Università
Ca'Foscari
Venezia

World Happiness Evaluation – Project Report

Kanan Mammadli

Introduction

The Analysis focuses on predicting Happiness Score (0-10) among countries and try to identify which factors are more important being happy in the world. It is obvious that money, family and democracy rating have huge impact our daily life, therefore analysis try to focus on these variables, in order to get depth result and clear answers.

About Dataset: The Dataset contain 12 variables and 160 observations in each variable where Happiness Score is our response variable and it is numerical. Rest of the are counted as explanatory variables and only 3 of them are factor variables. GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption describe the extent to which these factors contribute in evaluating the happiness in each country. There is also "Regions" variable which helps to decide Happiness globally.

Main decisions and Methodology: As mentioned above, Happiness Score is used as response variable and there are 2 model used to define "Score" by explanatory variables: Multiple Linear Regression model and Polynomial Regression model. Overall, both model fits very well and their accuracies are pretty enough high for analysis. In first Multiple Regression model, almost all variables took into account, whereas in 2nd one only explanatory variables were included to model. In Polynomial Regression model, response variable tried to explained by Life Expectancy. Exact decisions about models will be mentioned in modelling part. The summary() function, BIC, AIC, MSE methods were used during evaluation process. Additionally, some graphics plotted in order to show our result visually.

Exploratory Data Analysis and Data prepartion

In this section summary statistics give more depth information about dataset and some basic plots that show how do different variables effect “Happiness Score” and even each other. First of all, after checking missing values in dataset (Figure 1.1), surprisingly there were no NA. Summary data prove variables is normally distributed and no potential outliers exists.

```
> summary(happiness)
   Country             Region          HappinessRank    HappinessScore
 Length:158         Length:158         Min.   :  1.00   Min.   :2.839
 Class :character   Class :character   1st Qu.: 40.25   1st Qu.:4.526
 Mode  :character   Mode  :character   Median : 79.50   Median :5.232
                                       Mean   : 79.49   Mean   :5.376
                                       3rd Qu.:118.75   3rd Qu.:6.244
                                       Max.   :158.00   Max.   :7.587
 StandardError      GDPperCapita        Family       LifeExpectancy
 Min.   :0.01848   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.03727   1st Qu.:0.5458   1st Qu.:0.8568   1st Qu.:0.4392
 Median :0.04394   Median :0.9102   Median :1.0295   Median :0.6967
 Mean   :0.04788   Mean   :0.8461   Mean   :0.9910   Mean   :0.6303
 3rd Qu.:0.05230   3rd Qu.:1.1584   3rd Qu.:1.2144   3rd Qu.:0.8110
 Max.   :0.13693   Max.   :1.6904   Max.   :1.4022   Max.   :1.0252
    Freedom       GovernmentCorruption   Generosity      DystopiaResidual
 Min.   :0.0000   Min.   :0.00000      Min.   :0.0000   Min.   :0.3286
 1st Qu.:0.3283   1st Qu.:0.06168      1st Qu.:0.1506   1st Qu.:1.7594
 Median :0.4355   Median :0.10722      Median :0.2161   Median :2.0954
 Mean   :0.4286   Mean   :0.14342      Mean   :0.2373   Mean   :2.0990
 3rd Qu.:0.5491   3rd Qu.:0.18025      3rd Qu.:0.3099   3rd Qu.:2.4624
 Max.   :0.6697   Max.   :0.55191      Max.   :0.7959   Max.   :3.6021
```

Figure 1.1 – Summary statistics of data

This result help to get basic information about variables and their distribution. As mentioned in summary mean, median and specially min, max seems normal. However, in order to be sure, and get more depth info about response variable next plot explain distribution of “Happiness Score” and by graph (figure1.2) it is easily understandable that distribution is pretty normal and it does not need log transformation. When distribution right or left skewed, log transformation applied to make it normal distribution. Here mean for “Happiness Score” is 5.37, whereas median is 5.232. These are answers let to continue analysis process without any transformation.

