World Happiness Analysis

Andrea Porro

2024-10-15

World happiness analysis

The first project of introduction to data mining is focused on the analysis for the world happiness. Before starting work with the data-set, it is important to do an exhaustive analysis for each of our variable. Analyzing the variable help to learn more about their characteristics. This preliminary analysis is necessary to provide a correct interpretation of the data-set.

```
Data = read.csv("World-happiness-report-updated_2024.csv")
NewData=subset(Data,year==2023)
head(NewData)
```

```
##     Country.name year Life.Ladder Log.GDP.per.capita Social.support
## 15   Afghanistan 2023       1.446                 NA          0.368
## 31       Albania 2023       5.445              9.689          0.691
## 64     Argentina 2023       6.393              9.994          0.892
## 81       Armenia 2023       5.679              9.730          0.819
## 98     Australia 2023       7.025             10.846          0.896
## 114      Austria 2023       6.636             10.930          0.874
##     Healthy.life.expectancy.at.birth Freedom.to.make.life.choices Generosity
## 15                              55.2                        0.228         NA
## 31                              69.2                        0.872      0.068
## 64                              67.3                        0.832     -0.129
## 81                              68.2                        0.819     -0.179
## 98                              71.2                        0.876      0.187
## 114                             71.4                        0.874      0.209
##     Perceptions.of.corruption Positive.affect Negative.affect
## 15                      0.738           0.261           0.460
## 31                      0.855           0.597           0.314
## 64                      0.846           0.720           0.301
## 81                      0.681           0.575           0.423
## 98                      0.482           0.731           0.248
## 114                     0.529           0.712           0.240
```

The data-set if made by 11 variables:

1. **Country name**, is the name of the country.
2. **Year**, is the year when the information was taken.
3. **Life Ladder**, is the overall happiness score. This could be also considered as the target variable.
4. **Log GPD per capita**, represent the wealth of the country and is a major contributor to the happiness.
5. **Social support**, the measure of the individual's access to social support network such as having friends you can count on.

6. **Healthy life expectancy at birth**, the average number of years that an individual is expected to live in good health.
7. **Freedom to make life choices**, the freedom of each individual to make their own life choice.
8. **Generosity**, the generosity of the country's population that includes, for example, the charitable donations.
9. **Perceptions of corruption**, the perception of corruption in a country. This perception is made by the honesty of the government and business.
10. **Positive affect**, represent the frequency with which people experience positive emotion such as happiness, laughter… A high score of this variable means that people who lives in this country generally tent to experience positive emotion frequently.
11. **Negative affect**, in opposite to the variable "positive affect" this one shows the frequency with which people experience negative emotion such as sadness or anger. More this variable is high, more is frequently have people that experience negative emotions.

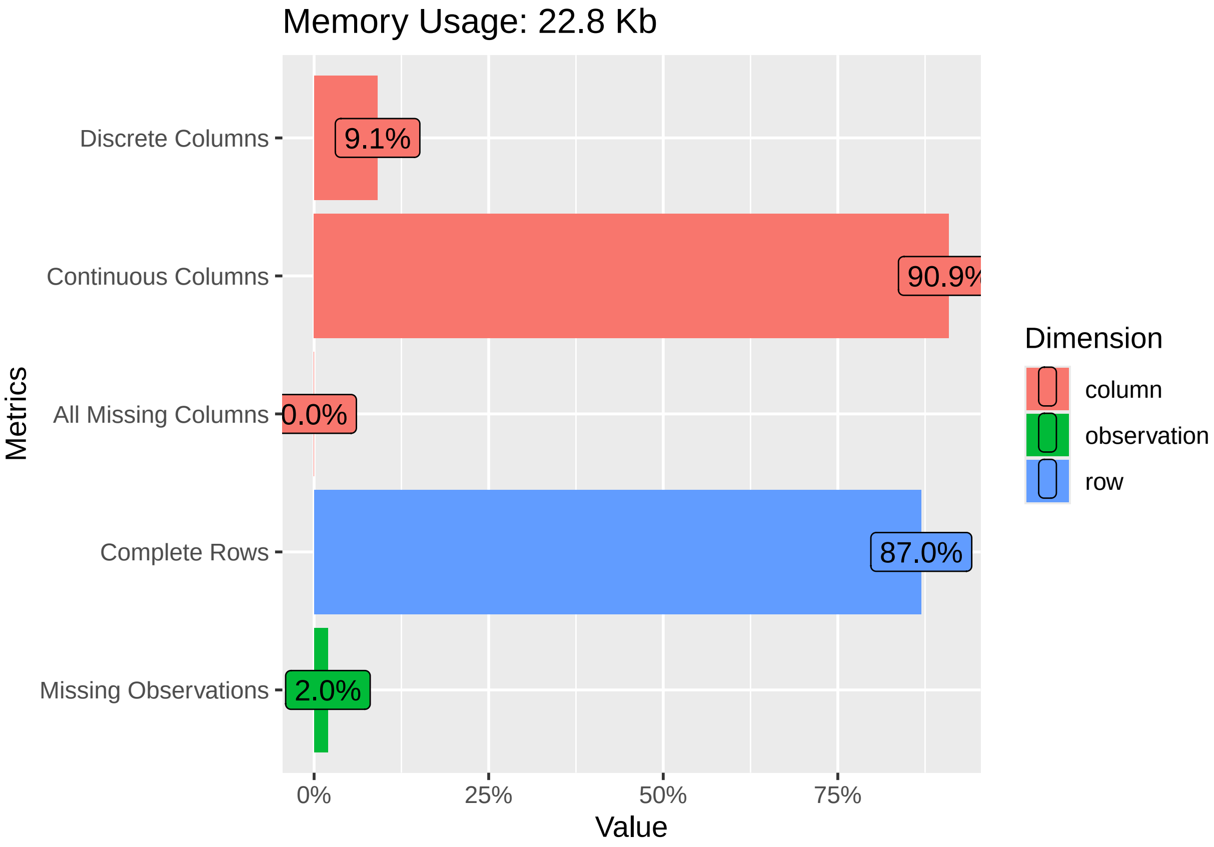
Before starting it is significant to know that the study is only analyzing the year 2023, since it is the most recent year available. However the study do not preclude the possibility to analyze how the happiness of a country may change during the years.

EDA - Exploratory data analysis

It is still important to deepen the understanding of the data-set.

First of all we will analyze in general our data-set building up a plot.

```
library(DataExplorer)
plot_intro(NewData)
```

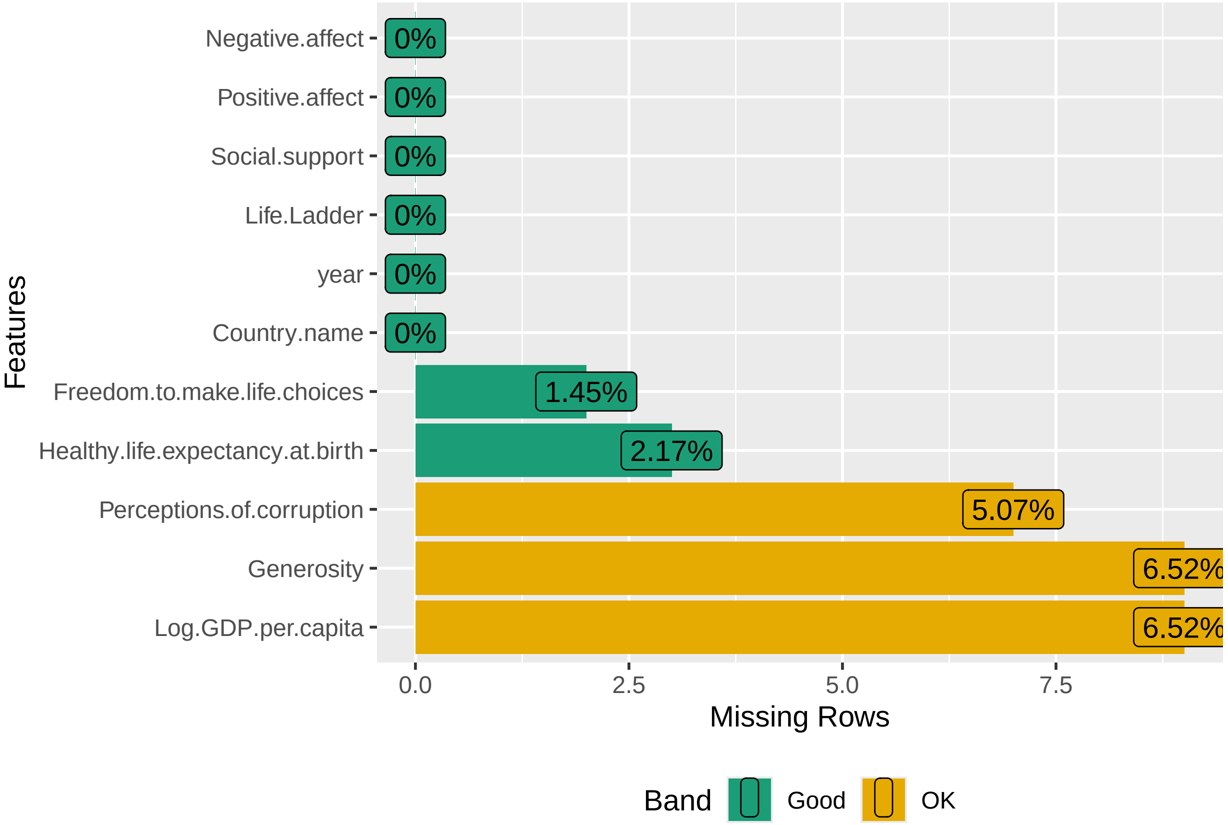


There are several considerations to do seeing this graph. The graph could be divided into three parts:

- **Columns**, the red part of our graph is focused on our columns. This divides the columns into two categories: numeric variables and categorical variables. From what can be seen from the graph, the data-set has 10 numerical variable, 90.9%, and only 1 categorical variable, 9.1%, that is obviously "country name".
- **Row**, the blue part of our graph that is focused on our rows. Differently from the previous part regarding the columns, in this portion there is only one graph bar. The complete row is equal to 87%, meaning that considering all the rows of the data-set just a 13% have missing values. Considering 138 observations, that means that only 18 of them have some missing values.
- **Missing observations**, the last part that is analyzed is about the missing values (NA). Analyzing the missing values is extremely important because knowing the presence of NA allows taking precautions during the study. The data-set about world happiness has only a 2% of missing values that must not be forgotten.

Since the study of missing values is so important the analysis propose another graph about the NA.

```
plot_missing(NewData)
```



The graph above gives to our project a clear vision of which variables have more missing values and which variables have less NA. As we did before, this graph could be divided into two different part. The first one is where the NA is lower than 5%, and it could be considered more than acceptable. To be more clear, the representation of this variable is where the bar is green. The other part is all the variables with a number of missing values higher than 5%. In this part there are the variables called generosity, perception of corruption and log GPD per capita and they are represented by the yellow bar.

The last general analyze the study will do about the variables before starting to answears to more specifics questions is the summary. The summary is one of the most important command on RStudio. This command allows to know important information for each variable. If the variables are numerical the command will print:

- The minimum, that is the lowest value
- The first quartile, means that the 25% of my observations reach up to the value of this quartile.
- The median, this is the central value. In addition the median represent also the second quartile meaning that the 50% of my observations reach up to the value of this quartile.
- The mean, that is the average value
- The third quartile, means that the 75% of my observations reach up to the value of this quartile.
- The maximum, the highest value for the variable.
- NA's, the number of missing values for the variable. If the variable hasn't any NA than there will not be this voice.

However if the variables are categorical the command will give other information.

```r
summary(NewData)
```

```
##  Country.name            year       Life.Ladder    Log.GDP.per.capita
##  Length:138         Min.   :2023   Min.   :1.446   Min.   : 7.076
##  Class :character   1st Qu.:2023   1st Qu.:4.680   1st Qu.: 8.620
##  Mode  :character   Median :2023   Median :5.863   Median : 9.637
##                     Mean   :2023   Mean   :5.621   Mean   : 9.517
##                     3rd Qu.:2023   3rd Qu.:6.487   3rd Qu.:10.504
##                     Max.   :2023   Max.   :7.699   Max.   :11.676
##                                                    NA's   :9
##  Social.support   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices
##  Min.   :0.3680   Min.   :52.20                    Min.   :0.2280
##  1st Qu.:0.7023   1st Qu.:60.70                    1st Qu.:0.7348
##  Median :0.8290   Median :66.10                    Median :0.8030
##  Mean   :0.7910   Mean   :65.19                    Mean   :0.7903
##  3rd Qu.:0.8898   3rd Qu.:69.60                    3rd Qu.:0.8762
##  Max.   :0.9790   Max.   :74.60                    Max.   :0.9650
##                   NA's   :3                        NA's   :2
##    Generosity      Perceptions.of.corruption Positive.affect  Negative.affect
##  Min.   :-0.2680   Min.   :0.1530            Min.   :0.2610   Min.   :0.1110
##  1st Qu.:-0.0710   1st Qu.:0.6620            1st Qu.:0.5813   1st Qu.:0.2293
##  Median : 0.0280   Median :0.7690            Median :0.6685   Median :0.2850
##  Mean   : 0.0336   Mean   :0.7211            Mean   :0.6521   Mean   :0.2934
##  3rd Qu.: 0.1380   3rd Qu.:0.8385            3rd Qu.:0.7355   3rd Qu.:0.3575
##  Max.   : 0.5900   Max.   :0.9480            Max.   :0.8430   Max.   :0.5160
##  NA's   :9         NA's   :7
```

The first variable shown using this command is "Country name," which is already well-known as a categorical variable. The command provides only the length, which is 138 (matching the number of observations), and both the class and the mode, which return "character," as expected.

