ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Maximizing Revenue: Insights into Customer Behaviour Across Supermarkets

(13)

Advanced Data Analysis

Student Giorgia Treglia

Identification Number 6042123

Academic Year 2024/2025

Contents

Introduction

The success of supermarkets is closely connected to their ability to understand customer behaviour and adapt that to their needs. The retail industry operates in a competitive environment where customer retention and satisfaction are key factors for driving revenue.

This report is based on a survey conducted to analyse the annual amount of money spent by customers in various supermarkets chains and to determine the factors that contribute to these expenditures. The dataset includes information about 18 large supermarket chains, customer spending, the use of a car for store access, customer evaluations of prices, the quality of fresh and packed products, queue lengths, the distance from the store to their homes, and demographic factors like gender. By leveraging this data, the aim is to uncover patterns and relationships that can guide supermarket management in making informed decisions to increase revenue.

This report specifically investigates whether there are differences in spending based on customer gender or price evaluations. Furthermore, it uses statistical models to analyse customer spending behaviour and identify actionable strategies for improving profit margins. Through this analysis, the goal is also to provide valuable insights for optimizing store operations and enhancing customer satisfaction. By studying pairwise relationships among variables, it would be easier to understand the dynamics of customer decision-making and their financial impact on supermarkets. Overall, the findings are expected to inform strategic decisions, such as pricing policies, store layout optimization, and improvements in product quality, helping supermarkets remain competitive in a challenging market.

Chapter 1

Descriptive and Exploratory Data Analysis

1.1 Summary Statistics

To gain a preliminary understanding of the dataset, a summary of the key variables is presented in *Table 1*. This analysis provides an overview of the distribution of variables, including customer demographics, spending behaviour and other characteristics. The summary statistics include the minimum, 1st quartile, median, mean, 3rd quartile, and maximum values for each variable.

Variable	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Shop	1	3	3	4.26	6	16
An. Spend	120	1,440	1,680	2,407	3,600	18,000
Car	0	0	1	0,71	1	1
Queues	1	4	5	4,37	5	5
Prices	1	4	5	4,51	5	5
Q. Fresh P.	1	5	5	4,57	5	5
Q. Product	1	4	5	4,34	5	5
Distance	1	2	2	1,88	2	4
Gender	0	0	0	0,22	0	1

Table 1 - Summary of Variables

The summary statistics provide a foundational understanding of the key variables influencing customer spending behaviour in supermarkets. The analysis highlights significant variability in annual spending, with a wide range from $120 to $18,000, suggesting diverse customer profiles and spending patterns. The high ratings for product quality, particularly fresh products, indicate customer satisfaction with offerings, while the relatively short queue lengths suggest efficient in-store operations. Demographic insights, such as the predominance of male customers and the frequent use of cars for store access, point to specific customer segments that supermarkets may target to optimize their strategies. Additionally, the generally short distances from customers' homes to stores imply convenience.

1.2 Univariate Analysis

The univariate analysis focuses on exploring the distribution of individual variables within the dataset. This step provides a foundational understanding of each variable's behaviour through summary statistics and visualizations.

1.2.1 Shop and Annual Spending Variables

The graphs below, shown in *Figure 1*, present the frequency distributions of two key variables: "Shop" and "Annual Spending." These visualizations provide valuable insights into customer behaviour and spending habits.

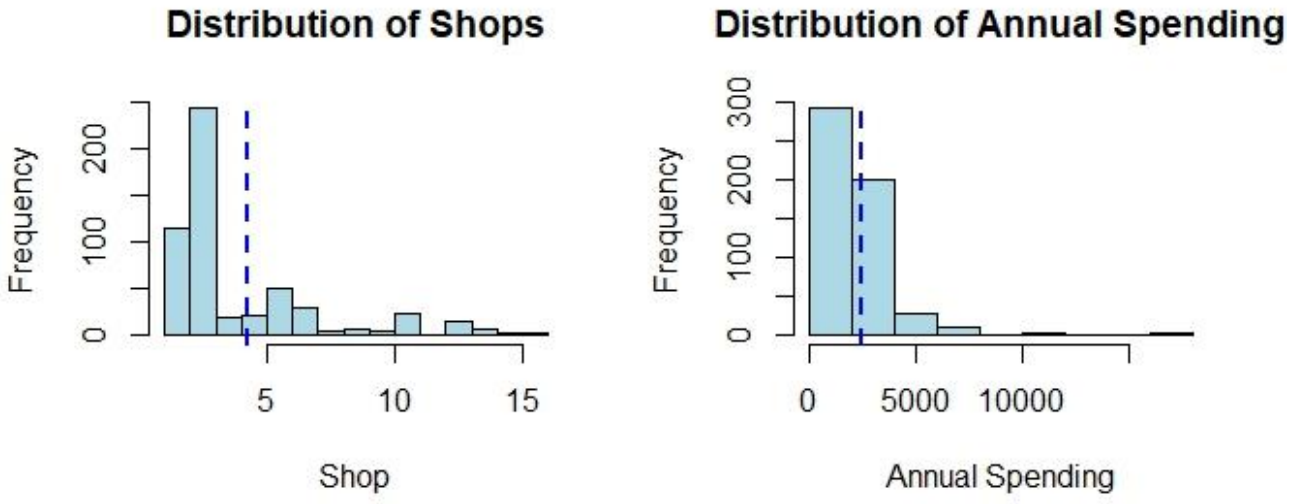


Figure 1 - Shop and Annual Spending Variables

The histogram for the "Shop" variable reveals that most customer visits are concentrated in stores with codes between 1 and 5, indicating that these stores are the main revenue contributors. Stores with codes above 10 have far fewer visits, indicating potential underperformance. High-performing stores should focus on maintaining customer satisfaction, while underperforming stores could benefit from targeted marketing or operational strategies. The distribution of "Annual Spending" highlights another critical aspect of customer behaviour. Most customers spend between $1,200 and $3,600 annually, with a smaller group of high spenders reaching up to $18,000. These high spenders may include customers with specific preferences, such as a demand for premium products.

1.2.2 Car, Distance and Gender Variables

This section explores Car Usage, Distance, and Gender to understand customer behaviour and demographics. The graphs below, *Figure 2*, illustrate their distributions.

