UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143

PROBABILITY & STATISTICAL DATA ANALYSIS

Section 05 2020/2021 Session

Project 2

Lecturer : **Dr. Nor Azizah Ali**

PREPARE BY **Group Kekw**

Name	Matric No.
Uh Guan Yong	**A20EC0230**
Tie Sing Hao	**A20EC0168**
Samuel Luk Kie Liang	**A20EC0224**
Mohamed Yasser Elrefaey	**A20EC9106**

DATE SUBMISSION:

3 JULY 2021

Table of Contents

1.0 Introduction and Background

The topic of this project is the World Happiness Report in 2021. The aim of the study is to investigate which of the six factors which are economic production, social support, life expectancy, freedom, absence of corruption, and generosity which influence more to making happiness.

We were interested in this question because we want to know why the happiest country is happy and want to understand which factor has the greatest impact on happiness. We were also divide the European countries into countries of the European Union and non-countries of the European Union to see if there is a difference between average happiness score between them using a t-test of 5% significance level.

2.0 Dataset

The main issue we want to study is the happiness index, in which the rankings of national happiness are based on a Cantril Ladder survey. Nationally representative samples of respondents are asked to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale. The report correlates the life evaluation results with various life factors. The first factor we considered is logged Gross domestic product per person (GDP per capita) measures the sum of marketed goods and services produced within the national boundary, averaged across everyone who lives within this territory. GDP per capita is calculated using a country's GDP in 2012 United States dollars (USD) which is then divided by the country's total population. The second variable is social support, which refers to the psychological and material resources provided by a social network to help individuals cope with stress. Social support can be measured as the perception that one has assistance available, the actual received assistance, or the degree to which a person is integrated into a social network. The third parameter is healthy life expectancy, which is defined as the average number of years that a person can expect to live in "full health" by taking into account years lived in less than full health due to disease and/or injury. The fourth factor is the freedom to make life choices, which is the national average of responses to the question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?". The fifth variable is generosity, which is the residual of regressing the national average of response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita. The last parameter is the perception of corruption, which measure is the national average of the survey responses to two questions in the GWP: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?". The possible outcomes we can predict is when the parameters of logged GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity increase, the happiness score will be also getting increased except for the parameter of corruption perceptions, which we believe it contributes negatively on the happiness score.

3.0 Data Analysis and Results

Hypothesis Testing 2-Sample

The number of EU countries is 27 while the number of Non-EU countries is 11. Non equality of variance will be assumed for both categories.

u1: The mean of happiness score for European Union countries.

u2: The mean of happiness score for Non- European Union countries.

H0: μ1=μ2

H1: μ1≠μ2

According to the t-test using R and also by using the following calculations:

$$\bar{X}_1 \approx 5.6851$$

$$\bar{X}_2 \approx 6.444$$

$$S_{X_1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2 \approx 1.0177$$

$$S_{X_2}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{2i} - \bar{X}_2\right)^2 \approx 0.5331$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{x}_2}} = \frac{6.444 - 5.6851}{0.2708} \approx 2.8026$$

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}} \approx 0.2708$$

d. o. $$f = \frac{\left(\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}\right)^2}{\frac{\left(\frac{X_1}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{X_2}{n_2}\right)^2}{n_2 - 1}} \approx 31$$

Since t = 2.803097>$t_{0.531}$ = 2.04 is therefore rejected, we reject the null hypothesis

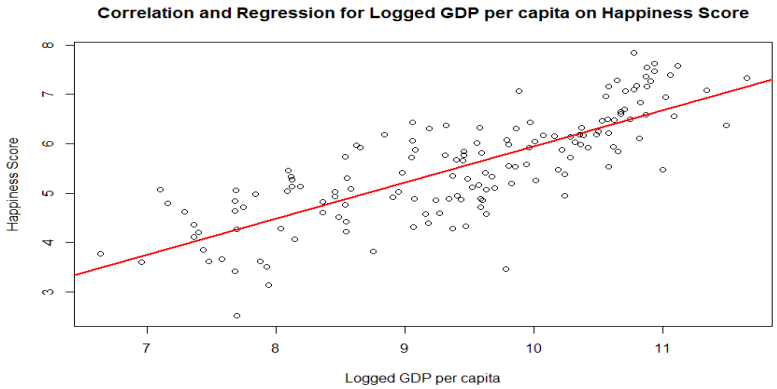
The absolute value of the calculated t exceeds the critical value, so the averages are significant otherwise, and we have evidence to conclude that the average of happiness achieves for EU countries is different from the average of happiness for non-EU countries.

