Statistical Analysis Report: Supermarket Sales Dataset

**Introduction**

Understanding customer behavior is very important in today’s competitive retail landscape for a business to thrive. The dataset chosen for this analysis contains transaction records and other data from a supermarket chain, detailing customer demographics, purchase information, and sales figures across different product categories. Analyzing this data can and will provide valuable insights into how different factors such as gender, membership, and city location influence customer purchasing behavior for the company. The business problem here revolves around optimizing marketing strategies, improving customer retention, and enhancing sales by understanding these underlying patterns in customer transactions using tools like R studio.

**Objective**

The primary objective of this assignment analysis is to perform a statistical examination of the supermarket sales data to uncover patterns and relationships that can influence and inform business decisions like: identifying customer spending factors, examine purchasing behavior, providing data for target marketing and promotions ETC.

**Business questions**

To reach the objective and analyze the data, there are some key questions to be analyzed:

1. Does gender affect the preference for certain product categories?
2. Is there a significant difference in purchasing behavior between members and nonmembers?
3. Does the city location influence sales and customer behavior?
4. What product categories generate the high sales and what factors influence the sales?

By answering questions like these through statistical techniques, the aim for providing insights for improving companies, business strategies can be reached.

**Exploratory data analysis EDA**

Loading the data set: the first step for performing the EDA in R studio it’s loading the data as shown below.

A screenshot of a computer

Description automatically generated

**Structure of the data set:** understanding the structure of the data set that you’re working with is also important. The command performed below shows the data set for you.

A screenshot of a computer code

Description automatically generated

The command perform below, shows the dataset that is uploaded in RStudio.

A screenshot of a computer screen

Description automatically generated

After loading the data set, we use the command below to check for the missing values. Through this command below we will make sure that there are no missing values, and the dataset is ready for the next steps.

A computer screen shot of a graph

Description automatically generated

The summary function below gives us a statistical summary of each column in the dataset.

A screenshot of a computer screen

Description automatically generated

**Summary statistics:**

The next step in analyzing the data said in R Studio is summary statistics which in this assignment will be computed for both numerical and categorical variables. First installation of psych package as required which will happen with the help of the command below. Functions, like summary and psych will be key here.

A screen shot of a computer

Description automatically generated

Descriptive statistics are used here to measure the central tendency, and other numerical variables present in the data set.

A screenshot of a computer screen

Description automatically generated

Table command is used to check the frequency of customer types.

A close-up of a computer screen

Description automatically generated

Table command is also used here for the frequency of product line.

A screen shot of a computer

Description automatically generated

**Data visualization:**

Data visualization graphs have been used here to display the information and present it better for the understanding of the data to the reader.

- Histogram graph has been used to show the unit price and its distribution off continuous variable.

- Bar graph has been used to show the product line and the product line distribution. The mian idea here is to visualize the frequency of categorical variables.

- Scatterplot graph is used to explore the relationship between unique prices and total. Job here is to show the relationship between the two continuous variables.

- The command sam() is used to check for the missing values.

A screenshot of a computer code

Description automatically generated

A graph of a distribution of unit price

Description automatically generatedA graph of green bars

Description automatically generatedA graph with blue dots

Description automatically generated

**Correlation analysis:**

Correlation analysis help in understanding the strength and direction of the relationship between the numerical variables. The command cor() is used below to perform death analysis.

A screenshot of a computer

Description automatically generated

**Hypothesis in RStudio:**

**Step 1: Inspect the Data**

First, the data should be looked at the average sales for each gender.

A close-up of a white background

Description automatically generated

**Step 2: Perform a T-test**

With the use of a T-test to check if there is a statistically significant difference between the means of two groups. Here, the groups are Male and Female customers.

A screenshot of a computer code

Description automatically generated

**Step 3: Interpret the Results**

* Look at the p-value in the output of the T-test.
  + If the p-value is less than 0.05, reject the null hypothesis (H0) and conclude that there is a significant difference in sales between Male and Female customers.
  + If the p-value is greater than 0.05, fail to reject the null hypothesis, meaning there is no significant difference.

