The World Happiness Report 2015

CS544 Final Project by Sabrina Minaya
Instructor: Heather Shappell
Facilitor: Teresa Filshtein
December 17, 2020

1.Objective

To study the objective factors that influence the level of happiness around the world such as, Economy (GDP growth), Family, government, health etc.

2. Preparation of Data

2.1 Import the data set into R:

The World Happiness Report contains the countries or regions that rank the highest or lowest in overall happiness. This report also identifies the factors that contribute to level of happiness in each region or country. The csv file "2015.csv" is downloaded from Kaggle website ("https://www.kaggle.com/unsdsn/world-happiness?select=2015.csv"). R function "read.csv is used to load the file into data frame.

```
> world.happiness<- read.csv(file = "2015.csv",stringsAsFactors = FALSE)
```

The data source consists of 12 variables, such as:

Country (or Region): is ranking by the highest to lowest level of happiness.

Happiness Score: Analyze the characteristics that influence the level of happiness. According to "World Happiness Report 2015", "on a scale running from 0 to 10, people in over 150 countries, surveyed by Gallup over the period 2012-2015, reveal an average score of 5.1(out of 10)" (World Happiness Report 2016, 1).

Economy (GDP per capita): Measure well-being, economic development, and go further, to contemplate the happiness of the population.

Family: Consist of the social support (or being able to count on someone in difficult times.

Health (Life Expectancy): Is defined as the number of free years of illness that, on average, a newborn can expect to live.

Freedom: Study if population is happy or unhappy with their freedom to decide what they want to do with their life.

Equally important, Generosity, Trust (Government) and Dystocia are other factors that influence the level of happiness in the report.

2.2 Preprocessing of dataset

The imported dataset was cleaned before starting the analysis as following:

- Created a list of all variables, variables labels and variable codes.

- Decided which variables are crucial to the analysis such as Country, Happiness Score, Economy (GDP per capita), family, and Health (Life Expectancy)

- Make a new variable categorical variable “Continent” to reduce large amount of content into manageable data.

- Imported different data visualization packages to represent a better data graphically in order to bring clarity during analysis and to communicate.

- Combined categorical and numeric data to improve the descriptive evaluation of the analysis.

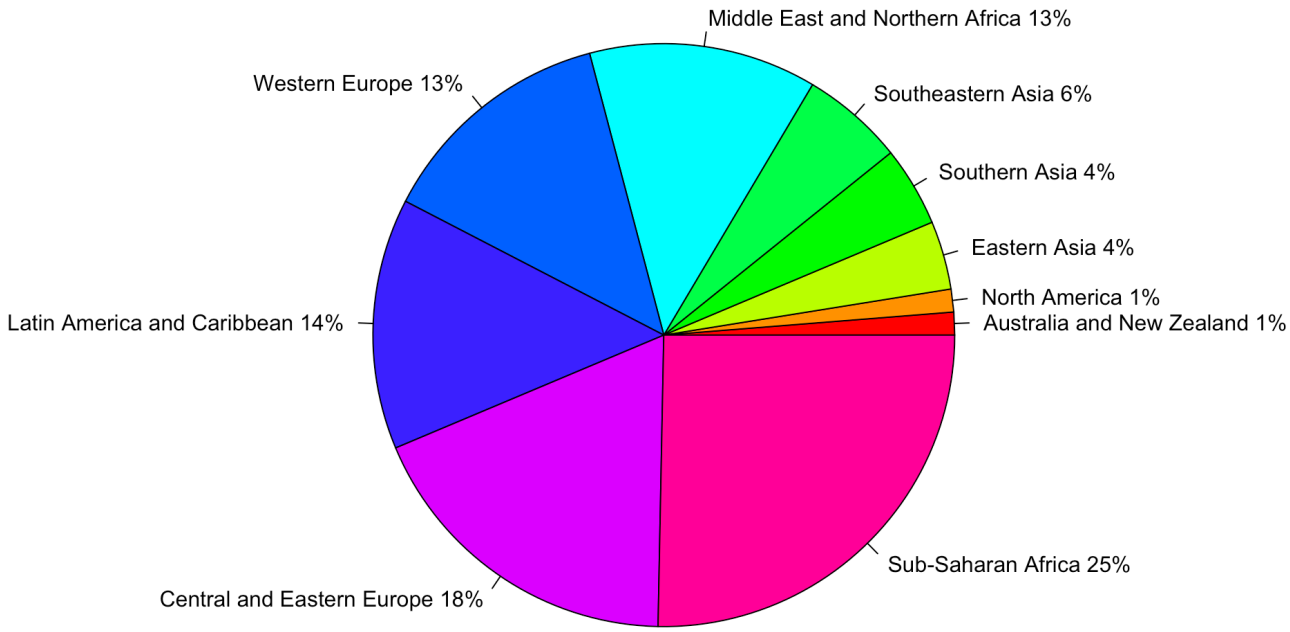
- Use the World Happiness Report 2015 as guideline to write-up the analysis https://worldhappiness.report/ed/2019/changing-world-happiness/

