AN ANALYSIS OF FACTORS THAT AFFECT HAPPINESS OF COUNTRIES OF THE WORLD

NAME OF THE COURSE : INST 627
PROJECT TEAM N0.2

SEMESTER : FALL 2022

Team-2, members and roles:

Name	Role	Contact Info
Sadaf Davre	ContactPerson/Project Manager	sdavre@umd.edu
Eitan Dunn	Subject Matter Expert	edunn123@umd.edu
Yogesh Boricha	Technical Expert	yogi@umd.edu
Ushasri Bhogaraju	Technical Expert	bushasri@umd.edu

INTRODUCTION :

Research question: ‘Does the perception of corruption affect happiness across the world?’

Why happiness? Because happiness is increasingly considered an important and useful way to guide public policy and measure its effectiveness. Because being happy is not just about feeling good. Research shows that it also makes us healthier, more productive – and nicer. The overall happiness level of a country shows the general wellbeing and standard of life for the citizens of that country.

We wanted to analyze a dataset that helped us understand how happy the world is and what factors are influencing it - We found a suitable dataset on ‘Kaggle’ that offered interesting variables on the same topic. The purpose of the dataset was to ascertain how happy the world is and which factors are influencing the ‘happiness score’

The research question talks about two variables- cpi and happiness_score. We thought that finding correlations between other variables and possibly finding which variable mostly affects the happiness rank would make it a good challenge.

The World Happiness Report (WHR) is based on the science of wellbeing, which uses quantitative methods to understand how different life experiences influence people’s happiness and quality of life. The purpose of this dataset is to calculate the world happiness score. World happiness is all about prosperity and economic growth. The real obstacle behind world economic growth and stability are means of corruption. Therefore asking the question ‘Does the perception of corruption affect happiness across the world?’ is paramount for understanding which parts of the world need to focus more on the wellbeing and standard of life of their citizens.

METHODS AND DATA

Data set : WorldHappiness_Corruption_2015_2020.csv @
https://www.kaggle.com/datasets/eliasturk/world-happiness-based-on-cpi-20152020

Author of the dataset: Mr.Elias Turk

About the dataset:

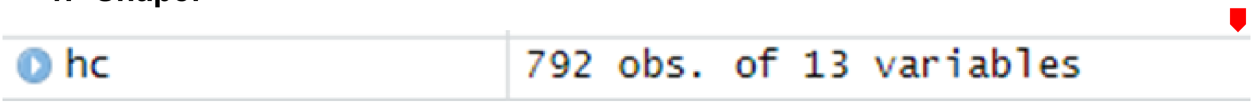
- The purpose of the dataset was to ascertain how happy the world is and which factors are influencing the ‘happiness score’.
- Respondents were asked to imagine the best life they could, and then rate their existing life using it as a baseline.
- The ‘World Happiness Score’ in this dataset was the mean of responses to the primary life evaluation question from the Gallup World Poll (GWP). GWP uses the **Cantril Ladder.**
- **The Cantril Self-Anchoring Scale,** developed by pioneering social researcher **Dr. Hadley Cantril**, consists of the following:
    - Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top.
    - The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you.
    - On which step of the ladder would you say you personally feel you stand at this time? (ladder-present)
    - On which step do you think you will stand about five years from now? (ladder-future)

- This data was compiled for 6 years from 2015 to 2020 and **corruption perception data was added to it from CPI index maintained by ‘transparency’. (transparency.org is an organization committed to removal of corruption in the public sector and promoting transparency of governments. It calculates and publishes cpi index every year for countries of the world, which was absorbed into the dataset by author)**

- **In the data set, a CPI score of ‘100’ indicates a clean government without public sector corruption and a ‘0’ score represents most corrupt government**

Description of the dataset:

1. **Shape:**



```
hc                         792 obs. of 13 variables
```

There are 792 observations and 13 variables.

