A REVIEW OF SUPERMARKET SALES

BY

FURKAN BERK DANIŞMAN

SILA İLYÜREK KILIÇ

March 2023

ABSTRACT

This research examines supermarket data to be a light for understanding the customers' behaviors considering the parametric and non-parametric data analysis approach. In addition, this research is conducted to investigate the uncertainty of which factors change the customer's total price and which variables affect or do not affect the customer's total price. In light of this problem, firstly, an analysis of the significant difference between the supermarket infrastructure and conducted data in the aspect of averages/medians has been done. Secondly, it was checked whether there was a significant difference between customers' supermarket expenditures according to payment method, membership type, and supermarket branch. Lastly, it has been investigated whether gender and product type affect total expenditures. As a result, these findings showed that supermarket expenditures might vary according to certain factors and are not affected by specific variables which are hypothesized in advance

1. Introduction

Supermarket companies, especially in big cities, have to understand customer behavior and create price ranges suitable for them due to the competition among them. At the same time, customers form their shopping habits by considering various factors such as product variety, price comparison, and payment methods while spending in their supermarkets. This mutual association between supermarket brands and customers is made possible by both stakeholder understanding of supermarket infrastructure. For this reason, the data has been shaped around various hypothesis tests for the consumer and supermarket infrastructure owner to understand this complex structure.

There are some obvious factors in the supermarket market that the total price will increase as the customer buys more products, or the linear relationship of the tax to the number of customer's expenses, but is the total price spent by the customer only consist of these facts?

The knowledge that a supermarket membership has an impact or not on the total price is an essential matter for customers and supermarket owners. The reasons are the business policy to be followed by the owner and whether the company should give priority to membership or not, which would be apparent through the profit results regarding the membership. The essentiality of this information also applies to the customers. In today's world, grocery shopping is becoming

a situation that must be considered more and more from the customer's point of view. Customers no longer buy a product with passion as they see something they are interested in. They need to consider their economic situation since everything is more expensive now. That is why they spend money more carefully. Therefore, knowing whether the membership has changed or not changed the customer's total price provides the customer with an economical relief option.

Another necessary information would be how products price change in different lines for the supermarket market. Every customer has an area of interest. Hence, the economic impact of these areas of interest is vital to the customers. It provides an idea of the future expenditures for the customer's interest through the information of more or fewer expenses compared to other interests. Customers may decide to suppress their interest and passion or be relieved about their interest if their interest is more economically viable than other expenditures through the information they acquire regarding product line expenses. Similarly, the product line is essential for supermarket owners; depending on which product line makes more money, the company can put those products line on high shelves or put them in the most visible place in the market.

Whether the supermarket market affects the customers' interest or whether membership has an effect or not on the customer's expenses is a curiosity. However, another matter of curiosity would be whether these expenditures are mostly made with cash or credit cards and whether this payment preference makes a difference regarding gender or not. The conducted supermarket data has the potential to eliminate these obscurities through the methods that are done in this research which are indicated in the following figure;

Methods	
Parametric	Non-parametric
One Sample T Test	Single-sample Sign Test
Two Sample T Test	Wilcoxon Rank Sum Test
Z Test	Chi-squared Test
Simple Linear Regression	Rank Based Estimation Regression
Multiple Linear Regression	Rank Based Estimation Regression
One Way ANOVA	Kruskal-Wallis Test

Figure 1

1.1. Data description

The supermarket data set includes 17 distinct variables and 1000 observations collected over three months by three separate firms. It includes factors such as the locations in which the firms are headquartered, customer membership categories, customer genders, customer payment methods, product types, and total customer expenditures. It has 8 continuous variables, 5 categorical variables, and 1 discrete variable. It is illustrated in the following figure:

Attributes			
Invoice.id	Computer generated sales slip invoice identification number	**Tax**	5% tax fee for customer buying
Branch	Branch of supercenter (3 branches are available identified by A, B and C).	**Total**	Total price including tax
City	Location of supercenters	**Time**	Purchase time (10am to 9pm)
Customer.type	Type of customers, recorded by Members for customers using member card and Normal for without member card.	**Payment**	Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet)
Gender	Gender type of customer	**Cogs**	Cost of goods sold
Product.line	General item categorization groups – Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel	**Gross Margin**	Gross margin percentage
Unit.price	Price of each product in $	**Gross Income**	Gross income
Quantity	Number of products purchased by customer	**Rating**	Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

Figure 2

Data Source: https://zdataset.com/free-dataset/supermarket-sales/

1.2. Research questions

1. Supermarket sales data has a random sample of the total price of all customers. We measure the total price for a sample of supermarket sales data. We claim that the average/median of the total price is approximately 270$. At α =0.05 can be concluded that the unknown mean/medianof the total price for customers is different from 270$.

2. Supermarket data consists of two types of customers which are member customers and normal customers. Our measurement is the customers' total prices. We claim that the mean/median total prices for member customers are greater than the mean/median total prices for normal customers at $\alpha = 0.05$.

