World Happiness Report Analysis

2024-10-27

Libraries

```
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
```

```
library(ggplot2)
```

Loading In the Dataset

The first step involves loading in the 2019.csv dataset which is stored locally on my desktop.

```
df <-read.csv("~/Desktop/2019.csv", sep=",",header=TRUE)
```

Introducing the Dataset

```
head(df)
```

```
##   Overall.rank Country.or.region Score GDP.per.capita Social.support
## 1            1           Finland 7.769          1.340          1.587
## 2            2           Denmark 7.600          1.383          1.573
## 3            3            Norway 7.554          1.488          1.582
## 4            4           Iceland 7.494          1.380          1.624
## 5            5       Netherlands 7.488          1.396          1.522
## 6            6       Switzerland 7.480          1.452          1.526
##   Healthy.life.expectancy Freedom.to.make.life.choices Generosity
## 1                   0.986                        0.596      0.153
## 2                   0.996                        0.592      0.252
## 3                   1.028                        0.603      0.271
## 4                   1.026                        0.591      0.354
## 5                   0.999                        0.557      0.322
## 6                   1.052                        0.572      0.263
##   Perceptions.of.corruption
## 1                     0.393
## 2                     0.410
## 3                     0.341
## 4                     0.118
## 5                     0.298
```

```
## 6                    0.343
```

About the Dataset

It is first essential to understand and comprehend the dataset. In essence, the dataset captures and provides a well rounded outlook of the Happiness levels on a global scale alongside of a multitude of related factors which contribute to the overall happiness of a nation. The dataset consists of 9 columns: the Happiness score, Country or Region, Overall Rank, GDP Per Capita, Social Support, Healthy Life, Freedom to make Life Choices, Generosity and Perceptions of Corruption. Moreover, there are 156 rows which represent countries all the way from Finland to South Sudan. Another interesting thing to note is that this dataset is from pre-Covid era (2019) which gives us a glimpse of global happiness before the pandemic had struck.

Aim of The Project

The primary aim of this project is to deeply examine how economic, social and governmental factors impact the overall happiness levels among the given countries. In other words, the extent to which these factors impact the happiness levels is the primary focus of this study. I will be deeply examining four variables, particularly, the GDP per capita, social support, life expectancy and freedom to make life choices. At the end of the analysis, I aim to rank these four factors from most important to least important on the basis of their impact on the happiness levels of a country. The primary objective is subdivided into smaller problems which all sum up to provide the bigger picture. Once the dataset is preprocessed, it will be analyzed and visualized in order to better understand just how much these factors influence the overall happiness of a nation. Also, the others variables such as generosity and level of corruption will not be observed because they are far more complex and controversial factors which might complicate our analysis.

Data Preprocessing

Before analyzing the data and derving relationships, it is first necessary to ensure that it is suitable for use. The following steps were taken by me to ensure this:

Structure and Summary of the Data Frame

Let's start off by inspecting the structure and summary of the data frame. This will help us understand the dataset better and assist us in the wrangling/munging/cleaning process.

```
str(df)
```

```
## 'data.frame':    156 obs. of  9 variables:
##  $ Overall.rank                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country.or.region           : chr  "Finland" "Denmark" "Norway" "Iceland" ...
##  $ Score                       : num  7.77 7.6 7.55 7.49 7.49 ...
##  $ GDP.per.capita              : num  1.34 1.38 1.49 1.38 1.4 ...
##  $ Social.support              : num  1.59 1.57 1.58 1.62 1.52 ...
##  $ Healthy.life.expectancy     : num  0.986 0.996 1.028 1.026 0.999 ...
##  $ Freedom.to.make.life.choices: num  0.596 0.592 0.603 0.591 0.557 0.572 0.574 0.585 0.584 0.532 ..
##  $ Generosity                  : num  0.153 0.252 0.271 0.354 0.322 0.263 0.267 0.33 0.285 0.244 ...
##  $ Perceptions.of.corruption   : num  0.393 0.41 0.341 0.118 0.298 0.343 0.373 0.38 0.308 0.226 ...
```

```
summary(df)
```

```
##   Overall.rank    Country.or.region      Score       GDP.per.capita
##  Min.   :  1.00   Length:156         Min.   :2.853   Min.   :0.0000
##  1st Qu.: 39.75   Class :character   1st Qu.:4.545   1st Qu.:0.6028
##  Median : 78.50   Mode  :character   Median :5.380   Median :0.9600
##  Mean   : 78.50                      Mean   :5.407   Mean   :0.9051
##  3rd Qu.:117.25                      3rd Qu.:6.184   3rd Qu.:1.2325
```

```
##  Max.   :156.00                      Max.   :7.769   Max.   :1.6840
##  Social.support  Healthy.life.expectancy Freedom.to.make.life.choices
##  Min.   :0.000   Min.   :0.0000          Min.   :0.0000
##  1st Qu.:1.056   1st Qu.:0.5477          1st Qu.:0.3080
##  Median :1.272   Median :0.7890          Median :0.4170
##  Mean   :1.209   Mean   :0.7252          Mean   :0.3926
##  3rd Qu.:1.452   3rd Qu.:0.8818          3rd Qu.:0.5072
##  Max.   :1.624   Max.   :1.1410          Max.   :0.6310
##    Generosity     Perceptions.of.corruption
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1087   1st Qu.:0.0470
##  Median :0.1775   Median :0.0855
##  Mean   :0.1848   Mean   :0.1106
##  3rd Qu.:0.2482   3rd Qu.:0.1412
##  Max.   :0.5660   Max.   :0.4530
```

