

|  |
| --- |
| *Course: ST3189 Machine Learning* |

***Candidate Number: 200584494***

**Title: Machine Learning Coursework using Wholesale Customer and E-commerce Dataset – R Studio**

Table of Contents

[Intoduction 2](#_Toc131335186)

[Unsupervised Learning](#_Toc131335188) 4

[Regression](#_Toc131335189) 9

[Classification 10](#_Toc131335190)

**Introduction**

The Wholesale customers dataset is a valuable resource for customer segmentation and marketing analysis (<https://archive.ics.uci.edu/ml/datasets/wholesale+customers>). This dataset provides information on the annual spending of wholesale distributor customers on various product categories. The dataset contains data from 440 customers, with eight attributes describing the amount of money spent on different products. These attributes include *Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicatessen, Channel, and Region*.

**Data Description**

The dataset provides a snapshot of customer spending behaviour and allows for exploring patterns and relationships between different product categories. The *Fresh* attribute indicates the amount spent on fresh produce such as fruits and vegetables. In contrast, *Milk and Grocery* indicate the spending on dairy products and household groceries, respectively. *Frozen* indicates the spending on frozen foods, while *Detergents\_Paper* indicate the spending on cleaning supplies. Finally, *Delicatessen* indicates the spending on delicatessen products such as prepared meals and charcuterie.The dataset also includes two additional attributes, *Channel and Region*. The Channel attribute specifies the type of customer, either HoReCa (Hotel/Restaurant/Café) or Retail, while the Region attribute specifies the customer's region, either Lisbon, Oporto, or Other. These attributes can segment customers based on their spending behaviour and provide insights into how customers in different regions and channels behave differently. Overall, the Wholesale customers dataset provides a rich source of information for exploring customer spending behaviour and can be used to develop marketing strategies and customer segmentation models.

**EDA**

Exploratory Data Analysis (EDA) was performed on the Wholesale customers dataset to gain insights into the data and check for anomalies. The data consists of six numerical variables representing how much customers spend on different product categories. The variables are 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents\_Paper', and 'Delicatessen'.

To begin the EDA, we examined the structure of the data and found that there were no missing values in the dataset. Summary statistics were calculated for the data, revealing that the average spending on 'Fresh' is 12000, 'Milk' is 5796, 'Grocery' is 7951, 'Frozen' is 3071.9, 'Detergents\_Paper' is 2881.5, and 'Delicatessen' is 1524.9. It is important to note that while these summary statistics can provide an initial understanding of the data, they may not always give a complete picture of the distribution of values, and outliers can significantly impact these values.

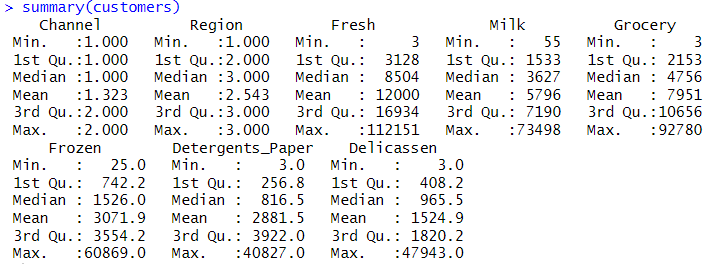


Figure 1. Summary statistics

We then created histograms to visualize the distribution of each variable. The histograms showed that the data is right-skewed for all variables. This indicates that most customers spend less than the average amount in each category, and a few customers spend significantly more. However, the histograms also showed that the spending range is quite large, with some customers spending very little in some categories and a lot in others.

