STATISTICS FOR MANAGEMENT – DECEMBER HOMEWORK

Tommaso Cecchellero (ID: 2203926)

INTRODUCTION:

I divided my analysis into two main steps.
Firstly, I sourced a suitable dataset considering the tasks that I had to follow. Then, I reported its reference (Link of the web page) and a quick explanation about its attributes.
Secondly, I performed all the kinds of analysis that we observed in the last part of the statistics' course, and I commented the results that I obtained.

DATASET:

The dataset that I used to perform the following analysis is about the "Sales of a Supermarket". It can be found on the following web page: https://www.kaggle.com/datasets/lovishbansal123/sales-of-a-supermarket?resource=download .

The attributes of the dataset are:

- 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).

METHODS AND RESULTS:

We can divide the methods and results section in 3 main parts:

1) HYPOTHESIS TESTING
In this part I performed all the types of hypothesis testing that we observed during our lectures.

I started with ONE SAMPLE T-TEST. I tried to predict the average unit price of the dataset, setting the value at 40$. Then, I computed the t-test using the t.test() function. I discovered that my idea was completely different from the real situation, where the average unit price is 55,67$.
Also, I tried to predict the average customer satisfaction of the dataset, setting the value at 7. Then, I computed the t-test using the t.test() function. I discovered that my idea was quite close from the real situation, where the average customer satisfaction is 6,97.

I continued with TWO SAMPLE T-TEST. I tried to understand if there is a difference in gross income between male and female customers. So, I computed the t-test using the t.test() function. I discovered that there isn't a significant difference in the gross income related to the gender of the customers.
Also, I tried to understand if there is a difference in customer satisfaction between male and female customers. So, I computed the t-test using the t.test() function. I discovered that neither in this case there is a significant difference in the customer satisfaction related to the gender of the customers.

I would continue with PAIRED T-TEST, but I noticed that in the dataset that I used there aren't two variables that could be useful for this type of hypothesis testing because the dataset should have a variable referred to a situation before a certain event and the same kind of variable referred to a situation after a certain event.
For example, to find if there is a significant difference in the customer satisfaction for purchases made before and after a promotional campaigns.

I continued with ONE PROPORTION Z-TEST. I tried to predict if the proportion of customers that use E-wallet as a payment method is 60%. So, I found the real value of the proportion that I tried to predict and I used it to compute the z-test using the prop.test() function. I discovered that my idea was very different compared to the real situation, where the real proportion of the customers that use E-wallet is about 34%.
Also, I tried to predict if the proportion of customers that buy 'Electronic Accessories' is 40%. So, I found the real value of the proportion that I tried to predict and I used it to compute the z-test using the prop.test() function. I discovered that my idea was very different compared to the real situation, where the real proportion of the customers that bought 'Electronic Accessories' is about 17%.

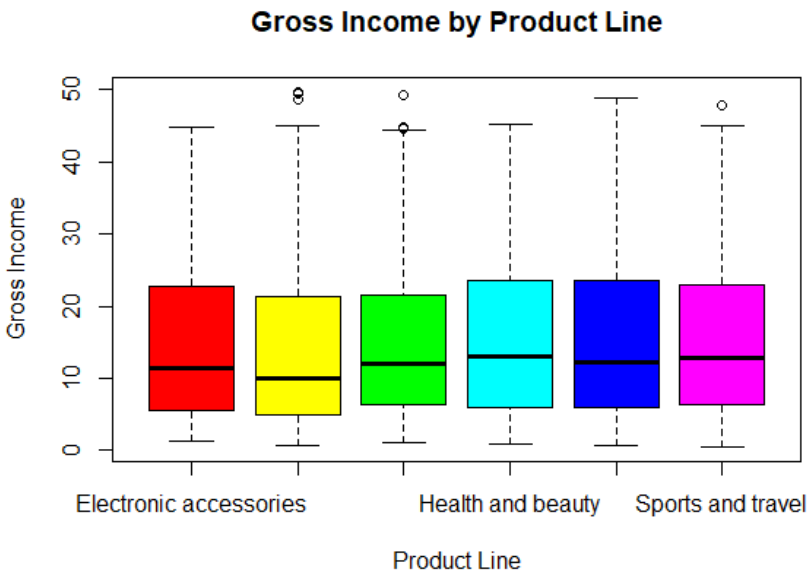
I continued with TWO PROPORTION Z-TEST. I tried to understand if there is a difference in the proportion of male and female customers using Credit Card as a payment method. So, I found the real number of male and female customers that use the Credit Card and the total amount of male and female customers and, then, I computed the z-test using the prop.test() function. I discovered that there isn't a very significant difference considering the two proportions.
Also, I tried to understand if there is a difference in the proportion of male and female customers that are member of the supermarket. So, I found the real number of male and female customers that are Members of the supermarket and I found the total amount of male and female customers and, then, I computed the z-test using the prop.test() function. I discovered that there isn't a very significant difference considering the two proportions.

