Netflix Data Analysis

Case Study

Objective

Netflix is a popular streaming service that offers a vast catalog of movies, TV shows, and original content. The purpose of this Netflix data analysis is to enhance Netflix's Content Strategy by using data-driven insights. This will allow to

1. Optimize content offerings across countries and genres.
2. Good understanding on audience preferences and the patterns
3. Assist in making better decisions about which genres, countries and types of content to expand

Netflix Dataset

The Netflix dataset from Kaggle contains information about movies and TV shows on the Netflix platform. Dataset Link: https://www.kaggle.com/datasets/shivamb/netflix-shows

Key info about the dataset

1. Content added to Netflix from 2008 to 2021
2. The number of row is 8807 with 12 columns
3. Data types of columns are shown.
4. Used Python to prepare and clean the dataset
5. Interactive reporting visuals created on the PowerBi tool

```
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
```

Purpose

The data analysis aims to understand the distribution of content types, popular genres, countries of origin and audience ratings. This will allow us to address key questions related to Netflix's content distribution to uncover opportunities to enhance its competitive edge in the market.

Key areas to focus on are

1. Analyze data pattern over Netflix's content type
2. Examine content distribution by country
3. Evaluating contributions of specific directors, actors or genres
4. Understanding the trends in genre diversity with duration

Data Cleansing

To prepare the dataset for visualisation, the following steps have been taken to clean the data

1. Handling manual errors
2. Handling null values and duplicate data
3. Cleaning column data value to have appropriate data type
4. Adding columns for later use
5. Changing data types
6. Creating a new subset of data
7. Python script to load data into PBI

```
# Correcting manual input by replacing rating to duration
df.loc[df['director']=='Louis C.K.','duration'] =df['rating']
df.loc[df['director']=='Louis C.K.','rating'] ='Unknown'

# dropping rows fpr small percentages if nulls
df.dropna(subset=['rating','date_added'],axis=0, inplace=True)

# Replace nan values with appropriate values
df['country'].replace(np.nan,'United States',inplace = True)
df['director'].replace(np.nan, 'Unknown', inplace = True)
df['cast'].replace(np.nan,'Unknown', inplace = True)

# Dropping duplicates
df.drop_duplicates(inplace=True)

# Removing characters from duration
df.duration=df.duration.apply(lambda x: x.replace (' min', '') if 'min' in x else x)
df.duration=df.duration.apply(lambda x: x.replace (' Season', '') if 'Season' in x else x)
df.duration=df.duration.apply(lambda x: x.replace ('s', '') if 's' in x else x)

# Correcting data types
df.loc[:,['duration']]=df.loc[:,['duration']].apply(lambda x: x.astype('int64'))
df['date_added'] = pd.to_datetime(df['date_added'])

# Add new columns for time analysis
df['date_added_year']=pd.DatetimeIndex(df['date_added']).year
df['date_added_month']=pd.DatetimeIndex(df['date_added']).month

# Mapping tables

# Creating a new list of directors with showid
director_s = df[['show_id','director']]
director_s = (director_s.drop('director', axis=1)
        .join (director_s.director.str.split(', ', expand = True).stack().reset_index(drop=True

# Creating a new list of cast with showid
cast_s = df[['show_id','cast']]
```

Data Analysis

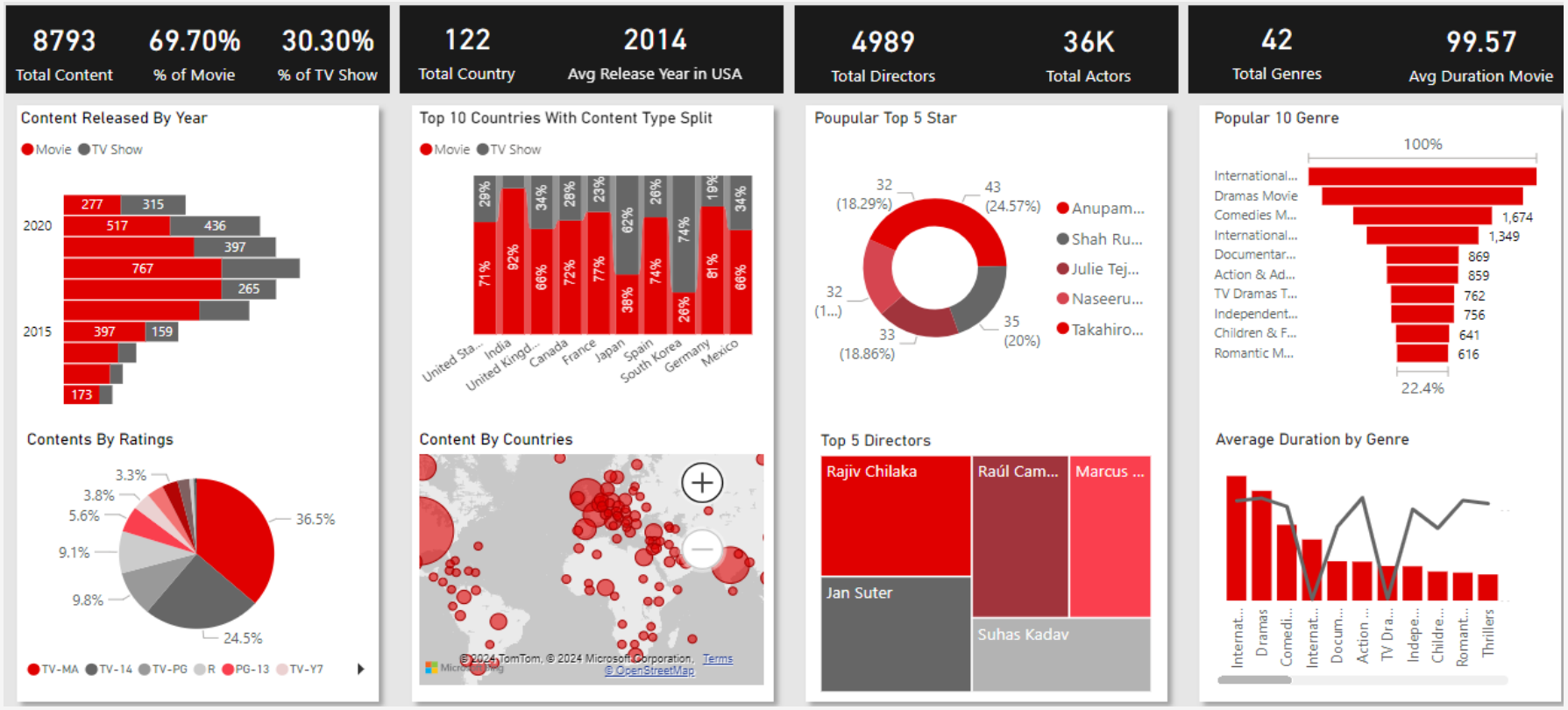
To understand trends and patterns following measures and KPIs used in visualisation

