DATA 606 Data Project Proposal

Vinayak Kamath

3/27/2020

Data Preparation

Life Expectancy and Happiness; We will use the following data-sets from Kaggle on relating happiness with life expectancy. Life expectancy could be linked to many factors like monetary or physical needs or living condition or external factors with politics or like. Over here we will aim to correlate life expectancy with happiness as a key factor.

Life Expectancy (WHO)
World Happiness Report

```
library(readr)
library(knitr)
library(dplyr)
library(tidyr)
library(ggplot2)
```

```
# Load data from the life-expectancy csv file:
theUrl.lifeExpectancy <- 'https://raw.githubusercontent.com/kamathvk1982/Data606-FinalProject/master/da
lifeExp.df <- read_csv(theUrl.lifeExpectancy,  na = c("", "NA","N/A"))

# Load data from the world-happiness csv file for 2018 and 2019:
theUrl.happ.2018 <- 'https://raw.githubusercontent.com/kamathvk1982/Data606-FinalProject/master/data/wo
happ.2018 <- read_csv(theUrl.happ.2018,  na = c("", "NA","N/A"))
theUrl.happ.2019 <- 'https://raw.githubusercontent.com/kamathvk1982/Data606-FinalProject/master/data/wo
happ.2019 <- read_csv(theUrl.happ.2019,  na = c("", "NA","N/A"))
```

```
# View a summary of the Data Frame along with Row counts:
dim(lifeExp.df)
```

```
## [1] 19028     4
```

```
kable(head(lifeExp.df,10))
```

Entity	Code	Year	Life expectancy (years)
Afghanistan	AFG	1950	27.638
Afghanistan	AFG	1951	27.878
Afghanistan	AFG	1952	28.361
Afghanistan	AFG	1953	28.852
Afghanistan	AFG	1954	29.350

Entity	Code	Year	Life expectancy (years)
Afghanistan	AFG	1955	29.854
Afghanistan	AFG	1956	30.365
Afghanistan	AFG	1957	30.882
Afghanistan	AFG	1958	31.403
Afghanistan	AFG	1959	31.925

```
dim(happ.2018)
```

```
## [1] 156   9
```

```
kable(head(happ.2018,10))
```

Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make
1	Finland	7.632	1.305	1.592	0.874	
2	Norway	7.594	1.456	1.582	0.861	
3	Denmark	7.555	1.351	1.590	0.868	
4	Iceland	7.495	1.343	1.644	0.914	
5	Switzerland	7.487	1.420	1.549	0.927	
6	Netherlands	7.441	1.361	1.488	0.878	
7	Canada	7.328	1.330	1.532	0.896	
8	New Zealand	7.324	1.268	1.601	0.876	
9	Sweden	7.314	1.355	1.501	0.913	
10	Australia	7.272	1.340	1.573	0.910	

```
dim(happ.2019)
```

```
## [1] 156   9
```

```
kable(head(happ.2019,10))
```

Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make
1	Finland	7.769	1.340	1.587	0.986	
2	Denmark	7.600	1.383	1.573	0.996	
3	Norway	7.554	1.488	1.582	1.028	
4	Iceland	7.494	1.380	1.624	1.026	
5	Netherlands	7.488	1.396	1.522	0.999	
6	Switzerland	7.480	1.452	1.526	1.052	
7	Sweden	7.343	1.387	1.487	1.009	
8	New Zealand	7.307	1.303	1.557	1.026	
9	Canada	7.278	1.365	1.505	1.039	
10	Austria	7.246	1.376	1.475	1.016	

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

1. Country in top 10 rank for happiness, have higher `Life Expectancy (years)` ?
2. Does higher `Life Expectancy (years)` countries are also most Happy (For 2019)?
3. Are people, in a given country, happier in 2019 then they were 2018?
4. Does being `Generous` make people happy ?
5. Does having `Freedom to make life choices` make people happy ?

Cases

What are the cases, and how many are there?

