Final Project: World Happiness Report

Lizeth Andrea Castellanos Beltran

December 2017

Introduction

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

Description of the Data

The happiness scores and rankings use data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The columns following the happiness score estimate the extent to which each of six factors:

1. Economic production.
2. Social support.
3. Life expectancy.
4. Freedom.
5. Absence of corruption.
6. Generosity.

All factors contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors.

Download the data

I am providing the dataset for the year 2015 in the github directory https://github.com/lacbeltran/lps/blob/master/tasks/2005.csv

If you want to download the data from the kaggle site, you have to be logged. The URL for the CSV file is https://www.kaggle.com/unsdsn/world-happiness/downloads/2015.csv

Read the Data

- Read the data(If you have the file in the local directory)

```
library(readr)
df <- read_csv("2015.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Region = col_character(),
##   `Happiness Rank` = col_integer(),
##   `Happiness Score` = col_double(),
##   `Standard Error` = col_double(),
##   `Economy (GDP per Capita)` = col_double(),
##   Family = col_double(),
##   `Health (Life Expectancy)` = col_double(),
##   Freedom = col_double(),
##   `Trust (Government Corruption)` = col_double(),
##   Generosity = col_double(),
##   `Dystopia Residual` = col_double()
## )
```

```
df;
```

```
## # A tibble: 158 x 12
##        Country                    Region `Happiness Rank`
##          <chr>                     <chr>            <int>
##  1 Switzerland            Western Europe                1
##  2     Iceland            Western Europe                2
##  3     Denmark            Western Europe                3
##  4      Norway            Western Europe                4
##  5      Canada             North America                5
##  6     Finland            Western Europe                6
##  7 Netherlands            Western Europe                7
##  8      Sweden            Western Europe                8
##  9 New Zealand Australia and New Zealand                9
## 10   Australia Australia and New Zealand               10
## # ... with 148 more rows, and 9 more variables: `Happiness Score` <dbl>,
## #   `Standard Error` <dbl>, `Economy (GDP per Capita)` <dbl>,
## #   Family <dbl>, `Health (Life Expectancy)` <dbl>, Freedom <dbl>, `Trust
## #   (Government Corruption)` <dbl>, Generosity <dbl>, `Dystopia
## #   Residual` <dbl>
```

Metadata description

The dataset contains 12 columns and 158 countries. The next are the descriptions of every colunm according to the column metadata information:

1. Country: Name of the country.
2. Region: Region the country belongs to.
3. Happiness Rank: Rank of the country based on the Happiness Score.
4. Happiness Score: A metric measured in 2015 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".
5. Standard Error.
6. Economy (GDP per Capita): The extent to which GDP contributes to the calculation of the Happiness Score.
7. Family: The extent to which Family contributes to the calculation of the Happiness Score.
8. Health (Life Expectancy): The extent to which Life expectancy contributed to the calculation of the Happiness Score.
9. Freedom: The extent to which Freedom contributed to the calculation of the Happiness Score.

10. Trust (Government Corruption): The extent to which Perception of Corruption contributes to Happiness Score.
11. Generosity: The extent to which Generosity contributed to the calculation of the Happiness Score.
12. Dystopia Residual: The extent to which Dystopia Residual contributed to the calculation of the Happiness Score.

The column 6 to the colum 11 describe the extent to which these factors contribute in evaluating the happiness in each country. The Dystopia Residual metric actually is the Dystopia Happiness Score(1.85) + the Residual value or the unexplained value for each country as stated in the previous answer. Dystopia is an imaginary country that has the world's least-happy people. The purpose in establishing Dystopia is to have a benchmark against which all countries can be favorably compared. If you add all these factors up, you get the happiness score so it might be un-reliable to model them to predict Happiness Scores.

Critical analysis of the data

Is possible to see that the countries are already ranked, as you can see in the column "Happiness Rank". However, It would be interesting to check whether a certain factor contribute more to the overall happiness than others. There are 12 variables in the dataset to describe every observation (country happiness rank). The website provides a dataset dictionary with the list and the description of each variable. The variable "Standard Error" have not a description and for me was not clear the way that this variable was calculated. The data is present in a tidy form and there are not missing data in the cells.

