Final project

Code

Isabella Liu

EDA:

1. Guiding question: What factors impact on people's happiness score the most?

2. Become acquainted with your data sources:

- Where did you find them?
  I found the data sources on World Happiness Report official website.
- Who collected/maintains them?
  Acoording to their official website, the data was collected from the Gallup World Poll, and supported by the Ernesto Illy Foundation, illycaffè, Davines Group, Blue Chip Foundation, the William, Jeff, and Jennifer Gross Family Foundation, and Unilever's largest ice cream brand Wall's.
- When & Why were they originally collected?
  The data was collected in 2015. The intention was to measure the happiness of people and to use the data to help guide public policy.
- What does a case represent in each data source, and how many total cases are available?

Hide

```
df2015 <- read.csv("2015.csv")

df2015
```

Country <chr>	Region <chr>	Happiness.Rank <int>	Happiness.Score <dbl>
Switzerland	Western Europe	1	7.587
Iceland	Western Europe	2	7.561
Denmark	Western Europe	3	7.527
Norway	Western Europe	4	7.522
Canada	North America	5	7.427
Finland	Western Europe	6	7.406
Netherlands	Western Europe	7	7.378
Sweden	Western Europe	8	7.364
New Zealand	Australia and New Zealand	9	7.286
Australia	Australia and New Zealand	10	7.284

1-10 of 158 rows | 1-4 of 12 columns Previous **1** 2 3 4 5 6 … 16 Next

There are 158 cases in the dataset. A case in represents a country's happiness index.

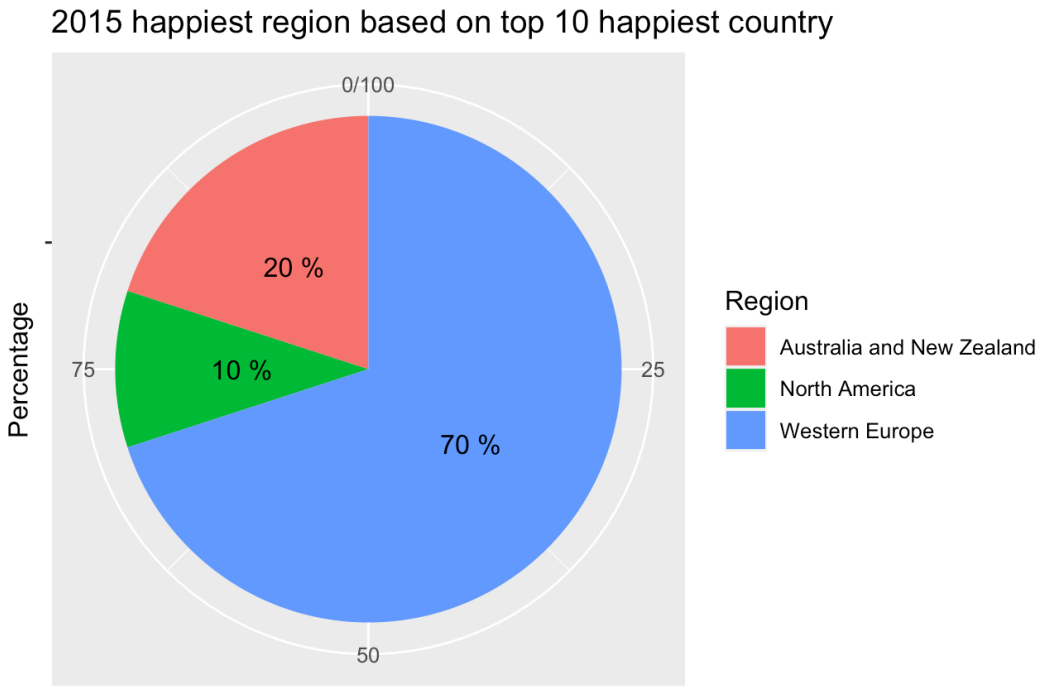
- What are some of the variables that you plan to use?
  I'm going to use `Country`, `Region`, `happiness.score`, `Economy..GDP.per.Capita`, `Family`, `Health..Life.Expectancy`, `Freedom`, `Trust..Government.Corruption`, and `Generosity`

3. Explore intuition related to the research question:

- Create some informative plots and summary statistics

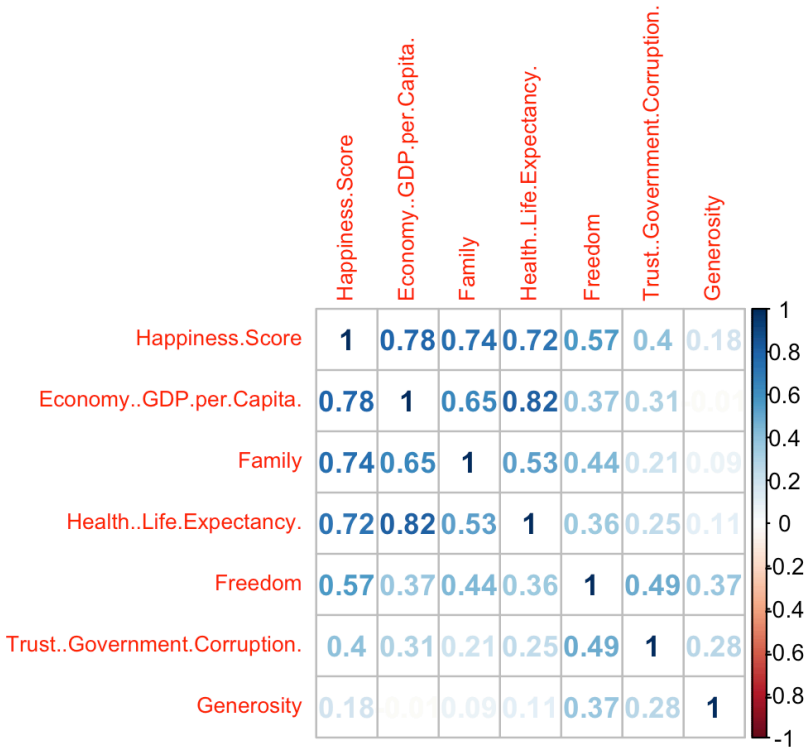
Hide

```
library(DataComputing)
#summary(df2015)

Region <- head(df2015,10) %>%
  group_by(Region) %>%
  summarise(n = n()) %>%
  mutate( percentage = signif(100 * ( n /sum(n)),2))

ggplot(data = Region) +
  geom_bar(mapping  =  aes(x = "", y = percentage, fill = Region), stat = "identity") +
  geom_text ( aes(x = c(1,1,1), y = c(100 -( percentage/2 +c(0,cumsum(percentage)[-length(Region)]))),label = pas
te(percentage,"%"))) +
  ggtitle("2015 happiest region based on top 10 happiest country") +
  labs(x = "Percentage",y = "")+
  coord_polar("y")
```



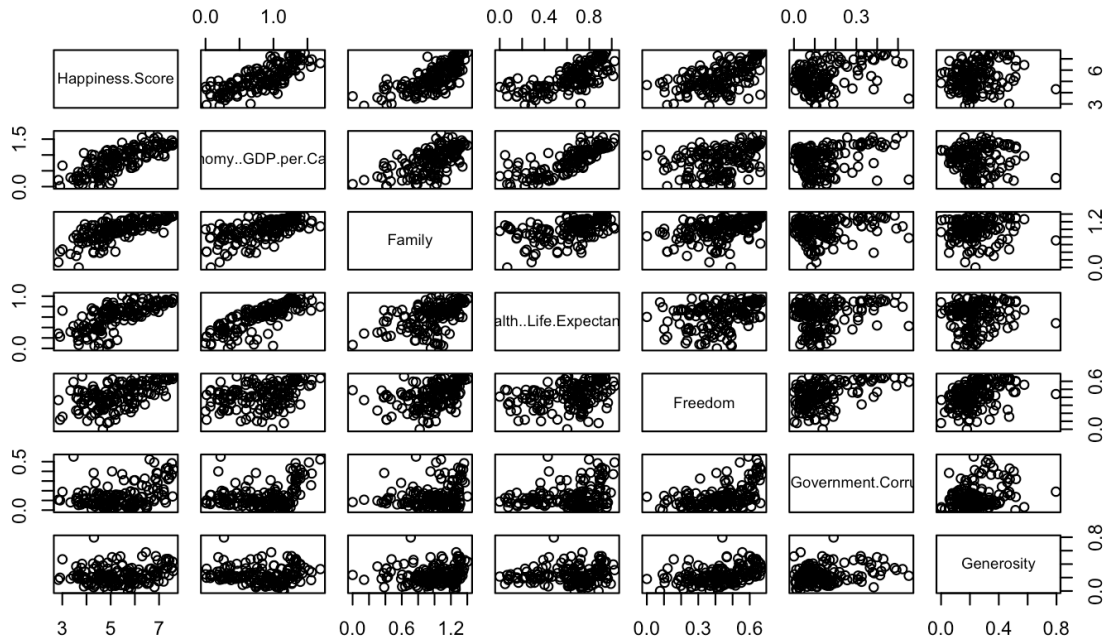
Hide

```
my_data <- df2015[, c(4,6,7,8,9,10,11)]
cormatrix <- cor(my_data)
library(corrplot)
corrplot(cormatrix, method = "number",tl.cex = 0.75)
```



Hide

```
pairs(~Happiness.Score + Economy..GDP.per.Capita. + Family + Health..Life.Expectancy. + Freedom + Trust..Governme
nt.Corruption. + Generosity, data=df2015)
```



- Describe preliminary observations and intuition about the research question

According to the output above, there are some positive relationship between the happiness score and Economy..GDP.per.Capita, Family, Health..Life.Expectancy, Freedom, Trust..Government.Corruption, and Generosity. Among all the variables, happiness score is most related to Economy GDP per Capita. The higher GDP, the higher happiness score. The intuition about the research question is to find out the relations between the variables to come up with the result that which field need to be improved to boost the happiness score and make people happier.

4. TWO Data Sources

I'm going to use the dataset from World Happiness Report official website, which is the world happiness report of 2015. And the other I'm going to use is WorldMap from the DataCOmputing package.

Introduction

In this project, I'm going to compare and contrast 10 countries with highest happiness score and 10 counties with lowest score using the data achieved in 2015 and 2019 (because 2015 was the first year the survey system became relatively completed and the most recent data available was collected in 2019. Considering there is a 5-year gap between these two datasets, the result from the analysis would be convincing and we would be able to use some patterns we observed throughout the project to predict future trend. The datasets are available on World Happiness Report official website. The data was collected from the Gallup World Poll, and supported by the Ernesto Illy Foundation, illycaffè, Davines Group, Blue Chip Foundation, the William, Jeff, and Jennifer Gross Family Foundation, and Unilever's largest ice cream brand Wall's. The survey was conducted yearly after 2012, and the intention was to measure the happiness of people and to use the data to help guide public. policy.