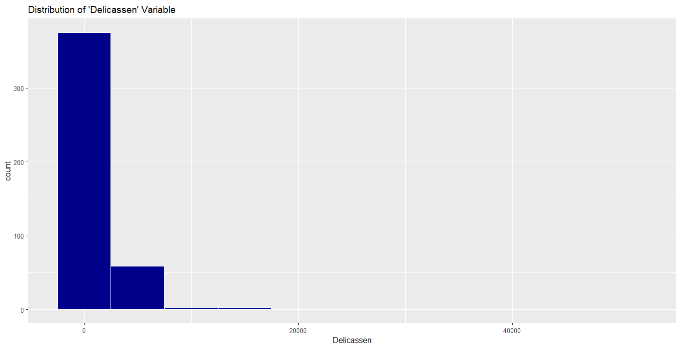
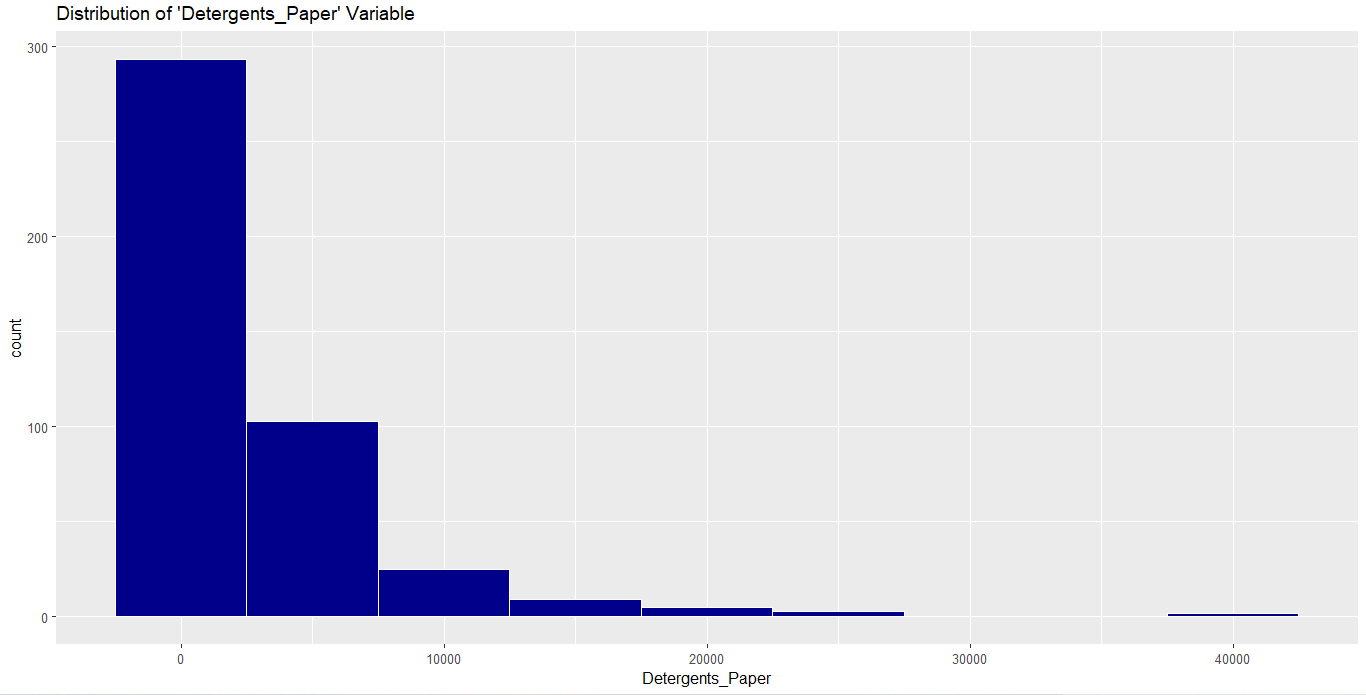
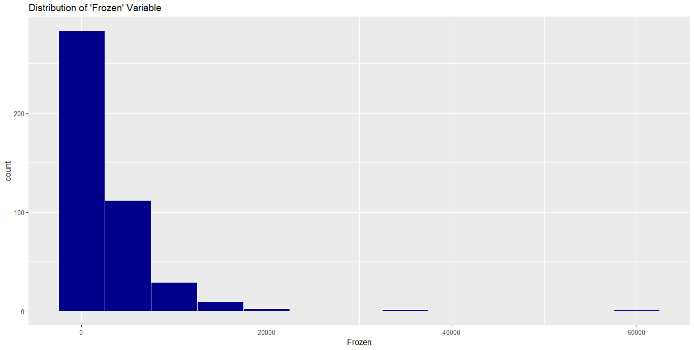
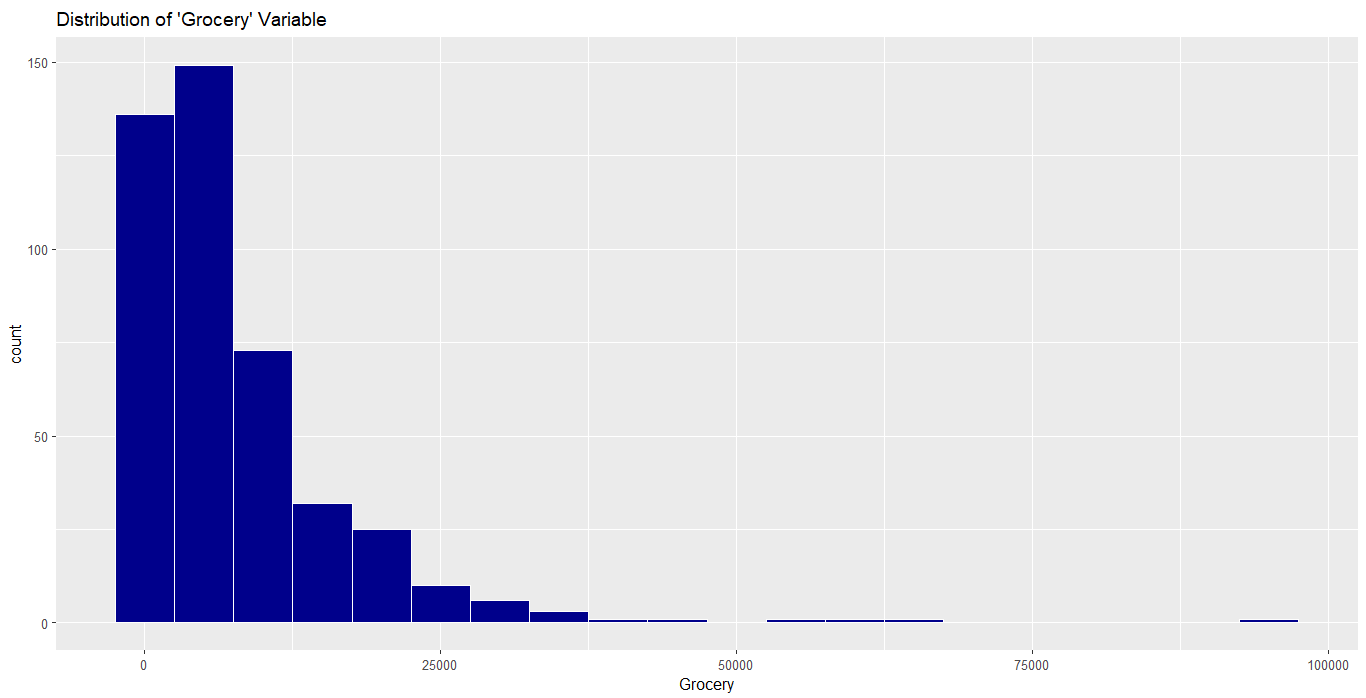
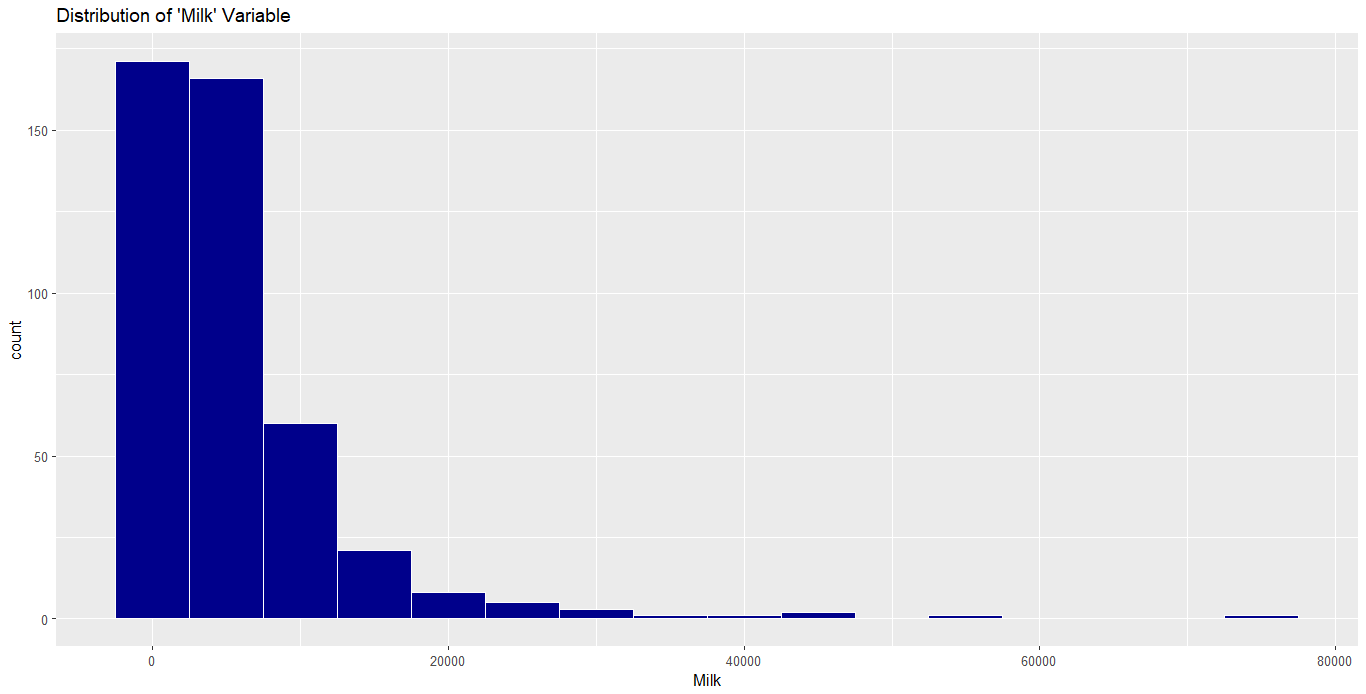
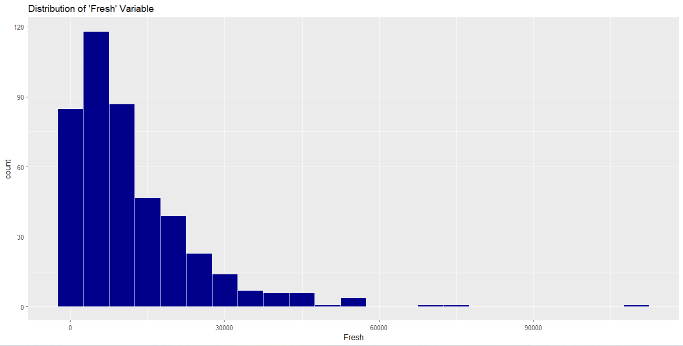