Goal and Question

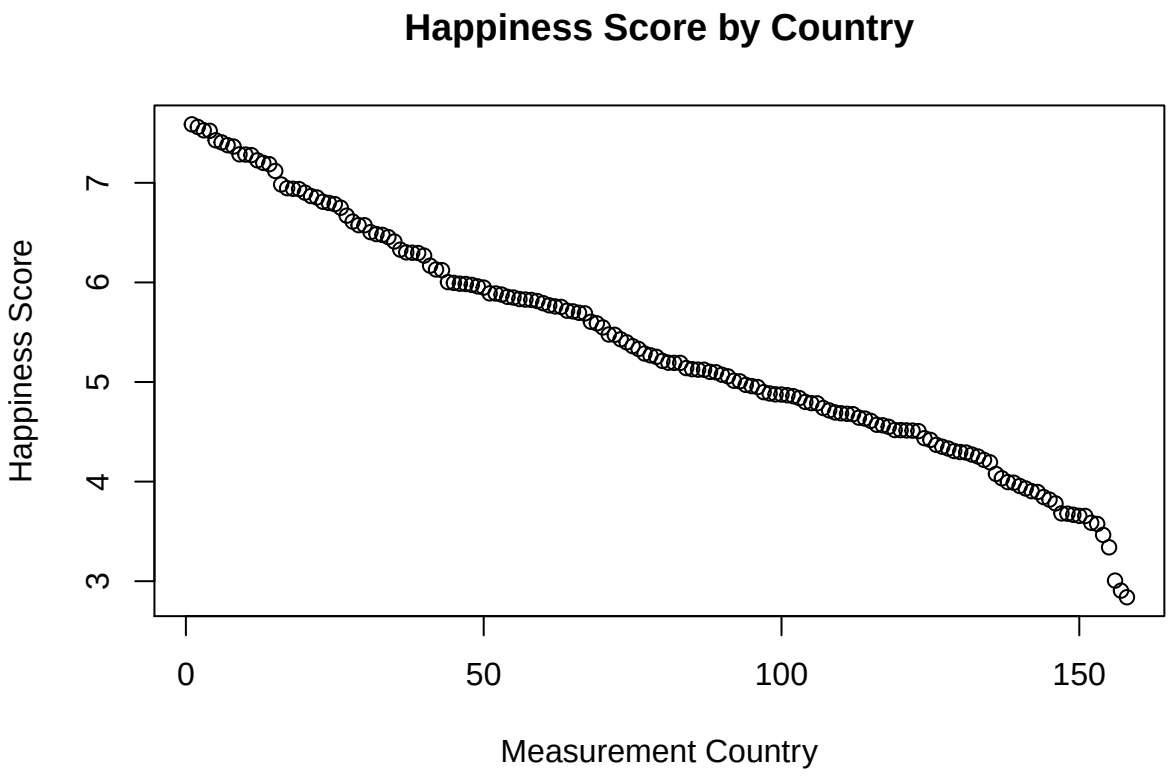
I will to try to answer the next specific question:

- What countries or regions rank the highest in overall happiness and each of the six factors contributing to happiness?

Analysis and Answer of the Question

The first thing I will do is plot the values of Happiness Score for all the 158 countries to have an idea about the distribution of this variable. It was to be expected the decreasing behaviour of the Happiness Score due that the countries are already ranked. One interesting thing to see in next figure is that the score gradually decrease. Nearby countries in the rank have very similar scores.

```
plot(df$`Happiness Score`, main = "Happiness Score by Country", ylab="Happiness Score", xlab="Measureme
```



Is possible to see that the mean estimation is `5.376`, while the median gives `5.232`. There is a big difference between the happiest and the least happy country in the world: `7.587` and `2.839` respectively.

```
summary(df$`Happiness Score`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.839   4.526   5.232   5.376   6.244   7.587
```

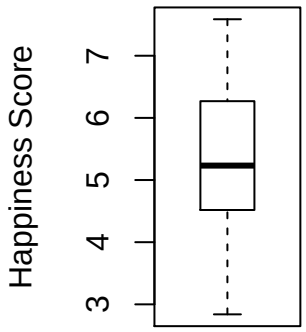
The standard deviation shows tha the variability of the Happiness Score is `1.14501`.

```
sd(df$`Happiness Score`)
```

```
## [1] 1.14501
```

The boxplot representation is useful to understand the distribution of the Happiness Score values.

```
boxplot(df$`Happiness Score`, ylab="Happiness Score")
```



The amount of countries is 158 so I will concentrate the analysis on the first ten and the last ten countries of the ranking. In others words I will select the top10 happiest and the top10 least happy countries in the world. The idea is plot their Happiness Score in a first overview. In a second way I will plot every factor to check whether a certain factor contribute more to the overall happiness than others.

