

---

## Assignment 1

Féliz LUBERNE- 22508775

---

### Table des matières

<b>Question 1: Data Exploration and Visualization (20 points)</b> .....	2
<b>(a)</b> Read the dataset brand_ratings.csv into R. Construct a histogram plot (as below) using variable perform.....	2
<b>(b)</b> Load the dataset churn.arff into R. Create a bar plot using the variable REPORTED_SATISFACTION. Your output should look similar as the below graph.....	3
<b>Question 2 Describe Data (40 Points)</b> .....	4
<b>(a)</b> How many observations in this data set? What are the types (numeric, integer, etc.) of these variables?.....	4
<b>(b)</b> Which variable(s) belong to the discrete variable? Check the unique values for these discrete variables. Which variable(s) belong to the continuous variable? Check the values of mean, standard deviation, and range for these continuous variables. ....	5
<b>(c)</b> Construct a frequency table as below. ....	7
<b>(d)</b> Is variable X4 normally distributed? Use ggplot2 to create a QQ plot to help answer this question.....	8
<b>(e)</b> Recreate the following boxplot for variable X3 across the different levels of X2. The result should look like the below. ....	9
<b>(f)</b> Create a new variable X6 which is the sum of X3 and X4. Visualize the distribution of X6 as below. ....	10
<b>Question 3: Describe Data (40 Points)</b> .....	11
<b>(a)</b> Read the file marketing_campaign.csv in R and construct a subset named df2_sub where the variable Income contains no missing value, and variables NumStorePurchases and NumWebPurchases are not equal to 0. How many observations and variables are in this subset? ..	11
<b>(b)</b> What are the values of 10%, 50%, 80% percentile for variable Income?.....	12
<b>(c)</b> What Write a named function to compute the ratio of the interquartile value against the range of a variable. Apply that function to three variables in the dataset.....	13
<b>(d)</b> Write an anonymous function to solve the above question. ....	14

### Question 1: Data Exploration and Visualization (20 points)

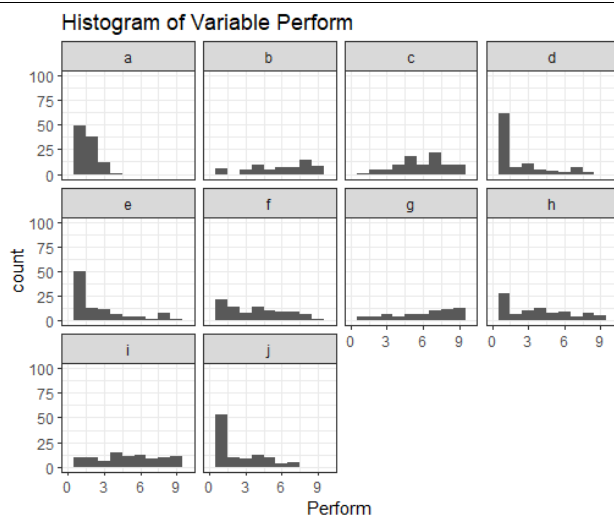
(a) Read the dataset `brand_ratings.csv` into R. Construct a histogram plot (as below) using variable `perform`.

#### Code to be entered

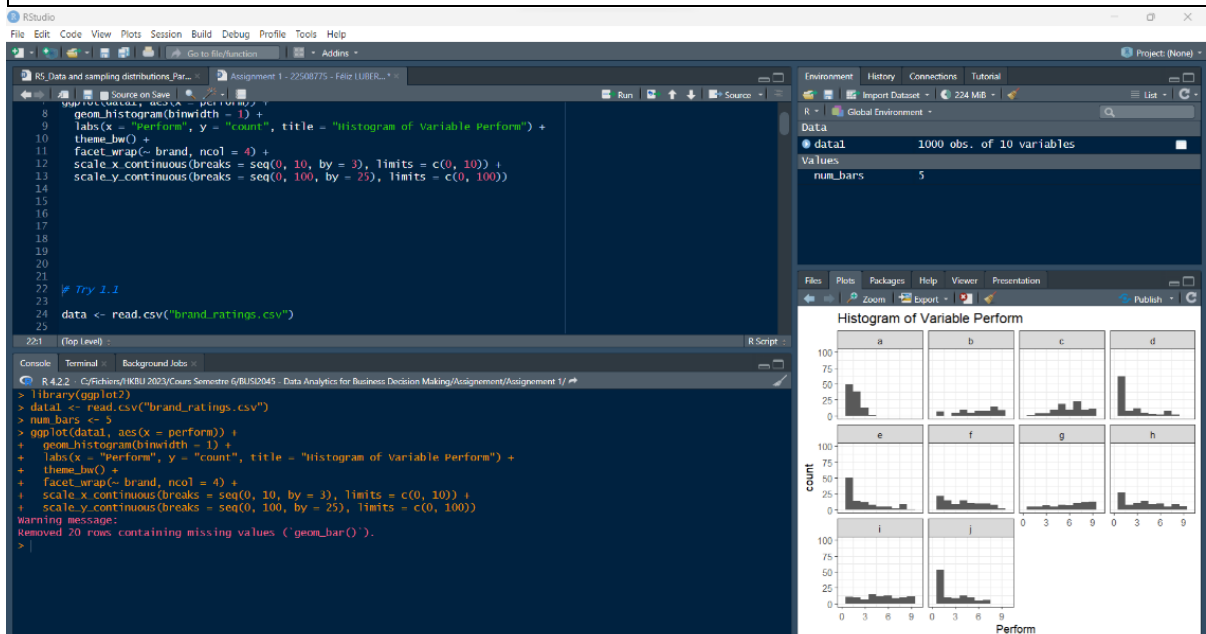
```
library(ggplot2)
data1 <- read.csv("brand_ratings.csv")

num_bars <- 5
ggplot(data1, aes(x = perform)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Perform", y = "count", title = "Histogram of Variable
Perform") +
  theme_bw() +
  facet_wrap(~ brand, ncol = 4) +
  scale_x_continuous(breaks = seq(0, 10, by = 3), limits = c(0, 10)) +
  scale_y_continuous(breaks = seq(0, 100, by = 25), limits = c(0,
100))
```

#### Results



#### Full Screen



**(b)** Load the dataset churn.arff into R. Create a bar plot using the variable REPORTED\_SATISFACTION. Your output should look similar as the below graph.

