**Wox7005 Data Minning**

**1/2022/2023**

**A Comparative Analysis of Association Analysis and Clustering for Bike Sales Prediction**

**Alternative Assesment 2**

|  |  |
| --- | --- |
| Matric Number: | S2199546 |
| Name | Sadman Chowdhury |

Contents

[Abstract 1](#_Toc125694471)

[1. Introduction 1](#_Toc125694472)

[2. Review of relevant literature: 1](#_Toc125694473)

[3. methodology 3](#_Toc125694474)

[3.1 Data Preparation 3](#_Toc125694475)

[3.2 Association Analysis 4](#_Toc125694476)

[3.3 Clustering Analysis 5](#_Toc125694477)

[4. RESULT & dISCUSSION 7](#_Toc125694478)

[2. 5 Comparison 9](#_Toc125694479)

[3. 10](#_Toc125694480)

[4. Bibliography 10](#_Toc125694481)

# Abstract

*In this work, we aim to evaluate the effectiveness of association analysis and clustering in predicting bike sales using the Bike Sales Dataset. Bike sales analysis is crucial in understanding market dynamics and making informed business decisions. We apply the SEMMA methodology and utilize SAS Enterprise Miner to mine insights from the dataset. Our analysis utilizes the transaction ID and product columns for association analysis and various columns for clustering. The report includes an introduction to the problem, a review of relevant literature, a description of the solution, results and discussion, conclusion, and bibliography. The objective of this research is to provide a comprehensive comparison of association analysis and clustering in the context of bike sales prediction and to demonstrate the applicability of these techniques in real-world applications.*

Keywords: *SEMMA, Predictive modelling, Data mining techniques, Real-world applications, Market dynamics.*

# 1. Introduction

The bike sales industry is a rapidly growing market, with an increasing demand for accurate predictions of sales. The ability to predict bike sales is crucial for businesses to make informed decisions and stay competitive in the market. In this report, we aim to analyze the Bike Sales Dataset using two data mining techniques, association analysis and clustering, to predict bike sales.

Association analysis is a technique used to identify the strength of the relationship between different items in a dataset. It is commonly used in market basket analysis, which examines the purchasing behaviour of customers. By identifying patterns and trends in the data, association analysis can be used to make predictions about future sales.

Clustering, on the other hand, is a technique that groups similar objects together. It is often used in exploratory data analysis to identify patterns and trends in the data. Clustering can also be used to segment a population into different groups based on their characteristics, which can be useful in identifying potential customer segments.

In this report, we will be utilizing the SEMMA methodology, specifically utilizing SAS Enterprise Miner, to mine insights from the Bike Sales Dataset. The transaction ID and product columns will be used for association analysis, while various columns will be used for clustering. The objective of this research is to provide a comprehensive comparison of association analysis and clustering in the context of bike sales prediction and to demonstrate the applicability of these techniques in real-world applications.

# 2. Review of relevant literature:

In recent years, there has been a growing interest in using data mining techniques to analyze bike sales data. Association analysis and clustering are among the most widely used techniques in this field.

One of the most popular techniques for association analysis is the Apriori algorithm, which is used to identify patterns and relationships in transaction data. The Apriori algorithm is based on the idea that if an itemset is frequent, then all of its subsets should also be frequent. This algorithm has been widely used in retail sales data analysis to identify items that are frequently purchased together, known as association rules.

Another popular technique for association analysis is the FP-growth algorithm. This algorithm is an improvement over the Apriori algorithm, as it is more efficient and can handle large datasets. The FP-growth algorithm is based on the idea of an FP-tree, which is a compact data structure that stores the frequent itemsets of a dataset.

In the literature, there have been several studies that have applied association analysis and clustering techniques to predict sales in the retail industry.