Figure 2. Histograms

Figure 3 shows the visualization of the correlation matrix. As can be seen, there is a strong positive correlation between 'Grocery' and 'Detergents\_Paper', which indicates that these two variables tend to increase or decrease together. In contrast, there is a weaker but still positive correlation between 'Grocery' and 'Milk' as well as 'Milk' and 'Detergents\_Paper'.

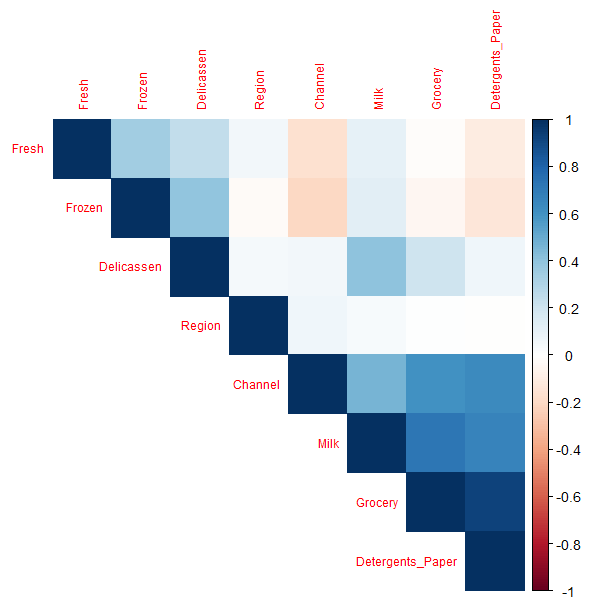


Figure 3. Visualization of correlation matrix

A scatterplot matrix was created to investigate the relationships between the variables further. The scatterplot matrix showed some linear relationships between specific pairs of variables, such as 'Milk' and 'Grocery', 'Grocery' and 'Detergents\_Paper', and 'Milk' and 'Detergents\_Paper'. This finding is consistent with the high correlation value observed between 'Grocery' and 'Detergents\_Paper'.