#### Code to be entered

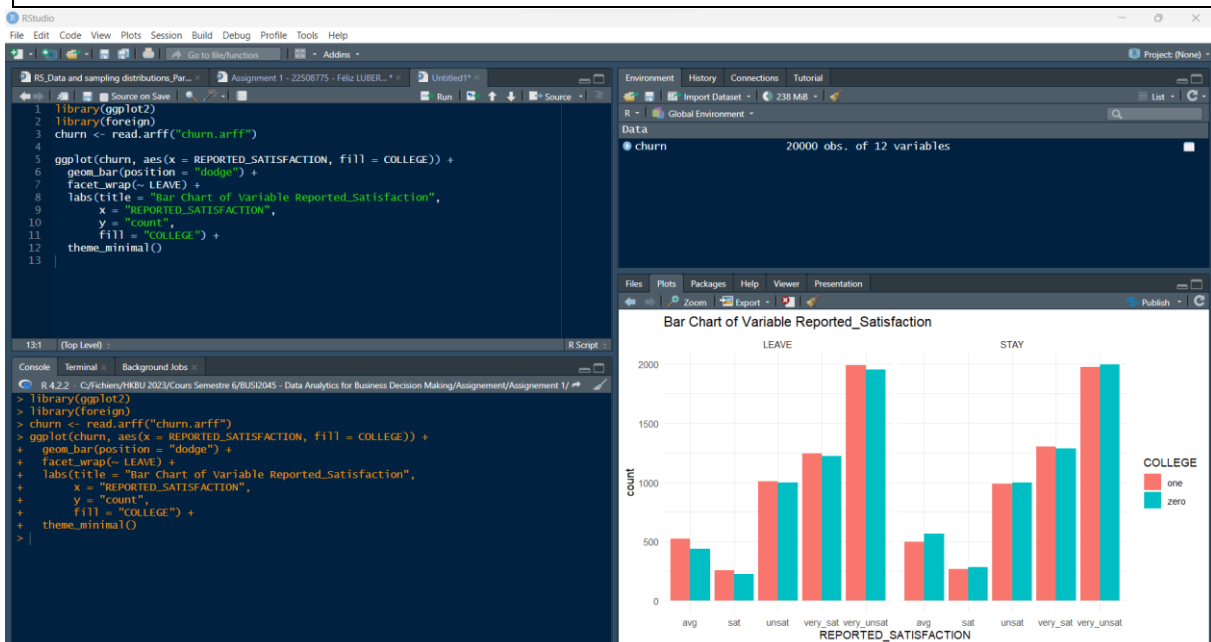
```
library(ggplot2)
library(foreign)
churn <- read.arff("churn.arff")

ggplot(churn, aes(x = REPORTED_SATISFACTION, fill = COLLEGE)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ LEAVE) +
  labs(title = "Bar Chart of Variable Reported_Satisfaction",
       x = "REPORTED_SATISFACTION",
       y = "count",
       fill = "COLLEGE") +
  theme_minimal()
```

#### Results



#### Full Screen



## Question 2 Describe Data (40 Points)

(a) How many observations in this data set? What are the types (numeric, integer, etc.) of these variables?

### Code to be entered

```
Q2_data <- read.csv("Assignment1_Q2.csv", header = TRUE)
Q2_data <- Q2_data[complete.cases(Q2_data$X3), ]

n_obs <- dim(Q2_data)[1]
cat("Number of observations in the data set: ", n_obs, "\n")

str(Q2_data)
```

### Results

```
> cat("Number of observations in the data set: ", n_obs, "\n")
Number of observations in the data set: 117
> str(Q2_data)
'data.frame':   117 obs. of  5 variables:
 $ X1: chr  "North" "West" "East" "South" ...
 $ X2: chr  "High" "Medium" "Medium" "Medium" ...
 $ X3: num  -6.66 5.22 11.68 -15.77 6.43 ...
 $ X4: num   3.94 9.59 5 5.54 -2.02 ...
 $ X5: int   0 1 0 1 0 1 0 1 0 1 ...
```

Answer: There is 117 observations in the data set with three types, which are:

- character ("chr"), the class of an object that holds character strings,
- numeric ("num"), which is used to convert a character vector into a numeric vector.
- Integer ("int"), is used to create integer data type in R, as by default, R shows the class of an Integer as Numeric.

### Full Screen

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains the R code for reading the CSV file, handling missing values, and displaying the number of observations and the structure of the data frame.
- Console:** Shows the output of the code execution, including the number of observations (117) and the structure of the data frame (117 obs. of 5 variables).
- Environment:** Lists the objects in the environment, including the data frame 'Q2\_data' with 117 observations and 5 variables.
- Files:** Shows the file explorer with various files and folders, including the CSV file 'Assignment1\_Q2.csv'.

**(b)** Which variable(s) belong to the discrete variable? Check the unique values for these discrete variables. Which variable(s) belong to the continuous variable? Check the values of mean, standard deviation, and range for these continuous variables.

