* **Introduction**

The preprocessing and cleaning procedures for the Netflix dataset are described in this report, which also offers insights via an interactive Tableau dashboard.

**Objectives** : Briefly state the goals, like gaining insights into Netflix content, understanding trends, and creating an easy-to-navigate visual representation for data analysis.

* **Data Description**

The Netflix dataset offers details on the different films and TV series that are accessible on the service. Important characteristics that aid in classifying, characterizing, and timestamping each title are included in this data. An overview of the dataset's primary characteristics is provided below**:**

1. **show\_id**: A unique identifier for each title (movie or TV show) in the dataset. This helps distinguish titles when performing analysis.
2. **type:** Specifies the type of content, either “Movie” or “TV Show.” This classification is useful for content comparison and understanding the overall balance of movies and shows on Netflix.
3. **title:** The title or name of the movie or TV show. This is a textual identifier representing each piece of content.
4. **director:** The name(s) of the director(s) associated with the movie or TV show. This field often has missing values, particularly for TV shows, where a director may not be specified.
5. **cast:** The main cast members involved in the movie or TV show. Like the director field, this attribute sometimes has missing values. However, it provides insights into popular actors and possible content preferences of the audience.
6. **country:** The country or countries where the movie or TV show was produced. This attribute helps in understanding regional content availability and Netflix’s global content strategy**.**
7. **date\_added:** The date when the title was added to Netflix’s catalog. This date is critical for analyzing content trends over time, such as the frequency of content additions in different years.
8. **release\_year:** The year the movie or TV show was initially released. This attribute helps analyze how recent or old the content on Netflix is and may reflect Netflix’s focus on newer or classic content.
9. **rating:** The maturity rating assigned to the movie or TV show (e.g., “PG-13,” “TV-MA,” “R”). This classification provides insights into the target audience age group and content suitability.
10. **duration:** Specifies the length of the content. For movies, it represents the runtime in minutes (e.g., "90 min"), and for TV shows, it shows the number of seasons (e.g., "3 Seasons").
11. **listed\_in:** A list of genres or categories under which the movie or TV show falls (e.g., “Dramas,” “Comedies,” “Action & Adventure”). This is useful for genre-based analysis and understanding Netflix’s content diversity**.**
12. **description:** A brief summary of the movie or TV show. This text provides a quick overview of the storyline or theme.

* **Initial Observations:**

1. **Missing Values:** Some attributes, such as director, cast, and nation, have missing values. This is probably because different types of content have different data availability (for example, TV shows may not have director information). Preprocessing requires cautious management of them.
2. **Duplicates:** To maintain data integrity, any possible duplicate rows in the dataset would be eliminated.
3. **Data Types:** The majority of fields are object-type, but in order to facilitate time-based analysis, columns such as date\_added should be transformed to datetime format.
4. **Text Fields:** Textual columns with leading and trailing spaces, such as the title and description, should be standardized.

**Summary of Data Size:**

* Number of Records: Approximately 6235 rows.
* Number of Attributes: 12 attributes, covering both categorical and textual data, as well as date-based and numerical fields.
* **Description of the Tool (Tableau)**

Tableau is a robust business intelligence (BI) and data visualization application that helps users turn unstructured data into visually engaging reports that can be shared. Its ability to link to a wide range of data sources makes it popular, and both novice and expert users may easily utilize it because to its intuitive interface. By providing aesthetically appealing reports and dashboards, Tableau aims to simplify and expedite data analysis while assisting stakeholders at all levels in making data-driven choices.

Tableau was selected for this project due to its user-friendly interface, speedy processing of huge datasets, and ease of use. It also offers a large selection of interactive visualization capabilities, which makes it perfect for clearly displaying insights and examining intricate data sets. Because they enable deeper insights into the dataset and dynamic data exploration, these features are particularly helpful when building dashboards.

**Key Features Used:**

**Data Connections:** xcel, SQL databases, and cloud-based data sources like Google Sheets and Amazon Redshift are just a few of the data sources that Tableau can connect to. You can work with a variety of datasets on a single platform thanks to its ability to link with many data sources, which facilitates data consolidation and analysis. In here, we have loaded the dataset form PC itself.  
  
**Options for Data Visualization:** Tableau provides a large selection of visuals. You may have utilized the following important features.  
1. Bar charts are helpful when comparing amounts in different categories.  
2. For monitoring trends over time, such shifts in views or other data, line graphs are perfect.  
3. Pie charts are useful for displaying distributions and proportions.

**Dashboard Functionality:** The ability to create interactive dashboards is one of Tableau's most potent capabilities. With these dashboards, Users can use filters to limit the data to particular sections, like time periods or regions.

* **Data Cleaning Process:**

**Methodology followed:**

import pandas as pd

# Load the dataset

file\_path = 'netflix\_titles.csv'

df = pd.read\_csv(file\_path)

# Display initial information about the dataset

print("Initial Data Information:\n", df.info())

print("\nMissing Values Per Column:\n", df.isnull().sum())

**# Step 1**: Drop duplicate rows, if any

df.drop\_duplicates(inplace=True)

**# Step 2:** Handle missing values

# Replace missing 'director' and 'cast' with 'Unknown' as they might not be essential

df['director'].fillna('Unknown', inplace=True)

df['cast'].fillna('Unknown', inplace=True)

# Drop rows where 'title' or 'date\_added' are missing (these might be essential fields)

df.dropna(subset=['title', 'date\_added'], inplace=True)

**# Step 3:** Convert data types

# Convert 'date\_added' to datetime format

df['date\_added'] = pd.to\_datetime(df['date\_added'], errors='coerce')

**# Step 4:** Standardize categorical text data

# Convert 'type' to lowercase for consistency

df['type'] = df['type'].str.lower()