The data set has no duplicate or missing values (missing values were replaced with zeros by author of the dataset)

The variable names in the dataset and what they represent:

1**. Gross Domestic Product per Capita (GDP per Capita)** - Gross Domestic Product per Capita (GDP per Capita) for each country.

2. **Family** - Family Satisfaction Rank

3. **Life Expectancy** - the average number of years one can expect to live.

4. **Freedom** - Measuring people's perceptions of freedom

5. **Generosity** - A numerical value calculated based on poll participants' perceptions of generosity in their country.

6. **Government Corruption/Trust** - A measurement of the public's trust in their governments.

7. **Dystopia Residual** - A score based on a hypothetical comparison to the world's saddest country.

8. **Continent** - Region of the country.

9. **CPI score** - The Corruption Perceptions Index (CPI) is an index which ranks countries "by their perceived levels of public sector corruption, as determined by expert assessments and opinion surveys. **The higher the score, the better and happier the country is.**

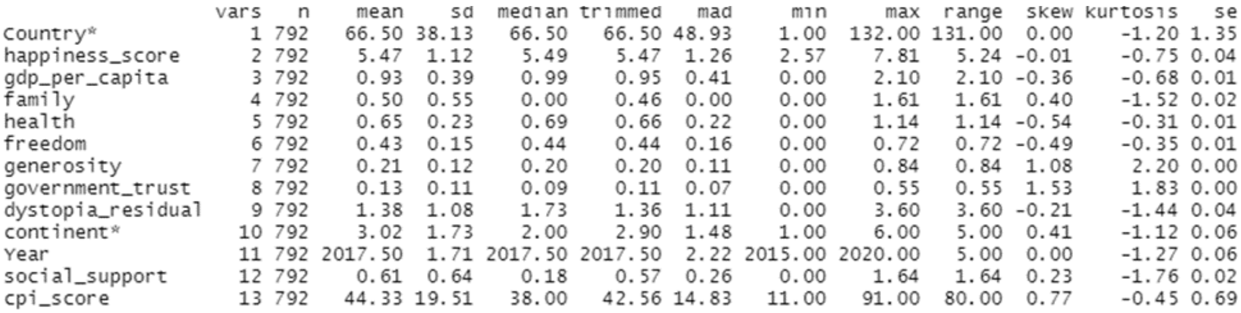
10. **Year** - Name of the year for which the observations were collected

11. **Country**: Name of the country

12. **social support** - Social support is the perception and actuality that one is cared for, has assistance available from other people, and most popularly, that one is part of a supportive social network.

Examining the dataset:

```
install.packages("psych")
library(psych)
 hc = read.csv("HappinessCorruption.csv", stringsAsFactors = T)
describe(hc)
```



```
                  vars   n    mean    sd  median trimmed   mad     min     max  range  skew kurtosis   se
Country*             1 792   66.50 38.13   66.50   66.50 48.93    1.00  132.00 131.00  0.00    -1.20 1.35
happiness_score      2 792    5.47  1.12    5.49    5.47  1.26    2.57    7.81   5.24 -0.01    -0.75 0.04
gdp_per_capita       3 792    0.93  0.39    0.99    0.95  0.41    0.00    2.10   2.10 -0.36    -0.68 0.01
family               4 792    0.50  0.55    0.00    0.46  0.00    0.00    1.61   1.61  0.40    -1.52 0.02
health               5 792    0.65  0.23    0.69    0.66  0.22    0.00    1.14   1.14 -0.54    -0.31 0.01
freedom              6 792    0.43  0.15    0.44    0.44  0.16    0.00    0.72   0.72 -0.49    -0.35 0.01
generosity           7 792    0.21  0.12    0.20    0.20  0.11    0.00    0.84   0.84  1.08     2.20 0.00
government_trust     8 792    0.13  0.11    0.09    0.11  0.07    0.00    0.55   0.55  1.53     1.83 0.00
dystopia_residual    9 792    1.38  1.08    1.73    1.36  1.11    0.00    3.60   3.60 -0.21    -1.44 0.04
continent*          10 792    3.02  1.73    2.00    2.90  1.48    1.00    6.00   5.00  0.41    -1.12 0.06
Year                11 792 2017.50  1.71 2017.50 2017.50  2.22 2015.00 2020.00   5.00  0.00    -1.27 0.06
social_support      12 792    0.61  0.64    0.18    0.57  0.26    0.00    1.64   1.64  0.23    -1.76 0.02
cpi_score           13 792   44.33 19.51   38.00   42.56 14.83   11.00   91.00  80.00  0.77    -0.45 0.69
```