- Load the necessary packages:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(ggplot2)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

The top10 happiest countries in the world

I will show the 10 first countries of the ranking:

```
df %>%
  head(n=10) %>%
 select(Country, Region, `Happiness Rank`, `Happiness Score`);
```

```
## # A tibble: 10 x 4
##        Country                    Region `Happiness Rank`
##          <chr>                     <chr>            <int>
##  1 Switzerland           Western Europe                1
##  2     Iceland           Western Europe                2
##  3     Denmark           Western Europe                3
##  4      Norway           Western Europe                4
##  5      Canada            North America                5
##  6     Finland           Western Europe                6
##  7 Netherlands           Western Europe                7
##  8      Sweden           Western Europe                8
##  9 New Zealand Australia and New Zealand               9
## 10   Australia Australia and New Zealand              10
## # ... with 1 more variables: `Happiness Score` <dbl>
```

As shown below, we can see that 7 of the 10 happiest countries are from Western Europe. I would suppose that maybe there are more developed countries in this region and this affects Happiness Score.

```
df %>%
  head(n=10) %>%
  group_by(Region) %>% summarize(occurrence=n());
```

```
## # A tibble: 3 x 2
##                       Region occurrence
##                        <chr>      <int>
## 1 Australia and New Zealand          2
## 2             North America          1
## 3           Western Europe          7
```

Now I will plot the Top10 Happiest Countries in a geom_bar to show the country and the score:

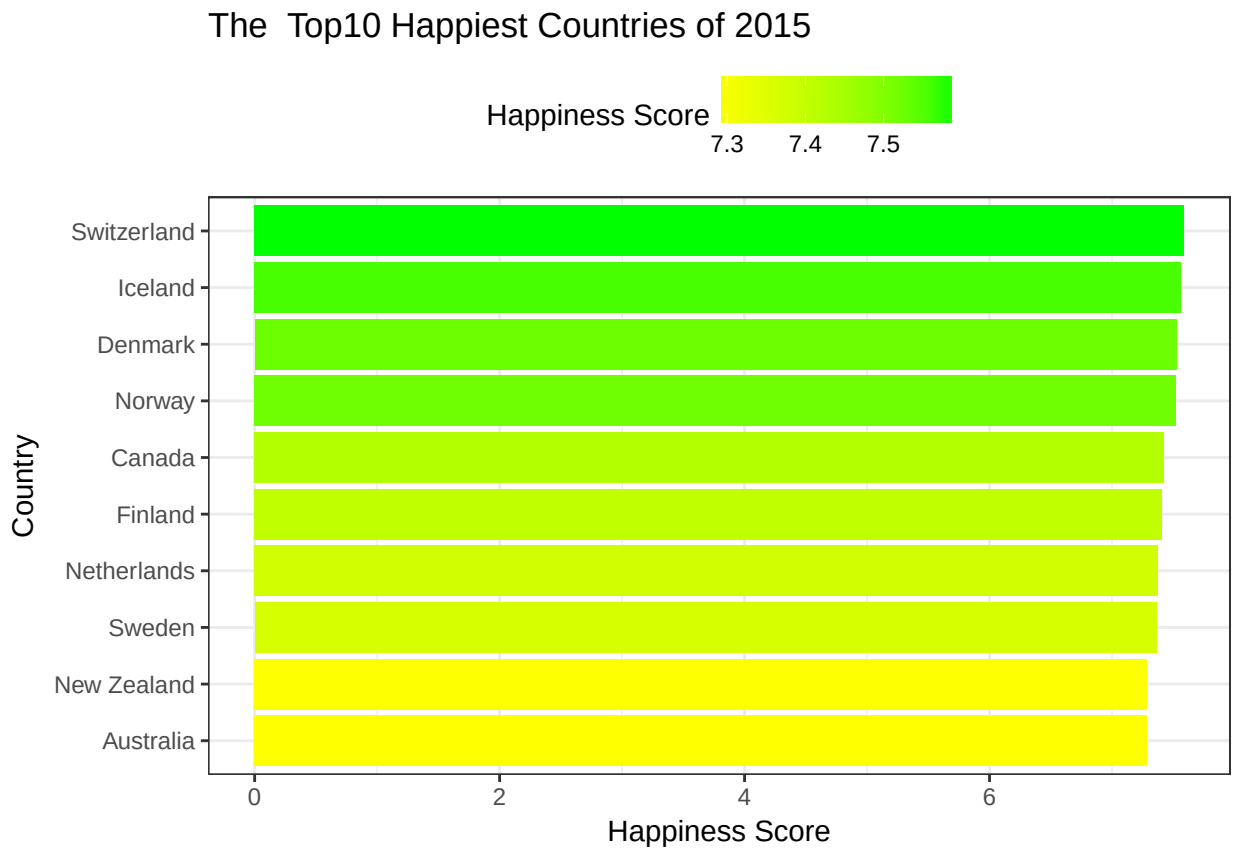
```
library(readr)
library(ggplot2)
df <- read_csv("2015.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Region = col_character(),
##   `Happiness Rank` = col_integer(),
##   `Happiness Score` = col_double(),
##   `Standard Error` = col_double(),
##   `Economy (GDP per Capita)` = col_double(),
##   Family = col_double(),
##   `Health (Life Expectancy)` = col_double(),
##   Freedom = col_double(),
##   `Trust (Government Corruption)` = col_double(),
##   Generosity = col_double(),
##   `Dystopia Residual` = col_double()
## )
```

```
df;
```

```
## # A tibble: 158 x 12
##        Country                    Region `Happiness Rank`
##          <chr>                     <chr>            <int>
##  1 Switzerland            Western Europe                1
##  2     Iceland            Western Europe                2
##  3     Denmark            Western Europe                3
##  4      Norway            Western Europe                4
##  5      Canada             North America                5
##  6     Finland            Western Europe                6
##  7 Netherlands            Western Europe                7
##  8      Sweden            Western Europe                8
##  9 New Zealand Australia and New Zealand                9
## 10   Australia Australia and New Zealand               10
## # ... with 148 more rows, and 9 more variables: `Happiness Score` <dbl>,
## #   `Standard Error` <dbl>, `Economy (GDP per Capita)` <dbl>,
## #   Family <dbl>, `Health (Life Expectancy)` <dbl>, Freedom <dbl>, `Trust
## #   (Government Corruption)` <dbl>, Generosity <dbl>, `Dystopia
## #   Residual` <dbl>
```

```r
df %>%
 #arrange(`Happiness Rank`) %>%
 head(10) %>%
 mutate(Country = factor(Country, levels = rev(Country))) %>%
 ggplot(aes(x=Country, y=`Happiness Score`,fill = `Happiness Score`)) +
 geom_bar(stat = "identity") + #position = position_stack(reverse = TRUE)) +
 coord_flip() + theme_bw() +
 ggtitle("The  Top10 Happiest Countries of 2015") +
 scale_fill_gradient(low = "yellow ", high = "green ")+
 theme(legend.position = "top")
```



