R Statistical Analysis

Name: Yi

2024-07-01

```
library(ggplot2)
library(tidyselect)
library(dplyr)
```

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

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

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

World Happiness Statistical Analysis

The World Happiness Report consists of data on global happiness from over 156 countries released by the United Nations. The report has six indicators affecting the happiness index: GDP per capita, social support, healthy life expectancy, freedom, generosity, and corruption perception. The project aims to use filtering techniques, statistical modeling, data clustering, and regression models to examine the data set with the LaTeX and R programming languages.

• Load the `world_happiness.csv` file into R and rename the variables with appropriate headers.

```
# import the datasets, rename the variables, and show the top ten
happiness <- read.csv("world_happiness.csv",sep=",")
colnames(happiness) <- c("rank","country","score","income", "support", "health", "freedo
m", "generosity", "perceptions")

head(happiness, 10)
```

```
##    rank     country score income support health freedom generosity perceptions
## 1     1     Finland 7.769  1.340   1.587  0.986   0.596      0.153       0.393
## 2     2     Denmark 7.600  1.383   1.573  0.996   0.592      0.252       0.410
## 3     3      Norway 7.554  1.488   1.582  1.028   0.603      0.271       0.341
## 4     4     Iceland 7.494  1.380   1.624  1.026   0.591      0.354       0.118
## 5     5 Netherlands 7.488  1.396   1.522  0.999   0.557      0.322       0.298
## 6     6 Switzerland 7.480  1.452   1.526  1.052   0.572      0.263       0.343
## 7     7      Sweden 7.343  1.387   1.487  1.009   0.574      0.267       0.373
## 8     8 New Zealand 7.307  1.303   1.557  1.026   0.585      0.330       0.380
## 9     9      Canada 7.278  1.365   1.505  1.039   0.584      0.285       0.308
## 10   10     Austria 7.246  1.376   1.475  1.016   0.532      0.244       0.226
```

• Display the number of attributes in the World Happiness database counted by columns and rows.

```
ncol(happiness)
```

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

```
nrow(happiness)
```

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

• Generate an overview with the `summary` function and specify whether there are zero values.

```
summary(happiness)
```

```
##       rank          country              score           income      
##  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  
##     support          health          freedom         generosity    
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1.056   1st Qu.:0.5477   1st Qu.:0.3080   1st Qu.:0.1087  
##  Median :1.272   Median :0.7890   Median :0.4170   Median :0.1775  
##  Mean   :1.209   Mean   :0.7252   Mean   :0.3926   Mean   :0.1848  
##  3rd Qu.:1.452   3rd Qu.:0.8818   3rd Qu.:0.5072   3rd Qu.:0.2482  
##  Max.   :1.624   Max.   :1.1410   Max.   :0.6310   Max.   :0.5660  
##   perceptions    
##  Min.   :0.0000  
##  1st Qu.:0.0470  
##  Median :0.0855  
##  Mean   :0.1106  
##  3rd Qu.:0.1412  
##  Max.   :0.4530  
```

```
sum(is.na(happiness))
```

```
## [1] 0
```

• Calculate the mean and standard deviation for the World Happiness database.

- The mean is the average calculated by adding up all the values and dividing by the total numbers using the formula $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. The standard deviation (sd) is the distance from the mean with the formula $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$. The low sd value of 1.11312 means data points are closer to the mean compared to points with +2 or +3 deviation.

```
mean(happiness$score)
```

```
## [1] 5.407096
```

```
sd(happiness$score)
```

```
## [1] 1.11312
```

• Display the correlation between the World Happiness attributes in the datasets.

