Wealth Analysis of Billionaires and Economic Indicators

Name: Kirtana Sridharan

Topic: Wealth Analysis of Billionaires and Economic Indicators

Why I chose this topic:

I chose to explore this topic because I am interested in the relationship between wealth inequality and economic growth. Billionaires represent the extreme upper end of the wealth spectrum, and their wealth can have a significant impact on the economy. For example, billionaires may invest their wealth in businesses, which can create jobs and boost economic growth. However, they may also use their wealth to influence public policy in ways that benefit themselves at the expense of the general public.

By analyzing the wealth distribution of billionaires and its correlation with economic indicators, I hope to gain insights into the socioeconomic landscape and identify potential areas for policy intervention.

Data selection and refinement:

To select the data for my analysis, I considered the following factors:

Relevance: The data should be relevant to my research questions, which are:

a) What are the factors influencing billionaire wealth?

b) What is the relationship between billionaire wealth and economic indicators?

The data should cover a wide range of billionaires and economic indicators, to ensure that my analysis is robust.

Quality: The data should be from reliable sources and should be well-maintained.

I identified a comprehensive dataset that includes information about billionaires' net worth, gender, birthplace, and various economic indicators such as GDP, CPI, and tax rates. I then cleaned the data, handled missing values, and converted relevant columns to numeric formats for statistical analysis. The link to the dataset can be found below: https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset

Choice of graphs and why:

I chose the following graph types:

1. Bar charts: Bar charts are useful for comparing different categories of data. For example, I used a bar chart to compare the number of billionaires in different countries.

2. Pie charts: Pie charts are useful for showing the distribution of data parts of a whole. For example, I used a pie chart to show the distribution of self-made vs. inherited wealth among billionaires.

3. Scatter plots: Scatter plots are useful for identifying correlations between two variables. For example, I used a scatter plot to identify the correlation between billionaire wealth and GDP.

4. Geospatial distribution: Geospatial distribution maps are a valuable tool for visualizing and analyzing geographic data. I used a geospatial distribution of billionaires map to show the geographic distribution of billionaire wealth around the world.

