Eitan and Yuval Bakirov

–June 2022–

Introduction

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012 and since then every year these reports have been distributed regularly.
The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

World Happiness Report - Website (https://worldhappiness.report/)

Goals

Our goal is to see the correlation between the happiness of countries to other statistical data, such as, GDP per capita of each country, its' healthy life expectancy, generosity etc.

In our project we will focus on:

- Tidy our data set
- Visualizations
- Statistical Models and methods learned during the course

The methods which we will use in this research are:

1. Hypothesis test - difference in means - we want to test the assumption that the median happiness score went up over the years. For this reason we will perform a T-Test.
2. Model of multiple regression - we want to examine the effect of explanatory variables on the level of happiness (the explained variable). We will perform tests to draw conclusions using Summary Statistics Table.

Part One - Data Import And Tidying

Data import

```
library(tidyverse)
library(gridExtra)
library(ggplot2)
library(dplyr)
library(corrplot)
library(ggcorrplot)
library(car)
library(DT)
library(rio)


whr2015 <- read_csv("2015.csv")
whr2019 <- read_csv("2019.csv")
```

Data Tidy:

We will start by arranging the tables so that they will be easy to read.

Delete irrelevant data columns and countries that do not appear in one of the tables.

(All the data is taken from Kaggle - World Happiness Report - Kaggle (https://www.kaggle.com/datasets/unsdsn/world-happiness))

```
colnames(whr2015) <- c('Country', 'Region', 'CUT' , 'Happiness_Score', 'CUT', 'GDP_per_capita', 'CUT',
                       'Life_Expectency', 'Freedom','Corruption', 'Generosity', 'CUT' )

colnames(whr2019) <- c('CUT', 'Country', 'Happiness_Score', 'GDP_per_capita', 'CUT', 'Life_Expectency',
                       'Freedom','Generosity', 'Corruption' )

whr2019$Region <- whr2015$Region[match(whr2019$Country, whr2015$Country)]

whr2015 <- whr2015[ , -which(names(whr2015) %in% c('CUT'))]
whr2019 <- whr2019[ , -which(names(whr2019) %in% c('CUT'))]

whr2015 <- whr2015[, c(1,2,3,4,5,6,8,7)]
whr2019 <- whr2019[, c(1,8,2,3,4,5,6,7)]

common <- intersect(whr2015$Country, whr2019$Country)

whr2015 = filter(whr2015, Country %in% common)
whr2019 = filter(whr2019, Country %in% common)


datatable(whr2019, rownames = FALSE,
          options = list(columnDefs = list(list(className = 'dt-center', targets = 5)),
                         pageLength = 5,lengthMenu = c(5, 10, 15, 20)))
```

Show 5 entries Search:

Country	Region	Happiness_Score	GDP_per_capita	Life_Expectency	Freedom	Generosity	Corruption
Iceland	Western Europe	7.494	1.38	1.026	0.591	0.354	0.118
Finland	Western Europe	7.769	1.34	0.986	0.596	0.153	0.393
Norway	Western Europe	7.554	1.488	1.028	0.603	0.271	0.341
Denmark	Western Europe	7.6	1.383	0.996	0.592	0.252	0.41

Country	Region	Happiness_Score	GDP_per_capita	Life_Expectency	Freedom	Generosity	Corruption
New Zealand	Australia and New Zealand	7.307	1.303	1.026	0.585	0.33	0.38

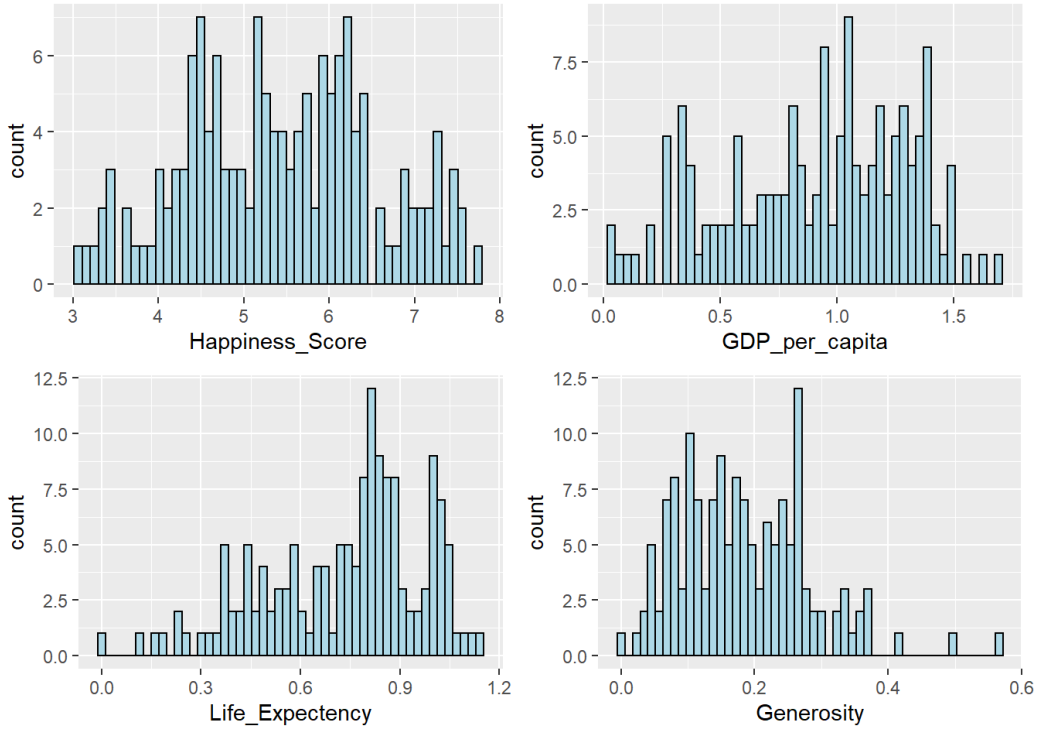
Showing 1 to 5 of 149 entries Previous 1 2 3 4 5 ... 30 Next

