DATA ANALYSIS & STATISTICS

ASSIGNMENT II: WORLD HAPPINESS REPORT

U2058657
PRABANCHAN CHANTHERAN VETHATHIRI

Prabanchan Chantheran Vethathiri
u2058657@unimail.hud.ac.uk

INTRODUCTION

The World Happiness Report is an annual publication of the Sustainable Development Solutions Network for the United Nations. It includes papers and national satisfaction rankings based on respondent ratings of their own lives, which are also associated with different life factors in the study. In July 2011, resolution 65/309 Happiness: Towards a Comprehensive Concept of Growth was adopted by the UN General Assembly, inviting Member States to measure the happiness of their citizens and to use the data to direct public policy. This was followed by the first UN High Level Meeting on April 2, 2012, called Health and Happiness: Establishing a New Economic Paradigm, chaired by UN Secretary General Ban Ki moon and Prime Minister Jigme Thinly of Bhutan, a nation that has adopted gross national happiness as its main measure of growth instead of gross domestic product. As a fundamental text for the UN High Level Meeting: Wellbeing and Happiness: Establishing a New Economic Model, attracting international attention, the first World Happiness Report was published on April 1, 2012. The study outlined the state of world happiness, causes of happiness and suffering, and case studies illustrated policy implications. The second World Happiness Study was released in 2013 and has been distributed on an annual basis since then, apart from 2014. In this report, we will discuss the Problem Statement, Dataset, Results and Conclusion.

PROBLEM STATEMENT

World Happiness is the sum of happiness of citizens of all the countries. Happiness of the citizens in a country is dependent on various factors like Freedom, Economy, Family, Government policies and so on.

The data set we chose is a World Happiness Report which ranks 156 countries with respect to their Happiness Score which is dependent on 6 other factors. In this report we are predicting the factors that impacts the Happiness Score and Rank the most. And categorizing the countries to their respective continents for exploratory data analysis. We have conducted the following tests and analysis to predict the highest impacting factors and predict a model.

1) t Test.
2) ANOVA.
3) Correlation and linear regression.

By running t Test, we can understand whether there is a significant difference or relation between the means of two groups with the help of the p value obtained by running the test. We assume a null hypothesis and if the p value is less than 0.05, it is rejected, and we accept the alternative hypothesis with a 95% confidence interval, we only have a 5% chance of being wrong.

ANOVA test can be helpful in determining the inferential statistical difference between the means of two or more groups. We assume there is no difference in the means in the null hypothesis of ANOVA as well and reject or accept the hypothesis based on the value obtained by running the test.

Correlation is a statistical measure that expresses the extent to which two or more variables are linearly related. This method will help us in determining the factors most impacting the Happiness Score and Rank.

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine which variables are significant predictors of the outcome variable.

DATA SET

A seminal survey of the state of global happiness is the World Happiness Study.As governments, organisations and civil society are continually using

happiness metrics to guide their policy
making decisions, the study continues to gain global attention. Leading expertsin thefields of economics, psychology, survey research, national statistics,

health, public policy and more explain how well
being metrics can be efficiently used to determine nations' development.

The studies review the state of happiness in today's world and illustrate how

personal and national differences in happiness are clarified by the modern science of happiness.

Information from the Gallup World Survey is included in the happiness scores

and rankings. The ratings are based on responses to the key question asked in the poll about life assessment. This question, referred to as the Cantril ladder, asks respondents to think of a ladder with a 10 being the best life possible for

them and a 0 being the worst life possible and to score their own current lives on that scale.

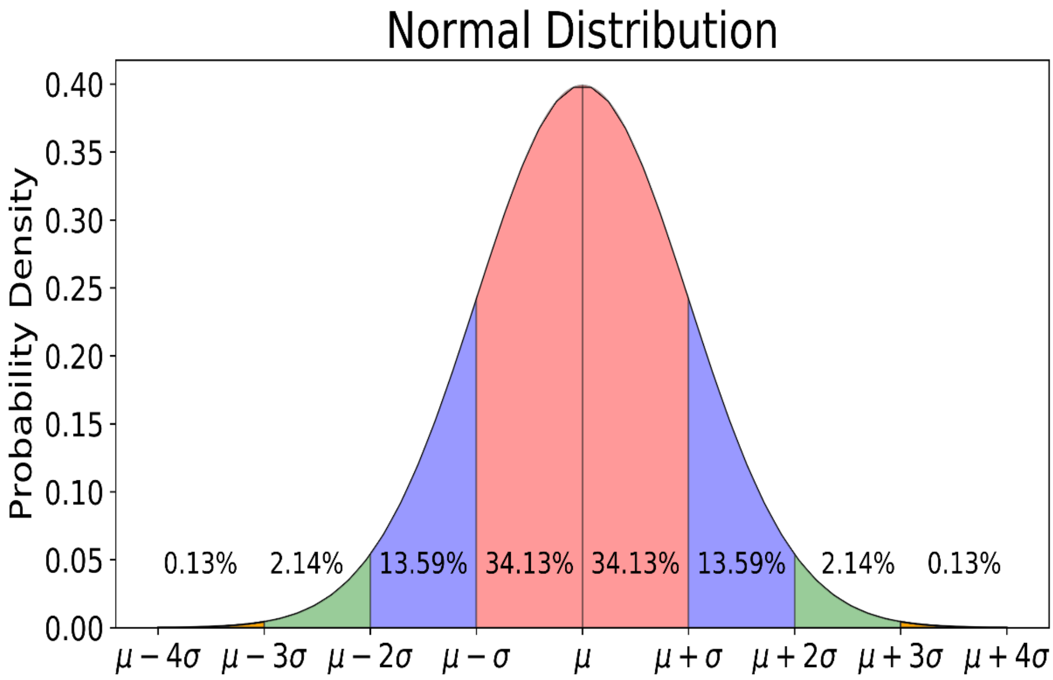
The World Happiness Report data set has 156 rows and 9 columns which provides rankings for 156 countries(rows) which are the ordinal transformation of Happiness Score which is dependent on the factors(columns):

1)GDP per capita 2) Social support 3) Healthy life expectancy 4) Freedom to make life choices 5) Generosity 6) Perceptions of corruption.

Reference: Kaggle. https://www.kaggle.com/unsdsn/world-happiness

Normal Distribution

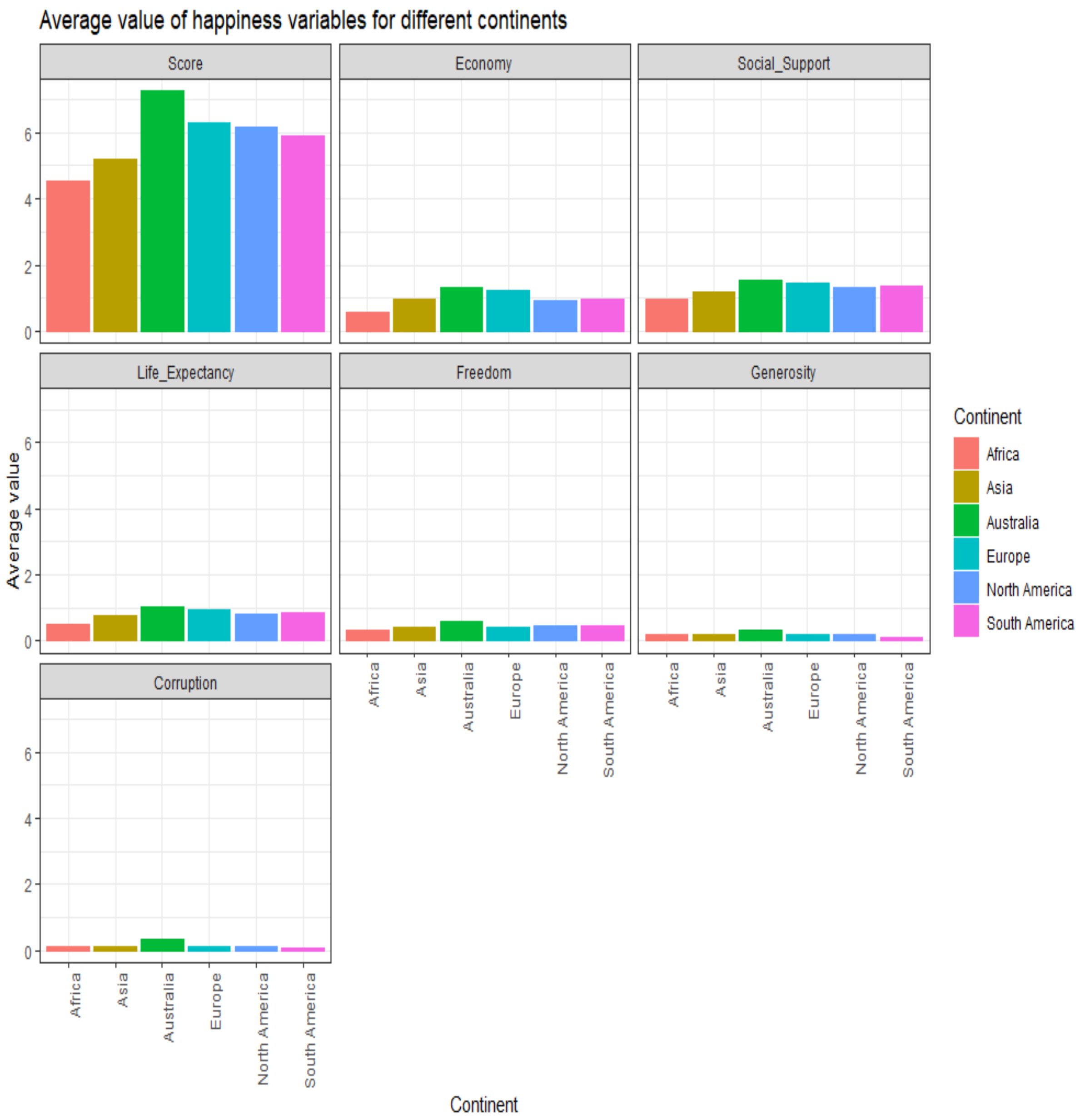
We performed normal distribution for all the variables and found it to be normally distributed.



