Introduction

The happiness rankings are a very useful countervailing force in a world that tends to take GDP as a proxy for a country's success, and in the past five years, some governments have launched initiatives aimed at improving the life satisfaction of their citizens.
As an example, the UAE has appointed a Minister of happiness, the United Kingdom has appointed a Minister for loneliness and New Zealand has revised its national budget based on how government spending will affect people's well-being.
The video above was very interesting for me to study more deeply the happiness rankings. The main purpose of choosing this kind of topic is to find out which factors are more important to live a happier life. As a consequence, countries and people can focus on the more important factors to achieve a higher happiness score. Also in this study we will consider different continents to find out if there are different trends for them regarding which factors play a meaningful role in obtaining a higher happiness score.

Characteristics of the Dataset

The presented dataset of numbers gives the happiness score the happiness rank according to 7 factors from 155 countries of the world (Freedom, Economy, Life.Expectancy, Trust, Family, Generosity and Dystopia.Residual).
The sum of the value on these 7 factors gives the happiness score. The higher the happiness score in this case, the lower the happiness rank. Factor "Dystopia" will be seen as a benchmark for other countries to show how far they are from being the poorest country in terms of happiness.

```
getwd()
#Reading the file
happy <- read.csv(file = 'C:/Users/emindryukova/Desktop/R/Data/2017.csv', header = TRUE)
```

Actually, our dataset has 155 observations and 12 variables. Some of the variable names are not so clear. We will change a little bit the name of several of them.

```
#Changing the name of columns
colnames (happy) <- c("Country", "Happiness.Rank", "Happiness.Score",
                      "Whisker.Low", "Whisker.High", "Life.Expectancy", "Economy", "Family",
                      "Freedom", "Trust", "Generosity", "Dystopia.Residual")
```

Also will remove the high- and low-whisker variables from the dataset because these variables only give the upper and lower confidence intervals of happiness scores, and there is no need to use them for visualization and prediction.

```
#Deleting useless columns (Whisker.low and Whisker.high)
```

Next, we will add another column to the dataset, which is conditional. We will study different continents to find out if there are different trends for them regarding which factors play a meaningful role in obtaining a higher happiness score.
After that we will move the position of the continent column to the 2nd column because with that position dataset looks much better.

```
happy <- happy[, -c(4,5)]
happy$Continent <- NA

happy$Continent[which(happy$Country %in% c("Israel", "United Arab Emirates", "Singapore", "Thailand",
"Vietnam", "Taiwan Province of China", "Saudi Arabia", "Kuwait", "Bahrain", "Iraq","Malaysia", "Uzbekistan", "Japan",
"South Korea", "Turkmenistan", "Kazakhstan", "Turkey", "Hong Kong S.A.R., China", "Philippines", "Jordan", "China",
"Yemen", "Pakistan", "Indonesia", "Qatar", "Azerbaijan", "Lebanon", "Tajikistan", "Bhutan", "Kyrgyzstan", "Mongolia",
"Palestinian Territories", "Armenia", "Iran", "Bangladesh", "Myanmar", "Sri Lanka", "India", "Georgia", "Cambodia",
"Afghanistan", "Syria", "Nepal"))] <- "Asia"

happy$Continent[which(happy$Country %in% c("Norway", "Denmark", "Iceland", "Switzerland",
"Finland","Netherlands", "Sweden", "Austria", "Ireland","Russia", "Bulgaria", "Romania", "Germany", "Belgium",
"Luxembourg", "United Kingdom", "Czech Republic", "Malta", "France", "Spain", "Slovakia", "Hungary", "Poland", "Italy",
"Lithuania", "Latvia", "Moldova", "Slovenia", "North Cyprus", "Belarus", "Cyprus", "Estonia", "Serbia", "Croatia", "Kosovo",
"Montenegro", "Greece", "Portugal", "Bosnia and Herzegovina", "Macedonia", "Albania", "Ukraine"))] <- "Europe"

happy$Continent[which(happy$Country %in% c("Canada", "Belize", "Costa Rica", "United States", "Mexico",
"Panama", "Trinidad and Tobago", "El Salvador", "Guatemala", "Jamaica", "Nicaragua", "Dominican Republic", "Honduras",
"Haiti"))] <- "North America"

happy$Continent[which(happy$Country %in% c("Chile", "Brazil", "Argentina", "Uruguay","Colombia",
"Ecuador", "Bolivia", "Peru", "Paraguay", "Venezuela"))] <- "South America"

happy$Continent[which(happy$Country %in% c("New Zealand", "Australia"))] <- "Australia"

happy$Continent[which(is.na(happy$Continent))] <- "Africa"

# Changing position of the continent column in the data to the 2 column
happy <- happy %>% select(Country,Continent, everything())

# Changing column's Continent to factor
happy$Continent <- as.factor(happy$Continent)
```

Now we have a ready-made ideal structure of 155 observations and 11 variables.
Country and continent – factor variables, Happiness rank – integer, other variables – numeric type.
Let's have a look at the data.

```
# Top 10 the happiest countries in 2017
head(happy)
```