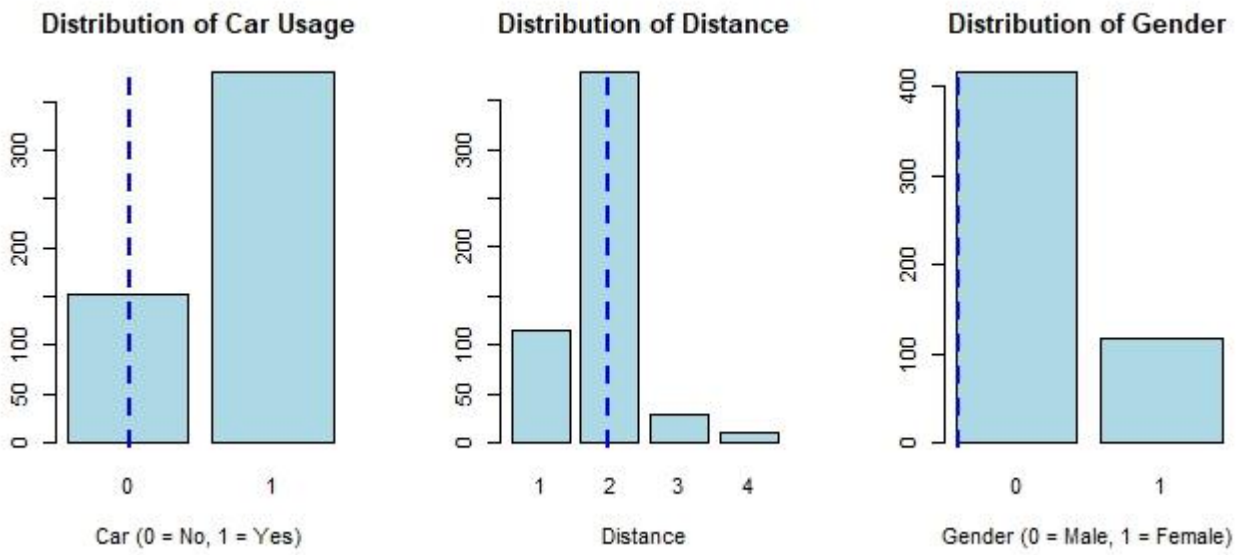


Figure 2 - Car, Distance and Gender Variables

The first graph reveals that most customers use a car to go to shops (Car = 1), indicating that transportation plays a critical role in shopping behaviour. This suggests that store accessibility by car is important for attracting and retaining customers, especially those located farther from the store. Instead, the second chart shows that most customers live within 2 units of distance from the stores. This highlights the importance of proximity in customer decision-making. Stores with customers at greater distances may need targeted strategies, such as delivery services or promotions.

The third graph examines the gender distribution, revealing a higher prevalence of male customers (Gender = 0) compared to female customers (Gender = 1). This demographic trend provides an opportunity to design targeted campaigns or product offerings tailored to the dominant customer group. However, it also highlights the potential to attract more female customers by introducing gender-specific promotions, products, or services.

1.2.3 Prices, Queues, Fresh and Packaged Product Variables

The following graphs, *Figure 3*, analyse customer ratings for four critical aspects of the shopping experience: Prices, Queues, Fresh Products, and Packaged Products.

Figure 3 - Prices, Queues, Fresh and Packaged Product Variables

The graphs show that customer satisfaction is generally high across all four areas, with most ratings concentrated at 4 and 5. Prices are perceived as fair and competitive, which likely contributes positively to customer loyalty and repeat visits. Queue ratings reflect efficient store operations, as shorter wait times appear to meet customer expectations and enhance the shopping experience. Similarly, the high ratings for fresh and packaged products indicate strong satisfaction with product quality. However, the presence of lower ratings in each category highlights opportunities for improvement. By addressing these areas, stores can further enhance customer satisfaction.

1.3 Boxplot

This boxplot visualizes the distribution of all numerical variables in the dataset, excluding Annual Spending. Removing this variable allows for a clearer comparison of the remaining features without the influence of its larger scale and outliers.

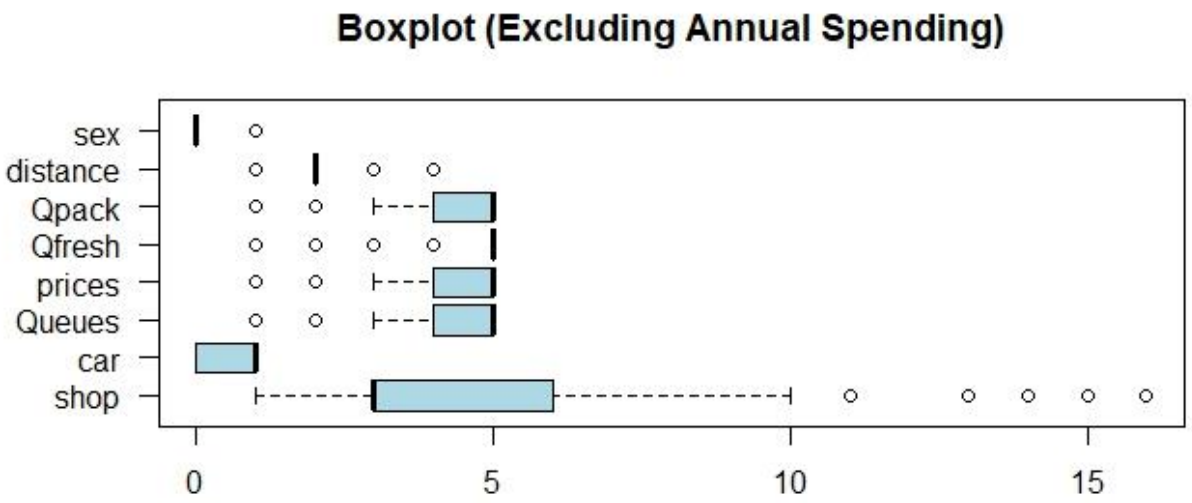


Figure 4 - Boxplot (Excluding Annual Spending)

The boxplot reveals that most variables have limited variability, reflecting consistent ratings or characteristics. For example, Queues, Prices, QFresh, and QProducts exhibit minimal spread, indicating that customer ratings are generally concentrated near high values, such as 4 or 5. Such consistency can be interpreted as a company's efforts in maintaining quality standards, whether through competitive pricing, fresh product offerings, or efficient queue management.

The presence of outliers in some variables, such as Distance, highlights unique cases that could represent valuable insights. For instance, customers traveling far distances may indicate high-value or loyal shoppers, or they could signal potential inefficiencies in the store network's location strategy. Understanding these outliers could guide decisions on improving store accessibility or refining targeted marketing efforts. Overall, the data shows that most stores are performing well and keeping customers happy. At the same time, the outliers offer an opportunity to understand unique cases and make the business even better.

