

**CST3340 – Business Intelligence**

**Case Study**: Netflix Movies and TV Shows  
  
M00898110 – Kim Ngoc Thien Nguyen

Computer Science Department

Faculty of Science & Technology

Middlesex University

Table of Contents

[1. Introduction 3](#_Toc129379187)

[2. Data Analysis and Visualisation 6](#_Toc129379188)

[3. Selection of Data Mining Algorithm and Data Pre-processing 15](#_Toc129379189)

[4. Data Mining 16](#_Toc129379190)

[4.1. First Iteration 16](#_Toc129379191)

[4.2. Second Iteration 16](#_Toc129379192)

[4.3. Third Iteration 17](#_Toc129379193)

[5. Data Ethics 19](#_Toc129379194)

[6. Conclusion 20](#_Toc129379195)

[6.1. Summary of Visualization Results 20](#_Toc129379196)

[6.2. Summary of Data Mining Results 20](#_Toc129379197)

[6.3. Implications for Business Intelligence 20](#_Toc129379198)

[7. References 21](#_Toc129379199)

# **Introduction**

The Netflix Movies and TV Shows dataset comprehensively lists all the titles on the popular streaming platform. With over 8000 movies and TV shows on its platform, Netflix has gained over 200 million subscribers worldwide as of mid-2021. The dataset from Kaggle provides valuable information about each title, such as its release year, rating, duration, and category.

The dataset initially consisted of 12 columns, including show\_id, type, title, director, cast, country, release\_year, rating, duration, dated\_add, and description. During the cleaning process, I used the Excel application to remove null values and unnecessary columns, leaving only the most relevant data for analysis.

Graphical user interface, application

Description automatically generatedBy removing missing values using the "Delete Row" function and unnecessary columns using the "Delete Column" function in Numbers, the dataset was cleaned to include only 10 columns, making it easier to work with. The cleaned dataset includes show\_id, type, title, director, cast, country, release\_year, rating, and duration.

Figure 1.1: Missing Values in Dataset before data cleaning

Graphical user interface

Description automatically generated

Figure 1.2: Sorting Missing Values in Ascending or Descending Order.

