Netflix Data Analysis

Agenda

Topics Covered

Data Cleaning with python

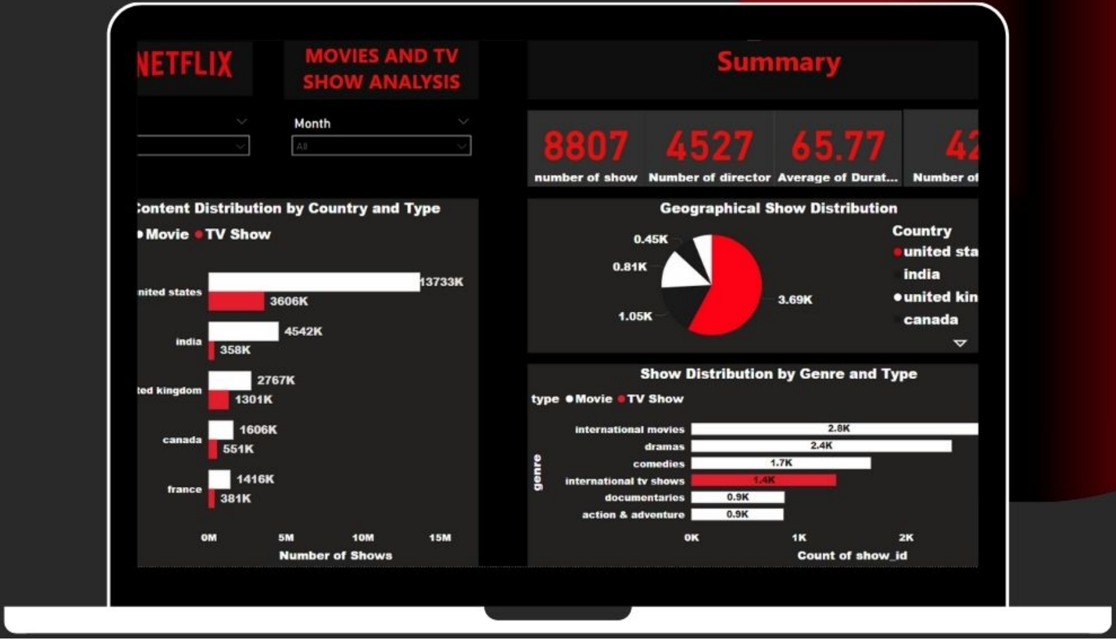
Database Connection with MySql

Data manuplation by power bi Dax

Data Visualization by power bi

Project Objective:

- Primary Goal: To gain insights into viewer preferences and trends on Netflix by analyzing show types, genres, and viewer ratings across different regions.
- Secondary Goals: To understand how these trends can influence content creation and marketing strategies for Netflix.
- Tools and Methods Used
- Mention the use of Python for data cleaning and manipulation, SQL for data retrieval and management, and Power BI for visualization.
-



Data cleaning by python

What is Data Cleaning

Briefly explain the importance of data cleaning in ensuring accurate, reliable analysis. Mention that clean data helps in making sound decisions based on the analysis.

Challenges in the Netflix Dataset

Inconsistent Formats: Describe how data entries like dates and show IDs had inconsistent formats that required standardization.

Missing Information: Highlight issues like missing director or cast details and how they impact analysis.

Duplicate Records: Point out any duplicate entries found in the dataset and the need to remove them to avoid skewed results.

Data cleaning by python

Import Library

```
import pandas as pd
import numpy as np
```

Pandas: Pandas is a powerful and popular open-source Python library specifically designed for data manipulation and analysis.

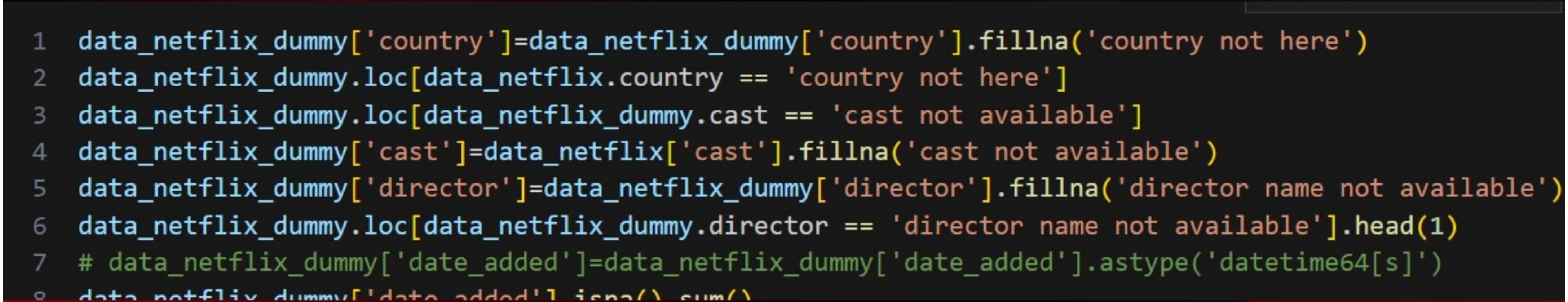
Numpy:a Python library that provides support for multidimensional arrays and mathematical functions for scientific computing. It's short for "numerical Python"

Data cleaning by python

Steps Taken for Data Cleaning

Standardization: Discuss the methods used to standardize data formats, such as converting all dates to a single format. Changing data type by “astype” and “Dtype”

