Assignment 2

Féliz LUBERNE- 22508775

Table des matières

Question 1: Correlation (30 points)

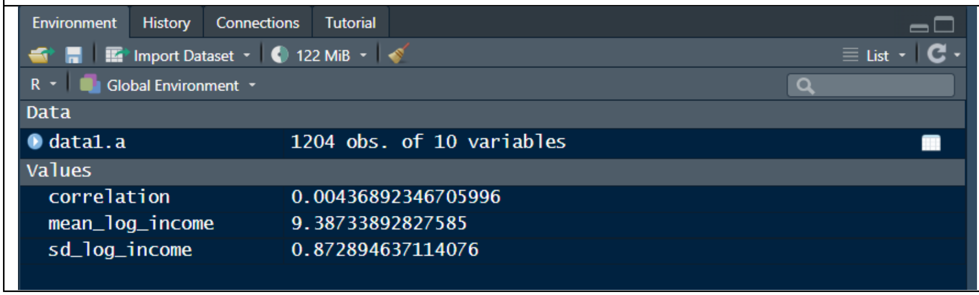
Load the data NILT2012GR_SUBSET.csv and answer the following questions. The data set contains 9 variables for 1204 citizens, which comes from Queen's University in Belfast (North Ireland) and is based on the Northern Ireland Life and Times Survey (NILT) 2012.

(a) Create a new variable named log_Income which takes log transformation of the variable persinc2 and calculate its mean and standard deviation. Note that the variable persinc2 measures personal income before tax and national insurance contributions. Then calculate the correlation coefficient between log_Income and rage. (Hints: note that the two variables contain NA values).

Code to be entered

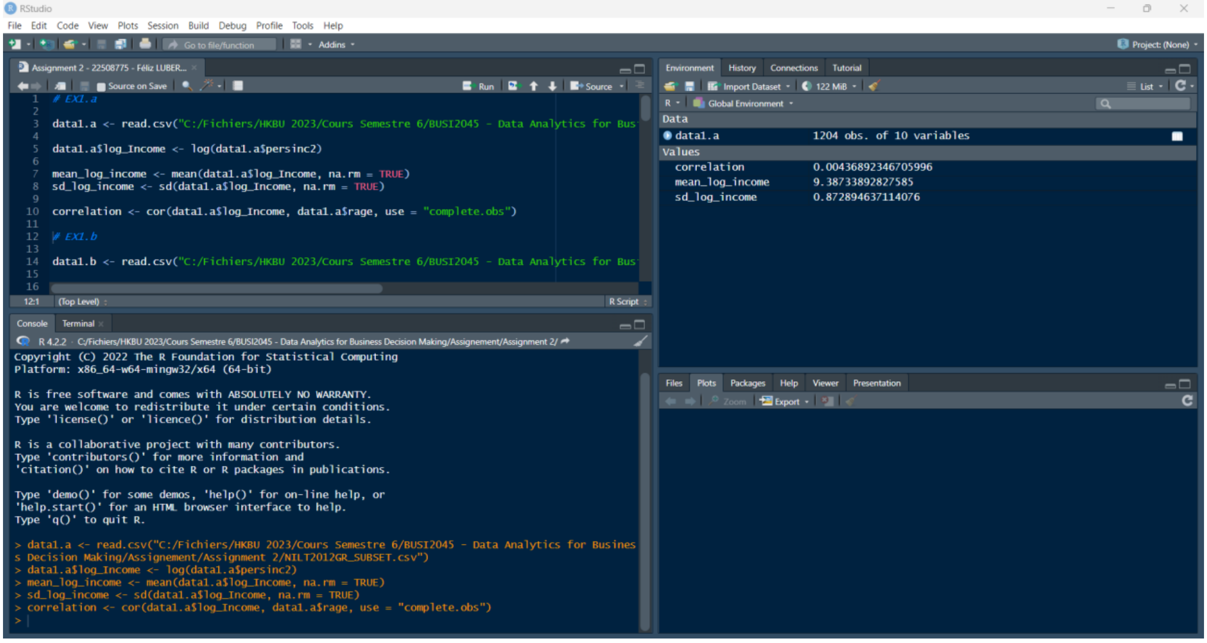
```
data1.a <- read.csv("NILT2012GR_SUBSET.csv")

data1.a$log_Income <- log(data1.a$persinc2)

mean_log_income <- mean(data1.a$log_Income, na.rm = TRUE)
sd_log_income <- sd(data1.a$log_Income, na.rm = TRUE)

correlation <- cor(data1.a$log_Income, data1.a$rage, use = "complete.obs")
```

Results



```
Data
data1.a            1204 obs. of 10 variables
Values
  correlation      0.00436892346705996
  mean_log_income  9.38733892827585
  sd_log_income    0.872894637114076
```

Full Screen



```
# EX1.a

data1.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Bus

data1.a$log_Income <- log(data1.a$persinc2)

mean_log_income <- mean(data1.a$log_Income, na.rm = TRUE)
sd_log_income <- sd(data1.a$log_Income, na.rm = TRUE)

correlation <- cor(data1.a$log_Income, data1.a$rage, use = "complete.obs")

# EX1.b

data1.b <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Bus
```

R 4.2.2 · C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Making/Assignement/Assignment 2/

```
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data1.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Busines
s Decision Making/Assignement/Assignment 2/NILT2012GR_SUBSET.csv")
> data1.a$log_Income <- log(data1.a$persinc2)
> mean_log_income <- mean(data1.a$log_Income, na.rm = TRUE)
> sd_log_income <- sd(data1.a$log_Income, na.rm = TRUE)
> correlation <- cor(data1.a$log_Income, data1.a$rage, use = "complete.obs")
> 
```

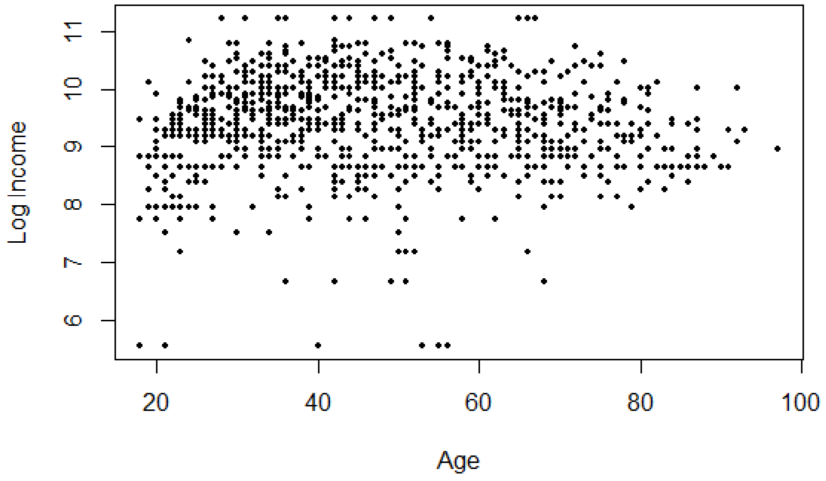
```
Data
data1.a            1204 obs. of 10 variables
Values
  correlation      0.00436892346705996
  mean_log_income  9.38733892827585
  sd_log_income    0.872894637114076
```