Country	Continent	Happiness.Rank	Happiness.Score	Economy	Family	Life.Expectancy	Freedom	Generosity	Trust	Dystopia.Residual
Norway	Europe	1	7.5	1.616	1.53	0.7967	0.635	0.362	0.3160	2.28
Denmark	Europe	2	7.5	1.482	1.55	0.7926	0.626	0.355	0.4008	2.31
Iceland	Europe	3	7.5	1.481	1.61	0.8336	0.627	0.476	0.1535	2.32
Switzerland	Europe	4	7.5	1.565	1.52	0.8581	0.620	0.291	0.3670	2.28
Finland	Europe	5	7.5	1.444	1.54	0.8092	0.618	0.245	0.3826	2.43
Netherlands	Europe	6	7.4	1.504	1.43	0.8107	0.585	0.470	0.2827	2.29
Canada	North America	7	7.3	1.479	1.48	0.8346	0.611	0.436	0.2874	2.19
New Zealand	Australia	8	7.3	1.406	1.55	0.8168	0.614	0.500	0.3828	2.05
Sweden	Europe	9	7.3	1.494	1.48	0.8309	0.613	0.365	0.3644	2.10
Australia	Australia	10	7.3	1.484	1.51	0.8439	0.602	0.478	0.3012	2.07

#verifying the summary statistics to get an idea of the dataset
summary(happy)

```
   Country                   Continent  Happiness.Rank Happiness.Score    Economy          Family
 Length:155          Africa       :44   Min.    :  1    Min.   :2.7     Min.   :0.00   Min.   :0.00
 Class :character    Asia         :43   1st Qu.: 40    1st Qu.:4.5     1st Qu.:0.66   1st Qu.:1.04
 Mode  :character    Australia    : 2   Median : 78    Median :5.3     Median :1.06   Median :1.25
                     Europe       :42   Mean   : 78    Mean   :5.4     Mean   :0.98   Mean   :1.19
                     North America:14   3rd Qu.:116    3rd Qu.:6.1     3rd Qu.:1.32   3rd Qu.:1.41
                     South America:10   Max.   :155    Max.   :7.5     Max.   :1.87   Max.   :1.61
 Life.Expectancy     Freedom       Generosity       Trust        Dystopia.Residual
 Min.   :0.00    Min.   :0.00   Min.   :0.00   Min.   :0.00    Min.   :0.38
 1st Qu.:0.37    1st Qu.:0.30   1st Qu.:0.15   1st Qu.:0.06    1st Qu.:1.59
 Median :0.61    Median :0.44   Median :0.23   Median :0.09    Median :1.83
 Mean   :0.55    Mean   :0.41   Mean   :0.25   Mean   :0.12    Mean   :1.85
 3rd Qu.:0.72    3rd Qu.:0.52   3rd Qu.:0.32   3rd Qu.:0.15    3rd Qu.:2.14
 Max.   :0.95    Max.   :0.66   Max.   :0.84   Max.   :0.46    Max.   :3.12
> |
```

The dataset is looking great now. We can note that the average Happiness.score around the world is **5.3.**

# Considering the data types
str(happy)

```
'data.frame':   155 obs. of  11 variables:
 $ Country           : chr  "Norway" "Denmark" "Iceland" "Switzerland" ...
 $ Continent         : Factor w/ 6 levels "Africa","Asia",..: 4 4 4 4 4 4 5 3 4 3 ...
 $ Happiness.Rank    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Happiness.Score   : num  7.54 7.52 7.5 7.49 7.47 ...
 $ Economy           : num  1.62 1.48 1.48 1.56 1.44 ...
 $ Family            : num  1.53 1.55 1.61 1.52 1.54 ...
 $ Life.Expectancy   : num  0.797 0.793 0.834 0.858 0.809 ...
 $ Freedom           : num  0.635 0.626 0.627 0.62 0.618 ...
 $ Generosity        : num  0.362 0.355 0.476 0.291 0.245 ...
 $ Trust             : num  0.316 0.401 0.154 0.367 0.383 ...
 $ Dystopia.Residual: num  2.28 2.31 2.32 2.28 2.43 ...
> |
```

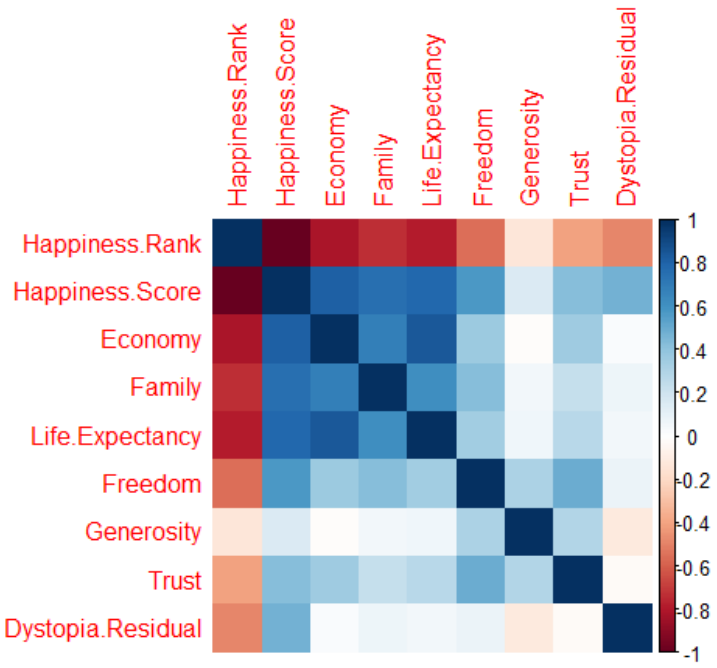
#checking if there are any missing values we should be worried about
sum(is.na(happy))

```
> sum(is.na(happy))
[1] 0
```

The data seems pretty much clean. Additionally, there are no missing values, which makes analyzing our dataset much easier. We don't have to put any time to handle the missing values.

