'HISTORICAL ANALYSIS OF OTT DATA' IN POWER BI DATA VISUALIZATION

BRIDGEWATER STATE UNIVERSITY
Department of Computer Science
Spring 2023
May 8, 2023

Sri Vaishnavi Kumbar
skumbar@student.bridgew.edu
Mentor: Haleh Khojasteh

ACKNOWLEDGEMENT

ABSTRACT

Data visualization is essential to assist organizations in quickly identifying trends in the data, which might in other case be a hassle. The pictorial or graphical representation of data sets allows analysts to visualize concepts and new patterns. With the increase in data day to day, experiencing the quintillion bytes of data is impossible without Data Proliferation, which includes data visualization. All the mediums used for visualizing and understanding data are a dashboard, graph, infographics, map, chart, video, slide, etc. Just as a picture is worth a thousand words, a visual is worth a thousand data points. Visualizing the data allows decision-makers to interrelate the data to find better insights and acquire the importance of data visualization. Visualization quality and readability will depend on the tool we use and, therefore, it is important to choose an application that suits our needs. Power BI is an application designed for business analysis and enables access to data from any device. Power BI is a predictive and intelligent system which has the capacity to transform the most complex data into visualizations or reports with attractive designs that promote understanding and legibility. It contains queries, different visualization settings, data models, reports, and other data which is added by the user.

This Power BI project includes visualizing various information related to OTT platforms such as Netflix. By accessing OTT media platform dataset from Kaggle and using the Query Editor in Power BI for data cleaning and preparation. Once data is ready for visualization, displaying the visuals using various plots, graphs, cards, such as pie charts, bar graphs, line charts, tree maps, doughnut Charts, etc. to build the dashboard. The goal of this directed study involves visualizing various information related to an OTT platform such as Netflix, Hotstar, Amazon Prime, etc., by using the Query Editor in Power BI for data cleaning and transformation and visualize different patterns, spot trends, identify outliers and create the dashboard using the data visualization tool - Power BI.

Table of Contents

1 INTRODUCTION

In simple terms, PowerBI is a data visualization tool that allows users to transform raw data into meaningful and actionable insights. With its advanced analytical capabilities, Power BI is an ideal platform for conducting historical analysis of OTT data. By analyzing historical data, businesses can identify patterns and trends, predict future performance, and make data-driven decisions.

Historical analysis of OTT data can be visualized using PowerBI dashboards and reports, designed to provide valuable insights and inform data-driven decisions with their general requirements such as:

Data source: Power BI supports a wide range of data sources, including Excel spreadsheets, .csv files, SQL databases, SharePoint lists, and more and should have access to the data that should be visualized in Power BI.

Data modeling: Before creating a dashboard or report, data needs be modelled in a way that is optimized for analysis. This includes defining relationships between tables, creating calculated columns and measures, and ensuring that the data is clean and structured.

Visualization design: The design of your visualizations should be based on the specific insights you want to communicate. It's important to choose the right visualizations, such as charts, tables, and maps, and to use color, font, and layout effectively to highlight key data points.

User experience: Dashboard or report should be designed with the end user in mind. This means making it easy to navigate, interactive, and intuitive. Also consider the device on which the dashboard or report will be viewed, such as a desktop computer or mobile device.

Performance and security: Power BI dashboards and reports need to be designed for optimal performance, which includes considerations such as data refresh frequency, query performance, and minimizing data load times. Additionally, you need to ensure that the data is secure and that only authorized users can access it.

By following these requirements, you can create an effective Power BI dashboard or report that solves common problems such as limited access to data, difficulty in visualizing complex data, and inefficient decision-making processes.

2   PROBLEM STATEMENT

The rise of over-the-top (OTT) streaming services has dramatically changed the entertainment industry. As the OTT market continues to grow, so does the amount of data generate by these services. In order to make informed business decisions and drive growth and success in the competitive entertainment industry, it is critical to be able to effectively analyze and visualize this data.

However, the vast amount of data generated by an OTT platform like Netflix can be difficult to manage and analyze without the right tools and techniques. Data cleaning and transformation using the Query Editor in Power BI is a key step in this process, as it allows for the removal of irrelevant or erroneous data and the creation of clean and structured datasets.

The visualization of this data using Power BI is equally important, as it allows for the identification of patterns, trends, and outliers that can inform business decisions. However, creating effective and informative visualizations can be challenging without a deep understanding of the platform and the data it generates.

Therefore, we can effectively use PowerBI to visualize and analyze data related to an OTT platform such as Netflix using the Query Editor in Power BI for data cleaning and transformation and create informative and effective dashboards using Power BI's data visualization tools.

3   OBJECTIVES

- To develop a dashboard of historical analysis of OTT(Netflix) data.
- To Provide clear and concise insights.
- To identify patterns, trends, and outliers of data.
- To implement dashboards and reports based on the key performance metrics.

4   PROPOSED FEATURES

1. **Data Analysis and Visualization**
   - Interactive visualization

- Data filtering and drill-down

- Historical analysis

- Content analysis

- Predictive analysis

2. **User Engagement Metrics and Feedback**

- User engagement metrics

- User feedback analysis

5   PROPOSED SOFTWARE AND PLATFORMS

- **Front-End:** Power BI Query Editor for data preparation and cleaning.

- **Back-end:** SQL in Power BI for data modeling and analysis.

- **Data sourcing:** Kaggle for high-quality datasets.

- **Data storage:** OneDrive for cloud-based file hosting and sharing.

- **Visualization tool:** Power BI for interactive visualizations and insights.

6   POWERBI ARCHITECTURE & LAYOUT

Power BI architecture is a cloud-based business analytics service that enables users to analyze and visualize data. It consists of several different components that work together to provide a powerful and scalable data analysis platform.
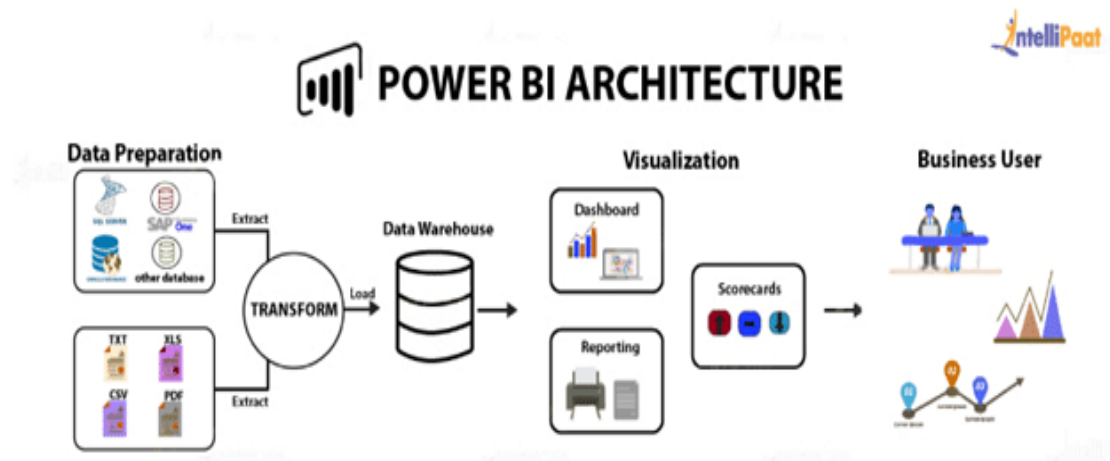
Figure 1

From the figure 1, the Power BI architecture can be divided into four main components:

Data Sources: Power BI supports a wide range of data sources, including cloud-based sources such as Azure SQL Database, SharePoint Online, and OneDrive for Business, as well as on-premises sources such as SQL Server and Oracle databases.