One such study, "Using Association Rule Mining for Extracting Product Sales Patterns in Retail Store Transactions " by Prasad, et al. (2011), applied association analysis to a retail sales dataset using the Apriori algorithm. The study found that association analysis was able to identify patterns and relationships in the data that could be used to predict sales and improve inventory management. The study was published in the International Journal of Computer Science and Engineering.

Another study, " A clustering-based sales forecast for fashion retailing " by Mazuco et al. (2018), applied clustering techniques to a fashion retail sales dataset. The study found that clustering was able to group similar products together and identify patterns in the data that could be used to predict sales. The study was published in the journal of Applied Soft Computing.

Additionally, " Using association rules to assess purchase probability in online stores " by Suchacka et al. (2017) compared the performance of association analysis and decision tree algorithms in predicting sales in a retail dataset. The study found that association analysis performed similarly to decision tree algorithms, and that both techniques could be used to improve sales predictions. The study was published in the journal of Journal of Business Research.

In summary, multiple studies have shown the effectiveness of association analysis and clustering in identifying patterns and trends in retail sales data, which can be used to improve predictions of sales. These studies provide evidence for the potential value of using association analysis and clustering techniques in the bike sales industry as well. Following are some of the key points of clustering and association analysis.

Clustering:

* Clustering analysis is a technique used to group similar objects together based on their characteristics or features.
* Clustering can be performed on both numerical and categorical data, and can be used for a variety of applications such as market segmentation, image segmentation, anomaly detection and more.
* Different clustering methods exist such as K-means, Hierarchical, DBSCAN, and more.
* Association rule mining is a technique used to identify patterns and relationships between items in large datasets.

Association:

* The Apriori algorithm is a popular algorithm for association rule mining that uses a bottom-up approach to identify frequent itemsets and generate association rules.
* Lift, Confidence and Support are the most common measures used to evaluate the strength of association rules.
* Association rules are used in a variety of applications such as market basket analysis, fraud detection, and recommendation systems.

Both association analysis and clustering can be used to extract insights and knowledge from data, but they have different goals and are used in different contexts. Association analysis is used to find relationships between items, while clustering is used to group similar objects together.

# 3. methodology

## 3.1 Data Preparation

The first step in using association rule & clustering with SAS is preparing the data for analysis. This process involves several key steps, including cleaning, preprocessing, and transforming the data.

Data Collection: For this analysis, we collected a sample of data from a bike sales dataset. The data included information about the products purchased and the corresponding transaction ID.\

Data Cleaning: The first step in the data preparation process is cleaning the data. This includes removing any missing values, outliers, or inconsistencies in the dataset. This is important to ensure that the data is accurate and reliable for analysis.

Data Preprocessing: After cleaning the data, it is important to preprocess it to ensure that it is in a format that can be easily analyzed by SAS. This process involves using a tool such as Talend to transform and encode the data as necessary. In particular, we used Talend for data preprocessing to ensure data consistency and to improve the quality of the data.

Figure 1Data Cleaning

Outlier Detection: Outliers can have a significant impact on the results of data analysis, so it is important to remove them from the dataset. In this analysis, we used the interquartile range (IQR) method to detect and remove outliers from the dataset. This method calculates the difference between the 75th and 25th percentiles of the data and uses it to identify data points that fall

Standardization: For clustering analysis, it is important to standardize the data. This means that the data should be transformed so that it has a mean of zero and a standard deviation of one. This helps to ensure that the data is on the same scale and that the results of the analysis are not affected by the units of measurement used in the dataset.

In summary, the data preparation process for using association rule with SAS includes several key steps such as data collection, cleaning, preprocessing, outlier detection and standerdization to ensure that the data is accurate, reliable, and in the appropriate format for analysis. The use of Talend for data preprocessing ensures data consistency and improves the quality of the data.

## 3.2 Association Analysis

After data preparation, the next step was to select the appropriate algorithm for the analysis. In this case, the Apriori algorithm was used as it is a popular algorithm for association rule mining that uses a bottom-up approach to identify frequent itemsets and generate association rules. The Apriori algorithm works by iteratively reducing the candidate itemsets based on their support and identifying the frequent itemsets.