I chose these graph types because they are effective in communicating the information in the charts in a clear and concise way. Bar charts are good for comparing different categories of data, pie charts are good for showing the distribution of data parts of a whole, and scatter plots are good for identifying correlations between two variables.

```
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
```

```
library(ggplot2)
library(leaflet)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
library(readr)

data <- read_csv('/Users/kirtanasridharan/Desktop/COLLEGE WORK/FA-550/Billionaires_Statistics_Dataset.cs
```

```
## Rows: 2640 Columns: 35

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (18): category, personName, country, city, source, industries, countryOf...
## dbl (16): rank, finalWorth, age, birthYear, birthMonth, birthDay, cpi_countr...
## lgl  (1): selfMade
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df <-data.frame(data)
summary(df)
```

```
##       rank        finalWorth       category          personName
##  Min.   :   1   Min.   :  1000   Length:2640        Length:2640
##  1st Qu.: 659   1st Qu.:  1500   Class :character   Class :character
##  Median :1312   Median :  2300   Mode  :character   Mode  :character
##  Mean   :1289   Mean   :  4624
##  3rd Qu.:1905   3rd Qu.:  4200
##  Max.   :2540   Max.   :211000
##
##       age           country              city              source
##  Min.   : 18.00   Length:2640        Length:2640        Length:2640
##  1st Qu.: 56.00   Class :character   Class :character   Class :character
##  Median : 65.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 65.14
##  3rd Qu.: 75.00
##  Max.   :101.00
##  NA's   :65
##   industries        countryOfCitizenship organization        selfMade
##  Length:2640        Length:2640          Length:2640        Mode :logical
##  Class :character   Class :character     Class :character   FALSE:828
##  Mode  :character   Mode  :character     Mode  :character   TRUE :1812
##
##
##
##
##     status             gender           birthDate           lastName
##  Length:2640        Length:2640        Length:2640        Length:2640
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    firstName            title               date              state
##  Length:2640        Length:2640        Length:2640        Length:2640
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  residenceStateRegion   birthYear      birthMonth       birthDay
##  Length:2640          Min.   :1921   Min.   : 1.00   Min.   : 1.0
##  Class :character     1st Qu.:1948   1st Qu.: 2.00   1st Qu.: 1.0
##  Mode  :character     Median :1957   Median : 6.00   Median :11.0
##                       Mean   :1957   Mean   : 5.74   Mean   :12.1
##                       3rd Qu.:1966   3rd Qu.: 9.00   3rd Qu.:21.0
##                       Max.   :2004   Max.   :12.00   Max.   :31.0
##                       NA's   :76     NA's   :76      NA's   :76
##    cpi_country     cpi_change_country gdp_country
##  Min.   : 99.55   Min.   :-1.900     Length:2640
##  1st Qu.:117.24   1st Qu.: 1.700     Class :character
##  Median :117.24   Median : 2.900     Mode  :character
##  Mean   :127.76   Mean   : 4.364
##  3rd Qu.:125.08   3rd Qu.: 7.500
```

```
##  Max.   :288.57   Max.   :53.500
##  NA's   :184      NA's   :184
##  gross_tertiary_education_enrollment gross_primary_education_enrollment_country
##  Min.   :  4.00                      Min.   : 84.7
##  1st Qu.: 50.60                      1st Qu.:100.2
##  Median : 65.60                      Median :101.8
##  Mean   : 67.23                      Mean   :102.9
##  3rd Qu.: 88.20                      3rd Qu.:102.6
##  Max.   :136.60                      Max.   :142.1
##  NA's   :182                         NA's   :181
##  life_expectancy_country tax_revenue_country_country total_tax_rate_country
##  Min.   :54.30           Min.   : 0.10               Min.   :  9.90
##  1st Qu.:77.00           1st Qu.: 9.60               1st Qu.: 36.60
##  Median :78.50           Median : 9.60               Median : 41.20
##  Mean   :78.12           Mean   :12.55               Mean   : 43.96
##  3rd Qu.:80.90           3rd Qu.:12.80               3rd Qu.: 59.10
##  Max.   :84.20           Max.   :37.20               Max.   :106.30
##  NA's   :182             NA's   :183                 NA's   :182
##  population_country  latitude_country longitude_country
##  Min.   :3.802e+04   Min.   :-40.90   Min.   :-106.35
##  1st Qu.:6.683e+07   1st Qu.: 35.86   1st Qu.: -95.71
##  Median :3.282e+08   Median : 37.09   Median :  10.45
##  Mean   :5.102e+08   Mean   : 34.90   Mean   :  12.58
##  3rd Qu.:1.366e+09   3rd Qu.: 40.46   3rd Qu.: 104.20
##  Max.   :1.398e+09   Max.   : 61.92   Max.   : 174.89
##  NA's   :164         NA's   :164      NA's   :164
```

```
dim(df)
```

```
## [1] 2640   35
```

```
sum(is.na(df))
```

```
## [1] 10812
```

```
# But removing the NA values would drop the number drastically.
# So, for now, I'll keep the dirty data, in order to mantain the data abundance

df_clean <- na.omit(df)

# Print the column names with numeric data type
numeric_columns <- sapply(df, is.numeric)
print(names(df[numeric_columns]))
```

```
##  [1] "rank"
##  [2] "finalWorth"
##  [3] "age"
##  [4] "birthYear"
##  [5] "birthMonth"
##  [6] "birthDay"
##  [7] "cpi_country"
##  [8] "cpi_change_country"
```