Exploratory Data Analysis

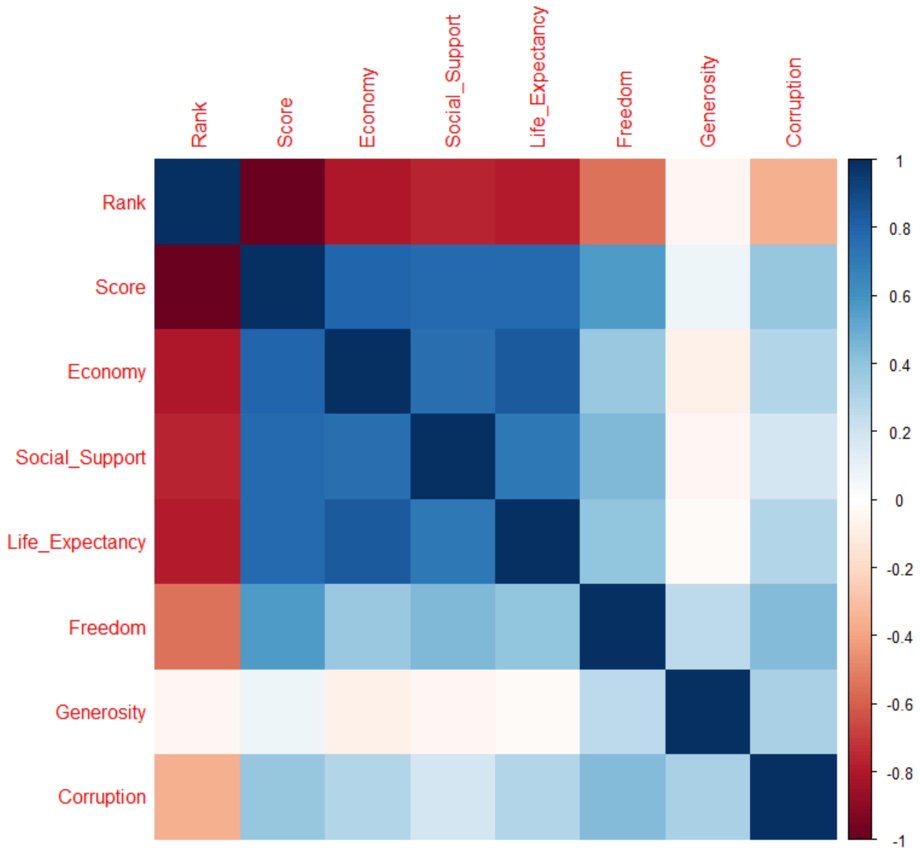
Average value of Happiness variables for different continents

We have categorized the countries to their respective continents and used bar plots for continent wise distribution with factors like Happiness Score, Economy, Social Support, Life expectancy, Freedom to make life choices, Generosity and Perceptions of corruption, respectively. From the below plot we can see that Australia has approximately the highest average in all fields.



Correlation between various factors:

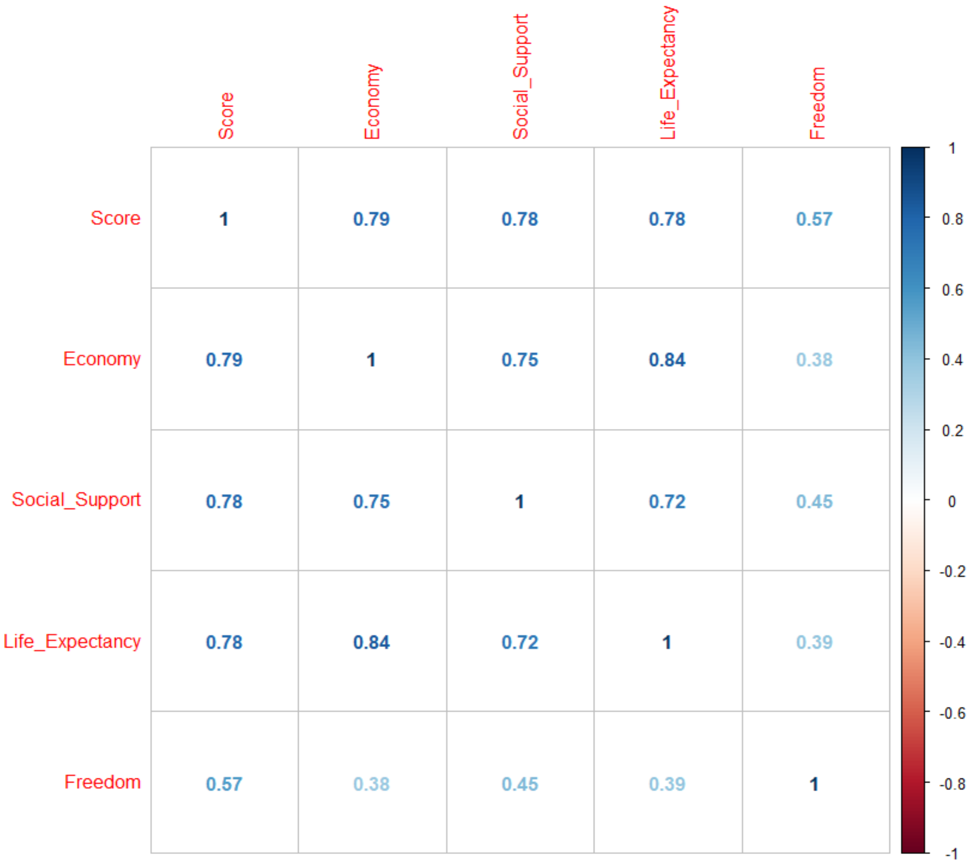
The below plot is a correlation matrix for the factors such as Rank, Score, Economy, Social Support, Life expectancy, Freedom, Generosity, Corruption.



The reverse association of "happiness rank" with all others is present.

In other words, the lower the rank of happiness, the higher the score of happiness, and the higher the other seven variables that add to happiness.

So, let us drop the rank of happiness, and again see the link.



The economy, life expectancy and family play the major role in contributing to happiness according to the above story. The lowest effect on the satisfaction level is on trust and kindness.

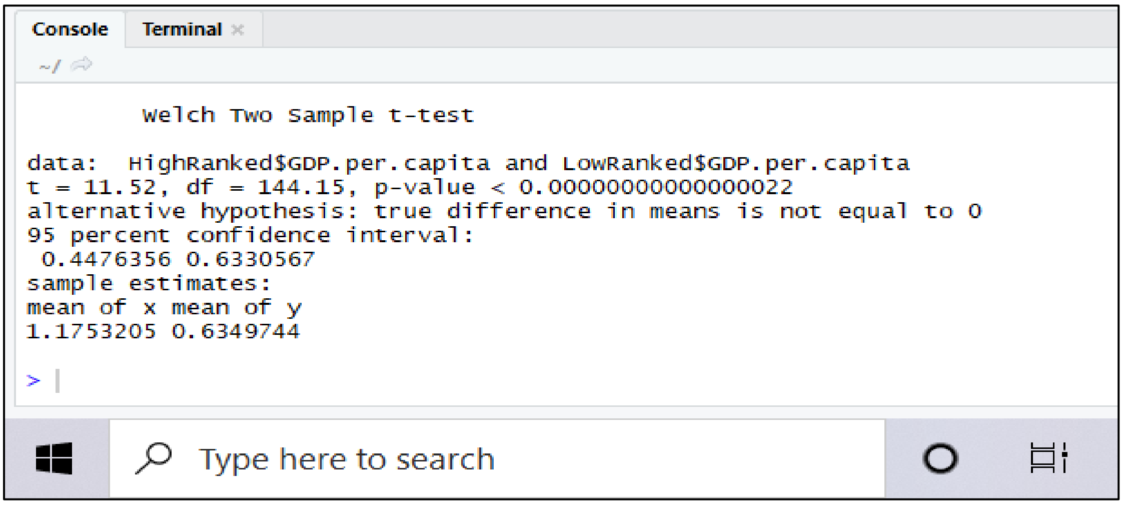
DATA Pre-processing

This report on World Happiness which we are working on does not have missing values. The pre-processing which we carried out are changing the column names, creating a new column named continents and categorized all the countries to their respective continents, rearranging the columns by moving the continent variable to the second column in the dataset, changing continent variable to factor, for the sake of conducting t Test we categorized the Happiness score(continuous variable) into High & Low (ordinal) and for ANOVA we categorized the Happiness Score into High, Mid-High, Mid-Low, Low.

RESULTS

We have performed t Test, ANOVA, Correlation and Linear Regression in the World Happiness Report for predicting the factors that impacts the Happiness score and rank the most and predicted a model.