- The Pearson correlation coefficient measures the correlation between two values. The values closer to 1 present a strong relationship, and a correlation closer to 0 suggests a weak relationship. The example shows a strong correlation between the happiness score and GDP per capita, suggesting that more income leads to a higher level of happiness.

```
cor(happiness$score, happiness$income)
```

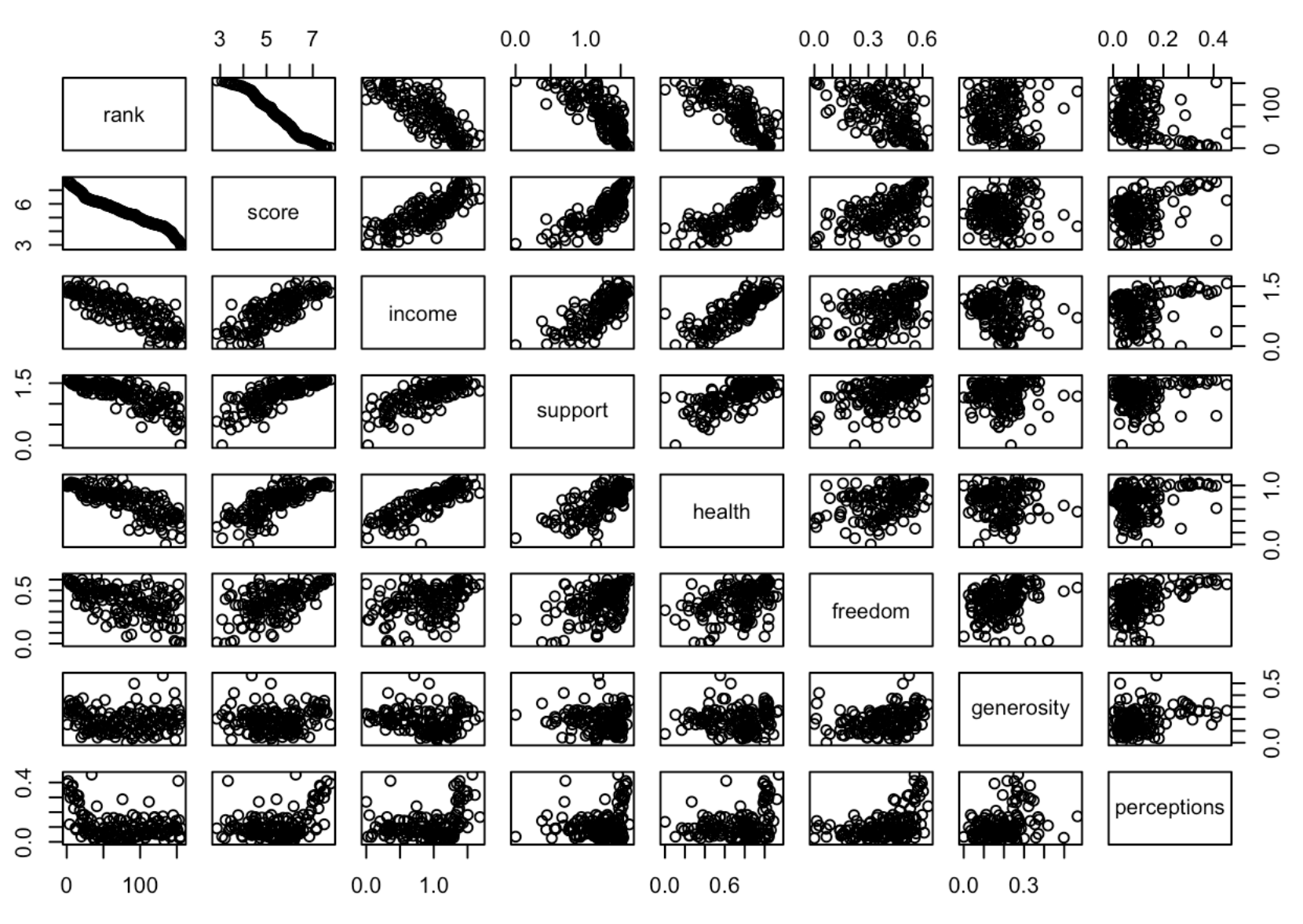
```
## [1] 0.7938829
```

Included Plots

• Create a pairwise scatterplot in R that displays the relationship between the continuous variables.

