Analysis of World Happiness Report

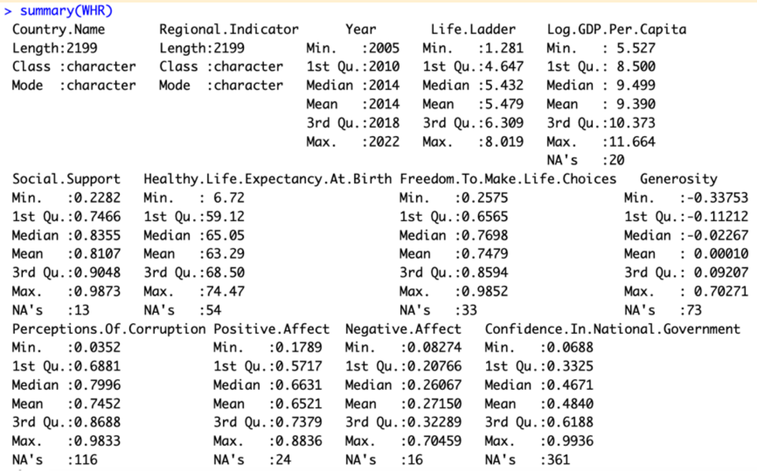
Yu-Lun Tsai

✓ Introduction

Description of data

The dataset was downloaded from Kaggle(https://www.kaggle.com/datasets/usamabuttar/world-happiness-report-2005-present/data), which has 2199 rows and 13 columns. The World Happiness Index has multiple calculation indexes: GDP per capita, Social Support, Healthy Life Expectancy, Freedom to make Life Choices, Generosity, and Perception of Corruption. In the dataset, it also has the columns of Regional Indicator, Positive Affect, Negative Affect, and Confidence of National Government.

Below is the summary of the whole data:



```
> summary(WHR)
 Country.Name       Regional.Indicator      Year        Life.Ladder    Log.GDP.Per.Capita
 Length:2199        Length:2199        Min.   :2005   Min.   :1.281   Min.   : 5.527
 Class :character   Class :character   1st Qu.:2010   1st Qu.:4.647   1st Qu.: 8.500
 Mode  :character   Mode  :character   Median :2014   Median :5.432   Median : 9.499
                                       Mean   :2014   Mean   :5.479   Mean   : 9.390
                                       3rd Qu.:2018   3rd Qu.:6.309   3rd Qu.:10.373
                                       Max.   :2022   Max.   :8.019   Max.   :11.664
                                                                      NA's   :20
 Social.Support   Healthy.Life.Expectancy.At.Birth Freedom.To.Make.Life.Choices   Generosity
 Min.   :0.2282   Min.   : 6.72                    Min.   :0.2575               Min.   :-0.33753
 1st Qu.:0.7466   1st Qu.:59.12                    1st Qu.:0.6565               1st Qu.:-0.11212
 Median :0.8355   Median :65.05                    Median :0.7698               Median :-0.02267
 Mean   :0.8107   Mean   :63.29                    Mean   :0.7479               Mean   : 0.00010
 3rd Qu.:0.9048   3rd Qu.:68.50                    3rd Qu.:0.8594               3rd Qu.: 0.09207
 Max.   :0.9873   Max.   :74.47                    Max.   :0.9852               Max.   : 0.70271
 NA's   :13       NA's   :54                       NA's   :33                   NA's   :73
 Perceptions.Of.Corruption Positive.Affect  Negative.Affect   Confidence.In.National.Government
 Min.   :0.0352            Min.   :0.1789   Min.   :0.08274   Min.   :0.0688
 1st Qu.:0.6881            1st Qu.:0.5717   1st Qu.:0.20766   1st Qu.:0.3325
 Median :0.7996            Median :0.6631   Median :0.26067   Median :0.4671
 Mean   :0.7452            Mean   :0.6521   Mean   :0.27150   Mean   :0.4840
 3rd Qu.:0.8688            3rd Qu.:0.7379   3rd Qu.:0.32289   3rd Qu.:0.6188
 Max.   :0.9833            Max.   :0.8836   Max.   :0.70459   Max.   :0.9936
 NA's   :116               NA's   :24       NA's   :16        NA's   :361
```

For this project, the main goal I want to know is: If there are any relations between Xs and Y? and which X affects Y the most?

