World Happiness Ranking Report

MinhChau

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve
3WBa
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ lubridate 1.9.4     ✔ tibble    3.2.1
## ✔ purrr     1.0.4     ✔ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks rstatix::filter(), stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

Context

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012. The report is based on statistical analyses of Gallup World Poll data which specifically monitors performance in six particular categories: gross domestic product per capita, social support, healthy life expectancy, freedom to make your own life choices, generosity of the general population, and perceptions of internal and external corruption levels.

The data `HP.csv` for this question is from the 2019 report and has 156 observations on 9 variables.

- `Country` : Country
- `Region` : Region
- `Score` : happiness score of the country (ranging from 0 to 10 with 10 being the happiest)
- `Score2019` : happiness score of the country in 2019 (ranging from 0 to 10 with 10 being the happiest)
- `Score2018` : happiness score of the country in 2018 (ranging from 0 to 10 with 10 being the happiest)
- `GDP` : gross domestic product of the country
- `Family` : indicator that shows family support to each citizen in the country
- `Life.Expectancy` : shows the healthiness level of the country
- `Freedom` : indicator that shows the citizen freedom to choose their life path, job or etc

- `Trust` : shows the level of trust from the citizen in the government (influenced by the corruption level and performance of the government)
- `Generosity` : indicator that shows the generosity level of the citizen of the country

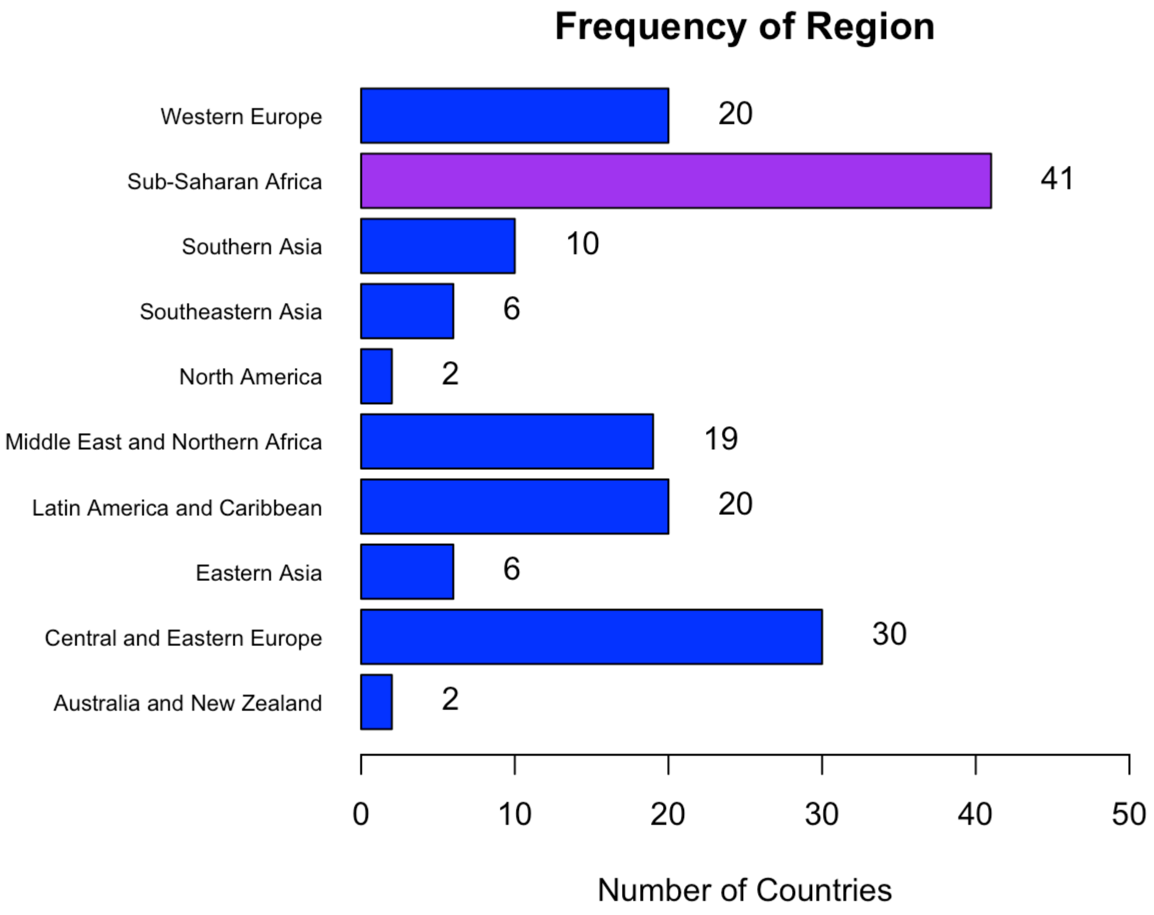
The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors.

1. Frequency Distribution for Region

Find the region that have the most and least number of countries in the data.

Frequency of Region

Region	n
Australia and New Zealand	2
Central and Eastern Europe	30
Eastern Asia	6
Latin America and Caribbean	20
Middle East and Northern Africa	19
North America	2
Southeastern Asia	6
Southern Asia	10
Sub-Saharan Africa	41
Western Europe	20



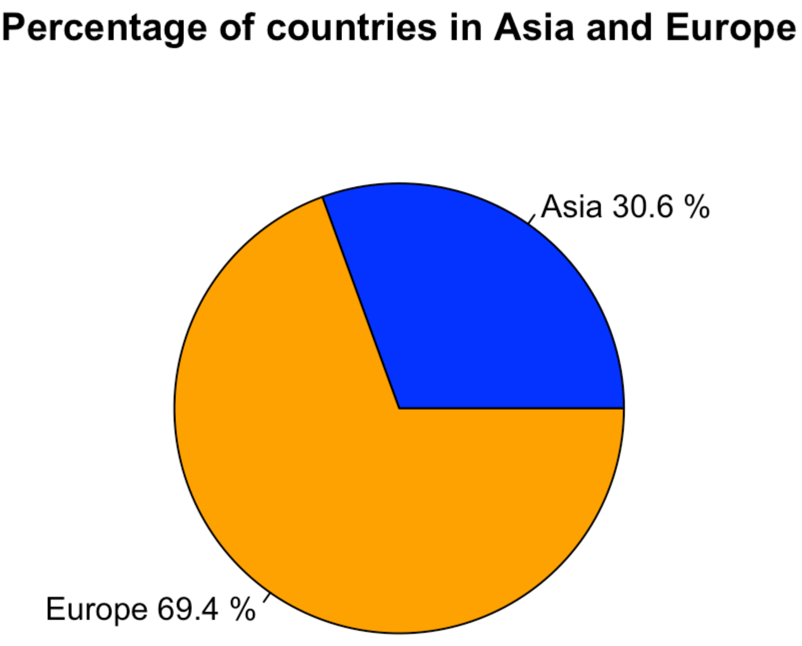
2. Frequency Distribution for Europe and Asia