Dataset and strategy for analysis:

- We have 3 categorical variables and 10 numeric variables.
- To address the research question whether ‘perception of corruption affects happiness of countries across the world’, we have 2 predictor variables closely related to each other namely government trust and cpi_score, and the happiness_score which is the outcome variable.
- We find that according to the dataset author, “Government *trust may already be taken from CPI scores , but CPI scores make it clearer to understand and contrast it with the dependent variable in our case happiness* score”.
- Hence we decided to use cpi_score for our DV and happiness_score as our IV
- There are many zero values in observations for variables such as family, dystopia_residual and social_support. This might become a limitation when all the variables are thrown into a model.
- In view of the above, the research question can be best analyzed by understanding the correlation between cpi_score and happiness_score variables.

Suitability of the Correlation test and alternative methods:

1. Both our IV and DV are numeric variables. The IV is numeric discrete and DV is numeric continuous. This lends to easy interpretation of the relationship between them when a correlation test is used.
2. Correlation test reveals not just the existence of a relationship or a lack thereof, but also the strength and direction of the relationship
3. The dataset provides other interesting variables that may have a greater correlation with the outcome variable than corruption perception score.
4. So, for further analysis we will introduce other variables and perform regression analysis.

We first compiled the descriptive statistics of our IV and DV to understand the nature of the distributions.

Our Dependant variable is "happiness_score"

Mean of happiness_score: 5.47

SD of happiness_score: 1.12

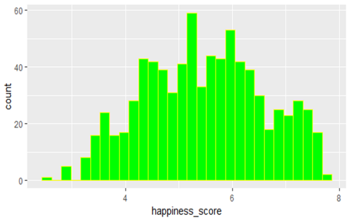
Our Independent variable is "cpi_score"

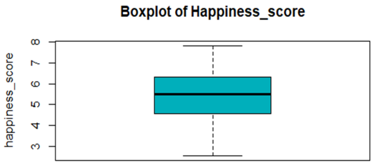
Mean of cpi_score: 44.33

SD od cpi_score: 19.51

Using R, we produced histograms and box plots for both the above variables, to understand the nature of the distributions.

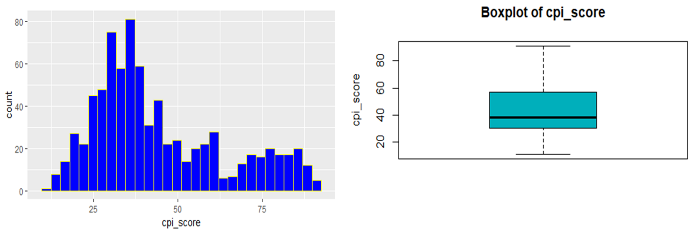
Histogram & Box plot of DV happiness_score:




Boxplot of Happiness_score

Histogram and Boxplot for our DV establish that the observations for 'happiness_score' are distributed normally.