Process of data cleaning

From the summary above, we can observe each column has missing values. For the Regional Indicator, I use the functions below to fill up the region for countries.

```
rows <- which(WHR$Country.Name == "Angola")
WHR$Regional.Indicator[rows] <- "Sub-Saharan Africa"
rows <- which(WHR$Country.Name == "Belize")
WHR$Regional.Indicator[rows] <- "Latin America and Caribbean"
rows <- which(WHR$Country.Name == "Bhutan")
WHR$Regional.Indicator[rows] <- "South Asia"
```

For other numeric values, I used the country’s median to fill up the values.

```
#ConfidenceInNationalGovernment
```

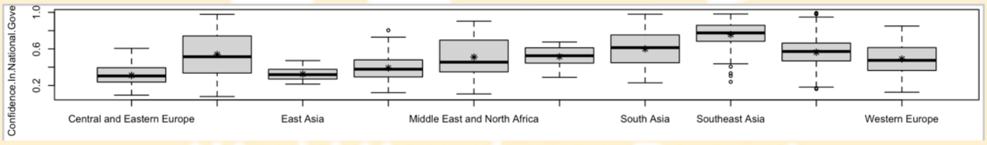
```
data_with_median <- WHR %>%
   group_by(Country.Name) %>%
   mutate(Confidence_In_National_Government_Median =
median(Confidence.In.National.Government, na.rm = TRUE))
WHR <- data_with_median %>%
   mutate(Confidence.In.National.Government
=ifelse(is.na(Confidence.In.National.Government),
Confidence_In_National_Government_Median,
Confidence.In.National.Government)) %>%
   select(-Confidence_In_National_Government_Median)
```

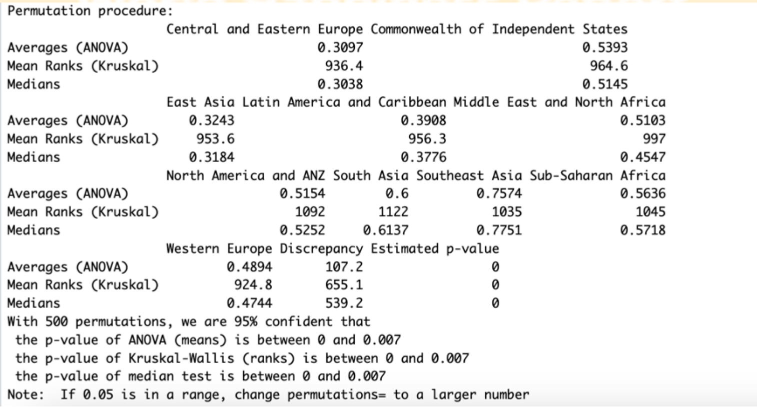
After these two types of processes, if there are still missing values such as a country having no data about a specific column, I delete the country for analysis normally.

```
WHR <- WHR[complete.cases(WHR), ]
```

✓ Association Analysis & Regression Models

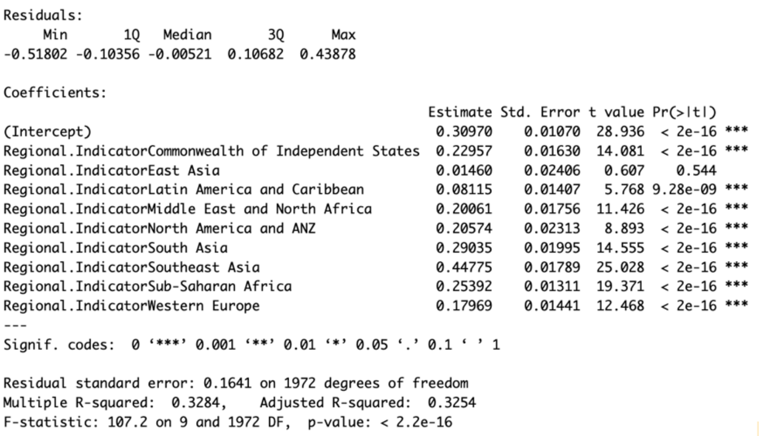
Region v.s. Confidence of National Government





```
Permutation procedure:
                     Central and Eastern Europe Commonwealth of Independent States
Averages (ANOVA)                         0.3097                             0.5393
Mean Ranks (Kruskal)                      936.4                              964.6
Medians                                  0.3038                             0.5145
                     East Asia Latin America and Caribbean Middle East and North Africa
Averages (ANOVA)        0.3243                      0.3908                       0.5103
Mean Ranks (Kruskal)     953.6                       956.3                          997
Medians                 0.3184                      0.3776                       0.4547
                     North America and ANZ South Asia Southeast Asia Sub-Saharan Africa
Averages (ANOVA)                    0.5154        0.6         0.7574             0.5636
Mean Ranks (Kruskal)                  1092       1122           1035               1045
Medians                             0.5252     0.6137         0.7751             0.5718
                     Western Europe Discrepancy Estimated p-value
Averages (ANOVA)             0.4894       107.2                 0
Mean Ranks (Kruskal)          924.8       655.1                 0
Medians                      0.4744       539.2                 0
With 500 permutations, we are 95% confident that
 the p-value of ANOVA (means) is between 0 and 0.007
 the p-value of Kruskal-Wallis (ranks) is between 0 and 0.007
 the p-value of median test is between 0 and 0.007
Note:  If 0.05 is in a range, change permutations= to a larger number
```