Switzerland is on the top of the ranking, followed vy more european countries. Only one country of America is on the top 10, this was Canada. Australia and New Zeland are in the top 10 too.

The top10 least happy countries in the world

I will show the 10 last countries of the ranking:

```
df %>%
  tail(n=10) %>%
 select(Country, Region, `Happiness Rank`, `Happiness Score`);
```

```
## # A tibble: 10 x 4
##          Country                           Region `Happiness Rank`
##            <chr>                            <chr>             <int>
##  1         Chad              Sub-Saharan Africa               149
##  2       Guinea              Sub-Saharan Africa               150
##  3  Ivory Coast              Sub-Saharan Africa               151
##  4 Burkina Faso              Sub-Saharan Africa               152
##  5  Afghanistan                   Southern Asia               153
##  6       Rwanda              Sub-Saharan Africa               154
##  7        Benin              Sub-Saharan Africa               155
##  8        Syria Middle East and Northern Africa               156
##  9      Burundi              Sub-Saharan Africa               157
## 10         Togo              Sub-Saharan Africa               158
## # ... with 1 more variables: `Happiness Score` <dbl>
```

As shown below, we can see that 8 of the 10 least happy countries are from Sub-Saharan Africa. African countries are associated with lack of basic human needs. So I think that in fact the geographical ubication affects the happiness score.

```
df %>%
  tail(n=10) %>%
  group_by(Region) %>% summarize(occurrence=n());
```

```
## # A tibble: 3 x 2
##                             Region occurrence
##                              <chr>      <int>
## 1 Middle East and Northern Africa          1
## 2                   Southern Asia          1
## 3              Sub-Saharan Africa          8
```

