A visualization and analysis report of Happiness Report and Terror Attack in 2015, 2016, and 2017

Gen Wang
28424379
Gwan0010@student.monash.edu
23 Apr. 19

1 Introduction:

This report will try to analysis and visualise World Happiness Score by using features provided by World Happiness Report and World terror attack report during 2015,2016, and 2017. Various visualization, statistical, and machine learning methods will be used to display, analyse and predict the datasets. The main content of this report consists of Data wrangling, Data check, Data exploration, conclusion, and reflection.

1.1 Problem description:

World Happiness Score is an index published by the United Nation annually. People had heard of the happiness score and the ranking according to the countries. However, people seldomly aware of the factors that contribute to the happiness score and the method used to calculate the score.

1.2 Questions:

The report will mainly try to answer the following question to provide insight into the happiness score.

1. How does the terrorist attack affect happiness index within different countries?
2. What is the distribution of the happiness index in different continents?
3. What is the relationship between happiness index with other feature provided in the happiness report?
4. Can we predict the happiness index by given the other features in the happiness report and terrorist attack? If we can, find out the relationship.

1.3 Motivation:

Generally, the motivation of this report is to assist people to understand the relationship between terrorist attack and happiness score, the distribution of the happiness index in a different continent, the formation of the happiness score, and the relationship between happiness score and other factors included in the world happiness report. By trying to predict the happiness score, we will be able to gain further understanding of the calculation of the happiness score.

2 Data wrangling:

The following part of the report will have a brief description of the dataset, the available links, and the data wrangling process before analysis and visualising the data.

2.1 Data overview:

There are four datasets to support this report. They are the Global Terrorism Database; The World Happiness Report in 2015, 2016, and 2017.

The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on international terrorist incidents that have occurred during this time period. We will use group by the count to return the number of the terrorist attack in a specific country in a different year.

The World Happiness Report is a landmark survey of the state of global happiness. The scores are from nationally representative samples for the years 2015 – 2017, and use the Gallup weights to make the estimates representative. The main features will be used in this report are listed as follow:

1. **Happiness Score:** A metric measured by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."
2. **Country：** Name of the country.
3. **Region：** Region the country belongs to.
4. **Happiness Rank：** Rank of the country based on the Happiness Score.
5. **Economy (GDP per Capita):** The extent to which GDP contributes to the calculation of the Happiness Score.
6. **Family:** The extent to which Family contributes to the calculation of the Happiness Score
7. **Health (Life Expectancy):** The extent to which Life expectancy contributed to the calculation of the Happiness Score
8. **Freedom:** The extent to which Freedom contributed to the calculation of the Happiness Score.
9. **Trust (Government Corruption):** The extent to which Perception of Corruption contributes to Happiness Score.
10. **Generosity:** The extent to which Generosity contributed to the calculation of the Happiness Score.
11. **Dystopia Residual：** The extent to which Dystopia Residual contributed to the calculation of the Happiness Score.
12. **Attack Count:** The number of terrorist attack happen in that country.

The datasets are available in the following links.

The Global Terrorism Database: https://www.kaggle.com/START-UMD/gtd

The World Happiness Report: https://www.kaggle.com/unsdsn/world-happiness

2.2 Wrangling process:

The process is executed under the Python environment. Pandas is the wrangling library in this project.

2.2.1 Wrangling steps:

1. Read data to dataframe, insert a column of the year to identify the year of the rows. and unify columns in the happiness report for 3 years.
2. Concatenate dataframe of 2015, 2016, and 2016.
3. Fill in the Nan value in region column from 2017 report, by correlating the country name with region provided in the 2015 and 2016 report.
4. Change the column name and Group the terrorist attack dataframe by using ‘Country’ and ‘Year’ as the key to sum the number of the terrorist attack in specific country and year. The detail is shown in figure 4.

5. Check the data type for the merging key. Unify the data type with the key, merge two dataframe, and write to CSV file.

3 Data checking:

The data checking process includes checking the error and nan value in the dataset. Reformat the dataframe for merging.

The data check error and problem include in the data checking process are listed as follow:

1. The Nan value in the dataset. (Process includes in step 3 of wrangling steps)
2. An uneven number of the column between different datasets. (Process includes in step 1 of wrangling steps)
3. Disunity column names. (Process includes in step 1, 4 of wrangling steps)
4. Disunity data type. (Process includes in step 5 of wrangling steps)
5. Synonyms. (Process includes in step 1, 3 of wrangling steps)
6. Unnecessary rows and columns in the dataset. (Process includes in step 4 of wrangling steps)

The correction method to reformat the dataset. (The reformatting and correcting process correspond to the previous list of errors and problems)

1. Using corresponding values in another dataset as a reference to find in the nan value. (Process includes in step 3 of wrangling)
2. Get different column by using the function defined in step 1. Drop the columns in each dataframe to unify.
3. Get different column names by using the function defined in step 1. Rename the column to prepare to merge.
4. Check the data type in different columns from different dataframe. Unify the data type for the column and use it to operate merging.
5. Use the function defined in step 1 to find the synonyms and change synonyms manually.
6. Select the rows that are meaningful for both datasets to perform merging.

The tools used in the previous process is a Python library “Pandas”. According to the documentation of “Pandas”, it is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools.

4 Data Exploration:

In this section of the report, we will discuss the questions raised in the introduction Questions section. We will try to answer four questions. They are

1. How does a terrorist attack affect happiness index within a different country?
2. What is the distribution of the happiness index in a different continent?
3. What is the relationship between happiness index with other feature provided in the happiness report?
4. Can we predict the happiness index by given the other features in the happiness report and terrorist attack? If we can, find out the relationship.

4.1 How does the terrorist attack affect happiness index within different countries?

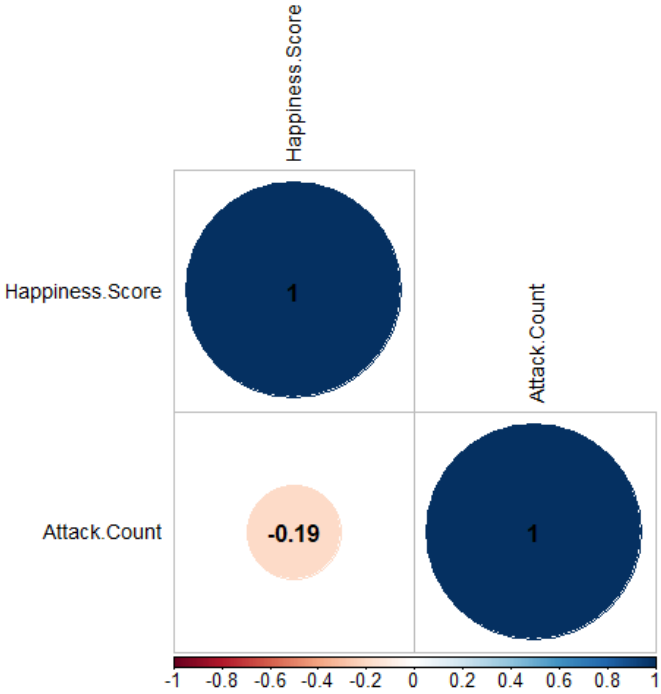


Figure 1

The correlation table in figure 1 shows the correlation coefficient between the terrorist attack and happiness score. According to the table, there is a negative correlation of **(-0.19)** between terrorist attack and happiness score, which means the higher number of the terrorist attack in a country will reduce the happiness score. However, the score is relatively small compared with other factors such as the Economy. GDP per Capita, and Health. Life Expectancy. In another word, the terrorist attack does not have a strong influence on the happiness score in different countries.

