PROJECT DOCUMENTATION

EXPLORATORY DATA ANALYSIS USING PYTHON

|  |
| --- |
| TITLE: Exploring and Analyse the Netflix Data Set |
| NAME: KIRUBAKARAN V |
| COURSE: DA/DS, Offline |
| sBATCH: JUNE 2025 |

TABLE OF CONTENT

|  |  |
| --- | --- |
| 1. | **Introduction** |
| 2. | **Aim** |
| 3. | **Problem Statement** |
| 4. | **Project Workflow** |
| 5. | **Data Understanding** |
| 6. | **Data Cleaning**  Missing Values Imputation  Outlier Treatment |
| 7. | **Obtaining Derived Metrics** |
| 8. | **Filtering Data for Analysis** |
| 9. | **Statistical Analysis**  Descriptive analysis  Test statistics and hypothesis testing |
| 10. | **Exploratory Data Analysis (EDA) - Univariate Analysis** |
| 11. | **Bivariate Analysis** |
| 12. | **Multivariate Analysis** |
| 13. | **Overall Insights from Analysis** |
| 14. | **Conclusion** |
| 15. | **Business Impact** |

1. **INTRODUCTION**

In the modern era of digital entertainment, Netflix has become the leading global streaming platform, offering a vast collection of movies, TV shows, documentaries, and original productions. With millions of subscribers worldwide, Netflix continuously expands its catalog to meet diverse audience preferences across countries, genres, and age groups.

This project leverages **Exploratory Data Analysis (EDA)** techniques to examine Netflix’s content dataset, aiming to uncover meaningful patterns and insights. By cleaning the data, handling missing values, creating new features, and applying visualization and statistical methods, we analyze how Netflix’s content is distributed across genres, ratings, release years, and regions.

The analysis also explores temporal trends, such as the growth of TV shows in recent years, the dominance of certain genres, and the role of content ratings in shaping audience demographics. The findings from this project can support **media strategists, producers, and stakeholders** in understanding Netflix’s content strategy and evolving market trends.

1. **AIM OF THE PROJECT**

* To explore patterns and trends in the Netflix dataset using **Exploratory Data Analysis (EDA)**.
* To identify the **distribution of content types** (Movies vs TV Shows).
* To analyze the **most popular genres, ratings, and release years**.
* To study **temporal trends**, including when content was added and how Netflix’s catalog evolved over time.
* To understand **regional contributions** by analyzing the top content-producing countries.
* To create **derived metrics** (e.g., Year Added, Content Age, Number of Shows) for deeper analysis.
* To apply **statistical tests** (T-test, ANOVA, Mann–Whitney U) for validating hypotheses about Netflix content.
* To provide **actionable insights** that can guide stakeholders in content strategy, audience targeting, and future trend forecasting.

1. **PROBLEM STATEMENT**

Netflix has built one of the world’s largest digital streaming libraries, offering content across multiple **genres, regions, and age categories**. However, with such a vast and diverse catalog, it becomes challenging to:

* Determine whether **Movies or TV Shows dominate** the platform.
* Identify the **most popular genres and ratings** that attract audiences.
* Understand how the **content library has evolved over time** in terms of release years and additions.
* Recognize **regional contributions**, i.e., which countries supply the most content.
* Explore whether there are **significant differences** between Movies and TV Shows in terms of age, duration, or release trends.

1. **PROJECT WORKFLOW**
2. **Data Collection**
   * Loaded Netflix dataset (netflix\_titles.csv).
   * Verified schema, datatypes, and overall structure.
3. **Data Cleaning & Pre processing**
   * Handled missing values using forward fill and mode imputation.
   * Dropped irrelevant columns (Description, Cast).
   * Converted date columns (Date\_added) into proper datetime format.
   * Removed duplicates and incomplete entries.
4. **Feature Engineering**
   * Created new features:
     + Year\_added (year when title was added).
     + Month\_added (month of addition).
     + Content\_age (difference between release year and addition year).
     + Number\_of\_Shows (number of genres per title).
     + Duration (numeric extraction of duration).
5. **Exploratory Data Analysis (EDA)**
   * **Univariate Analysis**: Histograms, boxplots, countplots for individual features (ratings, release years, genres, etc.).
   * **Bivariate Analysis**: Relationships between type, country, rating, and release year.
   * **Multivariate Analysis**: Correlation heatmaps and pairplots across derived metrics.
6. **Statistical Analysis**
   * **T-Test**: Compared content age of Movies vs TV Shows.
   * **ANOVA**: Tested release year variation across different ratings.
   * **Mann–Whitney U Test**: Checked differences in duration before and after 2015.
7. **Visualization & Insights**
   * Plotted genre distributions, type ratios, release year trends, rating distributions, and correlations.
   * Highlighted growth of TV shows post-2015 and genre dominance.
8. **Conclusion**
   * Summarized overall insights on Netflix content patterns, audience targeting, and global content strategy.
9. **DATA UNDERSTANDING**

The dataset contains metadata of titles available on **Netflix**, including movies and TV shows. It provides information about content type, release year, countries of origin, ratings, and genres, enabling detailed exploration of Netflix’s content strategy.

