THE WORLD HAPPINESS REPORT DATA ANALYSIS

BY KELLY FOX

CSC558 – DR. PARSON

MARCH 2020

OVERVIEW

Happiness - the state of being happy. That is how happiness is defined in the dictionary, but what does happiness mean to you? There are many factors that partake in someone’s happiness. With the use of The World Happiness Reports of 2015, 2016, 2017, 2018 and 2019 datasets, I will be analyzing the data from The World Happiness Report to determine which factors have the greatest impact in each year.

HOW IS HAPPINESS SCORED?

Happiness scores are measured by how people rate each of the following criteria. The background question people are given is, “Please imagine a ladder, with steps numbered from 0 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?” This measure is also called the Cantril Ladder. Not all of the attributes are measured by a question, some are taken directly from other sources.

Attribute	How it is Measured
Rank	Countries ranked in order of the best Happiness Score to the lowest Happiness Score.
Country/Region	Country/Region Name.
Economy (GDP Per Capita)	This is based off of the GDP-per-capita time series from 2018 to 2019 using country- specific forecasts of real GDP growth in 2019 first from the OECD Economic Outlook No 106 (Edition November 2019) and then, if missing, forecasts from World Bank’s Global Economic Prospects.
Social Support	National average of the binary responses (either 0 or 1) to the GWP question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”
Healthy Life Expectancy	Healthy life expectancies at birth are based on the data extracted from the World Health Organization’s (WHO) Global Health Observatory data repository.
Freedom to Make Life Choices	National average of responses to the GWP question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”
Generosity	Residual of regressing national average of response to the GWP question “Have you donated money to a charity in the past month?”
Perceptions of Corruption	This is the measure of the national average of the survey responses to two questions in the GWP: “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?” The

	overall perception is just the average of the two responses.
Score	This is the national average response to the question of life evaluations which are above.

Table 1

Explanations from: https://happiness-report.s3.amazonaws.com/2020/WHR20_Ch2_Statistical_Appendix.pdf

4.2.A DATA SOURCES

The data set used was found on Kaggle:
https://www.kaggle.com/unsdsn/world-happiness#2019.csv
(There are links to the other years from this Kaggle dataset.)

The data was collected from a Gallop World Poll.

4.2.B DATA

4.2.B.1 CHOOSING A DATASET

I originally decided to work with The World Happiness Report dataset because it really intrigued me. I think happiness is very important for someone's mental well-being. Happiness affects our health, decisions, relationships, work and goals. If you are not happy, then you have to identify areas that are causing unhappiness. I am very curious as to how countries around the world rank their happiness. After reading how happiness is scored it had me thinking about how I would rank these aspects living in America. I think this dataset is very relatable and gets you thinking about your own happiness.

4.2.B.2 PROBLEM

When I navigated to Kaggle about a week after choosing my data set, I realized a new World Happiness Report dataset was just released for the year 2020. They released the 2020 dataset on March 20, 2020, which is international day of happiness. For a moment in time I almost changed my dataset to 2020, but after farther inspection, the 2020 dataset was much different than the 2019 dataset. The 2020 dataset had already taken the data and analyzed it. They had taken the factors that were determining the happiness overall score and estimated how each one was related to the happiness score. After taking these items into consideration, I proceeded with the datasets for the years 2015-2016 and figured out my main goal for this assignment.

4.2.B.3 GOAL

I have one main goal for this data. I want to determine what influenced The World Happiness rankings the most for the years 2015-2019. In this data set, many factors go into the overall happiness ranking. I would like to figure out which factor has the greatest impact on the happiness scores.

4.2.B.4 NEW OR PREVIOUS ANALYSIS

I have never worked with this data before, this will be a new analysis.

4.2.C STEPS

4.2.C.1 STEP 1

After downloading the 5 datasets (2015, 2016, 2017, 2018, 2019) datasets from Kaggle, I opened them in Excel. From briefly looking over them on Kaggle I did not see any errors or problems. Upon farther inspection in Excel, there was no empty values or out of place characters in the datasets for 2015, 2016, 2017 or 2019. There was only one N/A value in the dataset for 2018, which I changed to a 0. The datasets were very clean.

