PREDICTING WORLD HAPPINESS

EXECUTIVE SUMMARY

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

The happiness scores and rankings use data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2013-2016 and use the Gallup weights to make the estimates representative. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

TASK DETAIL

This dataset shows the happiest countries on earth, which is great info when you're looking for your next move but what if you wanted to create a new country with the goal of having the happiest citizens? What if you're a president looking to improve their country? How would you do that?

The goal of this task is to find out what factors contribute to happiness.

VARIABLES

Happiness Rank: Rank of the country based on the Happiness Score.

Happiness Score: A metric measured in 2015 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."

Country: Name of the country

Region: Region the country belongs to.

Economy (GDP per Capita): The extent to which GDP contributes to the calculation of the Happiness Score. The equation uses the natural log of GDP per capita, as this form fits the data significantly better than GDP per capita.

Family: The extent to which Family contributes to the calculation of the Happiness Score. It is the national average of the binary responses (0=no, 1=yes) to the Gallup World Poll (GWP) question, "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

Health (Life Expectancy): The extent to which Life expectancy contributed to the calculation of the Happiness Score. The time series of healthy life expectancy at birth are constructed based on data from the World Health Organization (WHO) Global Health Observatory data repository, with data available for 2005, 2010, 2015, and 2016. To match this report's sample period, interpolation and extrapolation are used

Freedom: The extent to which Freedom contributed to the calculation of the Happiness Score. Freedom to make life choices is the national average of binary responses to the GWP question, "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

Trust: The extent to which Perception of Corruption contributes to Happiness Score. It is the average of binary answers to two GWP questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?" Where data for government corruption are missing, the perception of business corruption is used as the overall corruption-perception measure.

Generosity: The extent to which Generosity contributed to the calculation of the Happiness Score. Generosity is the residual of regressing on GDP per capita the national average of GWP responses to the question, "Have you donated money to a charity in the past month?"

Dystopia Residual: The extent to which Dystopia Residual contributed to the calculation of the Happiness Score. Dystopia Residual metric actually is the Dystopia Happiness Score (1.85) + the Residual value or the unexplained value for each country

PREPARING AND DESCRIBING THE DATA

Let's begin with importing the data into R studios. The data is open source data on Kaggle (link below). It is separated into five files, each file pertaining to a different year. It would be interesting to observe the data for the current year. Once the data is imported, we would remove all the columns that appear to be unnecessary to this analysis.

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

```
# Reading the file
happinessScore <- read.csv("2019.csv")
dim(happinessScore)
str(happinessScore)
head(happinessScore)

# Summarising the dataset
summary(happinessScore)
```

The dataset contains **156 records** and **9 variables** regarding happiness score in 2019 report. Here each record pertains to a distinct country.

```
> summary(happinessScore)
  Overall.rank      Country.or.region     Score       GDP.per.capita   Social.support  Healthy.life.expectancy
 Min.   :  1.00   Afghanistan:  1     Min.   :2.853   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.: 39.75   Albania    :  1     1st Qu.:4.545   1st Qu.:0.6028   1st Qu.:1.056   1st Qu.:0.5477
 Median : 78.50   Algeria    :  1     Median :5.380   Median :0.9600   Median :1.272   Median :0.7890
 Mean   : 78.50   Argentina  :  1     Mean   :5.407   Mean   :0.9051   Mean   :1.209   Mean   :0.7252
 3rd Qu.:117.25   Armenia    :  1     3rd Qu.:6.184   3rd Qu.:1.2325   3rd Qu.:1.452   3rd Qu.:0.8818
 Max.   :156.00   Australia  :  1     Max.   :7.769   Max.   :1.6840   Max.   :1.624   Max.   :1.1410
                  (Other)    :150
 Freedom.to.make.life.choices   Generosity     Perceptions.of.corruption
 Min.   :0.0000               Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.3080               1st Qu.:0.1087   1st Qu.:0.0470
 Median :0.4170               Median :0.1775   Median :0.0855
 Mean   :0.3926               Mean   :0.1848   Mean   :0.1106
 3rd Qu.:0.5072               3rd Qu.:0.2482   3rd Qu.:0.1412
 Max.   :0.6310               Max.   :0.5660   Max.   :0.4530
```

The mean happiness score for all the countries is **5.407**, the median is slightly lower than the mean suggesting that **the distribution** of scores is slightly skewed towards the right.