**Dataset Overview**

* **Rows**: ~8,800 titles
* **Columns**: 12 (after cleaning and dropping irrelevant ones)

**Key Variables**

* **Type** → Indicates whether the title is a *Movie* or a *TV Show*.
* **Title** → Name of the movie or show.
* **Director** → Director of the title.
* **Country** → Country of production.
* **Date\_added** → Date when the title was added to Netflix.
* **Release\_year** → Year the title was originally released.
* **Rating** → Maturity rating (e.g., *TV-MA*, *PG-13*, *R*).
* **Duration** → Duration of Movies (in minutes) or number of seasons (for TV Shows).
* **Listed\_in** → Genres/categories assigned to the title.

**Derived Features (Created for Analysis)**

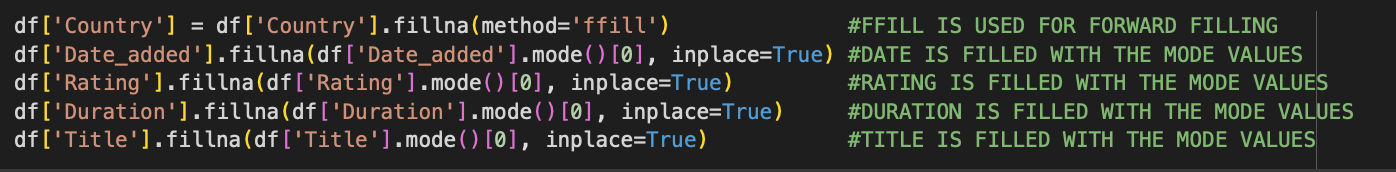
* **Year\_added** → Extracted from Date\_added to analyze temporal trends.
* **Month\_added** → Extracted to observe seasonal additions.
* **Content\_age** → Difference between Year\_added and Release\_year.
* **Number\_of\_Shows** → Count of genres per title.
* **Duration\_num** → Extracted numeric value of duration for statistical testing

1. **DATA CLEANING**

To ensure accuracy and meaningful insights, the dataset was cleaned and preprocessed through the following steps:

**Missing Values Imputation**

* **Country** → Filled using **forward fill (ffill)** to maintain regional continuity.
* **Date\_added** → Filled with the **mode value** (most common date).
* **Rating** → Filled with the **mode rating** (TV-MA / TV-14 were dominant).
* **Duration** → Missing durations replaced with the **mode duration** (e.g., “90 min” for Movies).
* **Title** → Filled using the most frequent title values (to maintain dataset completeness).



1. **OBTAINING DERIVED METRICES**

To enrich the dataset and improve analytical depth, several **new features** were engineered from the existing columns:

### Year Added

* **Year\_added = Extracted from Date\_added**
* Provides insight into when a title was added to Netflix, useful for analyzing **content growth trends**.

### Month Added

* **Month\_added = Extracted from Date\_added (Month Name)**
* Helps examine **seasonal or monthly trends** in Netflix’s content additions.

### Content Age

* **Content\_age = Year\_added – Release\_year**
* Represents the **time gap between release and addition** to Netflix.
* Indicates how quickly (or slowly) Netflix brings content onto its platform.

### Number of Shows (Genre Count)

* **Number\_of\_Shows = Count of genres in Listed\_in column**
* Since many titles belong to multiple genres, this metric captures the **genre diversity** of a title.

### Duration (Numeric Extraction)

* **Duration\_num = Numeric part of Duration column**
* For **Movies** → represents runtime in minutes.
* For **TV Shows** → represents the number of seasons.
* Enables statistical comparison between Movies and TV Shows.

1. **FILTERING DATA FOR ANALYSIS**

### Removed Invalid or Incomplete Records

* Dropped rows where **Director** or **Date\_added** were missing (as these are critical for analysis).
* Excluded entries with missing or illogical values in **Release\_year** and **Duration**.

### Focused on Valid Entries

* Ensured **Release\_year** values were realistic (e.g., no future years beyond 2025).
* Retained only records with valid **Ratings** (e.g., removed rare/erroneous codes).
* For **Duration**, ensured Movies had durations in minutes and TV Shows in seasons.