Guiding question: How did the factors impact on the top and bottom 10 happiest countries' happiness score in 2015 and 2019?

Attributes intepretation

Happiness_Rank: Rank of the country based on the Happiness Score
Country: Name of the country
Happiness_Score: A metric measured in 2015 and 2019 by asking the sampled people the question: "How would you rate your happiness"
Economy_GDP_per_Capita: The extent to which GDP contributes to the calculation of the Happiness Score
Social_Support: The extent to which Family contributes to the calculation of the Happiness Score
Health_Life_Expectancy: The extent to which Life expectancy contributed to the calculation of the Happiness Score
Freedom: The extent to which Freedom contributed to the calculation of the Happiness Score.
Generosity: The extent to which Generosity contributed to the calculation of the Happiness Score.
Trust_Government_Corruption: The extent to which Perception of Corruption contributes to Happiness Score.

Data Access

Hide

```
# Loading packages
library(DataComputing)
library(mosaic)
library(tidyverse)
library(lubridate)
library(ggplot2)
library(ggalt)
library(corrplot)
library(plotly)
library(gridExtra)
library(caTools)
library(rpart)
library(rpart.plot)
```

Hide

```
# Data sources
df2015 <- read.csv("2015.csv")
df2019 <- read.csv("2019.csv")

# Data inspection
head(df2015, 10)
```

	Country <chr>	Region <chr>	Happiness.Rank <int>	Happiness.Score <dbl>	Standard.Error <dbl>
1	Switzerland	Western Europe	1	7.587	0.03411
2	Iceland	Western Europe	2	7.561	0.04884
3	Denmark	Western Europe	3	7.527	0.03328
4	Norway	Western Europe	4	7.522	0.03880
5	Canada	North America	5	7.427	0.03553
6	Finland	Western Europe	6	7.406	0.03140
7	Netherlands	Western Europe	7	7.378	0.02799
8	Sweden	Western Europe	8	7.364	0.03157
9	New Zealand	Australia and New Zealand	9	7.286	0.03371
10	Australia	Australia and New Zealand	10	7.284	0.04083

1-10 of 10 rows | 1-6 of 12 columns

Hide

```
head(df2019, 10)
```

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

1-10 of 10 rows | 1-7 of 9 columns

Hide

```
summary(df2015)
```

```
   Country             Region          Happiness.Rank   Happiness.Score
 Length:158         Length:158         Min.   :  1.00   Min.   :2.839
 Class :character   Class :character   1st Qu.: 40.25   1st Qu.:4.526
 Mode  :character   Mode  :character   Median : 79.50   Median :5.232
                                       Mean   : 79.49   Mean   :5.376
                                       3rd Qu.:118.75   3rd Qu.:6.244
                                       Max.   :158.00   Max.   :7.587
 Standard.Error    Economy..GDP.per.Capita.     Family       Health..Life.Expectancy.
 Min.   :0.01848   Min.   :0.0000           Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.03727   1st Qu.:0.5458           1st Qu.:0.8568   1st Qu.:0.4392
 Median :0.04394   Median :0.9102           Median :1.0295   Median :0.6967
 Mean   :0.04788   Mean   :0.8461           Mean   :0.9910   Mean   :0.6303
 3rd Qu.:0.05230   3rd Qu.:1.1584           3rd Qu.:1.2144   3rd Qu.:0.8110
 Max.   :0.13693   Max.   :1.6904           Max.   :1.4022   Max.   :1.0252
    Freedom       Trust..Government.Corruption.   Generosity     Dystopia.Residual
 Min.   :0.0000   Min.   :0.00000               Min.   :0.0000   Min.   :0.3286
 1st Qu.:0.3283   1st Qu.:0.06168               1st Qu.:0.1506   1st Qu.:1.7594
 Median :0.4355   Median :0.10722               Median :0.2161   Median :2.0954
 Mean   :0.4286   Mean   :0.14342               Mean   :0.2373   Mean   :2.0990
 3rd Qu.:0.5491   3rd Qu.:0.18025               3rd Qu.:0.3099   3rd Qu.:2.4624
 Max.   :0.6697   Max.   :0.55191               Max.   :0.7959   Max.   :3.6021
```

Hide

```
summary(df2019)
```

```
  Overall.rank    Country.or.region      Score       GDP.per.capita   Social.support
 Min.   :  1.00   Length:156         Min.   :2.853   Min.   :0.0000   Min.   :0.000
 1st Qu.: 39.75   Class :character   1st Qu.:4.545   1st Qu.:0.6028   1st Qu.:1.056
 Median : 78.50   Mode  :character   Median :5.380   Median :0.9600   Median :1.272
 Mean   : 78.50                      Mean   :5.407   Mean   :0.9051   Mean   :1.209
 3rd Qu.:117.25                      3rd Qu.:6.184   3rd Qu.:1.2325   3rd Qu.:1.452
 Max.   :156.00                      Max.   :7.769   Max.   :1.6840   Max.   :1.624
 Healthy.life.expectancy Freedom.to.make.life.choices   Generosity
 Min.   :0.0000          Min.   :0.0000                Min.   :0.0000
 1st Qu.:0.5477          1st Qu.:0.3080                1st Qu.:0.1087
 Median :0.7890          Median :0.4170                Median :0.1775
 Mean   :0.7252          Mean   :0.3926                Mean   :0.1848
 3rd Qu.:0.8818          3rd Qu.:0.5072                3rd Qu.:0.2482
 Max.   :1.1410          Max.   :0.6310                Max.   :0.5660
 Perceptions.of.corruption
 Min.   :0.0000
 1st Qu.:0.0470
 Median :0.0855
 Mean   :0.1106
 3rd Qu.:0.1412
 Max.   :0.4530
```

Hide

```
nrow(df2015)
```

```
[1] 158
```

Hide

```
ncol(df2015)
```

```
[1] 12
```

Hide

```
nrow(df2019)
```

```
[1] 156
```

Hide

```
ncol(df2019)
```

```
[1] 9
```

Observation: There are 158 entries and 12 columns in df2015; 156 entries and 9 columns in df2019.
Used ‘summary’ function to get general statistics, such as mean, median, minimum and maximum score.

I noticed that although the number of rows and columns are not the same in df2015 and df2019, there are not many differences, and I'm going to select the features they both have for future analysis.
Features appear in both tables: happiness rank, country, happiness score, GDP, family/social support, Healthy life expectancy, freedom, generosity, and perception of corruption.

Data Wrangling

Hide

```
# Create column 'Year'
Year <- c(2015)
Year2015 <- data.frame(Year, df2015)
# Rename the columns to formalize
names(Year2015) <- c("Year", "Country", " Region", "Happiness_Rank","Happiness_Score","Standard_Error",  "Economy
_GDP_per_Capita","Social_Support","Health_Life_Expectancy","Freedom","Trust_Government_Corruption", "Generosity",
"Dystopia_Residual")


Year2 <- c(2019)
Year2019 <- data.frame(Year2, df2019)
names(Year2019) <- c("Year","Happiness_Rank","Country","Happiness_Score","Economy_GDP_per_Capita","Social_Support
","Health_Life_Expectancy","Freedom","Generosity","Trust_Government_Corruption")

New2015 <- Year2015 %>%
  select("Happiness_Rank","Country","Happiness_Score","Economy_GDP_per_Capita","Social_Support","Health_Life_Expe
ctancy","Freedom","Generosity","Trust_Government_Corruption", "Year")

# convert character to numeric
# combine data from 2015 and 2019 for future use
FinalTable <- rbind(New2015, Year2019, by= "Country")
FinalTable[3:10] <- lapply(FinalTable[3:10], as.numeric)
head(New2015)
```

	Happiness_Rank <int>	Country <chr>	Happiness_Score <dbl>	Economy_GDP_per_Capita <dbl>	Social_Support <dbl>
1	1	Switzerland	7.587	1.39651	1.34951
2	2	Iceland	7.561	1.30232	1.40223
3	3	Denmark	7.527	1.32548	1.36058
4	4	Norway	7.522	1.45900	1.33095
5	5	Canada	7.427	1.32629	1.32261
6	6	Finland	7.406	1.29025	1.31826

6 rows | 1-6 of 10 columns

Hide

```
head(Year2019)
```

	Year <dbl>	Happiness_Rank <int>	Country <chr>	Happiness_Score <dbl>	Economy_GDP_per_Capita <dbl>	Social_Support <dbl>
1	2019	1	Finland	7.769	1.340	1.587
2	2019	2	Denmark	7.600	1.383	1.573
3	2019	3	Norway	7.554	1.488	1.582
4	2019	4	Iceland	7.494	1.380	1.624
5	2019	5	Netherlands	7.488	1.396	1.522
6	2019	6	Switzerland	7.480	1.452	1.526

6 rows | 1-7 of 10 columns

Hide

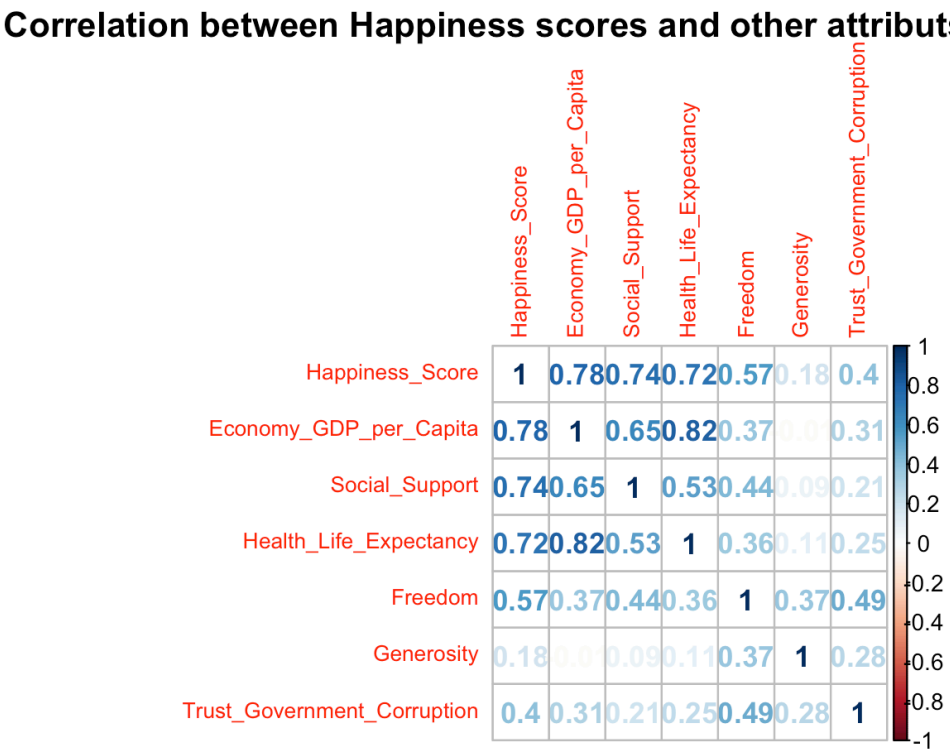
```
FinalTable <- drop_na(FinalTable)
data2015 <- FinalTable %>%
  filter(Year=="2015")

data2019 <- FinalTable %>%
  filter(Year=="2019")
```