NA values

```
colSums(is.na(df))
```

```
##                 Overall.rank            Country.or.region
##                            0                            0
##                        Score               GDP.per.capita
##                            0                            0
##               Social.support      Healthy.life.expectancy
##                            0                            0
## Freedom.to.make.life.choices                   Generosity
##                            0                            0
##    Perceptions.of.corruption
##                            0
```

Fortunately, there are no missing values in the dataset which is a good start.

Duplicate Values

```
sum(duplicated(df))
```

```
## [1] 0
```

There are no duplicate values either. This is a positive sign as well since it implies that the dataset is well mapped and clean.

Zero Values

Although there are no NA values, the summary of the dataset indicates the presence of 0 values. I am somewhat opposed to having zero values for variables such as GDP Per Capita because there is absolutely no functioning country in the world that has a Per Capita income of zero. The same logic applies for other variables as well which have zero values. Let's remove the zero values from all of these variables entirely. For this purpose, I will be using filter from dplyr and then use summary to reinspect the dataframe.

```
df<-df%>%
  filter(GDP.per.capita!=0,Social.support!=0,Healthy.life.expectancy!=0,
         Freedom.to.make.life.choices!=0,Generosity!=0,Perceptions.of.corruption!=0)
summary(df)
```

```
##   Overall.rank    Country.or.region      Score       GDP.per.capita
```

```
##  Min.   :  1.00   Length:150         Min.   :2.853   Min.   :0.0460
##  1st Qu.: 38.25   Class :character   1st Qu.:4.566   1st Qu.:0.6248
##  Median : 76.50   Mode  :character   Median :5.428   Median :0.9840
##  Mean   : 76.91                      Mean   :5.450   Mean   :0.9210
##  3rd Qu.:115.75                      3rd Qu.:6.197   3rd Qu.:1.2377
##  Max.   :156.00                      Max.   :7.769   Max.   :1.6840
##  Social.support  Healthy.life.expectancy Freedom.to.make.life.choices
##  Min.   :0.378   Min.   :0.1680          Min.   :0.0100
##  1st Qu.:1.061   1st Qu.:0.5590          1st Qu.:0.3140
##  Median :1.285   Median :0.7965          Median :0.4220
##  Mean   :1.225   Mean   :0.7378          Mean   :0.3989
##  3rd Qu.:1.455   3rd Qu.:0.8832          3rd Qu.:0.5078
##  Max.   :1.624   Max.   :1.1410          Max.   :0.6310
##    Generosity     Perceptions.of.corruption
##  Min.   :0.0250   Min.   :0.0040
##  1st Qu.:0.1092   1st Qu.:0.0505
##  Median :0.1775   Median :0.0860
##  Mean   :0.1863   Mean   :0.1117
##  3rd Qu.:0.2520   3rd Qu.:0.1417
##  Max.   :0.5660   Max.   :0.4530
```

As the summary indicates, the zero values have been removed. The revised dataframe seems much more realistic now. The dataset is fairly clean from the get go which means that there isn't much preprocessing which is required. Note: I have still included the Generosity and Percpetion.of.corruption variables in order to provide a complete view of the data. However, these variables will not be observed individually.