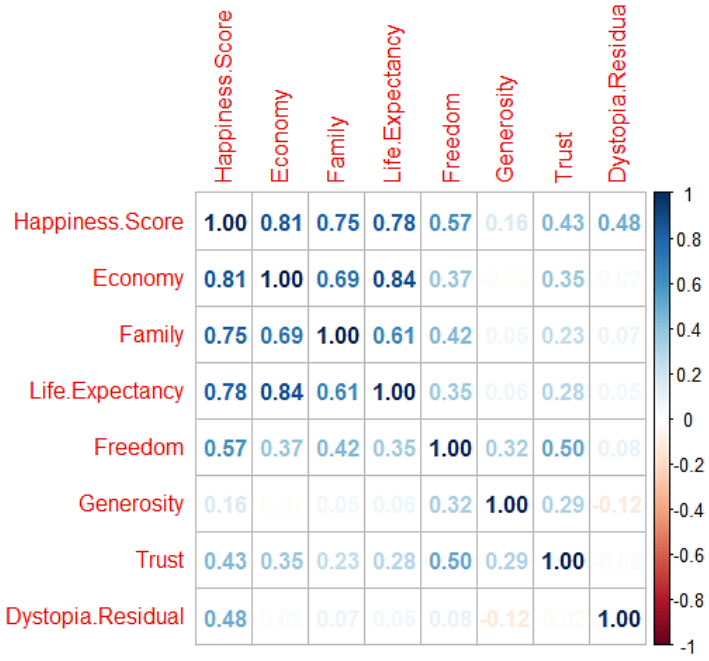
Correlation between Each Variable

```
# determination of correlation between numerical columns
Num.cols <- sapply(happy, is.numeric)
Cor.data <- cor(happy[, Num.cols])
corrplot(Cor.data, method = 'color')
```



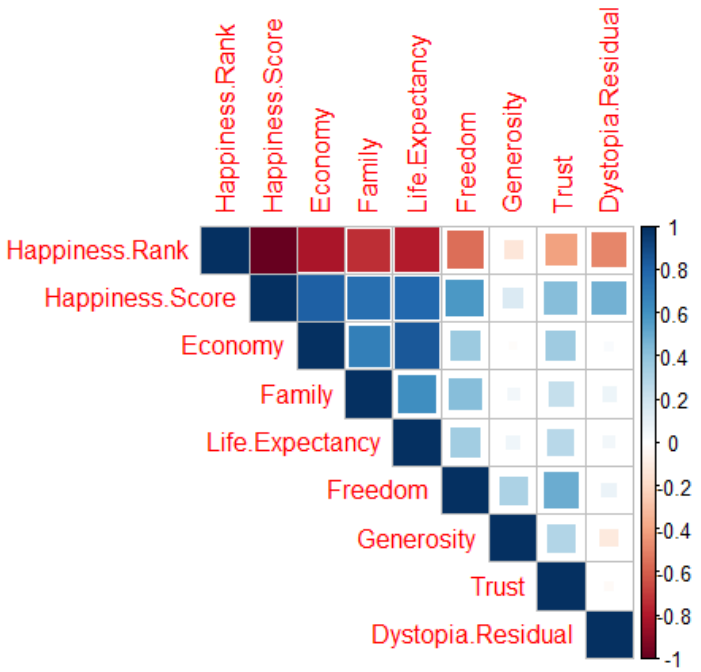
Most likely, there is an inverse correlation between the "Happiness Rank" and all other numerical v ariables. In this way the lower the happiness rank, the higher the happiness score and the higher th e other 7 factors contributing to happiness. Therefore, we will remove the rank of happiness and lo ok at the correlation again.

```
# Create a correlation plot
newdatacor = cor(happy[c(4:11)])
corrplot(newdatacor, method = "number")
```



As we can see from the above correlation plot **Economy, Family, Life expectancy** play the most important role in contributing to happiness. **Generosity and Trust** have the lowest impact on the happiness score.
Let's have a look at the correlation matrix

```
#Create a Correlogram
corrplot::corrplot(Cor.data,type = "upper",method = "square",mar = c(0,0,1,0))
```