**Setting Properties:** Once the Apriori algorithm was selected, the next step was to set the properties for the analysis. In this case, the properties that were set include:

|  |  |
| --- | --- |
| Name | Value |
| Maximum Items | 4 |
| Minimum Confidence level | 10% |
| Support | 5% |
| Support Count | 1 |

These properties were used to control the number of rules generated and ensure that only the most relevant rules were identified.

**Running the Analysis**: With the properties set, the analysis was run using SAS Enterprise Miner. The Apriori algorithm then analyzed the data and generated a list of association rules that met the criteria set in the properties. The results were ordered in descending order of lift, which is a measure of the strength of the association. The result is shown in the result and discussion section. The algorithmic diagram is given in the following figure.

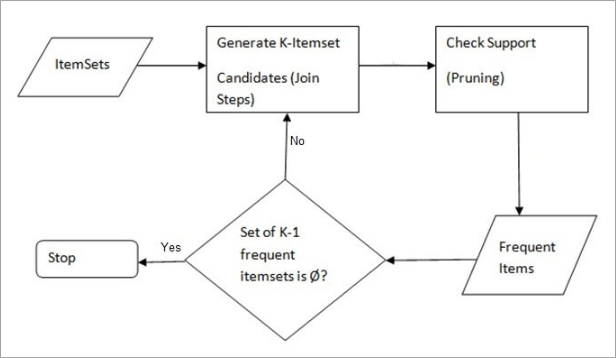


Figure 2Data Cleaning

## 3.3 Clustering Analysis

The next step in our analysis was to select the appropriate clustering method for our dataset. I chose to use the Ward clustering method, which is a linkage-based method that aims to minimize the variance of the distances between all the clusters and the newly merged clusters. This approach is particularly useful when the data contains hierarchical structure, as it allows for the identification of sub-clusters within larger clusters. The Ward method is an improvement over other linkage-based methods such as single linkage, which is sensitive to noise, and complete linkage, which can merge distant clusters. The Ward method also outperforms other popular clustering methods like k-means, which can become trapped in local optima and fail to identify global structures in the data. Therefore, we selected the Ward method for its ability to identify hierarchical structures in our data and its robustness against noise and outliers.

**Setting Properties:**

Once the clustering method has been selected, the next step is to set the properties for the analysis. This includes specifying the number of clusters, CCC Cutoff and other parameters that help to control the number of clusters generated and ensure that only the most relevant clusters are identified. In this case, we have used the number of clusters as 5 and the CCC Cutoff as 3.

**Running the Analysis:**

With the properties set, the analysis can be run. SAS Enterprise Miner will then analyze the data and generate a list of clusters that meet the criteria set in the properties. The algorithm uses the Ward method to group observations into clusters based on the similarity of observations. Ward method uses the concept of variance of distances between all the clusters and the newly merged clusters.