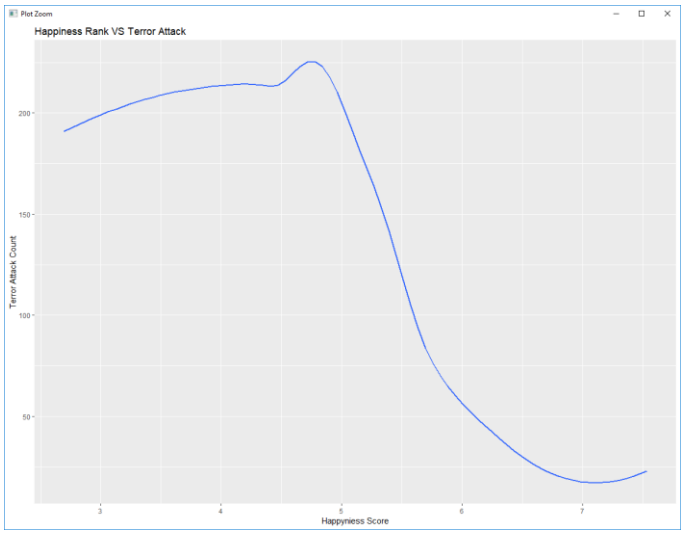


Figure 2

According to figure 2, we can clearly see that the increasing number of terrorist attack led to a significant decrease in the happiness score among countries with a score from **4.7 to 7**. However, the direction changes at happiness score from **0 to 4.7 and 7 to higher**, the **increase** of terrorist attack lead to a **higher** happiness score. This also proves that the small correlation between terrorist attack and happiness score.

4.2 What is the distribution of the happiness index in different continents?

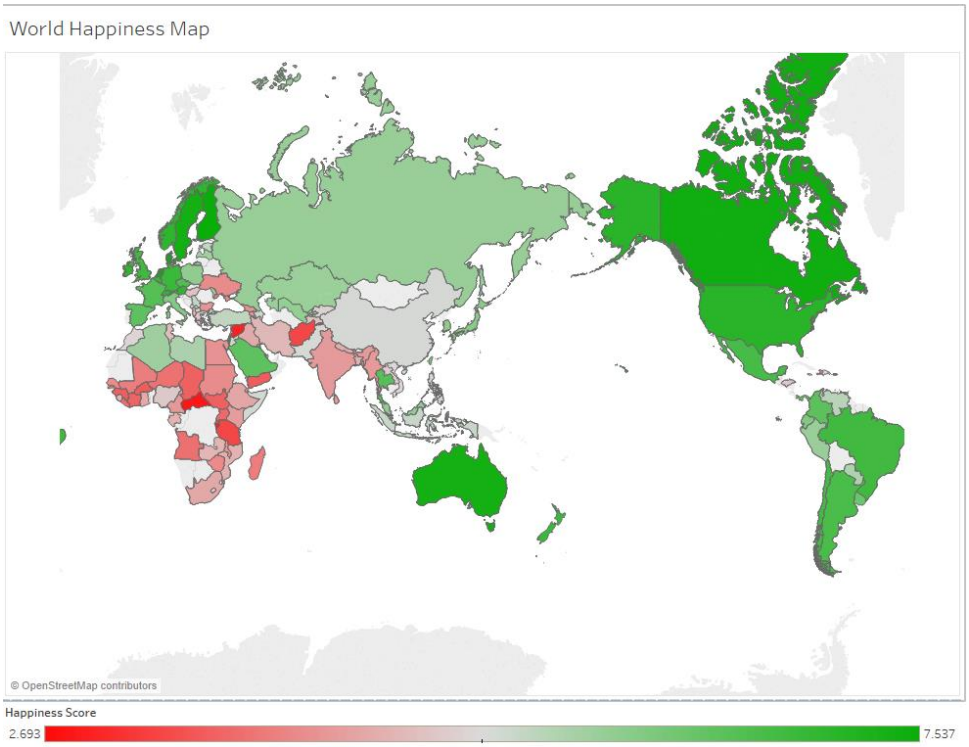


Figure 3

In figure 3, we will have a general picture of the distribution of the happiness score in the world. We can observe that most of the **unhappy countries** are located in **Africa, Middle East and Eastern Europe**.