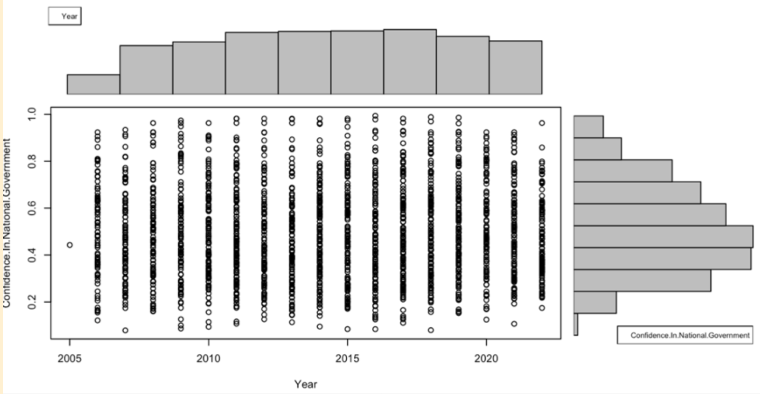
The box plot shows that the media, Q1, and Q3 have noticeable differences between regions. The p-value is 0 and the p-value of ANOVA is between 0-0.007, which indicates it is conclusive and statistically significant.

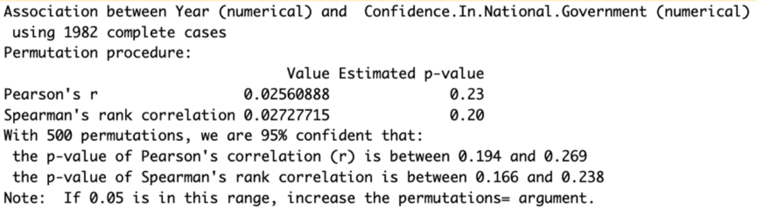


```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.51802 -0.10356 -0.00521  0.10682  0.43878 

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                      0.30970    0.01070  28.936  < 2e-16 ***
Regional.IndicatorCommonwealth of Independent States  0.22957    0.01630  14.081  < 2e-16 ***
Regional.IndicatorEast Asia                      0.01460    0.02406   0.607    0.544    
Regional.IndicatorLatin America and Caribbean    0.08115    0.01407   5.768 9.28e-09 ***
Regional.IndicatorMiddle East and North Africa   0.20061    0.01756  11.426  < 2e-16 ***
Regional.IndicatorNorth America and ANZ          0.20574    0.02313   8.893  < 2e-16 ***
Regional.IndicatorSouth Asia                     0.29035    0.01995  14.555  < 2e-16 ***
Regional.IndicatorSoutheast Asia                 0.44775    0.01789  25.028  < 2e-16 ***
Regional.IndicatorSub-Saharan Africa             0.25392    0.01311  19.371  < 2e-16 ***
Regional.IndicatorWestern Europe                 0.17969    0.01441  12.468  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1641 on 1972 degrees of freedom
Multiple R-squared:  0.3284,	Adjusted R-squared:  0.3254 
F-statistic: 107.2 on 9 and 1972 DF,  p-value: < 2.2e-16
```

The regression analysis of Region, The intercept is 0.3097 and all regions positively correlate with the y-variable. From there we can see here that only East Asia has no significant difference. And it is the question that I would get the answer and write it at my conclusion. RMSE is 0.1641, which means the observed values are close to the fitted regression line. R squared is 32.84%, which is have pretty high percentage that Region will affect the index of Confidence of National Government.

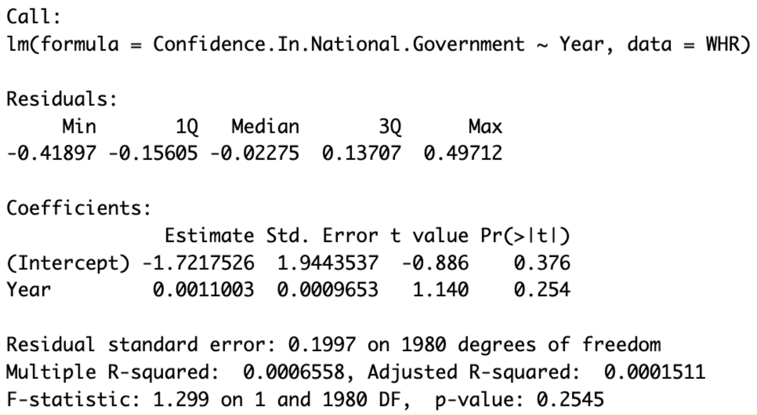
Year v.s. Confidence of National Government





```
Association between Year (numerical) and  Confidence.In.National.Government (numerical)
 using 1982 complete cases
Permutation procedure:
                                  Value Estimated p-value
Pearson's r                  0.02560888              0.23
Spearman's rank correlation 0.02727715              0.20
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0.194 and 0.269
 the p-value of Spearman's rank correlation is between 0.166 and 0.238
Note:  If 0.05 is in this range, increase the permutations= argument.
```