- The scatterplot shows a positive correlation between GDP per capita and the attributes (social support, healthy life expectancy, freedom, and corruption perceptions). The linear slope (upwards or downwards) between the data points indicates a positive correlation. The points that are scattered across the plot suggest a weaker relationship.

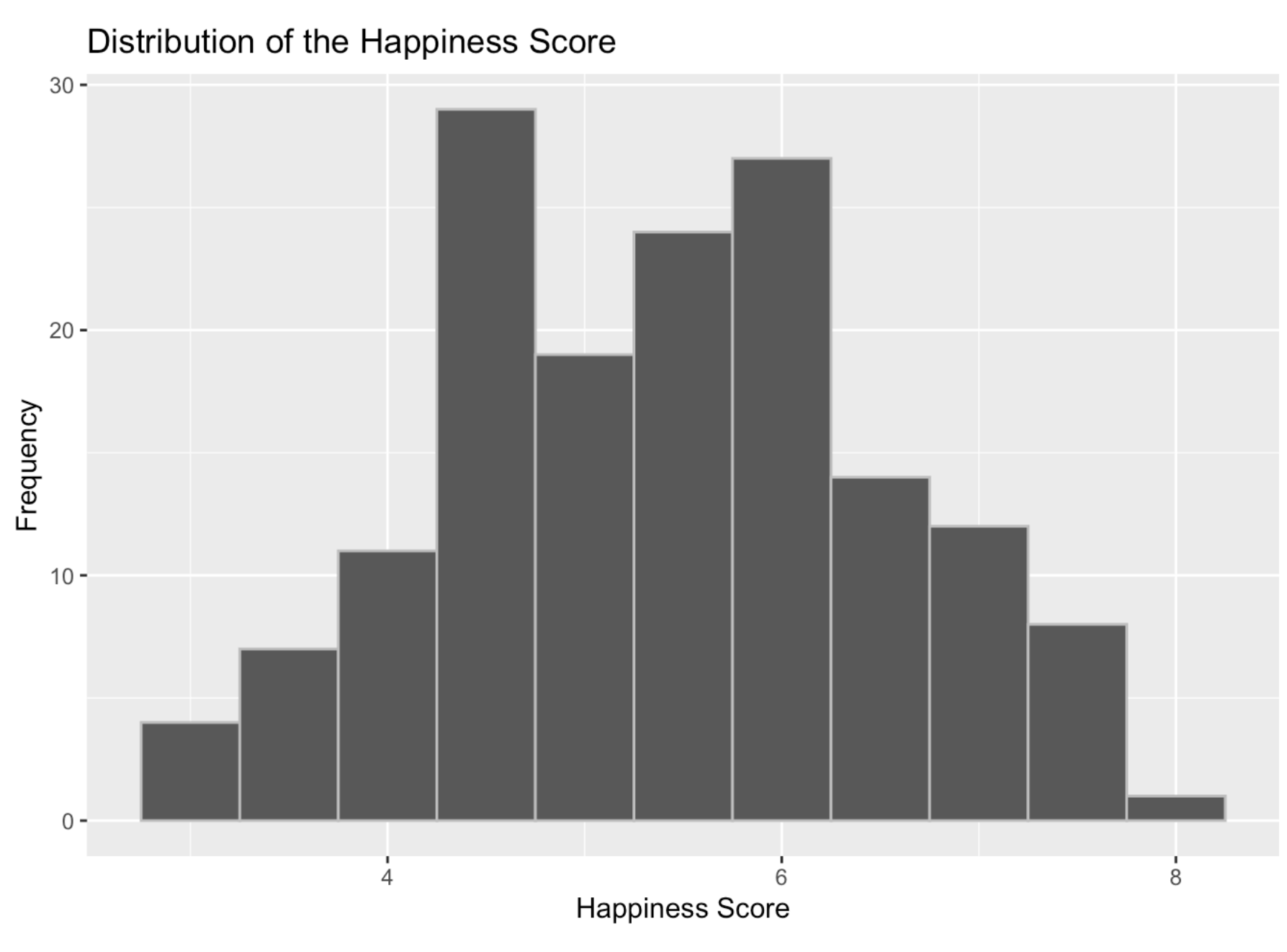
```
pairs(happiness[c("rank","score","income", "support", "health", "freedom", "generosity",
"perceptions")])
```



• Create a scatter plot and histogram to show the relationship between the variables.

