IST 707 Applied Machine Learning

HW2: Tell a Data Story

Growth of supermarkets in most populated cities is increasing and market competition is high. The following dataset contains historical sales of a supermarket across 3 different branches for a period of 3 months.

Assume you are working as a data scientist for the supermarket. **Tell the story** of this data by using appropriate data exploration and transformation techniques.

You are required to provide insights into sales data across branches. For example, what's the gross income distribution over different branches? Furthermore, are there gender differences in each branch?

Generally, you need to explore all highlighted variables below and tell the story (or stories) in this data. Don't forget relationships between or combinations of variables.

Find the story, tell it visually and, above all, truthfully.

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 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)

Source: https://www.kaggle.com/aungpyaeap/supermarket-sales?select=supermarket_sales+-+Sheet1.csv

Read the Data

Our dataset has 17 attributes. From this 17 our professor has requested to analyze only 9. The attributes that I have used in this homework's analysis, are: Branch, Customer type, Gender, Product line, Quantity, Time, Payment, Gross income and Rating.

Access Data Quality

Before analyzing each of the attributes, I have checked for null and duplicated values.
Removing null values and duplicates is necessary to ensure that the results of a data analysis are accurate, reliable, and relevant. It also helps to ensure that the data is well-suited for the specific analysis techniques being used and reduces the risk of encountering issues or errors during the analysis process. There are no null values and no duplicated values.

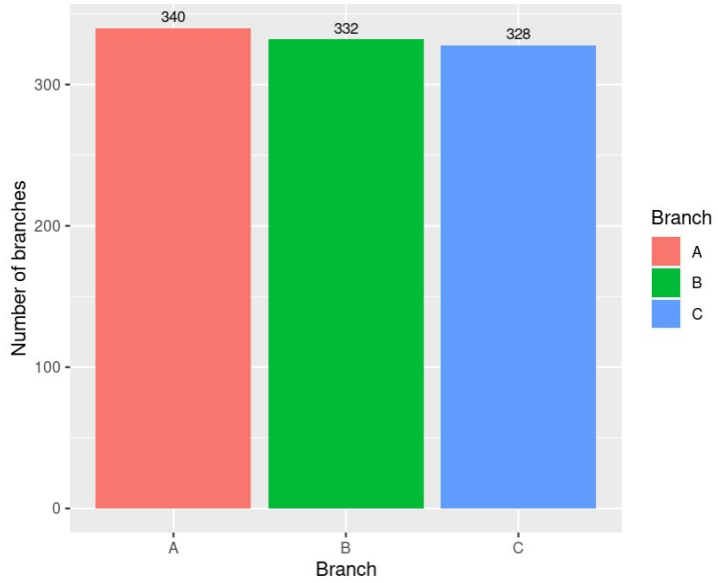
Transform Data for Analysis

To have a proper analysis for qualitative data, we need to transform data from char data type to factors. Factors are useful in statistical modeling and data analysis as they allow for convenient and efficient representation of categorical variables, while also providing methods for transforming, summarizing, and plotting the data. By converting a character variable to a factor, you can ensure that R treats the variable as categorical data, and use factor-specific functions and methods in my analysis. Branch, Customer type, Gender, Product line, Payment and Rating are all categorical attributes.

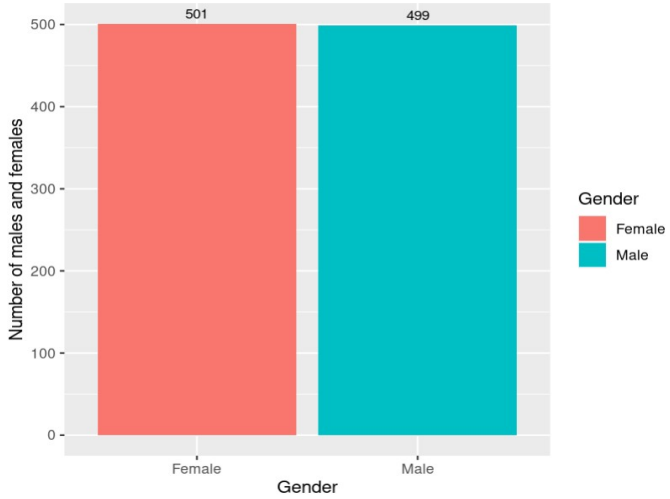
Analyzing the data

After cleaning and transforming the data, the next step is to analyze the data.
Initially, each attribute is analyzed individually to assess their distribution.
For the attributes of type factor, we verify how are they distributed in the dataset. So, for the following attributes we have these distributions:

Branch attribute is divided into 3 categories, Branch A, B and C. The super center has 340 A branches, 332 B branches and 328 C branches as also shown in the chart below.

