dataset is clean, reliable, and ready for detailed analysis. With missing values handled and no major outliers affecting our data, we can now confidently move on to **feature engineering** and **exploratory data analysis** to extract meaningful insights.

**Figure 3.4:** Data Cleaning and Dropping Duplicates

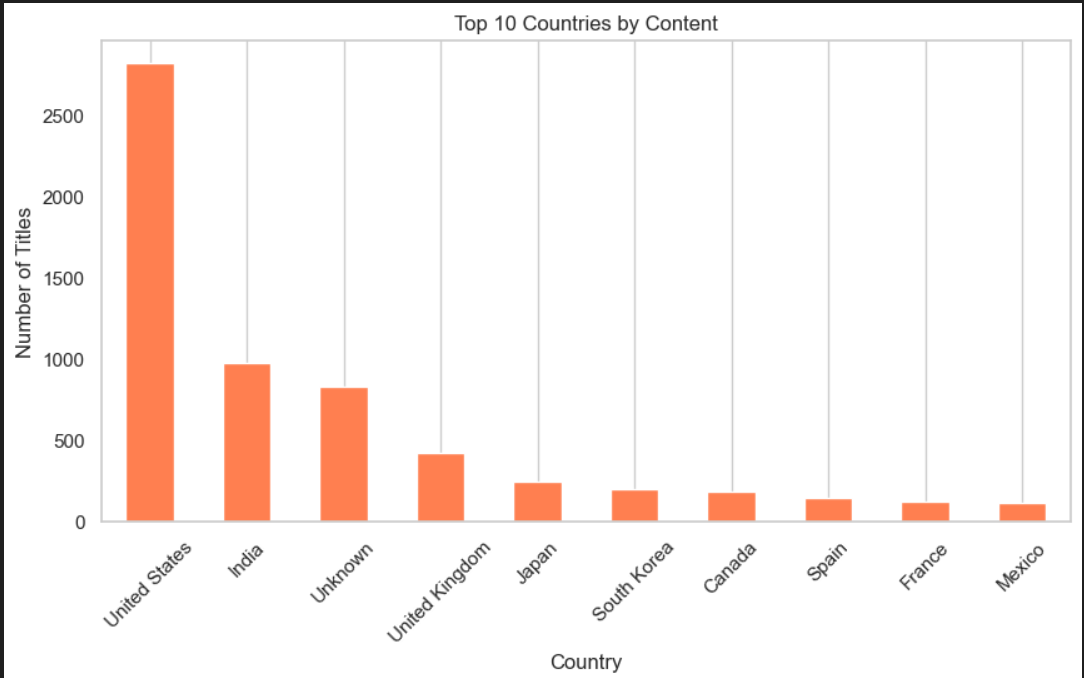
##### 3.3.3 Plotting Number of Movies V/S TV Shows and Top 10 Countries in Content