All other variables are numerical, as seen previously in the plot_intro. We will analyze just a few of them, as the interpretation is similar for each one.

- **Year**
  As mentioned earlier, the study focuses on analyzing only the most recent year available. For this reason, the output shows "2023" for each index. This confirms that the previously used code operated correctly.

- **Life ladder**
  This is one of the most important variables in the dataset. Using the command, it was discovered that the lowest value is 1.446 and the highest is 7.669. The average value of the happiness score is 5.621. Interestingly, the mean is lower than the median, suggesting that the distribution might have a negative skew, meaning some countries have very low Life Ladder values, but most countries have values close to the mean. However, since there is no graph, this remains only a hypothesis.

  Another important point is the variability. The difference between the lowest and highest values is 6.253, meaning that the perception of happiness varies greatly between countries. Let's now analyze some quartiles. The first quartile of the Life Ladder is 4.680, meaning that 25% of the countries have a happiness level lower than this value. The second quartile, represented by the median, is 5.863, meaning that 50% of the countries have a happiness level below 5.863. The third quartile value is 6.487. Countries with a value higher than this have a relatively good perception of life quality.

- **Log GDP per capita**
  As before, let's start by analyzing the minimum and maximum. The lowest value of Log GDP per capita is 7.096, and the highest is 11.676. As noted in plot_missing, this variable has one of the highest numbers of missing values, and the summary gives the exact number, which is 9. The mean and median are close, as seen in the previous variable, which could suggest a normal distribution with negative skewness. However, until a graph is produced, this remains speculative.

  Moreover, the values are more symmetrically distributed; the median and mean are almost equidistant from the minimum and maximum values. This differs from the Life Ladder variable, where the minimum is much farther from the mean and median than the maximum is. In the quartile analysis, 25% of countries have a Log GDP per capita lower than 8.620, 50% lower than 9.637, and 75% lower than 10.504.

Correlation

The papar start analyzing the correlation plot which is one of the most important plot for statistics studies.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
NewData_numeric = NewData %>% select_if(is.numeric)
NewData_numeric = NewData_numeric %>% select(-year)
```

First of all it is necessary to create a new data-set which have only numerical variable. The new data-set is called "NewData_numeric". This new data contain 138 observation and the 9 numerical variable that was analyzed before. Now it can proceed building the correlation matrix.

```r
corr_matrix=cor(NewData_numeric, use = "complete.obs")
print(corr_matrix)
```

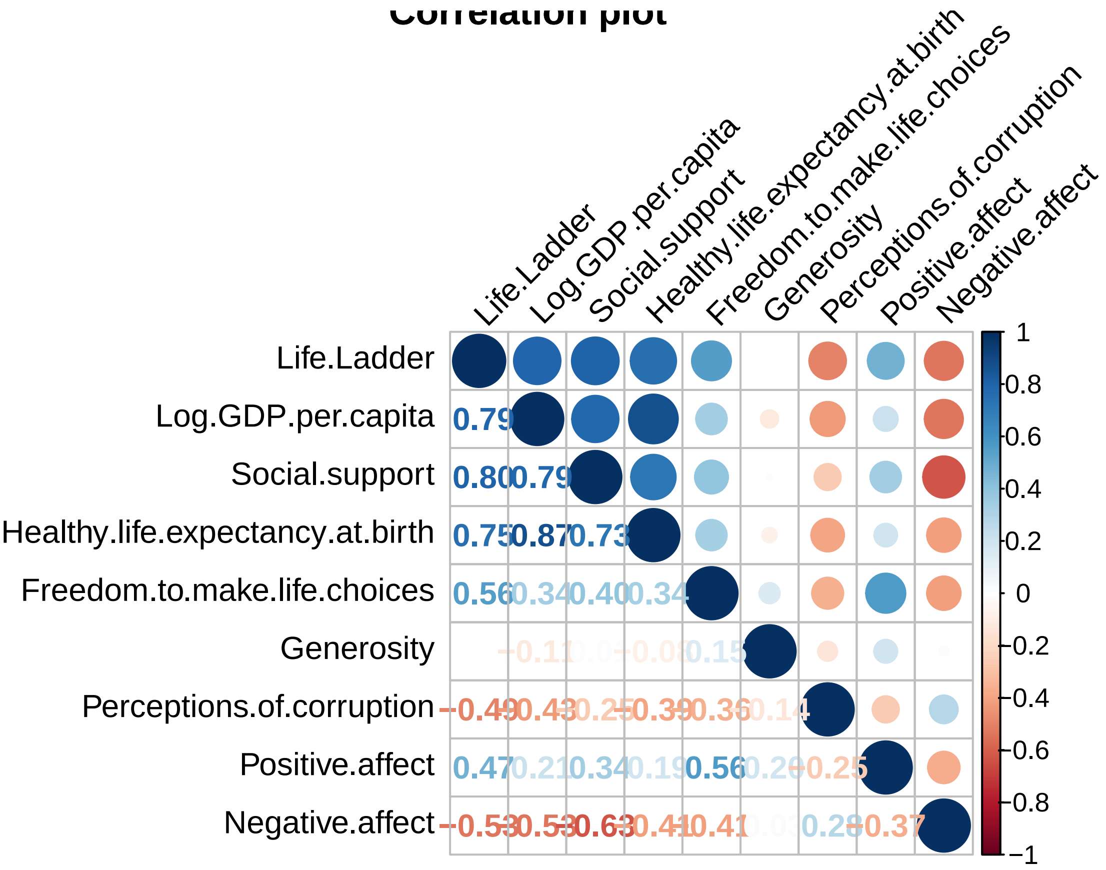
```
##                                  Life.Ladder Log.GDP.per.capita Social.support
## Life.Ladder                      1.000000000          0.7920684     0.80111354
## Log.GDP.per.capita               0.792068356          1.0000000     0.78632872
## Social.support                   0.801113535          0.7863287     1.00000000
## Healthy.life.expectancy.at.birth 0.753117200          0.8709759     0.72562344
## Freedom.to.make.life.choices      0.555625825          0.3442152     0.39860844
## Generosity                         0.007218658         -0.1116968     0.01127926
## Perceptions.of.corruption         -0.491998781         -0.4262296    -0.25127019
## Positive.affect                    0.471947564          0.2121678     0.34404266
## Negative.affect                   -0.530789460         -0.5337187    -0.62576206
##                                  Healthy.life.expectancy.at.birth
## Life.Ladder                                             0.75311720
## Log.GDP.per.capita                                      0.87097590
## Social.support                                          0.72562344
## Healthy.life.expectancy.at.birth                        1.00000000
## Freedom.to.make.life.choices                            0.33988725
## Generosity                                             -0.07922411
## Perceptions.of.corruption                              -0.39268731
## Positive.affect                                         0.19137776
## Negative.affect                                        -0.41020688
##                                  Freedom.to.make.life.choices   Generosity
## Life.Ladder                                         0.5556258  0.007218658
## Log.GDP.per.capita                                  0.3442152 -0.111696792
## Social.support                                      0.3986084  0.011279261
## Healthy.life.expectancy.at.birth                    0.3398872 -0.079224108
## Freedom.to.make.life.choices                        1.0000000  0.152698266
## Generosity                                          0.1526983  1.000000000
## Perceptions.of.corruption                          -0.3554442 -0.135788463
## Positive.affect                                     0.5623004  0.195693819
## Negative.affect                                    -0.4116225  0.029436278
##                                  Perceptions.of.corruption Positive.affect
## Life.Ladder                                     -0.4919988       0.4719476
## Log.GDP.per.capita                              -0.4262296       0.2121678
## Social.support                                  -0.2512702       0.3440427
## Healthy.life.expectancy.at.birth                -0.3926873       0.1913778
## Freedom.to.make.life.choices                    -0.3554442       0.5623004
## Generosity                                      -0.1357885       0.1956938
## Perceptions.of.corruption                        1.0000000      -0.2545744
## Positive.affect                                 -0.2545744       1.0000000
## Negative.affect                                  0.2841795      -0.3663516
##                                  Negative.affect
## Life.Ladder                          -0.53078946
## Log.GDP.per.capita                   -0.53371874
## Social.support                       -0.62576206
## Healthy.life.expectancy.at.birth     -0.41020688
## Freedom.to.make.life.choices         -0.41162248
## Generosity                            0.02943628
## Perceptions.of.corruption             0.28417949
## Positive.affect                      -0.36635156
## Negative.affect                       1.00000000
```

The correlation matrix gives the correlation value between the variable, however, it is much better to create a plot using those values. Create a plot may help to understand more easily and faster which variable are correlated and if they are positive or negative correlated.

```r
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```
corrplot.mixed(corr_matrix, lower="number", upper="circle", tl.pos = "lt",
               tl.col="black", tl.srt=45, diag="u", title="Correlation plot")
```



Let's focus our attention on the upper part of the graph, which can be divided into two distinct sections:

- The first section is where the bubbles are dark blue. If the correlation between two variables is represented by a blue bubble, it means the correlation is positive. In the study, almost all positive correlations are medium to high, which is also evident from the size of the bubbles. The larger the circle, the higher the correlation.
- The second section contains the red bubbles. A red bubble indicates a negative correlation. In the study, there are only two variables that consistently show medium negative correlations: "perceptions of corruption" and "negative affect." However, it was already expected that these variables would have negative correlations with others. In fact, the higher the perception of corruption in a country, the harder it is for people to feel happy. The same applies to the "negative affect" variable: the more negative emotions people experience, the harder it is for them to feel happy in that country.

Now, let's analyze the variables further. As previously mentioned, the plot shows that many variables have a high positive correlation, which indicates a potential issue with multicollinearity. Specifically, a high correlation can be observed between "Log GDP per capita" and "Social support" (with a correlation of 0.79) and between "Log GDP per capita" and "Healthy life expectancy at birth" (with a correlation of 0.87).

Another interesting point the study highlights is the correlation between "Generosity" and the other variables. It is clear from the plot that this variable has a very low correlation with all the others, suggesting that "Generosity" may not have a strong connection to the overall happiness analysis.

It is possible to create a cycle that give all the information for high or low correlations, an example of this command is:

```
corr_GPD_Healpth=corr_matrix["Log.GDP.per.capita","Healthy.life.expectancy.at.birth"]

if (abs(corr_GPD_Healpth) < 0.5) {
  cat("Age and balance have a weak or no linear relationship.\n")
} else {
  cat("Age and balance are moderately or strongly correlated.\n")
}
```

```
## Age and balance are moderately or strongly correlated.
```

In conclusion it can be said:

- The variables “Log GDP per capita”, “Social support”, “Healthy life expectancy at birth” are strong correlated to the happiness.
- The variables “Freedom to make life choices” and “Positive affect” are medium correlated with the target variable.
- The variables “Perceptions of corruption” and “Negative affect” are correlated with the happiness in a negative way.
- The variable “Generosity” is too low correlated to each of our variable.

To improve the happiness of their people, each country should focus on increasing life expectancy by investing in hospitals and supporting healthcare workers. Another important aspect to address is increasing GDP per capita. Raising this variable is more challenging, nations should work to stabilize prices, invest in infrastructure, promote innovation and technology, and support small businesses. At the same time, they must also fight corruption by educating the public, raising awareness, and making stricter anti-corruption laws.

Multicollinearity

Since the correlation between some variables is high, it is important to study the presence or absence of multicollinearity.

```
modello = lm(formula= Life.Ladder ~ . -Country.name - year, data=NewData)

library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
vif(modello)
```

```
##               Log.GDP.per.capita                   Social.support
##                         5.997095                         3.698791
## Healthy.life.expectancy.at.birth     Freedom.to.make.life.choices
##                         4.491613                         1.710043
##                       Generosity        Perceptions.of.corruption
##                         1.138818                         1.454177
##                  Positive.affect                  Negative.affect
##                         1.598047                         1.893460
```

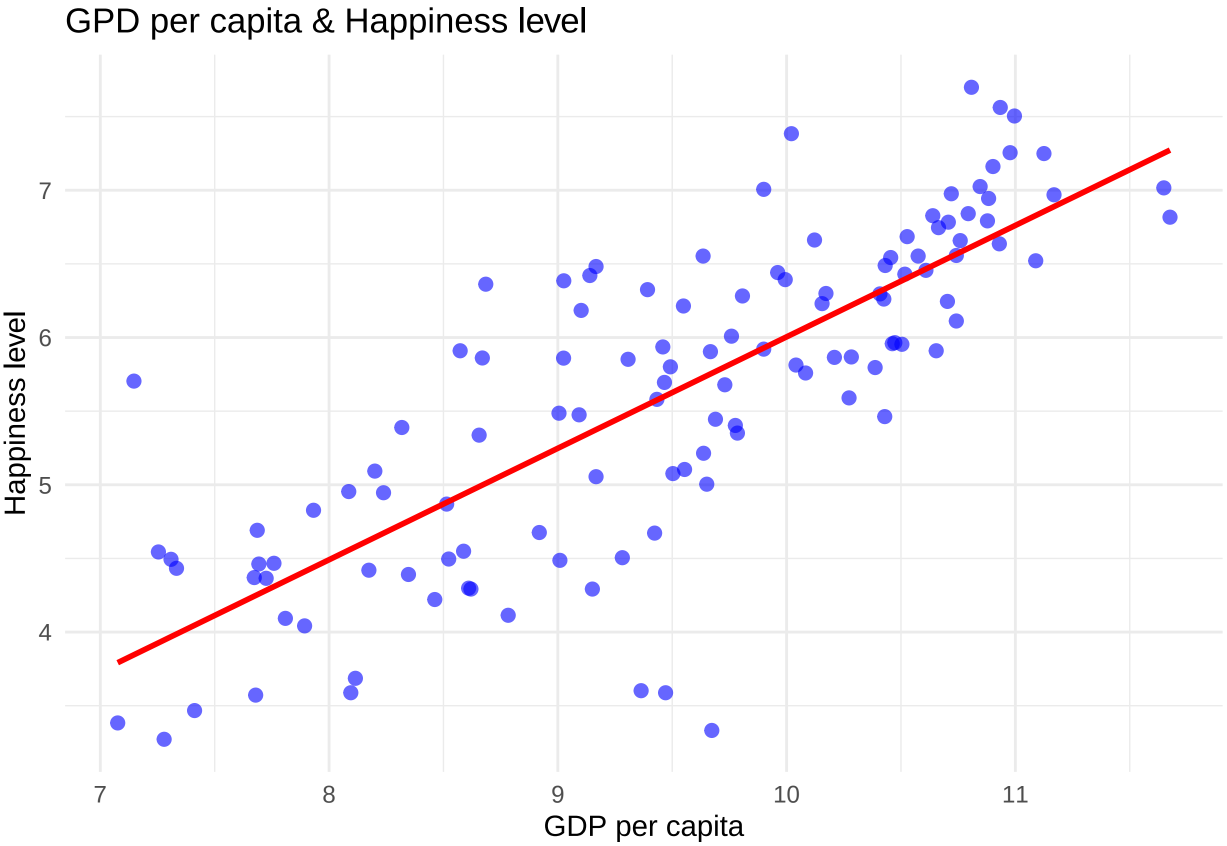
As we can see from the results obtained from the VIF, all values are <10, so we can conclude that there is no multicollinearity.

GDP per capita analysis

To analyze the correlation between two variables more precisely, it is possible to create a graph called scatter plot.