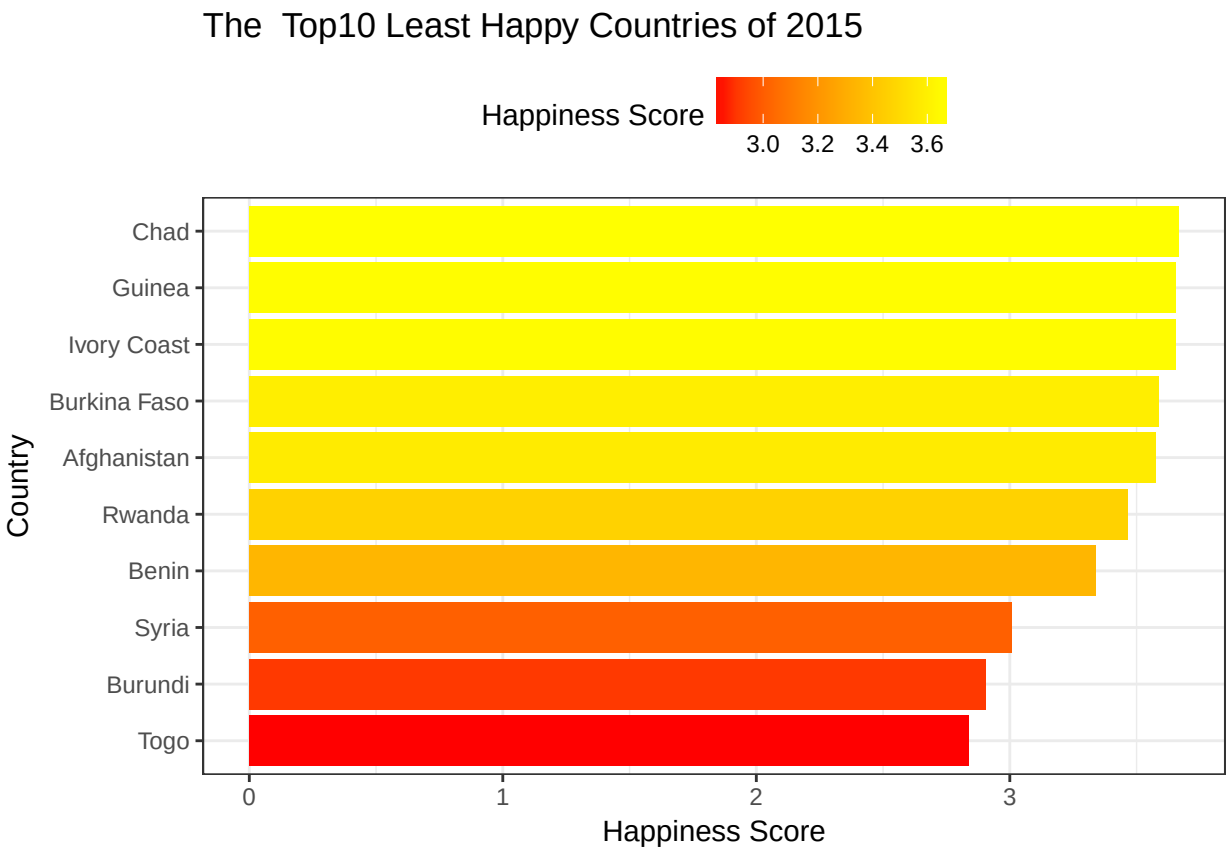
Now I will plot the Top10 Least Happy Countries in a geom_bar to show the country and the score:

```
library(readr)
df <- read_csv("2015.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Region = col_character(),
##   `Happiness Rank` = col_integer(),
##   `Happiness Score` = col_double(),
##   `Standard Error` = col_double(),
##   `Economy (GDP per Capita)` = col_double(),
##   Family = col_double(),
##   `Health (Life Expectancy)` = col_double(),
##   Freedom = col_double(),
##   `Trust (Government Corruption)` = col_double(),
##   Generosity = col_double(),
##   `Dystopia Residual` = col_double()
## )
```

```
#df;

 df %>%
  #arrange(`Happiness Rank`) %>%
  tail(10) %>%
  mutate(Country = factor(Country, levels = rev(Country))) %>%
  ggplot(aes(x=Country, y=`Happiness Score`,fill = `Happiness Score`)) +
    geom_bar(stat = "identity") + #position = position_stack(reverse = TRUE)) +
  coord_flip() + theme_bw() +
  ggtitle("The  Top10 Least Happy Countries of 2015") +
  scale_fill_gradient(low = "red ", high = "yellow")+
  theme(legend.position = "top")
```