Handling Missing Values



```
data_netflix_dummy['country']=data_netflix_dummy['country'].fillna('country not here')
data_netflix_dummy.loc[data_netflix.country == 'country not here']
data_netflix_dummy.loc[data_netflix_dummy.cast == 'cast not available']
data_netflix_dummy['cast']=data_netflix['cast'].fillna('cast not available')
data_netflix_dummy['director']=data_netflix_dummy['director'].fillna('director name not available')
data_netflix_dummy.loc[data_netflix_dummy.director == 'director name not available'].head(1)
# data_netflix_dummy['date_added']=data_netflix_dummy['date_added'].astype('datetime64[s]')
data_netflix_dummy['date_added'].isna().sum()
```

Data cleaning by python

Normlization

the process of organizing data in a database to make it more flexible and protect it. It involves: Creating tables, Establishing relationships between tables, Eliminating redundancy, and Removing inconsistent dependency

Problem in My Data Set

Genre Column

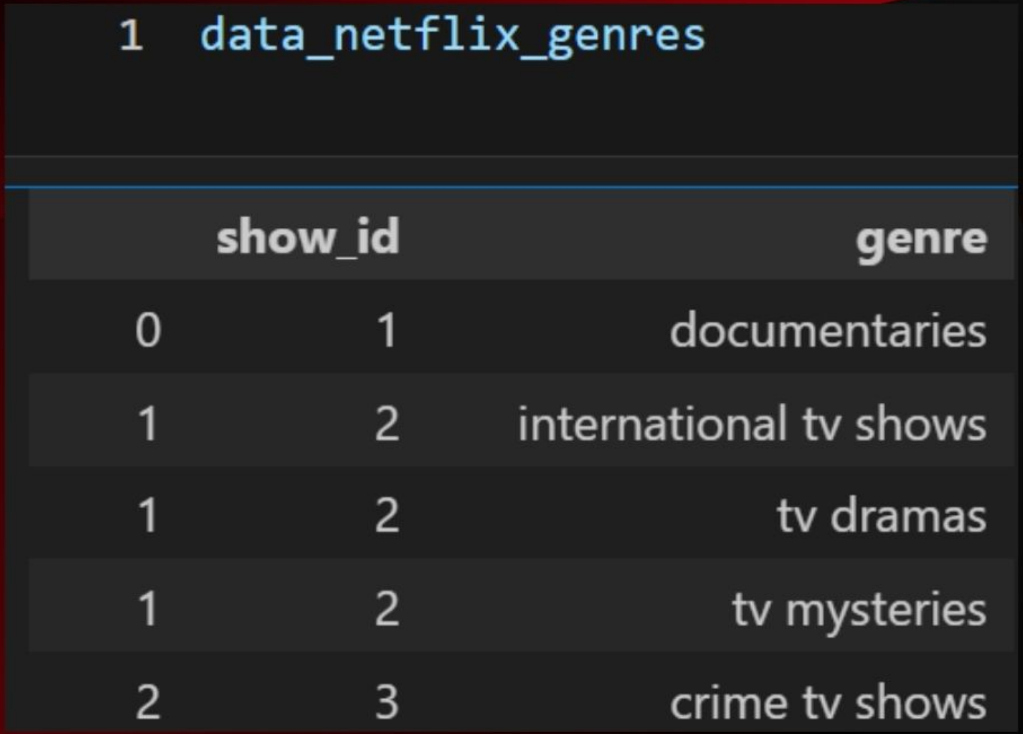
Action & Adventure	128
Action & Adventure, Anime Features	2
Action & Adventure, Anime Features, Children & Family Movies	12
Action & Adventure, Anime Features, Classic Movies	6
Action & Adventure, Anime Features, Horror Movies	

As you see above a Action & Adventure are repeated in many rows but when we do group by where as Action & Adventure is single its count not with other so solved this.

