Practical Techniques for Big Data Processing-18CSC403

Assignment 1

Name: HARSHITHA.T

Roll no: CB.SC.I5DAS21079

Installation of packages

```
#Install packages

install.packages('dplyr')

install.packages('ggplot2')


library(dplyr)

library(ggplot2)


getwd()

[1] "D:/Int Msc Data Science/Sem-7/Practical Techniques for Big Data Proce
ssing/Assignment-Lab"
>
```

Loading dataset

```
> df <- read.csv('2019.csv',TRUE,',')
> head(df)
  Overall.rank Country.or.region Score GDP.per.capita Social.support Healt
hy.life.expectancy
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
  Freedom.to.make.life.choices Generosity Perceptions.of.corruption
```

```
1                         0.596      0.153                     0.393
2                         0.592      0.252                     0.410
3                         0.603      0.271                     0.341
4                         0.591      0.354                     0.118
5                         0.557      0.322                     0.298
6                         0.572      0.263                     0.343
> tail(df)
    Overall.rank        Country.or.region Score GDP.per.capita Social.supp
ort
151          151                    Yemen 3.380          0.287          1.
163
152          152                   Rwanda 3.334          0.359          0.
711
153          153                 Tanzania 3.231          0.476          0.
885
154          154              Afghanistan 3.203          0.350          0.
517
155          155 Central African Republic 3.083          0.026          0.
000
156          156              South Sudan 2.853          0.306          0.
575
    Healthy.life.expectancy Freedom.to.make.life.choices Generosity Percep
tions.of.corruption
151                   0.463                        0.143      0.108
0.077
152                   0.614                        0.555      0.217
0.411
153                   0.499                        0.417      0.276
0.147
154                   0.361                        0.000      0.158
0.025
155                   0.105                        0.225      0.235
0.035
156                   0.295                        0.010      0.202

> colnames(df) <- gsub(" ", "_", colnames(df))
> colnames(df)
[1] "Overall.rank"                "Country.or.region"
[3] "Score"                       "GDP.per.capita"
[5] "Social.support"              "Healthy.life.expectancy"
[7] "Freedom.to.make.life.choices" "Generosity"
[9] "Perceptions.of.corruption"
        0.091
> dim(df)
[1] 156   9
> names(df)
```

```
[1] "Overall.rank"                 "Country.or.region"
[3] "Score"                        "GDP.per.capita"
[5] "Social.support"               "Healthy.life.expectancy"
[7] "Freedom.to.make.life.choices" "Generosity"
[9] "Perceptions.of.corruption"
> str(df)
'data.frame':  156 obs. of  9 variables:
 $ Overall.rank                : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Country.or.region           : chr  "Finland" "Denmark" "Norway" "Icelan
d" ...
 $ Score                       : num  7.77 7.6 7.55 7.49 7.49 ...
 $ GDP.per.capita              : num  1.34 1.38 1.49 1.38 1.4 ...
 $ Social.support              : num  1.59 1.57 1.58 1.62 1.52 ...
 $ Healthy.life.expectancy     : num  0.986 0.996 1.028 1.026 0.999 ...
 $ Freedom.to.make.life.choices: num  0.596 0.592 0.603 0.591 0.557 0.572
0.574 0.585 0.584 0.532 ...
 $ Generosity                  : num  0.153 0.252 0.271 0.354 0.322 0.263
0.267 0.33 0.285 0.244 ...
 $ Perceptions.of.corruption   : num  0.393 0.41 0.341 0.118 0.298 0.343 0
.373 0.38 0.308 0.226 ...
>
```

Data sumarization

```
> summary(df)
  Overall.rank    Country.or.region      Score       GDP.per.capita   Soci
al.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   Medi
an :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
```

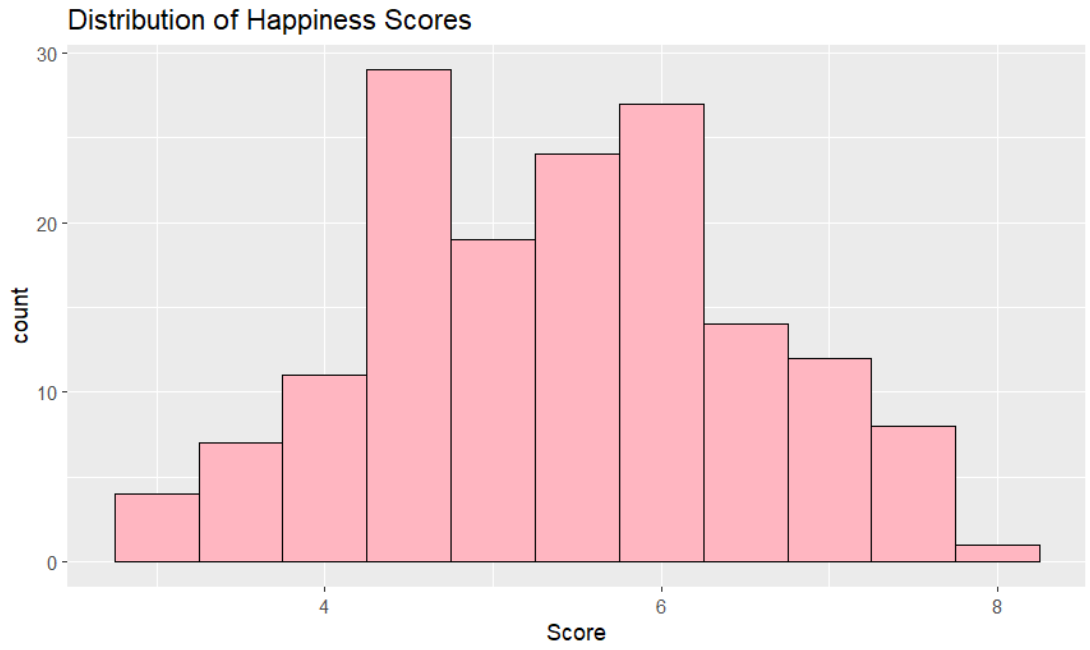
```
 1st Qu.:0.0470
 Median :0.0855
 Mean   :0.1106
 3rd Qu.:0.1412
 Max.   :0.4530
>
```

