UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

Subject:

Probability & Statistical Data Analysis

Code Subject:

SECI2143

PROJECT 2

"Student Placement in University"

NAME	Nor Armirani binti Mohd Mazlan
MATRIC NO	A19EC0122
SECTION	02
LECTURER'S NAME	Dr. Chan Weng Howe

Submission date : 27th June 2020

Table of Contents

Introductions

R Studio is an integrated environment (IDE) that is used for R, it provides open source tools and software for data science teams to develop their own works. By using R Studio, we can analyse data further and conclude the data by ourselves. For examples, R Studio can be used to plot graphs using the data collected thus users can make a conclusion from there.

The data chosen for this project is Placement of Students in Jain University Bangalore that is collected by Ben Roshan D. The purpose of this data set is to investigate the factors that affected the placement of students in the campus. The data set can be found at https://www.kaggle.com/benroshan/factors-affecting-campus-placement .

A group of 215 students answered the survey and the data is collected in order to be analysed further thus making the conclusion of the which factor affected the placement the most. From this data, several test will be used in order to analyse the data such as hypothesis testing, regression, correlation and chi square test. A few variables are selected according to the suitable test and relationship between the variables can be known. These are a few variables that are used in the test :-

Variables	Type
Degree Percentage	Ratio
Gender	Nominal
Status of Placement	Nominal
Secondary Education Percentage	Ratio
Higher Secondary Percentage	Ratio

Contents – Data Analysis

1. **Hypothesis testing (1 sample)**

   The variable chosen for hypothesis testing is Degree Percentage and the first step is to calculate the population mean by dividing the total of Degree percentage with the total number of students from the survey. Then, the hypothesis is made using the value of mean obtained. The H0 is mean equal to 2.41573 and the H1 is mean is not equal to 2.41573 or in this calculation mean is greater than 2.41573. The next step is to find the mean and standard deviation of the sample data then the value of z is calculated by using the formula z = (xbar – mu) / (s/sqrt(n)), xbar is the sample mean, mu is the population mean, s is the sample standard deviation and n is the total number of sample data. By using the z value obtained, we can find the p-value by referring to the normal distributions table then compare it with the alpha, significance level which is 0.05. In this test, the z value obtained is -4.7105. **Since (pvalue = 1.2356×10e-6 < α=0.05), therefore reject H0. There is enough evidence that mean is less than 2.41573.**

2. Chi Square Test

H0: There is no relationship between gender and status of placement

H1: Relationship exists between gender and status of placement

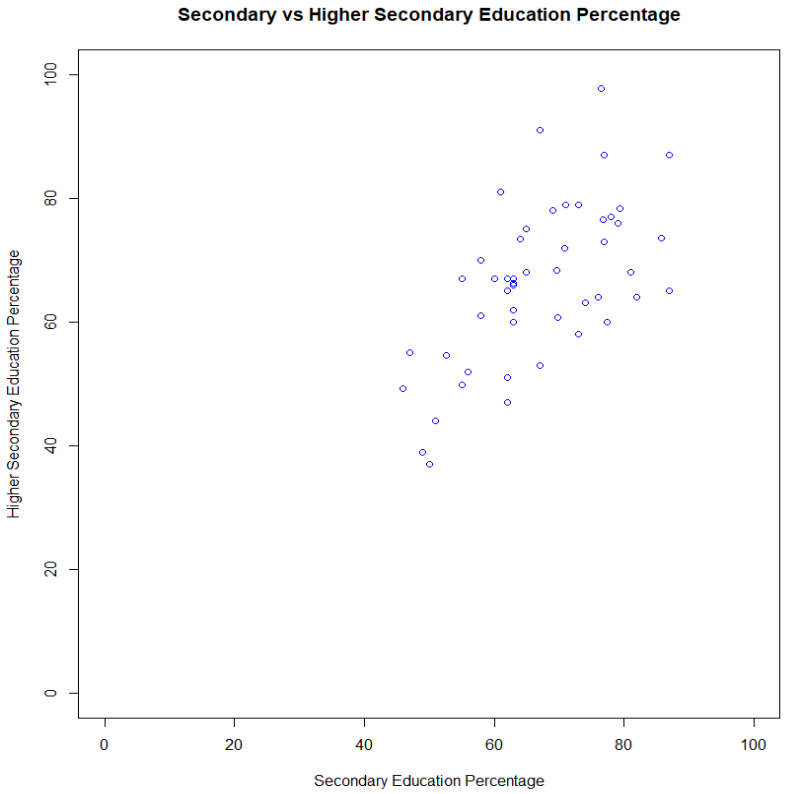
Gender	Status of Placement			
	Placed	**Expected**	**Not Placed**	**Expected**
M	19	E: (27×32)/50 = 17.28	8	E: (27×18)/50 = 9.72
F	13	E: (23×32)/50 = 14.72	10	E: (18×23)/50 = 8.28

