MATH1324 Introduction to Statistics Assignment 3

Does a higher GDP imply a country’s happiness? : 2019 Review

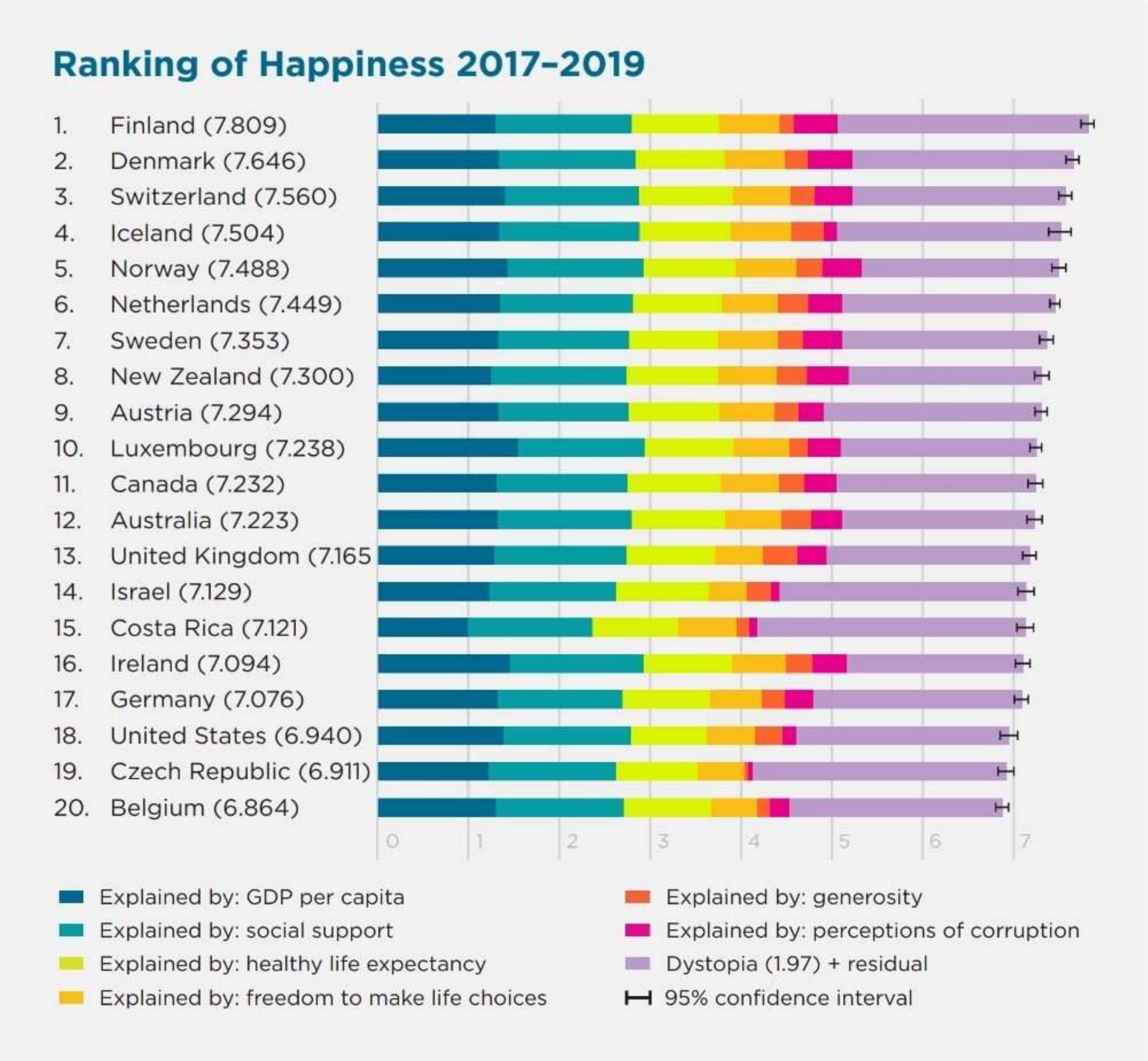
Anwesha Dutta (s3790886)