```
library(ggplot2)
ggplot(na.omit(NewData),aes(x=Log.GDP.per.capita, y=Life.Ladder)) +
        geom_point(size=2, alpha=0.6, color="blue") +
        geom_smooth(method="lm", color="red",se=F) +
        labs(title="GPD per capita & Happiness level",
            x = "GDP per capita",
            y = "Happiness level") +
        theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



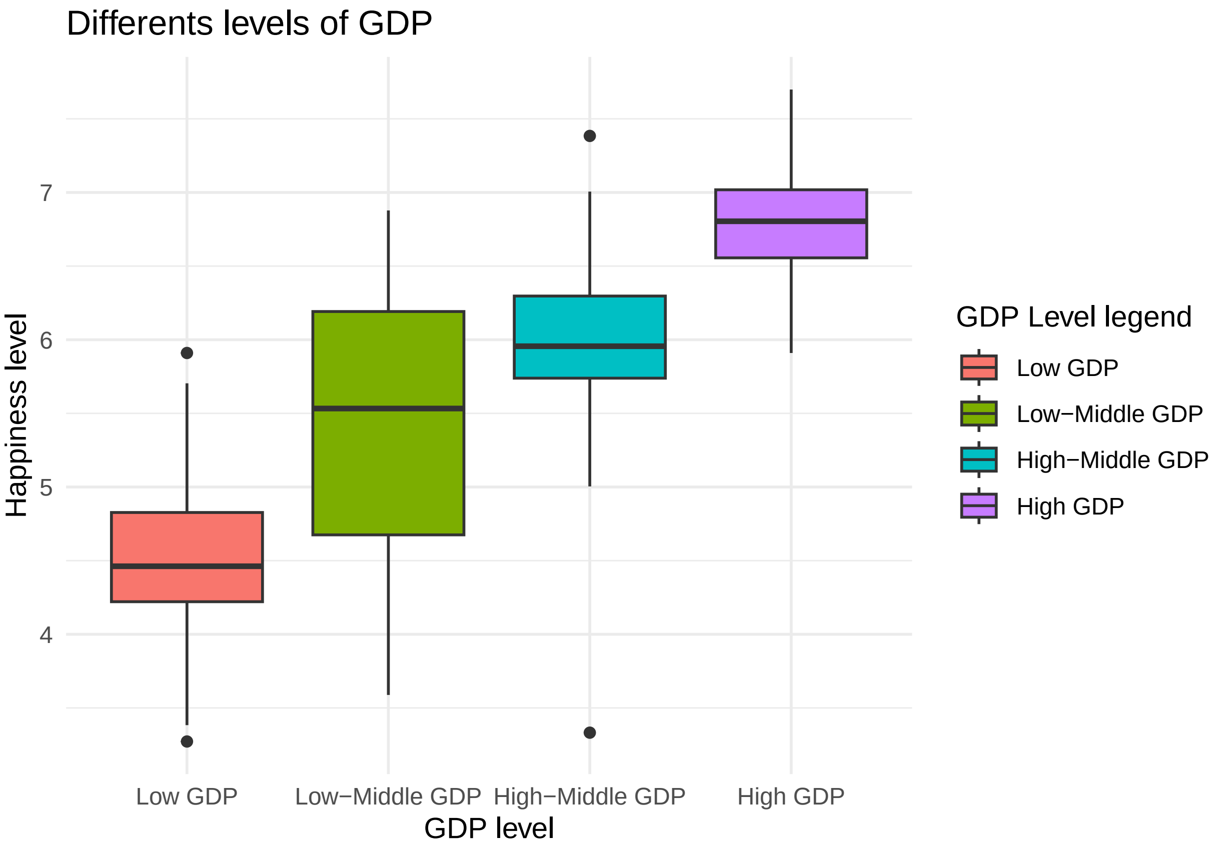
As observed in the correlation analysis, the two variables are correlated with a value of 0.79, and this graph confirms what was founded: GDP per capita has a significant impact on a country's happiness. However, it can also be seen that some countries have the same level of GDP per capita but different levels of happiness.

The study proceeds by creating a new variable called "GDP_Levels," which divides each country into quartiles based on their GDP. In this way, countries with a GDP lower than the first percentile, which is 8.620, are classified as "Low GDP." Countries between the first and the second percentiles are classified as "Lower-Middle GDP," and so on.

```
world_happiness_data=NewData %>%
  mutate (GDP_levels = cut (Log.GDP.per.capita, breaks=quantile(Log.GDP.per.capita, probs=seq(0,1,0.25)


world_happiness_data = world_happiness_data %>% filter(!is.na(GDP_levels))
world_happiness_data = world_happiness_data %>% filter(!is.na(Life.Ladder))

ggplot(world_happiness_data, aes(x=GDP_levels, y=Life.Ladder, fill=GDP_levels)) +
      geom_boxplot() +
      labs(title="Differents levels of GDP",
           x = "GDP level",
           y = "Happiness level",
           fill = "GDP Level legend") +
      theme_minimal()
```



Let's start analyzing each level and then giving a more general conclusion. As the it is well clear from the graph there are four different level:

1. **Low GDP (red)**. The box for this first group ranges from approximately 4.3 to 4.8, indicating that 50% of the countries with low GDP have similar happiness levels. This group has some outliers, with values above 5.6 and below 3.4. The median happiness level for countries in the low GDP group is around 4.5.
2. **Low-middle GDP (green)**. The range box for this second group is from 4.7 to 6.2, suggesting more variation in happiness levels among countries in this group. This group does not have any outliers, and the median happiness level is 5.5.
3. **High-middle GDP (light blue)**. Now looking at the high-middle GDP group, the box range is from

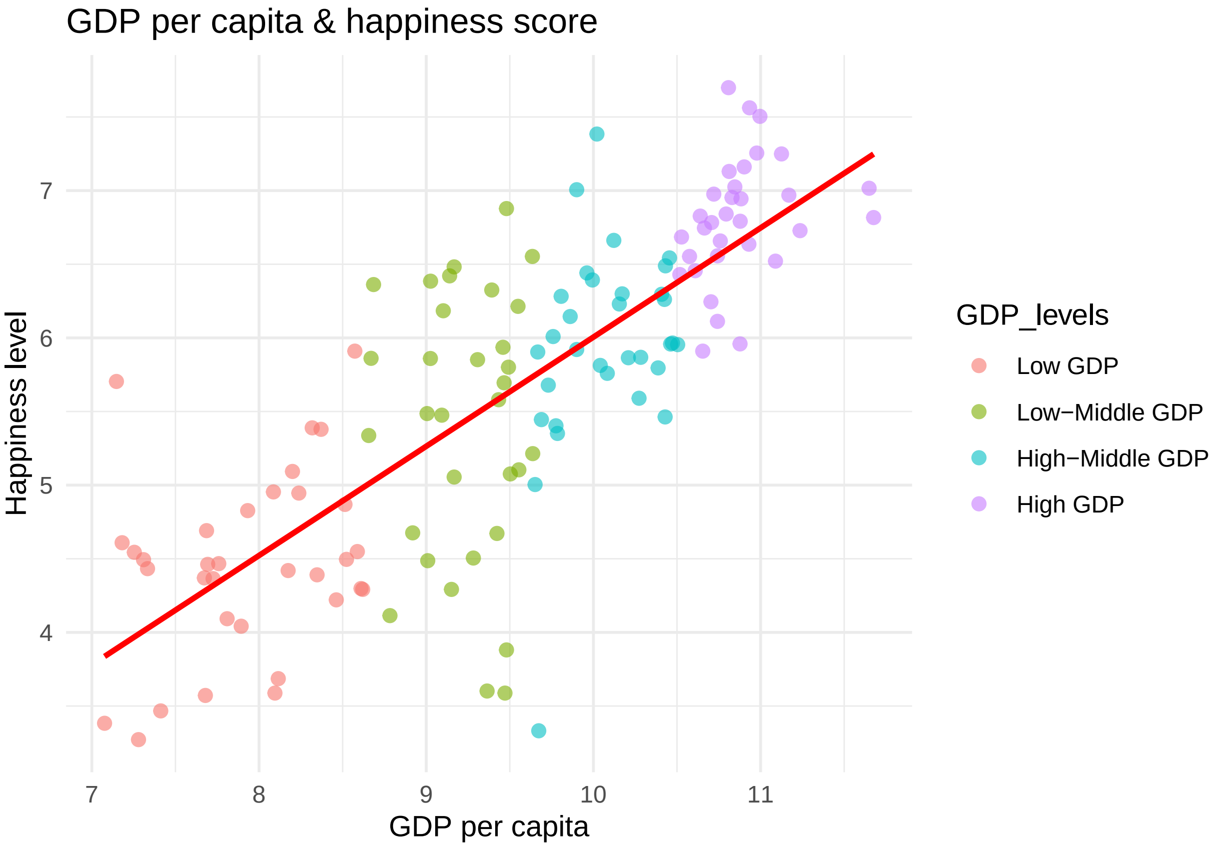
5.7 to 6.3, indicating that, like the low GDP group, countries in this group also have relatively similar happiness levels. This group has some outliers, with happiness levels above 7 and below 5. The median is 5.9.

4. **High GDP (purple)**. As expected, countries with high GDP levels have the highest happiness levels. The box is narrow, ranging from 6.6 to 7, showing very little variability. The median is at 6.8, and this group, like the lower-middle GDP group, does not have any outliers.

One conclusion is that the lower-middle GDP group has the most variability in happiness levels, while the high GDP group shows the least variability. Both the low GDP and high-middle GDP groups have some outliers.

```
ggplot(world_happiness_data, aes(x=Log.GDP.per.capita, y=Life.Ladder))+
  geom_point(aes(color=GDP_levels), size=2,alpha=0.6) +
  geom_smooth(method="lm", color="red", se=F) +
  labs(title="GDP per capita & happiness score",
       x = "GDP per capita",
       y = "Happiness level") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This is the final graph from our study on GDP. It is the same scatter plot made before, but the observation are divided by levels of GDP per capita. This makes the results clearer and easier to understand.

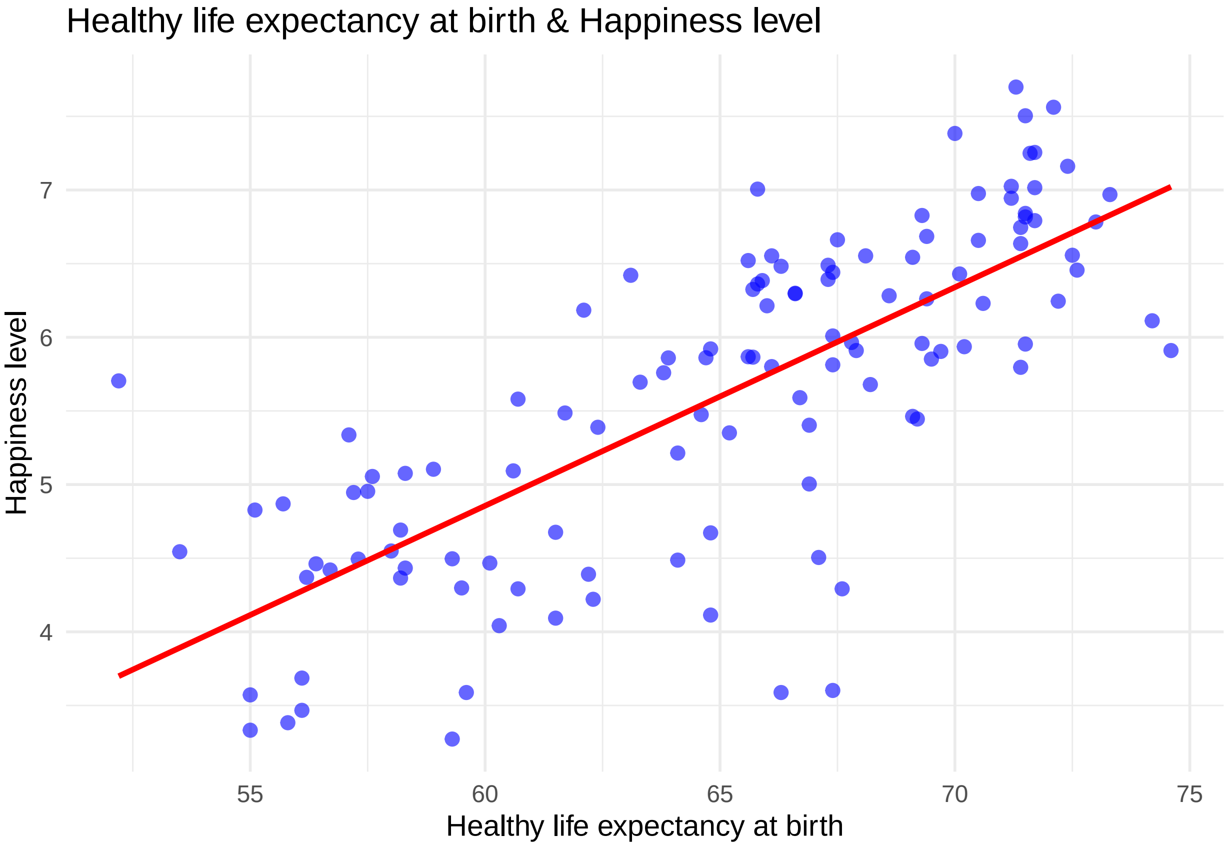
Healthy life expectancy at birth analysis

Another important question our study want to resolve is know if also the healthy life expectancy at birth has an impacts to the happiness of a country. Undertanding the relationship between these two variables is essential for determining if a nation should invest more in healthcare.

Similar to the analysis of GDP per capita, the study proceeds by creating a scatter plot.