Data Analysis and Visualization

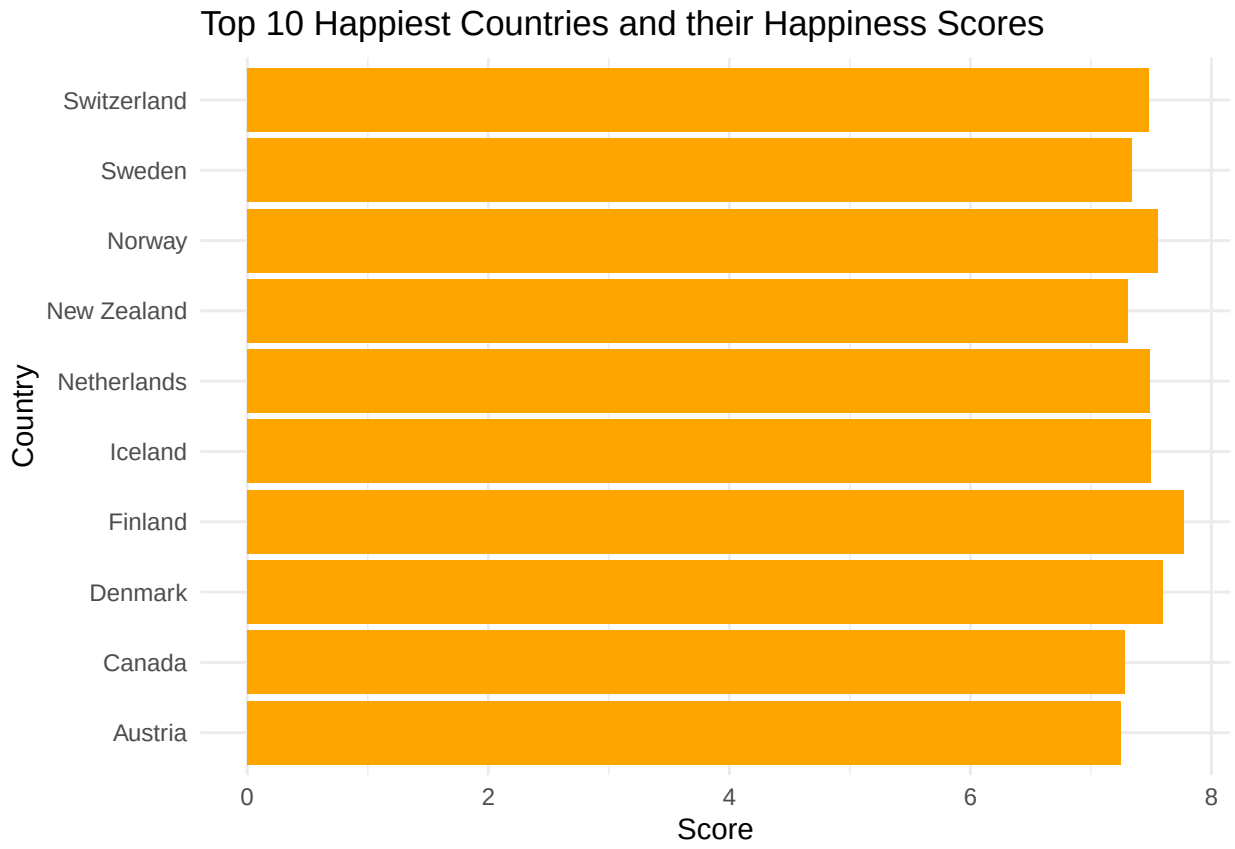
In this section, we will be deeply analyzing the revised dataset to gain a better insight about the problem that we are trying to solve. As mentioned earlier, the bigger problem is subdivided into smaller problems. This section will explore the smaller problems so that we can hopefully derive the answer to our primary objective.

What are the top 10 happiest countries and their scores? What is the average statistics of the top 10 happiest countries in the world?