# Standardize 'rating' to uppercase (e.g., to match any future encoding or analysis requirements)

df['rating'] = df['rating'].str.upper()

**# Step 5:** Clean text columns

# Remove leading/trailing whitespace from the 'description' column

df['description'] = df['description'].str.strip()

**# Step 6:** Handle multiple genres

# Split the 'listed\_in' column into a list of genres

df['listed\_in'] = df['listed\_in'].str.lower()  # Standardize to lowercase

df['genres'] = df['listed\_in'].str.split(', ')  # Split into a list of genres

**# Step 7:** Extract additional features

# Extract the year from 'date\_added' to analyze content addition over time

df['year\_added'] = df['date\_added'].dt.yea**r**

**# Step 8:** Remove unrealistic dates

# Retain rows with 'year\_added' between 1900 and the current year only

current\_year = pd.Timestamp.now().year

df = df[(df['year\_added'] >= 1900) & (df['year\_added'] <= current\_year)]

**# Step 9:** Rename columns for consistency

# Make column names lowercase and replace spaces with underscores

df.columns = [col.lower().replace(" ", "\_") for col in df.columns]

# Display the cleaned dataset information and first few rows

print("\nProcessed Data Information:\n", df.info())

print("\nFirst few rows of the cleaned dataset:\n", df.head())

# Save the cleaned dataset if needed

#df.to\_csv('/mnt/data/cleaned\_netflix\_titles.csv', index=False)

**Importance of Data cleaning:**

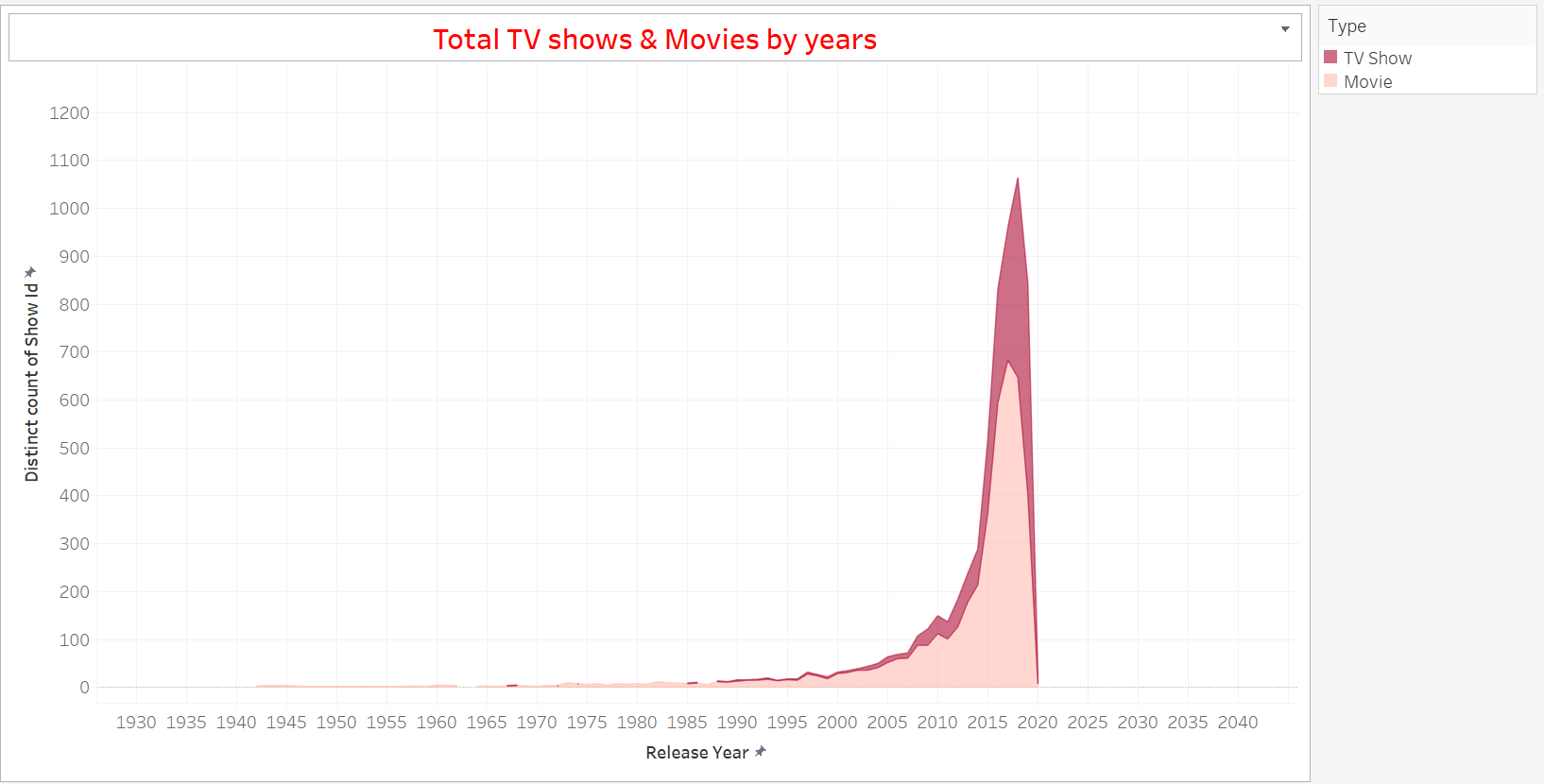
1. **Ensuring Data Quality:** Raw data is frequently unreliable, jumbled, or lacking. Data cleansing guarantees that the information you deal with is correct, trustworthy, and prepared for analysis. Any study or conclusions made in the absence of clean data may be inaccurate or misleading.
2. **Dealing with Missing Data:** Results may be distorted by missing or insufficient data. For instance, the dataset may be less representative or the analysis may be skewed if significant columns (such as title or date\_added) include missing values. Biased judgments can be avoided by appropriately managing missing data.
3. **Improving Consistency:** There may be differences in data from various formats or sources. For instance, text fields with names like "type" or "rating" may have differing naming rules or be mixed-case. By ensuring uniformity, standardizing these variables facilitates data analysis and visualization.
4. **Enabling Data Analysis:** More precise statistical analysis, modeling, or data visualization can be carried out with clean data. Well-structured data makes it easier to perform trend analysis, apply machine learning algorithms, and produce useful results.