```r
ggplot(na.omit(NewData),aes(x=Healthy.life.expectancy.at.birth, y=Life.Ladder)) +
        geom_point(size=2, alpha=0.6, color="blue") +
        geom_smooth(method="lm", color="red",se=F) +
        labs(title="Healthy life expectancy at birth & Happiness level",
            x = "Healthy life expectancy at birth",
            y = "Happiness level") +
        theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



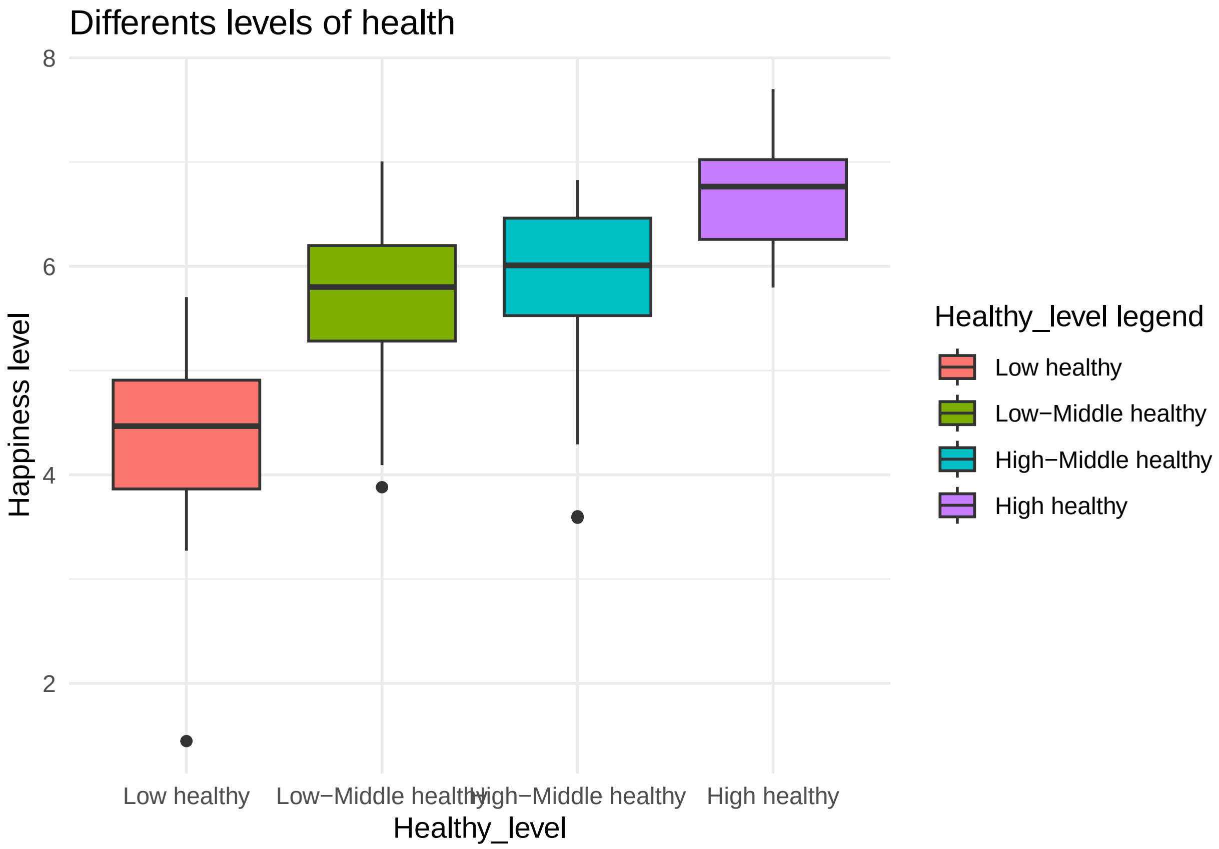
It is already known that the correlation between “Healthy life expectancy at birth” and “Happiness level” is 0.75. The graph suggest that the country where people live longer, with healthier lives have also higher levels of happiness.

Like for the analysis of GDP let’s proceed creating the box plot of this variable.

```r
world_happiness_data1=NewData %>%
  mutate (Healthy_level = cut(Healthy.life.expectancy.at.birth,
                              breaks=quantile(Healthy.life.expectancy.at.birth,
                                              probs = seq(0,1,0.25),na.rm=T),
                              include.lowest = T,
                              labels=c("Low healthy", "Low-Middle healthy", "High-Middle healthy", "Hig

world_happiness_data1 = world_happiness_data1 %>% filter(!is.na(Healthy.life.expectancy.at.birth))
world_happiness_data1 = world_happiness_data1 %>% filter(!is.na(Life.Ladder))
```

```
ggplot(world_happiness_data1, aes(x=Healthy_level, y=Life.Ladder, fill=Healthy_level)) +
  geom_boxplot() +
  labs(title="Differents levels of health",
       x = "Healthy_level",
       y = "Happiness level",
       fill = "Healthy_level legend") +
  theme_minimal()
```



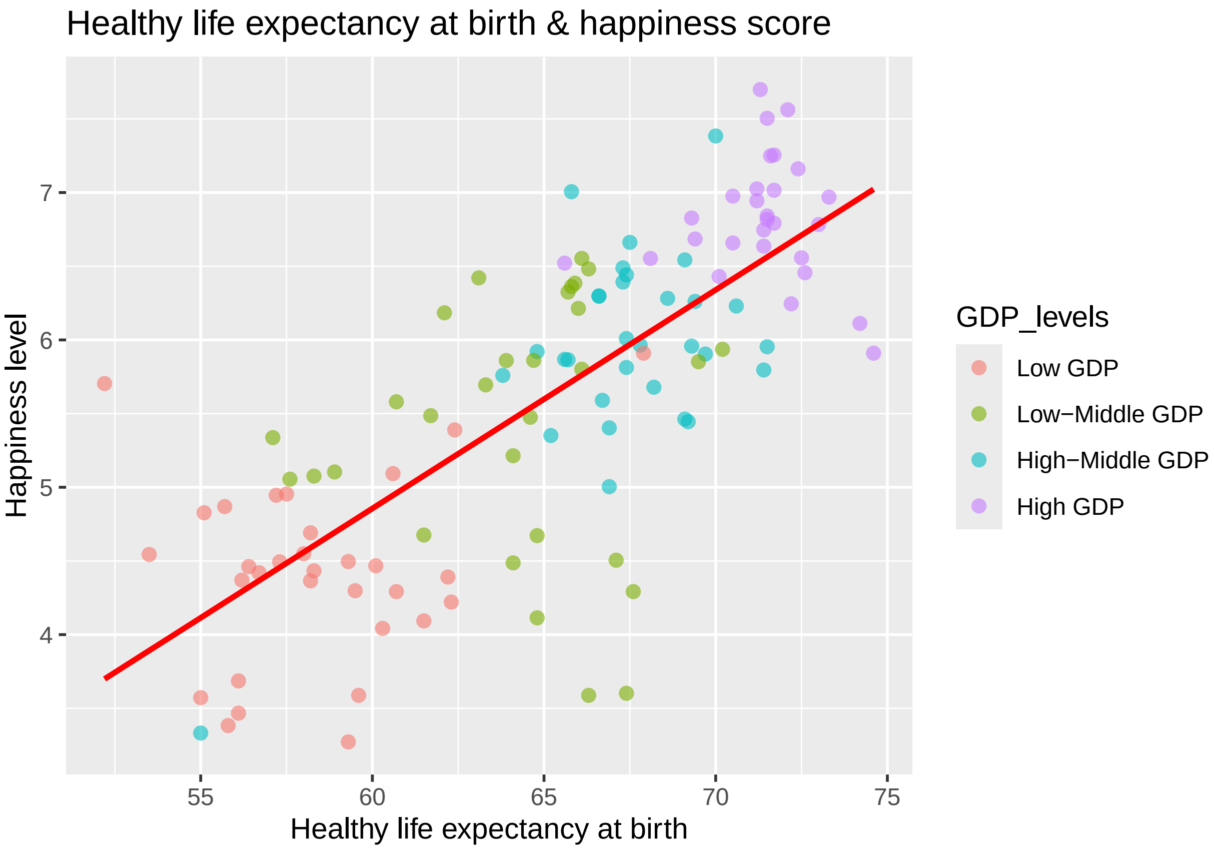
The graph shows that improving health increase happiness, especially for the country where the health level are high. The variability across all the four groups is low, with outliers present only in the first three groups. In conclusion, it can be said that an investing in the healthcare sector, such as research companies or nursing homes, may improve the happiness.

Are nations that have high levels of GDP also those with high levels of life expectancy at birth?

For the practice another interesting point is study the scatter plot between the "healthy life expectancy at birth" and the "happiness level".

```
ggplot(na.omit(world_happiness_data), aes(x=Healthy.life.expectancy.at.birth, y=Life.Ladder))+
  geom_point(aes(color=GDP_levels), size=2,alpha=0.6) +
  geom_smooth(method="lm", color="red", se=F) +
  labs(title="Healthy life expectancy at birth & happiness score",
       x = "Healthy life expectancy at birth",
       y = "Happiness level")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



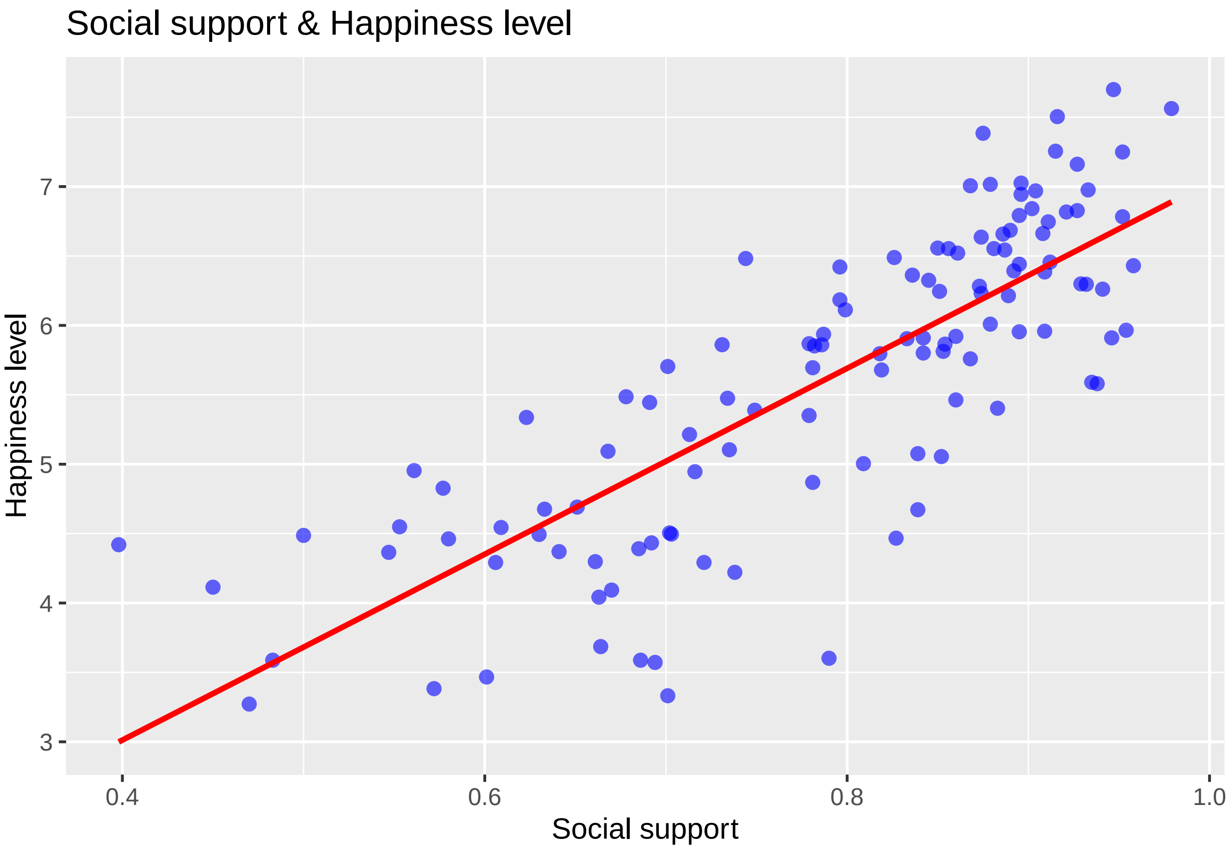
The graph represents the distribution of observations between the variables “Life.Ladder” and “Healthy.life.expectancy.at.birth” and the color of each observation is releted to the level of the GDP. As shown in the graph, countries with low levels of health and happiness tend to also have low GDP levels. This may be explained by the fact that countries with low GDP often cannot invest sufficiently in healthcare. Since both health and GDP are strongly correlated with happiness, it is likely that countries with low GDP per capita and low life expectancy at birth also tend to have lower happiness levels. On the other hand, nations with higher GDP and better healthcare generally have higher levels of happiness.

Social support analysis

This time the paper will analyze the information between social support and life ladder.