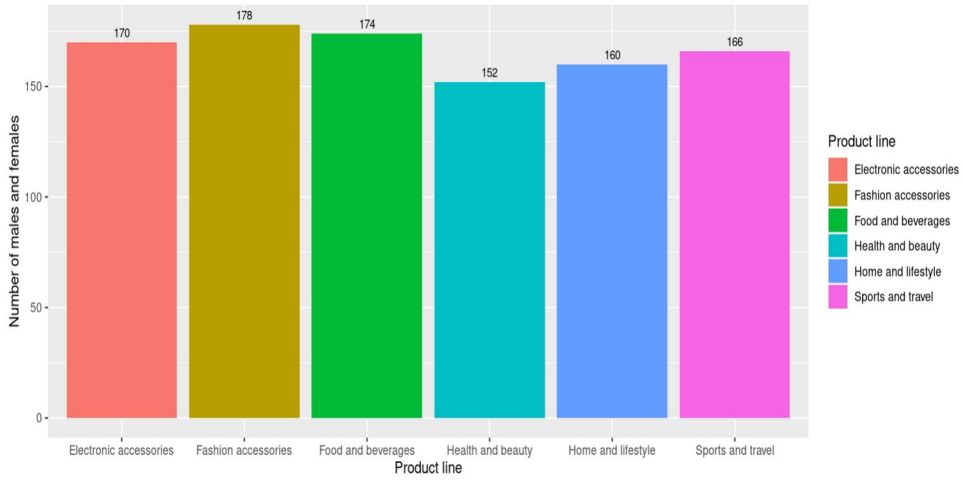


Another information that I am providing is the total number of females and males that work within the company. As we see the number of females is only 2 more than the number of males.

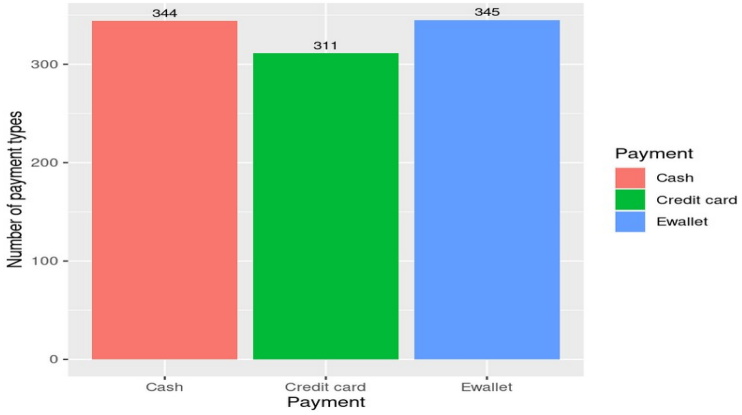


Product line shows the general item categorization group. There are 6 categories in this attribute.
Electronic accessories (170 obs);
Fashion accessories (178 obs);
Food and beverages (174 obs);
Health and beauty (152 obs);
Home and lifestyle (160 obs);

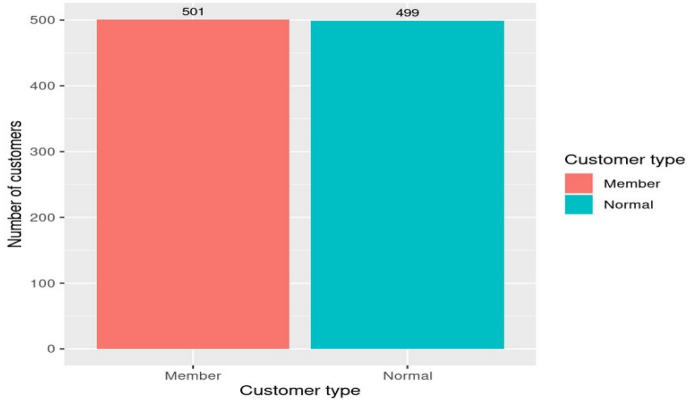
Sports and travel (166 obs)