This question can give us somewhat of a general overview of how the top 10 happiest countries fare on average in terms of the variables we wish to observe.

```
happy<-df%>%
arrange(desc(Score))%>%
head(10)

ggplot(data=happy,aes(x=Country.or.region,y=Score))+
  geom_bar(stat="identity",fill="orange")+
  labs(title="Top 10 Happiest Countries and their Happiness Scores",
       x="Country",y="Score")+coord_flip()+theme_minimal()
```




```
stats<-happy %>%
    summarize(
      happiness=mean(Score),gdp=mean(GDP.per.capita),
      support=mean(Social.support), life=mean(Healthy.life.expectancy),
      freedom=mean(Freedom.to.make.life.choices),generosity=mean(Generosity),
      corruption=mean(Perceptions.of.corruption))
  print(stats)
```

```
##   happiness   gdp support   life freedom generosity corruption
## 1    7.4559 1.387  1.5438 1.0177  0.5786     0.2741      0.319
```

We can see that in general, the top 10 happiest countries have above average levels of GDP Per Capita, Social Support, Life Expectancy and Freedom to make life choices.

How do economic factors such as GDP Per Capita relate to the happiness score of a country? Is it always the case that higher per capita is equal to higher happiness?

Lets first calculate the summary of the gdp.per.capita variable to understand it better.

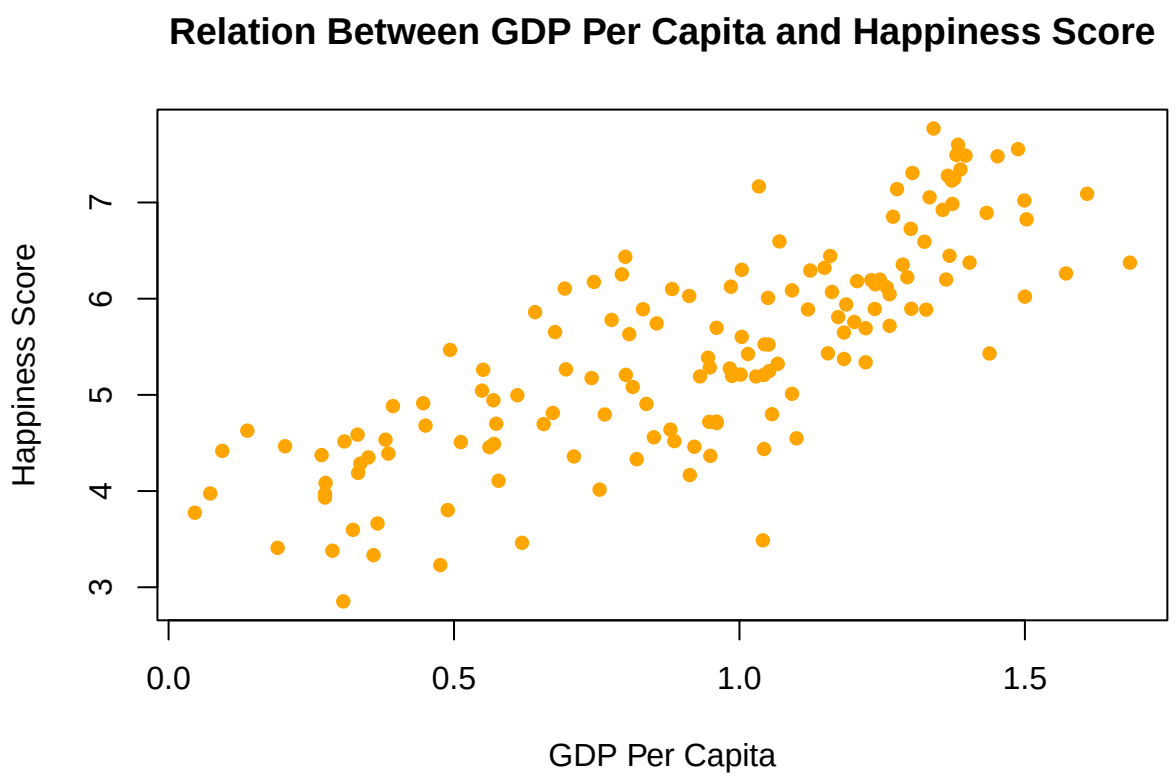
```
summary(df$GDP.per.capita)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0460  0.6248  0.9840  0.9210  1.2377  1.6840
```