Descriptive statistics

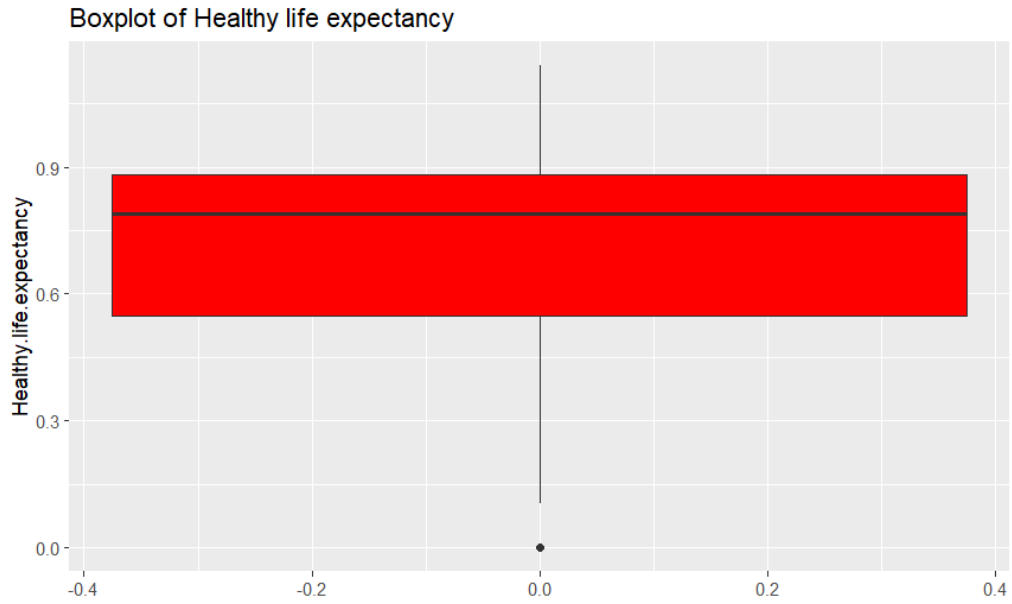
```
> descriptive_stats <- df %>%
+   summarise(
+     Mean_Score = mean(Score, na.rm = TRUE),
+     Median_Score = median(Score, na.rm = TRUE),
+     SD_Score = sd(Score, na.rm = TRUE),
+     Min_Score = min(Score, na.rm = TRUE),
+     Max_Score = max(Score, na.rm = TRUE),
+     Q1_Score = quantile(Score, 0.25, na.rm = TRUE),
+     Q3_Score = quantile(Score, 0.75, na.rm = TRUE)
+   )
> print(descriptive_stats)
  Mean_Score Median_Score SD_Score Min_Score Max_Score Q1_Score Q3_Score
1   5.407096       5.3795  1.11312     2.853     7.769   4.5445   6.1845
>
> #Clean the dataset
> anyNA(df)
[1] FALSE
> sum(is.na(df))
[1] 0
>
```

EXPLORATORY DATA ANALYSIS

```
> # 1.What is the distribution of happiness scores across all countries?
> ggplot(df, aes(x = Score)) +
+   geom_histogram(binwidth = 0.5, fill = "lightpink", color = "black") +
+   labs(title = "Distribution of Happiness Scores")
>
```




```
> # Boxplot
> ggplot(df, aes(y = Healthy.life.expectancy)) +
+   geom_boxplot(fill = "red") +
+   labs(title = "Boxplot of Healthy life expectancy")
```




```
>
```