3. Analysis of Data

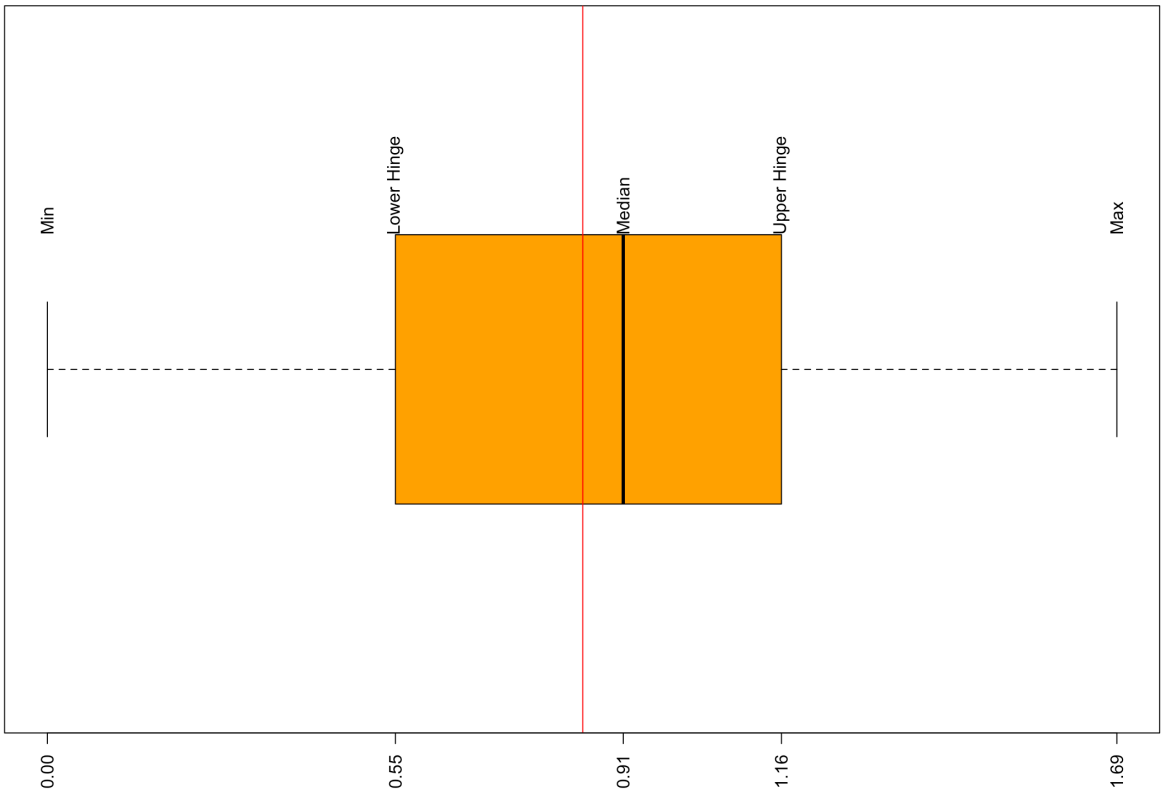
3.1 Regional Analysis: Number of Countries per Region

The general sense of happiness and satisfaction of people has hardly increased in the past years. This one of the data collected by the “World Happiness Report 2015”, according which, although in poor countries the economic level is crucial element for happiness, in developed countries the equivalence between money and well-being is not clear. To illustrate, Pie chart below shows that Sub-Saharan Africa consist of the greatest percentage of countries with 25%. This is followed by Central and Eastern European with 18% and Latin America and Caribbean with 14%. Then, these are closely followed by Western Europe and Middle East and Northern Africa with an equal percentage of 13%. Asia, on the other hand, comprise to only 14% of the overall distribution of countries. The lowest percentage of countries per region is attributed to

North and Australia and New Zealand with 1% apiece. From the graph, Sub Saharan Africa and Central and Eastern Europe comprise almost half of the percentages. Asia accounts to only 14% while North America and Australia and New Zealand account to only 2%. The statistics demonstrate that the level of happiness in some countries or region not only depends on the economic factor. There are other conditioning factors such as family and health which increase the level of happiness in the Region.