- i.Keep only countries that are in Europe and Asia and store them in a dataframe called `dfEA` .
- ii. Create a new variable in dfEA, `Region2` , which contains only the values “Asia” and “Europe”.
- iii. Plot the table and pie chart displaying the frequency (for table) and percentage (for pie chart) of countries in Europe and Asia. Give a title to the table and pie chart.

```
dfEA <- HP[HP$Region %in% c("Western Europe", "Central and Eastern Europe", "Southern
Asia", "Southeastern Asia","Eastern Asia"), ]
dfEA$Region2 <- ifelse(dfEA$Region %in% c("Western Europe", "Central and Eastern Euro
pe"), "Europe", "Asia")
freq_Reg2 <- table(dfEA$Region2)
print(freq_Reg2)
```

```
##
##   Asia Europe
##     22     50
```

```
pie1 <- prop.table(freq_Reg2)*100
pie(pie1,
    labels = paste(names(pie1), round(pie1, 1), "%"),
    main = "Percentage of countries in Asia and Europe",
    col = c("blue", "orange")
    )
```



3. Singapore Happiness Data (2 marks)

i. extract the row of data for Singapore and display it in a table.
ii. find the rank of Singapore in terms of its happiness score with Rank 1 being the happiest country

```
singapore <- HP[HP$Country == "Singapore", ]
HP$rank <- rank(-HP$Score, ties.method = "min")
sgrank <- HP$rank[HP$Country == "Singapore"]
cat("Singapore happiness ranking", sgrank, "\n")
```

```
## Singapore happiness ranking 34
```

4. Happiness Ranking

In the `HP` dataframe, create a column called `Rank` which will contain the happiness ranking for each country based on their Happiness Score where Rank will be 1 for the country with the highest score. There are no ties so the values will be from 1 to 156.

Identify which are the 5 happiest countries and 5 least happiest countries. Present your answers in two separate tables: one for the 5 happiest countries and another table for 5 least happiest countries. Each table should have three columns, namely `Country` , `Rank` and `Score` . Sort the countries in decreasing score for the 5 happiest countries and in increasing score for the 5 least happy countries. (2 marks)

```
HP$Rank <- rank(-HP$Score, ties.method = "first")
happiest <- HP[order(-HP$Score), c("Country", "Rank", "Score")][1:5, ]
least_happy <- HP[order(HP$Score), c("Country", "Rank", "Score")][1:5, ]

print(happiest)
```

```
##         Country Rank Score
## 44      Finland    1 7.769
## 37      Denmark    2 7.600
## 106      Norway    3 7.554
## 58      Iceland    4 7.494
## 99  Netherlands    5 7.488
```

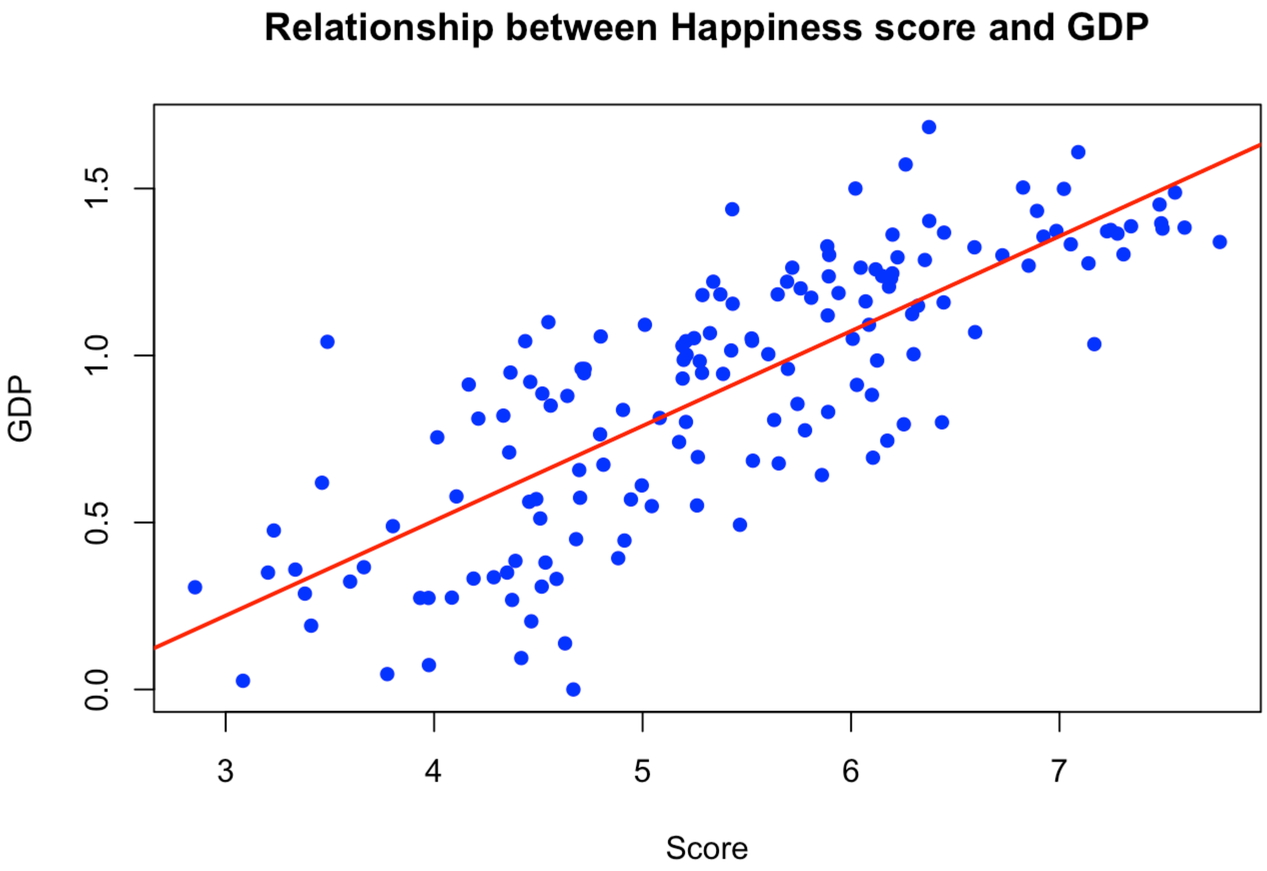
```
print(least_happy)
```

```
##                      Country Rank Score
## 129              South Sudan  156 2.853
## 25  Central African Republic  155 3.083
## 1                Afghanistan  154 3.203
## 138                 Tanzania  153 3.231
## 118                   Rwanda  152 3.334
```