The summary gives us some important insights about the GDP.per.capita variable. We can see that there is a large variation between the min and max which indicates a large range of values. The mean is 0.92 and median is 0.98 which indicates that most of the countries are likely to be middle-income. A simple scater plot

can visualize this:

```
plot(df$GDP.per.capita,df$Score,xlab="GDP Per Capita",ylab="Happiness Score",
     main="Relation Between GDP Per Capita and Happiness Score",
     col="orange",pch=16)
```



It can clearly be seen that there is a direct relation between the GDP per Capita and the Happiness Score. Generally speaking, the happiest countries tend to have wealthier individuals as opposed to less happier countries which have lesser wealth per capita. However, the plot also indicates that this is not always a one to one relation because there are a considerable amount of outliers. Upon closer inspection, it can be seen that there are several middle-income to higher-income countries that score lower on the happiness score than some lower-income ones.

Finally, lets make a simple linear regression model to further analyze the extent to which the per capita income impacts the happiness score.

```
gdp_reg<-lm(Score~GDP.per.capita,data=df)
summary(gdp_reg)
```

```
##
## Call:
## lm(formula = Score ~ GDP.per.capita, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23103 -0.46894  0.00699  0.46013  1.46365
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      3.3866     0.1410   24.02   <2e-16 ***
## GDP.per.capita   2.2405     0.1411   15.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6699 on 148 degrees of freedom
## Multiple R-squared:   0.63,  Adjusted R-squared:  0.6275
## F-statistic:   252 on 1 and 148 DF,  p-value: < 2.2e-16
```

This model reveals a multitude of important statistics which can help us in better understanding the correlation between the given factors. It can be observed that the p value is on the lower range which indicates a high level of significance between happiness and GDP Per Capita. We can also see that the R squared value is 0.63 which falls on the higher range. Another strong indicator is the slope which is positive and decently high in our case. To sum it all up, there is a signifcant level of relevance between the per capita income and the happiness score but this is not a one to one relation.

To what extent does the life expectancy impact a nation's overall happiness level? Is this a one to one relationship?

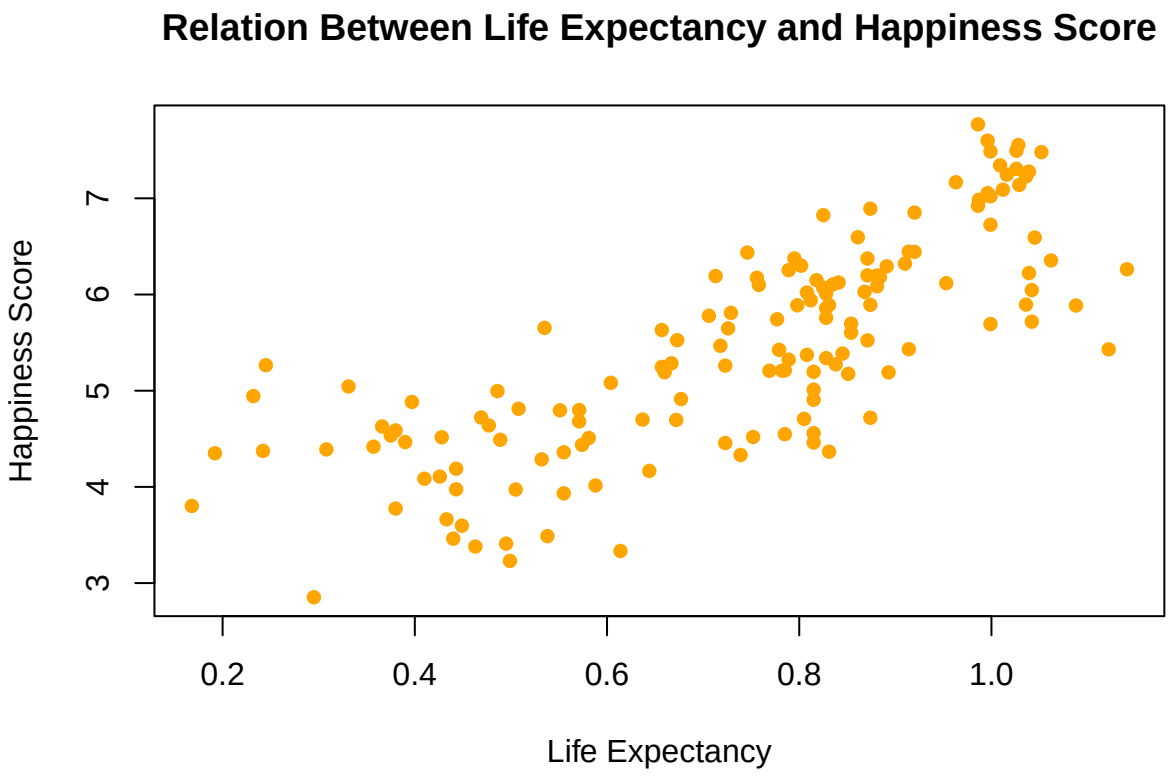
The next thing that we are going to be exploring is the relation between happiness and life expectancy.

```
summary(df$Healthy.life.expectancy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1680  0.5590  0.7965  0.7378  0.8832  1.1410
```