Data Visualization

Overall correlation among variables in 2015 and 2019

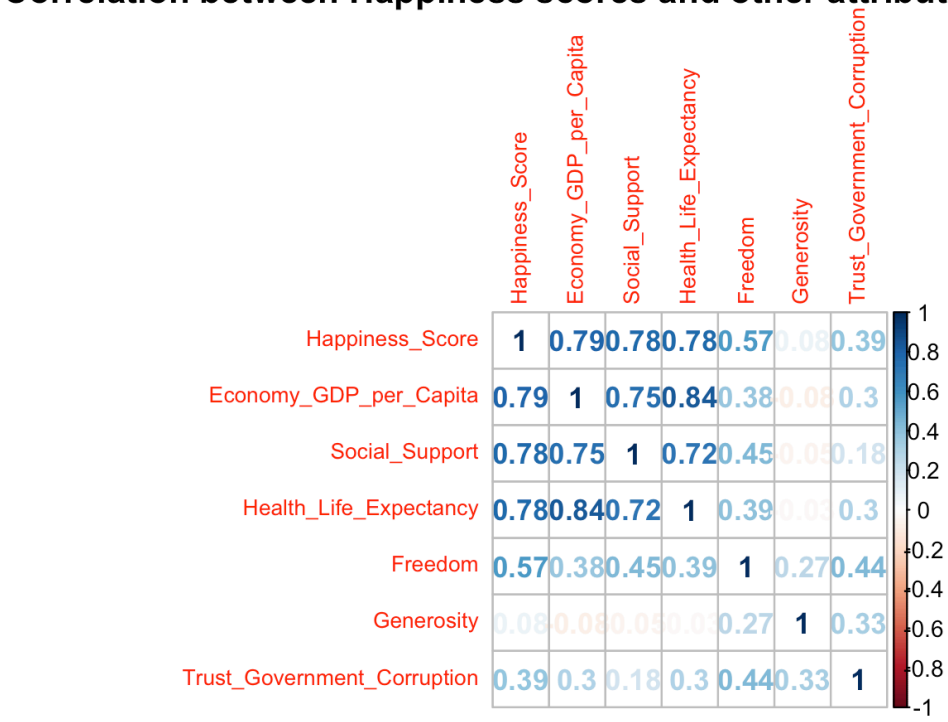
Hide

```
cormatrix2015 <- cor(data2015[3:9],use="pairwise.complete.obs")
corrplot(cormatrix2015, method = "number",tl.cex = 0.75, title="Correlation between Happiness scores and other at
tributs 2015",mar=c(0,0,1,0))
```



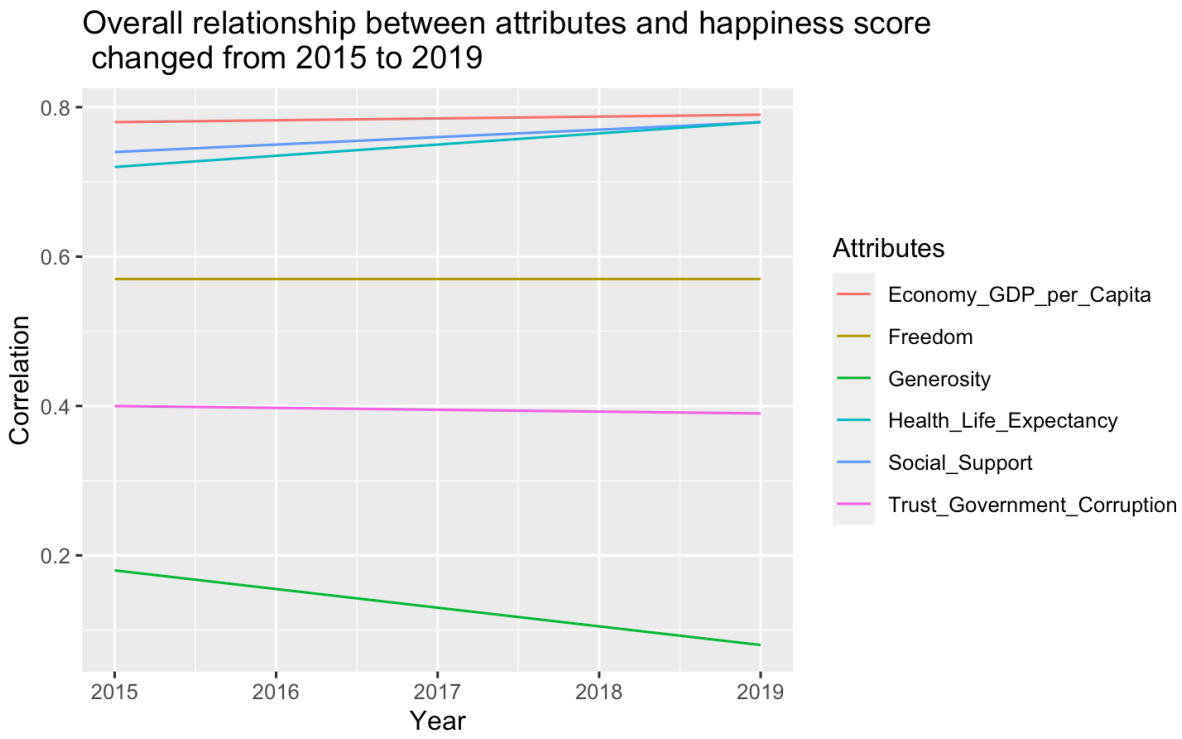
Hide

```
cormatrix2019 <- cor(data2019[3:9],use="pairwise.complete.obs")
corrplot(cormatrix2019, method = "number",tl.cex = 0.75, title="Correlation between Happiness scores and other at
tributs 2019",mar=c(0,0,1,0))
```



Hide

```
Overall <- data.frame(x0 = c(2015,2015,2015,2015,2015,2015),
                      y0 = c(0.78,0.74,0.72,0.57,0.18,0.4),
                      x1 = c(2019,2019,2019,2019,2019,2019),
                      y1 = c(0.79,0.78,0.78,0.57, 0.08, 0.39))
Overall$Attributes <- c('Economy_GDP_per_Capita','Social_Support','Health_Life_Expectancy','Freedom', 'Generosity
', 'Trust_Government_Corruption ')

ggplot(Overall) +
    geom_segment(aes(x = x0,y = y0,xend = x1,yend = y1,colour = Attributes)) + ggtitle("Overall relationship betw
een attributes and happiness score \n changed from 2015 to 2019") + xlab("Year") + ylab("Correlation")
```



According to the graph, we can see that, as time goes by, the correlation between generosity and happiness score decreases, meaning that generosity doesn't impact on people's happiness as much as before. Meanwhile, the positive relationship between health life expectancy and happiness score, and social support and happiness score are increasing.

Side-by-side boxplot to show how factors related to happiness score have changed over time:

Hide

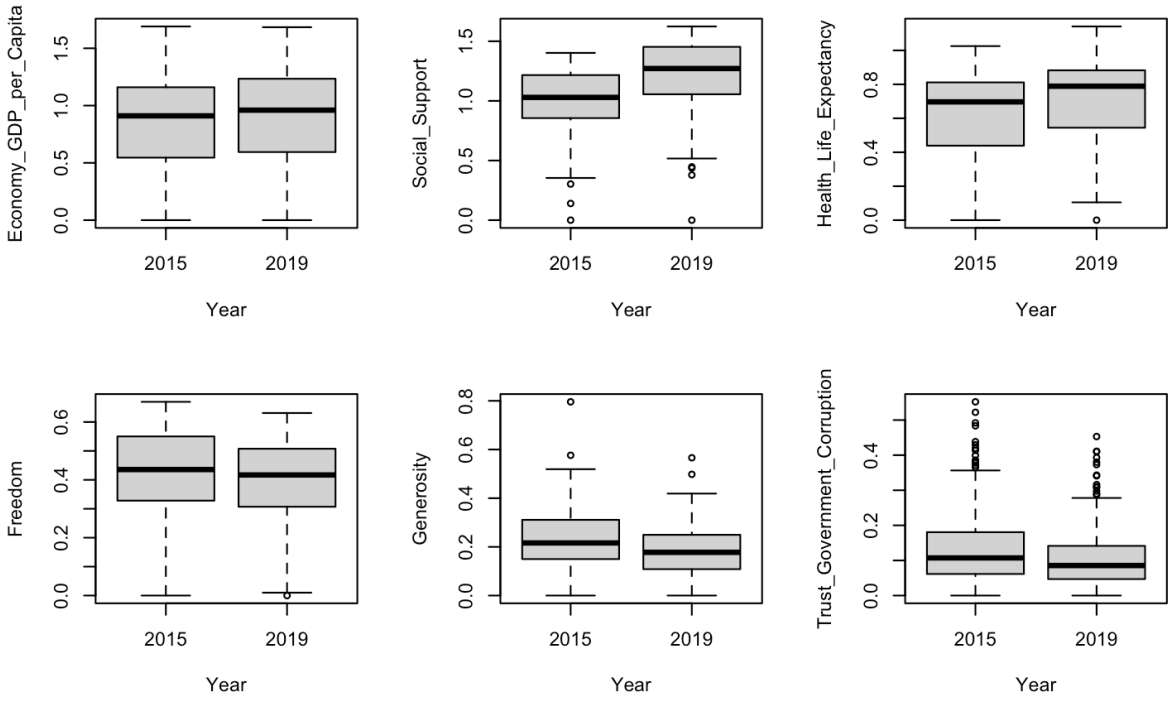
```
par(mfrow=c(2,3))
boxplot(Economy_GDP_per_Capita~Year,data=FinalTable)
boxplot(Social_Support~Year,data=FinalTable)
```

Hide

```
boxplot(Health_Life_Expectancy~Year,data=FinalTable)
boxplot(Freedom~Year,data=FinalTable)
```

