UTM
UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF SOCIAL SCIENCES
AND
HUMANITIES

SECI2143 (Probability and Statistical Data Analysis)

SEMESTER 2, 2019/2020

Final Project

Understanding and analyzing the Suicide problem in Japan

PSDA-SEC 05.

Hamzeh Wahed Bajbouj -A19EC4009.

Table of content:

1) Introduction…

There is no doubt that suicide is one of the biggest mistakes a person may do when he decides to end his life. Suicide is one of the major social issues of the modern time. There are many causes of suicide, including what may be psychological, or what may be due to financial pressures or psychological trauma. In this research, we will try to understand the nature of suicides and whether the suicide numbers differ between men and women. Are suicides increasing over the years, or not? Is there a direct link between suicides and the financial condition?

2) Why Japan?

Japan is one of the countries with the most suicide rates in the world, where it had the sixth highest suicide rate according OECD [1], Japan has a long history with suicide, since the suicide was in the past considered an honorable act for the warrior to do.

3) About the Data...

The data set used in this analysis project was downloaded from *Kaggle*. what is *Kaggle* !?!?Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals.

The dataset is derived from another large dataset (Suicide Rates Overview 1985 to 2016) which was collected by multiple organizations:

1. United Nations Development Program.
2. World Bank.
3. World Health Organization.

All these data that were collected for different purposes and then were combined together in one large dataset to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.

I used the dataset to analyze the nature of suicide in Japan as well as to linked it with the increasing of GDP per capita. From the original dataset I derived another small dataset that I used it in the analyzing of the correlation and regression.

This table shows the variables that is in the original data set.

Variable-Type	The variables names
Nominal	Age, Country, Gender, country-year, generation.
Ordinal	Null.
Interval	Year.
Ratio	suicides_no, population, suicides/100k pop, HDI for year, gdp_for_year ($), gdp_per_capita ($).

For the data set that I made from the original data set

Variable	Description
year_date	Show all years from 1985-2015,each in the data set will have all the data belongs to the year.
female_suicides_no	The sum for all suicides number for all female age groups in specific year.
male_suicides_no	The sum for all suicides number for all male age groups in specific year..
total_suicides_no	The sum for the male_suicides_no and female_suicides_no, so it's the total suicide number for all the year
gdp_per_capita	The GDP per capita for each year.
gdp_for_year	The GDP for year , for each year.

4) **What we aim to and our study cases.**

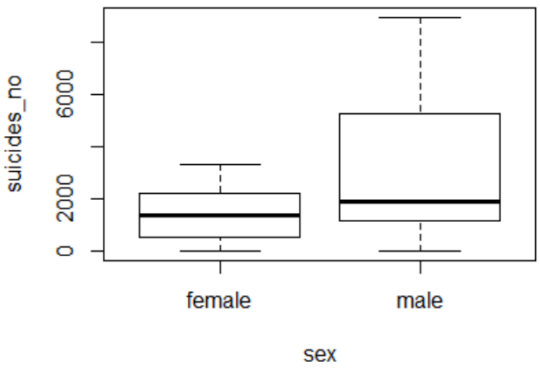
In the next statistical tests, we are aiming to test these claims.

1. To test the mean suicide number between females and males
2. Show the relationship between the time and suicides number for females.
3. Show the relationship between the time and suicides number for males.
4. Show the relationship between the time and total suicides number
5. Regression analysis between suicides number for females and GDP per capita
6. Regression analysis between suicides number for males and GDP per capita
7. ANOVA test, to see is the mean suicide differ based of the age

5) **Hypothesis two-sample test**

In this hypothesis test we want to know is there any difference in the suicide mean for men and women in Japan or not?

I used T statistical test to test the claim that the suicide mean for men and women is equal. Where I assumed that he samples variance is not equal based on two things.