Data cleaning by python

Solution Of Multivalued Column

Making other table is solution of multivalued is called 1 NF

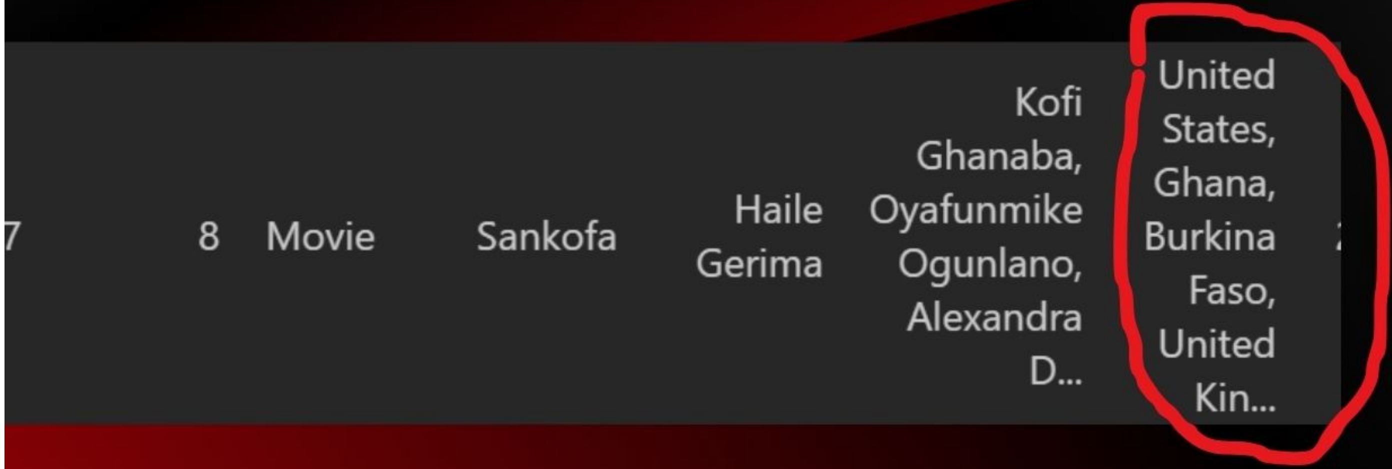


```python
data_netflix_genres
```

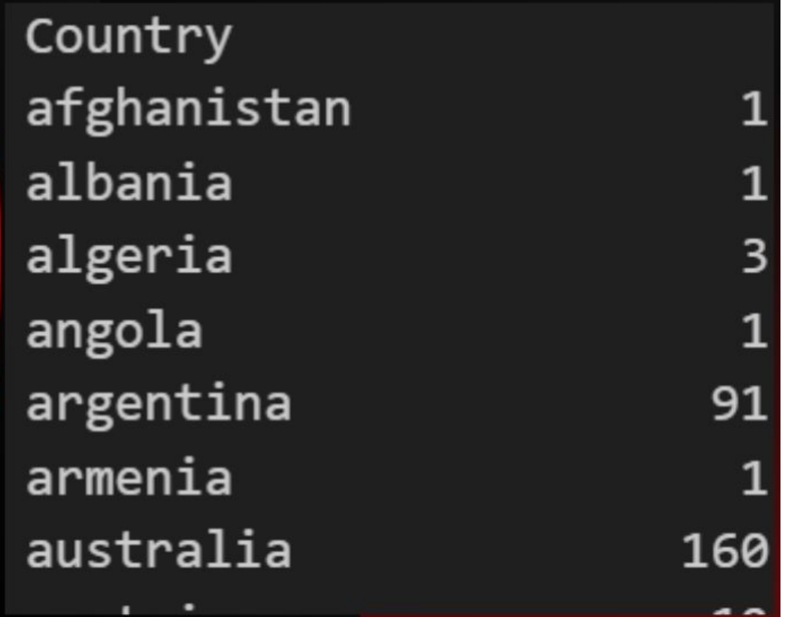
	show_id	genre
0	1	documentaries
1	2	international tv shows
1	2	tv dramas
1	2	tv mysteries
2	3	crime tv shows

Data cleaning by python

Same as in country column



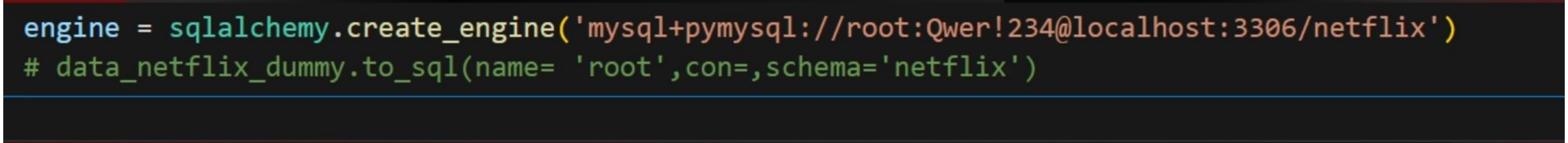
7	8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...



```
Country
afghanistan      1
albania          1
algeria          3
angola           1
argentina       91
armenia          1
australia      160
```