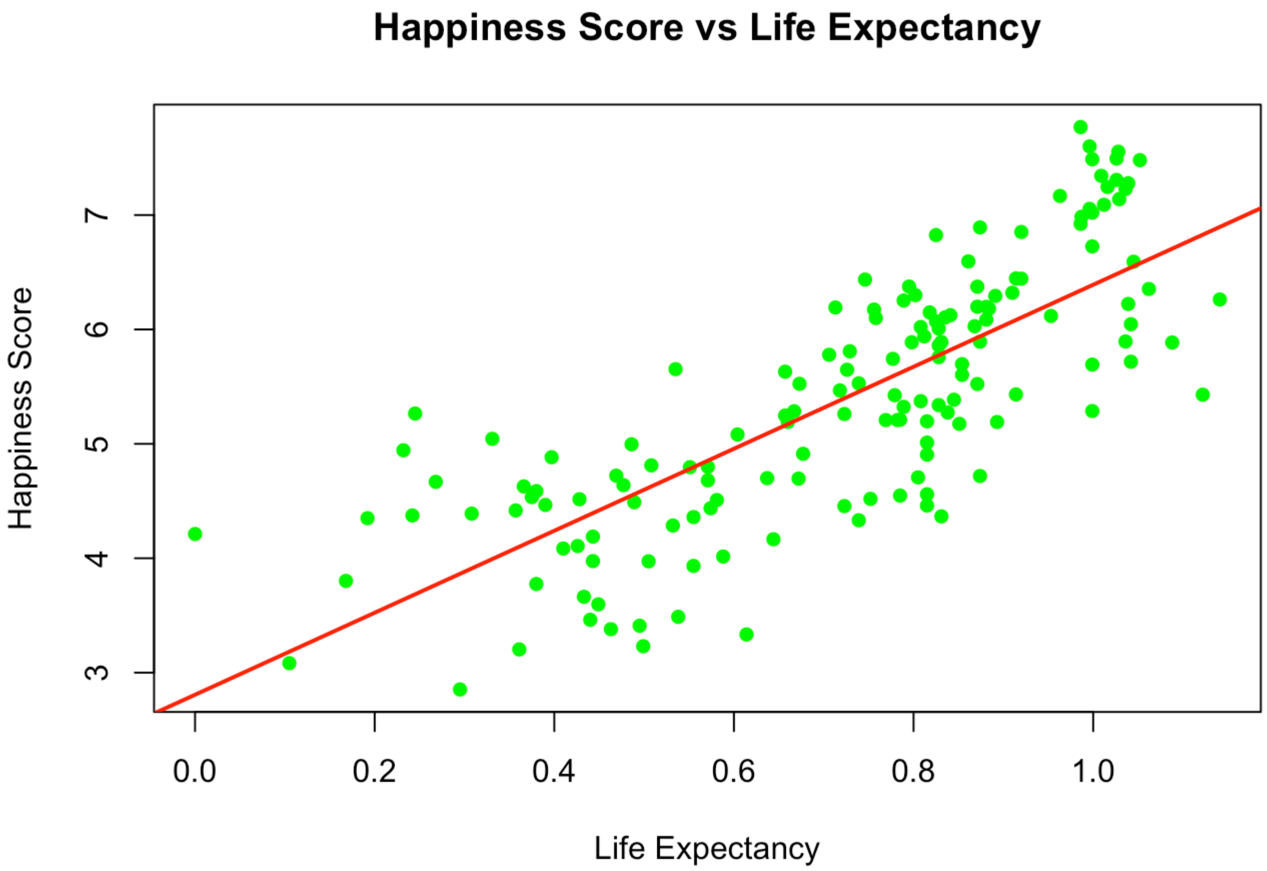
5. Relationship between Life.Expectancy and Score

Plot a chart to explore the linear relationship between Score and GDP, as well as between Score and Life.Expectancy. Describe your findings. (3 marks)

```
plot(HP$Score,
     HP$GDP,
     main = "Relationship between Happiness score and GDP",
     xlab = "Score",
     ylab = "GDP",
     col = "blue",
     pch = 16)
abline(lm(GDP ~ Score, data = HP), col = "red", lwd = 2)
```



```
plot(HP$Life.Expectancy, HP$Score,
     main = "Happiness Score vs Life Expectancy",
     xlab = "Life Expectancy",
     ylab = "Happiness Score",
     col = "green", pch = 16)
abline(lm(Score ~ Life.Expectancy, data = HP), col = "red", lwd = 2)
```

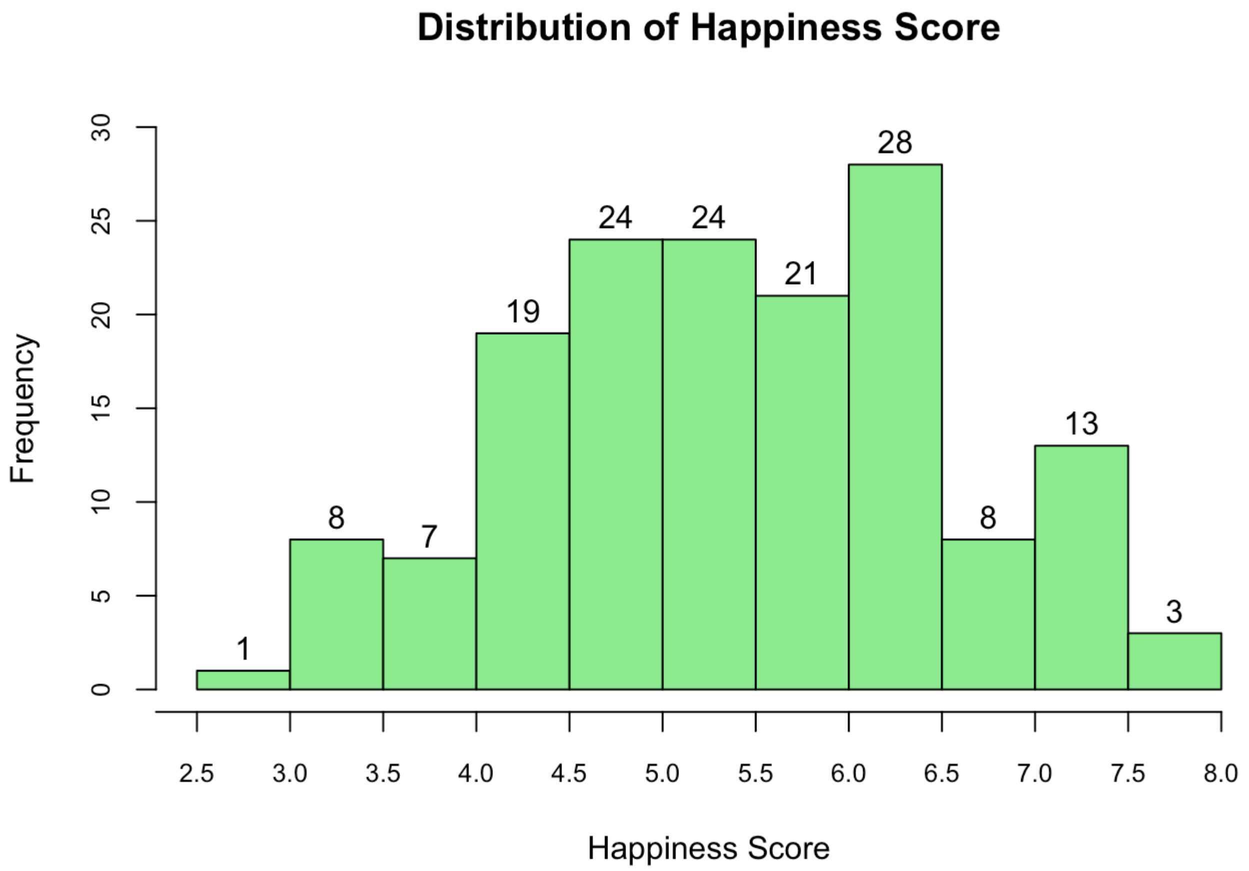


Happiness Score vs GDP - positive correlation: higher GDP per capita generally leads to higher Happiness Score. This suggests economic prosperity contributes to happiness, but the relationship may not be perfectly linear. Happiness Score vs Life Expectancy - positive correlation: higher Life Expectancy is associated with higher Happiness Score. Countries with better healthcare and living conditions tend to have higher happiness levels.