Data Transformation: Once data is brought into Power BI, it can be transformed and cleaned using Power Query, which is an intuitive data transformation tool. Power Query allows users to perform a range of transformations on data, including filtering, merging, and pivoting, among others.

Data Modeling: Power BI includes a powerful data modeling engine that enables users to create relationships between different tables and define hierarchies and calculations. The data modeling engine also supports the creation of calculated columns and measures, which can be used to perform complex calculations on data.

Data Visualization: Power BI provides a range of data visualization options, including charts, tables, maps, and gauges. Users can create custom dashboards and reports that display their data in a visually appealing and easy-to-understand format.
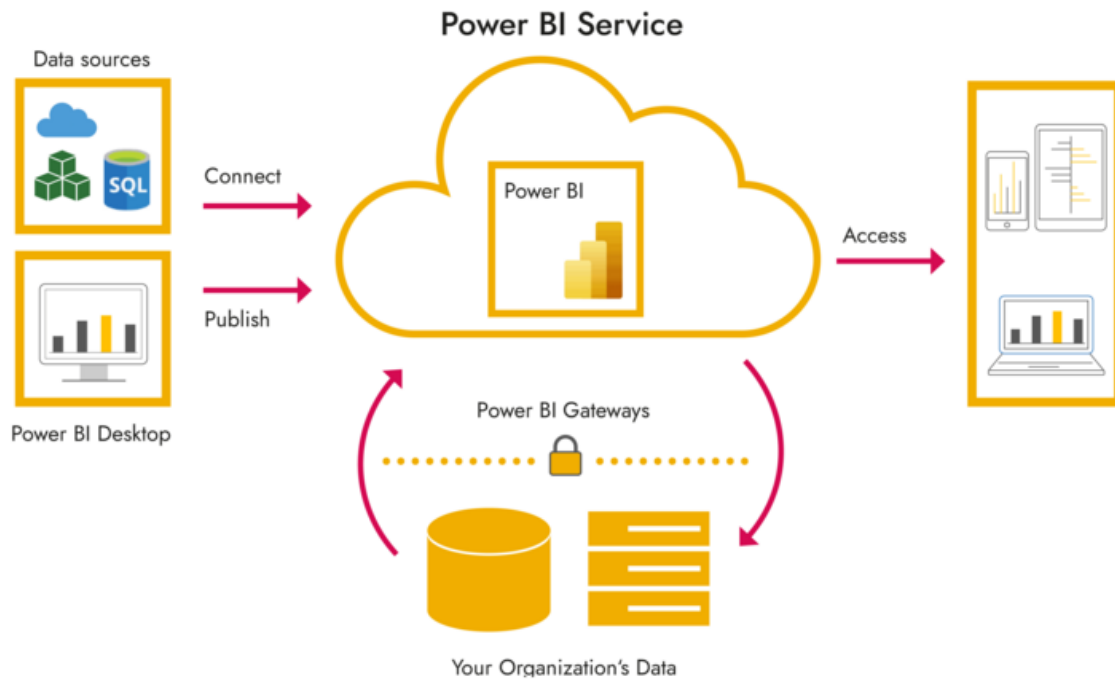
- PowerBI Service:

Figure 2

Power BI Service is an integral part of the Power BI ecosystem that facilitates the sharing and collaboration of reports and dashboards created using Power BI Desktop. As a web-based platform, it provides a centralized location to access and manage reports, collaborate with other users, and create insightful dashboards.

The Power BI Service is offered in three versions, namely, Free, Pro, and Premium, catering to the needs of various users. With its advanced features such as natural language Q&A and alerts, the Power BI Service enables users to gain valuable insights and take proactive actions based on the data.

Also referred to as Power BI.com, Power BI Workspace, Power BI Site, and Power BI Web Portal, this powerful tool offers a seamless experience for designing, implementing, and sharing data visualizations. As a steppingstone towards achieving data-driven decision-making, the Power BI Service empowers users to uncover hidden patterns, identify trends, and make informed decisions based on data.

Overall, the Power BI architecture is designed to be flexible, scalable, and easy-to-use, making it an ideal choice for businesses of all sizes looking to analyze and visualize their data.

- PowerBI End-End Diagram :

  A Power BI end-to-end diagram provides an overview of the entire process of creating a Power BI report, from data source acquisition to report publication. It typically includes steps such as data source identification, data cleaning and transformation, data modeling, report creation, and report publishing. The diagram helps to visualize the different components involved in the process and how they are interconnected. By following the end-to-end diagram, users can ensure that they are properly executing each step in the process and can more easily identify any potential issues or areas for improvement.
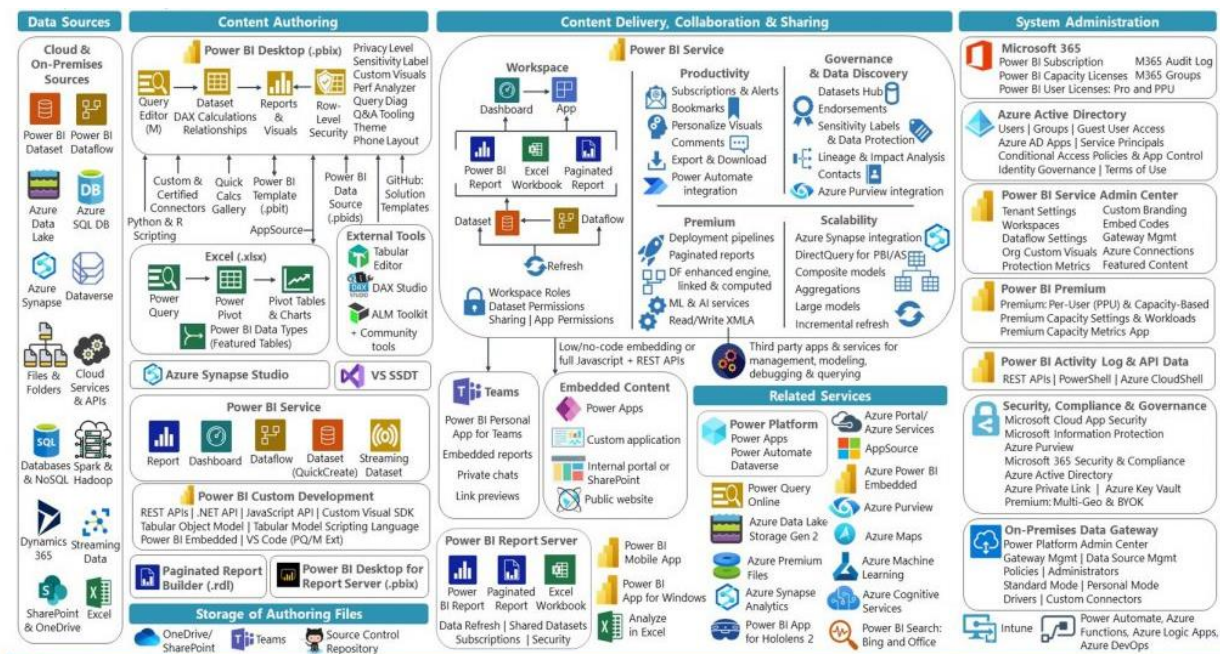

Figure 3

7 IMPLEMENTATION

In this section, the primary focus is on the various approaches used during the project creation. The main objective was to concentrate on creating dashboards, loading, extracting, and transforming data. The data analyzed in this project was Netflix data, which includes six sheets containing information on Movies, Genres, IMDB Score, IMDB Votes, and Production Countries. To visualize the data and identify trends, Power BI was used as one of the best data visualization tools available.