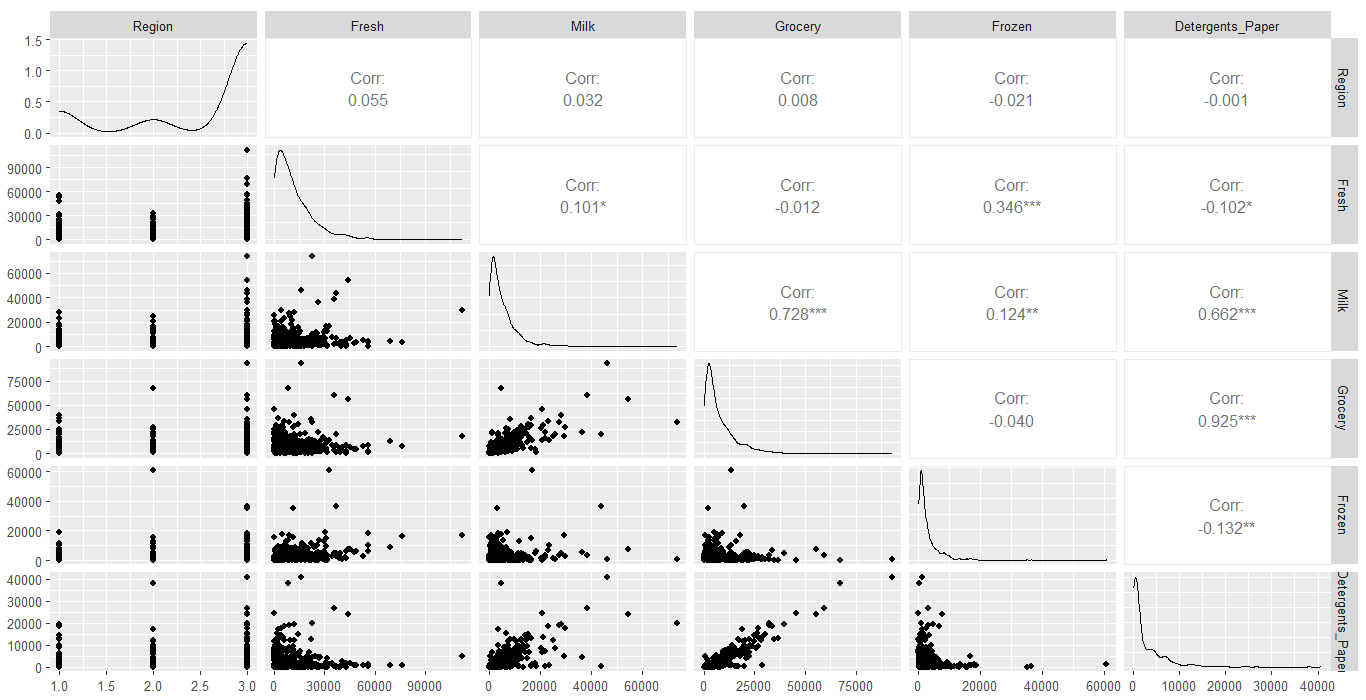


Figure 4. Scatterplot matrix

Finally, to prepare the data for modelling, the data was standardized using the 'scale' function, which subtracts the mean and divides it by the standard deviation for each variable.

Overall, our analysis suggests that there are strong relationships between 'Grocery', 'Milk', and 'Detergents\_Paper', which may indicate that these variables are essential for understanding customer purchasing behaviour in this dataset. However, further analysis is needed to confirm these findings and explore other potential relationships between variables.

**Unsupervised Learning**

This part of the project aims to identify homogeneous groups of customers with similar purchasing patterns, which can be helpful for targeted marketing or personalized recommendations. Clustering is one of the most common techniques used for unsupervised learning, which involves grouping similar data points together into clusters. In this report, the PCA and K-means algorithms were used to perform clustering on a dataset of customers' purchasing behavior.

**K-Means**

Two distinct techniques were employed to find out the best number of clusters. The first method was the elbow method, which comprised plotting the within-cluster sum of squares (WSS) against the number of clusters and identifying the point where the rate of decrease in WSS slows down as the ideal number of clusters. Using this method, the optimal number of clusters was determined to be five, as illustrated in Figure 5.