Last updated: 2022-06-10

RPubs LINK INFORMATION

This presentation is available online for viewing on RPubs by going to the URL below :

http://rpubs.com/data_pojoy/happiness19



Top 20 Happy Nations (*click for source image*)

INTRODUCTION

This year, the 7th Annual **World Happiness Report** was released on March 20 on the occasion of International Day of Happiness by the Sustainable Development Solutions Network of the United Nations in collaboration with Gallup and the Ernesto Illy Foundation.

- It is a landmark survey of the state of global happiness that ranks 156 countries by how happy their citizens perceive themselves to be, on a scale of zero to 10, from the worst possible life to the best possible life.
- The Happiness Index ranks countries on the basis of questions from the Gallup World poll and the results are correlated to income, healthy life longevity, generosity, freedom, social support and absence of corruption.

While GDP per capita remains to be a deciding statistic for the Happiness Index, questions still arise: does increased GDP indicate a higher average happiness for the nations? Even when people are relatively better off than before, why has their happiness not increased? Clearly the correlation between happiness and wealth is subject to examination. This very question is explored in this report based on statistical investigation using linear regression model and efficient data visualization in R.

PROBLEM STATEMENT

Question: Can GDP per capita of a country be used to determine the happiness of its people ?

Methods:

- Linear regression model is applied to examine the relationship between two selected variables.
- Hypothesis testing is performed for the overall linear reggresion model and parameters (α and β).
- Assumptions for independence, linearity, normality of residuals, homoscedasticity are made.
- Pearson correlation coefficient is applied to measure the strength of the linear relationship.

Open dataset used : https://s3.amazonaws.com/happiness-report/2019/Chapter2OnlineData.xls

ANALYSIS OF DATA SOURCE

1. UNSDSN : The information was gathered from a single direct primary source: The World Happiness Report produced by the United Nations Sustainable Development Solutions Network (UNSDSN) in partnership with the Ernesto Illy Foundation. While the possibility of a bias is always possible, these companies likely stand to profit more from accurate international information pertaining to people's desires, so it seems very unlikely that an agenda is being pushed meaning the data is very likely accurate.

2. Gallup, Inc.: Gallup, Inc. is an American analytics and advisory company known for its public opinion polls conducted worldwide. The happiness score was received from Gallup's data, which was done by performing randomised phone surveys (that were still representative of the overall demographics of the country above 15 years old) in countries where that was a proven method. In countries where it wasn't or where phone lines weren't readily available country-wide, face to face surveys were conducted. Phone surveys were 15-30 minutes, face to face ones 30-60 minutes usually.The information is unclear on how many people were interviewed.

- The dataset consists of 156 observations (representing 156 countries) and 11 variables.
- Variables *GDP per capita*, *Social support*, *Healthy life expectancy*, *Freedom to make life choices*, *Generosity*, *Perceptions of corruption* and *Dystopia* are taken from the population (by survey) to determine a nation's happiness score.