* **Dashboard Explanation**

**About various sheets:**

1. **Sheet 1:**

Sheet 1 helps understand total TV shows & movies released by the year.

****

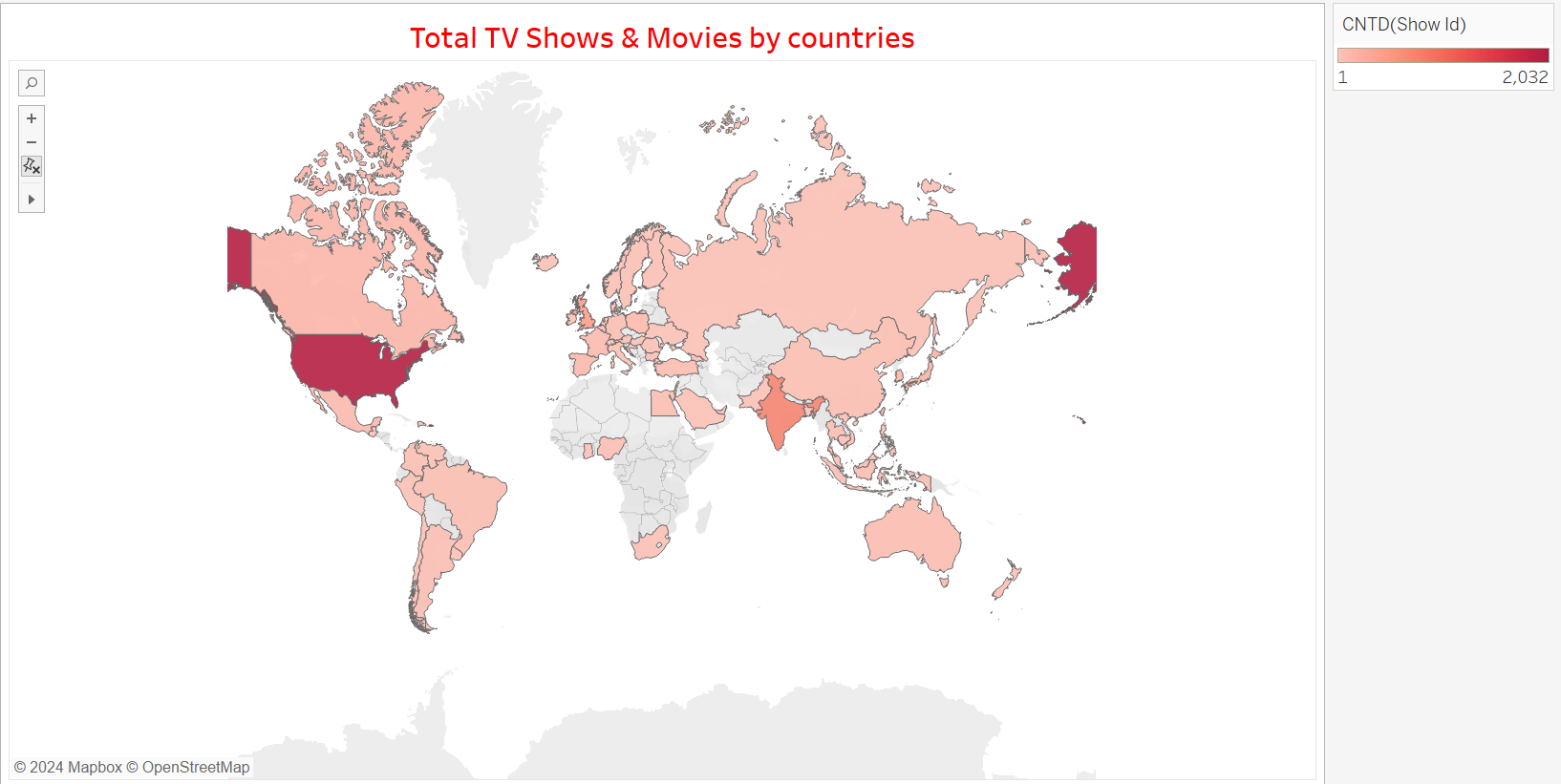
**How it can be used:**

Trend Analysis: The entertainment industry's trends can be found with the aid of this data. For instance, a sharp jump in releases in recent years may be a sign of a surge in content creation, which could be connected to the emergence of streaming services.

Content Strategy: By knowing when people are most interested in consuming material, entertainment organizations may use this knowledge to plan their release schedules.

1. **Sheet 2:**

Help under stand how many shows and the movies are released by the particular contry.