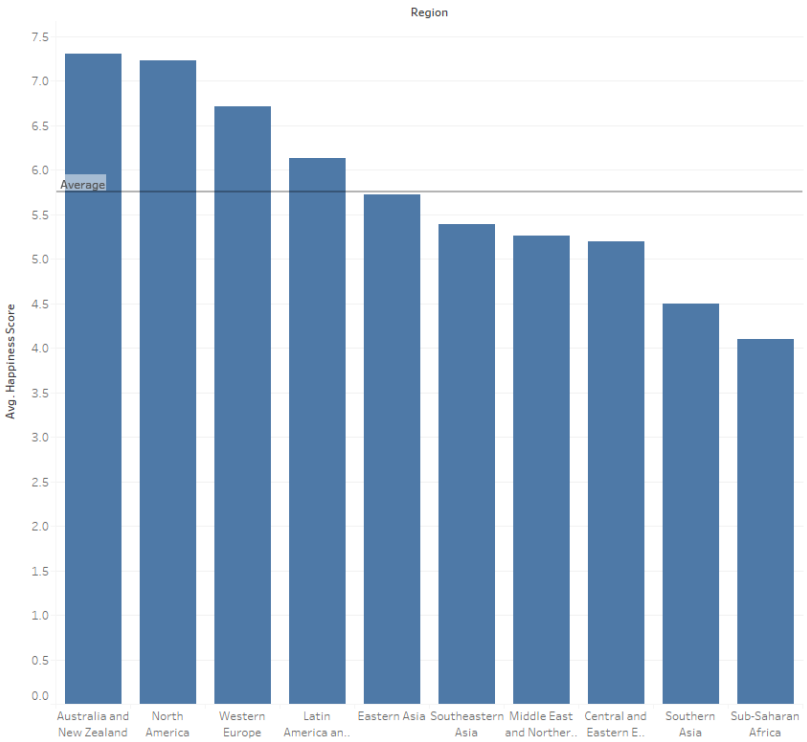


Figure 4

In figure 4, the “Average” line indicates the world average happiness score, and each of the bars shows the average happiness score in that region/continent. We can find that only Australia and New Zealand, North America, Western Europe, and Latin America and Caribbean exceed the average happiness score. Sub-Saharan Africa region has the lowest happiness score among all.

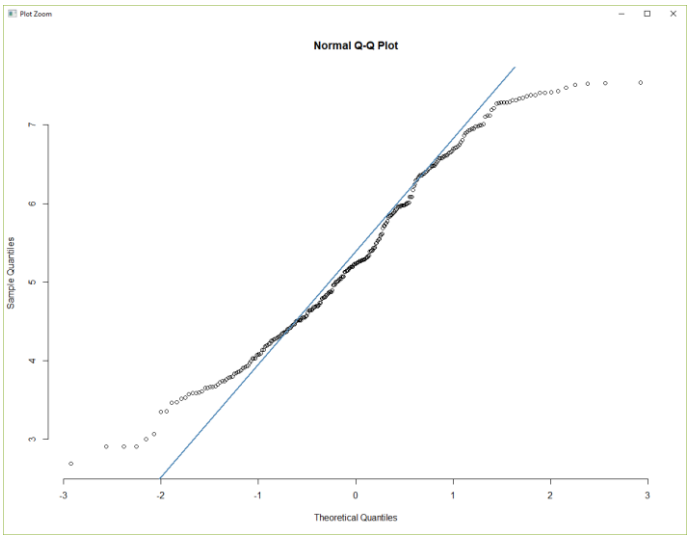


Figure 5

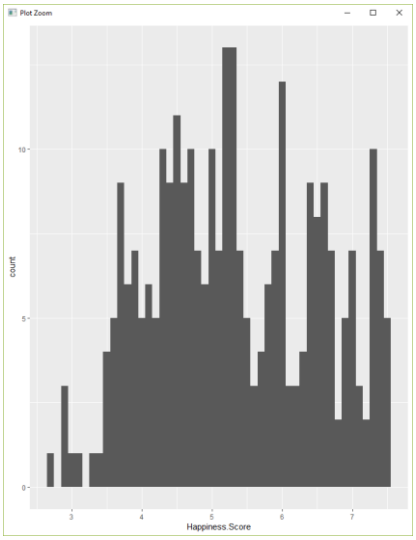


Figure 6

In figure 5 and figure 6, we can notice that the distribution of the happiness score has a **fat tail** in both ends.