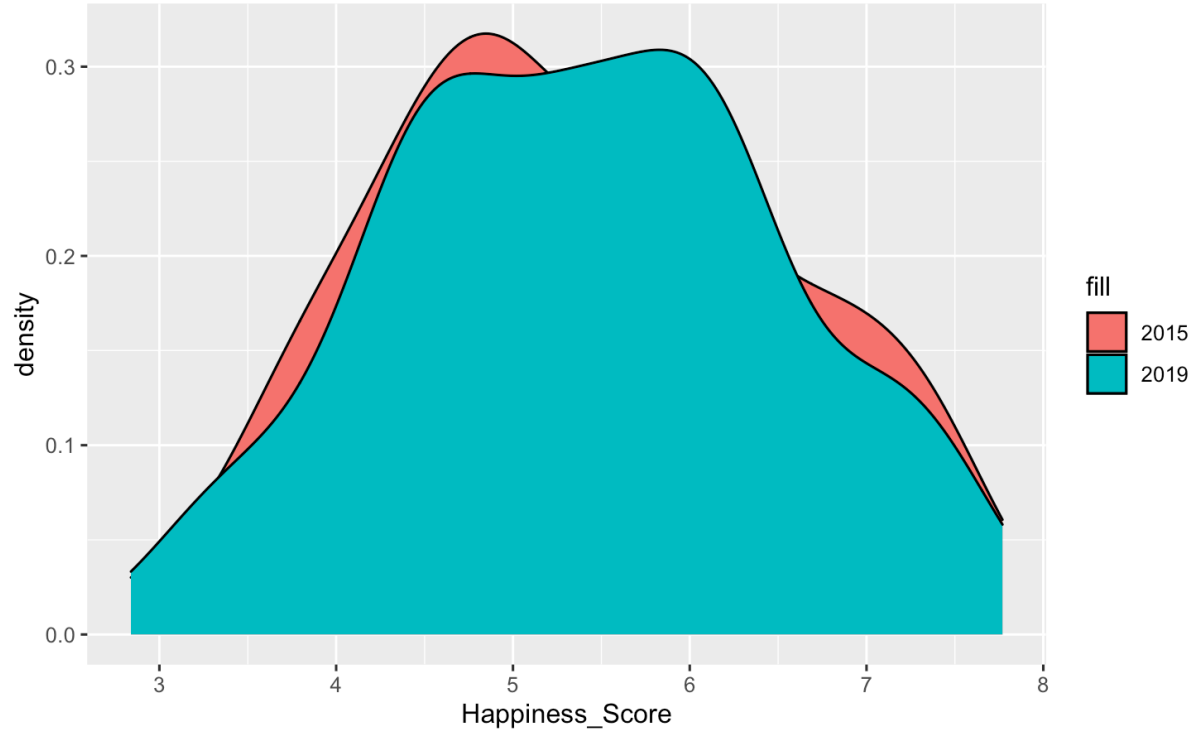
Hide

```
boxplot(Generosity~Year,data=FinalTable)
boxplot(Trust_Government_Corruption~Year,data=FinalTable)
```



Hide

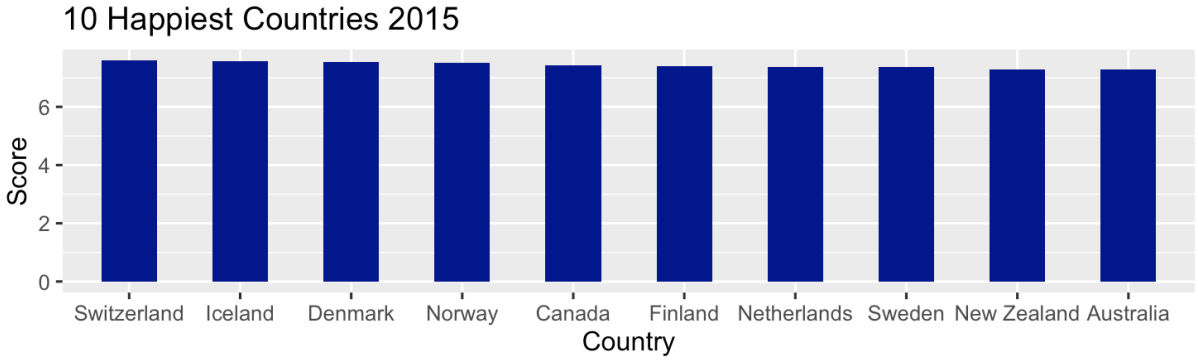
```
# plot the density of happiness score for 2015 and 2019
ggplot(data2015, aes(x = Happiness_Score, fill="2015")) + geom_density()+ geom_density(data=data2019, aes(x=Happi
ness_Score, fill="2019"))
```

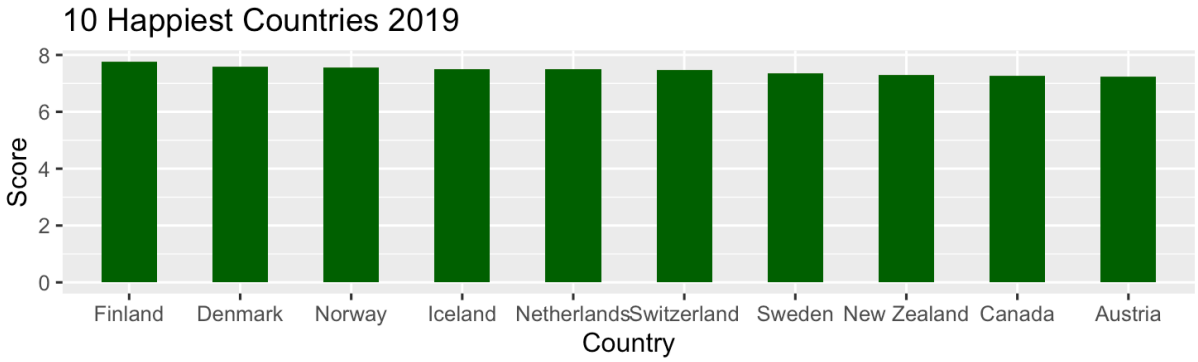


Now, let's list 10 happiest and 10 unhappiest countries in 2015 and 2019.

Hide

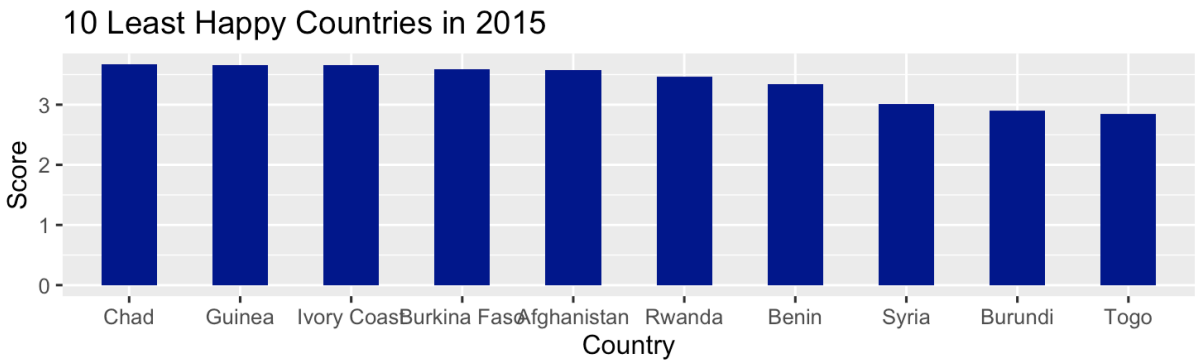
```
#top 10
Top2015 <- data2015 %>% head(n=10)
Top2019 <- data2019 %>% head(n=10)


Top2015Graph <- ggplot(Top2015,aes(x=factor(Country, levels=Country), y=Happiness_Score)) + geom_bar(stat="identi
ty", width = 0.5, fill="darkblue") +labs(title="10 Happiest Countries 2015", x="Country", y="Score")
Top2019Graph <- ggplot(Top2019,aes(x=factor(Country, levels=Country), y=Happiness_Score)) +geom_bar(stat="identit
y", width = 0.5, fill="darkgreen")+labs(title="10 Happiest Countries 2019", x="Country", y="Score")

# Bottom 10
Bottom2015 <- data2015 %>% tail(n=10)
Bottom2019 <- data2019 %>% tail(n=10)

Bottom2015Graph <- ggplot(Bottom2015,aes(x=factor(Country, levels=Country), y=Happiness_Score)) + geom_bar(stat="
identity", width = 0.5, fill="darkblue") +labs(title="10 Least Happy Countries in 2015", x="Country", y="Score")
Bottom2019Graph <- ggplot(Bottom2019,aes(x=factor(Country, levels=Country), y=Happiness_Score)) +geom_bar(stat="i
dentity", width = 0.5, fill="darkgreen")+labs(title="10 Least Happy Countries in 2019", x="Country", y="Score")

grid.arrange(Top2015Graph, Top2019Graph)
```

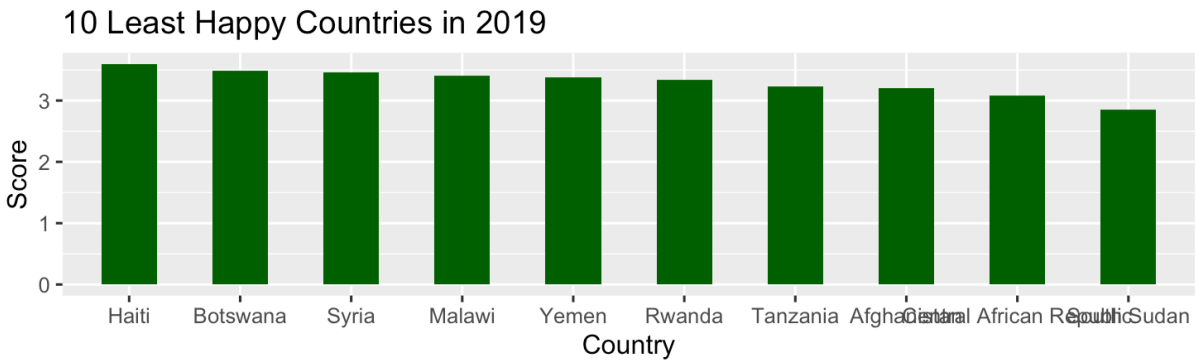




Hide

```
grid.arrange(Bottom2015Graph, Bottom2019Graph)
```





Hide

```
# Mean of happiness score for top and bottom 10 happiest countries from 2015 to 2019
mean(Top2015$Happiness_Score)
```

```
[1] 7.4342
```