#### Code to be entered

```
supply(Q2_data, class)
```

```
unique(Q2_data$X1)
unique(Q2_data$X2)
unique(Q2_data$X3)
unique(Q2_data$X4)
unique(Q2_data$X5)
```

```
summary(Q2_data$X3)
summary(Q2_data$X4)
summary(Q2_data$X5)
```

```
sd(Q2_data$X3)
range(Q2_data$X3)
```

```
sd(Q2_data$X4)
range(Q2_data$X4)
```

```
sd(Q2_data$X5)
range(Q2_data$X5)
```

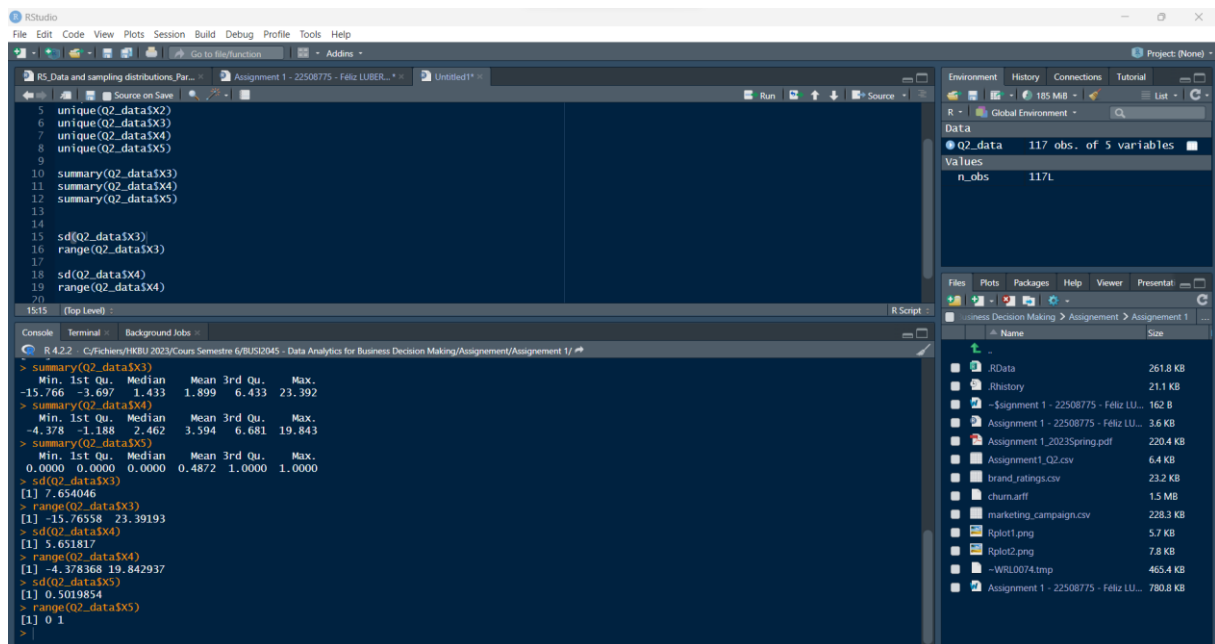
#### Results

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains the R code entered for the assignment.
- Console:** Displays the output of the code execution, including the class of each variable and the unique values, summary, standard deviation, and range for each variable.
- Environment Pane:** Shows the objects created in the environment, including 'Q2\_data' (117 observations of 5 variables) and 'values' (117).
- Files Pane:** Lists the files in the project, including 'RData', '.Rhistory', and various assignment files.

**Console Output:**

```
> supply(Q2_data, class)
      X1      X2      X3      X4      X5
"character" "character" "numeric" "numeric" "integer"
> unique(Q2_data$X1)
[1] "North" "West" "East" "South"
> unique(Q2_data$X2)
[1] "High" "Medium" "Low"
> unique(Q2_data$X3)
 [1] -6.65652600  5.21943394 11.67552941 -15.76558162  6.43299751  7.04844714 -1.59791968 -1.37305485 -1.51561599 -4.12630263
[11] -0.81794160 -4.98709156 -3.21003116  3.51567054 10.67595247  2.11771604 -1.08807605 -4.28956333 -3.69737344 22.32668143
[21] 4.07270576 -0.52438298  6.67671553 -2.54976198 -8.58563928  7.59804577  2.87889360 -4.48758881 11.81838037 -0.80474463
[31] -2.67552030 -1.01006448 -10.03274775 -6.34095410 -14.44031719 -7.72794554  0.64564913 -0.72718032 14.59597012 -5.54914179
[41] -3.84291707  0.75501598 -4.95472061 -4.74811454 -5.85854554 -7.01588709 -1.19062495 -0.97479966 -11.44825005 -1.65660740
[51] -5.87111700 -5.11969608  1.70152381  7.50444655 16.18253978 -3.18682739 15.84727703 -6.26246838  8.25270771 23.39192857
[61] 2.72191688 -2.35706864  2.93016195 17.21667559 -6.10886189 13.94261744 13.63651833  5.69178238  3.05514271 -0.64374991
[71] 0.06780854  8.18629254 19.56216689 1.77281270 -8.12560757 -2.78865422  5.06609410 1.57768034  1.64004739 -7.97841509
[81] 1.60970264  9.80185806  8.58086969  7.39997881 -0.22185580  1.46724984 -6.55622304  2.57472945  5.04156801 16.64771206
[91] 11.01210602 -0.96466754  5.84440238 -6.07686435 10.02562902 10.78333403 19.96893684  6.31618827 -0.79774779  3.52794795
[101] -1.01982226 -3.60798870  4.23591424 -4.17011701  4.34548310  5.83974609  2.58315907  1.43252305 -2.19253801 -5.87813785
[111] 9.79419362  3.17890021  9.64912494 -6.95430281  4.35221131  8.38533046  2.78978899
> unique(Q2_data$X4)
 [1] 3.94200469  9.59108824  5.00387282  5.54239125 -2.02317972 -1.82642299 -1.95116399 -4.23026480 -1.34034890 -3.43377726  2.45121172
[12] 8.71645841  1.22800154 -1.57706654 -4.37836792 -3.8600176 18.91084625  2.93861754 -1.43480128 -1.08383511  5.21712609  6.02329005
[23] 1.89403190  9.71608282 -1.22915155 -2.96608026 -0.06005701 -1.26128278 12.14647386 -3.98755244  0.03563899 -1.66679683 -1.47794970
[34] -2.07453147  0.86183334  5.94139073 13.53472231 -3.41347397 13.24136740  6.59611925 19.84293750  1.75667727 -2.68743506  1.94676671
[45] 14.43959114 11.57479026 11.30695354  4.35530958  2.04824987 -1.18828117 -0.56566753  6.53800597 16.49189603  0.92621111 -3.06507244
[56] 3.80783234 -0.21941380  0.75547030  0.81004146  0.78348981  7.95162580  6.88326098  5.84998146 -0.81912383  0.65884361  1.62788827
[67] 3.78637201 13.94174809  9.01059277 -1.46908410  4.48885208  8.14742539  8.81041727 16.84781974  4.90166474 -1.32302931  2.46195446
[78] -3.78199011  3.16892498 -4.27385238  3.17729771  4.48477783  1.63526418  0.62845767 -2.54348826  7.94491942  2.15653768  7.81798432
```