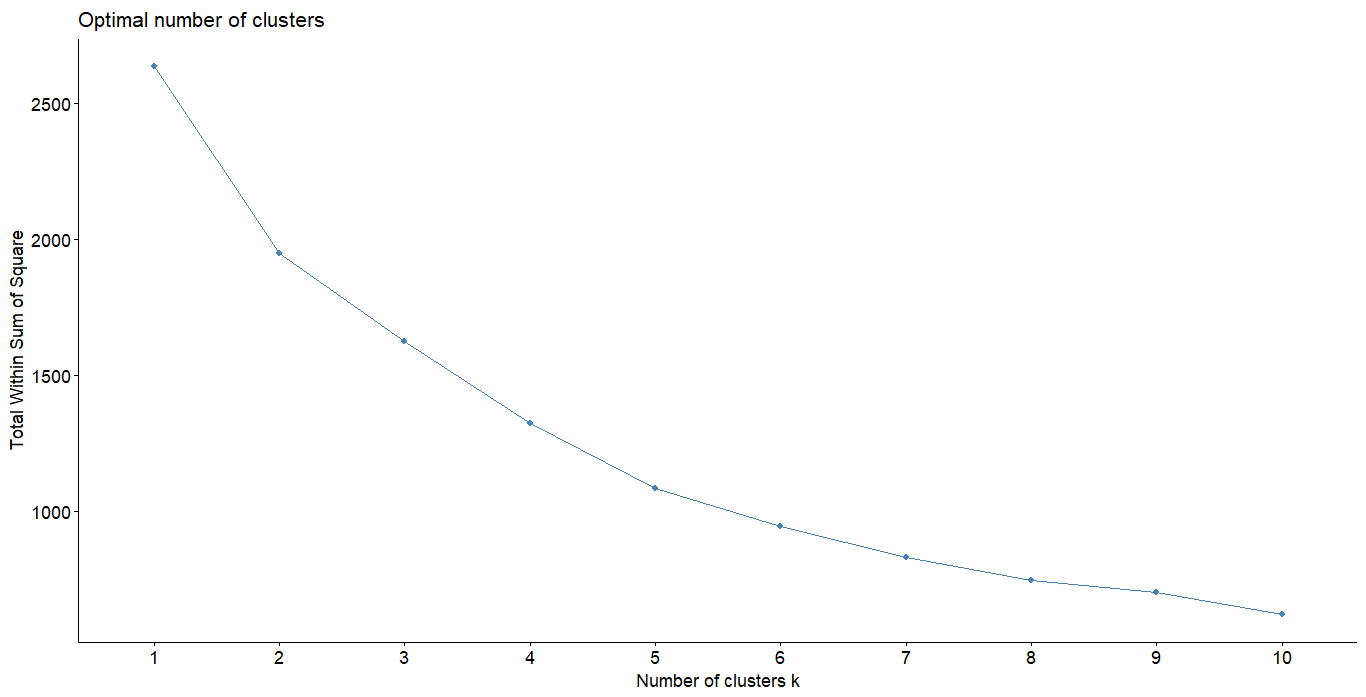


Figure 5. Elbow method

The second method we used to determine the optimal number of clusters was the gap statistic method, which compares the WSS of the clustering solution with the expected WSS of a reference distribution generated by a null model. The optimal number of clusters according to this method was found to be three. (Figure 6)

Изображение выглядит как диаграмма

Автоматически созданное описание

Figure 6. Gap statistic method

After determining the optimal number of clusters, we used the K-means algorithm to perform clustering on the scaled data. We performed K-means clustering for both K = 3 (Figure 7) and K = 5 (Figure 8).

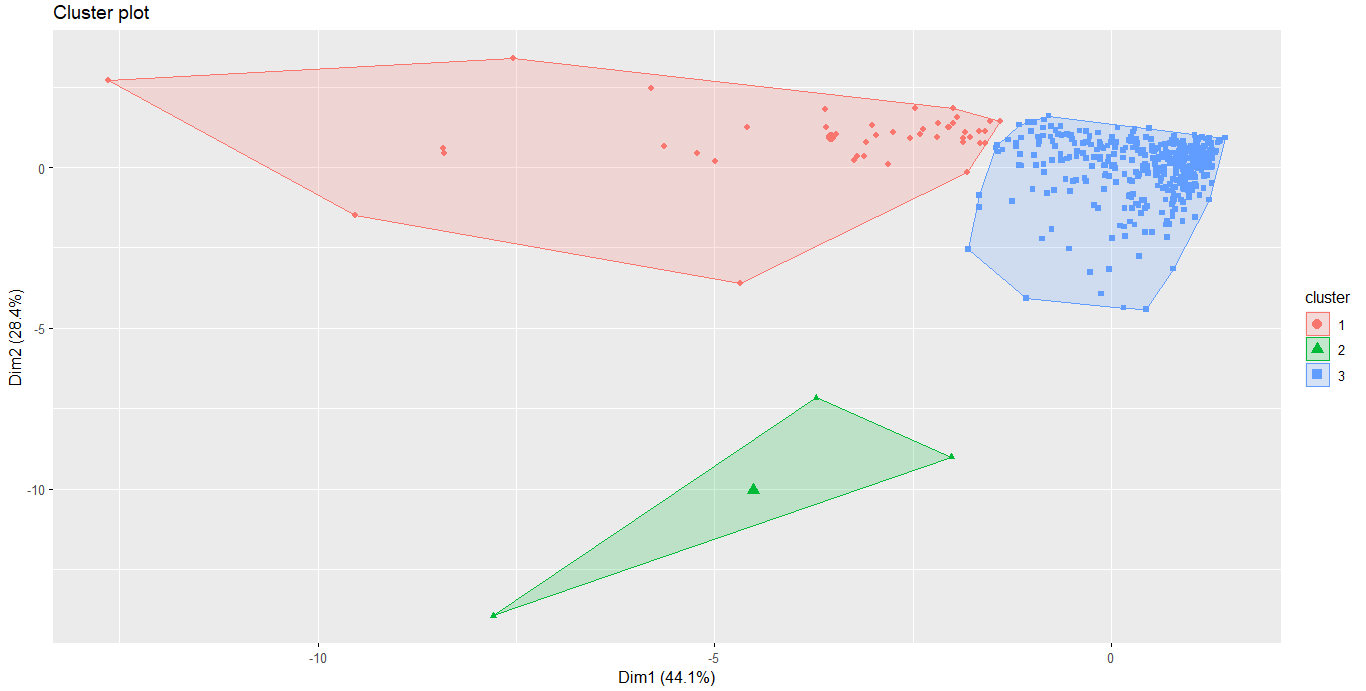
****

Figure 7. K-Means for K=3

**Изображение выглядит как диаграмма

Автоматически созданное описание**

Figure 8. K-Means for K=5

So K =3 was chosen as the optimal cluster size for our analysis. This choice will separate the high variation observations into a separate group, which can include potential high-spending customers.

