Module 6: Final Project

Exploring Content Trends and User Preferences through Data Analysis of Netflix

Sean McLean

Catherine Smereena Dommaty

Susheel Reddy Nelvoy

Northeastern University

ALY 6040 – Data Mining Applications

Professor Reilly

May 14th, 2024

Introduction

Netflix is a well-known streaming service that provides its users with a wide selection of films and television series. In this report, we examine a dataset that we downloaded from Kaggle that includes details on the platform material that is available. The dataset has 8,807 entries with 12 variables, each containing information about a movie or television shows, including the director, release year, rating, runtime, genre, and synopsis. The primary goal is to utilize viewer preferences and recent data to enhance content selection, recommendation systems, and platform features. This includes the overarching objective of attracting and retaining subscribers and fostering sustained growth for Netflix. By deriving actionable insights, Netflix can elevate user satisfaction, bolster engagement, and solidify its market position in the streaming industry. This process will be conducted through means of exploratory data analysis followed by several methods of data mining.

Business Questions

1. **Content Strategy & Viewer Retention**: What content mix (movies vs. television series) is optimal for attracting and retaining subscribers, considering trends in viewer preferences over time?
2. **Recommendation System Enhancement**: How can Netflix improve its recommendation system by incorporating preferred genres, content ratings, and viewing patterns?
3. **Content Acquisition & Success Forecasting**: How can recent data aid in identifying profitable genres or formats and predict the success of upcoming content?
4. **Platform Growth & User Experience**: How can Netflix sustain growth by utilizing data insights to attract new subscribers and enhance user experience for improved customer loyalty?

We will address the business inquiries by initially conducting exploratory data analysis on the dataset imported into R, followed by employing data mining techniques such as association mining and clustering. These approaches will offer insights into how user history over time and viewership patterns influence the acquisition of content on the streaming provider.

Exploratory Data Analysis (EDA)

In the context of exploration, we loaded the dataset, examined its structure, and performed data preprocessing tasks. We aimed to understand the data's characteristics, detect any missing values, and clean the data appropriately. After the initial exploration, we conducted EDA to delve deeper into the dataset. After loading and exploring the Netflix dataset, we performed data preprocessing, converting 'type' and 'rating' to factor variables and 'date_added' to date type. Missing values in the 'director' and 'cast' columns were handled by replacing them with an "unknown" metric. The dataset overall is relatively clean, as it contained no duplicates and no significant outliers. However, some unusual entries, such as zero minutes of runtime, were identified, so it revealed some valuable insights into the streaming dataset.

Most of the content on the platform consists of movies, with a smaller portion being television series. The most common ratings are TV-MA, TV-14, and TV-PG, and most of the content has a runtime between 90 to 120 minutes. An initial bar plot depicting the highest-rated streaming content in descending order was generated to show the data (See Appendix A). However, this visualization was amended upon observation that eight entries in the 'rating' variable were either empty or contained inaccurate information. Despite the scarcity of movies and television shows with specific ratings, they were retained considering the likelihood that their availability on Netflix is attributed to their popularity.

The number of releases on Netflix has increased over the years, with a significant surge after 2010. Directors Raúl Campos, Jan Suter, and Marcus Raboy have the most content on the platform, and international movies, dramas, and comedies are the most common genres that are available to watch. This information can help with content acquisition and recommendation methods and is essential for comprehending the impact of directors on the platform. These findings are also significant as they provide insights into user preferences, enabling Netflix to enhance its recommendation system, tailor its content offerings, and increase user engagement and subscriber retention.

The Netflix rating distribution shows how frequently certain content ratings are offered on the network. The most common ratings suggest that there is more content aimed at older audiences. Comprehending this distribution facilitates the classification of material and recommendation algorithms, enabling Netflix to customize content suggestions according to users' inclinations and suitability for their age. By creating recommendations based off user preferences and making sure that the content is in line with the tastes and sensitivities of the audience, this method improves the user experience.

Key Findings

The dataset includes a substantially higher proportion of movies than television programming, so Netflix and content makers need to understand this distribution to invest and buy more movies than TV series based off these preferences. The distribution of ratings provides information about the target demographic and age suitability of the Netflix material. This discovery is crucial for enhancing recommendation engines and classifying material. The streaming provider may adjust its content offerings to reflect the tastes and sensitivities of its user base by knowing the most popular ratings.

