Assignment 1

Féliz LUBERNE- 22508775

Table des matières

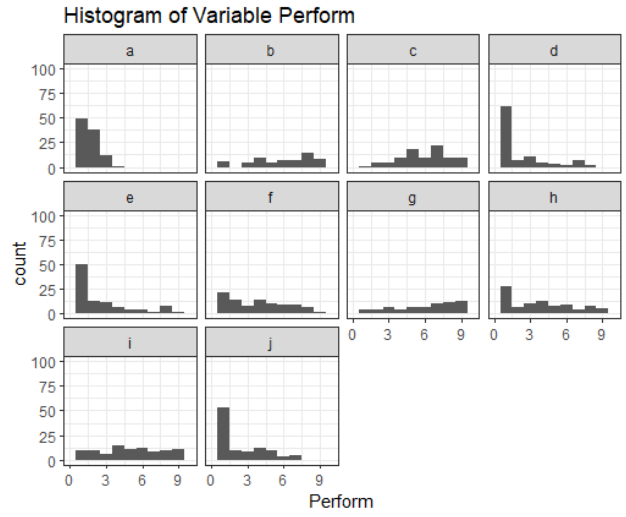
Question 1: Data Exploration and Visualization (20 points)

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

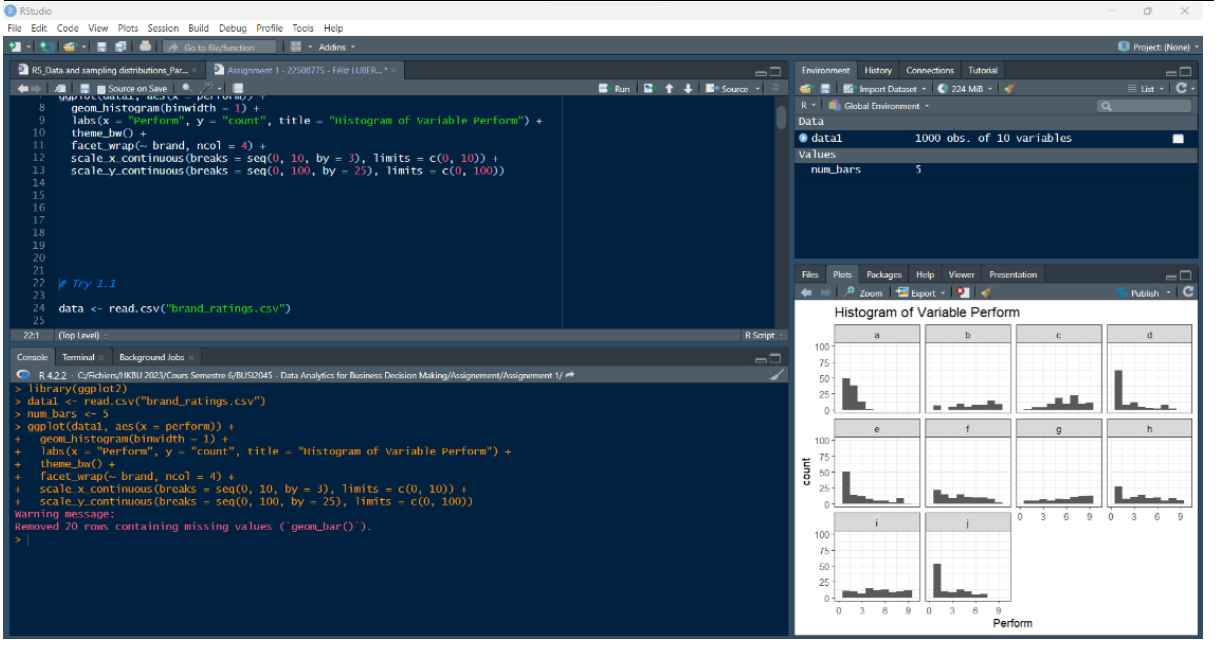
Code to be entered

```