Part Two - Visualizations

In order to get to understand our data better we will visualize it using GGPlot package.

First of all, we will plot the different variables using histograms:

```
happiness_score_dist <- ggplot(whr2019, aes(x=Happiness_Score)) +
  geom_histogram(bins = 50, color="black", fill="lightblue")

gdp_dist <- ggplot(whr2019, aes(x=GDP_per_capita)) +
  geom_histogram(bins = 50, color="black", fill="lightblue")

life_expectency_dist <- ggplot(whr2019, aes(x=Life_Expectency)) +
  geom_histogram(bins = 50, color="black", fill="lightblue")

generosity_dist <- ggplot(whr2019, aes(x=Generosity)) +
  geom_histogram(bins = 50, color="black", fill="lightblue")

grid.arrange(happiness_score_dist, gdp_dist,life_expectency_dist,generosity_dist)
```



A slightly different look at the table grouped by regions.

```
whr2019 %>%
  group_by(Region)%>%
  summarize(
    mean_Score = mean(Happiness_Score),
    mean_GDP = mean(GDP_per_capita),
    mean_LE = mean(Life_Expectency),
    mean_Generosity = mean(Generosity)
  )
```

```
## # A tibble: 10 x 5
##    Region                          mean_Score mean_GDP mean_LE mean_Generosity
##    <chr>                                <dbl>    <dbl>   <dbl>           <dbl>
##  1 Australia and New Zealand             7.27    1.34    1.03            0.331
##  2 Central and Eastern Europe            5.57    1.02    0.808           0.141
##  3 Eastern Asia                          5.69    1.24    0.953           0.173
##  4 Latin America and Caribbean           5.94    0.909   0.817           0.143
##  5 Middle East and Northern Africa       5.24    1.06    0.751           0.153
##  6 North America                         7.08    1.40    0.956           0.282
##  7 Southeastern Asia                     5.27    0.93    0.745           0.302
##  8 Southern Asia                         4.53    0.650   0.617           0.235
##  9 Sub-Saharan Africa                    4.31    0.452   0.412           0.187
## 10 Western Europe                        6.90    1.36    1.01            0.221
```

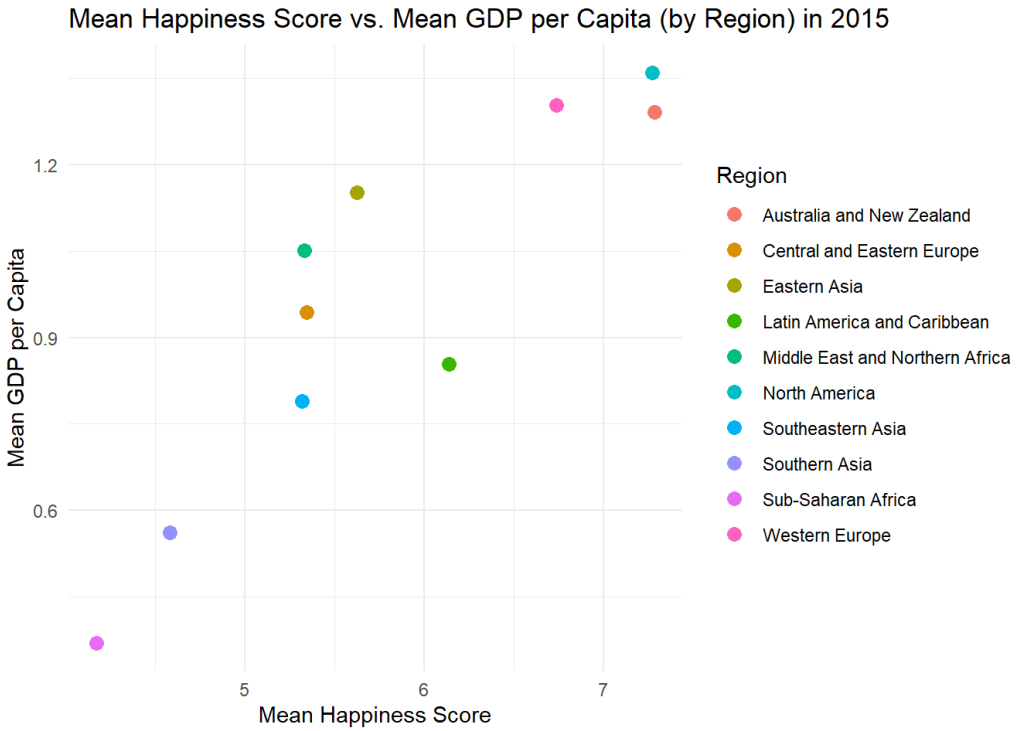
```
whr2015 %>%
  group_by(Region)%>%
  summarize(
    mean_Score = mean(Happiness_Score),
    mean_GDP = mean(GDP_per_capita),
    mean_LE = mean(Life_Expectency),
    mean_Generosity = mean(Generosity)
  )
```

```
## # A tibble: 10 x 5
##    Region                          mean_Score mean_GDP mean_LE mean_Generosity
##    <chr>                                <dbl>    <dbl>   <dbl>           <dbl>
##  1 Australia and New Zealand             7.28    1.29    0.920           0.455
##  2 Central and Eastern Europe            5.34    0.943   0.718           0.150
##  3 Eastern Asia                          5.63    1.15    0.877           0.226
##  4 Latin America and Caribbean           6.14    0.854   0.713           0.215
##  5 Middle East and Northern Africa       5.33    1.05    0.703           0.189
##  6 North America                         7.27    1.36    0.884           0.430
##  7 Southeastern Asia                     5.32    0.789   0.677           0.419
##  8 Southern Asia                         4.58    0.560   0.541           0.341
##  9 Sub-Saharan Africa                    4.17    0.370   0.277           0.218
## 10 Western Europe                        6.74    1.30    0.908           0.304
```