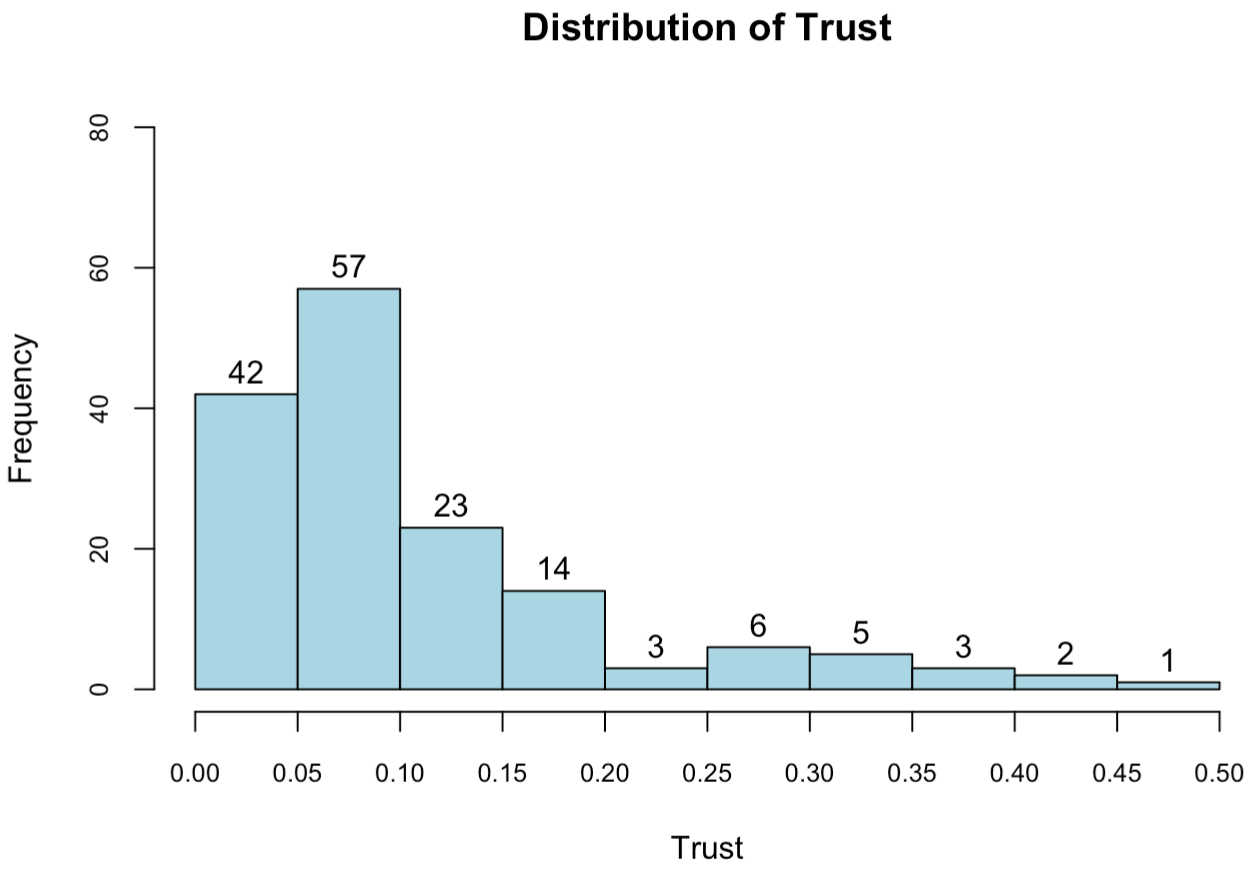
6. Data Checking

- i. Check the distribution for `Score` and `Trust` by plotting a histogram, including value labels for each bar. Analize if the variables approximately normal.
- ii. Evaluate and explain if there are any outliers for these two variables Checking the data and getting it ready for further analyses.

```
#I
H2 <- hist(HP$Score,
          ylim = c(0,30),
          xaxp = c(2,8,12),
          labels = TRUE,
          main = "Distribution of Happiness Score",
          xlab = "Happiness Score",
          ylab = "Frequency",
          col = "lightgreen",
          cex.axis=0.8)
```



```
H1 <- hist(HP$Trust,
          ylim = c(0,80),
          labels = TRUE,
          xaxp=c(0,0.5,10),
          main = "Distribution of Trust",
          xlab = "Trust",
          ylab = "Frequency",
          col = "lightblue",
          cex.axis=0.8)
```



```
shapiro.test(HP$Score)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  HP$Score
## W = 0.9872, p-value = 0.1633
```

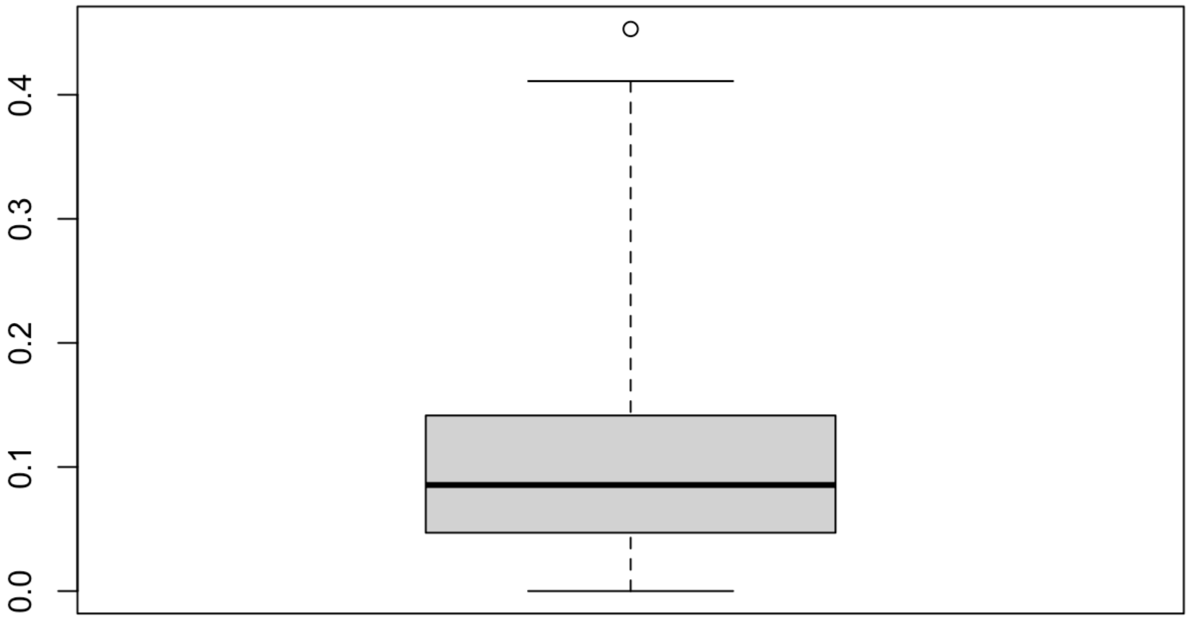
```
#W = 0.9872, p-value = 0.1633 > 0.05 => normally distributed

shapiro.test(HP$Trust)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  HP$Trust
## W = 0.8228, p-value = 1.813e-12
```

```
#W = 0.8228, p-value = 1.813e-12 < 0.05 => not normally distributed

#II
bt3 <- boxplot(HP$Trust,
        range = 3)
```



```
bt3$out
```

```
## [1] 0.453
```