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

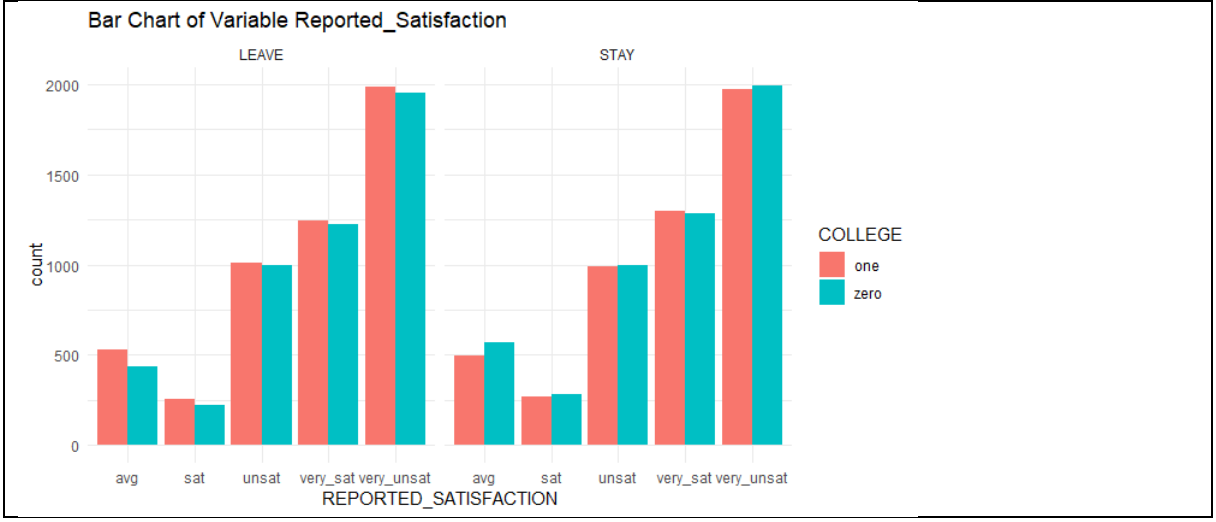


(b) Load the dataset churn.arff into R. Create a bar plot using the variable REPORTED_SATISFACTION. Your output should look similar as the below graph.

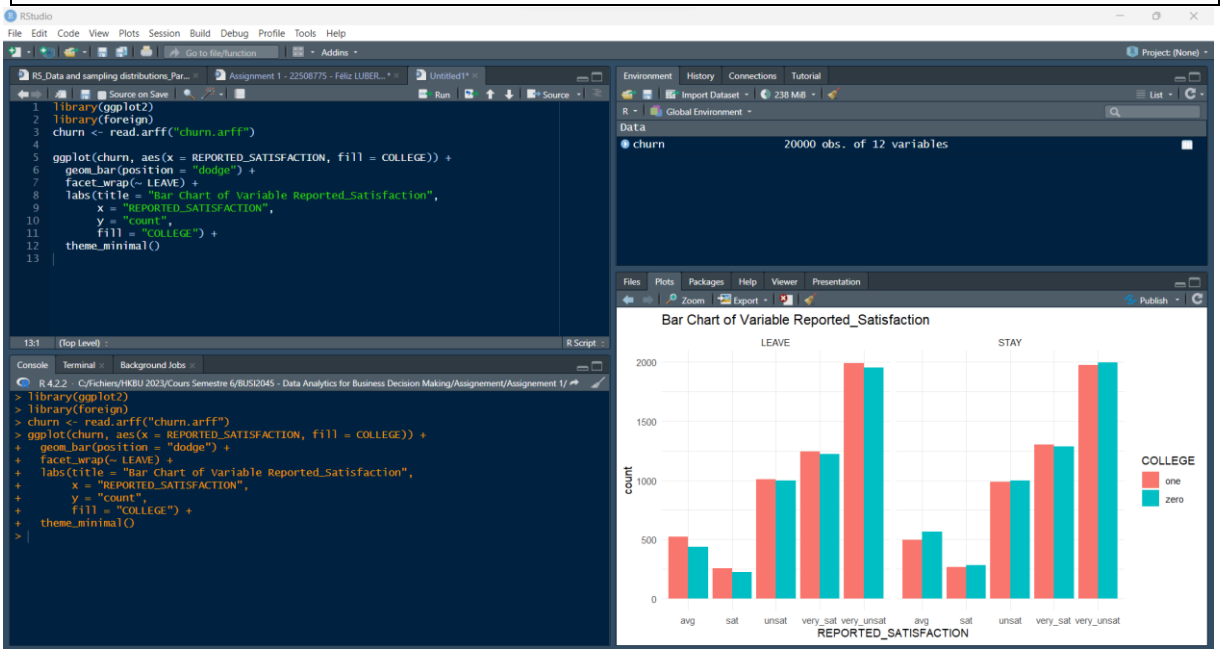
Code to be entered

```
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)
```

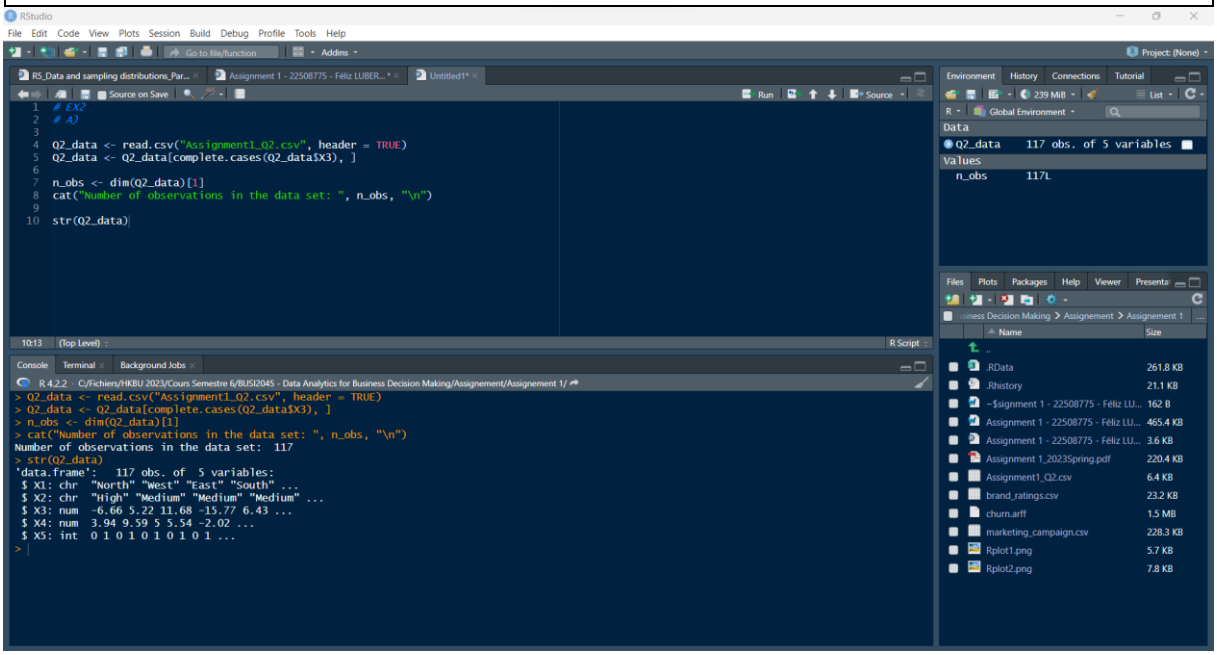
Results

```
> 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

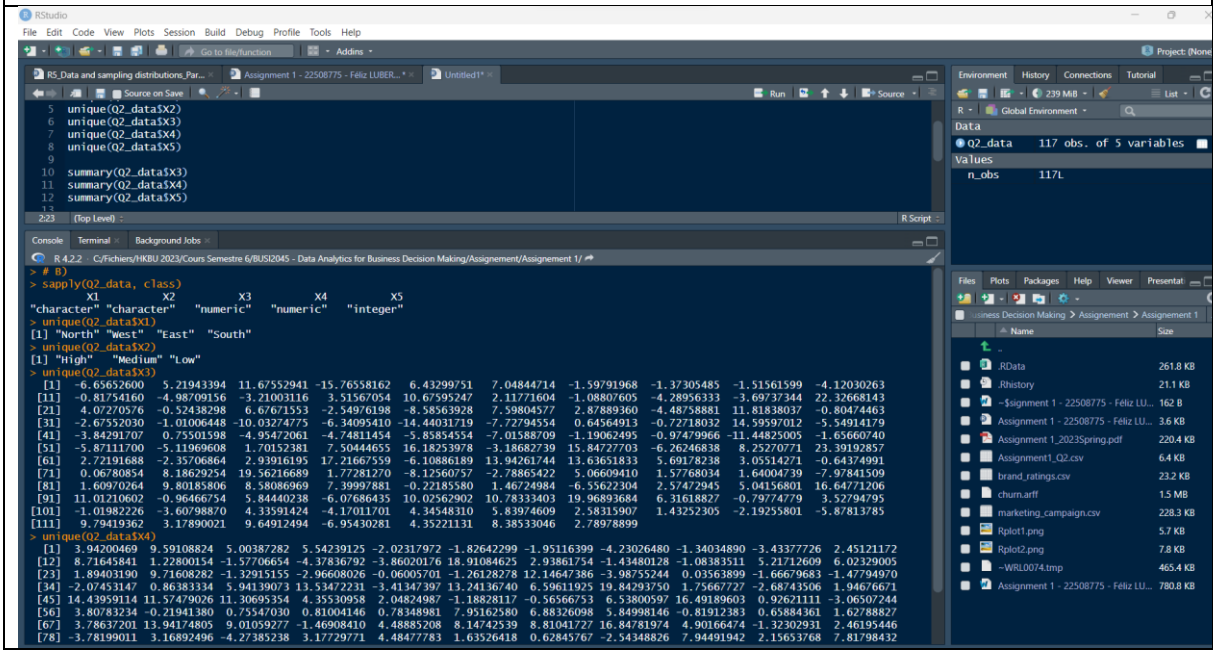


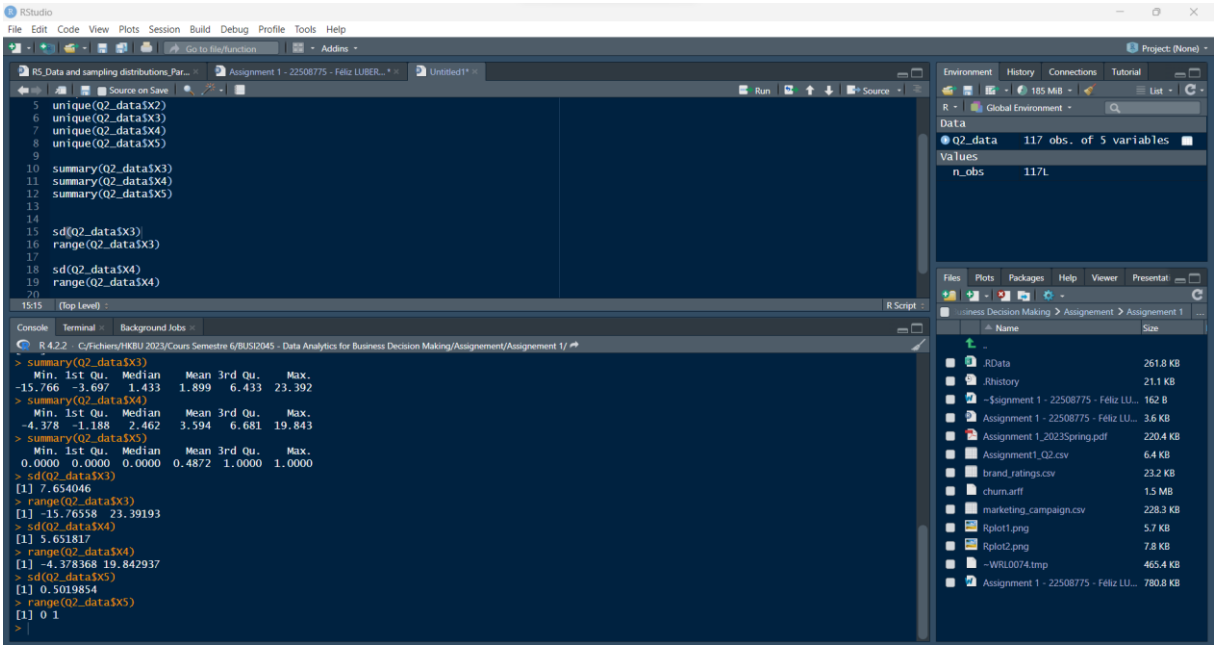
(b) Which variable(s) belong to the discrete variable? Check the unique values for these discrete variables. Which variable(s) belong to the continuous variable? Check the values of mean, standard deviation, and range for these continuous variables.