Although both the categorized countries are located in Europe, but they do not enjoy the same level of happiness and this can be due to many factors such as GDP per capita, freedom ... etc.

Correlation and Regression Test

1. Dependent variable, y: Happiness score

Independent variable, x: Logged GDP per capita



Strength of correlation coefficient (r) is moderate which is: 0.7897597

```
> cor(data$`Ladder score`,data$`Logged GDP per capita`)
[1] 0.7897597
```

Relationship between x and y is described by a positive linear function which when x increases, y also increases.

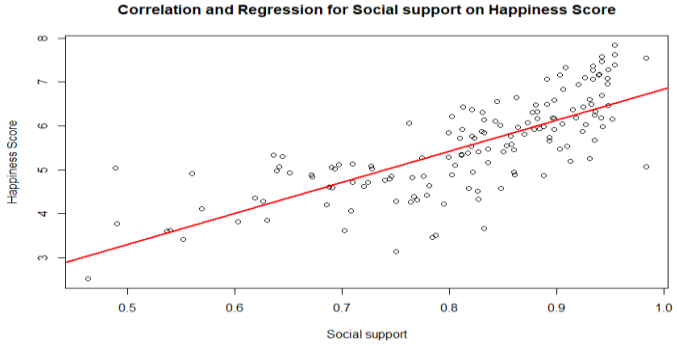
```
Call:
lm(formula = data$`Ladder score` ~ data$`Logged GDP per capita`)

Coefficients:
                  (Intercept)  data$`Logged GDP per capita`
                       -1.372                         0.732
```

The regression equation is: $\hat{y} = -1.372 + 0.732x$

2. Dependent variable, y: Happiness score

Independent variable, x: Social support



Strength of correlation coefficient (r) is moderate which is: 0.7568876

```
> cor(data$`Ladder score`,data$`Social support`)
[1] 0.7568876
```

Relationship between x and y is described by a positive linear function which when x increases, y also increases.

```
Call:
lm(formula = data$`Ladder score` ~ data$`Social support`)

Coefficients:
          (Intercept)  data$`Social support`
              -0.2315                 7.0750
```

The regression equation is: $\hat{y} = -0.2315 + 7.0750x$

3. Dependent variable, y: Happiness score

Independent variable, x: Healthy life expectancy

Strength of correlation coefficient (r) is moderate which is: 0.7680995

```
> cor(data$`Ladder score`,data$`Healthy life expectancy`)
[1] 0.7680995
```

Relationship between x and y is described by a positive linear function which when x increases, y also increases.

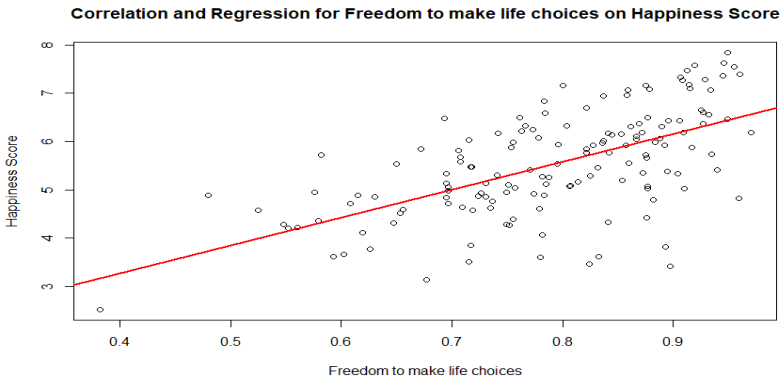
```
Call:
lm(formula = data$`Ladder score` ~ data$`Healthy life expectancy`)

Coefficients:
                    (Intercept)  data$`Healthy life expectancy`
                         -2.395                           0.122
```

The regression equation is: $\hat{y} = -2.395 + 0.122x$

4. Dependent variable, y: Happiness score

Independent variable, x: Freedom to make life choices



Strength of correlation coefficient (r) is moderate which is: 0.6077531

```
> cor(data$`Ladder score`,data$`Freedom to make life choices`)
[1] 0.6077531
```

Relationship between x and y is described by a positive linear function which when x increases, y also increases.