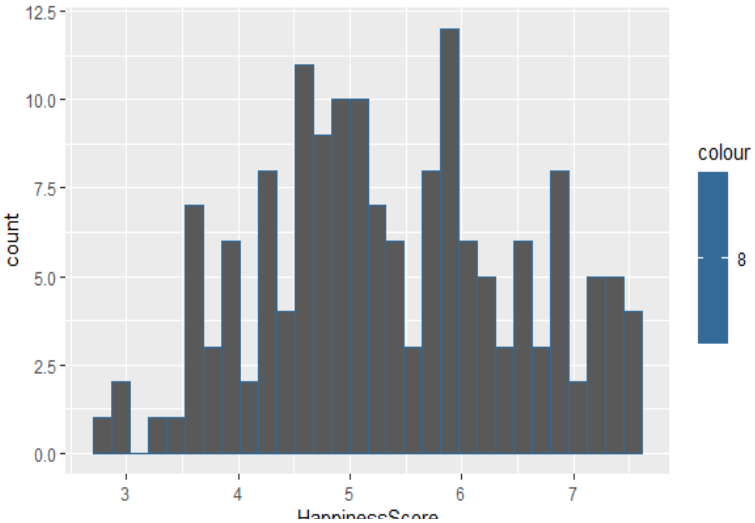

Figure 1.2 – Variable “Happiness” normally distributed

Even if analysis focuses on to make prediction about “Happiness Score”, some other investigation was made in order to find relationship between explanatory variables. For example, relationship between “GovernmentCorruption” (it would be useful to mention that more Corruption Score shows how Government clear is) and “GdpPerCapita” subjectively interesting and before making plot, it seems there is real relationship between them. However basic scatter plot (figure 1.3) proves that there is not clear relationship between them, even some highly corrupted governments have huge GDP. Moreover, after some coding it was clear that, Rwanda, Qatar, Singapore, Denmark, Sweden are countries that with the highest GDP per capita and the lowest corruption rate.

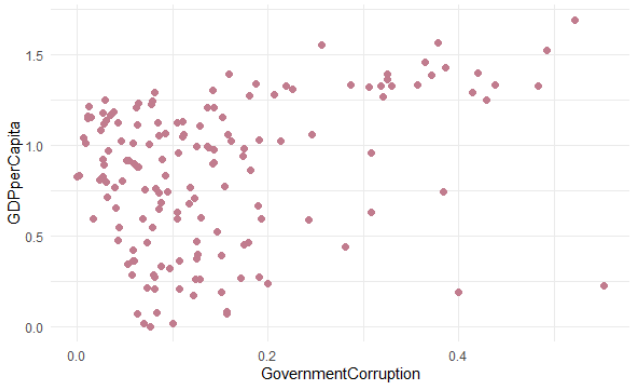

Figure 1.3 – Relationship between Corruption and GDP per capita.

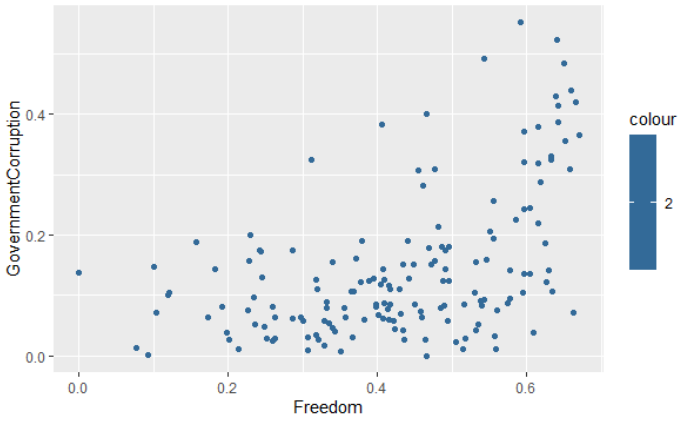

Figure 1.4 – Relationship between Freedom and Corruption rate

Vice versa to Figure 1.3, Figure 1.4 proves more Free countries have less Corruption Rate in Governments.