The distribution of release years sheds light on how Netflix's content has changed over time. The notable increase of releases after 2010 suggests the platform's growth and heightened expenditure on content generation and procurement (Bansal, 2020). This discovery is noteworthy since it demonstrates Netflix's ongoing efforts to update and improve its catalogue. The runtime range can also be an indicator of how long users prefer to watch movies before seeing a decrease in interest (See Appendix B).

Identifying the top directors can help Netflix understand who are most appealing to its audience, allowing them to acquire more content directed by these individuals. Understanding which genres viewers are watching will enhance user preferences and content categorization depending on the results. By identifying the most popular genres, Netflix can increase its investment in these areas and better cater its content offerings to its user base's tastes.

Deeper study would be required in the next stages to comprehend the relationships between the various features. For example, determining whether runtime and user ratings are correlated, or whether the content's rating is influenced by the year of release. Further insights can be gained by investigating whether specific directors prefer any given genre. Sentiment analysis on the descriptions might also be done to comprehend user preferences more deeply.

Data Mining Techniques: Analysis

The application of **association mining** unveils prevalent patterns in customer selection behaviors for building recommendation systems. It was used for this dataset because it is a pivotal technique in data mining, exploring intriguing relationships among items within a dataset, unveiling frequent co-occurrences of attribute-value conditions (Cakir, Aras, 2012). Association rule mining also entails the identification of latent connections among distinct items (Chonyy, 2020). Numerous approaches exist for conducting association rule mining. Initially, the Netflix

dataset is imported into R using the 'arules' library, followed by summarizing its statistics to grasp its attributes. Conversion of the dataset into a transaction object enables association mining analysis, revealing frequent items and transaction sizes. The 'apriori' algorithm is utilized to establish parameters and return item sets and transactions, with results inspected for support, confidence, and lift values (Chonyy, 2020). Additionally, the 'eclat' function identifies and provides visualizations for the most frequent items, offering insights into item frequencies based on total transaction counts.

After the dataset has been converted for association mining analysis, the summary shows that the average transactions are 11.11 and the most frequent items are ‘Other’ by a far margin followed by ‘Movie,’ ‘TV-MA,’ and ‘TV-Show.’ For further analysis and a deeper look at the top transactions, the parameters of the association rule mining are adjusted to 0.08 for support and 0.4 for confidence. The association rules imposed on the dataset show the correlations between item sets with the ‘other’ item that is blank when the code is executed, and ‘movie’ item relates to high values in the support, confidence, coverage, and lift measures. Whatever the other item is means that it leads to an occurrence of the movie item which could be a rating, a director, cast member, or some other attribute. This could be part of the process in how a recommendation system is built for a subscriber looking for a movie to view (Wang, 2022). Inspecting the other top itemsets shows that movies are most correlated with release dates, ratings, and the location of the movie. The only relationship with television content that appears among the top transactions is the show's duration of one season. The counts of each item set are high enough that we can be confident in the metrics of each correlation of the rules.

The ‘eclat’ function is implemented that provides all the support metrics between rules and items that have a value of at least 0.07, and the maximum length of frequent item sets

parameter is set to 15. The highest on the list is the movies item at 0.69, and the other items that appear include release dates that are all current, locations like the United States and India, and several ratings for all types of content. Like the ‘apriori’ function, the rules that match the parameter requirements in the ‘eclat’ function are mainly movies with release dates, ratings, and locations and television shows with ratings, location, and duration of the show in seasons. All the rules and items have adequate count totals to feel confident that the metrics are accurate. The highest count total by a significant margin is movies followed by the rating ‘TV-MA’ and television shows (See Appendix C).

The ‘apriori’ function is used again with the parameters set to 0.01 for support and 0.8 for confidence. The rules with television content all correlate with the program's genre and duration at four seasons, while all the movie content correlates with the runtime of the feature. All six of the rules that match the parameters have very low count totals and support values but indicates that when they occur together the confidence levels are all very high. A few of these rules also have some of the highest lift values, and the other rules with the highest lift values are all internationally based programming. The parameters of the dataset applying the ‘apriori’ function are fine-tuned again and alleviate all the subset rules that will remove any redundancy. Of the 3,852 rules in the dataset there were 3,254 subset rules that are jettisoned.

