**Assignment 1**

**Féliz LUBERNE - 22508775**

Table des matières

[**Question 1**: Data Exploration and Visualization (20 points) 2](#_Toc127879874)

[**(a)** Read the dataset brand\_ratings.csv into R. Construct a histogram plot (as below) using variable perform. 2](#_Toc127879875)

[**(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](#_Toc127879876)

[**Question 2** Describe Data (40 Points) 4](#_Toc127879877)

[**(a)** How many observations in this data set? What are the types (numeric, integer, etc.) of these variables? 4](#_Toc127879878)

[**(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](#_Toc127879879)

[**(c)** Construct a frequency table as below. 7](#_Toc127879880)

[**(d)** Is variable X4 normally distributed? Use ggplot2 to create a QQ plot to help answer this question. 8](#_Toc127879881)

[**(e)** Recreate the following boxplot for variable X3 across the different levels of X2. The result should look like the below. 9](#_Toc127879882)

[**(f)** Create a new variable X6 which is the sum of X3 and X4. Visualize the distribution of X6 as below. 10](#_Toc127879883)

[**Question 3:** Describe Data (40 Points) 11](#_Toc127879884)

[**(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? 11](#_Toc127879885)

[**(b)** What are the values of 10%, 50%, 80% percentile for variable Income? 12](#_Toc127879886)

[**(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. 13](#_Toc127879887)

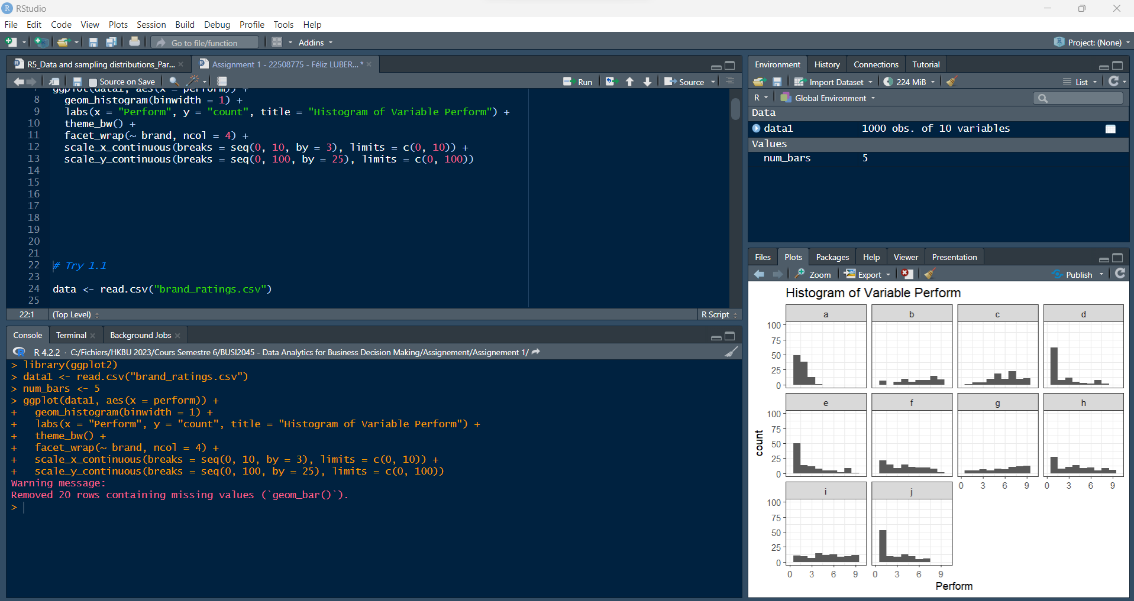
[**(d)** Write an anonymous function to solve the above question. 14](#_Toc127879888)

# **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 |
|  |

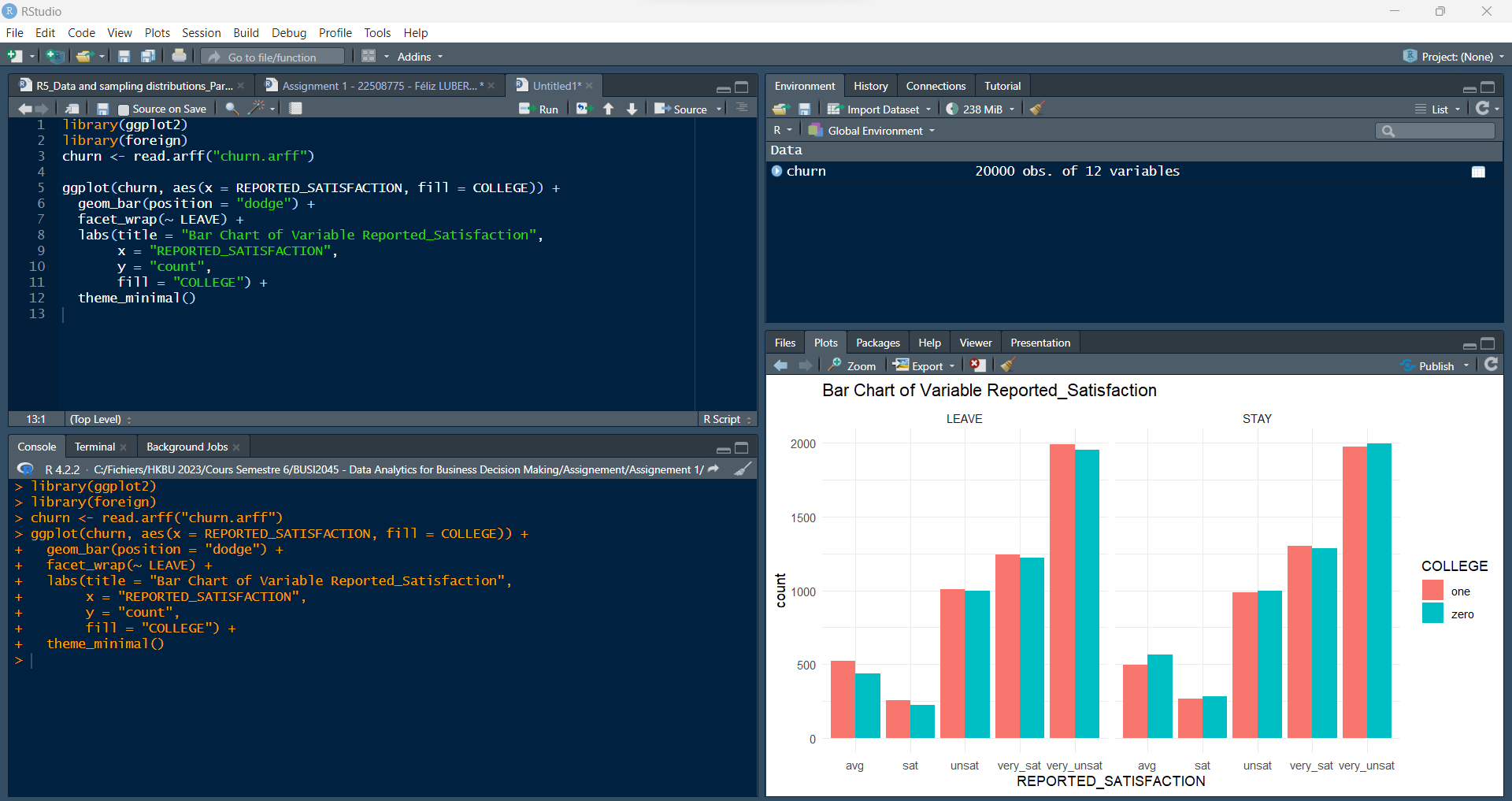
|  |
| --- |
| Full Screen |



## **(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)

## 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.

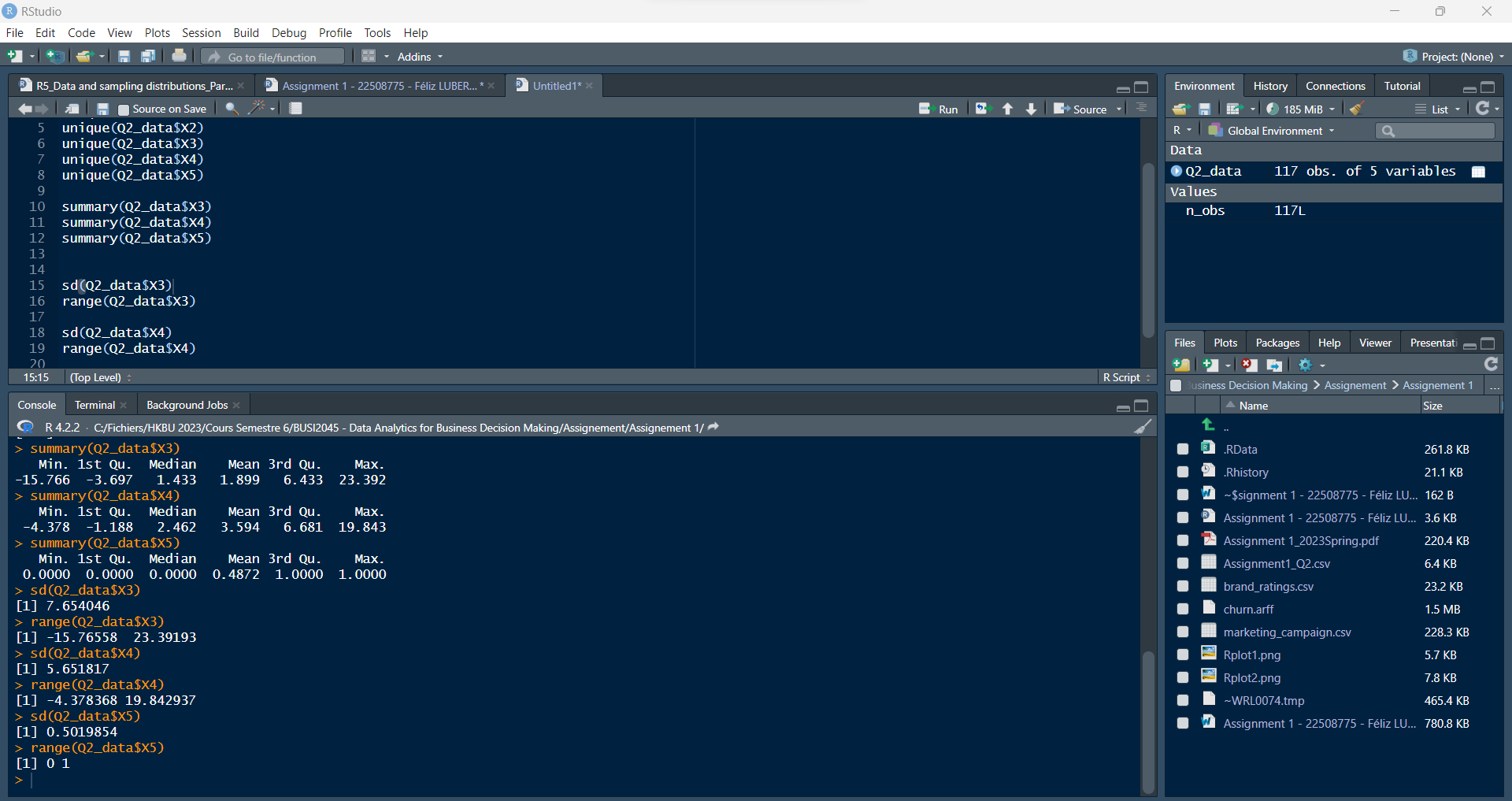
|  |
| --- |
| Full Screen |

Une image contenant texte, capture d’écran, ordinateur, moniteur

Description générée automatiquement

## 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 |
|  |

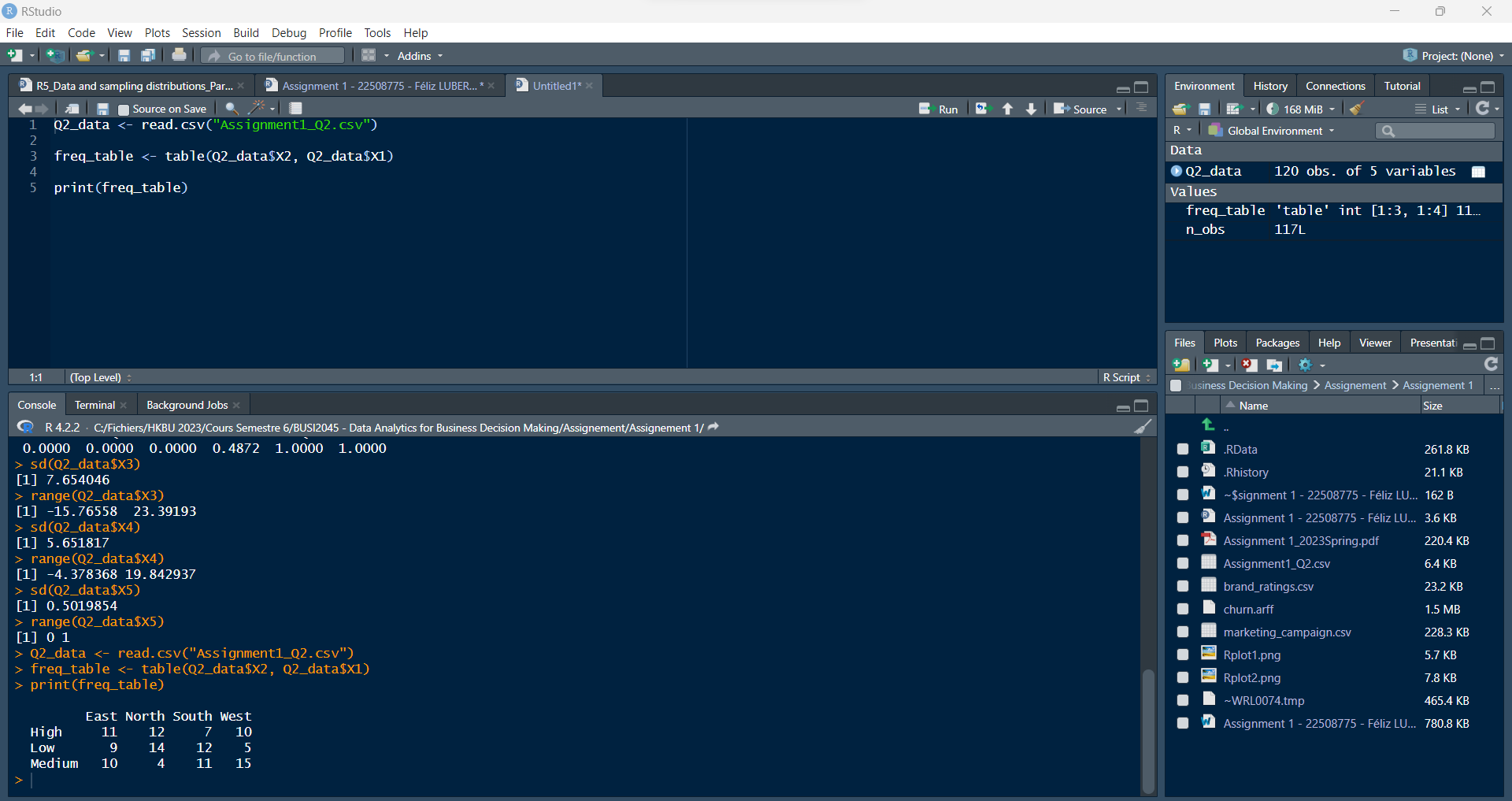


Answer: X1 and X2 are discrete variables, with unique values (please see the first screenshot above). X3, X4 and X5 are continuous variables with