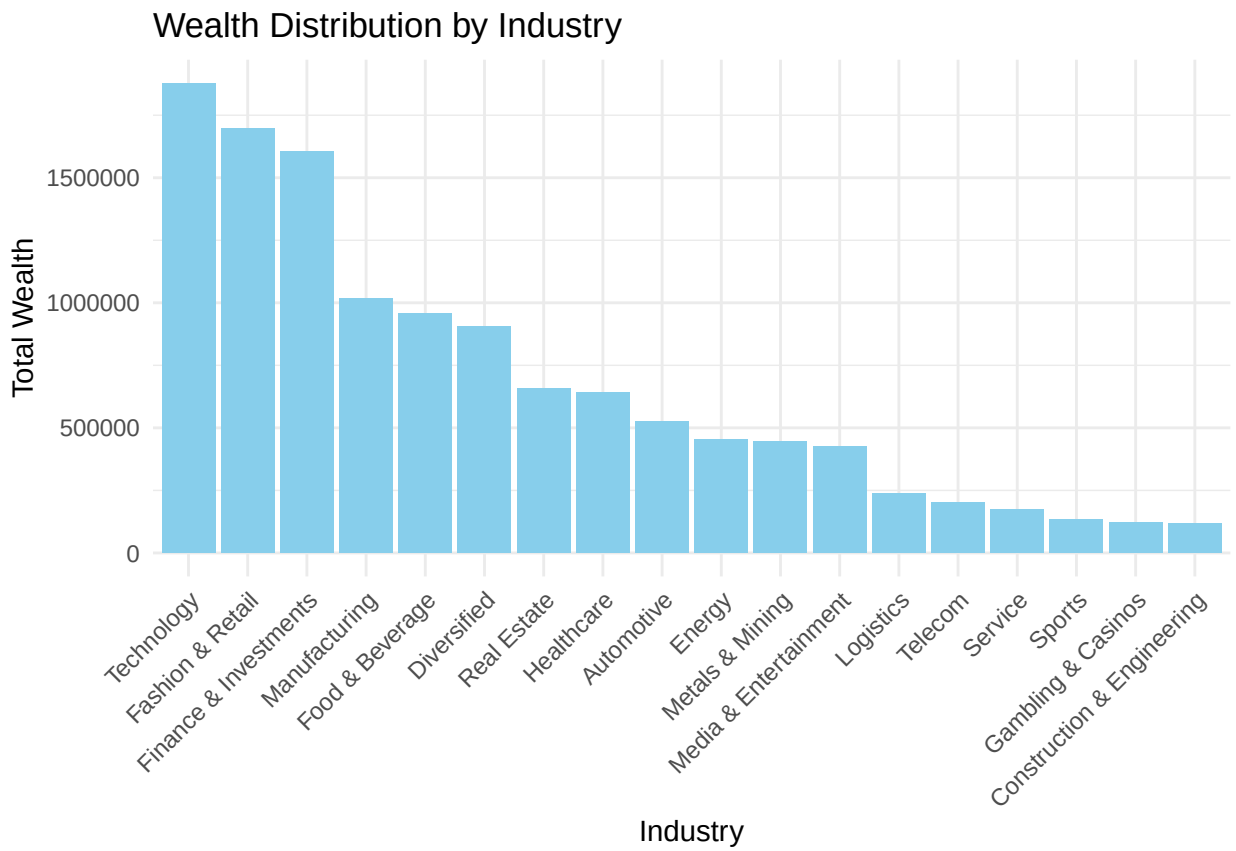
```
##  [9] "gross_tertiary_education_enrollment"
## [10] "gross_primary_education_enrollment_country"
## [11] "life_expectancy_country"
## [12] "tax_revenue_country_country"
## [13] "total_tax_rate_country"
## [14] "population_country"
## [15] "latitude_country"
## [16] "longitude_country"
```

```
# Converting gdp_country from character to numeric type
df$gdp_country <- as.numeric(gsub("[^0-9.]", "", df$gdp_country))

# Check the structure of the data frame
str(df$gdp_country)
```

```
##  num [1:2640] 2.72e+12 2.14e+13 2.14e+13 2.14e+13 2.14e+13 ...
```