(b) Build a scatter plot to visualize the relationship between log_Income and rage (which measures age for each person). What is the relationship between log_Income and rage based on the plot?7

Code to be entered

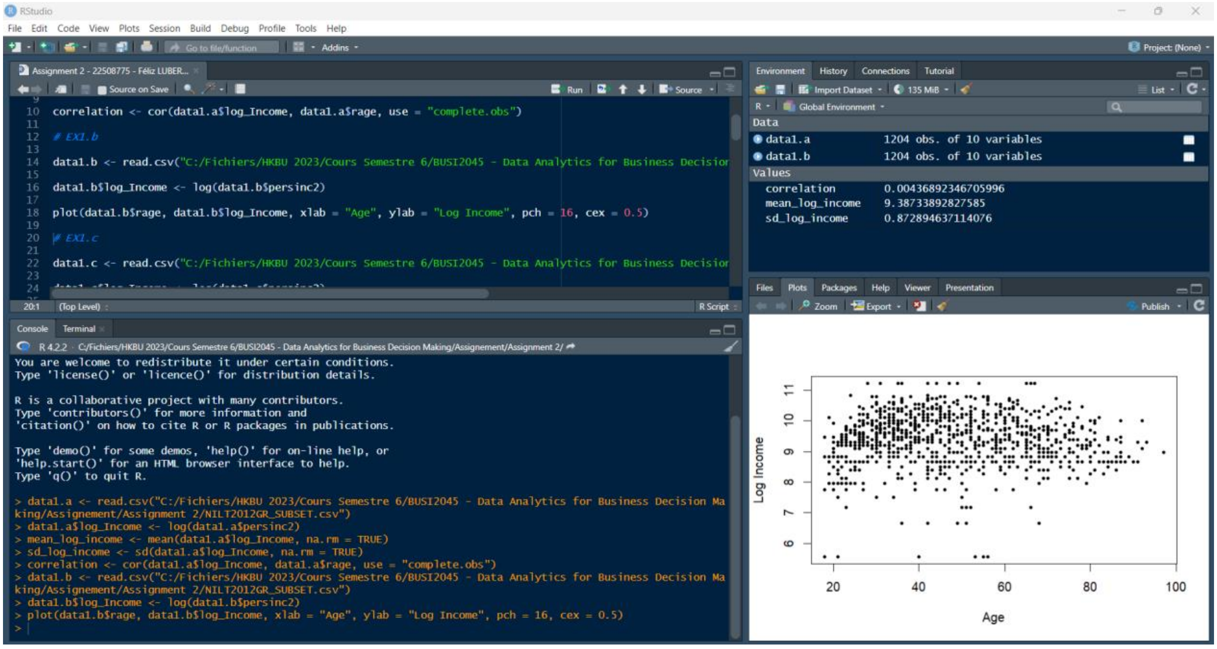
```
data1.b <- read.csv("NILT2012GR_SUBSET.csv")

data1.b$log_Income <- log(data1.b$persinc2)

plot(data1.b$rage, data1.b$log_Income, xlab = "Age", ylab = "Log Income", pch = 16, cex = 0.5)
```

Results



Answer : According to the graphic, there seems to be a tenuous positive correlation between log_Income and age. This indicates that, although the association between age and log_Income tends to be weak, it tends to increase as well. Additionally, log_Income appears to vary somewhat with age. Furthermore, It's worth noting that there are some extreme values of log_Income at the higher end of the age range, but it's unclear from the plot whether these are valid or outliers.

Full Screen



```
correlation <- cor(data1.a$log_Income, data1.a$rage, use = "complete.obs")

# EX1.b

data1.b <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decisior

data1.b$log_Income <- log(data1.b$persinc2)

plot(data1.b$rage, data1.b$log_Income, xlab = "Age", ylab = "Log Income", pch = 16, cex = 0.5)

# EX1.c

data1.c <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decisior
```

```
Data
data1.a            1204 obs. of 10 variables
data1.b            1204 obs. of 10 variables
Values
  correlation      0.00436892346705996
  mean_log_income  9.38733892827585
  sd_log_income    0.872894637114076
```

```
R 4.2.2 · C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Making/Assignement/Assignment 2/
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data1.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Ma
king/Assignement/Assignment 2/NILT2012GR_SUBSET.csv")
> data1.a$log_Income <- log(data1.a$persinc2)
> mean_log_income <- mean(data1.a$log_Income, na.rm = TRUE)
> sd_log_income <- sd(data1.a$log_Income, na.rm = TRUE)
> correlation <- cor(data1.a$log_Income, data1.a$rage, use = "complete.obs")
> data1.b <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Ma
king/Assignement/Assignment 2/NILT2012GR_SUBSET.csv")
> data1.b$log_Income <- log(data1.b$persinc2)
> plot(data1.b$rage, data1.b$log_Income, xlab = "Age", ylab = "Log Income", pch = 16, cex = 0.5)
> 
```

(c) When we conduct a statistical test on whether there is a linear association between log_Income and rage, what would be the null and alternative hypothesis? Implement this statistical test and interpret the result.