Hide

```
mean(Bottom2015$Happiness_Score)
```

```
[1] 3.3695
```

Hide

```
mean(Top2019$Happiness_Score)
```

```
[1] 7.4559
```

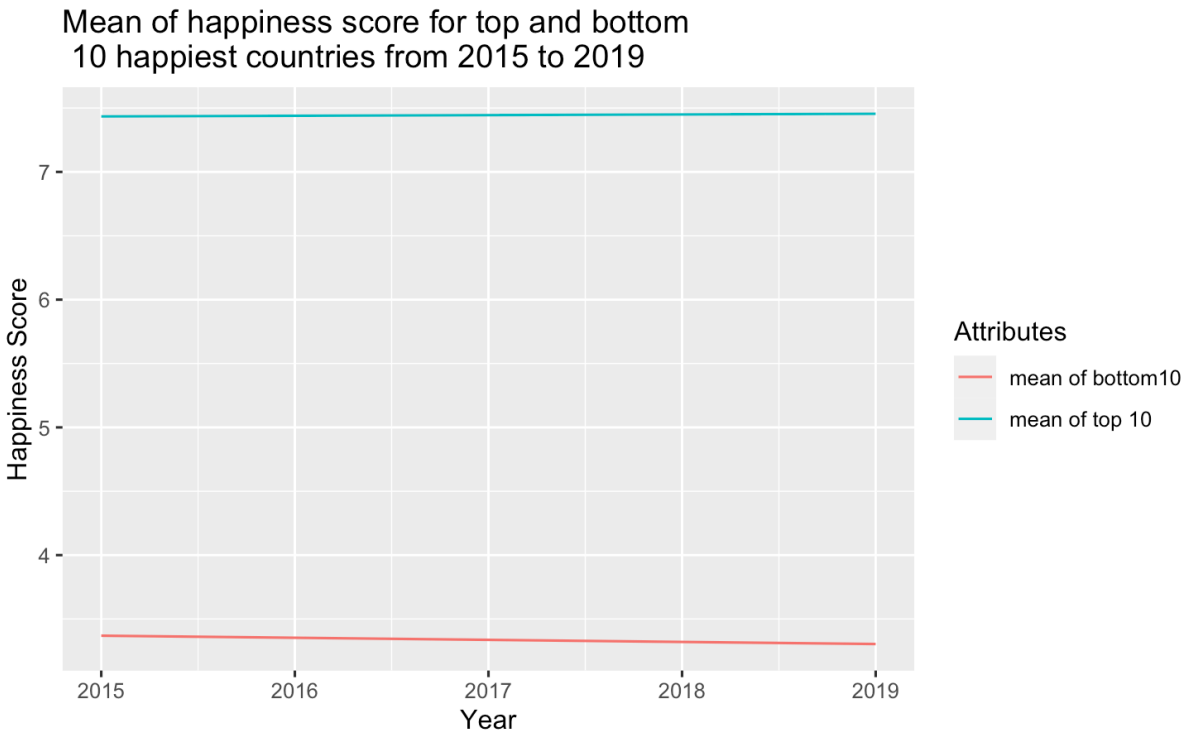
Hide

```
mean(Bottom2019$Happiness_Score)
```

```
[1] 3.3041
```

Hide

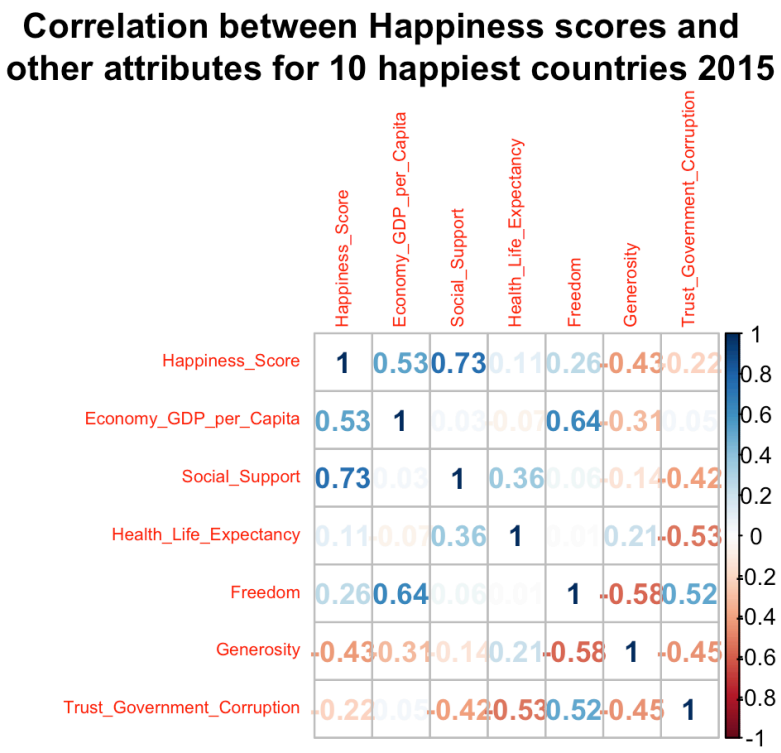
```
Change <- data.frame(x0 = c(2015,2015),
                     y0 = c(7.4342, 3.3695),
                     x1 = c(2019,2019),
                     y1 = c(7.4559, 3.3041))
Change$Attributes <- c('mean of top 10','mean of bottom10')

ggplot(Change) +
    geom_segment(aes(x = x0,y = y0,xend = x1,yend = y1,colour = Attributes)) + ggtitle("Mean of happiness score f
or top and bottom \n 10 happiest countries from 2015 to 2019") + xlab("Year") + ylab("Happiness Score")
```



The mean happiness score for the top 10 happiest countries has slightly improved as time goes by, while the mean of the bottom 10 happiest countries has slightly decreased.

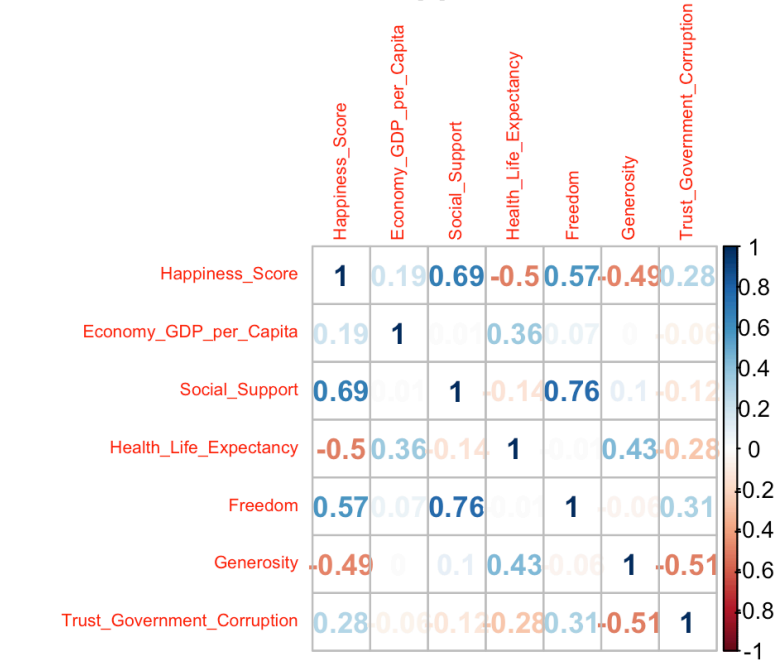
Hide

```
cormatrix2015T <- cor(Top2015[3:9],use="pairwise.complete.obs")
corrplot(cormatrix2015T, method = "number",tl.cex = 0.6, title="Correlation between Happiness scores and \n other
attributes for 10 happiest countries 2015",mar=c(0,0,2.5,0))
```



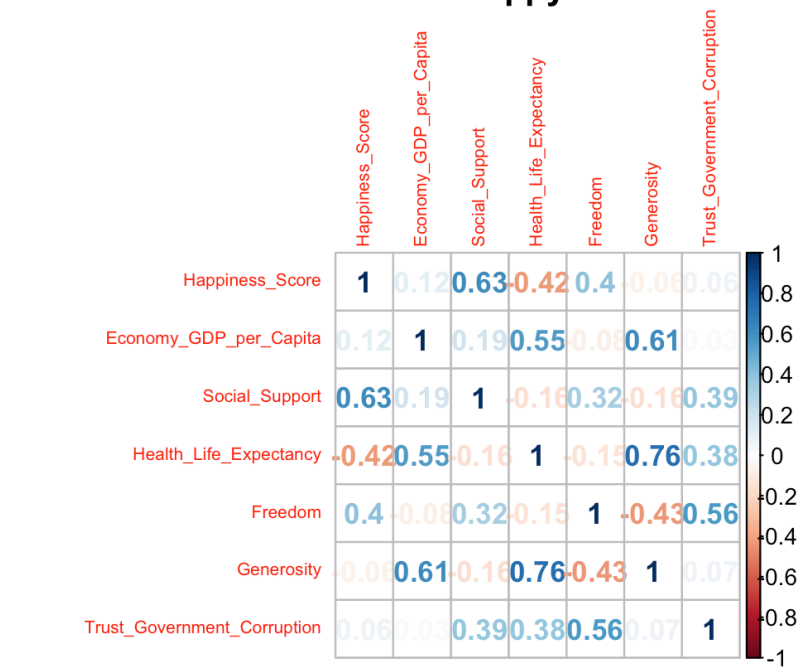
Hide

```
cormatrix2019T <- cor(Top2019[3:9],use="pairwise.complete.obs")
corrplot(cormatrix2019T, method = "number",tl.cex = 0.6, title="Correlation between Happiness scores and \n other
attributes for 10 happiest countries 2019",mar=c(0,0,2.5,0))
```



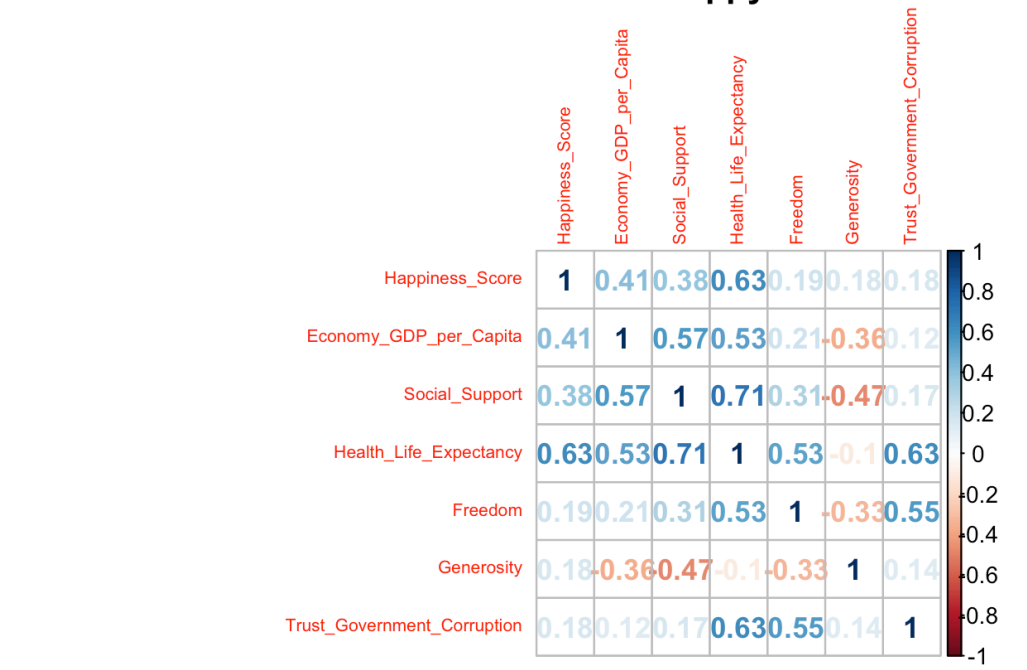
Hide

```
cormatrix2015B <- cor(Bottom2015[3:9],use="pairwise.complete.obs")
corrplot(cormatrix2015B, method = "number",tl.cex = 0.6, title="Correlation between Happiness scores and \n other
attributes for 10 least happy countries 2015",mar=c(0,0,2.5,0))
```