EXPLORING DATASET

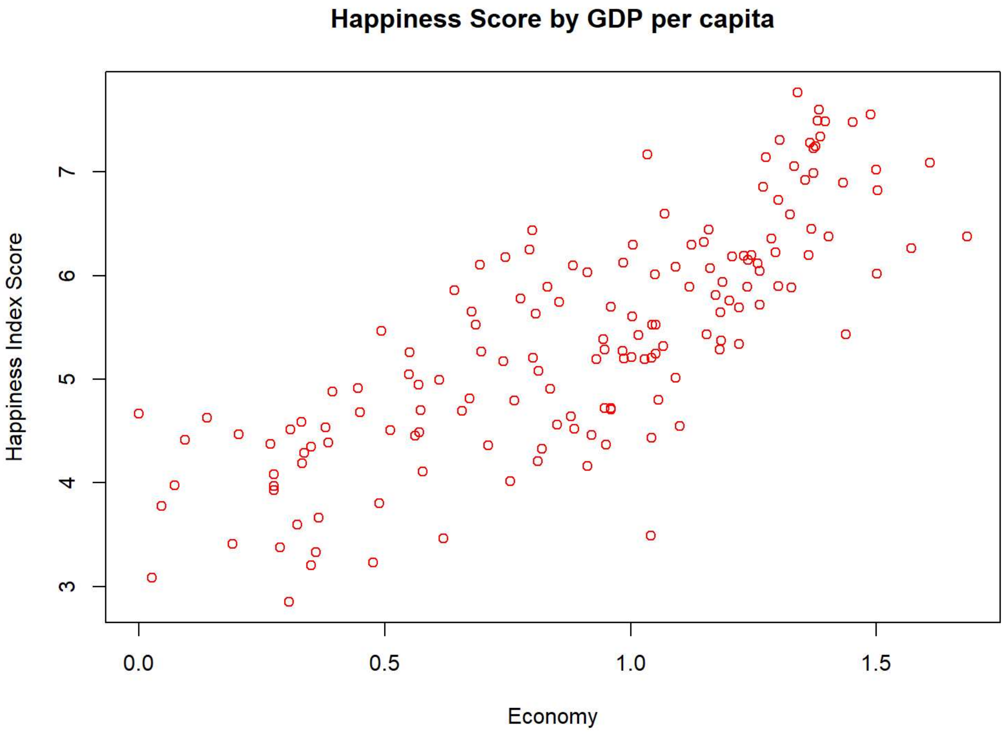
- *Happiness Score* : How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest?
- *Economy* : GDP per capita
- Variables *Country* ,*Happiness score* and *Explained by: GDP per Capita* are subset to a new data frame and column names are changed to **Happiness Score** and **Economy** to make the data easier to read.

```
#setwd("~/Project") #set working directory

#read data and make columns readable
Happiness_Index_19 <- read_excel ("Chapter2OnlineData.xls", sheet = 2)
names(Happiness_Index_19) <-make.names(names(Happiness_Index_19),unique = TRUE)

happiness<- Happiness_Index_19[,c(1,2,6)] #subset the data

colnames(happiness)[c(2,3)]<- c("Happiness_Score", "Economy") #change column names
head(happiness)
```

INITIAL VISUALISATION

Scatter plot is used to visualise the relationship between Economy and Happiness Score. It depicts that as GDP increases, the Happiness Score also increases, indicating a positive relationship.

```
happiness %>% plot(Happiness_Score ~ Economy, data=., col="red",
                   main = "Happiness Score by GDP per capita",
                   xlab = "Economy",ylab = "Happiness Index Score")
```



DESCRIPTIVE STATISTICS

Best line of fit : $Happiness\ Score = 3.3991 + 2.2185 * Economy$

```
#use lm() function to fit the linear regression model
happiness_model<- lm(Happiness_Score ~ Economy, data=happiness)
happiness_model %>% summary()
```

```
##
## Call:
## lm(formula = Happiness_Score ~ Economy, data = happiness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22004 -0.48410  0.00838  0.48443  1.47382
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3991     0.1353   25.12   <2e-16 ***
## Economy       2.2185     0.1369   16.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.679 on 154 degrees of freedom
## Multiple R-squared:  0.6303, Adjusted R-squared:  0.6279
## F-statistic: 262.6 on 1 and 154 DF,  p-value: < 2.2e-16
```

HYPOTHESIS TESTING : F-TEST FOR OVERALL MODEL

$$H_0 : Data\ does\ not\ fit\ the\ linear\ regression\ model$$

$$H_A : Data\ fits\ the\ linear\ regression\ model$$

```
happiness_model %>% summary() %>% coef()
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 3.399088  0.1353126 25.12027 2.438723e-56
## Economy     2.218513  0.1369037 16.20491 4.234917e-35
```

- We found p< 0.001. So, we reject H_0, as the p-value is less than the 0.05 level of significance. There is statistically significant evidence that the data fits a linear regression model.
- R^2-statistic is reported from the model as **0.6303**.**63%** of variability in the happiness score can be explained by a linear relationship with economy. Higher value of R^2 reflects the goodness of fit for linear regression, i.e., the line fits the data good.