Industry wise wealth distribution

```
industry_distribution <- df %>%
  group_by(industries) %>%
  summarize(total_wealth = sum(finalWorth, na.rm = TRUE))

ggplot(industry_distribution, aes(x = reorder(industries, -total_wealth), y = total_wealth)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Wealth Distribution by Industry", x = "Industry", y = "Total Wealth") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Analysis: The top five industries with the highest wealth concentration are:

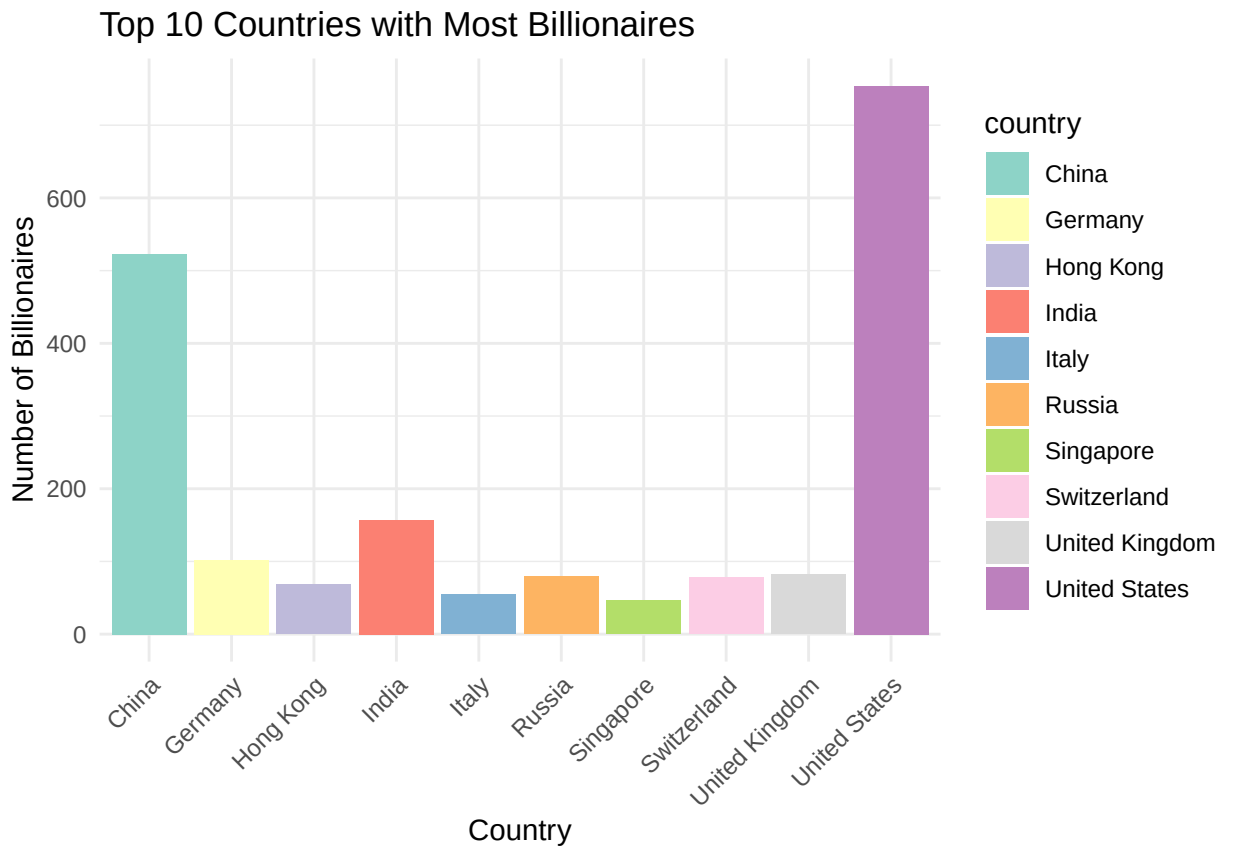
1. Technology
2. Fashion & Retail
3. Finance & Investments
4. Manufacturing
5. Food & Beverage

The chart suggests that billionaires are disproportionately concentrated in a few key industries, such as technology and fashion & retail. This is likely due to the high profits and low barriers to entry in these industries.

Plotting the 10 countries with the most number of billionaires

```
top_countries <- df %>%
  count(country) %>%
  arrange(desc(n)) %>%
  head(10)

ggplot(top_countries, aes(x = country, y = n, fill = country)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 10 Countries with Most Billionaires", x = "Country", y = "Number of Billionaires") +
  scale_fill_brewer(palette = "Set3") +
```

```
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Analysis: The chart shows that the United States and China are the two countries with the most billionaires. This is likely due to the size and economic growth of these two countries. The United States has the largest economy in the world, and China has the second-largest economy. The United States and China are the two countries with the most billionaires, followed by India, Germany, and the United Kindgom.