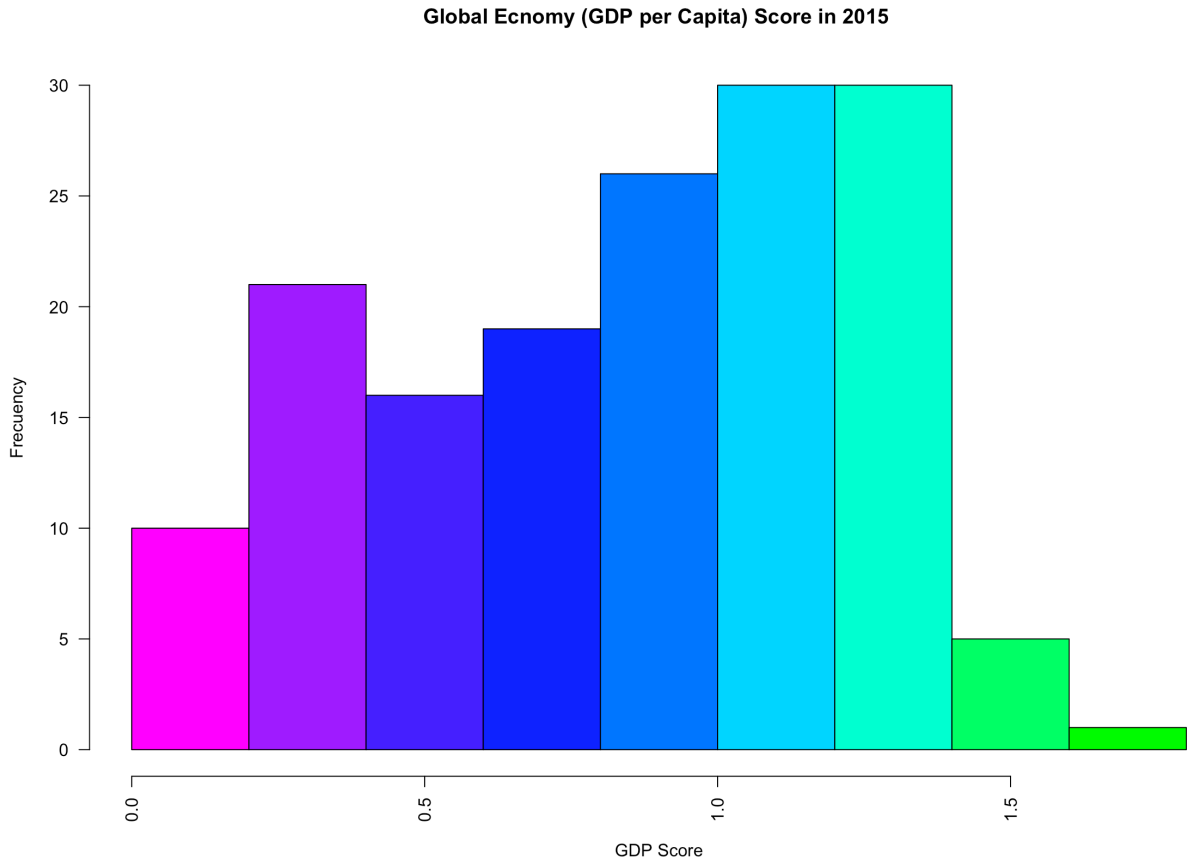


3.2 Global Economy (GDP per Capita) Summary



To explain the levels of happiness of regions, the Report takes a detailed look about what role does the factor of economy play in determining the happiness of a country. According to the "World Happiness Report 2015", Economy (GDP per Capita) "refers to the value of production within a country without considering the capital depreciation". The value of GDP per capita are obtained from the "World Bank's Global Economic Prospects". As illustration, the boxplot shows the five-number summary of the GDP score. There are no outliers present in the box- plot as local minima for the plot is 0.0 which will be equal to original min value in the dataset that is

0 and local maxima for the plot is 1.6900 which is equal to original max value. Therefore, the points will be plot inside that range. Hence, outliers are missing in the plot or feature. The Median value is 0.9105 and Mean value is round 0.87 which means that median is greater than mean. GDP feature is left skewed that is most of the values are present on the left side of the median which is closed to Upper hinge or 75 percentiles. Further, the value for 25 percentiles is 0.5460 which means that 25% of the values are lower than 0.5460 and median value is 0.9105. The observation conclude that outliers are absent in the GDP feature, and the skewness is present in the GPD feature (left Skewed).