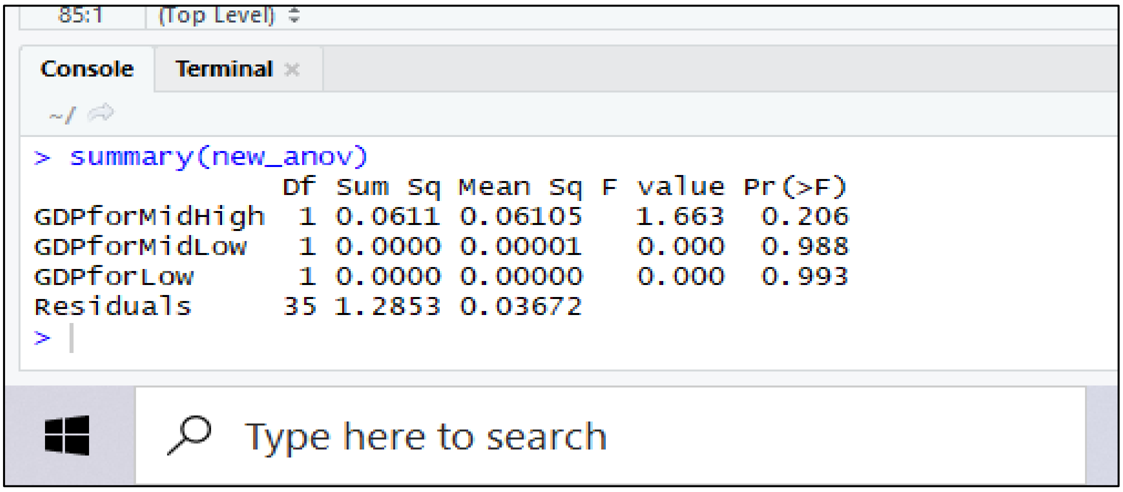
t Test



```
        Welch Two Sample t-test

data:  HighRanked$GDP.per.capita and LowRanked$GDP.per.capita
t = 11.52, df = 144.15, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4476356 0.6330567
sample estimates:
mean of x mean of y
1.1753205 0.6349744

>
```

In the above result, the value of t-statistics is 11.52 with 144.15 degrees of freedom. These are used to derive the p-value which in this case is even less than 0.00000000000000022. This implies that even if we repeat the experiment multiple times, there is very high chance that the mean will vary. The difference between the means is significant and is not because of chance. The 95% of the confidence interval tells that if we repeat over and over, 95% of the times, it will give the same result.

ANOVA



```
> summary(new_anov)
               Df Sum Sq Mean Sq F value Pr(>F)
GDPforMidHigh   1 0.0611 0.06105   1.663  0.206
GDPforMidLow    1 0.0000 0.00001   0.000  0.988
GDPforLow       1 0.0000 0.00000   0.000  0.993
Residuals      35 1.2853 0.03672
>
```

The ANOVA is run for the four groups:

1) countries having High Happiness scores
2) countries having Mid-High Happiness scores
3) countries having Mid-Low Happiness scores
4) countries having Low Happiness scores

ANOVA is run for the 3 groups against the High Happiness Score group.

The result tells the degrees of freedom, sum of squares, mean square values and F-statistics for each variable. The degrees of freedom are 1, because it is comparing each variable against the variable considered (i.e. countries having high Happiness scores). Using these, the p-value for each variable is computed.

P-value for GDP/capita for mean comparison between High and Mid-High is 0.206

P-value for GDP/capita for mean comparison between High and mid-Low is 0.988

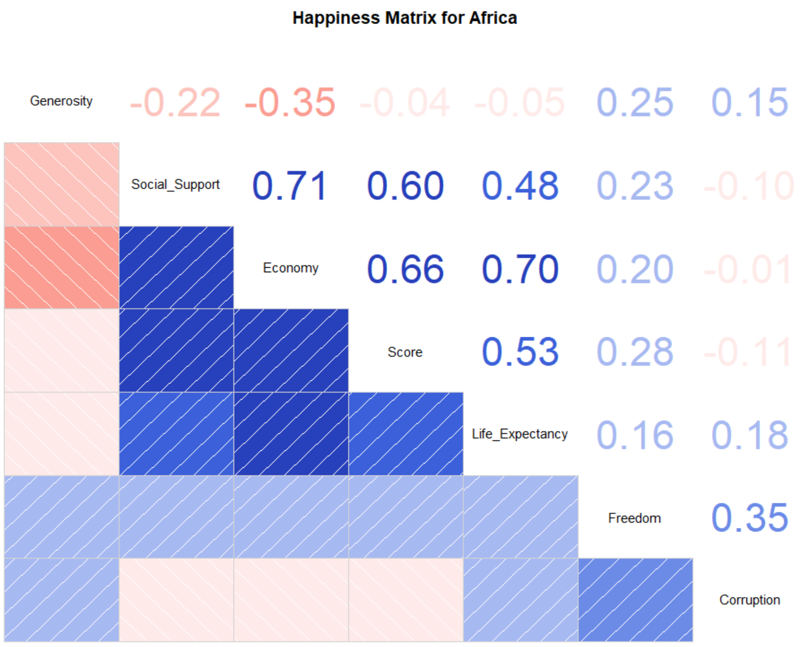
P-value for GDP/capita for mean comparison between High and Low is 0.993

Our assumed α value is 0.05. Since all the 3 computed p-values are > assumed α value, we can safely assume that if the experiment were to repeat over and over, 95% of the times the means of the various groups will not vary significantly with the high happiness scores' group. But a low value of F-statistics tells that p-value cannot be considered a strong indicator of the group mean difference.

Correlation plot for each continent

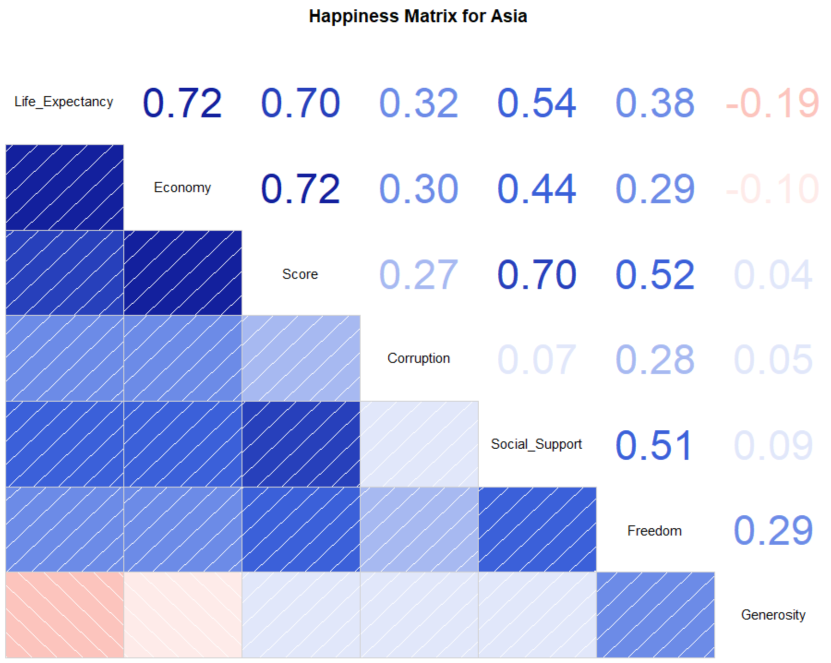
Correlation between "Happiness Score" and the other variables in Africa:

- Economy (0.66) > Social Support (0.60) > Life Expectancy (0.53) > Freedom (0.28)
- There is no correlation between happiness score and generosity.
- There is an inverse correlation between happiness score and corruption.



Correlation between "Happiness Score" and the other variables in Asia:

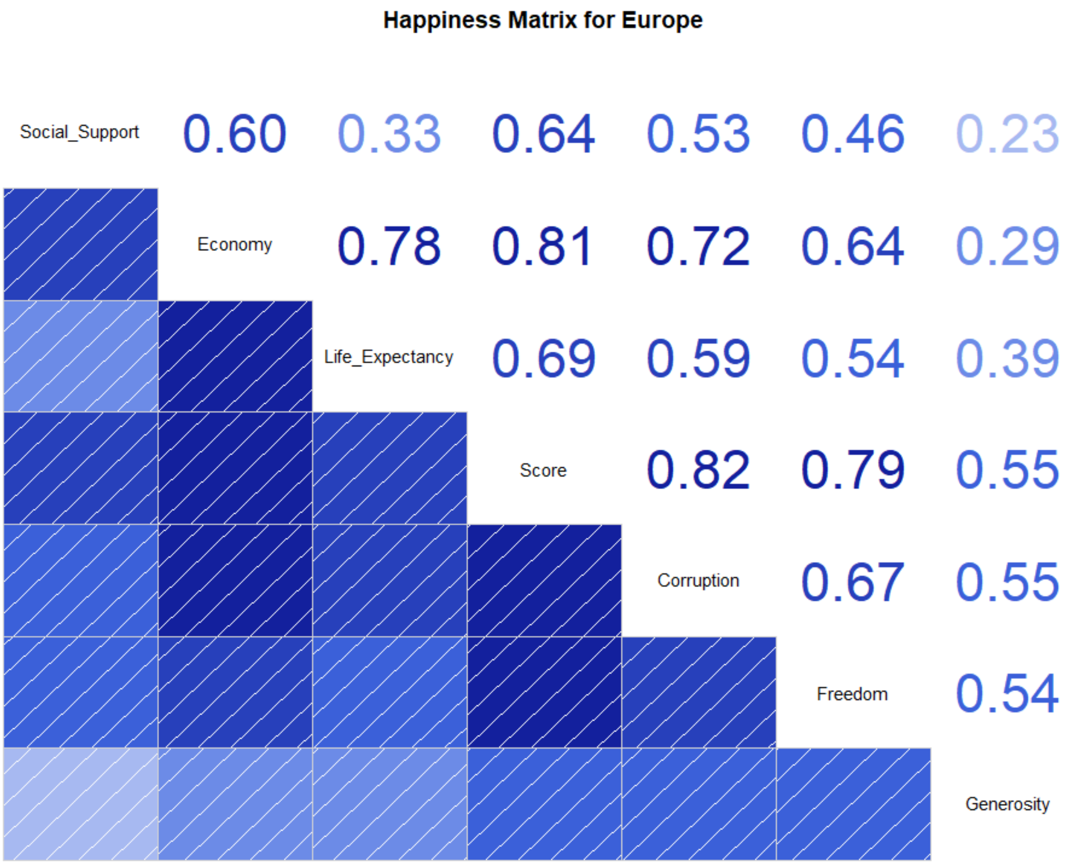
- Economy (0.72) > Social support = Life Expectancy (0.70) > Freedom (0.52) > Corruption (0.27)
- There is no correlation between happiness score and generosity (shown by the number0.04).