3. Supermarket data consists of 1000 customers and it includes that 344 of the customers do their supermarket shopping with cash. According to the US Federal Reserve (2019), 30%of all payments were made with cash in 2017. Can it be concluded that the proportion of the sample of payments being cash by customers is different from 0.3 at $\alpha = 0.05$?

4. Supermarket data consists of 1000 customers which 163 of 501 women pay with credit cards, while 148 of 499 men pay with credit cards. At 0.05 level, is there a difference betweenthe two proportions?

5.1 In this research, we wonder if gender is significantly relevant to the variation in the total price of a customer's supermarket expenses at $\alpha = 0.05$.

5.2 In this research, we wonder if gender and product line is significantly associated with thevariation in the total price of customers' supermarket expenses at $\alpha = 0.05$.

6.0 In this research, we wonder whether the branch of a supermarket makes a difference in terms of the total price of customer's expenses or other words, whether the means/median of different supermarket branches is equal or not at $\alpha = 0.05$.

1.3 Aim of the study

The purpose of this research is to help supermarket owners and customers understand the supermarket infrastructure and construct connection points between supermarket key elements. Three of these key elements, especially membership, product line, and the relationship with the total price of the payment method are the main elements of this research.

The focus of the research was on the difference between the normal and member status of the customer on the price of their shopping at the supermarket, the relationship between the selected product line and the total price paid by the customer, and how the total price payment proportion differs from customers according to the payment method. In addition to these, whether there is a difference between the total price of the three brands and whether there is a difference in the total price averages according to gender has been investigated with various hypotheses and statistical tests. We hope that testing these ideas will reveal whether there is a difference or not between the variables that are mentioned above.

2. Methodology/Analysis

In the first stage of the research, we created diagnostic plots with the histogram of the determined response variable (total price) and put it into the Shapiro test and the Anderson-Darling test to test whether the data is normally distributed or not. After deciding that the data is not normally distributed, we applied the box-cox method to understand which transformation method is necessary to apply. Then, we applied the determined transformation method to our data. We created the diagnostic plots along with the histogram of the transformed data and put it to the Shapiro test and Anderson-Darling test. We decided that the transformed data is also not normally distributed. Then, since the sample size of the data was large enough, we assumed that the data belonged to the population with a normal distribution and applied parametric tests, and also we assumed that we did not know about the population distribution and applied non-parametric tests.

Under parametric tests, we first applied the t-test method to inference about the mean and comparison of the means. To check the assumptions of this test, we applied the F test and interpreted the difference between the variances. Then we applied the Z test to have an effect on the proportions and to the comparison of the proportions. Then, we applied simple linear regression and multiple linear regression methods for categorical variables. After that, to meet the assumptions of one-way Anova, we checked if the variances are equal or not through the variance

ratio test and also use diagnostic plots to clarify the normality. Finally, we implement a one-way-Anova test and Tukey multiple comparison tests to decide if the means are significantly different from each other or not.

Under non-parametric tests, we first applied the signed test and Wilcoxon rank-sum test methods to make an inference about the median and to compare the medians. Then we applied the Chi-squared method to have information about the proportions and comparison of the proportions. Then, we applied the Rank Based Estimation Regression method for categorical variables. Finally, we used the Kruskal-Wallis method.

3. Discussion/Conclusion

As a result of the findings obtained in this research, firstly, it was determined that the data was not normal through the non-belly curved histogram shape and the QQ plot. In the QQ-Plot, values do not lie along the line, which indicates that there are outliers and that data is not normally distributed. In addition, the Anderson-darling normality test has been used to support the interpretation of this graph. As a result, the test indicates that there is non-normality through the p-value that is smaller than 0.05. Since there is a non-normality indication, the box-cox transformation method is used. It shows that the delta is close to zero. Hence, it indicates that log transformation should be used. After the implementation of the log transformation, the histogram and Shapiro test method is used. Since still, the histogram is a non-belly curved shape, and the Shapiro test's p-value is smaller than 0.05, there is an indication of non-normality even though there is a transformation. As a result, since the sample size of the data was large enough, we assumed that the data belonged to the population with a normal distribution and applied parametric tests. Also, we assumed that we did not know about the population distribution and applied non-parametric tests. If the results have any contradiction, we accept that the non-parametric test result is more reliable since in non-parametric test's median is not affected by outliers as much as means.