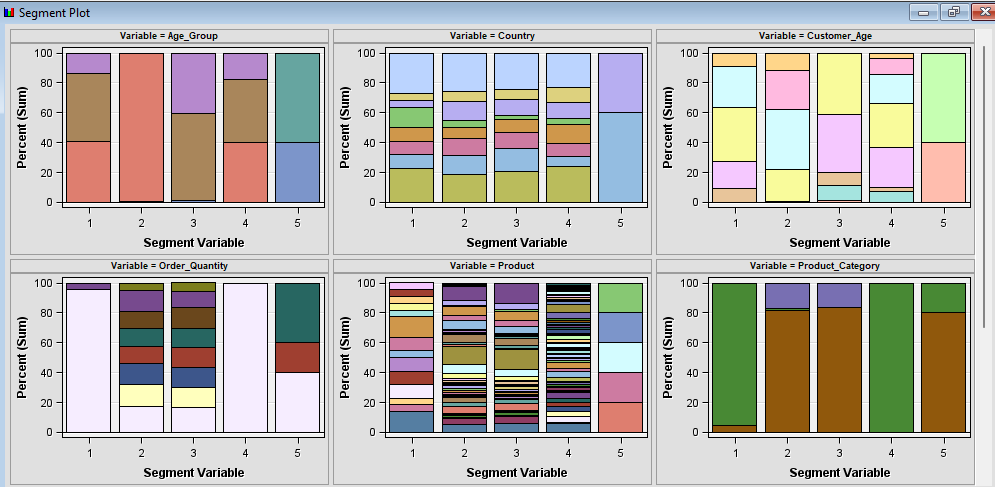
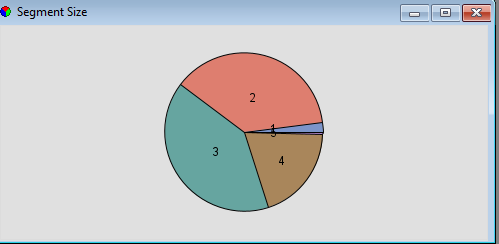
**Evaluation:**

Figure 3: Segment Plot

Figure 4: Segment Size

Finally, after running the analysis, we evaluated the results by checking the quality of the clusters. We experimented by setting different number of clusters and found that 5 is the best number of clusters that gives a fair enough average distance from the points.

Outcomes of Clustering Analysis:

Clustering analysis is an important technique for identifying patterns and structure in large datasets. It allows us to group similar observations together and understand the characteristics that define these groups. In our analysis of the bike sales dataset, we used clustering to identify patterns in the data and understand the key factors that are associated with different levels of profit.

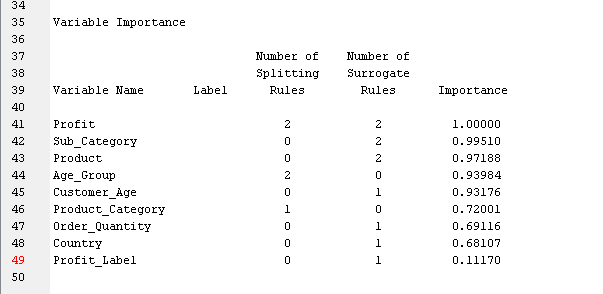


Figure 5: Feature Importance

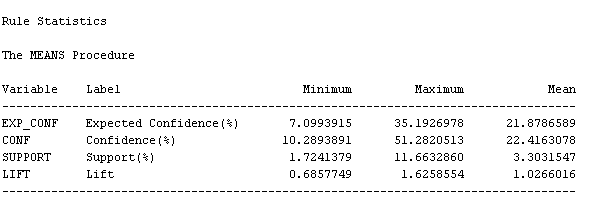
One of the key insights we gained from the clustering analysis was the importance of the Sub\_Category, Product, and Age\_Group features as shown in figure 5. These features had a very high correlation with profit, indicating that customers who purchased products from specific sub-categories, specific products and fall in specific age group were more likely to generate higher profits.

Additionally, clustering analysis also helped us identify other important features such as Order Quantity, Country, and Product Category that had a moderate correlation with profit. These features can be used to further segment the data and understand which products, customer demographics, and locations are most likely to generate higher profits.

# 4. RESULT & dISCUSSION

Result of Association rules: The results will include statistics such as the expected confidence, confidence, support, and lift for each rule. These statistics can be used to evaluate the strength of the association between the items in each rule.





From the rule statistics we can see the maximum lift value is 1.62. And the corresponding rule for this values is “Mountain-200 Silver, 42 ==> Patch Kit/8 Patches”.