mean, standard deviation, and range (see the second screenshot above).

## 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 |

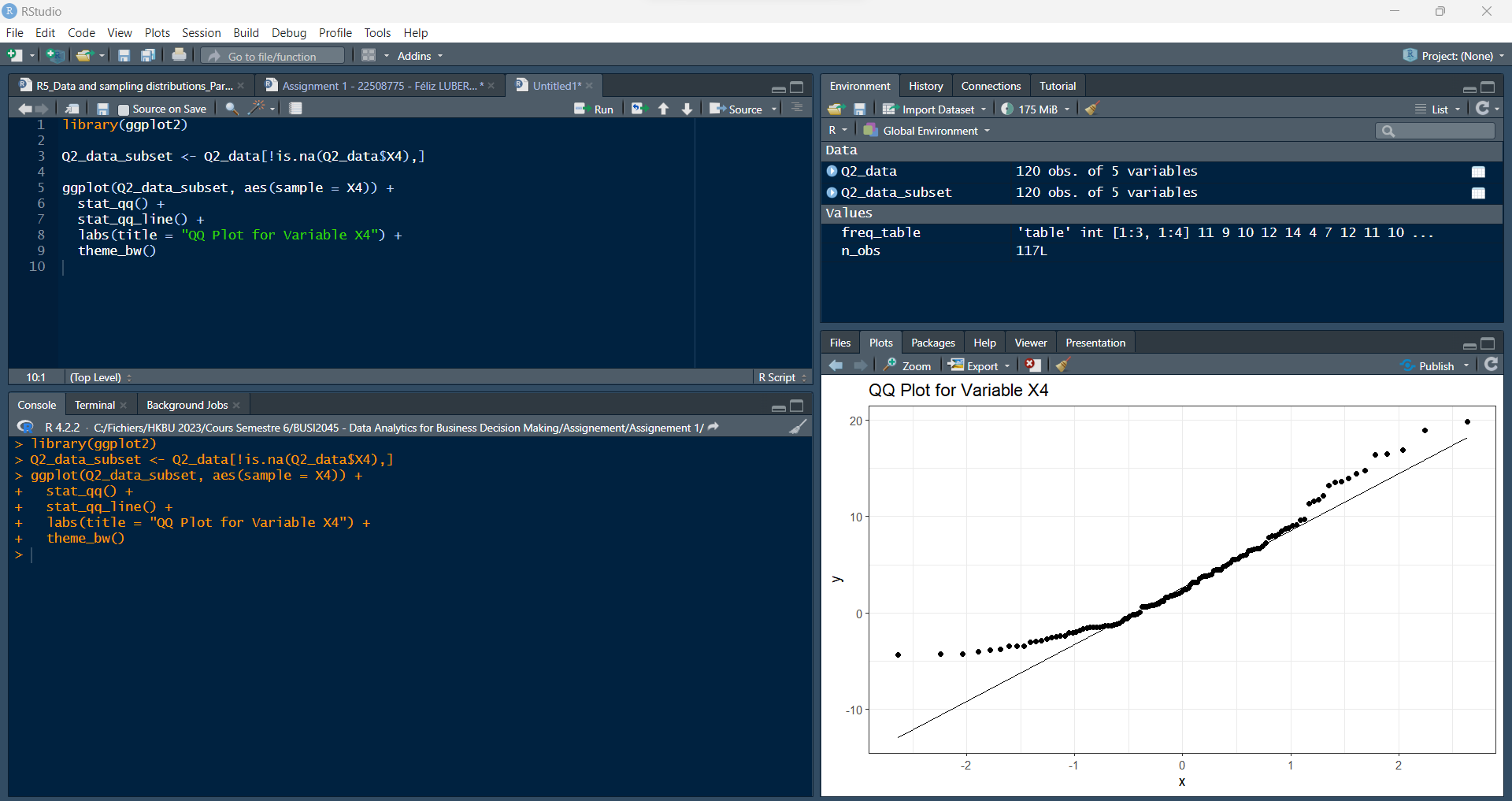


## Is variable X4 normally distributed? Use ggplot2 to create a QQ plot to help answer this question.

|  |
| --- |
| 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.