Code to be entered

```
sapply(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





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.

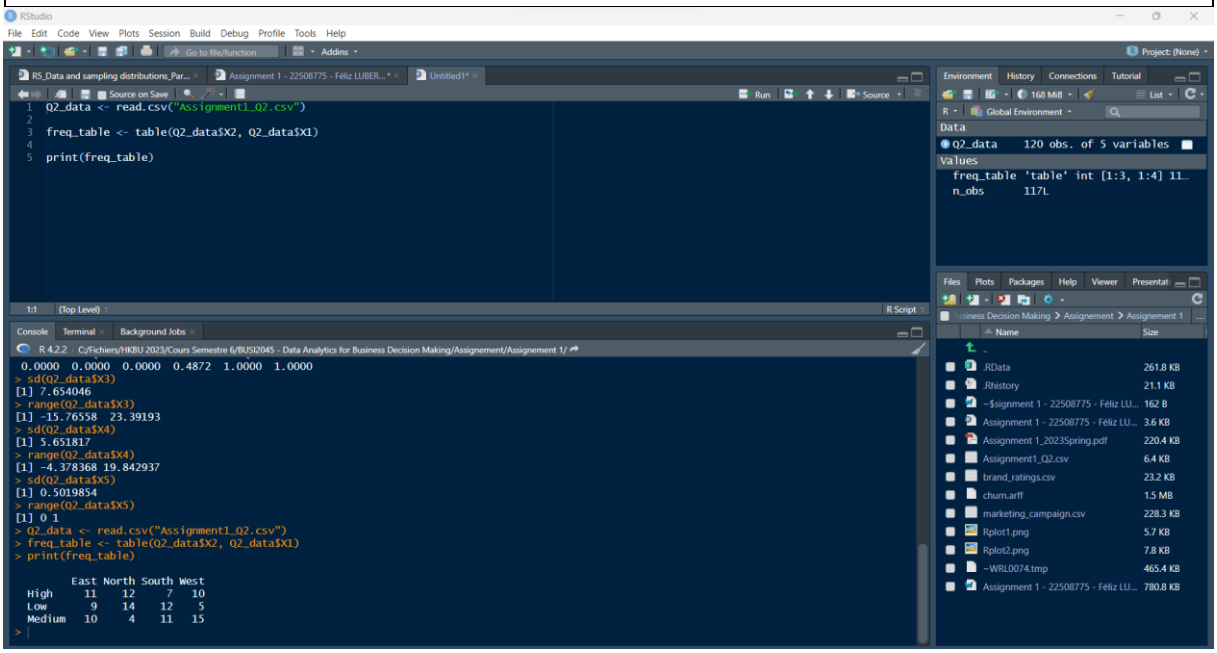
Code to be entered

`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

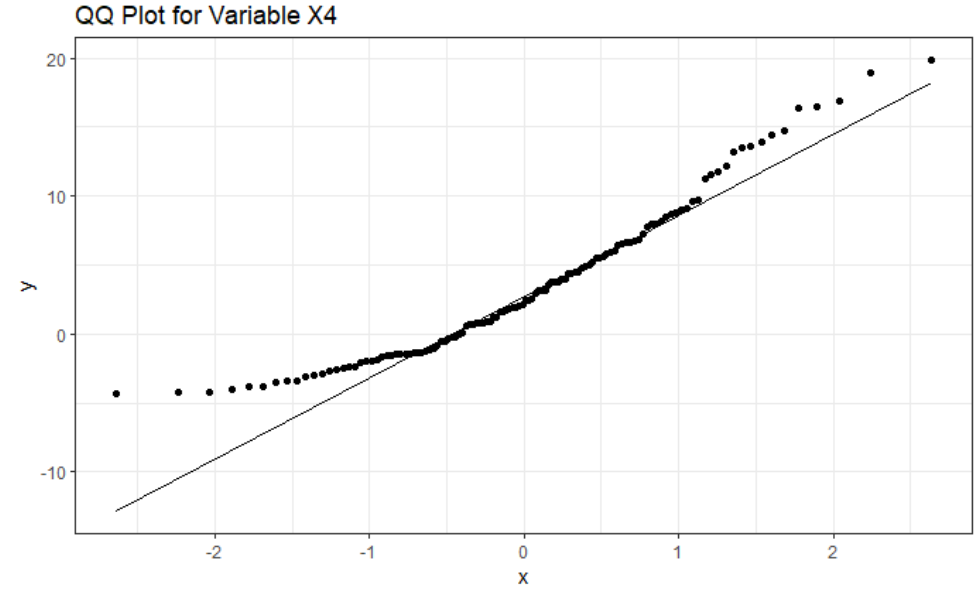


(d) Is variable X4 normally distributed? Use ggplot2 to create a QQ plot to help answer this question.

