Data Analysis Project

Supermarket Sales Data Analysis

Data Analysis: techniques and tools - Academic Year 2025/2026

Calonghi Vittoria, Cristofolini Claudia, Di Palma Benedetta

January 22, 2026

Contents

1 Introduction

1.1 From data to knowledge

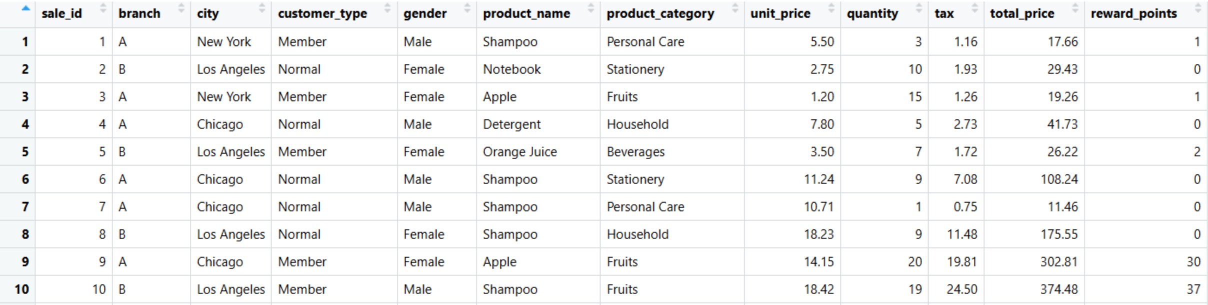
This report shows the analysis of a dataset extrapolated from Kuddle about Retail Sales. The goal of this report is to turn raw Data into useful Knowledge. According to the course criteria, data consists of objective facts (transactions), while knowledge arises from finding patterns and relationships through systematic analysis. We will evaluate the dataset based on criteria such as relevance, consistency, and reliability to make sure that the derived knowledge is valid. This analysis aims to extract meaningful patterns from the retail environment to provide actionable insights for business strategy.

1.2 Dataset description

The dataset analyzed, "sales.csv", contains records of 1.000 transactions from a retail activity across two supermarket branches where the sale occurred (named A and B) in three major cities: New York, Los Angeles and Chicago. The data structure consists of 12 variables, classified as follows: e consists of 12 variables, classified as follows:

- nominal attributes (categorical): *branch*, *city*, *customer_type* (Member or Normal), *gender*, *product_name* and *product_category*.
- numerical attributes: *sale_id*, *unit_price*, *quantity*, *tax*, *total_price* and *reward_points*.

Sale_id refers to a unique sales identifier for each transaction; *tax* considered 7% sales tax for each product; *customer_type* refers to member customers (customers who adopt a loyalty card) and normal customers (who do not); and *reward_points* are earned only by member customers based on the total transaction amount.



	sale_id	branch	city	customer_type	gender	product_name	product_category	unit_price	quantity	tax	total_price	reward_points
1	1	A	New York	Member	Male	Shampoo	Personal Care	5.50	3	1.16	17.66	1
2	2	B	Los Angeles	Normal	Female	Notebook	Stationery	2.75	10	1.93	29.43	0
3	3	A	New York	Member	Female	Apple	Fruits	1.20	15	1.26	19.26	1
4	4	A	Chicago	Normal	Male	Detergent	Household	7.80	5	2.73	41.73	0
5	5	B	Los Angeles	Member	Female	Orange Juice	Beverages	3.50	7	1.72	26.22	2
6	6	A	Chicago	Normal	Male	Shampoo	Stationery	11.24	9	7.08	108.24	0
7	7	A	Chicago	Normal	Male	Shampoo	Personal Care	10.71	1	0.75	11.46	0
8	8	B	Los Angeles	Normal	Female	Shampoo	Household	18.23	9	11.48	175.55	0
9	9	A	Chicago	Member	Female	Apple	Fruits	14.15	20	19.81	302.81	30
10	10	B	Los Angeles	Member	Male	Shampoo	Fruits	18.42	19	24.50	374.48	37

Table 1: First rows of the Dataset

2 Data quality and preparation

First, we prepared the data by verifying that the dataset did not contain missing values which would compromise our analysis. Then we converted all categorical variables into factors to allow a correct exploratory data analysis, enable supervised models such as decision trees to work properly, and improve data visualization by making it easier to order and compare categories.

3 Exploratory Data Analysis (EDA)

We begin the EDA following the methodology outlined in the course, using different types of graphs to better visualize data and identify some insights. We consider the fact that the dataset has not real data and try to answer these three questions (from a business point of view):

- Where do we sell best? (Geographic Analysis: City & Branch efficiency)
- Who are we selling to? (Customer Profiling: Gender & Loyalty Status)
- What are we selling and what is it worth? (Merchandising Analysis: Product Volume vs. Revenue Contribution)

3.1 Summary statistics and KPIs

Starting with the summary statistics, we display an overall view of data into a set of Key Performance Indicators (KPIs) that we typically use in Business Courses. These indicators show a Total Revenue of about €118.584, with an Average Order Value of €118,6, indicating a relatively consistent spending pattern across transactions. A total of 10.337 units were sold, with an average of 10.3 units per sale, suggesting that customers often purchase multiple items per transaction. The maximum value of a single sale (€433.99) highlights the presence of high-value transactions. Moreover, customers accumulated 6,057 reward points, reflecting active participation in the loyalty program.

3.2 Geographical and Demographic Distribution

We analyzed sales performance per city, identifying the quantity sold in each city, divided into Branches. From the plot it is evident that Chicago and Branch A are the ones with the highest sales performance. In the code, while creating the graph, we also include the possibility of having more than one branch per city, because, in a real context, the code would be already set-up and work if the dataset would be expanded (e.g., New York opens a branch B).

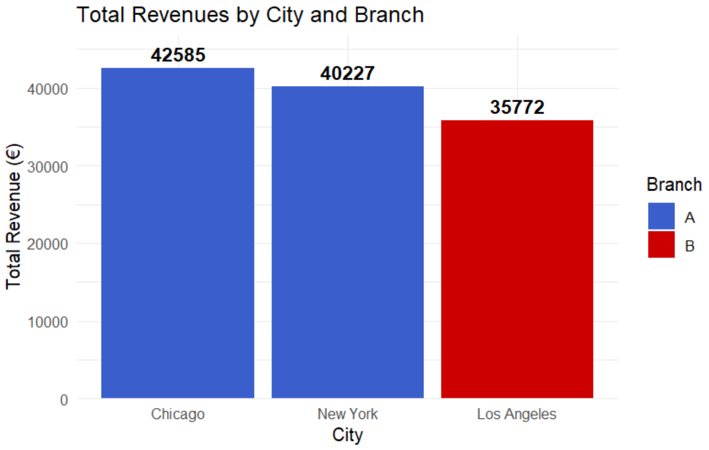


Figure 1: Total Revenues by City and Branch

3.3 Customer Demographics and Loyalty Program

To identify the target audience we analyzed the customer base through two main dimensions: gender and membership status.
The analysis of sales volume by gender reveals a balanced distribution between male and female customers (Figure 2). This suggests that the products appear almost equally to female and male customers, implying that marketing strategies should focus more on ”universal” needs rather than gender-specific targeting.

Regarding the Loyalty Program (Figure 3): we examined the ratio between ”Member” and ”Normal” customers. The graph shows a near-even split (approximately 50/50), meaning that a solid 51.6% of customers adopted the loyalty card: this is useful for supermarkets that can collect this data for conducting purchasing habits campaigns.