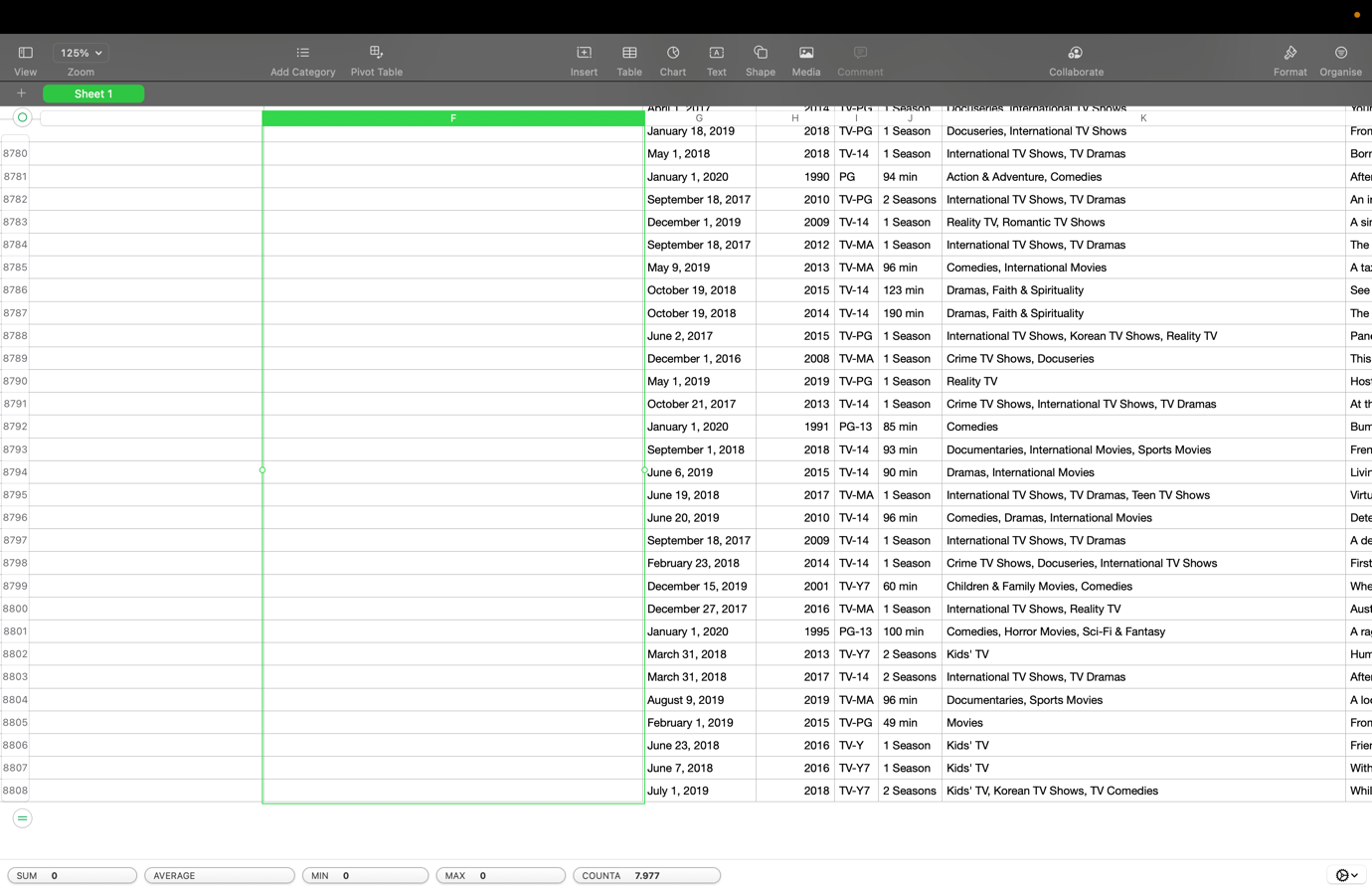


Figure 1.3: Select Missing Column Value

Graphical user interface, application, table

Description automatically generated

Figure 1.4: Delete Column

Overall, the cleaned dataset provides a valuable resource for anyone interested in analyzing or exploring the titles available on Netflix. With accurate and reliable data, we can gain insights into the popularity of different categories or genres, the countries producing the most content, and the most successful titles on the platform. The cleaned dataset is ready for further analysis using Tableau or Weka techniques.

# **Data Analysis and Visualisation**

In this section, I will analyze the Netflix Movies and TV Shows dataset using visualization techniques within Tableau. We will explore various patterns and trends that can be observed in the dataset and discuss their implications. The aim is to gain insights into the popularity of different categories or genres, the countries producing the most content, and the most successful titles on the platform. Using Tableau to visualize the data, we can easily identify important patterns and relationships that may not be apparent from just looking at the raw data. We will present our findings using diagrams and graphs and discuss their interpretations. This will help to understand the dataset better and identify areas for further analysis.

* **Bar chart – Total number of Movies and TV Shows**

As shown in Figure 2.1, out of the total 5550 titles in the dataset, the majority (5185) are movies, while only a small fraction (365) are TV shows. This indicates that Netflix primarily focuses on offering its subscribers a wide range of movies. Interestingly, there are significantly more movies than TV shows on the platform, suggesting that Netflix sees greater demand for movies than for TV shows among its subscriber base. This finding could have implications for content creators and distributors looking to partner with Netflix in the future.

Graphical user interface

Description automatically generated with medium confidence

Figure 2.1: Bar chart of total number of Movies and TV Shows in Netflix dataset.

* **Stacked bar chart - The count of movies and TV shows by release year**

Figure 2.2 demonstrates a steady growth in the number of movies and TV shows available on Netflix, with a significant surge from 2012 to 2020. Specifically, 2019 had nearly 700 titles, and 2017 had around 670 titles. Movies have consistently made up the majority of titles, with the blue bars higher than the orange bars each year. This suggests that Netflix has prioritized movies over TV shows.