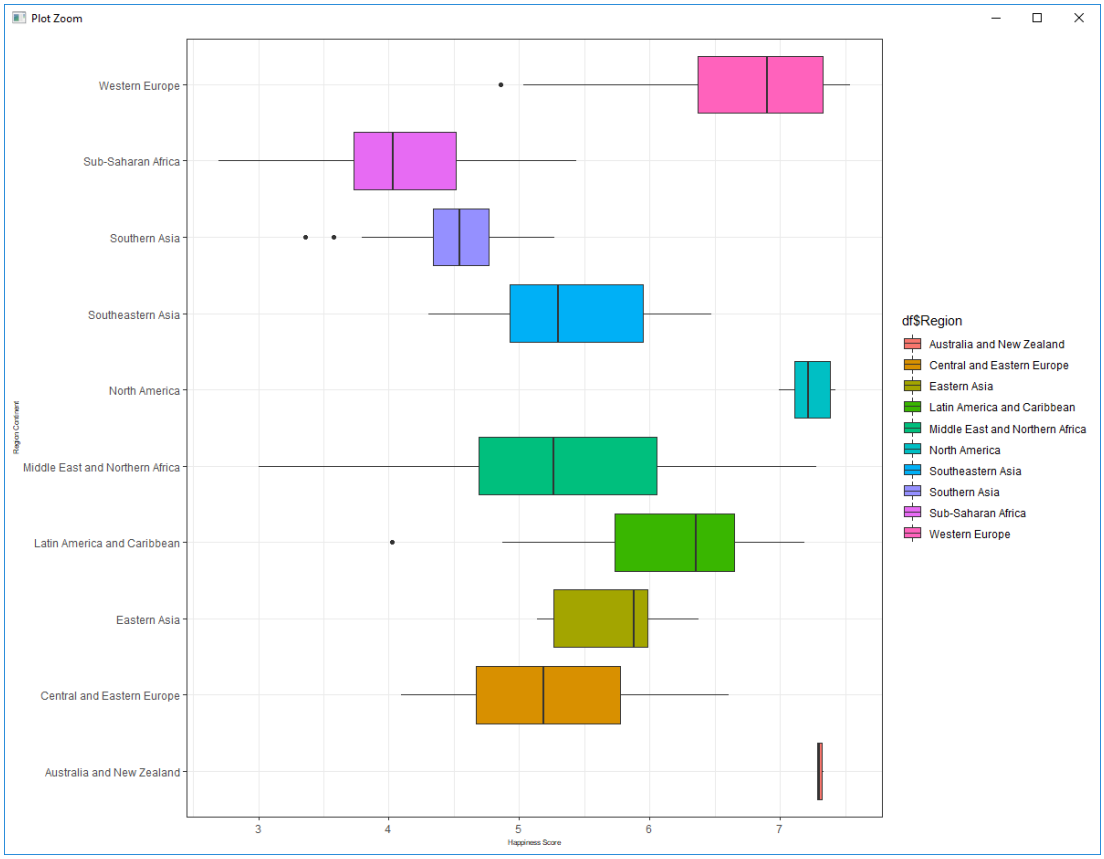


Figure 7

It shows in figure 7 that the most evenly distributed (approximately normally distributed) happiness score occurs in North America, and Central and Eastern Europe, Australia and New Zealand, and Southern Asia. This means that most of the countries in these regions/continents are sharing **similar median and average happiness score**. Western Europe, Latin American and Caribbean, and East Asia is left-skewed, meaning that there are more countries in this region/ continent have **higher happiness score compare with regional/continent average**. Sub-Saharan Africa, Middle East and North Africa, and South-eastern Asia are slightly right-skewed. This means that there are more countries in this region/ continent have **lower happiness score compare with regional/continent average**.

Another interesting fact that we find out in this boxplot is that there is no conceptually developed region has a right-skewed distribution.