Code to be entered

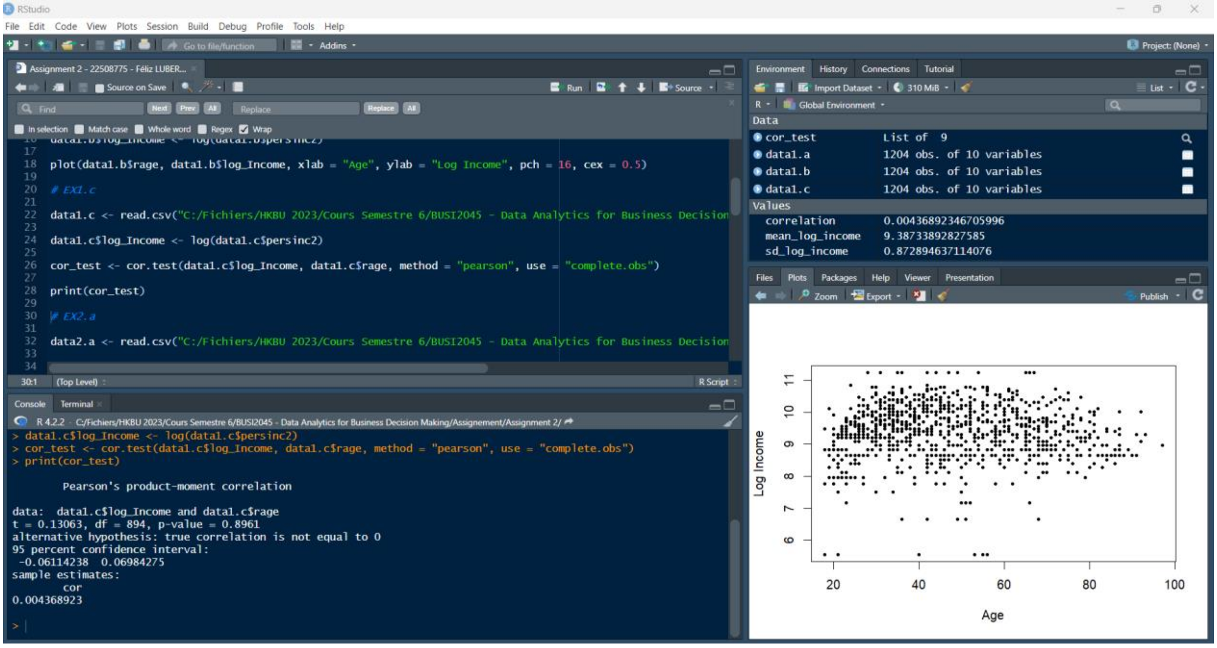
```
data1.c <- read.csv("NILT2012GR_SUBSET.csv")

data1.c$log_Income <- log(data1.c$persinc2)

cor_test   <-   cor.test(data1.c$log_Income,   data1.c$rage,   method   =
"pearson", use = "complete.obs")

print(cor_test)
```

Results

```
Pearson's product-moment correlation

data:  data1.c$log_Income and data1.c$rage
t = 0.13063, df = 894, p-value = 0.8961
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06114238  0.06984275
sample estimates:
        cor
0.004368923
```

Answer: The correlation coefficient (cor) is 0.004, indicating a very weak positive relationship between log_Income and rage. The p-value is 0.8961, which is greater than 0.05, indicating weak evidence against the null hypothesis. Therefore, we fail to reject the null hypothesis and conclude that there is insufficient evidence to support the claim of a statistically significant linear association between log_Income and rage. It's important to note that the confidence interval (-0.061, 0.070) contains 0, further supporting the lack of evidence for a linear relationship between the two variables

Full Screen



```r
data1.b$log_Income <- log(data1.b$persinc2)

plot(data1.b$rage, data1.b$log_Income, xlab = "Age", ylab = "Log Income", pch = 16, cex = 0.5)

# EX1.c

data1.c <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision

data1.c$log_Income <- log(data1.c$persinc2)

cor_test <- cor.test(data1.c$log_Income, data1.c$rage, method = "pearson", use = "complete.obs")

print(cor_test)

# EX2.a

data2.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision
```

```
R 4.2.2 · C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Making/Assignement/Assignment 2/
> data1.c$log_Income <- log(data1.c$persinc2)
> cor_test <- cor.test(data1.c$log_Income, data1.c$rage, method = "pearson", use = "complete.obs")
> print(cor_test)

	Pearson's product-moment correlation

data:  data1.c$log_Income and data1.c$rage
t = 0.13063, df = 894, p-value = 0.8961
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06114238  0.06984275
sample estimates:
        cor 
0.004368923 

> 
```

```
Data
cor_test          List of  9
data1.a           1204 obs. of 10 variables
data1.b           1204 obs. of 10 variables
data1.c           1204 obs. of 10 variables
Values
  correlation       0.00436892346705996
  mean_log_income   9.38733892827585
  sd_log_income     0.872894637114076
```

Question 2 Compare Groups (40 Points)

Read the data marketing_campaign.csv in R. Assume the data is a random sample from a population and each row represents a customer, answer the following questions.

(a) Create a subset in which the variable Education only contains “Graduation”, “Master” , and “PhD” values, and the variable Marital_Status only contains “Single” and “Married” values. Check how many observations left in the subset.

Use the subset to answer the following questions.

Code to be entered

```
data2.a <- read.csv("marketing_campaign.csv")

data_subset <- subset(data2.a, Education %in% c("Graduation", "Master",
"PhD") & Marital_Status %in% c("Single", "Married"))

nrow(data_subset)
```

Results

```
[1] 1188
```

Answer: There is 1188 observations in the subset

(b) Which education group has the highest number of customers? Which education group has the highest marriage rate?

Code to be entered

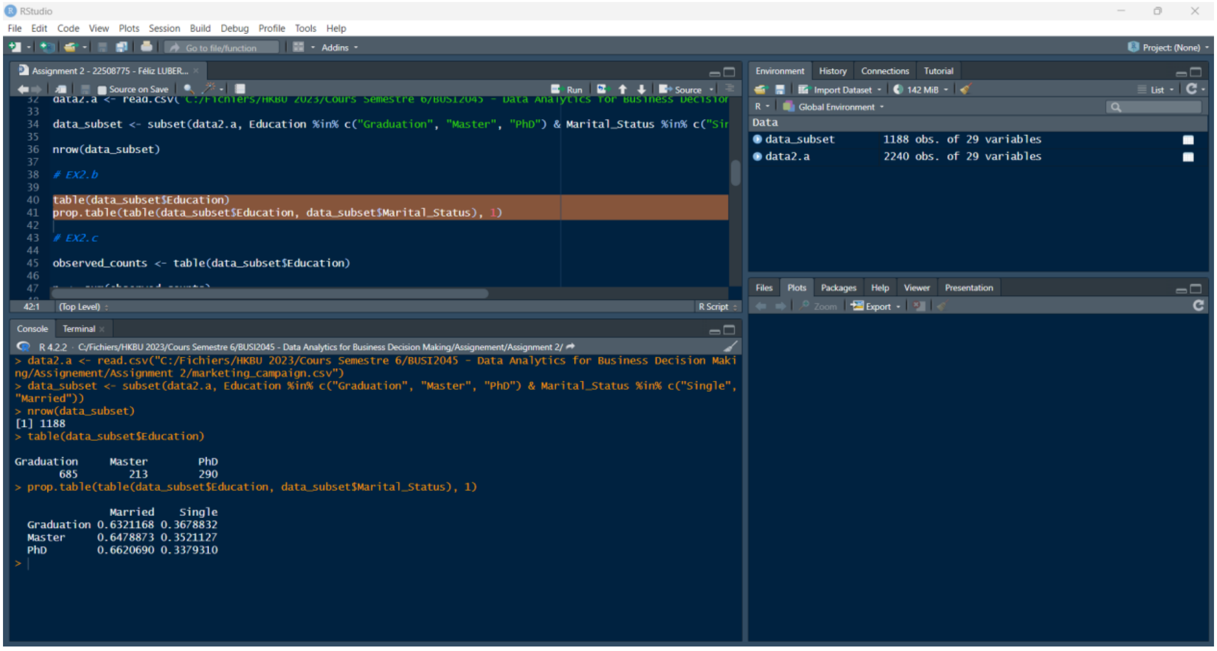
```
table(data_subset$Education)
prop.table(table(data_subset$Education,    data_subset$Marital_Status),
1)
```

Results

```
> table(data_subset$Education)

Graduation      Master        PhD
       685         213        290
> prop.table(table(data_subset$Education, data_subset$Marital_Status),
1)

               Married    Single
  Graduation 0.6321168 0.3678832
  Master     0.6478873 0.3521127
  PhD        0.6620690 0.3379310
```

Answer: The group with the highest number of customers is the graduation group, and the education group with the highest marriage rate is the PhD group.