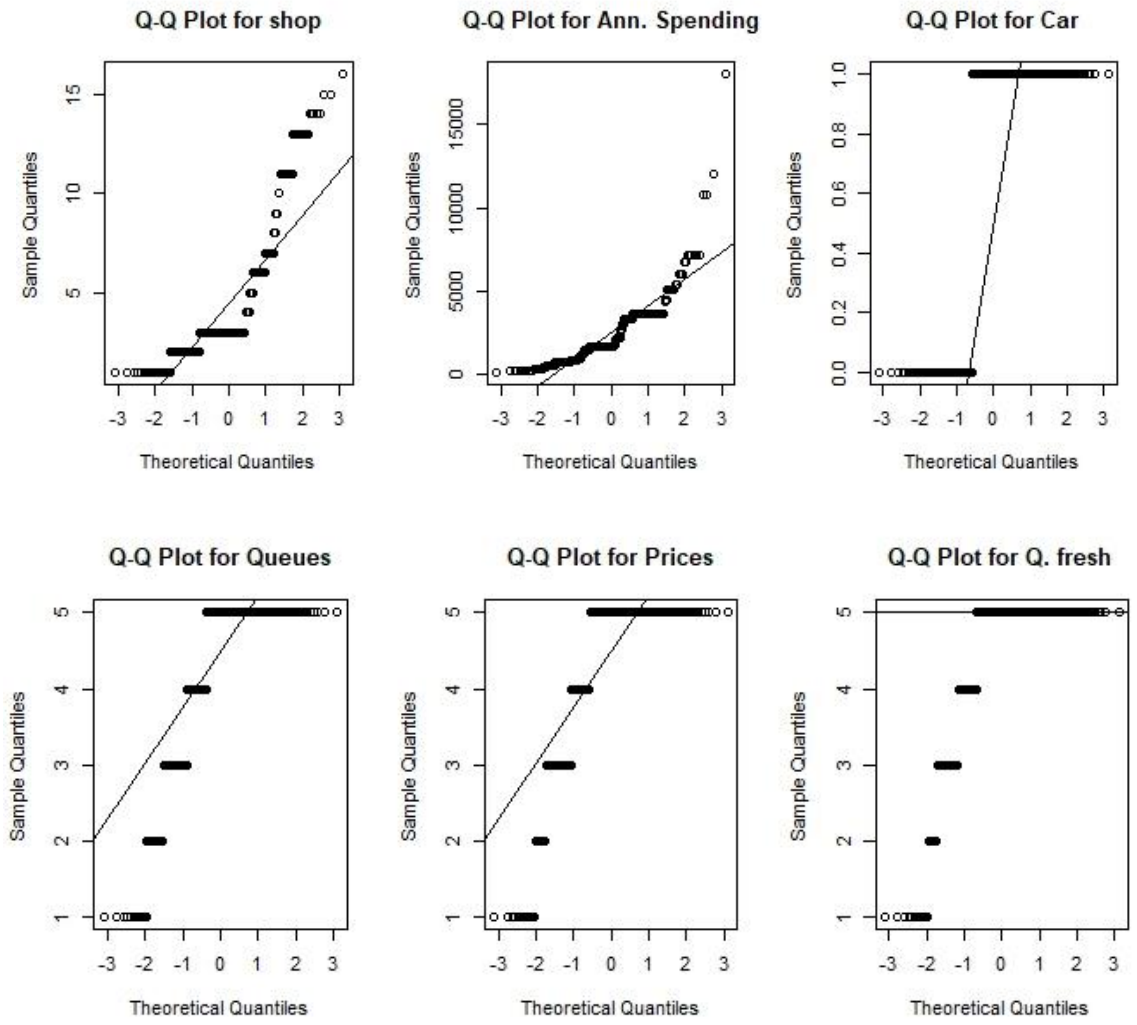
Chapter 2

Pairwise Comparisons

In this section, we explore the relationships between pairs of variables to identify significant associations and trends that impact the target variable, Annual Spending. The aim is to highlight key interactions between variables and their potential influence on customer behaviour and revenue.

2.1 Testing for Normality (Q-Q plot)

To evaluate the assumption of normality for the variables in our dataset, which is essential for performing a t-test, Q-Q plots were created for each variable. These plots compare the quantiles of the observed data with the ones of a normal distribution, helping to assess how closely the data follow a normal distribution. Below are the Q-Q plots, *Figure 5*, for the selected variables.



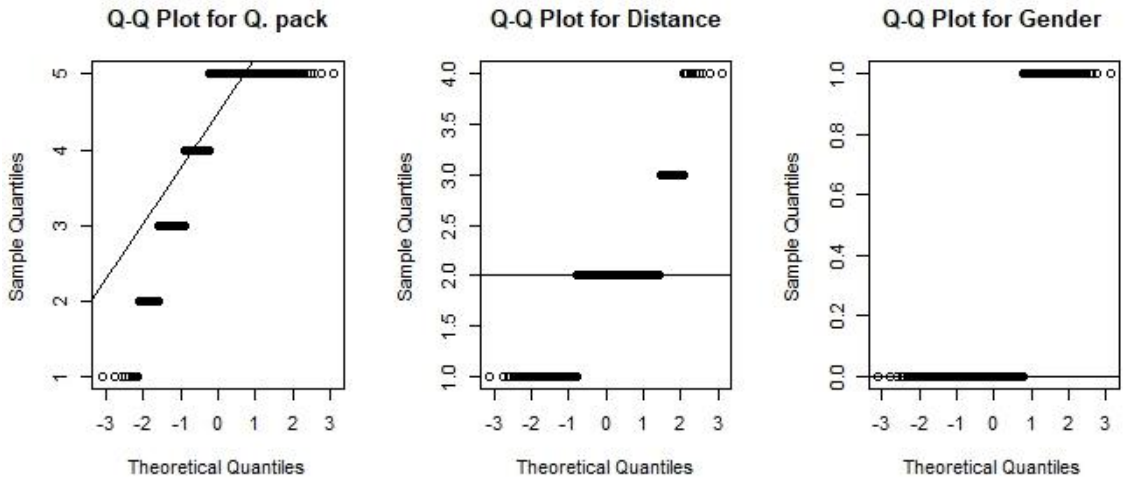


Figure 5- Q-Q plots for checking Normality

The Q-Q plots demonstrate that the assumption of normality is not satisfied for any of the analysed variables. It is evident that none of the variables meet the criteria for normal distribution. The plots display significant deviations from the diagonal reference line, particularly in the tails of the distribution, which suggest the presence of skewness or kurtosis in the data.

2.2 Analysis of Variable Significance Using T-Test and Wilcox Test

Given that the variables in our analysis did not meet the assumption of normality, we utilized both parametric and non-parametric methods to ensure the robustness of our results.