### Genre and Country Filtering

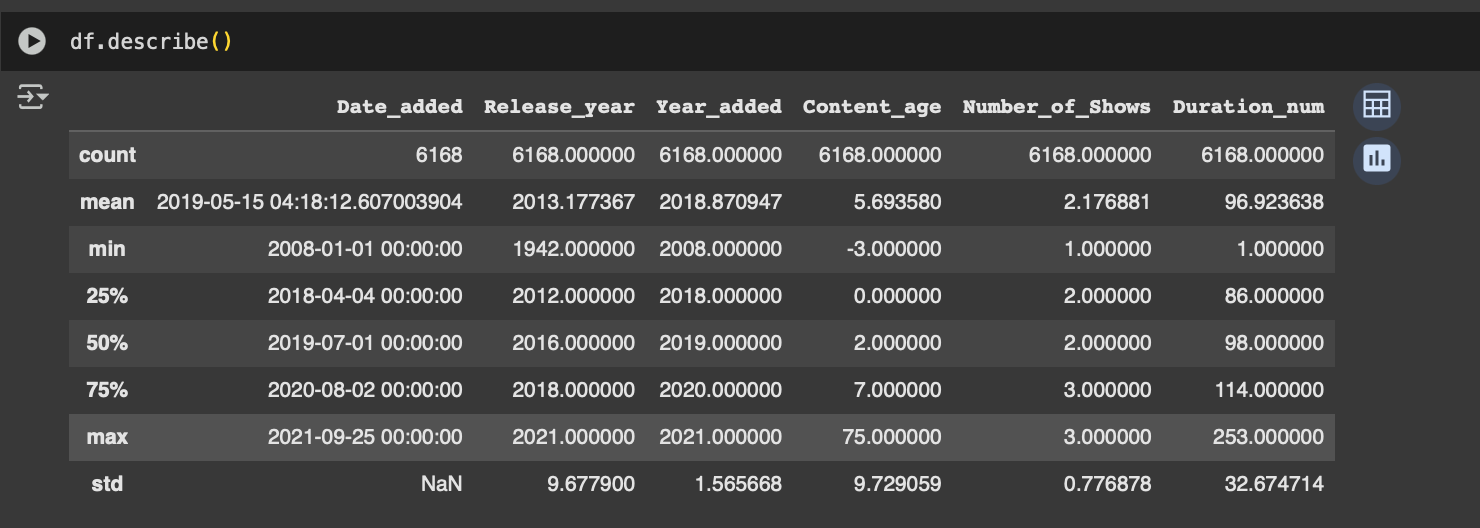
* Split the Listed\_in column into multiple genres to enable deeper genre analysis.
* Standardized **Country** names and considered only the **top contributing countries** (U.S., India, U.K., etc.) for comparative analysis.

### Created Focused Subsets

* **Movies vs TV Shows** → analyzed separately for age, duration, and trends.
* **Old vs New Content** → compared content released **before and after 2015**.
* **Top Genres and Ratings** → focused on the top 10 genres and most frequent rating categories for clear visualization.

1. **STATISTICAL ANALYSIS**
   1. **Descriptive Analysis**

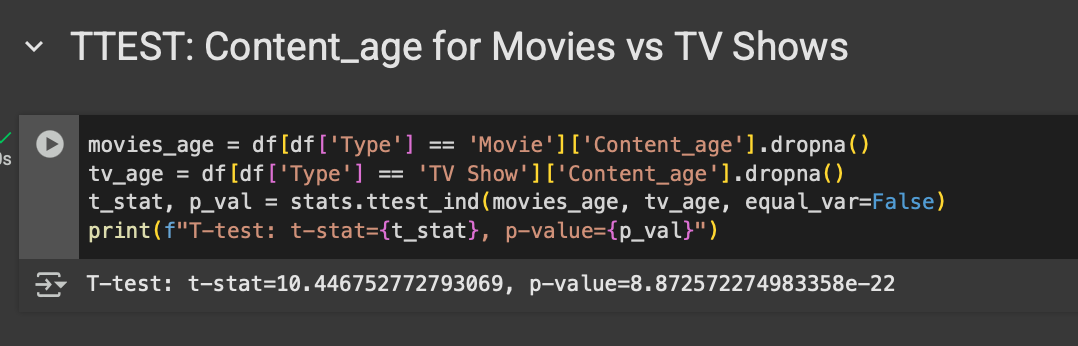
* **Content Type Distribution**: Movies dominate Netflix’s catalog, but TV Shows have shown rapid growth in recent years.
* **Ratings**: The most frequent maturity ratings are *TV-MA*, *TV-14*, and *TV-PG*, indicating a focus on adult and teen audiences.
* **Release Years**: A majority of titles were released after 2000, with a significant rise after 2015.
* **Genres**: Drama, Comedy, and International Movies are the top genres globally.