Full Screen



```
data2.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decisior

data_subset <- subset(data2.a, Education %in% c("Graduation", "Master", "PhD") & Marital_Status %in% c("Sir

nrow(data_subset)

# EX2.b

table(data_subset$Education)
prop.table(table(data_subset$Education, data_subset$Marital_Status), 1)

# EX2.c

observed_counts <- table(data_subset$Education)
```

```
Data
data_subset       1188 obs. of 29 variables
data2.a           2240 obs. of 29 variables
```

R 4.2.2 · C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Making/Assignement/Assignment 2/

```
> data2.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Maki
ng/Assignement/Assignment 2/marketing_campaign.csv")
> data_subset <- subset(data2.a, Education %in% c("Graduation", "Master", "PhD") & Marital_Status %in% c("Single",
"Married"))
> nrow(data_subset)
[1] 1188
> table(data_subset$Education)

Graduation     Master        PhD 
       685        213        290 
> prop.table(table(data_subset$Education, data_subset$Marital_Status), 1)
            
               Married    Single
  Graduation 0.6321168 0.3678832
  Master     0.6478873 0.3521127
  PhD        0.6620690 0.3379310
> 
```

(c) Conduct a statistical test to explore whether the number of customers is the same across education groups. What is the null and alternative hypothesis? What is your conclusion based on the result?

Code to be entered

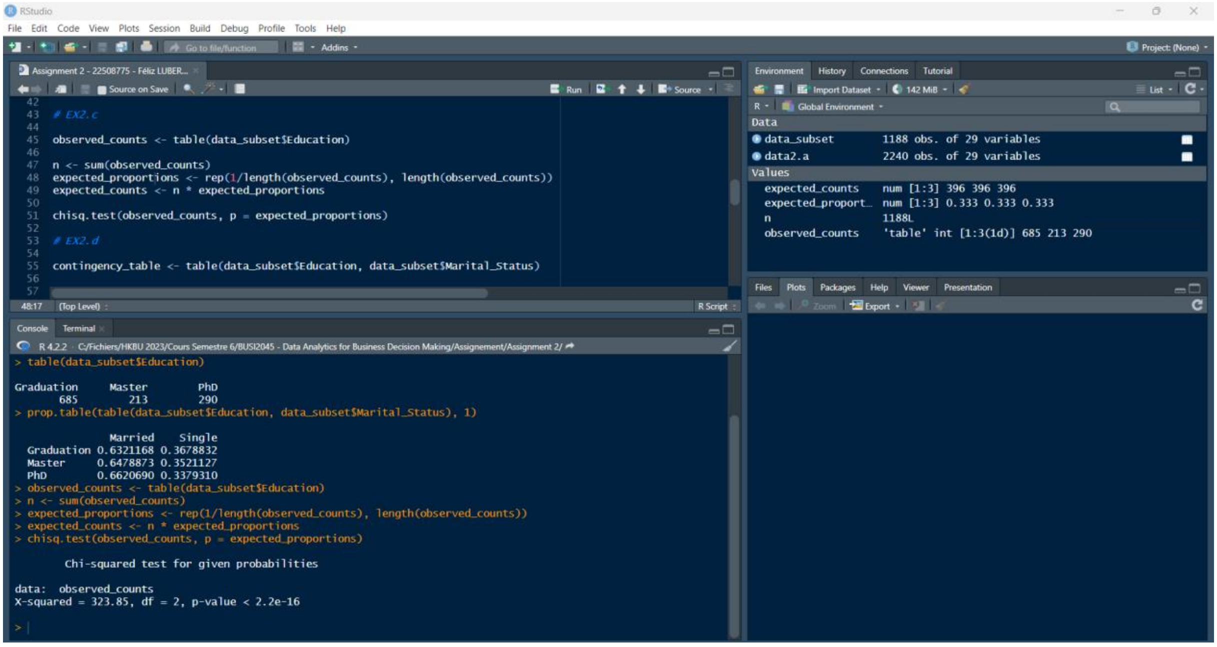
```
observed_counts <- table(data_subset$Education)

n <- sum(observed_counts)
expected_proportions <- rep(1/length(observed_counts), 
length(observed_counts))
expected_counts <- n * expected_proportions

chisq.test(observed_counts, p = expected_proportions)
```

Results

```
> chisq.test(observed_counts, p = expected_proportions)

	Chi-squared test for given probabilities

data:  observed_counts
X-squared = 323.85, df = 2, p-value < 2.2e-16
```

Asnwer: This shows that the degrees of freedom (df) are 2 and the test statistic (X-squared) is 323.85. The p-value is lower than 2.2e-16, which is less than the usual threshold of 0.05 for significance. Since there is a statistically significant variation in the proportion of consumers across education groups, the null hypothesis that the proportion of customers in each education group is equal can be rejected.

Full Screen



```
# EX2.c

observed_counts <- table(data_subset$Education)

n <- sum(observed_counts)
expected_proportions <- rep(1/length(observed_counts), length(observed_counts))
expected_counts <- n * expected_proportions

chisq.test(observed_counts, p = expected_proportions)

# EX2.d

contingency_table <- table(data_subset$Education, data_subset$Marital_Status)
```

```
> table(data_subset$Education)

Graduation     Master        PhD 
       685        213        290 
> prop.table(table(data_subset$Education, data_subset$Marital_Status), 1)
            
               Married    Single
  Graduation 0.6321168 0.3678832
  Master     0.6478873 0.3521127
  PhD        0.6620690 0.3379310
> observed_counts <- table(data_subset$Education)
> n <- sum(observed_counts)
> expected_proportions <- rep(1/length(observed_counts), length(observed_counts))
> expected_counts <- n * expected_proportions
> chisq.test(observed_counts, p = expected_proportions)

	Chi-squared test for given probabilities

data:  observed_counts
X-squared = 323.85, df = 2, p-value < 2.2e-16

> 
```

```
Data
data_subset          1188 obs. of 29 variables
data2.a              2240 obs. of 29 variables
Values
  expected_counts      num [1:3] 396 396 396
  expected_proport…    num [1:3] 0.333 0.333 0.333
  n                    1188L
  observed_counts      'table' int [1:3(1d)] 685 213 290
```

(d) We'd like to know whether Marital_Status is related with Education. What is the null and alternative hypothesis? What is your conclusion based on statistical test?

