The United Nations Study on Happiness

Carter Rogers, Nicholas Hunter, Shivam K C

12/5/2019

Our Problem and Hypothesis:

Some countries are happier than others. We want to find what factors best predict happiness. Therefore, we hypothesize the following: (i) We expect factors such as life expectancy and GDP to be positively correlated with happiness. In our dataset, Life Ladder is the happiness score of a nation gotten by averaging the national response to survey, Cantril Ladder, to a single number. (ii) We expect factors such as corruption to be negatively correlated with happiness. (iii) We expect Europe to be the happiest continent on average.

Exploratory Data Analysis

Distribution of Happiness Score

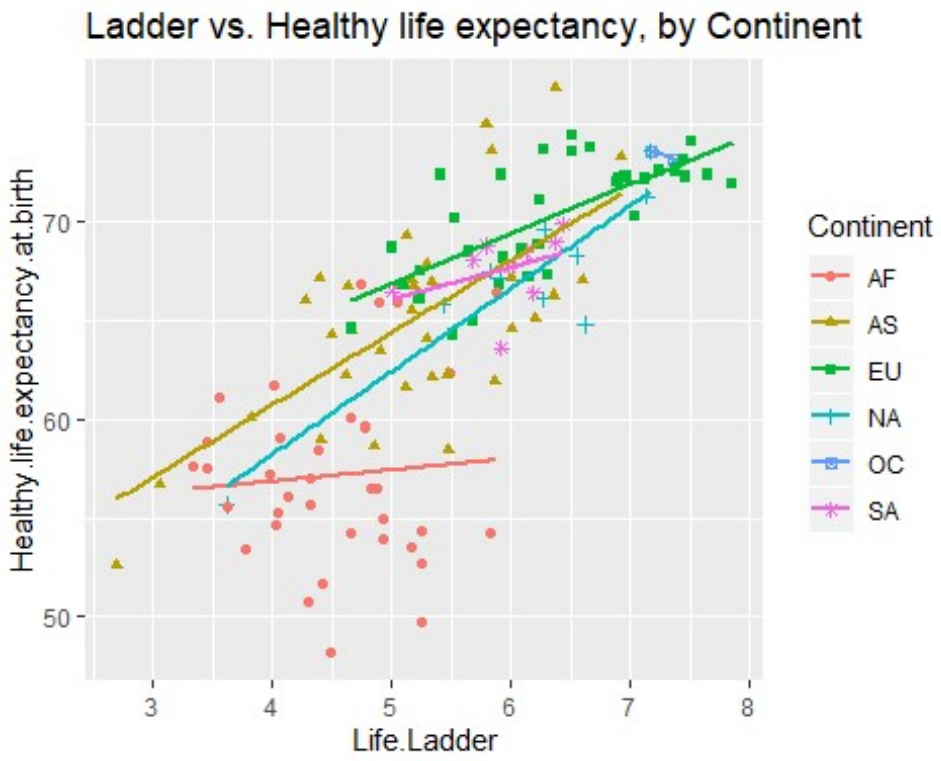
We plotted this graph to observe the distribution of happiness score. Here, the distribution of happiness score is symmetric. This means that there are many countries with average happiness scores.

Statistics of Happiness Score

Min.	Median	Mean	Max.
2.694	5.409	5.469	7.858

We are using this table to check the statistics (spread and mean) of Happiness Score which is our dependent variable. We observe that the mean Happiness Score of the World in 2018 was 5.469. The maximum score was 7.86 and the minimum score was 2.69.