1. Each case in the Happiness dataset is for a given Country and shows its rank, score and other parameter score. There are 156 cases in each 2018 and 2019 dataset.
2. Each case in the Life Expectancy dataset is for a Given country and year and shows its life expectancy for the year. There are 19028 cases in the dataset.

Data collection

Describe the method of data collection.
The data was collected from Kaggle, but the original source for the data is World Bank Data.

Type of study

What type of study is this (observational/experiment)?

This is an observational study. We are looking at data for year 2018 and 2019 for the countries in the world.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

The data was collected from below sites:
Life Expectancy (WHO) data is found here.
World Happiness Report data is found here.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?
The response variable is the `Score` variable and its quantitative (numerical).

Independent Variable

You should have two independent variables, one quantitative and one qualitative.
The independent variables are the `Life expectancy (years)` variable and the other variable is `Generosity` in the happiness dataset; Both these are quantitative (numerical).

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
summary(lifeExp.df)
```

1. Summary Statistics for each dataframe is as below:

```
##     Entity              Code                Year      Life expectancy (years)
##  Length:19028       Length:19028       Min.   :1543   Min.   :17.76
##  Class :character   Class :character   1st Qu.:1961   1st Qu.:52.31
##  Mode  :character   Mode  :character   Median :1980   Median :64.71
##                                        Mean   :1975   Mean   :61.75
##                                        3rd Qu.:2000   3rd Qu.:71.98
##                                        Max.   :2019   Max.   :86.75
```

```
summary(happ.2018)
```

```
##   Overall rank    Country or region      Score       GDP per capita
##  Min.   :  1.00   Length:156         Min.   :2.905   Min.   :0.0000
##  1st Qu.: 39.75   Class :character   1st Qu.:4.454   1st Qu.:0.6162
##  Median : 78.50   Mode  :character   Median :5.378   Median :0.9495
##  Mean   : 78.50                      Mean   :5.376   Mean   :0.8914
##  3rd Qu.:117.25                      3rd Qu.:6.168   3rd Qu.:1.1978
##  Max.   :156.00                      Max.   :7.632   Max.   :2.0960
##
##  Social support  Healthy life expectancy Freedom to make life choices
##  Min.   :0.000   Min.   :0.0000          Min.   :0.0000
##  1st Qu.:1.067   1st Qu.:0.4223          1st Qu.:0.3560
##  Median :1.255   Median :0.6440          Median :0.4870
##  Mean   :1.213   Mean   :0.5973          Mean   :0.4545
##  3rd Qu.:1.463   3rd Qu.:0.7772          3rd Qu.:0.5785
##  Max.   :1.644   Max.   :1.0300          Max.   :0.7240
##
##    Generosity     Perceptions of corruption
##  Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.1095   1st Qu.:0.051
##  Median :0.1740   Median :0.082
##  Mean   :0.1810   Mean   :0.112
##  3rd Qu.:0.2390   3rd Qu.:0.137
##  Max.   :0.5980   Max.   :0.457
##                   NA's   :1
```

```
summary(happ.2019)
```

```
##   Overall rank    Country or region      Score       GDP per capita
##  Min.   :  1.00   Length:156         Min.   :2.853   Min.   :0.0000
```

```
##  1st Qu.: 39.75   Class :character   1st Qu.:4.545   1st Qu.:0.6028
##  Median : 78.50   Mode  :character   Median :5.380   Median :0.9600
##  Mean   : 78.50                      Mean   :5.407   Mean   :0.9051
##  3rd Qu.:117.25                      3rd Qu.:6.184   3rd Qu.:1.2325
##  Max.   :156.00                      Max.   :7.769   Max.   :1.6840
##  Social support  Healthy life expectancy Freedom to make life choices
##  Min.   :0.000   Min.   :0.0000          Min.   :0.0000
##  1st Qu.:1.056   1st Qu.:0.5477          1st Qu.:0.3080
##  Median :1.272   Median :0.7890          Median :0.4170
##  Mean   :1.209   Mean   :0.7252          Mean   :0.3926
##  3rd Qu.:1.452   3rd Qu.:0.8818          3rd Qu.:0.5072
##  Max.   :1.624   Max.   :1.1410          Max.   :0.6310
##    Generosity     Perceptions of corruption
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1087   1st Qu.:0.0470
##  Median :0.1775   Median :0.0855
##  Mean   :0.1848   Mean   :0.1106
##  3rd Qu.:0.2482   3rd Qu.:0.1412
##  Max.   :0.5660   Max.   :0.4530
```