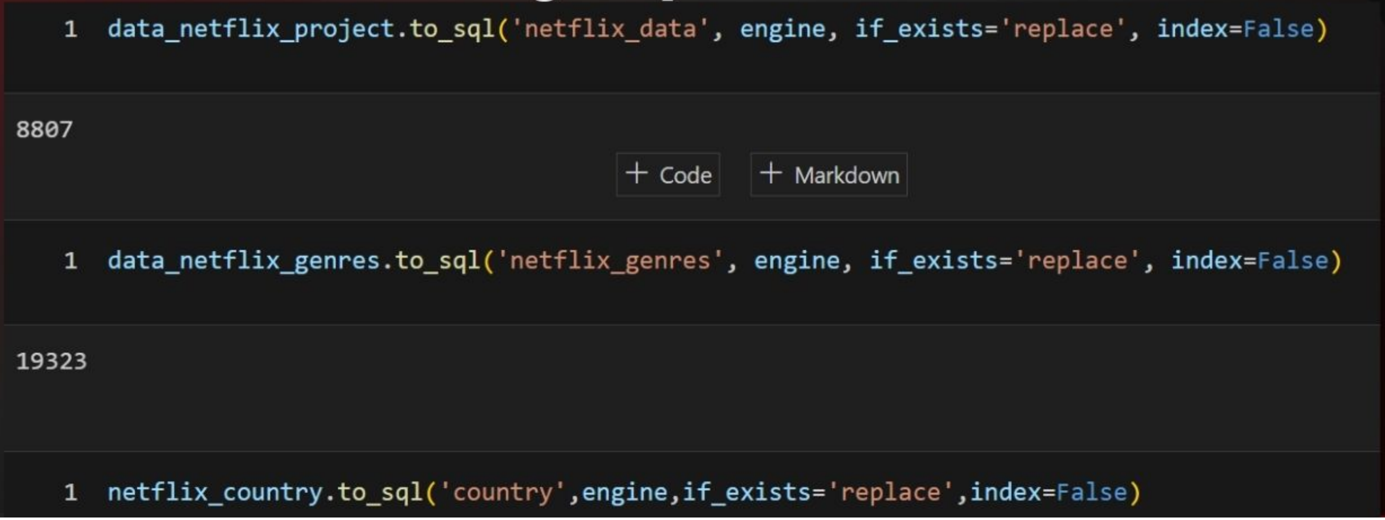
Database Connection with MySql

import sqlalchemy



```
engine = sqlalchemy.create_engine('mysql+pymysql://root:Qwer!234@localhost:3306/netflix')
# data_netflix_dummy.to_sql(name= 'root',con=,schema='netflix')
```

Connection with MySql



```
data_netflix_project.to_sql('netflix_data', engine, if_exists='replace', index=False)
```

```
8807
```

```
data_netflix_genres.to_sql('netflix_genres', engine, if_exists='replace', index=False)
```

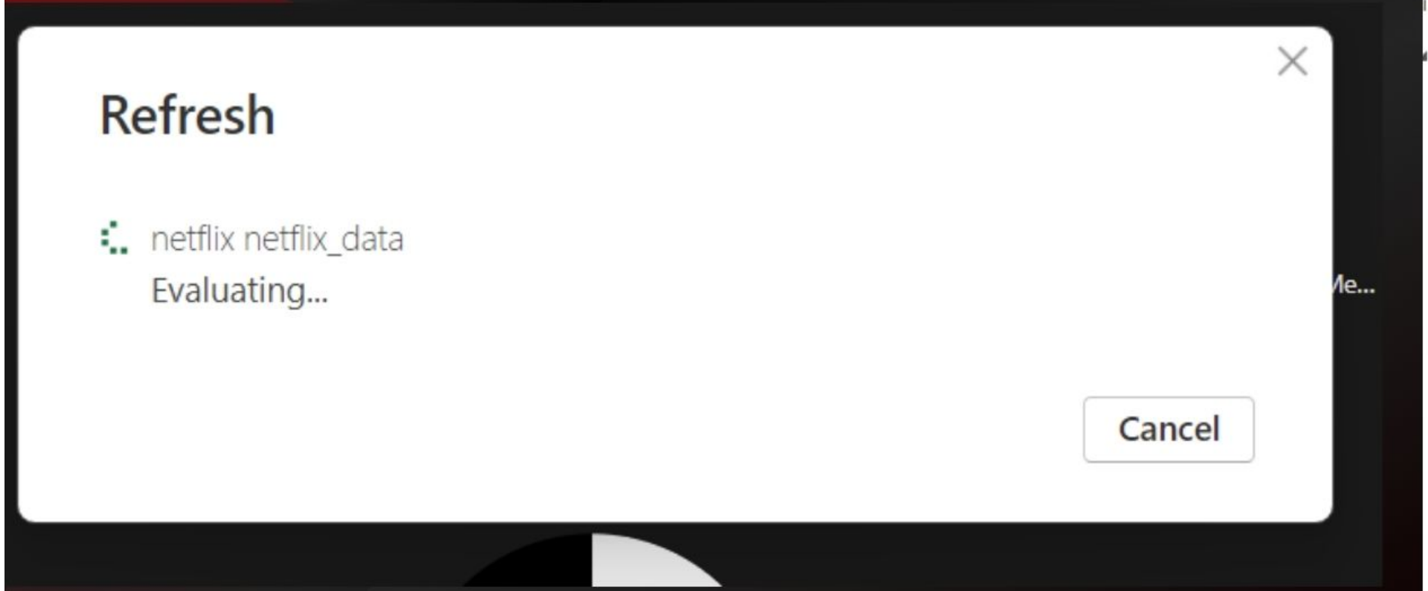
```
19323
```

```
netflix_country.to_sql('country',engine,if_exists='replace',index=False)
```