Finally, for this first part, I performed the CHI-SQUARE TEST. I tried to understand if there is a dependency between the type of customers and the payment method. So, I created a contingency table with the two considered variables and I computed the chi-square test using the chisq.test() function. I discovered that the two variables are quite dependent, probably because Members might have higher spending methods as they are more likely to be regular shoppers.
Also, I tried to understand if there is a dependency between the gender of the customers and the number of products that are bought. So, I created a contingency table with the two considered variables and I computed the chi-square test using the chisq.test() function. I discovered that the two variables are quite dependent, probably because women might shop more frequently for items in categories like Fashion Accessories or Home & Lifestyle, leading to more products overall.

2) ANALYSIS of VARIANCE (ANOVA)
In this second part I performed an ANOVA analysis to understand two features.

Firstly, I tried to understand if gross income differs across product lines. So, I computed the ANOVA test using the aov() function. I discovered that there isn't a significant difference across the groups.

Graphically:



Secondly, I tried to understand if customer type differs across customer satisfaction. So, I computed the ANOVA test using the aov() function. I discovered that there isn't a significant difference across the groups.

Graphically:

3) REGRESSION ANALYSIS

In this third and last part, I performed all the types of regression analysis that we observed during our lectures.

I started with the LINEAR REGRESSION. I tried to predict the gross income using one predictor: Quantity. So, I computed the regression using the lm() function. I discovered that quantity is a good predictor of the gross income, which makes sense because the more products are bought, the higher is the value of gross income.

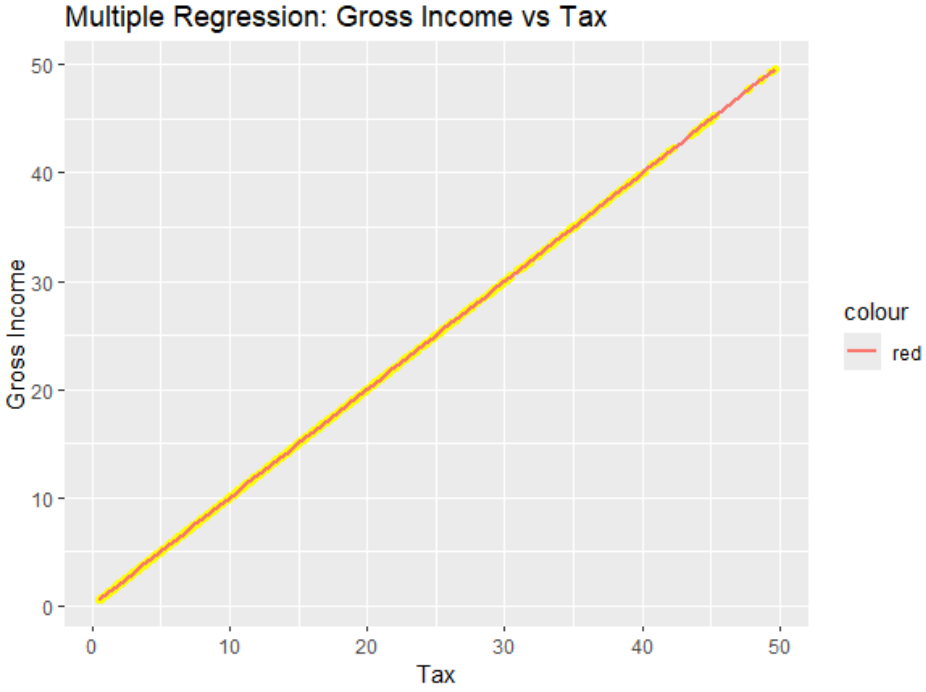
Graphically:

Also, I tried to predict the customer satisfaction using one predictor: Quantity. So, I computed the regression using the lm() function. I discovered that quantity is a good predictor of the customer satisfaction, which makes sense because the more a customer is satisfied, the higher is his willingness to buy more products.

Graphically:

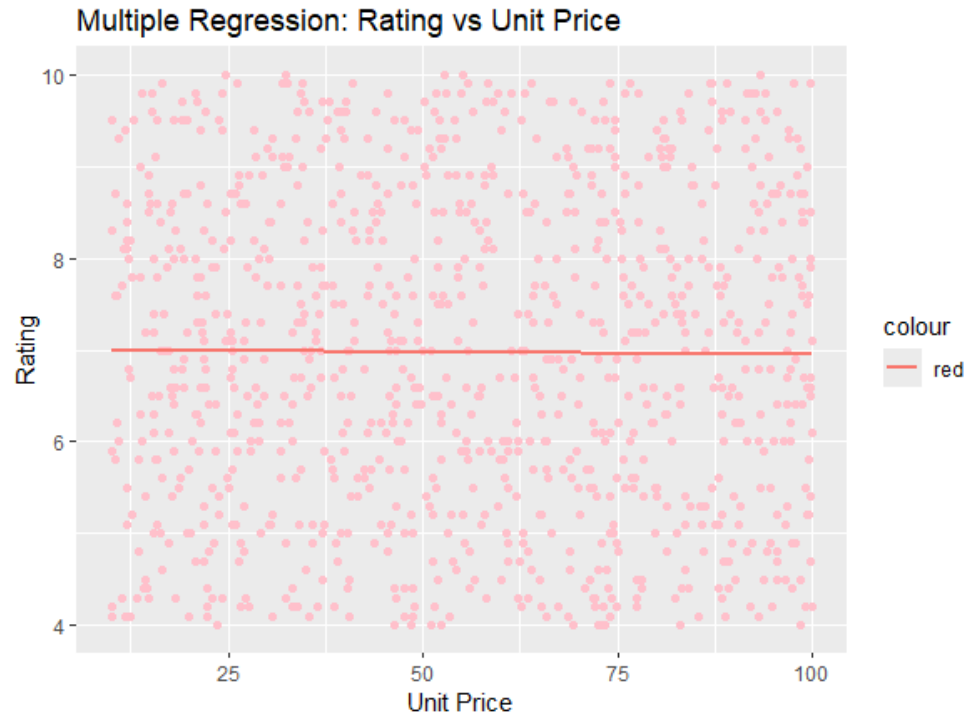
I continued with the MULTIPLE REGRESSION. I tried to predict the gross income using three predictors: Quantity, Unit.price and Tax. So, I computed the regression using the lm() function. I discovered that all the predictors that I used are good to understand the value of gross income, which makes sense because the gross income is bigger, the greater is the quantity of products that allows to increase the unit.price and the amount of tax.