Using geometric points to draw a conclusion about the level of happiness and GDP product by regions.

We can understand which areas have a higher level of happiness and product.

```
# Calculate the mean values for whr2015
mean_data_2015 <- whr2015 %>%
  group_by(Region) %>%
  summarize(
    mean_Score = mean(Happiness_Score),
    mean_GDP = mean(GDP_per_capita)
  )

# Create the scatterplot with only mean points
ggplot(mean_data_2015, aes(x = mean_Score, y = mean_GDP, color = Region)) +
  geom_point(size = 3, shape = 19) +  # Plot mean points
  labs(title = "Mean Happiness Score vs. Mean GDP per Capita (by Region) in 2015",
    x = "Mean Happiness Score",
    y = "Mean GDP per Capita") +
  theme_minimal()
```

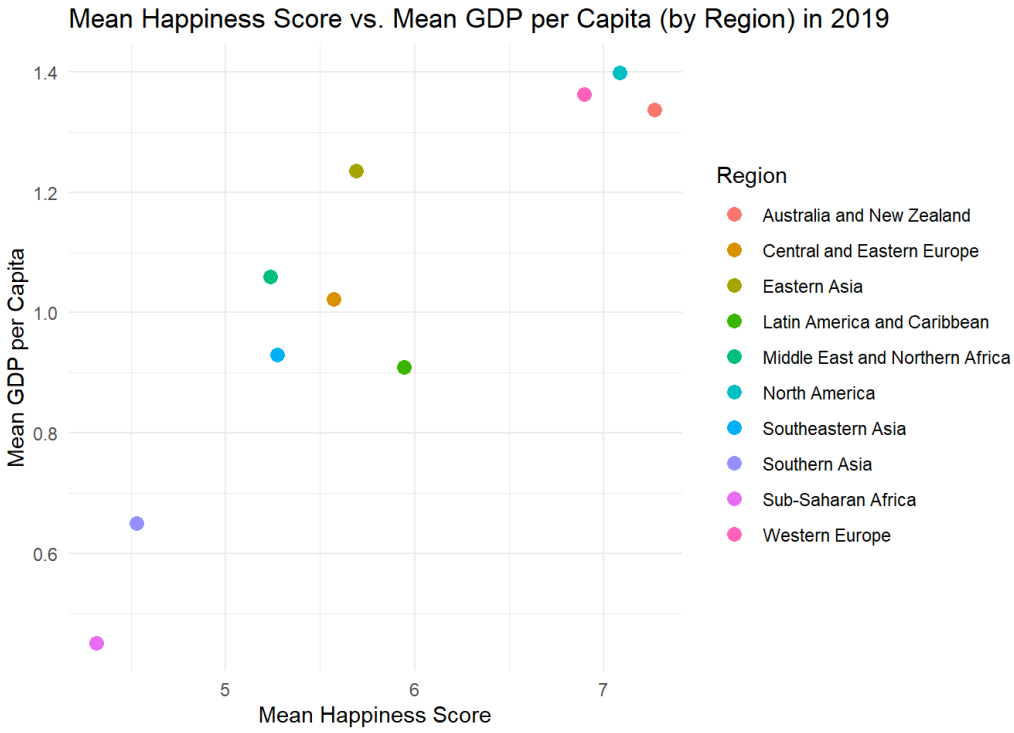


Western Europe, Australia and New Zealand and North America have a have a higher level of happiness and product. Whereas Sub-Saharan Africa and Southern Asia are on the lower level in both metrics.

We can also conclude from these mean points that as the level of happiness rises, the GDP per capita rises as well and vise versa.

