UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING

FACULTY OF ENGINEERING

SECI2143-04 PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2 : REPORT

COURSE : COMPUTER NETWORK AND SECURITY

COURSE CODE : 1SECR

SECTION : O4

TEAM MEMBERS :

NO	NAME	MATRIC CARD
1.	NOR FARAHZIBA BINTI HAMADUN	A19EC0123

LECTURER : DR SUHAILA MOHD YUSOF

DATE OF SUBMISSION : 27TH JUNE 2020.

Table Of Contents

I. Introduction

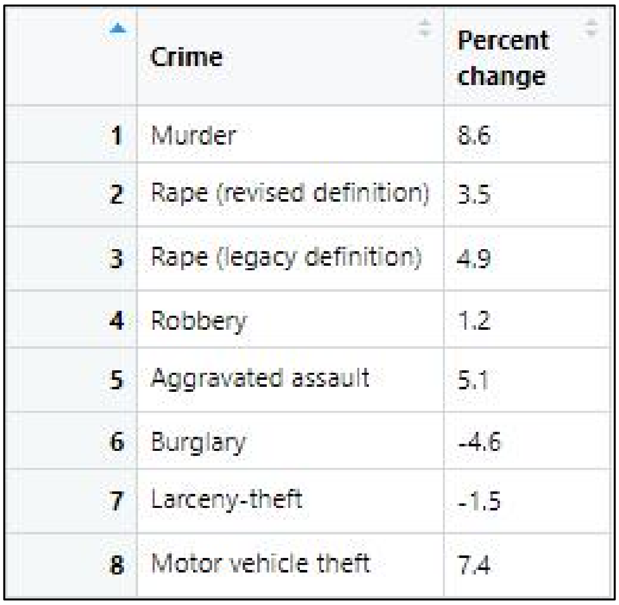
The data-set chosen by me retrieved from the FBI government portal with the title “2016 Crime in the United States”. However, the topic of the data-set is Crime in the United States by Region, Geographic Division, and State, 2015–2016. This data-set was collected by Uniform Crime Reporting (UCR). The purpose is to analyze the crime in the United States by Region, Geographic Division, and State in 2015–2016. The data-set provides the type of crime, the rate (per 100,000) of crime, and the percentage change between two years in each region, geographic division, and state in 2015 and 2016. Murder and non-negligent manslaughter, rape (revised definition), rape (legacy definition), robbery, and aggravated assault are categorized as violent crime. Meanwhile, burglary, larceny-theft, and motor vehicle theft are categorized as property crimes. The table also consists of the population of each area in the United State.

So, based on this data-set, I want to study the percentage change from 2015 to 2016 for the number of crimes in each region in the United States. Either its increase or decrease, all depend on the data-set provided and the test analysis. Hence, I need to analyze the data-set to find a conclusion for my case study. For the specification of the target population, I choose all ages, male and female, all-region in the United States, and all types of violent crime and property crime.