From the boxplot we can see that the variance for each gender is not equal. As well I used (var) command to check from the sample variance

```
> var(suicides_no[sex=="male"])
[1] 7161685
> var(suicides_no[sex=="female"])
[1] 949492.6
```

My hypothesis:

#H0: mean suicide for male = suicide number mean for female

#H1: mean suicide for male != suicide number mean for female

With 95% confidence level

$H_0 : \mu_1 - \mu_2 = 0$

$H_A : \mu_1 - \mu_2 \neq 0$

```
        Welch Two Sample t-test

data:  suicides_no by sex
t = -7.8174, df = 233.21, p-value = 1.846e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2043.911 -1221.056
sample estimates:
mean in group female   mean in group male
            1352.849             2985.333
```

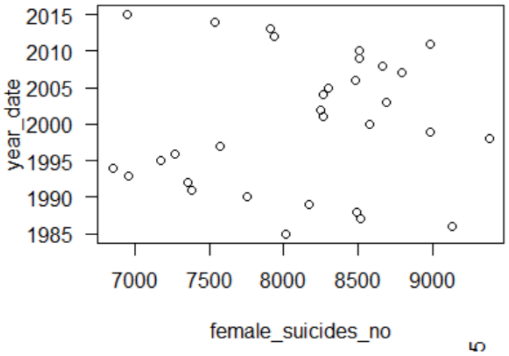
We will reject H_0 if t-value is < T(0.05,233)= **- 1.970198** or > T(0.05,233)= **+ 1.970198**

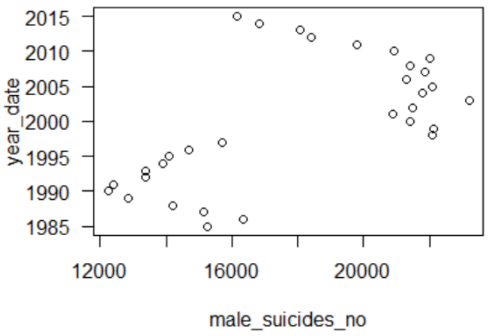
Which in our case T-value is less than -1.97, so now based the on previous test we know that the mean suicide is differ from men to women where is mean suicide for men is much larger than the women mean suicide, which means that most of the suicide cases are done by men.

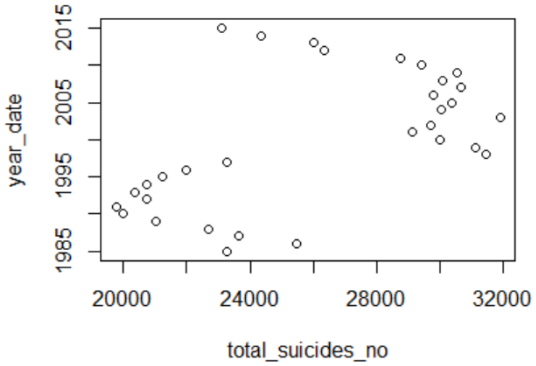
6) **Correlation and Regression**

In this part the purpose in understand more about the nature of the suicide where we are going to see and test if the suicide number are increasing with the time or not, and whether the GDP per capita has a direct effect on the suicide numbers.

In the first part we want to test the relationship between the time and the suicide number, the purpose is to check whether the suicide numbers are increasing each year of not.







For the first scatter plot which shows the association relationship between the female suicide number with time. where correlation coefficient value, r = 0.09017026 which means that the relationship between time and suicide number is very week, therefore we cannot be certain that the suicides numbers for females are increasing or decreasing each year.

To test this claim, we will assume that H0: ρ = 0 , by using the T-statistic test we will test that there is not (no linear correlation) between the years and the female suicide numbers. To reject H0 T-value should be > T(0.05,29) = **+2.04523** or < T(0.05,29) = **-2.04523**

```
        Pearson's product-moment correlation

data:  female_suicides_no and year_date
t = 0.48757, df = 29, p-value = 0.6295
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2728889  0.4307476
sample estimates:
       cor
0.09017026
```

Since T-value =0.48 which is < T(0.05,29) = **+2.04523** so we fail to reject H0 and there is not linear correlation between *years* and *female suicide number.*

For the second scatter plot the same idea as the first scatter plot where we want to see the linear correlation between years and male suicide numbers. where correlation coefficient value is r =

0.6106414 which mean there is moderate positive linear relationship between years and male suicide number, which we can say that the suicide numbers are increasing each year.to test that there is (linear relationship) we assumed that the

H0: ρ = 0 (no linear correlation) H1: ρ !=0 (there is a linear correlation).

```
        Pearson's product-moment correlation

data:  male_suicides_no and year_date
t = 4.1525, df = 29, p-value = 0.000264
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3270712 0.7933258
sample estimates:
      cor
0.6106414
```

To reject H0 T-value should be > T (0.05,29) = **+2.04523** or < T(0.05,29) = **-2.04523** which in our case T-value is > T(0.05,29) = **+2.04523** we reject H0 therefore our claim that the suicide numbers for males are increasing each year is true.