Now let's look at 2019 data:

```
# Calculate the mean values for whr2015
mean_data_2019 <- whr2019 %>%
  group_by(Region) %>%
  summarize(
    mean_Score = mean(Happiness_Score),
    mean_GDP = mean(GDP_per_capita)
  )

# Create the scatterplot with only mean points
ggplot(mean_data_2019, aes(x = mean_Score, y = mean_GDP, color = Region)) +
  geom_point(size = 3, shape = 19) +  # Plot mean points
  labs(title = "Mean Happiness Score vs. Mean GDP per Capita (by Region) in 2019",
    x = "Mean Happiness Score",
    y = "Mean GDP per Capita") +
  theme_minimal()
```

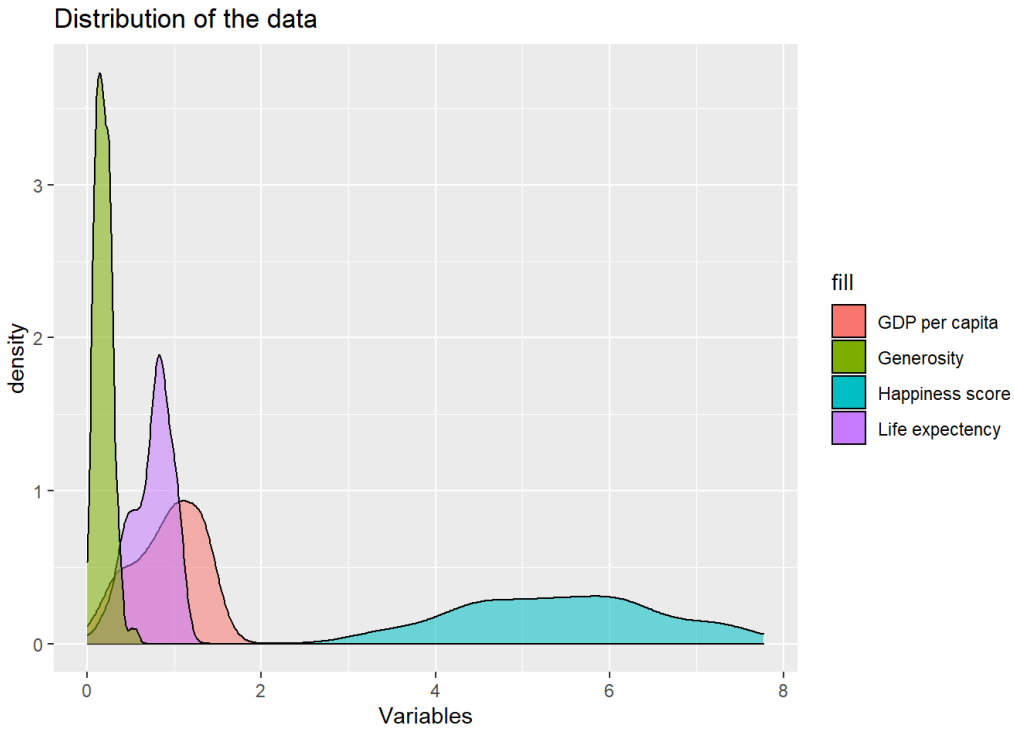


Seems like not much changed in 2019.

Density of the different variables:

In order to be more confident about our distribution of the data, we will use geom_density function to visualize the bell shaped graph.

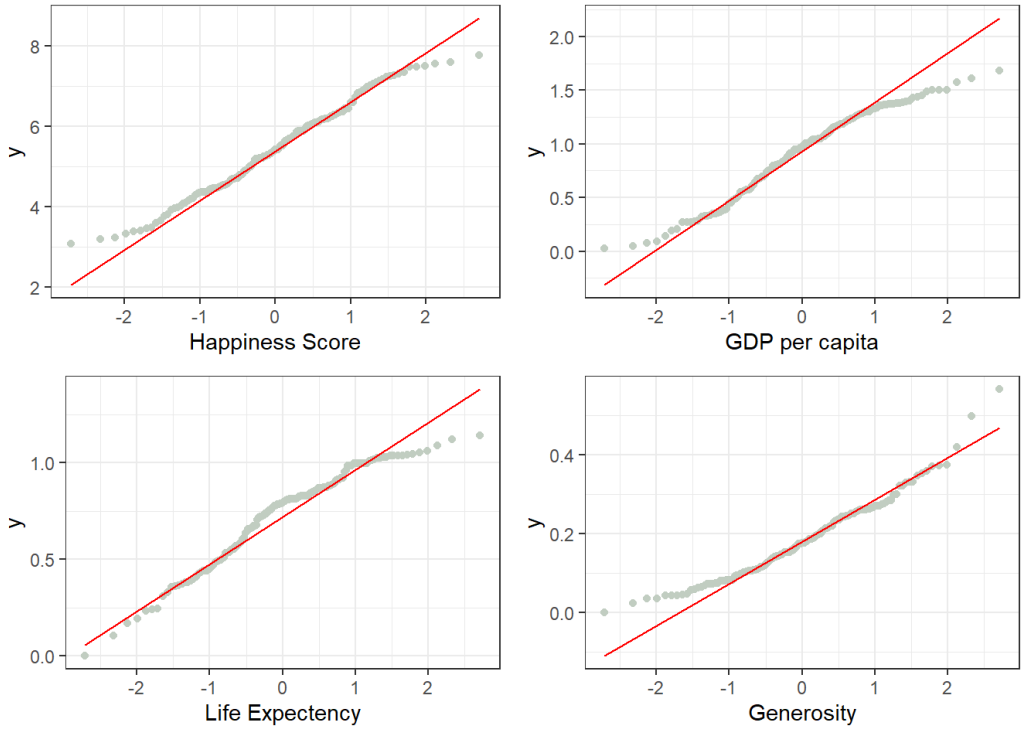
```
data_density <- ggplot(whr2019) +
  geom_density(aes(Happiness_Score, fill="Happiness score", alpha=0.1)) +
  geom_density(aes(GDP_per_capita, fill="GDP per capita", alpha=0.1)) +
  geom_density(aes(Life_Expectency, fill="Life expectency", alpha=0.1)) +
  geom_density(aes(Generosity, fill="Generosity", alpha=0.1)) +
    scale_x_continuous(name = "Variables") +
  ggtitle("Distribution of the data") +
  guides(alpha = FALSE)  # Remove alpha from legend

plot(data_density)
```