An example of a transaction after the subsets have been removed with the same parameters as any shows with the rating of ‘TV-MA’ is displayed with the correlations of other items and the metrics per item. Most of the top rules and items with these low set parameters pertain to international television programming that is available on the streaming provider. The parameters were revised again to show what users that were watching programs with the “TV-MA” rating were also watching. This was done by changing the confidence minimum value to

0.15 and switching the sides of the “TV-MA” and “default” rules in their association. The viewers that watched these types of programs also watched a variety of content that premiered in recent years, was one season in length, and was shot in the United States.

After importing a few libraries for data visualizations of the association mining results, the parameters of the ‘apriori’ function are edited again with support set to 0.01 and confidence set to 0.2. The rules are then ordered from one to 10, and an inspection shows that the top five rules are blank or default on the left side with the correlation being to adult content that was taped in the United States. The top-rated rule that does not have a default on either side is a rule that shows the correlation between reality television and television shows. A bar plot of the top 10 rules and a scatterplot of the 392 rules within the set parameters are then executed (See Appendix D). The scatterplot displays the points with support values on the x-axis and the lift values on the y-axis, and the shade of the data points in red shows the confidence values. Most of the points with the highest confidence values are below 0.2 in support and below five in lift value. All the outliers in the plot have a high lift value, low support values, and vice versa.

Additional evaluation of the frequencies of the item relationships by support measure and count shows that movies and television shows are still quite common. The ratings of the content, the year the content was released, and the location of the production is always prevalent. The release dates are all recent, so viewers are seeking newer content over older movies and television shows. The only other country besides the United States is India which indicates that there is content from there that has a popular viewership following on Netflix. The top three items in terms of support measure values are movies, the rating TV-MA, and television movies, so adults are likely to be looking for those items with that rating over other demographics and preferences on the platform.

Additionally, the exploration of **clustering** analysis plays a vital role in our team's efforts, aimed at categorizing data into distinct groups based on similarities among observations. By employing this powerful technique, we aim to uncover latent structures and patterns within the Netflix dataset that might elude conventional exploratory data analysis (EDA) methods alone. It was used in the data mining analysis because recommendation systems employ clustering algorithms to group users with similar preferences, enabling personalized suggestions based on past interactions across various online platforms (Rajah, 2023). This approach allows us to identify clusters indicative of recurring themes, genres, or viewer preferences, providing valuable insights for targeted marketing initiatives, content recommendation systems, and content development decisions.

In our analysis of Netflix's content library, we leverage both EDA and clustering methodologies to extract insights crucial for addressing key business inquiries. Through the utilization of features such as release year and duration, we aim to identify patterns and categorize similar content, aiding in content strategy, recommendation system enhancement, content acquisition, and platform expansion. Our clustering techniques, including k-means and hierarchical clustering, are complemented by methodologies such as the Elbow Method, Silhouette Method, and Gap Statistic Method, allowing us to determine optimal cluster numbers and reveal meaningful content categories and trends.

Another technique for figuring out how many clusters in a dataset is ideal is to use the gap statistic. The within-cluster sum of squares (WSS) of the observed data is compared to reference datasets that were produced at random. This assists in determining the statistical significance of the observed clustering structure by creating reference datasets that have similar features with the original data.

The number of clusters (k) being analysed within the given range is displayed on the x-axis (See Appendix E). The gap statistic is computed for each value of k that each tick mark on the x-axis corresponds to. The gap statistic, which shows the difference between the observed and expected WSS for each value of k, is plotted on the y-axis. Greater evidence for the existence of separate clusters in the data is indicated by a bigger gap statistic (Mallikarjun, A J, 2022).

Interpretations and Recommendations

The analysis of the dataset using association mining methods shows that the content that are viewed the most frequently (support) and have high confidence values encompasses movies and TV programming for adults that was released recently and was filmed in the United States. These results were similar when the 'eclat' function was executed with some international content also among the most viewed programming. When the parameters were revised with content that has been watched less frequently but had high confidence values, the types of programs were different. This indicates that there is a subgroup of users that are watching certain types of content from the item sets even though they are transpiring less frequently than some of the most popular content available. When the lift metrics were adjusted to show the strength of associations in transactions, the content was similar but international programming was rated high by this metric. When the subsets were removed and the rating of 'TV-MA' was used as an example, international programming was also prevalent in viewership. This suggests that the streaming provider has some variety in its content that is being watched from particular rules and itemsets but with less frequency.