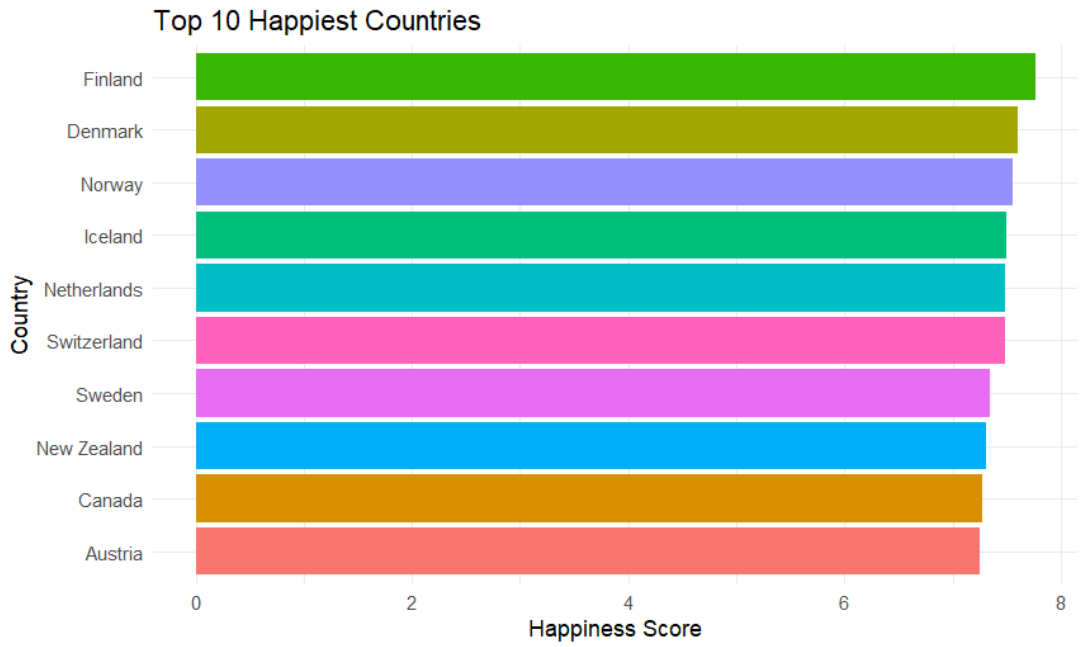
ANALYSIS QUESTIONS

1. Which countries have the highest and lowest happiness scores?
2. How generosity varies among different countries?
3. What is the relationship between social support and happiness scores?
4. Does higher life expectancy correlate with higher happiness scores?
5. How does happiness scores affect the freedom to make life choices among the least happiest countries?
6. Which countries have the lowest and highest GDP?
7. What is the average happiness score across all countries?
8. What is the range of GDP per capita?
9. Which region has the highest average social support?
10. What is the distribution of corruption perceptions among the top 10 happiest countries in the world?

```
>
> # 1. Which countries have the highest and lowest happiness scores?
> top_10_happy <- df %>% arrange(desc(Score)) %>% head(10)
> bottom_10_happy <- df %>% arrange(Score) %>% head(10)
> top_10_happy
   Overall.rank Country.or.region Score GDP.per.capita Social.support Heal
thy.life.expectancy
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
```

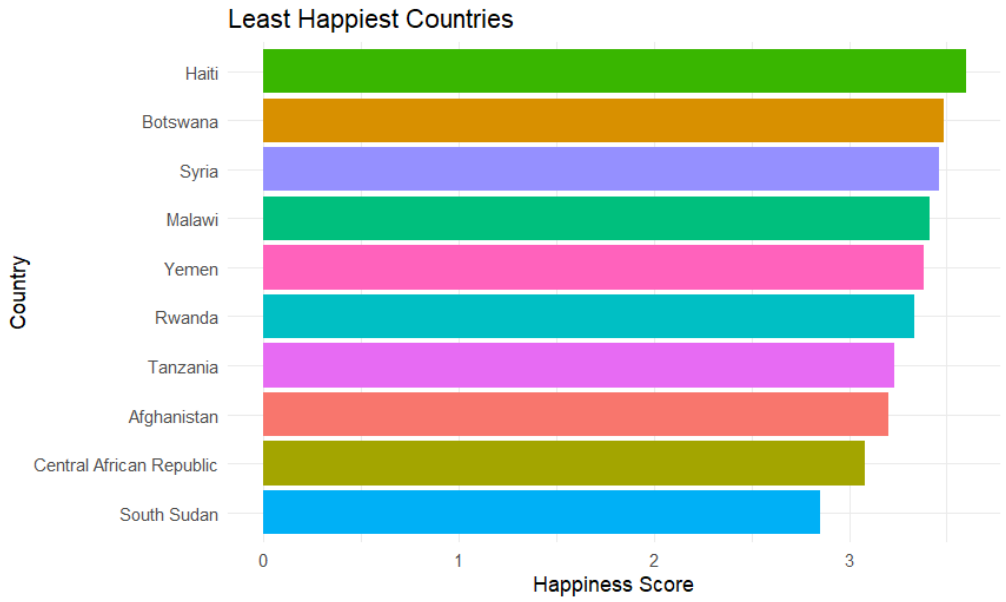
```
7              7             Sweden 7.343          1.387          1.487
1.009
8              8        New Zealand 7.307          1.303          1.557
1.026
9              9             Canada 7.278          1.365          1.505
1.039
10            10            Austria 7.246          1.376          1.475
1.016
   Freedom.to.make.life.choices Generosity Perceptions.of.corruption
1                         0.596      0.153                     0.393
2                         0.592      0.252                     0.410
3                         0.603      0.271                     0.341
4                         0.591      0.354                     0.118
5                         0.557      0.322                     0.298
6                         0.572      0.263                     0.343
7                         0.574      0.267                     0.373
8                         0.585      0.330                     0.380
9                         0.584      0.285                     0.308
10                        0.532      0.244                     0.226
```

```
ggplot(top_10_happiest, aes(x = reorder(`Country.or.region`, Score), y = S
core, fill = `Country.or.region`)) +
  geom_bar(stat = 'identity') +
  coord_flip() +
  labs(title = 'Top 10 Happiest Countries', x = 'Country', y = 'Happiness
Score') +
  theme_minimal() +
  theme(legend.position = 'none')
```