Next, let's focus on countries and regions. It is possible to see from graph (Figure 1.5), as expected, Western European, American and Oceanian countries have more average happiness score. These regions are highly democratized, therefore it is assumed that democracy has huge role Happiness directly and or indirectly. However, it is also acceptable that GDP per capita is not only indicator for happiness score, because even if southern and eastern Asian countries have enough high GDP per capita, their happiness indicator is not high like top 3 region.

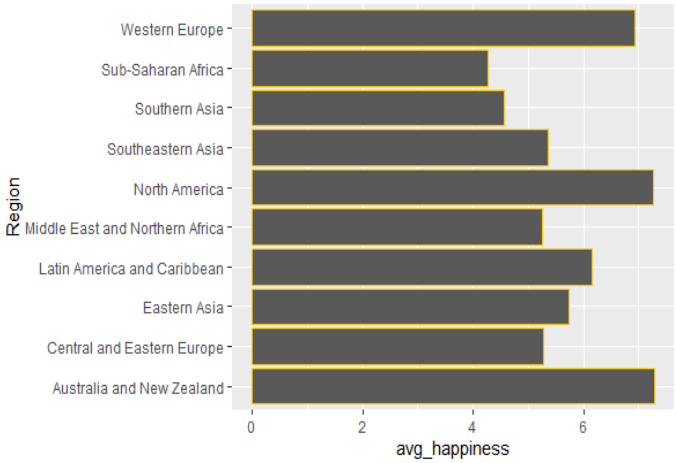

Figure 1.5 – Regions by average Happiness Score

Let's dive in more depth and see countries separately, which are among the last and first 10 happiest countries (Figure1.6) . Here bar chart indicate that Switzerland, Iceland, Denmark and Norway are the most happiest countries whereas second red bar chart indicate people in Chad, Guinea, Ivory Coast and Burkina Faso are not happy and live in worst conditions.