Specifically, we applied the T-Test, *Figure 6*, a parametric method, and the Wilcoxon Test, a non-parametric alternative, as shown in *Figure 7*. The T-Test is typically used for comparing means between two groups under the assumption of normality. However, because the assumption of normality was not satisfied, we also used the Wilcoxon Test, which does not require this assumption and is appropriate for ordinal or continuous variables that are not normally distributed.

Additionally, the variable "Shop" was excluded from the analysis due to issues with the distribution of observations across its categories. Specifically, several groups defined by this variable contained insufficient sample sizes, which made it unsuitable for performing robust statistical tests.

The following graphs, *Figure 6* and *Figure 7*, shows the p-values of the analysed variables.

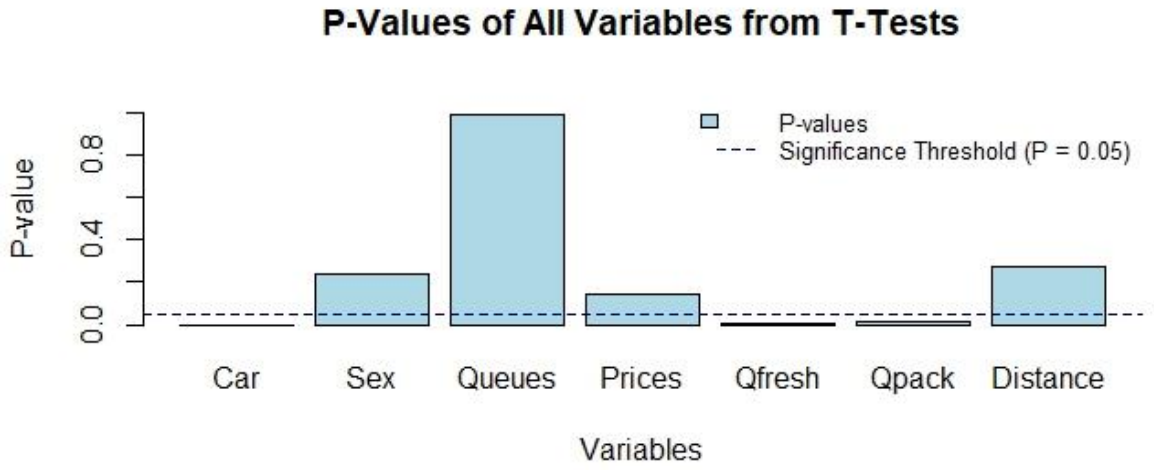


Figure 6 - P-value of all Variables from T-Test

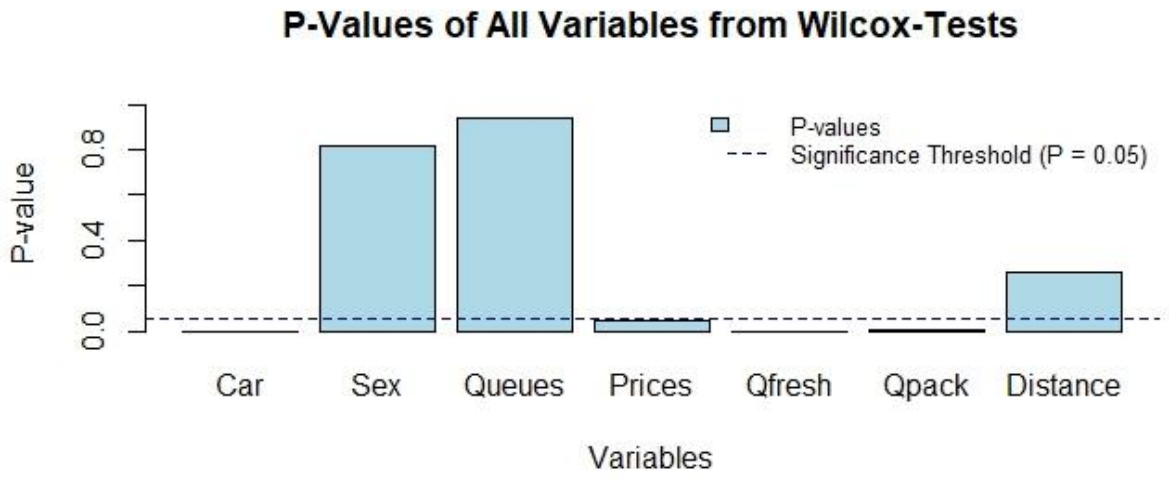


Figure 7- P-Values of All Variables from Wilcox Test

The results from the T-Test and Wilcoxon Test largely align, confirming the robustness of the analysis despite the non-normality of the data. Both tests identify Car, Qpack, and Qfresh as significant variables, with p-values below the threshold of 0.05, indicating a strong association with annual revenues. Additionally, the Wilcoxon Test highlights Price as significant, emphasizing its potential impact on customer spending. These findings suggest that Car, Qpack, Qfresh, and Price play meaningful roles in influencing differences in annual spending among customers.

2.3 Pairwise multiple with Kruskal-Wallis Test

The chart below, *Figure 8*, displays the p-values obtained from the Fligner-Killeen test, performed to assess the homogeneity of variances for the variable spen_yr across different variables, critical for determining whether to proceed with standard ANOVA or alternative statistical methods.

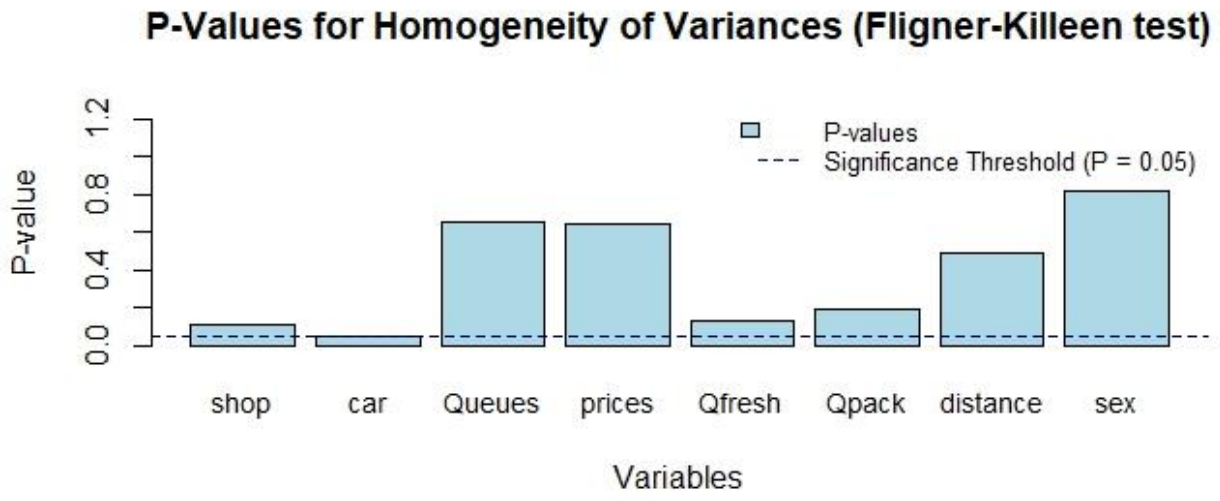


Figure 8 - P-Values for Homogeneity of Variances (Fligner-Killeen test)

Since the dataset did not meet the assumption of normality, and the Fligner-Killeen test revealed that the variable "Car" exhibit unequal variance; instead of ANOVA, the Kruskal-Wallis test was conducted to evaluate the differences in the annual spending across various groups.