Answer: X1 and X2 are discrete variables, with unique values (please see the first screenshot above). X3, X4 and X5 are continuous variables with mean, standard deviation, and range (see the second screenshot above).

(c) Construct a frequency table as below.

### Code to be entered

```
Q2_data <- read.csv("Assignment1_Q2.csv")

freq_table <- table(Q2_data$X2, Q2_data$X1)

print(freq_table)
```

### Results

	East	North	South	West
High	11	12	7	10
Low	9	14	12	5
Medium	10	4	11	15

### Full Screen

The screenshot displays the RStudio environment with the following components:

- Script Editor:** Contains the R code:

```
1 Q2_data <- read.csv("Assignment1_Q2.csv")
2
3 freq_table <- table(Q2_data$X2, Q2_data$X1)
4
5 print(freq_table)
```
- Console:** Shows the output of the code execution:

```
R 4.2.2 C:\Fichiers\HKBU 2023\Cours Semestre 6\BUSI2045 - Data Analytics for Business Decision Making\Assignment\Assignment 1/ #
0.0000 0.0000 0.0000 0.4872 1.0000 1.0000
> sd(Q2_data$X3)
[1] 7.654046
> range(Q2_data$X3)
[1] -15.76558 23.39193
> sd(Q2_data$X4)
[1] 5.651817
> range(Q2_data$X4)
[1] -4.378368 19.842937
> sd(Q2_data$X5)
[1] 0.5019854
> range(Q2_data$X5)
[1] 0 1
> Q2_data <- read.csv("Assignment1_Q2.csv")
> freq_table <- table(Q2_data$X2, Q2_data$X1)
> print(freq_table)

      East North South West
High    11    12     7    10
Low     9    14    12     5
Medium 10     4    11    15
```
- Environment:** Lists the objects in the global environment:
  - Q2\_data: 120 obs. of 5 variables
  - freq\_table: 'table' int [1:3, 1:4] 11...
  - n\_obs: 117L
- Files:** Shows the project files, including the data file 'Assignment1\_Q2.csv' (6.4 KB).