II. Hypothesis Testing

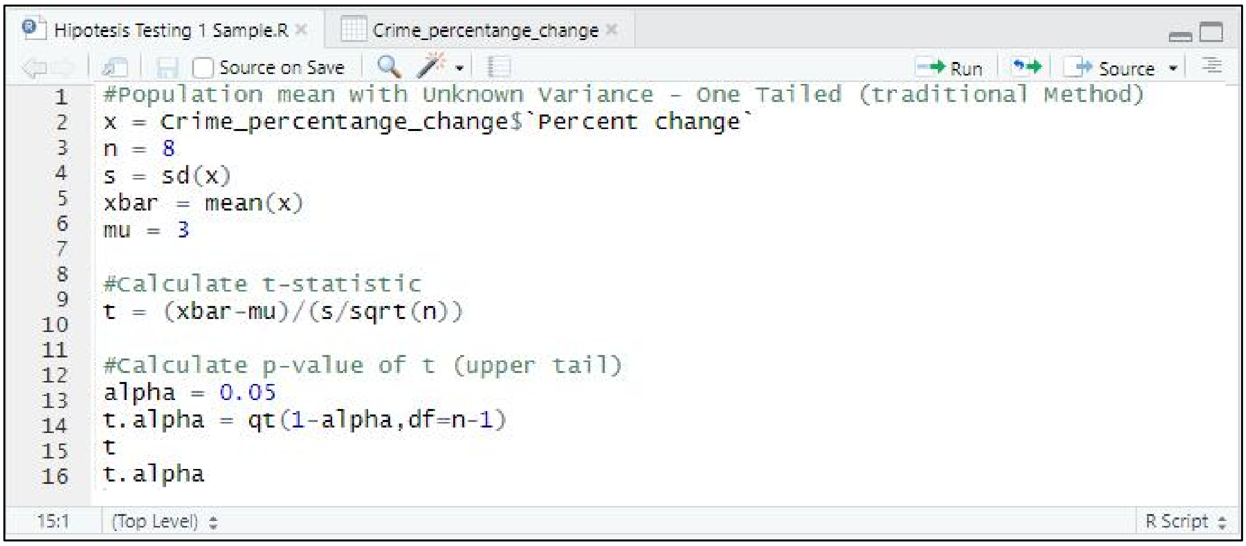
1. *Hypothesis Testing 1 Sample*

Given the sample size, n is 8, and the mean is 1000000 with unknown variance. Based on the data from Table 1, can we reject the null hypothesis that the mean of percentage change for each type of crime at 0.05 significant level?



	Crime	Percent change
1	Murder	8.6
2	Rape (revised definition)	3.5
3	Rape (legacy definition)	4.9
4	Robbery	1.2
5	Aggravated assault	5.1
6	Burglary	-4.6
7	Larceny-theft	-1.5
8	Motor vehicle theft	7.4

Table 1: Percentage Change For Each Type of Crime



```
#Population mean with Unknown Variance - One Tailed (traditional Method)
x = Crime_percentange_change$`Percent change`
n = 8
s = sd(x)
xbar = mean(x)
mu = 3

#Calculate t-statistic
t = (xbar-mu)/(s/sqrt(n))

#Calculate p-value of t (upper tail)
alpha = 0.05
t.alpha = qt(1-alpha,df=n-1)
t
t.alpha
```

Hypothesis Testing 1 Sample Coding Based On Table 1

H0: $\mu = 3$

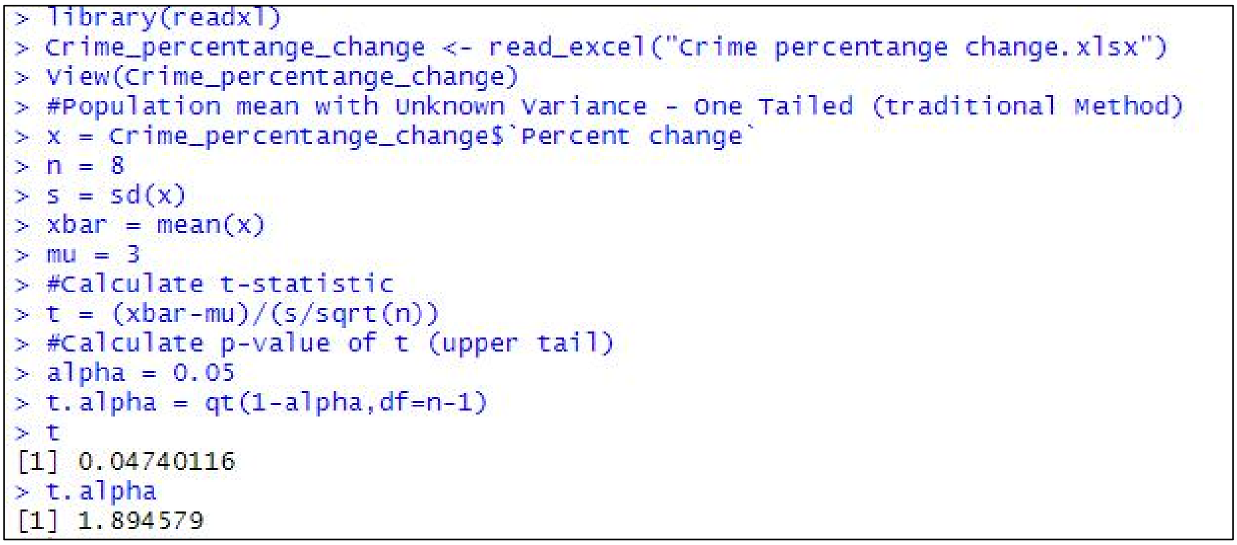
H1: $\mu > 3$

Critical value: $t_{(7,0.05)} = 1.894579$

Test Statistic: 0.04740116

Decision: Fail reject H0 (null hypothesis)

Conclusion: Since 0.04740116 (test statistic) < 1.894579 (critical value),its fail to reject H0. There is insufficient evidence that we can reject the null hypothesis that the mean of percentage change for each type of crime at 0.05 significant level.