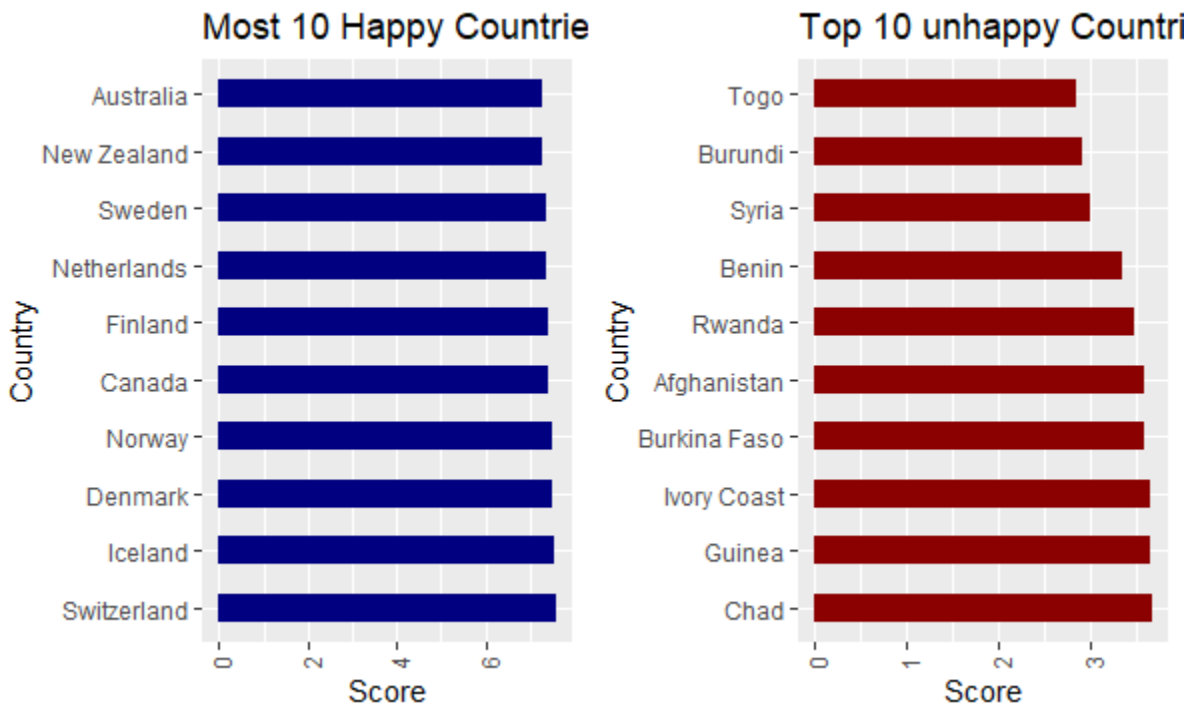

Figure 1.6 – Top 10 happy and unhappy country

Correlation and Correlogram

One of the most important parts of data preparation process is finding highly correlated variables of data. Before applying correlation between variables, factor variables have to be dropped. Only then that would be right to using cor() function that share correlation degree of variables between -1 and 1. When independent variables are highly correlated, change in one variable would cause change to another and so the model results fluctuate significantly. The model results will be unstable and vary a lot given a small change in the data or model. Correlogram shows correlation (Figure 1.7) between variables (positive correlation with green , negative correlation with red) such as, Life Expectancy and GDP per capita (82%) , GDP per capita and Family (more than 60%), however dropping these variables could be wrong decision, because in modelling part analysis proves that these variables are very important to predicting "Happiness Score". Moreover, it is interesting that, there are almost no

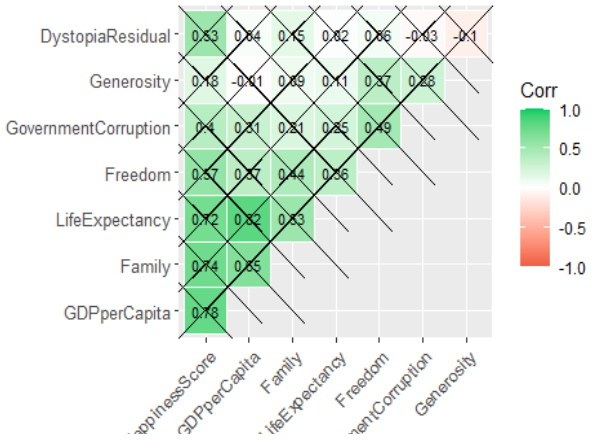

Figure 1.7 - Correlogram

negative correlation between variables and at the same time, almost all variables (except Generosity) have at least 40% correlation with “Happiness Score”.

Data Modelling

As mentioned above, we used Multiple Linear Regression and Polynomial regression models in our modelling part.

Before applying models, dataset we divide into training and test parts. Training part contain 75% of total observations whereas test is 25%. In first model of Multiple Linear Regression was created with all variables (“GDPperCapita”, “Family”, “LifeExpectancy”, “Freedom”, “GovernmentCorruption”,”Generosity”)

```
## The model ##
linear_fit1 <- lm(HappinessScore ~GDPperCapita+Family+ LifeExpectancy+
Freedom+GovernmentCorruption +Generosity, data = train_hp)
```

```
     Min       1Q   Median       3Q      Max
-1.38654 -0.32048 -0.02817  0.33017  1.43971

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.0051     0.2068   9.694  < 2e-16 ***
GDPperCapita         0.9383     0.2631   3.567 0.000535 ***
Family               1.3046     0.2473   5.274 6.64e-07 ***
LifeExpectancy       0.8332     0.3665   2.273 0.024931 *
Freedom              1.4474     0.4242   3.412 0.000901 ***
GovernmentCorruption 0.8620     0.5171   1.667 0.098358 .
Generosity           0.2269     0.4184   0.542 0.588754
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5311 on 111 degrees of freedom
Multiple R-squared:  0.8011,    Adjusted R-squared:  0.7903
F-statistic:  74.5 on 6 and 111 DF,  p-value: < 2.2e-16
```