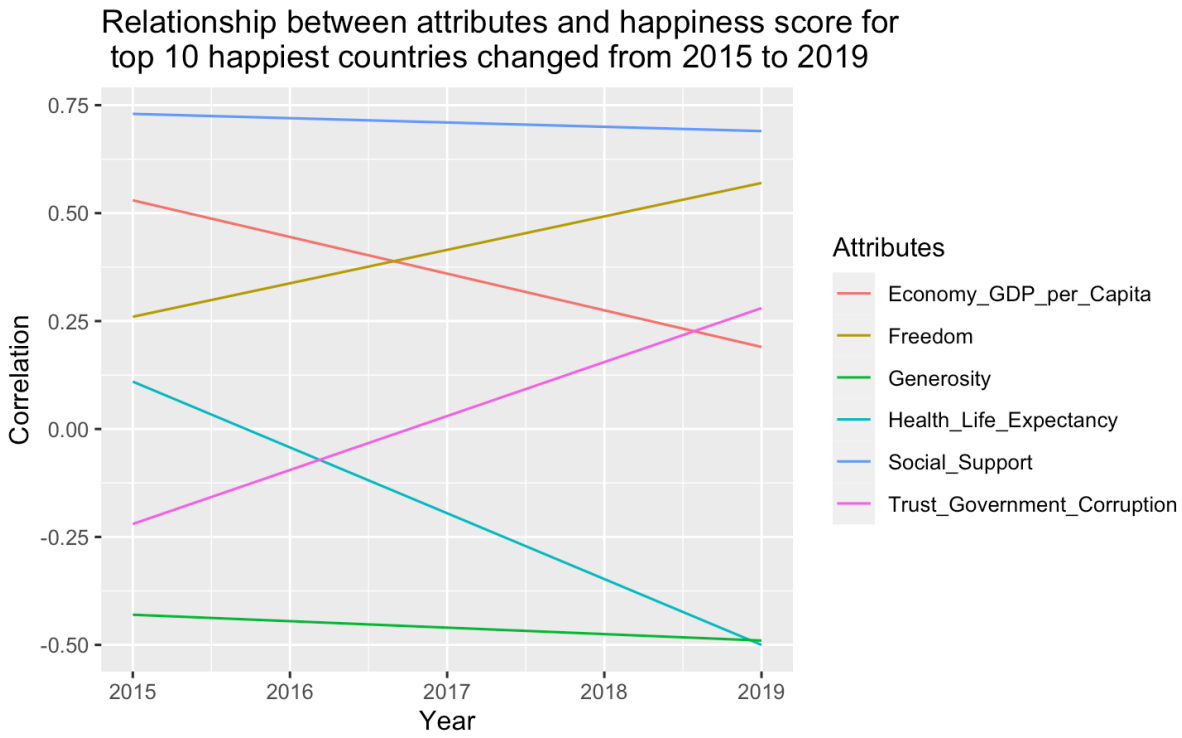
Hide

```
cormatrix2019B <- cor(Bottom2019[3:9],use="pairwise.complete.obs")
corrplot(cormatrix2019B, method = "number",tl.cex = 0.6, title="Correlation between Happiness scores and \n other
attributes for 10 least happy countries 2019",mar=c(0,0,2.5,0))
```



Hide

```
Happiest1 <- data.frame(x0 = c(2015,2015,2015,2015,2015,2015),
                        y0 = c(0.53, 0.73,0.11,0.26, -0.43, -0.22),
                        x1 = c(2019,2019,2019,2019,2019,2019),
                        y1 = c(0.19, 0.69, -0.5, 0.57, -0.49, 0.28))
Happiest1$Attributes <- c('Economy_GDP_per_Capita','Social_Support','Health_Life_Expectancy','Freedom', 'Generosi
ty', 'Trust_Government_Corruption ')

ggplot(Happiest1) +
    geom_segment(aes(x = x0,y = y0,xend = x1,yend = y1,colour = Attributes)) + ggtitle("Relationship between attr
ibutes and happiness score for \n top 10 happiest countries changed from 2015 to 2019") + xlab("Year") + ylab("Co
rrelation")
```

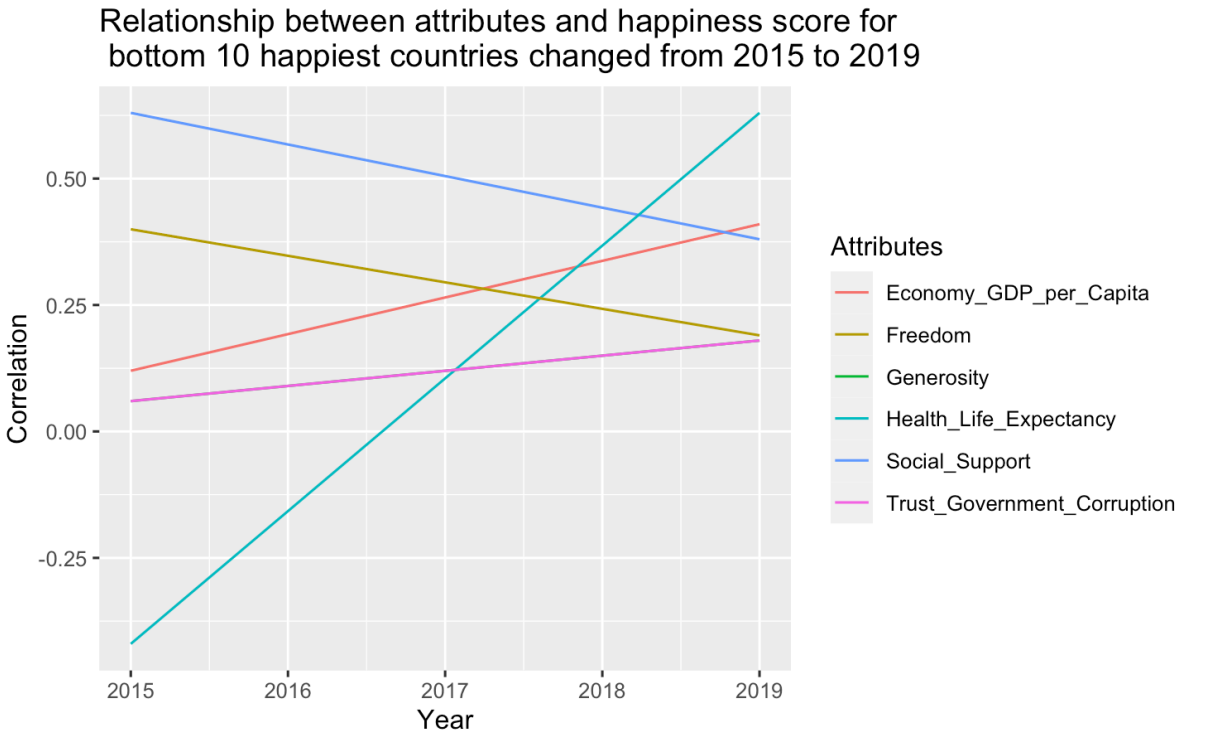


For top 10 happiest countries from 2015 to 2019:

There's a positive relationship between social support and happiness score, but it has slightly decreased overtime.
There's a positive relationship between GDP and happiness score but it has significantly dropped overtime.
There's a positive relationship between freedom and happiness score, but it has increased overtime.
There was a weak positive relationship between health life expectancy and happiness score when in 2015, but the correlation has become negative since 2016. On the contrast, the correlation between trust government corruption and happiness score has become positive in 2019 while it was negative in 2015.
There's negative relationship between generosity and happiness score and it has slightly decreased as time goes by.

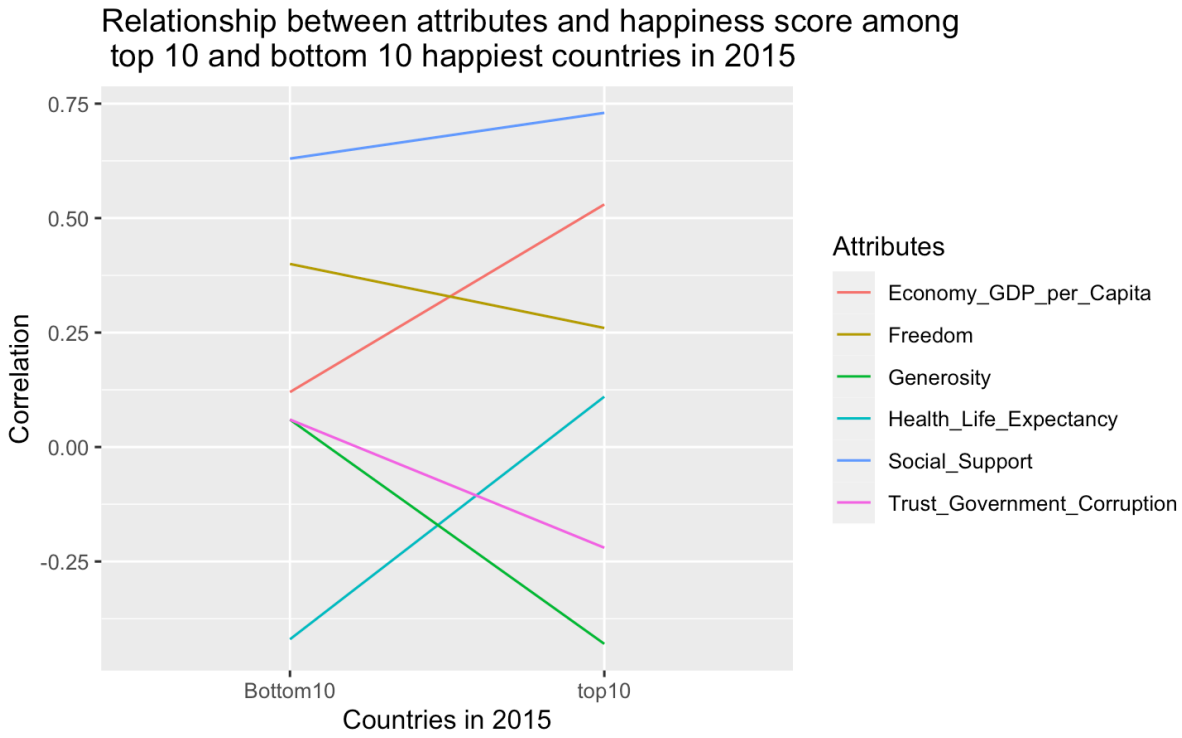
Hide

```
Least1 <- data.frame(x0 = c(2015,2015,2015,2015,2015,2015),
                     y0 = c(0.12,0.63,-0.42,0.4,0.06,0.06),
                     x1 = c(2019,2019,2019,2019,2019,2019),
                     y1 = c(0.41,0.38,0.63,0.19,0.18,0.18))
Least1$Attributes <- c('Economy_GDP_per_Capita','Social_Support','Health_Life_Expectancy','Freedom', 'Generosity'
, 'Trust_Government_Corruption ')

ggplot(Least1) +
    geom_segment(aes(x = x0,y = y0,xend = x1,yend = y1,colour = Attributes)) + ggtitle("Relationship between attr
ibutes and happiness score for \n bottom 10 happiest countries changed from 2015 to 2019") + xlab("Year") + ylab(
"Correlation")
```