Testing Hypothesis 1



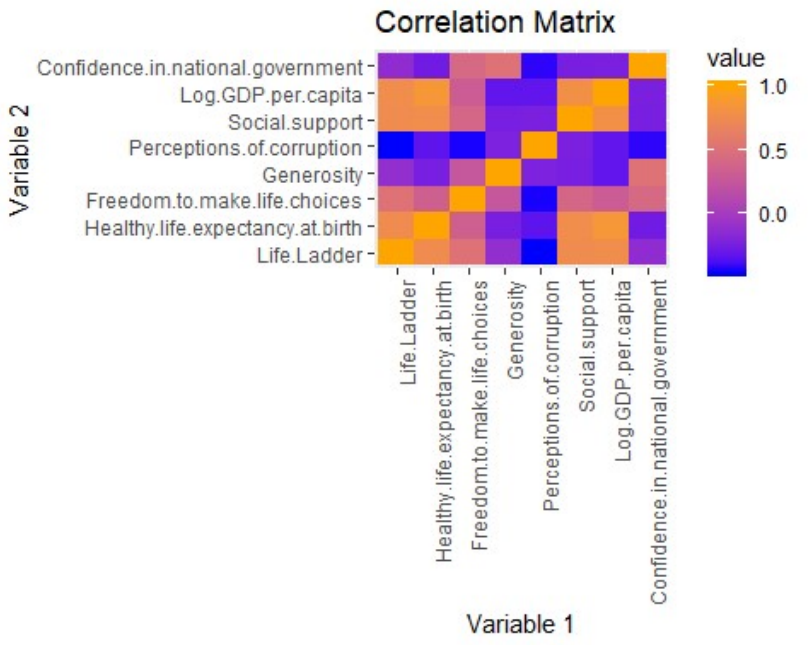
We constructed this plot in order to examine the relationship between healthy life expectancy for each country and that respective country's happiness score. We see that European and Asian countries are strongly and positively correlated with happiness. African countries are not as strongly correlated, as the line of best fit has a lower slope than the others. Similar result was observed when GDP was substituted in for Life expectancy (result can be observed in the Rscript).

Testing Hypothesis 3

	Continent	count	ladder_avg
1	AF	37	4.52
2	AS	34	5.17
3	EU	35	6.36
4	NA	10	6.08
5	OC	2	7.27
6	SA	9	5.94

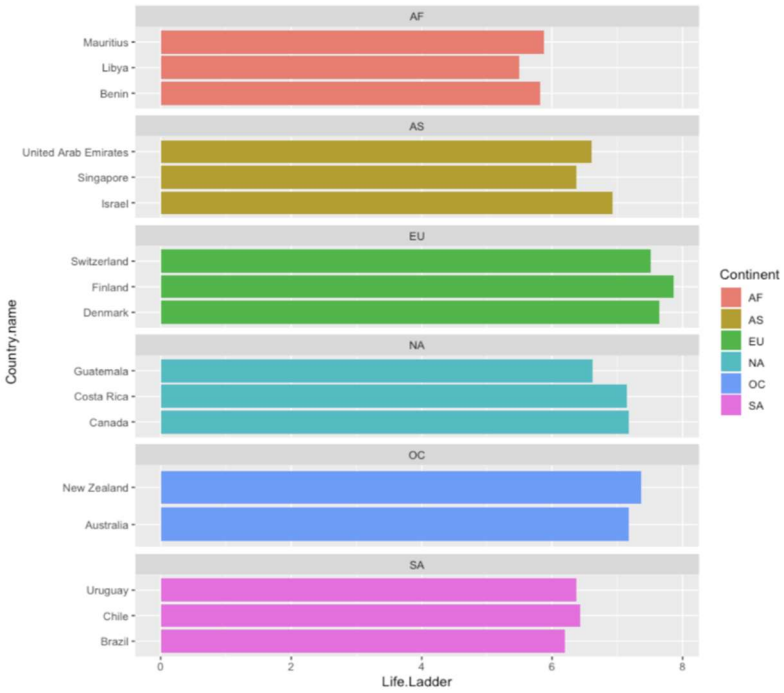
We made this table to test our hypothesis 3. We observe Oceanic to have the highest Happiness score, but this score only considers 2 countries. We do observe (excluding OC) Europe to be the happiest continent.

Testing Hypothesis 2



This plot shows the correlations between a selected set of variables which we believe could relate to our hypothesis. Here we see stronger correlations between Life.Ladder, which is our self reported happiness score, and GDP per capita, Social Support, Freedom, and Life expectancy. We see weaker correlations with Confidence in Government and Generosity. We also see a negative correlation between happiness and Perceptions of Corruption

Comparing Top 3 Happiest Countries from Each Continent



We plotted this graph to compare top 3 happiest countries from each continent. We observe that top 3 happiest countries from Europe are happier than or at the (very least) same level as top 3 happiest countries from other continents.