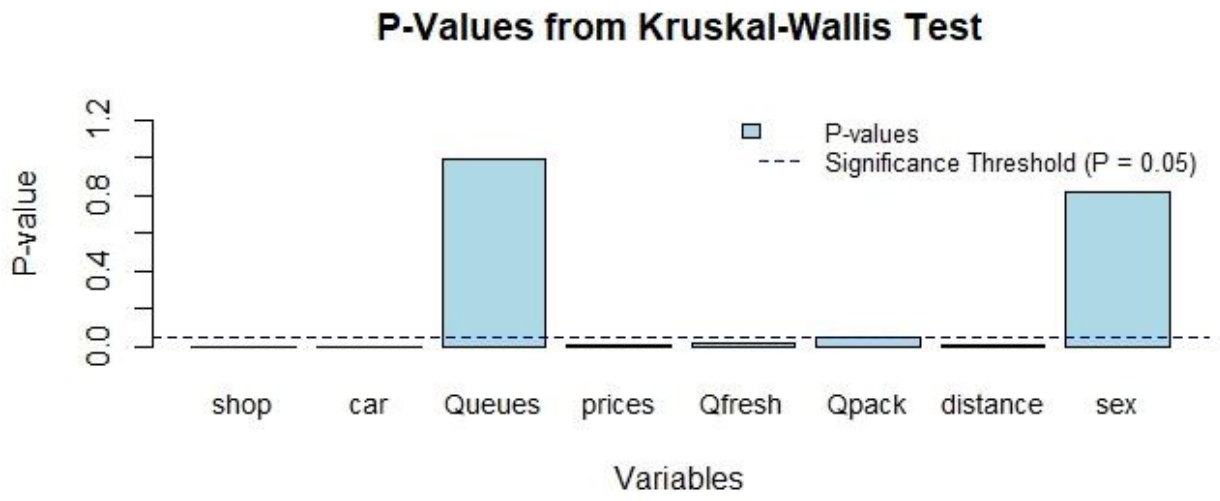


Figure 9 - P-Values from Kruskal-Wallis Test

The analysis revealed that variables such as Shop, Prices, Qpack, Qfresh and Distance showed p-values below the 0.05 threshold. This indicates significant differences in annual spending between the groups for these factors. Spending patterns may vary across supermarket chains or due to perceptions of pricing and convenience. On the other hand, Queues and sex do not show statistically significant differences (p-value > 0.05), suggesting a limited impact on spending behaviour.

2.3.1 Post-hoc Analysis for Significant Variables

To identify specific group differences a Dunn's post-hoc test with Bonferroni correction was applied. The *table 2* below presents the significant comparisons between groups for the variables analysed, focusing on adjusted p-values that are below the significance threshold of 0.05.

Variables	Groups Compared	P-value (Unadjusted)	P-value (Adjusted)
Shop	11 - 13	0.00004	0.0043
Shop	11 - 2	0.000463	0.0486
Shop	11 - 4	0.00012	0.0128
Car	0 - 1	<0.00001	<0.00001
Prices	2 - 5	0.0028	0.0278
Prices	4 - 5	0.0029	0.0288
Quality Fresh	3 - 5	0.0036	0.0357
Distance	1 - 3	0.00061	0.0037

Table 2 - Summary of Significant Group Comparisons

For the variable Shop, group 11 shows significant differences in spending when compared to groups 13, 2, and 4, suggesting distinct characteristics of the supermarket chains associated with these groups. Similarly, the variable Car shows a highly significant result indicating that car ownership strongly impacts spending behaviour. Also, the variables Prices and Quality Fresh reveal specific group-level differences. Lastly, for the variable Distance, a significant difference in spending was observed indicating that proximity to the supermarket impacts customer behaviour. In contrast, the variable Qpack did not show any significant differences between groups, suggesting that packaging quality has a limited impact on annual spending. These results show that factors such as the choice of supermarket, whether customers own a car, how they perceive pricing, the freshness of products, and the distance to the store all have a noticeable impact on how much people spend annually.

2.4 Exploration of Correlations Among Variables

To better understand the relationships between the variables in the dataset, a correlation analysis was performed, *Table 3*. This visualization highlights the strength and direction of the relationships among the variables, offering insights into potential patterns that might influence annual spending.

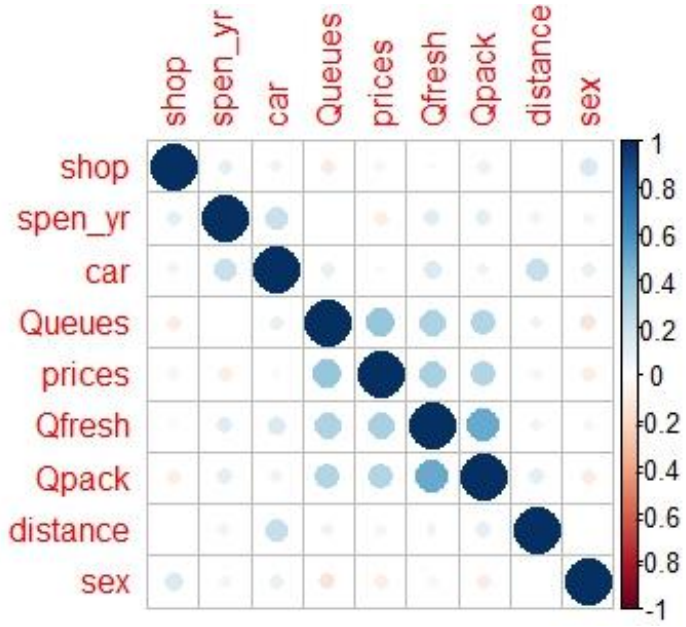


Table 3 - Correlation

The correlation matrix reveals several important relationships among the variables. A weak positive correlation is observed between spen_yr and car, indicating that car ownership might slightly increase annual spending, likely due to the convenience of transporting larger purchases. Instead, distance has a very weak negative correlation with spen_yr, suggesting that proximity to the supermarket has only a minor influence on spending patterns. A strong positive correlation is observed between prices and Qfresh, highlighting that customers who perceive products as fresh are also more likely to view pricing as fair. Similarly, Qfresh and Qpack show a moderate positive correlation, indicating that customers associate high-quality packaging with product freshness. Other variables, such as Queues and sex, show negligible correlations with spen_yr and other variables, indicating they have little to no impact on spending behaviour.