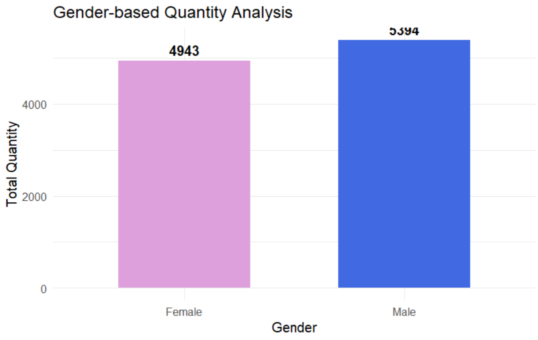


Figure 2: Total Quantity sold based on gender

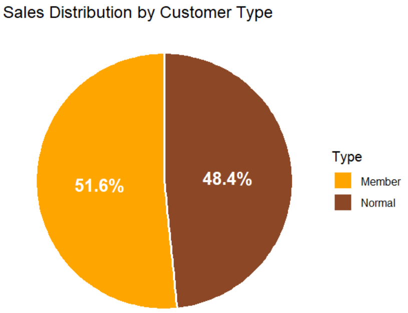


Figure 3: Sales distribution by customer type

3.4 Product and Category Performance

In this part of the analysis we focus on the merchandise mix. We distinguished between Operational Performance (Sales Volume) and Financial Performance (Revenue Contribution) to uncover deeper insights into profitability.

By analyzing the ”Top-selling Products by Quantity” (Figure 4), we identified the items with the highest rotation: Shampoo and Orange Juice. Understanding these high-velocity items is crucial for inventory management to prevent stock-outs that could lead to lost sales.

The results are reflected into Top-selling product categories, where the most sold categories are Fruits and Personal care. However, digging into the product categories, we found a relevant contrast between sales popularity and actual financial impact, specifically when comparing Fruits and Personal Care (Figure 5 & 6). Even though Fruits has the highest sales volume (2,286 quantities), it ranks second in terms of value (22.1%); while Personal Care sold slightly fewer units (2,278) but generates the highest revenue contribution (22.8%). Even if the difference is minimal, this insight indicates that, from a strategic perspective, Personal Care is more efficient because it generates more revenue with less logistical effort.

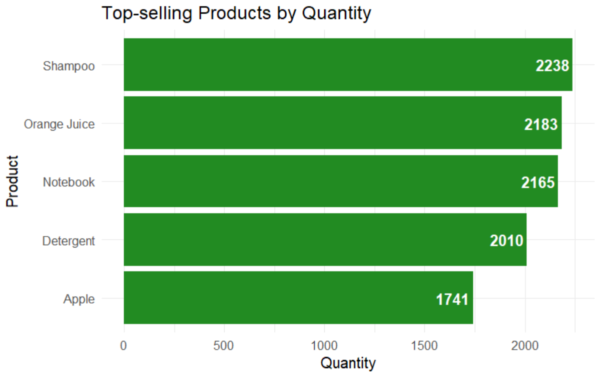


Figure 4: Top-selling Products by Quantity

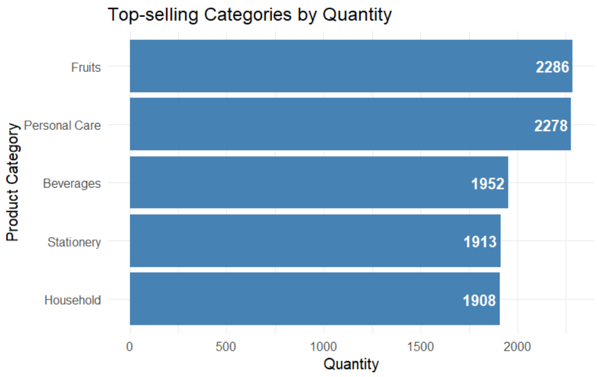


Figure 5: Top-selling Categories by Quantity

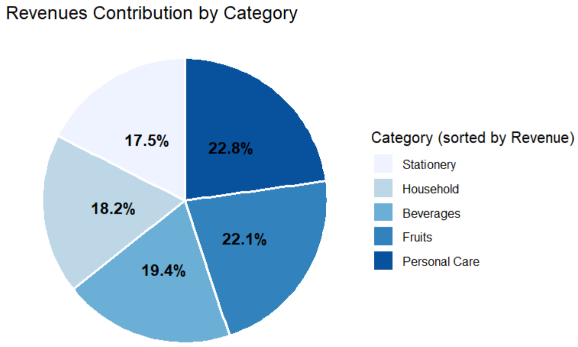


Figure 6: Revenues Contribution by Categories

3.5 Correlation Analysis

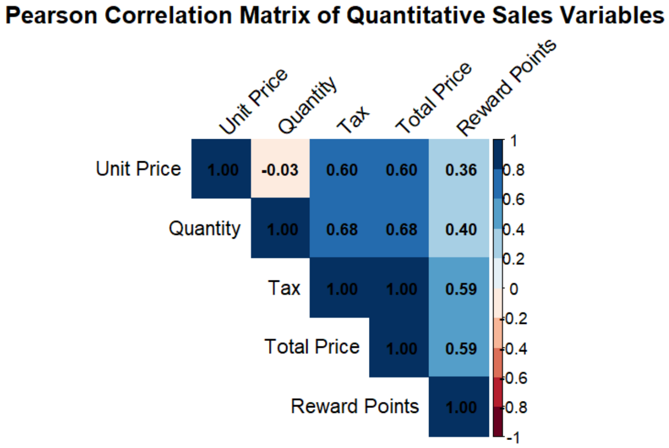
We performed the Pearson correlation analysis in order to observe the relationships among numerical variables, visualizing through a Heatmap (Figure 7). First, we selected just numeric variables, excluding *sales_id* (which is an index), then we created a matrix to test the correlation between *Unit price*, *Quantity*, *Tax*, *Total price* and *Reward points*, using colors to highlight a stronger correlation (Blue and Red) or a weaker correlation (lighter colors like White) between variables.

The matrix highlights three key findings:

1. The perfect positive correlation (r = 1) between *Tax* and *Total Price* confirms the data integrity, as tax is a fixed percentage of the price.

2. A strong positive correlation between *Total Price* and *Reward Points* confirms that the loyalty program is based on how much customers spent: the more they spend, the more they earn points.

3. The most relevant regards the near-zero correlation between *Unit Price* and *Quantity*. This indicates the absence of a strong linear relationship between the two variables. This may suggest that customer demand is relatively insensitive to price changes, or that price effects differ across Branches and Cities and therefore cancel out when data are aggregated. Overall, purchases appear to be driven more by necessity or store-specific factors than by price alone.



Pearson Correlation Matrix of Quantitative Sales Variables

	Unit Price	Quantity	Tax	Total Price	Reward Points
Unit Price	1.00	-0.03	0.60	0.60	0.36
Quantity		1.00	0.68	0.68	0.40
Tax			1.00	1.00	0.59
Total Price				1.00	0.59
Reward Points					1.00

Figure 7: Pearson Correlation Matrix

To validate the correlation matrix results, we visualized the relationship between *Unit Price* and *Quantity* through a scatter plot (Figure 8) where branches are in red or blue and cities are represented by different forms. The scatterplot shows that there is not a clear linear relationship between the two variables, and this result is coherent with the Pearson's correlation near to zero. The linear regression lines for Branches show a slightly different behavior: for branch A (city of Chicago and New York) the line is almost horizontal, confirming the weakness of the relationship between *Price* and *Quantity*; for Branch B the line is slightly negative, suggesting that an increase of *Price* would lead to a moderate decrease of *Quantity* purchased.