Histogram & Box plot of cpi_score:

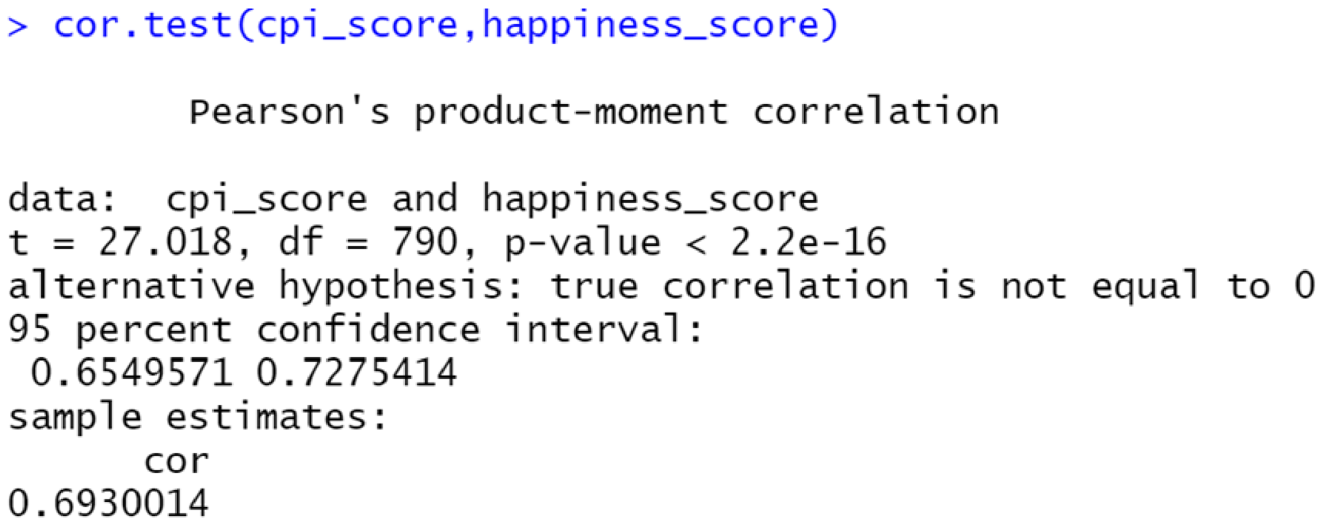


Histogram establishes that the observations for ‘cpi_score’ are right-skewed

RESULTS :

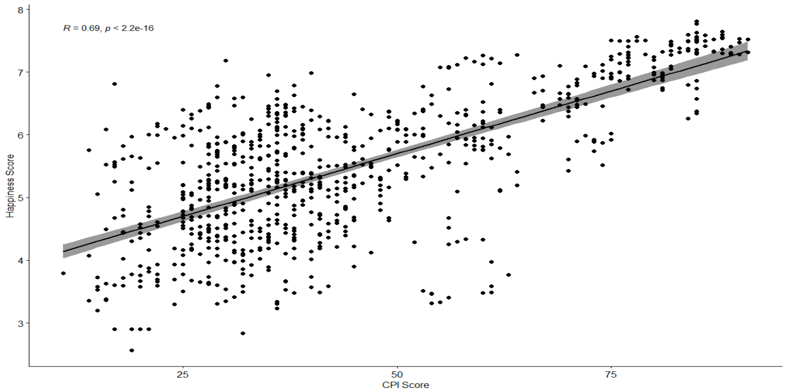
Examining correlation between the IV and DV:

Label-1: Correlation test output from R



```
> cor.test(cpi_score,happiness_score)

        Pearson's product-moment correlation

data:  cpi_score and happiness_score
t = 27.018, df = 790, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6549571 0.7275414
sample estimates:
      cor
0.6930014
```