Code to be entered

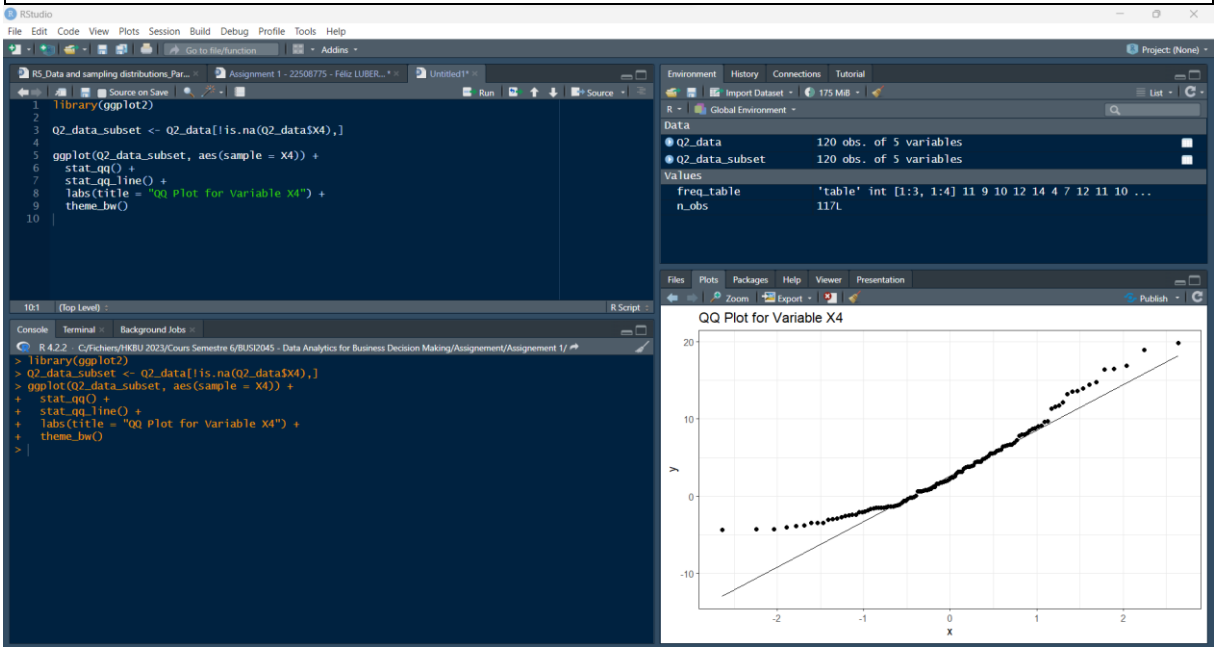
```
library(ggplot2)

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

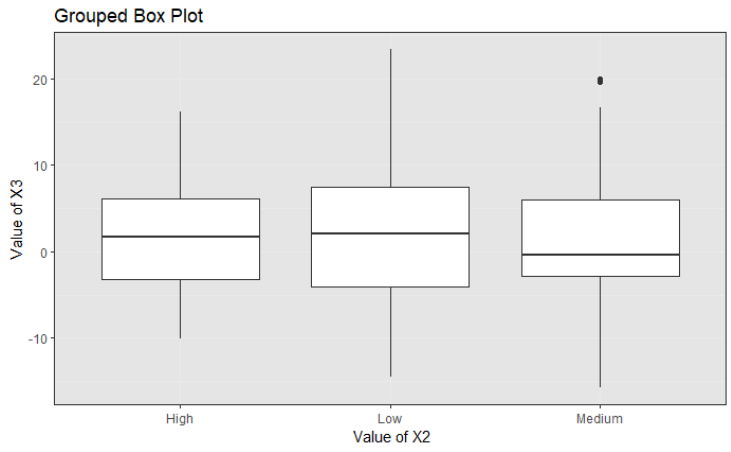


(e) Recreate the following boxplot for variable X3 across the different levels of X2. The result should look like the below.