However, in both cases the slope is minimal, suggesting a low sensibility of demand with respect to price. In this scenario, price is not the principal driver of purchased quantity, and the purchasing behaviour could be influenced by other factors (e.g, customers' habits or marketing campaigns).

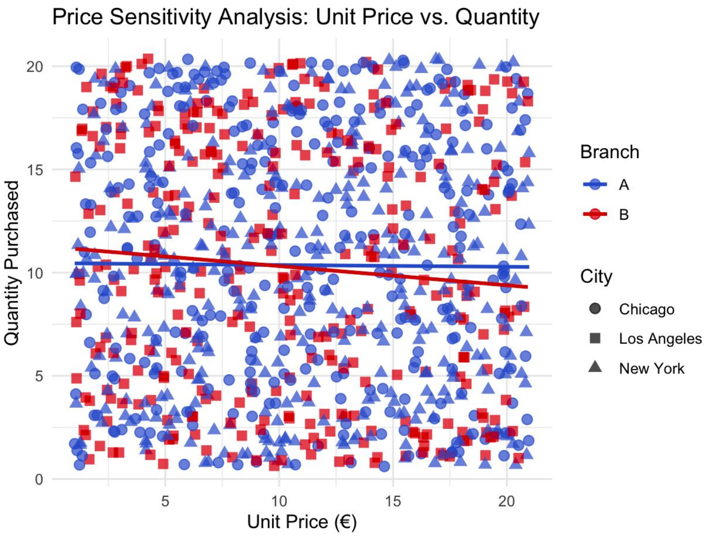


Figure 8: Price sensitivity Analysis between Price and Quantity

4 Clustering

Clustering is an unsupervised learning method that divides observations into similar groups based on specific similarity measures. In this analysis, two complementary clustering techniques have been applied: Hierarchical Clustering and K-Means Clustering. The dataset consists of supermarket sales transactions and includes both numerical and categorical variables. However, since distance-based clustering methods require numerical inputs, only quantitative variables are considered. To ensure fairness and prevent larger magnitude variables, such as *total_price*, from dominating the results, the numerical variables *unit_price*, *quantity*, *tax*, *total_price*, and *reward_points* were standardized using the scale() function prior to clustering. Standardization is essential because distance-based methods are highly sensitive to differences in scale. The clustering process uses the Euclidean distance metric, which provides a straightforward measure of dissimilarity in multidimensional space by calculating the straight-line distance between observation points.

4.1 Hierarchical Clustering

Hierarchical clustering builds a hierarchy of nested groupings by iteratively merging similar observations. For this analysis, an agglomerative approach was utilized: each observation initially starts as its own cluster, and clusters are merged step-by-step based on their proximity until a single cluster remains. One major advantage of hierarchical clustering is that it does not require specifying the number of clusters in advance. A dendrogram visually represents the process, showing how clusters merge at different levels of dissimilarity.

The heights at which clusters merge reflect their distances and are helpful for determining an appropriate number of clusters with meaningful interpretations.

4.1.1 Linkage Methods Comparison

Hierarchical clustering constructs a dendrogram by using linkage methods to define how the distance between clusters is measured. Four linkage strategies were evaluated to identify the most suitable approach for this dataset:

- Single Linkage defines the distance between two clusters as the smallest distance between any pair of points across clusters. While computationally efficient, it tends to create elongated chains in the dendrogram due to its "chaining" effect, leading to poorly defined clusters. This made single linkage unsuitable for capturing clearly separated customer segments in our analysis.

- Complete Linkage uses the maximum distance between points in different clusters, resulting in more compact and spherical clusters. The dendrogram shows improved separation with cluster heights extending up to approximately 8 units, indicating more compact, spherical clusters of similar size.

- Average Linkage computes the average distance between all pairs of observations from different clusters, offering a compromise between single and complete linkage. The dendrogram displays a well-balanced hierarchical structure with heights reaching approximately 4 units, providing clear cluster boundaries while maintaining reasonable within-cluster homogeneity.

- Ward's Method minimizes the total within-cluster variance at each merge step. The dendrogram shows a very clear separation, with high merge levels, indicating distinct groups. However, it assumes spherical clusters of equal variance and tends to create groups of similar size. Given the observed heterogeneity in transaction patterns, ranging from small routine purchases to large bulk orders, this assumption may impose excessive structural constraints.

4.1.2 Dendrogram Analysis

After comparing the dendrograms obtained using different linkage criteria, average linkage was selected as the most suitable method. This choice was motivated by its flexibility in capturing the natural variability of customer purchasing behavior, while providing clear vertical separation between clusters and maintaining overall structural stability. The dendrogram revealed three well-defined customer segments, and the tree was therefore cut at k=3 using the *cutree*() function, assigning each transaction to one of the three clusters. The dendrogram (Figure 9) illustrates the hierarchical relationships among observations, with red rectangles indicating the selected cutting level. The dendrogram was cut to obtain three clusters, representing an optimal compromise between interpretability and internal cohesion. A smaller number of clusters would have led to excessive aggregation and loss of information, while a higher number would have reduced clarity and practical usefulness. At this cutting level, each transaction is assigned to one of three distinct groups, resulting in 718 observations in Cluster 1, 134 in Cluster 2, and 148 in Cluster 3.

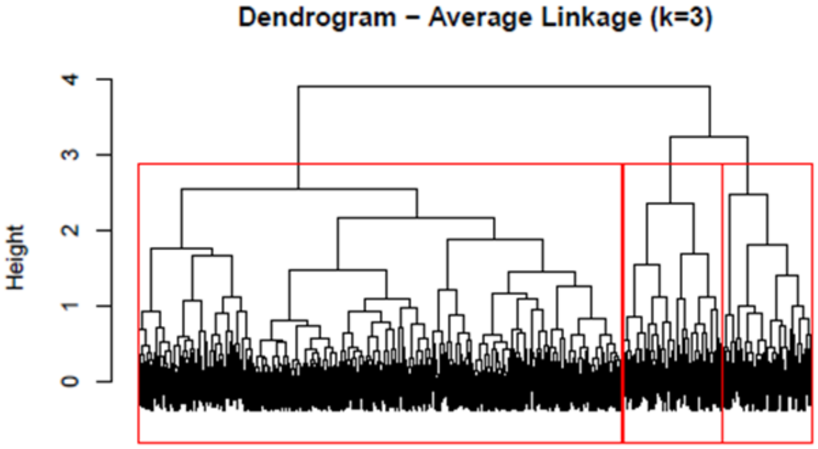


Figure 9: Dendrogram – Average Linkage (cut tree, k=3)

4.1.3 Cluster Distribution Analysis

To understand the composition of each cluster, we examined their distribution across cities and gender. To visualize the clustering results in the original feature space, we created scatter plots mapping quantity versus total price, with observations colored by cluster membership and shaped by city (Figure 10) and gender (Figure 11).

```
table(data[,c(3,13)])

  city         1    2    3
  Chicago      228  47   55
  Los Angeles  243  43   40
  New York     247  44   53
```