7.1 Data Acquisition

The data used for this project was collected from the Kaggle dataset "The Ultimate Netflix TV Shows and Movies Dataset" by user "The Devastator". The dataset contains information on over 7,000 TV shows and movies available on Netflix as of 2021.

To access the dataset, I visited the Kaggle website and searched for "Netflix dataset". The search results led me to "The Ultimate Netflix TV Shows and Movies Dataset". After reviewing the details of the dataset, I downloaded the CSV file containing the data.

The dataset included six different sheets: Movies, TV Shows, Genres, IMDb Score, IMDb Votes, and Production Countries. Each sheet contained relevant information for the project, such as the title, genre, and IMDb rating of each TV show or movie.

Before using the data in Power BI, I reviewed the dataset to ensure that it was clean and did not contain any missing or incorrect information. I also checked for duplicates and removed them as necessary. Overall, the data collection process was straightforward, and the dataset provided a solid foundation for the project.

7.2 Data Preparation

Import the dataset: To import the dataset into Power BI, the user needs to open a workspace, select the new toolbar, and then select the data flow option from the dropdown menu. The user should then define new tables and select the option to add new tables.
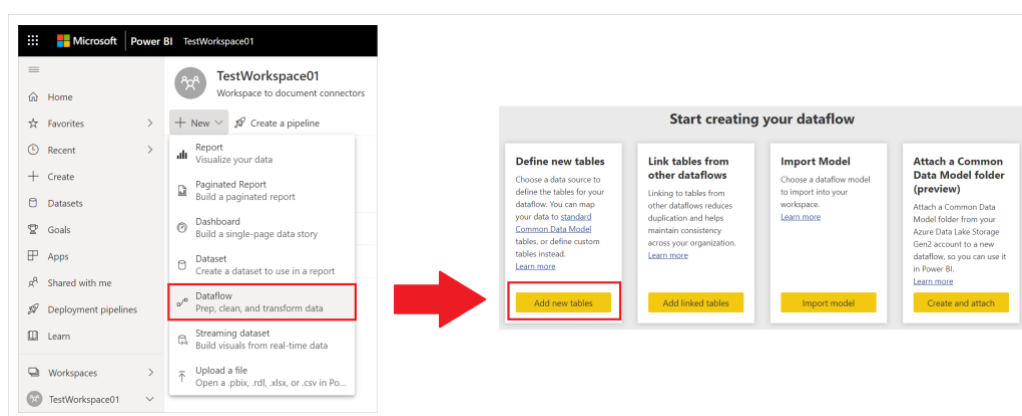

Figure 4

Next, the user should navigate to the Home tab and select "Get Data," followed by selecting the appropriate file type, such as CSV. It is important to ensure that the correct file type is selected to avoid any errors during the import process. By following these

steps, the user can successfully import the dataset into Power BI for analysis and visualization.
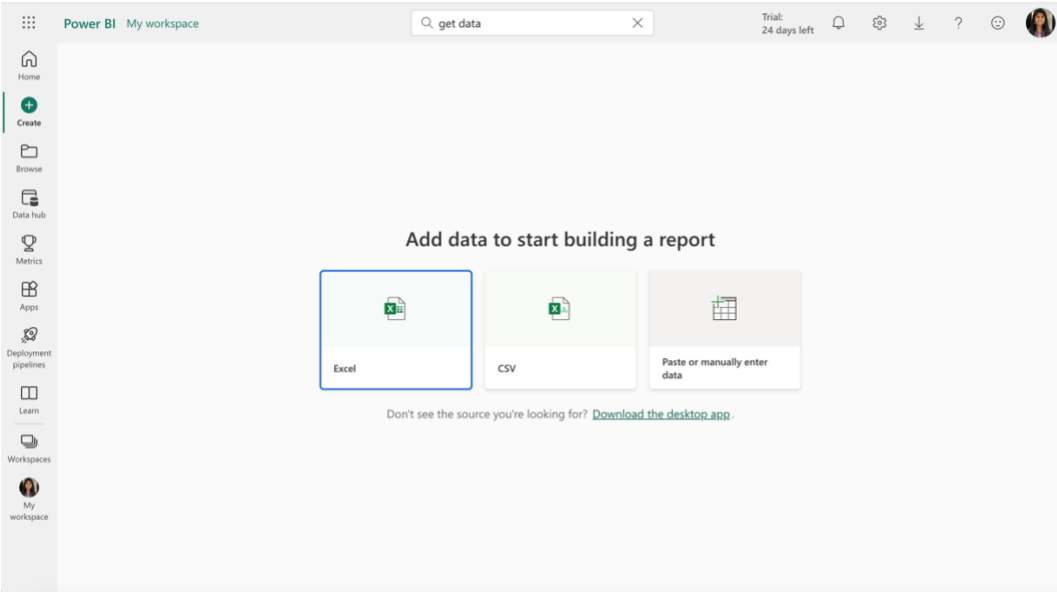

Figure 5

Figure 6

7.3   Data Transformation

Data transformation is a critical aspect of data analysis and involves converting raw data into a more suitable format for analysis. This process typically includes a series of operations, such as cleaning, filtering, aggregating, and structuring data to make it more useful for analysis. In many cases, the data obtained for analysis may be in a format that is not compatible with the analysis tools or may contain missing or inaccurate information that needs to be corrected. Data transformation enables

analysts to convert data into a format that can be analyzed effectively, making it easier to uncover insights and identify patterns. With the ever-increasing amount of data being generated in today's digital age, data transformation has become a critical skill for data analysts and scientists.
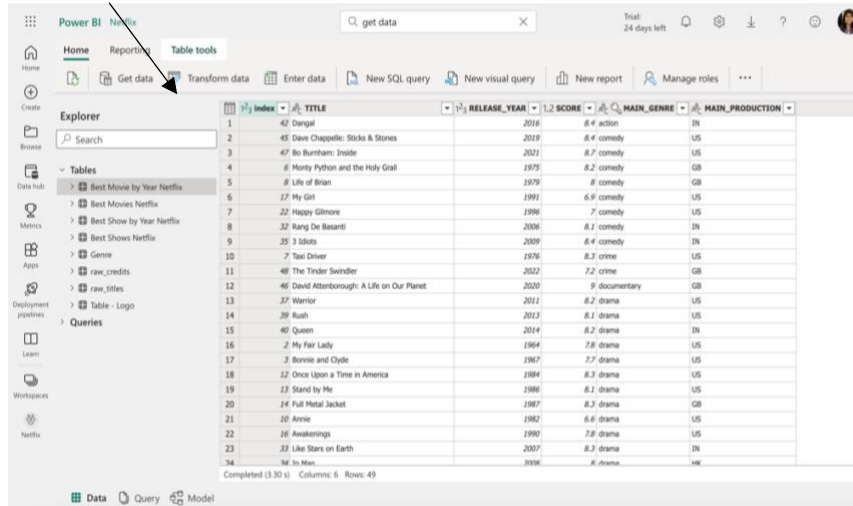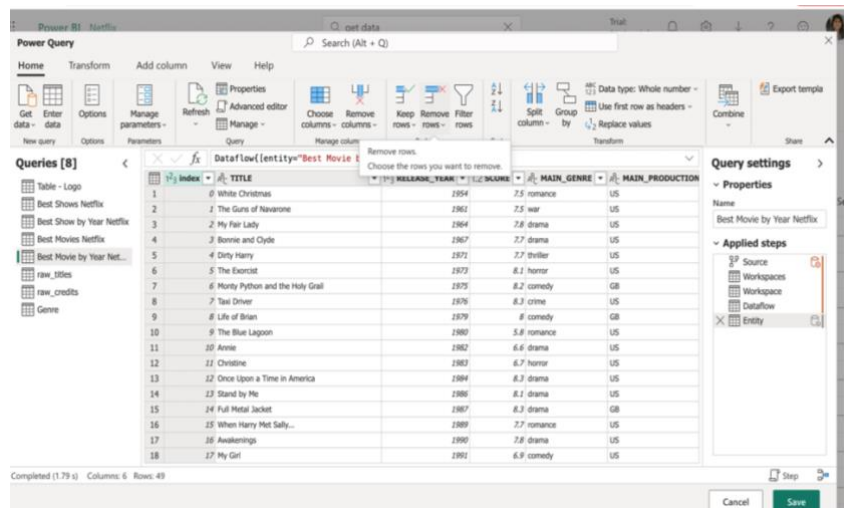