- **Findings of Research Question – 1)**

In one sample hypothesis research question, There is a claim that the average/median of the total price is different from 270$ at a=0.05. The analysis of this claim is tested through t-test and sign-test. And the p-value is calculated as less than 0.05 in the t-test and greater than 0.05 in the sign-test. Since the sign-test median is evaluated, the outliers have less effect on the result. Therefore, the sign-test result is more reliable than the test. As a result, we fail to reject the null hypothesis, and it can be concluded that even though there is a difference between the sample mean and the claimed mean, there is no difference between the sample median and the claimed median at 0.05 level

- **Findings of Research Question – 2)**

In the two-sample hypothesis research question, There is a claim that the mean/median total prices for member customers are greater than the mean/median total prices for normal customers at a=0.05. The analysis of this claim is tested through a t-test and Wilcoxon rank-sum test. In the t-test, to meet the assumptions, variances are checked for two samples. Through the F-test, a p-value greater than 0.05 is calculated, and the ratio is close to 1. Hence we conclude that variances are equally distributed. Hence, the t-test and Wilcoxon rank-sum test's p-values are evaluated. Since the p-value for both of the tests is greater than 0.05, there is no contradiction. Hence, we fail to reject the null hypothesis, and it is concluded that there is not enough evidence to support that the total price of members is greater than the total price of normal customers at the 0.05 level.

- **Findings of Research Question – 3)**

In a one-sample proportion hypothesis research question, According to the US Federal Reserve, 30% of all payments were made with cash in 2017. The claim is that the proportion of the sample of payments being cash by the customer is different from the US Federal Reserve results at a=0.05. The analysis of this claim is tested through a z-test and chi-square test. The z-test value is greater than the critical value of 1.96, and the Chi-square test value is greater than the critical value of 3.841. Hence, we reject the null hypothesis, and it is concluded that there is enough evidence to support that the proportion of sample payments being cash by the customer is different from the US Federal Reserve results at a=0.05.

- **Findings of Research Question – 4)**

In the two-sample proportion hypothesis research question, there is a claim that the proportion of women that pays with a credit card is different from the proportion of men that pays with a credit card at 0.05 level. The analysis of this claim is tested through a z-test and chi-square test. The z-test value is less than the critical value of 1.96, and the Chi-square test value is less than the critical value of 3.841. Hence, we fail to reject the null hypothesis, and it is concluded that there is not enough evidence to support that the proportion of women that pays with credit card is different from the proportion of men at the 0.05 level.

- **Findings of Research Question – 5.1)**

In the regression part of the research, the effect of a different gender on the variation in the total price of customers' supermarket expenses is analyzed through simple linear regression and rank-based estimation regression at a=0.05. For both of the regression types, the p-values are calculated. Since the p-value is greater than 0.05 for both versions and there is no contradiction, it can be concluded that gender is not significantly relevant to the variation in the total price of customers' supermarket expenses at 0.05 level.

- **Findings of Research Question – 5.2)**

In the multiple regression part of the research, the effect of different gender and product lines on the variation in the total price of customers' supermarket expenses is analyzed through multiple linear regression and rank-based estimation regression at a=0.05. For both of the regression types, the p-values are calculated. Furthermore, since the p-value for all variables are greater than 0.05 for both versions, and there is no contradiction, we can conclude that taking the product variable into account does not change the significance of gender. Moreover, it can be seen that gender and product line is not significantly associated with the variation in the Total Price of customers' supermarket expenses at 0.05 level.

- **Findings of Research Question – 6)**

In the variances difference analysis part of the research, there is a claim that the branch of a supermarket makes a difference or not in terms of the total price of customer's expenses or, in other words, whether the means/medians of different supermarkets branches are equal or not at a=0.05. The analysis of this curiosity is tested through the One Way ANOVA, Krusko-Wallis Rank Sum Test, and Tukey Multiple Comparision Test. To meet the assumptions of One-way-Anova, the QQ plot and variance ratio test is used. In the QQ plot, the variables do not lie in the

line, which indicates non-normality, but since the sample is large enough, we assumed that the population is normally distributed. In addition, from the ratio test, we find that the variances are equally distributed since the ratio is closer to 1. As a result, we can implement a one-way-ANOVA test. For both of the variance analysis tests, the p-values are calculated. Since the p-value for both tests is greater than 0.05, and there is no contradiction, we fail to reject the null hypothesis. It can be concluded that there is not enough evidence to support the claim that the branch of a supermarket makes a difference in terms of the total price of customer's expenses or, in other words, the means/medians of different supermarket branches are equal at 0.05 level. To support the findings from ANOVA and Kruskal-Wallis and show that if we analyzed companies two by two separately, we still do not find enough evidence to show the difference, we used Tukey Multiple Comparision Test. In the Tukey Multiple Comparision Test, since all adj p values are greater than 0.05, we fail to reject the null hypothesis, and we conclude that there is no difference in means/medians between any two branches.

References

1. FENG, C., WANG, H., LU, N., CHEN, T., HE, H., LU, Y., & TU, X. (2014). *Log-transformation and its implications for data analysis*. PubMed Central (PMC).
2. Kloke, J. (2012). *Rfit: Rank-based Estimation for Linear Models*. Journal.r-project.org.
3. *The Fed - The 2019 Federal Reserve Payments Study*. Board of Governors of the Federal Reserve System.
4. Zach, V. (2020). *How to Perform Tukey's Test in R - Statology*. Statology.
5. Zach, V. (2020). *How to Perform a Box-Cox Transformation in R (With Examples)*. Statology.