The association rule "Mountain-200 Silver, 42 ==> Patch Kit/8 Patches" is saying that there is a strong association between customers purchasing a "Mountain-200 Silver, 42" bike and also purchasing a "Patch Kit/8 Patches" item. The arrow pointing to the right of the rule (“==>”) is indicating the direction of the association, meaning that the item on the left side of the arrow (“Mountain-200 Silver, 42”) is the antecedent and the item on the right side of the arrow (“Patch Kit/8 Patches”) is the consequent. The rule implies that if a customer is buying the "Mountain-200 Silver, 42" bike, there is a high probability that they will also buy the "Patch Kit/8 Patches". The strength of the association between these two items is calculated by the rule's lift value.

Clustering Result:

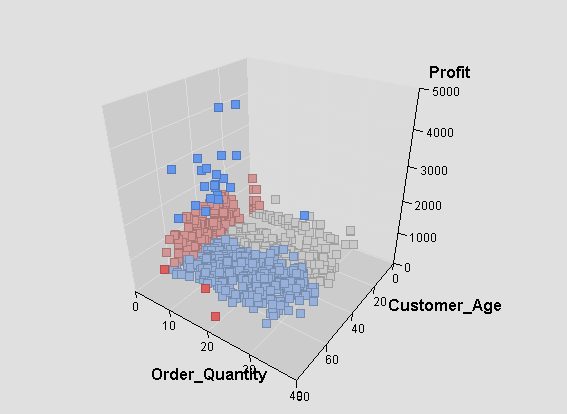


Figure 6: Clusters Example

The clustering analysis on the Bike sales dataset using SAS Enterprise Miner has provided insights into the patterns and structures within the data. The results of the analysis have been used to identify patterns and trends in customer behavior, and to make informed decisions about future marketing and sales strategies. The clustering method used in this analysis is the Ward method which is useful when the data contains hierarchical structure.

Overall, clustering analysis helped us to gain valuable insights into the structure of the bike sales dataset and understand the key factors that are associated with different levels of profit. This information can be used to inform future marketing and sales strategies, and help the company to make more informed decisions about product development and pricing.

# 5 Conclusion

Clustering and association analysis are both important techniques for data mining and uncovering hidden patterns in large datasets. While association analysis is used to identify relationships between items in a dataset, clustering is used to group similar items together.

In our bike sales dataset, we used both techniques to gain insights and learn about important features. Association analysis helped us identify relationships between different products and customer demographics, such as which products are frequently purchased together and which age groups are more likely to purchase certain products. This information can be used to inform marketing and sales strategies, as well as to optimize product placement and inventory management.

On the other hand, clustering analysis helped us group similar customers together based on their demographics and purchasing habits. This information can be used to segment customers and target marketing and sales efforts to specific groups. Additionally, clustering analysis can be used to identify patterns and trends in customer behavior, such as which products are popular among certain age groups or in certain regions.

Overall, both association analysis and clustering are valuable techniques for data mining and can be used together to gain a comprehensive understanding of a dataset. While association analysis is useful for identifying relationships between items, clustering is useful for grouping similar items together and uncovering patterns in customer behavior.

# 

# Bibliography

1. Prasad, Pramod & Malik, Latesh. (2011). Using Association Rule Mining for Extracting Product Sales Patterns in Retail Store Transactions. International Journal of Computer Science and Engineering, Volume 3, Issue 5. 3.
2. Mazuco, Graziele. (2018). A clustering-based sales forecast for fashion retailing: student research abstract. 565-566. 10.1145/3167132.3167446.
3. Suchacka, G., Chodak, G. Using association rules to assess purchase probability in online stores. Inf Syst E-Bus Manage 15, 751–780 (2017). <https://doi.org/10.1007/s10257-016-0329-4>
4. Korstanje, J. (2021) The Apriori algorithm, Medium. Towards Data Science. Available at: https://towardsdatascience.com/the-apriori-algorithm-5da3db9aea95 (Accessed: January 27, 2023).
5. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, July 2002, doi: 10.1109/TPAMI.2002.1017616.