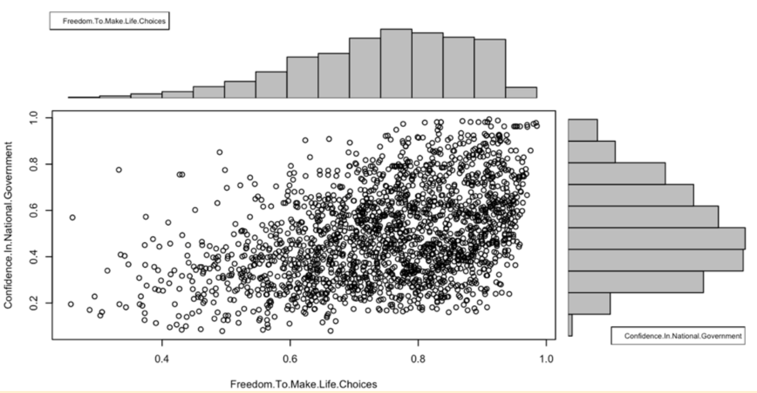
From the scatter plot, we can observe that it is a non-linear relationship, so I chose Spearman's correlations to be the correlation to calculate. The value is 0.27, and the p-value range of Spearman's rank correlation does not include 0.05, so we can conclude that it is conclusive but not statistically significant. It is an interesting finding for me, because it means the year and confidence of the national government did not have a significant relationship. It means the outcome is totally rebut my thoughts which in certain years of social unrest or economic weakness, people's confidence in their national governments would decrease.

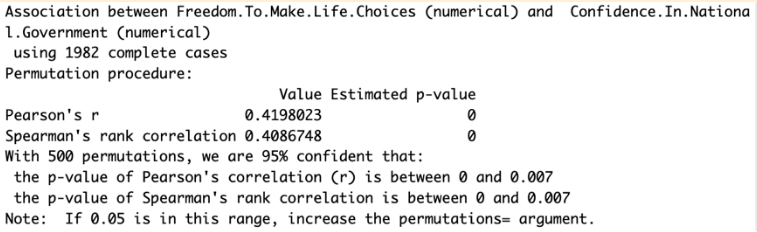


```
Call:
lm(formula = Confidence.In.National.Government ~ Year, data = WHR)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.41897 -0.15605 -0.02275  0.13707  0.49712 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.7217526  1.9443537  -0.886    0.376
Year         0.0011003  0.0009653   1.140    0.254

Residual standard error: 0.1997 on 1980 degrees of freedom
Multiple R-squared:  0.0006558, Adjusted R-squared:  0.0001511 
F-statistic: 1.299 on 1 and 1980 DF,  p-value: 0.2545
```

Regarding regression, there has been a slight increase with Year, and the base starts from -1.7217526. Each Year increases by 0.0011003 unit confidence index. RMSE is 0.1997, which is also small and close to the fitted regression line. R squared is 0.06%, meaning it only affects the y-variable with a very tiny influence.

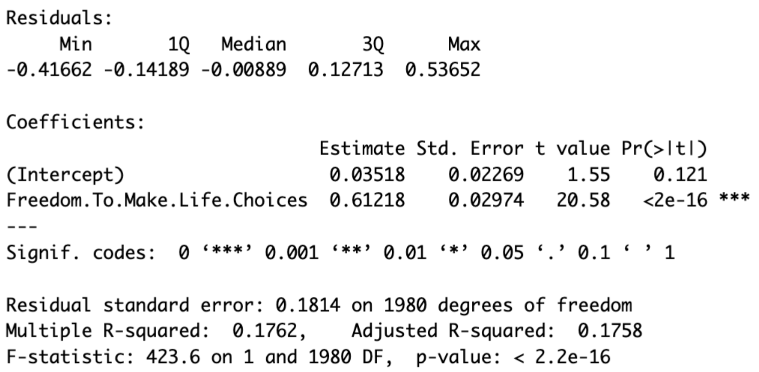
Freedom to Make Life Choices v.s. Confidence of National Government





```
Association between Freedom.To.Make.Life.Choices (numerical) and  Confidence.In.Nationa
l.Government (numerical)
 using 1982 complete cases
Permutation procedure:
                             Value Estimated p-value
Pearson's r              0.4198023                 0
Spearman's rank correlation 0.4086748              0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
Note:  If 0.05 is in this range, increase the permutations= argument.
```

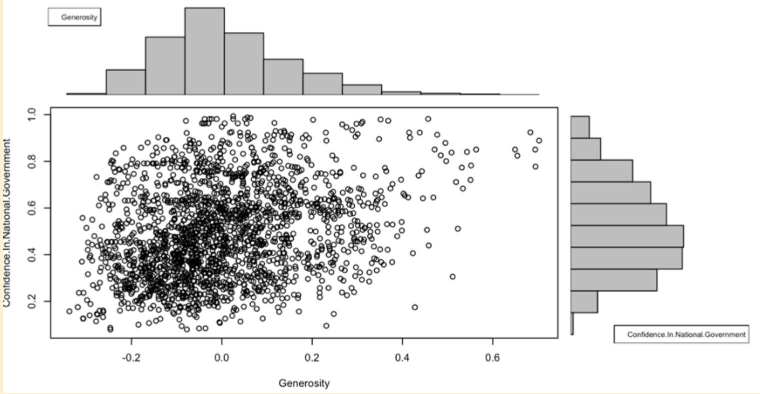
We can see the plot as a clear Non-linear and monotonic plot, in which the plot points are positive. Pearson's correlation is 0.41, with a p-value of 0 and the p-vlaue of Spearman’s rank correlation is between 0-0.007, which can conclude that it is conclusive and statistically significant.