Data manuplation by power bi Dax

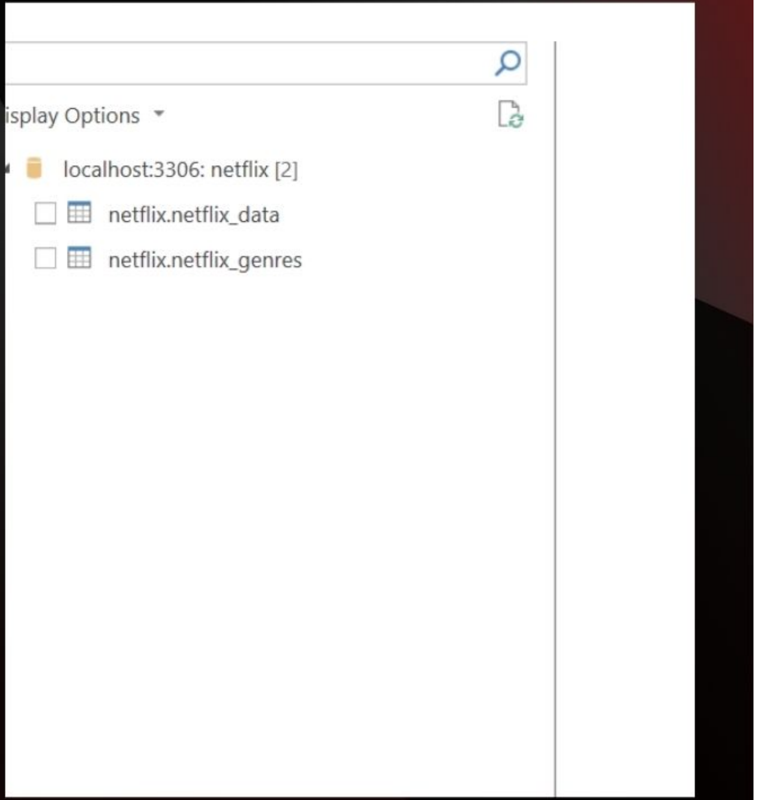
Connection Sql with Power Bi

Benifit of these changing in data set or real time data set we work easily.



Refresh

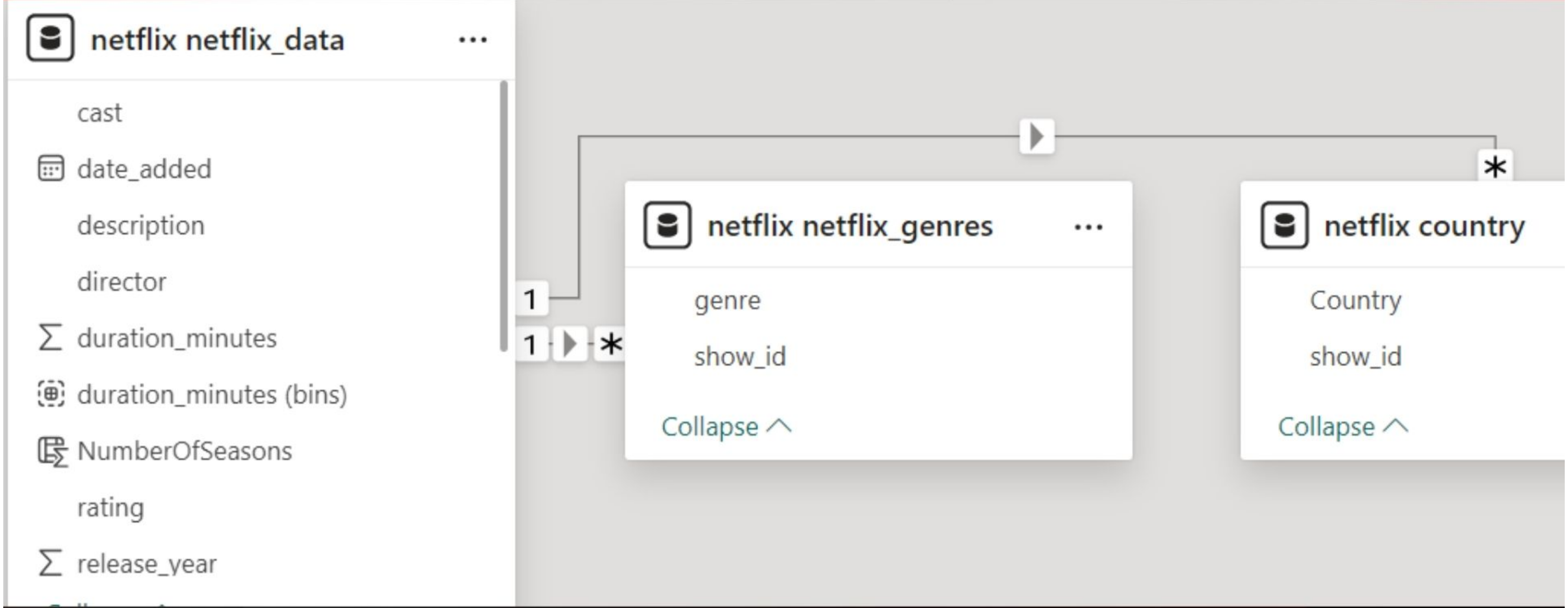
netflix netflix_data
Evaluating...

