**Assignment 2**

**Féliz LUBERNE - 22508775**

Table des matières

[**Question 1**: Correlation (30 points) 2](#_Toc130815009)

[**(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). 2](#_Toc130815010)

[**(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 3](#_Toc130815011)

[**(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. 4](#_Toc130815012)

[**Question 2** Compare Groups (40 Points) 6](#_Toc130815013)

[**(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. 6](#_Toc130815014)

[**(b)** Which education group has the highest number of customers? Which education group has the highest marriage rate? 6](#_Toc130815015)

[**(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? 7](#_Toc130815016)

[**(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? 8](#_Toc130815017)

[**(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? 9](#_Toc130815018)

[**Question 3:** Compare Groups (30 Points) 11](#_Toc130815019)

[**(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. 11](#_Toc130815020)

[**(b)** What is the average income across different education groups (Education)? Please display the result with both a statistic summary and a bar plot. 12](#_Toc130815021)

[**(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. 13](#_Toc130815022)

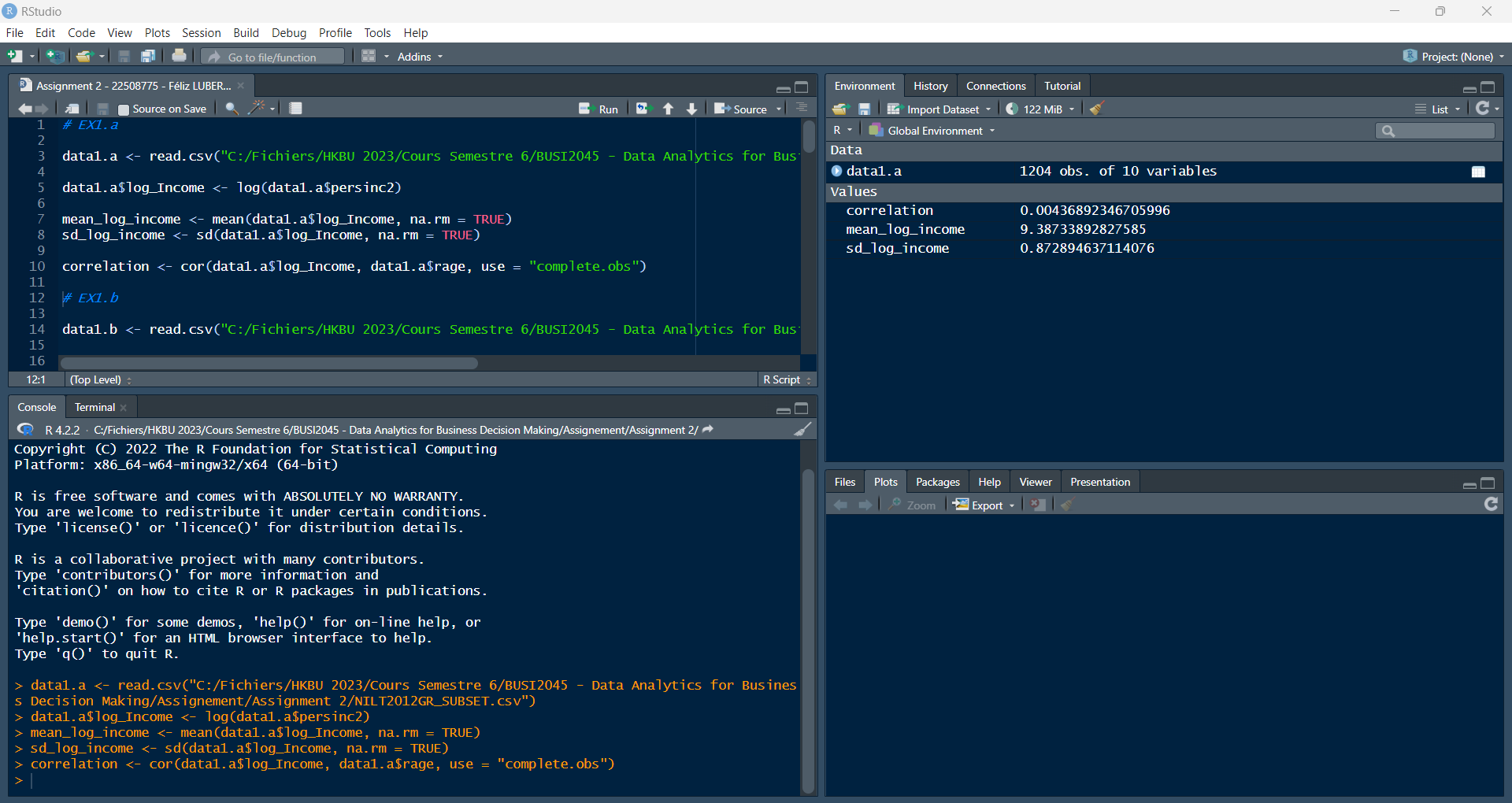
# **Question 1**: Correlation (30 points)