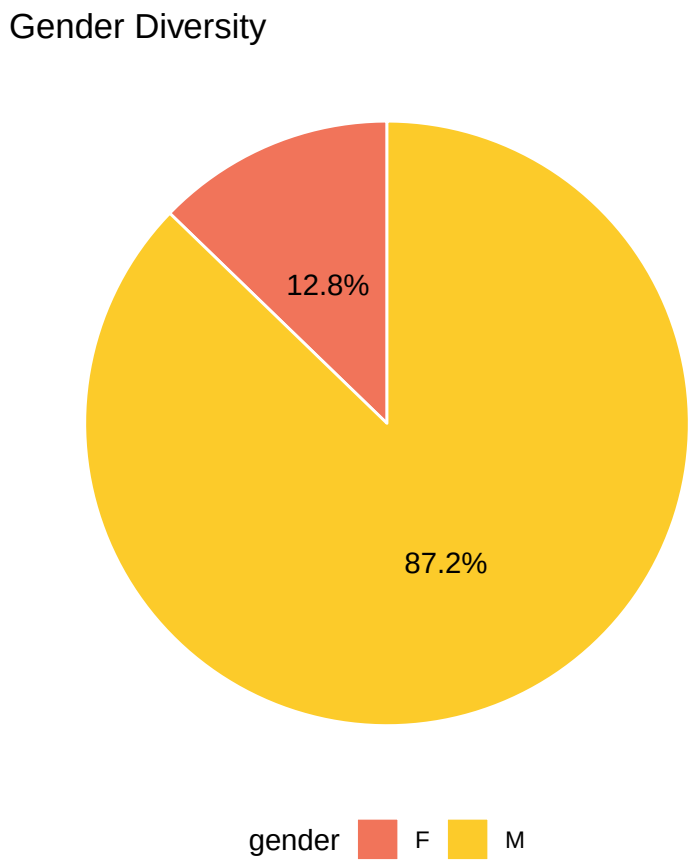
The other countries in the top 10 are also all developed countries with strong economies. However, the number of billionaires in these countries is significantly lower than in the United States and China.

The chart also shows that the distribution of billionaires is uneven across the world. The vast majority of billionaires are concentrated in a small number of developed countries.

Demographic Analysis - Gender diversity and Age Distribution

```
gender_diversity <- as.data.frame(table(df$gender))

# Rename the columns
names(gender_diversity) <- c("gender", "count")

ggplot(gender_diversity, aes(x = "", y = count, fill = gender)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y") +
  scale_fill_manual(values = c("#F1745A", "#FCCB2A")) +
```

```
  theme_void() +
  theme(legend.position = "bottom") +
  ggtitle("Gender Diversity") +
  geom_text(aes(label = sprintf('%1.1f%%', (count/sum(count)*100))), position = position_stack(vjust = 
```