Label-2: Scatter plot of IV and DV with a fitted regression line



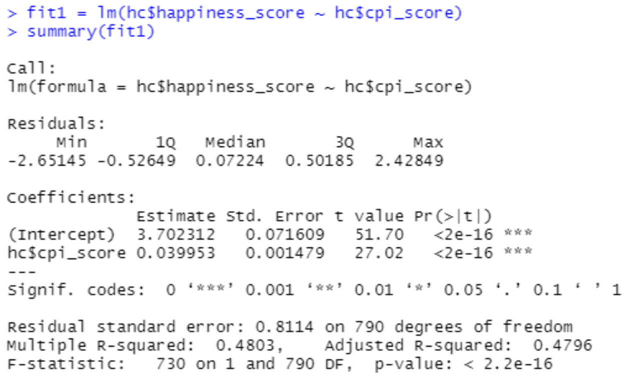
Interpretation : The test result is reported as (t= 27.018, df = 790, p value 2.2e-16) which is statistically significant. The correlation coefficient r = 0.693 which indicates that for a unit increase in cpi_score, there is a 0.693 increase in happiness_score.

To analyze the effect of other interesting variables on the happiness_score of countries we used the following Linear Regression models:

Before we add the other variables we first did a linear regression model on the cpi score to happiness score. This linear regression model had a low R-squared of 0.4796

Label-3: Regression analysis of IV and DV

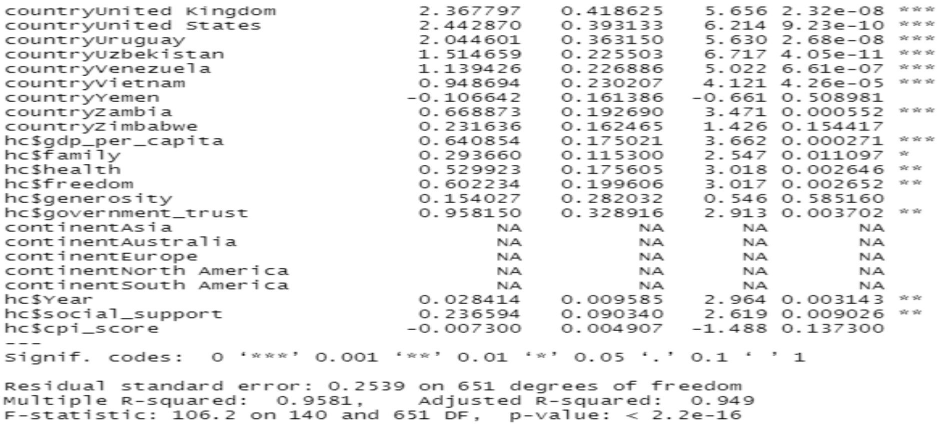
.



```
> fit1 = lm(hc$happiness_score ~ hc$cpi_score)
> summary(fit1)

Call:
lm(formula = hc$happiness_score ~ hc$cpi_score)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.65145 -0.52649  0.07224  0.50185  2.42849 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.702312   0.071609   51.70   <2e-16 ***
hc$cpi_score 0.039953   0.001479   27.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8114 on 790 degrees of freedom
Multiple R-squared:  0.4803,	Adjusted R-squared:  0.4796 
F-statistic:   730 on 1 and 790 DF,  p-value: < 2.2e-16
```

Regression Model 1:

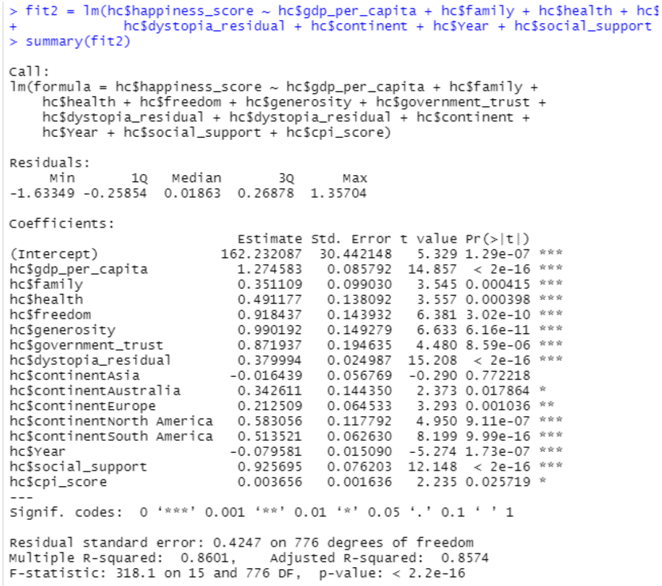
The first model we used all the available variables except happiness_score as independent variables. We found that many countries were not significant predictors so we decided to remove country from the model. The model with the countries did have a very high adjusted R-squared of 0.949



```
countryUnited Kingdom            2.367797    0.418625    5.656 2.32e-08 ***
countryUnited States             2.442870    0.393133    6.214 9.23e-10 ***
countryUruguay                   2.044601    0.363150    5.630 2.68e-08 ***
countryUzbekistan                1.514659    0.225503    6.717 4.05e-11 ***
countryVenezuela                 1.139426    0.226886    5.022 6.61e-07 ***
countryVietnam                   0.948694    0.230207    4.121 4.26e-05 ***
countryYemen                    -0.106642    0.161386   -0.661 0.508981
countryZambia                    0.668873    0.192690    3.471 0.000552 ***
countryZimbabwe                  0.231636    0.162465    1.426 0.154417
hc$gdp_per_capita                0.640854    0.175021    3.662 0.000271 ***
hc$family                        0.293660    0.115300    2.547 0.011097 *
hc$health                        0.529923    0.175605    3.018 0.002646 **
hc$freedom                       0.602234    0.199606    3.017 0.002652 **
hc$generosity                    0.154027    0.282032    0.546 0.585160
hc$government_trust              0.958150    0.328916    2.913 0.003702 **
continentAsia                          NA          NA       NA       NA
continentAustralia                     NA          NA       NA       NA
continentEurope                        NA          NA       NA       NA
continentNorth America                 NA          NA       NA       NA
continentSouth America                 NA          NA       NA       NA
hc$Year                          0.028414    0.009585    2.964 0.003143 **
hc$social_support                0.236594    0.090340    2.619 0.009026 **
hc$cpi_score                    -0.007300    0.004907   -1.488 0.137300
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2539 on 651 degrees of freedom
Multiple R-squared:  0.9581,    Adjusted R-squared:  0.949
F-statistic: 106.2 on 140 and 651 DF,  p-value: < 2.2e-16
```

Regression Model 2:

In the second model, we used all remaining independent variables except ‘country’ as predictor variables and happiness_score as outcome variable