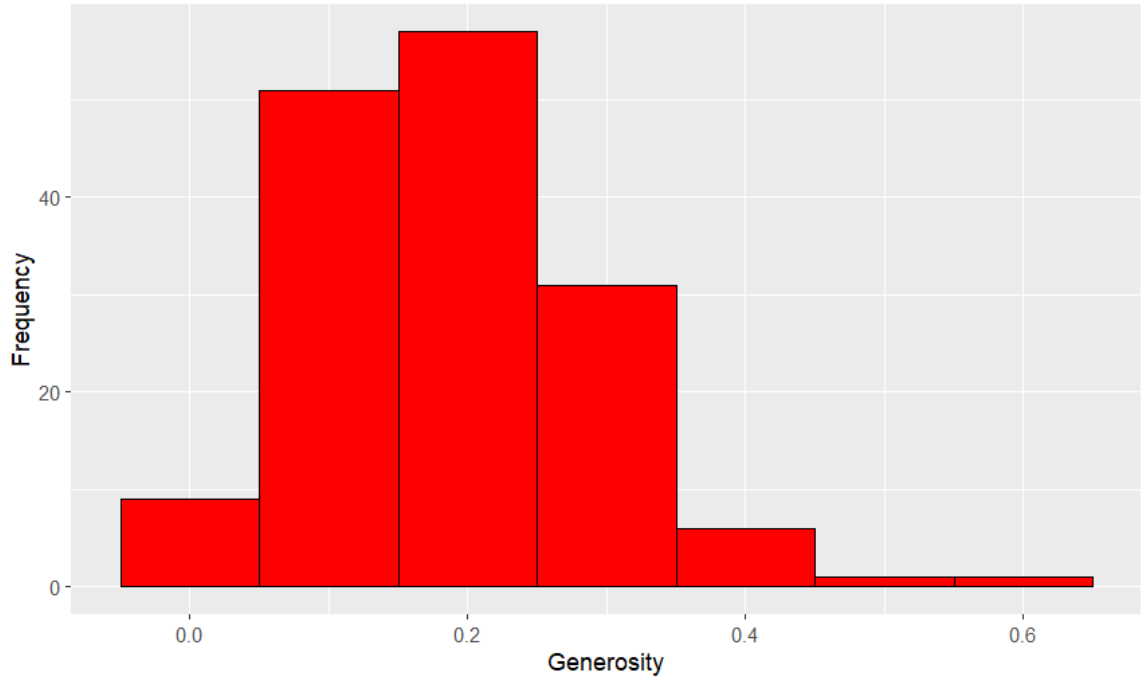


```
> bottom_10_happy
   Overall.rank        Country.or.region Score GDP.per.capita Social.suppo
rt
1           156              South Sudan 2.853          0.306          0.5
75
2           155 Central African Republic 3.083          0.026          0.0
00
3           154              Afghanistan 3.203          0.350          0.5
17
4           153                 Tanzania 3.231          0.476          0.8
85
5           152                   Rwanda 3.334          0.359          0.7
11
6           151                    Yemen 3.380          0.287          1.1
63
7           150                   Malawi 3.410          0.191          0.5
60
8           149                    Syria 3.462          0.619          0.3
78
9           148                 Botswana 3.488          1.041          1.1
45
10          147                    Haiti 3.597          0.323          0.6
88
   Healthy.life.expectancy Freedom.to.make.life.choices Generosity Percept
ions.of.corruption
1                    0.295                         0.010      0.202
0.091
2                    0.105                         0.225      0.235
0.035
3                    0.361                         0.000      0.158
0.025
4                    0.499                         0.417      0.276
0.147
5                    0.614                         0.555      0.217
0.411
6                    0.463                         0.143      0.108
0.077
7                    0.495                         0.443      0.218
0.089
8                    0.440                         0.013      0.331
0.141
9                    0.538                         0.455      0.025
0.100
10                   0.449                         0.026      0.419
0.110
>
```

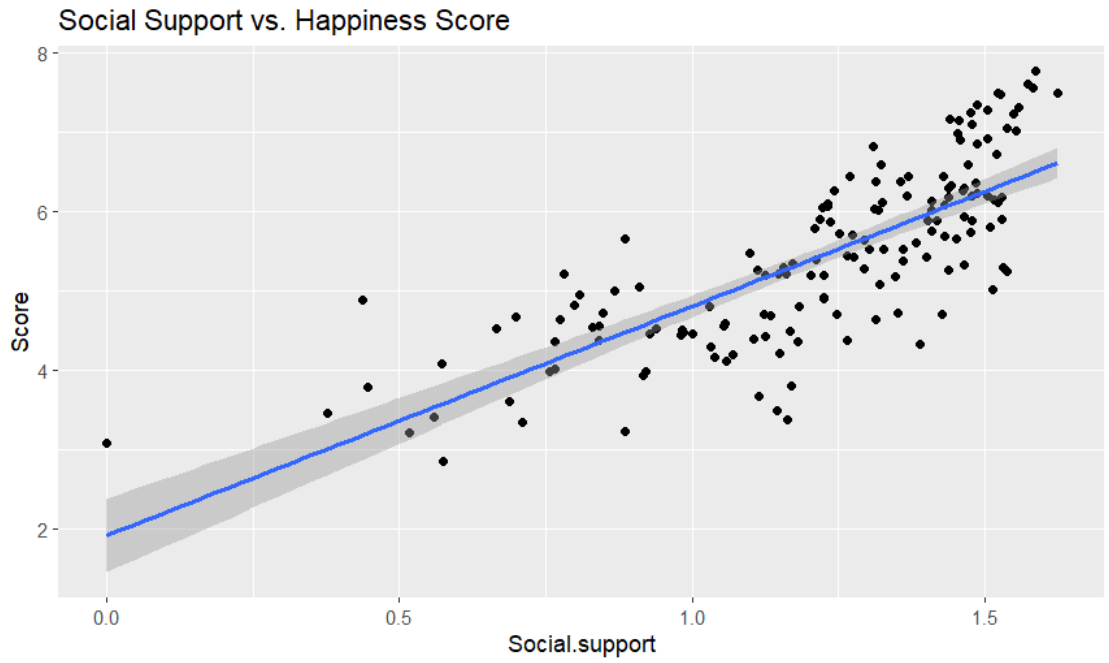
```
 > ggplot(bottom_10_happy, aes(x = reorder(`Country.or.region`, Score), y
= Score, fill = `Country.or.region`)) +
+   geom_bar(stat = 'identity') +
+   coord_flip() +
+   labs(title = 'Least Happiest Countries', x = 'Country', y = 'Happiness
Score') +
+   theme_minimal() +
+   theme(legend.position = 'none')
```