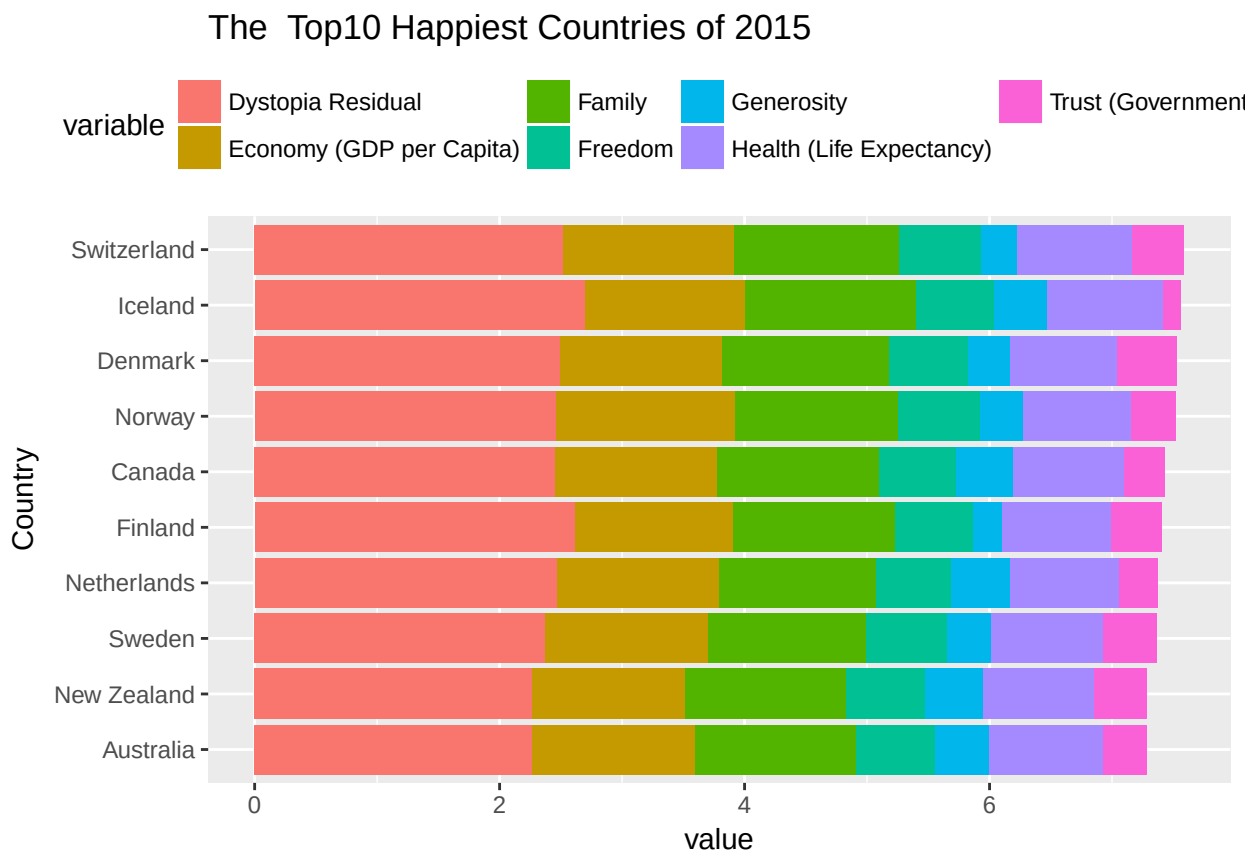


Togo is on the last position of the ranking. Only one country of Asia is on the ranking of the least happy countries. Neither Europe nor America are in the lasts positions.