Figure 7

Figure 8

Replace missing data: The dataset may contain missing or null values that need to be addressed. One approach is to replace missing values with a default value such as "Null". This can be done by selecting the column header, clicking "Replace Values" from the Transform tab, and entering the default value.
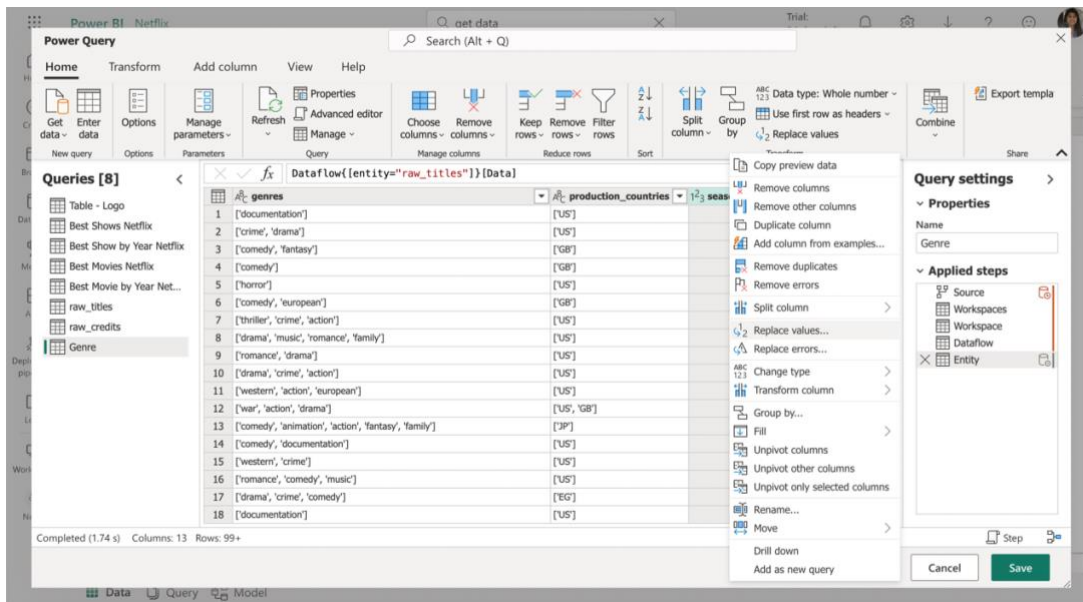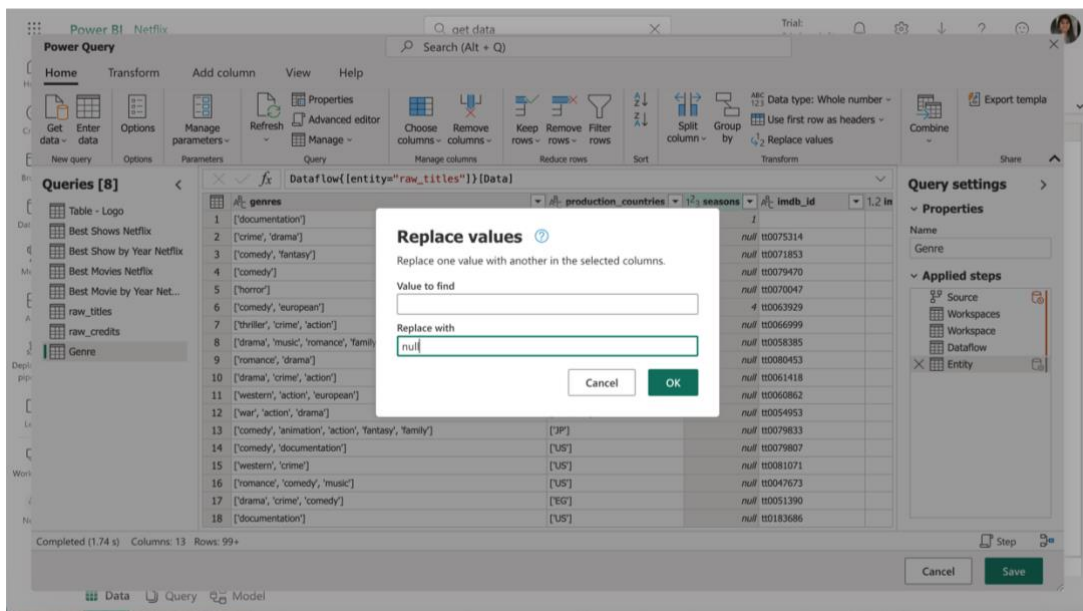
Figure 9

