BUSI2045 – Homework 2

Note: Please include both R codes and results in your report. (You may use the "Compile Report" function under Menu "File" in RStudio to generate a Word/PDF report)

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 contains NA values*).

(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?

(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.

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.

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

(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?

(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?

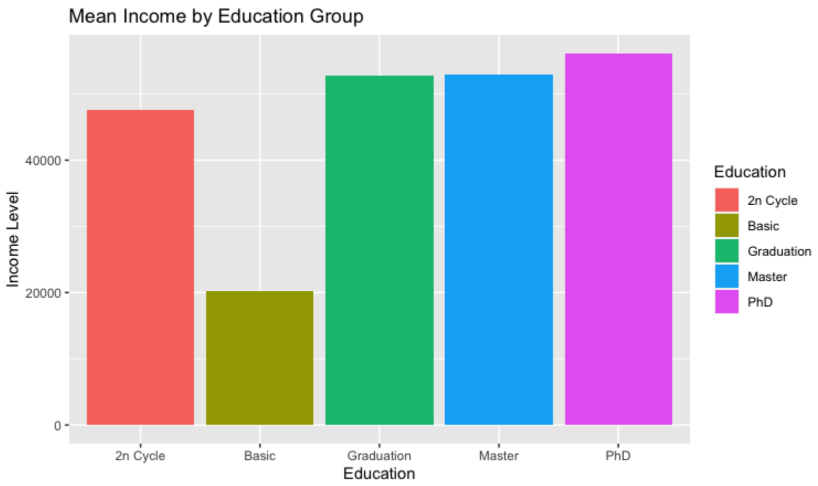
(e) What is the marriage rate in general? Given the observed marriage rate, can we say that in the population the true marriage rate is 60%? Why?

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 average income truly different in the population? State your null and alternative hypotheses, implement the hypothesis test, and interpret the result.
Hint: you may create a subset where the variable Marital_Status only include 'Married' and "Single".

(b) What is the average income across different education groups (*Education*)? Please display the result with both a statistic summary and a bar plot. The bar plot should look like similar as below.



(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.