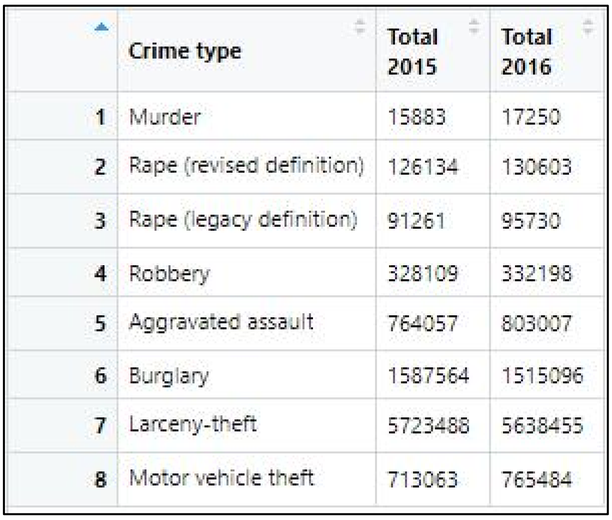


```
> library(readxl)
> Crime_percentange_change <- read_excel("Crime percentange change.xlsx")
> View(Crime_percentange_change)
> #Population mean with Unknown Variance - One Tailed (traditional Method)
> x = Crime_percentange_change$`Percent change`
> n = 8
> s = sd(x)
> xbar = mean(x)
> mu = 3
> #Calculate t-statistic
> t = (xbar-mu)/(s/sqrt(n))
> #Calculate p-value of t (upper tail)
> alpha = 0.05
> t.alpha = qt(1-alpha,df=n-1)
> t
[1] 0.04740116
> t.alpha
[1] 1.894579
```

Output of The Coding

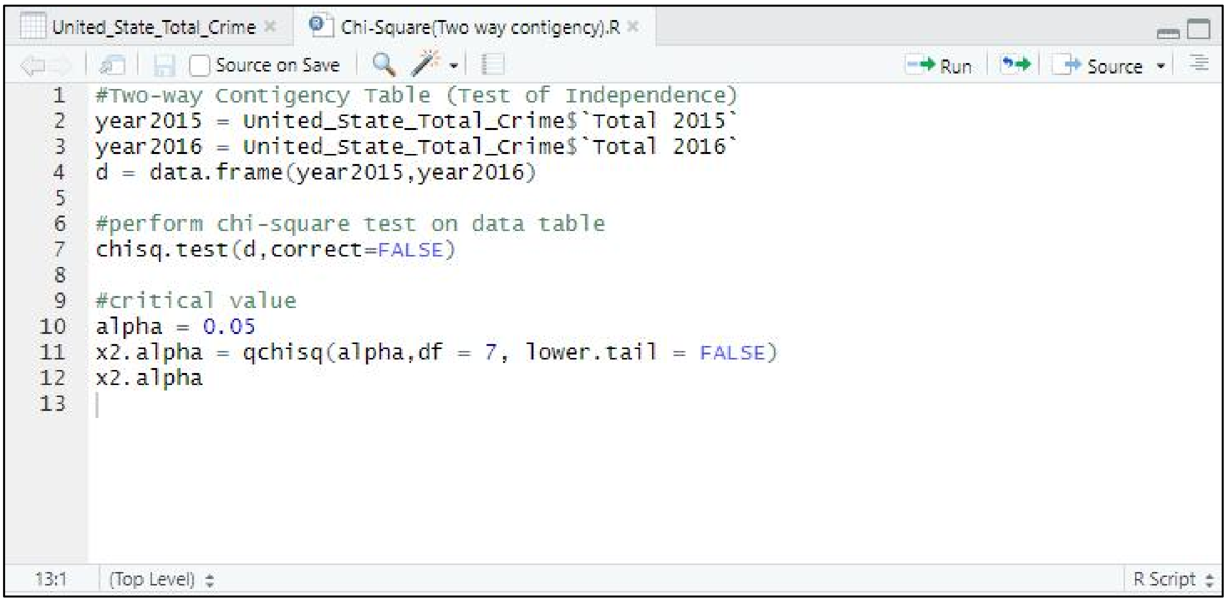
2. *Chi-Square*

From Table 2, we test the hypothesis whether the type of crime has a relationship with years at the 0.05 significance level.