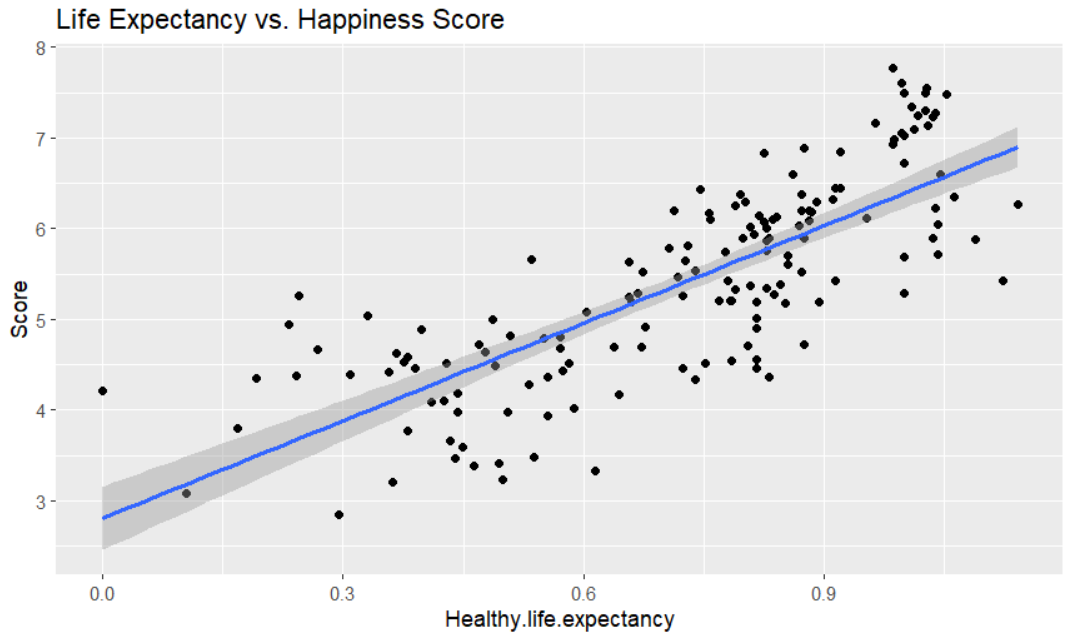

```
>
>
> # 2.How generosity varies among different countries?
> ggplot(world_happiness, aes(x = Generosity)) +
  geom_histogram(binwidth = 0.1, color = "black", fill = "skyblue") +
  labs(x = "Generosity", y = "Frequency")
```



```
>
> # 3.What is the relationship between social support and happiness scores
?
> ggplot(df, aes(x = `Social.support`, y = Score)) +
+   geom_point() +
+   geom_smooth(method = "lm") +
+   labs(title = "Social Support vs. Happiness Score")
`geom_smooth()` using formula = 'y ~ x'
```