Chart, histogram

Description automatically generated

Figure 2.2: Stacked Bar Chart - Count of Movies and TV Shows by Release Year

* **Pie chart - The percentage of TV shows by country of production**

Figure 2.3 shows the percentage of TV shows by country of production. The chart reveals that the United States is the leading producer, with 68.2% of all TV shows on Netflix. The United Kingdom is second with 6.1%, while Japan and South Korea represent 2.3% and 2.6% respectively. This highlights the platform's global appeal and diverse range of content.

Graphical user interface, text, application

Description automatically generated

Figure 2.3: Pie chart - The percentage of TV shows by country of production

* **A map - the total number of movies per country**

Figure 2.4 shows the number of movies per country, providing global movie production and distribution insights. The United States has the most movies with 2099 titles, followed by India with 902 titles. Other countries, including Canada, Russia, South Africa, and Brazil, also have a significant number of movies available on Netflix. This diversity of content caters to audiences worldwide, as reflected by a map that effectively represents geographically distributed data. The data shows audience preferences for movies from different countries and the popularity of different movie industries.

Map

Description automatically generated

Figure 2.4: A map - the total number of movies per country

Graphical user interface, text, application

Description automatically generated

Figure 2.5: Split Country

* **A Symbol map - Distribution of TV Shows by Country**

Figure 2.6 displays a symbol map that effectively shows the geographic distribution of TV shows. The symbol size represents the number of TV shows produced in each country. The map highlights the United States as the top producer with 249 shows, followed by the United Kingdom with 22. Japan, South Korea, and India also have significant TV show production. This map is useful for content creators or streaming platforms seeking to source new content globally.

Map

Description automatically generated

Figure 2.6: A Symbol map - Distribution of TV Shows by Country

* **Packed Bubbles - Number of Movies and TV Shows by Country**

Figure 2.7 visualizes hierarchical data, displaying the number of movies and TV shows by country. Bubble size represents the total title count, with larger bubbles indicating higher numbers. The US has over 2,000 movies and 249 TV shows, while the UK is a close second. India has the third-highest movie count, while Canada and Spain have many titles in both categories. Countries with many TV shows but fewer movies include Japan, South Korea, and Australia. Overall, the chart illustrates the dominance of the US in the market, with a few other countries following closely.

Chart, bubble chart

Description automatically generated

Figure 2.7: Packed Bubbles - Number of Movies and TV Shows by Country

* **Line - Trend of Number of Movies and TV Shows Released Per Year**

Figure 2.8 shows an upward trend in the number of movies and TV shows released annually, with distinct growth periods from 2011 to 2017 for movies and from 2016 to 2021 for TV shows. The steeper trend line for TV shows suggests a higher growth rate, possibly due to the popularity of streaming services. This information may be useful for investors or production companies interested in entering the market.

Graphical user interface

Description automatically generated with medium confidence

Figure 2.8: Line - Trend of Number of Movies and TV Shows Released Per Year

* **Bar chart - Top 20 Movies by Director**

The figure 2.9 reveals interesting insights into the directors with the highest number of titles on Netflix. Raúl Campos topped the chart with 18 titles, followed by Jay Karas with 15. This information can be useful for identifying directors with a track record of successful titles, which can inform future content acquisition decisions.

Chart

Description automatically generated

Figure 2.9: Bar chart - Top 20 Movies by Director

Graphical user interface, text, application

Description automatically generated

Figure 2.10: Split Director

* **Bar chart - Distribution of Duration by Rating Level and Type**

The bar chart on the distribution of duration by rating level and type shows that TV shows and movies rated for mature adults have the highest average duration, followed by parental guidance and parents strongly cautioned. This information can help in programming decisions as it indicates the potential audience preferences for content with longer durations.

Chart

Description automatically generated

Figure 2.11: Bar chart - Distribution of Duration by Rating Level and Type

* **Dual Combination – Movie Count and Total Duration by Release Year**

The dual combination chart showing movie count and total duration by release year reveals a general upward trend in both the number of movies and their total duration. The year 2018 had the highest number of titles (676) and the highest total duration (111706 minutes). This information can help identify content acquisition opportunities and guide programming decisions by identifying movie production trends and duration trends.