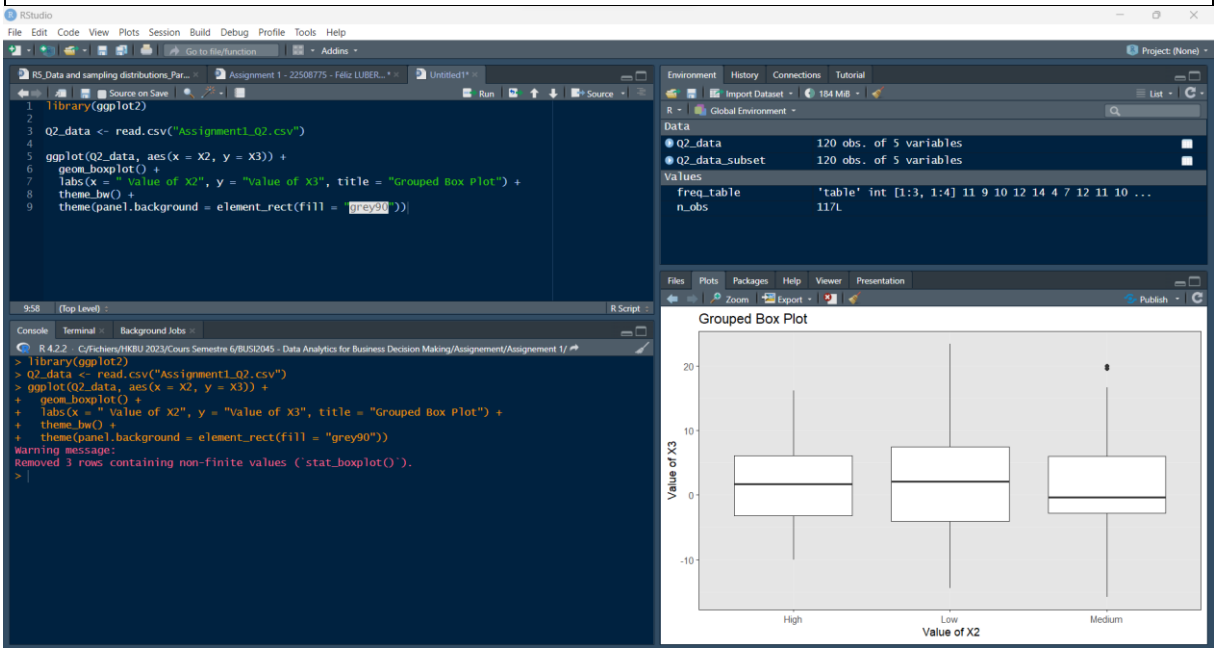
Code to be entered

```
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

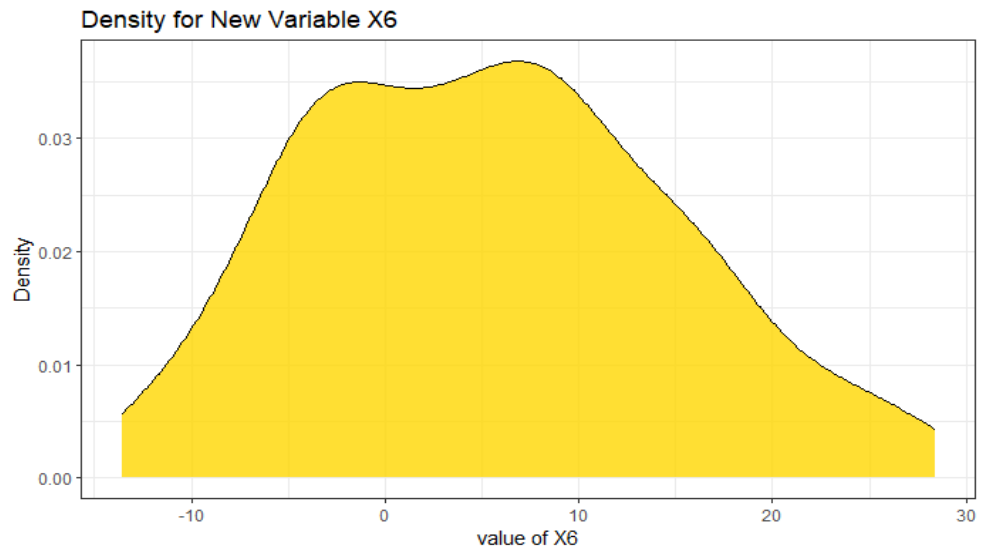


(f) Create a new variable X6 which is the sum of X3 and X4. Visualize the distribution of X6 as below.

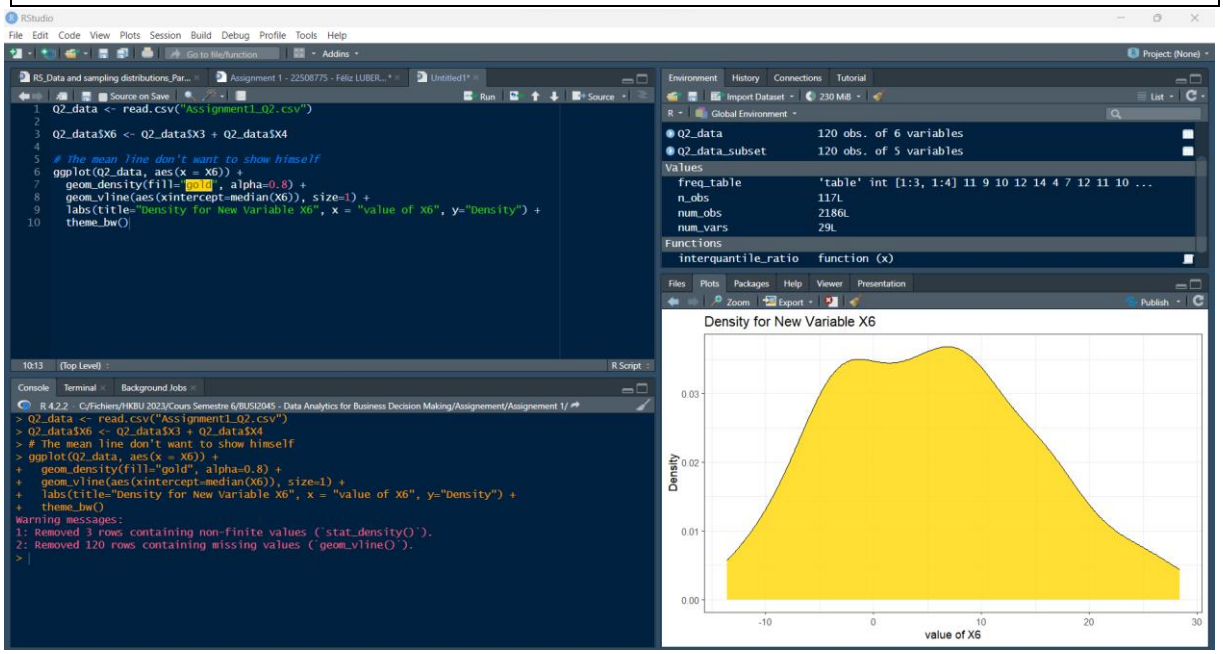
Code to be entered