- The histogram displays the distribution of the happiness score based on frequency. The Normal Distribution shows that the data points are symmetrically distributed around the center, and the standard deviation is spread evenly. The normal distribution suggests that most countries have an average happiness score of 5 to 7.

```
# Histogram to show the frequency of happiness
ggplot(happiness, aes(x = score)) +
  geom_histogram(binwidth = 0.5, col='gray',) +
  labs(title = "Distribution of the Happiness Score", x = "Happiness Score", y = "Freque
ncy")
```



• Create a regression analysis between the variables in the database.

- The regression model shows the relationship between the dependent and the independent variables. The distribution of the residuals is near zero indicating that the data point is approximate to a linear model. The low coefficient p-value implies that the model is statistically significant. The high Multiple R-squared values of 0.7425 and Adjusted R-squared of 0.7374 indicate a well-fit model.

```
model <- lm(score ~ income + health + freedom, data = happiness)
summary(model)
```

```
##
## Call:
## lm(formula = score ~ income + health + freedom, data = happiness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94349 -0.35976  0.07486  0.42184  1.02208
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4201     0.1667  14.519  < 2e-16 ***
## income        1.1781     0.2104   5.599 9.84e-08 ***
## health        1.4578     0.3480   4.189 4.73e-05 ***
## freedom       2.1993     0.3492   6.298 3.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5704 on 152 degrees of freedom
## Multiple R-squared:  0.7425, Adjusted R-squared:  0.7374
## F-statistic: 146.1 on 3 and 152 DF,  p-value: < 2.2e-16
```

```
model
```

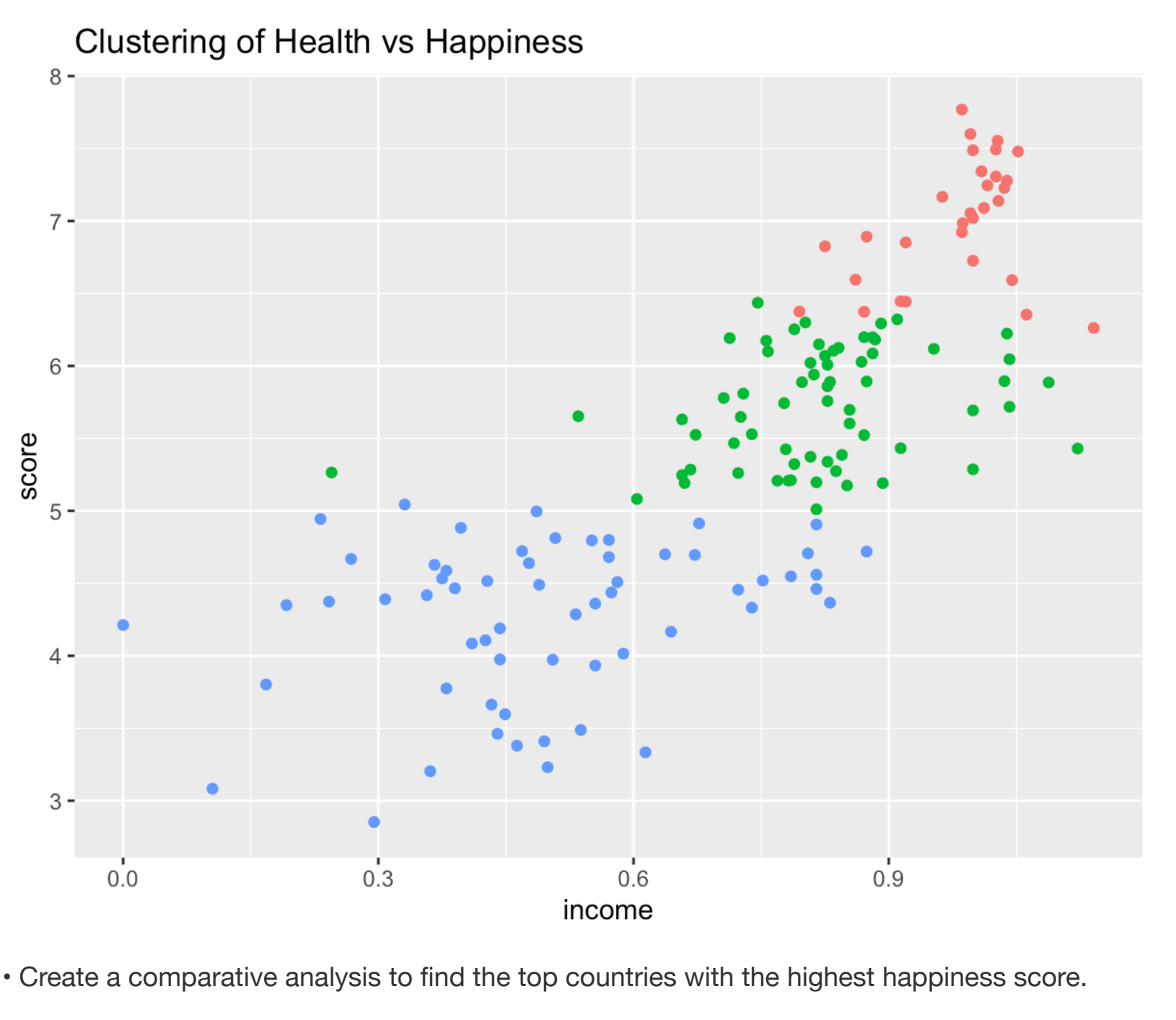
```
##
## Call:
## lm(formula = score ~ income + health + freedom, data = happiness)
##
## Coefficients:
## (Intercept)       income       health      freedom
##       2.420        1.178        1.458        2.199
```