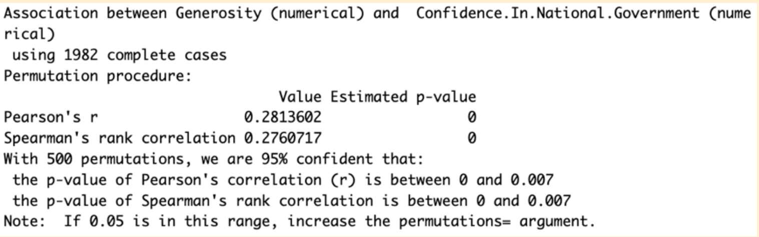


```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.41662 -0.14189 -0.00889  0.12713  0.53652 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   0.03518    0.02269    1.55    0.121    
Freedom.To.Make.Life.Choices  0.61218    0.02974   20.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1814 on 1980 degrees of freedom
Multiple R-squared:  0.1762,	Adjusted R-squared:  0.1758 
F-statistic: 423.6 on 1 and 1980 DF,  p-value: < 2.2e-16
```

We can see that the trend of the Confidence index gradually increases with the degree of freedom. Each freedom index increases by 0.6 unit confidence index and the intercept is 0.03518. RMSE is 0.1814, which is also small and close to the fitted regression line. R squared is 17.62%, which is a relatively high probability of affecting the y-variable.

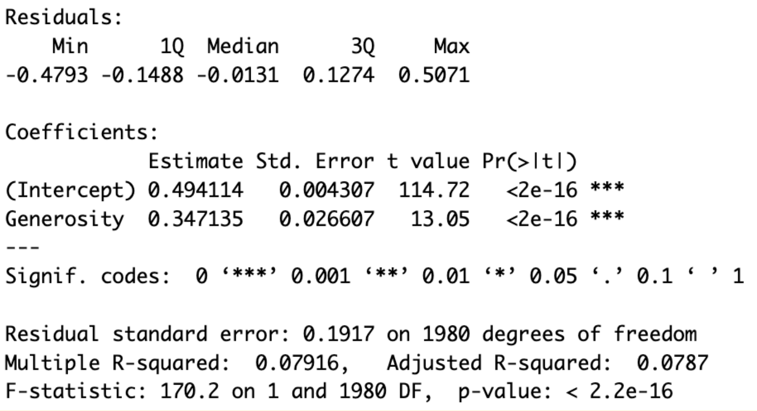
Generosity v.s. Confidence of National Government





```
Association between Generosity (numerical) and  Confidence.In.National.Government (nume
rical)
 using 1982 complete cases
Permutation procedure:
                                 Value Estimated p-value
Pearson's r                  0.2813602                 0
Spearman's rank correlation 0.2760717                 0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
Note:  If 0.05 is in this range, increase the permutations= argument.
```

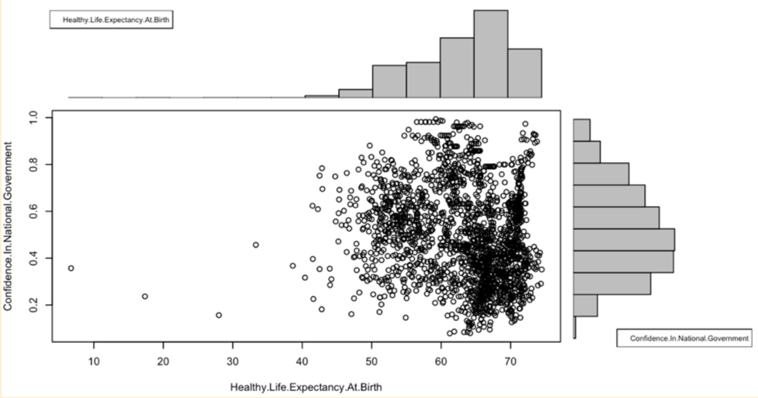
The plot is Non-linear and monotonic, in which the plot points are in the positive trend. Pearson's correlation is 0.28, with a p-value of 0 and the p-value of Pearson's rank correlation is between 0-0.007, which can conclude that it is conclusive and statistically significant.

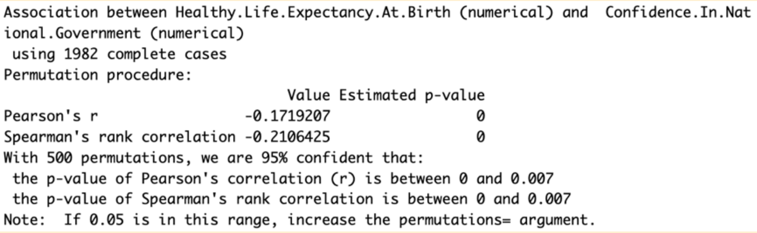


```
Residuals:
    Min      1Q  Median      3Q     Max 
-0.4793 -0.1488 -0.0131  0.1274  0.5071 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.494114   0.004307  114.72   <2e-16 ***
Generosity  0.347135   0.026607   13.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1917 on 1980 degrees of freedom
Multiple R-squared:  0.07916,	Adjusted R-squared:  0.0787 
F-statistic: 170.2 on 1 and 1980 DF,  p-value: < 2.2e-16
```

The regression of Generosity is in a positive correlation. Each Generosity index increases by 0.35 unit confidence index. RMSE is small as well and close to the fitted regression line. R squared is 7.9%, which does not that affect y compared to the previous index.