4.3 What is the relationship between happiness index with other feature provided in the happiness report?

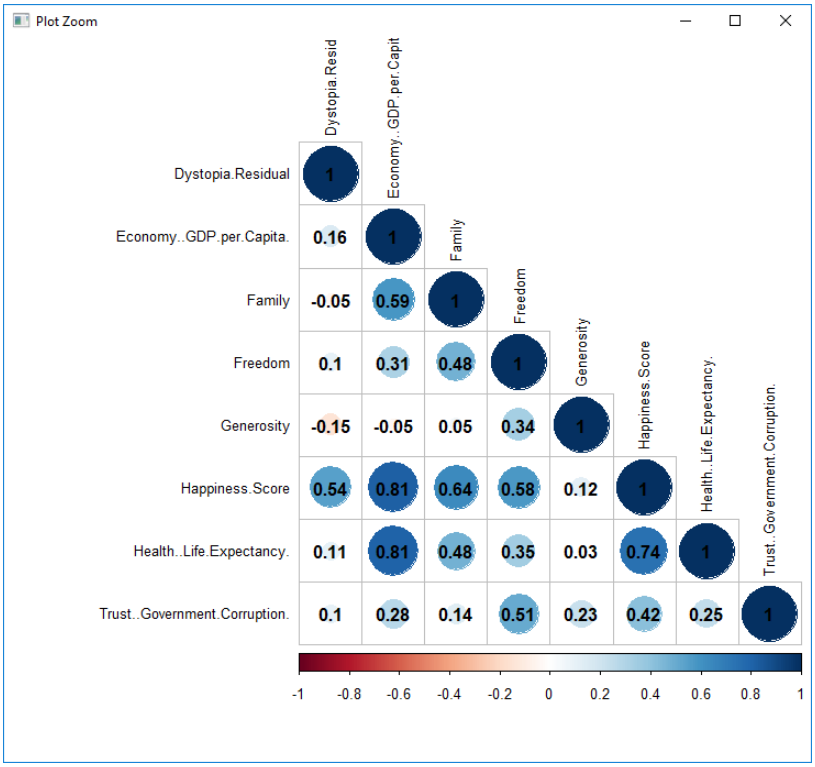


Figure 8

We can conclude in figure 8 that there is a strong positive relationship between happiness score and Economy GDP per Capita, Health Life Expectancy, freedom, Dystopia Residual, and Family. It means that it is highly likely that the happiness score can be generated by using these features.

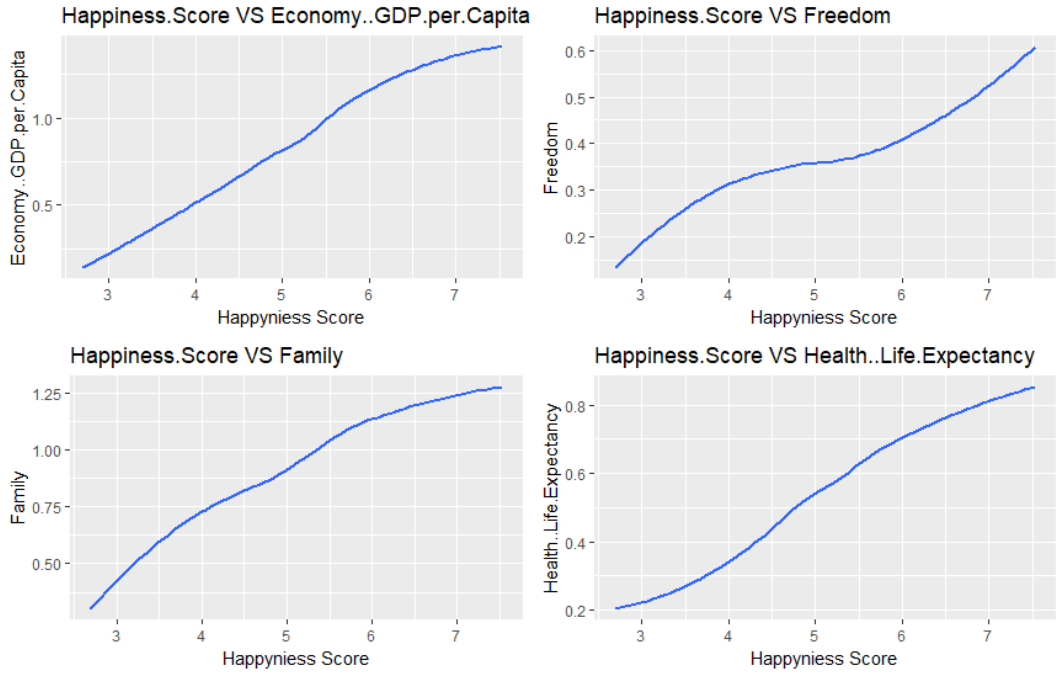
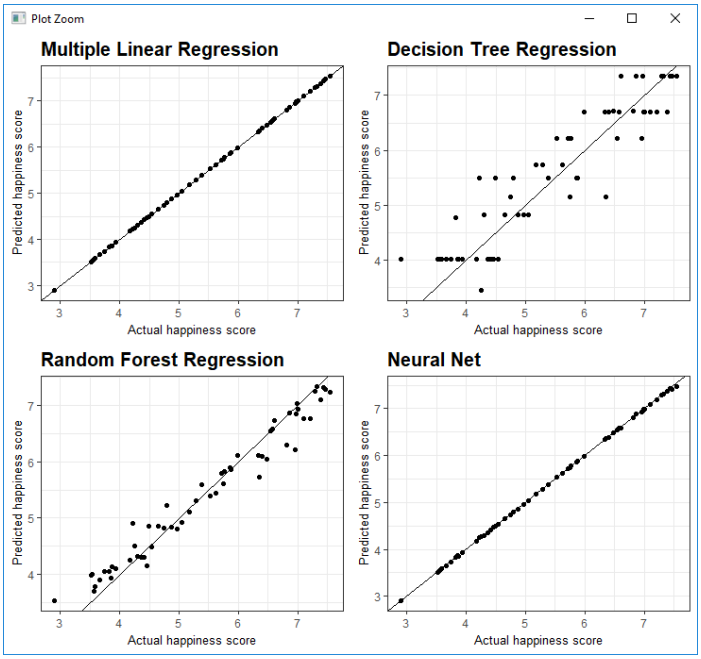