Rules that are less viewed but with high confidence values can also be used to bring in new subscribers and can be analyzed more in the future. These factors will also be vital in content acquisition and forecasting what types of programming will be relevant and popular. The blank sides of some of the rules with high values might indicate that there are so many options that they are combined into one side that correlates with the other. With the high-count totals in the dataset, I recommend removing subsets and then building training and testing sets for a deeper analysis that would hopefully eliminate any blank sides in the rules and itemsets.

Overall, when the rules with the highest supports values and the highest confidence values were executed, it showed that more general programming like newer content for adults that was filmed in United States was more frequent. The highest confidence values though were content from other types of programming that were viewed less frequently. These findings from this type of data mining can help address the business questions in that it can help find the right balance in content that can retain its current subscribers and used to help attract new users. It can also help strengthen recommendation systems for viewers by selecting content they will most likely watch by factoring in rules that are most frequent.

I also recommend that Netflix continues to look beyond the national landscape for quality content because from the data there is a market for quality television and movies from other countries. In places like India and South Korea the association mining results show that viewership is high from shows that are released from those parts of the globe (Netflix.com, 2024).

From the clustering mining, the two clusters may naturally arise from the data, which could indicate distinct genres or audience preferences. These clusters could be used by the data owner to customise marketing campaigns or suggestions for various target segments. Additional

variables like viewer ratings or outside sourced genre classifications could be incorporated to improve the accuracy and insights of clustering.

The overarching objectives of the analysis from clustering techniques include refining content strategy, improving recommendation systems, guiding content acquisition decisions, and facilitating platform growth. By incorporating additional features and techniques such as genres, ratings, popularity metrics, temporal analysis, collaborative filtering, and predictive modelling, we seek to enrich our understanding of user preferences and behaviors, ultimately driving user engagement and platform success.

Integrating word tagging from descriptions into the algorithm could also introduce more diversity into recommendation systems, ensuring that a broader range of content is suggested. For instance, if a user is interested in climate change, the system should recommend all relevant content on the topic. Trending content should be emphasized to attract new subscribers, with recommendation parameters fine-tuned regularly to adapt to evolving viewing preferences.

It is essential to promote both trending and popular content to new subscribers, focusing on what resonates with the current user base. This approach not only sustains growth but also caters to the diverse interests of existing and potential customers. Additionally, leveraging association mining metrics can help identify high-confidence content to attract new audiences. This strategy may involve introducing a currently popular movie that might appeal to a new demographic, even if it does not rank high among current users on Netflix.

These findings answer most of the business questions in offering what the best combination of movie and television streaming options available on Netflix. From past user history the algorithms can be fine-tuned to build the most comprehensive and suitable recommendation system for each subscriber. The plethora of data also provides the statistics on

what features are most popular and can build the best streaming library based off user preferences. The other question that focuses on future growth can be researched more in the next week to find more insights into how the platform can continue to provide the best results to its customers. A recommendation would be to look at past data mining results for patterns and trends that could provide more answers to this business question. More analysis through all data mining processes could also be executed to identify more behaviors and inclinations in user viewing habits.

Conclusion

In brief, the analysis uncovered valuable insights to assist the data custodian in decision-making, such as classifying content, segmenting users, and improving recommendation systems. By effectively mining the structure of the Netflix dataset, stakeholders can make informed decisions to enhance user satisfaction, attract new subscribers, improve content suggestions, and expand the business. Further exploration and experimentation with various clustering algorithms and association rule mining techniques, tailored to specific business inquiries, offer the potential to discover additional insights and fine-tune advancement strategies.

References

1. Bansal, S. (2020). *Netflix Show and Movie – Exploratory Analysis*. Kaggle. https://www.kaggle.com/shivamb/netflix-shows-and-movies-exploratory-analysis.

2. Cakir, O., and Efe Aras, M. (2012). *A recommendation engine by using association rules.* Elsevier Ltd. https://www.sciencedirect.com/science/article/pii/S187704281203515X?ref=cra_js_challenge&fr=RR-1

3. Chonyy. (2020, Oct. 25th) *Apriori, - Association Rule Mining In-depth Explanation and Python Implementation.* Medium. https://towardsdatascience.com/apriori-association-rule-mining-explanation-and-python-implementation-290b42afdfc6