Code to be entered

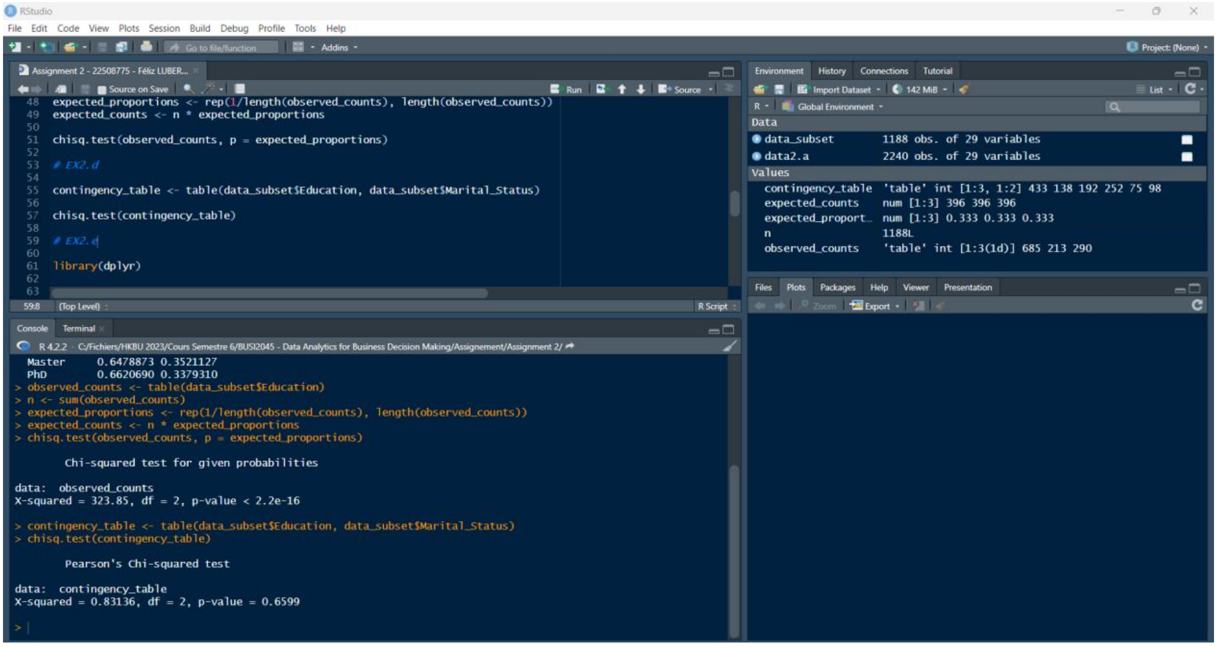
```
contingency_table          <-          table(data_subset$Education,
data_subset$Marital_Status)

chisq.test(contingency_table)
```

Results

```
> chisq.test(contingency_table)

Pearson's Chi-squared test

data:  contingency_table
X-squared = 0.83136, df = 2, p-value = 0.6599
```

Answer: The output indicates that a Pearson's chi-squared test was conducted on the contingency table created from the education and marital status variables in the data subset. The test resulted in a chi-squared value of 0.83136 with 2 degrees of freedom and a p-value of 0.6599. This suggests that there is no evidence of a significant association between education and marital status in the data subset.

Full Screen



```r
expected_proportions <- rep(1/length(observed_counts), length(observed_counts))
expected_counts <- n * expected_proportions

chisq.test(observed_counts, p = expected_proportions)

# EX2.d

contingency_table <- table(data_subset$Education, data_subset$Marital_Status)

chisq.test(contingency_table)

# EX2.e

library(dplyr)

```

```
R 4.2.2 · C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Making/Assignement/Assignment 2/
  Master     0.6478873 0.3521127
  PhD        0.6620690 0.3379310
> observed_counts <- table(data_subset$Education)
> n <- sum(observed_counts)
> expected_proportions <- rep(1/length(observed_counts), length(observed_counts))
> expected_counts <- n * expected_proportions
> chisq.test(observed_counts, p = expected_proportions)

	Chi-squared test for given probabilities

data:  observed_counts
X-squared = 323.85, df = 2, p-value < 2.2e-16

> contingency_table <- table(data_subset$Education, data_subset$Marital_Status)
> chisq.test(contingency_table)

	Pearson's Chi-squared test

data:  contingency_table
X-squared = 0.83136, df = 2, p-value = 0.6599

> 
```

```
Data
data_subset          1188 obs. of 29 variables
data2.a              2240 obs. of 29 variables
Values
  contingency_table  'table' int [1:3, 1:2] 433 138 192 252 75 98
  expected_counts    num [1:3] 396 396 396
  expected_proport…  num [1:3] 0.333 0.333 0.333
  n                  1188L
  observed_counts    'table' int [1:3(1d)] 685 213 290
```

(e) What is the marriage rate in general? Given the observed marriage rate, can we say that in the population the true marriage rate is 60%? Why?

Code to be entered

```
marriage_rate  <-  sum(data_subset$Marital_Status  ==  "Married")  /
nrow(data_subset)
summary_data <- data.frame(marriage_rate)

print(summary_data)
```

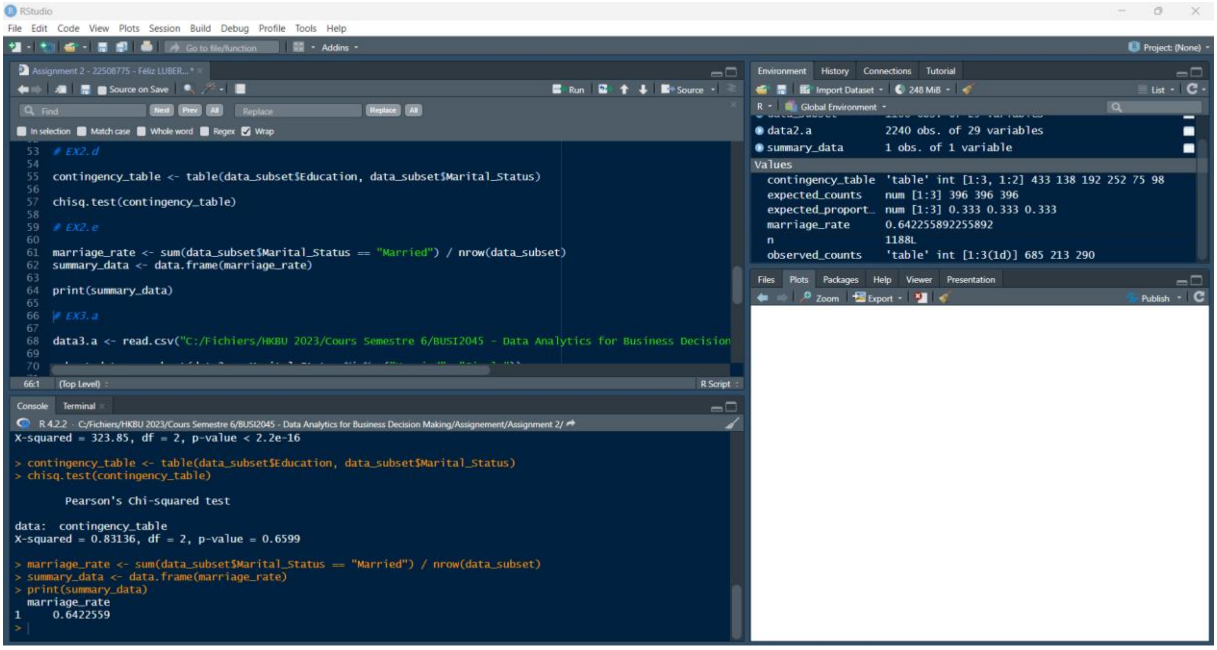
Results

```
marriage_rate
1      0.6422559
```