Figure 10

Merge related data: The dataset may contain related data that needs to be merged. For example, the "Genres" and "Production Countries "columns may contain multiple values separated by commas. These columns can be split into separate rows by selecting the column header, clicking "Split Columns" from the Transform tab, and selecting "By Delimiter".
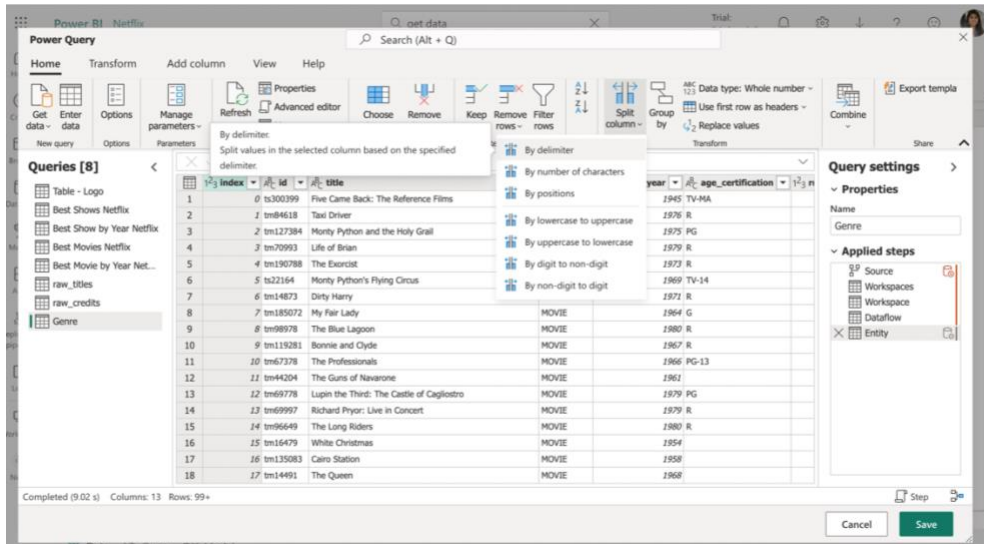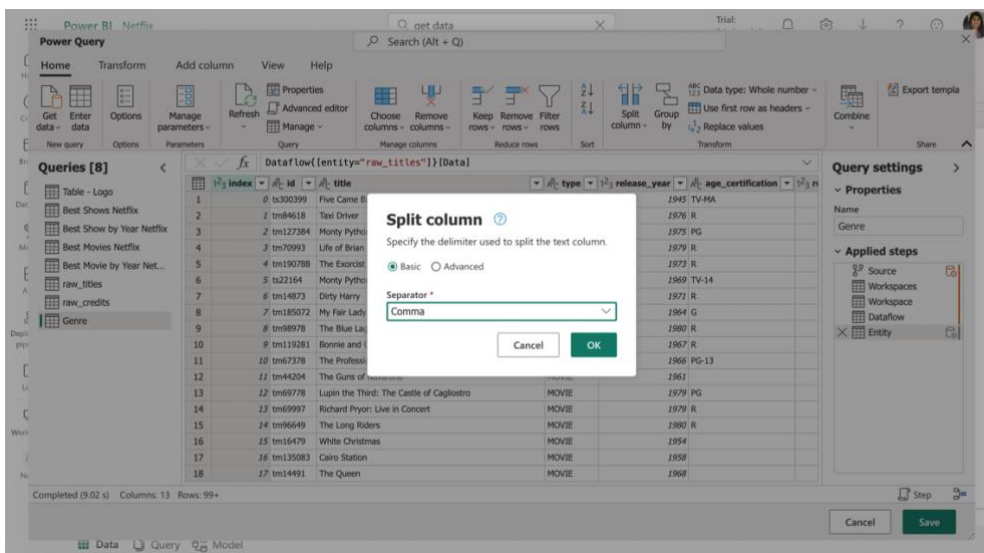
Figure 11

Figure 12

Remove unnecessary columns/rows: The dataset may contain columns/rows that are not needed for the analysis. These columns/rows can be removed by selecting the column header and clicking "Remove Columns"/ "Remove Rows" from the Modeling tab.
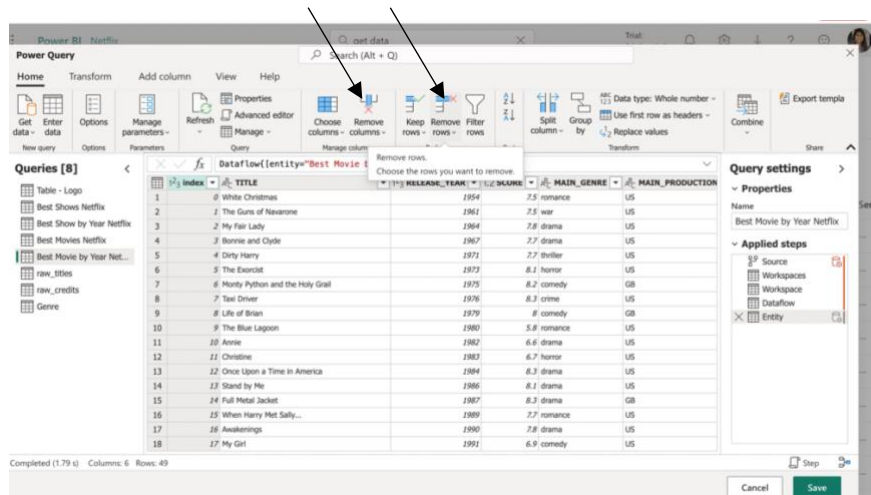
Figure 13

Trim columns: To trim columns in Power BI, first open the query editor window by selecting "Transform data" from the Home tab. Next, select the column or columns that need to be trimmed. Then, click on the "Transform" tab and select the "Trim" option. This will remove any leading or trailing spaces in the selected columns. It is important to note that this transformation is non-destructive, meaning it does not alter the original data in the data source, but rather creates a new trimmed version of the data to be used in subsequent analyses. Once the transformation is complete, the trimmed data can be loaded into a new table or merged with other data sources to create a more comprehensive data model for analysis.
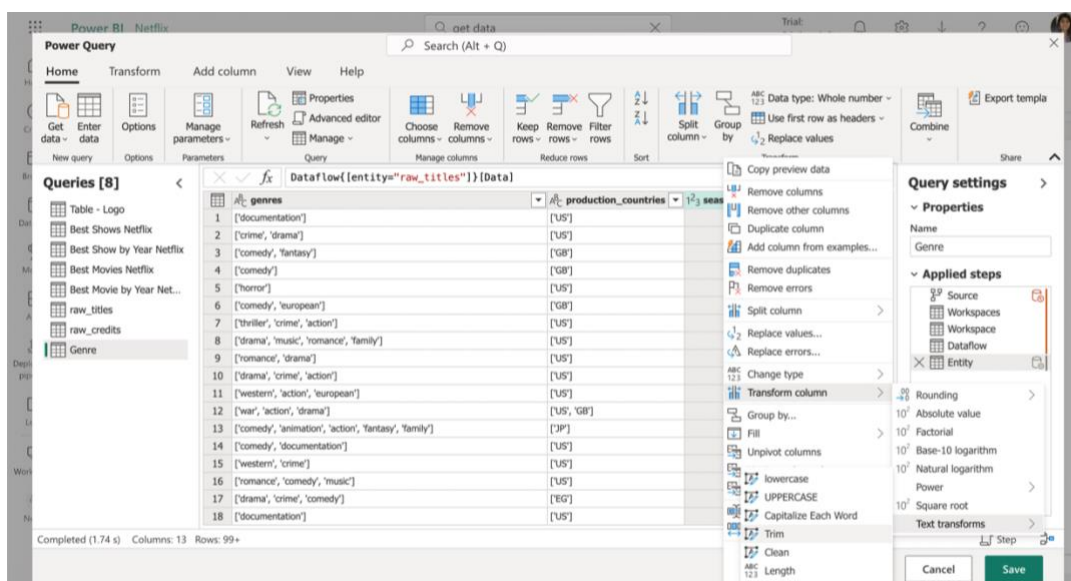

Figure 14

Filter data: The dataset may contain data that is not relevant to the analysis. This data can be filtered out by selecting the column header, clicking "Filter Rows" from the Transform tab, and selecting the desired filter criteria.

Rename columns: The dataset may contain columns with long or unclear names. These columns can be renamed by selecting the column header, clicking "Rename Columns" from the Modeling tab, and entering the desired name.

7.4   Data Analysis

After collecting and preparing the data, the next step is to analyze it to gain insights and identify trends. Here are some of the data analysis techniques that can be used on the Netflix dataset:

Descriptive analysis: This involves examining the dataset to obtain a summary of its main features. This includes the total number of movies and TV shows, the most common genres, ratings, etc.