```
# We will get the Top 10 ranked Happy countries in 2018:
top.10.2018 <- happ.2018 %>%
  filter(happ.2018$"Overall rank" < 11 ) %>%
  select("Overall rank", "Country or region") %>%
  mutate(Year = 2018)

# Next, We will get the Top 10 ranked Happy countries in 2019:
top.10.2019 <- happ.2019 %>%
  filter(happ.2018$"Overall rank" < 11 )  %>%
  select("Overall rank", "Country or region")%>%
  mutate(Year = 2019)

# We will bind the above tow dataset into one:
top.10.new <- rbind(top.10.2018, top.10.2019)

# Next, we work on the Life Expectancy Dataset to get 2018 and 2019 data:
lifeExp.df.new <- lifeExp.df %>%
  filter(lifeExp.df$Year == 2018 | lifeExp.df$Year == 2019 )

# Now, joining the above tow final data sets of Happieness and Life Expectancy
# we can check the life expectany for happier countries:
top.10.hap.life <- inner_join(lifeExp.df.new, top.10.new,
                              by = c(Entity="Country or region" , "Year" = "Year") ) %>%
  arrange(Entity, Year)

kable(top.10.hap.life)
```

2. Getting Top ten happy country in 2018 and 2019 and getting their life expectancy:

Entity	Code	Year	Life expectancy (years)	Overall rank
Australia	AUS	2018	83.281	10
Austria	AUT	2019	81.544	10
Canada	CAN	2018	82.315	7
Canada	CAN	2019	82.434	9
Denmark	DNK	2018	80.784	3
Denmark	DNK	2019	80.898	2
Finland	FIN	2018	81.736	1
Finland	FIN	2019	81.908	1
Iceland	ISL	2018	82.855	4
Iceland	ISL	2019	82.993	4
Netherlands	NLD	2018	82.143	6
Netherlands	NLD	2019	82.283	5
New Zealand	NZL	2018	82.145	8