For the third scatter plot we want to see the linear correlation relationship between all the suicide cases for male and females and have an overview about the suicide numbers generally in Japan and to see are they in case of increase or decrease. our correlation coefficient value is r 0.5525227 which means there is indeed a linear correlation relationship between the *years* and *total suicide numbers* where the r shows that there is moderate positive linear relationship. Which means the number of the suicide cases are in increase.

To test our claim that there is linear correlation relationship

H0: ρ = 0 (no linear correlation)

H1: ρ !=0 (there is a linear correlation).

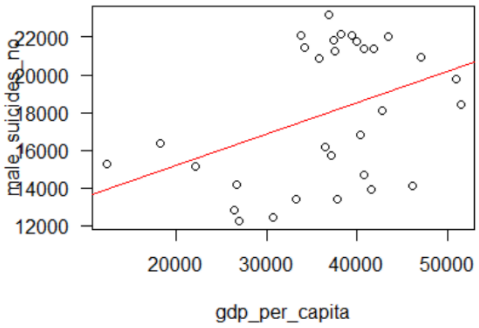
To reject H0 T-value should be > T (0.05,29) = **+2.04523** or < T(0.05,29) = **-2.04523**

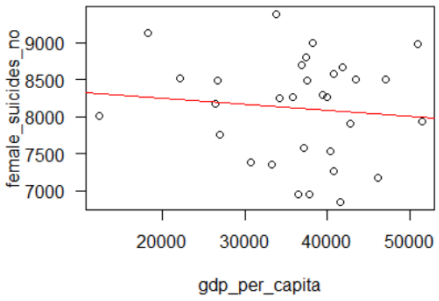
So since T-value should be > T (0.05,29) = **+2.04523** we reject H0 and our claim that there is linear correlation relationship is true.

```
        Pearson's product-moment correlation

data:  total_suicides_no and year_date
t = 3.5698, df = 29, p-value = 0.001268
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2464286 0.7583854
sample estimates:
      cor
0.5525227
```

For the regression analysis , I conducted two regression analyses to show does the GDP per capita has a direct affect on the suicide numbers for female and males ? where we wanted to predict the (y) values since we assumed that the changes in (y) are assumed to be caused by changes in (x), and our (y) for the first regression analysis is male suicide numbers with (x) which is gdp_per_capita, and in the second regression analysis changed (y) to be female suicide numbers.





In the first scatter plot the dependent variable is (male_suicide_no) and our independent variable is (gdp_per_capita).

My hypothesis:

The null hypothesis, Ho: $\beta 1 = 0$

The alternative hypothesis, H1: $\beta 1 \neq 0$

From the results we know that

Y= (11810.56) + (0.167) x

```
Call:
lm(formula = male_suicides_no ~ gdp_per_capita)

Residuals:
    Min      1Q  Median      3Q     Max
-5449.9 -2854.3  -378.5  3115.2  5189.4

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.181e+04  2.673e+03   4.419 0.000127 ***
gdp_per_capita 1.676e-01  7.145e-02   2.346 0.026013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3438 on 29 degrees of freedom
Multiple R-squared:  0.1595,    Adjusted R-squared:  0.1305
F-statistic: 5.505 on 1 and 29 DF,  p-value: 0.02601
```

The value of intersection coefficient (β_0) is (11810.56) which is the estimated average value of (Male suicide numbers) when the value of (GDP per capita) is zero.

whereas our value of slope coefficient (β_1) is (0.167) which is the estimated change in the average value of (Male suicide numbers) as a result of a one-unit change in of (GDP per capita).

Our Coefficient of Determination, R^2 = 0,1595 (R Squared is the square of the correlation coefficient = 0.399) so the It means that the predictors (x) explain 16.0% of the variance of Y, which indicates there is positive week relationship between the male suicide numbers and the GDP per capita, so therefor we conclude that, we cannot guarantee that the increasing of the GDP per Capita will decrease the suicide cases for males but it's the opposite.