• Create a clustering model to display the relationship between the variables.

- The K-means clustering model aims to group similar cluster points in the World Happiness database. The data points for the four attributes: happiness score, GDP per capita, social support, and healthy life expectancy are scattered in a linear direction. This suggests that as the GDP Per Capita increases, other attributes like social support, and healthy life expectancy also increase.

```
set.seed(123)
clusters <- kmeans(happiness[, c("score", "income", "support", "health")], centers = 3)
happiness$cluster <- as.factor(clusters$cluster)

ggplot(happiness, aes(x = health, y = score, color = cluster)) +
  geom_point() +
  labs(title = "Clustering of Health vs Happiness", x = "income", y = "score")
```



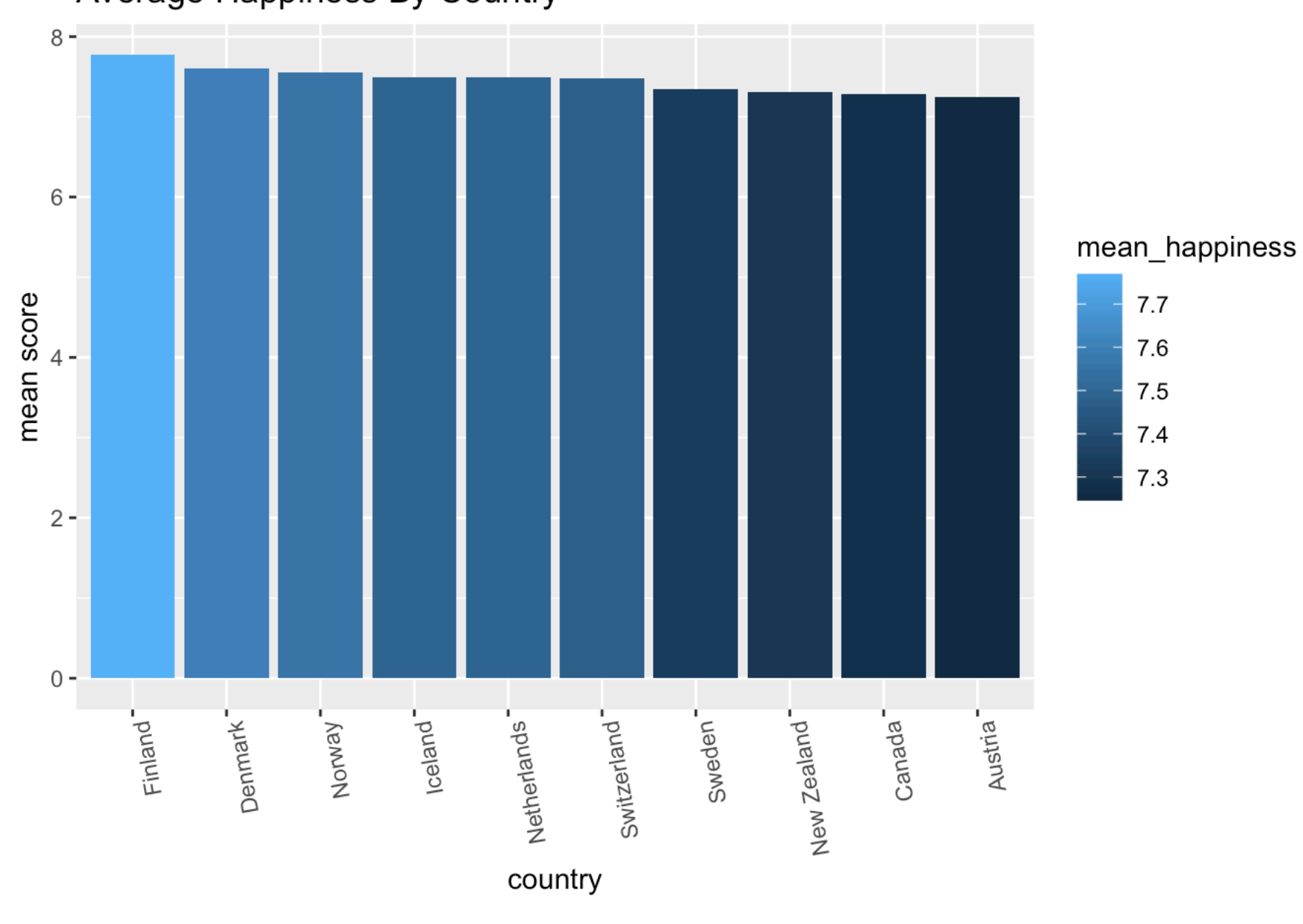
• Create a comparative analysis to find the top countries with the highest happiness score.

- The plot shows that the places with the highest happiness score are regions in Europe including Finland, Denmark, Norway, etc. followed by other economically developed countries.

```
country_happiness <- happiness %>%
  group_by(country) %>%
  summarise(mean_happiness = mean(score))

top_regions <- country_happiness %>%
  arrange(desc(mean_happiness)) %>%
  head(10)

ggplot(top_regions, aes(x = reorder(country, -mean_happiness), y = mean_happiness, fill=
mean_happiness)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Happiness By Country", x = "country", y = "mean score") +
  theme(axis.text.x = element_text(angle = 100, hjust = 1))
```