Answer: The marriage rate in the subset of the data is approximately 64.23%. However, this does not necessarily represent the true marriage rate in the population, as it is only based on a sample of the data.

We cannot say whether the true marriage rate in the population is 60% based solely on the observed marriage rate in the subset of the data. We can use statistical tests to evaluate the evidence for (or against) a specific value of the true marriage rate, but we cannot definitively determine the true marriage rate from a single sample of data.

Full Screen



```r
# EX2.d

contingency_table <- table(data_subset$Education, data_subset$Marital_Status)

chisq.test(contingency_table)

# EX2.e

marriage_rate <- sum(data_subset$Marital_Status == "Married") / nrow(data_subset)
summary_data <- data.frame(marriage_rate)

print(summary_data)

# EX3.a

data3.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision
```

R 4.2.2 · C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Making/Assignement/Assignment 2/

```
X-squared = 323.85, df = 2, p-value < 2.2e-16

> contingency_table <- table(data_subset$Education, data_subset$Marital_Status)
> chisq.test(contingency_table)

	Pearson's Chi-squared test

data:  contingency_table
X-squared = 0.83136, df = 2, p-value = 0.6599

> marriage_rate <- sum(data_subset$Marital_Status == "Married") / nrow(data_subset)
> summary_data <- data.frame(marriage_rate)
> print(summary_data)
  marriage_rate
1     0.6422559
> 
```

```
data2.a             2240 obs. of 29 variables
summary_data        1 obs. of 1 variable
Values
  contingency_table 'table' int [1:3, 1:2] 433 138 192 252 75 98
  expected_counts   num [1:3] 396 396 396
  expected_proport… num [1:3] 0.333 0.333 0.333
  marriage_rate     0.642255892255892
  n                 1188L
  observed_counts   'table' int [1:3(1d)] 685 213 290
```

Question 3: Compare Groups (30 Points)
Continue with the original data marketing_campaign.csv and answer the following questions. Note the below questions are based on the entire dataset, not the subset created in 2(a)

(a) What is the average income (variable Income) for the single and married group? Are their averages income truly different in the population? State your null and alternative hypotheses, implement the hypothesis test, and interpret the result.

Code to be entered

```
data3.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 -
Data Analytics for Business Decision Making/Assignement/Assignment
2/marketing_campaign.csv")

subset_data <- subset(data3.a, Marital_Status %in% c("Married",
"Single"))
income_avg <- aggregate(Income ~ Marital_Status, data = subset_data,
mean)
t.test(Income ~ Marital_Status, data = subset_data)
```

Results

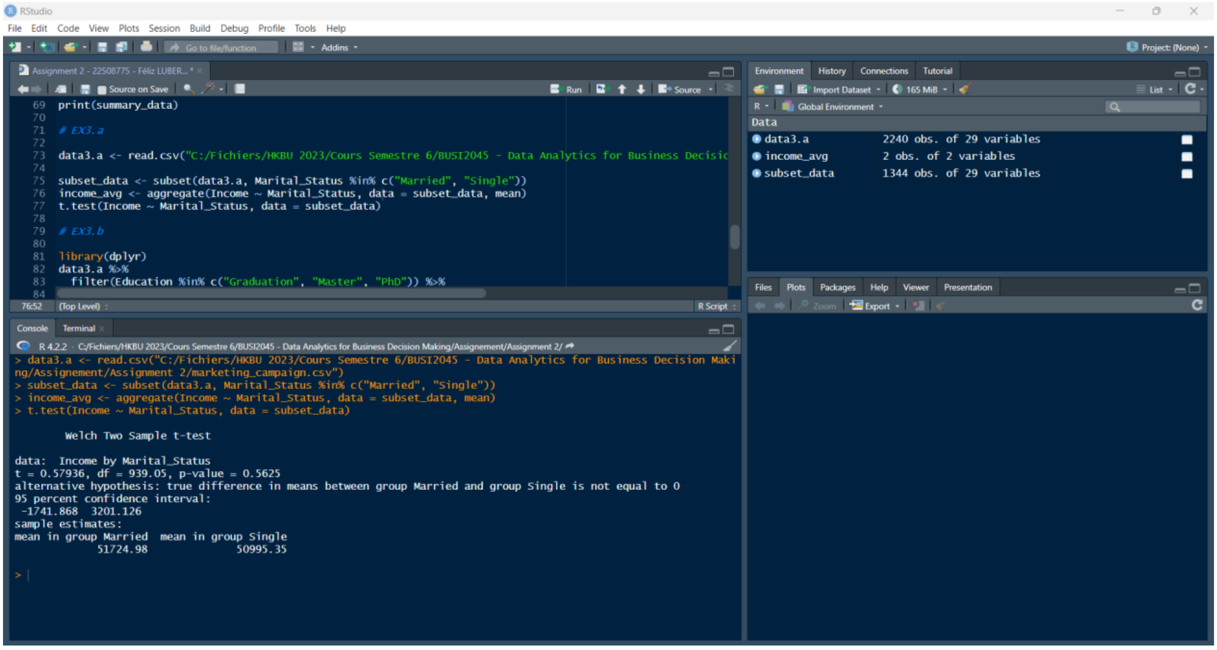
```
Welch Two Sample t-test

data:  Income by Marital_Status
t = 0.57936, df = 939.05, p-value = 0.5625
alternative hypothesis: true difference in means between group Married
and group Single is not equal to 0
95 percent confidence interval:
 -1741.868  3201.126
sample estimates:
mean in group Married  mean in group Single
             51724.98              50995.35
```

Answer: The average income for the single group is 50995.35, and the average income for the married group is 51724.98.

The null hypothesis is that there is no significant difference in the mean income between the single and married groups in the population. The alternative hypothesis is that there is a significant difference in the mean income between the two groups.