Chapter 3

Predictive and Descriptive Modelling

In this chapter, we develop and evaluate statistical models to understand which factors affect annual spending and to make predictions. The analysis includes model selection, assessment of goodness of fit, and out-of-sample predictions to ensure robust and generalizable results.

3.1 Model Selection

To model annual spending, several transformations of the dependent variable were considered, including linear, polynomial, square-root, and logarithmic forms. Among these, the mixed transformations, *Table 4*, proved to be the most suitable, it demonstrated better performance in terms of goodness-of-fit compared to alternative specifications, making it both robust and practical.

Coefficients	Estimate	Std. Error	T value	Pr(>|t|)	Significance
Intercept	7.1161215	0.2247684	31.660	< 2e-16	***
Shop	0.0068504	0.0102776	0.667	0.505364	
poly(prices, 2)1	-3.0524396	0.7805724	-3.911	0.000104	***
poly(prices, 2)2	-1.2260121	0.7300404	-1.679	0.093681	.
poly(distance, 2)1	0.3890924	0.7164846	0.543	0.587323	
poly(distance, 2)2	-0.3594769	0.7191537	-0.500	0.617385	
Car	0.3375823	0.0708773	4.763	2.48e-06	***
Queues	-0.0001804	0.0337002	-0.005	0.995732	
poly(Qfresh, 2)1	2.5699704	0.8388487	3.064	0.002300	**
poly(Qfresh, 2)2	1.0516345	0.7463165	1.409	0.159405	
Qpack	0.0413792	0.0387093	1.069	0.285579	
Gender	-0.0558725	0.0745214	-0.750	0.453744	

Table 4 - Regression Model Summary

*Legend: Signif. codes: 0 ‘***’0.001 ‘**’0.01 ‘*’0.05 ‘.’0.1 ‘ ’1*

The regression results highlight several factors influencing annual spending, which has been modelled on a logarithmic scale. Car ownership emerges as significant, with individuals who own a car tending to spend more annually, in fact the coefficient of 0.33 for the "Car" variable suggests

that owning a car is associated with an approximate 40% increase in annual spending, holding other factors constant. The level of prices also shows a significant relationship with spending, where higher prices are associated with lower spending. Additionally, the quality and freshness of products appear to influence spending decisions, emphasizing the importance of consumer preferences for these attributes. Variables such queues, distance, packaged product and gender did not show a statistically significant impact. For example, for every one-unit increase in the value of Qpack, the log of annual spending is estimated to increase by 0.0414, assuming all other variables constant.

Residual Standard Error	0.6889 on 519 degrees of freedom
Multiple R-squared	0.1202
Adjusted R-squared	0.1015
F-statistic	6.445 on 11 and 519 DF, p-value: 4.474e-10

Table 5 - Performance Metrics

The model (*Table 5*) shows an Adjusted R-squared of 0.1015, which indicates that 10.15% of the variation in annual spending can be explained by the predictors in the model. The F-statistic suggests that the model is statistically significant and the relatively low RSE suggests that the model fits the data well, although there is still some unexplained variance.

3.2 Optimized model

To refine the analysis and focus on the most impactful variables, an optimized version of the model was developed. This model retains only the predictors that demonstrated significant associations with annual spending in the original analysis.

Coefficients	Estimate	Std. Error	T value	Pr(>|t|)	Significance
Intercept	7.30148	0.05629	129.717	< 2e-16	***
poly(prices, 2)1	-2.93487	0.72988	-4.021	6.64e-05	***
poly(prices, 2)2	-1.36495	0.71895	-1.899	0.0582	.
Car	0.35201	0.06690	5.262	2.08e-07	***
poly(Qfresh, 2)1	2.99091	0.73866	4.049	5.92e-05	***
poly(Qfresh, 2)2	1.28011	0.71705	1.785	0.0748	.

Table 6 - Regression Optimized Model Summary

The optimized model focuses on the variables "prices," "car," and "Qfresh," which were identified as critical drivers of annual spending. As indicate also in the original model, the coefficients show that higher prices are associated with reduced spending, having a car shows a positive and statistically significant association with spending, suggesting that car owners are likely to spend more annually. Similarly, the freshness of products exhibits a strong positive effect on spending.

Residual Standard Error	0.6868 on 525 degrees of freedom
Multiple R-squared	0.1154
Adjusted R-squared	0.1070
F-statistic	13.7 on 5 and 525 DF, p-value: 1.372e-12

Table 7 - Performance Metrics Optimized Model

Despite a slight reduction in the model's R2 and Adjusted R2 compared to the original, the optimized model offers a more focused approach, prioritizing variables with meaningful and statistically significant contributions.

3.3 Checking Assumptions and Outliers

Before proceeding with model estimation and evaluation, it is crucial to verify the underlying assumptions of the statistical optimized model. These assumptions include linearity, normality of residuals, independence of residuals, and homoscedasticity.

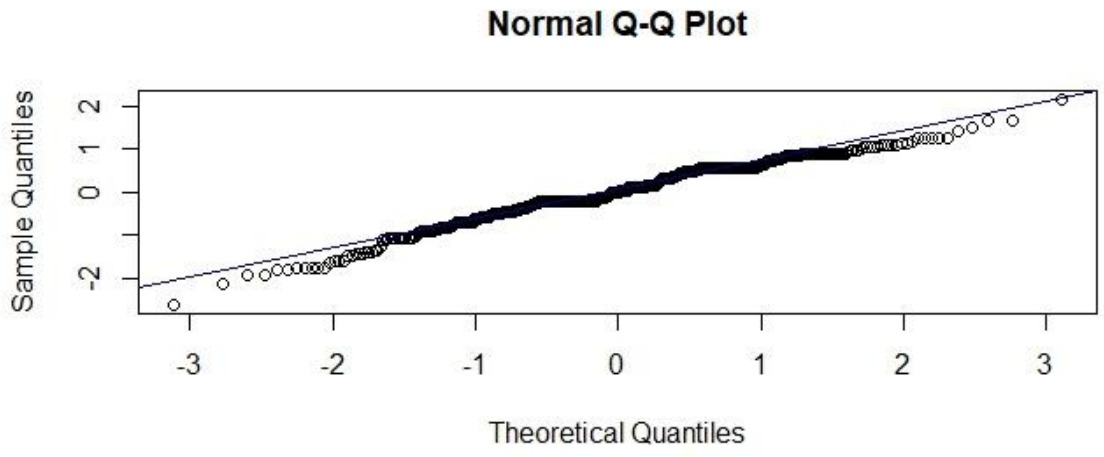


Figure 10 - Normal Q-Q Plot

The Q-Q plot (*Figure 10*) illustrates the distribution of the residuals compared to a theoretical normal distribution. While the central points closely follow the reference line, deviations are evident at the tails, particularly at the extremes. This suggests that the residuals do not fully follow the assumption of normality. However, given the large sample size (n > 500), this deviation isn't critical for the validity of the coefficient estimates.