How the factors contribute th the ranking

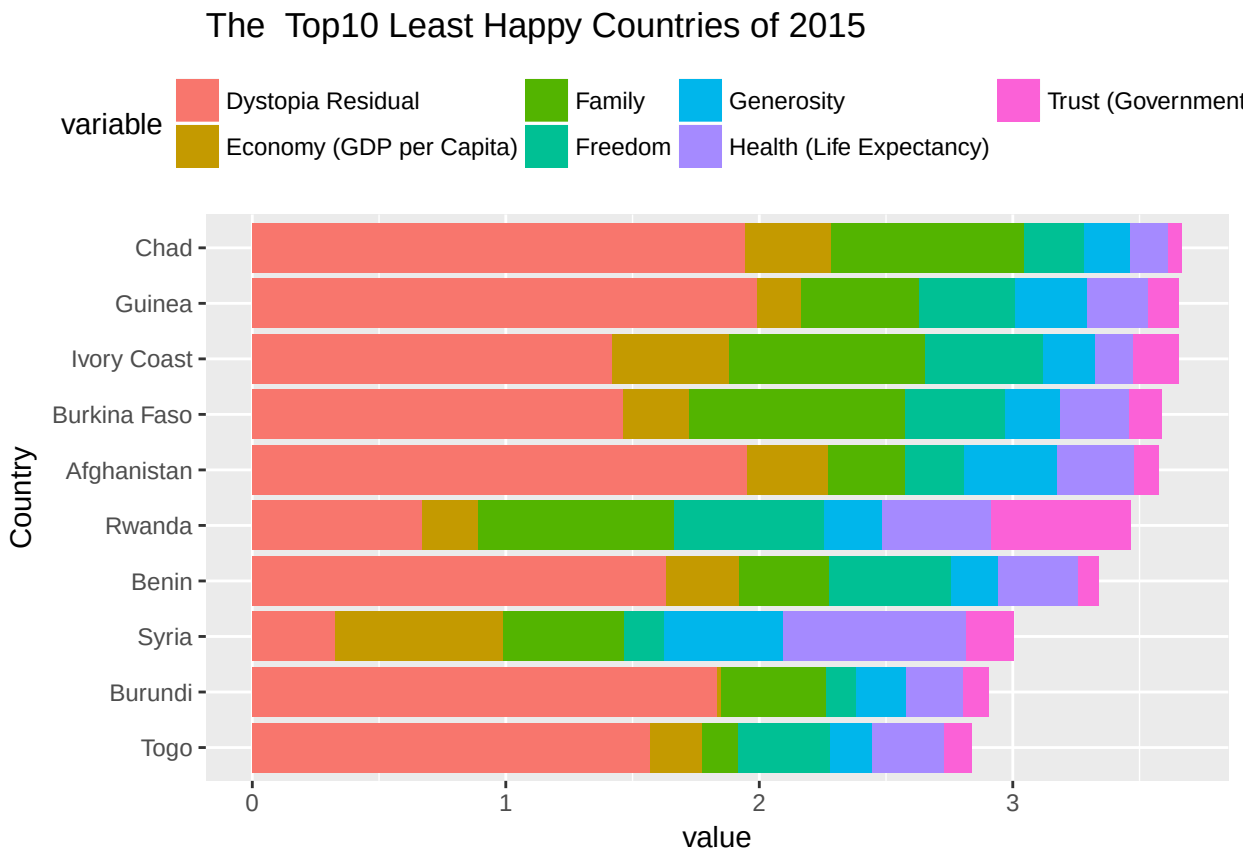
- Now, It would be interesting to check whether a certain factor contribute more to the overall happiness than others. So I need to plot all the factors in a way that I can compare the difference between countries.

```
library(tidyr)
library(magrittr)
library(dplyr)
library(ggplot2)
df %>%
  arrange(`Happiness Rank`) %>%
  #sample_n(10) %>%
  head(10) %>%
  mutate(Country = factor(Country, levels = rev(Country))) %>%
  select(1,6:12)%>%
  gather(variable, value, -Country) %>%
  ggplot(aes(Country))+
  geom_bar(aes(y=value, fill = variable), position = position_stack(reverse = TRUE), stat="identity") +
  coord_flip() +
  ggtitle("The  Top10 Happiest Countries of 2015") +
  theme(legend.position = "top")
```



For the Top10 Happiest Countries we can see that all the factors are very similar.