Interpretation of Results

The T-test results show:

Mean Total for Female customers: 335.0957

Mean Total for Male customers: 310.7892

p-value: 0.1181

Since the p-value is 0.1181, which is greater than the significance level of 0.05, fail to reject the null hypothesis (H0). This means there isn't enough statistical evidence to conclude that there is a significant difference in the average Total sales between Male and Female customers.

Data Pre-processing, Sampling, and Cleaning:

To achieve a processed clean sample the steps below are required:

Check for missing values: check for any missing data in the dataset.

Drop any unnecessary columns: Let's assume the Invoice ID column isn't necessary for the analysis, then drop it.

Sample the data: then take a random sample of 200 rows from the dataset to make analysis quicker.

View the cleaned and sampled data and if the result is achieved in R studio, we move to the next step.

A screenshot of a computer code

Description automatically generated

**Application of Descriptive and Inferential Statistics:**

**Step 1: Descriptive Statistics**

calculate basic descriptive statistics for numerical variables like Total, Quantity, and Tax 5%.

A screenshot of a computer

Description automatically generated

**Step 2: Inferential Statistics**

For inferential statistics, conduct a T-test to see if there's a significant difference between Total sales based on Customer type (Member vs. Normal).

**Hypothesis:**

Null Hypothesis (H₀): There is no significant difference in the mean total sales between Member and Normal customers.

Alternative Hypothesis (H₁): There is a significant difference in the mean total sales between Member and Normal customers.

After doing so the print command is used to show the result.

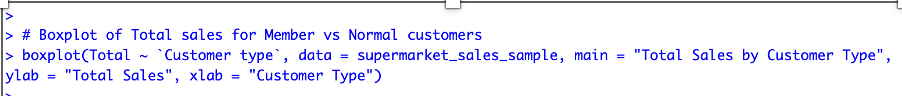
A screenshot of a computer

Description automatically generated

After getting done with the two steps, the conclusion below is made:

The t.test compares the mean Total sales between two customer types (Member and Normal).

The result will show the p-value and confidence interval, helping to determine whether there's a statistically significant difference.



A graph of a chart

Description automatically generated with medium confidence

**Performing T-Test and Printing Results:**

**The output of the T-test gives you:**

p-value: If the p-value is less than 0.05, we reject the null hypothesis, meaning there is a significant difference between Member and Normal customers.

**Confidence Interval:** Shows the range within which the true mean difference lies with 95% confidence.

**Mean of each group:** The average Total sales for Member and Normal customers.

The T-test results indicate that Members tend to spend more on average than Normal customers in total sales

A screenshot of a computer

Description automatically generated

**Conclusion:**

Through many steps and process applied in R Studio in the assignment above the conclusion below has been made.

The analysis of the supermarket sales from the dataset above shows no significant difference in the total sales between member and non-member customers. While member customers have a slightly higher mean total sales, the t-test indicates that this difference is not statistically significant.

**Business Implication**:

The supermarket could focus less on driving membership for sales increases since there is no large difference in spending between members and non-members.

However, further analysis could explore other factors, such as product lines or customer loyalty programs, to drive sales.

**Limitations**:

The analysis was limited to total sales and customer type due to it being for a mini research and analysis paper. Future work could involve examining other variables such as product category, payment method, or branch location for more comprehensive insights.

**GITHUB Link:**



Assessment Submission Form

|  |  |
| --- | --- |
| **Student Number**  (If this is group work, please  include the student numbers of all group participants) | GH1024093 |
| **Assessment Title** | Statistical Analysis Report: Supermarket Sales Dataset |
| **Module Code** | B105 |
| **Module Title** | Applied statistical modelling |
| **Module Tutor** | William Morrison |
| **Date Submitted** | 22/09/2024 |

**Declaration of Authorship**

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

I fully understand that the unacknowledged inclusion of another person’s writings or ideas or works in this work may be considered plagiarism and that, should a formal investigation process confirms the allegation, I would be subject to the penalties associated with plagiarism, as per GISMA Business School, University of Applied Sciences’ regulations for academic misconduct.

Signed………………Ali Jawed Delawari……………. Date ………22/09/2024………………