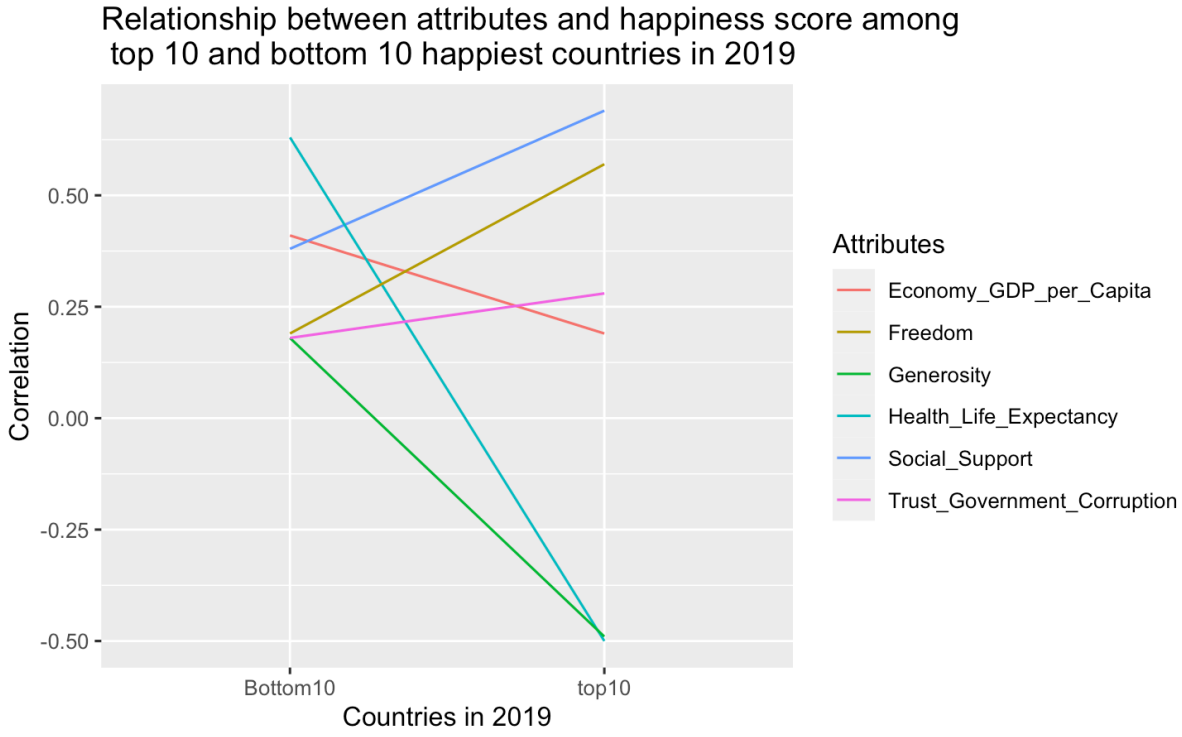
Hide

```
H_and_L_2015 <- data.frame(x0 = c("top10","top10","top10","top10","top10","top10"),
                     y0 = c(0.53, 0.73,0.11,0.26, -0.43, -0.22),
                     x1 = c("Bottom10","Bottom10","Bottom10","Bottom10","Bottom10","Bottom10"),
                     y1 = c(0.12,0.63,-0.42,0.4,0.06,0.06))
H_and_L_2015$Attributes <- c('Economy_GDP_per_Capita','Social_Support','Health_Life_Expectancy','Freedom', 'Gener
osity', 'Trust_Government_Corruption ')

ggplot(H_and_L_2015) +
    geom_segment(aes(x = x0,y = y0,xend = x1,yend = y1,colour = Attributes)) + ggtitle("Relationship between attr
ibutes and happiness score among \n top 10 and bottom 10 happiest countries in 2015") + xlab("Countries in 2015")
+ ylab("Correlation")
```



Hide

```
H_and_L_2019 <- data.frame(x0 = c("top10","top10","top10","top10","top10","top10"),
                      y0 = c(0.19, 0.69, -0.5, 0.57, -0.49, 0.28),
                      x1 = c("Bottom10","Bottom10","Bottom10","Bottom10","Bottom10","Bottom10"),
                      y1 = c(0.41,0.38,0.63,0.19,0.18,0.18))
H_and_L_2019$Attributes <- c('Economy_GDP_per_Capita','Social_Support','Health_Life_Expectancy','Freedom', 'Gener
osity', 'Trust_Government_Corruption ')

ggplot(H_and_L_2019) +
    geom_segment(aes(x = x0,y = y0,xend = x1,yend = y1,colour = Attributes)) + ggtitle("Relationship between attr
ibutes and happiness score among \n top 10 and bottom 10 happiest countries in 2019") + xlab("Countries in 2019")
+ ylab("Correlation")
```



In 2019, there is a positive relationship between Health_Life_Expectancy and Happiness score for the bottom 10 countries, whereas the relationship between Health_Life_Expectancy and Happiness score for the top 10 countries is negative. To specify, in 2019, for bottom 10 countries, the higher the Health_Life_Expectancy is, the higher the happiness score will be. For top 10 countries, the higher

health_Life_Expectancy matches with lower happiness score. Similarly, the correlation between generosity and happiness score for bottom 10 happiest countries is positive while that of top 10 countries in 2019 is negative.

Machine learning models

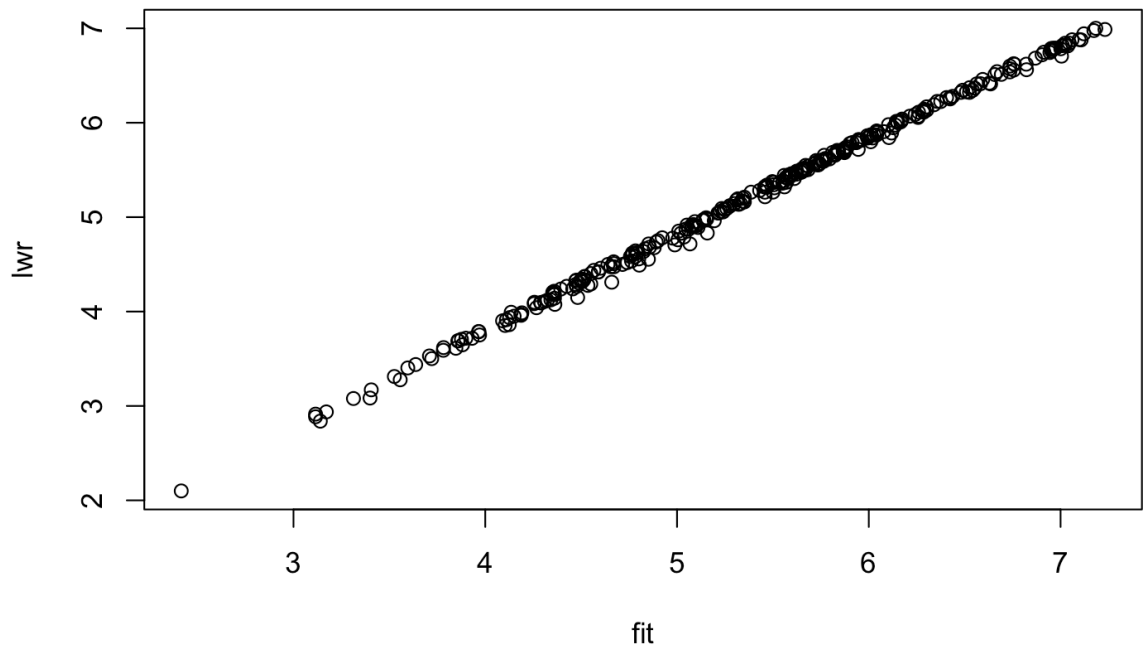
Hide

```
#install.packages('caTools')
set.seed(200)
dataset <- FinalTable[3:10]
split = sample.split(dataset$Happiness_Score, SplitRatio = 0.8)
training_set = subset(dataset, split == T)
test_set = subset(dataset, split == F)
dataFit = lm(formula = Happiness_Score ~ ., data = training_set)

summary(dataFit)
```

```

Call:
lm(formula = Happiness_Score ~ ., data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.60263 -0.33990  0.03263  0.33817  1.45939 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                122.43808   39.31197   3.115 0.002064 ** 
Economy_GDP_per_Capita       0.78789    0.16548   4.761 3.31e-06 ***
Social_Support               1.17843    0.16735   7.042 1.94e-11 ***
Health_Life_Expectancy       0.92656    0.23745   3.902 0.000123 ***
Freedom                      1.66207    0.29174   5.697 3.51e-08 ***
Generosity                   0.24526    0.31449   0.780 0.436227    
Trust_Government_Corruption  0.85120    0.36446   2.335 0.020332 *  
Year                        -0.05973    0.01951  -3.062 0.002447 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.523 on 243 degrees of freedom
Multiple R-squared:  0.7749,    Adjusted R-squared:  0.7684 
F-statistic: 119.5 on 7 and 243 DF,  p-value: < 2.2e-16
```

Hide

```
a <- predict(dataFit,dataset,interval="confidence",level=0.95)
plot(a)
```