Figure 2.1 - First Multiple Linear Regression model

Interpretation: if independent variables differed by 1%, “Happiness Score “will increase by “Estimate” units, on average

Happiness = 2.0051 + 0.9383xGDPperCapita + 1.3046xFamily + 0.8332xLifeExpectany+1.4474xFreedom+0.8620xGovernmentCorruption+ 0.2269xGenerosity + E

Summary function shows about how strong the first model is. First thing we have to take into account should be p-values. All variables’ p-values of variables – “GDPperCapita”,”Family” , and “Freedom” is strongly significant, because their values is less than 0.001 and only

"LifeExpectancy" counted as significant. Only "Government Corruption" and "Generosity"not significant. After considered overall p-value, we reject null hypothesis (explanatory variables have no impact on our response variable). Moreover, our Adjusted R-squared is 79%, which means that our response 79.03% explained by independent variables. That is pretty enough high result. Additionally, our R-squared is 80.11% which means our variance of response variable, that percentage explained by independent variables. AIC and BIC values are 194.301 and 216.4664 respectively. AIC and BIC values should be small in order to prove model performance.

In second Multiple Linear Regression we just try something new for experiment. Here we dropped unsignificant value to see what would happen. So, dropping "Generosity" and "Government Corruption" is not that good idea, because summary output says Adjusted R squared decreased about 1%.

```
linear_fit2 <- lm(HappinessScore ~GDPperCapita+Family+ LifeExpectancy+ Freedom,
data = train_hp)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.4289 -0.3285 -0.0093  0.3370  1.4465

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.0203     0.2043   9.889  < 2e-16 ***
GDPperCapita   1.0021     0.2495   4.016 0.000107 ***
Family         1.2504     0.2449   5.105 1.35e-06 ***
LifeExpectancy 0.8217     0.3628   2.265 0.025412 *
Freedom        1.8364     0.3670   5.003 2.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5344 on 113 degrees of freedom
Multiple R-squared:  0.795,     Adjusted R-squared:  0.7877
F-statistic: 109.5 on 4 and 113 DF,  p-value: < 2.2e-16
```

Figure 2.2 – Multiple Linear Regression after dropping some variables

After checking 2 models, we could say that there is no big difference over models. However, despite of high Adjusted R squared and Multiple R squared we continue with second model because of less variables as predictor and plot diagnostics of model (figure 2.3**).** Normal Q-Q plot – The points follow the straight line, meaning that the standardized residuals follow a standard normal distribution. Moreover, Residual versus fitted values – We can see a slightly (so small) parabolic pattern in our case, meaning that there is a non-linear relationship that was not explained by the model and was left out in the residuals

(that is why we going to use polynomial regression as next model). Scale-Location – The residuals seem to spread the closer it gets to 0 along the x-axis.