At this stage, *'Overall rank'* and *'country or region'* columns are removed, as they won't be necessary in this analysis.

```
# Removing country name and happiness rank from the dataset
happinessScoredf<-happinessScore[,-c(1,2)]
str(happinessScoredf)
```

DATA VISUALISATION

This visual gives us a more appealing view of factors affecting the happiness score in World ranking report. Let's create a histogram to understand the distribution of happiness score across all the countries.

```
# Distribution of Happiness Scores
hist(happinessScore$Score,
     main="Histogram",
     xlab="Happiness Score",
     xlim=c(0,10),
     col="darkseagreen",
     freq=FALSE)
```

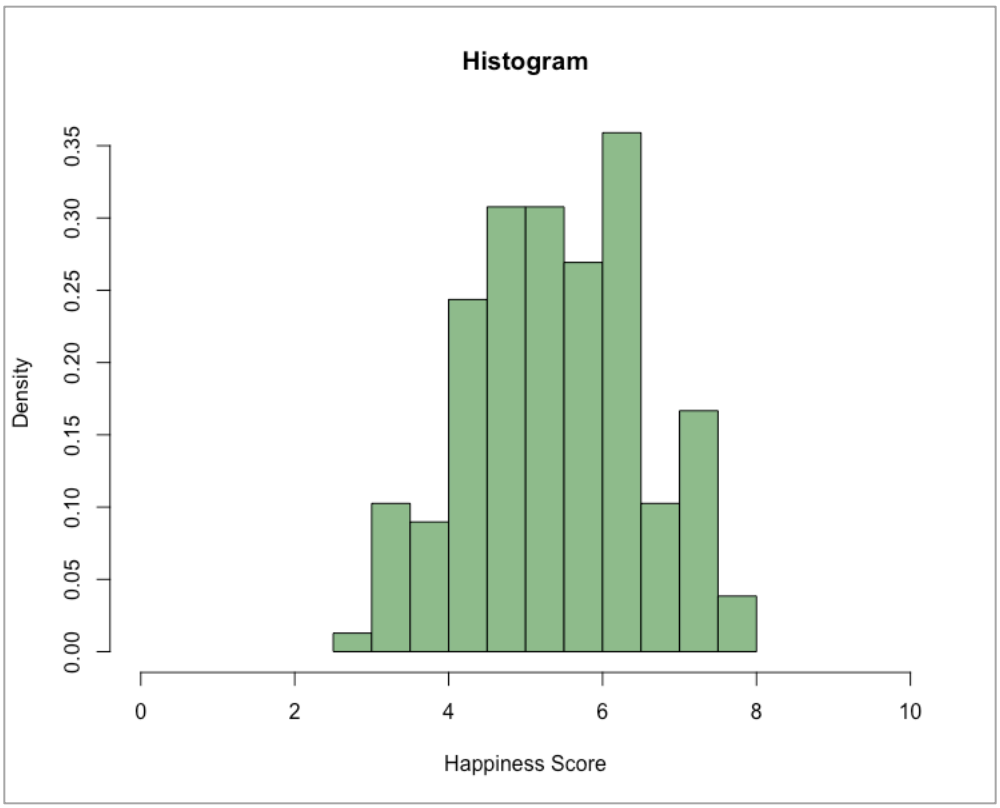

Fig 1. Histogram for Happiness Score

It can be observed from Fig 1 that the data for happiness score is slightly skewed towards the right.

Which of the six variables most affect a country's "happiness"? Is it a country's government or economy that makes its citizens the happiest? We can create scatter plots and perform regression to see how correlated a variable is with happiness score.

```
# Relationship between GDP per Capita and Happiness Score
p <- ggplot(happinessScore, aes(x=Score, y=GDP.per.capita)) +  geom_point(color = 'deepskyblue3')+
  geom_smooth(method=lm, color='gray46')

p + ggtitle("GDP vs Score") + theme(plot.title = element_text(hjust = 0.5))+
  xlab("GDP per capita") + ylab("Score")
```