Figure 9

By observing figure 9, we can define the positive correlation between happiness score with the top four features that correlate with happiness score defined in figure 8.
In the next question, we will try to predict the happiness score by using these features.

4.4 Can we predict the happiness index by given the other features in the happiness report and terrorist attack? If we can, find out the relationship.

We will try to use varies method to predict the happiness score using the feature mention in the previous question. In order to do that, we will randomly split the dataset into training and testing with the proportion of 80:20. We will try to use four common machine learning methods (Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, and Neural Net) to predict the happiness score.



According to figure 10, Multiple Linear Regression and Neural Net are the best models for predicting the happiness score with a feature provided in the happiness report.

Figure 10

```
> coefficients(lr)
                  (Intercept)             Dystopia.Residual  Economy..GDP.per.Capita.                          Family
                 7.177873e-05                  9.999767e-01                  1.000091e+00                  1.000050e+00
                      Freedom                    Generosity  Health..Life.Expectancy.  Trust..Government.Corruption.
                 9.999721e-01                  1.000143e+00                  9.997031e-01                  9.998834e-01
```

Figure 11

According to the model in Multiple Linear Regression, we can further write down the formula to calculate the happiness score by retrieving the coefficients of the Multiple Linear Regression in figure 11.
That is:
0.00007177873 + 0.9999767*Dystopia. Residual + Economy.GDP.Capita +
Family + 0.9999721*Freedom + 1.000143*Generosity + 0.9997031*Health.Life. Expectancy+ 0.9998834*Trust Government Corruption

5 Conclusion:

To conclude, by exploring the datasets, we find out that the terrorist attack does not have a significant influence on the happiness score. Generally, the happiness score is not normally distributed in different regions/continents. There are four regions have higher average happiness score than the worldwide average, and all of the regions are in Western Europe, America, and Australia and New Zealand. Happiness score is highly influenced by Economy GDP per Capita, Health Life Expectancy, freedom, Dystopia Residual, and Family. We find out the formula to calculate the happiness score in the last question we propose. The report has fully answered all the question proposed initially.

6 Reflection:

From this report, I had learnt the appropriate procedure of data analysis and visualisation. It includes a data check, data wrangling, and data exploration. By identifying the pattern of the dataset with the statistical method, we can generate a model or formula to predict the result with similar inputs.

If I could have finished this project differently, I would like to include more features in the prediction to find out the formula to calculate the indexing feature in the happiness report, so that I could have the chance to use raw data to predict the happiness score.

References:

Deselecting a column by name. (1, 7). Retrieved from https://stackoverflow.com/questions/9805507/deselecting-a-column-by-name

Global Terrorism Database. (n.d.). Retrieved from https://www.kaggle.com/START-UMD/gtd

How to split data into training/testing sets using sample function. (10, 5). Retrieved from https://stackoverflow.com/questions/17200114/how-to-split-data-into-training-testing-sets-using-sample-function

Plotting two variables as lines using ggplot2 on the same graph. (7, 8). Retrieved from https://stackoverflow.com/questions/3777174/plotting-two-variables-as-lines-using-ggplot2-on-the-same-graph

Python Data Analysis Library — pandas: Python Data Analysis Library. (n.d.). Retrieved from https://pandas.pydata.org/

Python Pandas: group by in group by and average? (11, 3). Retrieved from https://stackoverflow.com/questions/30328646/python-pandas-group-by-in-group-by-and-average

Rotating and spacing axis labels in ggplot2. (8, 9). Retrieved from https://stackoverflow.com/questions/1330989/rotating-and-spacing-axis-labels-in-ggplot2

World Happiness Report. (n.d.). Retrieved from https://www.kaggle.com/unsdsn/world-happiness