Hide

```
anova(dataFit)
```

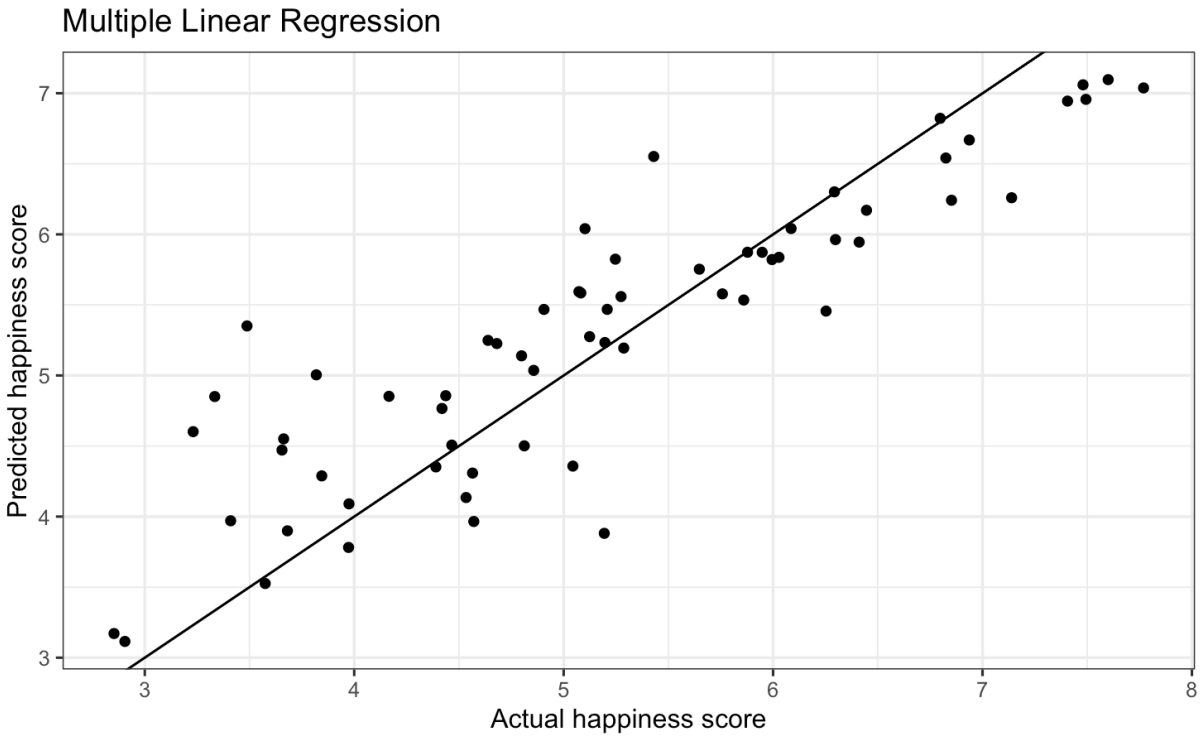
```
Analysis of Variance Table

Response: Happiness_Score
                            Df  Sum Sq Mean Sq  F value    Pr(>F)
Economy_GDP_per_Capita       1 179.051 179.051 654.7094 < 2.2e-16 ***
Social_Support               1  18.005  18.005  65.8377 2.437e-14 ***
Health_Life_Expectancy       1   3.823   3.823  13.9796 0.0002305 ***
Freedom                      1  22.760  22.760  83.2239 < 2.2e-16 ***
Generosity                   1   0.873   0.873   3.1907 0.0753040 .
Trust_Government_Corruption  1   1.674   1.674   6.1201 0.0140499 *
Year                         1   2.564   2.564   9.3747 0.0024473 **
Residuals                  243  66.456   0.273
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The equation of the multiple linear regression line $\hat{HappinessScore} = 122.43808 + 0.78789 EconomyGDPperCapita + 1.17843 SocialSupport + 0.92656 HealthLifeExpectancy + 1.66207 Freedom + 0.24526 Generosity + 0$

$R^2 = 0.7749$, meaning about 77.49% of the variation in happiness score is captured by the regression line based on the 6 variables we chose above.

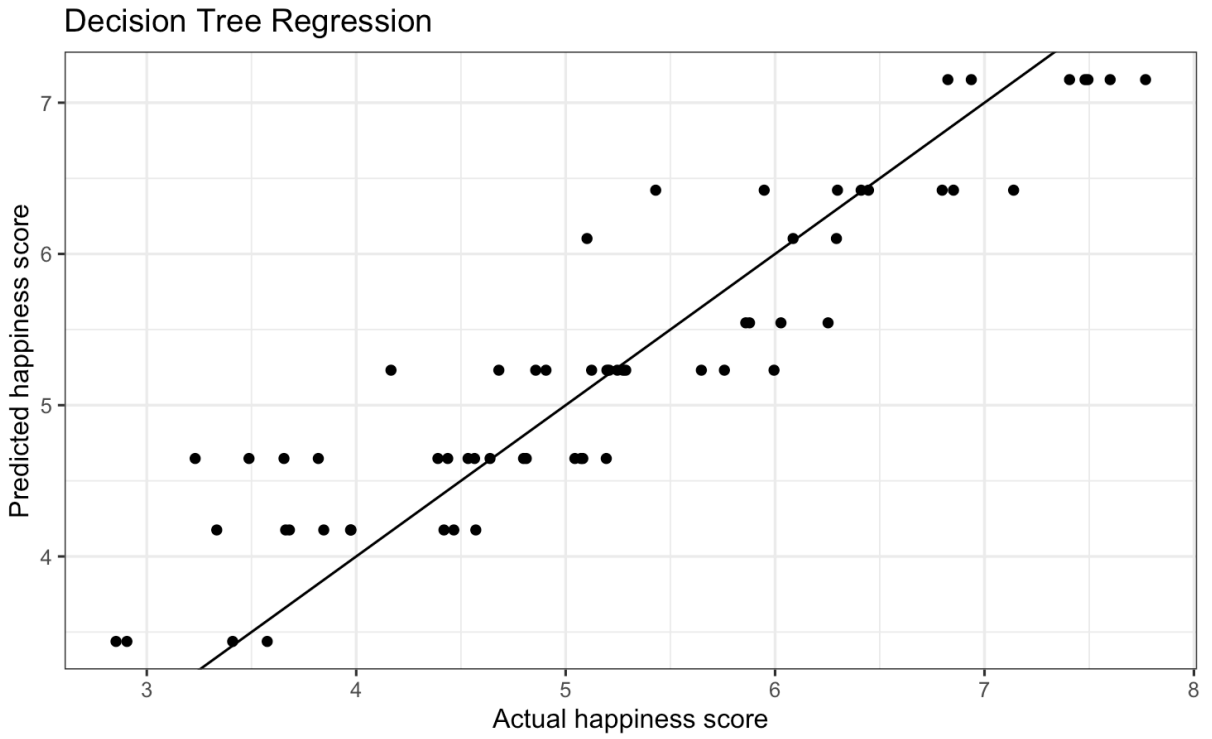
Hide

```
PredYFit = predict(dataFit, newdata = test_set)

PredActualFit <- as.data.frame(cbind(Prediction = PredYFit, Actual = test_set$Happiness_Score))

FitGraph <- ggplot(PredActualFit, aes(Actual, Prediction)) +
  geom_point() + theme_bw() + geom_abline() +
  labs(title = "Multiple Linear Regression", x = "Actual happiness score", y = "Predicted happiness score")
FitGraph
```



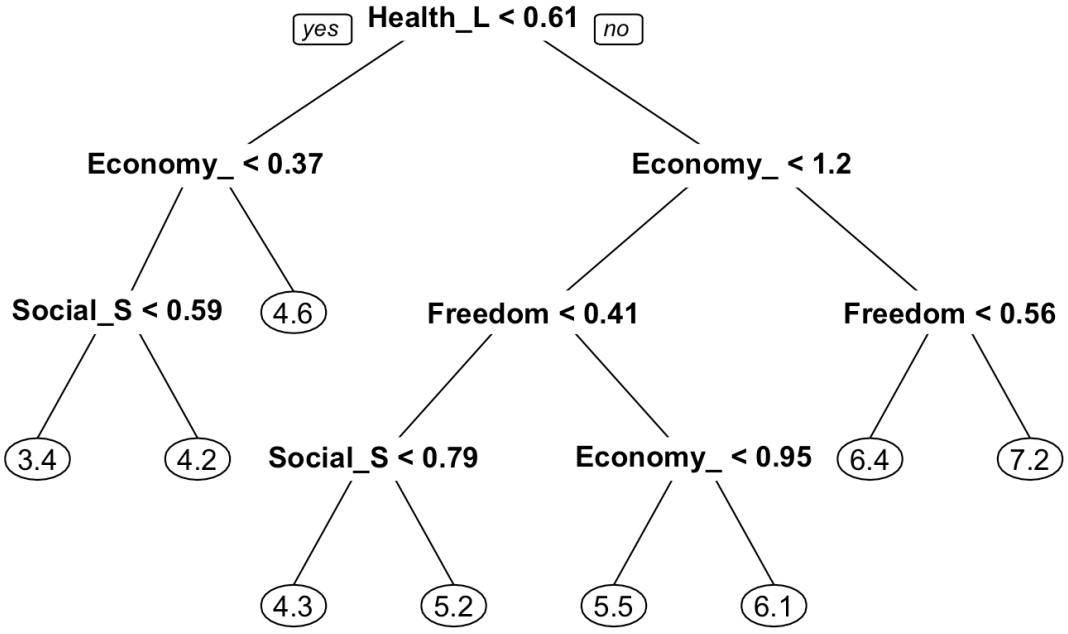
Hide

```
# Fitting Decision Tree Regression
TreeFit = rpart(formula = Happiness_Score ~ .,data = dataset,control = rpart.control(minsplit = 20))

PredTreeFit = predict(TreeFit, newdata = test_set)

PredActualTree <- as.data.frame(cbind(Prediction = PredTreeFit, Actual = test_set$Happiness_Score))


TreeGraph <- ggplot(PredActualTree, aes(Actual, Prediction )) +
  geom_point() + theme_bw() + geom_abline() +
  labs(title = "Decision Tree Regression", x = "Actual happiness score", y = "Predicted happiness score")
TreeGraph
```



Hide

```
prp(TreeFit)
```



From the decision tree regression, we noticed that the decision tree won't be a good fit. Choose multiple linear regression model instead.

Conclusion

From the overall relationship between attributes and happiness score from 2015 to 2019, the correlation between happiness score and generosity and trust government corruption is decreasing overtime, while the correlation between happiness score and social support and health expectation has increased. Overall, GDP, social support, and health expectation are the top three attributes impact on happiness score the most.

For top 10 happiest countries from 2015 to 2019, social support has always been the one that impact on happiness score the most. The positive relationship between freedom and happiness score has become stronger, same as trust government corruption and happiness score. GDP has become much less important for people's happiness.
For bottom 10 happiest countries from 2015 to 2019, health life expectancy has become the most important attribute affects people's happiness. Relatively, freedom and social support became less important, which are opposite to that of top 10 happiest countries.

In 2015, there's a stronger positive relationship between GDP and happiness score for top 10 happiest countries than that of bottom 10 happiest countries. On the opposite, the positive relationship between freedom and happiness score for top 10 happiest countries is weaker than that of bottom 10 happiest countries. When the correlation between health life expectancy and happiness score for top 10 happiest countries is positive, that of bottom 10 happiest countries are negative. Social support has always been the attribute that has the strongest positive relationship with happiness score for both top and bottom 10 happiest countries.

In 2019, there's strong negative correlation between generosity and happiness score and health life expectancy and happiness score for top 10 happiest countries, while the correlation for bottom 10 happiest countries are positive and quite strong. The positive relationship between GDP and happiness score for bottom 10 countries are stronger than that of top 10 countries.