4.2.C.2 STEP 2

After examining the data in Excel, I had to get them into Weka. The datasets are originally CSV files, so I had to convert them into .arff files to be able to work with them in Weka. I did the following in order to convert the .cvc to .arff files for all of the datasets: open the Weka GUI Chooser window → selecting Tools along the top → ArffViewer → File → Open → Changing the Files of Type selection at the bottom to CSV → Select the dataset on my computer → Open → File → Save As → Type a file name → Hit Save. After completing those steps for each dataset, I had 5 .arff files to work with.

4.2.C.3 STEP 3

I opened my 2015.arff, 2016.arff, 2017.arff, 2018.arff and 2019.arff files in Weka. So far, I have not encountered any problems, the data was clean and easy to read.

4.2.C.4 STEP 4

After looking at the columns in excel, I knew I would have to get rid of some attributes. There were a few attributes that occurred in some years but did not occur in other years. Different years had different attributes, but there were seven attributes that remained the same in all of the data sets that I would be able to compare through the years. The attributes I would be keeping are overall rank, country or region, happiness score or ladder, GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption. In order to make things easier for me, I changed the names and order of the attributes in all datasets to the same thing.

WEKA STEPS TO RENAME ATTRIBUTES:

Under the preprocess tab → choose → Weka → filters → unsupervised → attribute → RenameAttribute. After clicking RenameAttribute, input the attributeIndices you want to rename and type the new name in "replace". Select OK then Apply.

WEKA STEPS TO CHANGE SCORE TO THE LAST ATTRIBUTE:

The target attribute I will be working with is "score". Originally this attribute was not the last attribute in the list for any of the datasets, so I had to move it to the last attribute on the list for each dataset. In order to move "score" to last attribute, in the preprocess window under Filter I selected Choose → Weka → filters → unsupervised → attribute → Reorder. After selecting reorder, type the order of the attributes you want into attributeIndices → OK → Apply. Now we have the target attribute "score" at the end and the other attributes are untouched.

I CHANGED THE NINE NAMES AND THE ORDER OF THE ATTRIBUTES IN EACH DATASET TO:

Rank
Country/Region
Economy (GDP Per Capita)
Social Support
Life Expectancy
Freedom
Generosity
Perceptions of Corruption
Happiness Score

THE FOLLOWING IS EXACTLY WHAT I CHANGED IN EACH YEAR DATASET TO MAKE THEM ALL UNIFORM:

2015:

Deleted Attributes:
region, standard error, dystopia residual
Attribute name changes:
Change Country → Country/Region
Happiness Rank → Rank
Family → social support
Healthy Life Expectancy → Life Expectancy
Trust (Government Corruption) → Perceptions of Corruption

2016:

Deleted Attributes:

region, Lower Confidence Interval, Upper Confidence Interval, Dystopia Residual

Attribute name changes:

Change Country → Country/Region
Happiness Rank → Rank
Family → social support
Healthy Life Expectancy → Life Expectancy
Trust (Government Corruption) → Perceptions of Corruption

2017:

Deleted Attributes:

whisker high, whisker low, dystopia residual

Attribute name changes:

Country → Country/Region
Happiness.Rank → Rank
Happiness.Score → Happiness Score
Economy…GDP.per.Capita. → Economy (GDP Per Capita)
Family → Social Support
Health…Life.Expectancy → Life Expectancy
Trust…Government.Corruption. → Perceptions of Corruption

2018:

Deleted Attributes:

No Deletes

Attribute name changes:

Overall Rank → Rank
Country or region → Country/Region
Score → Happiness Score
GDP per Capita → Economy (GDP Per Capita)
Healthy Life Expectancy → Life Expectancy
Freedom to make life choices → Freedom

2019:

Deleted Attributes:

No Deletes

Attribute name changes:

Overall Rank → Rank
Country or region → Country/Region
Score → Happiness Score
GDP per Capita → Economy (GDP Per Capita)
Healthy Life Expectancy → Life Expectancy
Freedom to make life choices → Freedom

4.2.C.5 STEP 5

For the final step, I saved my new attribute lists. For this I select Save → renamed it to cleanYEAR.arff → select save. Now I have five new .arff files. I did not want to overwrite my first files just in case I want to use them later on.

4.2.C.6 STEP 6 – FINAL ATTRIBUTES

The following are the attributes I will be performing my analysis on. The target attribute is Happiness Score located at the end of this list.

Attribute Name	Description	Attribute Type
Rank	Rank of the country or region based on the Happiness Score	Numeric
Country/Region	Name of the country or region.	Nominal
Economy (GDP Per Capita)	The extent to which GDP contributes to the calculation of the happiness score.	Numeric
Social Support	The extent to which family and friends contributes to the calculation of the happiness score	Numeric
Life Expectancy	The extent to which life expectancy contributed to the calculation of the happiness score	Numeric
Freedom	The extent to which Freedom contributed to the calculation of the happiness score	Numeric
Generosity	The extent to which Generosity contributed to the calculation of the happiness score	Numeric
Perceptions of Corruption	The extent to which Perception of Corruption contributes to happiness score	Numeric
Happiness Score	A metric measured in 2019 by asking sampled people the question: “How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest?”	Numeric

Table 2

4.2.D FUTURE USES

I believe that this data tells us a lot about our society. Governments around the world can use this analyzation to figure out areas of improvement in their countries. If the analysis finds any strong correlations, those areas can be looked into farther. If there are any stand out analyses, I would love to dive deeper into why that country has such a strong correlation between that attribute and happiness. There are a lot of things happening in many countries around the world to play a role in that particular attribute. I expect a strong correlation between economy and happiness in underdeveloped countries. I think that next year's report would be very interesting to analyze. Seeing any jumps in rank of each country from this year to next year would be beneficial in figuring out the long-lasting effects of COVID-19 on happiness for 2020.

4.2.E TECHNIQUE AND TOOL

4.2.E.1 TECHNIQUE

For my analysis I plan on using numeric calculations. The only non-numeric attribute is country name/region and that is not what I am using for my analysis. I will be using linear regression on this project. Linear regression allows me to figure out which factor has the most weight for the happiness scores.

4.2.E.2 WEKA AND FUTURE

For completing my analysis, I plan to use Weka. Weka will help determine which attribute is the strongest correlated. In the future, I think it would be a great analysis to compare the years 2015 and 2019 in Python. I would love to compare countries rank from 2015 to 2019 to see if any countries have moved a substantial amount in rank and look into why that may have occurred. As mentioned before, governments could use data like this to determine if a historical event has affected a countries happiness score.

4.2.F OTHER ASPECTS

As of right now, there are no other aspects of this project I feel is important to communicate. This could change later down the road.

5.2.A ADDITIONAL DATA

I did not collect any additional data for my analysis.

5.2.B GOAL

After my analysis on the data, my intended goal was achieved. My analysis successfully allowed me to see which attribute contributed the most to world happiness for each year. Based on the table below, you can see how each attribute ranked in comparison to the world happiness overall score. To produce these results, I used the correlation ranking filter. It is interesting to see that the rank of the attributes are the same for every year. For each year, the attribute that correlated the most to the world happiness overall score was economy.

	2015	2016	2017	2018	2019
Ranked 1st	Economy (GDP per Capita) 0.781	Economy (GDP per Capita) 0.79	Economy (GDP Per Capita) 0.812	Economy (GDP Per Capita) 0.802	Economy (GDP Per Capita) 0.7939
Ranked 2nd	Social Support 0.741	Social Support 0.739	Social Support 0.753	Social Support 0.746	Social Support 0.7771
Ranked 3rd	Life Expectancy 0.724	Life Expectancy 0.765	Life Expectancy 0.782	Life Expectancy 0.776	Life Expectancy 0.7799
Ranked 4th	Freedom 0.568	Freedom 0.567	Freedom 0.57	Freedom 0.544	Freedom 0.5667
Ranked 5th	Generosity 0.18	Generosity 0.157	Generosity 0.155	Generosity 0.136	Generosity 0.0758
Ranked 6th	Perceptions of Corruption 0.395	Perceptions of Corruption 0.402	Perceptions of Corruption 0.429	Perceptions of Corruption 0.392	Perceptions of Corruption 0.3856

Table 3

5.2.C MACHINE LEARNING STEPS

The first step I took when starting my analysis was to remove the attributes rank and country/region. I did this so they would not affect the scores. Rank was the countries numbered by their happiness score compared to the other countries. Country/Region was just the country or region name. I was not testing either of these attributes compared to the overall happiness score. Neither of these attributes would make sense to test. I did not remove them originally in case I wanted to use them later down the road. I did not use either of these in my analysis for this project. After removing both of these attributes, I had seven attributes left. The attributes left were economy (GDP per Capita), social support, life expectancy, freedom, perception of corruption, generosity and happiness score.

The second step in my analysis was running LinearRegression for each year with the target attribute as happiness score. I wanted to do LinearRegression overall just to confirm my thoughts that these attributes are closely correlated. The following are the outcomes of LinearRegression for each year. As you can see, my thought was correct. These attributes are pretty closely correlated. They each have a correlation coefficient greater that .86.

2015:

```
Correlation coefficient                  0.8668
Mean absolute error                       0.437
Root mean squared error                      0.5694
Relative absolute error                 45.0394 %
Root relative squared error                49.2519 %
Total Number of Instances                  158
```

2016:

```
Correlation coefficient                  0.8781
Mean absolute error                       0.4108
Root mean squared error                      0.5446
Relative absolute error                 42.4393 %
Root relative squared error                47.2836 %
Total Number of Instances                  157
```

2017:

```
Correlation coefficient                  0.886
Mean absolute error                       0.3952
Root mean squared error                      0.523
Relative absolute error                 41.8883 %
Root relative squared error                46.0391 %
Total Number of Instances                  155
```

2018:

```
Correlation coefficient                  0.8693
Mean absolute error                       0.4289
Root mean squared error                      0.5525
Relative absolute error                 45.7606 %
Root relative squared error                48.8791 %
Total Number of Instances                  156
```

2019:

```
Correlation coefficient                  0.8713
Mean absolute error                       0.4283
Root mean squared error                      0.5447
Relative absolute error                 46.158  %
Root relative squared error                48.4405 %
Total Number of Instances                  156
```

The next step in my analysis was to run LinearRegression for each attribute for each year. By completing this step, I could see which attribute was the most correlated to the happiness score. The number I was most interested in was the correlation coefficient. In order to complete this step, I had to temporarily remove every attribute that I was not testing. For example, to get the LinearRegression results for Economy (GDP per Capita), I had to temporarily remove the attributes social support, life expectancy, freedom, generosity and perceptions of corruption. I kept economy (GDP per Capita) and happiness score, then ran LinearRegression. After running LinearRegression, I selected undo to bring back the temporarily removed attributes and completed the steps for the next attribute. I did this for every year. The following are my results:

2015:

Economy (GDP per Capita) and Happiness Score	Correlation coefficient	0.7748
Social Support and Happiness Score	Correlation coefficient	0.7228
Life Expectancy and Happiness Score	Correlation coefficient	0.7167
Freedom and Happiness Score	Correlation coefficient	0.5485
Generosity and Happiness Score	Correlation coefficient	0.08
Perceptions of Corruption and Happiness Score	Correlation coefficient	0.3473

2016:

Economy (GDP per Capita) and Happiness Score	Correlation coefficient	0.784
Social Support and Happiness Score	Correlation coefficient	0.7279
Life Expectancy and Happiness Score	Correlation coefficient	0.7583
Freedom and Happiness Score	Correlation coefficient	0.55
Generosity and Happiness Score	Correlation coefficient	-0.0758
Perceptions of Corruption and Happiness Score	Correlation coefficient	0.3658

2017:

Economy (GDP per Capita) and Happiness Score	Correlation coefficient	0.8061
Social Support and Happiness Score	Correlation coefficient	0.7468
Life Expectancy and Happiness Score	Correlation coefficient	0.7744
Freedom and Happiness Score	Correlation coefficient	0.5524
Generosity and Happiness Score	Correlation coefficient	-0.02
Perceptions of Corruption and Happiness Score	Correlation coefficient	0.3937

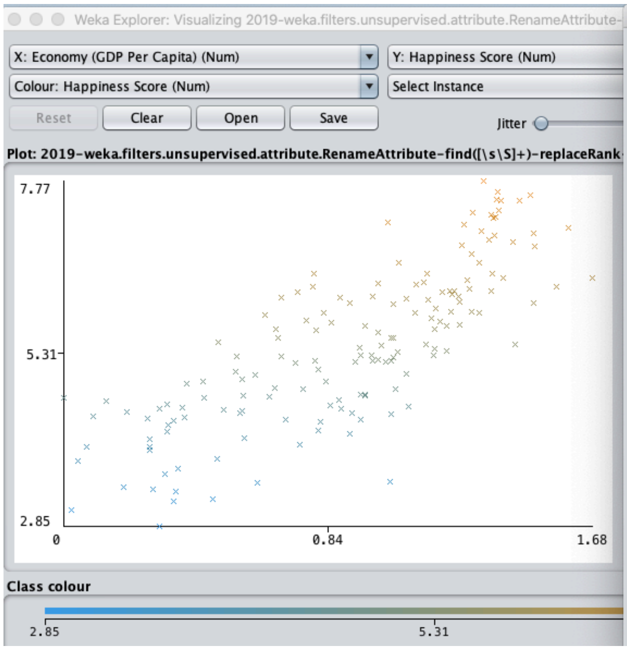
2018:

Economy (GDP per Capita) and Happiness Score	Correlation coefficient	0.7959
Social Support and Happiness Score	Correlation coefficient	0.739
Life Expectancy and Happiness Score	Correlation coefficient	0.7679
Freedom and Happiness Score	Correlation coefficient	0.5279
Generosity and Happiness Score	Correlation coefficient	-0.0652
Perceptions of Corruption and Happiness Score	Correlation coefficient	0.3537

2019:

Economy (GDP per Capita) and Happiness Score	Correlation coefficient	0.7852
Social Support and Happiness Score	Correlation coefficient	0.772
Life Expectancy and Happiness Score	Correlation coefficient	0.7717
Freedom and Happiness Score	Correlation coefficient	0.5497
Generosity and Happiness Score	Correlation coefficient	-0.3413
Perceptions of Corruption and Happiness Score	Correlation coefficient	0.3461

After looking at these results, two things stood out to me. The first is that all of the correlation coefficients for the attributes are very close throughout the years. Economy (GDP per capita) correlation coefficients are all around 0.79. Social support correlation coefficient scores are right around 0.75. Life expectancy correlation coefficient scores are around 0.74. Freedom correlation coefficients are right around 0.55. Generosity correlation coefficients is the attribute that ranges the most. These scores range from -0.3 to 0.08. The correlation coefficients for perceptions of corruption are right around 0.36. When looking at these overall, we can see that there is not much change in the five-year span. The second stand out to me is that generosity resulted in negatives in four out of the five years. Since there are more negative scores than positive, this means that generosity and the happiness score are a little opposite of each other. These five values also span over 0, meaning there is no correlation between them. I found this very interesting because economy (GPD per capita), social support and life expectancy values are all over 0.7 for each year, meaning that have a pretty high correlation with the happiness score. When looking back at the questions of the two lowest values, I can understand why these two values are the least correlated with the happiness score. For generosity, people are asked, “Have you donated money to a charity in the past month?” A lot of people do not have money to donate to others, especially in third world countries so I understand why this value received the lowest score. The second lowest score is perceptions of corruption. At first this result surprised me, but after reviewing the question I understood why. The question is “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?” I believe a lot of people would not give this a high rating, even if it were true due to fear. For each of the years, I looked at the visualization for the best performing attribute and the worst. The following visualization is for the year 2019. The left visualization is for economy (GDP per capita), as you can see, there is a pretty good correlation of LinearRegression. The right visualization is for generosity. As you can see, there is not a strong correlation of LinearRegression and it produced a negative number.



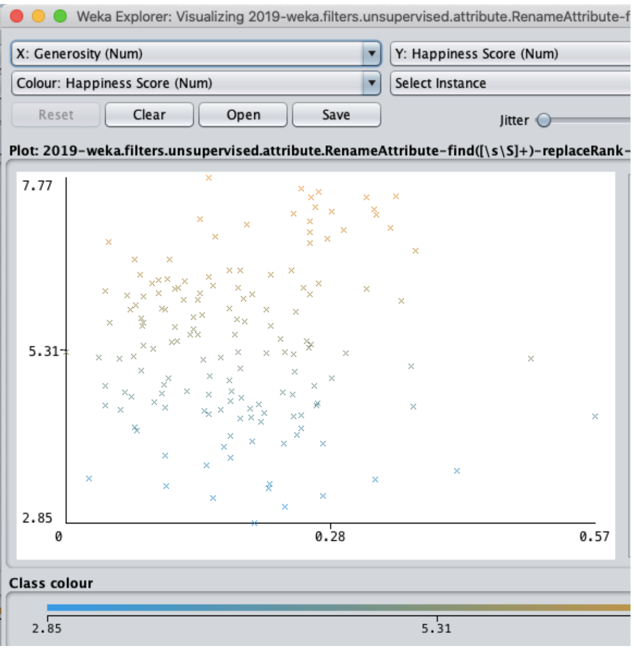
Weka Explorer: Visualizing 2019-weka.filters.unsupervised.attribute.RenameAttribute-