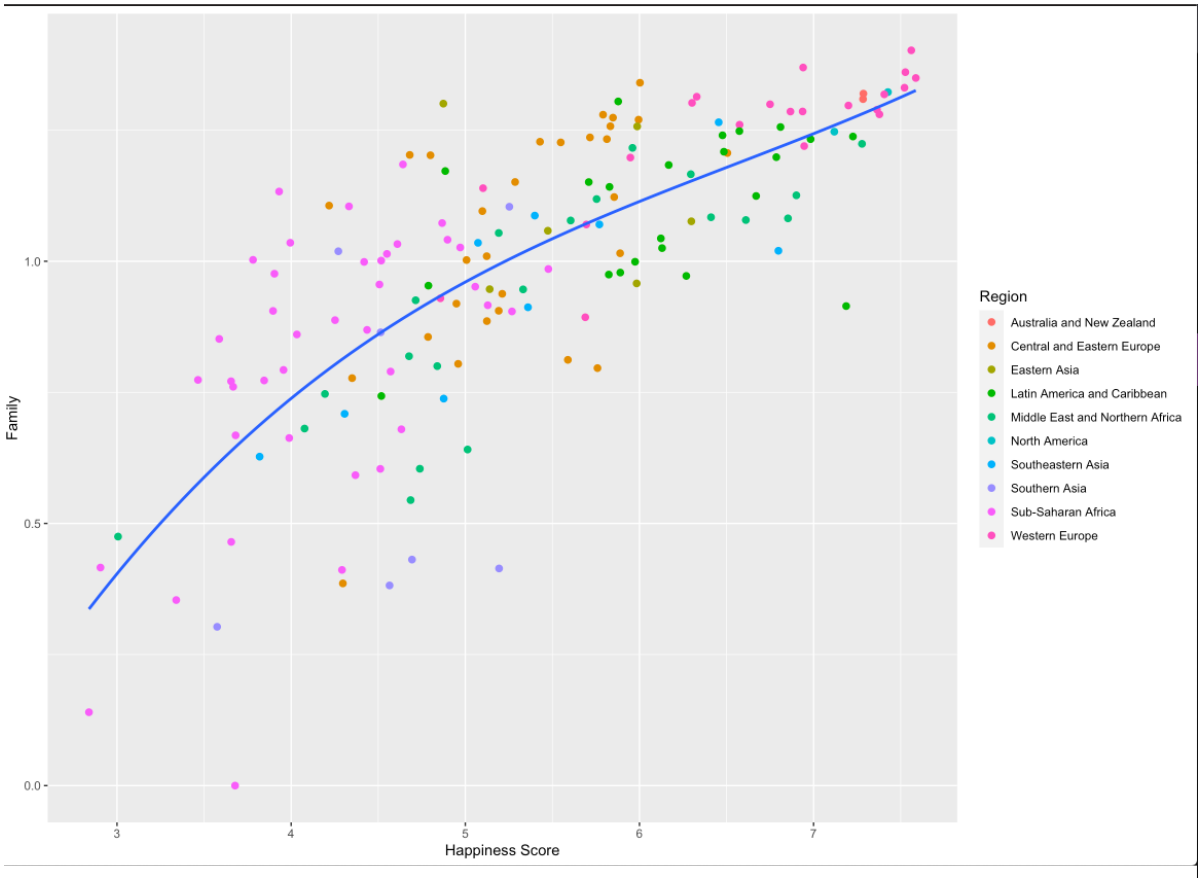
The histogram above shows that major group of people who contributed in high GDP growth is between (0.7 to 1.5). Likewise, highest GDP contribution frequency is 30 and the lowest GDP contribution frequency is 2. The graph is normally distributed, but the data points are more influence in the left skewed.

3.3 Top Factors that Influence the Level of Happiness

The correlation plot above shows that the Economic GDP score has the highest impact to happiness score. Even though GDP is considered as one of the most important factors for the level of happiness in the report of 2015, wealth is not a sufficient or necessary condition to feel

happy. For example, Family score is super closed to GDP score since many people consider that social support from family or friends in the event of problems can help to increase the level of happiness. One bigger concern is an important factor as Trust has the lowest scores of all conditions looked at. When population of a country no trust their government, they feel deprived of their rights, and aren't able to take the life choices as they wish, which is demonstrate in the correlation between low trust score and low freedom score.