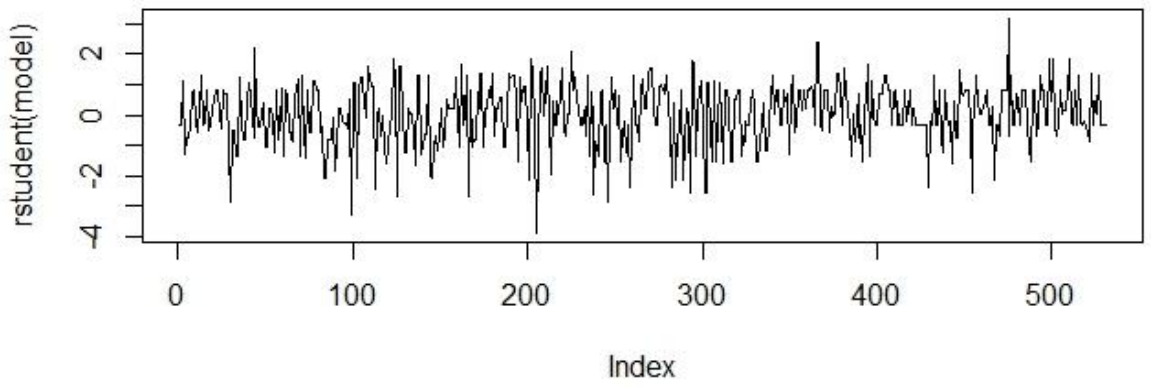


Figure 11 - Residual Plot

The Durbin-Watson statistic (at 1.71) indicate only a mild deviation from the assumption of independence. The residual plot (*Figure 11*) supports this, showing residuals that fluctuate randomly around zero, with no clear cyclic trends or systematic patterns, while some degree of autocorrelation may be present, it appears limited and unlikely to significantly compromise the model's reliability.

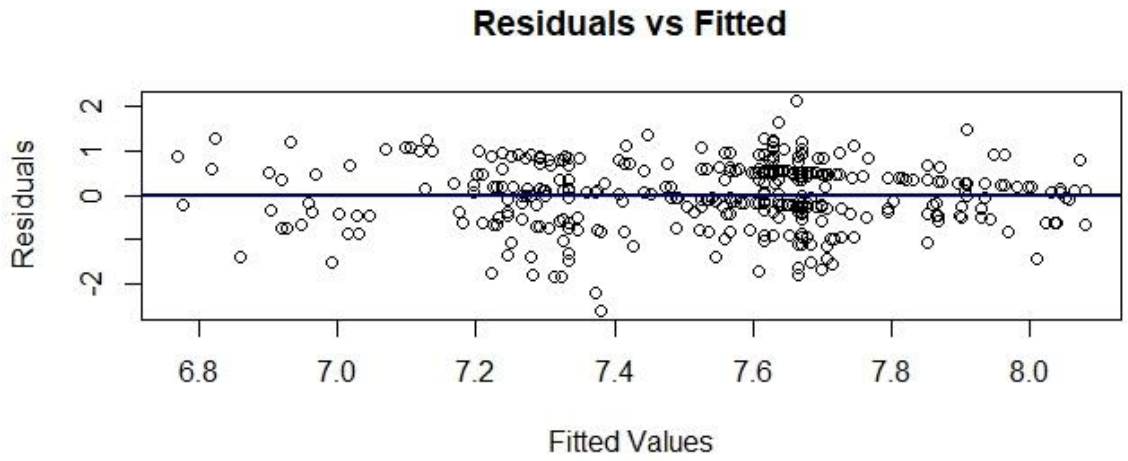


Figure 12 - Residual vs Fitted Plot

The Residuals vs. Fitted plot (*Figure 12*) is used to verify the linearity assumption in the model. In this plot, the residuals are scattered around the horizontal zero line without a clear pattern, indicating that the assumption of linearity is reasonably met.

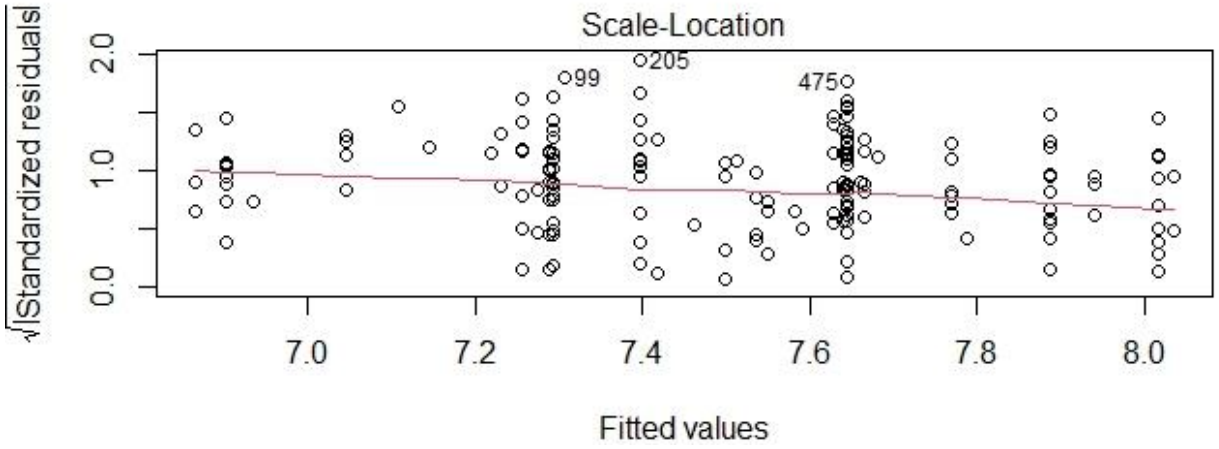


Figure 13 - Scale - Location

The Scale-Location plot (*Figure 13*) was used to assess the assumption of homoscedasticity. Ideally, the points should be randomly scattered around the horizontal line, indicating equal variance. In this case, the plot shows some mild patterns and uneven spread, suggesting potential heteroscedasticity, a result consistent with the Breusch-Pagan test (p = 0.008). While this violation might influence the efficiency of the coefficient estimates, it does not render the model invalid.