Genre analysis: The dataset contains information on the genres of the movies and TV shows available on Netflix. Analyzing the distribution of genres can provide insights into the preferences of Netflix users. For instance, which genres are more popular, and which ones are not.

Ratings analysis: The dataset includes information on ratings from the International Movie Database (IMDB) and Netflix. Analyzing these ratings can provide insights into the quality of the movies and TV shows available on Netflix. For instance, which movies have the highest ratings, and which ones have the lowest.

Country analysis: The dataset also includes information on the countries where the movies and TV shows were produced. Analyzing this data can provide insights into the diversity of content available on Netflix.

By conducting these types of analysis on the Netflix dataset, we can gain valuable insights into the preferences of Netflix users, the quality of the content available, and the trends in the types of content being added over time.

- **Data Model:**

In Power BI, creating relationships is a crucial step in data modeling that involves linking tables based on common fields or keys to establish a relationship between them. This process helps to combine data from different tables and create a unified view for analysis and visualization. To create relationships, users need to identify the fields that can connect tables, define the type of relationship, and ensure that the data is aligned properly. By doing this, users can easily navigate and explore complex datasets, gain deeper insights, and make more informed decisions. The process of creating relationships involves adding connections between table parameters as shown in below figure.
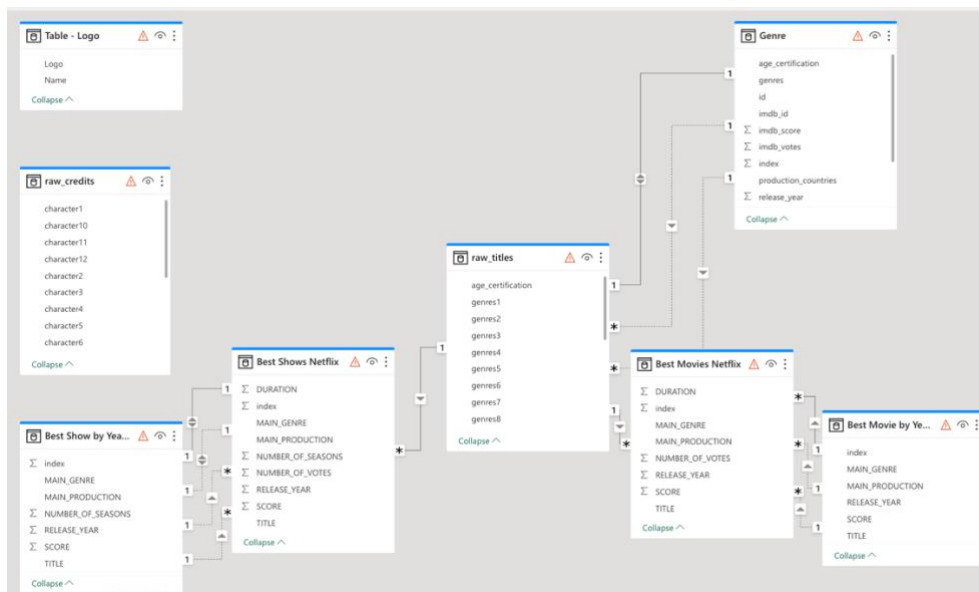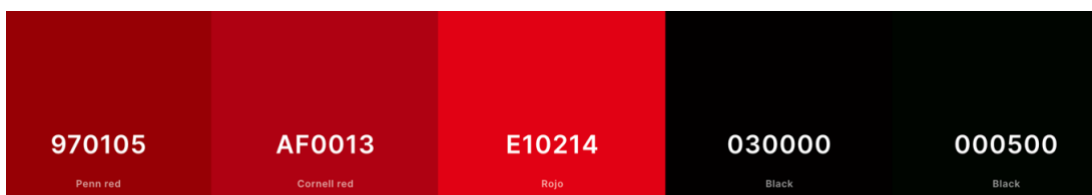

Figure 15

- **Color Palette:**

Figure 16

The color palette used in this project is picked from the Netflix Logo using Coolors.co. An effective color scheme can help users focus on important information and reduce

distractions. This project's color palette is highly relevant to the topic of Netflix, allowing users to quickly identify with the brand and gain a better understanding of the data presented.

- **Queries:**

  SQL queries are used for few data modelling to find distinct characters, distinct countries, distinct countries by Year, distinct countries by Type, and Genres in the Netflix Data.

  Distinct Characters:

```
SELECT DISTINCT distinct_characters
FROM (
  SELECT DISTINCT REPLACE(name, ' ', '') AS distinct_characters
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character1
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character2
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character3
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character4
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character5
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character6
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character7
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character8
  FROM raw_credits
  UNION ALL
  SELECT DISTINCT character9
  FROM raw_credits
) AS distinct_characters_table
WHERE distinct_characters IS NOT NULL AND distinct_characters <> '';;
```

  Distinct Countries:

```
SELECT DISTINCT production_countries1 AS unique_production_countries
FROM raw_titles
WHERE production_countries1 IS NOT NULL
UNION
SELECT DISTINCT production_countries2 AS unique_production_countries
FROM raw_titles
WHERE production_countries2 IS NOT NULL
UNION
SELECT DISTINCT production_countries3 AS unique_production_countries
FROM raw_titles
```

WHERE production_countries3 IS NOT NULL;

Distinct Countries by Year:

SELECT production_country, release_year, type, COUNT(*) AS num_titles
FROM (
  SELECT production_countries1 AS production_country, release_year, type
  FROM raw_titles
  WHERE production_countries1 IS NOT NULL AND type IN ('movie', 'show')
  UNION ALL
  SELECT production_countries2 AS production_country, release_year, type
  FROM raw_titles
  WHERE production_countries2 IS NOT NULL AND type IN ('movie', 'show')
  UNION ALL
  SELECT production_countries3 AS production_country, release_year, type
  FROM raw_titles
  WHERE production_countries3 IS NOT NULL AND production_countries3 <> '' AND type IN ('movie', 'show')
) AS countries
WHERE release_year IS NOT NULL
GROUP BY production_country, release_year, type
ORDER BY production_country, release_year, num_titles DESC;

Distinct Countries by Type:

SELECT production_country, type, COUNT(*) AS num_titles
FROM (
  SELECT production_countries1 AS production_country, type
  FROM raw_titles
  WHERE production_countries1 IS NOT NULL AND type IN ('movie', 'show')
  UNION ALL
  SELECT production_countries2 AS production_country, type
  FROM raw_titles
  WHERE production_countries2 IS NOT NULL AND type IN ('movie', 'show')
  UNION ALL
  SELECT production_countries3 AS production_country, type
  FROM raw_titles
  WHERE production_countries3 IS NOT NULL AND production_countries3 <> '' AND type IN ('movie', 'show')
) AS countries
GROUP BY production_country, type
ORDER BY production_country, num_titles DESC;

Genre:

SELECT DISTINCT genre
FROM (
  SELECT genres1 AS genre FROM raw_titles WHERE genres1 IS NOT NULL
  UNION
  SELECT genres2 AS genre FROM raw_titles WHERE genres2 IS NOT NULL
  UNION
  SELECT genres3 AS genre FROM raw_titles WHERE genres3 IS NOT NULL
  UNION
  SELECT genres4 AS genre FROM raw_titles WHERE genres4 IS NOT NULL
  UNION

```
SELECT genres5 AS genre FROM raw_titles WHERE genres5 IS NOT NULL
UNION
SELECT genres6 AS genre FROM raw_titles WHERE genres6 IS NOT NULL
UNION
SELECT genres7 AS genre FROM raw_titles WHERE genres7 IS NOT NULL
UNION
SELECT genres8 AS genre FROM raw_titles WHERE genres8 IS NOT NULL
) AS all_genres;
```

7.5   Data Visualization

One of the main objectives of this project was to create visually appealing dashboards that could help identify trends and insights from the data. To achieve this, Power BI was used to create a range of interactive visualizations, including:

Simple image visualization for Netflix Logo, where it is taken from the Logo table by logo in Image URL and Name in Alt text.