	Crime type	Total 2015	Total 2016
1	Murder	15883	17250
2	Rape (revised definition)	126134	130603
3	Rape (legacy definition)	91261	95730
4	Robbery	328109	332198
5	Aggravated assault	764057	803007
6	Burglary	1587564	1515096
7	Larceny-theft	5723488	5638455
8	Motor vehicle theft	713063	765484

Table 2: Total number of offence for each type of crime in 2015 and 2016



```r
#Two-way Contigency Table (Test of Independence)
year2015 = United_State_Total_Crime$`Total 2015`
year2016 = United_State_Total_Crime$`Total 2016`
d = data.frame(year2015,year2016)

#perform chi-square test on data table
chisq.test(d,correct=FALSE)

#critical value
alpha = 0.05
x2.alpha = qchisq(alpha,df = 7, lower.tail = FALSE)
x2.alpha

```

Chi-Square (Two-way Contingency Table) Coding Based On Table 2

H0: Type of crime has no relationship with years

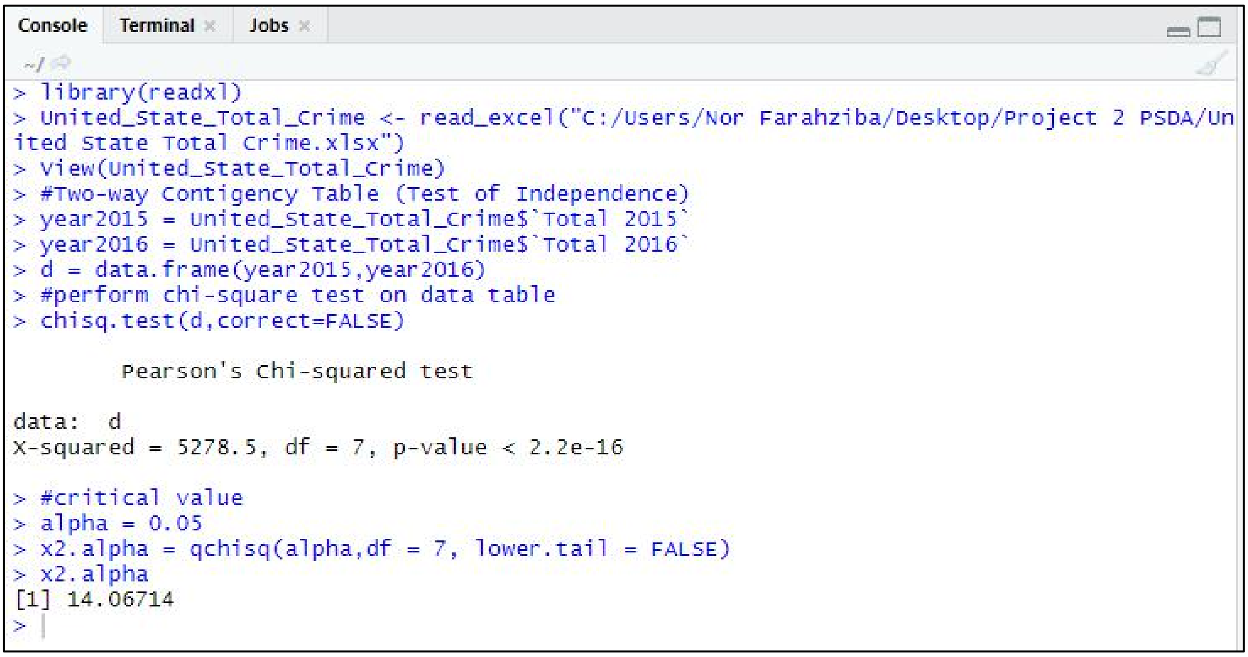
H1: Type of crime has relationship with years

Critical value: $\chi^2_{(7,0.05)} = 14.06714$

Test Statistic: 5278.5

Decision: Reject H0 (null hypothesis)

Conclusion: Since 5278.5 (test statistic) > 14.06714 (critical value), its reject H0. There is sufficient evidence to claim the type of crime has a relationship with years at the 0.05 significance level.

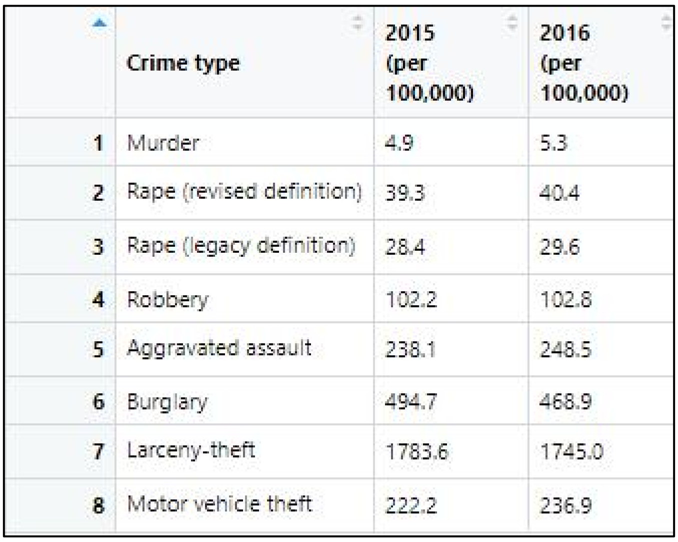


```
> library(readxl)
> United_State_Total_Crime <- read_excel("C:/Users/Nor Farahziba/Desktop/Project 2 PSDA/Un
ited State Total Crime.xlsx")
> View(United_State_Total_Crime)
> #Two-way Contigency Table (Test of Independence)
> year2015 = United_State_Total_Crime$`Total 2015`
> year2016 = United_State_Total_Crime$`Total 2016`
> d = data.frame(year2015,year2016)
> #perform chi-square test on data table
> chisq.test(d,correct=FALSE)

        Pearson's Chi-squared test

data:  d
X-squared = 5278.5, df = 7, p-value < 2.2e-16

> #critical value
> alpha = 0.05
> x2.alpha = qchisq(alpha,df = 7, lower.tail = FALSE)
> x2.alpha
[1] 14.06714
> 
```