Correlation between "Happiness Score" and the other variables in Europe:

- Corruption (0.82) > Economy (0.81) > Freedom (0.79) > Life Expectancy (0.69) > Social Support (0.64) > Generosity (0.55)

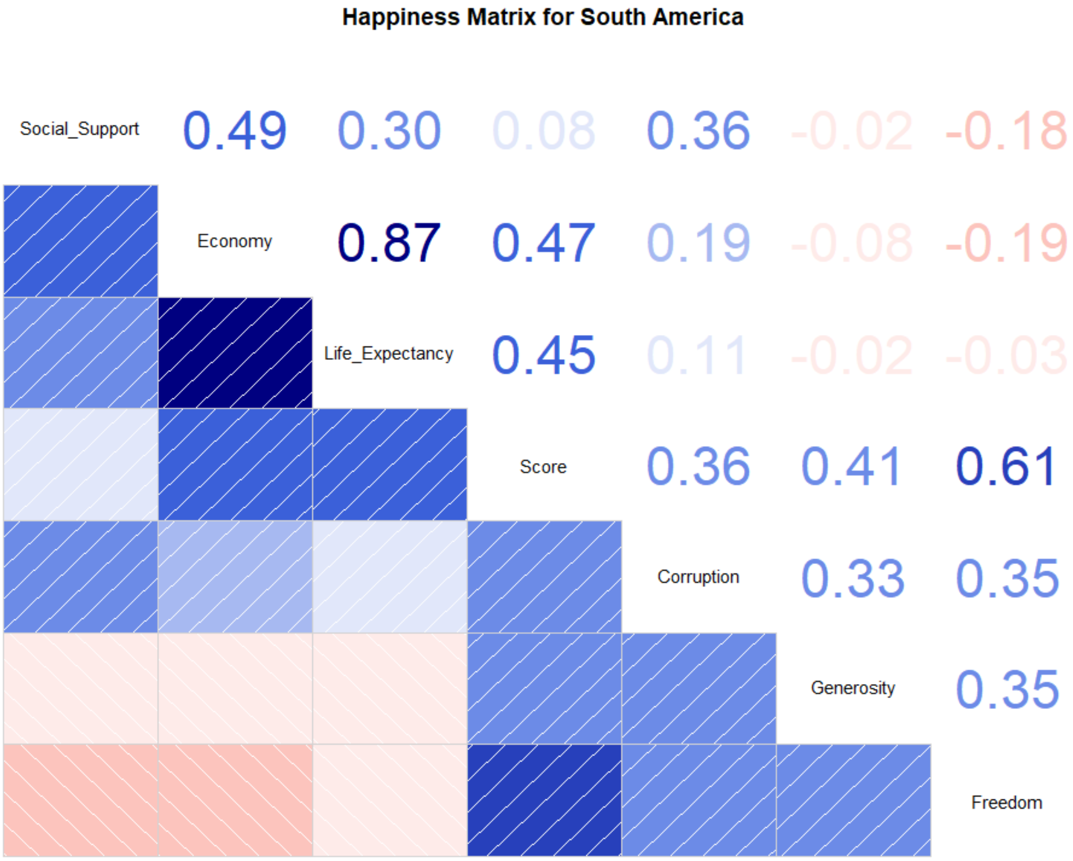
- The highest correlation between generosity and happiness score took place in Europe.



Correlation between "Happiness Score" and the other variables in South America:

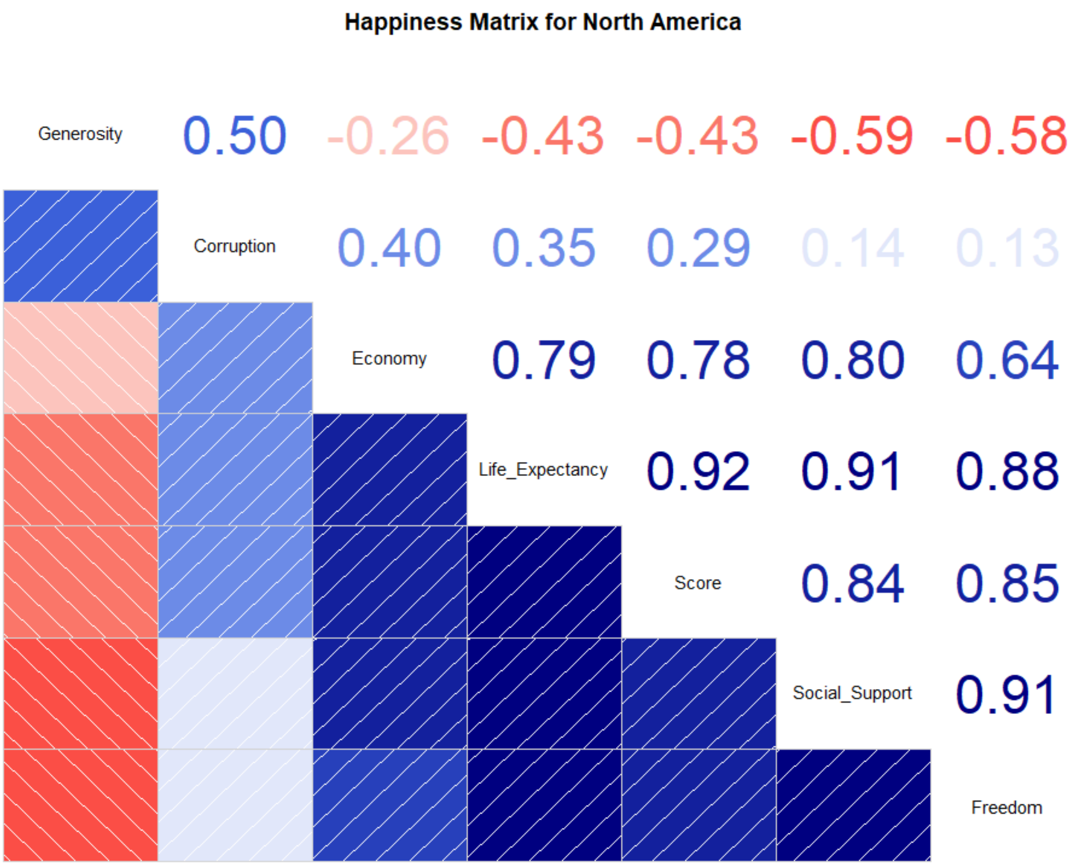
- Freedom (0.61) > Economy (0.47) > Life Expectancy (0.45) > Generosity (0.41) > Corruption (0.36)

- Social Support (0.08) seems to be the least significant factor in South America.



Correlation between "Happiness Score" and the other variables in North America:

- Life Expectancy (0.92) > Freedom (0.85) > Social Support (0.84) > Economy (0.78) > Corruption (0.29)
- There is an inverse correlation between happiness score and generosity (-0.43).



Scatter plot with regression line

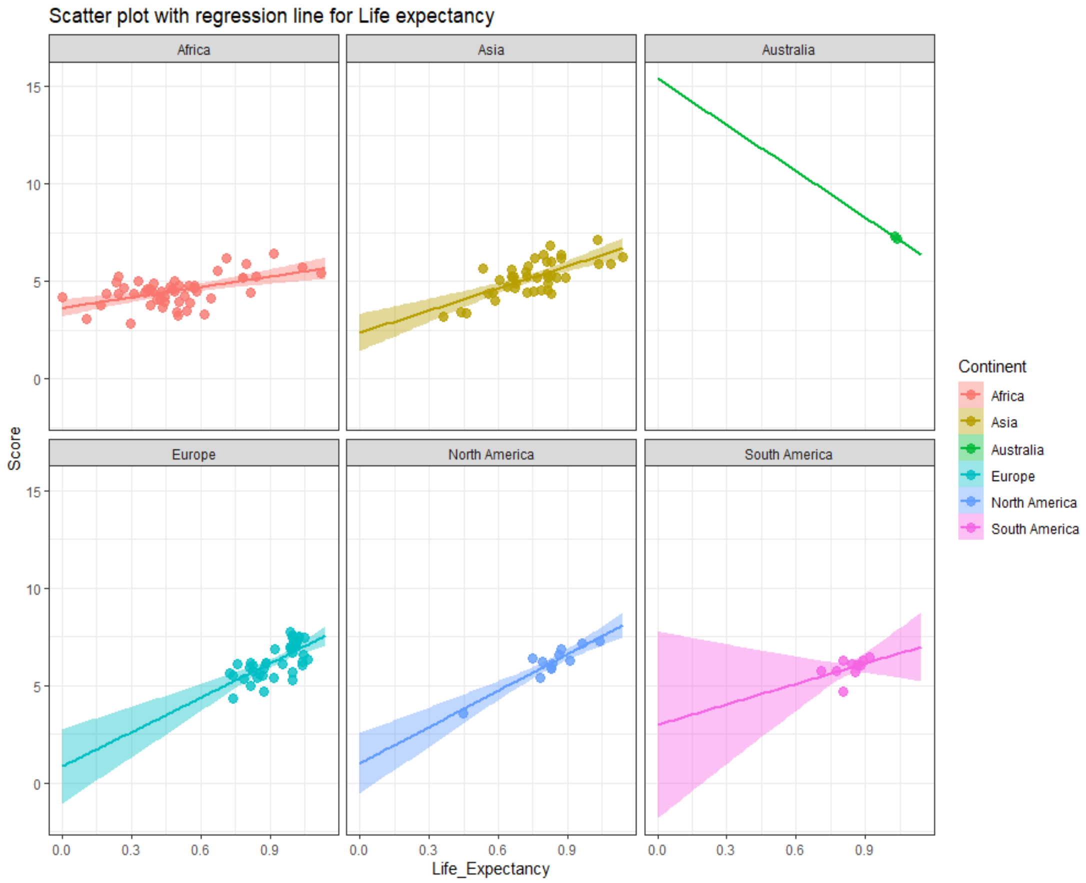
Let us see the correlation between happiness score and the other seven factors in the happiness dataset for different continents by creating a scatter plot.