**How It can be used:**

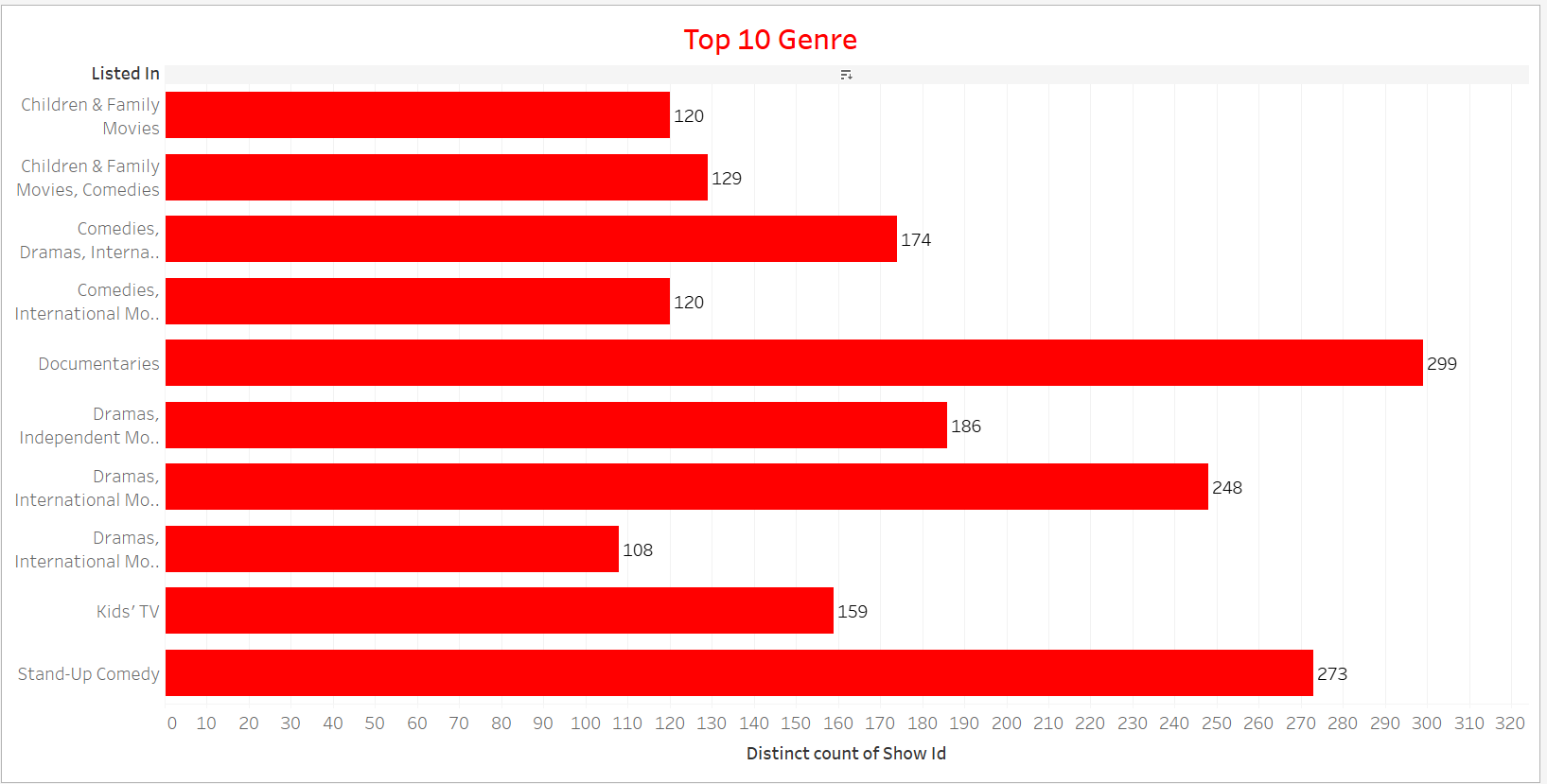
Regional Content Analysis: It helps analyze which countries are producing the most content, giving insights into global entertainment markets.

Market Opportunities: For streaming platforms and media companies, this information can identify regions with underrepresentation in content production. Companies may then focus on these regions for expansion or partnerships.

Cultural Preferences: By understanding the number of releases per country, it can also highlight cultural trends, such as the preference for certain types of content in specific regions**.**

1. **Sheet 3:**

Gives the information about the top 10 genre throughout the data.