Figure 17

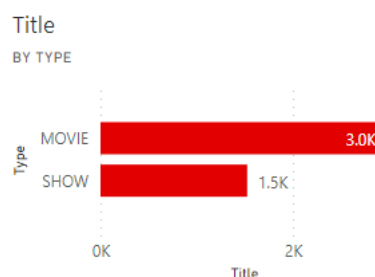Stacked Bar Chart for the total number of titles by Movies and TV Shows.

Figure 18

Card Visual used for Total number of Titles, Average Runtime and Average IMDB Votes.

Total No. of Titles	Avg. Runtime(in ...	Avg. IMDB Votes
4501	77.36	28.40K

Figure 19   *Figure 20*   *Figure 21*

Gauge for predicting the average IMDB score of total Netflix data.

Avg. IMDB Score

6.54

0.00   10.0

Figure 22

Tree Map to predict the Titles which has greater than 70,000 votes, where Inception has most votes.

No. of Votes by Title

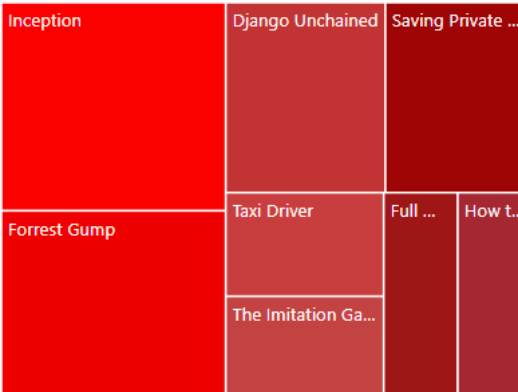| Inception	Django Unchained	Saving Private ...
Forrest Gump	Taxi Driver	Full ... / How t..
	The Imitation Ga...	

Figure 23

21

Line and Stacked column chart for predicting the content added per year by type i.e, movie/tv show.
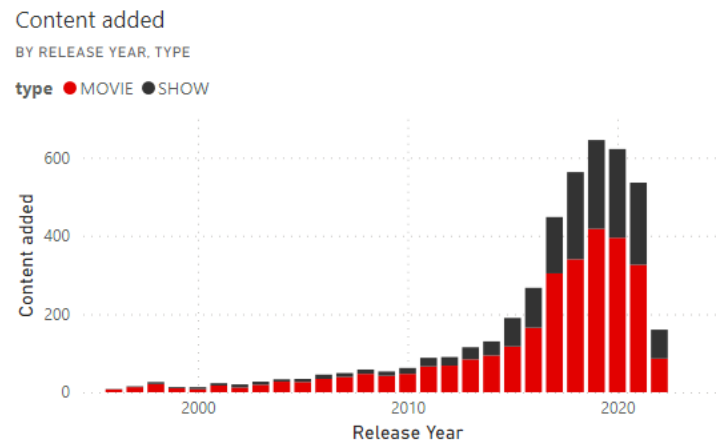

Figure 24

Bar Chart for predicting the Age certification by type i.e, movie/tv show
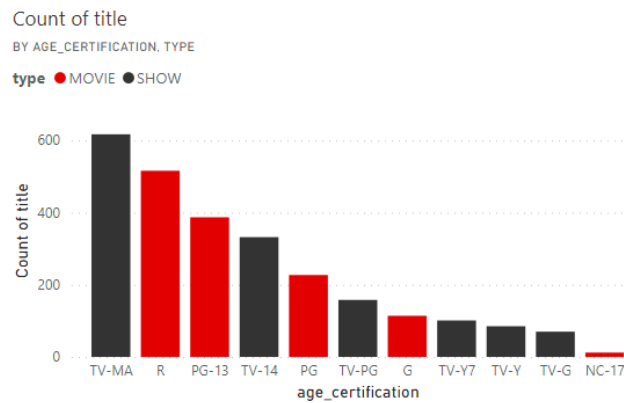

Figure 25

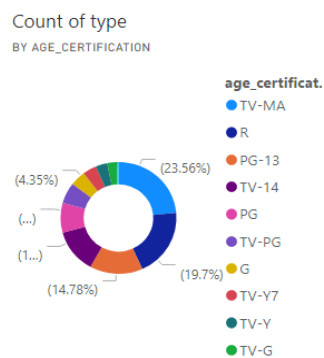Doughnut chart for the total count of age certifications.

Figure 26

22

Clustered Column charts to show the number of movies and TV shows on Netflix by country and year.
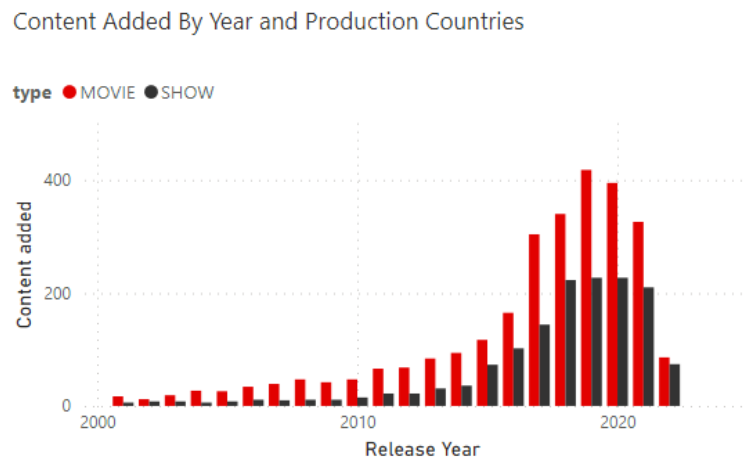

Figure 27

Pie Chart and Slicer to show the distribution of genres in the Netflix library.
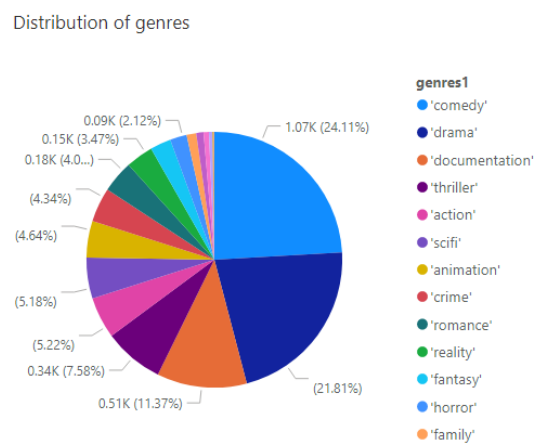