As we can see the different variables do have a bell shaped distribution.

But to be even more confident we will now use QQPlot Graph.

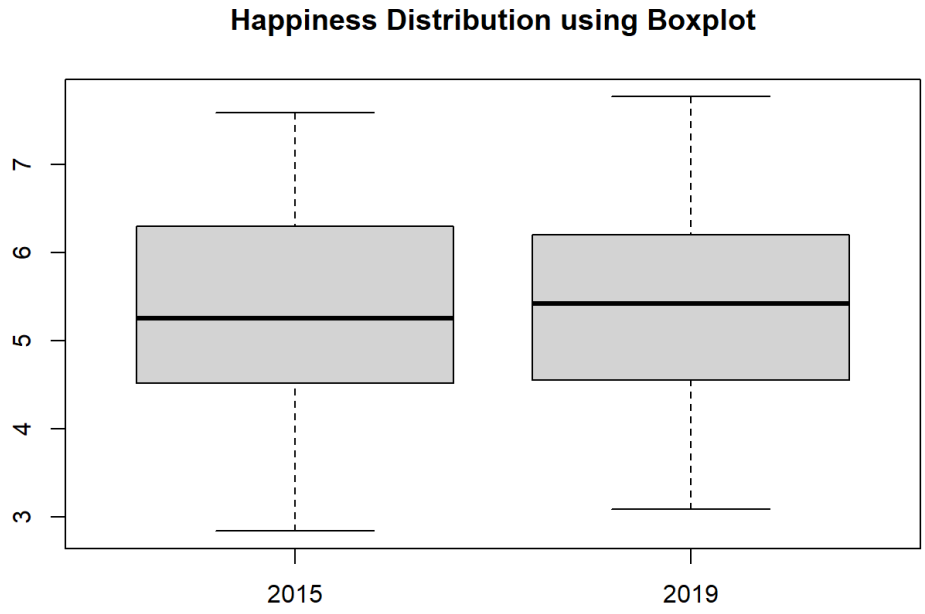
```
happy_qq <- ggplot(whr2019, aes(sample=Happiness_Score)) +
  geom_qq(color = "honeydew3") + geom_qq_line(col="red") + theme_bw() +
    labs(x= "Happiness Score")

gdp_qq <- ggplot(whr2019, aes(sample=GDP_per_capita)) +
  geom_qq(color = "honeydew3") + geom_qq_line(col="red") +theme_bw() +
    labs(x= "GDP per capita")

life_expectency_qq <- ggplot(whr2019, aes(sample=Life_Expectency)) +
  geom_qq(color = "honeydew3") + geom_qq_line(col="red") + theme_bw() +
    labs(x= "Life Expectency")

generosity_qq <- ggplot(whr2019, aes(sample=Generosity)) +
  geom_qq(color = "honeydew3") + geom_qq_line(col="red") + theme_bw() +
    labs(x= "Generosity")

grid.arrange(happy_qq, gdp_qq,life_expectency_qq,generosity_qq)
```



From these graphs we could be pretty sure that our data is indeed normally distributed.

Lets take a look on the happiness score with boxplot.

```
boxplot(whr2015$Happiness_Score, whr2019$Happiness_Score,
        names = c("2015", "2019"), main = "Happiness Distribution using Boxplot")
```



Part Three - Modeling

After we got to know our data better we can now move on to carry out our research.

1. Hypothesis test - difference in means

The happiness index in the world is expected to be the same. We will perform a system of hypotheses to test whether the happiness score has increased over the years and the world is moving towards a better future. We will perform a calculation at a significance level of 5 percent.

Since we compare two samples with the same parameters for the same countries.
Each observation in the first sample is paired with a single observation in the second sample and that is why these are paired samples and therefore we will conduct a paired t-test on them.