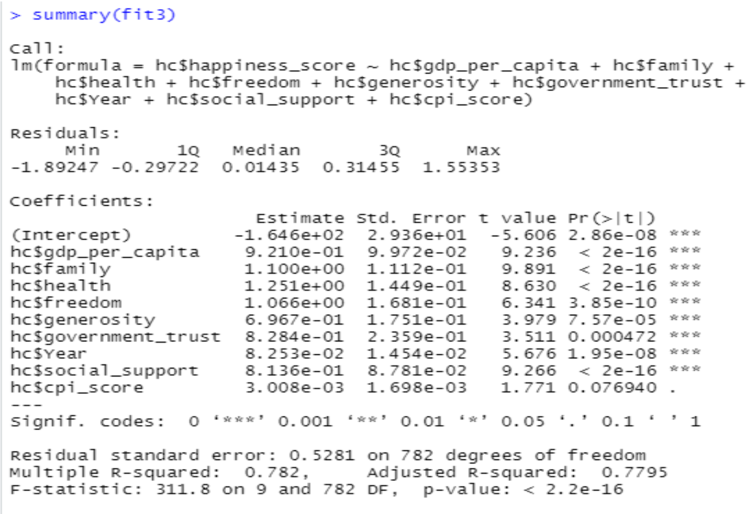


```
> fit2 = lm(hc$happiness_score ~ hc$gdp_per_capita + hc$family + hc$health + hc$
+            hc$dystopia_residual + hc$continent + hc$Year + hc$social_support
> summary(fit2)

Call:
lm(formula = hc$happiness_score ~ hc$gdp_per_capita + hc$family + 
    hc$health + hc$freedom + hc$generosity + hc$government_trust + 
    hc$dystopia_residual + hc$dystopia_residual + hc$continent + 
    hc$Year + hc$social_support + hc$cpi_score)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.63349 -0.25854  0.01863  0.26878  1.35704 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             162.232087  30.442148   5.329 1.29e-07 ***
hc$gdp_per_capita         1.274583   0.085792  14.857  < 2e-16 ***
hc$family                 0.351109   0.099030   3.545 0.000415 ***
hc$health                 0.491177   0.138092   3.557 0.000398 ***
hc$freedom                0.918437   0.143932   6.381 3.02e-10 ***
hc$generosity             0.990192   0.149279   6.633 6.16e-11 ***
hc$government_trust       0.871937   0.194635   4.480 8.59e-06 ***
hc$dystopia_residual      0.379994   0.024987  15.208  < 2e-16 ***
hc$continentAsia         -0.016439   0.056769  -0.290 0.772218    
hc$continentAustralia     0.342611   0.144350   2.373 0.017864 *  
hc$continentEurope        0.212509   0.064533   3.293 0.001036 ** 
hc$continentNorth America  0.583056   0.117792   4.950 9.11e-07 ***
hc$continentSouth America  0.513521   0.062630   8.199 9.99e-16 ***
hc$Year                  -0.079581   0.015090  -5.274 1.73e-07 ***
hc$social_support         0.925695   0.076203  12.148  < 2e-16 ***
hc$cpi_score              0.003656   0.001636   2.235 0.025719 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4247 on 776 degrees of freedom
Multiple R-squared:  0.8601,    Adjusted R-squared:  0.8574 
F-statistic: 318.1 on 15 and 776 DF,  p-value: < 2.2e-16
```

Interpretation: As can be seen in this model most predictor variables were statistically significant. The only part that was not was the continent of Asia. This model had an adjusted R - squared value of 0.8574.

Regression Model 3:

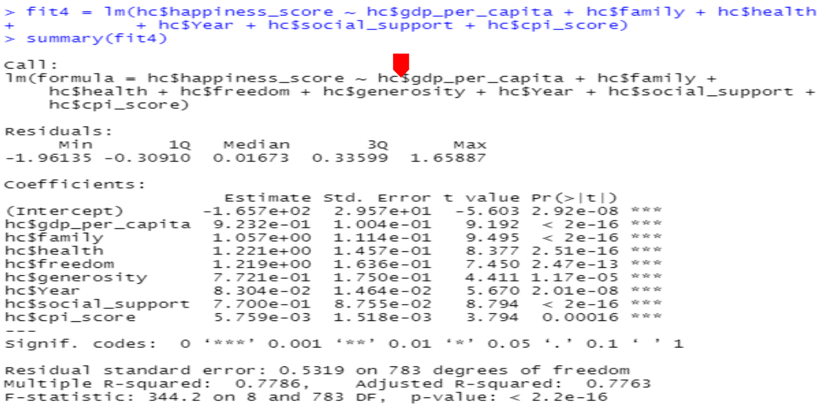
Since Asia was statistically insignificant we wanted to run a model without the continents to see if that would improve upon the model.



```
> summary(fit3)

Call:
lm(formula = hc$happiness_score ~ hc$gdp_per_capita + hc$family + 
    hc$health + hc$freedom + hc$generosity + hc$government_trust + 
    hc$Year + hc$social_support + hc$cpi_score)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.89247 -0.29722  0.01435  0.31455  1.55353 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -1.646e+02  2.936e+01  -5.606 2.86e-08 ***
hc$gdp_per_capita    9.210e-01  9.972e-02   9.236  < 2e-16 ***
hc$family            1.100e+00  1.112e-01   9.891  < 2e-16 ***
hc$health            1.251e+00  1.449e-01   8.630  < 2e-16 ***
hc$freedom           1.066e+00  1.681e-01   6.341 3.85e-10 ***
hc$generosity        6.967e-01  1.751e-01   3.979 7.57e-05 ***
hc$government_trust  8.284e-01  2.359e-01   3.511 0.000472 ***
hc$Year              8.253e-02  1.454e-02   5.676 1.95e-08 ***
hc$social_support    8.136e-01  8.781e-02   9.266  < 2e-16 ***
hc$cpi_score         3.008e-03  1.698e-03   1.771 0.076940 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5281 on 782 degrees of freedom
Multiple R-squared:  0.782,	Adjusted R-squared:  0.7795 
F-statistic: 311.8 on 9 and 782 DF,  p-value: < 2.2e-16
```