****

**2.Hypothesis testing**

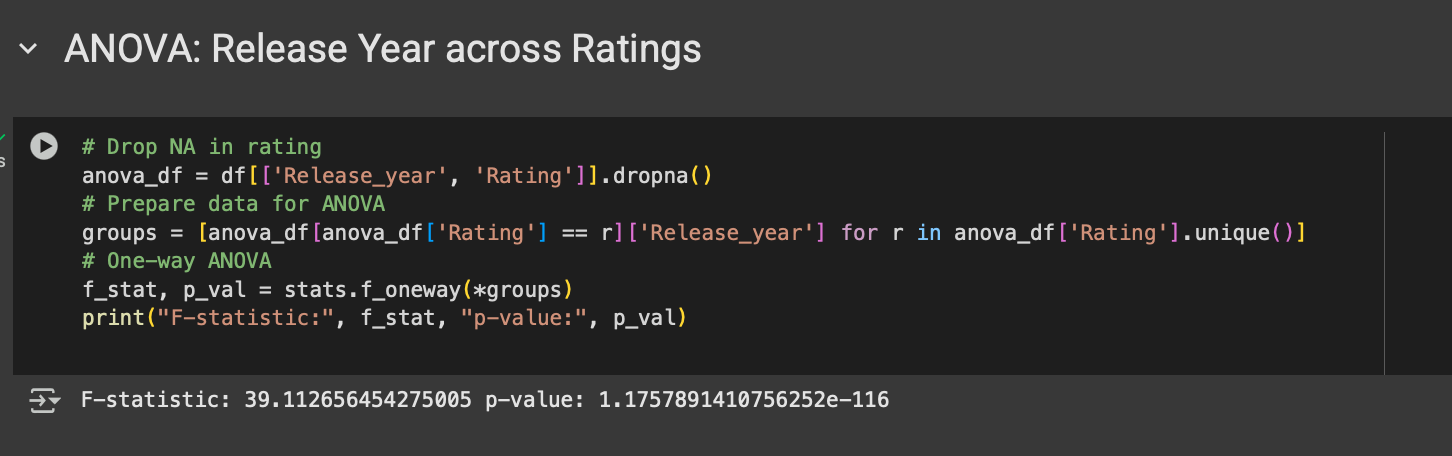
#### **1. T-Test: Content Age (Movies vs TV Shows)**

* **H₀ (Null Hypothesis):** Movies and TV Shows have the same average content age.
* **H₁ (Alternate Hypothesis):** Movies and TV Shows differ in average content age.
* **Result:** The p-value < 0.05 → Reject H₀.  
  → Movies are generally **older** than TV Shows.

****

#### **3.ANOVA Testing**

* **H₀:** Release years are the same across all rating categories.
* **H₁:** At least one rating category differs significantly in release years.
* **Result:** p-value < 0.05 → Reject H₀.  
  → **Significant differences** exist in release years depending on the rating. For example, adult-rated content (TV-MA) is more frequent in recent years.