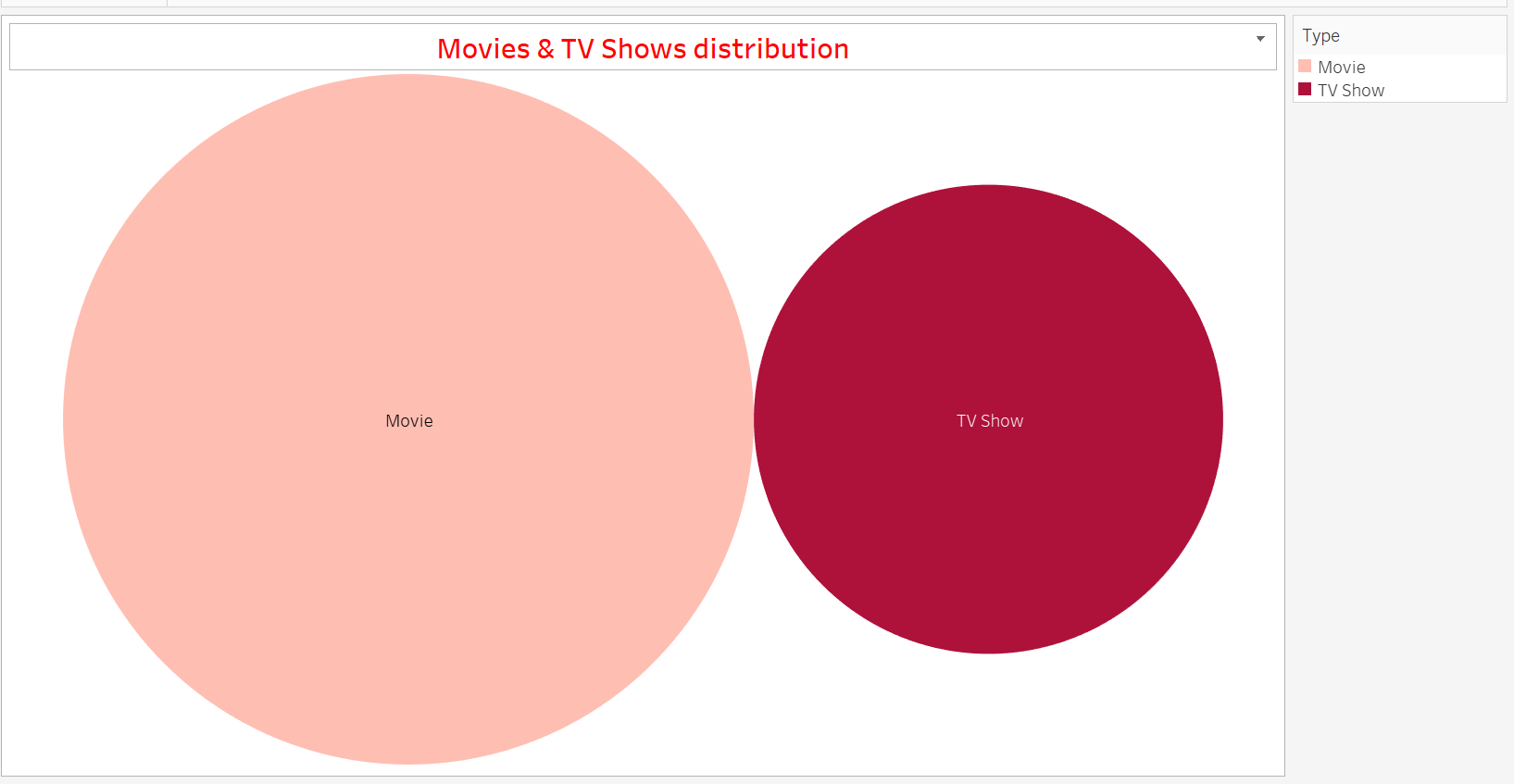
**How it can be used:**

Audience Preferences: This data provides valuable insights into consumer preferences. Knowing which genres are most popular helps content creators and marketers tailor their offerings to meet demand.

Targeted Marketing: Streaming platforms or TV networks can use this data to develop targeted marketing campaigns based on genre popularity.

1. **Sheet 4:**

Gives an information about the percentage of the show and the movies by total, by that, we can look for distribution.

****

**How it can be used:**

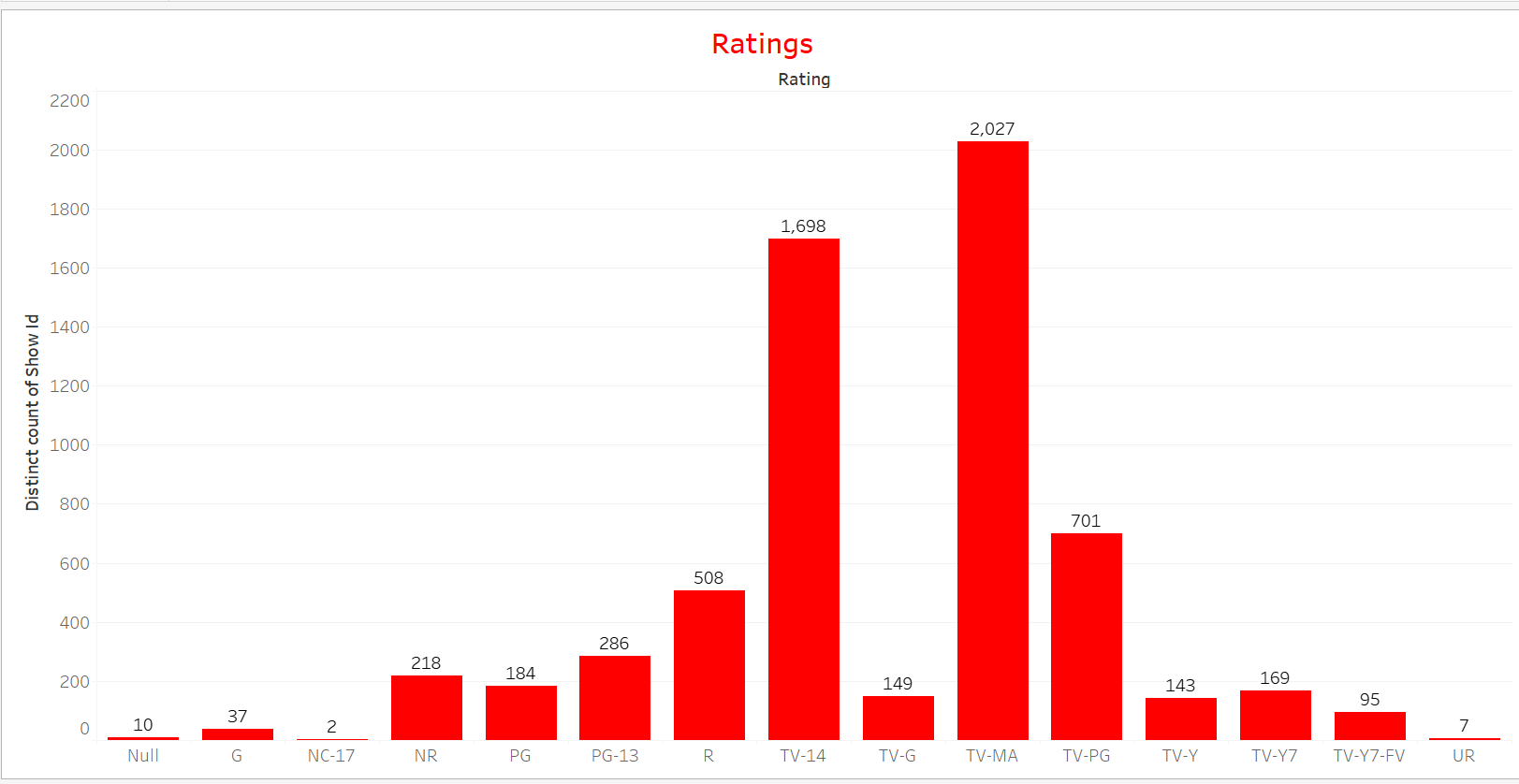
Content Distribution: It gives a clear picture of how much TV content versus movies are being produced and released. This can help production companies decide whether to focus more on TV shows or movies based on industry trends.