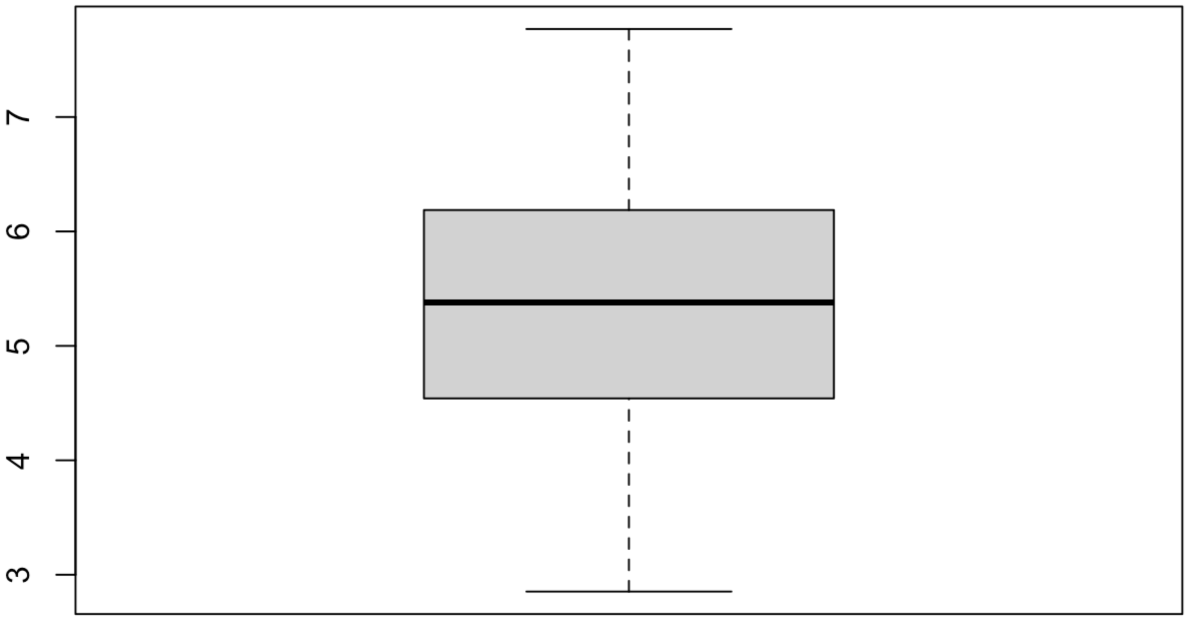
```
length(bt3$out)
```

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

```
min(bt3$out)
```

```
## [1] 0.453
```

```
#no outliners
bs3 <- boxplot(HP$Score,
        range = 3)
```



```
#there might be outliners
```

From histogram, distribution look approximately normal, relatively symmetrical and follows a bell shaped curve. From histogram, Trust is not normally distributed. It has a right skewed distribution. => Based on both histograms, we do not see any points that are far from the rest.Hence it does not look like there are any outliers

For the boxplot One extreme outlier with value of 0.453 based on rules of thumb that is further than 3IQR from Q3 but this could also be due to the fact that the data is right skewed. No extreme outliers for score

7. Descriptive statistics for Happiness Score and Trust Dashboard

- i. Generate the descriptive statistics for `Score` and `Trust` in a table, including only these statistics: mean, sd, min, max, skew, kurtosis.
- ii. Interpret the skew and kurtosis results. Is this aligned with your observation above from the histograms?

CODE

```
dfScore <- HP %>%
  summarise(
    vars = "Score",
    mean = mean(Score, na.rm = TRUE),
    sd = sd(Score, na.rm = TRUE),
    min = NA,
    max = NA,
    skew = e1071::skewness(Score, na.rm = TRUE),
    kurtosis = e1071::kurtosis(Score, na.rm = TRUE)
  )

dfTrust <- HP %>%
  summarise(
    vars = "Trust",
    mean = mean(Trust, na.rm = TRUE),
    sd = sd(Trust, na.rm = TRUE),
    min = min(Trust, na.rm = TRUE),
    max = max(Trust, na.rm = TRUE),
    skew = e1071::skewness(Trust, na.rm = TRUE),
    kurtosis = e1071::kurtosis(Trust, na.rm = TRUE)
  )

dfST <- rbind(dfScore, dfTrust)
kable(dfST, row.names = FALSE, digits = 2,
      caption = "Descriptive Statistics for Score and Trust")
```

Descriptive Statistics for Score and Trust

vars	mean	sd	min	max	skew	kurtosis
Score	5.41	1.11	NA	NA	0.01	-0.66
Trust	0.11	0.09	0	0.45	1.62	2.23

Interpret Mean, sd, skew For Score: • Skewness = 0.01, which is very close to 0. • This indicates that the score variable is approximately symmetrical, confirming our observation from the histogram that it follows a bell-shaped curve. For Trust: • Skewness = 1.62, which is significantly positive. • This suggests that the trust variable has a right-skewed distribution, meaning there are more lower values with a long tail on the right. • This aligns with the histogram observation that trust is not normally distributed and is right-skewed.

The kurtosi() function from psych calculates excess kurtosis. For Score: • Excess Kurtosis = -0.66 • This indicates the tails are lighter than a normal distribution, and there are fewer extreme values. For Trust: • Excess Kurtosis = 2.23 • This suggests the distribution has heavier tails than a normal distribution.

8. Score by Region Analyses Dashboard

Let’s develop a dashboard that allows us to drill down further into the data by Region.

- i. Generate a table that displays the number of countries, mean, standard deviation, min and max for `Score` for each region.
- ii. Add a column to the table that shows the “Coefficient of Variation” for each region and display the data in descending order of COV. Describe findings with respect to COV
- iii. Plot a chart that displays and allows easy comparison of mean Score for each Region. Describe your findings

```
#I
tabS <- HP %>% group_by(`Region`) %>%
  summarise(n(),
            Mean_Score = round(mean(Score), 1),
            SD_Score = sd(Score),
            min = min(Score),
            max = max(Score))
kable(tabS)
```

Region	n()	Mean_Score	SD_Score	min	max
Australia and New Zealand	2	7.3	0.0558614	7.228	7.307
Central and Eastern Europe	30	5.6	0.5738283	4.332	6.852
Eastern Asia	6	5.7	0.4759401	5.191	6.446
Latin America and Caribbean	20	5.9	0.7409150	3.597	7.167
Middle East and Northern Africa	19	5.2	1.0603215	3.380	7.139
North America	2	7.1	0.2729432	6.892	7.278
Southeastern Asia	6	5.2	0.6023525	4.360	6.008
Southern Asia	10	4.8	0.8553081	3.203	6.262
Sub-Saharan Africa	41	4.3	0.6843572	2.853	6.192
Western Europe	20	6.9	0.6796091	5.287	7.769