Output of the coding

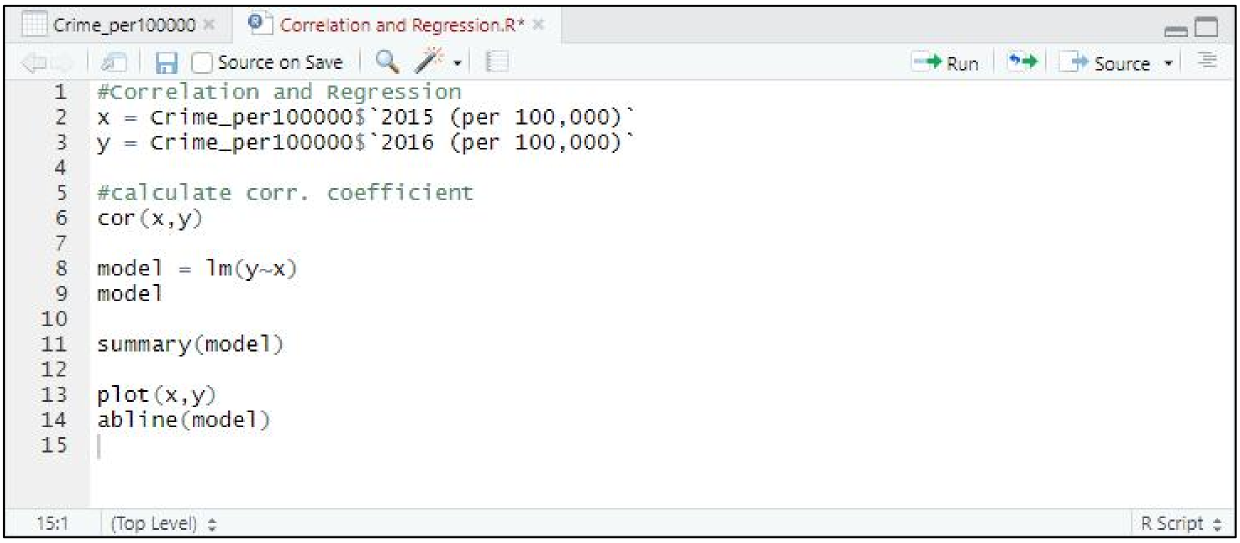
3. *Correlation and Regression*

The following data on the rate in 2015 and the rate in 2016 for the 8 types of crime committed. Test the claim that the crime rate in 2015 will affect the crime rate in 2016 by using α = 0.05.



	Crime type	2015 (per 100,000)	2016 (per 100,000)
1	Murder	4.9	5.3
2	Rape (revised definition)	39.3	40.4
3	Rape (legacy definition)	28.4	29.6
4	Robbery	102.2	102.8
5	Aggravated assault	238.1	248.5
6	Burglary	494.7	468.9
7	Larceny-theft	1783.6	1745.0
8	Motor vehicle theft	222.2	236.9