New Zealand	NZL	2019	82.288	8
Norway	NOR	2018	82.271	2
Norway	NOR	2019	82.404	3
Sweden	SWE	2018	82.654	9
Sweden	SWE	2019	82.797	7
Switzerland	CHE	2018	83.630	5
Switzerland	CHE	2019	83.779	6

=> We can see that the happier the country is the higher the life expectancy it has.

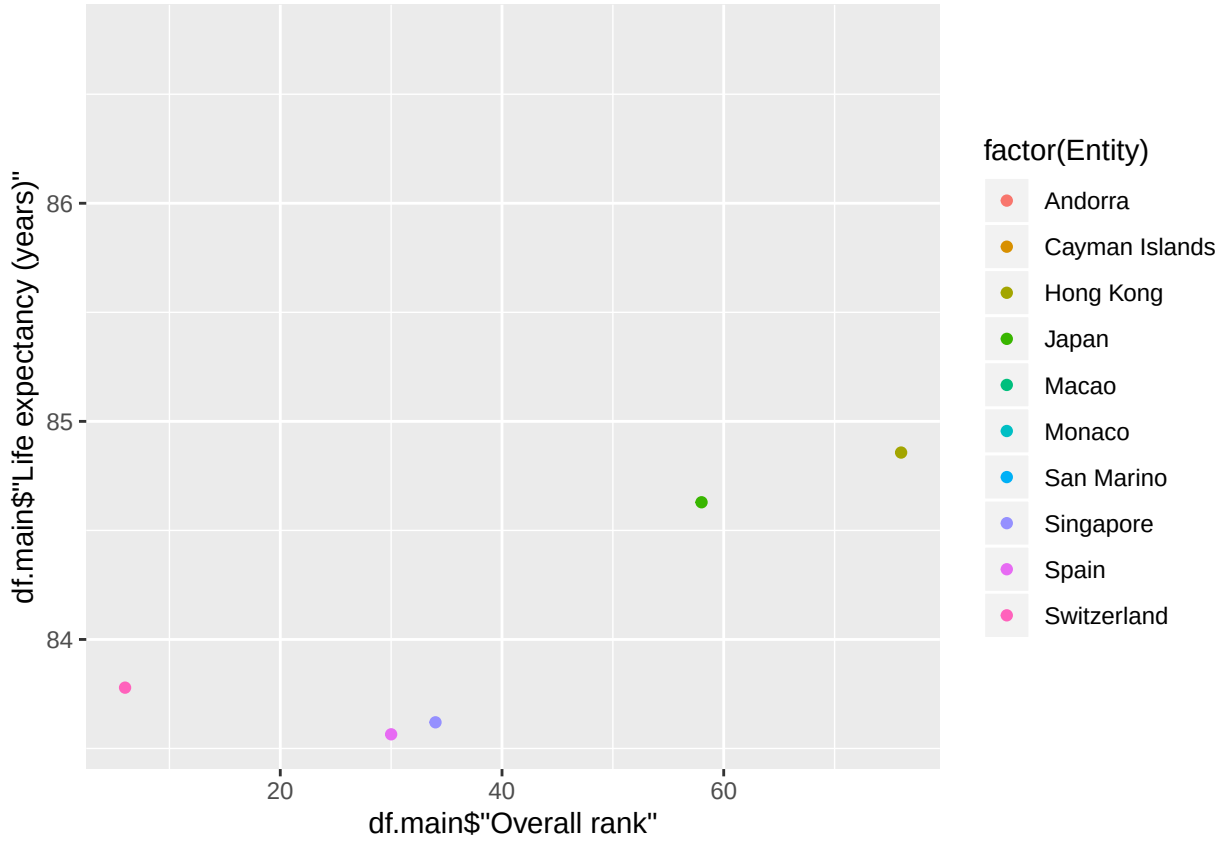
```
# Getting the 2019 data and then getting top 10 life expentancy countries:
lifeExp.df.2019 <- lifeExp.df %>%
  filter(Year == 2019)

top.10.life <- top_n(lifeExp.df.2019, 10, lifeExp.df.2019$"Life expectancy (years)")

# We are using Left join here as not all countries data is avialble in Happ.2019:
df.main <- left_join(top.10.life,happ.2019,  by=c( "Entity"="Country or region" ))

kable(df.main)
```

3. Does higher Life Expectancy (years) countries are also most Happy (for 2019) :

Entity	Code	Year	Life expectancy (years)	Overall rank	Score	GDP per capita	Social support	Heal
Andorra	AND	2019	83.732	NA	NA	NA	NA	
Cayman Islands	CYM	2019	83.924	NA	NA	NA	NA	
Hong Kong	HKG	2019	84.857	76	5.430	1.438	1.277	
Japan	JPN	2019	84.629	58	5.886	1.327	1.419	
Macao	MAC	2019	84.244	NA	NA	NA	NA	
Monaco	MCO	2019	86.751	NA	NA	NA	NA	
San Marino	SMR	2019	84.972	NA	NA	NA	NA	
Singapore	SGP	2019	83.620	34	6.262	1.572	1.463	
Spain	ESP	2019	83.565	30	6.354	1.286	1.484	
Switzerland	CHE	2019	83.779	6	7.480	1.452	1.526	

```
ggplot(df.main) +
  geom_point(aes(x=df.main$"Overall rank", y=df.main$"Life expectancy (years)"
                 , colour = factor(Entity)) )
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