1. Count the number of Movies vs TV Shows
2. Find the most common rating for movies and TV shows
3. Identify understated genres on Netflix
4. Analyze country-wise contributions to Netflix's library
5. I identify the most frequent directors, actors and genres
6. Find the average release year for content produced in a specific country
7. Identify the longest movie and TV show duration

Dashboard Reporting

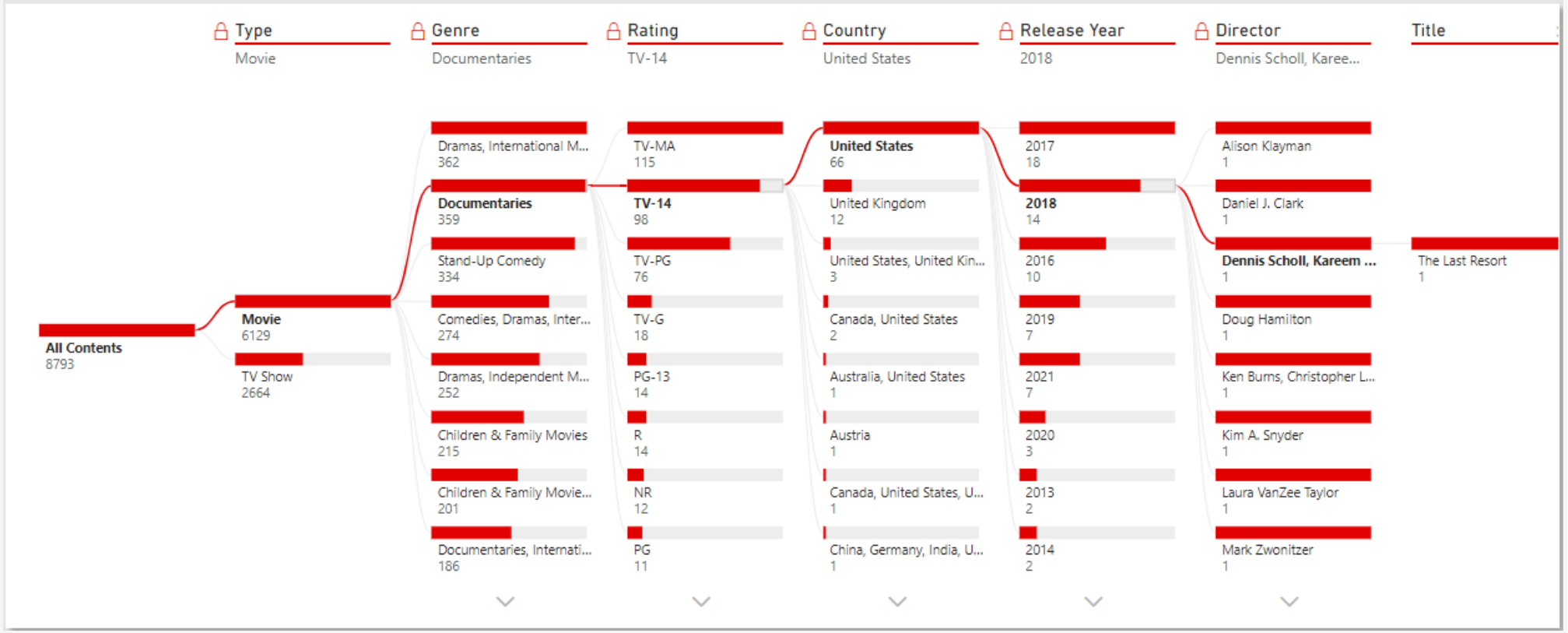
An interactive report indicates the key performance indicators over content types, distribution, and origin of countries, and a few visuals analyse the patterns and trends over different categorisations.