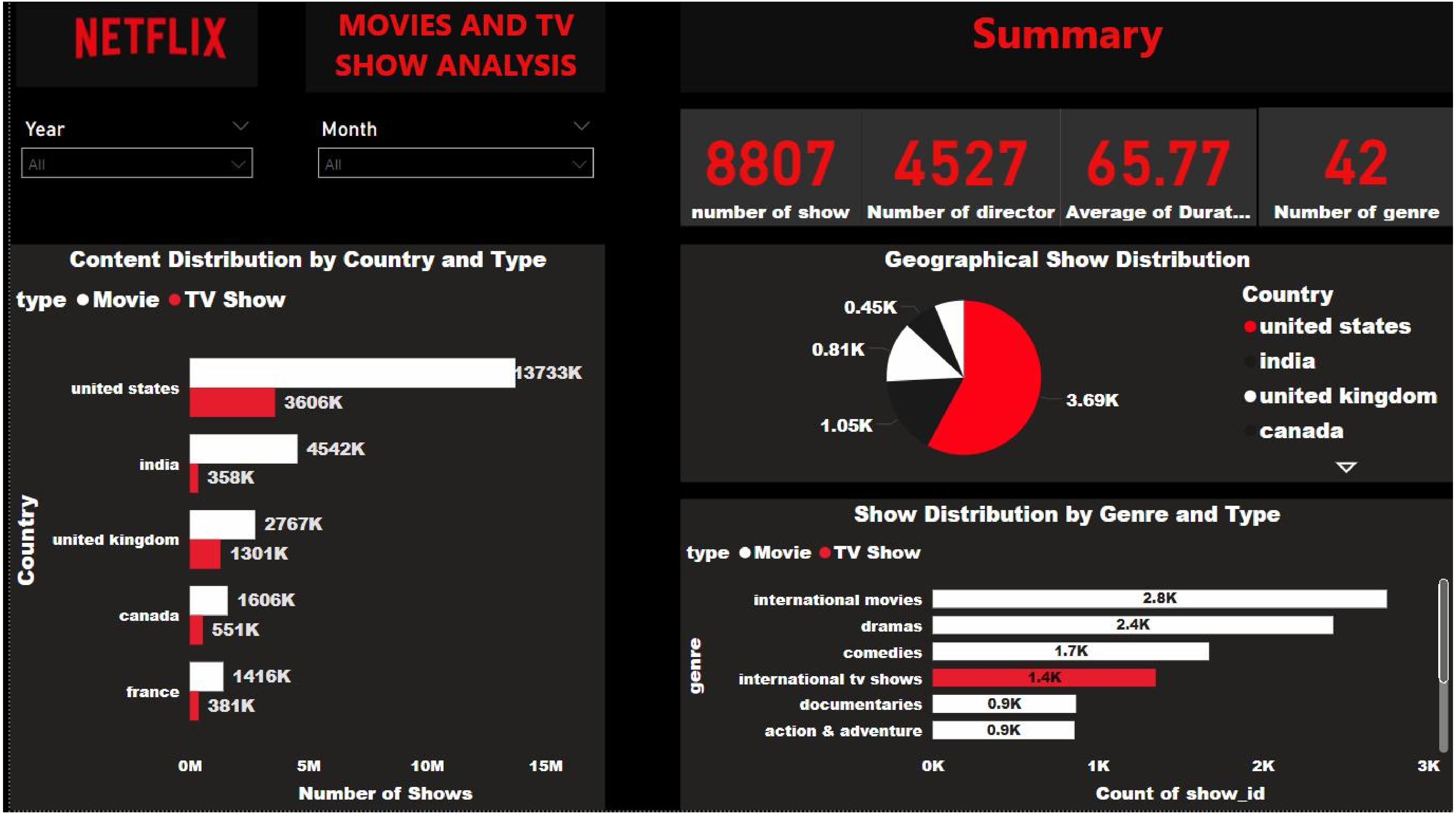


localhost:3306: netflix [2]
☐ netflix.netflix_data
☐ netflix.netflix_genres

Data manuplation by power bi Dax

Relationship between tables





NETFLIX

MOVIES AND TV SHOW ANALYSIS

Summary

Year: All

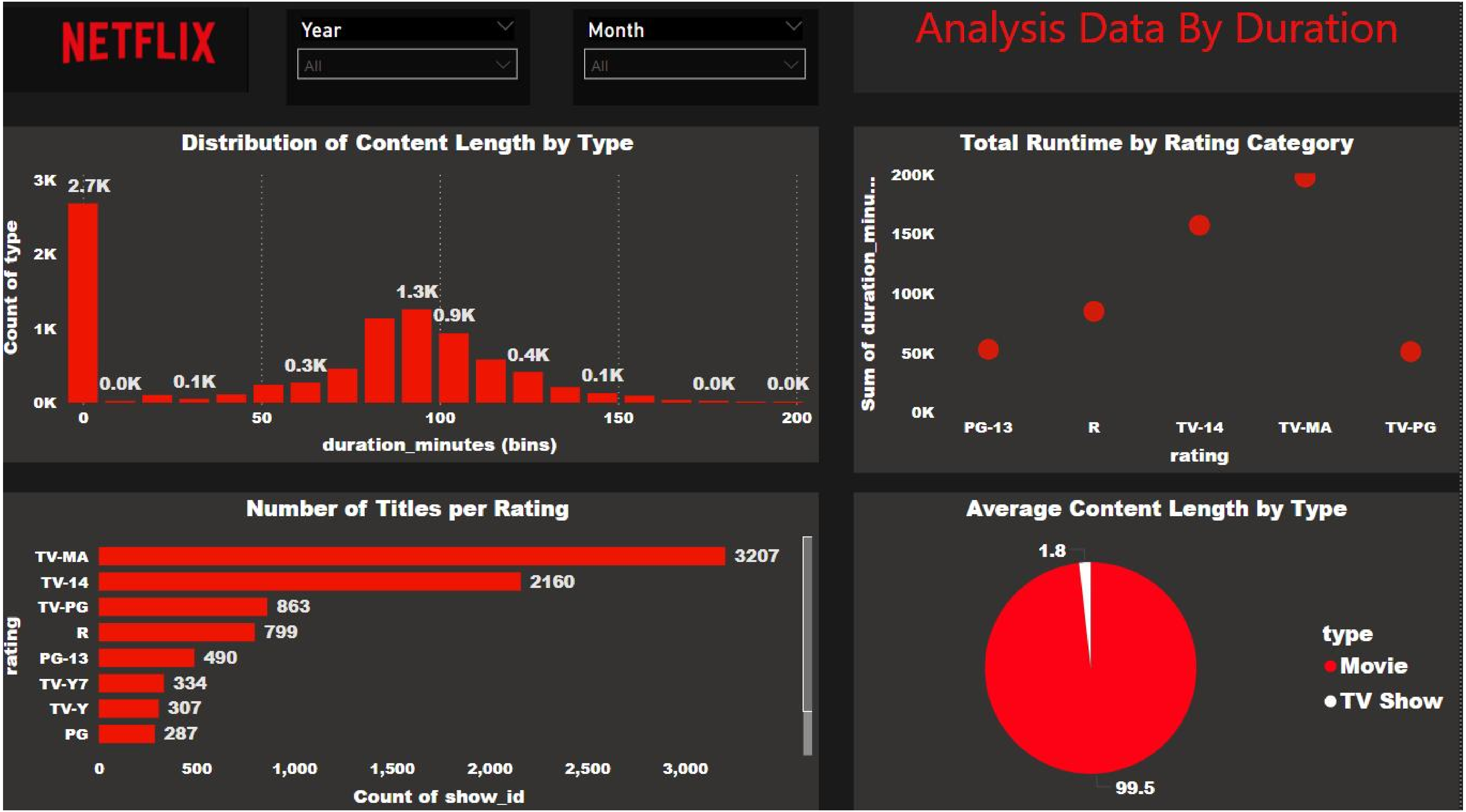
Month: All

8807	4527	65.77	42
number of show	Number of director	Average of Durat...	Number of genre