4. Mallikarjun, AJ. (2022). *Gap_Statistics.* Kaggle. https://www.kaggle.com/code/mallikarjunaj/gap-statistics

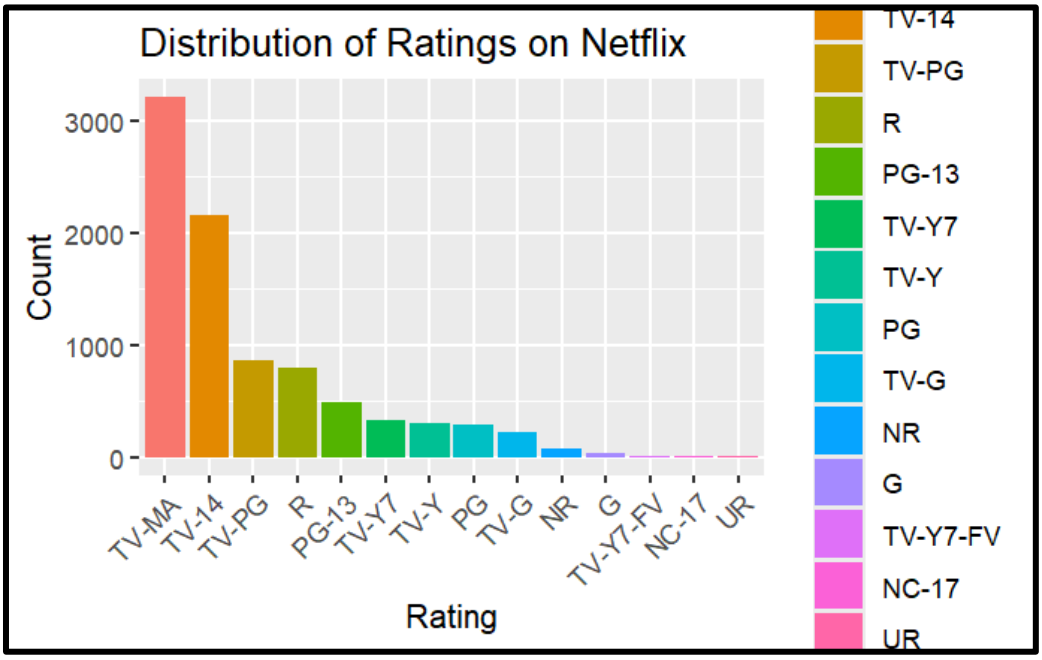
5. Netflix. (2024). *About Netflix*. Netflix. https://about.netflix.com/en.

6. Rajah, K. (2023, Feb. 5th). *Clustering Based Algorithms in Recommendation Systems.* Medium. https://medium.com/@Karthickk_Rajah/clustering-based-algorithms-in-recommendation-system-205fcb15bc9b

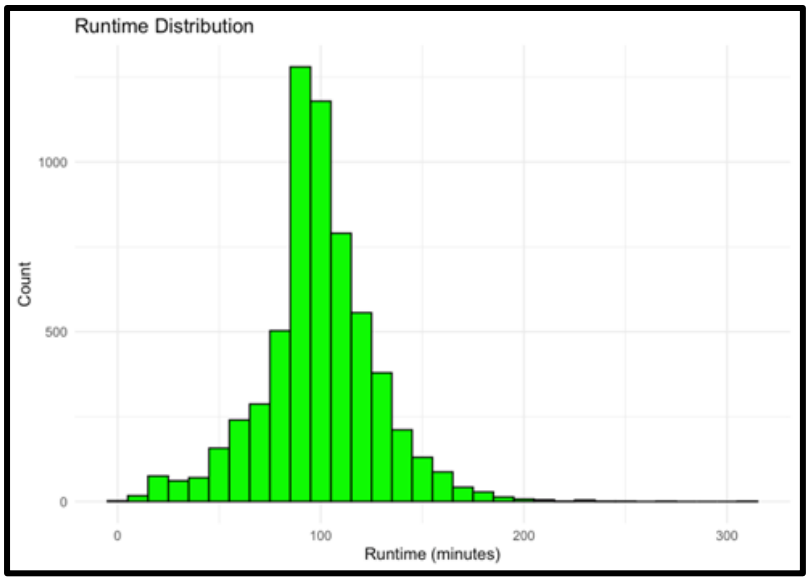
7. Wang, A. (2022, Feb 14th). *Netflix's Recommendation Systems: Entertainment Made for You.* Illumin Magazine. https://illumin.usc.edu/netflixs-recommendation-systems-entertainment-made-for-you/