Market Segmentation: Streaming platforms can use this information to segment their content offerings, ensuring they maintain a balance that appeals to a broad audience.

Investment Decisions: Investors and stakeholders in the entertainment industry can use this data to evaluate whether more resources should be directed towards TV shows or movies based on current industry distribution.

1. **Sheet 5:**

Give information about the rating through the different platform.

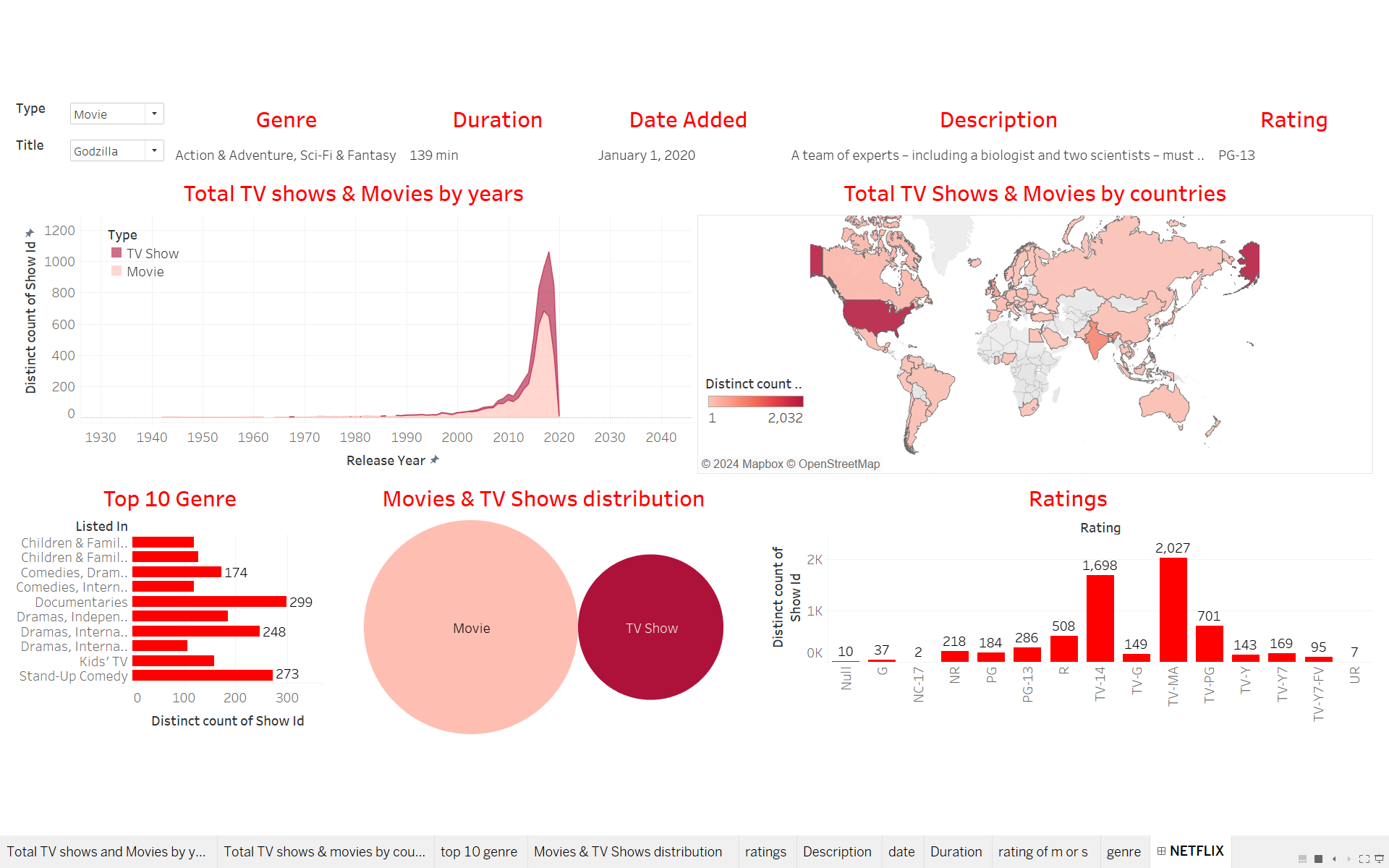


**How it can be used:**

Consumer Preferences: Knowing the ratings across platforms can guide viewers in choosing which service to subscribe based on the quality of content offered.

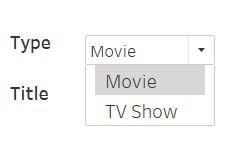
Performance Metrics: This data can be used by production companies or distributors to evaluate how well their content performs on various platforms, allowing them to optimize their distribution strategies.