μ_1 : The average score of 2019 report
μ_2 : The average score of 2015 report

Hypothesis test:
$H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_1 - \mu_2 > 0$

```
t.test(whr2019$Happiness_Score, whr2015$Happiness_Score, paired = TRUE,
       alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  whr2019$Happiness_Score and whr2015$Happiness_Score
## t = 0.96123, df = 148, p-value = 0.169
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.04013068         Inf
## sample estimates:
## mean of the differences
##              0.05558389
```

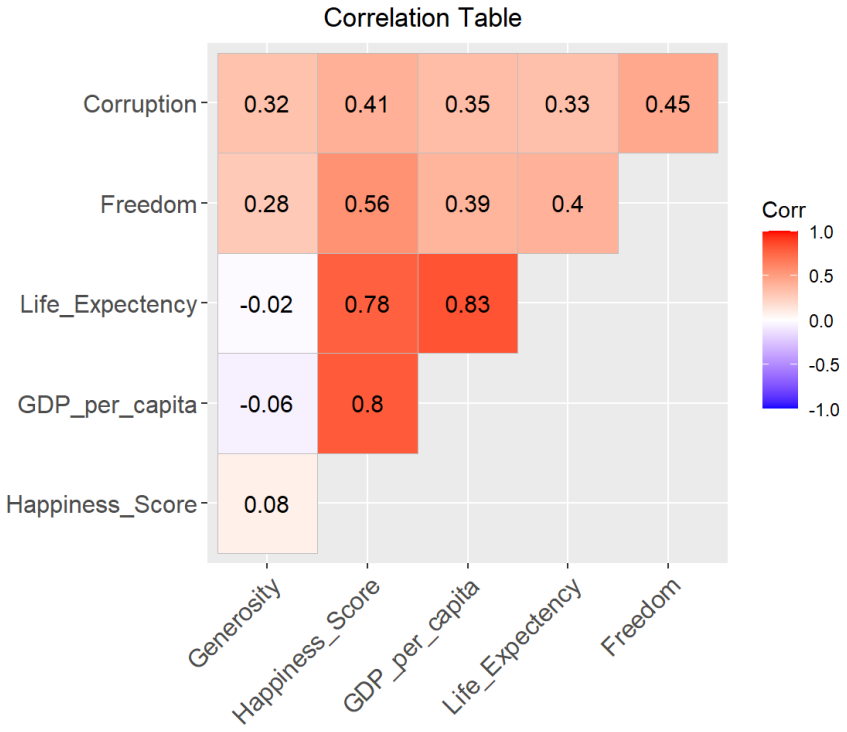
Conclusion:

We thought that the overall happiness will increase over the years. We compared P-value to the Alpha value and saw that Alpha < P-Value. Therefore we will not reject the null hypothesis at a significance level of 5 percent - the differences between the two years are insignificant.
We will conclude that there has been no significant change in the level of happiness in the world.

2. Multiple Regression

First, we want to see the correlation between the different variables and find simple association rules.

```
numberic_whr <- whr2019 %>%
  select(Happiness_Score, GDP_per_capita, Life_Expectency, Freedom, Generosity, Corruption)

whr_corr <- cor(numberic_whr)

ggcorrplot(whr_corr, title ="               Correlation Table",
           hc.order = TRUE, type = "upper", lab = T,
           ggtheme = ggplot2::theme_gray)
```

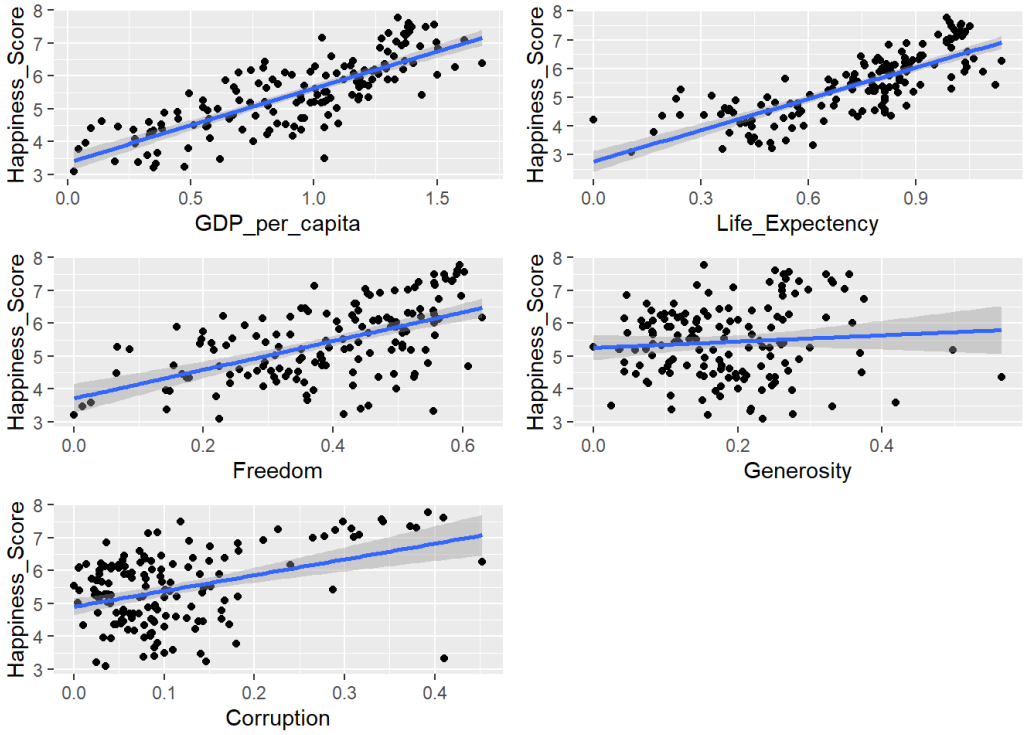


We can conclude from the chart that there is a high correlation between **happiness score**, **life expectancy** and the **GDP index**. That is, there is an almost perfect match between these three variables.