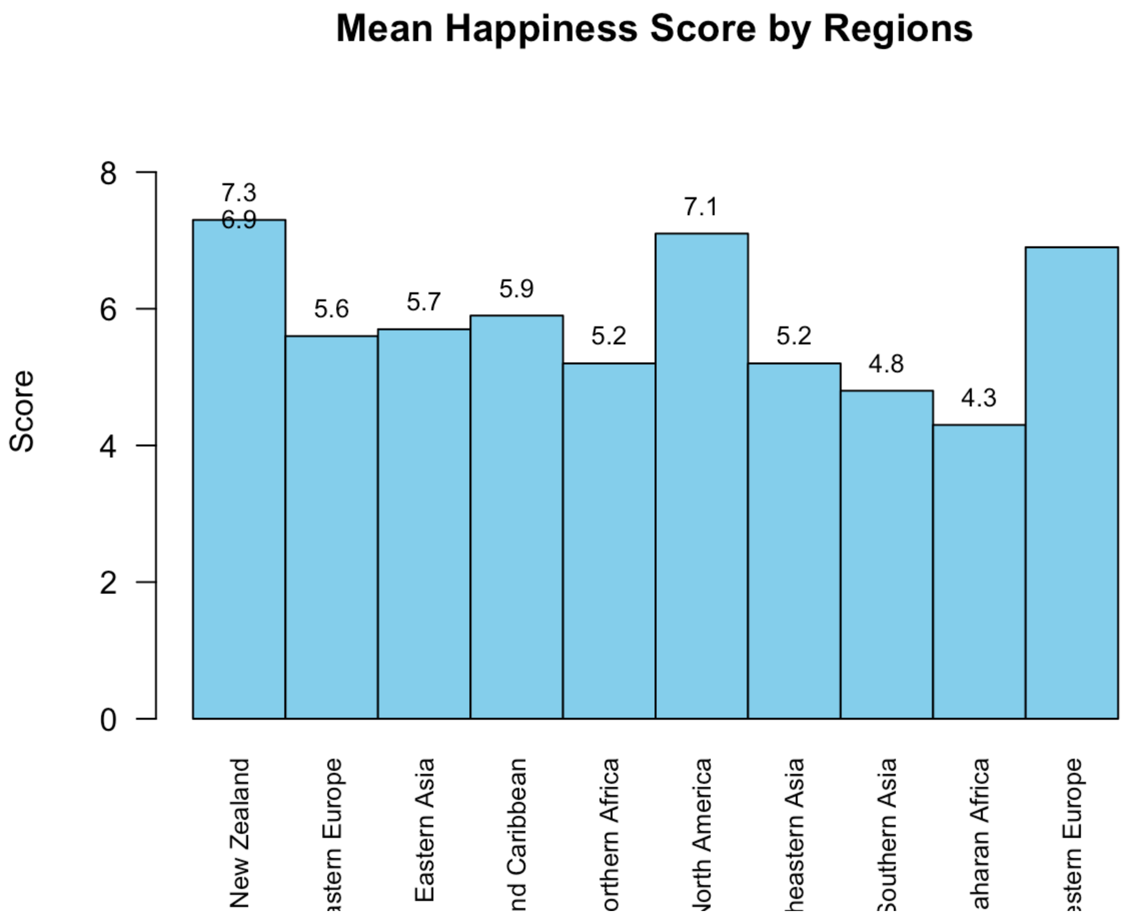
```
#II
tabS %>% mutate (COV = (SD_Score/Mean_Score)*100) %>%
arrange(desc(COV))
```

```
## # A tibble: 10 × 7
##    Region                          `n()` Mean_Score SD_Score   min   max    COV
##    <chr>                           <int>      <dbl>    <dbl> <dbl> <dbl>  <dbl>
##  1 Middle East and Northern Africa    19        5.2   1.06    3.38  7.14 20.4
##  2 Southern Asia                      10        4.8   0.855   3.20  6.26 17.8
##  3 Sub-Saharan Africa                 41        4.3   0.684   2.85  6.19 15.9
##  4 Latin America and Caribbean        20        5.9   0.741   3.60  7.17 12.6
##  5 Southeastern Asia                   6        5.2   0.602   4.36  6.01 11.6
##  6 Central and Eastern Europe         30        5.6   0.574   4.33  6.85 10.2
##  7 Western Europe                     20        6.9   0.680   5.29  7.77  9.85
##  8 Eastern Asia                        6        5.7   0.476   5.19  6.45  8.35
##  9 North America                       2        7.1   0.273   6.89  7.28  3.84
## 10 Australia and New Zealand           2        7.3   0.0559  7.23  7.31  0.765
```

Middle East & Northern show the greatest variability in happiness, meaning the happiness levels vary significantly among countries. Australia & New Zealand have the lowest COV, indicating that happiness levels are quite consistent across countries in these regions.

Since we are comparing countries in different regions, their data may be on different scales. • Variance alone is misleading when comparing datasets with different means. • CV standardizes variability, making it a better metric for comparing variation across different regions.

```
#III
par(mar=c(5,7,4,2))
bar <- as.matrix(tabS$Mean_Score)
barplot(bar,
        names.arg = tabS$Region,
        beside=TRUE,
        col= "skyblue",
        main = "Mean Happiness Score by Regions",
        cex.names= 0.8,
        las = 2,
        ylab = "Score",
        ylim = c(0,9)
        )
text(x = 1.5:length(tabS$Region),
     y = tabS$Mean_Score,
     labels= round(tabS$Mean_Score,1),
     pos = 3,
     cex = 0.8,
     col = "black")
```



9. Correlation Analyses

Let’s explore the linear relationship between `Score` and the other 6 factors.

- i. Create a correlation matrix for the 7 variables of interest using corr.test
- ii. Describe the strength, direction and significance of the linear relationship between Happiness Score and the other 6 variables.

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:rcompanion':
##
##     phi
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
#I
correlation_matrix <- corr.test(HP[3:9])
print(correlation_matrix$r, digits = 2)
```

```
##                 Score   GDP Family Life.Expectancy Freedom Trust Generosity
## Score           1.000  0.79  0.777            0.78    0.57  0.39      0.076
## GDP             0.794  1.00  0.755            0.84    0.38  0.30     -0.080
## Family          0.777  0.75  1.000            0.72    0.45  0.18     -0.048
## Life.Expectancy 0.780  0.84  0.719            1.00    0.39  0.30     -0.030
## Freedom         0.567  0.38  0.447            0.39    1.00  0.44      0.270
## Trust           0.386  0.30  0.182            0.30    0.44  1.00      0.327
## Generosity      0.076 -0.08 -0.048           -0.03    0.27  0.33      1.000
```