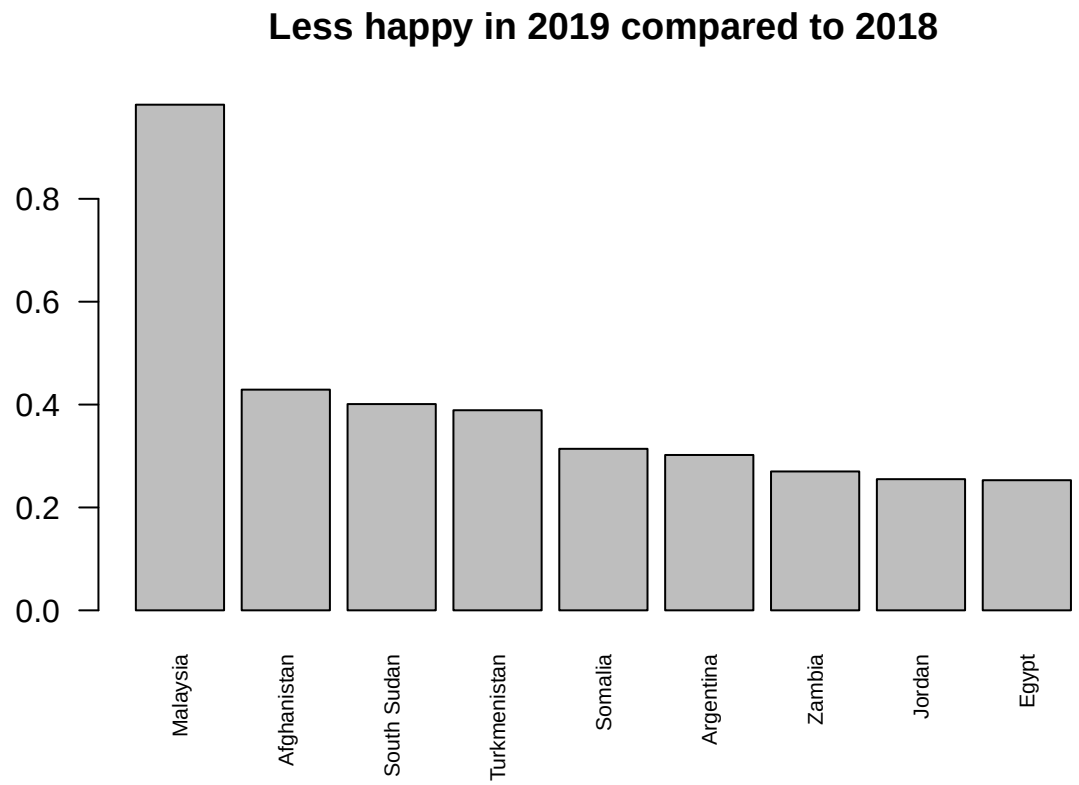
=> By checking the Rank and Score for these 5 countries we can say that higher life expectancy does not lead to Higher Happiness.

```
# We can check this by comparing score of the country in 2018 vs that in 2019 and see
# if they dropped by a margin of more then 0.25 basis:
df.main1 <- inner_join(happ.2018 , happ.2019, by=c("Country or region" = "Country or region")
           ,  suffix=c(".2018",".2019")  ) %>%
  filter(Score.2018 - Score.2019 >= 0.25) %>%
  select ("Country or region", "Overall rank.2018","Score.2018", "Overall rank.2019", "Score.2019") %>%
  mutate(Margin.Diff = Score.2018 - Score.2019) %>%
  arrange(desc(Margin.Diff))

kable(df.main1)
```

