**Project #1**

**Sales Report for the supermarket**

**Prepared by :**

Omar Sami Mousa

**Overview**

**This sheet shows the results of sales for the supermarket to** 1000 **invoices in** 3 **branches .**

**# Of customers in Branch A, Yangon city: 340**

* The member customers 167 (80 F, 87 M).
* The normal customers 173 (81 F,92 M).

**# Of Customers in Branch B, Mandalay city:** **332**

* The member customers 165 (85 F, 80 M).
* The normal customers 167 (77 F, 90 M).

**# Of Customers in Branch** **C, Naypyitaw:** **328**

* The member customers 169 (96 F, 73 M).
* The normal customers 159 (82 F, 77 M).

**The supermarket have many products such as (food and beverages , health and beauty , sports and travel , fashion accessories , home and lifestyle , electronic accessories).**

**We have** 1000 **rows and** 17 **columns the types of columns**

* Invoice ID , Branch , City, Customer type , Gender , Product line **,** Date , Time , Payment **: character.**
* Unit price **: numeric** , this column has a range from 10-100**.**
* Quantity **: numeric** , the range from 1-10**.**
* Tax **: numeric**  ,the range from 0-50**.**
* Total **: numeric ,** the range from 0-1100**.**
* Cogs and Rating **: numeric ,** (Rating) the range from 5-10.
* Gross margin percentage **: numeric ,** the range from 0-5.
* Gross income **: numeric ,** the range from 0-50**.**

library(readxl)

library(tidyverse)

library(ggplot2)

path = 'C:\\Users\\user\\Desktop\\Statistical for data science\\Project\\supermarket\_sales.csv'

my\_dataset = read.csv(path)

my\_dataset

# 1

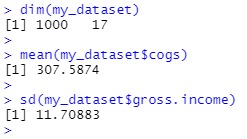
dim(my\_dataset)

# 2

mean(my\_dataset$cogs)

# 3

sd(my\_dataset$gross.income)



# 1 : to show how many columns and rows .

# 2 : to get the mean from column “cogs”(cost of goods sold).

#3 : get the Standard Deviation from column “gross.income”

\_\_\_\_\_\_\_\_\_\_

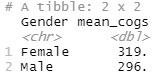
# 4

my\_dataset %>% select(Gender,cogs) %>% group\_by(Gender) %>%

summarise(

mean\_cogs = mean(cogs)

)



# 4 : to find the mean for column “cogs” for all each male and female.

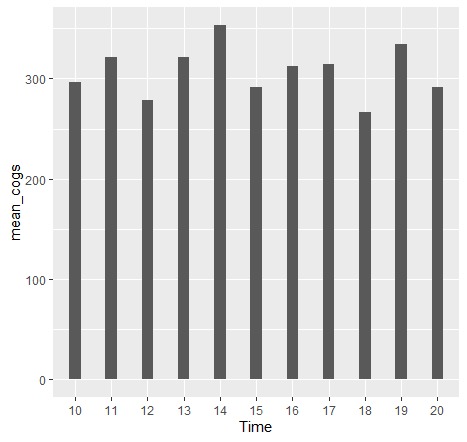
# 5

my\_dataset %>% mutate(Time = substr(Time,start = 1,stop = 2)) %>% select(Time, cogs) %>%

group\_by(Time) %>% summarise(mean\_cogs = mean(as.double(cogs))) -> time\_vs\_cogs

time\_vs\_cogs %>% ggplot(aes(x = Time,y = mean\_cogs)) + geom\_bar(stat = 'identity',width

= 0.32)



# 5 : in the visualization show the high time to cost of goods sold the time best at 2, and low time at 6

\_\_\_\_\_\_\_\_\_\_

# 6

dtatFram\_for\_Male = data.frame(my\_dataset %>% filter(Gender == 'Male') %>% select(gross.income)) %>%

mutate(gross\_income\_for\_male = gross.income) %>%

select(-gross.income)

head(dtatFram\_for\_Male,5)

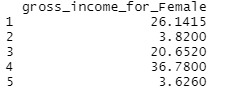
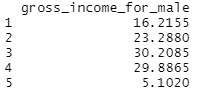
dtatFram\_for\_Female = data.frame(my\_dataset %>% filter(Gender == 'Female') %>% select(gross.income)) %>%

mutate(gross\_income\_for\_Female = gross.income) %>%

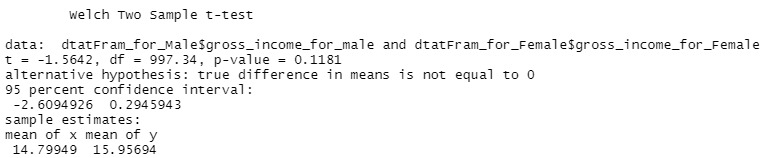
select(-gross.income)

head(dtatFram\_for\_Female, 5)

t.test(dtatFram\_for\_Male$gross\_income\_for\_male, dtatFram\_for\_Female$gross\_income\_for\_Female)



# 6 : show each group male and female for gross income



# 6 : notice the p-value 0.1181 is greater than 0.05 and this means that we can’t refuse the H0 which states that the averages are the same.