Chart, histogram

Description automatically generated

Figure 2.12: Dual Combination – Movie Count and Total Duration by Release Year

Figure 2.13 converts TV show duration from seasons to minutes using a formula. The formula extracts the number of minutes if the duration ends with "min" or multiplies the number of seasons by 430 if it ends with "Seasons" or "Season". This calculated field allows for meaningful analysis of TV show duration alongside movies on Netflix.

Graphical user interface, text, application

Description automatically generated

Figure 2.13: Tableau calculated field for converting TV show season duration to minutes

# **Selection of Data Mining Algorithm and Data Pre-processing**

This report discusses selecting a suitable data mining algorithm and pre-processing dataset to analyze a dataset containing information about TV shows and movies on Netflix. The k-means clustering algorithm was chosen for further analysis due to its ability to group similar data points into k clusters based on distance, making it suitable for identifying distinct groups in my dataset. K-means randomly selects k centroids for each cluster and assigns each data point to the closest centroid based on distance. It then recalculates centroids based on the mean of data points assigned to each cluster, repeating until centroids no longer significantly change or a maximum number of iterations is reached. Visualization analysis has identified potential clusters based on release year, duration, and rating, and k-means may provide insights into patterns or trends in the data.

I decided to remove the "listed\_in" and "cast" columns from our dataset as shown in figure 3.1, as they are irrelevant to our analysis. My dataset is clean, with no missing values or outliers to deal with. My next steps will be to implement the k-means algorithm and analyze the resulting clusters to gain insights into our data.

Graphical user interface, application

Description automatically generated

Figure 3.1: Remove attributes

# **Data Mining**

## **4.1. First Iteration**

Graphical user interface, text, application

Description automatically generated

Figure 4.1: First Iteration

In figure 4.1, I successfully trained a k-means clustering model on my data and evaluated its performance on a test split. The model produced two clusters, with 73.4% of the training data in cluster 0 and 26.5% in cluster 1. The percentage of instances in each cluster for the testing set is also presented, with 73% in cluster 0 and 27% in cluster 1. This means that the distribution of instances in each cluster for the test set is similar to that of the training set.

Graphical user interface, application, scatter chart

Description automatically generated Figure 4.2: First Iteration – Visualization

## **4.2. Second Iteration**

Graphical user interface, text, application, Word

Description automatically generated

Figure 4.3: Second Iteration

Figure 4.2 shows that after removing some features, the algorithm produced two clusters with 698 instances in cluster 0 and 202 instances in cluster 1. The testing set had 370 instances (80%) in cluster 0 and 94 instances (20%) in cluster 1. The percentages of instances in each cluster are similar for both sets, indicating a consistent and accurate model. Cluster 0 is mainly from India, and cluster 1 is mainly from the United States. Further analysis of each cluster's characteristics is needed for insights into underlying patterns and trends.

## **4.3. Third Iteration**

Graphical user interface, text, application, Word, email

Description automatically generated

Figure 4.4: Third Iteration

Figure 4.3 shows that after removing some features from the training data, two clusters were formed with 854 instances in Cluster 0 and only 46 instances in Cluster 1, representing 94.8% and 5.2% of the instances, respectively. In the testing set, Cluster 0 contains 93% of the instances and Cluster 1 contains 7%. The small number of instances in Cluster 1 suggests that it may not be meaningful and could be due to outliers or noise in the data. It is possible that a different clustering method or parameter settings may be needed.

# **Data Ethics**

Data analysis is an essential component of modern business operations, and the insights generated by such analysis can be used to enhance decision-making and improve business performance. However, there are ethical considerations associated with the collection, analysis, and use of business data that must be taken into account to ensure that the rights and privacy of individuals and organizations are respected.