Content Distribution by Country and Type

Geographical Show Distribution

Show Distribution by Genre and Type



NETFLIX

Year: All

Month: All

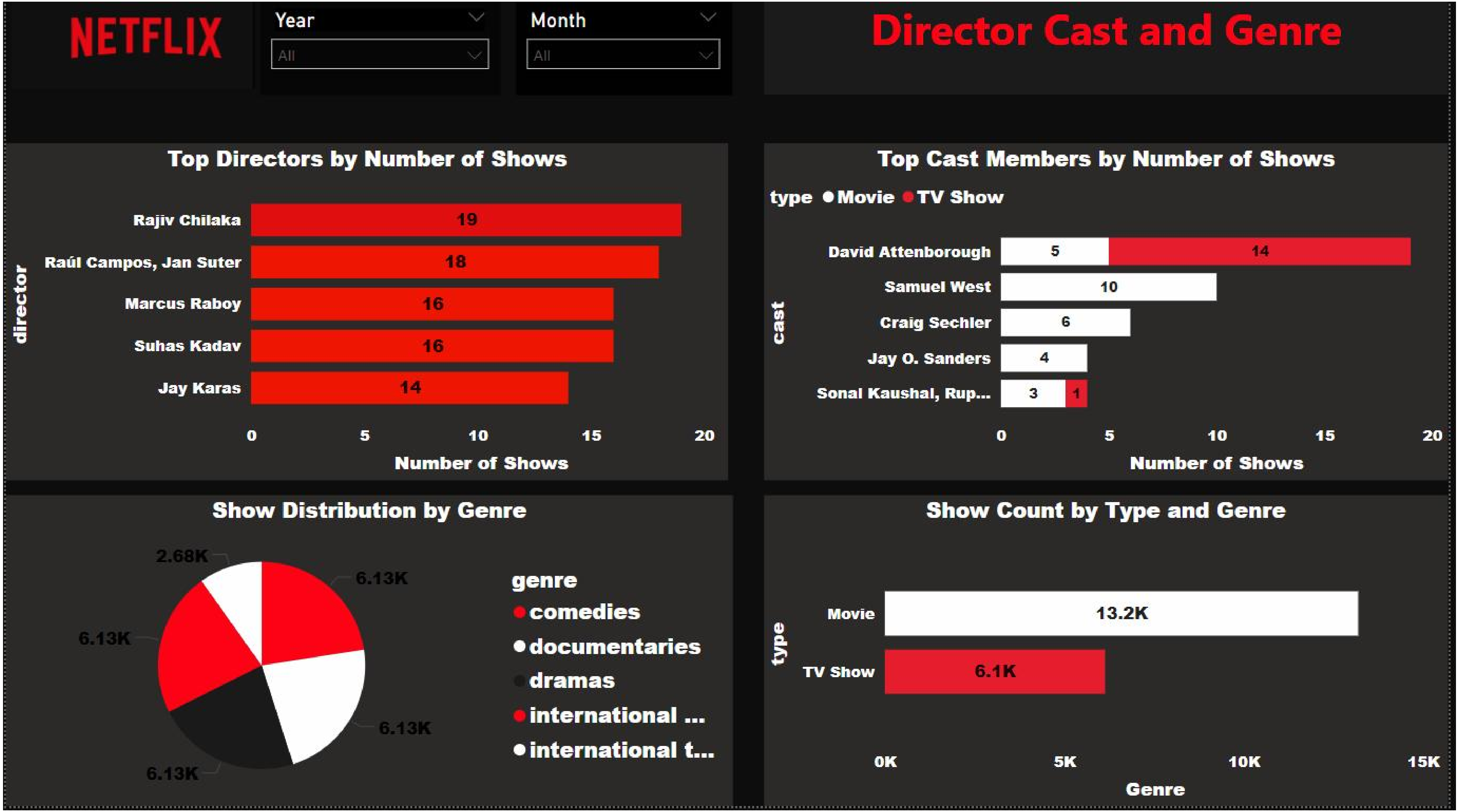
Analysis Data By Duration

Distribution of Content Length by Type

Total Runtime by Rating Category

Number of Titles per Rating

Average Content Length by Type



NETFLIX

Year
All

Month
All

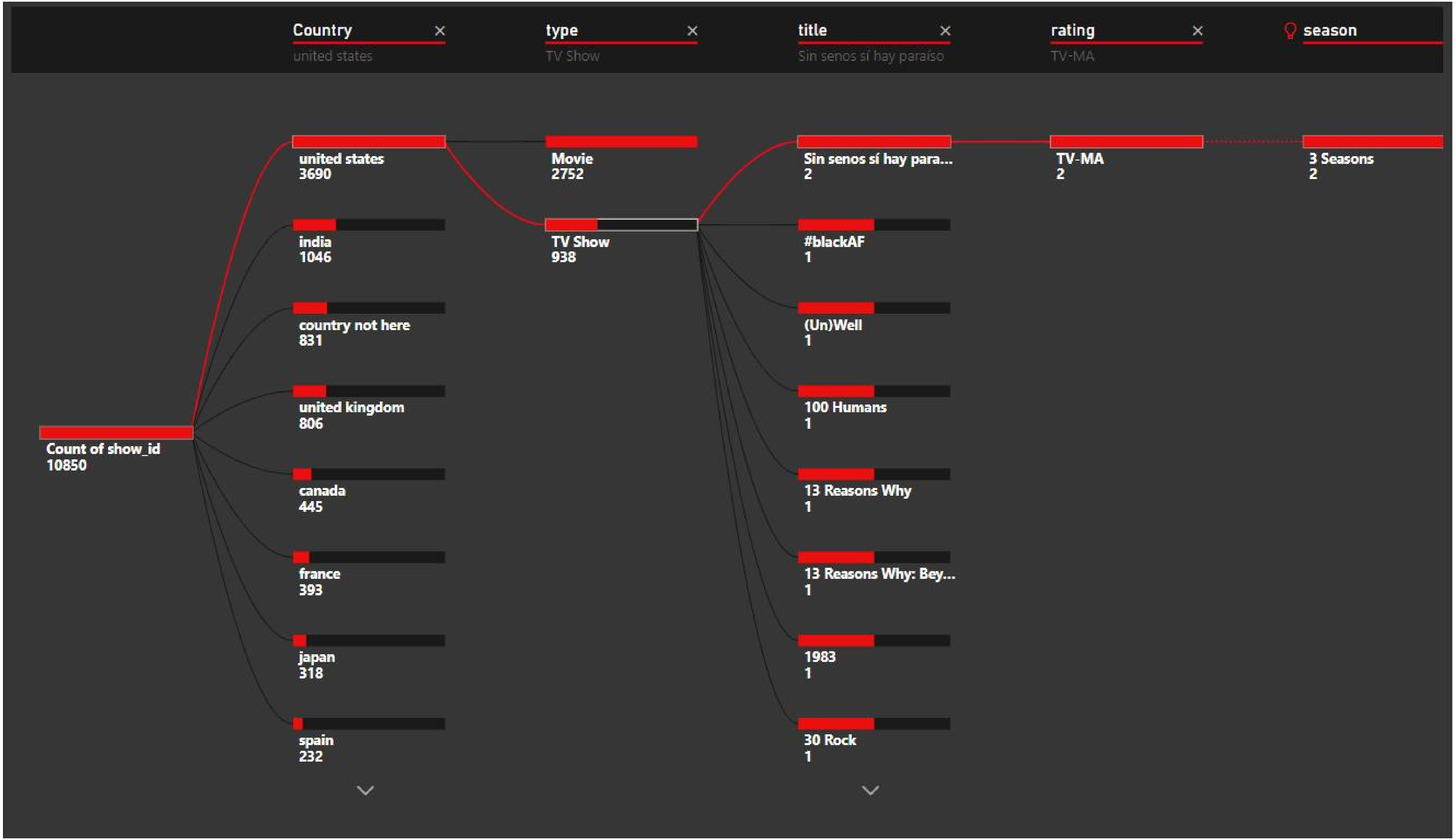
Director Cast and Genre

Top Directors by Number of Shows

Top Cast Members by Number of Shows

Show Distribution by Genre

Show Count by Type and Genre



Country ✕ united states

type ✕ TV Show

title ✕ Sin senos sí hay paraíso

rating ✕ TV-MA

season

Insights from Netflix Data Analysis

Dominance of United States Content:

The United States offers significantly more content compared to other countries, showcasing its leading role in the content availability on Netflix.

Preference for Movies Over TV Shows in the US:

Analysis data by duration using different charts.

Popularity of International Movies:

International movies constitute a substantial portion of Netflix’s catalog, highlighting its diverse global audience appeal.

Notable Figures:

- David Attenborough is prominently featured across various content, emphasizing a strong viewer interest in nature and documentary genres.
- Rajiv Chilaka holds the title for directing the most shows

Thank you!

Kush Shukla