* **The complete Dashboard:**

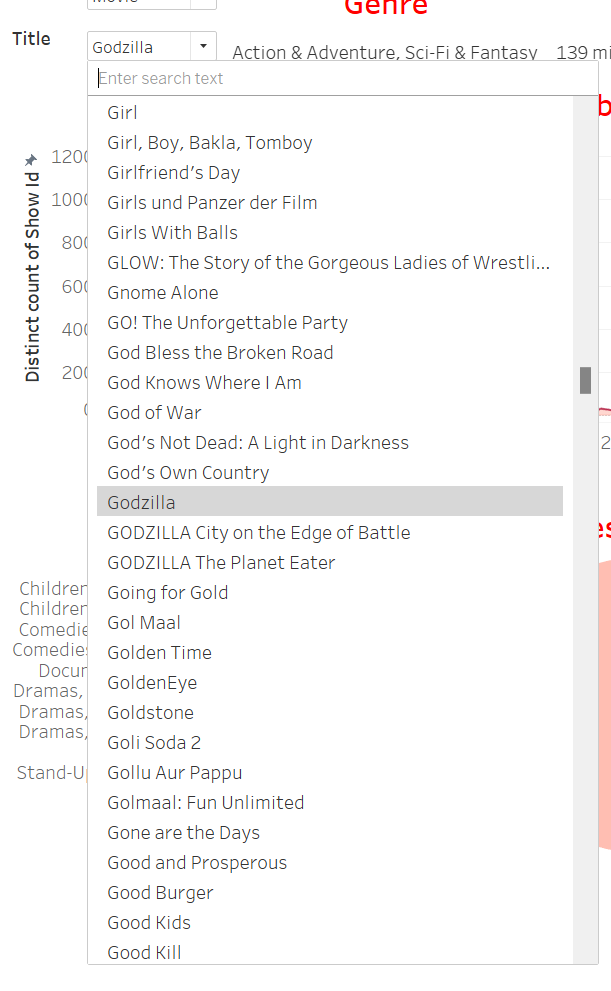
****

In this dashboard, from the top right we can choose the movie or show accordingly, and can have the genre, duration of the movie or the show. The date it was originally released, and the description and the rating of accordingly.

**Example:**



Select any two of them to search,



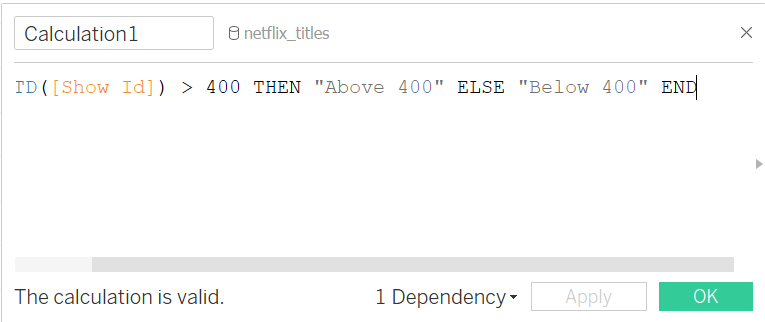
Select any of the movies or the shows, then you will find all the information about the particular show or the movie.

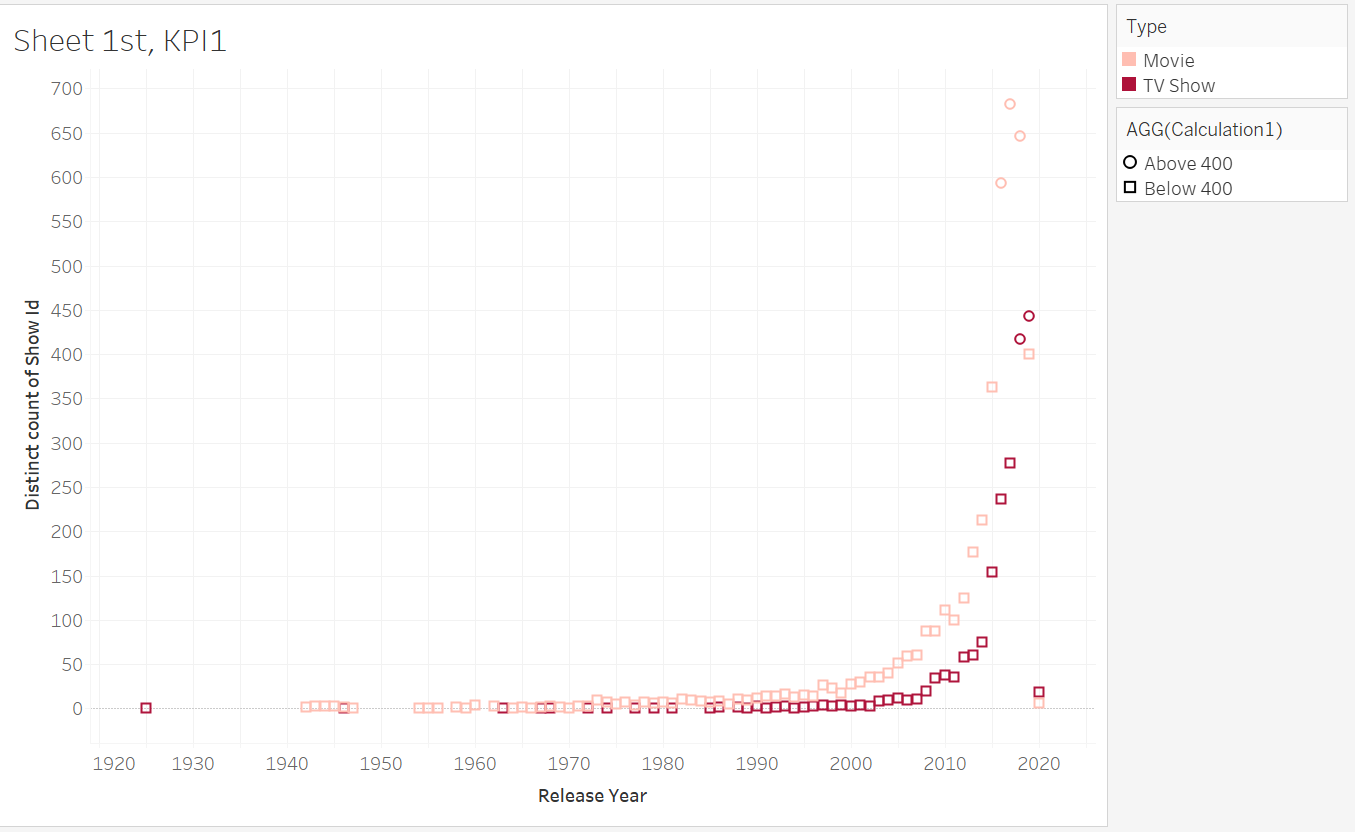


* **KPIs (Key Performance Indicators)**

**KPI 1:**

This approach will enable Tableau to dynamically show only the years where the number of TV shows and movies meets or exceeds the defined KPI threshold, helping to focus on periods of higher production.