Appendices

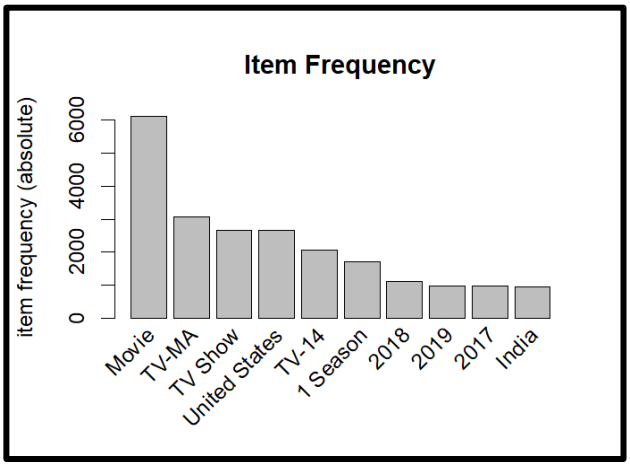
Appendix A



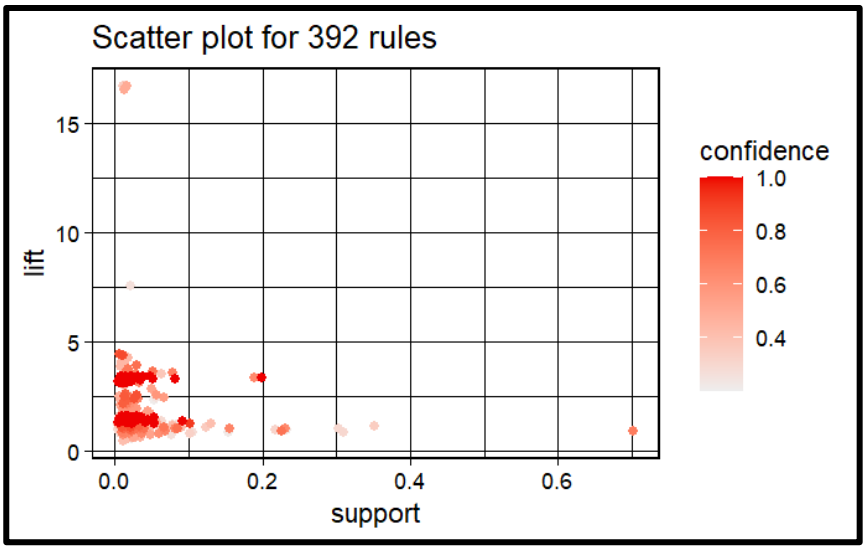
Appendix B



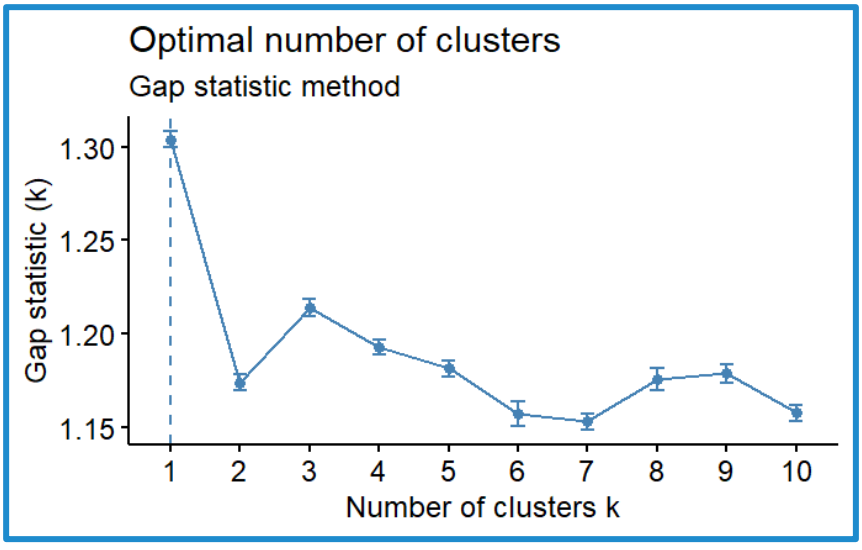
Appendix C



Appendix D


Scatter plot for 392 rules
lift
15
10
5
0
0.0
0.2
0.4
0.6
support
confidence
1.0
0.8
0.6
0.4

Appendix E


Optimal number of clusters
Gap statistic method
Gap statistic (k)
1.30
1.25
1.20
1.15
1
2
3
4
5
6
7
8
9
10
Number of clusters k