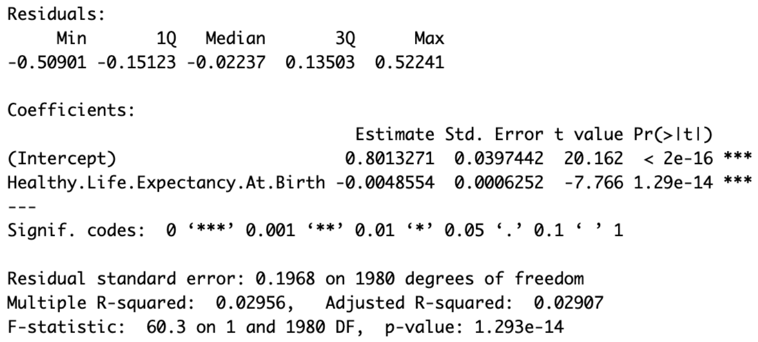
Healthy Life Expectancy v.s. Confidence of National Government





```
Association between Healthy.Life.Expectancy.At.Birth (numerical) and  Confidence.In.Nat
ional.Government (numerical)
 using 1982 complete cases
Permutation procedure:
                                 Value Estimated p-value
Pearson's r                 -0.1719207                 0
Spearman's rank correlation -0.2106425                 0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
Note:  If 0.05 is in this range, increase the permutations= argument.
```

The outcome of which is very interesting. We can see it is a non-linear plot, and the correlation is negative! It is the only index that performed the negative correlation. The Pearson's rank correlation ranks between 0-0.007, a conclusive and statistically significant relation. Is that because the residents of developed countries have lower confidence in their national government? Or are there any lurking variables? The research I did will be included in the conclusion.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.50901 -0.15123 -0.02237  0.13503  0.52241 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       0.8013271  0.0397442  20.162  < 2e-16 ***
Healthy.Life.Expectancy.At.Birth -0.0048554  0.0006252  -7.766 1.29e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1968 on 1980 degrees of freedom
Multiple R-squared:  0.02956,	Adjusted R-squared:  0.02907 
F-statistic:  60.3 on 1 and 1980 DF,  p-value: 1.293e-14
```

The regression of Healthy Life Expectancy is in a negative correlation. Each Healthy Life Expectancy index (which also means one year old) decreases by 0.005 unit confidence index. RMSE is tiny as well and close to the fitted regression line. R squared is 2.96%, which does not affect y compared to other indexes.

✓ Additional Packages

According to my own needs in data analysis, the new packages I applied focus on improving analysis accuracy. The data for this project is based on a country-by-country breakdown, which means I need to find a way to group data by country, so I found and applied these two additional packages in my code: dplyr and magrittr. I mainly used pipe operator, group_by, mutate, and select to organize my data. The pipe operator helps me simplify my code and helps me better understand this code I just learned. Group_by is used to group data by country, which I can calculate the median by each country, mutate to add the new columns I need in the following actions, and select used to delete the columns I created for calculation.