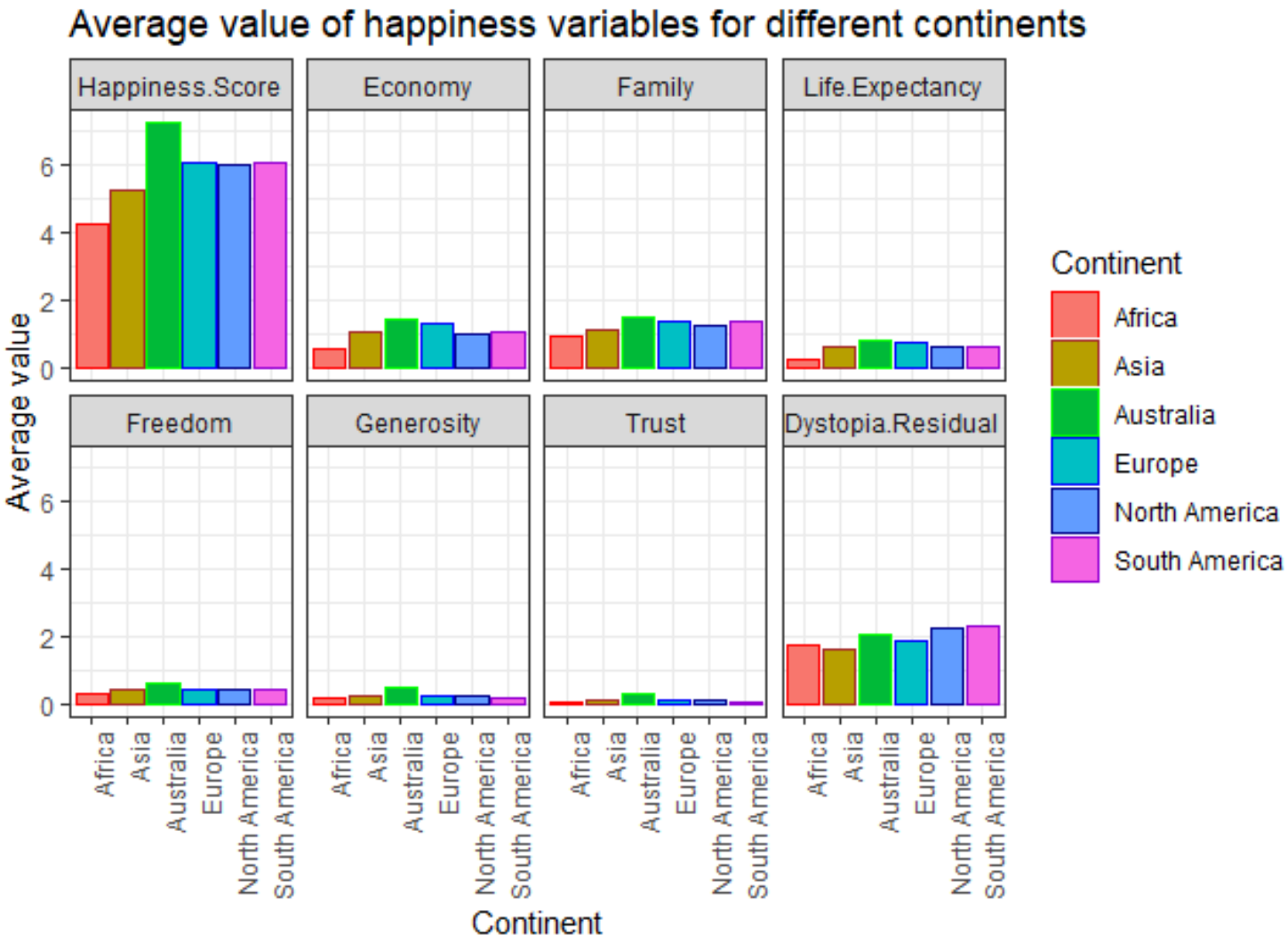
Happiness Score was highly correlated with Economy, Family, Life Expectancy and Freedom

Comparison of different continents regarding their happiness variables

To compare different continents relative to their happiness variables we will calculate the average happiness score and the average of the other 7 variables for each continent. Then we will melt it to have variables and values in separate columns. Eventually, we will use ggplot to show the difference between continents.

```
#calculating the average happiness score and for each continent the average of the other 7 variables
happy.Continent <- happy %>%
  select(-3) %>%
  group_by(Continent) %>%
  summarise_at(vars(-Country), funs(mean(., na.rm=TRUE)))
happy.Continent.melt <- reshape2::melt(happy.Continent)

ggplot(happy.Continent.melt, aes(y=value, x=Continent, color=Continent,  fill=Continent)) +
  geom_bar( stat="identity") +
   scale_color_manual(values = c('red','brown','green','blue','dark blue', 'dark violet'))+
  facet_wrap(~variable, nrow = 2) + theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Average value of happiness variables for different continents",
       y = "Average value")
```



The diagram shows us that Australia has the highest average in all fields (except dystopia residual). North America, South America and Europe, are almost the same regarding happiness score and the other 7 factors. As for Africa and Asia, it is the lowest scores in all fields.