**(d)** Is variable X4 normally distributed? Use ggplot2 to create a QQ plot to help answer this question.

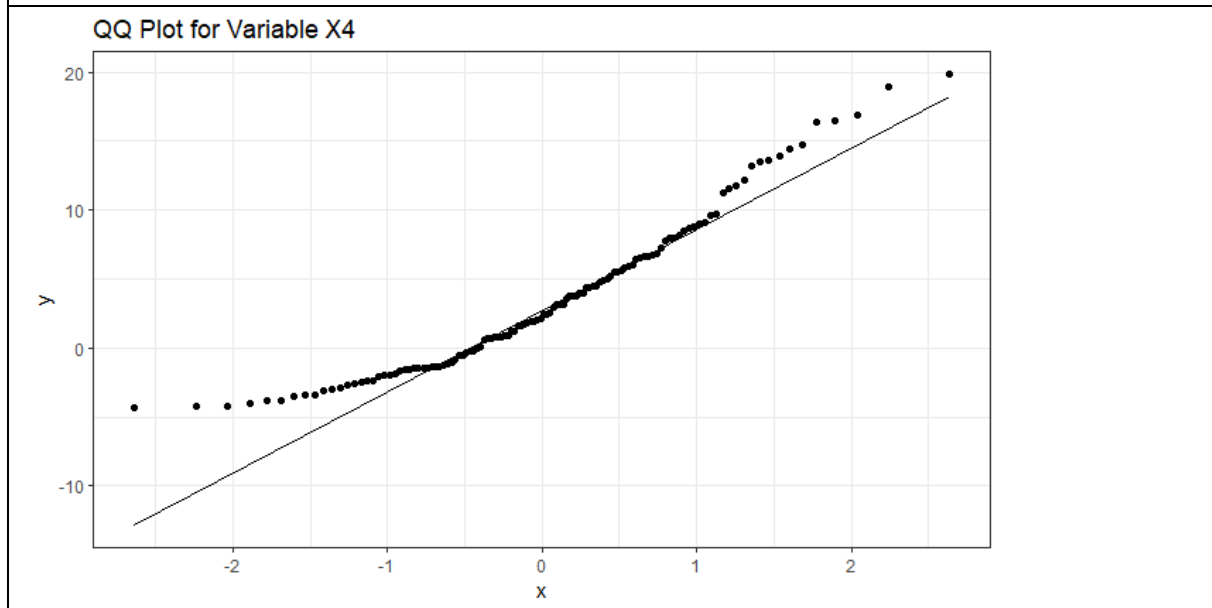
#### Code to be entered

```
library(ggplot2)

Q2_data_subset <- Q2_data[!is.na(Q2_data$X4),]

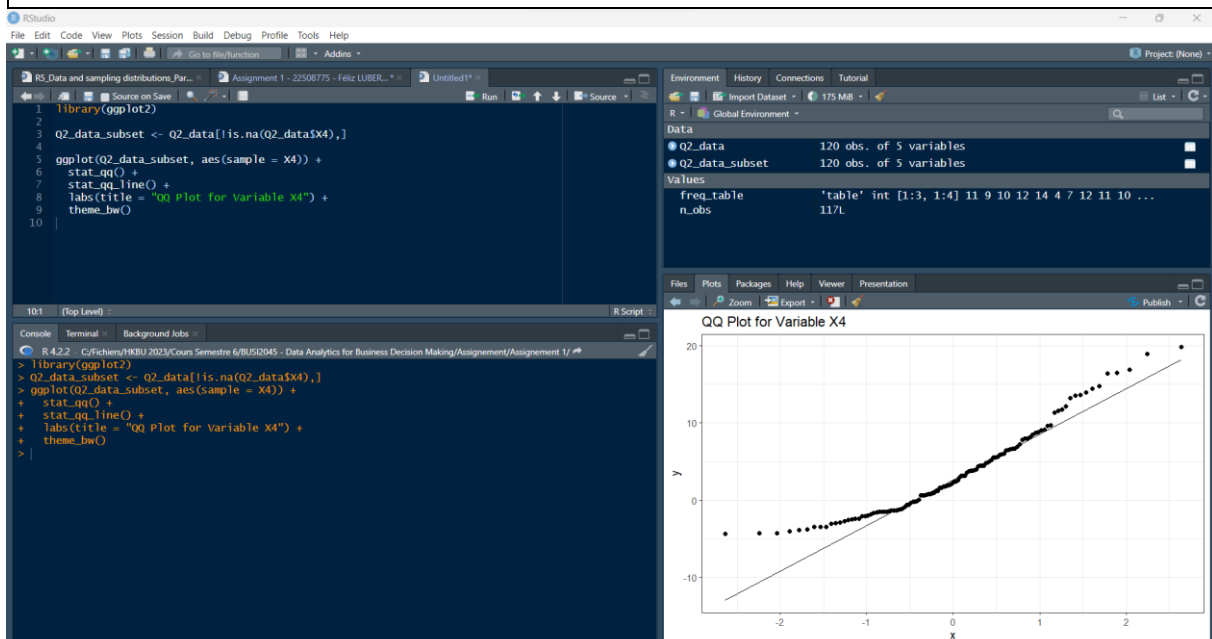
ggplot(Q2_data_subset, aes(sample = X4)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "QQ Plot for Variable X4") +
  theme_bw()
```

#### Results



Answer: As we can see on the table QQ Plot above, the variable X4 is normally distributed.

#### Full Screen





**(e)** Recreate the following boxplot for variable X3 across the different levels of X2. The result should look like the below.

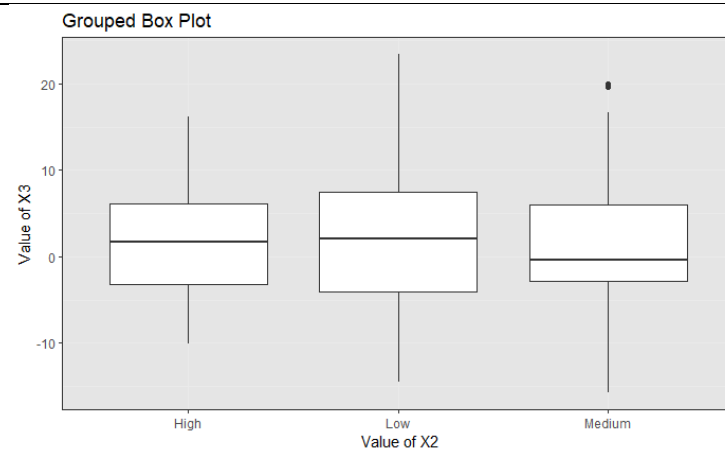
#### Code to be entered

```
library(ggplot2)

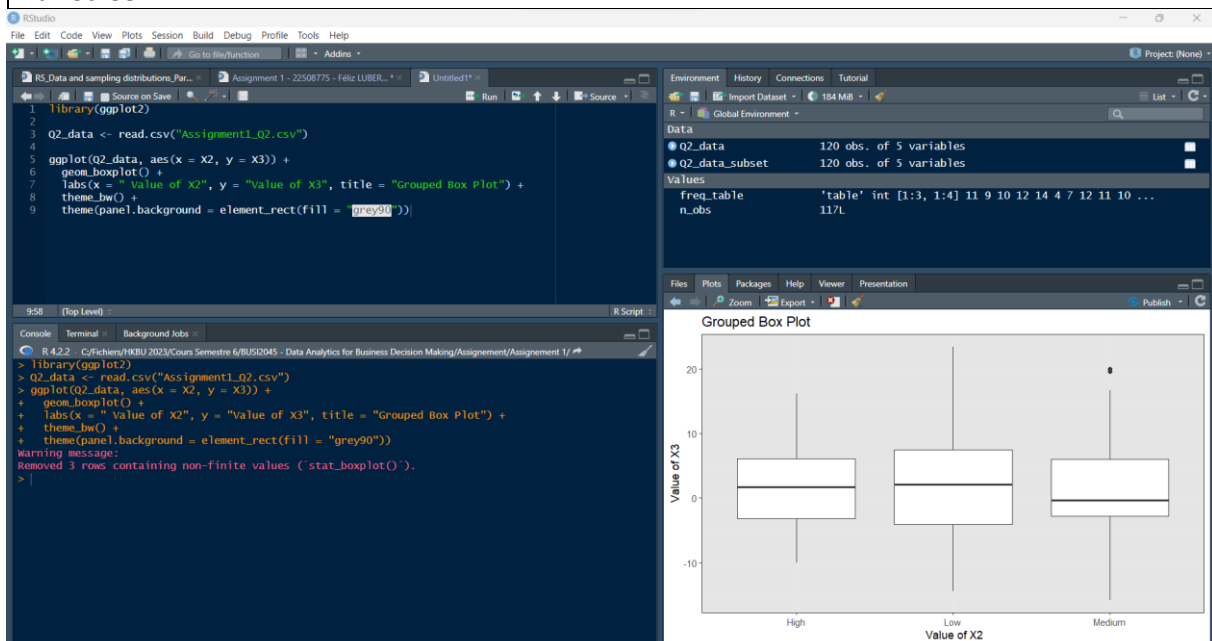
Q2_data <- read.csv("Assignment1_Q2.csv")

ggplot(Q2_data, aes(x = X2, y = X3)) +
  geom_boxplot() +
  labs(x = " Value of X2", y = "Value of X3", title = "Grouped Box Plot") +
  theme_bw() +
  theme(panel.background = element_rect(fill = "grey90"))
```