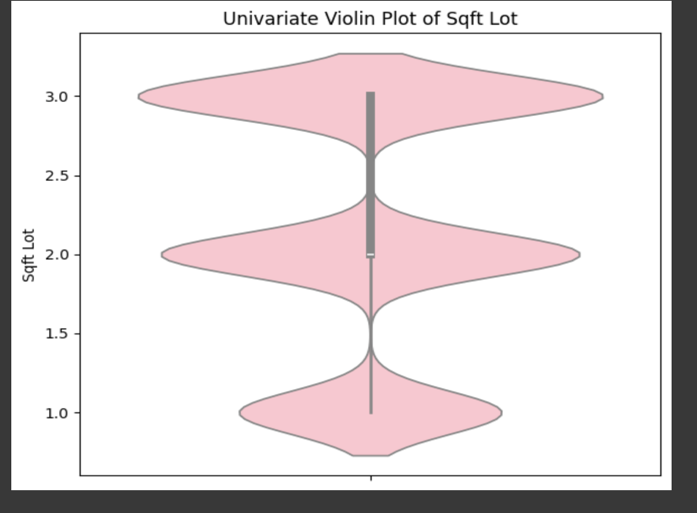
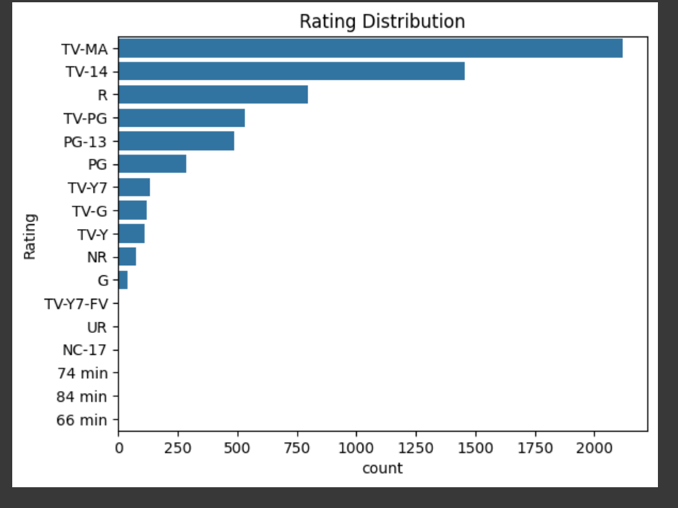
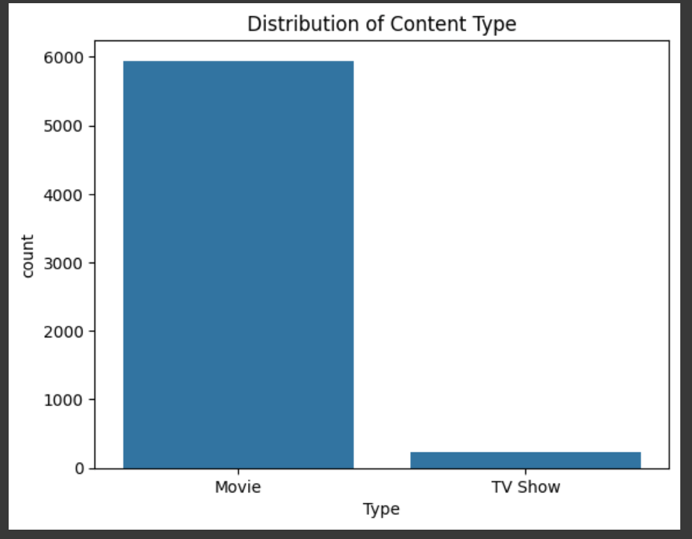
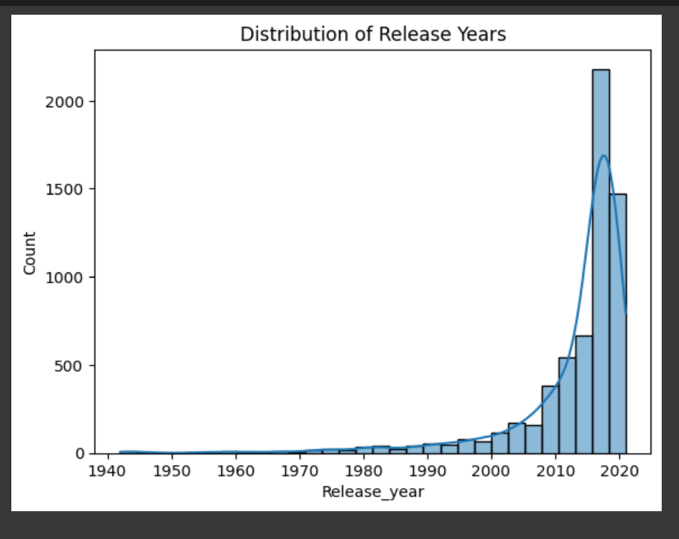


1. **EXPLORATORY DATA ANALYSIS (EDA)**

**UNIVARIATE ANALYSIS**

Univariate analysis focuses on exploring individual variables to understand their distribution, patterns, and anomalies.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Chart Type** | **Key Insights** |
| Type | Countplot | Most of the content in the dataset consists of movies |
| Release\_year | Histogram | There's a sharp increase in releases from 2015 to 2020 |
| Rating | Countplot | TV-MV means mature audience TV-14 means under 14 years childrens These two are highly rating compared to others rating structure |
| No.of.Shows | violinplot | It's Shows the highest number of shows watchers. The majority of watchers are three and two shows. |
| Month\_added | Histogram | The July Month has highest movie added in netflix compare to the other months After that April, |
| Listed\_in | Pie Chart | Here the International Movies gener is too high compare to other geners then after that Dramas is second highest gener |