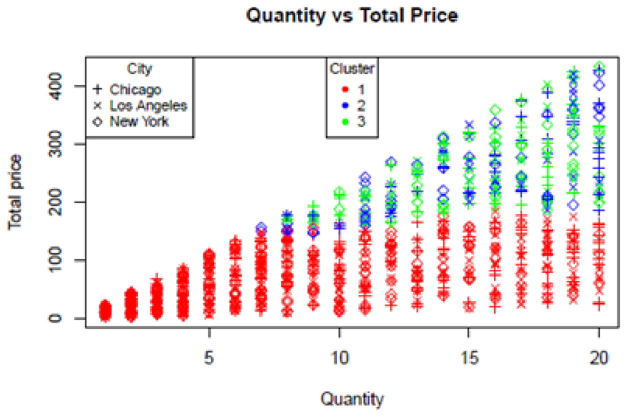


Figure 10: Scatterplot quantity vs. total price – city observations

```
table(data[,c(5,13)])

  gender    1     2    3
  Female    350   54   68
  Male      368   80   80
```

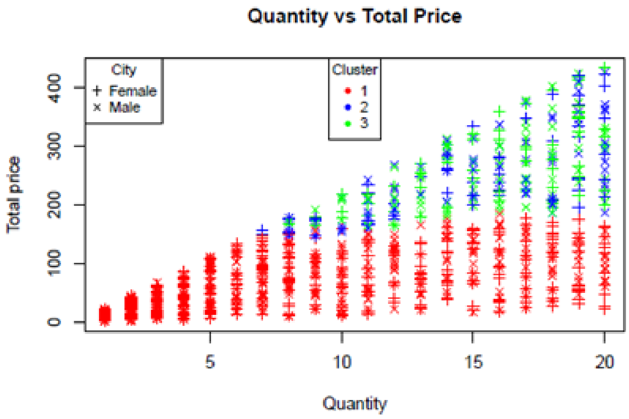


Figure 11: Scatterplot quantity vs. total price – gender observations

The scatter plot reveals three distinct purchasing patterns:

- Cluster 1 (red) concentrated in the lower-left quadrant with low to moderate quantities (1-15 units) and prices below €200. This represents everyday shoppers making routine purchases across all cities. The dense clustering suggests consistent, predictable buying behavior.

- Cluster 2 (blue) positioned in the upper-right corner with high quantities (12-20 units) and elevated prices (€250-400+). This segment represents bulk buyers or high-value transactions, showing strong purchasing power regardless of location. These customers likely represent either large households or small business purchases.

- Cluster 3 (green) occupies the middle region with moderate to high quantities (8-18 units) and prices ranging €150-300. This intermediate segment bridges routine shoppers and bulk buyers, potentially representing larger families or customers stocking up during promotional periods.

The overlap between clusters in the 10-15 unit range suggests that quantity alone is insufficient to discriminate customer segments; the total price dimension provides crucial additional information about product mix and unit values. All clusters are evenly distributed across cities and genders, indicating that segmentation is driven by purchasing behavior rather than demographics. This suggests that marketing strategies should prioritize behavioral factors, such as purchase volume and value, over demographic targeting.

4.2 K-Means Clustering

K-Means clustering represents an alternative partitioning approach that assigns observations to clusters by minimizing the within-cluster sum of squared distances from cluster centroids. The algorithm iteratively assigns each observation to the nearest cluster centroid and then recalculates centroids as the mean of all points in each cluster. This process continues until cluster assignments stabilize. Unlike hierarchical clustering, K-Means requires specifying the number of clusters *a priori*. Based on the hierarchical analysis, we set k = 3 clusters.

```
data_km <- kmeans(data_scaled, 3, nstart = 50)
```

The $nstart = 50$ parameter instructs the algorithm to execute 50 independent runs with random initial centroid positions, selecting the solution with the lowest total within-cluster variance. This mitigates the algorithm's sensitivity to initialization and increases the likelihood of finding the global optimum.

4.2.1 Cluster Characteristics and Centroids

The K-Means algorithm converged after 4 iterations ($data\_km\$iter = 4$), producing three clusters with the following sizes:

- Cluster 1: 256 observations
- Cluster 2: 309 observations
- Cluster 3: 435 observations

Each cluster in k-means is represented by a centroid, which corresponds to the mean value of each standardized variable within the cluster. The standardized cluster centroids reveal the defining characteristics of each segment:

	unit_price	quantity	tax	total_price	reward_points
Cluster 1	0.884	0.853	1.418	1.418	1.012
Cluster 2	-0.814	0.671	-0.254	-0.254	-0.249
Cluster 3	0.058	-0.978	-0.654	-0.654	-0.419

- Cluster 1 represents high-value, loyalty-engaged customers, purchasing expensive items in large quantities, resulting in elevated total price, tax, and reward points. It is the most profitable segment.

- Cluster 2 combines low unit prices with moderate quantities, indicating customers buying many low-cost items. Negative reward points suggest predominantly non-member shoppers making frequent, medium-value purchases, likely reflecting price-sensitive behavior.
- Cluster 3 shows very low quantities and near-zero unit prices, reflecting occasional or convenience shoppers with minimal basket sizes and low total sales.

The clustering quality is evaluated through variance decomposition:

```
Within cluster sum of squares: 896.85, 442.90, 707.92
Between_SS / Total_SS = 59.0%
```

Each within-cluster sum of squares measures the compactness of a cluster: the total squared distance of points from their centroid. Cluster 2 (442.90) is the most compact, while Cluster 1 (896.85) shows greater internal dispersion. The between-cluster variance accounts for 59% of total variance, indicating that the three clusters capture substantial heterogeneity in purchasing patterns. The remaining 41% represents within-cluster variability, suggesting reasonable homogeneity within segments while maintaining distinct separation between them.

4.2.2 K-Means Visualization and Interpretation

Hull Plot: Quantity vs Total Price (*Figure 12*)

The convex hull visualizations enclose all points within each cluster, providing a geometric perspective on cluster boundaries. The minimal overlap between hulls confirms good cluster separation, particularly between Cluster 1 and the other segments. The hull plot of quantity versus total price clearly illustrates the three-cluster separation. Cluster 1 occupied the upper-right area with high values on both dimensions, confirming these are heavy spenders. Cluster 2 concentrates in the lower-right region, and the widespread distribution suggests higher internal variability among low-spend customers. Cluster 3 dominates the lower-left corner with low quantities and low spending, representing the largest segment of small-basket shoppers. The centroid positions (marked as C1, C2, C3) precisely capture the central tendency of each cluster.