#### Results



#### Full Screen



(f) Create a new variable X6 which is the sum of X3 and X4. Visualize the distribution of X6 as below.

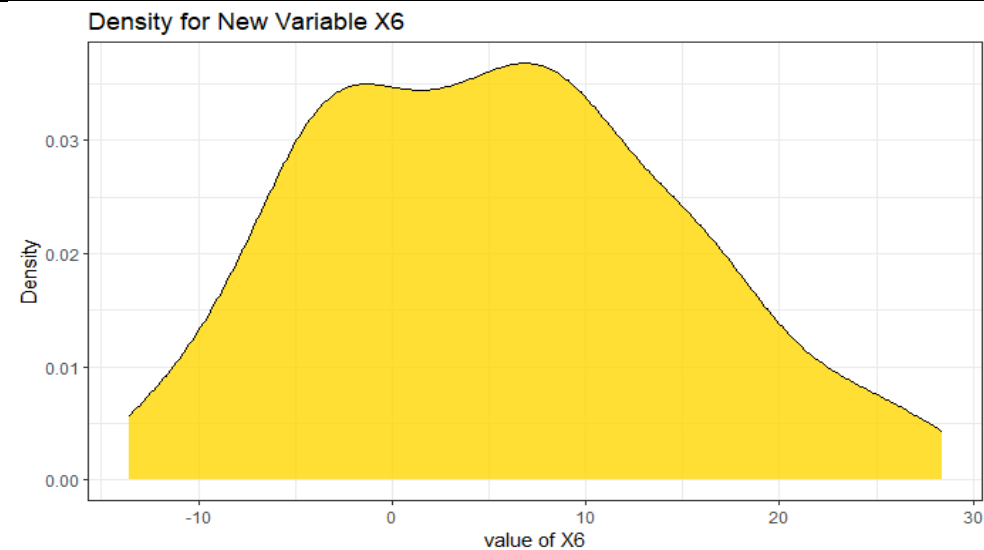
#### Code to be entered

```
Q2_data <- read.csv("Assignment1_Q2.csv")

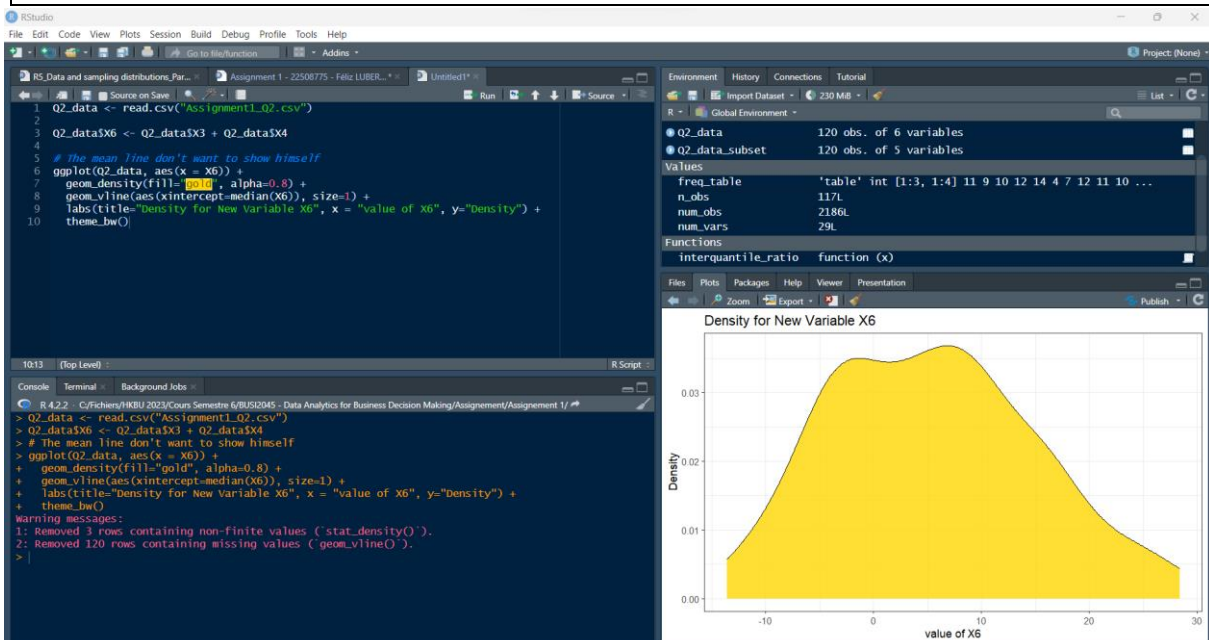
Q2_data$X6 <- Q2_data$X3 + Q2_data$X4

# The mean line don't want to show himself
ggplot(Q2_data, aes(x = X6)) +
  geom_density(fill="gold", alpha=0.8) +
  geom_vline(aes(xintercept=median(X6)), size=1) +
  labs(title="Density for New Variable X6", x = "value of X6",
y="Density") +
  theme_bw()
```