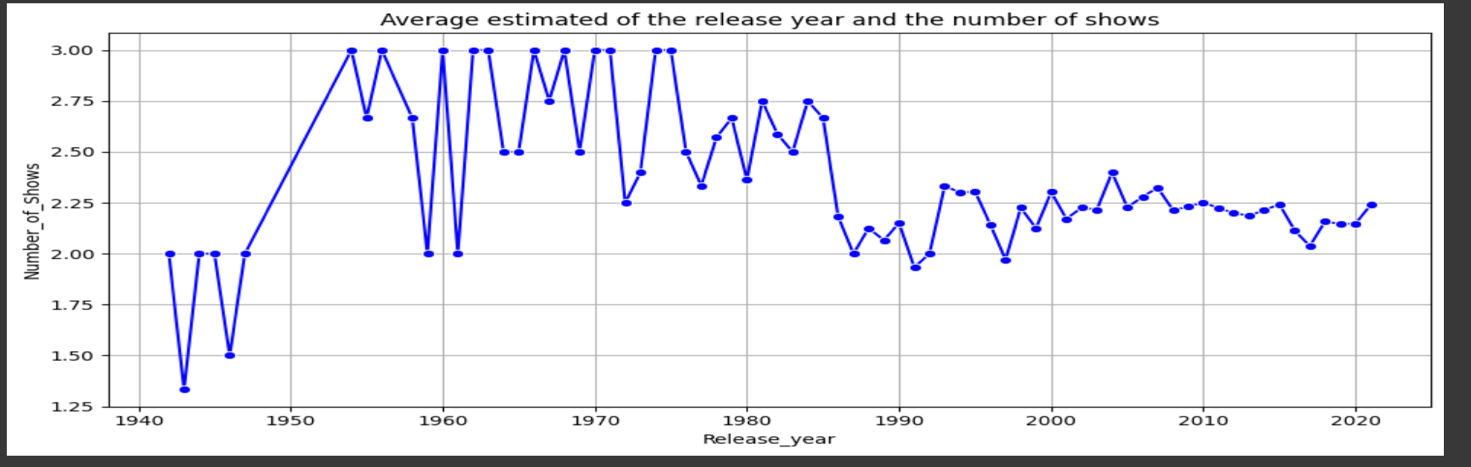
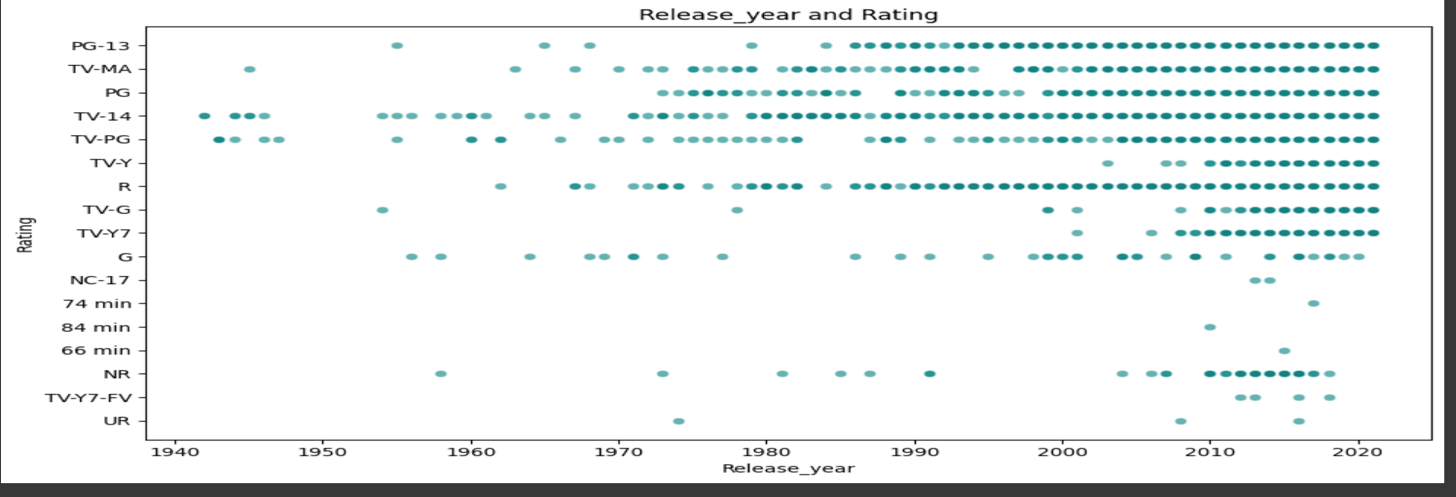
**Insights Gained:**