It can be seen that there is a large range of values since it starts from 0.168 and goes all the way up to 1.141. Since the mean and median are relatively close to each other, it can be concluded that the values are evenly distributed and without many extremes.

```
plot(df$Healthy.life.expectancy,df$Score,xlab="Life Expectancy",ylab="Happiness Score",
     main="Relation Between Life Expectancy and Happiness Score",
     col="orange",pch=16)
```



As expected, the graph shows an overall positive trend between life expectancy and the happiness index. Most of the points are clustred in the middle which indicates that countries that fall in the middle range of life expectancy usually also fall in the middle range of happiness. However, there are a decent amount of outliers which indicates that there is no outright way of predicting a country's happiness by its life expectancy alone.

```
life_reg<-lm(Score~Healthy.life.expectancy,data=df)
summary(life_reg)
```

```
##
## Call:
## lm(formula = Score ~ Healthy.life.expectancy, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65087 -0.46304  0.09428  0.52820  1.66731
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.6767     0.1911   14.00   <2e-16 ***
## Healthy.life.expectancy   3.7593     0.2476   15.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6886 on 148 degrees of freedom
## Multiple R-squared:  0.609,  Adjusted R-squared:  0.6064
## F-statistic: 230.5 on 1 and 148 DF,  p-value: < 2.2e-16
```

It can be observed that the p value is on the lower range which indicates a high level of significance between

happiness and Life Expectancy. We can also see that the R squared value falls on the higher range. The slope is positive and high which indicates a positive trend. To sum it all up, there is a signifcant level of relevance between the life expectancy and the happiness score.

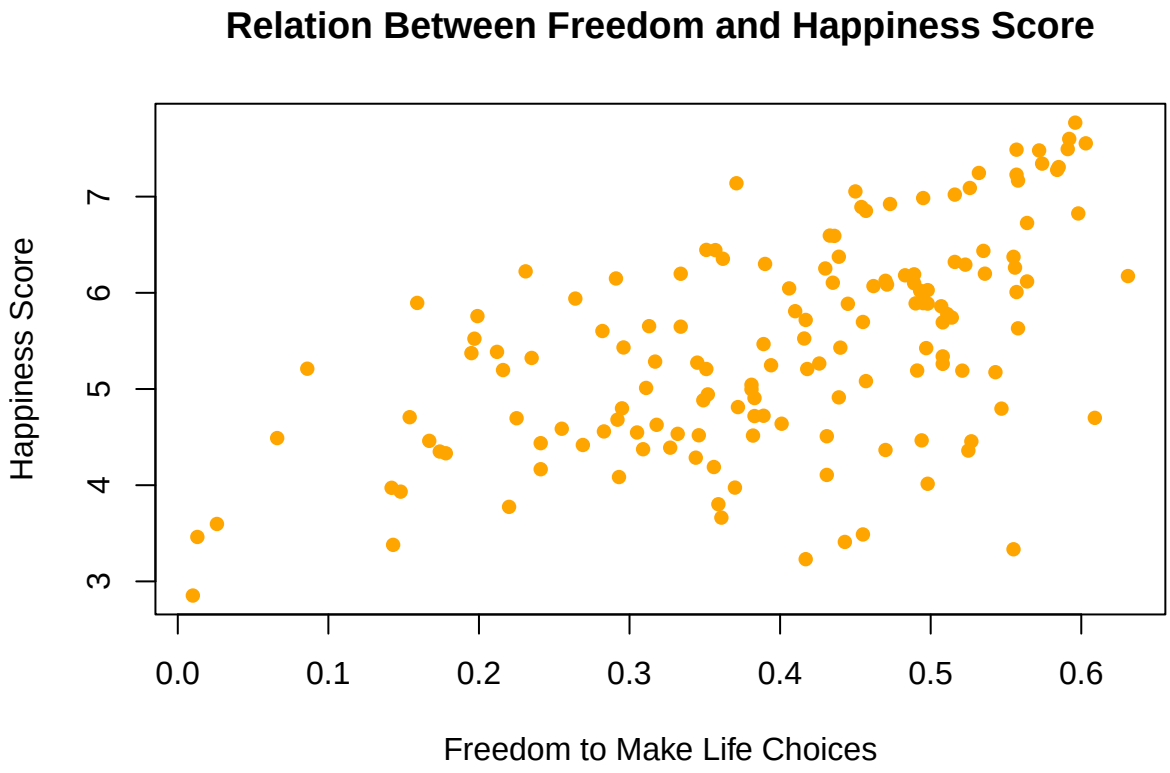
To what extent does the freedom to make life choices impact the happiness score of a country?

```
summary(df$Freedom.to.make.life.choices)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.3140  0.4220  0.3989  0.5078  0.6310
```