Modeling

Hypothesis: We believe that GDP (Log.GDP.per.capita), Life Expectancy (Healthy.life.expectancy.at.birth), Corruption (Perceptions.of.corruption), and Continent predict a country's happiness (Life.Ladder).

Predictor/s: GDP (Log.GDP.per.capita), Life Expectancy (Healthy.life.expectancy.at.birth), Corruption (Perceptions.of.corruption), and Continent

Type of variable: All the variables except Continent are numeric variables. Continent is a categorical variable.

Response: Happiness (Life.Ladder).

Type of variable: Numeric Variable.

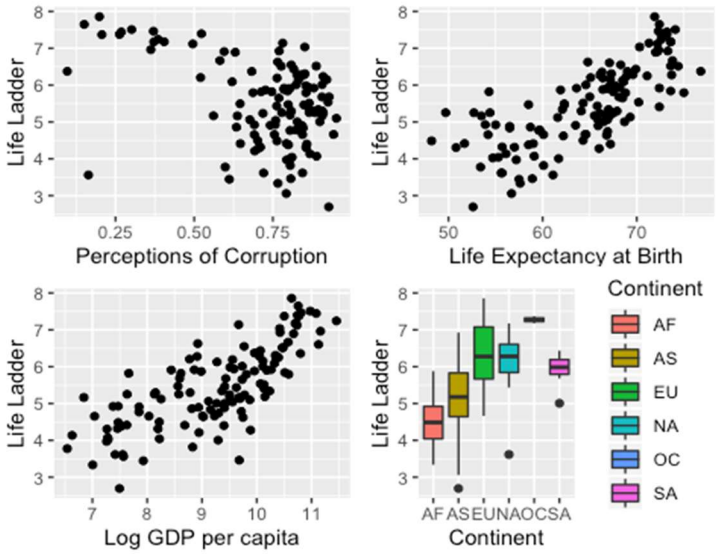
Type of model used: Multiple Linear Regression Model.

```
model <- lm(Life.Ladder ~ Perceptions.of.corruption + Healthy.life.expectancy.at.birth + Continent+Log.GDP.per.capita, data = whr2018_w_continents)
summary(model)

##
## Call:
## lm(formula = Life.Ladder ~ Perceptions.of.corruption + Healthy.life.expectancy.at.birth +
##     Continent + Log.GDP.per.capita, data = whr2018_w_continents)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.74114 -0.33736  0.02974  0.31889  1.47870
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.64894    0.95493   0.680 0.498274
## Perceptions.of.corruption      -1.33380    0.35286  -3.780 0.000261 ***
## Healthy.life.expectancy.at.birth  0.03008    0.02010   1.497 0.137422
## ContinentAS                    -0.10185    0.19942  -0.511 0.610626
## ContinentEU                     0.39686    0.24920   1.593 0.114263
## ContinentNA                     0.73259    0.26516   2.763 0.006767 **
## ContinentOC                     0.59522    0.51596   1.154 0.251274
## ContinentSA                     0.55034    0.28201   1.952 0.053658 .
## Log.GDP.per.capita              0.39855    0.10254   3.887 0.000178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6114 on 105 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.7189, Adjusted R-squared:  0.6975
## F-statistic: 33.57 on 8 and 105 DF,  p-value: < 2.2e-16
```

Answers to questions:

1. Is there a relationship between the variables and the response?



We can see from the plots that there is probably a relationship between GDP (Log.GDP.per.capita), Life Expectancy (Healthy.life.expectancy.at.birth), Corruption (Perceptions.of.corruption), and Continent vs happiness (Life.Ladder).

2. How strong is the relationship between the predictors and the response?

Since the R-squared of the fitted model is 0.7189, which is close to 1, we say the relationship is relatively strong.

3. Do the predictors contribute to the response?

Since the p-values of Perceptions.of.corruption, ContinentNA, and Log.GDP.per.capita are < 0.05, we say that these predictors contribute to the response. The other predictors do not contribute to the response.