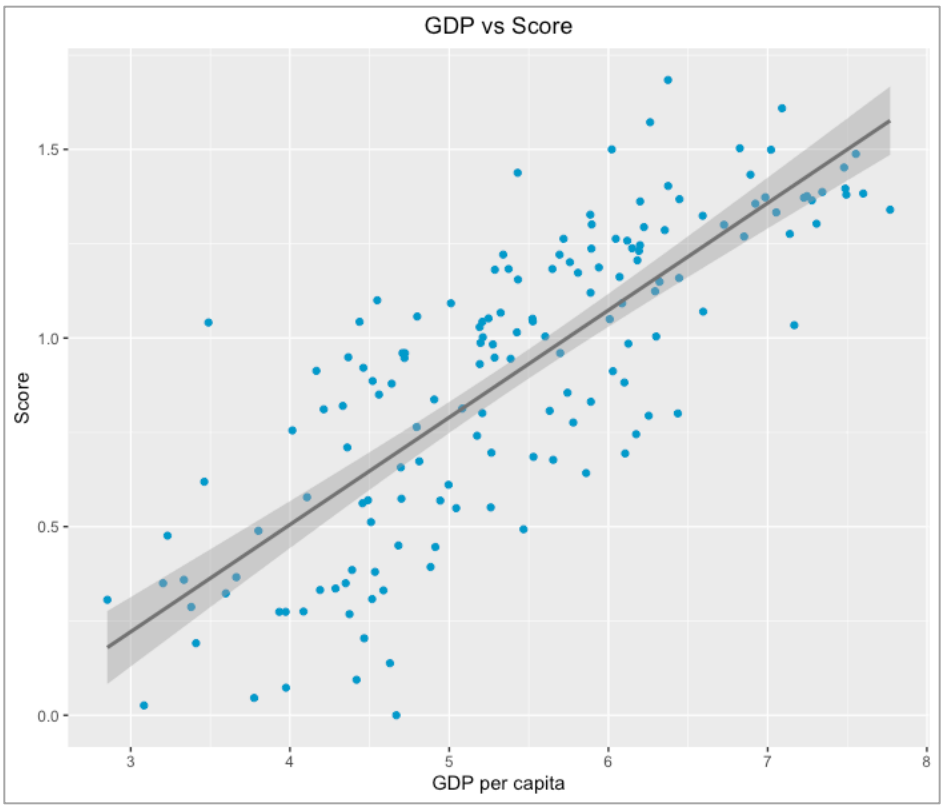

Fig 2. Scatterplot for GDP vs Happiness Score

There is a strong positive correlation between GDP per capita and Happiness Score; as the GDP per capita for a country increases the happiness score also increases and vice versa.

```
# Relationship between Social Support and Happiness Score
p <- ggplot(happinessScore, aes(x=Score, y=Social.support)) +  geom_point(color = 'deepskyblue3')+
  geom_smooth(method=lm, color='gray46')

p + ggtitle("Social Support vs Score") + theme(plot.title = element_text(hjust = 0.5))+
  xlab("Social Support") + ylab("Score")
```