```
#p value
print(correlation_matrix$p, digits = 2)
```

```
##                   Score     GDP  Family Life.Expectancy Freedom   Trust
## Score           0.0e+00 8.6e-34 1.6e-31         7.2e-32 1.9e-13 7.3e-06
## GDP             4.3e-35 0.0e+00 8.8e-29         1.5e-40 1.1e-05 1.2e-03
## Family          9.0e-33 5.2e-30 0.0e+00         6.6e-25 6.7e-08 1.2e-01
## Life.Expectancy 3.8e-33 7.0e-42 4.1e-26         0.0e+00 5.6e-06 1.3e-03
## Freedom         1.2e-14 1.1e-06 4.8e-09         4.7e-07 0.0e+00 1.3e-07
## Trust           6.7e-07 1.5e-04 2.3e-02         1.8e-04 1.0e-08 0.0e+00
## Generosity      3.5e-01 3.2e-01 5.5e-01         7.1e-01 6.6e-04 3.2e-05
##                 Generosity
## Score              1.00000
## GDP                1.00000
## Family             1.00000
## Life.Expectancy    1.00000
## Freedom            0.00397
## Trust              0.00029
## Generosity         0.00000
```

```
#II
# if r > 0 (eg r= 0.850) => Strength: Strong positive relationship, Positive directio
n (as Factor1 increases, HappinessScore tends to increase). p <0.05 (eg: p= 0.001, so
the relationship is statistically significant.

#SIGNIFICANCE
#Score vs. GDP: p-value: 8.6e-34 (≈ 0.000) => Significance: Statistically significan
t.
#Score vs. Family: p-value: 1.6e-31 (≈ 0.000) => Significance: Statistically signific
ant.
#Score vs. Life.Expectancy: p-value: 7.2e-32 (≈ 0.000) => Significance: Statistically
significant.
#Score vs. Freedom: p-value: 1.9e-13 (≈ 0.000) => Significance: Statistically signifi
cant.
#Score vs. Trust: p-value: 7.3e-06 (≈ 0.000) => Significance: Statistically significa
nt.
#Score vs. Generosity: p-value: 1.00000 => Significance: Not statistically significan
t.

#STRENGTH & DIRECTION
#Score vs. GDP: Correlation (r): 0.794 => Strength: Strong positive relationship.Dire
ction: Positive (as GDP increases, Score tends to increase).
#Score vs. Family: (r): 0.777 => Strong positive relationship. Positive Direction (as
Family increases, Score tends to increase).
#Score vs. Life.Expectancy: (r): 0.780 => Strong positive relationship. Positive Dire
ction (as Life.Expectancy increases, Score tends to increase).
#Score vs. Freedom: (r): 0.567 => Moderate positive relationship. Positive Direction
(as Freedom increases, Score tends to increase).
#Score vs. Trust (r): 0.386 => Weak positive relationship.Positive Direction (as Trus
t increases, Score tends to increase slightly).
#Score vs. Generosity (r): 0.076 => Very weak or negligible relationship.Positive Dir
ection (as Generosity increases, Score tends to increase very slightly).
```

If r > 0 (eg r= 0.850) => Strength: Strong positive relationship, Positive direction (as Factor1 increases, HappinessScore tends to increase). p <0.05 (eg: p= 0.001, so the relationship is statistically significant.

We observe a strong positive correlation between Happiness Score and GDP. With a p-value smaller than 0.05, this correlation is statistically significant (significantly different from 0) GDP • Correlation: 0.79 (strong positive correlation) • Significance: p = 0.00 (highly significant) Freedom Correlation: 0.57 (moderate positive correlation) • Significance: p = 0.00 (highly significant) Generosity • Correlation: 0.08 (weak positive correlation) • Significance: p = 0.35 (not significant) …

STRENGTH & DIRECTION - Score vs. GDP: Correlation (r): 0.794 => Strength: Strong positive relationship.Direction: Positive (as GDP increases, Score tends to increase). - Score vs. Family: (r): 0.777 => Strong positive relationship. Positive Direction (as Family increases, Score tends to increase).

10.Computing Proportions and Probabilities for Sub-Saharan Africa region

- i. What proportion of countries have a happiness score that is lower than the average score in the region?
- ii. Assuming the Score data is normally distributed, what is the probability of a country having a Score more than 5?

```
#i
dfa <- HP[HP$Region == "Sub-Saharan Africa", ]
avgscore <- mean(dfa$Score)
P1 <- dfa %>% filter(Score<avgscore)
proportion <- nrow(P1)/nrow(dfa)
#ANSWER: the proportion of countries have a happiness score that is lower than the av
erage score in the region is 41,46%
#ii
sd_score <- sd(dfa$Score)
z_score <- (5 - avgscore) / sd_score
probability <- 1 - pnorm(z_score)
probability
```

```
## [1] 0.1696988
```

```
#ANSWER: the probability of a country having a Score more than 5 is 16,97%
```

Can use : pnorm(5,mean=mScore, sd(HPssa$Score),lower.tail = F) - pnorm(quantile, mean, sd) : Find probability from quantile - qnorm(probability, mean, sd):Find quantile from probability

```
## [1] "/Users/minhchau/Documents/BT1101 Past Year Papers/2223SEM1-BT1101"
```

```
##       Country                          Region Rank Score.2019   GDP Family
## 1 Afghanistan                   Southern Asia  154      3.203 0.350  0.517
## 2     Albania      Central and Eastern Europe  107      4.719 0.947  0.848
## 3     Algeria Middle East and Northern Africa   88      5.211 1.002  1.160
## 4   Argentina     Latin America and Caribbean   47      6.086 1.092  1.432
## 5     Armenia      Central and Eastern Europe  116      4.559 0.850  1.055
## 6   Australia       Australia and New Zealand   11      7.228 1.372  1.548
##   Life.Expectancy Freedom Trust Generosity Score.2018
## 1           0.361   0.000 0.025      0.158      3.632
## 2           0.874   0.383 0.027      0.178      4.586
## 3           0.785   0.086 0.114      0.073      5.295
## 4           0.881   0.471 0.050      0.066      3.795
## 5           0.815   0.283 0.064      0.095      6.388
## 6           1.036   0.557 0.290      0.332      4.321
```

11.Comparing Average Score in 2019 across Regions

i. Create a new categorical variable `Region2` to contain 4 possible values (Africa, American, Asia, Europe). `Region2` should be assigned the value of “Asia” if `Region` is “Australia and New Zealand” or contains “Asia”. `Region2` should be assigned the value of “Europe” if `Region` contains “Europe”. `Region2` should be assigned the value of “America” if `Region` contains “America” and `Region2` should be assigned the value of “Africa” if `Region` contains “Africa”.
ii. Display the mean, standard deviation, min and max of score in 2019 for each of the 4 regions (defined by `Region2`) in a table. Describe any interesting patterns/observations you can make from the table.
iii. Now, conduct the appropriate test(s) to assess if there is any difference in mean 2019 score across regions (as defined by `Region2` . State your hypotheses clearly, ensuring all symbols and groups are defined. Describe your conclusion with reference to the results. You may assume score to be normally distributed.

CODE

```
#i
dregion <- d2 %>% mutate(Region2 = case_when(
  Region == "Australia and New Zealand" ~ "Asia",
  grepl("Asia", Region) ~ "Asia",
  grepl("Europe", Region) ~ "Europe",
  grepl("America", Region) ~ "America",
  grepl("Africa", Region) ~ "Africa",
))

#ii
tabs <- dregion %>% group_by(`Region2`) %>%
  summarise(n(),
            Mean_Score = round(mean(Score.2019), 1),
            SD_Score = sd(Score.2019),
            min = min(Score.2019),
            max = max(Score.2019))
kable(tabs)
```

Region2	n()	Mean_Score	SD_Score	min	max
Africa	60	4.6	0.9138628	2.853	7.139
America	22	6.0	0.7830867	3.597	7.278
Asia	24	5.3	0.9506733	3.203	7.307
Europe	50	6.1	0.8990315	4.332	7.769

```
#iii

anova_result <- aov(Score.2019 ~ Region2, data = dregion)
anova_summary <- summary(anova_result)
anova_summary
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Region2       3  69.51  23.169   28.74 8.9e-15 ***
## Residuals   152 122.54   0.806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

II. Higher Mean in Europe & America => these regions have better performance. High Standard Deviation in Africa and Asia => This indicates large variations in scores. Low SD score in America and Europe => scores in these regions are more consistent Africa has the lowest min score, it means disparities.