For Life Expectancy

The correlation between life expectancy and happiness score North America, and Asia is more significant than the other continents. In Asia, North America and Africa, the life expectancy values are scattered between 0 and 1 (that may be because we have good amount of data points for these two continents) and we can conclude that as the life expectancy increases, the happiness scores also increase

With Europe the life expectancy values are more on the higher side (somewhere > 0.7, which can be seen in the graph).

Worth mentioning that we will not take Australia into account because there are just two countries in Australia (namely Australia and New Zealand) and creating scatter plot with the regression line for this continent will not give us any insight.

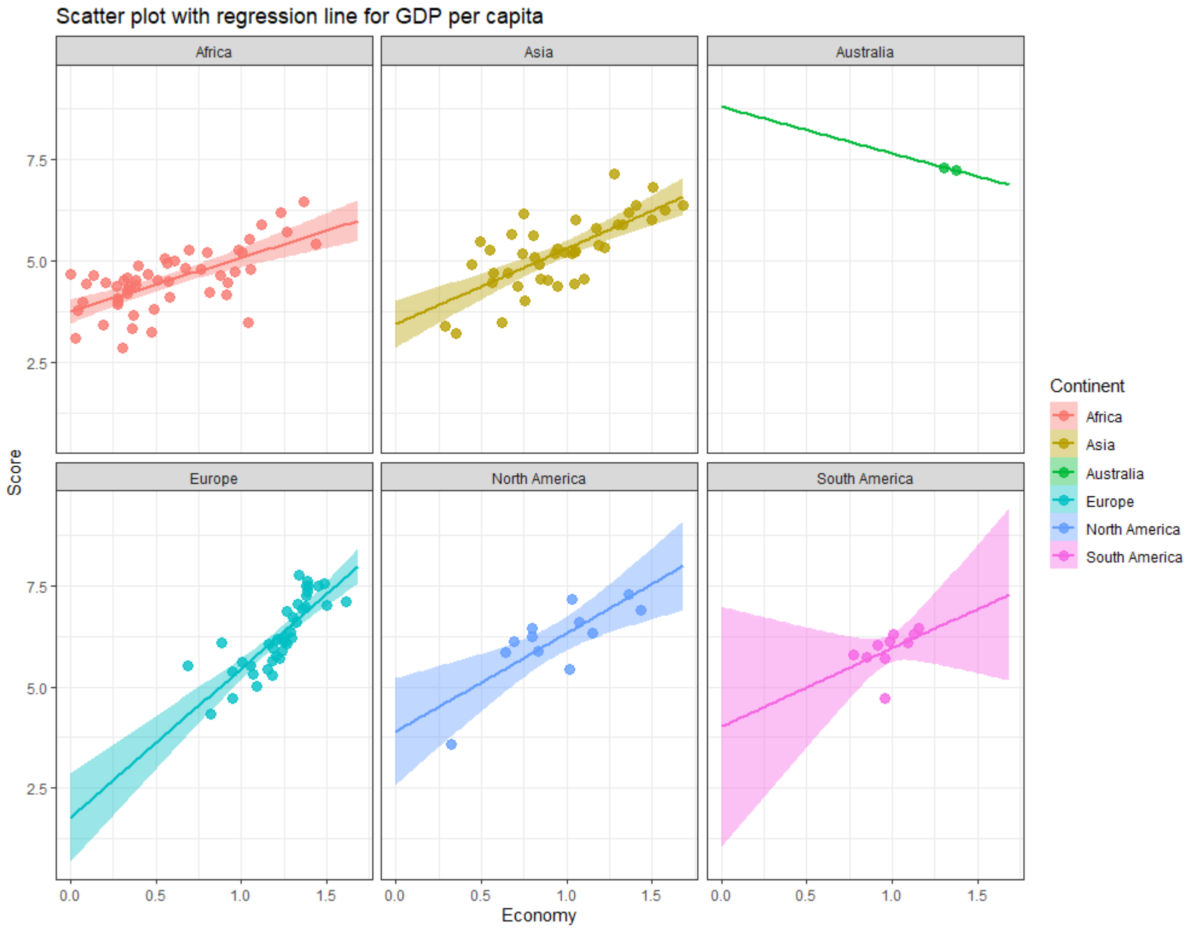


For

Economy (GDP per capita)

We can pretty much see the same result here for the correlation between happiness score and economy. Europe and North America have high correlation in this regard.

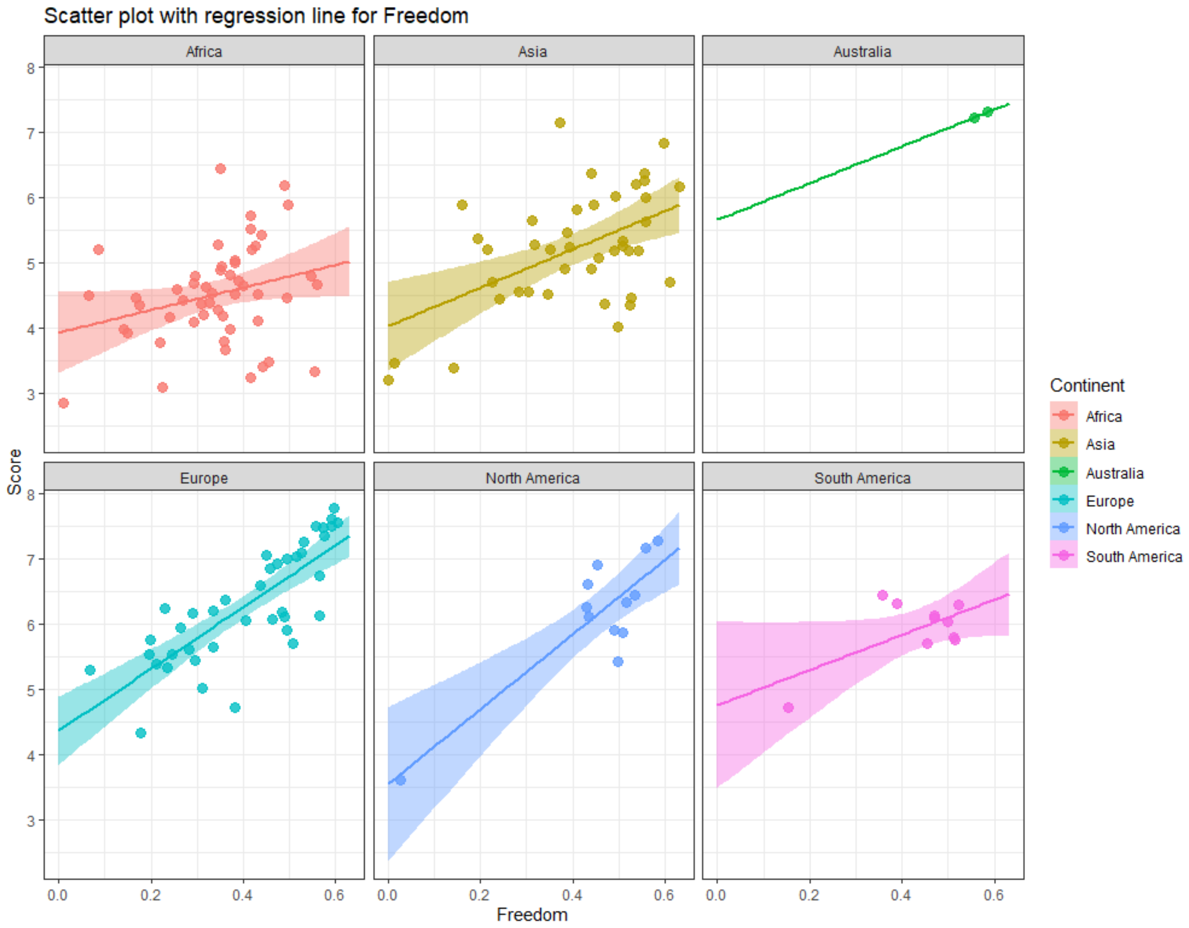
Asia, Africa and South America also have positive correlation.



For Freedom

Freedom in Europe and North America is more correlated to happiness score than any other continents.