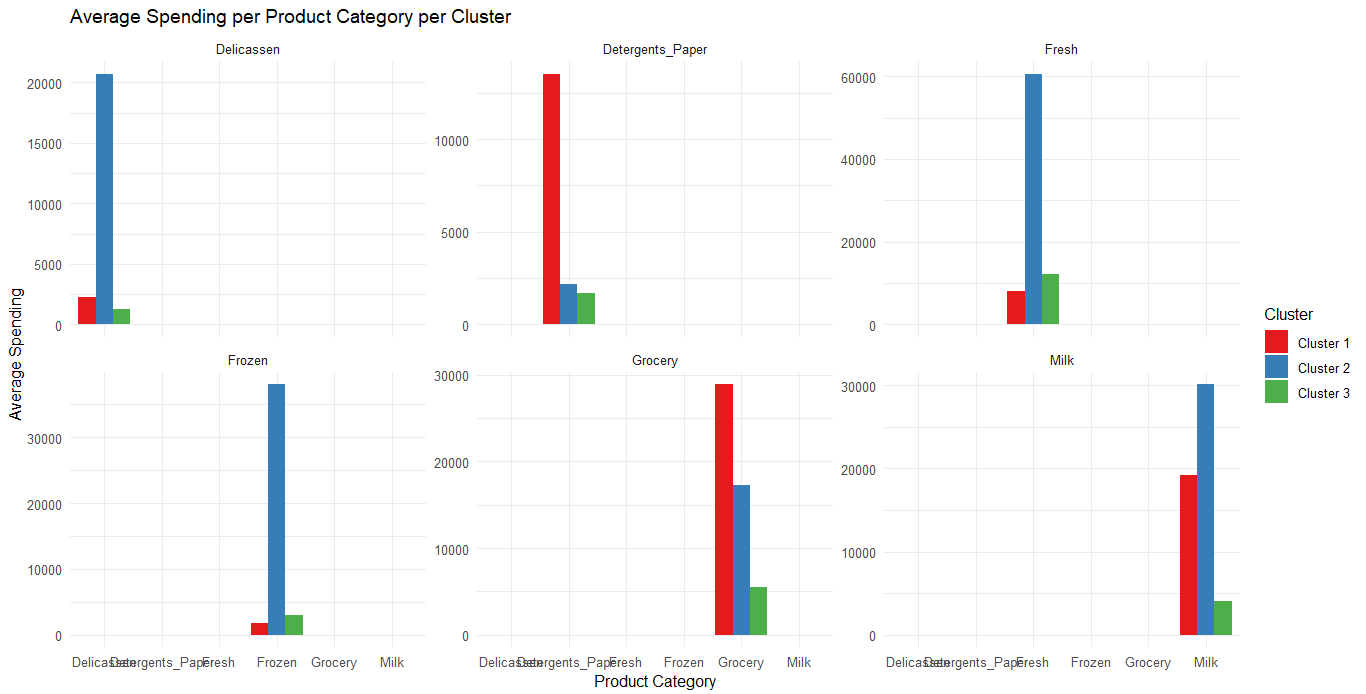


Figure 9. Visualization of the average spending per product category per cluster

Изображение выглядит как диаграмма

Автоматически созданное описание

Figure 10. Visualization of average product spending per customer segment

**Results**

The clustering analysis results indicate three distinct clusters among our customers. The size of each cluster is 44, 3, and 393 customers, respectively. The cluster means for each attribute are as follows:

*Cluster 1* includes only Retail customers who spend a lot on Groceries, Milk, Detergents\_Paper, and Fresh products. These customers prioritize buying household staples for personal use.

*Cluster 2* has a few Hotel/Restaurant/Cafe customers who spend more on Fresh and Frozen products, followed by Milk and Groceries.

*Cluster 3* comprises mainly Hotel/Restaurant/Cafe customers and some Retail customers who spend reasonably on Fresh products, followed by Groceries and Milk, but very little on Detergents\_Paper and Delicassen compared to other groups. These customers prioritize purchasing fresh food for commercial use.

Overall, this analysis provides valuable insights into our customers' spending habits and enables us to categorize them based on similar behaviours. This information can help in developing targeted marketing strategies and enhancing customer satisfaction.

**PCA**

One common unsupervised learning technique is Principal Component Analysis (PCA), which is used to identify the essential variables in a dataset and reduce its dimensionality.

To be more precise, the PCA examination revealed that the first three principal components captured a significant portion of the overall variance in the data, with a total explained variance of 84.8%. Among the three components, the first principal component was responsible for the largest amount of variability (44.1%), followed by the second component (28.4%) and the third component (12.3%). The scree plot, a graphical tool used to determine the number of principal components to keep, supported retaining these three components..