For the second Scatter plot which shows the Linear Regression between (Female suicide numbers) and (GPD per capita)

```
Call:
lm(formula = female_suicides_no ~ gdp_per_capita)

Residuals:
    Min      1Q  Median      3Q     Max
-1223.3  -542.9   142.4   483.9  1241.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.412e+03  5.407e+02   15.56 1.31e-15 ***
gdp_per_capita -8.093e-03  1.445e-02   -0.56     0.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 695.6 on 29 degrees of freedom
Multiple R-squared:  0.01069,   Adjusted R-squared:  -0.02342
F-statistic: 0.3135 on 1 and 29 DF,  p-value: 0.5799
```

My hypothesis:

The null hypothesis, Ho: $\beta 1 = 0$

The alternative hypothesis, H1: $\beta 1 \neq 0$

From the results we know that

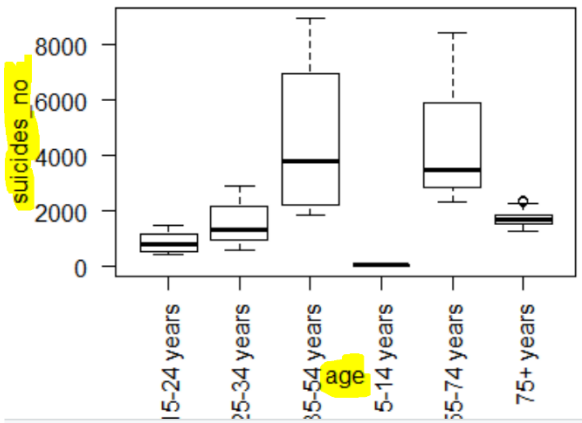
Y= (8412) + (-0.008093) x

The value of intersection coefficient, $\beta 0$ is (8412) which indicates the estimated average value of (Female suicide numbers) when the value of (GDP per capita) is zero, whereas the value of slope coefficient (β_1) is (-0.008093), which indicates the estimated change in the average value of (Female suicide numbers) as a result of a one-unit change in of (GDP per capita).

Our Coefficient of Determination, R^2 =0.01069 (R Squared is the square of the correlation coefficient = -0.1034) so the It means that the predictors (x) explain 1.1% of the variance of Y, whereas the (correlation coefficient) indicates there is a very week relationship between the Female suicide numbers and GPD per capita, since the slope coefficient is negative that means there is an inverse relationship, so when the GDP increase the female suicide numbers will decrease little bit, since the Coefficient of Determination is 0.01.

7) **ANOVA test**

In the dataset , the number of suicides was classified by age group, so I conducted a ANOVA test to check are all the mean suicide for all the age group are equal or not , the purpose from the test, is to prove that the mean is not equal, since by logic it's impossible the mean for all the age groups will be equal we cannot say that the mean suicide for age (age5-14 years) is equal with (age25-34 years), therefor I will use the ANOVE test to prove my claim.



In the dataset there are (six different age groups)

- 75+ years / 55-74 years
- 35-54 years / 25-34 years
- 15-24 years / 5-14 years

```
> summary(ANOVATEST)
             Df    Sum Sq   Mean Sq F value Pr(>F)
age           5 1.073e+09 214666272   116.4 <2e-16 ***
Residuals   366 6.751e+08   1844486
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The hypothesis:

H0: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5 = \mu 6$

H1: at least one mean is different.

From the test we know that F(5,366) =116.4 , Where as the P-value = (2*10^-16)

So to reject our Null hypothesis the F(5,366) should be > the F critical value for (5,366) which is 2.601, And in fact the value of our F staticst test is F(5,366) =116.4 which is > F critical value (5,366)=2.601.Which indecates that the mean suicide for all aga groups is not euqal.

8) **Conclusion…**

Understanding the nature of siucide may help us to decreae siucide rates, as well to find out the real issues that cause people to commit siucide, in this report we took a general overview about the suicide cases in one country, where I tried to do give a clear view about the suicide nature in Japan.

We conclude and proved that the suicde rates differe based on the geneder, where the most suicides were commited by men more than women, as well we prvoed a logiacal claim that suicides commited is different based on the age group. Where

We also prvoen that there is no a strong direct affect from GPD per capita on the suicide rates for both genders as the most people think that the GPD can increase the suicide rate if the GPD per capita decreased, so therefor the people in Japan don't commit suicide due to finaical issues.

9) **Appendix**

[1] https://data.oecd.org/healthstat/suicide-rates.htm

THE DATASET USED: https://www.kaggle.com/yeasin3437/japan-suicide-rates-overview-1985-to-2015/data