HYPOTHESIS TESTING : MODEL PARAMETERS (α)

$$Null\ Hypothesis:\ H_0: \alpha = 0$$

$$Alternate\ Hypothesis:\ H_A: \alpha \neq 0$$

- The intercept is reported as **3.3991**. This hypothesis is tested using a t statistic, reported as **t = 25.12, p<0.001**.
- The constant is statistically significant at 0.05 significance level. So, there is statistically siginificant evidence that the constant is not 0.

In order to confirm that p < 0.001, 95% CI is calculated for α by using `confint()` function:

```
happiness_model %>% confint()
```

```
##                2.5 %   97.5 %
## (Intercept) 3.131780 3.666397
## Economy     1.948061 2.488964
```

- 95% CI for α is reported to be **[3.132, 3.666]**.
- $H_0: \alpha = 0$ is clearly not captured by this interval, therefore, it is rejected.

HYPOTHESIS TESTING : MODEL PARAMETERS (β)

$$Null\ Hypothesis:\ H_0 : \beta = 0$$

$$Alternate\ Hypothesis:\ H_A : \beta \neq 0$$

- Slope of the regression line was reported as $\beta = 2.2185$. This means that one unit increase in GDP is related to an average increase in happiness score of **2.2185** units, i.e., a positive change.
- The slope is tested using a t statistic, reported as **t = 16.205, p<0.001**.
- Since the slope is statistically significant at 0.05 level, there is statistically siginificant evidence that $\beta \neq 0$.
- `confint()` function used in previous slide shows the 95% CI for slope to be **[1.948, 2.489]**. As the 95% CI does not capture H_0, it is rejected.

ASSUMPTIONS

1. **Independence :** All the survey results are independent of each other and there are no multiple measurements for the same country.

2. **Linearity :** We have checked and confirmed the linearity in the begining of the report by using scatter plot. The scatter plot shows a positive relationship : as GDP increases, so does the happiness scores.

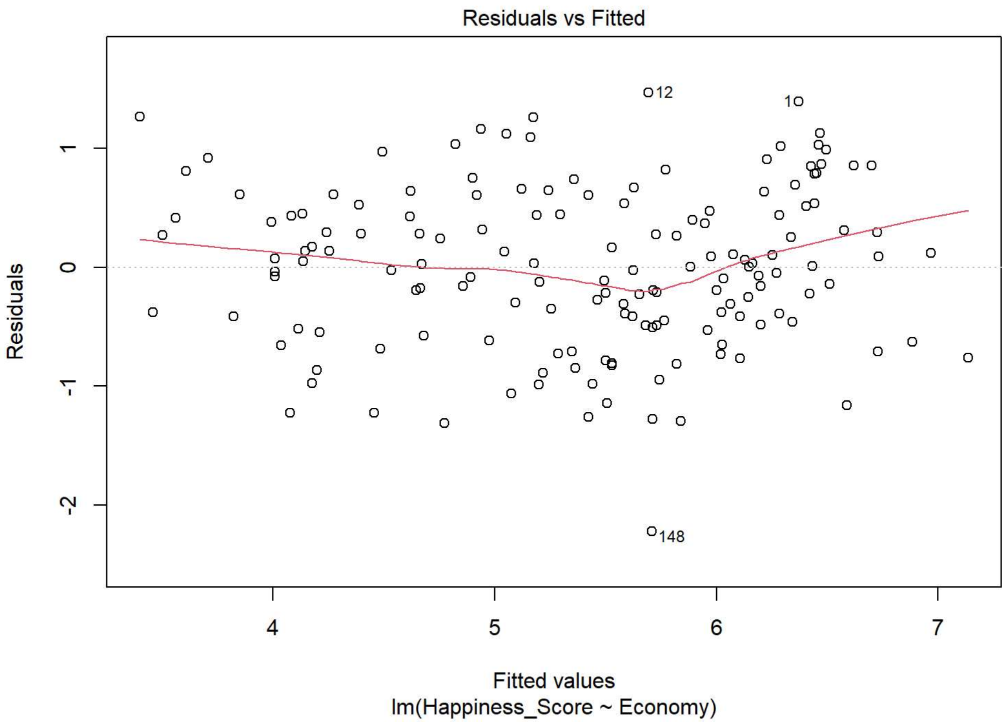
3. **Normality of residuals and Homoscedasticity :** The `plot()` function is used to obtain a series of plots for checking the diagnostics of a fitted regression model in the following slides.