4. What is the effect of each predictor on the response?

When Perceptions.of.corruption increases by 1 unit, Life.Ladder decreases by 1.3338 units. When Healthy.life.expectancy.at.birth increases by 1 unit, Life.Ladder increases by 0.03008 unit. When Log.GDP.per.capita increases by 1 unit, Life.Ladder increases by 0.39855 unit. The slope for ContinentEU is 0.39686. This means that ContinentEU is 0.39686 more happier than ContinentAF on average. The slope for ContinentAS is -0.10185. This means that ContinentAS is 0.10185 less happier than ContinentAF on average. The slope for ContinentNA is 0.73259. This means that ContinentNA is 0.73259 more happier than ContinentAF on average. The slope for ContinentOC is 0.59522. This means that ContinentOC is 0.59522 more happier than ContinentAF on average. The slope for ContinentSA is 0.55034. This means that ContinentSA is 0.55034 more happier than ContinentAF on average.

5. How accurately can we predict the response using the predictor?

Residual standard error: 0.6114 on 105 degrees of freedom

6. Use a model selection procedure to select the best model

#Cross validation:

```
##   nvmax     RMSE  Rsquared      MAE
## 1     2 0.7177028 0.5825633 0.5560871
## 2     3 0.6644969 0.6399593 0.5234397
## 3     4 0.6644969 0.6399593 0.5234397
```

```
summary(Best_LOOCV$finalModel)
```

```
## Subset selection object
## 3 Variables  (and intercept)
##                              Forced in Forced out
## Perceptions.of.corruption          FALSE     FALSE
## Healthy.life.expectancy.at.birth     FALSE     FALSE
## Log.GDP.per.capita                 FALSE     FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: backward
##          Perceptions.of.corruption Healthy.life.expectancy.at.birth
## 1  ( 1 ) " "                       " "
## 2  ( 1 ) "*"                       " "
## 3  ( 1 ) "*"                       "*"
##          Log.GDP.per.capita
## 1  ( 1 ) "*"
## 2  ( 1 ) "*"
## 3  ( 1 ) "*"
```

Here, RMSE of the model is lowest when the predictors are Perceptions.of.corruption and Log.GDP.per.capita. Thus, the following is the best model (excluding Continent):

```
##
## Call:
## lm(formula = Life.Ladder ~ Perceptions.of.corruption + Log.GDP.per.capita,
##     data = data_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22182 -0.41915  0.08447  0.43161  1.42850
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.54721    0.64491   0.849 0.397968
## Perceptions.of.corruption -1.41935    0.36262  -3.914 0.000156 ***
## Log.GDP.per.capita          0.64888    0.05531  11.732  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6753 on 112 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.6384, Adjusted R-squared:  0.632
## F-statistic: 98.88 on 2 and 112 DF,  p-value: < 2.2e-16
```

When the Continent is included, following model is obtained.

```
##
## Call:
## lm(formula = Life.Ladder ~ Perceptions.of.corruption + Log.GDP.per.capita +
##     Continent, data = data_2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.80896 -0.33101  0.07587  0.35023  1.45374
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.65257    0.72456   2.281 0.024542 *
## Perceptions.of.corruption -1.39984    0.35293  -3.966 0.000132 ***
## Log.GDP.per.capita        0.49042    0.07599   6.454 3.27e-09 ***
## ContinentAS               0.04505    0.18611   0.242 0.809218
## ContinentEU               0.60363    0.22648   2.665 0.008885 **
## ContinentNA               0.94662    0.23924   3.957 0.000137 ***
## ContinentOC               0.84494    0.50381   1.677 0.096441 .
## ContinentSA               0.75906    0.26042   2.915 0.004336 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6179 on 107 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7108, Adjusted R-squared:  0.6918
## F-statistic: 37.56 on 7 and 107 DF,  p-value: < 2.2e-16
```

Here, Perceptions.of.corruption, Log.GDP.per.capita, and Continent predicts Life.Ladder better than Perceptions.of.corruption and Log.GDP.per.capita as determined by RSE.

7. Is the selected model the same as the model with all the variables?

The selected model is not the same as the model with all the variables.

8. How is the selected model different from the original model?

The original model used Perceptions.of.corruption, Log.GDP.per.capita, Continent, and Healthy.life.expectancy.at.birth to predict Life.Ladder.

The selected model used Perceptions.of.corruption, Log.GDP.per.capita and Continent to predict Life.Ladder. Thus, according to our analysis, perceptions of corruption, GDP per capita, and continent a country belongs to, help us predict the country's happiness.