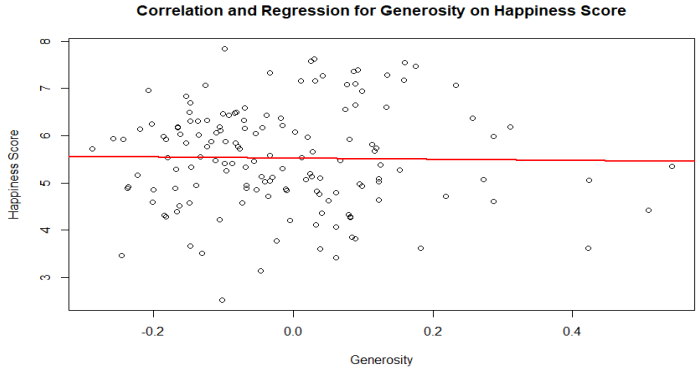
```
Call:
lm(formula = data$`Ladder score` ~ data$`Freedom to make life choices`)

Coefficients:
                          (Intercept)  data$`Freedom to make life choices`
                                0.974                                5.759
```

The regression equation is: $\hat{y} = 0.974 + 5.759x$

5. Dependent variable, y: Happiness score

Independent variable, x: Generosity



Strength of correlation coefficient (r) is weak which is: -0.01779928

```
> cor(data$`Ladder score`,data$Generosity)
[1] -0.01779928
```

There is no relationship between x and y.

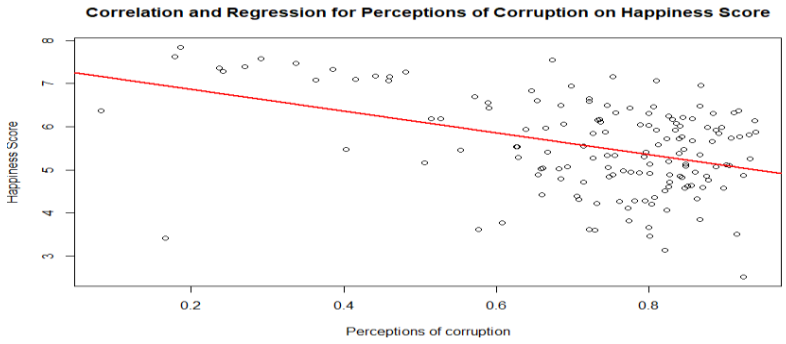
```
Call:
lm(formula = data$`Ladder score` ~ data$Generosity)

Coefficients:
    (Intercept)  data$Generosity
         5.5309          -0.1269
```

The regression equation is: $\hat{y} = 5.5309 + -0.1269x$

6. Dependent variable, y: Happiness score

Independent variable, x: Perception of corruption



Strength of correlation coefficient (r) is weak which is: -0.42114

```
> cor(data$`Ladder score`,data$`Perceptions of corruption`)
[1] -0.42114
```

Relationship between x and y is described by a negative linear function which when x increases, y decreases.

```
Call:
lm(formula = data$`Ladder score` ~ data$`Perceptions of corruption`)

Coefficients:
                    (Intercept)  data$`Perceptions of corruption`
                          7.369                            -2.523
```