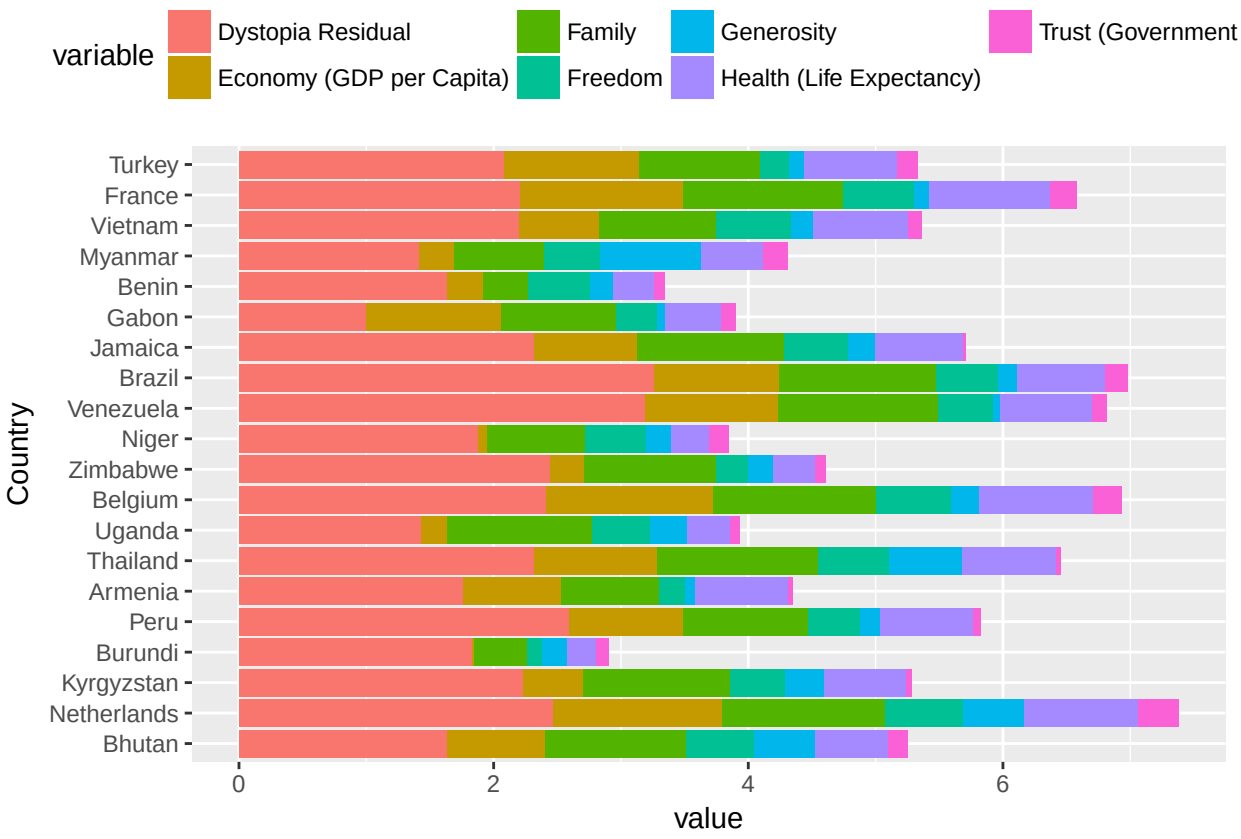
```
library(tidyr)
library(magrittr)
library(dplyr)
library(ggplot2)
df %>%
  arrange(`Happiness Rank`) %>%
  #sample_n(10) %>%
  tail(10) %>%
  mutate(Country = factor(Country, levels = rev(Country))) %>%
  select(1,6:12)%>%
  gather(variable, value, -Country) %>%
  ggplot(aes(Country))+
  geom_bar(aes(y=value, fill = variable), position = position_stack(reverse = TRUE), stat="identity") +
  coord_flip() +
  ggtitle("The  Top10 Least Happy Countries of 2015") +
  theme(legend.position = "top")
```



However, for the Top10 Least Happy Countries we can see some differences between the factors. In this case, Dystopia Residual and Family are the factors that more contribute to the happiness. Trust in the goverment, life expectancy and generosity are the factors that contribute least to the happiness.

Now, I will sample 20 countries of the ranking to see differences between the others contries. Curiously, trust in the goverment and generosity are the factor that contribute least to the happiness. Dystopia Residual, economy and family are the factors than more contribute to the happiness.

```
library(tidyr)
library(magrittr)
library(dplyr)
library(ggplot2)
df %>%
  #arrange(Happiness Rank) %>%
  sample_n(20) %>%
  #head(10) %>%
  mutate(Country = factor(Country, levels = rev(Country))) %>%
  select(1,6:12)%>%
  gather(variable, value, -Country) %>%
  ggplot(aes(Country))+
  geom_bar(aes(y=value, fill = variable), position = position_stack(reverse = TRUE), stat="identity") +
  coord_flip() +
  theme(legend.position = "top")
```


variable
Dystopia Residual
Economy (GDP per Capita)
Family
Freedom
Generosity
Health (Life Expectancy)
Trust (Government
Country
Turkey
France
Vietnam
Myanmar
Benin
Gabon
Jamaica
Brazil
Venezuela
Niger
Zimbabwe
Belgium
Uganda
Thailand
Armenia
Peru
Burundi
Kyrgyzstan
Netherlands
Bhutan
0
2
4
6
value