Moreover, we can infer that the level of **generosity** in countries has almost no effect on the **life expectancy** and the **GDP per capita**.

Let’s take a look on the linear regression of each variable with the happiness score:

```
gg_hs_gdp <- ggplot(whr2019, aes(x=GDP_per_capita, y=Happiness_Score)) +
  geom_point() +
  stat_smooth(method = "lm")

gg_hs_le <- ggplot(whr2019, aes(x=Life_Expectency, y=Happiness_Score)) +
  geom_point() +
  stat_smooth(method = "lm")

gg_hs_free <- ggplot(whr2019, aes(x=Freedom, y=Happiness_Score)) +
  geom_point() +
  stat_smooth(method = "lm")

gg_hs_genr <- ggplot(whr2019, aes(x=Generosity, y=Happiness_Score)) +
  geom_point() +
  stat_smooth(method = "lm")

gg_hs_corr <- ggplot(whr2019, aes(x=Corruption, y=Happiness_Score)) +
  geom_point() +
  stat_smooth(method = "lm")

grid.arrange(gg_hs_gdp, gg_hs_le, gg_hs_free, gg_hs_genr, gg_hs_corr)
```



As observed and stated before GDP per capita, life expectancy and freedom are indeed correlated to the happiness score, whereas generosity and corruption are not that correlated…

Now we can create the linear model using Summary Statistics Table:

```
model <- lm(Happiness_Score ~ GDP_per_capita + Life_Expectency + Freedom +
            Generosity + Corruption, data = whr2019)
summary(model)
```

```
##
## Call:
## lm(formula = Happiness_Score ~ GDP_per_capita + Life_Expectency +
##     Freedom + Generosity + Corruption, data = whr2019)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85095 -0.33158  0.08602  0.39151  0.97172
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.3842     0.1924  12.394  < 2e-16 ***
## GDP_per_capita    1.2228     0.2222   5.503 1.67e-07 ***
## Life_Expectency   1.4688     0.3603   4.076 7.57e-05 ***
## Freedom           1.8372     0.4015   4.575 1.02e-05 ***
## Generosity        0.4582     0.5427   0.844    0.400
## Corruption        0.4427     0.5945   0.745    0.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5764 on 143 degrees of freedom
## Multiple R-squared:   0.74,  Adjusted R-squared:  0.7309
## F-statistic:  81.4 on 5 and 143 DF,  p-value: < 2.2e-16
```

The first step in interpreting the multiple regression analysis is to examine the F-statistic and the associated p-value, at the bottom of model summary.

In this model, it can be seen that p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable.

By looking at the coefficients table of the variables we can see that there is significant association between **GDP per capita, life expectancy** and **freedom** with the outcome variable - **the happiness score**.

But, **generosity** and **corruption** variables are not significant in the model.

Which means that the 74% (R-Squared) of the score of happiness can explained by these variables.

Because generosity and corruption are not significant we can remove them from the model and get even more accurate model:

```
model2 <- lm(Happiness_Score ~ GDP_per_capita + Life_Expectency + Freedom, data = whr2019)
summary(model2)
```

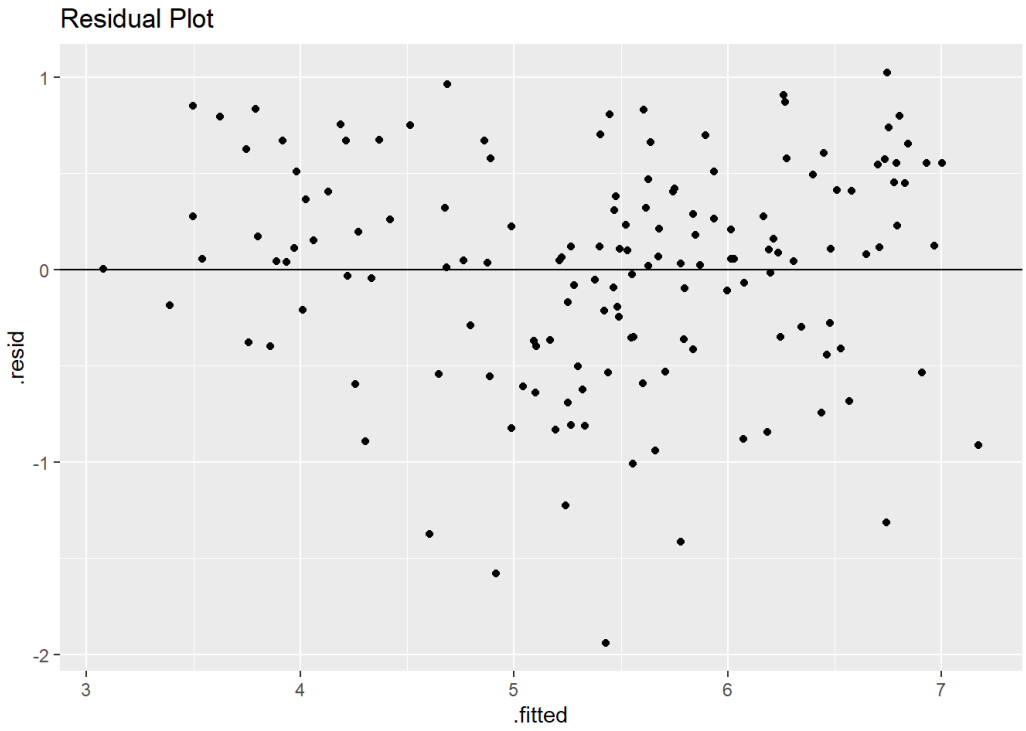
```
##
## Call:
## lm(formula = Happiness_Score ~ GDP_per_capita + Life_Expectency +
##     Freedom, data = whr2019)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93954 -0.36786  0.05693  0.41276  1.02537
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.4297     0.1745  13.926  < 2e-16 ***
## GDP_per_capita    1.2172     0.2178   5.589 1.10e-07 ***
## Life_Expectency   1.4783     0.3599   4.107 6.66e-05 ***
## Freedom           2.0558     0.3655   5.625 9.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.576 on 145 degrees of freedom
## Multiple R-squared:  0.7368, Adjusted R-squared:  0.7314
## F-statistic: 135.3 on 3 and 145 DF,  p-value: < 2.2e-16
```