TESTING ASSUMPTIONS : RESIDUALS VS. FITTED

- The trend-line between fitted values and residuals is flat, which is a good indication that we are modelling a linear relationship.

- The variability on y-axis is constant across the range of values on the x-axis and there is no distinct pattern in variablitiy, therefore, this is a sign of homoscedasticity or constant variance.

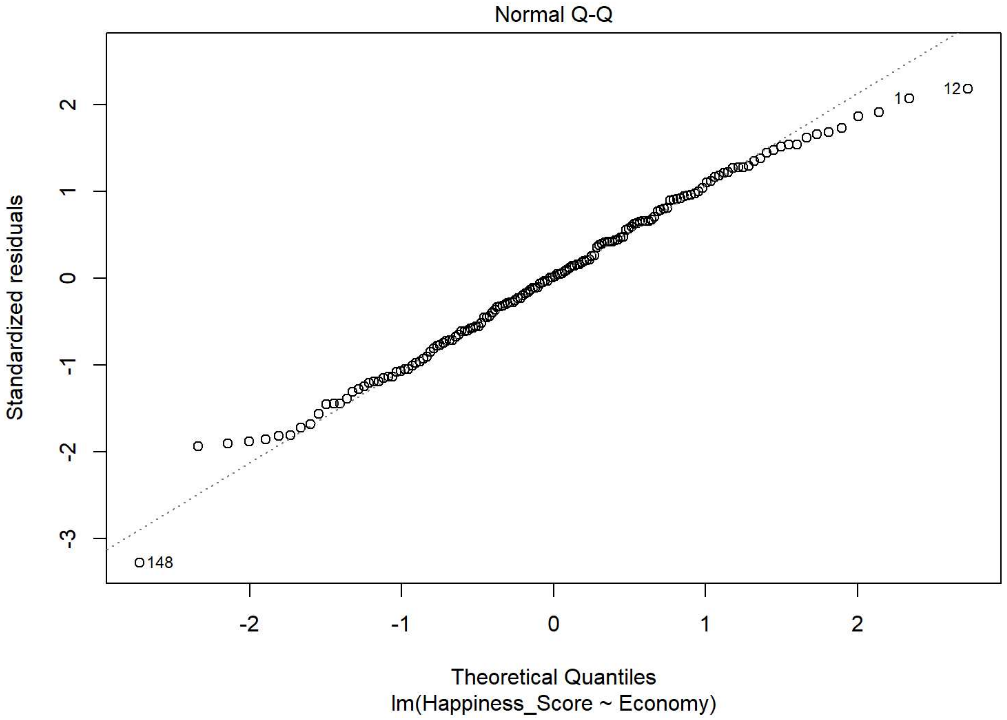
```
happiness_model %>% plot(which =1)
```



TESTING ASSUMPTIONS : NORMAL Q-Q

- The residuals fall close to the line.
- There are no major deviations from normality. So it is safe to assume that the residuals are approximately normally distributed.