#### Results



#### Full Screen



### Question 3: Describe Data (40 Points)

- (a) Read the file `marketing_campaign.csv` in R and construct a subset named `df2_sub` where the variable `Income` contains no missing value, and variables `NumStorePurchases` and `NumWebPurchases` are not equal to 0. How many observations and variables are in this subset?

#### Code to be entered

```
df2 <- read.csv("marketing_campaign.csv", header = TRUE, na.strings =
c("Unknown", " ", ""))

df2_sub <- df2[complete.cases(df2$Income) & df2$NumStorePurchases != 0
& df2$NumWebPurchases != 0,]

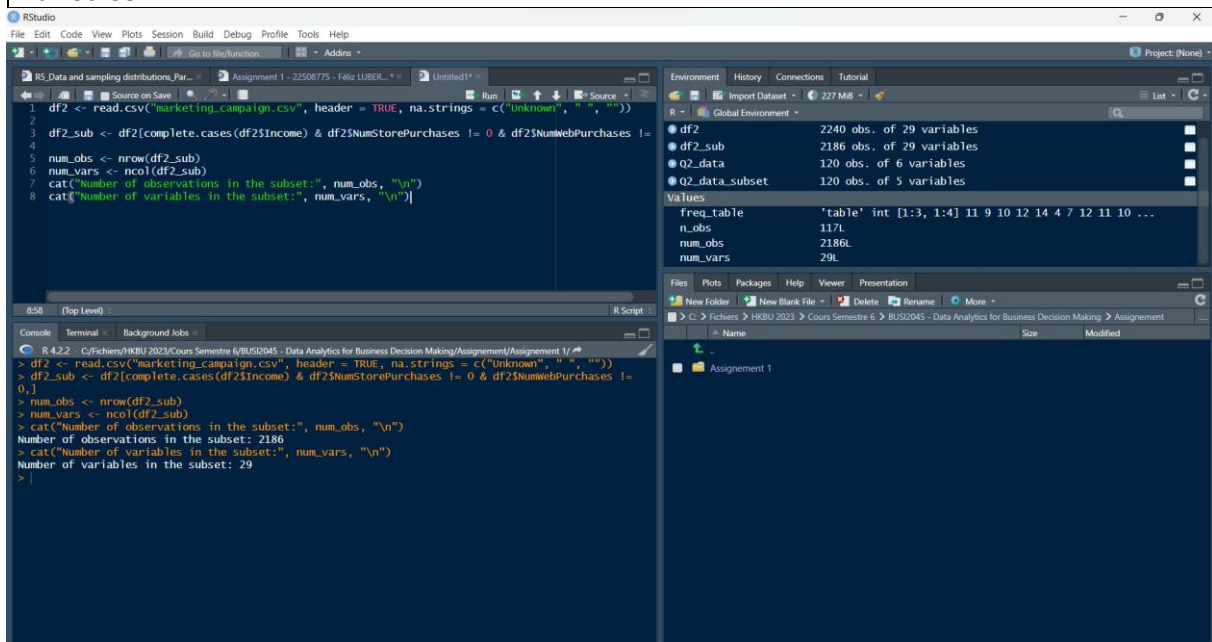
num_obs <- nrow(df2_sub)
num_vars <- ncol(df2_sub)
cat("Number of observations in the subset:", num_obs, "\n")
cat("Number of variables in the subset:", num_vars, "\n")
```

#### Results

```
> cat("Number of observations in the subset:", num_obs, "\n")
Number of observations in the subset: 2186
> cat("Number of variables in the subset:", num_vars, "\n")
Number of variables in the subset: 29
```

Answer: There is 29 variables in the data set and 2186 observations.

#### Full Screen



**(b)** What are the values of 10%, 50%, 80% percentile for variable Income?

#### Code to be entered

```
df2 <- read.csv("marketing_campaign.csv")
df2_sub <- df2[complete.cases(df2$Income) & df2$NumStorePurchases != 0
& df2$NumWebPurchases != 0,]

quantile(df2_sub$Income, probs = c(0.1, 0.5, 0.8))
```