X: Economy (GDP Per Capita) (Num)

Y: Happiness Score (Num)

Colour: Happiness Score (Num)

Plot: 2019-weka.filters.unsupervised.attribute.RenameAttribute-find([\s\S]+)-replaceRank



Weka Explorer: Visualizing 2019-weka.filters.unsupervised.attribute.RenameAttribute-f

X: Generosity (Num)

Y: Happiness Score (Num)

Colour: Happiness Score (Num)

Plot: 2019-weka.filters.unsupervised.attribute.RenameAttribute-find([\s\S]+)-replaceRank-

5.2.B.1 PROBLEMS

Honestly, I did not really run into any problems. LinearRegression really performed how I expected.

5.2.D OTHER TECHNIQUE

A technique that was not used in assignments 1-3, but I used, was CorrelationAttributeEval. This evaluation technique ranks the attributes in how closely correlated they are. To use this evaluation, it is under the “select attribute” tab. The results of this evaluator is in the table 3 above where I had the attributes ranked, but I will break down each one below:

2015:

```
Attribute Evaluator (supervised, Class (numeric): 7 Happiness Score):
        Correlation Ranking Filter
Ranked attributes:
 0.781  1 Economy (GDP per Capita)
 0.741  2 Social Support
 0.724  3 Life Expectancy
 0.568  4 Freedom
 0.395  6 Perceptions of Corruption
 0.18    5 Generosity
        Selected attributes: 1,2,3,4,6,5 : 6
```

2016:

```
Attribute Evaluator (supervised, Class (numeric): 7 Happiness Score):
        Correlation Ranking Filter
Ranked attributes:
 0.79    1 Economy (GDP per Capita)
 0.765  3 Life Expectancy
 0.739  2 Social Support
 0.567  4 Freedom
 0.402  6 Perceptions of Corruption
 0.157  5 Generosity
        Selected attributes: 1,3,2,4,6,5 : 6
```

2017:

```
Attribute Evaluator (supervised, Class (numeric): 7 Happiness Score):
        Correlation Ranking Filter
Ranked attributes:
 0.812  1 Economy (GDP Per Capita)
 0.782  3 Life Expectancy
 0.753  2 Social Support
 0.57    4 Freedom
 0.429  6 Perceptions of Corruption
 0.155  5 Generosity
```

```
        Selected attributes: 1,2,3,4,6,5 : 6
```

2018:

```
Attribute Evaluator (supervised, Class (numeric): 7 Happiness Score):
        Correlation Ranking Filter
Ranked attributes:
 0.802  1 Economy (GDP Per Capita)
 0.776  3 Life Expectancy
 0.746  2 Social Support
 0.544  4 Freedom
 0.392  6 Perceptions of Corruption
 0.136  5 Generosity
        Selected attributes: 1,3,2,4,6,5 : 6
```

2019:

```
Attribute Evaluator (supervised, Class (numeric): 7 Happiness Score):
        Correlation Ranking Filter
Ranked attributes:
 0.7939  1 Economy (GDP Per Capita)
 0.7799  3 Life Expectancy
 0.7771  2 Social Support
 0.5667  4 Freedom
 0.3856  6 Perceptions of Corruption
 0.0758  5 Generosity
        Selected attributes: 1,3,2,4,6,5 : 6
```

When comparing these results of the attribute ranking and LinearRegression, the values are very close, which really solidifies my results.

5.2.E USE OF ANALYSIS

I believe that these results would be very helpful to countries and governments. I think happiness has a major impact on everyone's life. Governments could take these findings and question themselves on how to make the results even better. The attribute that was correlated the most to happiness was economy. Governments and countries could ensure their economy is good in order to help with happiness. If a country or region has unhappy people, there are many negative effects it could have. By having the attributes ranked, a country or government could go through the findings and figure out ways to improve their scores. I think these scores and results would be beneficial to any country/region.

5.2.F OTHER ASPECTS

There are no other aspects I feel are important to communicate at this time.