Interpretation: The accuracy from removing continents went down significantly with only an adjusted r-squared of 0.7795. Also CPI became an insignificant variable. We thought that this might be due to the close relationship between government trust and cpi score so we ran one last model that removed government trust.

Regression Model 4:



```
> fit4 = lm(hc$happiness_score ~ hc$gdp_per_capita + hc$family + hc$health 
+           + hc$Year + hc$social_support + hc$cpi_score)
> summary(fit4)

Call:
lm(formula = hc$happiness_score ~ hc$gdp_per_capita + hc$family + 
    hc$health + hc$freedom + hc$generosity + hc$Year + hc$social_support + 
    hc$cpi_score)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.96135 -0.30910  0.01673  0.33599  1.65887 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.657e+02  2.957e+01  -5.603 2.92e-08 ***
hc$gdp_per_capita  9.232e-01  1.004e-01   9.192  < 2e-16 ***
hc$family          1.057e+00  1.114e-01   9.495  < 2e-16 ***
hc$health          1.221e+00  1.457e-01   8.377 2.51e-16 ***
hc$freedom         1.219e+00  1.636e-01   7.450 2.47e-13 ***
hc$generosity      7.721e-01  1.750e-01   4.411 1.17e-05 ***
hc$Year            8.304e-02  1.464e-02   5.670 2.01e-08 ***
hc$social_support  7.700e-01  8.755e-02   8.794  < 2e-16 ***
hc$cpi_score       5.759e-03  1.518e-03   3.794  0.00016 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5319 on 783 degrees of freedom
Multiple R-squared:  0.7786,	Adjusted R-squared:  0.7763 
F-statistic: 344.2 on 8 and 783 DF,  p-value: < 2.2e-16
```

Interpretation: Removing government trust made cpi score a significant variable again and did not change the adjusted r-squared of the 2nd model by much. Model 4 has an adjusted r-squared of 0.7763.

DISCUSSION/CONCLUSION :

Weaknesses in the analysis:

While analyzing the dataset to answer our research question we came across a few weaknesses in the data set. We found there were a lot of 0 values in the social_support, family, and dystopia_residual columns. Also, the size of the dataset was smaller than we would like since it only spanned over 6 years. The last weakness we came across was that there are possible similarities between cpi_score and government_trust variables which was shown in our model analysis as well.

Practical Implications:

Our analysis has a number of real world practical implications. Our analysis can be used to help a government or organizations working towards the well being of its citizens find what factors lead to happiness in a country. Through this analysis we found that the CPI score as well as other variables lead to a higher happiness since the higher the CPI score the more transparent the government. This analysis can be used to measure how transparent the government is through the CPI score.

Conclusions:

Through our analysis we found that while cpi_score is a significant variable to happiness score it does not have as high of an estimated value in comparison to the other variables. Also, we found that cpi_score was positively correlated to happiness score so the higher the cpi_score the higher the happiness score. This relationship between the two variables is contrary to what one might think initially since you might expect a high perception of corruption to be a sign of an unhappy country. This was shown to not be the case in our analysis since it showed that the higher the cpi_score was the lower the actual corruption is due to the increase in awareness.