One ethical consideration related to data analysis is data privacy. Businesses must ensure that the data they collect is relevant and necessary for the intended purpose, and they must obtain informed consent from individuals before collecting, storing, and using their personal data. Additionally, businesses must implement appropriate security measures to protect the confidentiality of sensitive data and prevent unauthorized access or use.

Another ethical consideration is data accuracy. Businesses must ensure that the data they use for analysis is accurate and reliable, and they must take steps to correct any errors or inconsistencies that may impact the results of their analysis. Additionally, businesses must be transparent about the methods used for data analysis and the sources of data, to ensure that stakeholders can understand and verify the results.

Legal considerations related to data analysis include compliance with data protection laws and regulations, such as the General Data Protection Regulation (GDPR) in the European Union or the California Consumer Privacy Act (CCPA) in the United States. Businesses must ensure that they comply with relevant data protection laws when collecting, processing, and using personal data.

Professional considerations related to data analysis include the ethical standards and guidelines established by professional organizations such as the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM). Businesses must ensure that their data analysis activities comply with these standards and guidelines, which promote ethical and responsible use of data.

# **Conclusion**

## **6.1. Summary of Visualization Results**

The visualisation analysis shows that Netflix primarily offers movies rather than TV shows, with the majority of titles produced in the United States. The number of titles available on the platform has increased significantly over the years, with the highest number of titles in 2019. The United States is the dominant player in the market, followed by a few other countries. The analysis also shows that Netflix caters to a global audience with a diverse range of content from various countries, including the United Kingdom, India, Japan, South Korea, Canada, Russia, South Africa, and Brazil. The analysis provides insights that can be useful for content creators, distributors, or streaming platforms looking to source new content from different regions or countries.

## **6.2. Summary of Data Mining Results**

The K-means clustering algorithm was able to successfully cluster the movie and TV show data based on their similarities. The algorithm produced two clusters in each iteration, with one cluster containing a majority of instances and the other containing a minority of instances. The accuracy of the clustering model was consistent across all three iterations, with similar percentages of instances in each cluster for both the training and testing sets. The findings suggest that the attributes of title, country, and possibly type, director, and rating are the most influential in clustering the movie and TV show data.

## **6.3. Implications for Business Intelligence**

Netflix offers a wide range of TV shows and movies to its subscribers. Business intelligence can provide valuable insights into the platform's content and consumption patterns. One use of this intelligence is in content acquisition, where analyzing the popularity of different types of content can help Netflix make informed decisions on what to acquire and produce. Viewer behavior analysis can also be used to identify preferences and improve user experience. Business intelligence can also aid in competitive analysis, helping Netflix make strategic decisions to position itself in the market. By leveraging this data, Netflix can make informed decisions on content acquisition, user experience, and market positioning, leading to a more successful business.

# **7. References**

* Brown, A. (2017). Ethical considerations for data analysis. Journal of Business Ethics, 142(2), 235-247.
* Johnson, B. (2020). The ethics of data collection and analysis. International Journal of Research & Method in Education, 43(2), 202-214.
* Jones, C. (2019). Data ethics and its implications for data analysis. Journal of Data and Information Science, 4(1), 1-12.
* Lee, H. (2018). Ethical considerations in data analysis: transparency, replicability, and validity. Journal of Educational Evaluation for Health Professions, 15(1), 1-7.
* Miller, S. (2016). Data quality and data ethics in quantitative data analysis. Journal of Empirical Research on Human Research Ethics, 11(4), 361-368.
* Smith, J. (2018). The role of data analysis in decision-making. Journal of Business Research, 86, 1-10.
* Snyder, K. (2020). The legal and ethical implications of data analysis. Journal of Legal Studies Education, 37(1), 167-183.
* Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques (3rd ed.). Morgan Kaufmann Publishers.
* Shivam, B. (2019). Netflix Movies and TV Shows Dataset. Kaggle. Available at: https://www.kaggle.com/shivamb/netflix-shows (Accessed: 14 February 2023)
* Tableau Software. (n.d.). Available at: https://www.tableau.com/ (Accessed: 14 February 2023).
* Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. ACM SIGKDD Explorations Newsletter, 11(1), 10-18.