Table 3: Rate In 2015 And Rate In 2016 For Each Type Of Crime

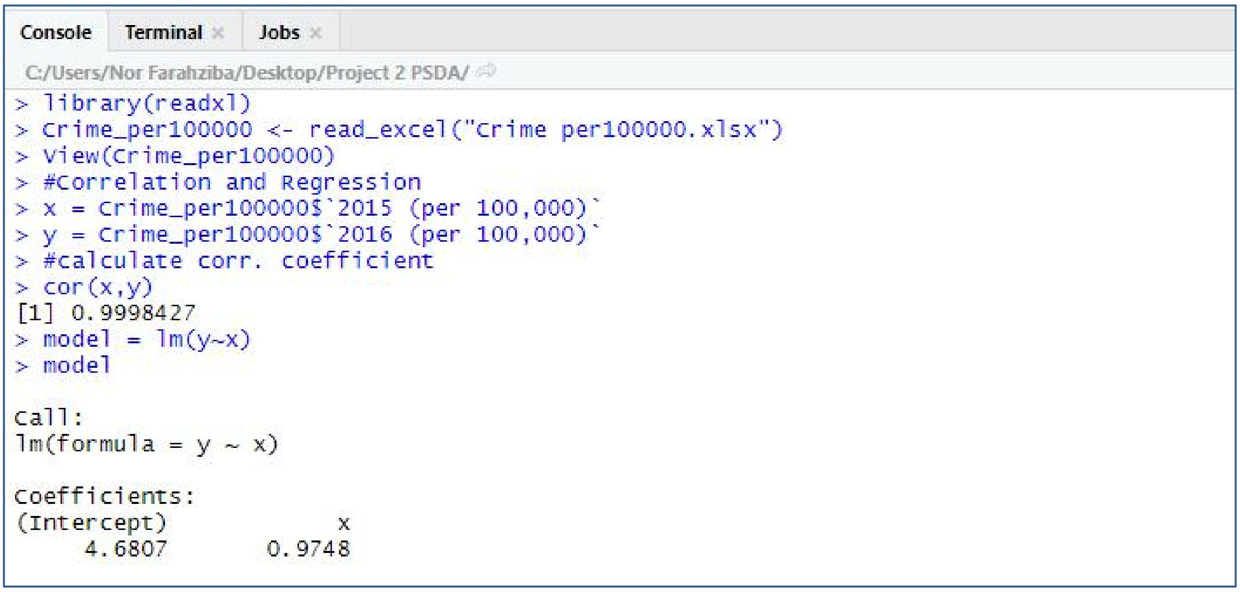


```
#Correlation and Regression
x = Crime_per100000$`2015 (per 100,000)`
y = Crime_per100000$`2016 (per 100,000)`

#calculate corr. coefficient
cor(x,y)

model = lm(y~x)
model

summary(model)

plot(x,y)
abline(model)

```

Correlation And Regression Coding Based On Table 3

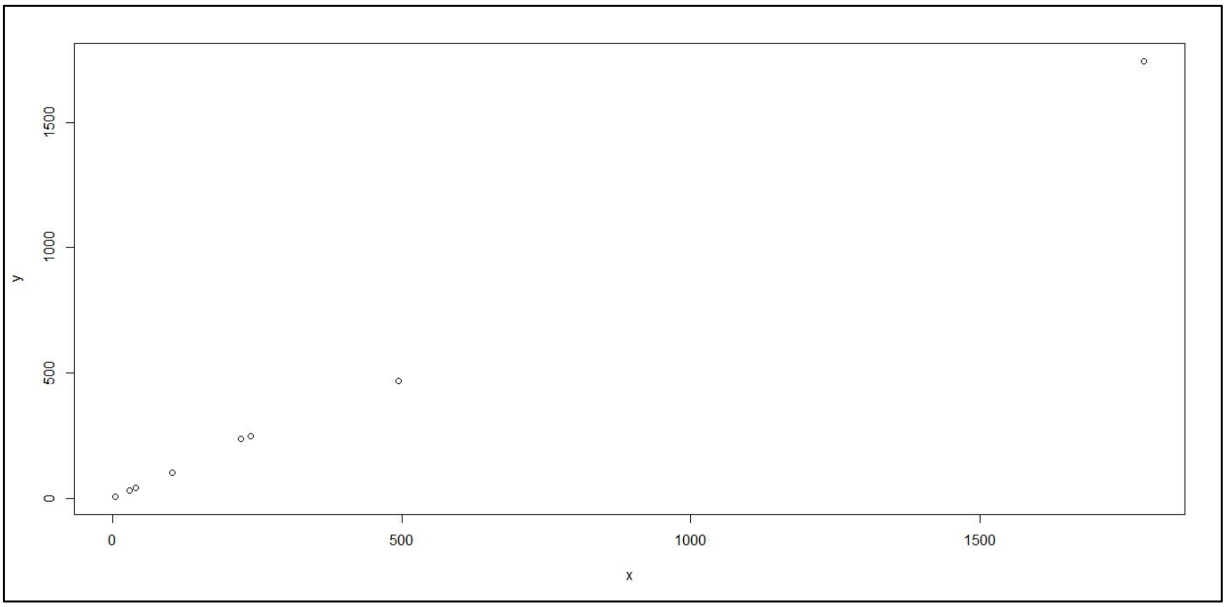
- **Correlation**

By using cor() function we can determine the correlation coefficient (r). Regarding on Table 3, the correlation coefficient is 0.9998427. So, the strength of the relationship between two variables (rate in 2015 and rate in 2016) is strong. The equation of the least-square regression line for rate in 2015 and rate in 2016 is $\hat{y} = 4.6807 + 0.9748x$.



```
> library(readxl)
> Crime_per100000 <- read_excel("Crime per100000.xlsx")
> View(Crime_per100000)
> #Correlation and Regression
> x = Crime_per100000$`2015 (per 100,000)`
> y = Crime_per100000$`2016 (per 100,000)`
> #calculate corr. coefficient
> cor(x,y)
[1] 0.9998427
> model = lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     4.6807       0.9748
```