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 |
|  |

|  |
| --- |
| Full Screen |

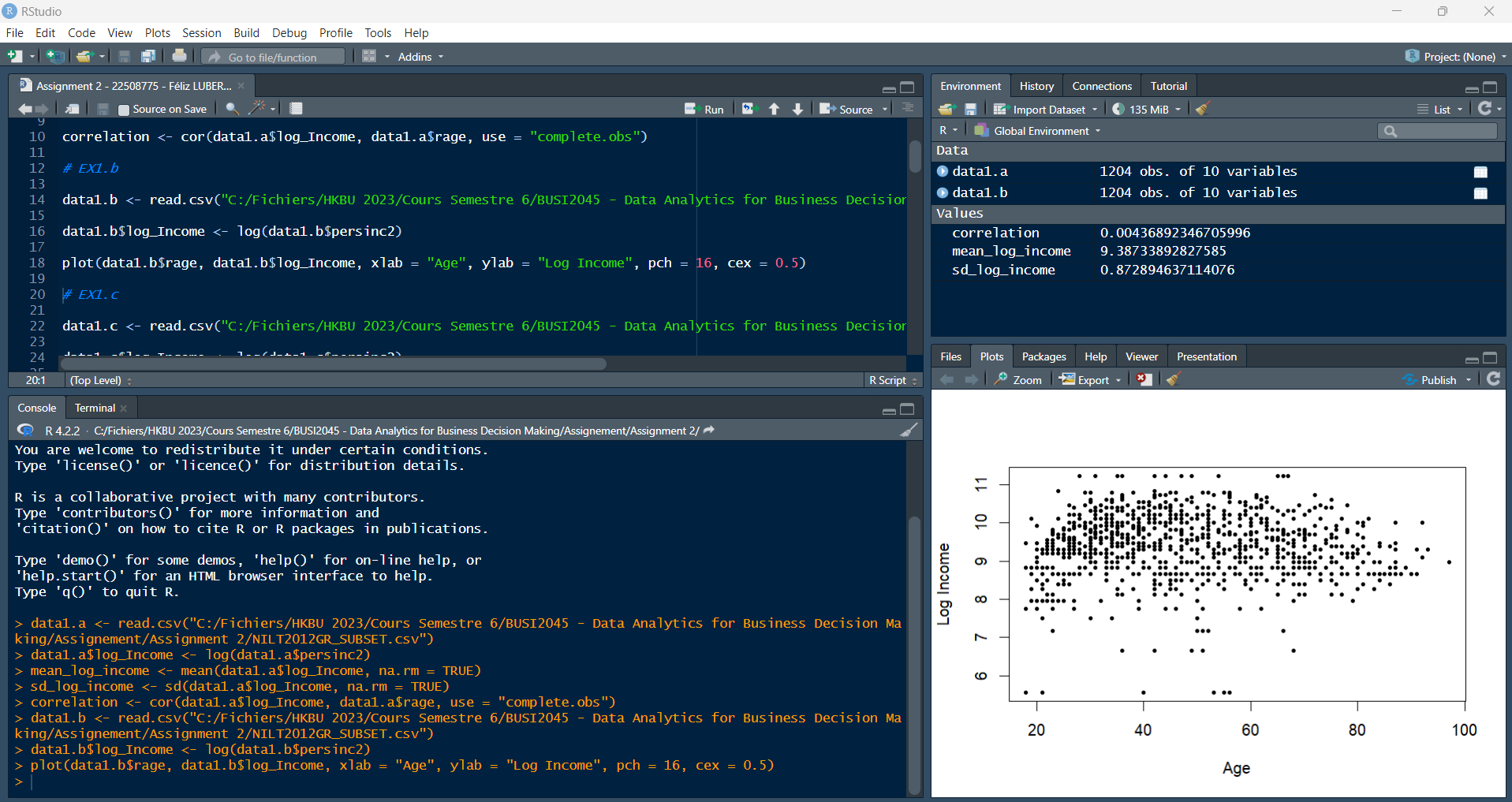


## **(b)** Build a scatter plot to visualize the relationship between log\_Income and rage (which measures age for each person). What is the relationship between log\_Income and rage based on the plot?7

|  |
| --- |
| 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 |



## **(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 |

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

Description générée automatiquement

# **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 |

Une image contenant texte, capture d’écran, intérieur

Description générée automatiquement

## **(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 |

Une image contenant texte, capture d’écran, écran, noir

Description générée automatiquement

## **(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 |

Une image contenant texte

Description générée automatiquement

## **(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 |

Une image contenant texte, capture d’écran, écran

Description générée automatiquement

# **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 |

Une image contenant texte, capture d’écran, écran, intérieur

Description générée automatiquement

## **(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 |

Une image contenant texte

Description générée automatiquement

## **(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 |