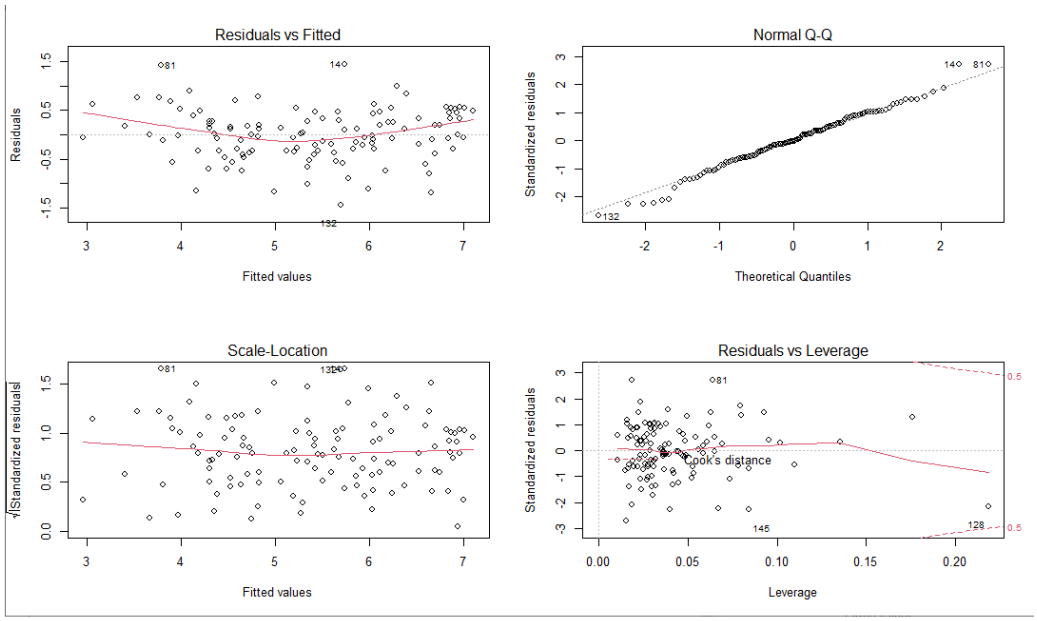

Figure 2.3 – Diagnostics of Multiple Regression Model

Polynomial Regression

We also understood from last model that GDP per capita is the most important variable that effect "Happiness Score" and linear regression significantly explain relationship among "HappinessScore" and "GPDperCapita". Therefore ,we plot "Life Expectence" for see relationship with "HappinessScore" and we going to see relationship among them is not so linear. That is why we going to try next polynomial regression "Life Expectence" as predictor and "HappinessScore" as response variable.

First of all, We create an empty numeric vector, cverror, where we will save the LOOCV MSE for each value of degree dd. We give a grid of possible values for dd in degree, from 1 to 5.

```
> cbind(degrees, cverror)
     degrees    cverror
[1,]       1 0.6592463
[2,]       2 0.5713308
[3,]       3 0.5644461
[4,]       4 0.5681287
[5,]       5 0.5832172
>
```

Figure 2.4 – Polynomial Degree

We choose 2nd degree because, there is only clear difference from 1st degree. As we continue with 2nd degree of polynomial regression, we fit our test data and it fits very well. We can now realize that why "Life Expectence" is not strongly significant in multiple linear model (it does not show linearity, instead shows slightly upward curve, as showed in Figure 2.5)