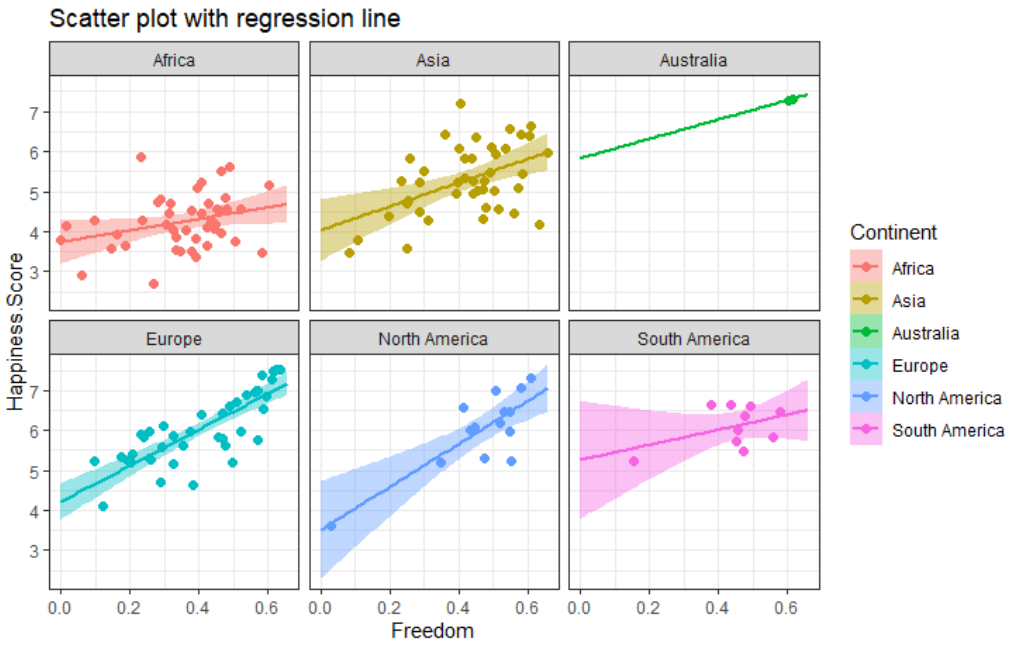
Scatter plot with regression line

Let's have a look at the correlation between happiness score and the other 7 factors in the happy dataset for different continents by a scatter plot.

It is worth noting that it was possible to ignore Australia, since there are only 2 countries there (Australia, New Zealand), and creating a scatter plot with a regression line for this continent does not give us any comprehension.

1. #Scatter plot:Factor – Freedom

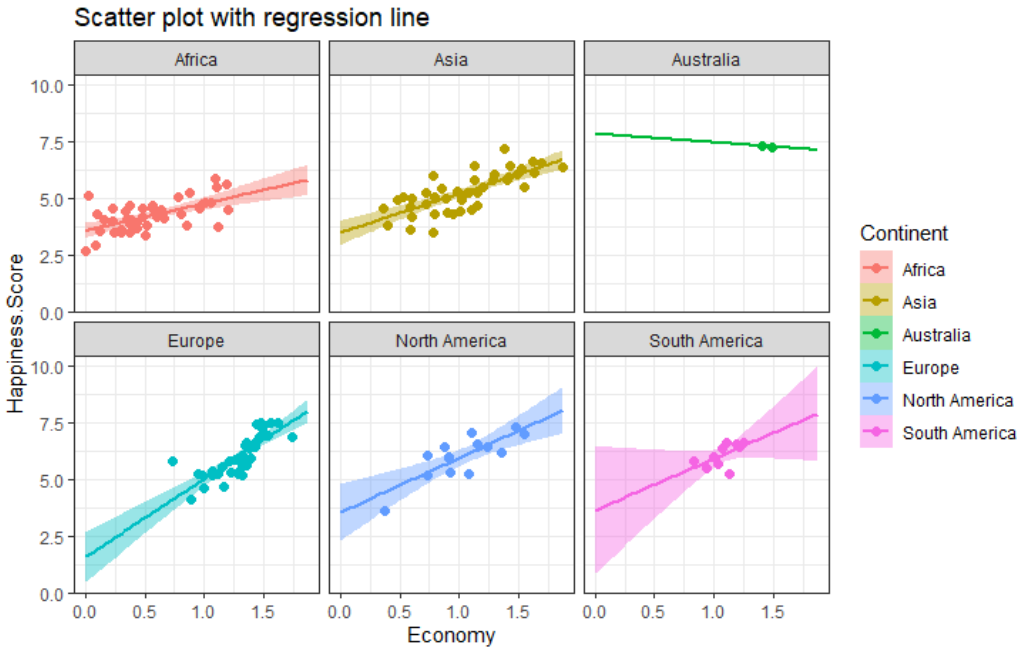
```
ggplot(subset(happy), aes(x = Freedom, y = Happiness.Score)) +
  geom_point(aes(color=Continent), size = 2, alpha = 1) +
  geom_smooth(aes(color = Continent, fill = Continent),
              method = "lm", fullrange = TRUE) +
  facet_wrap(~Continent) +
  theme_bw() + labs(title = "Scatter plot with regression line")
```



Freedom in Europe and North America strongly correlates with happiness, unlike other continents.

2. #Scatter plot:Factor – Economy

```
ggplot(subset(happy), aes(x = Economy, y = Happiness.Score)) +
  geom_point(aes(color=Continent), size = 2, alpha = 1) +
  geom_smooth(aes(color = Continent, fill = Continent),
              method = "lm", fullrange = TRUE) +
  facet_wrap(~Continent) +
  theme_bw() + labs(title = "Scatter plot with regression line")
```