```
#Take healthy Life Expectancy for example
data_with_median <- WHR %>%
        group_by(Country.Name) %>%
        mutate(HL_Median = median(Healthy.Life.Expectancy.At.Birth, na.rm=TRUE))
WHR <- data_with_median %>%
        mutate(Healthy.Life.Expectancy.At.Birth=ifelse(is.na(Healthy.Life.Expectan
                                                cy.At.Birth),
                                                HL_Median,
                                                Healthy.Life.Expectancy.At.Birth)) %>%
          select(-HL_Median)
```

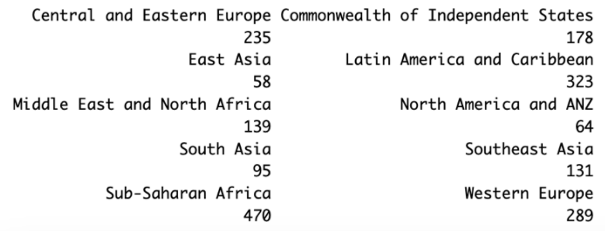
✓ Conclusion

In the finished analysis I successfully found out the answer to my main goal: Region is the x-variable that influences the confidence index of national government the most. Sort by influence: Region, Freedom to make life choices, Generosity, Healthy life expectancy, and the variable affect the least is year.

I also put all calculation indexes in the World Happiness Index to calculate the RMSE. For all indexes, sort by influence: Region, Perception of Corruption, Freedom to make life choices, Generosity, GDP per capita, Healthy life expectancy, Social Support, and the variable affect the least still is year.

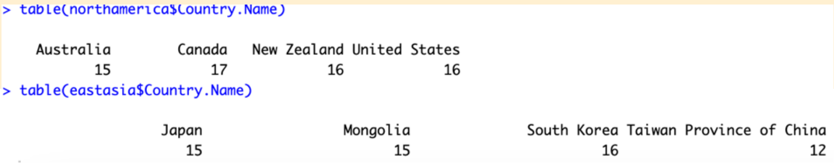
After the whole analysis, I popped out with two additional questions: Why is it that only East Asia has a higher p-value in the Region Index? Why does the Healthy life expectancy negatively correlate with the Confidence of the national government? Is this the phenomenon that residents of developed countries have lower confidence in their national government?

First, I want to determine why only East Asia has a high p-value. Below is the table of regional indicators:

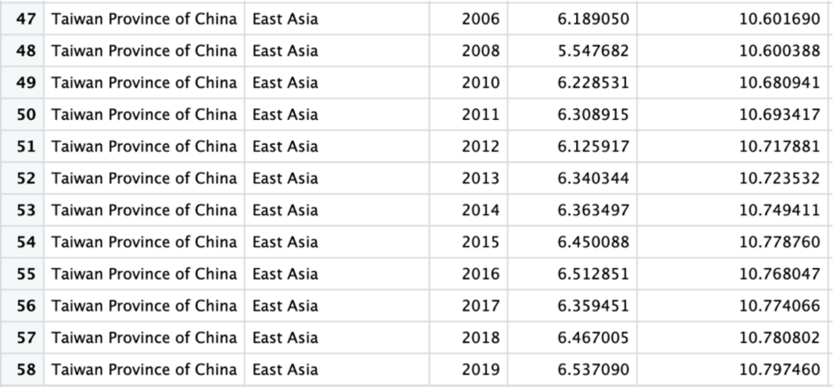


```
  Central and Eastern Europe Commonwealth of Independent States 
                         235                                178 
                   East Asia        Latin America and Caribbean 
                          58                                323 
  Middle East and North Africa             North America and ANZ 
                         139                                 64 
                  South Asia                     Southeast Asia 
                          95                                131 
          Sub-Saharan Africa                     Western Europe 
                         470                                289 
```

In the table, we can observe that East Asia has the smallest number of values, which may cause the problem that the p-value's outcome is inaccurate. Next, I tried to look at East Asia and North America, these two of the smallest subsets, to find why the North America is statically significant but East Asia is not.



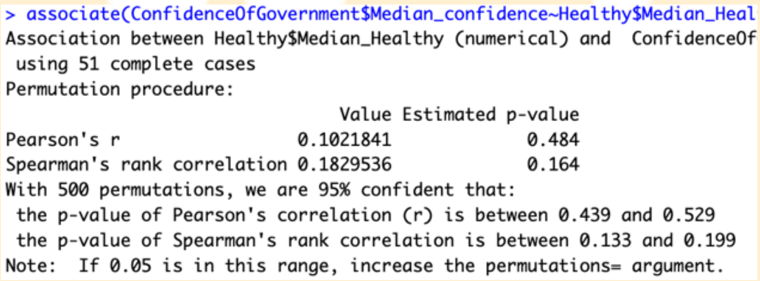
```
> table(northamerica$Country.Name)

    Australia        Canada   New Zealand United States 
           15            17            16            16 
> table(eastasia$Country.Name)