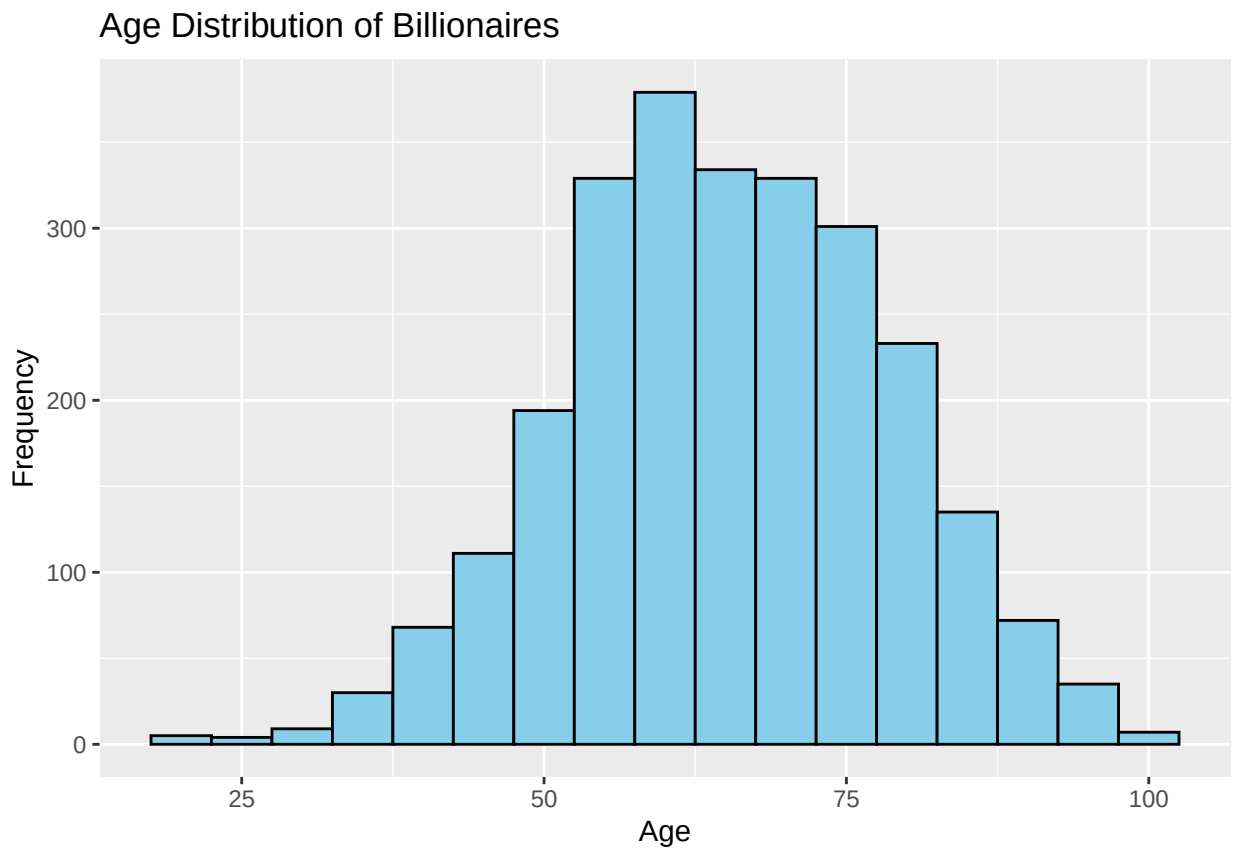
Analysis: Women are significantly underrepresented among the world's billionaires.

There are a number of possible explanations for this gender gap, such as unpaid care work, workplace discrimination, and a lower likelihood of starting businesses.

The gender gap among billionaires is a matter of concern, as it suggests that women are not being given equal opportunities to achieve financial success.

```
ggplot(df, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Age Distribution of Billionaires", x = "Age", y = "Frequency")
```

```
## Warning: Removed 65 rows containing non-finite values (stat_bin).
```



Analysis: The age distribution of billionaires is skewed towards the older ages, with the majority of billionaires being between the ages of 50 and 75.

This suggests that most billionaires have had time to accumulate wealth through their careers and investments.

However, there is a significant number of billionaires under the age of 50 and over the age of 75, which suggests that it is possible to become and remain a billionaire at any age, but it is relatively rare.