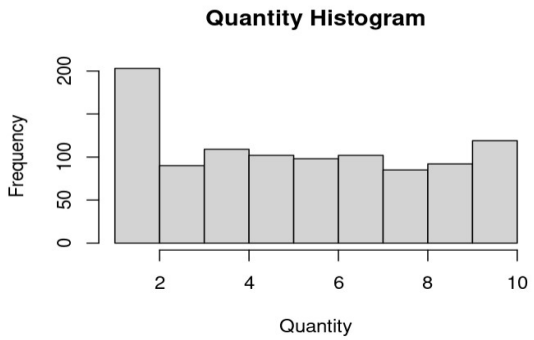
There are 3 different types of payments cash, credit card and Ewallet. From what we see from the chart below, people prefer to pay more with Ewallet, followed by cash and credit card in the end.

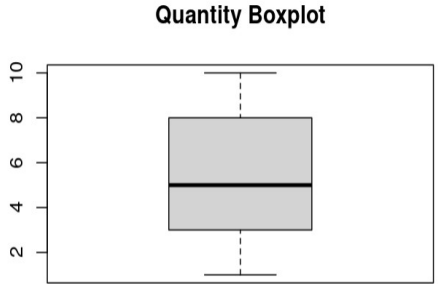


There are 2 customer types in our dataset. The ones that are members and just the normal customers. From what we see, the number of normal and member customers is almost the same.

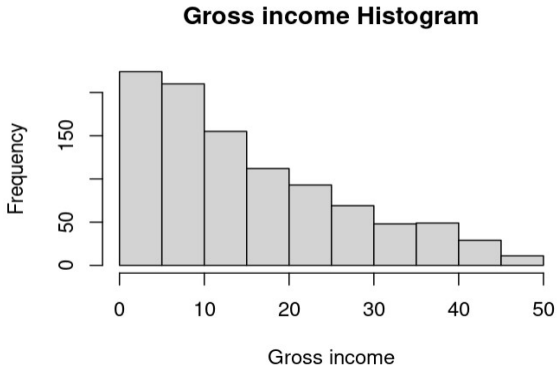


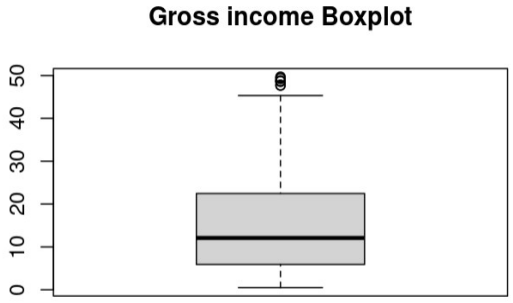
Next step is to check the distribution for the numerical variables.
Quantity: the histogram indicates that most of the records have a quantity of 2 products. The minimum of quantity per record is 1 and the maximum is 10. Most of the values fall in quantity between 3 to 8, but in average the quantity of product in invoices is 5.51.