```r
ggplot(na.omit(NewData),aes(x=Social.support, y=Life.Ladder)) +
  geom_point(size=2, alpha=0.6, color="blue") +
  geom_smooth(method="lm", color="red",se=F) +
  labs(title="Social support & Happiness level",
       x = "Social support",
       y = "Happiness level")
```

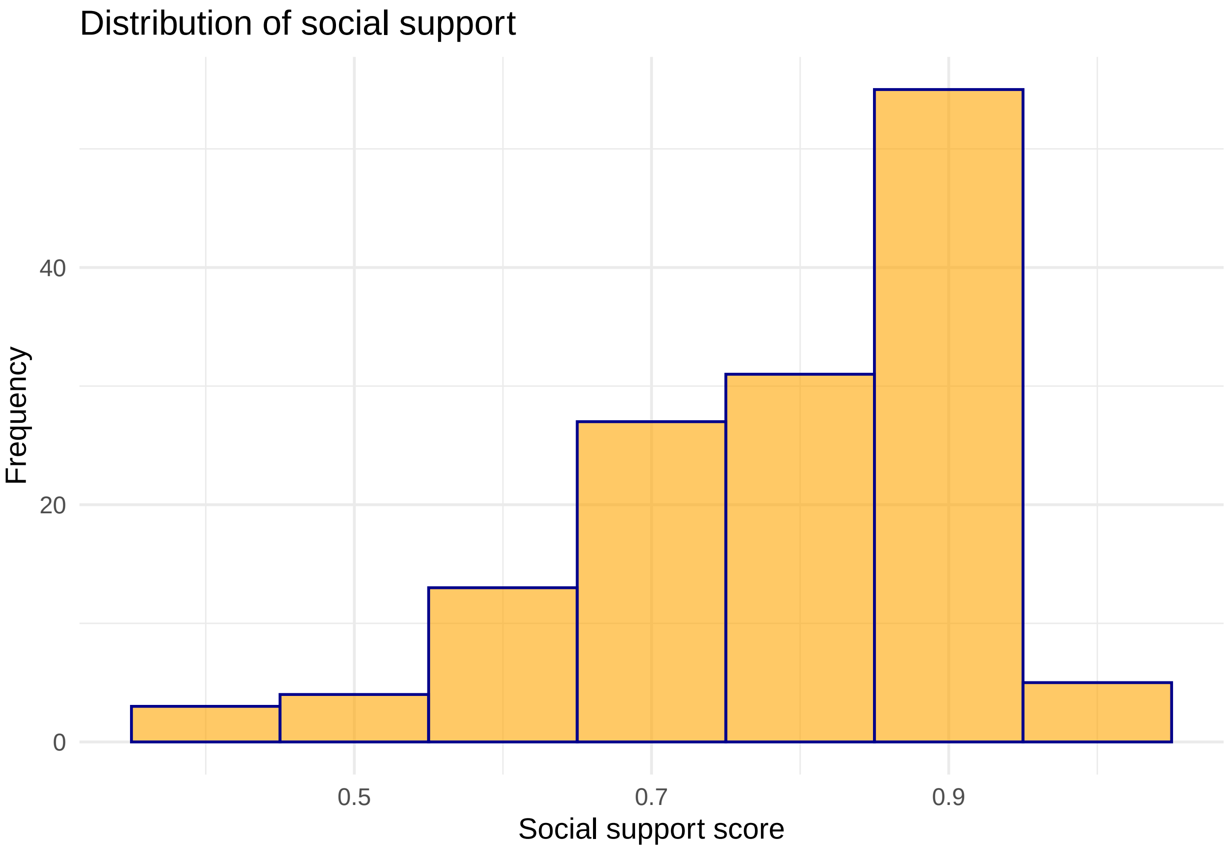
```
## `geom_smooth()` using formula = 'y ~ x'
```



As with the previous two variables in our case study, there is also a strong correlation between ‘Life Ladder’ and ‘Social support.’ In this case, we observe the highest correlation between the target variable and the explanatory variables, which is 0.80. Looking at the scatter plot, we can see that there are few observations with a social support value below 0.6. Easier access to social support is associated with nations where people are generally happier.

Now, let’s continue by creating an histrogram:

```
ggplot(NewData, aes(x = na.omit(Social.support))) +
  geom_histogram(binwidth = 0.1, fill = "orange", alpha = 0.6, color = "darkblue") +
  labs(title = "Distribution of social support",
       x = "Social support score",
       y = "Frequency") +
  theme_minimal()
```



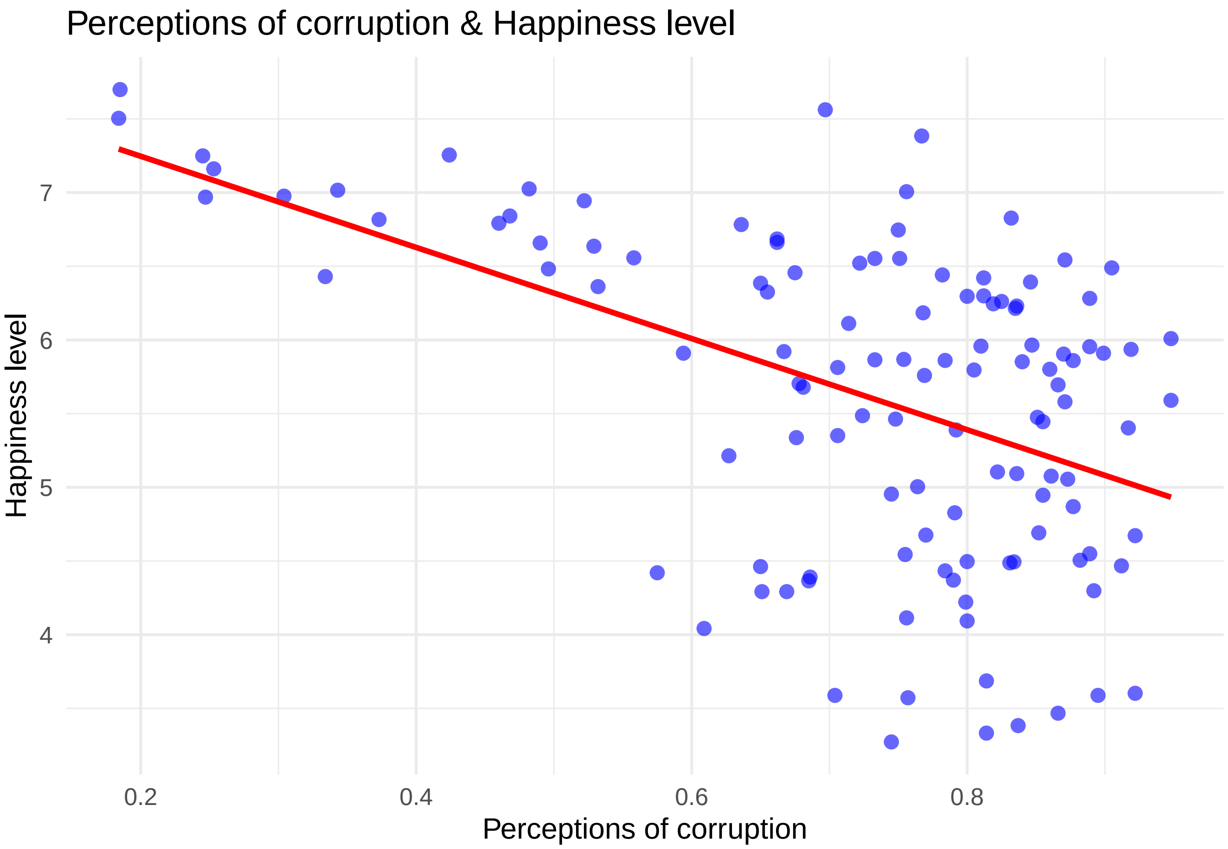
The histogram appears to follow a normal distribution with positive asymmetry. Most countries have "Social support" values around 0.9. However, only a few countries have scores below 0.5, which may indicate that, in most cases, people feel comfortable with their level of social support.

Perceptions of corruption analysis

An important aspect on which it is important to focus is that not all the variable are positive correlated with the happiness score. The perception of corruption is one of those variable who are negative correlate. Now, let's create the scatter plot:

```
ggplot(na.omit(NewData),aes(x=Perceptions.of.corruption, y=Life.Ladder)) +
  geom_point(size=2, alpha=0.6, color="blue") +
  geom_smooth(method="lm", color="red",se=F) +
  labs(title="Perceptions of corruption & Happiness level",
       x = "Perceptions of corruption",
       y = "Happiness level") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



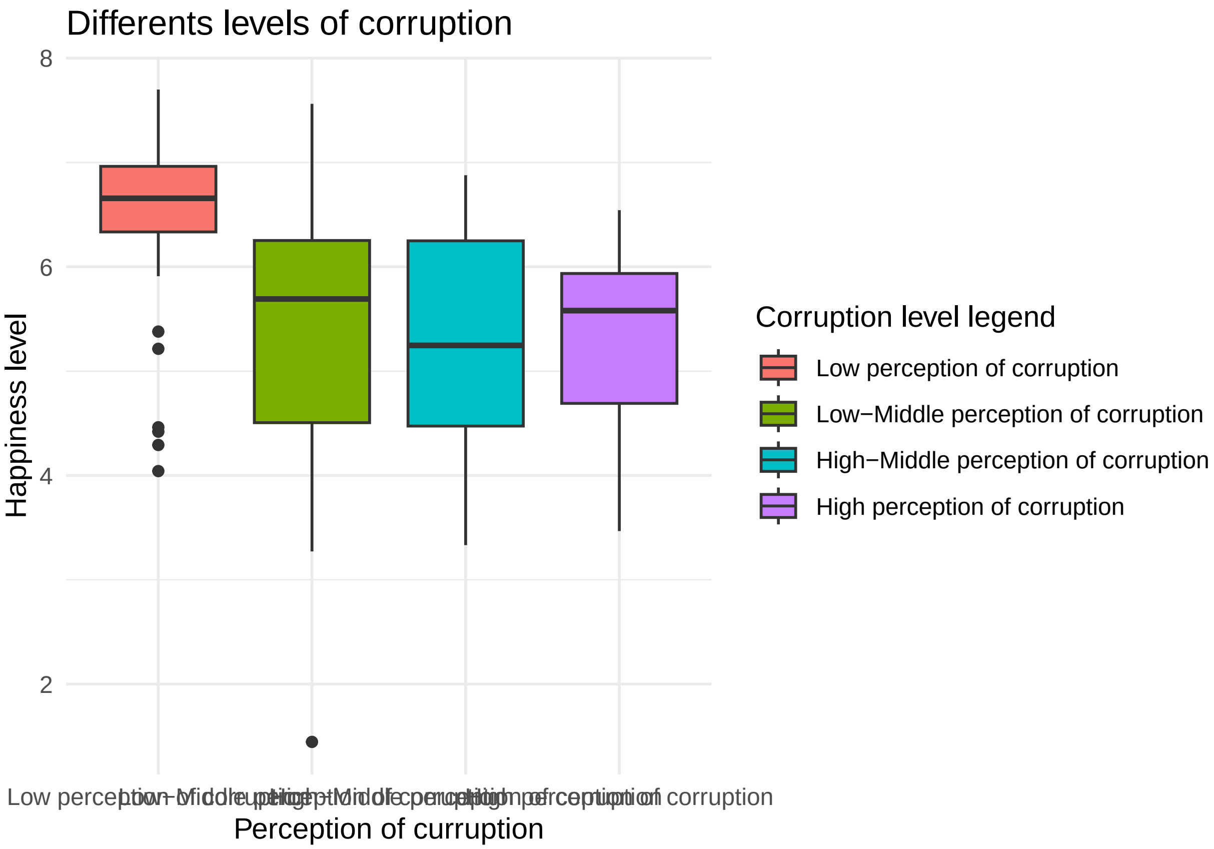
The graph show the relationship between perceptions of corruption and happiness levels. The red line demonstrates a negative correlation, indicating that higher perceptions of corruption are associated with lower happiness levels.The perception of corruption exceeds 0.5 for many observation, which may suggest low trust in government and institutions. The correlation coefficient of -0.49 signifies that as corruption perceptions increase, happiness levels tend to decline.