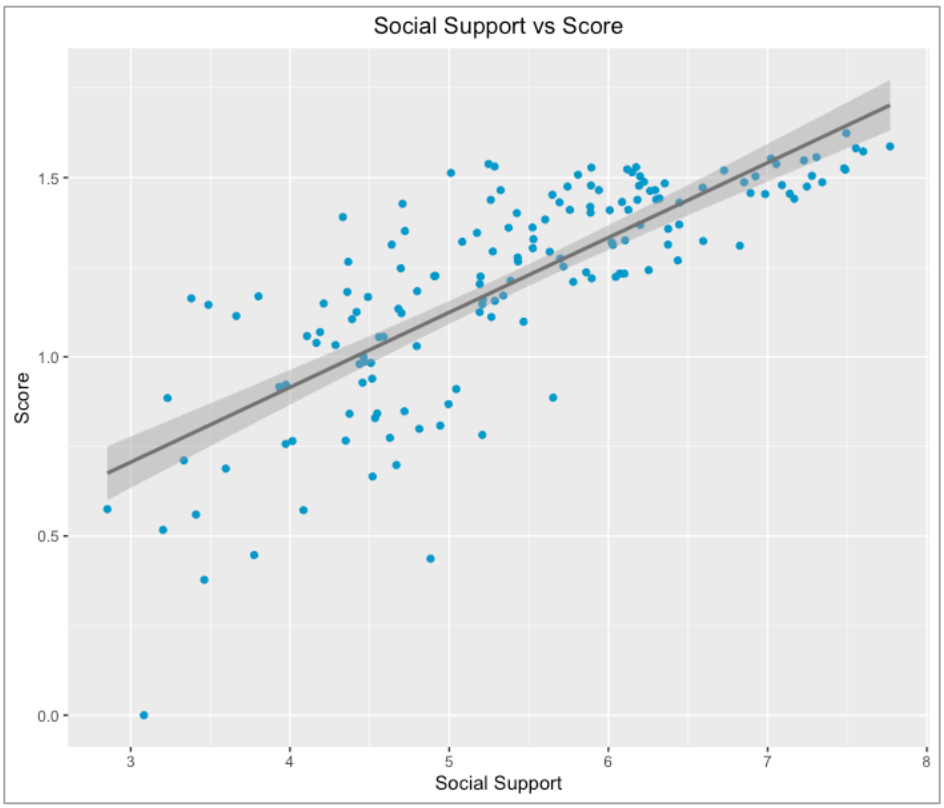

Fig 3. Scatterplot for Social Support (Family) vs Happiness Score

There is a strong positive correlation between Social Support and Happiness Score; as the Social Support for a country increases the happiness score also increases and vice versa.

```
# Relationship between Healthy Life Expectancy and Happiness Score
p <- ggplot(happinessScore, aes(x=Score, y=Healthy.life.expectancy)) +  geom_point(color = 'deepskyblue3')+
  geom_smooth(method=lm, color='gray46')

p + ggtitle("Life Expectancy vs Score") + theme(plot.title = element_text(hjust = 0.5))+
  xlab("Healthy Life Expectancy") + ylab("Score")
```

Fig 4. Scatterplot for Healthy Life Expectancy vs Happiness Score

There is a strong positive correlation between Healthy Life Expectancy and Happiness Score; as the Life Expectancy for a country increases the happiness score also increases and vice versa.

```
# Relationship between Freedom to make life choices and Happiness Score
p <- ggplot(happinessScore, aes(x=Score, y=Freedom.to.make.life.choices)) +  geom_point(color = 'deepskyblue3')+
  geom_smooth(method=lm, color='gray46')

p + ggtitle("Freedom vs Score") + theme(plot.title = element_text(hjust = 0.5))+
  xlab("Freedom to make life choices") + ylab("Score")
```

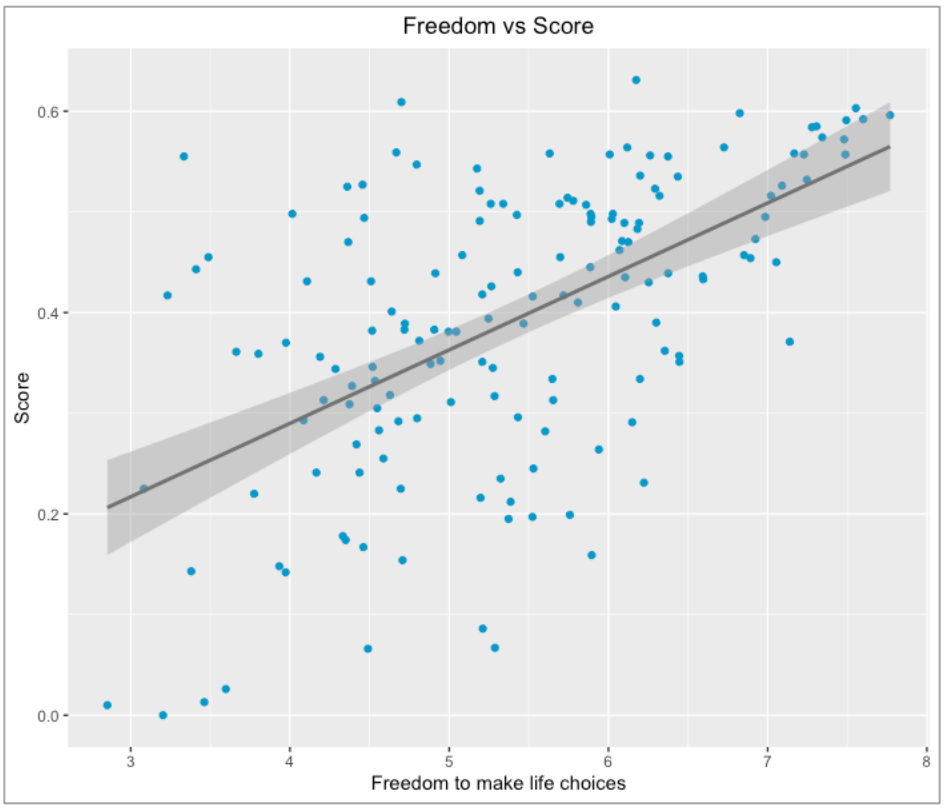

Fig 5. Scatterplot for Freedom to make choices vs Happiness Score

There is a strong positive correlation between Freedom to make choices and Happiness Score; as the Freedom to make choices for a country increases the happiness score also increases and vice versa.

```
# Relationship between Generosity and Happiness Score
p <- ggplot(happinessScore, aes(x=Score, y=Generosity)) +  geom_point(color = 'deepskyblue3')+
  geom_smooth(method=lm, color='gray46')

p + ggtitle("Generosity vs Score") + theme(plot.title = element_text(hjust = 0.5))+
  xlab("Generosity") + ylab("Score")
```