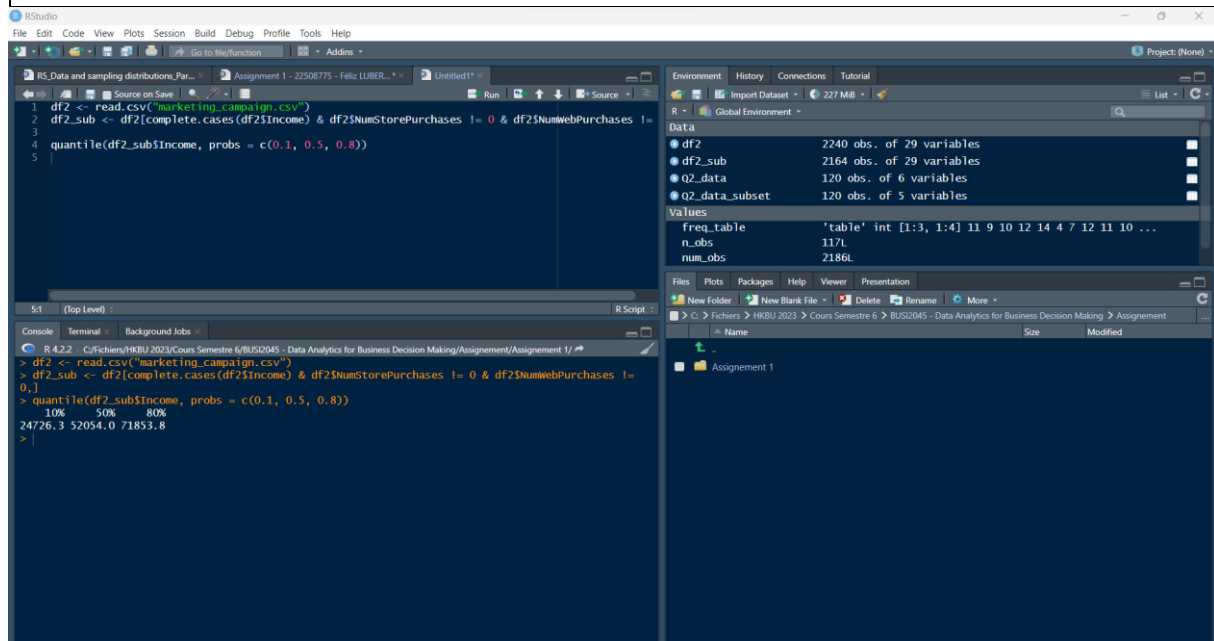
#### Results

```
> quantile(df2_sub$Income, probs = c(0.1, 0.5, 0.8))
      10%      50%      80%
24726.3 52054.0 71853.8
```

Answer: The values for:

- 10%: 24726.3
- 50%: 52054.0
- 80%: 71853.8

#### Full Screen



(c) What Write a named function to compute the ratio of the interquantile value against the range of a variable. Apply that function to three variables in the dataset.

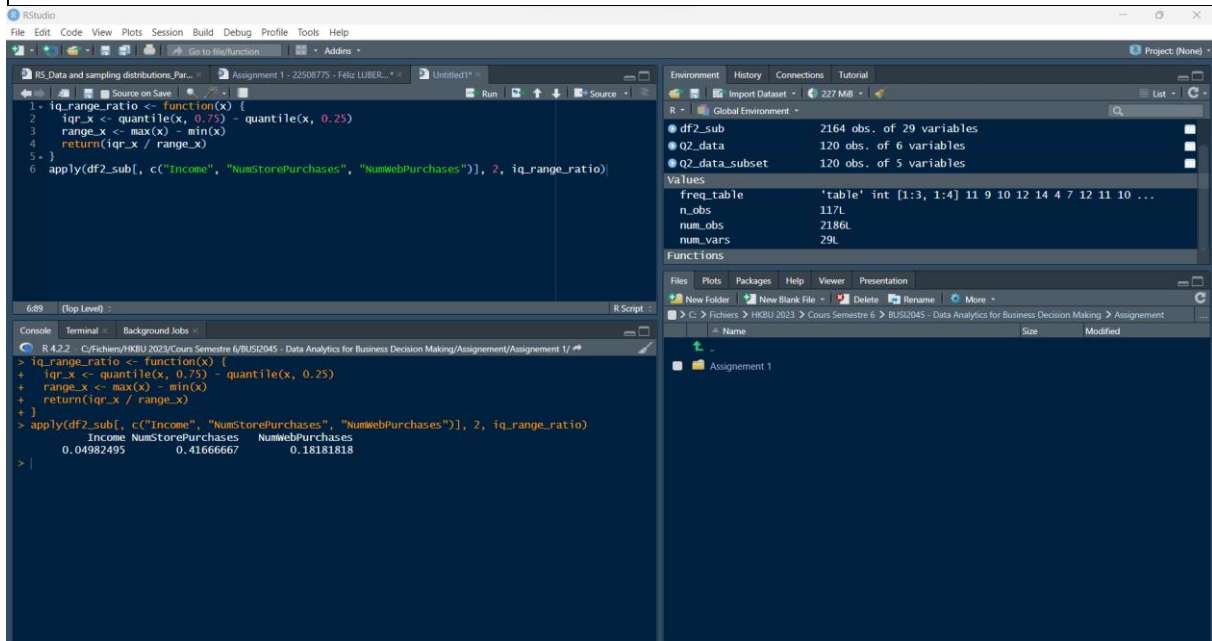
Code to be entered

```
iq_range_ratio <- function(x) {  
  iqr_x <- quantile(x, 0.75) - quantile(x, 0.25)  
  range_x <- max(x) - min(x)  
  return(iqr_x / range_x)  
}  
apply(df2_sub[, c("Income", "NumStorePurchases", "NumWebPurchases")],  
2, iq_range_ratio)
```

Results

```
> apply(df2_sub[, c("Income", "NumStorePurchases",  
"NumWebPurchases")], 2, iq_range_ratio)  
      Income NumStorePurchases  NumWebPurchases  
0.04982495      0.41666667      0.18181818
```

Full Screen



(d) Write an anonymous function to solve the above question.

#### Code to be entered

```
interquantile_ratio <- function(x) {  
  IQR(x, na.rm = TRUE) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))  
}  
  
apply(df2_sub[, c("Income", "NumStorePurchases", "NumWebPurchases")],  
2, interquantile_ratio)
```

#### Results

```
> apply(df2_sub[, c("Income", "NumStorePurchases",  
"NumWebPurchases")], 2, interquantile_ratio)  
      Income NumStorePurchases NumWebPurchases  
0.04982495      0.41666667      0.18181818
```

#### Full Screen