As expected the R-Squared went down by a small amount since we dropped two variables from the model. Also, the Adjusted R-Squared indeed went up (the closer to 1 the better - we were punished less).

But before even making final conclusions we must first check whether our model does meet the assumptions:
- The residuals are homoscedastic.
- The residuals are distributed normally.
- No multicollinearity between the explanatory variables.

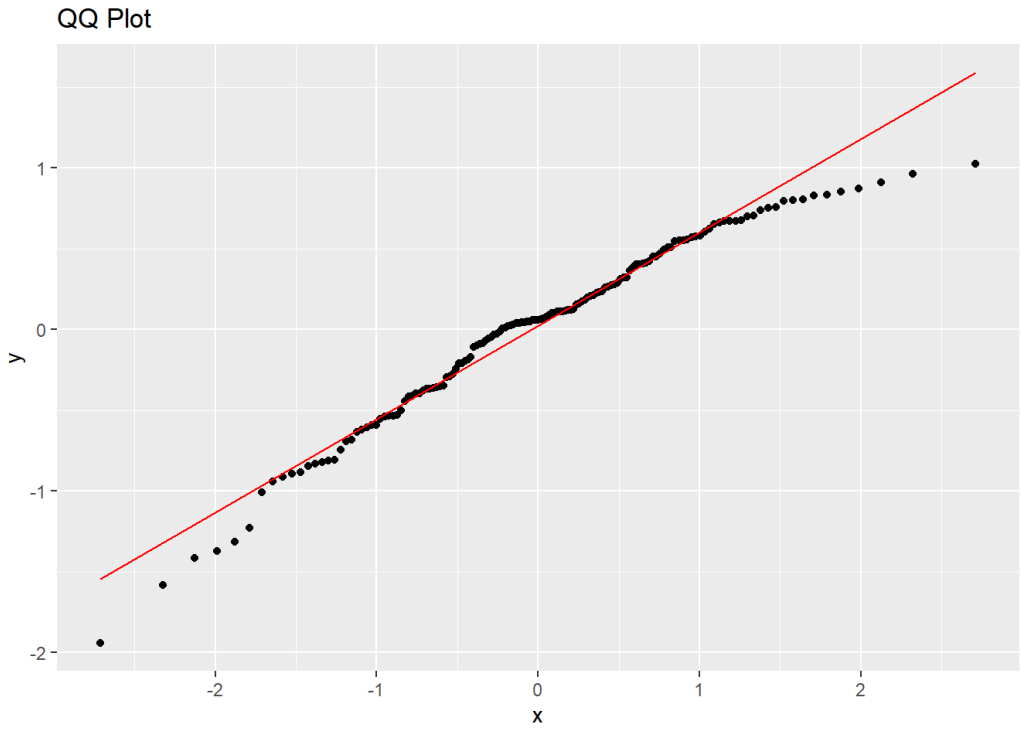
Residual Plot

```
residual_upd_model <- model2$residuals

model2 %>% ggplot(aes(x=.fitted,y=.resid)) +
  geom_point() + geom_hline(yintercept=0) +
  labs(title="Residual Plot")
```



Distribution of the residuals using QQ Plot

```
model2 %>% ggplot(aes(sample=.resid)) +
  geom_qq() + geom_qq_line(col="red") +
  labs(title="QQ Plot")
```



VIF between the explanatory variables

```
car::vif(model2)
```

```
##  GDP_per_capita Life_Expectency         Freedom
##        3.262036        3.288904       1.206238
```

As we can see:

- the residuals are hetroscedastic.
- the residuals are not distributed normally.
- the VIF between the independent variables is lower than 5 which means they are not correlated.

It can be seen that assumptions 1 and 2 do not hold, therefore, the model we have created seems to be unsuitable for predicting models.

Conclusion:

In the project we wanted to study the happiness index among the countries of the world. We selected a database, presented the results and researched them. We studied the variables in each country and their effects on the happiness index. We used a number of models - Examining hypotheses for the level of happiness between different countries over the years. A multiple regression model on the happiness index with the explanatory variables, and the effect of different variables on the data.

We have come to interesting conclusions about the different countries, their index of happiness and the variables that affect and are affected by happiness.

The project helped us to deeply understand the different tests we learned during the course, how to build them, how to work with them and most important - how to draw interesting conclusions from them.


FINDING
HAPPINESS