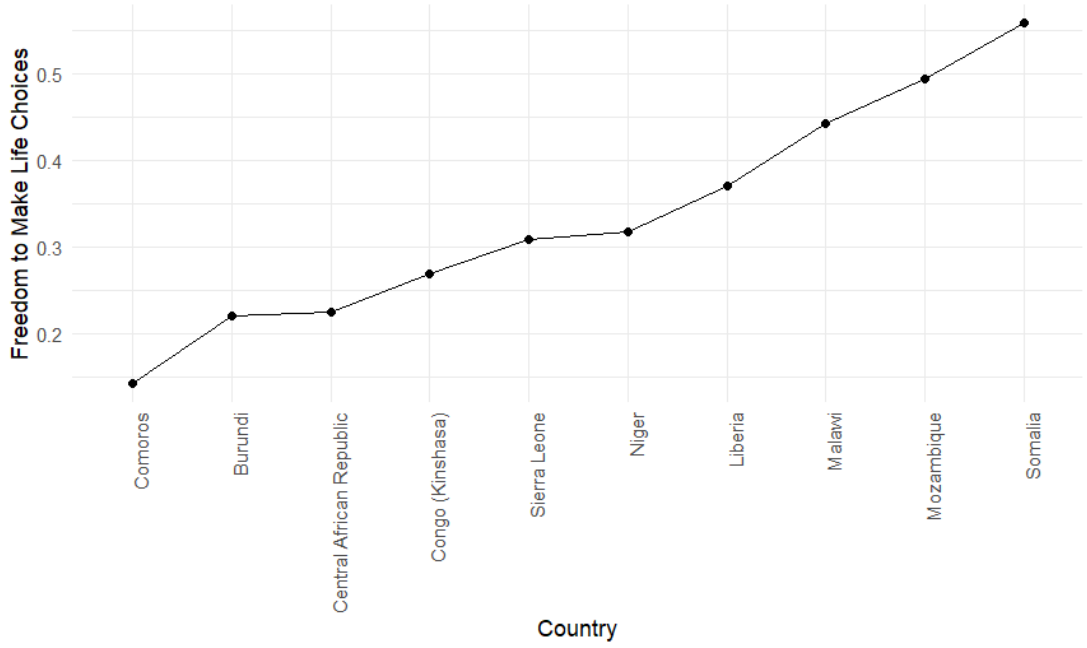



```
>
> # 4.Does higher life expectancy correlate with higher happiness scores?
> ggplot(df, aes(x = `Healthy.life.expectancy`, y = Score)) +
+   geom_point() +
+   geom_smooth(method = "lm") +
+   labs(title = "Life Expectancy vs. Happiness Score")
`geom_smooth()` using formula = 'y ~ x'
```




```
>
> # 5.How does happiness scores affect the freedom to make life choices am
ong the least happiest countries?

> bottom_10_gdp <- df %>%
  arrange(GDP.per.capita) %>%
  slice(1:10)
ggplot(bottom_10_gdp, aes(x = reorder(Country.or.region, Freedom.to.make.l
ife.choices), y = Freedom.to.make.life.choices, group = 1)) +
  geom_line() +
  geom_point() +
  labs(x = "Country", y = "Freedom to Make Life Choices") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

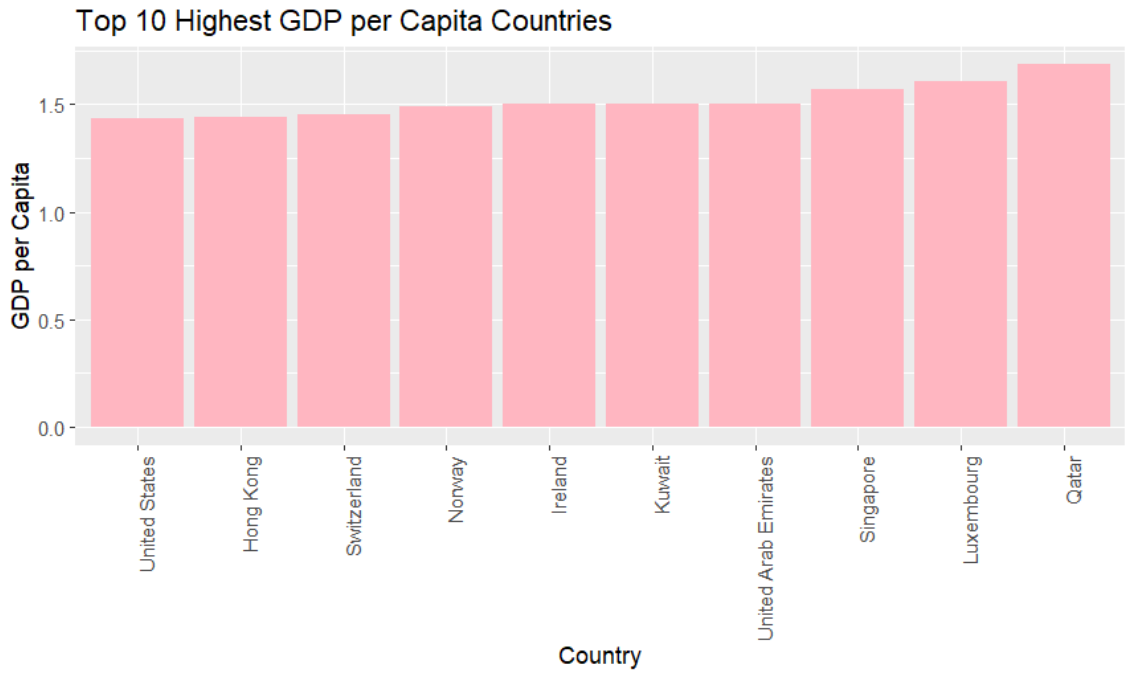



```
> # 6.Which countries have lowest and highest GDP?
> top_10_gdp <- df %>% arrange(desc(GDP.per.capita)) %>% head(10)
> bottom_10_gdp <- df %>% arrange(GDP.per.capita) %>% head(10)
> top_10_gdp

   Overall.rank    Country.or.region Score GDP.per.capita Social.support
1            29                Qatar 6.374          1.684          1.313
2            14           Luxembourg 7.090          1.609          1.479
3            34            Singapore 6.262          1.572          1.463
4            21 United Arab Emirates 6.825          1.503          1.310
5            51               Kuwait 6.021          1.500          1.319
6            16              Ireland 7.021          1.499          1.553
7             3               Norway 7.554          1.488          1.582
8             6          Switzerland 7.480          1.452          1.526
9            76            Hong Kong 5.430          1.438          1.277
10           19        United States 6.892          1.433          1.457
   Healthy.life.expectancy Freedom.to.make.life.choices Generosity Percept
ions.of.corruption
1                    0.871                        0.555      0.220
0.167
2                    1.012                        0.526      0.194
0.316
3                    1.141                        0.556      0.271
0.453
4                    0.825                        0.598      0.262
0.182
5                    0.808                        0.493      0.142
0.097
```