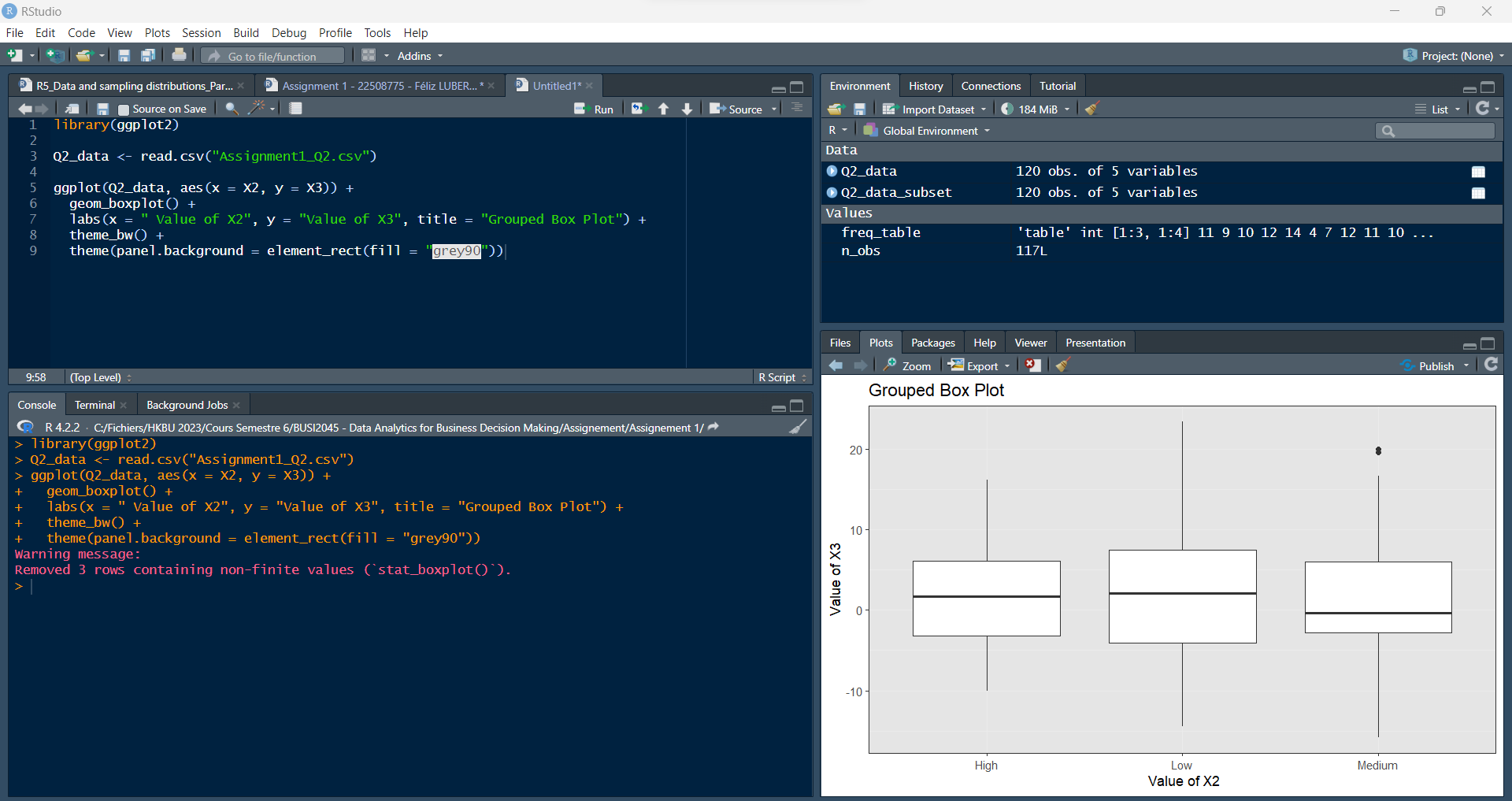
|  |
| --- |
| Full Screen |



## 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 |
|  |

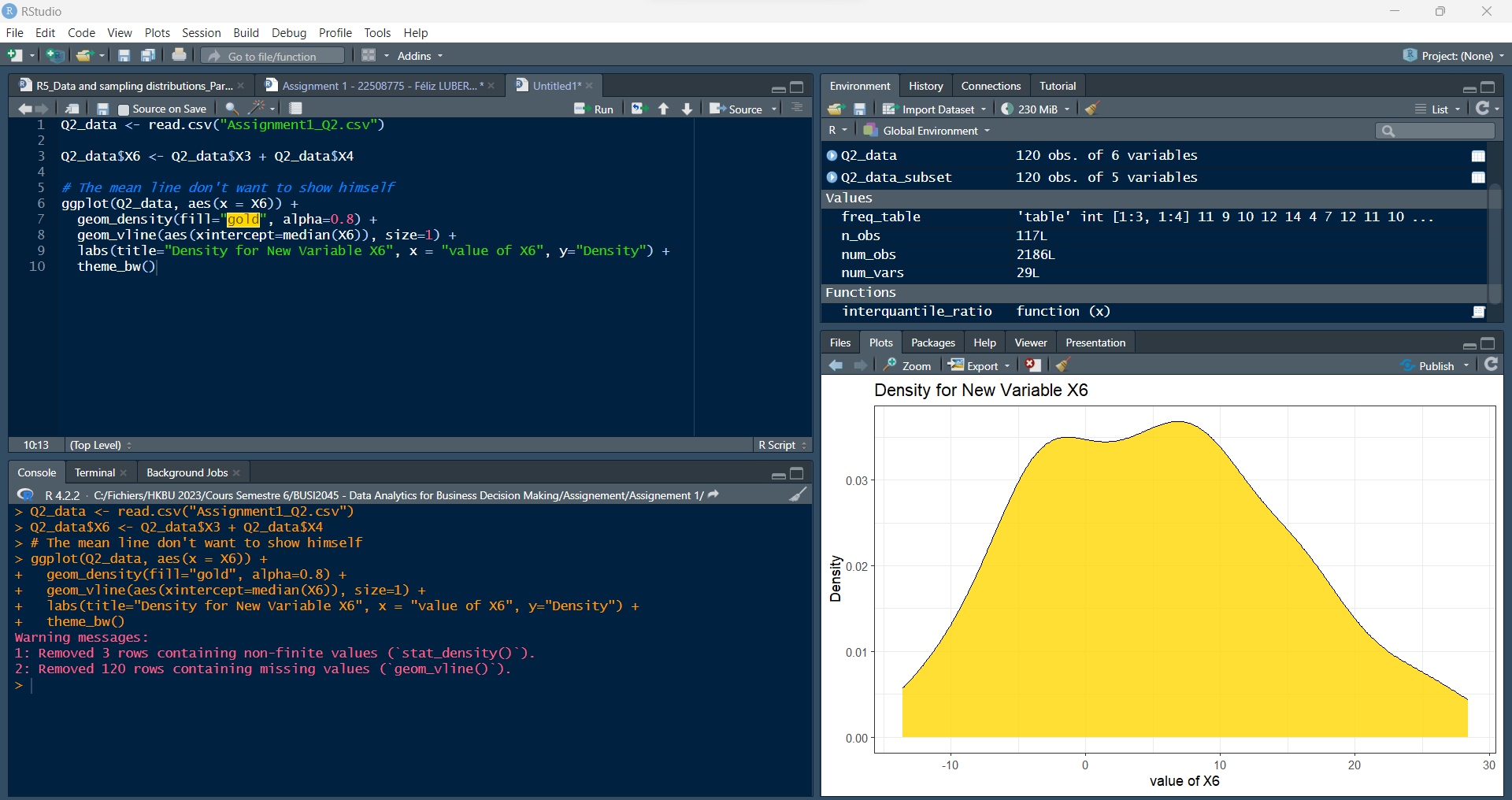