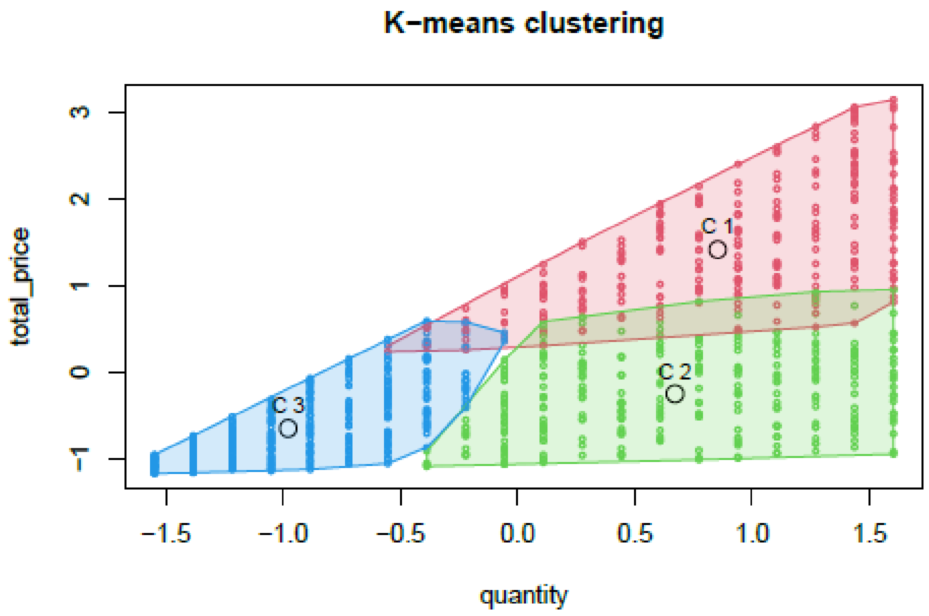


Figure 12: K-means clustering: quantity vs. total price

Hull Plot: Unit Price vs Quantity (*Figure 13*)

The hull plot of *unit_price* versus quantity reveals a different perspective on cluster structure. Cluster 1 maintained high quantity but also showed elevated unit prices, meaning these customers buy both premium products and larger volumes. Cluster 2 spread across lower unit prices while keeping positive quantities, which is consistent with value-seeking behavior. Cluster 3 stayed in the low quantity region across different price points, suggesting these customers consistently buy small amounts regardless of price level.

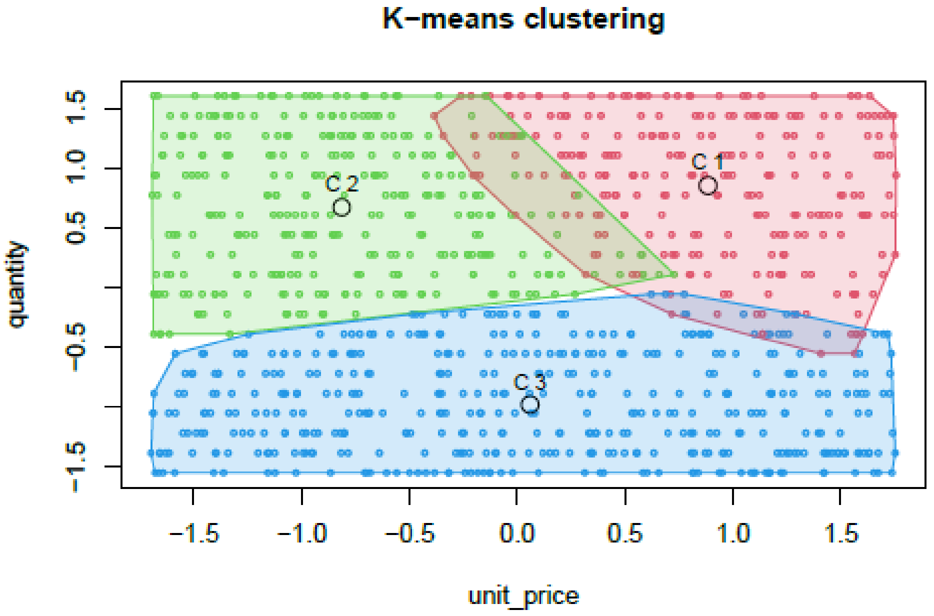


Figure 13: K-means clustering: quantity vs. unit price

4.3 Comparative Analysis

The clustering analysis identified three distinct customer segments within the supermarket transaction data using both hierarchical and K-means approaches. Hierarchical clustering, using average linkage, highlighted a clear segmentation based on transaction size. This approach categorized customers into routine shoppers (718 observations), medium-volume buyers (148 observations), and bulk purchasers (134 observations). Meanwhile, K-means clustering offered a complementary view of customer behavior, identifying high-value loyalty customers (256), price-sensitive frequent shoppers (309), and occasional convenience buyers (435). The three-cluster solution explained 59% of the total variance, adding depth to the segmentation insights.

The analysis shows that customer segmentation is driven primarily by purchasing behavior rather than demographics, indicating that marketing strategies should prioritize purchase volume, value, and frequency over traditional demographic factors.

5 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique focused on reducing the dimensionality of a dataset while preserving as much as possible of the original variance of the data. The resulting eigenvectors define the directions of maximum variance, known as principal components, while the corresponding eigenvalues quantify the proportion of total variance explained by each component.

5.1 Data preparation for PCA

PCA relies on the computation of variances and covariances between variables, which are not defined for categorical or non-numeric data. Therefore, PCA was applied only to numerical attributes *unit_price*, *quantity*, *tax*, *total_price* and *reward_points*. Before performing the analysis, the selected numeric variables were standardized using the *scale*() function. This step ensures that each variable contributes equally to the principal components and prevents variables with larger values (as *total_price*) from dominating the analysis.

5.2 PCA Implementation and Explained Variance

The PCA itself was implemented with the *prcomp*() function, which performs singular value decomposition (SVD) on the scaled data to compute the principal components, their standard deviations and the rotation matrix. To examine the proportion of variance explained by each component, the *get_eigenvalue*() function was used, providing both the individual and cumulative contributions of the principal components.

```
eig.value=get_eigenvalue(data.pca)
eig.value
        eigenvalue variance.percent cumulative.variance.percent
Dim.1 3.301643e+00     6.603287e+01                    66.03287
Dim.2 1.034011e+00     2.068021e+01                    86.71308
Dim.3 5.713521e-01     1.142704e+01                    98.14012
Dim.4 9.299369e-02     1.859874e+00                   100.00000
Dim.5 8.702953e-08     1.740591e-06                   100.00000
```

As can be observed from the results, the first three principal components together explain approximately 98% of the total variance, justifying the dimensionality reduction from five original variables to three principal axes without significant loss of information.

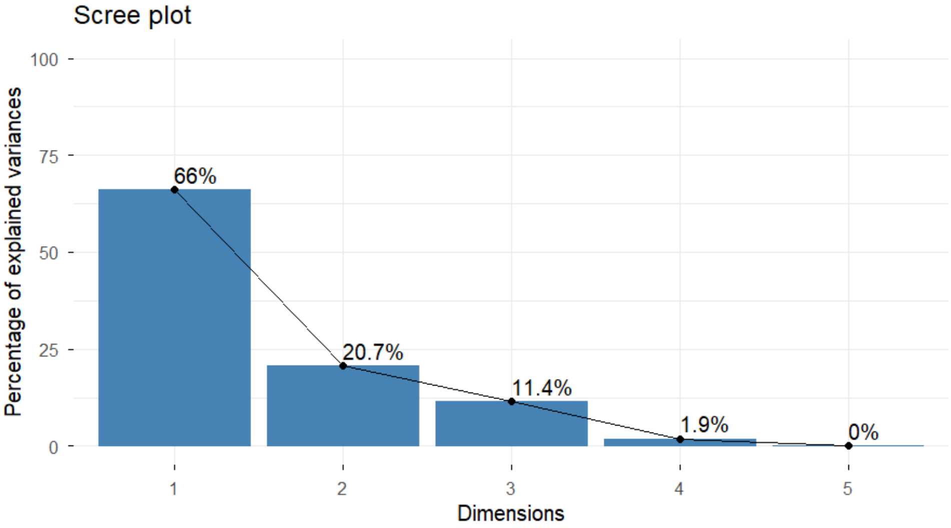


Figure 14: Scree plot of the variance

As can be seen from the scree plot, the first principal component explains 66% of the total variance, while the second and third components account for 20.7% and 11.4%, respectively. Beyond the third component, the marginal gain in explained variance becomes negligible, supporting the selection of the first three components.