```
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

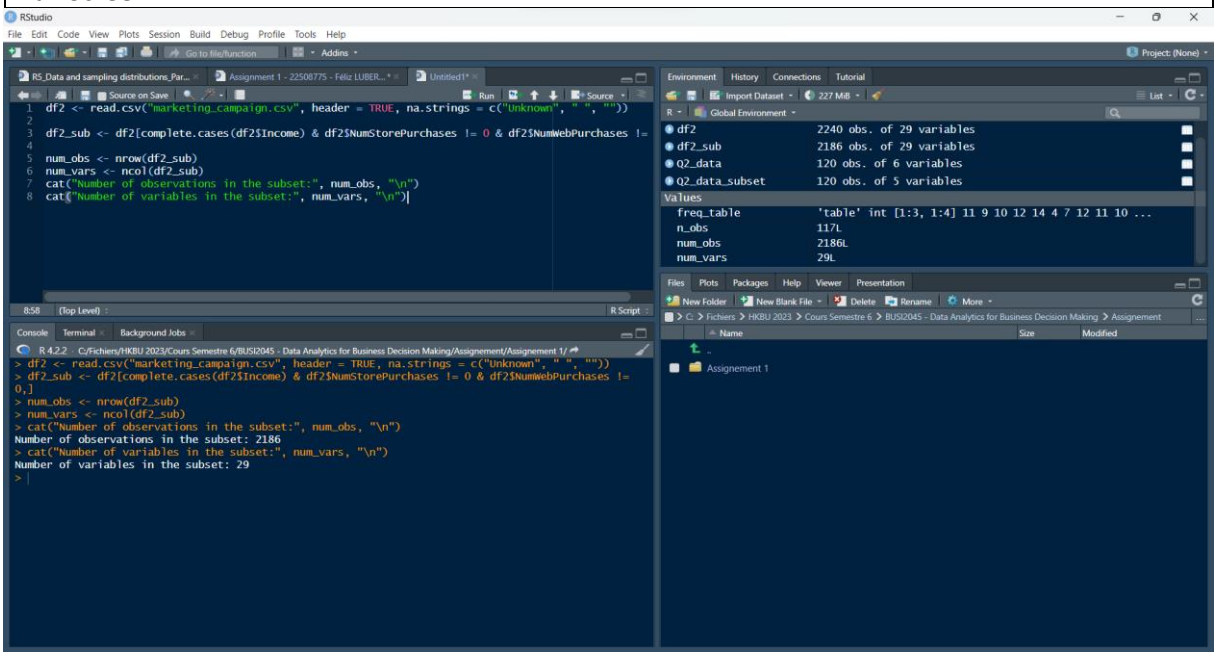
```
df2 <- read.csv("marketing_campaign.csv", header = TRUE, na.strings = c("Unknown", " ", ""))

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))`

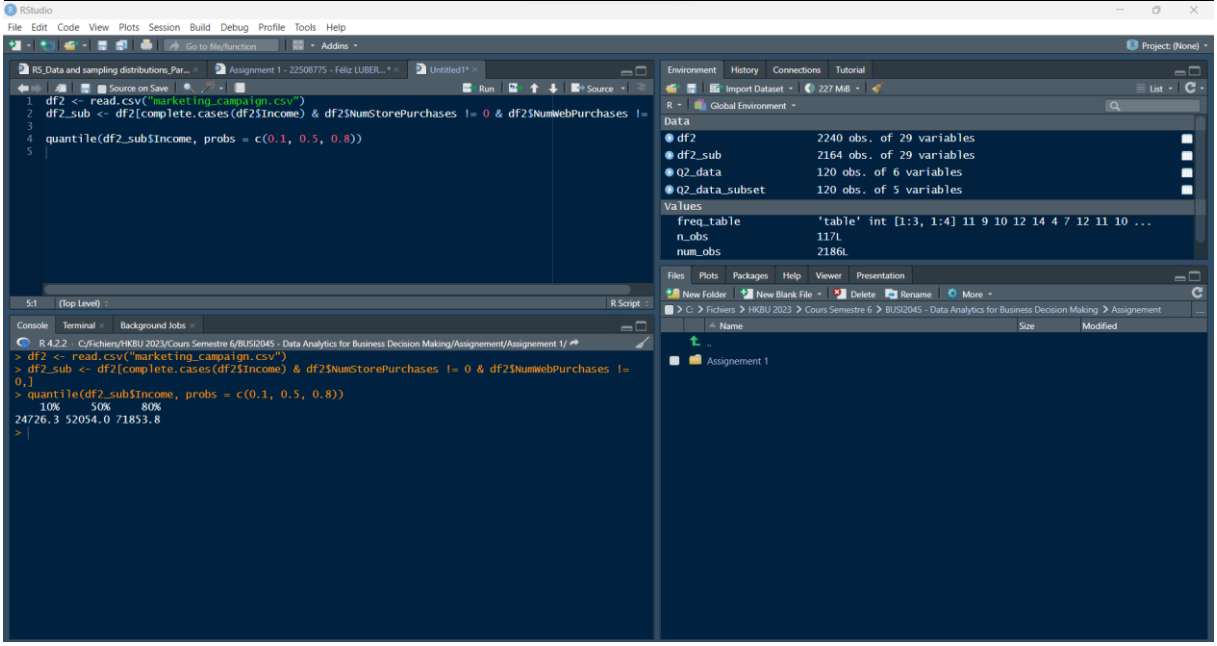
Results

`> 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.

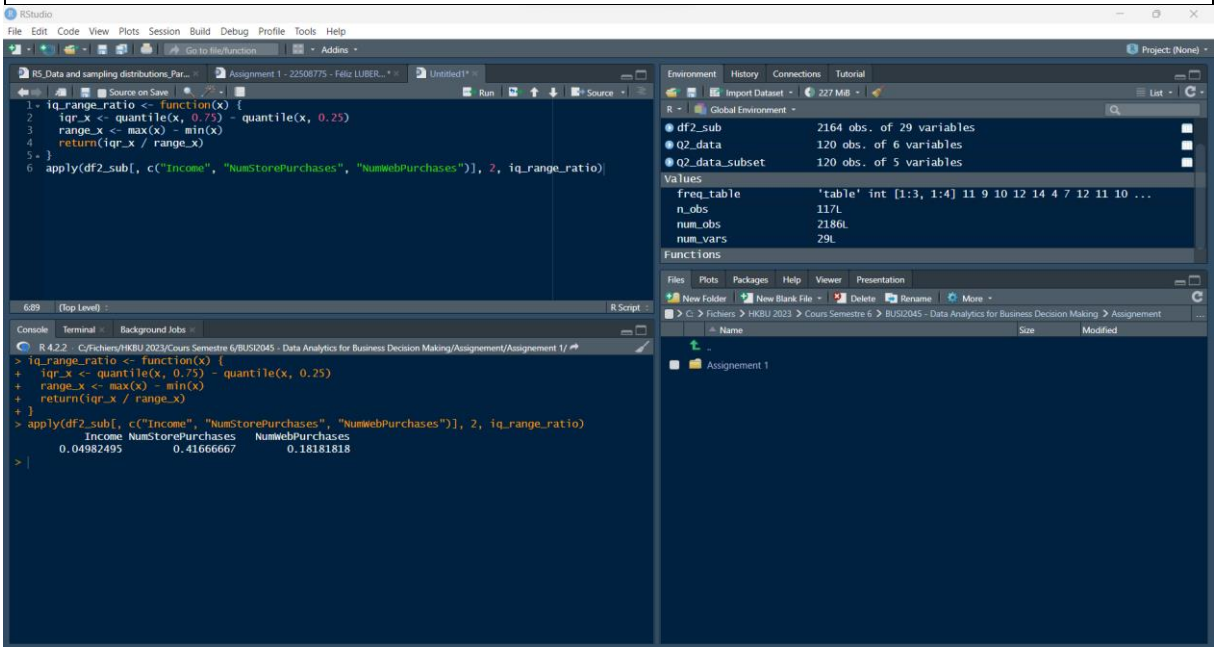
Code to be entered

```
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