Cell	Observed, O	Expected, E	$[O - E]^2 / E$
1,1	19	17.28	0.1712
1,2	8	9.72	0.3044
2,1	13	14.72	0.2010
2,2	10	8.28	0.3573
x^2 =			**1.0339**

$\boldsymbol{x^2_{k=1,\alpha=0.05}}$ **= 3.841**

Since test statistic (x^2 = 1.0339) < critical value ($x^2_{k=1,\alpha=0.05}$ = 3.841), therefore fail to reject H0. There is enough evidence that there is no relationship between gender and status of placement.

3. Correlation

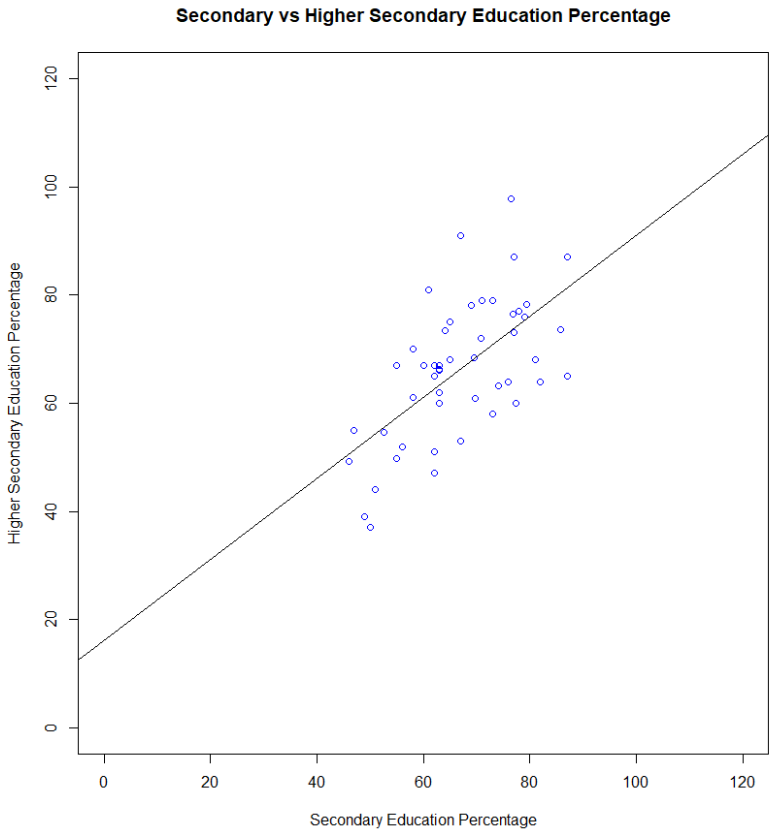


The scatter plot shows the relationship between the higher secondary education percentage against secondary education percentage. From the graph, we can see that both variables have a relationship since when secondary education percentage increases, the higher secondary education percentage will increase too, although there's some of the data that does not follow it.

```
cor (x, y)
[1] 0.6243862
```

From the above result, we can conclude that the correlation value is 0.6243862 and the strength of the linear relationship is moderate.

4. Regression




```
Coefficients:
(Intercept)            x
     16.083        0.749
```

From the scatter plot above, we can see that the type of regression model is positive linear relationship and based on the coefficients obtained, the estimated regression model is $\hat{y}=16.083+0.749x$.

Discussion

From this project, students can enhance their skills on analysing data using R Studio and learn more about how to conclude the analysed data based on the graph or results obtained. Although the process of this project is not that smooth sailing, we managed to overcome the issue and conclude the data collected. At first, there's some of the data that is not analysed properly and the results are not as expected, such as when we call the variables by using the wrong coding, the results would be wrong and the graph would be messed up. So, we need to be careful and choose the right ways to call or use variables so that we can avoid such situations. In my case, I've called the variables by using the wrong ways and the results won't come out while the program display error messages. So, after reviewing the tutorial notes, I've realized my mistakes and do the coding again. Fortunately, the problem was solved and I obtained the results for each tests of the data.

Conclusion

From the test done in this project, we can conclude that gender has no effect on the placement of students in this university. It shows that the probability of male and female to get placed in university is almost the same thus there is no relationship between gender and status of placement. Other than that, we can also conclude that the mean for degree percentage in the data collected is less than 2.41573 so the null hypothesis is rejected.