SUPERMARKET SALES DATASET

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

Context:

The growth of supermarkets in most populated cities are 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.

Description of columns:

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 Ewallet)

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)

Solve these questions

1. Read the ***supermarket_sales.csv*** dataset in your Google Colab as ***sales***.
2. Determine the size of ***sales***.
3. Display the last 3 rows of ***sales***.
4. Identify the data type of each column of ***sales*** and determine if there are any null values in them.
5. Identify the type of data represented by each column : categorical, numerical, index/identifier.
6. Check for missing values in the dataset.
7. Plot a histogram of the gross income of customers for all the sales (datapoints).
8. Which city is most represented in the dataset, illustrate your answer with bar chart?
9. Plot a visualization of the correlation matrix of the dataset using the heatmap() function from seaborn (as was done for the confusion matrix during the deep learning tutorial). What do you expect about the correlation between 'tax 5%' and 'total' variables?
10. Is there a gender bias in the data, illustrate with a bar chart again.
11. Compare the products bought and the payment method used by males and females. What do you conclude?
12. In which city does 'home and lifestyle' accounts for the biggest proportion of sales?
13. Illustrate the respective proportions of product bough for each city using pie charts.

Important Guidelines:

1. Group has to prepare one Presentation with 5-10 Slides answering all the questions asked above (maximum 10 mins).
2. Slides should include the followings (1-2 Slide(s) for each): Project Title and Student Names, Problem Statement, Data Description and Source, Answers to the Questions asked (2-4 slides), Name of the Methods used, What did you learn from this.
3. Present the project in front of a Jury followed by Question and Answer Session. Evaluation will be based on your work, presentation and answers to questions raised by the Jury.

BEST OF LUCK !