5.3 Variable Contributions and Interpretation

The PCA variable *biplot*, generated using the *FactoMineR* package and visualized through the function *fviz_pca_var*(), illustrates the relationship between the original numerical features and the first two principal components, which together account for 86.7% of the total variance. The first dimension (Dim1) is primarily defined by *total_price*, *tax* and *reward_points* attributes showing the strongest positive correlations and highest contribution levels (indicated by the warmer orange shades). In contrast, the second dimension (Dim2) captures pricing and volume, where *unit_price* and *quantity* are extremely relevant.

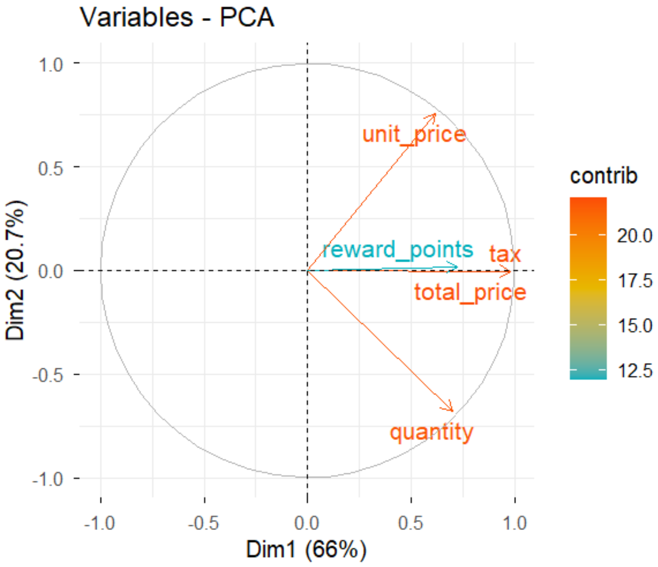


Figure 15: PCA variable *biplot*

The output of the function *get_pca_var*() gives us other specific information about the relationship be tween variables and the dimensions. In particular, the command $*contrib* indicates the contributions of the variables.

```
round((var$contrib[,1:5]/100),3)
              Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
unit_price    0.115 0.555 0.034 0.296   0.0
quantity      0.147 0.444 0.033 0.375   0.0
tax           0.290 0.000 0.046 0.164   0.5
total_price   0.290 0.000 0.046 0.164   0.5
reward_points 0.157 0.000 0.841 0.002   0.0
```

The first principal component *Dim*.1 is defined by the overall monetary volume of the transactions. This dimension is largely dominated by the variables *total_price* and *tax*, which together contribute nearly 60% to its definition. The second component *Dim*.2 is mainly driven by *unit_price* (55.5%) and *quantity* (44.4%), which together explain almost all of its variation. Finally, the third component *Dim*.3 is mostly driven by the *reward_points* variable, which contributes 84.1% to its construction.

5.4 PCA Projection and Cluster Analysis

Following the dimensionality reduction, the dataset was projected onto the subspace defined by the first three principal components *data.pca.*3. By utilizing these projected coordinates instead of the original raw variables, the hierarchical clustering effectively filters out noise. Using the *cutree* function, the observations were segmented into three distinct groups and the cluster assignments were integrated into the original dataset as the new variable *pca.*3.

To interpret these segments, we performed a comprehensive summary and cross-tabulation, analyzing the distribution of each cluster across cities and gender. The summaries reveal that Cluster 1 represents customers with the highest spending levels, Cluster 2 consists of occasional shoppers with smaller transaction sizes and Cluster 3 is characterized by strong engagement with the loyalty program. On the other hand, the cross-tabulation of the three clusters against demographic variables reveals a balanced distribution.

```
table(data$city,data$pca.3)
              1   2   3
  Chicago     242  27  61
  Los Angeles 258  15  53
  New York    264  21  59
```

```
table(data$gender,data$pca.3)
          1   2   3
  Female 369  31  72
  Male   395  32 101
```

Cluster 1 is the largest group and maintains a dominant presence in all three cities with New York showing the highest absolute frequency (264). Cluster 2 is the smallest segment in every city with its lowest presence in Los Angeles (15). Cluster 3 is well represented across all categories with Chicago leading in this category (61). While, clusters are mostly balanced between genders, except for cluster 3 which shows a higher concentration of male customers.

6 Decision Tree

To investigate the relationship between transaction features and their respective cities, we implemented a decision tree classification to predict the city of each transaction based on all available features. Initially, the necessary R packages *rpart* and *rpart.plot* were installed and loaded. These packages provide functions to construct recursive partitioning trees and visualize them effectively.

6.1 Training and test set

```
length(data)
nrow(data)
set.seed(2025)
data.idx=sample(1000,1000*.13)
data.train=data[data.idx,]
data.test=data[-data.idx,]
```

After ensuring reproducibility with *set.seed*(2025), the dataset was randomly split into a training set comprising 13% of the observations (130 rows) and a test set containing the remaining 87% of the data (870 rows). The training subset was used to fit a decision tree model using the *rpart*() function, with the dependent variable *city* and all other variables as predictors. The tree was visualized with *rpart.plot*(), providing a graphical representation of the decision rules used.

6.2 Decision tree construction and visualization

```
data.dc=rpart(city ~.,data=data.train)
data.dc
rpart.plot(data.dc)
```

The printed tree structure reveals hierarchical splits on variables such as *branch*, *product_category*, *unit_price*, *quantity*, *sale_id* and *reward_points*.

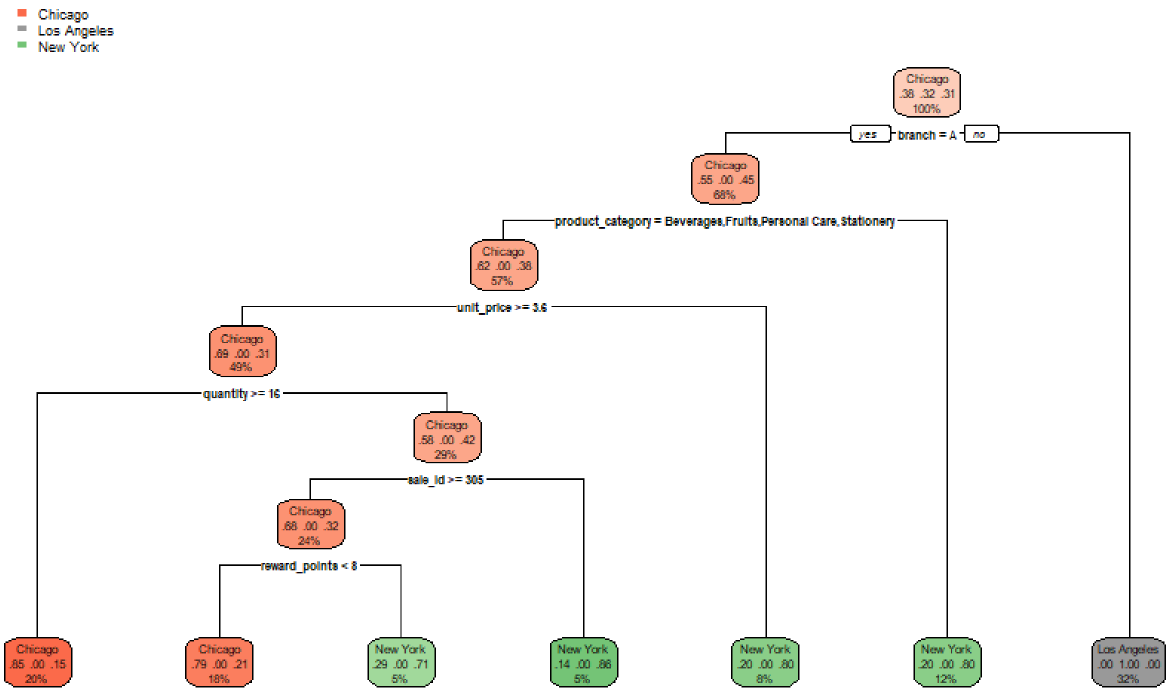


Figure 16: Decision tree classification

The hierarchy of the generated decision tree provides insights into the underlying structure of the data.

6.2.1 Root node

The model considers the entire training dataset of 130 observations, where the initial distribution is relatively balanced among Chicago (38%), Los Angeles (32%), and New York (30%). The attribute branch represents the most critical feature for the initial split. If the branch is B, the observation is classified with absolute certainty, whereas Branch A requires further attributes to resolve the classification.

6.2.2 First Level

The first level partitions the data based on the *branch* attribute. The branch B isolates the 32% of the total observations, where the probability of the city being Los Angeles reaches 100%, making this a terminal *pure* node. While branch A represents the 68% of the population. Within this subset, the probability of Chicago increases from 38% to 55%, while Los Angeles drops to 0% and New York rises to 45%.

6.2.3 Intermediate Nodes

For the Branch A population, the model evaluates *product_category*, where transactions involving Household items are strongly associated with New York at an 80% probability.

For other categories (Beverages, Fruits,...), the model performs a depth analysis of the transactional data. The attribute *unit_price* acts as a further filter: if the price is below 3.6, the probability of New York is 80%. If the price is higher, the model examines the variable *quantity*, suggesting that higher volume purchases ($quantity \geq 16$) are a strong indicator for Chicago with a 85% probability.

6.2.4 Last Level (Leaf Nodes)

The final level of the tree refines the classification using *sale_id* and *reward_points*. A relevant terminal node is reached when $sale\_id \geq 305$ and $reward\_points < 8$, which isolates a segment representing the 18% of the training data with a 79% probability for Chicago. However, if $reward\_points \geq 8$, the predicted class changes to New York with a 71% probability.

6.3 Accuracy of the model

```
data.dc.pred=predict(data.dc,data.test,type='class')
conf.matrix=table(data.test$city,data.dc.pred)
conf.matrix
             data.dc.pred
              Chicago Los Angeles New York
  Chicago         117           0      164
  Los Angeles       0         285        0
  New York        147           0      157

accuracy=sum(diag(conf.matrix)) /sum(conf.matrix)
accuracy
[1] 0.6425287
```

Predictions on the test set were generated using the $predict()$ function with $type =' class'$ to obtain discrete city labels. The resulting confusion matrix shows the distribution of predicted cities versus actual cities. The diagonal elements of the matrix represent the number of correctly predicted observations, while the remaining elements correspond to misclassifications. The model accurately classified all Los Angeles transactions (285/285), but the model showed variable performance for Chicago and New York: 117 Chicago observations were correctly classified, while 164 were misclassified as New York; similarly, 157 New York observations were correctly identified and 147 misclassified as Chicago. The overall accuracy of the model, computed as the ratio of correctly classified observations to total test samples, is approximately 64.25%. These results indicate that while the decision tree performs perfectly for Los Angeles, there is a substantial misclassification between Chicago and New York, suggesting similar patterns between these two cities that the current tree cannot fully separate.

7 Linear Regression

The aim of this linear regression analysis is to estimate and interpret the linear relationship between customer expenditure *total_price* and the number of loyalty points earned *reward_points*, while evaluating the effects of categorical variables such as *customer_type* and *product_category*.

7.1 Model construction

We constructed three nested models as follows.

```
model1=lm(reward_points ~ total_price, data = data)
```

Model 1 is a simple linear regression examining the direct effect of *total_price* on *reward_points*. This model estimates the baseline relationship between expenditure and points without considering any categorical segmentation.

```
model2=lm(reward_points ~ total_price + customer_type, data = data)
```

Model 2 represents an extended model adding *customer_type* to account for differences between *Member* and *Normal* customers. This allows the model to capture systematic differences in reward allocation based on membership category.

```
model3=lm(reward_points ~ total_price + customer_type + product_category, data = data)
```

Model 3 represents the full model and incorporates *product_category* to examine whether rewards differ across product category.

7.2 Statistical evaluation of model outputs

The summary of Model 1 revealed a statistically significant positive relationship between *total_price* and *reward_points* ($\beta \approx 0.055$, $p < 0.001$). The Multiple R-squared $R^2 = 0.35$ indicates that expenditure alone explains 35% of the variance in loyalty points. Moreover, the intercept was non significant ($p = 0.173$), which is consistent with the expectation that customers with zero expenditure earn zero points.

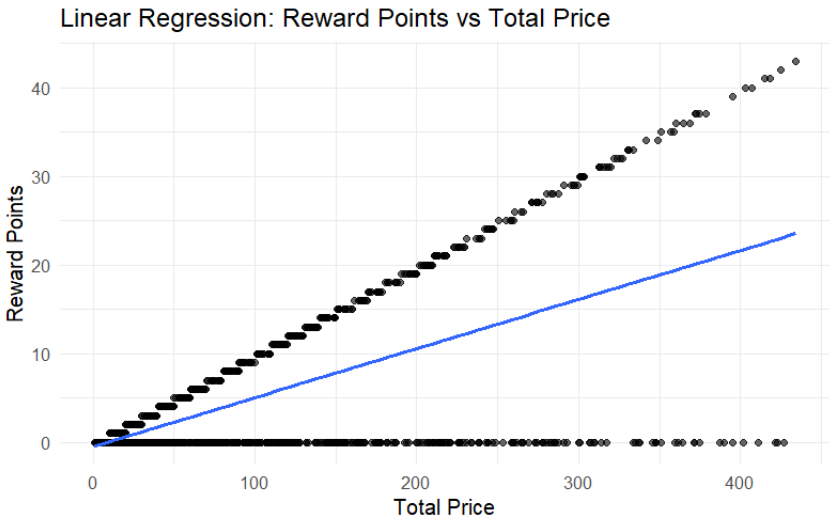


Figure 17: Model 1: $reward\_points \sim total\_price$

Including $customer\_type$ considerably improved model's goodness of fit. The Adjusted R-squared increased to 0.71, reflecting a higher proportion of variance explained rather than Model 1. The coefficient for $customer\_typeNormal$ is negative and highly significant ($\beta \approx -11.31$, $p < 0.001$), indicating that non members earn fewer reward points at any given expenditure level.