Coding Output



The strength of relationship between two variables are strong and positive correlation (Correlation Plot)

- **Regression**

H0: β = 0

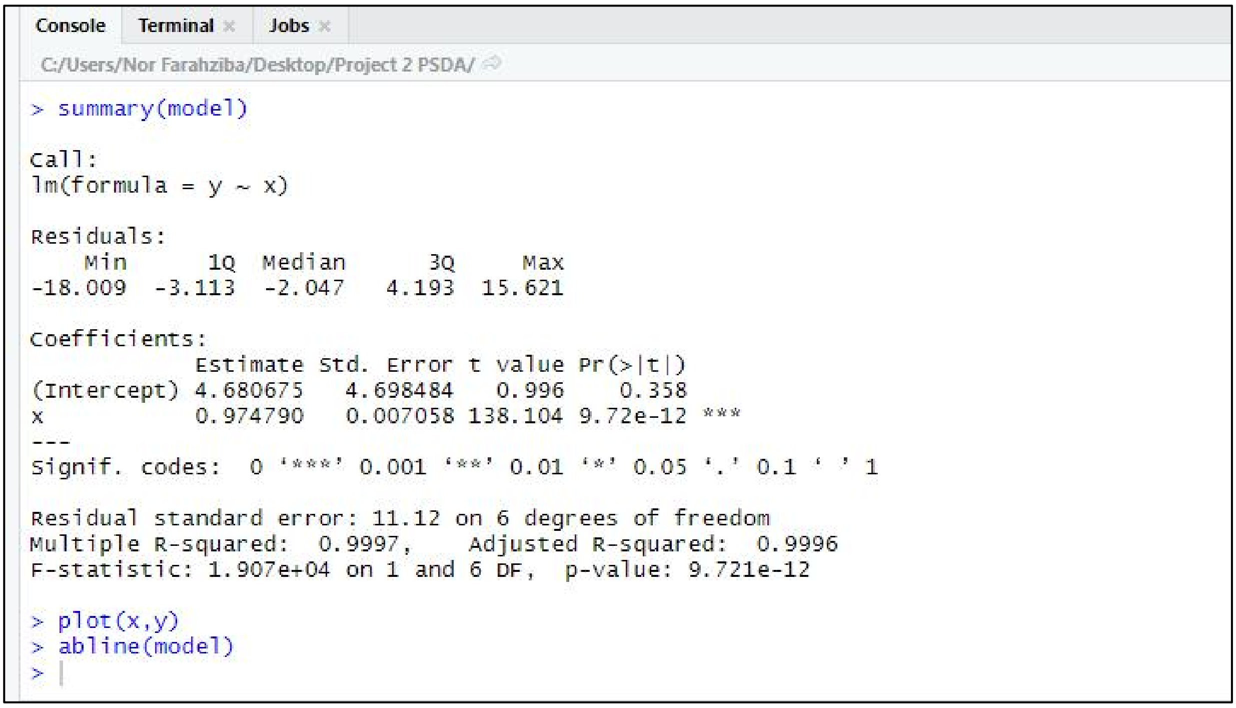
H1: β ≠ 0

Critical value: $t_{(6,0.05)} = \pm 2.447$

Test Statistic: 0.996

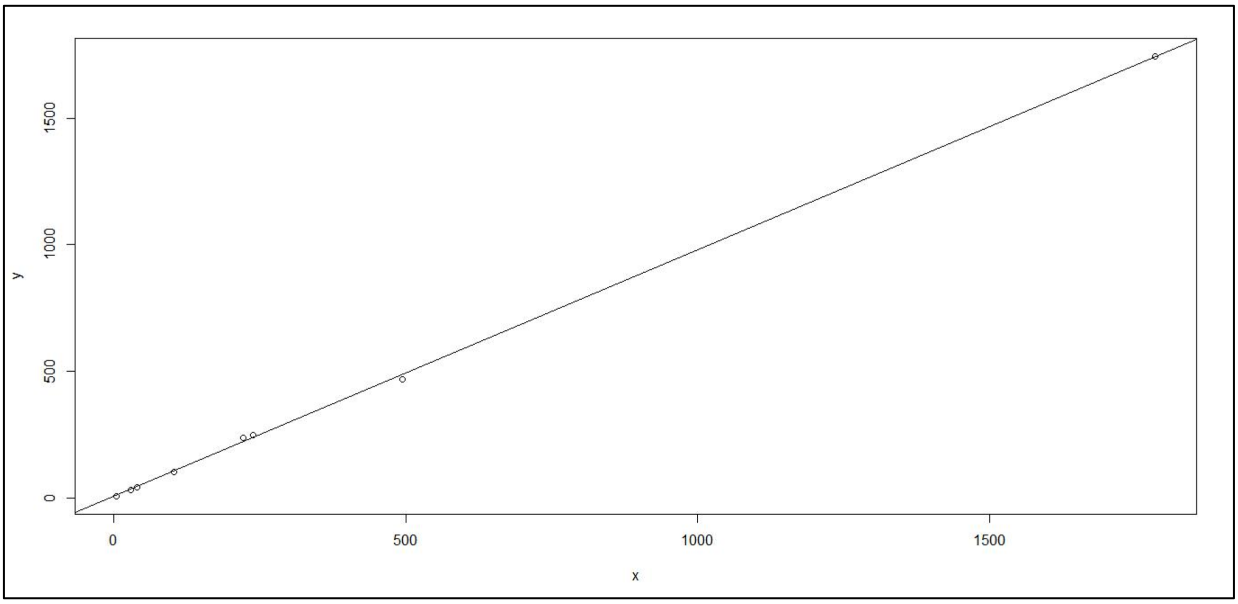
Decision: Fail to reject H0 (null hypothesis)

Conclusion: Since 0.996 (test statistic) < 2.447 (critical value),its fail to reject H0. There is insufficient evidence to conclude that the crime rate in 2015 affect the crime rate in 2016 at a significant level of 0.05.



```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.009  -3.113  -2.047   4.193  15.621 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.680675   4.698484   0.996    0.358    
x           0.974790   0.007058 138.104 9.72e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.12 on 6 degrees of freedom
Multiple R-squared:  0.9997,	Adjusted R-squared:  0.9996 
F-statistic: 1.907e+04 on 1 and 6 DF,  p-value: 9.721e-12

> plot(x,y)
> abline(model)
> 
```

Coding Output



Positive Linear Regression Relationship
(Regression Plot)

III. Discussion

Four types of tests were conducted to study the percentage change from 2015 to 2016 for the number of crimes in each region in the United States. The four tests are Hypothesis Testing 1 Sample, Chi-square, Correlation, and Regression test. The table for each test was created based on the data-set chosen regarding to my case study.

From the hypothesis testing 1 sample, its fail rejects the null hypothesis since the test statistic is smaller than the critical value. So, there is insufficient evidence that we can reject the null hypothesis that the mean of percentage change for each type of crime at 0.05 significant level.