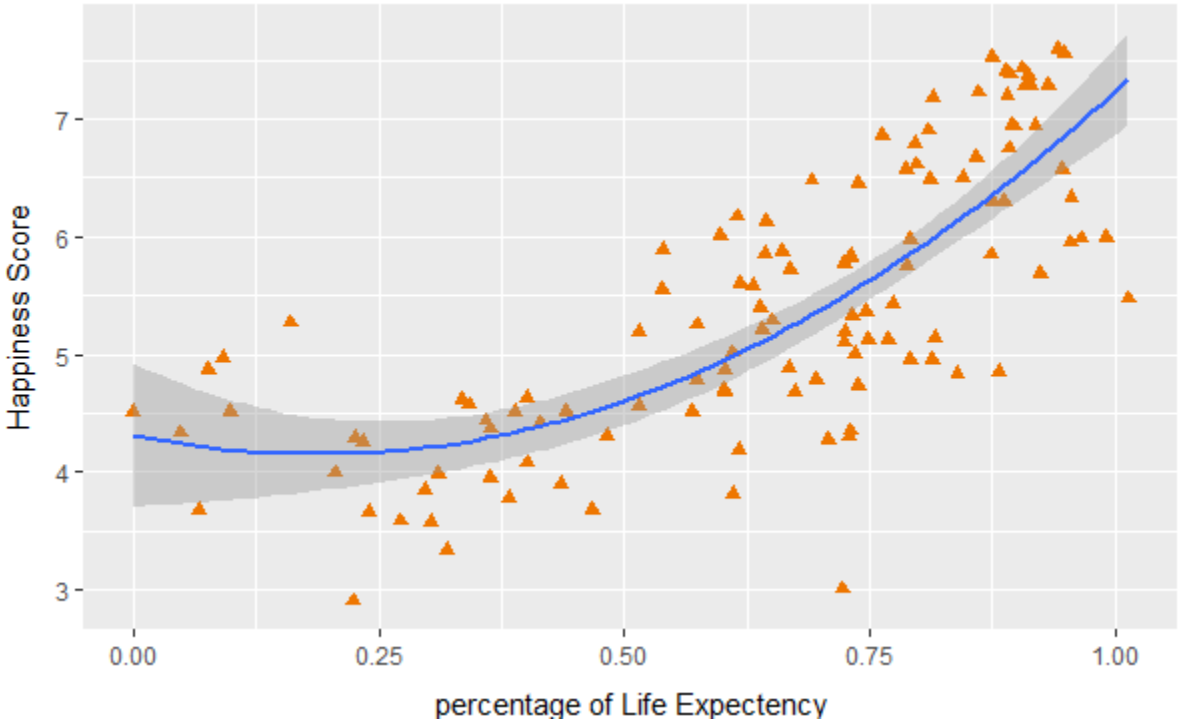

Figure 2.5 – Fitting of polynomial regression on training data

Summary statistics indicate (Figure 2.6) we have highly significant explanatory variables and p-values are strongly significant (less than 0.001). We can reject null hypothesis and accept Alternative one. However, our Adjusted R-squared percentage not that high.

```
lm(formula = HappinessScore ~ poly(LifeExpectancy, d_opt), data = train_hp)

Residuals:
     Min       1Q   Median       3Q      Max
-2.47888 -0.52734  0.08206  0.57923  1.20024

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5.3884     0.0687  78.438  < 2e-16 ***
poly(LifeExpectancy, d_opt)1  9.0845     0.7462  12.174  < 2e-16 ***
poly(LifeExpectancy, d_opt)2  3.2891     0.7462   4.408 2.36e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7462 on 115 degrees of freedom
Multiple R-squared:  0.5931,    Adjusted R-squared:  0.586
F-statistic: 83.82 on 2 and 115 DF,  p-value: < 2.2e-16
```

Figure 2.6 – summary of Polynomial model

At the end, we can check model performance with AIC, BIC, MAE and R2.

```
> performance.poly <- data.frame(
+   AIC = round(AIC(model.poly),4),
+   BIC = round(BIC(model.poly),4),
+   RMSE = RMSE(pred = pred.poly, obs = test_hp$HappinessScore),
+   R2 = R2(pred = pred.poly, obs = test_hp$HappinessScore),
+   MAE = MAE(pred = pred.poly, obs = test_hp$HappinessScore)
+ )
> performance.poly
      AIC       BIC       RMSE         R2        MAE
1 270.748 281.8307 0.7742017 0.5134305 0.6457476
>
```

Figure 2.7 – Performance of Polynomial Regression

From figure that we see above, tells us general performance of Polynomial Regression. R-squared is not enough good at the same time, RMSE value is high, which should be between 0.2-0.5, In order to be count as good. AIC and BIC values are 270.748 and 282.8307 respectively which are a bit bigger than multiple regression model's values.

Conclusion

To sum up, both models perform good (even polynomial not perfect), and their performance measurement results were pretty high.

Overall, variables such as "Generosity" and "GovernmentCorruption" have no clear impact to Happiness Score (as there was no clear relationship between GDP and Corruption rate). But there are some interesting trends were found out. For example, we can assume that, percentage of LifeExpectency starting to impact target after passing some percentage, not linearly.

Overtheless, "GDPperCapita","Family" , and "Freedom" are highly significant variables, specially "Family" is the most important explanatory variable.

When we have to choose model from Multiple Linear Regression, because of less variables, we would continue with 2nd model. Both model perform almost same performance.

Analysis, showed basic need for happiness is Family and Freedom. I subjectively support and believe that, there is nothing important in the world behind Family and Freedom.

Reference:

Data: World Happiness Report | Kaggle