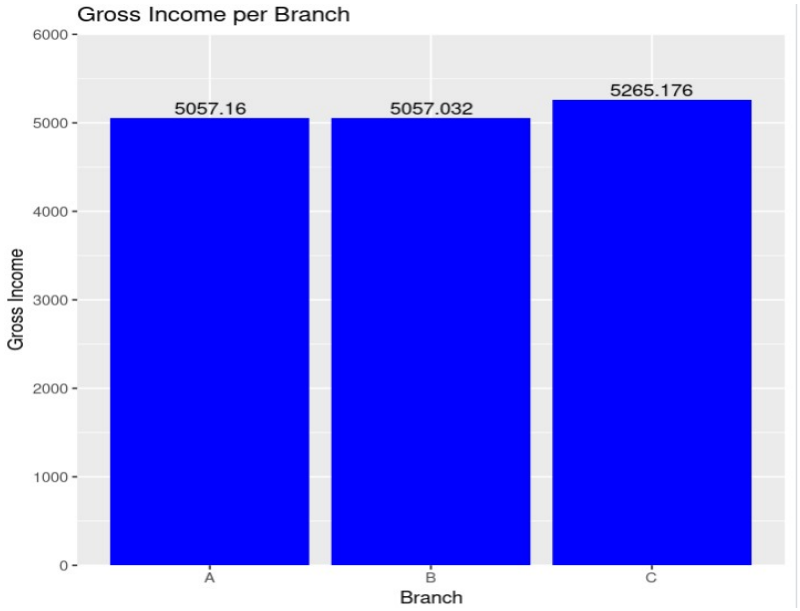
Gross income: The minimum is 0.5085 and maximum 49.65. Also, some values as outliers are identified. Most of the data from gross income fall in the range of 5.9 to 22.5 as we also see from the box plot below too.





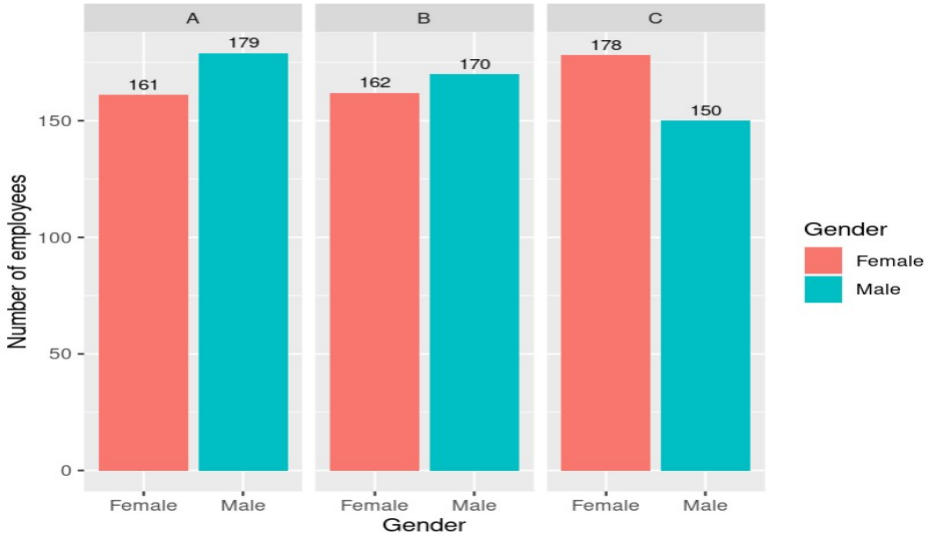
What is the gross income distribution over different branches?

There are 3 different branch types A, B, C. To find the gross income of each of the branch types, we need to group by each of the branch types and then summarize their incomes to find the gross income for each branch type. As we see also from the bar chart below, branch C has the highest gross income, followed by B with just a little difference from A.