4. Are people, in a given country, happier in 2019 then they were 2018 :

Country or region	Overall rank.2018	Score.2018	Overall rank.2019	Score.2019	Margin.Diff
Malaysia	35	6.322	80	5.339	0.983
Afghanistan	145	3.632	154	3.203	0.429
South Sudan	154	3.254	156	2.853	0.401
Turkmenistan	68	5.636	87	5.247	0.389
Somalia	98	4.982	112	4.668	0.314
Argentina	29	6.388	47	6.086	0.302
Zambia	125	4.377	138	4.107	0.270
Jordan	90	5.161	101	4.906	0.255
Egypt	122	4.419	137	4.166	0.253

```
barplot(df.main1$Margin.Diff, names.arg=df.main1$"Country or region", las=2,cex.names=0.7
        , main="Less happy in 2019 compared to 2018")
```



=> There are 9 (out of 156) countries that were less happy in 2019 compared to 2018.

```
# We can get the min, max and mean of the Score for first 10 ranked Countries
# and compare aginst the same for first 10 `Generous` countries:
```

```r
# Min, Max and Mean for top 10 ranked countries:
top_n( happ.2018, -10, happ.2018$"Overall rank" ) %>%
  select(Score) %>%
  summary(Score)
```

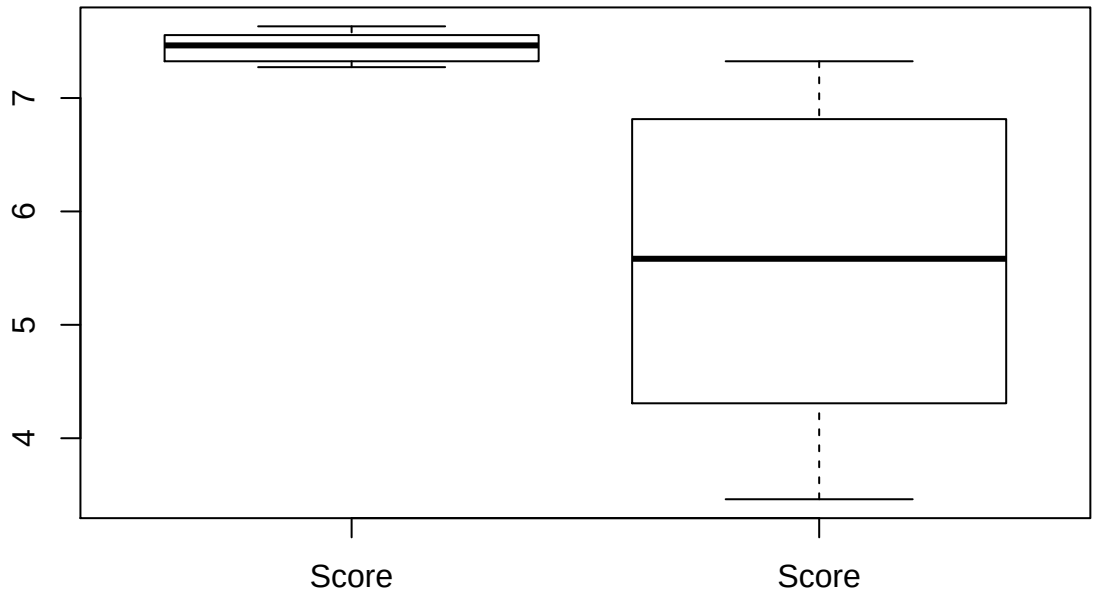
5. Checking `Generosity` :

```
##      Score
##  Min.   :7.272
##  1st Qu.:7.325
##  Median :7.464
##  Mean   :7.444
##  3rd Qu.:7.540
##  Max.   :7.632
```

```r
# Min, Max and Mean for top 10 `Generous` countries:
top_n( happ.2018, 10, Generosity ) %>%
  select(Score) %>%
  summary(Score)
```

```
##      Score
##  Min.   :3.462
##  1st Qu.:4.502
##  Median :5.582
##  Mean   :5.564
##  3rd Qu.:6.767
##  Max.   :7.324
```

```r
boxplot(c(top_n( happ.2018, -10, happ.2018$"Overall rank" ) %>%
  select(Score), top_n( happ.2018, 10, Generosity ) %>%
  select(Score)  ) )
```