3.4 Family & Happiness Score by Region

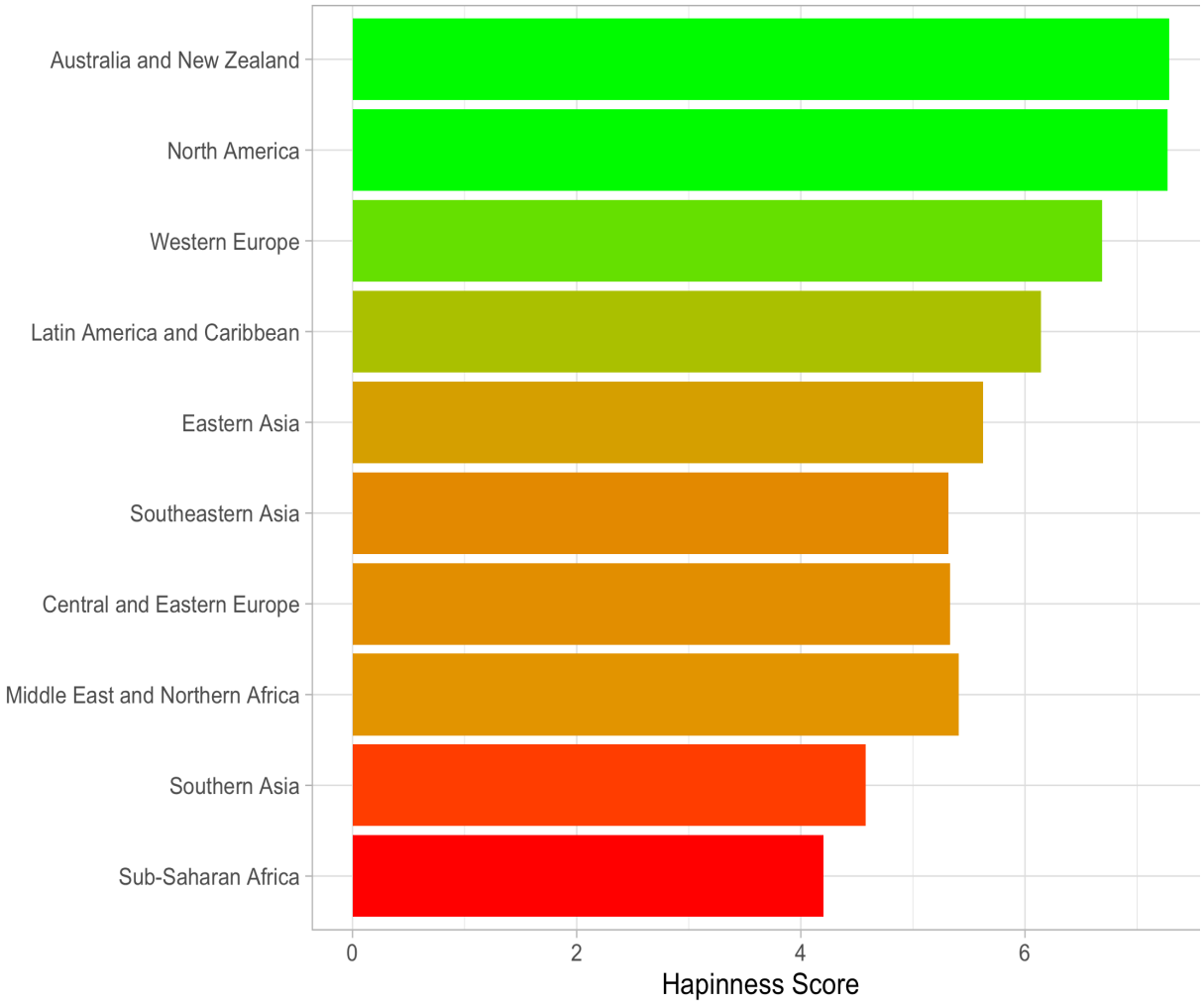


The scatterplot above shows an uphill pattern that is applied to all the regions. This indicates a positive relationship between happiness score and family. The plot also shows that Western Europe region has the Maximun family score while Sub-Saharan Africa has the minimun score.

This prove that people who have supportive and loving relationships in their life tend to have a higher level of happiness.