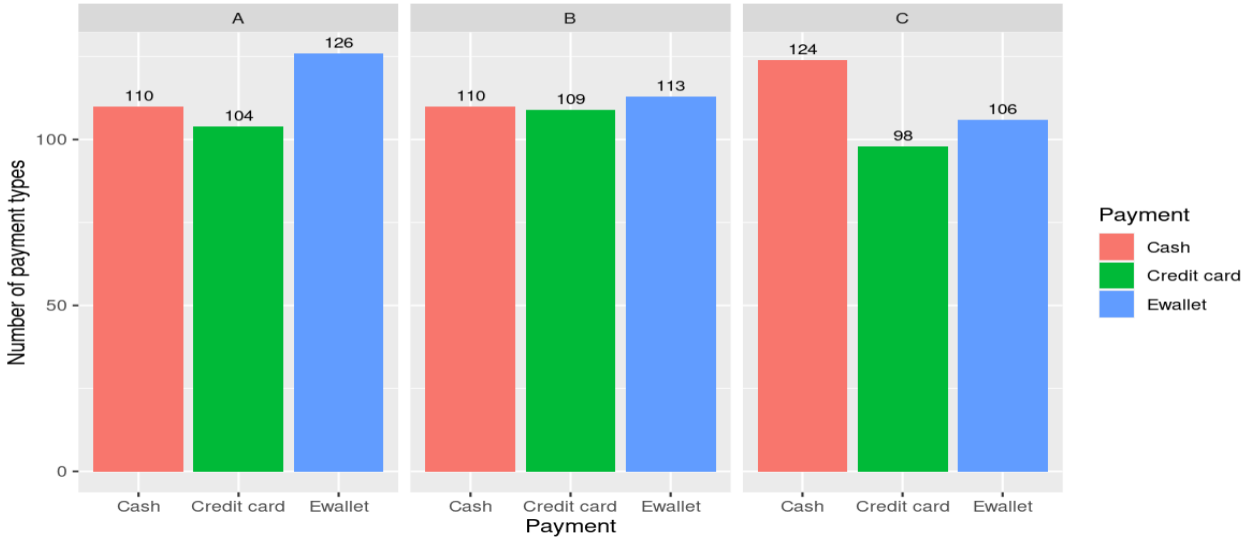
Are there gender differences in each branch?

As we see from the chart below, both branches A and B have more male workers than female workers. There are 18 and 8 male workers more in branches A and B respectively. While in brach C, there are 28 female workers more than male workers.



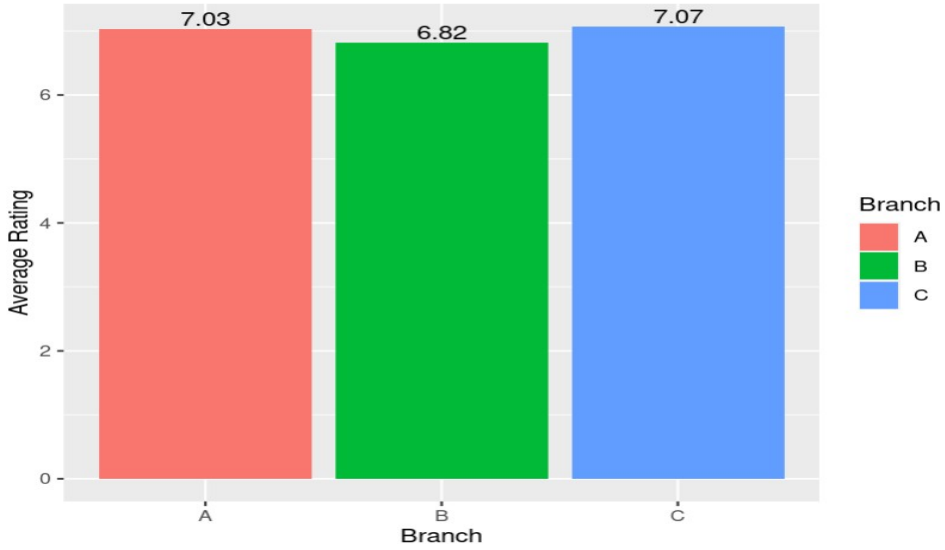
What are the most preferred payment types in each of the branches?

In this part of analysis we compare the payment types in each of the branches. As we see, people have preferred to pay with Ewallet more in branches A and B followed by cash and then credit cards. In branch C, people have preferred to pay with cash more, followed by Ewallet and credit card in the last place.



In which of the branches the customer are more satisfied?

To find in which branches the customers are more satisfied, I have taken the average for each branch types and as we see, customers in branch C are the most satisfied with the shopping experience, followed by A and in the end B with the lowest average of 6.82.



What are the sales distribution in each branch in each hour?

In the figure below we see the distribution of sales in each hour in each of the branches. From what is shown, for branch A, 10 AM in the morning, is the time, where most of the sales have occurred and 8 PM the time with the lowest sales 22 of them. For branch B, 9 PM is the time, when they have the highest number of sales and 4 PM the lowest. Branch C has the highest number of sales same as branch A at 10 AM and the lowest number of sales at 11 AM.