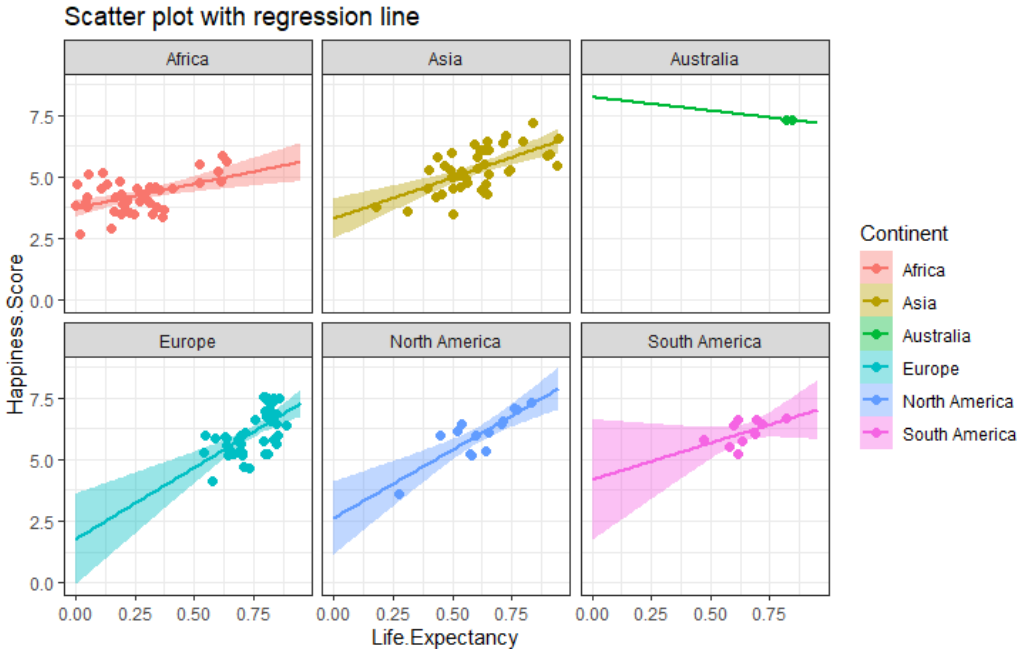


This graph shows us almost the same result in the correlation between the Happiness.Score and the economy. We see that Africa has the lowest rates.

3. #Scatter plot:Factor - Life.Expectancy

```
ggplot(subset(happy), aes(x = Life.Expectancy, y = Happiness.Score)) +
  geom_point(aes(color=Continent), size = 2, alpha = 1) +
  geom_smooth(aes(color = Continent, fill = Continent),
```

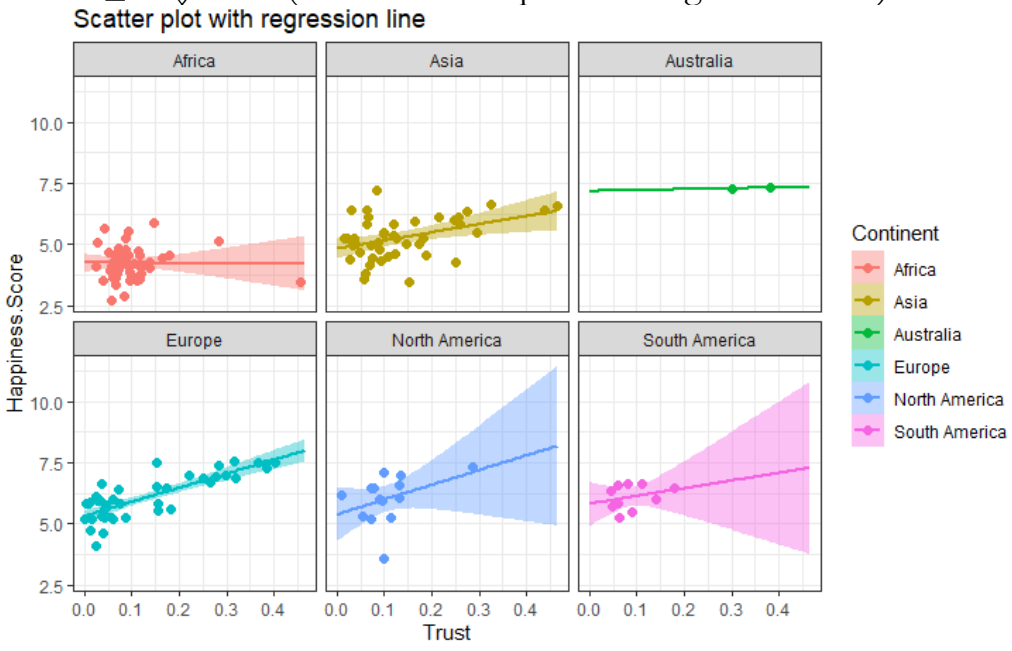
```
            method = "lm", fullrange = TRUE) +
  facet_wrap(~Continent) +
  theme_bw() + labs(title = "Scatter plot with regression line")
```



The graph shows us that he correlation is more significant between Life.expectancy and Happiness.Score in North America, Europe and Asia than the other continents.