=> By comparing the three attributes Min, Max and mean, we can say that being `Generous` does not lead to being Happy.

```
# We can get the min, max and mean of the Score for first 10 ranked Countries
# and compare aginst the same for first 10 `Freedom to make life choices` countries:

# Min, Max and Mean for top 10 ranked countries:
top_n( happ.2018, -10, happ.2018$"Overall rank" ) %>%
  select(Score) %>%
  summary(Score)
```

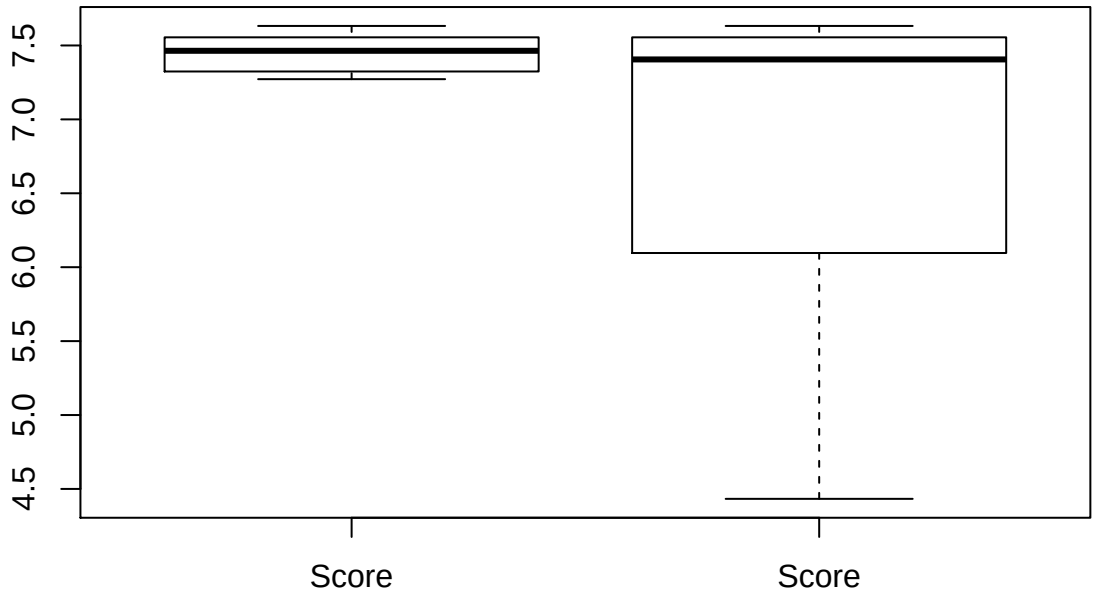
6. Checking `Freedom to make life choices` :

```
##      Score
##  Min.   :7.272
##  1st Qu.:7.325
##  Median :7.464
##  Mean   :7.444
##  3rd Qu.:7.540
##  Max.   :7.632
```

```r
# Min, Max and Mean for top 10 `Freedom to make life choices` countries:
top_n( happ.2018, 10, happ.2018$"Freedom to make life choices" ) %>%
  select(Score) %>%
  summary(Score)
```

```
##      Score
##  Min.   :4.433
##  1st Qu.:6.401
##  Median :7.405
##  Mean   :6.791
##  3rd Qu.:7.540
##  Max.   :7.632
```

```r
boxplot(c(top_n( happ.2018, -10, happ.2018$"Overall rank" ) %>%
  select(Score), top_n( happ.2018, 10, happ.2018$"Freedom to make life choices" ) %>%
  select(Score) ) )
```



=> By comparing the three attributes Min, Max and mean, we can say that having `Freedom to make life choices` does lead to being Happy. The mean is close to the top ranked mean and max score matches.

We can make a lot of linear regression comparisons between various variables, comparing it to happiness `Score`, and use various other statistical inference techniques to determine best variables to use.