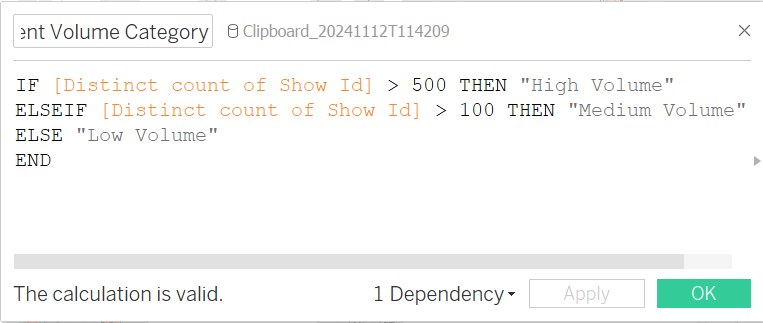
****

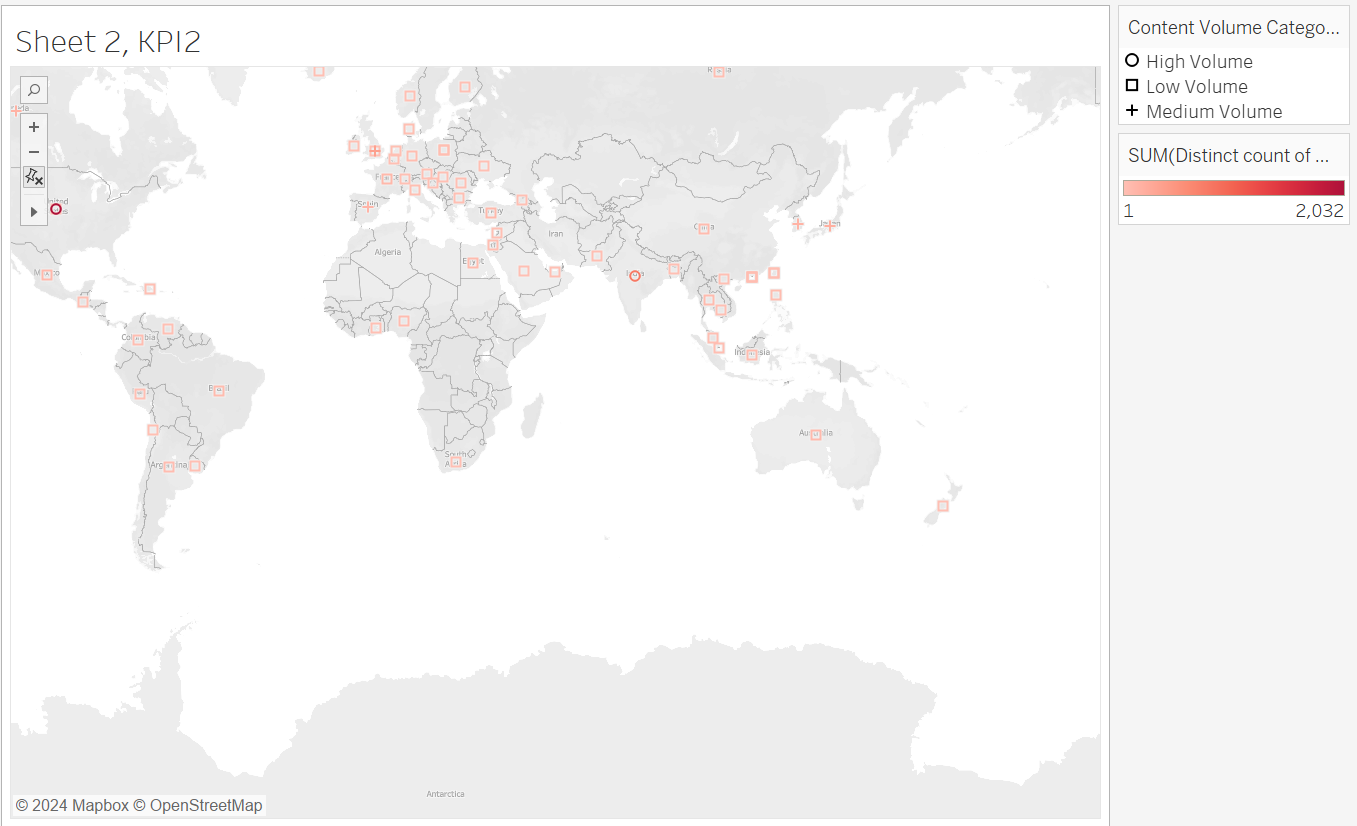
****

**KPI 2:**

Show only countries with a high count of distinct titles (e.g., above 500 titles).

Create a calculated field that categorizes countries based on content volume (e.g., "High Volume" if count > 500, "Medium" if 100-500, "Low" if below 100). Then apply this calculated field as a filter or add a color distinction based on the category.





* **Conclusion:**

The data cleaning and preprocessing procedures and a dashboard summary are all included in this report's exploration of Netflix content data. Tableau was used to examine the cleaned data, and visual representations of the findings helped to clarify distribution and trends in the content.