Economic Indicators

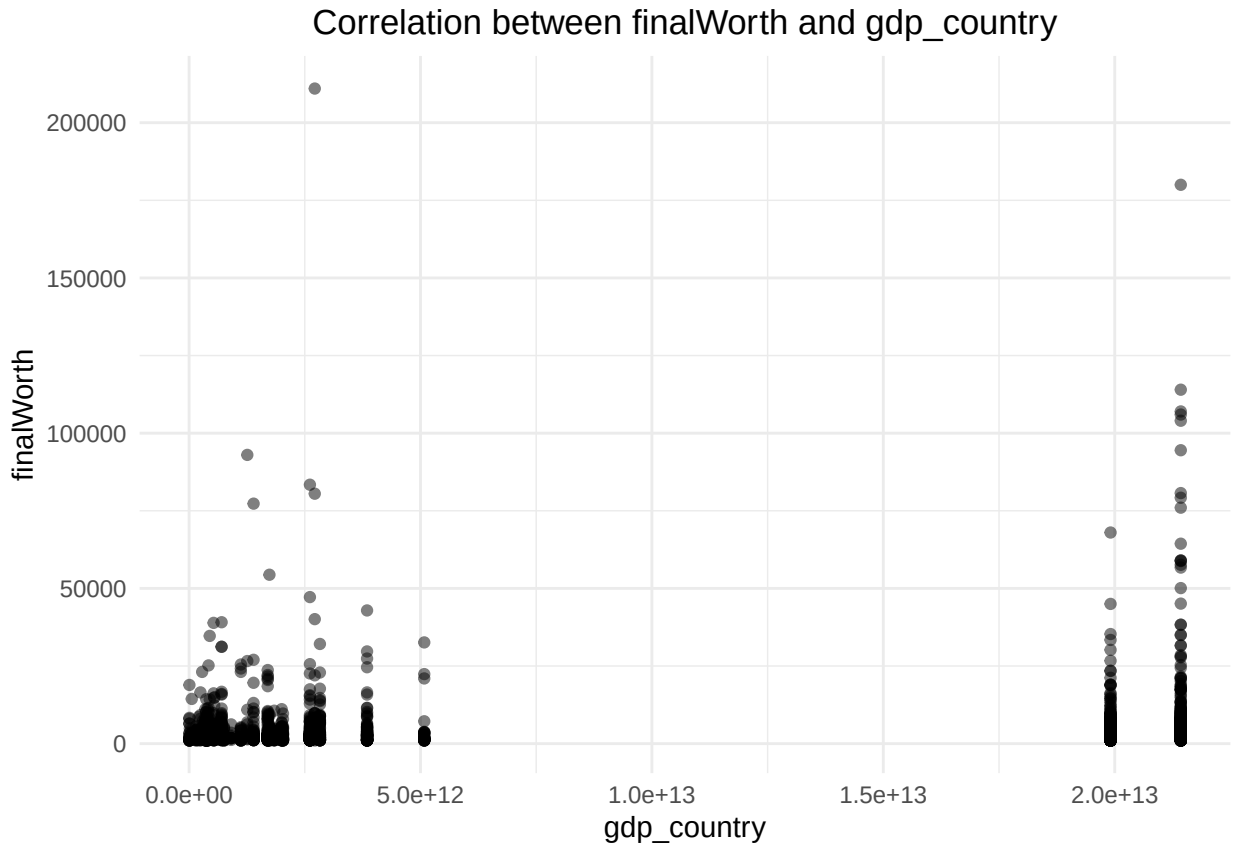
```
df_filtered <- df[c('finalWorth', 'gdp_country', 'cpi_country', 'total_tax_rate_country')]
df_filtered$gdp_country <- as.numeric(df_filtered$gdp_country)
df_filtered$cpi_country <- as.numeric(df_filtered$cpi_country)
df_filtered$total_tax_rate_country <- as.numeric(df_filtered$total_tax_rate_country)

columns <- c('gdp_country', 'cpi_country', 'total_tax_rate_country')

# Printing the results
for (column in columns) {
  correlation <- cor(df_filtered$finalWorth, df_filtered[[column]], use = "complete.obs")
  print(paste("Correlation between finalWorth and", column, ":", format(correlation, digits=2)))
}
```

```
## [1] "Correlation between finalWorth and gdp_country : 0.038"
```