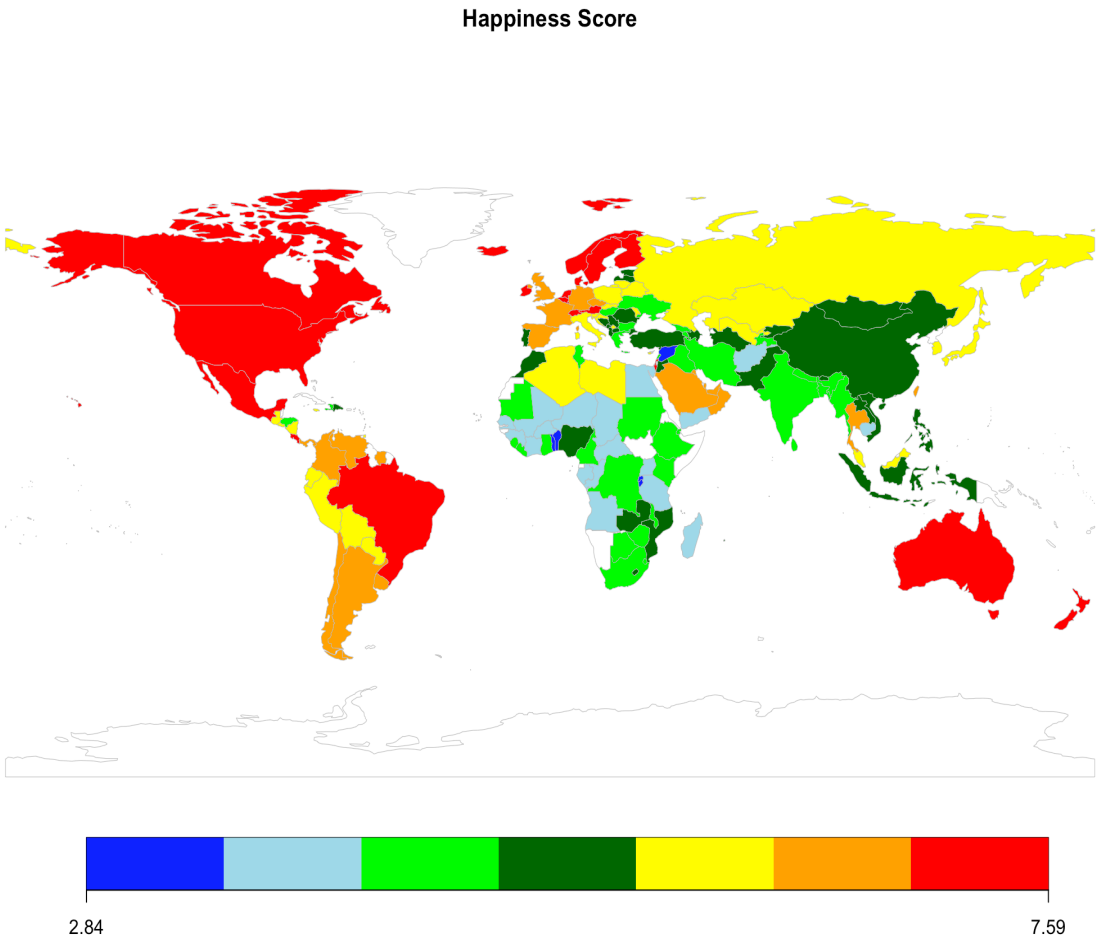
3.5 Global Happiness Rank in 2015

How did the “World Pappiness Report 2015” measure the level of Happiness? To measure happiness or the degree of satisfaction with life can be uncontrollable. Happiness Score is measure on scale from 0 to 10, where 0 is the worst possible life and 10 is the best possible life. . As illustration, the bar-chart below shows that Australia and New Zealand and North America tally the highest happiness scores among all regions with scores around 7.5. This is followed by European regions, such as Western Europe and Latin America and Caribbean, with 6.8 and 6.2, respectively. Asian regions such as Eastern and Southeastern Asia score about a 5.2 to 5.5. Central and Eastern Europe and Middle East and Northern Africa scored about 5.5. The region with the least happiness score is attributed to Sub Saharan Africa with a score of about 4.2. Generally, Australia and New Zealand regions score higher compared to the rest of the regions. North America also are tallied equivalently happy at the top spot. European countries are generally happy than Asian countries while the African regions are considered the least happy.