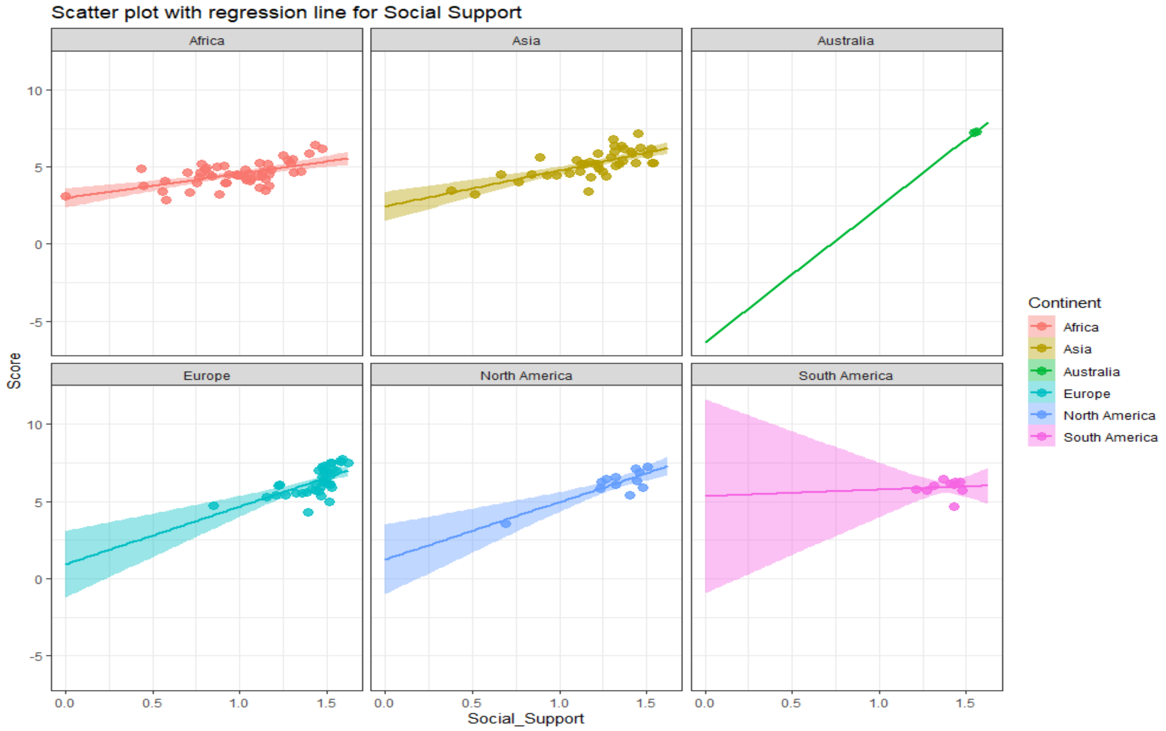
However, we can see how values are scattered around the line in the graph plotted above for Asia and Africa. This tells us that Freedom does not seem to be that significant for Africa and Asia and is rather random.



For Social Support

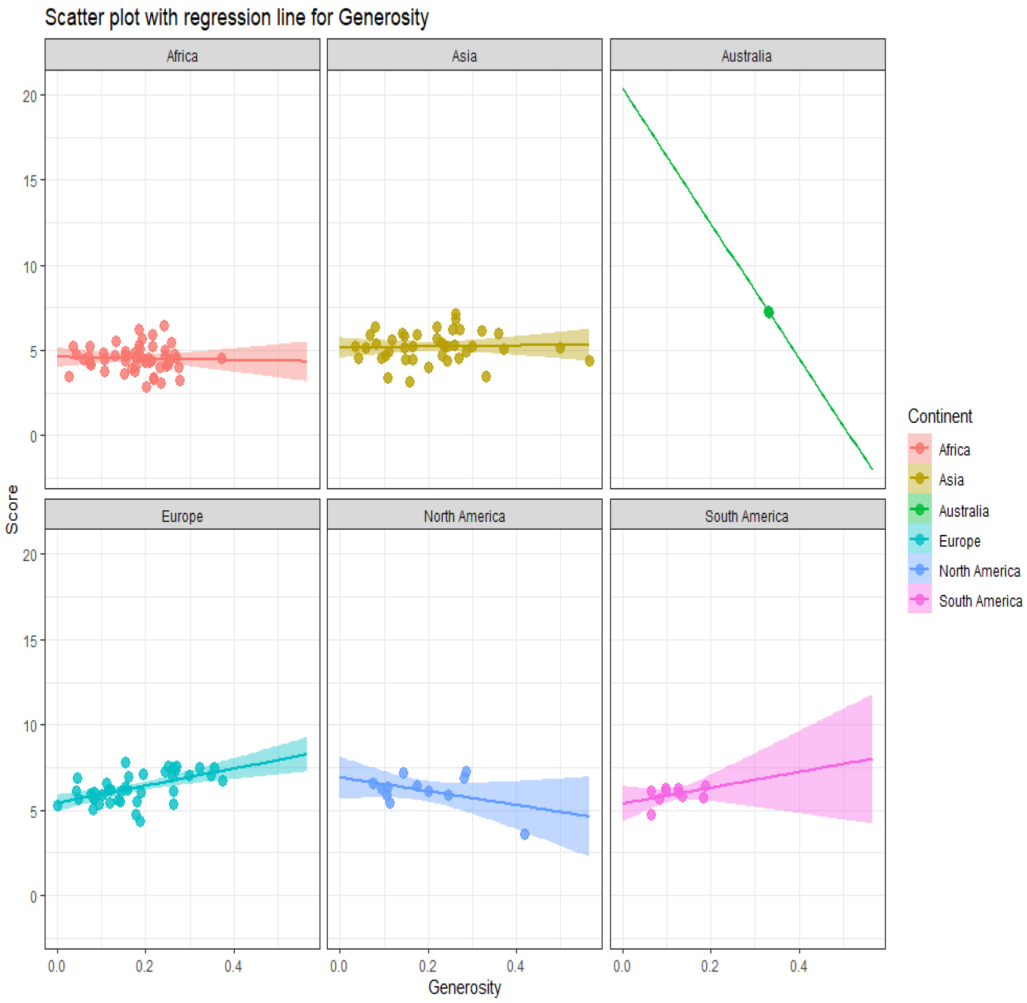
In South America with increase in the family score, the happiness score remains constant.

Similarly, with Asia and Africa, the y-axis range between 2 and 7 shows that there will be very less change in the happiness of people even if they get good social support.



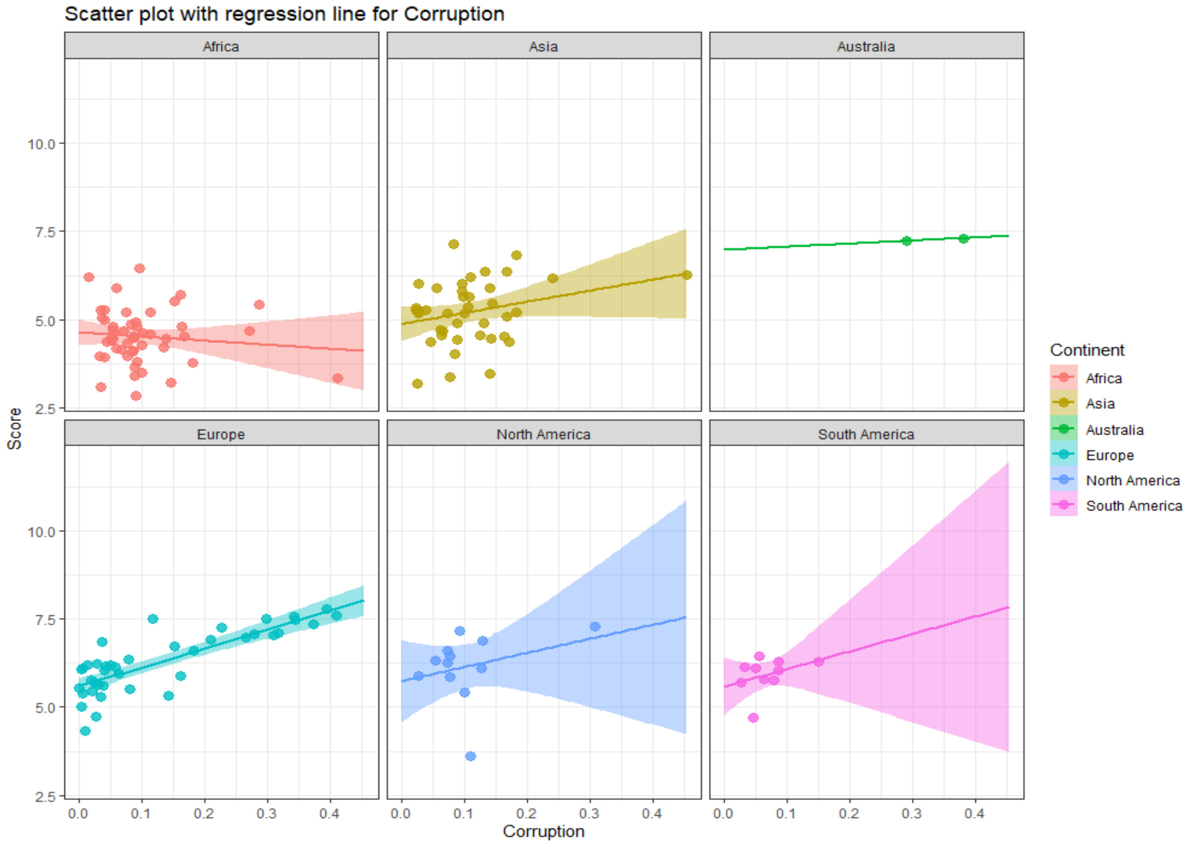
For Generosity

The regression line has a positive slope only for Europe and South America, however the slope for the linear plots is lower as compared to what we saw in life expectancy and GDP per capita above. This tells that though there is positive correlation, but the correlation is not strong enough. For Asia, the line is horizontal, implying that there is no correlation. Surprisingly, if the generosity is increased in Africa and North America, that will reduce people’s happiness, as can be seen through the negative slope line for both continents.