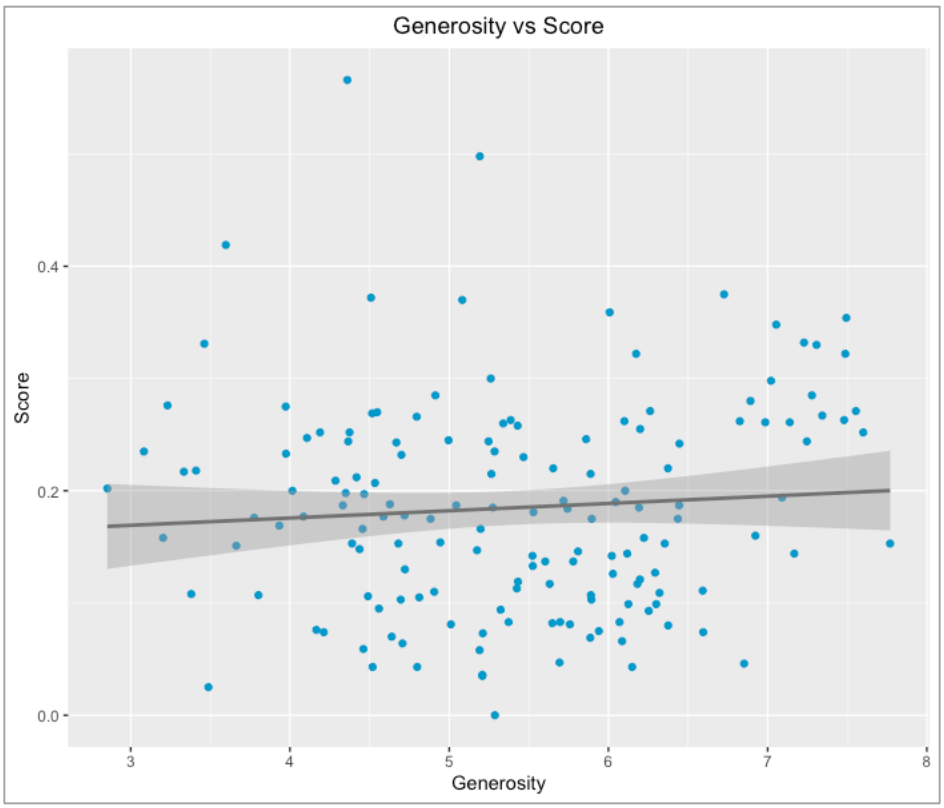

Fig 6. Scatterplot for Generosity vs Happiness Score

There is a positive correlation between Generosity and Happiness Score however the relationship is quite weak; as the Generosity for a country increases the happiness score also increases and vice versa.

```
# Relationship between Perception of corruption and Happiness Score
p <- ggplot(happinessScore, aes(x=Score, y=Perceptions.of.corruption)) +  geom_point(color = 'deepskyblue3')+
  geom_smooth(method=lm, color='gray46')

p + ggtitle("Corruption vs Score") + theme(plot.title = element_text(hjust = 0.5))+
  xlab("Perceptions of corruption") + ylab("Score")
```