The summary indicates a decent level of range and variability. Since the mean and median are not too far apart, the overall distribution is roughly even.

```
plot(df$Freedom.to.make.life.choices,df$Score,xlab="Freedom to Make Life Choices",
     ylab="Happiness Score",main="Relation Between Freedom and Happiness Score",
     col="orange",pch=16)
```



It can be observed that there is positive growth between freedom and happiness. That being said, the degree to which freedom impacts the happiness level isn't as strong as the other factors that we have observed because the points are more spaced out as opposed to clustered. This indicates that the relationship between these two factors isn't very strong but it is rather on the middle range.

Once again, a simple linear regression model will be made and then its summary will be inspected to derive insights

```
freedom_reg<-lm(Score~Freedom.to.make.life.choices,data=df)
summary(freedom_reg)
```

```
##
## Call:
## lm(formula = Score ~ Freedom.to.make.life.choices, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81740 -0.56474 -0.00365  0.67413  1.81406
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.6585     0.2279  16.051  < 2e-16 ***
## Freedom.to.make.life.choices   4.4916     0.5403   8.314 5.53e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9093 on 148 degrees of freedom
## Multiple R-squared:  0.3183, Adjusted R-squared:  0.3137
## F-statistic: 69.12 on 1 and 148 DF,  p-value: 5.53e-14
```

It can be concluded that there is a moderate level of significance between the happiness score and the freedom to make life choices. To begin with, the R squared value is around 0.32 which is on the lower to middle range. Moreover, the p value is fairly low which indicates that there is a decent level of significance between these two variables. Finally, the slope is positive and decently high which indicates a moderate level of significance.