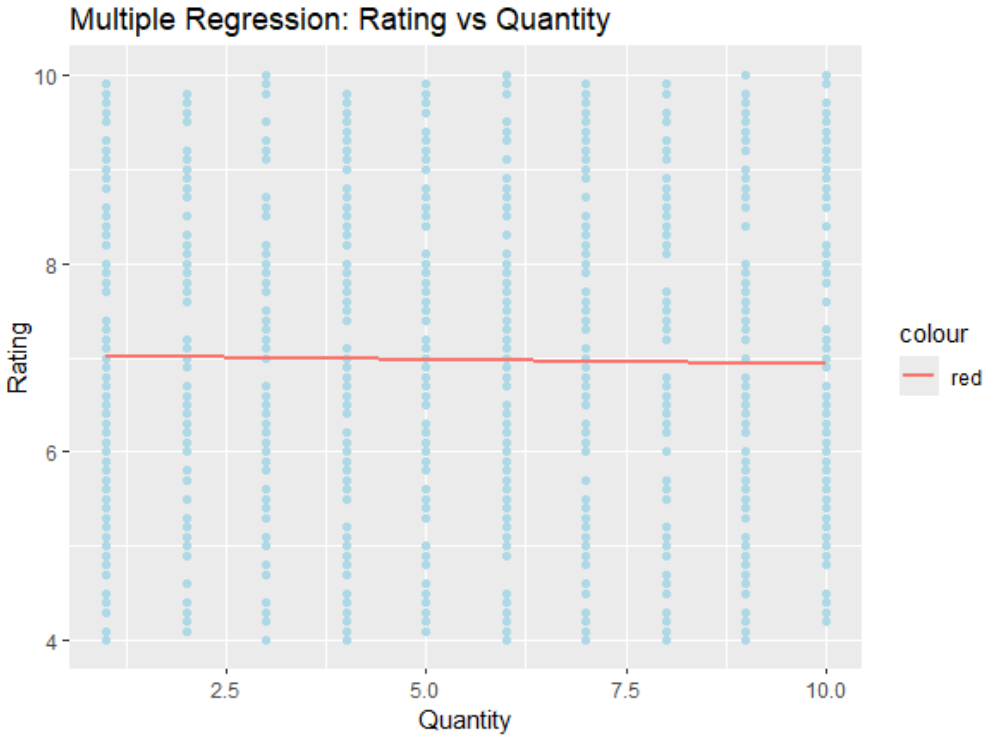
Graphically (more complex because there are more variables considered):



Also, I tried to predict the customer satisfaction using two predictors: Unit.price and Quantity. So, I computed the regression using the lm() function. I discovered that all the predictors that I used are not good to understand the value of customer satisfaction, which makes sense because if a customer spend a lot of money to buy a lot of products, it doesn't mean that he will be satisfied about its purchases.

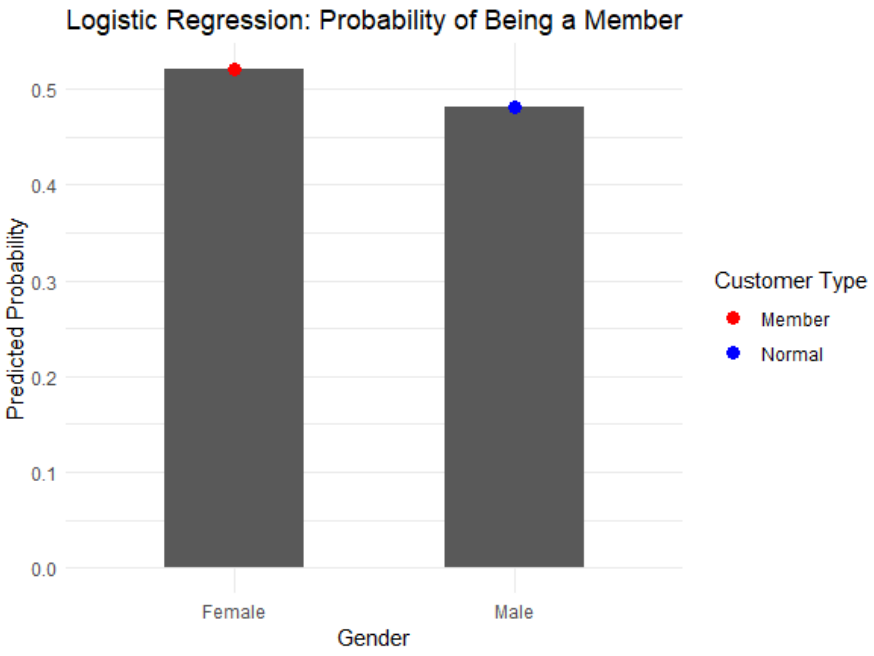
Graphically (more complex because there are more variables considered):