|  |
| --- |
| Full Screen |



## 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 |



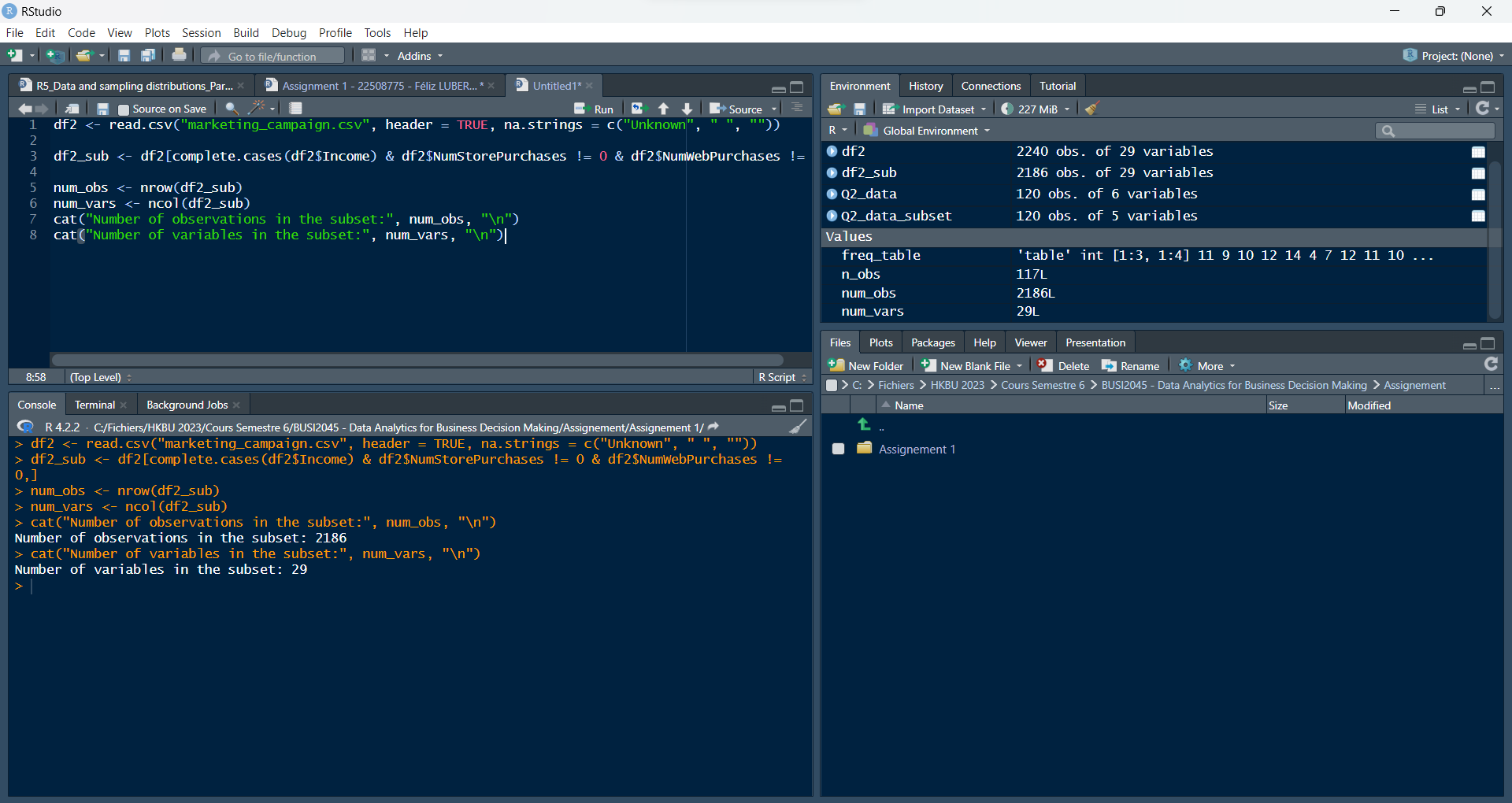
# **Question 3:** Describe Data (40 Points)