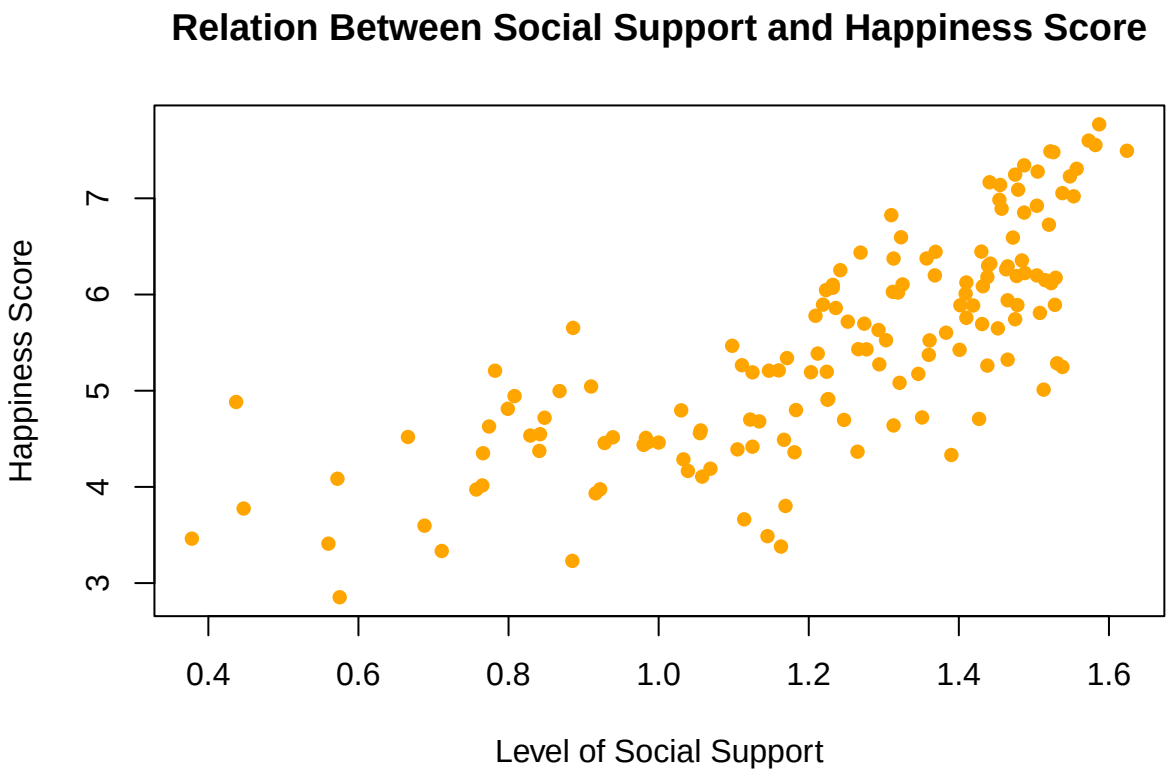
To what extent does social support impact the happiness level of a country? Is this a one to one relationship?

```
summary(df$Social.support)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.378   1.061   1.285   1.225   1.455   1.624
```

We can see that there is a high level of variability and range. The mean and median values are very close to each other which once again indicates an even distribution

```
plot(df$Social.support,df$Score,xlab="Level of Social Support",
     ylab="Happiness Score",main="Relation Between Social Support and Happiness Score",
     col="orange",pch=16)
```



The plot indicates a significantly high level of relation between the happiness socre and social support level. The graph is clearly indicating a positive trend and there is a fair level of variability. Once again, we can see clear outliers which means that the relationship between social support and happiness is not at all a one to one relation.

```
support_reg<-lm(Score~Social.support,data=df)
summary(support_reg)
```

```
##
## Call:
## lm(formula = Score ~ Social.support, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88271 -0.48411 -0.02491  0.57139  1.82075
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.7377     0.2585   6.722 3.62e-10 ***
## Social.support   3.0309     0.2058  14.728  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7014 on 148 degrees of freedom
## Multiple R-squared:  0.5944, Adjusted R-squared:  0.5917
## F-statistic: 216.9 on 1 and 148 DF,  p-value: < 2.2e-16
```

Like the plot, the model also indicates a fairly high level of relation between social support and happiness.

The p values is low which highlights the high level of significance. Moreover, the R squared value is moderate which indicates a decent level of variablity. Finally, the slope is positive and decently high which further proves the point.

Conclusion and Final Thoughts

After deeply analyzing and visualizing the given dataset, it can be concluded that there is a strong link between the happiness score and economic, social and governmental factors. However, we can also conclude that there is no certainty that if a country fares well in these variables, it will be undoubtedly happier than a country that performs poorly in the same factors. In other words, the relationship between the explored factors and happiness level is not one to one but it is more so of a general trend. This information is crucial to know because it is often a common misconception that a wealthier country with better social and governmental factors will always be happier regardless of other factors. Once again, in general terms, such a country would indeed be happier but this is not necessary in all cases as the high number of outliers in our analysis pointed out. The in depth analysis also provided the opportunity to better understand the extent to which each of these factors impact the overall happiness level of a country. The order of importance for each of these variables from most to least is as follows: GDP Per Capita, Healthy Life Expectancy, Social Support and Freedom to make life choices. This relationship was derived by carefully observing the summary, graph and simple regression model for each of these variables. Overall, the GDP per capita outshines all the other variables in terms of overall impact which means that the economic factor is by far the most important factor that we have observed. To sum it all up, the project served as a great learning experience as it provided me with the platform to showcase my data analysis, visualization and storytelling skills.

Dataset Citation

"World Happiness Report." Kaggle, 27 Nov. 2019, www.kaggle.com/datasets/unsdsn/world-happiness/code.