Supermarket Sales

Mohamed Ahmed Mohamed Abo Elfadl 42010595

Yousef Adel Ali 42010550

Ahmed Shaban Elsayed 42010530

Eslam Ashraf 42010488

Mahmoud Mohamed Abo Elnoor 42010353

Yousef Ali Kamal 42010565

Shimaa Nasser 42010621

What does data mean?

The growth of supermarkets in most populated cities is increasing and market competitions are also high. The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches for 3 months data. Predictive data analytics methods are easy to apply with this dataset.

Attribute information:

**Invoice ID**: Computer generated sales slip invoice identification number.

**Branch**: Branch of supercenter (3 branches are available identified by A, B and C).

**City**: Location of supercenters

**Customer type**: Type of customers, recorded by Members for customers using member card and Normal for without member card.

**Gender**: Gender type of customer

**Product line**: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel

**Unit price**: Price of each product in $

**Quantity**: Number of products purchased by customer

Tax: 5% tax fee for customer buying

**Total**: Total price including tax

**Date**: Date of purchase (Record available from January 2019 to March 2019)

**Time**: Purchase time (10am to 9pm)

**Payment**: Payment used by customer for purchase (3 methods are available – Cash, Credit card and E-wallet)

**COGS**: Cost of goods sold

**Gross margin percentage**: Gross margin percentage

**Gross income**: Gross income

**Rating**: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

Problem Definition:

**In this data set we had to deal with:**

• **Outliers**: Outliers in the 'Total' column and ‘cogs’ column.

• **Data Cleaning:** Drop Duplicates, gross income column and Outliers.

• **Missing values:** No Missing values.

• **Data types:** No Wrong data type.

Method:

Using python, the code handles:

* **Data cleaning:** The code is dropping the ‘Gross income column, which appears to be an extraneous column with no useful information.
* **Outlier:** The code is identifying and handling potential outliers in the 'Total, COGS' column by replacing them with the median value.
* **Data visualization:** Various visualizations, such as histograms, and box plots, are created to explore the relationships between different variables in the dataset.

Experiment:

Potential outliers in the 'Total' column are identified and handled by dropping them.

We have used some questions in data analysis to help us understand the dataset and the relationship of columns to each other.

References:

https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales