HarvardX: PH125.9x
Data Science

World Happiness Index Project

Compiled by Davis Varghese

4/29/2020

Overview

Introduction

The latest version of the World happiness report released in 2019 measured 156 countries on several socio economic parameters to score and rank each country and most notably include factors like links between government and happiness, the power of prosocial behaviour and changes in information technology. This report is the undisputed survey of the state of global happiness, first published in 2012 and is widely used as a reference by governments and organizations to influence their policy-making decisions based on the happiness indicators. The Gallup World Poll data is used to power the scores and rankings which is based on answers to the question called the main life evaluation question. Also, known as the Cantril ladder, this question asks those taking the survey to think of an imaginary ladder with steps where in the best possible life for them being at the top step or a 10 and the worst possible life being at the bottom step or a 0 and to rate their own current lives on a scale based on that premise.

The data thus collated is split into several columns listed below:

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

It's worth noting that previous reports until 2017 used to include a data column called Residual Dystopia.

Dystopia & Residual Dystopia
'Dystopia' which is the state opposite of 'Utopia' is perceived to be an imaginary nation where all inhabitants are generally unhappy when measured against the six key parameters like income, life expectancy, generosity, corruption, freedom and social support. It serves as a baseline to compare other performing countries against these six

parameters. The residuals which are usually the unexplained components differ for each country and have an average value of approximately zero over the entire set of countries.

Via this report, we will explore the co-relation of parameters on the overall happiness and analyse some models to conclude the best model which can be used to estimate overall happiness scores based on the RMSE value or Root Mean Square Error value as an indicator.

Methodology & Data Analysis

Initial Data Analysis

Datasets used for data analysis in this report have been made available as csv files for the years 2015 through 2019 as part of the kaggle.com datasets for World Happiness report. We will pick up the recent years i.e. 2018 and 2019 to do a comparative analysis of the parameters influencing World Happiness for both the years.

Accordingly, lets load the 2019 data into a dataframe and display the 10 highest ranked countries.

2019 Data with 10 highest scores:

```
#####################################################################
# Connect to github,import and load 2019 data into a dataframe and show 10
scores
#####################################################################
urlfile1 <-
'http://raw.githubusercontent.com/davivarg/CYOWorldHappiness/master/2019.csv'
data_2019 <- read.csv(url(urlfile1))
head(data_2019, n =10)
```

```
##    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
## 7             7            Sweden 7.343          1.387          1.487
## 8             8       New Zealand 7.307          1.303          1.557
## 9             9            Canada 7.278          1.365          1.505
## 10           10           Austria 7.246          1.376          1.475
##    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
## 7                    1.009                          0.574      0.267
## 8                    1.026                          0.585      0.330
## 9                    1.039                          0.584      0.285
## 10                   1.016                          0.532      0.244
##    Perceptions.of.corruption
## 1                      0.393
## 2                      0.410
## 3                      0.341
## 4                      0.118
## 5                      0.298
## 6                      0.343
## 7                      0.373
## 8                      0.380
## 9                      0.308
## 10                     0.226
```

Now, lets load the 2018 data into a dataframe and display the 10 highest ranked countries.

2018 Data with 10 highest scores:

```
#####################################################################
# Connect to github,import and load 2018 data into a dataframe and show 10
scores
#####################################################################
urlfile2 <-
'http://raw.githubusercontent.com/davivarg/CYOWorldHappiness/master/2018.csv'
data_2018 <- read.csv(url(urlfile2))
head(data_2018, n =10)
```

```
##    Overall.rank Country.or.region Score GDP.per.capita Social.support
## 1             1           Finland 7.632          1.305          1.592
## 2             2            Norway 7.594          1.456          1.582
## 3             3           Denmark 7.555          1.351          1.590
## 4             4           Iceland 7.495          1.343          1.644
## 5             5       Switzerland 7.487          1.420          1.549
## 6             6       Netherlands 7.441          1.361          1.488
## 7             7            Canada 7.328          1.330          1.532
## 8             8       New Zealand 7.324          1.268          1.601
## 9             9            Sweden 7.314          1.355          1.501
## 10           10         Australia 7.272          1.340          1.573
##    Healthy.life.expectancy Freedom.to.make.life.choices Generosity
## 1                    0.874                       0.681      0.202
## 2                    0.861                       0.686      0.286
## 3                    0.868                       0.683      0.284
## 4                    0.914                       0.677      0.353
```

```
## 5                    0.927                          0.660      0.256
## 6                    0.878                          0.638      0.333
## 7                    0.896                          0.653      0.321
## 8                    0.876                          0.669      0.365
## 9                    0.913                          0.659      0.285
## 10                   0.910                          0.647      0.361
##    Perceptions.of.corruption
## 1                      0.393
## 2                      0.340
## 3                      0.408
## 4                      0.138
## 5                      0.357
## 6                      0.295
## 7                      0.291
## 8                      0.389
## 9                      0.383
## 10                     0.302
```

As can be seen, the first 10 nations leading the Happiness Index table for both the years remain unchanged except for the rankings.

The histogram of the 2018 Happiness score and a summary can be obtained using the following code:

```
######################################################################
# Summary of 2018 Happiness scores
######################################################################
summary(data_2018$Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.905   4.454   5.378   5.376   6.168   7.632
```

```
#################################################################
# Histogram of 2018 Happiness scores
#################################################################
hist(data_2018$Score, col="blue", border="white", main = "Global Happiness
Score in 2018", xlab = "Happiness Score")
```

Similarly, the histogram of the 2019 Happiness score and a summary can be obtained using the following code:

```
#################################################################
# Summary of 2019 Happiness scores
#################################################################
summary(data_2019$Score)


##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.853   4.545   5.380   5.407   6.184   7.769
```

```
#################################################################
# Histogram of 2019 Happiness scores
#################################################################
hist(data_2019$Score, col="green", border="white", main = "Global Happiness
Score in 2019", xlab = "Happiness Score")
```

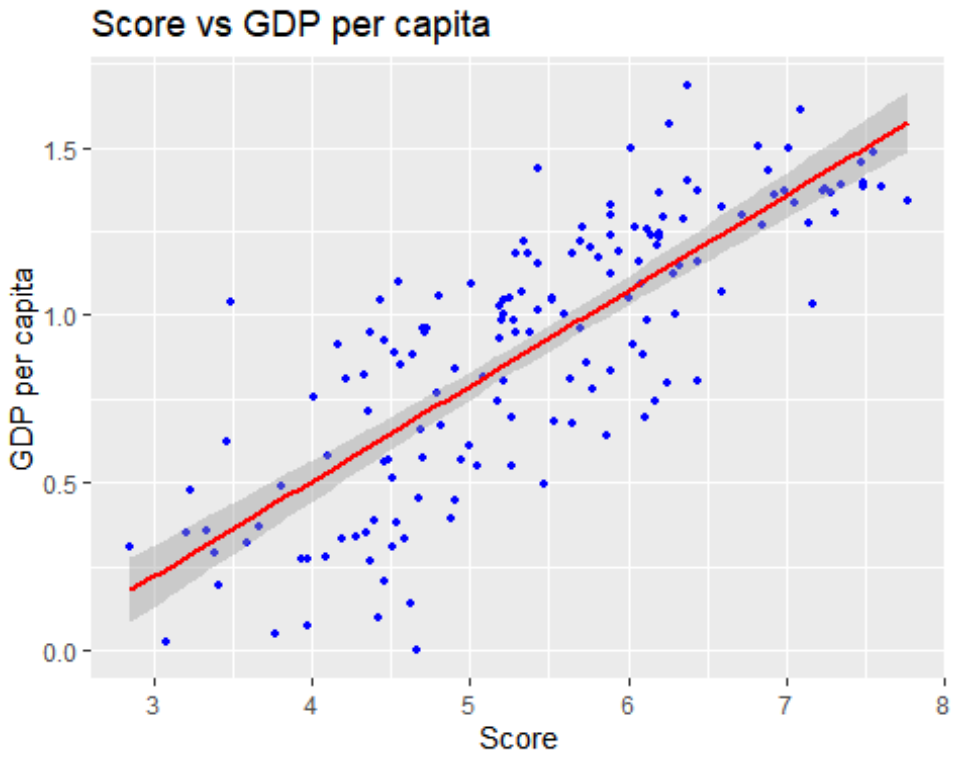
Comparing the mean and median for both years, we can see that the median is slightly on the higher side for 2018 indicating a slightly negatively skewed distribution whereas for 2019 it's on the positive side.

Now for 2019, let's try to find the correlation of each parameter with the score by plotting these parameters. The parameters in question being:

- GDP per capita
- Social support
- Healthy life expectancy
- Freedom to make life choices
- Generosity
- Perceptions of corruption

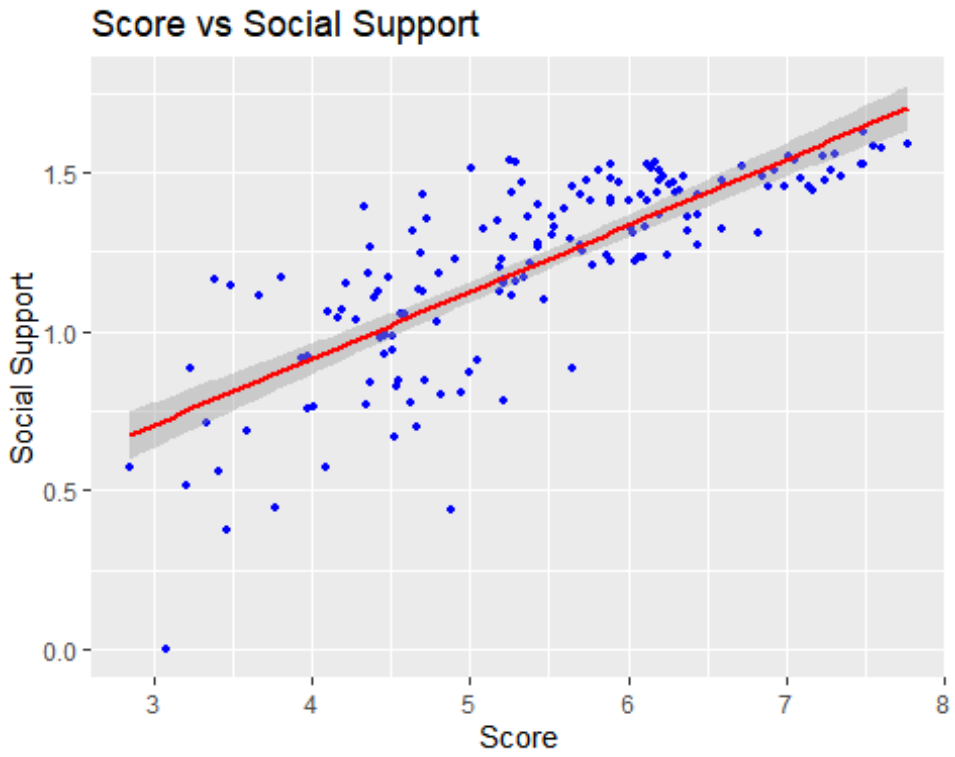
GDP per capita:
A plot of Score vs GDP per capita shown below shows a clear positive correlation i.e when GDP per capita increases, Happiness Score also increases.

```
#################################################################
# Plot of Score vs GDP per capita
#################################################################
  ggplot(data = data_2019, aes(x = Score, y = GDP.per.capita)) +
  geom_point(color='blue', size = 1) +
  labs(y = "GDP per capita")+
  labs(title = paste("Score vs GDP per capita"))+
  geom_smooth(method = "lm", color = "red", se = TRUE)
```



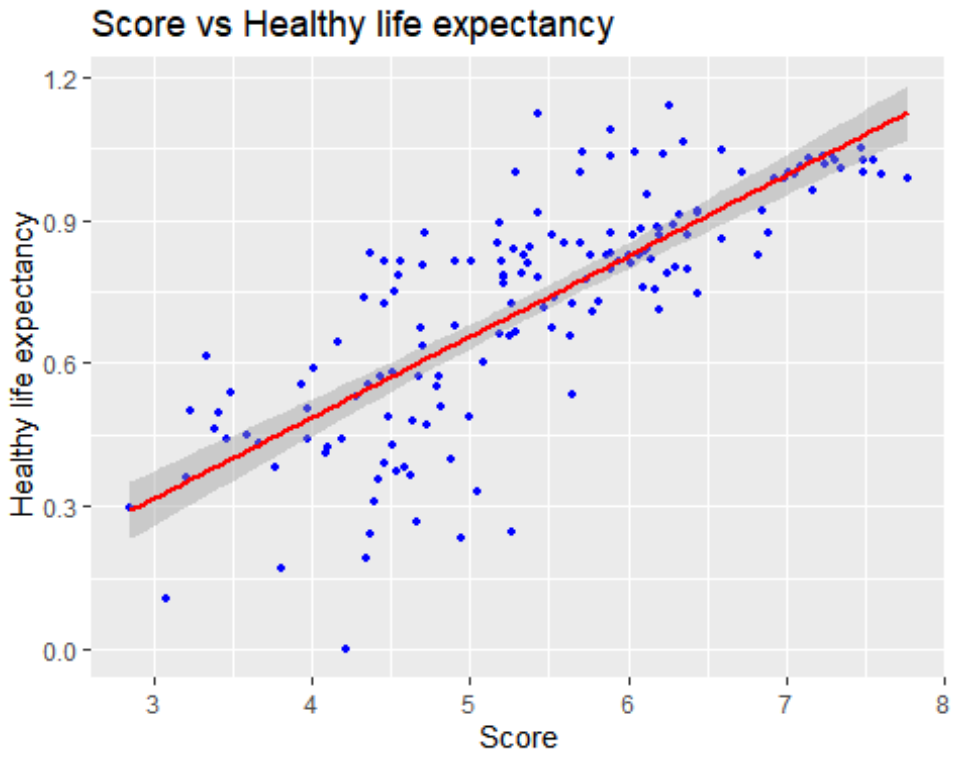
Social Support:
A plot of Score vs Social Support shown below shows a clear positive correlation i.e when Social Support increases, Happiness Score also increases.

```
#################################################################
# Plot of Score vs Social Support
#################################################################
  ggplot(data = data_2019, aes(x = Score, y = Social.support)) +
  geom_point(color='blue', size = 1) +
  labs(y = "Social Support")+
  labs(title = paste("Score vs Social Support"))+
  geom_smooth(method = "lm", color = "red", se = TRUE)
```



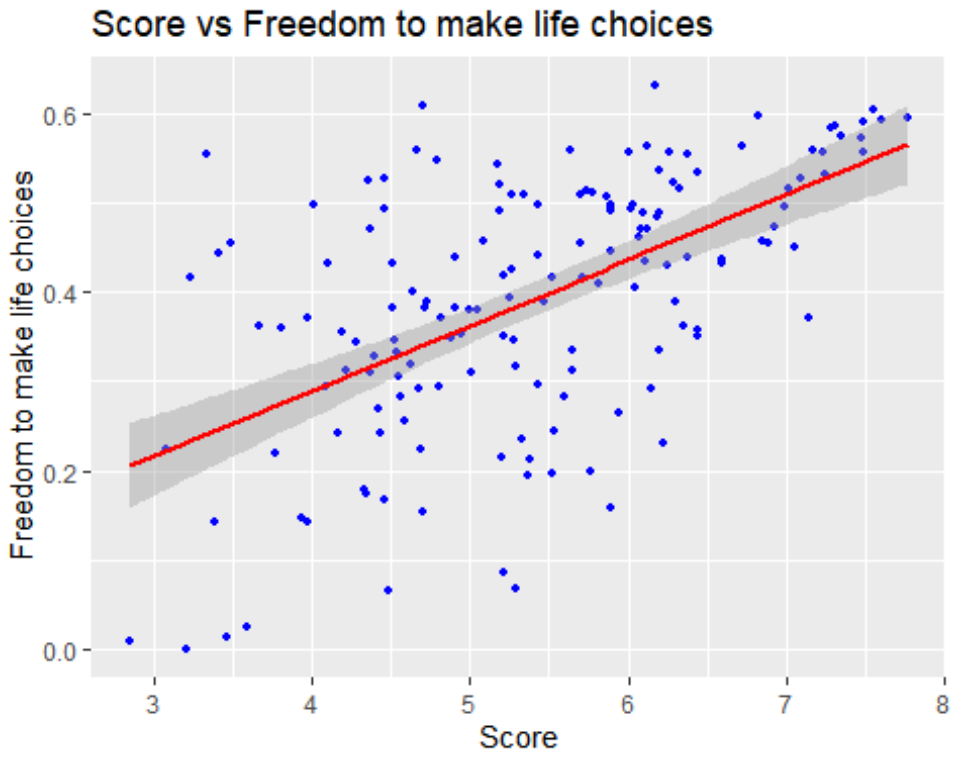
Healthy life expectancy:
A plot of Score vs Healthy life expectancy shown below shows a clear positive correlation i.e when Healthy life expectancy increases, Happiness Score also increases.

```
#################################################################
# Plot of Score vs Healthy life expectancy
#################################################################
  ggplot(data = data_2019, aes(x = Score, y = Healthy.life.expectancy)) +
  geom_point(color='blue', size = 1) +
  labs(y = " Healthy life expectancy")+
  labs(title = paste("Score vs Healthy life expectancy"))+
  geom_smooth(method = "lm", color = "red", se = TRUE)
```



Freedom to make life choices:
A plot of Score vs Freedom to make life choices shown below shows a clear positive correlation i.e when Freedom to make life choices increases, Happiness Score also increases.

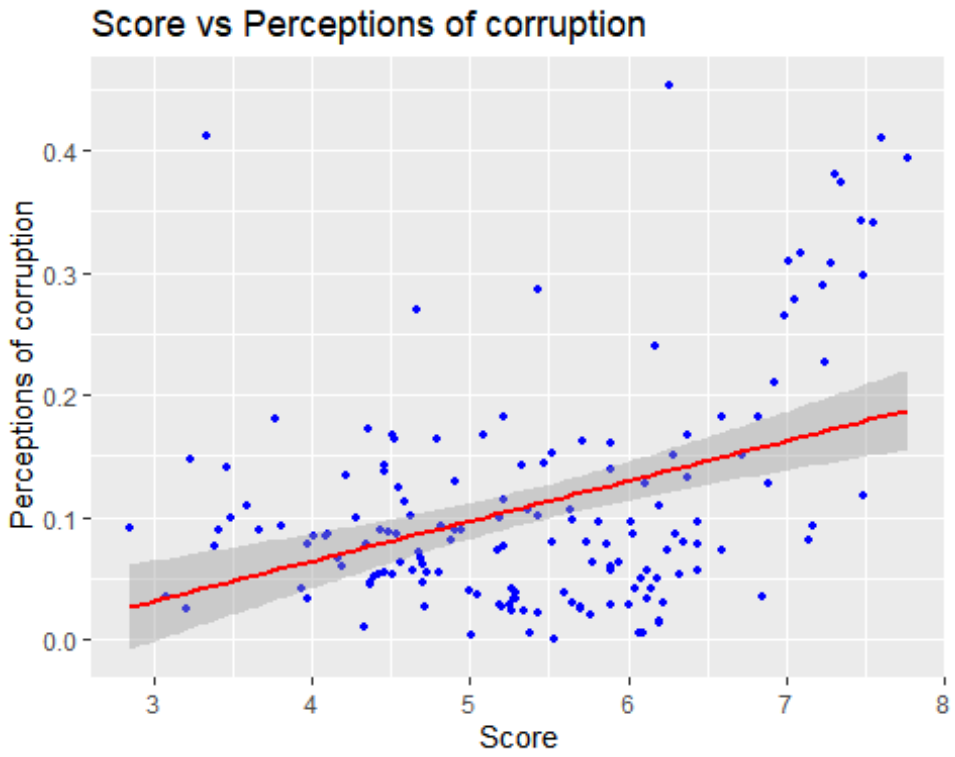
```
#####################################################################
# Plot of Score vs Freedom to make life choices
#####################################################################
  ggplot(data = data_2019, aes(x = Score, y = Freedom.to.make.life.choices))
+
  geom_point(color='blue', size = 1) +
  labs(y = " Freedom to make life choices")+
  labs(title = paste("Score vs Freedom to make life choices"))+
  geom_smooth(method = "lm", color = "red", se = TRUE)
```



Perceptions of corruption:

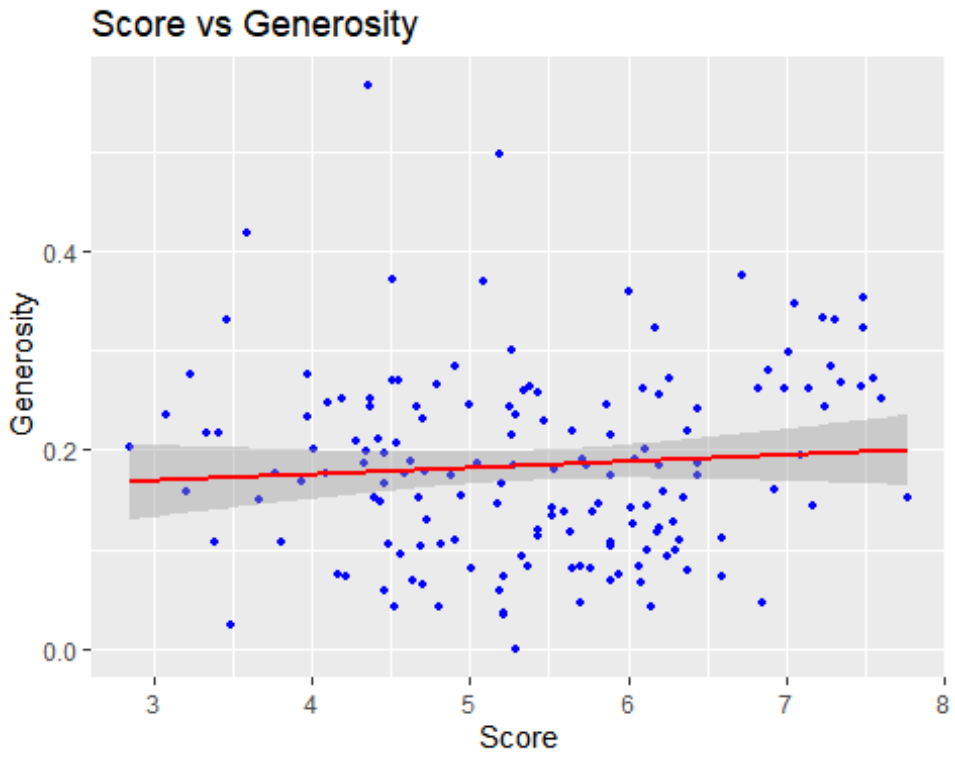
A plot of Score vs Perceptions of corruption shown below shows clear positive correlation i.e when Perceptions of corruption increases, Happiness Score also increases.

```
##################################################################
# Plot of Perceptions of corruption
##################################################################
  ggplot(data = data_2019, aes(x = Score, y = Perceptions.of.corruption)) +
  geom_point(color='blue', size = 1) +
  labs(y = " Perceptions of corruption")+
  labs(title = paste("Score vs Perceptions of corruption"))+
  geom_smooth(method = "lm", color = "red", se = TRUE)
```



Generosity:
Finally, a plot of Score vs Generosity shown below shows a lack of correlation with relatively high confidence bands.

```
#################################################################
# Plot of Score vs Generosity
#################################################################
  ggplot(data = data_2019, aes(x = Score, y = Generosity)) +
  geom_point(color='blue', size = 1) +
  labs(y = " Generosity")+
  labs(title = paste("Score vs Generosity"))+
  geom_smooth(method = "lm", color = "red", se = TRUE)
```



The above analysis infers that removing the Generosity parameter when building the models may improve the accuracy.

Now that we have the initial analysis done on the 2019 data let us focus on analysis of a few different models which can then be used to arrive at the most efficient model for this exercise to determine the best way to predict a happiness score. The deciding parameter will be the RMSE or the Root Mean Square Error which gives the absolute fit of the model to the data. It provides the difference which is the standard deviation between observed data to predicted values. The lower the value of the RMSE, the better the model will be, which is what we are trying to achieve.

Models Analysis & Evaluation

2019 Baseline model with all parameters

A classical baseline model would be to take into account the sum of all parameters and calculate the RMSE. These parameters influence the extent of contribution in evaluating the happiness in each country.

```
#################################################################
# 2019 Baseline model with all parameters and RMSE
#################################################################
Base_model <- data_2019 %>% mutate(rmse_score = GDP.per.capita +
                                   Social.support +
                                   Healthy.life.expectancy +
                                   Freedom.to.make.life.choices +
                                   Generosity +
                                   Perceptions.of.corruption,
                                 RMSE = RMSE(Score, rmse_score))

# Highest Ranked countries in the Baseline model
Base_model %>%
  filter(Overall.rank <= 10) %>%
  select(Overall.rank, Country.or.region, Score, rmse_score, RMSE)
```

```
##    Overall.rank Country.or.region Score rmse_score     RMSE
## 1             1           Finland 7.769      5.055 1.952387
## 2             2           Denmark 7.600      5.206 1.952387
## 3             3            Norway 7.554      5.313 1.952387
## 4             4           Iceland 7.494      5.093 1.952387
## 5             5       Netherlands 7.488      5.094 1.952387
## 6             6       Switzerland 7.480      5.208 1.952387
## 7             7            Sweden 7.343      5.097 1.952387
## 8             8       New Zealand 7.307      5.181 1.952387
## 9             9            Canada 7.278      5.086 1.952387
## 10           10           Austria 7.246      4.869 1.952387
```

The RMSE thus obtained is seemingly similar for all the nations i.e. **1.9523** which is too high a score to be considered accurate.

Now, earlier in this report we have mentioned about 'Dystopia', a hypothetical nation ranking lower than the lowest ranking country on the report where all inhabitants are generally unhappy and the average life is rated at 1.85 on the 0 to 10 scale.

2019 model with all parameters plus Dystopia value of 1.85

Since the Dystopia value have to be taken into consideration, the value is accordingly added to predicted scores of the 2019 baseline model since each factor has to be ranked against the comparative dystopia score.

```
################################################################ 2019
Baseline model with all parameters plus dystopia score and RMSE
#################################################################

Base_model_dys <- data_2019 %>% mutate(dys_rmse_score = GDP.per.capita +
                                Social.support +
                                Healthy.life.expectancy +
                                Freedom.to.make.life.choices +
                                Generosity +
                                Perceptions.of.corruption +
                  1.85,
                              RMSE_dys = RMSE(Score, dys_rmse_score))

# Highest Ranked countries in the Baseline model with dystopia included
Base_model_dys %>%
  filter(Overall.rank <= 10) %>%
  select(Overall.rank, Country.or.region, Score, dys_rmse_score, RMSE_dys)



##    Overall.rank Country.or.region Score dys_rmse_score  RMSE_dys
## 1             1           Finland 7.769          6.905 0.5280065
## 2             2           Denmark 7.600          7.056 0.5280065
## 3             3            Norway 7.554          7.163 0.5280065
## 4             4           Iceland 7.494          6.943 0.5280065
## 5             5       Netherlands 7.488          6.944 0.5280065
## 6             6       Switzerland 7.480          7.058 0.5280065
## 7             7            Sweden 7.343          6.947 0.5280065
## 8             8       New Zealand 7.307          7.031 0.5280065
## 9             9            Canada 7.278          6.936 0.5280065
## 10           10           Austria 7.246          6.719 0.5280065
```

Note that the RMSE has dropped to **0.5280** which is a much improved value.

2018-2019 model with dystopia value included

Now, let's analyse a combination baseline model of 2018 and 2019 data and evaluate the RMSE. A dataframe is first setup to hold the combined values of 2018 & 2019 data after converting the Perceptions of corruption column to numeric.

```
################################################
# 2018-2019 model with dystopia value included
################################################

# Convert Perceptions.of.corruption column to numeric
data_2018$Perceptions.of.corruption <-
as.numeric(as.character(data_2018$Perceptions.of.corruption))
data_2018[is.na(data_2018)] <- 0
# Bind 2018-2019 data to obtain a combination
data_1819 <- rbind (data_2019,data_2018)
#Calculate predicted score and RMSE
model_1819 <- data_1819 %>% mutate(score_1819 = GDP.per.capita +
Social.support +
Healthy.life.expectancy +
Freedom.to.make.life.choices +
Generosity +
Perceptions.of.corruption +
1.85,
RMSE_score = RMSE(Score, score_1819))
# RMSE Score for the 18-19 model
RMSE_1819 <- RMSE(model_1819$Score, model_1819$score_1819)
```

The RMSE score of the combined 2018-19 model thus obatined is **0.5257**.

Generalized linear model for 2019

What is a General Linear Model?
The General Linear Model (GLM) is used to compare how several variables affect different continuous variables and in its simplest form is shown as Data = Model + Error. It is called General since the procedure can accommodate a wide variety of variables, including a non-numerical one.

The formula for the general linear model is:
Y=β0+β1X

Where:

- Y is the dependent variable
- β0 is the intercept which is always a constant
- β1 is the coefficient (weight or slope)
- X is a variable.

To get a good fit before using machine learning algorithms, the available data must be split or partitioned into a training and test data set before the linear regression model is applied. Accordingly, a variety of split ratios can be used i.e. 80-20,75-25,70-30 or 60-40. While 80% data for training and 20% data for test might be too less a split or 60-40 might prove

to be on the higher side, we will consider a split of 70-30 (2/3rd Training and 1/3rd Test) for this evaluation.

```
###################################
# Generalized linear model for 2019
###################################
# Set seed
set.seed(1, sample.kind = "Rounding")
# Create a data partition on a 70-30 ratio split
index_2019 <- createDataPartition(data_2019$Score, times=1, p=0.70,
list=FALSE)
train_2019 <- data_2019[index_2019,]
test_2019 <- data_2019[-index_2019,]
# # Fit the Generalized linear model on all parameters
glmfit_2019 <- glm(Score ~ GDP.per.capita +
Social.support +
Healthy.life.expectancy +
Freedom.to.make.life.choices +
Generosity +
Perceptions.of.corruption,
data = train_2019)
# # Get the results into a dataframe and add the predicted scores
results_2019 <- test_2019 %>%
mutate(pred_score_2019 = predict.glm(glmfit_2019, newdata=test_2019))
# # Print the RMSE value
RMSE(results_2019$Score, results_2019$pred_score_2019)

## [1] 0.5718812
```

The RMSE thus obtained is **0.5718812**

The coefficients obtained by using the fitted model using the code shown is as below:

```
#################################################################
# Print coefficients of the model
glmfit_2019$coefficients
```

- Intercept - 1.8483663
- Social.support (SS) - 1.0155070
- GDP.per.capita (GD) - 0.8230753
- Healthy.life.expectancy (HL) - 1.0725756
- Freedom.to.make.life.choices (FC) - 1.5864416
- Generosity (GS) - 0.9939111

- Perceptions.of.corruption (CO) - 0.6122442

The formula for the general linear model can be thus translated using the coefficients as shown below:

y = 1.848 + 0.823βGD + 1.015βSS + 1.072βHL + 1.586βFC + 0.993βGS + 0.612βCO

Generalized linear model for 2019 without Generosity parameter

Let's evaluate the RMSE of the fitted model above but without the generosity parameter.

```
#####################################################################
#Generalized linear model for 2019 without Generosity parameter
#####################################################################
# Set seed
set.seed(1, sample.kind = "Rounding")
# Create a data partition on a 70-30 ratio split
index_2019 <- createDataPartition(data_2019$Score, times=1, p=0.70,
list=FALSE)
train_2019 <- data_2019[index_2019,]
test_2019 <- data_2019[-index_2019,]
# # Fit the Generalized linear model on all but without the generosity
parameter
glmfit_nogen <- glm(Score ~ GDP.per.capita +
Social.support +
Healthy.life.expectancy +
Freedom.to.make.life.choices +
Perceptions.of.corruption,
data = train_2019)
# # Get the results into a dataframe and add the predicted scores
results_nogen <- test_2019 %>%
mutate(pred_score_nogen = predict.glm(glmfit_nogen, newdata=test_2019))
# # Print the RMSE value
RMSE(results_nogen$Score, results_nogen$pred_score_nogen)

## [1] 0.5588842
```

The RMSE obtained is **0.5588842**

The coefficients obtained by using the fitted model without generosity using the code shown is as below:

```
#################################################################
# Print coefficients of the model
glmfit_nogen$coefficients
#################################################################
```

- Intercept - 1.9755643
- Social.support (SS) - 0.9743670
- GDP.per.capita (GD) - 0.7931440
- Healthy.life.expectancy (HL) - 1.0938625
- Freedom.to.make.life.choices (FC) - 1.8273152
- Perceptions.of.corruption (CO) - 0.8300638

The formula for the general linear model can be thus translated using the coefficients as shown below:

y = 1.975 + 0.793βGD + 1.974βSS + 1.093βHL + 1.827βFC + 0.830βCO

Results

The overall summary of results for the different models analyzed and their respective RMSE values are as follows:

1. 2019 Baseline model with all parameters - **1.9523**
2. 2019 model with all parameters plus Dystopia value of 1.85 - **0.5280**
3. 2018-2019 model with dystopia value included - **0.5257**
4. Generalized linear model for 2019 - **0.5718**
5. Generalized linear model for 2019 without Generosity parameter - **0.5588**

As can be seen the 2018-2019 model with dystopia value included produces the best RMSE value of **0.5257**.

Conclusion

The report explores the co-relation of various parameters and the influence of each on the Overall Happiness score. It can be clearly seen that as parameters such as GDP, Social Support, Life expectancy increased, the overall rank of happiness also increased. As an example, Burundi, 2018's least happy country jumped almost 10 places in 2019 due to improvements in these parameters. Thus, these parameters provide a substantial understanding of happiness at a macro level.
Five model approaches were evaluated to find the lowest RMSE score and hence arrive at the most optimal model to estimate happiness scores. With the 2018-2019 combination model, the algorithm was able to provide a RMSE score of 0.5257 which is lower than the results produced by the Generalized linear model.