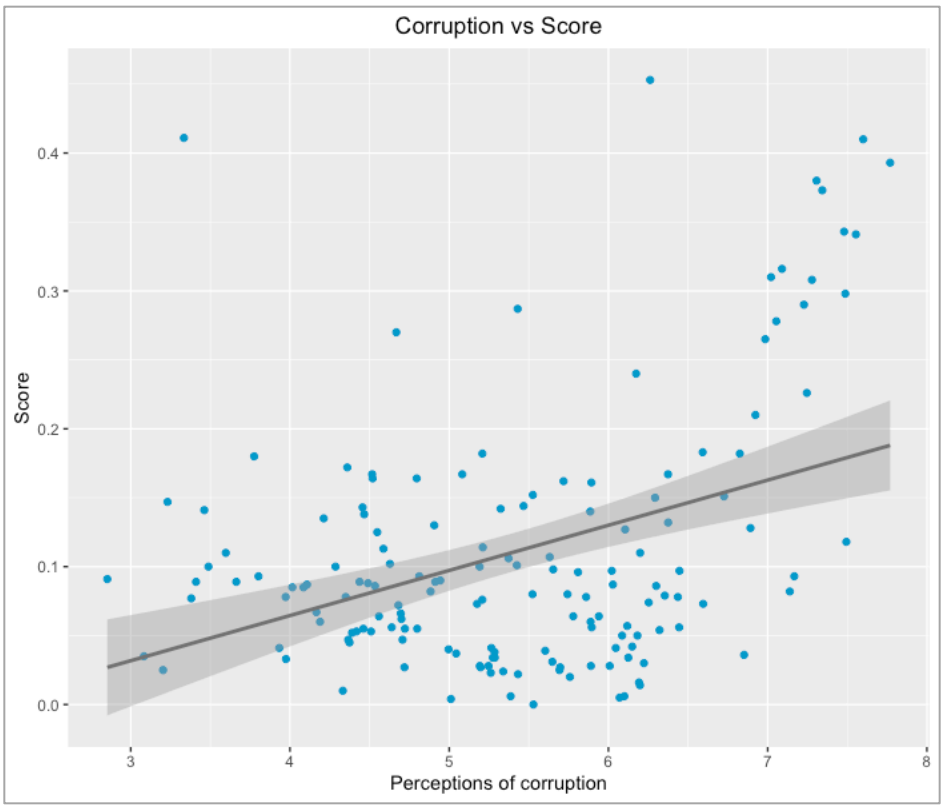

Fig 7. Scatterplot for Perception of Corruption vs Happiness Score

There is a slightly positive correlation between Perception of corruption and Happiness Score. as the Perception of corruption for a country increases the happiness score also increases and vice versa.

To proceed with developing a predictive model. Let's create a heatmap to get an overall idea of the correlation among all the variables. Since our focus is happiness score, let's concentrate on that column.

The lighter blue square depicts a stronger positive correlation, and obviously, variables will have a correlation of 1 with each other. We can see that happiness score is really strongly correlated with economy, and health, followed by family and freedom. Thus, in our model, we should see the coefficients of these variables must be higher compared to the coefficients of trust and generosity.

```
# HeatMap to see the correlation between the variables and the Happiness Score
# Removing country name and happiness rank from the dataset
happinessScoredf<-happinessScore[,-c(1,2)]
str(happinessScoredf)

cor.mat<-round(cor(happinessScoredf),2)
melted.cor.mat<-melt(cor.mat)

p<-ggplot(melted.cor.mat, aes(x = X1, y= X2, fill = value))+ geom_tile()+geom_text(aes(x=X1, y= X2, label = value))
p + ggtitle("Correlation heatmap for variables") + theme(plot.title = element_text(hjust = 0.5))+
  xlab("") + ylab("")
```