## 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 |



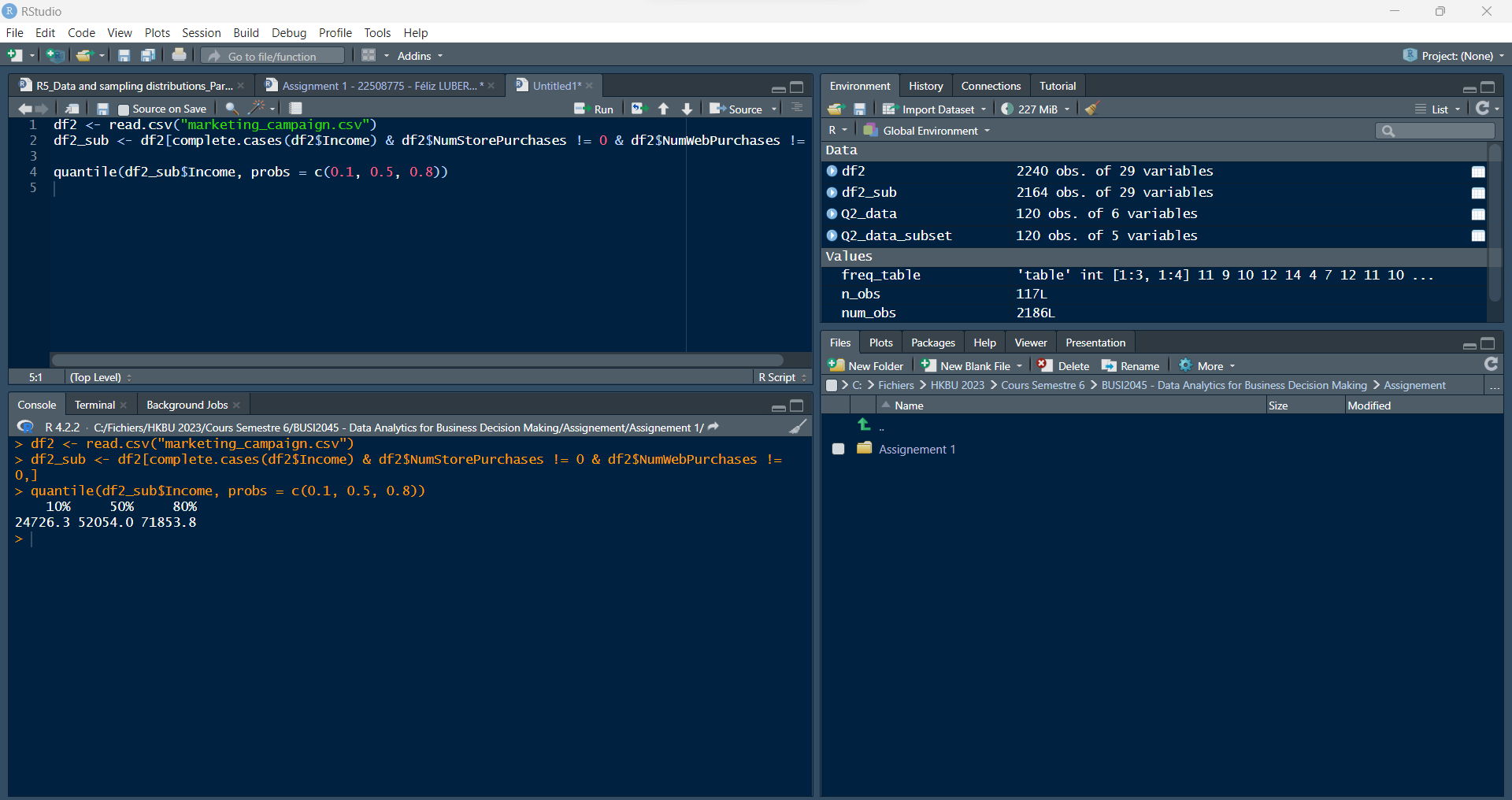
## What are the values of 10%, 50%, 80% percentile for variable Income?

