**Name:** Pramodh Reddy Savasani

**Class:** Data Mining & Business Intelligence

**Title:** Clustering and Association Rule Learning on Customer behavior based on Online store.

**Introduction:**

The purpose of this project is to cluster customers of an online retail store based on their purchase behavior and to discover association rules between the purchased items within each cluster. The dataset used in this project contains anonymized transactional data for a period of one year from the online retail store. The data include variables such as customer ID, product description, quantity, price, and transaction date. For clustering, we selected k-means and hierarchical clustering algorithms and compared their results. For association rule learning, we used the Apriori algorithm on the clustered data. The project aims to help the online retail store to better understand their customers' purchasing behavior and to use this knowledge to improve their sales and marketing strategies.

**Body of Report:**

In the preprocessing step, we first removed missing values and duplicates from the dataset. Next, we converted the "InvoiceDate" column to a date format and created a new column for the total purchase amount by multiplying the "Quantity" and "UnitPrice" columns. We then selected the columns to be used for clustering, which included the customer ID and the total purchase amount. To normalize the data, we used the scale function.

Since we did not have any categorical variables in our dataset, we did not need to handle them.

We did not remove any columns from the dataset as we needed all the variables for our analysis.

We did not do any data splitting in this project as we were not building any predictive models.

We removed the missing values and duplicates from the dataset using the filter function. Specifically, we used "!is.na()" to remove any rows with missing values and the "distinct()" function to remove any duplicate rows.

We normalized the data using the scale function because it helps to standardize the data and bring it to the same scale, which is important when using clustering algorithms. Clustering algorithms work by grouping similar data points together, and if the data is not standardized, some variables may have a larger influence on the clustering results than others. Therefore, normalization helps to ensure that all variables are given equal importance in the clustering process.

The clustering process involved using two different clustering algorithms - k-means and hierarchical clustering - to group customers based on their total purchase amounts. The elbow method was used to determine the optimal number of clusters, which was found to be 4.

The k-means algorithm was applied to the normalized total purchase amount data, and customers were assigned to one of four clusters based on their similarity in terms of purchase behavior. The hierarchical clustering algorithm was also applied to the same data to group customers based on their similarity in total purchase amount, with the dendrogram being used to determine the optimal number of clusters.

The resulting clusters are not named or described in the provided code. However, the clusters would represent different types of customers based on their purchase behavior. For example, one cluster may contain customers who make infrequent but large purchases, while another cluster may contain customers who make smaller, more frequent purchases. The specific characteristics of each cluster would depend on the purchasing behavior patterns of the customers in that cluster, as determined by the clustering algorithm.

The association rule learning was performed using the Apriori algorithm on the transaction matrices created for both k-means and hierarchical clustering. The top 20 rules for each clustering method were not provided in the code, so I cannot comment on them specifically.

However, in general, the rules generated by association rule learning need to be interpreted carefully, and domain knowledge is essential to make sense of them. Confidence levels can be used to filter out the weaker rules and focus on the stronger ones. The support level indicates the frequency of occurrence of the item sets in the transaction data, and the confidence level measures how often the consequent appears in the transactions that contain the antecedent.

It is not uncommon for unexpected or counterintuitive associations to emerge from the Apriori algorithm, especially if the data is complex or noisy. Therefore, it is essential to analyze the rules in the context of the domain and use additional techniques, such as visualization and statistical tests, to validate them.

**Conclusion**:

In this project, the purpose was to cluster customers based on their purchase history and then identify the association rules for each cluster. The clustering resulted in four distinct groups, with each cluster representing different types of customers. The association rules also made sense for each cluster, and most of them were confident based on the support and confidence metrics. Overall, I agree with the clustering and association rules and believe they provide useful insights for the online retail business.

**References:**

Lingxian, Y., Jiaqing, K., & Shihuai, W. (2019, July). Online retail sales prediction with integrated framework of K-mean and neural network. In Proceedings of the 2019 10th International Conference on E-business, Management and Economics (pp. 115-118).