#Scatter plot:Factor – Trust

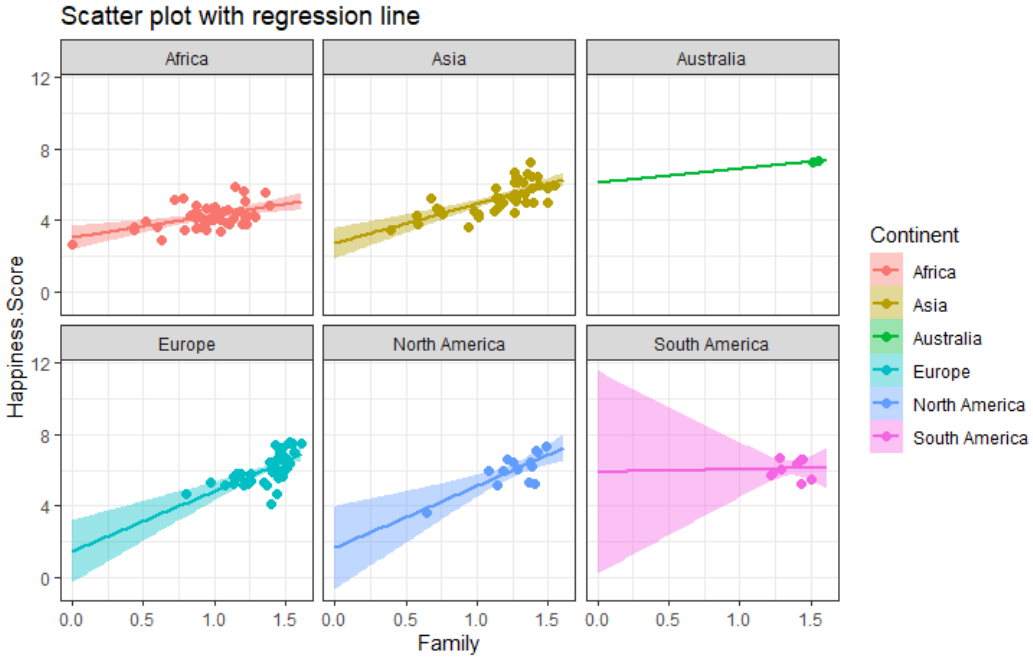
```
ggplot(subset(happy), aes(x = Trust, y = Happiness.Score)) +
  geom_point(aes(color=Continent), size = 2, alpha = 1) +
  geom_smooth(aes(color = Continent, fill = Continent),
              method = "lm", fullrange = TRUE) +
  facet_wrap(~Continent) +
  theme_bw() + labs(title = "Scatter plot with regression line")
```



In Africa there is almost no correlation between trust and Happiness.Score, because the line is horizontal.

4. #Scatter plot:Factor – Family

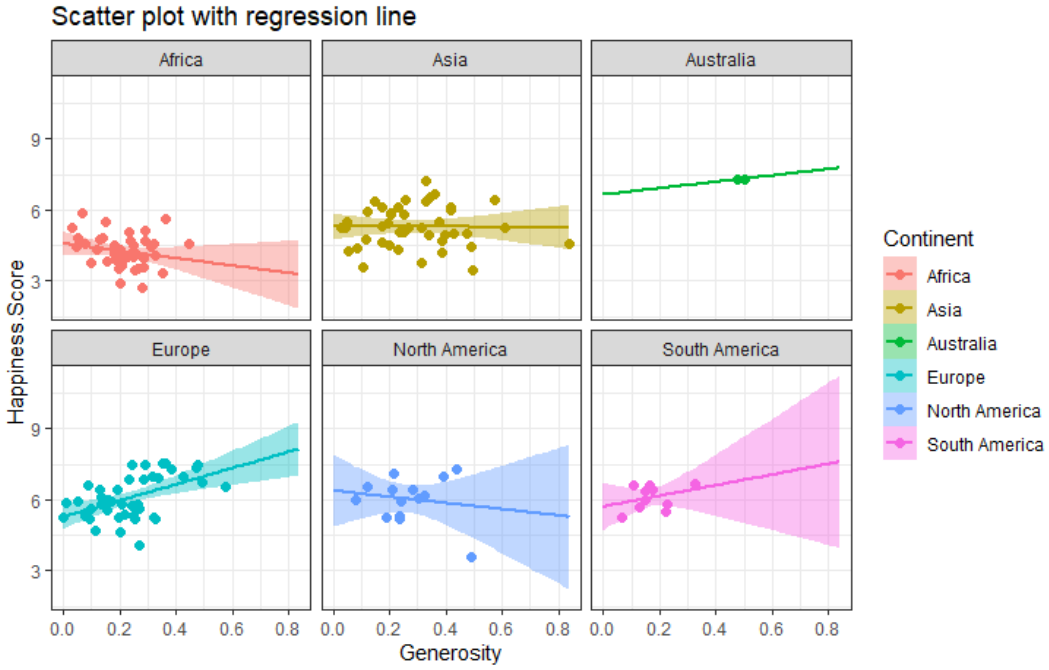
```
ggplot(subset(happy), aes(x = Family, y = Happiness.Score)) +
  geom_point(aes(color=Continent), size = 2, alpha = 1) +
  geom_smooth(aes(color = Continent, fill = Continent),
              method = "lm", fullrange = TRUE) +
  facet_wrap(~Continent) +
  theme_bw() + labs(title = "Scatter plot with regression line")
```



According to the graph in South America with increase the Family score, the happiness score remains.