Lets explore the study further by constructing the box plot for this variable as well.

```
world_happiness_data3=NewData %>%
  mutate (Corruption_level = cut(Perceptions.of.corruption,
                              breaks=quantile(Perceptions.of.corruption,
                                              probs = seq(0,1,0.25),na.rm=T),
                              include.lowest = T,
                              labels=c("Low perception of corruption", "Low-Middle perception of corrupti

world_happiness_data3 = world_happiness_data3 %>% filter(!is.na(Perceptions.of.corruption))
world_happiness_data3 = world_happiness_data3 %>% filter(!is.na(Life.Ladder))

ggplot(world_happiness_data3, aes(x=Corruption_level, y=Life.Ladder, fill=Corruption_level)) +
  geom_boxplot() +
  labs(title="Differents levels of corruption",
       x = "Perception of curruption",
       y = "Happiness level",
       fill = "Corruption level legend") +
  theme_minimal()
```



The box plot compares happiness levels across the different groups based on the perception of corruption. Let's analyze each group:

- **Low perception of corruption (red)**, this group have a small spread of the box meaning that people in countries with low perception of corruption tent to have high happiness levels. However this group present some outliers, indicating that not every country where the perception of corruption is low means people are happy.
- **Low-middle perception of corruption (green)**, the spread of the box in this group is pretty big, from 4.5 to 6.2, indicating more variability.
- **High-middle perception of corruption (blue)**, the box of this group appears very similar to the previously studied group.5 It can be seen that again the levels of variability are more or less the same since this box also has happiness values ranging from 4.5 to 6.2. However, in this last case it can be seen that the median value is significantly lower than in the previous group, with a value of 5.2 vs. 5.8, and this is also true for the maximum happiness values, which are also lower. Note that this is also the group with the lowest median value
- **High perception of corruption (purple)**. Finally, as the last group, we analyze the one where there is the highest perceived level of corruption. Surprisingly from what might have been expected this group has higher minimum values than the two groups studied earlier. The same median value is also higher than the "High-middle perception of corruption" group.

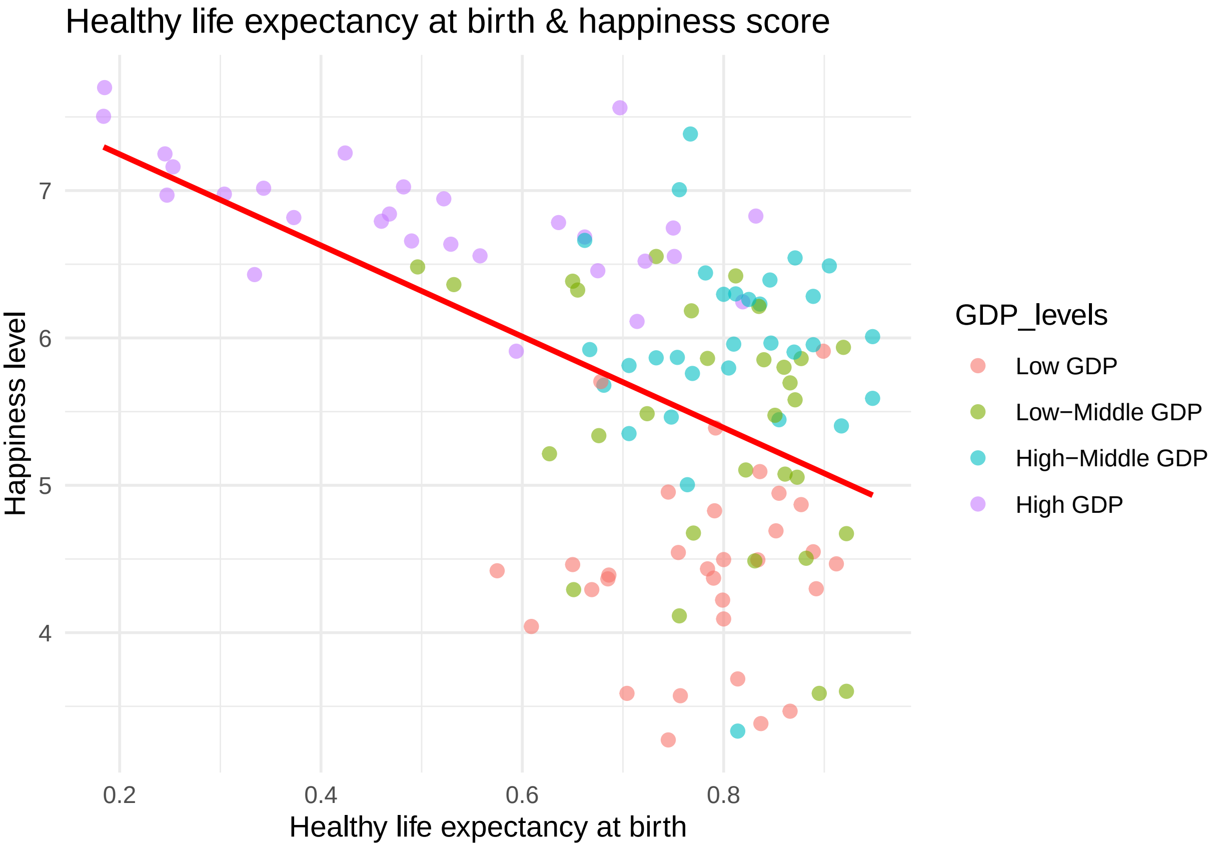
Some final considerations that can be made are that if perceived levels of corruption are very low then this significantly affects the happiness of a nation's residents (removed for some outliers). However, as noted from the box plots, if levels of perceived corruption are medium or high then this does not significantly impact a nation's happiness. A nation must be able to invest a portion in corruption prevention, stricter laws, transparently leaving access to information…

Is it possible that the level of corruption is also influenced by a nation's GDP level?

During the study, the possible correlation of a nation's wealth and corruption came to mind. Is it possible to think that poorer countries have bigger problems with corruption? does this mean that richer countries have fewer problems dealing with this issues?

```
ggplot(na.omit(world_happiness_data), aes(x=Perceptions.of.corruption, y=Life.Ladder))+
  geom_point(aes(color=GDP_levels), size=2,alpha=0.6) +
  geom_smooth(method="lm", color="red", se=F) +
  labs(title="Healthy life expectancy at birth & happiness score",
       x = "Healthy life expectancy at birth",
       y = "Happiness level") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

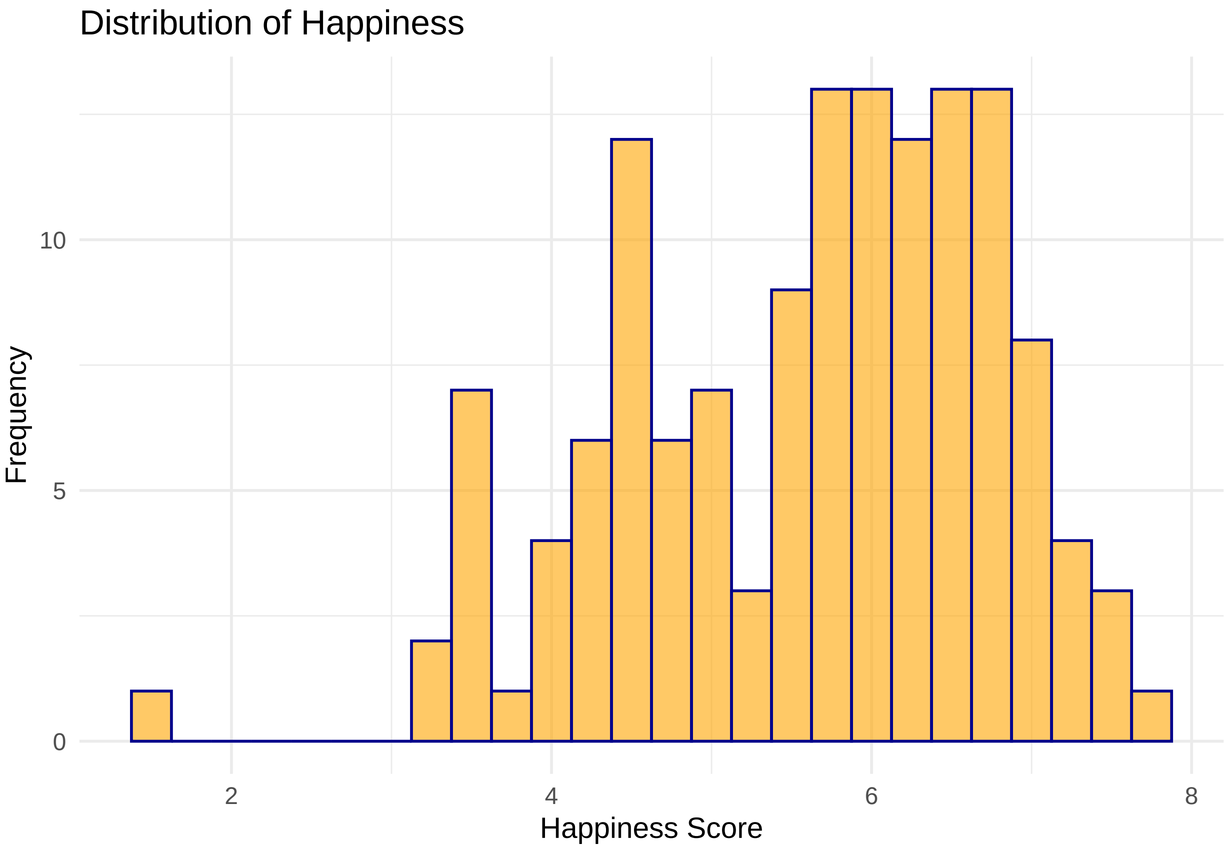


The hypotheses we had set ourselves were well founded. As can be seen from the scatter plot above, we note that all those observations before 0.5, concerning the value of "healthy life expectancy at birth", are also those that have higher levels of happiness. Conversely, when it comes to those countries where we have medium-low GDP, we note that they have higher levels of corruption, which also support lower levels of happiness. It can also be noted that the blue observations, despite having the same levels of corruption as the orange observations, have higher levels of happiness. This confirms that corruption is not the only element that affects happiness and that GDP has a greater influence.

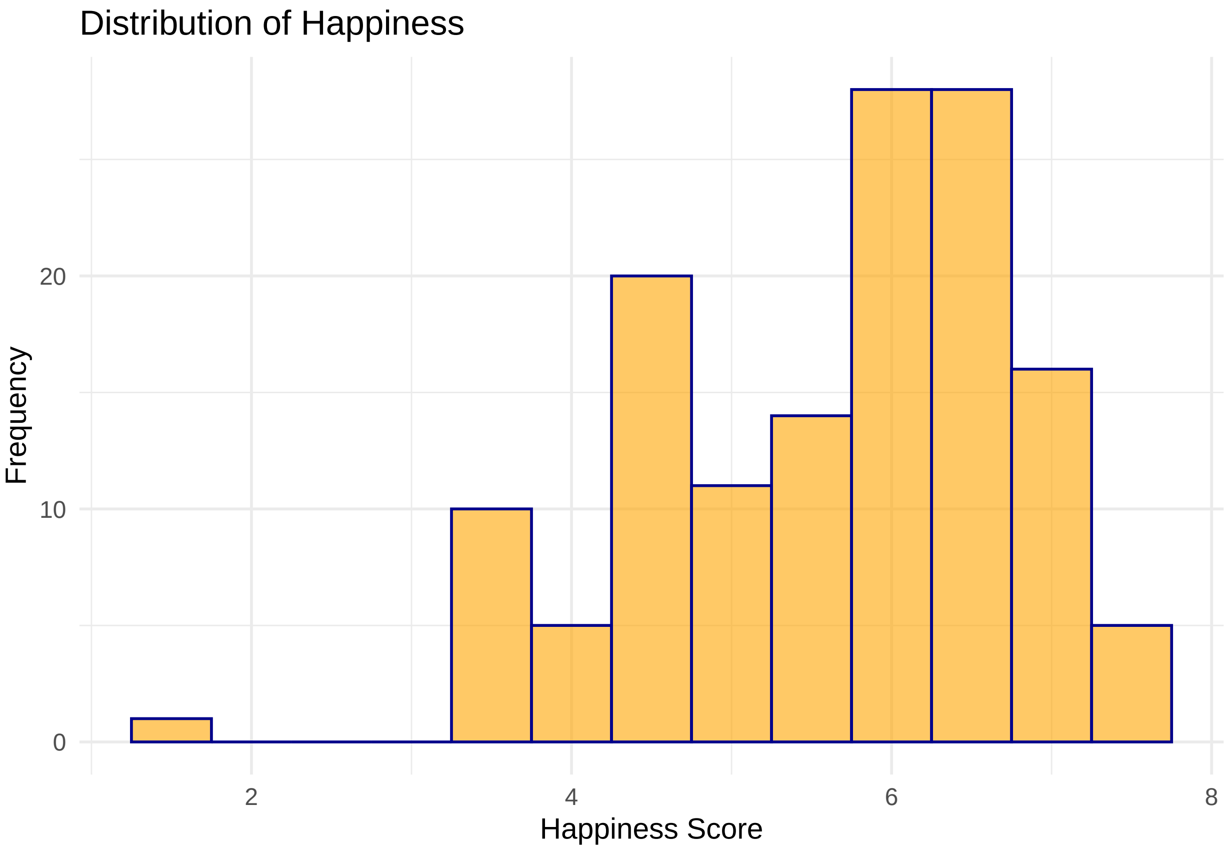
Happiness analysis

The last variable the study wants to focus on is our target variable "life ladder," which measures the happiness. Proceed by creating four histograms, each with a different total number of bars.

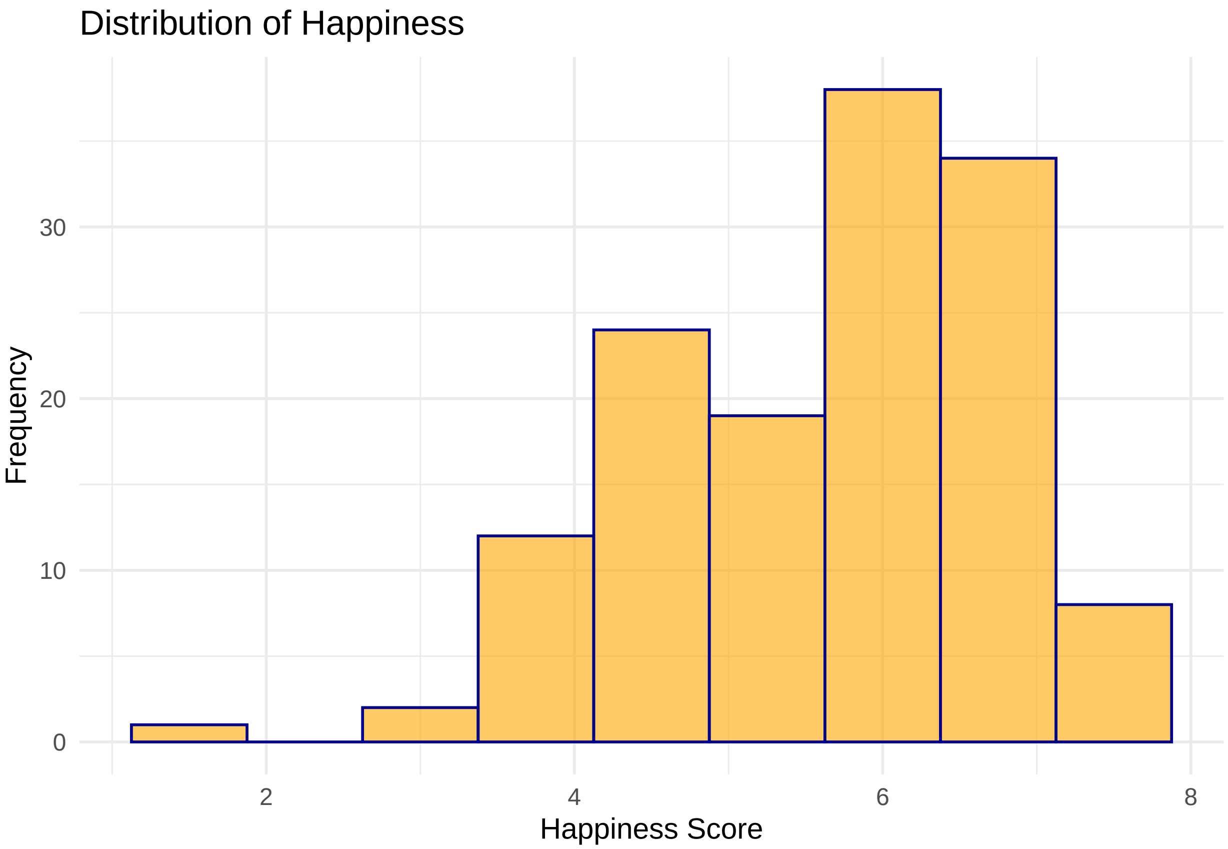
```r
ggplot(NewData, aes(x = na.omit(Life.Ladder))) +
  geom_histogram(binwidth = 0.25, fill = "orange", alpha = 0.6, color = "darkblue") +
  labs(title = "Distribution of Happiness",
       x = "Happiness Score",
       y = "Frequency") +
  theme_minimal()
```



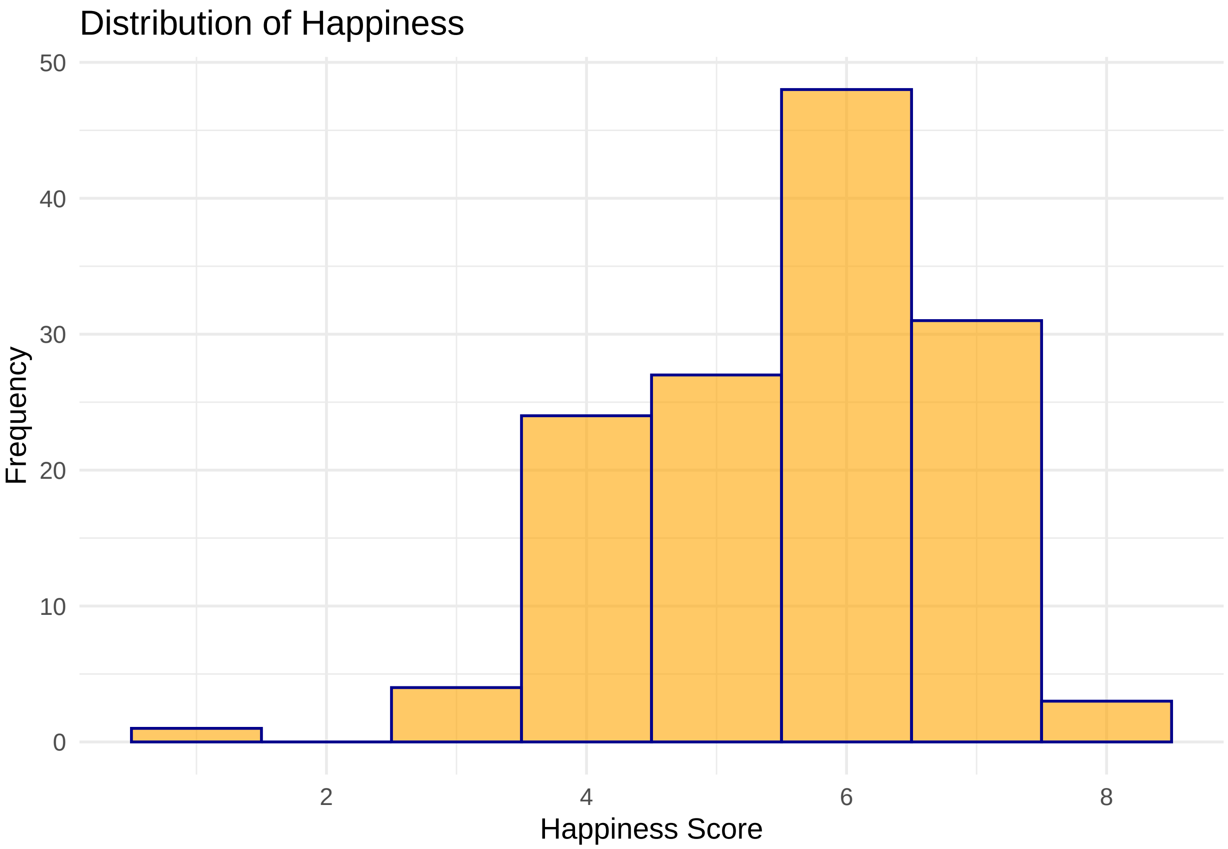
```r
ggplot(NewData, aes(x = na.omit(Life.Ladder))) +
  geom_histogram(binwidth = 0.5, fill = "orange", alpha = 0.6, color = "darkblue") +
  labs(title = "Distribution of Happiness",
       x = "Happiness Score",
       y = "Frequency") +
  theme_minimal()
```