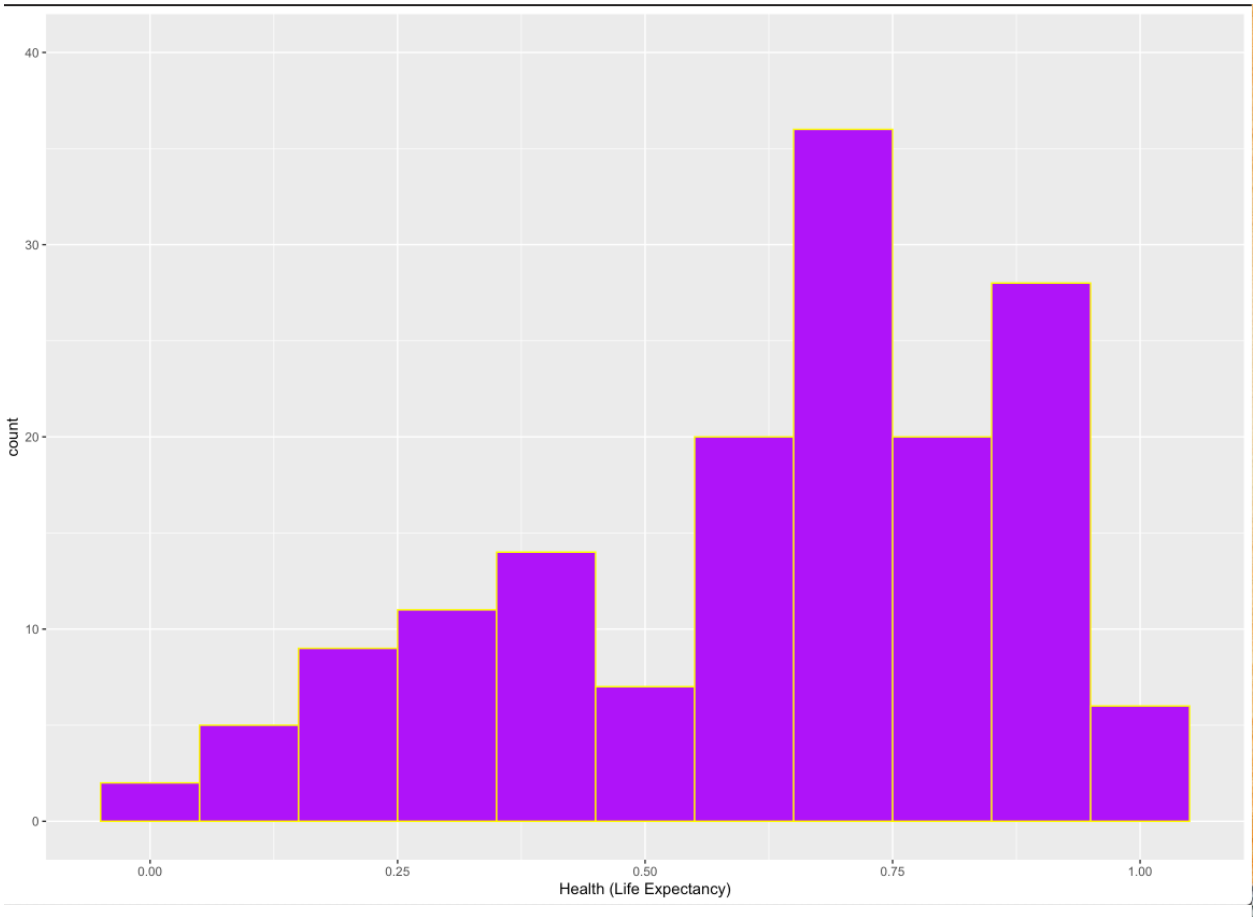


The map below is another great visual representation since it shows which countries are the happiest in the world in 2015. According to the report, the indicator to measure the happiness included 156 countries around the world. Then, to study the difference in the level of happiness of countries, an index composed of six factors is used as demonstrate in the correlation graph above. In 2015, the countries in the top ten were: Switzerland, Iceland, Denmark, Norway, Canada, Finland, Netherlands, Sweden, New Zealand, and Australia". All of these countries have the highest level of happiness since they promote the factor of happiness such as, income, health life expectancy, social relationships, freedom etc. On another hand, the ten countries with the

lowest score were: Chad, Guinea, Ivory Coast, Burkina Faso, Afghanistan, Rwanda, Benin, Syria, Burundi, Togo. These ten countries have the largest fall in the happiness index since they experienced some combination of economic, political and social stress.



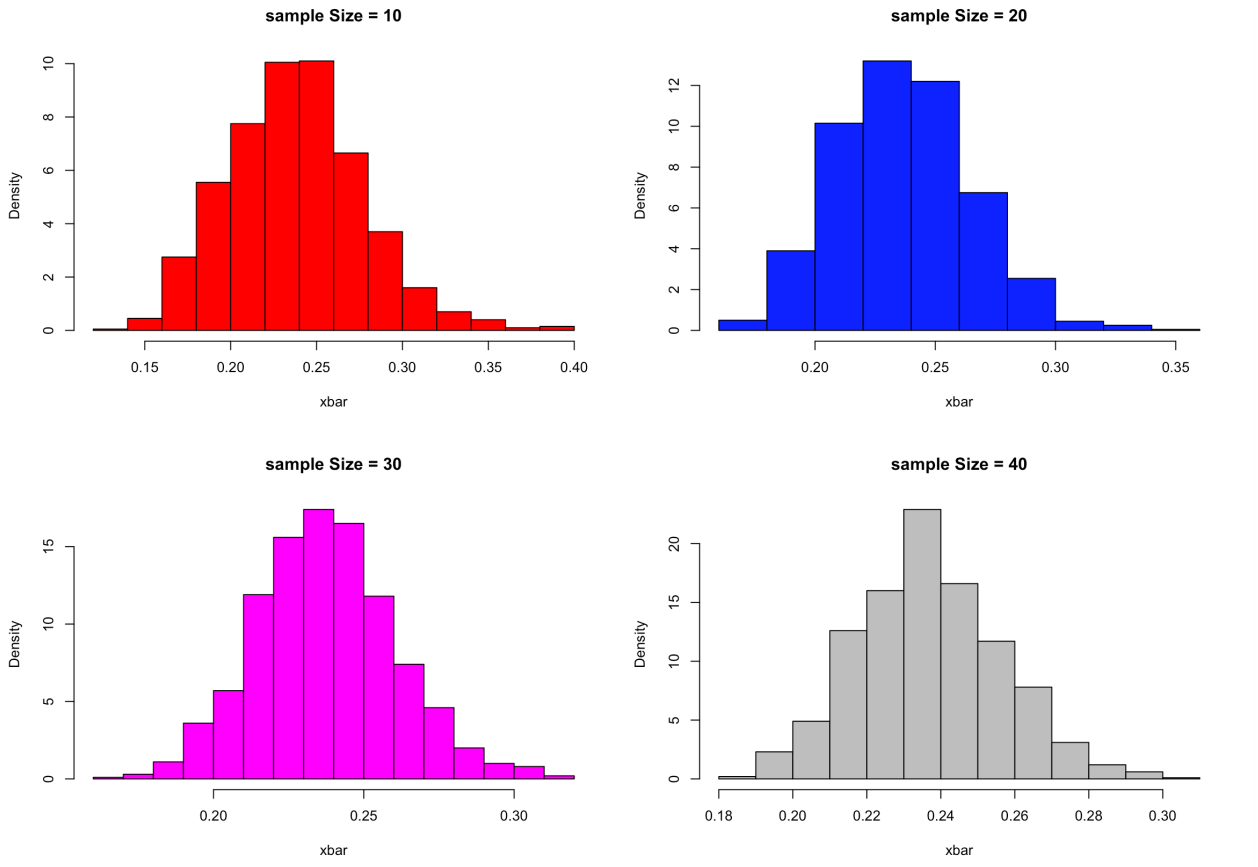
3.6 Examine Health (Life Expectancy) Score Distribution