The regression equation is: $\hat{y} = 7.369 + -2.523x$

Summary of multiple linear regression:

```
Call:
lm(formula = data$`Ladder score` ~ data$`Logged GDP per capita` +
    data$`Social support` + data$`Healthy life expectancy` +
    data$`Freedom to make life choices` + data$Generosity + data$`Perceptions of corruption`)

Residuals:
     Min       1Q   Median       3Q      Max
-1.85049 -0.30026  0.05735  0.33368  1.04878

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -2.23722    0.63049  -3.548 0.000526 ***
data$`Logged GDP per capita`      0.27953    0.08684   3.219 0.001595 **
data$`Social support`             2.47621    0.66822   3.706 0.000301 ***
data$`Healthy life expectancy`    0.03031    0.01333   2.274 0.024494 *
data$`Freedom to make life choices`  2.01046    0.49480   4.063 7.98e-05 ***
data$Generosity                   0.36438    0.32121   1.134 0.258541
data$`Perceptions of corruption`  -0.60509    0.29051  -2.083 0.039058 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5417 on 142 degrees of freedom
Multiple R-squared:  0.7558,    Adjusted R-squared:  0.7455
F-statistic: 73.27 on 6 and 142 DF,  p-value: < 2.2e-16
```

Multiple regression equation:

Happiness score = -2.23722 + 0.27953 (Logged GDP per capita) + 2.47621 (Social support) + 0.03031(Healthy life expectancy) + 2.01046 (Freedom to make life choices) + 0.36438 (Generosity) – 0.60509 (Perception of corruption)

The estimated mean happiness score is -2.23722 when logged GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption are 0. Multiple $R^2 = 0.7558$, approximately 75.58% of variation in happiness score

of different countries can be explained by variation in logged GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption.

Avona Test

Variables used: Logged GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity, Perceptions of corruption

i. Hypothesis statement

null hypothesis H_0: $\mu1 = \mu2 = \mu3 = \mu4 = \mu5 = \mu6$

All the means are the same

Alternative Hypothesis: H_1: At least one mean is different

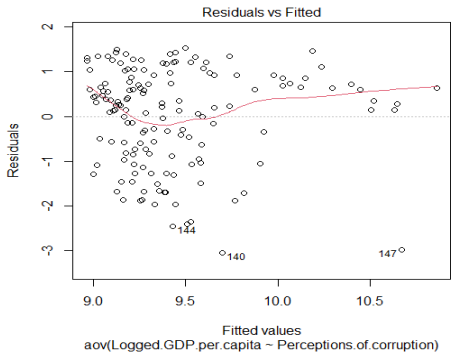
ii. Avona test

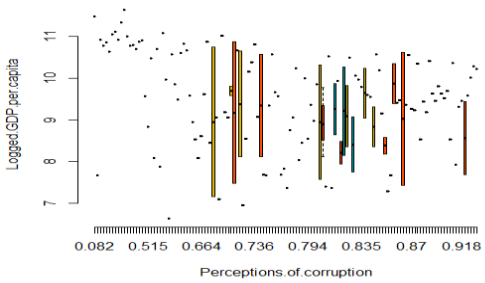
```
> summary(Anova1)
                            Df Sum Sq Mean Sq F value   Pr(>F)
Perceptions.of.corruption   1  23.28  23.283   19.52 1.92e-05 ***
Residuals                 147 175.39   1.193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>     )
```

F-statistic = 19.52

P-value = 1.92e-05

iii. Boxplot of the data:





iv. Result: Since the F test statistic = 19.52 > P-value = 1.92e-05. H_0 is rejected. There is at least one mean is not equal. There is sufficient evidence to claim that the different type of data have the different mean of ladder score

4.0 Discussion and Conclusion

After the hypothesis test, we can conclude that the average of happiness achieves for EU countries is different from the average of happiness for non-EU countries. After that, from the correlation test, we can conclude that the relationship between economic production and happiness is strongest among the 6 factors. Furthermore, from the regression test, we can conclude that the relationships between the dependent variable (Happiness score) and independent variables (logged GDP per capita, social support, healthy life expectancy, freedom to make life choices) are positively linear and the relationship between happiness score and perceptions of corruption is negatively linear. There is no relationship between happiness score and generosity. Lastly, from the ANOVA test we can conclude that different type of data have the different mean of ladder score.

5.0 References

https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021