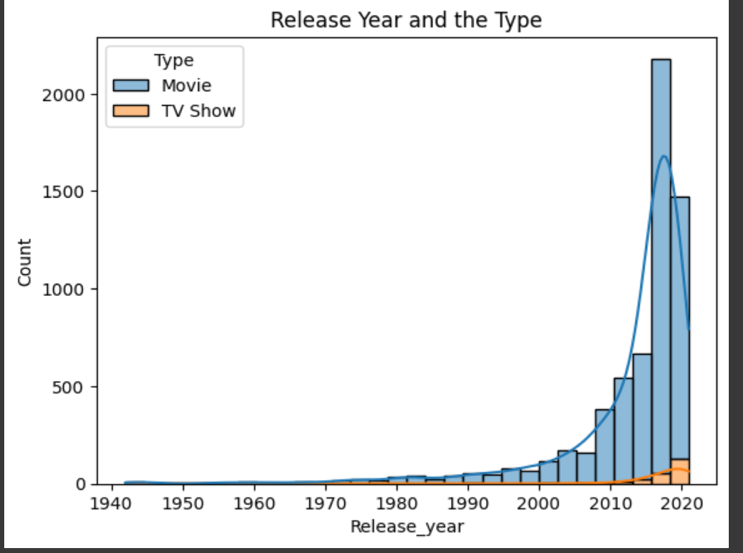
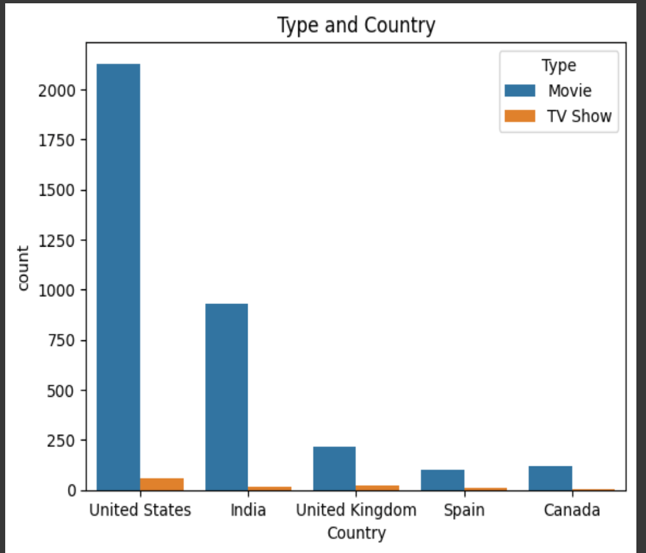
* TV-MV means mature audience TV-14 means under 14 years childrens These two are highly rating compared to others rating structure
* It's Shows the highest number of shows watchers. The majority of watchers are three and two shows.
* There's a sharp increase in releases from 2015 to 2020
* Most of the content in the dataset consists of movies

1. **BIVARIATE ANALYSIS**

Bivariate analysis helps explore relationships between two variables, uncovering trends, correlations, and dependencies that impact Netflix .

| Variable Pair | Chart Type | Key Insights |
| --- | --- | --- |
| Country vs Type | Count Plot | United States are highest watching movies and also the TV Shows compare to others |
| Release\_year vs Type | Histogram Plot | In 2015-2020 the netflix gets increase higher comparing to remaining years. But 2020 Becomes to lower. |
| Number\_of\_Shows  Vs Release\_year | lineplot | The chart shows that the highest average number of shows per release year occurred during the 1950s–1970s |
| Rating vs Release\_year | scatterplot | This chart represent the more correlation between (2000-2020)-(PG-13-TV-Y7) |