Figure 28

Scatterplots to explore the relationship between IMDb ratings and the number of IMDb votes for movies and TV shows.
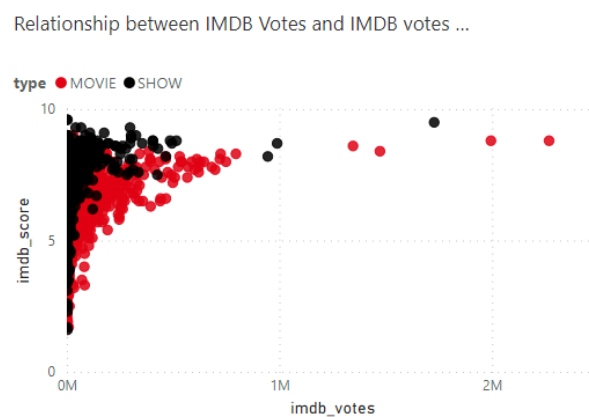

Figure 29

23

Word clouds to highlight the most common words in the titles and descriptions of movies and TV shows.
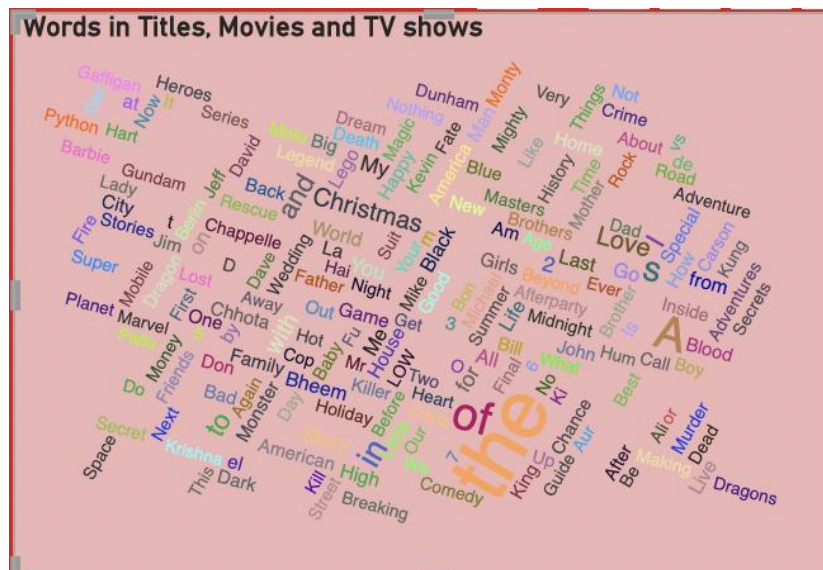
Figure 30

These visualizations allowed us to gain a better understanding of the Netflix library and identify key trends and insights. For example, we could see that the number of movies and TV shows on Netflix has been steadily increasing over the years, with the United States and India having the largest number of titles. We could also see that the most common genres on Netflix are international, dramas, comedies, and documentaries, and that there is a positive correlation between IMDb ratings and the number of IMDb votes.

Overall, the data visualization using Power BI allowed us to effectively communicate the findings of our data analysis and provide valuable insights for future decision-making.

7.6   Results and Insights

Please find the link to the dashboard site where you can access the visualization results.

Link: https://app.powerbi.com/reportEmbed?reportId=927aba2d-0a74-4f16-a91c-c46670fb01be&autoAuth=true&ctid=48ec3bf8-d165-4eab-bbee-f8d5307f46e1

By exploring the aspects of the data, we can gain a more comprehensive understanding of the underlying dynamics and make more informed decisions. Additionally, we can consider incorporating user feedback and suggestions to further enhance the value and usability of the dashboard.

Movies and Shows by Country

●MOVIE ●SHOW

Based on the data analysis and visualization, some key results and insights from the Netflix dataset are:

- The majority of the content on Netflix is movies, followed by TV shows.
- The most common genres on Netflix are Drama, Comedy, and Documentary.
- The countries with the most production of content on Netflix are the United States, India, and the United Kingdom.

- The average IMDb score on Netflix is 6.54, and average IMDb votes on Netflix is 28.40k.

- Netflix has been consistently adding more content to its platform over the years.

- The top 10 rated movies on Netflix are mainly drama and thriller genres, while the top 10 rated TV shows are mainly crime and drama genres.

- There is a positive correlation between the IMDb score and the number of IMDb votes, indicating that the more votes a title receives, the higher its score tends to be.

- These insights can be useful for Netflix to make decisions on what type of content to produce and acquire, which genres to focus on, and which countries to target for production. The insights can also be useful for users to find content they may be interested in based on their preferred genres and IMDb scores.

8 DEPLOYMENT

During the development of the Netflix data visualization project using Power BI, I utilized Power BI Pro service to create and share dashboards and reports with my supervisor for review and feedback. This process allowed for seamless collaboration and allowed my supervisor to provide input and suggestions throughout the development process. While the project did not have a traditional deployment and production environment, the use of Power BI Pro service allowed for efficient sharing and communication during the development phase, ensuring that the final product met the necessary requirements and provided valuable insights to the end-users.

9 FUTURE WORK

This Netflix data visualization project using PowerBI has proven to be a valuable tool for gaining insights into viewership and content trends. It has also highlighted the importance of using data visualization tools like Power BI to analyze and present complex data in a clear and concise manner.

To further enhance the project, future work can focus on expanding the analysis to compare data from other OTT platforms such as Amazon Prime, Hotstar, Aha, etc. This will provide a more comprehensive overview of the OTT industry trends, which can help identify growth opportunities for different platforms.

Another area of future work would be to incorporate sales and revenue data from Netflix and other OTT platforms to identify revenue trends and forecast future revenue. This will help evaluate the financial performance of the company and its competitors and understand the driving factors behind their success.

In addition, real-time data updates could be incorporated into the project to provide more up-to-date insights and trends. Advanced analytics techniques like predictive modeling, machine learning, and artificial intelligence could also be applied to the data to gain more sophisticated insights. These improvements will further strengthen the project's ability to provide valuable insights into the OTT industry.

10 CONCLUSION

Working on this project has been a rewarding experience as it has provided valuable insights into the viewership and content trends of Netflix data using the PowerBI dashboard. I always wanted to experience the challenges of being the data analyst. Therefore, Through the process of designing and implementing the Power BI dashboard, I have gained a deeper understanding of the importance of data visualization tools in analyzing and presenting complex data in a clear and concise manner.

This project has also served as a steppingstone for further analysis of data of the OTT industry, as the insights gained can be applied to compare data from other platforms such as Amazon Prime, Hotstar, Uber, Lyft and many more. This will provide a more comprehensive overview of the trends in the industry, which can be used to identify new opportunities for growth and development.

In conclusion, this project has highlighted the power of data visualization tools in making sense of complex data and has opened new avenues for future research and analysis in the OTT industry. I look forward to continuing to build on these insights and developing more sophisticated analytics techniques to gain even deeper insights into the world of OTT.

REFERENCES

- https://blog.bismart.com/en/data-visualization-with-power-bi
- https://medium.com/mobiotics/data-reporting-visualization-for-ott-6345c50abad5
- https://www.projectpro.io/article/power-bi-microsoft-projects-examples-and-ideas-for-practice/533#mcetoc_1gd0kj9qio4
- https://intellipaat.com/blog/power-bi-project-ideas/#no3
- https://www.ijrar.org/papers/IJRAR1CNP012.pdf
- https://bimentalist.com/2018/07/30/analyze-your-netflix-data-in-power-bi/
- https://www.kaggle.com/datasets/thedevastator/the-ultimate-netflix-tv-shows-and-movies-dataset
- https://www.kaggle.com/datasets/shivamb/netflix-shows
- https://studentwork.prattsi.org/infovis/visualization/a-glance-at-netflixs-database/
- https://powerbi.tips/2021/01/power-bi-end-to-end-architecture/