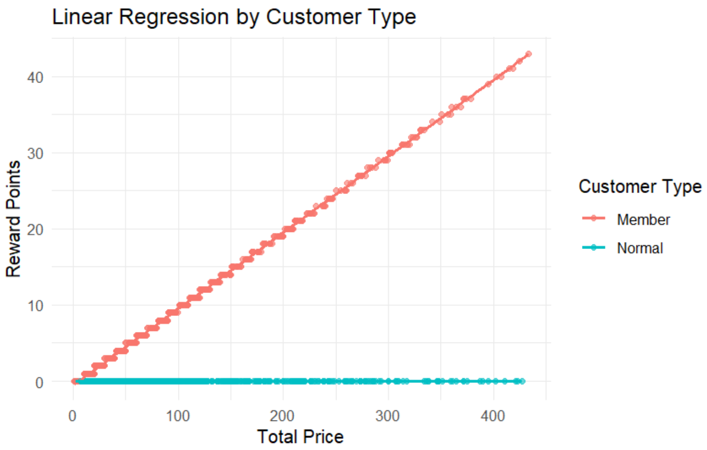


Figure 18: Model 2: $reward\_points \sim total\_price + customer\_type$

The full model, which incorporated $product\_category$, showed that categorical variables (Fruits, Household, ...) were not statistically significant ($p - values > 0.05$). Moreover,adding complexity results in a lower Adjusted R-squared (0.7137), suggesting that the additional variables do not improve the model's explanatory power. This shows that the reward system is applied uniformly across product sectors and product category does not contribute additional explanatory power beyond Model 2.

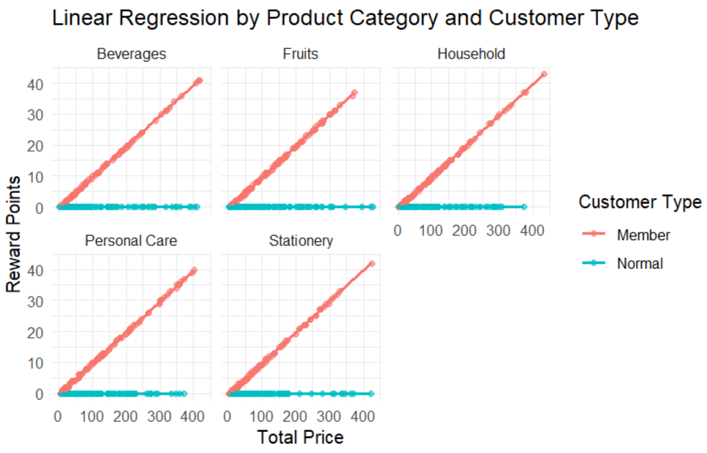


Figure 19: Model 3: $reward\_points \sim total\_price + customer\_type + product\_category$

7.3 Model Comparison: Analysis of Variance (ANOVA)

An Analysis of Variance (ANOVA) was performed to compare the three nested models and assess whether the increased complexity is statistically justified.

The ANOVA results confirm that moving from Model 1 to Model 2 is highly statistically significant ($F = 1273.87, p < 0.001$), whereas adding $product\_category$ in Model 3 does not provide a significant improvement ($p = 0.7269$). Consequently, Model 2 is identified as the most robust and efficient framework for explaining reward point distribution in our dataset, accounting for approximately 71.5% of the total variance.

7.4 Normal Q-Q plot

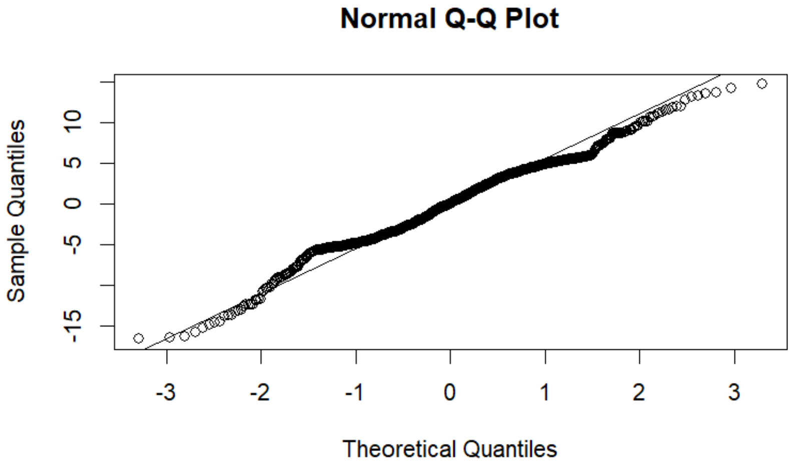


Figure 20: Normal Q-Q plot for model 2

Since *model 2* was identified as the best framework, we analyze its residuals to verify the model's reliability and to validate the assumptions of the linear regression. The Normal Q-Q plot for *model 2* exhibits a strong linear pattern along the diagonal reference line. This indicates that residuals generally follow the theoretical diagonal line, which confirms that error terms are approximately normally distributed. There are small deviations at the extreme tails, which may indicate outliers, but overall the residuals are close enough to the reference line to support the reliability of the estimated coefficients and p-values.

8 Conclusions

In conclusion, our exploratory data analysis highlighted the key patterns related to customer behavior, product performance and geographical distribution, in particular regarding sales volume and revenue value. Geographically Chicago and Branch A are the ones with the highest sales performance; while the large adoption of loyalty programs is useful to collect customers' info and conduct personalized marketing promotion and campaigns. The correlation analysis shows that while pricing variables are internally consistent and linked to rewards, the relationship between unit price and quantity is weak, suggesting that customer purchasing behavior is driven more by contextual or behavioral factors than by price alone. Here it is also important to remember that the dataset has some limitations regarding the truth of data: prices and quantities are not necessarily correlated for this reason. Another limit is the absence of temporal information or marketing promotion information that would affect results of our analysis.

The clustering analysis revealed three meaningful customer segments distinguished by purchasing behavior. Hierarchical clustering primarily differentiated customers by transaction size, while K-means highlighted differences in spending intensity and loyalty patterns. Both approaches confirmed that customer segmentation is driven more by behavioral factors than by demographic characteristics. These results underline the importance of focusing marketing strategies on transaction volume, spending behavior, and customer loyalty rather than relying on demographic information.

PCA effectively reduced the dimensionality of the dataset from five variables to three principal components, preserving approximately 98% of the total variance and filtering out noise for the subsequent clustering. We identified that customer behavior is primarily driven by the total spending, the balance between price and quantity and loyalty engagement (*reward points*). Significantly, PCA confirms that these behaviors are consistent across all cities and genders, showing that customers are more accurately characterized by their purchasing behavior than by their individual identity.

The Decision Tree classification demonstrated that transaction features can be effectively used to predict the city in which a sale occurs. The model performed very well for Los Angeles, while the observed misclassification between Chicago and New York indicates the presence of similar purchasing patterns across these two cities.

The linear regression analysis showed that total price and customer type are the main factors influencing the reward system with Model 2 explaining the 71.5% of the variance. The results indicate that non-member customers receive fewer points for the same level of spending, whereas product category is not statistically significant. In addition, the Normal Q–Q plot supports the reliability of the model by proving that the residuals are approximately normally distributed.

Finally, it is important to note that the dataset is not based on real data and has some limitations. In particular, it does not include time information or marketing and promotion variables, which limits deeper analysis. Despite these limits, the report shows how different data analysis techniques can be combined to extract useful insights from supermarket sales data.

References

[1] https://www.kaggle.com/datasets/chadwambles/supermarket-sales/versions/2/data