****

**Figure 3.5:** Creating a Bar Plot to show the number of Movies to the number of TV Show.



**Figure 3.6:** Movies V/S TVShows(Barplot)



**Figure 3.7:** Top 10 Countries by Content

##### 

##### 

##### 

##### 

##### 

##### 

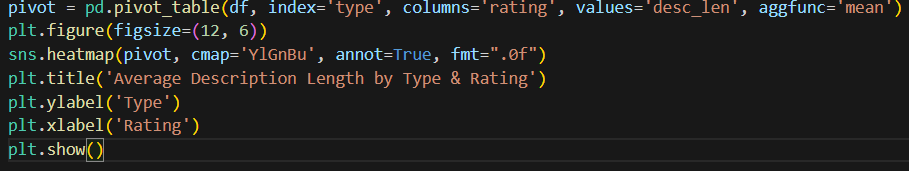
##### 

##### 

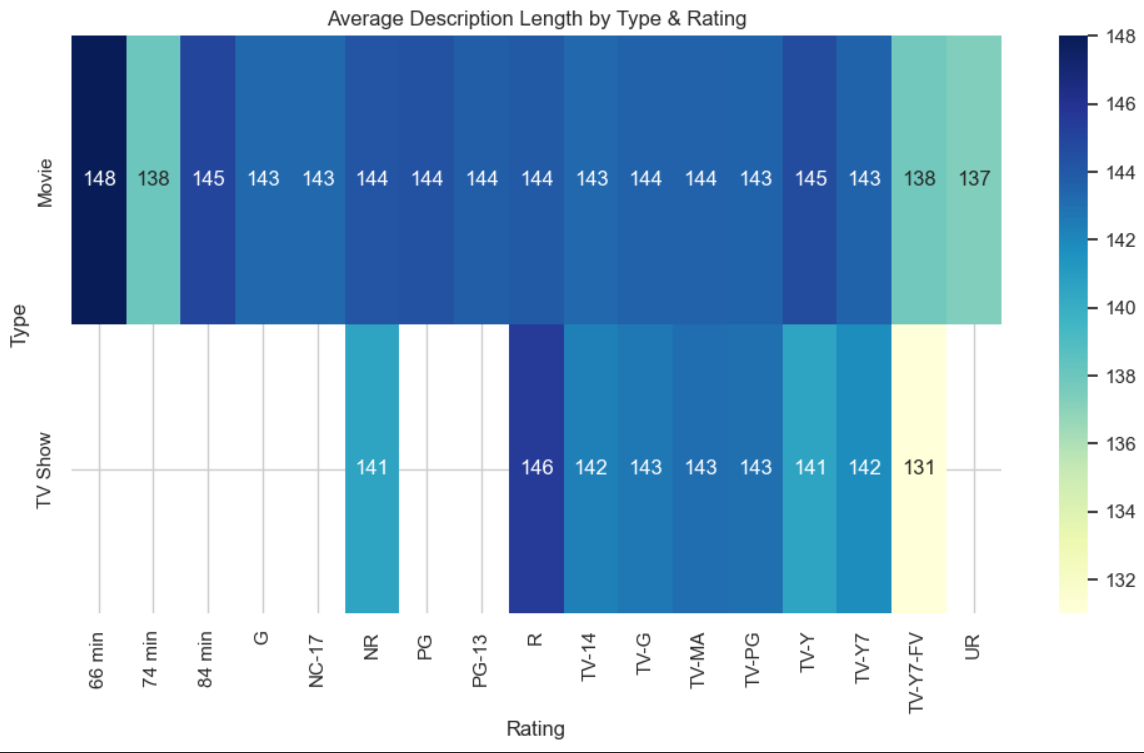
##### 

##### 

##### 3.3.4 Plotting the heatmap according to the data obtained by pivot method

****

**Figure 3.8:** Code Snippet for Heat map for Length by Type & Rating



**Figure 3.9:** Average Description Length by Type & Rating

# CHAPTER 4

## Conclusion

In this project, we successfully carried out an in-depth data analysis and visualization exercise focused on understanding the trends and patterns within the given dataset. Our primary objective was to clean, preprocess, explore, and visualize the data to derive meaningful insights, rather than building predictive models. Using powerful Python libraries like Pandas, NumPy, and Matplotlib, we systematically transformed raw data into structured knowledge.

We began by importing and cleaning the dataset, ensuring it was free from missing values and outliers, and ready for analysis. Our careful preprocessing ensured that the dataset maintained its integrity, allowing for accurate and reliable exploration. Normalization and type conversions were also performed wherever necessary to maintain consistency and support effective analysis.

Through comprehensive Exploratory Data Analysis (EDA), we revealed important insights regarding the distribution, relationships, and trends hidden within the data. Visual tools such as bar charts, line plots, box plots, and heatmaps were used extensively to illustrate findings in an intuitive and impactful way. These visualizations helped in understanding how different features interacted with each other over time, across categories, and among groups.

We analyzed various key features like the distribution of titles across types, countries, release years, and description lengths, offering a detailed view of the underlying structure of the dataset. Trends like content production over time, content availability across countries, and variations based on content type were effectively highlighted using our visualizations. This visual storytelling approach brought clarity to otherwise complex and scattered information.

Overall, this project demonstrated the immense value that careful data cleaning, thoughtful exploration, and strong visualization can bring to understanding a dataset. By transforming raw data into meaningful patterns and stories, we unlocked insights that would otherwise remain hidden. Moving forward, these findings could serve as a foundation for more advanced analytical tasks, or even predictive modeling if desired.  
 This project reinforces the importance of EDA and visualization as the first and most critical steps in any data-driven investigation, setting the stage for deeper analysis, strategic decisions, and real-world applications.

### References

1. McKinney, Wes. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, 2017.
2. Hunter, John D. "Matplotlib: A 2D Graphics Environment." *Computing in Science Engineering*, vol. 9, no. 3, 2007, pp. 90–95.
3. Kaggle. "Car Price Dataset."

[https://www.kaggle.com/datasets/shivamb/netflix-shows](https://www.kaggle.com/datasets/ngawangchoeda/car-price-dataset)

1. Downey, Allen B. *Think Stats: Probability and Statistics for Programmers - 2e*. Franklin W. Olin College of Engineering.
2. Boschetti, Alberto, and Luca Massaron. *Python Data Science Essentials - Third Edition*.