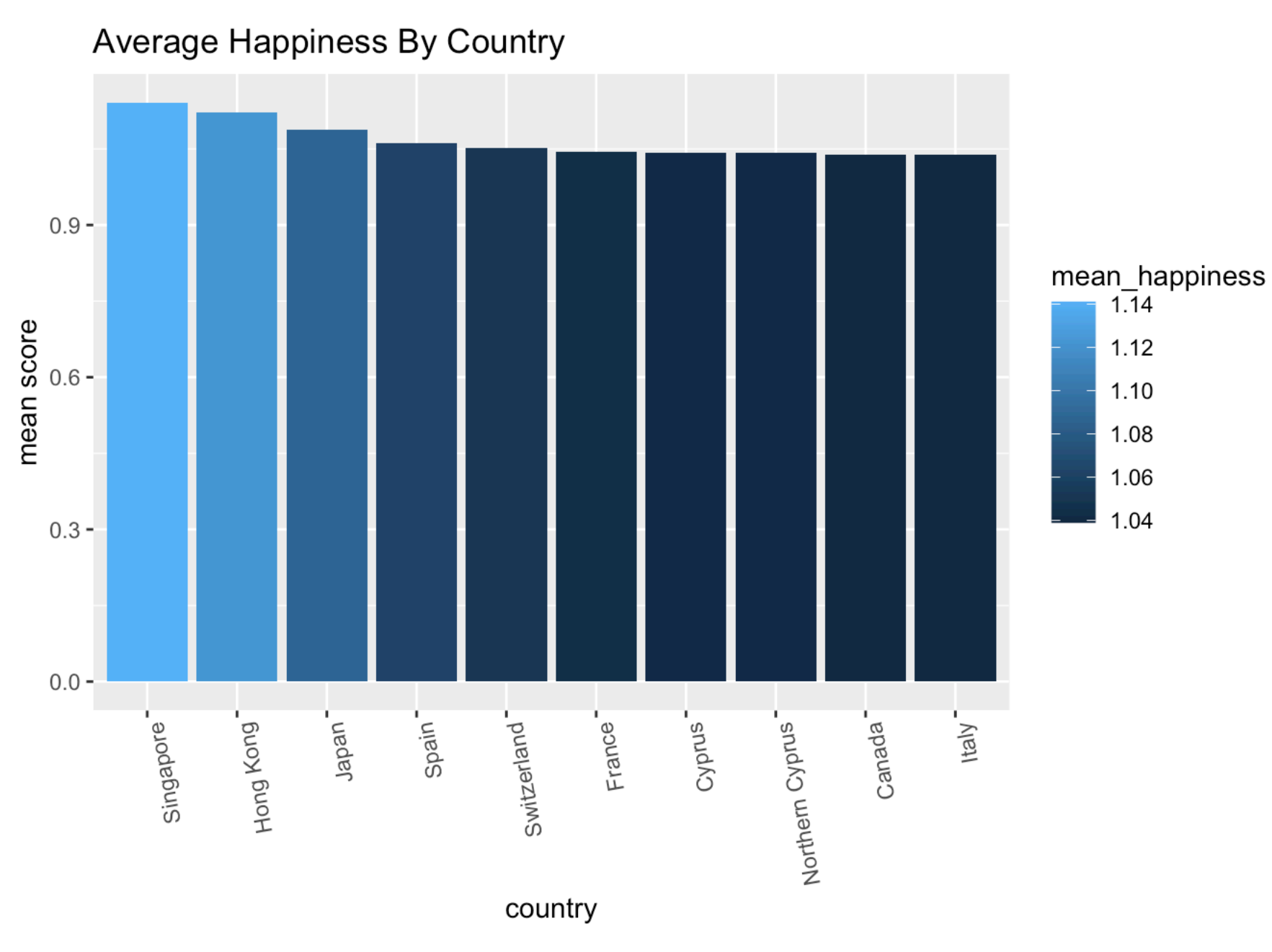
• Create a comparative analysis to find the countries with the highest health score.

- The plot shows that the places with the highest health scores are regions in East Asia including Singapore, Hong Kong, and Japan followed by Europe countries.

```
country_happiness <- happiness %>%
  group_by(country) %>%
  summarise(mean_happiness = mean(health))

top_regions <- country_happiness %>%
  arrange(desc(mean_happiness)) %>%
  head(10)

ggplot(top_regions, aes(x = reorder(country, -mean_happiness), y = mean_happiness, fill
= mean_happiness)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Happiness By Country", x = "country", y = "mean score") +
  theme(axis.text.x = element_text(angle = 100, hjust = 1))
```



Conclusion

In conclusion, the result shows that the attribute income (GDP per capita) positively influenced the happiness score in the country. This means that improving the GDP per capita for each country will also increase the country's happiness index, and permit individuals to access more resources. Economic resources in the country improve the quality of life for most people and influence other criteria including life expectancy, social support, freedom, and safety.