To test this hypothesis, we can perform a two-sample t-test. From the output, we see that the t-statistic is 0.57936 and the p-value is 0.5625. Since the p-value is greater than 0.05 (assuming a significance level of 0.05), we fail to reject the null hypothesis. This means that we do not have sufficient evidence to conclude that there is a significant difference in the mean income between the single and married groups in the population.

Full Screen



```
print(summary_data)

# EX3.a

data3.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decisic

subset_data <- subset(data3.a, Marital_Status %in% c("Married", "Single"))
income_avg <- aggregate(Income ~ Marital_Status, data = subset_data, mean)
t.test(Income ~ Marital_Status, data = subset_data)

# EX3.b

library(dplyr)
data3.a %>%
  filter(Education %in% c("Graduation", "Master", "PhD")) %>%
```

```
> data3.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Maki
ng/Assignement/Assignment 2/marketing_campaign.csv")
> subset_data <- subset(data3.a, Marital_Status %in% c("Married", "Single"))
> income_avg <- aggregate(Income ~ Marital_Status, data = subset_data, mean)
> t.test(Income ~ Marital_Status, data = subset_data)

	Welch Two Sample t-test

data:  Income by Marital_Status
t = 0.57936, df = 939.05, p-value = 0.5625
alternative hypothesis: true difference in means between group Married and group Single is not equal to 0
95 percent confidence interval:
 -1741.868  3201.126
sample estimates:
mean in group Married  mean in group Single 
             51724.98              50995.35 

> 
```

```
Data
data3.a        2240 obs. of 29 variables
income_avg     2 obs. of 2 variables
subset_data    1344 obs. of 29 variables
```

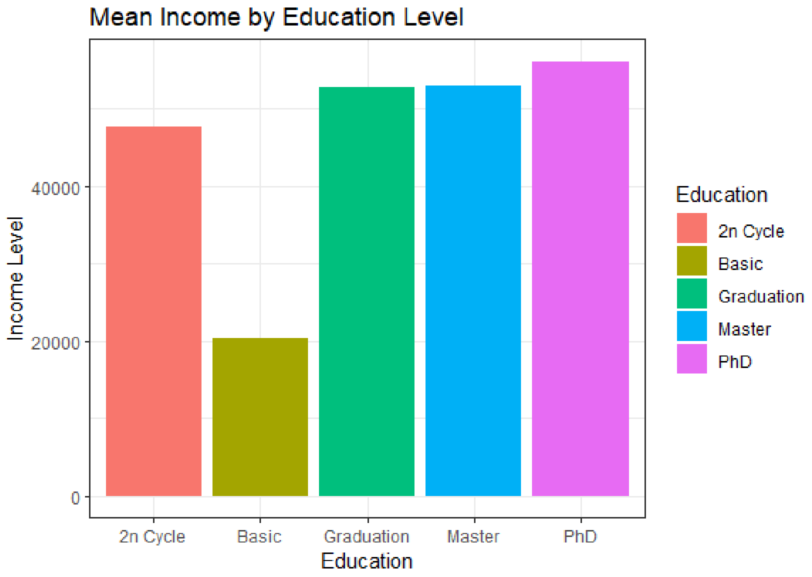
(b) What is the average income across different education groups (Education)? Please display the result with both a statistic summary and a bar plot.

Code to be entered

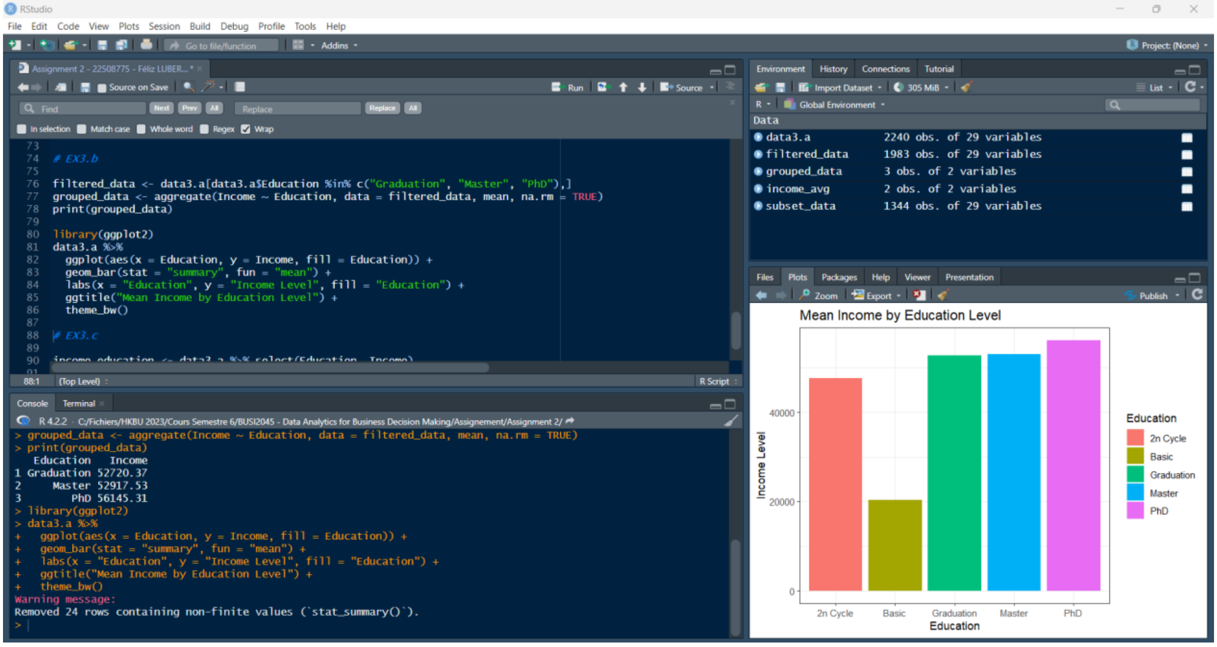
```
filtered_data <- data3.a[data3.a$Education %in% c("Graduation", "Master", "PhD"),]
grouped_data <- aggregate(Income ~ Education, data = filtered_data, mean, na.rm = TRUE)
print(grouped_data)

library(ggplot2)
data3.a %>%
  ggplot(aes(x = Education, y = Income, fill = Education)) +
  geom_bar(stat = "summary", fun = "mean") +
  labs(x = "Education", y = "Income Level", fill = "Education") +
  ggtitle("Mean Income by Education Level") +
  theme_bw()
```

Results

```
> filtered_data <- data3.a[data3.a$Education %in% c("Graduation", "Master", "PhD"),]
> grouped_data <- aggregate(Income ~ Education, data = filtered_data, mean, na.rm = TRUE)
> print(grouped_data)
   Education   Income
1 Graduation 52720.37
2     Master 52917.53
3        PhD 56145.31
```