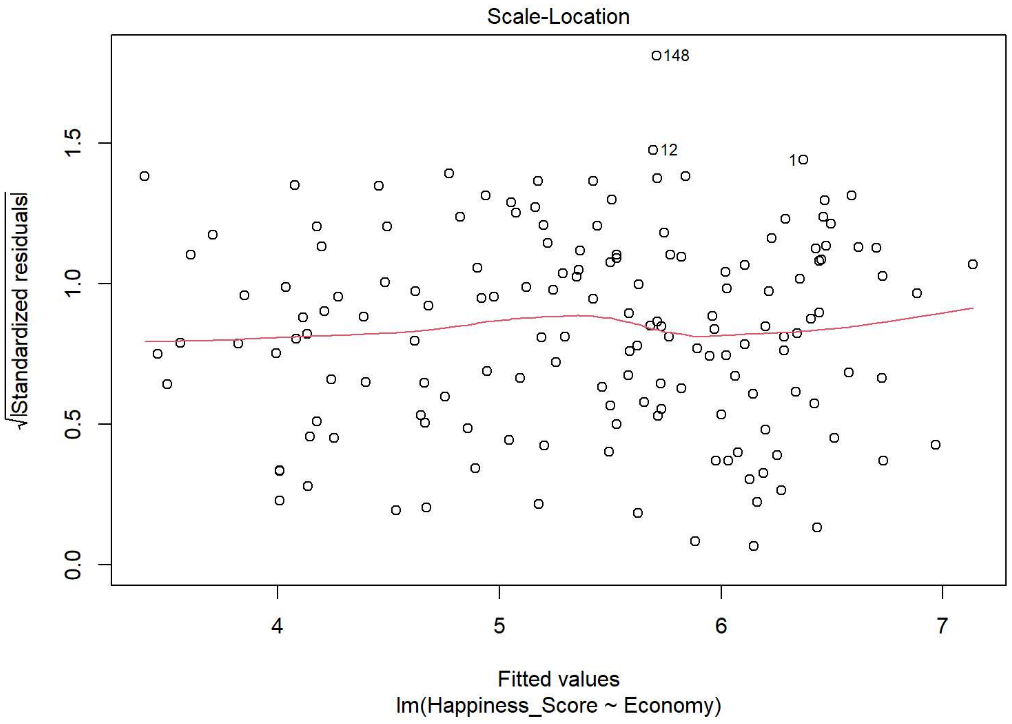
```
happiness_model %>% plot(which =2)
```



TESTING ASSUMPTIONS : SCALE-LOCATION

- The red line in the plot below is close to flat.
- The variance in the square root of the standardised residuals is consistent across the fitted values. Therefore, this is a sign of homoscedasticity.

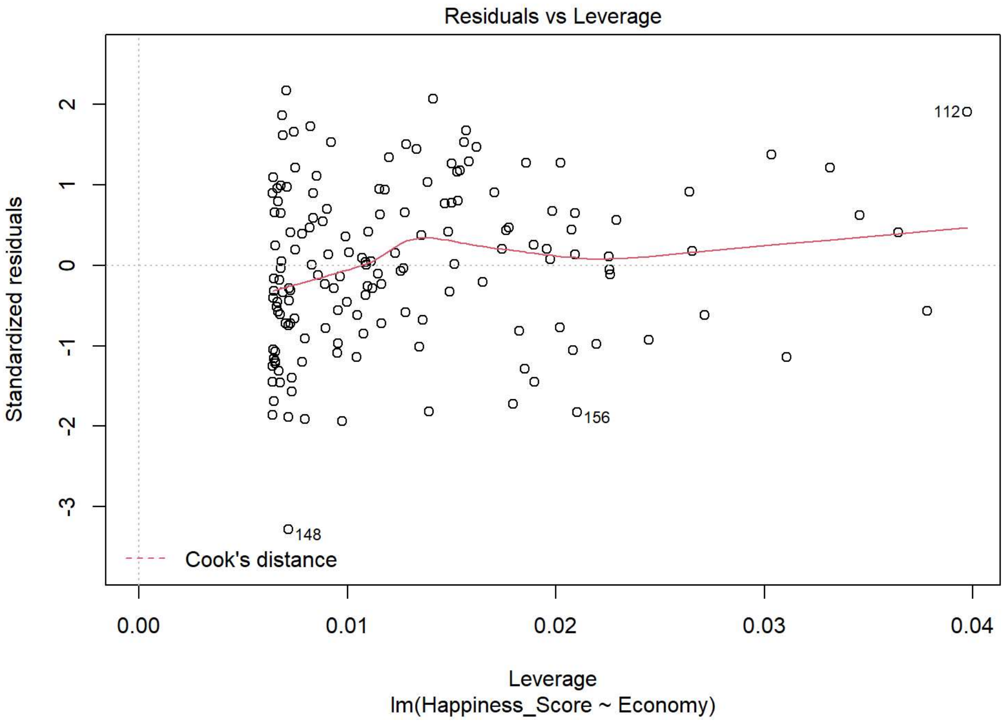
```
happiness_model %>% plot(which =3)
```