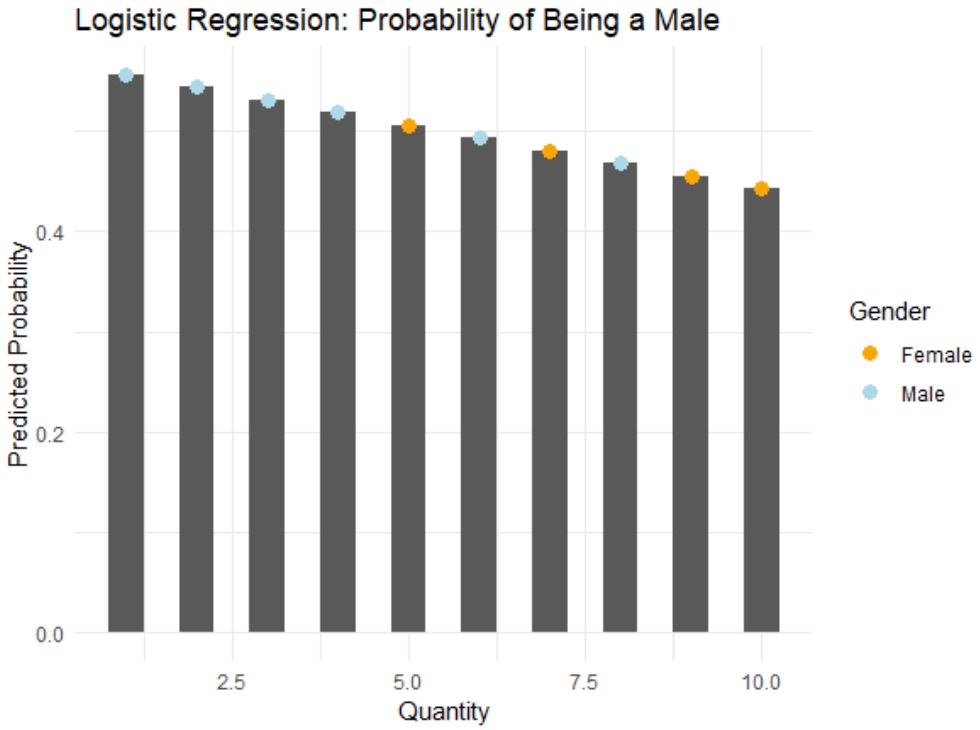
I continued with the LOGISTIC REGRESSION. I tried to predict the type of customer based on the gender. So, I rearranged the variable 'Customer.type' making it binary and computed the regression using the glm() function. I discovered that there is no significant evidence that suggest that the gender of the customers influences the type of customer, which makes sense because it's a subjective decision to become a Member of a specific supermarket, it's based on personal behaviours.

Graphically:



Also, I tried to predict if a customer is a male or a female based on the quantity of the products that he/she has bought. So, I rearranged the variable 'Gender' making it binary and computed the regression using the glm() function. I discovered that there is no significant evidence that suggest that the quantity of the products that a customer bought influences the gender of the customer, which is strange because, in my opinion, females are used to buy more products compared to males' customers.

Graphically:



Finally, I performed the POISSON REGRESSION. I tried to predict the number of items purchased (Quantity) based on the Gender of the customer. So, I computed the regression using the glm() function. I discovered that Gender is a significant predictor of the purchase count, which makes sense based on my personal experience.

This result is the opposite of the last logistic regression because, if we consider the AIC value, in the last case AIC_logistic = 1384.8, which is much lower compared to AIC_poisson = 5109.1. This means that my prediction fits well the Poisson Model. So, we have a more predictable result.

Graphically:

Also, I tried to predict the customer satisfaction based on the Gender of the customers. So, I computed the regression using the glm() function. I discovered that gender isn't a good predictor of customer satisfaction because this last variable is subjective, it's based on the individual purchase experience of each customer that doesn't depend on the Gender.