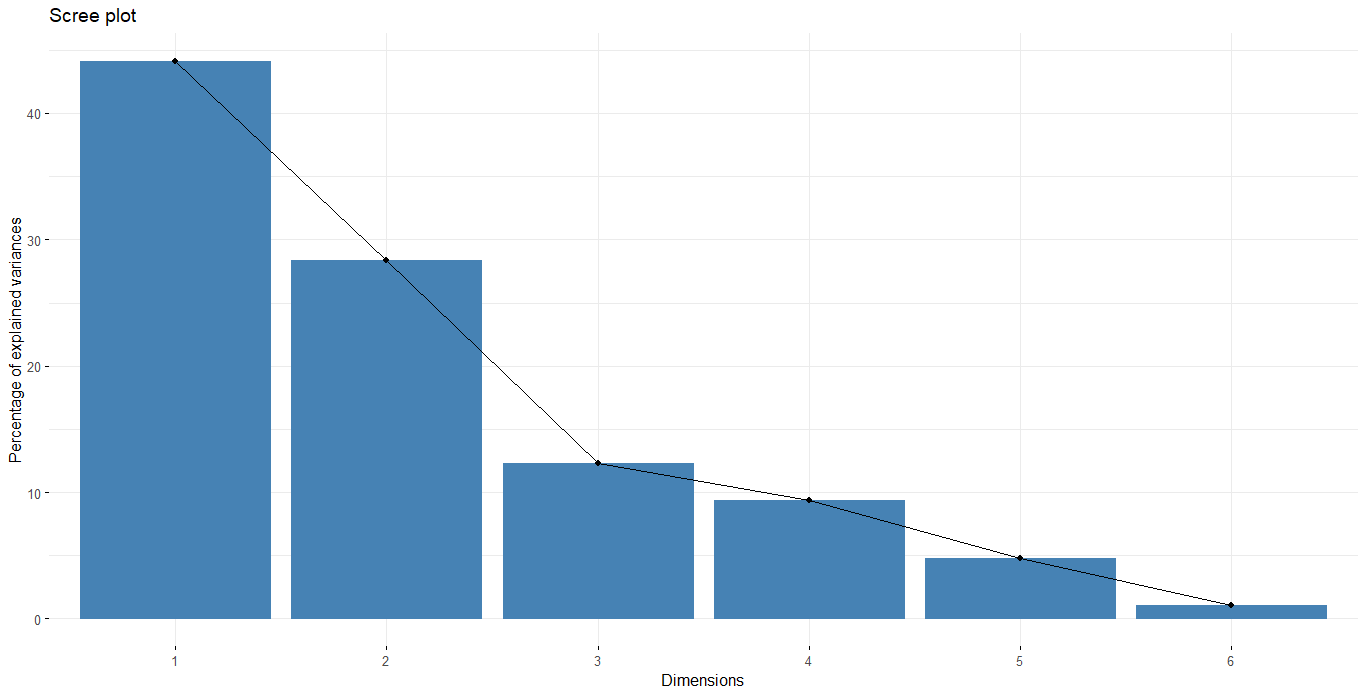


Figure 11. Scree plot to determine the number of principal components to retain

Examining the individual values, it was identified that customers who spend more on Milk, Grocery, and Detergents\_Paper tend to have higher scores on the first principal component, while those who spend more on Fresh and Frozen tend to have lower scores. On the second principal component, customers who spend more on Fresh, Frozen, and Delicassen have higher scores, while those who spend more on Milk, Grocery, and Detergents\_Paper tend to have lower scores. Finally, on the third principal component, customers who spend more on Delicassen tend to have higher scores, while those who spend more on Fresh, Frozen, and Detergents\_Paper tend to have lower scores.

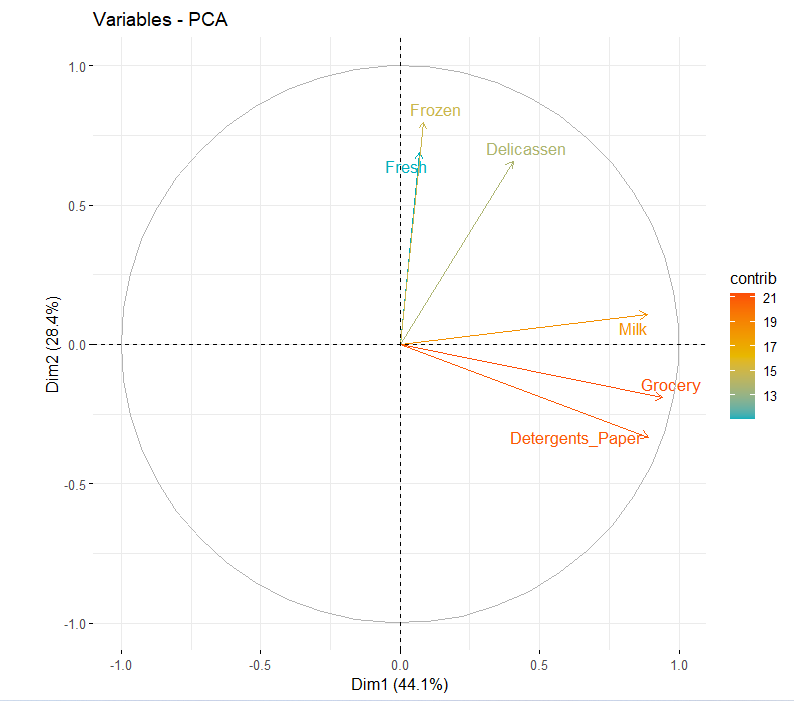


Figure 12. Visualization of the principal components

Изображение выглядит как диаграмма

Автоматически созданное описание

Figure 13. PCA plot of customer segment

Overall, this analysis demonstrates that three key factors drive customer spending habits, with different customers having different priorities across these factors. These insights can inform marketing and sales strategies that better target specific customer segments.

**Regression analysis on the Wholesale Customers Dataset**

We performed a regression analysis on the Wholesale Customers dataset using linear regression, decision trees, and random forests. This part aimed to predict the Total\_Spending variable based on the other variables in the dataset.

We first created a new variable called Total\_Spending, the sum of the other variables in the dataset. We then split the data into training and testing sets and fit the models to the training data.

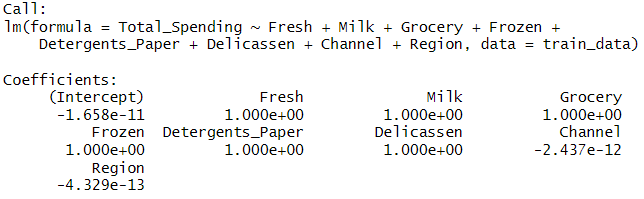


Figure 14. Summary of the Linear Regression of the WholeSale Customers Dataset

The***linear regression model***performed well with an RMSE of 2.27074e-11 and an R-squared of 1. This indicates that the model perfectly fits the data and can explain all the variation in the Total\_Spending variable. This could be since the Total\_Spending variable is simply the sum of the other variables, which are highly correlated with each other.

The ***decision tree and random forest models'*** performance was evaluated compared to the linear regression model. On the other hand, the decision tree and random forest models produced higher RMSE values of 18520.92 and 10447.96, respectively, suggesting that they could not fit the data as accurately as the linear regression model. However, it is essential to note that both models performed reasonably well, with the random forest model performing slightly better than the decision tree model.

Overall, adding the total\_spending variable improved the performance of all three models, as it provides a more comprehensive measure of customer spending behaviour than any individual featureThe tree-based models were less precise in capturing the connection between the features and the target variable than the linear regression model. This could be since there was a strong linear correlation between the features and the target variable.