For Perceptions of Corruption

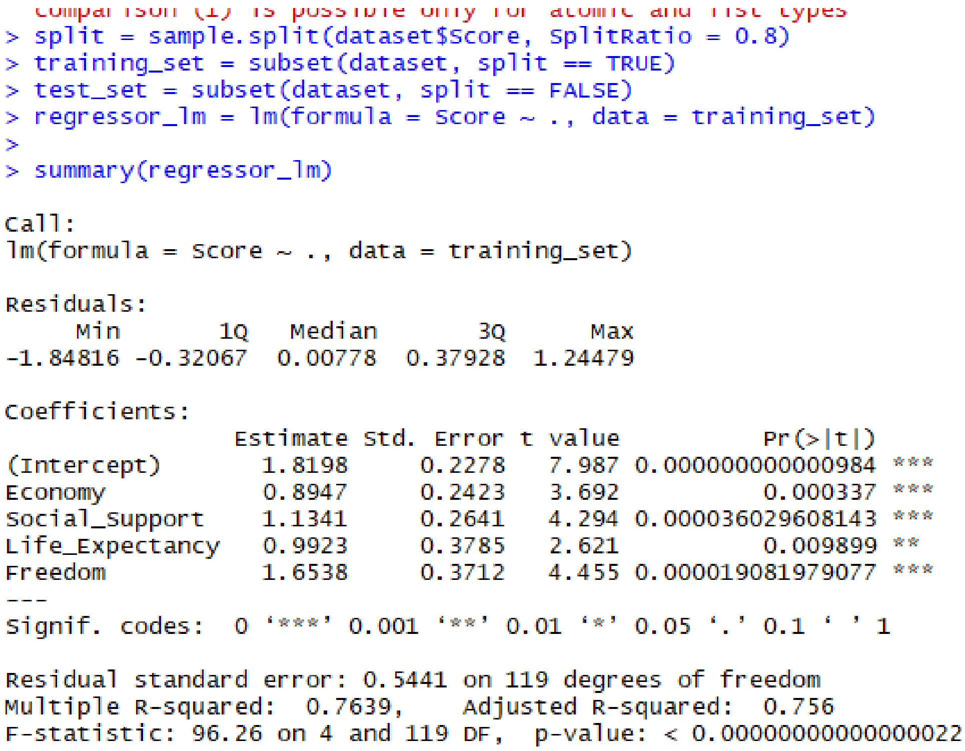
The scattered graphs for all the continents show that approximately there is no correlation between corruption and happiness scores.



Multiple Linear Regression

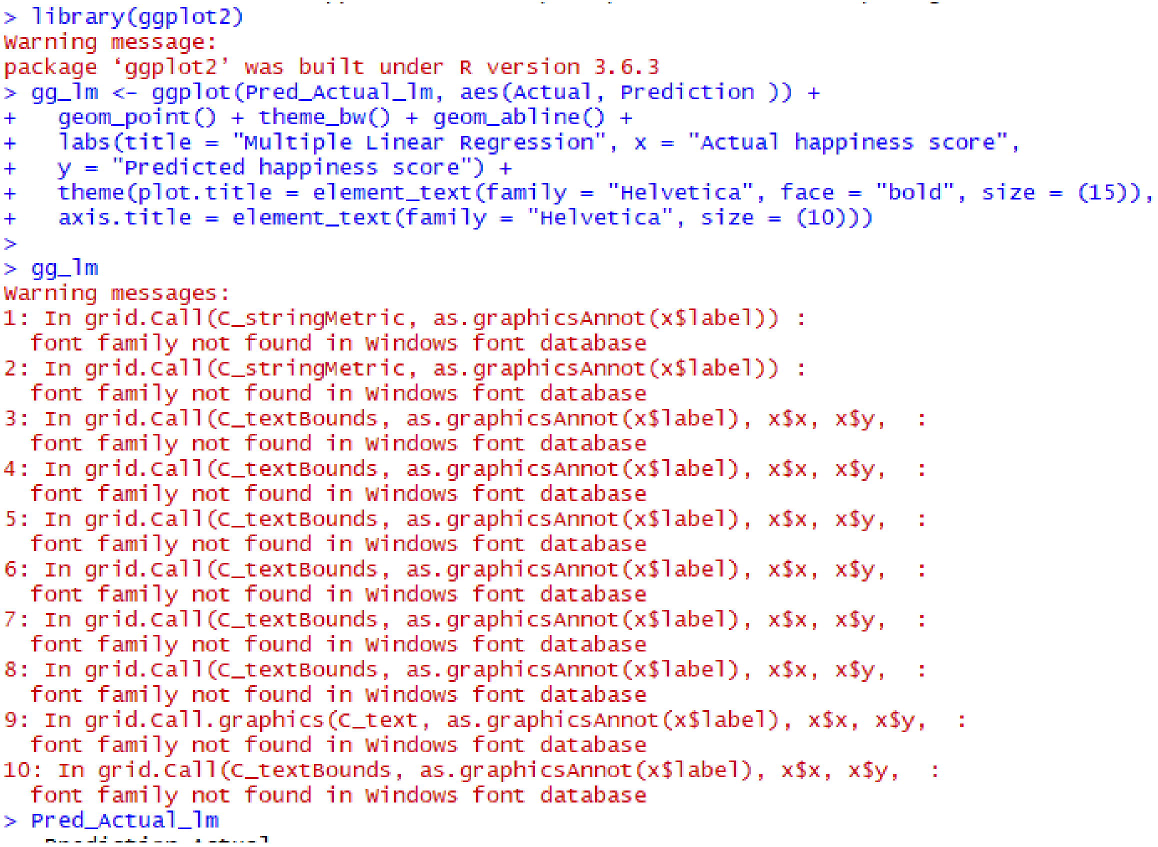
For running the linear regression, we first divide the dataset in train and test dataset, with a split of 80/20. We have 156 overall values in the total dataset, train dataset has 124 observations and test has 32 observations.
Below is the code snippet for splitting the dataset and running the regression model.

Under the “Call”: the formula ran for running the model is shown; “Residuals” describe the data, and the “Coefficients” tells the significance of the regression. The p-value of Economy, Social support, life expectancy and freedom are less than 0.05 and the stars on top of these values tell that all these variables have significant impact on the happiness score. The way to interpret this is, with each unit increase in social support, the happiness increases with a value of 1.1341 with a standard error of 0.2641.

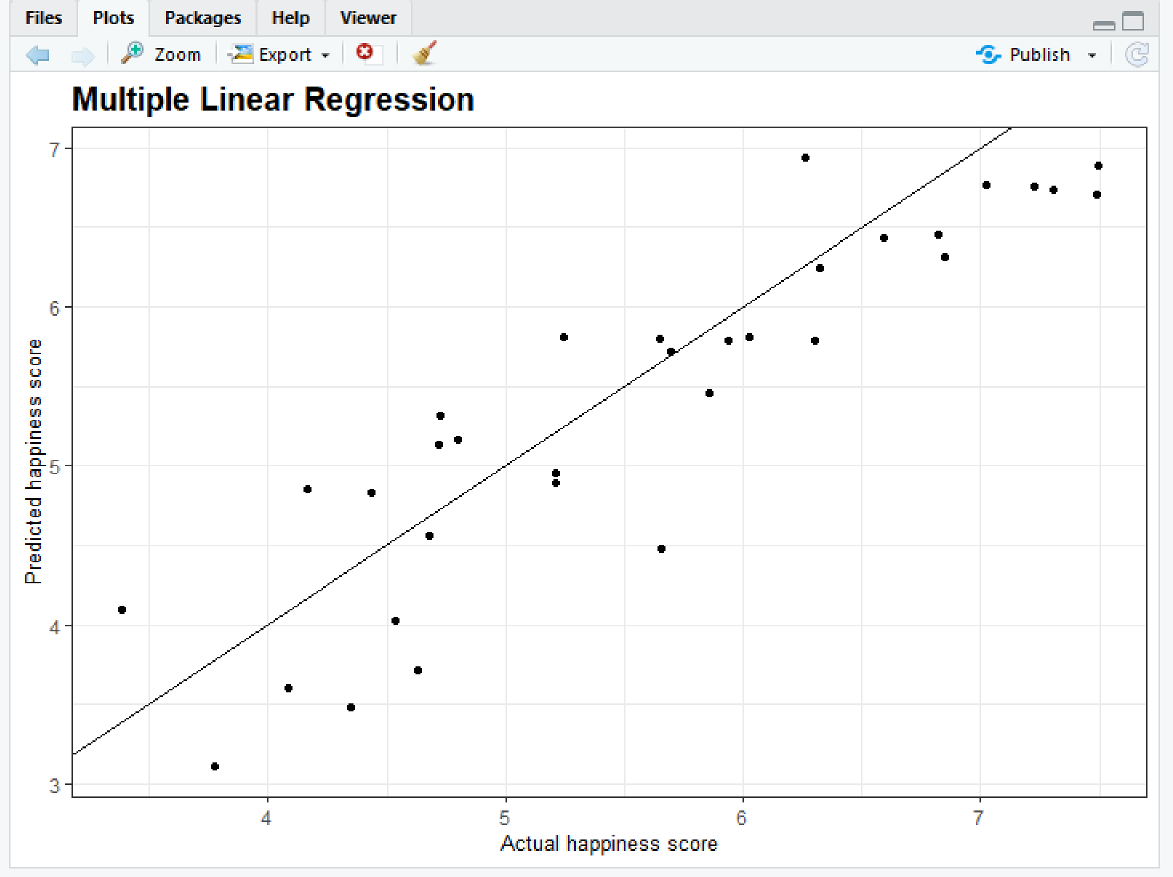


```
  comparison (1) is possible only for atomic and list types
> split = sample.split(dataset$Score, SplitRatio = 0.8)
> training_set = subset(dataset, split == TRUE)
> test_set = subset(dataset, split == FALSE)
> regressor_lm = lm(formula = Score ~ ., data = training_set)
> 
> summary(regressor_lm)

Call:
lm(formula = Score ~ ., data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.84816 -0.32067  0.00778  0.37928  1.24479 

Coefficients:
                Estimate Std. Error t value          Pr(>|t|)    
(Intercept)       1.8198     0.2278   7.987 0.000000000000984 ***
Economy           0.8947     0.2423   3.692          0.000337 ***
Social_Support    1.1341     0.2641   4.294 0.000036029608143 ***
Life_Expectancy   0.9923     0.3785   2.621          0.009899 ** 
Freedom           1.6538     0.3712   4.455 0.000019081979077 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5441 on 119 degrees of freedom
Multiple R-squared:  0.7639,	Adjusted R-squared:  0.756 
F-statistic: 96.26 on 4 and 119 DF,  p-value: < 0.00000000000000022
```