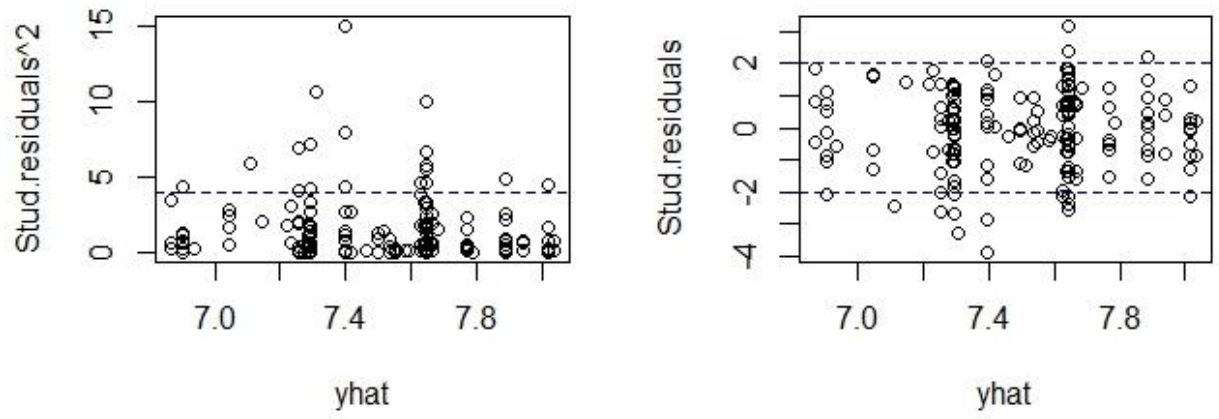


Figure 14 - Residual Analysis

To assess the presence of outliers and their potential impact on the model, studentized residuals were analysed. The left panel of the plot (*Figure 14*) displays the squared studentized residuals versus the fitted values allowing the detection of potential outliers. The horizontal dashed line indicates a threshold value of 4, above which points may be considered influential, a few points exceed the threshold, suggesting the presence of influential observations.

The second graph shows the predicted values on the x-axis and the studentized residuals on the y-axis. It shows some residuals outside the range of ± 2, confirming the presence of potential outliers. However, most residuals fall within acceptable ranges, indicating that the model's overall fit is not influenced by extreme observations, the limited number of outliers suggests that they do not substantially compromise the reliability of the model's conclusions.

3.4 Model Comparison and Goodness of Fit

To evaluate the performance of the optimized model compared to the original model, several goodness-of-fit metrics were calculated, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), R2 and adjusted R2.

Metric	Full Model	Optimized Model
AIC	1125.004	1115.866
BIC	1180.575	1145.789
R Squared	0.1201879	0.115433
Adjust R Squared	0.1015406	0.1070086

Table 8 - Comparison of Goodness-of-Fit Metrics

The results in *Table 8* show that the optimized model performs better than the full model. It has lower AIC (1115.866) and BIC (1145.789), which means it finds a better balance between accuracy and simplicity. While the full model explains slightly more variance ($R^2 = 0.1202$), the optimized model has a good Adjusted R^2 (0.1070) considering it uses fewer variables. This suggests that the optimized model is more efficient and easier to interpret. It achieves similar performance without unnecessary complexity.

3.5 Out-of-Sample Prediction

Maximizing revenue requires a clear understanding of the factors that influence annual spending. To ensure the reliability and generalizability of the models, the dataset was divided into training and test sets, as detailed in *Table 9*. This approach allows for out-of-sample prediction, where the model's performance is evaluated on data it has not seen during training.

Metric	Train Set	Test Set
Dimension	371	160
Variables	9	9

Table 9 - Train and Test Set Characteristics

Two models were developed: a full model including all predictors and an optimized model with a reduced subset of significant variables. Their performance was evaluated using RMSE, Root Squared Error, which measures the average difference between values predicted and the actual values; and MAE, Mean Absolute Error, which calculates the average absolute difference between predicted and actual values, as shown in *Table 10*.

Metric	Full Model	Optimized Model
RMSE	0.7387911	0.71877
MAE	0.5916493	0.5752587

Table 10 - Model Performance Metrics

The analysis demonstrates that the Optimized Model achieves slightly better performance, with a lower RMSE (0.71877) and MAE (0.5752587) compared to the Full Model. This suggests that simplifying the model by excluding less impactful variables can improve both predictive accuracy and model interpretability. The improved performance of the Optimized Model also highlights its potential to generalize better to new, unseen data.

Conclusion

The analysis shows that spending is influenced by key factors such as car accessibility, product freshness, and price perception. These factors not only reflect what customers prioritize but also reveal areas where strategies improvements can drive better results.

The study confirms that car ownership significantly impacts spending, with car users tending to spend approximately 40% more annually. This suggests that supermarkets could benefit from improving parking facilities or introducing promotions.

Also, product freshness emerged as a significant driver of higher spending. While the current analysis focuses on ratings for fresh products, a deep investigation into the types of fresh products could help supermarkets identify which categories most strongly influence customer preferences.

The findings also suggest that underperforming stores could benefit from focused interventions, particularly those with fewer visits, may benefit from operational strategies, such as localized promotions or improved accessibility. Pricing remains a delicate balance, while most customers perceive prices as fair, the analysis suggests that price sensitivity varies among customer segments. Offering dynamic pricing models or loyalty rewards can incentivize increased spending.

Moreover, while this report effectively models key variables, the relatively low adjusted R-squared indicates there is room for incorporating additional predictors.

In summary, the findings of this report suggest to prioritizing quality, leveraging data-driven insights, and addressing operational inefficiencies. By doing so, supermarkets can ensure long-term revenue maximization and customer loyalty.

Web References

https://www.datascienceblog.net/post/statistical_test/signed_wilcox_rank_test/

https://statistics.laerd.com/spss-tutorials/independent-t-test-using-spss-statistics.php

https://www.technologynetworks.com/informatics/articles/the-kruskal-wallis-test-370025

https://www.statology.org/dunns-test-in-r/

https://otexts.com/fppit/selecting-predictors.html

https://cran.r-project.org/doc/contrib/Ricci-regression-it.pdf

List of Tables

List of Figures