**Regression analysis on the E-Commerce Dataset**

It was decided to perform a regression analysis also on the Ecommerce Dataset. It provides data on customers who buy clothes online from a store that offers in-store style and clothing advice sessions. The dataset consists of several variables: Avg. Session Length, Time on Website, Time on App, Length of Membership, and Yearly Amount Spent.

Using linear regression, data was splitted into training and testing sets, builted a linear model based on the training data, and predicted the Yearly Amount Spent using the other variables. The resulting model had an R-squared value of 0.9881, indicating that it explains 98.81% of the variability in the data. Furthermore, when the model was tested on the testing data, it had a relatively low root mean squared error (RMSE) of 8.56, suggesting that it accurately predicted the Yearly Amount Spent.

The analysis found a strong positive relationship between Yearly Amount Spent and Average Session Length, Time on App, and Length of Membership. However, there was no significant relationship between Yearly Amount Spent and Time on Website. The model's coefficients showed that increasing Average Session Length by one minute is associated with a $25.74 increase in Yearly Amount Spent, while increasing Time on App by one minute is associated with a $38.95 increase, and increasing Length of Membership by one year is associated with a $61.48 increase.

It is important to note that the coefficients represent the average effect of each variable on Yearly Amount Spent, holding all other variables constant. Other factors, such as the type of clothing and store pricing strategies, were not included in this analysis and may affect Yearly Amount Spent.

Based on these findings, the company may want to focus on improving the mobile app experience, as it appears to have a stronger positive impact on Yearly Amount Spent than the website. In addition, they may also want to invest in strategies that increase Average Session Length and Length of Membership, as these variables have a significant positive relationship with Yearly Amount Spent.

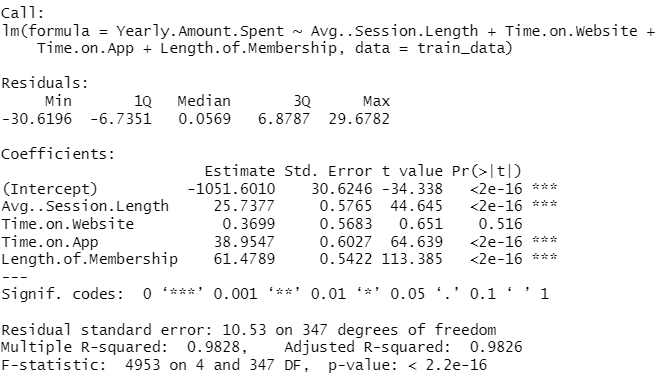


Figure 15. Summary of the Linear Regression for E-Commerce Dataset

**Classification**

The logistic regression model was used to predict the Channel variable (Retail or Horeca) based on the other variables in the Customers dataset. The dataset was splited into a training set (70% of the data) and a test set (30% of the data).

The logistic regression model results showed that several variables were significant in predicting the channel variable. The intercept, RegionOporto, and RegionOther significantly impacted the predicted channel variable, while Fresh, Milk, Grocery, Frozen, and Delicassen were insignificant. The coefficients for the significant variables indicate that the RegionOporto and RegionOther regions were more likely to be associated with the Horeca channel than the Retail channel. Additionally, higher spending on Detergents\_Paper was found to be associated with the Retail channel. To identify the best model, the stepwise variable selection was used. This method involves iteratively removing variables until the best subset is found. Then, the backward direction was chosen. The final model included Region, Frozen, and Detergents\_Paper variables and had an AIC value of 155.78. This suggests that the model could accurately predict the Channel variable for the customers in the test set. Also, the confusion matrix showed that out of 132 observations in the test set, the model correctly predicted 120 (87 Horeca and 33 Retail). This gives an accuracy of 90.91%.

Overall, the model performs well, with an accuracy of 90.91% and a Kappa value of 0.7822, indicating substantial agreement beyond chance.

Изображение выглядит как диаграмма

Автоматически созданное описание

Figure 16. Distribution of Probability Prediction Data

Also, four classification algorithms were evaluated: *K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Linear Discriminant Analysis/Quadratic Discriminant Analysis (LDA/QDA).*

**K-Nearest Neighbors (KNN)** is a non-parametric classification algorithm that classifies a new observation based on the majority class of its k nearest neighbours. We perform KNN classification on our data, varying the number of neighbours (k) and using cross-validation to tune the model. The optimal value of k is found to be 9. We predict the test data using this model.

**Logistic Regression** is a parametric classification algorithm that models the probability of an observation belonging to a particular class. We also performed logistic Regression with ridge regularization, using cross-validation to tune the regularization parameter.

**Naive Bayes** is a parametric classification algorithm that models the joint probability distribution of the predictors given the class. We perform Naive Bayes classification on our data.

**Linear Discriminant Analysis/Quadratic Discriminant Analysis (LDA/QDA)**

LDA and QDA are parametric classification algorithms that model the class-conditional densities of the predictors given the class.

**Results:** Among the four models, KNN achieved the highest accuracy (0.906) and Kappa (0.785), making it the best-performing model for predicting the customer channel. Naive Bayes was the second-best performing model, with an accuracy of 0.885 and Kappa of 0.7251. Logistic Regression and LDA/QDA performed slightly worse, with Logistic Regression achieving an accuracy of 0.8473 and Kappa of 0.6157 and LDA/QDA reaching an accuracy of 0.8414 and Kappa of 0.5887.

**Conclusion:** It is possible to accurately predict which channel a customer will use based on their spending behaviour. The K-Nearest Neighbors and Naive Bayes algorithms achieved the highest accuracy and Kappa, respectively, and are recommended for predicting the customer channel. The other models, Logistic Regression and LDA/QDA did not perform as well as KNN and Naive Bayes.