The below code is run to understand how good the predicted values are as compared to the actual values. The same is also plotted for better understanding.

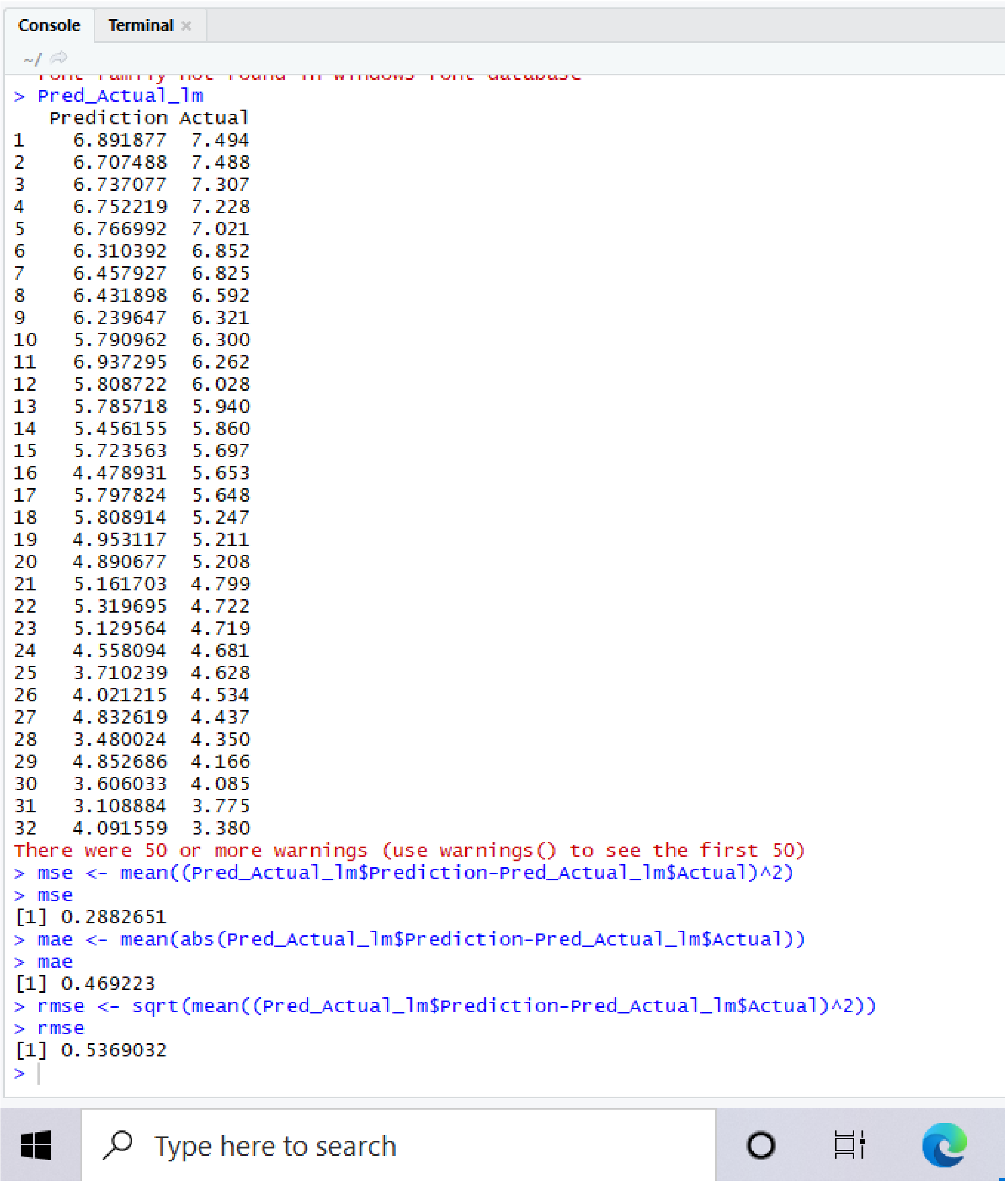


```
> library(ggplot2)
Warning message:
package ‘ggplot2’ was built under R version 3.6.3 
> gg_lm <- ggplot(Pred_Actual_lm, aes(Actual, Prediction )) +
+   geom_point() + theme_bw() + geom_abline() +
+   labs(title = "Multiple Linear Regression", x = "Actual happiness score",
+   y = "Predicted happiness score") +
+   theme(plot.title = element_text(family = "Helvetica", face = "bold", size = (15)), 
+   axis.title = element_text(family = "Helvetica", size = (10)))
> 
> gg_lm
Warning messages:
1: In grid.Call(C_stringMetric, as.graphicsAnnot(x$label)) :
  font family not found in Windows font database
2: In grid.Call(C_stringMetric, as.graphicsAnnot(x$label)) :
  font family not found in Windows font database
3: In grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y,  :
  font family not found in Windows font database
4: In grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y,  :
  font family not found in Windows font database
5: In grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y,  :
  font family not found in Windows font database
6: In grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y,  :
  font family not found in Windows font database
7: In grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y,  :
  font family not found in Windows font database
8: In grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y,  :
  font family not found in Windows font database
9: In grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y,  :
  font family not found in Windows font database
10: In grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y,  :
  font family not found in Windows font database
> Pred_Actual_lm
   Prediction Actual
```



The above plot is for the multi-linear regression between happiness score as given in the dataset and the happiness score predicted by out model. The lines are close to the linear line, which tells us that the model did justice to the predicted values.

For another check, we ran the model on the test dataset to see how the model performs. Below are the results of Mean Square Error, mean Absolute Error and Root Mean Square Error.



```
Font family not found in Windows font database
> Pred_Actual_lm
   Prediction Actual
1    6.891877  7.494
2    6.707488  7.488
3    6.737077  7.307
4    6.752219  7.228
5    6.766992  7.021
6    6.310392  6.852
7    6.457927  6.825
8    6.431898  6.592
9    6.239647  6.321
10   5.790962  6.300
11   6.937295  6.262
12   5.808722  6.028
13   5.785718  5.940
14   5.456155  5.860
15   5.723563  5.697
16   4.478931  5.653
17   5.797824  5.648
18   5.808914  5.247
19   4.953117  5.211
20   4.890677  5.208
21   5.161703  4.799
22   5.319695  4.722
23   5.129564  4.719
24   4.558094  4.681
25   3.710239  4.628
26   4.021215  4.534
27   4.832619  4.437
28   3.480024  4.350
29   4.852686  4.166
30   3.606033  4.085
31   3.108884  3.775
32   4.091559  3.380
There were 50 or more warnings (use warnings() to see the first 50)
> mse <- mean((Pred_Actual_lm$Prediction-Pred_Actual_lm$Actual)^2)
> mse
[1] 0.2882651
> mae <- mean(abs(Pred_Actual_lm$Prediction-Pred_Actual_lm$Actual))
> mae
[1] 0.469223
> rmse <- sqrt(mean((Pred_Actual_lm$Prediction-Pred_Actual_lm$Actual)^2))
> rmse
[1] 0.5369032
> 
```

CONCLUSION

Running the t-test gives shows that two groups (countries with high happiness scores vs low happiness scores) have significant difference in the means, while running ANOVA shows that the means of the groups do not vary with the group having high happiness scores.

The correlation graphs give us the important factors for various continents. Running regression shows which factors are significant overall (considering all continents' data) to give high happiness scores.

The same analysis could be improved running other regression models. For cross-checking the important factors, a Principal Component Analysis could be run which gives the most important factors.

REFERENCES

1) Gardener, M. (2017;2012;). *Statistics for ecologists using R and excel: Data collection, exploration, analysis and presentation* (Second ed.). Exeter, England: Pelagic Publishing.

2) *World Happiness Report*. https://en.wikipedia.org/wiki/World_Happiness_Report#2019_World_Happiness_Report

3) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013;2017;). An introduction to statistical learning: With applications in R. New York: Springer. doi:10.1007/978-1-4614-7138-7

4) Kaggle. https://www.kaggle.com/unsdsn/world-happiness