                   Japan                 Mongolia              South Korea Taiwan Province of China 
                      15                       15                       16                       12 
```



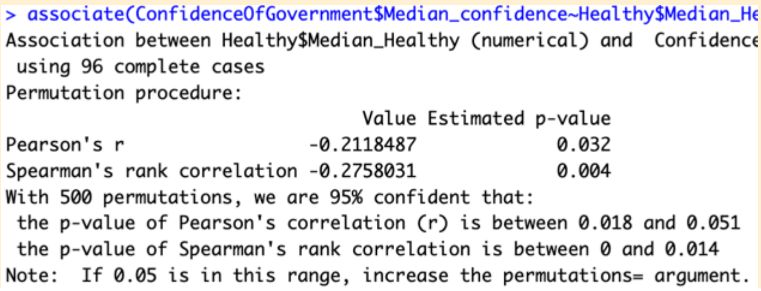
47	Taiwan Province of China	East Asia	2006	6.189050	10.601690
48	Taiwan Province of China	East Asia	2008	5.547682	10.600388
49	Taiwan Province of China	East Asia	2010	6.228531	10.680941
50	Taiwan Province of China	East Asia	2011	6.308915	10.693417
51	Taiwan Province of China	East Asia	2012	6.125917	10.717881
52	Taiwan Province of China	East Asia	2013	6.340344	10.723532
53	Taiwan Province of China	East Asia	2014	6.363497	10.749411
54	Taiwan Province of China	East Asia	2015	6.450088	10.778760
55	Taiwan Province of China	East Asia	2016	6.512851	10.768047
56	Taiwan Province of China	East Asia	2017	6.359451	10.774066
57	Taiwan Province of China	East Asia	2018	6.467005	10.780802
58	Taiwan Province of China	East Asia	2019	6.537090	10.797460

We can see both include four countries, the difference in sample size comes from the data for 2007, 2009, and the recent three years are unavailable in Taiwan. Because of the small sample size and the lack of current data, this may be why the association analysis in East Asia is not significant, so this analysis in East Asia may be inaccurate compared to other regions.

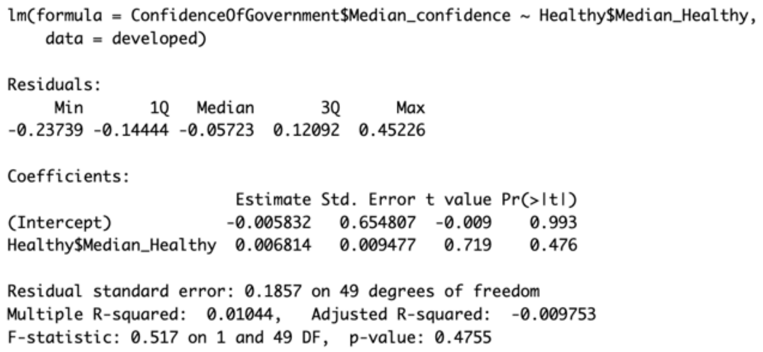
The second question is why the Healthy life expectancy negatively correlates with the Confidence of the national government. First, I use the GDP per capita to subset the developed/developing/least developed countries to analyze the data.



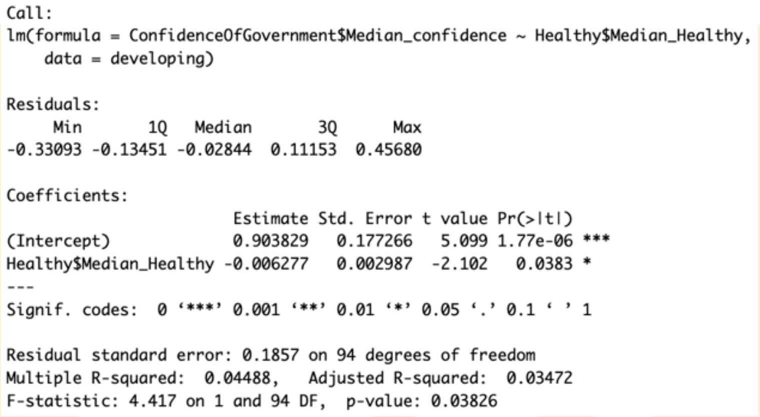
```
> associate(ConfidenceOfGovernment$Median_confidence~Healthy$Median_Heal
Association between Healthy$Median_Healthy (numerical) and  ConfidenceOf
 using 51 complete cases
Permutation procedure:
                                Value Estimated p-value
Pearson's r                 0.1021841             0.484
Spearman's rank correlation 0.1829536             0.164
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0.439 and 0.529
 the p-value of Spearman's rank correlation is between 0.133 and 0.199
Note:  If 0.05 is in this range, increase the permutations= argument.
```



```
> associate(ConfidenceOfGovernment$Median_confidence~Healthy$Median_He
Association between Healthy$Median_Healthy (numerical) and  Confidence
 using 96 complete cases
Permutation procedure:
                                 Value Estimated p-value
Pearson's r                 -0.2118487             0.032
Spearman's rank correlation -0.2758031             0.004
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0.018 and 0.051
 the p-value of Spearman's rank correlation is between 0 and 0.014
Note:  If 0.05 is in this range, increase the permutations= argument.
```



```
lm(formula = ConfidenceOfGovernment$Median_confidence ~ Healthy$Median_Healthy,
    data = developed)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.23739 -0.14444 -0.05723  0.12092  0.45226 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            -0.005832   0.654807  -0.009    0.993
Healthy$Median_Healthy  0.006814   0.009477   0.719    0.476

Residual standard error: 0.1857 on 49 degrees of freedom
Multiple R-squared:  0.01044,   Adjusted R-squared:  -0.009753 
F-statistic: 0.517 on 1 and 49 DF,  p-value: 0.4755
```



```
Call:
lm(formula = ConfidenceOfGovernment$Median_confidence ~ Healthy$Median_Healthy,
    data = developing)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.33093 -0.13451 -0.02844  0.11153  0.45680 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             0.903829   0.177266   5.099 1.77e-06 ***
Healthy$Median_Healthy -0.006277   0.002987  -2.102   0.0383 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1857 on 94 degrees of freedom
Multiple R-squared:  0.04488,   Adjusted R-squared:  0.03472 
F-statistic: 4.417 on 1 and 94 DF,  p-value: 0.03826
```

After dividing data into developed and developing countries, I surprisingly found out the negative correlation has come from developing countries. Because the samples of a developing country are more than developed, it leads to the whole dataset having a negative correlation with healthy life expectancy.