\_\_\_\_\_\_\_\_\_\_

# 7 : In column “Product.line” There are several products, and each customer wants quantities from column “Quantity” of these products, we use mean() statistics to find out which products have mean quantities.

# 7

data\_Product = my\_dataset %>%

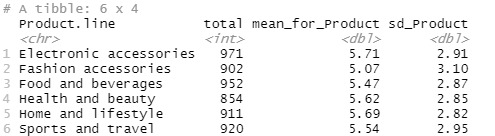
group\_by(Product.line) %>%

summarise(total = sum(Quantity), mean\_for\_Product = mean(Quantity), sd\_Product = sd(Quantity))

data\_Product

anova\_data1 = aov(Quantity~Product.line, data = my\_dataset)

summary(anova\_data1)



# 7 : The mean and standard deviation for each products is same average.



# 7 : sold we have found that the p value is 0.328 which means that we can’t refuse the H0.

\_\_\_\_\_\_\_\_\_\_

# 8 : every Branches need to collection the cost of goods sold and find the average :

# 8

data\_branch = my\_dataset %>%

group\_by(Branch) %>%

summarise(mean\_cog = mean(cogs), total\_of\_cogs = sum(cogs) , sd\_cogs = sd(cogs))

data\_branch

my\_dataset %>%

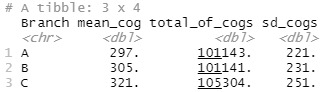
ggplot(aes(x = Branch, y = cogs)) +

geom\_boxplot()+

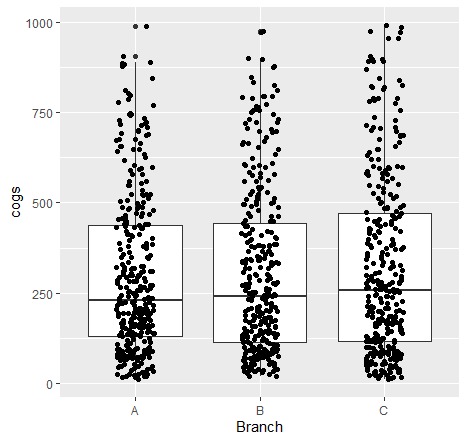
geom\_jitter(position = position\_jitter(0.15))

anova\_data2 = aov(cogs~Branch, data = my\_dataset)

summary(anova\_data2)



# 8 : branch ‘C’ have high average and biggest number for cost of goods sold and high standard deviation, while branches ‘A’ & ‘B’ they like each other.



# 8 : there are not different between branches, they have same cost of goods sold.



# 8 : After made the anova test on the type of “Branch” and the “cogs” sold we have found that the p value is 0.413 which means that we can’t refuse the H0.

\_\_\_\_\_\_\_\_\_\_

From column “Payment” customers have several ways to pay, Some pay cash, some pay with a credit card, and some pay with an electronic wallet, for each payment have rate from column “Rating”.

# 9

my\_dataset %>%

group\_by(Payment) %>%

summarise(mean\_for\_grossIncome = mean(gross.income), mean\_for\_rating = mean(Rating),

sd\_for\_grossIncome = sd(gross.income), sd\_for\_rating = sd(Rating))

anova\_data3 = aov(gross.income~Payment, data = my\_dataset)

summary(anova\_data3)

my\_dataset %>%

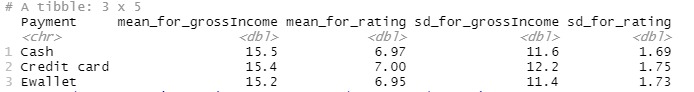
ggplot(aes(x = Payment, y = cogs)) +

geom\_boxplot()+

geom\_point()+

geom\_jitter(shape = 15,color = "steelblue", position = position\_jitter(0.15))+

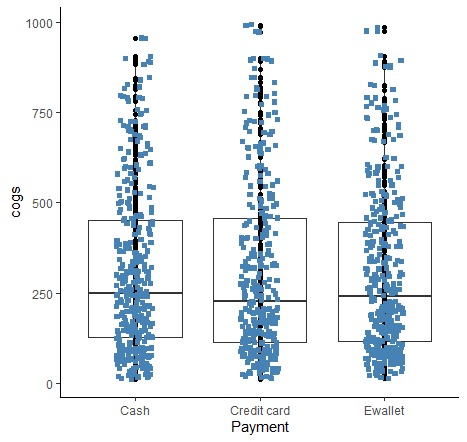
theme\_classic()



# 9 : the mean gross income and mean rating for each Payment is same, the standard deviation gross income and standard deviation rating for each Payment is same.



# 9 : After we made the anova test on the type of “Payment” and the “gross.income” sold we have found that the p value is 0.922 which means that we can’t refuse the H0.



# 9 : each paymrnt have same cost pf goods sold.

***The Sourse :***

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