```
6                    0.999                          0.516       0.298
0.310
7                    1.028                          0.603       0.271
0.341
8                    1.052                          0.572       0.263
0.343
9                    1.122                          0.440       0.258
0.287
10                   0.874                          0.454       0.280
                0.128
> top_10_gdp_countries <- df %>%
  arrange(desc(GDP.per.capita)) %>%
  head(10)

ggplot(top_10_gdp_countries, aes(x = reorder(Country.or.region, GDP.per.ca
pita), y = GDP.per.capita)) +
  geom_bar(stat = "identity", fill = "lightpink") +
  labs(title = "Top 10 Highest GDP per Capita Countries", x = "Country", y
= "GDP per Capita") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

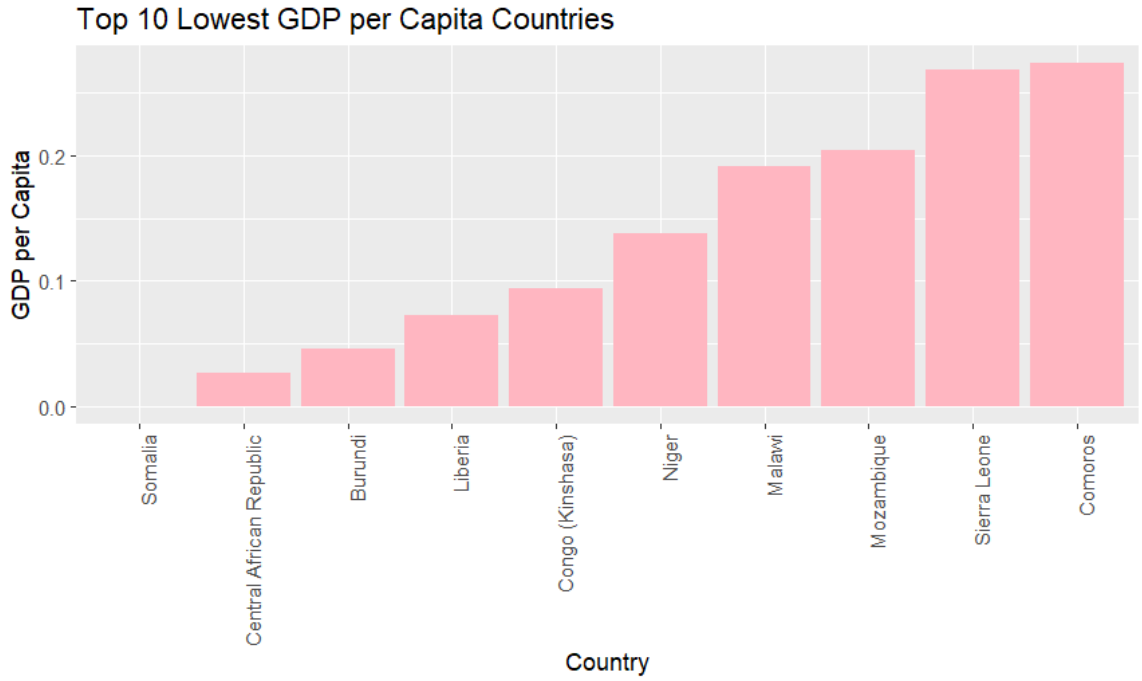



```
bottom_10_gdp
   Overall.rank        Country.or.region Score GDP.per.capita Social.suppo
rt
1           112                  Somalia 4.668          0.000          0.6
98
2           155 Central African Republic 3.083          0.026          0.0
00
```