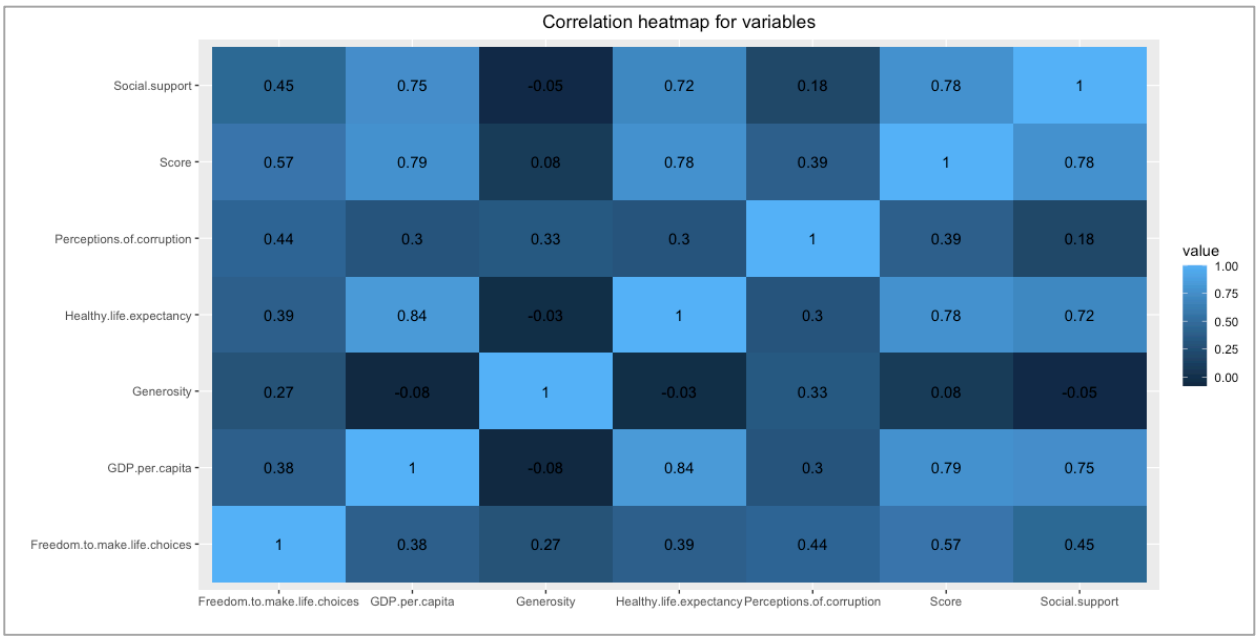

Fig 8. Correlation Heatmap for variables

Moving on, now that we have a bit of an understanding of the relationship between variables, we can start with training the model on training data and then testing the performance of the model on an unseen data. Let's partition the data first in training (60%) and testing (40%) sets.

```
# Linear Regression Model
# Partitioning data
set.seed(1)
train.index<-sample(rownames(happinessScoredf),dim(happinessScoredf)[1]*0.6)
test.index<-setdiff(rownames(happinessScoredf),train.index)

train.df<-happinessScoredf[train.index,]
test.df<-happinessScoredf[test.index,]

# Linear Regression Model training
linearmodel1<-lm(Score~.,data=train.df)
summary(linearmodel1)

# Linear Regression Model testing
linearmodel1.predict=predict(linearmodel1,test.df)
accuracy(linearmodel1.predict,test.df$Score)
```

Model Interpretation:

The coefficient value signifies how much the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. Looking at the model summary, we can observe that the happiness score for a country would be 1.8219 if all the six variables are zero. The coefficient for social support variable is highest among all the variables, with a value of 1.1795 which means that if there is an increase in the social support value by 1 unit, while keeping the other variables constant, the happiness score for a country would increase by 1.1795. The lowest coefficient is for generosity, which indicates that with increase in 1 unit in this variable the happiness score would be increased by the least amount compared to other 5 variables.

P value for GDP, social support, healthy life expectancy and freedom to make life choices is more than 0.05, which means that the sample data provides enough evidence to reject the null hypothesis for the entire population. The data

favor the hypothesis that there is a non-zero correlation. Changes in the independent variable are associated with changes in the response at the population level. This variable is statistically significant and probably a worthwhile addition to our regression model.

R-Squared evaluates the scatter of the data points around the fitted regression line. It is also called the coefficient of determination. The value of adjusted R-squared is 76.87%, it represents that the model explains approximately 77% of the variation in the response variable around its mean.

```
> linearmodel1<-lm(Score~.,data=train.df)
> summary(linearmodel1)

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.76456 -0.31822  0.07593  0.38334  1.23797

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.8219     0.2807   6.490 5.25e-09 ***
GDP.per.capita                0.8792     0.2780   3.162 0.002162 **
Social.support                1.1795     0.3143   3.753 0.000316 ***
Healthy.life.expectancy       0.9409     0.4299   2.189 0.031341 *
Freedom.to.make.life.choices  1.2282     0.5355   2.294 0.024253 *
Generosity                    0.5056     0.7454   0.678 0.499358
Perceptions.of.corruption     0.8722     0.7323   1.191 0.236942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5628 on 86 degrees of freedom
Multiple R-squared:  0.7838,    Adjusted R-squared:  0.7687
F-statistic: 51.95 on 6 and 86 DF,  p-value: < 2.2e-16
```

Here, is the performance summary/accuracy of the model on the testing data set. The table below shows the Mean Error, Root Mean Squared Error, Mean Average Error, Mean Percentage Error and Mean Average Eercentage error for the predicted values with respect to the actual happiness score.

```
> linearmodel1.predict=predict(linearmodel1,test.df)
> accuracy(linearmodel1.predict,test.df$Score)
                 ME      RMSE       MAE        MPE     MAPE
Test set 0.02261244 0.4951132 0.4020897 -0.3918947 7.672662
```