|  |
| --- |
| 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

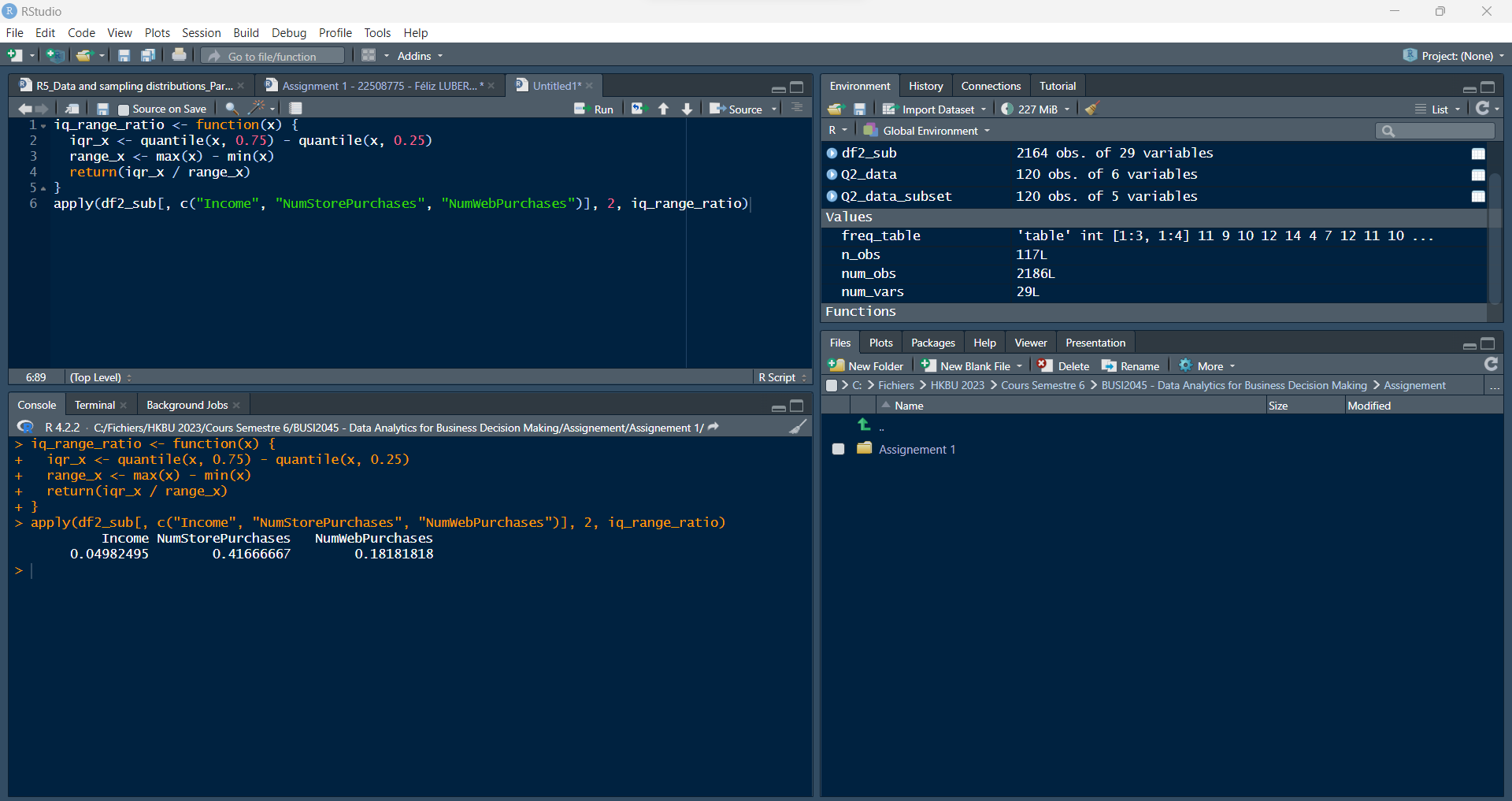
|  |
| --- |
| Full Screen |



## 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 |



## 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 |