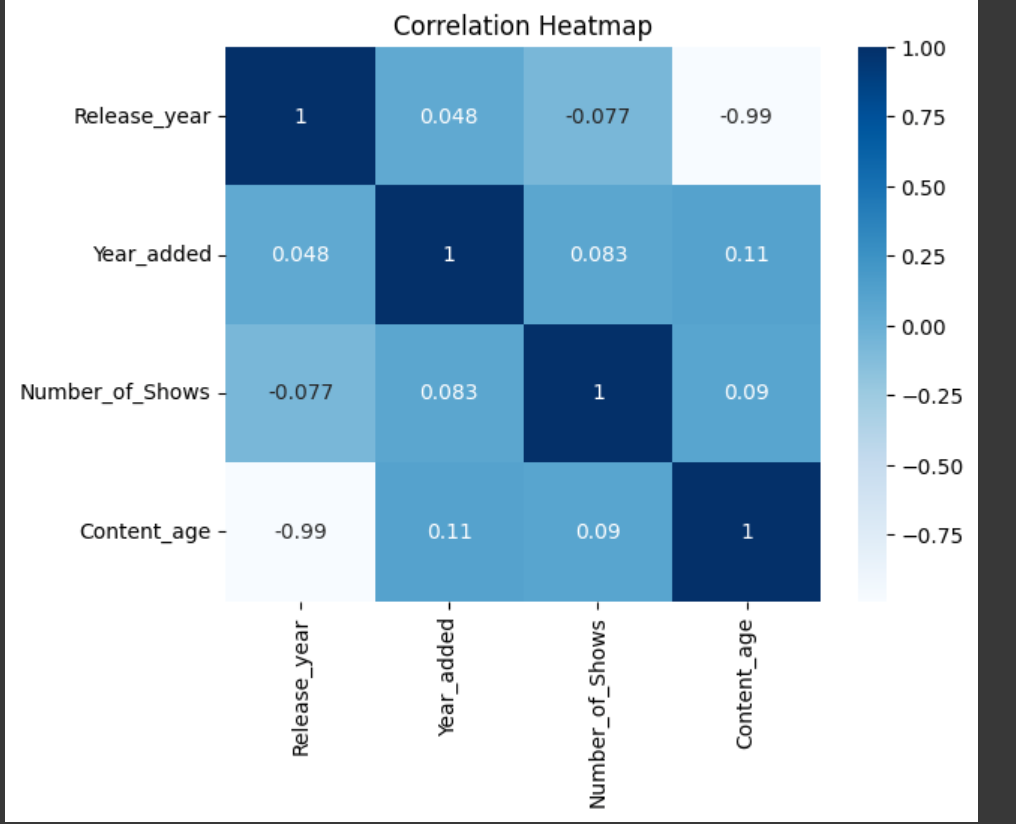


**Insights Gained:**

* United States are highest watching movies and also the TV Shows compare to others
* In 2015-2020 the netflix gets increase higher comparing to remaining years. But 2020 Becomes to lower.
* The chart shows that the highest average number of shows per release year occurred during the 1950s–1970s
* This chart represent the more correlation between (2000-2020)-(PG-13-TV-Y7)

1. **MULTIVARAIATE ANALYSIS**

Multivariate analysis was performed to explore relationships across **multiple variables simultaneously**, providing deeper insights into Netflix’s content strategy.



**Insights Gained:**

* Varied by the type of Shows Named as Movie and TV Show Y-Axis: 0,100,200,300,400,500,600. This chart represent the content rating by movie and tv show

1. **OVERALL INSIGHTS FROM ANALYSIS**
   1. **Univariate Analysis Insights**

* TV-MV means mature audience TV-14 means under 14 years childrens These two are highly rating compared to others rating structure
* It's Shows the highest number of shows watchers. The majority of watchers are three and two shows.
* The July Month has highest movie added in netflix compare to the other months After that April,
* Here the International Movies gener is too high compare to other geners then after that Dramas is second highest geners
* There's a sharp increase in releases from 2015 to 2020Most of the content in the dataset consists of movies

**2. Bivariate Analysis Insights**

* United States are highest watching movies and also the TV Shows compare to others
* In 2015-2020 the netflix gets increase higher comparing to remaining years. But 2020 Becomes to lower.
* The chart shows that the highest average number of shows per release year occurred during the 1950s–1970s
* This chart represent the more correlation between (2000-2020)-(PG-13-TV-Y7)

**3. Multivaraiate Analysis Insights**

* This chart represent the more correlation 0.083 and low correlation -0.99
* Varied by the type of Shows Named as Movie and TV Show Y-Axis: 0,100,200,300,400,500,600. This chart represent the content rating by movie and tv show
* This pair plot represent the pairwise relationships between Release\_year ,Year\_added ,Content\_age ,Number\_of\_Shows

1. **CONCLUSION**

This Netflix dataset project provided a comprehensive exploration of how the streaming platform structures its content library. Through **data cleaning, feature engineering, visualization, and statistical testing**, several key insights were uncovered

* Movies dominate Netflix’s catalog, though TV Shows have grown rapidly after 2015, reflecting Netflix’s shift toward episodic and binge-worthy content.
* Genres such as International Movies, Dramas, and Comedies represent the largest share of titles, highlighting Netflix’s focus on globally relatable and high-demand categories.
* Ratings distribution shows that most Netflix content is targeted at mature audiences (TV-MA, TV-14), aligning with global viewing trends.
* Country analysis indicates that the U.S. and India contribute the largest share of content, making them critical markets for Netflix.
* Statistical testing confirmed that:
* Movies are generally older than TV Shows (T-test).
* Ratings categories differ significantly in terms of release years (ANOVA).
* Content durations changed after 2015, particularly for TV Shows (Mann–Whitney U Test).
* Temporal insights revealed a surge of content additions between 2015–2020, with July being one of the peak months for new releases.

### **Business Impact**

These findings can help Netflix and other stakeholders:

* Optimize **content acquisition strategies** by focusing on trending genres and regions.
* Plan **targeted marketing campaigns** based on dominant ratings and genres.
* Anticipate **viewer demand trends**, especially with the rise of TV Shows and international content.

Overall, the analysis confirms that **Netflix’s growth strategy is heavily driven by international expansion, genre diversification, and the rising popularity of TV Shows**.