Let’s analyze the residuals for testing dataset. The histogram shows the distribution of residuals, the mean error is approx. zero. The maximum number of residuals are between the range -0.5 to +0.5. There are 20 records where the residuals are above or below one standard deviation from the mean. So, approximately 87% of the records have the value of residual under one standard deviation from the mean. Therefore, we can say that the model did pretty well on the testing data and the predicted values are quite close to the actual score.

```r
#Residual Analysis
test.res <- data.frame(test.df$Score, linearmodel1.predict,
                       residuals = test.df$Score - linearmodel1.predict)

hist(test.res$residuals,
     main="Histogram",
     xlab="Residuals",
     col="darkseagreen",
     freq=FALSE)

standard_dev_res <- sd(test.res$residuals)
mean_res <- mean(test.res$residuals)

standard_dev_res
mean_res
length(test.res$residuals[which(test.res$residuals>standard_dev_res | test.res$residuals< - standard_dev_res)])
```

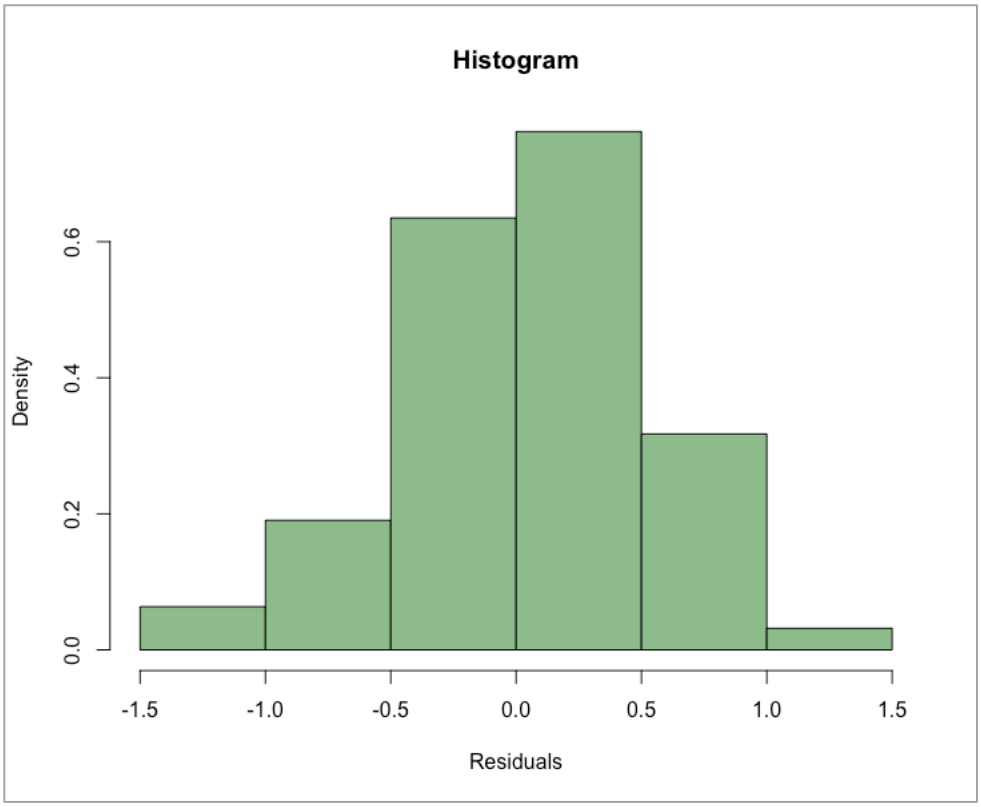

Fig 9. Histogram of residuals

```
> standard_dev_res
[1] 0.4985692
> mean_res
[1] 0.02261244
> length(test.res$residuals[which(test.res$residuals>standard_dev_res | test.res$residuals< - standard_dev_res)])
[1] 20
```

In conclusion, our happiness score for world happiness can be used using the model above. We have built a preliminary machine learning tool that will help us generate country scores, and the higher the score, the more highly ranked the happiness of that country will be. Of course, there is always tools and analysis you can do further to this model in order to make it more and more accurate, and better to use. It would be beneficial to further explore

a comparison between the five years in our report, and also look at comparisons between subcontinents. Although, we have a pretty good start in order to further investigate this data.