TESTING ASSUMPTIONS : RESIDUAL vs. LEVERAGE

- There are no values that fall in the upper and lower right hand side of the plot beyond the red bands based on Cook's distances. Notably, the bands are not even visible in the plot below. So, there is no evidence of influential cases.

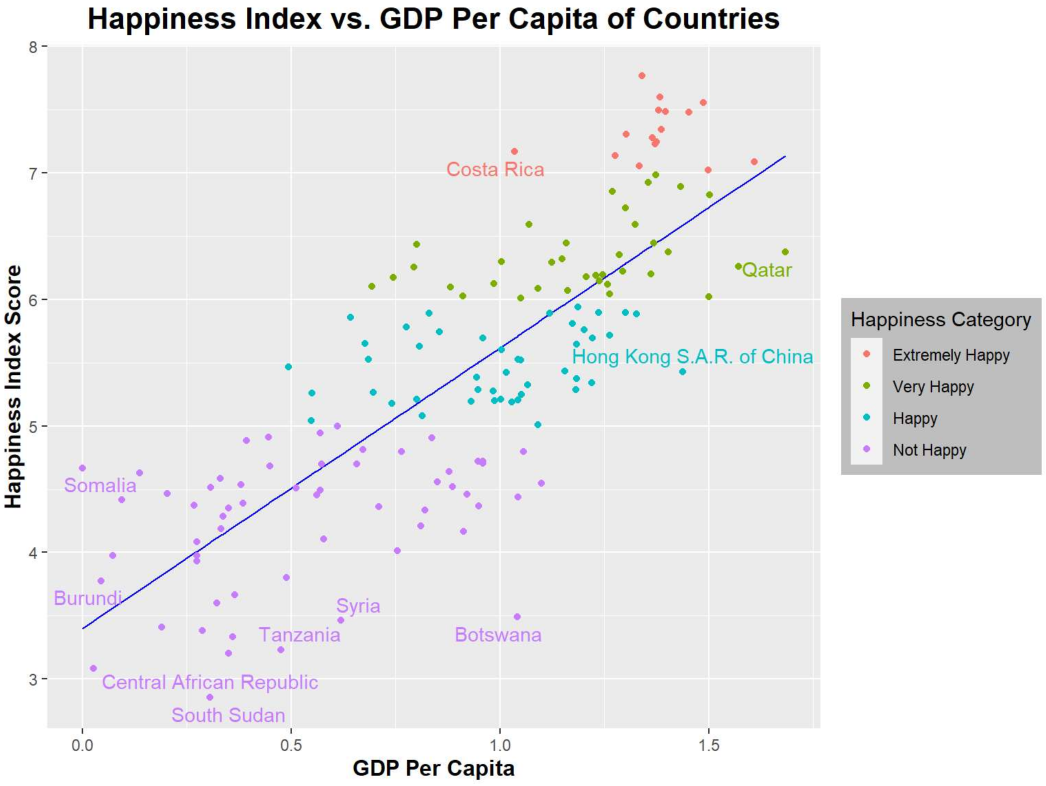
```
happiness_model %>% plot(which =5)
```