Full Screen



```
# EX3.b

filtered_data <- data3.a[data3.a$Education %in% c("Graduation", "Master", "PhD"),]
grouped_data <- aggregate(Income ~ Education, data = filtered_data, mean, na.rm = TRUE)
print(grouped_data)

library(ggplot2)
data3.a %>%
  ggplot(aes(x = Education, y = Income, fill = Education)) +
  geom_bar(stat = "summary", fun = "mean") +
  labs(x = "Education", y = "Income Level", fill = "Education") +
  ggtitle("Mean Income by Education Level") +
  theme_bw()

# EX3.c

income_education <- data3.a %>% select(Education, Income)
```

Data

data3.a	2240 obs. of 29 variables
filtered_data	1983 obs. of 29 variables
grouped_data	3 obs. of 2 variables
income_avg	2 obs. of 2 variables
subset_data	1344 obs. of 29 variables

R 4.2.2 · C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Making/Assignement/Assignment 2/

```
> grouped_data <- aggregate(Income ~ Education, data = filtered_data, mean, na.rm = TRUE)
> print(grouped_data)
   Education   Income
1 Graduation 52720.37
2     Master 52917.53
3        PhD 56145.31
> library(ggplot2)
> data3.a %>%
+   ggplot(aes(x = Education, y = Income, fill = Education)) +
+   geom_bar(stat = "summary", fun = "mean") +
+   labs(x = "Education", y = "Income Level", fill = "Education") +
+   ggtitle("Mean Income by Education Level") +
+   theme_bw()
Warning message:
Removed 24 rows containing non-finite values (`stat_summary()`).
> 
```

(c) Are the average incomes in the five education groups truly different in the population? Please state your null and alternative hypothesis, implement the hypothesis test, and interpret the result.

Code to be entered

```
income_education <- data3.a %>% select(Education, Income)
model <- aov(Income ~ Education, data = income_education)
summary(model)
```

Results

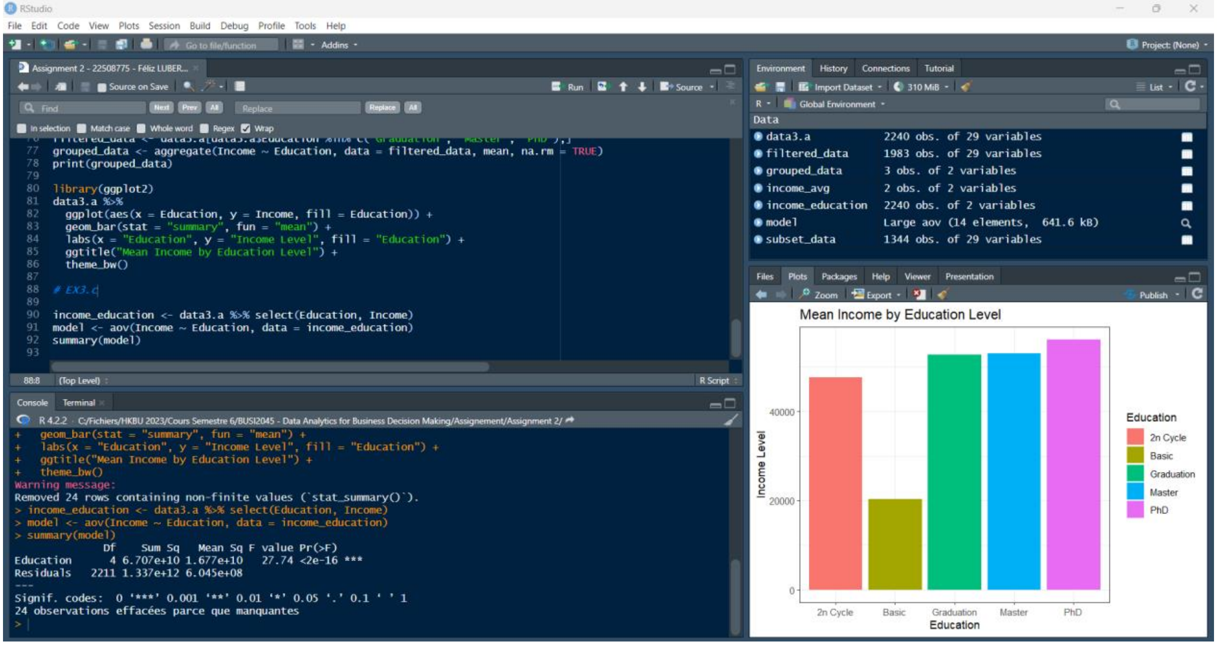
```
> income_education <- data3.a %>% select(Education, Income)
> model <- aov(Income ~ Education, data = income_education)
> summary(model)
                 Df     Sum Sq    Mean Sq F value Pr(>F)
```

```
Education      4 6.707e+10 1.677e+10   27.74 <2e-16 ***
Residuals   2211 1.337e+12 6.045e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Answer: The null hypothesis is that the mean income is the same across all education groups, and the alternative hypothesis is that the mean income is different across at least one education group.

We used ANOVA to test the null hypothesis. The ANOVA result shows that the p-value for Education is less than 0.05, which indicates that there is a significant difference in income across the five education groups. Therefore, we can reject the null hypothesis that the mean income is the same across all education groups.

Full Screen



```
filtered_data <- data3.a[data3.a$Education %in% c("Graduation", "Master", "PhD"),]
grouped_data <- aggregate(Income ~ Education, data = filtered_data, mean, na.rm = TRUE)
print(grouped_data)

library(ggplot2)
data3.a %>%
  ggplot(aes(x = Education, y = Income, fill = Education)) +
  geom_bar(stat = "summary", fun = "mean") +
  labs(x = "Education", y = "Income Level", fill = "Education") +
  ggtitle("Mean Income by Education Level") +
  theme_bw()

# EX3.c

income_education <- data3.a %>% select(Education, Income)
model <- aov(Income ~ Education, data = income_education)
summary(model)
```

```
+   geom_bar(stat = "summary", fun = "mean") +
+   labs(x = "Education", y = "Income Level", fill = "Education") +
+   ggtitle("Mean Income by Education Level") +
+   theme_bw()
Warning message:
Removed 24 rows containing non-finite values (`stat_summary()`).
> income_education <- data3.a %>% select(Education, Income)
> model <- aov(Income ~ Education, data = income_education)
> summary(model)
              Df    Sum Sq   Mean Sq F value Pr(>F)
Education      4 6.707e+10 1.677e+10   27.74 <2e-16 ***
Residuals   2211 1.337e+12 6.045e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
24 observations effacées parce que manquantes
>
```

```
Data
data3.a            2240 obs. of 29 variables
filtered_data      1983 obs. of 29 variables
grouped_data       3 obs. of 2 variables
income_avg         2 obs. of 2 variables
income_education   2240 obs. of 2 variables
model              Large aov (14 elements,  641.6 kB)
subset_data        1344 obs. of 29 variables
```