III. Hypotheses Let M1, M2, M3, M4 be the median Score_2019 for Asia, America, Europe, and Africa. Null Hypothesis (H0): There is no significant difference in Score_2019 across regions. H0: M1=M2=M3=M4 Alternative Hypothesis (H1): At least one region has a significantly different median score. H1:At least one M is different As p_value = 8.9e-15 < 0.05, we reject the null hypothesis H0. This means there is a significant difference in mean Score.2019 across regions (Region2).

12. Score in 2018 and 2019

Conduct the appropriate test to evaluate if there is any significant change in countries’ happiness scores from 2018 to 2019. State your hypotheses clearly, ensuring all symbols and groups are defined. Describe your conclusion with reference to the results. You may assume score to be normally distributed.

Hypotheses Let Md be the mean difference in happiness scores between 2018 and 2019. Null Hypothesis (H0): There is no significant change in happiness scores from 2018 to 2019, i.e., the mean difference is zero. H0: Md = 0 Alternative Hypothesis (H1): There is a significant change in happiness scores from 2018 to 2019, i.e., the mean difference is not zero. H1: Md =/ 0

CODE

```
d3 <- d2 %>% select(Country, Score.2019, Score.2018)
t_test_result <- t.test(d3$Score.2019, d3$Score.2018, paired = TRUE)
print(t_test_result)
```

```
## 
##  Paired t-test
## 
## data:  d3$Score.2019 and d3$Score.2018
## t = 0.37142, df = 155, p-value = 0.7108
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.1346465  0.1970055
## sample estimates:
## mean difference 
##      0.03117949
```

#As p value > 0.05 → Fail to reject H0 => There is no significant change in happiness scores from 2018 to 2019.

13. 2019 Score in Europe

Focusing only on countries in Europe, conduct the appropriate test to evaluate if average happiness scores in 2019 is the same for countries in “Western Europe” and those in “Central and Eastern Europe”. State your hypotheses clearly, ensuring all symbols and groups are defined. Describe your conclusion with reference to the results. You may assume score to be normally distributed.

Hypotheses Let M1 represent the mean happiness score for countries in Western Europe in 2019, and M2 represent the mean happiness score for countries in Central and Eastern Europe in 2019. Null Hypothesis (H0): There is no significant difference in the average happiness scores between Western Europe and Central and Eastern Europe => H0: M1 = M2 Alternative Hypothesis (H1): There is a significant difference in the average happiness scores between Western Europe and Central and Eastern Europe => H1: M1 =/ M2

CODE

```
d4 <- d2 %>% filter(Region %in% c("Western Europe", "Central and Eastern Europe"))

t_test_result <- t.test(Score.2019 ~ Region, data = d4)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  Score.2019 by Region
## t = -7.2146, df = 36.022, p-value = 1.714e-08
## alternative hypothesis: true difference in means between group Central and Eastern
Europe and group Western Europe is not equal to 0
## 95 percent confidence interval:
##  -1.706002 -0.957331
## sample estimates:
## mean in group Central and Eastern Europe
##                                 5.566733
##             mean in group Western Europe
##                                 6.898400
```

#As p value < 0.05 → strong evidence to reject H0 => There is no significant difference in the average happiness scores between Western Europe and Central and Eastern Europe.

14. 2019 Scores in Asia

Let’s narrow down our analyses to the countries in Asia (Defined by Region2). Assuming that these countries are a sample of countries in whole of Asia,

i. compute the 95% confidence interval for the average score in 2019 for countries in Asia.
ii. compute the 99% confidence interval for proportion of countries with 2019 score exceeding 5.5.
iii. compute the 90% prediction interval for score in 2019.

Describe briefly what each of the above interval estimates tell us? From your result, could you conclude that true mean happiness score for countries in Asia is 5.5?

CODE

```
#i
asia_data <- dregion %>% filter(Region2 == "Asia")
meanA <- mean(asia_data$Score.2019, na.rm = TRUE)
sdA <- sd(asia_data$Score.2019, na.rm = TRUE)
n <- nrow(asia_data)
t_value <- qt(0.025, df = n - 1)
lower_bound <- meanA + (t_value * sdA * sqrt(1 + 1/n))
upper_bound <- meanA - (t_value * sdA * sqrt(1 + 1/n))
cbind(lower_bound, upper_bound)
```

```
##      lower_bound upper_bound
## [1,]    3.318621    7.332962
```

```
# we 95% confident that the true average happiness score in 2019 for all countries in
Asia falls within the range of 3.32 to 7.33.

#ii
asia_data <- dregion %>% filter(Score.2019 > 5.5)
pd1 <- nrow(asia_data) / nrow(d2)
lcipd1 <- pd1 + (qnorm(0.005)*sqrt(pd1*(1-pd1)/nrow(d2)))
ucipd1 <- pd1 - (qnorm(0.005)*sqrt(pd1*(1-pd1)/nrow(d2)))
print(cbind(lcipd1, ucipd1), digits=3)
```

```
##      lcipd1 ucipd1
## [1,]  0.365  0.571
```

```
# we are 99% confident that the true proportion of countries in the dataset with a 20
19 score exceeding 5.5 lies between 36.5% and 57.1%.

#iii
mnscr <- mean(d2$Score.2019)
sdscr <- sd(d2$Score.2019)
n <- nrow(d2)
t_value <- qt(0.05, df = n - 1)
lpi_scr <- mnscr + (t_value * sdscr * sqrt(1 + 1/n))
upi_scr <- mnscr - (t_value * sdscr * sqrt(1 + 1/n))
cbind(lpi_scr, upi_scr)
```

```
##       lpi_scr  upi_scr
## [1,] 3.559274 7.254919
```

We are 90% confident that the next individual country selected from the same population will have a 2019 score within the range of 3.559 to 7.255.

Based on 95% Confidence Interval for Mean, 99% Confidence Interval for Proportion, 90% Prediction Interval => we cannot reject the possibility that the true mean happiness score for countries in Asia is 5.5. However, we can say that 5.5 is a plausible value for the true mean => we cannot conclusively state that the true mean happiness score for countries in Asia is exactly 5.5, but we can’t rule it out either, as 5.5 falls within the 95% confidence interval.