FINAL VISUALISATION

There is a statistically significant positive relationship between GDP per capita and happiness score for a nation. The linear relationship can be visualised using a plot as follows:

```
for (i in 1:length(happiness$Country)){
happiness$Category <-
with (happiness, ifelse(Happiness_Score >= 7, "Extremely Happy",
ifelse(Happiness_Score >= 6 & Happiness_Score < 7,"Very Happy",
ifelse(Happiness_Score >= 5 & Happiness_Score < 6,"Happy","Not Happy")))) }

happiness$Category <- factor(happiness$Category,
levels = c("Extremely Happy","Very Happy","Happy","Not Happy"))

#colour points by happiness category
ggplot(data=happiness, aes(Economy, Happiness_Score, label= happiness$Country, color= happiness$Category)) + scale_color_discrete(name="Happiness
         Category") +

#add linear regression line
geom_line(data = fortify(happiness_model), aes(x = Economy, y = .fitted), col="blue") +

#add a layer for the plot type (points)
geom_point() + theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid")) +

# label the outliers
stat_dens2d_filter(geom = "text_repel", keep.fraction = 0.06, nudge_x=0.06, nudge_y=0.3) +

#add title , label axes
labs(title="Happiness Index vs. GDP Per Capita of Countries",
       y = "Happiness Index Score", x = "GDP Per Capita") +

#format title and axes labels for better readability
theme(plot.title = element_text(color="black", size=16, face="bold", hjust = 0.5),
        axis.title.x = element_text( size=12, face="bold"),
        axis.title.y = element_text( size=12, face="bold"))
```



INTERPRETATION

Findings :

- Independence and linearity was assumed, normality of residuals and homoscedasticity were maintained with no influential cases.
- $R^2 = 0.63$, i.e., 63% variability in **Happiness Score**
- F-statistic $F(1, 154) = 262.6, p < 0.001$
- $\alpha = 3.3991$, $p < 0.001$, 95% CI **[3.132, 3.666]** ; $\beta = 2.2185, p < 0.001$, 95% CI **[1.948, 2.489]**

Decision :

- Overall model: Reject H_0
- Intercept α : Reject H_0
- Slope β : Reject H_0
- Estimated regression equation is : $Happiness\ Score = 3.3991 + 2.2185 * Economy$

DISCUSSION

- There was a statistically significant positive linear relationship between **Happiness Score** and **Economy**. Happiness score was estimated to explain up to 63% of the variability in economy. So, the estimated linear regression model can be used to predict the level of happiness of the countries that are not included in this report and compare the predicted happiness score with the actual happiness score.

- Referring to the outliers, it can be noticed that countries having high GDP's like Qatar (**1.684**) and Hong Kong (**1.438**) don't qualify among the happiest nations (i.e., Extremely Happy category). Whereas, a less affluent country like Costa Rica with a much lesser GDP of **1.034** appears to be much happier indicating that wealth is not primarily an indication of happiness.

Strengths: Choices of significant variables for simple linear regression model.

Limitation: Only one variable (**Economy**) from the dataset is used to investigate whether the happiness score of each country can be determined.

Directions for future investigations :

- The simple regression model should further be used for comparision in 2-3 years.

- The happiness score is recommended for determining the happiness level within a nation instead of comparing with other nations.

REFERENCES

1. Summary from : https://en.wikipedia.org/wiki/World_Happiness_Report

2. Original dataset retrieved from **Chapter2: Online Data** of *The World Happiness Report 2019* : https://s3.amazonaws.com/happiness-report/2019/Chapter2OnlineData.xls

3. Complete *World Happiness Report 2019* can be found here