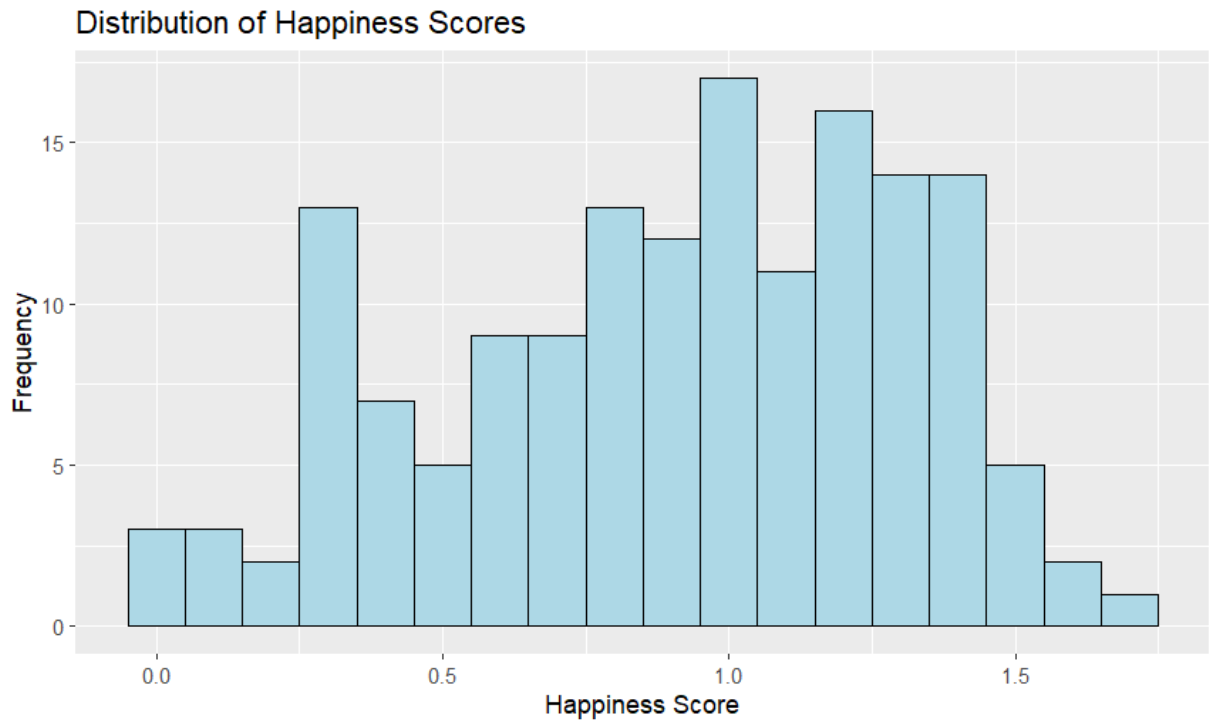
```
3           145                  Burundi 3.775          0.046          0.4
47
4           141                  Liberia 3.975          0.073          0.9
22
5           127         Congo (Kinshasa) 4.418          0.094          1.1
25
6           114                    Niger 4.628          0.138          0.7
74
7           150                   Malawi 3.410          0.191          0.5
60
8           123                Mozambique 4.466          0.204          0.9
86
9           129             Sierra Leone 4.374          0.268          0.8
41
10          142                  Comoros 3.973          0.274          0.7
57
   Healthy.life.expectancy Freedom.to.make.life.choices Generosity Percept
ions.of.corruption
1                    0.268                        0.559      0.243
0.270
2                    0.105                        0.225      0.235
0.035
3                    0.380                        0.220      0.176
0.180
4                    0.443                        0.370      0.233
0.033
5                    0.357                        0.269      0.212
0.053
6                    0.366                        0.318      0.188
0.102
7                    0.495                        0.443      0.218
0.089
8                    0.390                        0.494      0.197
0.138
9                    0.242                        0.309      0.252
0.045
10                   0.505                        0.142      0.275
0.078
> bottom_10_gdp_countries <- df %>%
  arrange(GDP.per.capita) %>%
  head(10)

ggplot(bottom_10_gdp_countries, aes(x = reorder(Country.or.region, GDP.per
.capita), y = GDP.per.capita)) +
  geom_bar(stat = "identity", fill = "lightpink") +
```

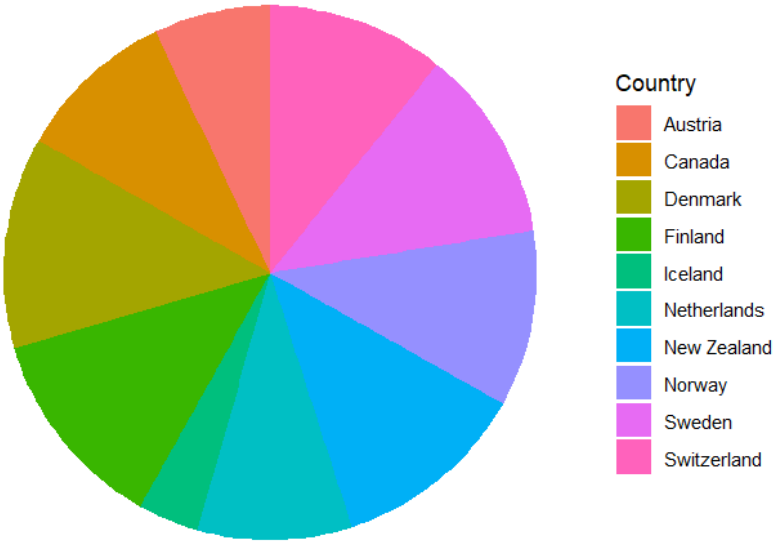
```
  labs(title = "Top 10 Lowest GDP per Capita Countries", x = "Country", y
= "GDP per Capita") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
> # 7.What is the average happiness score across all countries?
> average_happiness <- df %>% summarize(Average_Score = mean(Score))
> average_happiness
  Average_Score
1      5.407096
>
> # 8.What is the range of GDP per capita?
> gdp_range <- df %>% summarize(Range = max(GDP.per.capita) - min(GDP.per.
capita))
> gdp_range
  Range
1 1.684
> ggplot(df, aes(x =GDP.per.capita)) +
  geom_histogram(binwidth = 0.1, color = "black", fill = "lightblue") +
  labs(title = "Distribution of Happiness Scores", x = "Happiness Score",
y = "Frequency")
```




```
> # 9.Which region has the highest average social support?
> region_highest_social_support <- df %>%
+   group_by(Country.or.region) %>%
+   summarize(Average_Social_Support = mean(Social.support)) %>%
+   arrange(desc(Average_Social_Support)) %>%
+   head(1)
> region_highest_social_support
# A tibble: 1 × 2
  Country.or.region Average_Social_Support
  <chr>                              <dbl>
1 Iceland                             1.62
>
> # 10. What is the distribution of corruption perceptions among the top 10
happiest countries in the world?
>
> top_10_countries <- df %>%
+   arrange(desc(Score)) %>%
+   slice(1:10)
>
>
> ggplot(top_10_countries, aes(x = "", y = Perceptions.of.corruption, fill
= Country.or.region)) +
+   geom_bar(width = 1, stat = "identity") +
+   coord_polar("y", start = 0) +
+   theme_void() +
+   labs(fill = "Country", x = "", y = "Perceptions of Corruption")
```


Country
Austria
Canada
Denmark
Finland
Iceland
Netherlands
New Zealand
Norway
Sweden
Switzerland