```
ggplot(NewData, aes(x = na.omit(Life.Ladder))) +
  geom_histogram(binwidth = 0.75, fill = "orange", alpha = 0.6, color = "darkblue") +
  labs(title = "Distribution of Happiness",
       x = "Happiness Score",
       y = "Frequency") +
  theme_minimal()
```



```
ggplot(NewData, aes(x = na.omit(Life.Ladder))) +
  geom_histogram(binwidth = 1, fill = "orange", alpha = 0.6, color = "darkblue") +
  labs(title = "Distribution of Happiness",
       x = "Happiness Score",
       y = "Frequency") +
  theme_minimal()
```



Let’s analyze each of the graph created:

1. The first histogram shows a distribution with multiple bars, indicating a relatively spread-out range of happiness scores. The highest frequencies are located around a happiness score of 6. There are only a few countries (perhaps only one or two) with happiness scores below 3, suggesting that very low happiness scores are rare in this data-set.

2. In the second histogram, the number of classes (bins) has been significantly reduced, leading to a different representation of the data. The highest frequencies are still around 6.25. However, while the previous histogram showed frequencies around 12, this time the frequencies have increased significantly, with values exceeding 25. This suggests that reducing the number of bins creates a more concentrated distribution, making the central tendency around 6.25 more evident.This third graph reconfirms that the happiness level of most observations is around a value of 6, a value very close to the median of this variable which is 5.863 (discovered in the EDA).

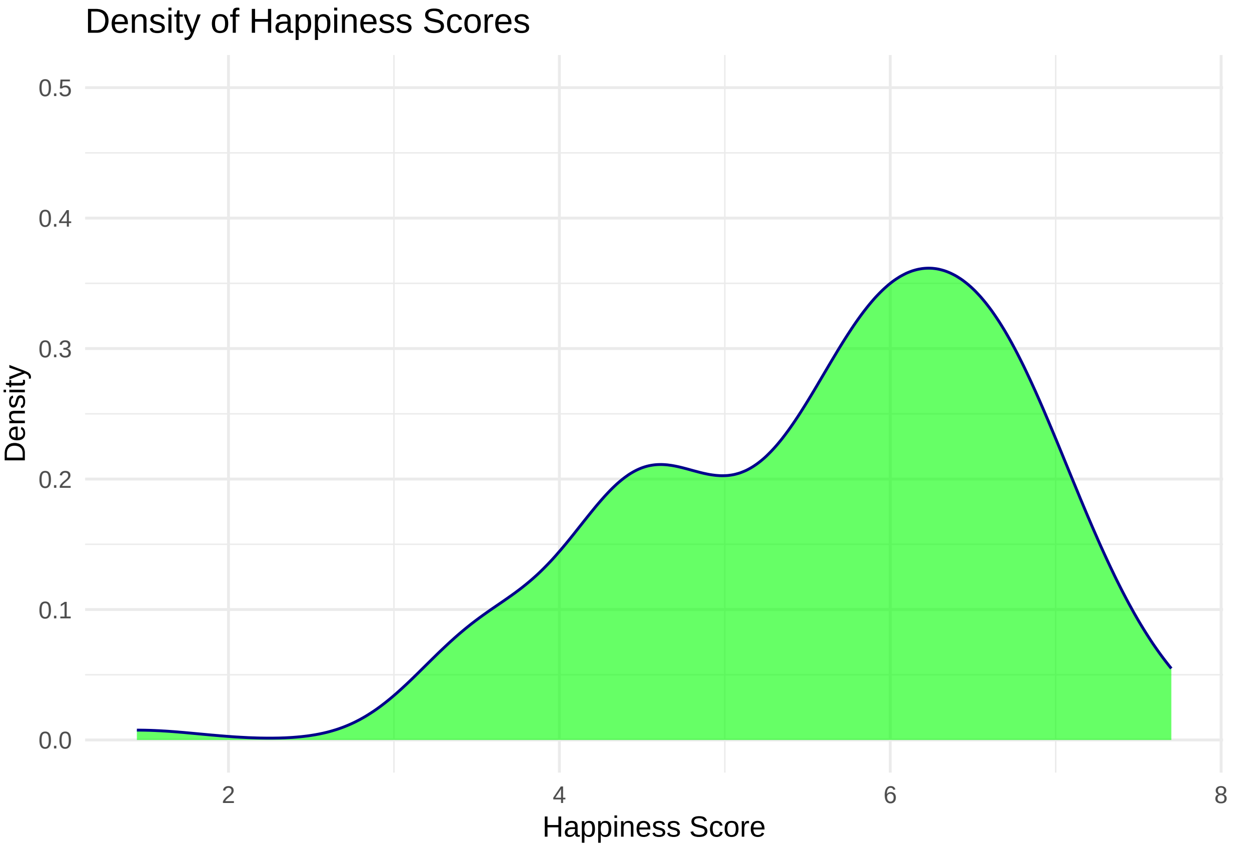
3. The last graph shows an histograms that it is close to a normal distribution, but it’s not perfectly symmetrical. Also this time the peak is around 6, which have around 47 observation. In conclusion, it can be said that this graph reminds one of a normal distribution with positive asymmetry, where most of the values are concentrated among the central values (4 to 6).

4. The fourth histogram presents a distribution that is close to a normal distribution, but it’s not perfectly symmetrical. The peak frequency occurs around a happiness score of 6, with approximately 47 observations at this value. Although the distribution is similar to a normal distribution, it exhibits positive asymmetry, where the majority of values are concentrated between 4 and 6, with a few higher values extending the tail to the right.

In conclusion across the four histograms, we observe a progression toward a more concentrated distribution of happiness scores. The first histogram shows a broader spread, with multiple peaks, while the others reveal a stronger concentration of scores around 6. For the purposes of this study, the second or third graph likely provides the most useful representation, as they highlight the central concentration without oversimplifying or overcomplicating the distribution.

A really good way to have a better understanding of how the data are really distributed is through the creation of the density graph.