The histogram above counts every distinct number of life expectancy in each of the countries. It's clear from the graph that the highest frequencies lie on the right side and the lowest on the left side. As a result, the distribution is negatively skewed (left tailed). That's because there's a long tail in the negative direction. The mean is 0.63, and the standard deviation is 0.25. The frequency of graph reflects that the number of years with happiness varies substantially between countries. Another key point is the fact that people in developed countries live longer doesn't necessarily mean that they live in good health. The magnitude of gender differences in life expectancy is substantial and varies between countries.

3.7 Health (Life expectancy): Central Limit Theorem

According to the module 4, the Central Limit Theorem states that the distribution of the sample means approaches a normal distribution, as the size the sample increases. The health life expectancy is a numerical variable, which will demonstrate this theory. To illustrate, the below four histograms are showing the sample means of 1000 random samples with sample size 10,20,30, and 40 that all normally distributed.



The output below shows that the means all the same for the four-sample size while the standard deviations has different samples sizes. These results prove the Central Limit Theorem which

states that the distribution of the sample means approaches a normal distribution as the sample size increase.

```
sample size =  10 mean = 0.24  sd = 0.04
sample size =  20 mean = 0.24  sd = 0.028
sample size =  30 mean = 0.24  sd = 0.023
sample size =  40 mean = 0.24  sd = 0.02
```

3.8 Sampling

A sample refers to a set of observations drawn from a population. The common methods used in probability sampling are simple random sampling, systematic sampling, stratified sampling, and cluster sampling. To illustrate, the following graphs compares distribution of the level of happiness from the population using the simple random sampling without replacement, systematic sampling, and cluster sampling.

A simple random sample of size 80 is draw from population of 158 without replacement. Every individual from the population has the same chance for selection in the sample as every other individual. As a result, the data of the selected sample shows the distribution and the frequency of the select Continents in each region as shown below.

```
> table(sample.1$Continent)

       Africa          Asia     Australia        Europe North America South America
           32            14             2            21             2             9
```

This show that every item or individual in the frame has the same chance 1/158 for selection.

Now, looking at proper table is a bit more effective than the table since it can compare the overall distribution of the regions for the whole dataset. The following table shows the percentages of these with respect to the sample size.

```
> round(prop.table(table(sample.1$Continent)),3)

      Africa         Asia    Australia       Europe North America South America
       0.400        0.175        0.025        0.262        0.025        0.112
```

A systematic sampling of sample size 80 is divided into 2 groups. From the first group, a random item is selected. The rows of the systematic sample are now computed by taking every 2nd item. The frequency and the percentages of select continents in each region is show below.

```
> table(sample.2$Continent)

      Africa         Asia    Australia       Europe North America South America
          30           11            2           25            1           11
```

```
> round(prop.table(table(sample.2$Continent)),3)

      Africa         Asia    Australia       Europe North America South America
       0.375        0.138        0.025        0.312        0.012        0.138
```

A stratifies sampling of sample size 80**.** The population of the whole frame are subdivided into 6 separate strata or subgroups base on the number of continents. Simple random samples are selected from each stratum and combine for the desired sample of size 80. The following tables shows that the number of samples selected from each stratum is proportional to the sizes based on the Continent variable. Also, it shows the frequencies for the selected continents, and the percentages respect to the sample size

```
> table(sample.3$Continent)

      Africa         Asia     Australia       Europe North America South America
          30           11            1           25             1            11
```

```
> round(prop.table(table(sample.3$Continent)),3)

      Africa         Asia     Australia       Europe North America South America
       0.380        0.139        0.013        0.316         0.013         0.139
```

Unfortunately, the table above shows that a selection bias occurs as a result of stratifies sampling since one element of the population was not selected. As a result, stratified sampling provides a faster way to select any item from frame, but it cannot be used with every data set.

Conclusion, these samples can be used instead of the whole dataset since is a useful method to determine which part of a population should be examined, in order to make inferences in the same population. However, we need to make sure to find the right sample to achieve an adequate representation of the population because some bias can occurs depending of the size of the we take from the population.