1 Content Analysis



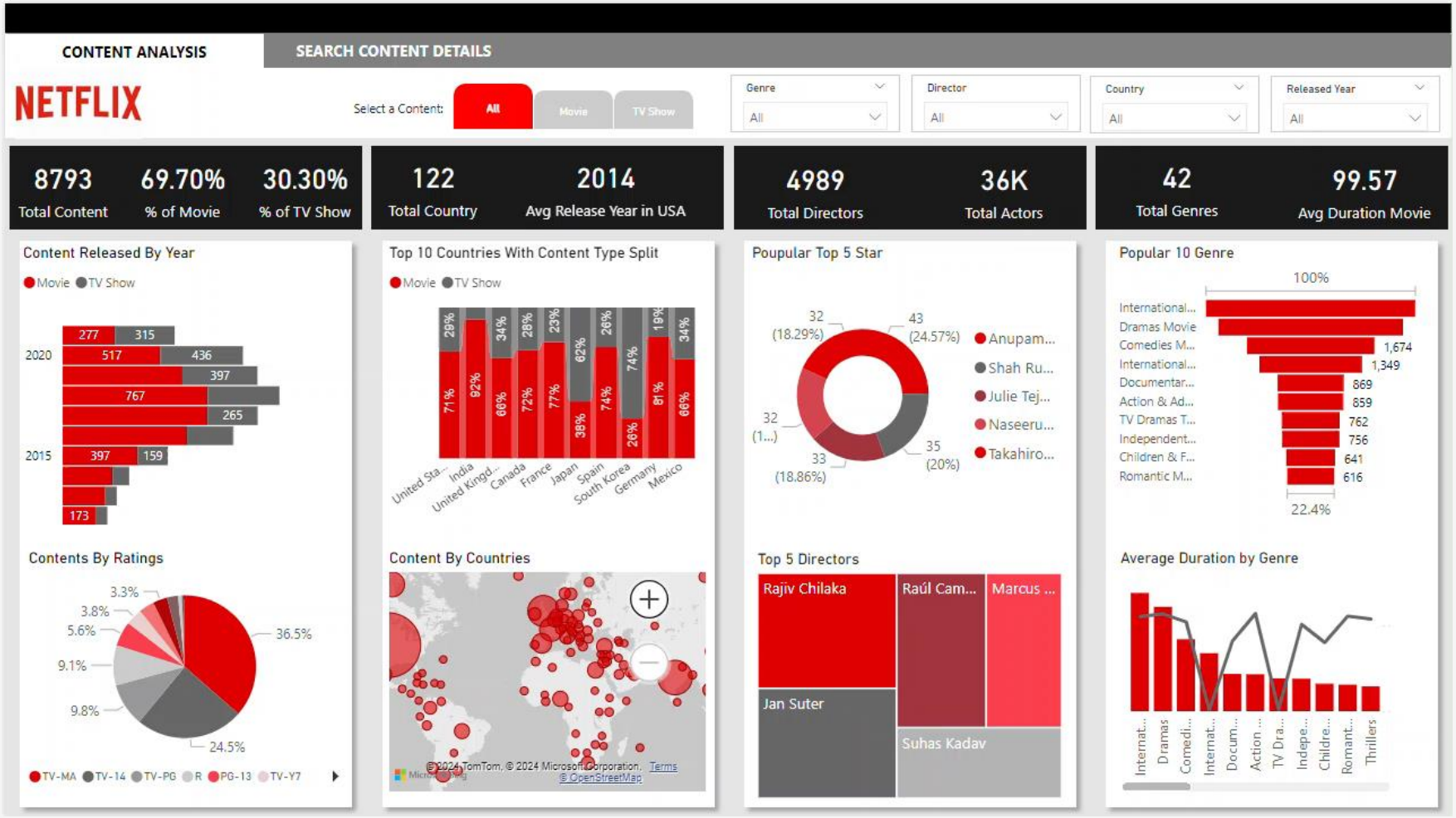
Dashboard Reporting

2 Search Content



Dashboard Reporting

3 Menu Navigation



Conclusion

In conclusion, Netflix's data analysis reveals key insights into user behaviour, content preferences, and operational trends that are essential for driving business decisions and enhancing user experience.

Followings are the key takeaways

1. Netflix has more Movies than TV shows
2. Most Movies & TV shows are produced in the United States, followed by India which has made the second number of movies on Netflix
3. 2018 is the year in which Netflix released a lot more Content as compared to other years
4. International Movies, Dramas & Comedies are the most popular Genres on Netflix
5. Netflix's movie duration is mostly between 80 to 120 minutes, and there is a drastic difference in preferred one season of TV shows.

END