5. #Scatter plot: Factor – Generosity

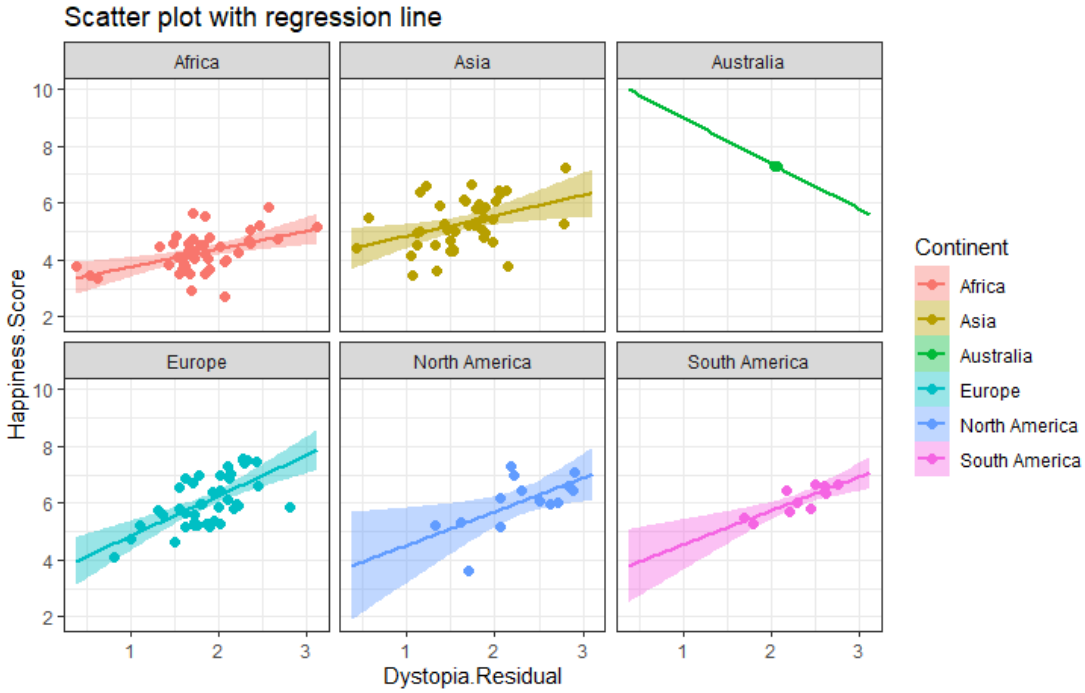
```
ggplot(subset(happy), aes(x = Generosity, y = Happiness.Score)) +
  geom_point(aes(color=Continent), size = 2, alpha = 1) +
  geom_smooth(aes(color = Continent, fill = Continent),
              method = "lm", fullrange = TRUE) +
  facet_wrap(~Continent) +
  theme_bw() + labs(title = "Scatter plot with regression line")
```



In this case, the graph shows us the regression line has a positive slope only for South America and Europe. As for Asia, the line is horizontal, and for Africa and North America, the slope is negative.

6. #Scatter plot: Factor – Dystopia.Residual

```
ggplot(subset(happy), aes(x = Dystopia.Residual, y = Happiness.Score)) +
  geom_point(aes(color=Continent), size = 2, alpha = 1) +
  geom_smooth(aes(color = Continent, fill = Continent),
              method = "lm", fullrange = TRUE) +
   facet_wrap(~Continent) +
   theme_bw() + labs(title = "Scatter plot with regression line")
```



According to the graph all continents act the same regarding Dystopia.residual, the regression line has a positive slope

Collinearity. Prediction

Let's to predict happiness score. We will split our dataset into Training.set and Test.set.
Dependent variable - happiness score
Independent variables: freedom, trust, family, economy, generosity, life expectancy and dystopia residual.

```
# Splitting the dataset into the Training_set and Test_aset
# install.packages('caTools')
library(caTools)
set.seed(123)
happyset <- happy[4:11]
split = sample.split(happyset$Happiness.Score, SplitRatio = 0.8)
training.set = subset(happyset, split == TRUE)
test.set = subset(happyset, split == FALSE)

#Multiple Linear Regression
#Fitting Multiple Linear Regression to the Training set
model = lm(formula = Happiness.Score ~ .,
           data = training.set)
```

```
> summary(model)

Call:
lm(formula = Happiness.Score ~ ., data = training.set)

Residuals:
      Min        1Q    Median        3Q       Max
-5.91e-04 -2.01e-04 -2.00e-07  2.51e-04  4.85e-04

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)       1.70e-04   1.51e-04     1.13     0.26
Economy           1.00e+00   1.30e-04  7690.84   <2e-16 ***
Family            1.00e+00   1.25e-04  7981.80   <2e-16 ***
Life.Expectancy   1.00e+00   2.12e-04  4711.65   <2e-16 ***
Freedom           1.00e+00   2.25e-04  4453.25   <2e-16 ***
Generosity        1.00e+00   2.31e-04  4330.04   <2e-16 ***
Trust             1.00e+00   3.33e-04  2997.19   <2e-16 ***
Dystopia.Residual 1.00e+00   5.45e-05 18343.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00028 on 116 degrees of freedom
Multiple R-squared:     1,      Adjusted R-squared:     1
F-statistic: 2.69e+08 on 7 and 116 DF,  p-value: <2e-16
```

According to the above summary all independent variables have a meaningful impact. Also we can note adjusted R2 is very high and equal to 1. Obviously there is a linear correlation between independent and dependent variables.
And a very important point the sum of the independent variables is equal to the dependent variable which is the Happiness.score. This explains that an adjusted R2 equals to 1. Therefore, here we can assume that Multiple Linear Regression will predict Happiness.scores with 100 % precision!

Conclusions

After analyzing data of Global Happiness Levels in the world we were able to discover the impact of each different factor in determining "happiness."
It seems safe to say that Economy, Family and Life expectancy play the most important role in contributing to happiness. And here Generosity and Trust have the lowest impact on the happiness score. And this is a big problem, especially for countries that are located in Africa. Trust has the lowest scores of all conditions looked at. Countries that have little to no trust and confidence in the governments, make it so that the citizens feel disenfranchised and are not able to take the life choices they wish, which is illustrated in the correlation between low trust and low Freedom scores.

By looking at and analyzing these data, we could understand what makes countries and their citizens happier, thus allowing us to focus on prioritizing and improving these aspects of each nation each continents. So that many countries, for example, can review its national budget based on how government spending would affect people's well-being.