For the chi-square test, since test statistic bigger critical value, its reject null hypothesis. Hence, there is sufficient evidence to claim the type of crime has a relationship with years at the 0.05 significance level.

In the correlation test, I found out the strength of the relationship between two variables (rate in 2015 and rate in 2016) is strong because the correlation coefficient calculated through the coding is 0.9998427. Besides, the equation of the least-square regression line for rate in 2015 and rate in 2016 is $\hat{y} = 4.6807 + 0.9748x$.

The last test is a regression test. From the test, its failure to reject H0 since test statistic smaller critical value. Therefore, there is insufficient evidence to conclude that the crime rate in 2015 affect the crime rate in 2016 at a significant level of 0.05.

IV. Conclusion

Referring to the test, we can conclude that the mean of percentage change for each type of crime is 3, which indicate total crime in United State increase by 3 from 2015 to 2016. Moreover, the type of crime has a relationship with years. As the years increase, the total crime in the United state will also increased. The strength of the relationship between the rate in 2015 and the rate in 2016 is strong. However, the crime rate in 2015 do not affect the crime rate in 2016.

Briefly, the purpose of this study achieved. From this project, I learn how to conduct the four tests which are Hypothesis Testing 1 Sample, Chi-Square, Correlation, and Regression using R language. Furthermore, I also learn to make a conclusion for each test handled in order to accomplish the purpose of the case study.

References

1.

Themes, &. (n.d.). *Explaining the lm() Summary in R*. Retrieved June 25, 2020, from http://www.learnbymarketing.com/tutorials/explaining-the-lm-summary-in-r/

2.

Stat Trek. (n.d.). *Hypothesis Test for Regression Slope*. Retrieved June 25, 2020, from https://stattrek.com/regression/slope-test.aspx