```
ggplot(NewData, aes(x = na.omit(Life.Ladder))) +
  geom_density(fill = "green", alpha = 0.6, color = "darkblue") +
  labs(title = "Density of Happiness Scores",
       x = "Happiness Score",
       y = "Density") +
  ylim(0,0.5) +
  theme_minimal()
```



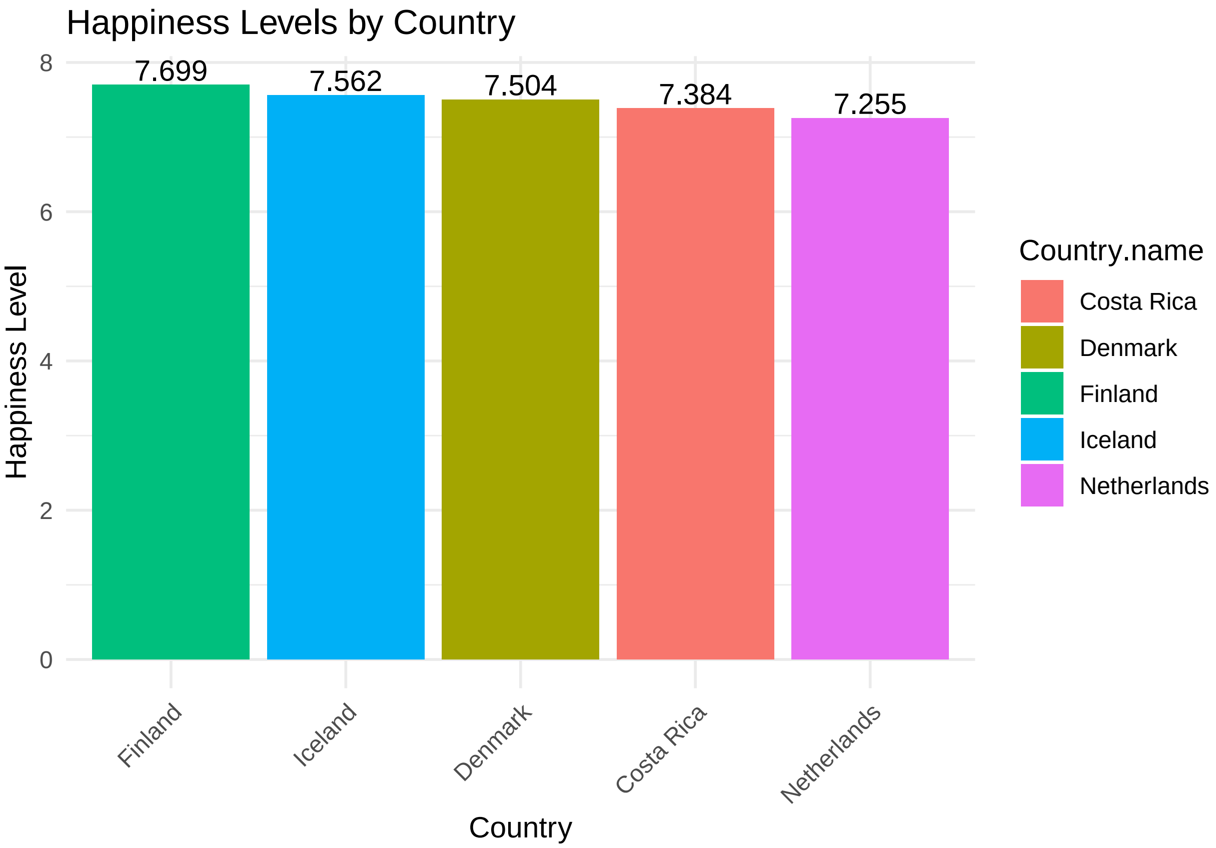
This graph reinforces our previous hypotheses: the peak occurs around 6.25. Additionally, we observe that the distribution tends to increase until it reaches his peak, except for a slight dip around 4.5, where the frequency decreases temporarily before rising again. This pattern is also clearly visible in graph 1.

Best and worst nations in terms of happiness

Before concluding the project, it is important to find out which are the five happiest and five saddest countries and also proceed to compare the happiest and saddest nation as well. This will make it possible to give more accurate and effective conclusions.

```
top_5_nations = NewData %>%
  arrange(desc(Life.Ladder)) %>%
  head(5)

ggplot(top_5_nations, aes(x = reorder(Country.name, -Life.Ladder), y = Life.Ladder, fill = Country.name
  geom_bar(stat = "identity") +
  geom_text(aes(label = Life.Ladder), vjust = -0.2) +
  theme_minimal() +
  labs(title = "Happiness Levels by Country",
       x = "Country",
       y = "Happiness Level") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The five happiest countries in the world in order are Finland, Iceland, Denmark, Costa rica, Netherlands. Let's try to understand what these five countries have in common.

```
print(top_5_nations)
```

```
##   Country.name year Life.Ladder Log.GDP.per.capita Social.support
## 1      Finland 2023       7.699             10.808          0.947
## 2      Iceland 2023       7.562             10.934          0.979
## 3      Denmark 2023       7.504             10.996          0.916
## 4   Costa Rica 2023       7.384             10.021          0.875
## 5  Netherlands 2023       7.255             10.977          0.915
##   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices Generosity
## 1                             71.3                        0.943     -0.001
## 2                             72.1                        0.918      0.299
## 3                             71.5                        0.923      0.089
## 4                             70.0                        0.933     -0.067
## 5                             71.7                        0.847      0.223
##   Perceptions.of.corruption Positive.affect Negative.affect
## 1                     0.185           0.717           0.173
## 2                     0.697           0.793           0.185
## 3                     0.184           0.757           0.229
## 4                     0.767           0.806           0.282
## 5                     0.424           0.693           0.202
```

First, it can be said that 4 of 5 of these countries are located in Europe, more specifically in northern Europe, while Costa Rica is located in Central America.

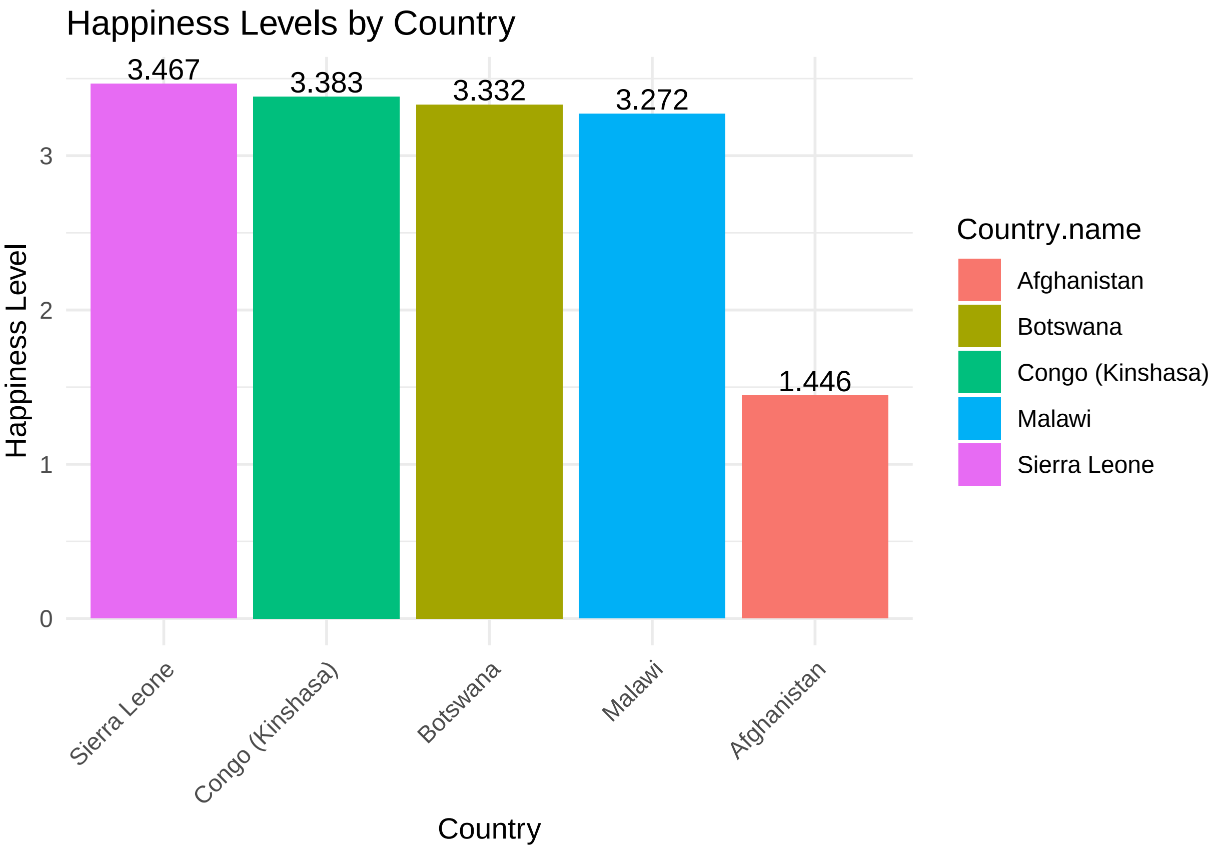
We observe that all countries have extremely high GDP values, exceeding 10. Similarly high values are also found for all variables positively correlated with “Life Ladder”. For almost all the countries in this study, the life expectancy at birth is around 71 years. Now, let’s focus on the variable “generosity.” The level of generosity varies across these countries, with some having values as high as 0.3, while others have negative values, going as low as -0.07. This confirms that, since generosity is weakly correlated with our target variable, its values, whatever they may be, do not significantly impact the happiness of a country.

Another aspect worth considering is corruption. Despite these being the five happiest countries, it can be seen that this variable also exhibits different values. For the countries Finland, Denmark, and the Netherlands, the values are medium-low, for Iceland, the value is medium-high, and finally for Costa Rica, the value is very high at 0.767. This confirms that, even if the perception of corruption in a country is high, it does not necessarily have a significant impact on happiness.

We now proceed with the analysis of the least happy countries

```
least_5_nations = NewData %>%
  arrange(Life.Ladder) %>%
  head(5)

ggplot(least_5_nations, aes(x = reorder(Country.name, -Life.Ladder), y = Life.Ladder, fill = Country.nam
  geom_bar(stat = "identity") +
  geom_text(aes(label = Life.Ladder), vjust = -0.2) +
  theme_minimal() +
  labs(title = "Happiness Levels by Country",
       x = "Country",
       y = "Happiness Level") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



In this study the five least happy countries are “Sierra Leone”, “Congo”, “Botswana”, “Malawi” and “Afganistan”. Let’s try to understand more.

```
print(least_5_nations)
```

```
##       Country.name year Life.Ladder Log.GDP.per.capita Social.support
## 1      Afghanistan 2023       1.446                 NA          0.368
## 2           Malawi 2023       3.272              7.279          0.470
## 3         Botswana 2023       3.332              9.673          0.701
## 4 Congo (Kinshasa) 2023       3.383              7.076          0.572
## 5     Sierra Leone 2023       3.467              7.412          0.601
##   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices Generosity
## 1                             55.2                        0.228         NA
## 2                             59.3                        0.738      0.014
## 3                             55.0                        0.741     -0.264
## 4                             55.8                        0.687      0.152
## 5                             56.1                        0.694      0.101
##   Perceptions.of.corruption Positive.affect Negative.affect
## 1                     0.738           0.261           0.460
## 2                     0.745           0.520           0.338
## 3                     0.814           0.657           0.247
## 4                     0.837           0.546           0.497
## 5                     0.866           0.504           0.430
```

In this case, the countries are located in Central Africa, Southern Africa, and South Asia.

Let’s proceed with a more depth analysis. We immediately notice that the country considered the least happy does not have data for the GDP variable, which is also the one that has the greatest impact on the happiness level. For the other countries, the GDP values are around 7, except for Botswana, which has a value of around 9.6. The life expectancy at birth is extremely low, with values around 55 years. Regarding the "Perception of corruption" variable, all nations have relatively high values.

Let's proceed by analyzing the happiest and the saddest country.

```
NewDataOrder = NewData %>%
  arrange(desc(Life.Ladder))

first_observation = head(NewDataOrder,1)
last_observation = tail(NewDataOrder, 1)

Best_and_worse = rbind(first_observation,last_observation)
print(Best_and_worse)
```

```
##     Country.name year Life.Ladder Log.GDP.per.capita Social.support
## 1        Finland 2023       7.699             10.808          0.947
## 138  Afghanistan 2023       1.446                 NA          0.368
##     Healthy.life.expectancy.at.birth Freedom.to.make.life.choices Generosity
## 1                               71.3                        0.943     -0.001
## 138                             55.2                        0.228         NA
##     Perceptions.of.corruption Positive.affect Negative.affect
## 1                       0.185           0.717           0.173
## 138                     0.738           0.261           0.460
```

As already evident from the previously created histograms, the happiest country is Finland, while the saddest is Afghanistan. The difference in happiness scores is 6.253, indicating a significant gap. It is not possible to compare the GDP levels between the two countries, as there is no GDP data available for Afghanistan. Being born in a country like Finland allows to live approximately 16 years longer than in Afghanistan. Additionally, there is a notable difference in the "social support" variable. The freedom to make choices in Afghanistan is significantly lower compared to a country like Finland, where corruption is also perceived to be lower, with a score of 0.185 compared to 0.738 in Afghanistan.

Conclusions

In conclusion, we can conclude that countries with higher life expectancy, higher GDP per capita, greater freedom in making life decisions, and lower perceived corruption are the ones where people are happier to live.

Based on the five happiest countries and the saddest ones, it can be hypothesized that Central and Northern American countries, European nations, and Central Asian countries are the happiest. Conversely, countries located in South America, Africa, the Middle East, and South Asia are likely to be places where it is more difficult to be happy.

Every government should invest a portion of its resources to combat corruption by making data more accessible, promoting prevention in schools, and enforcing stricter laws. There should also be investments in healthcare, such as hospitals and nursing homes. Finally, each country should strive to increase its GDP.