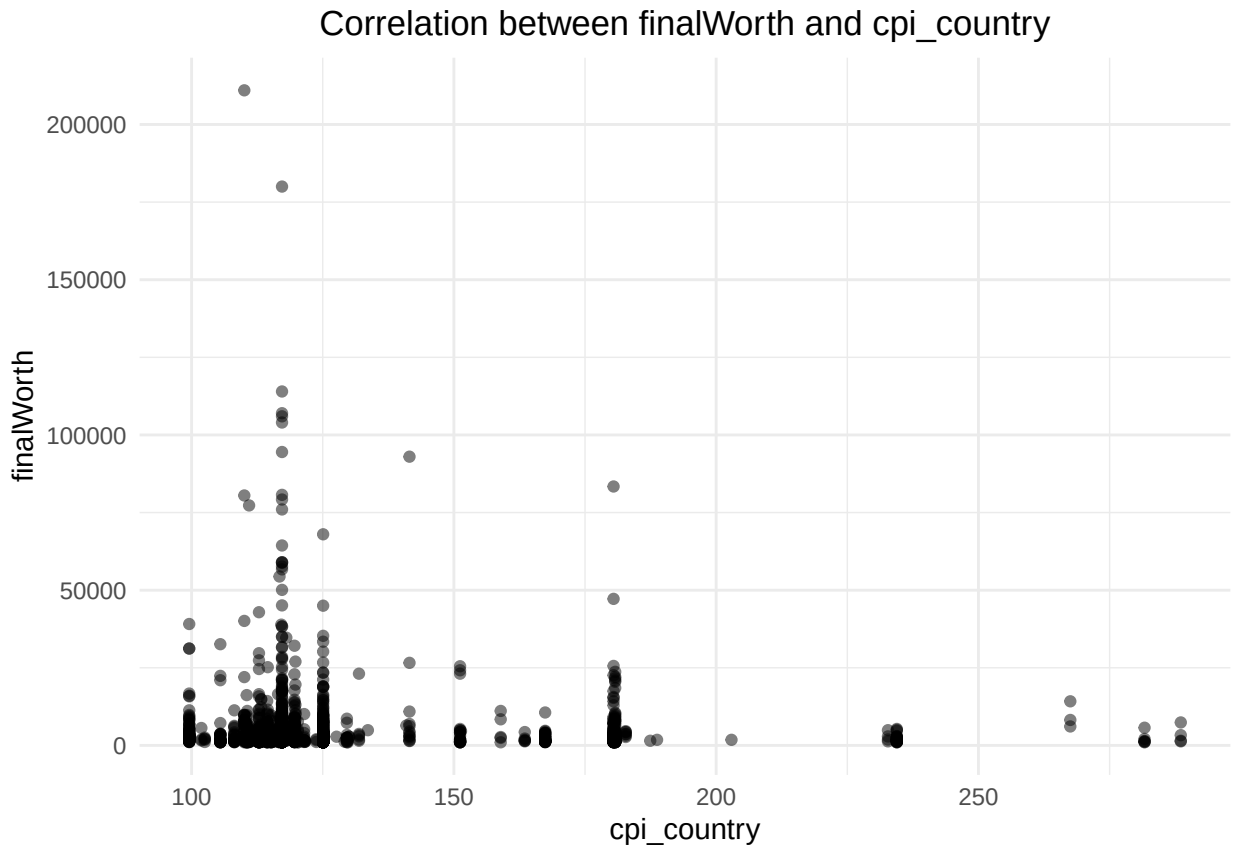
```
## [1] "Correlation between finalWorth and cpi_country : -0.043"
## [1] "Correlation between finalWorth and total_tax_rate_country : -0.036"
```

```r
columns <- c('gdp_country', 'cpi_country', 'total_tax_rate_country')

for (column in columns) {
  # Create a scatter plot
  plot <- ggplot(df_filtered, aes_string(x = column, y = 'finalWorth')) +
    geom_point(alpha = 0.5) +
    labs(title = paste("Correlation between finalWorth and", column),
         x = column,
         y = 'finalWorth') +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5)) +

    theme_minimal() + theme(plot.title = element_text(hjust = 0.5))

  print(plot)
}
```

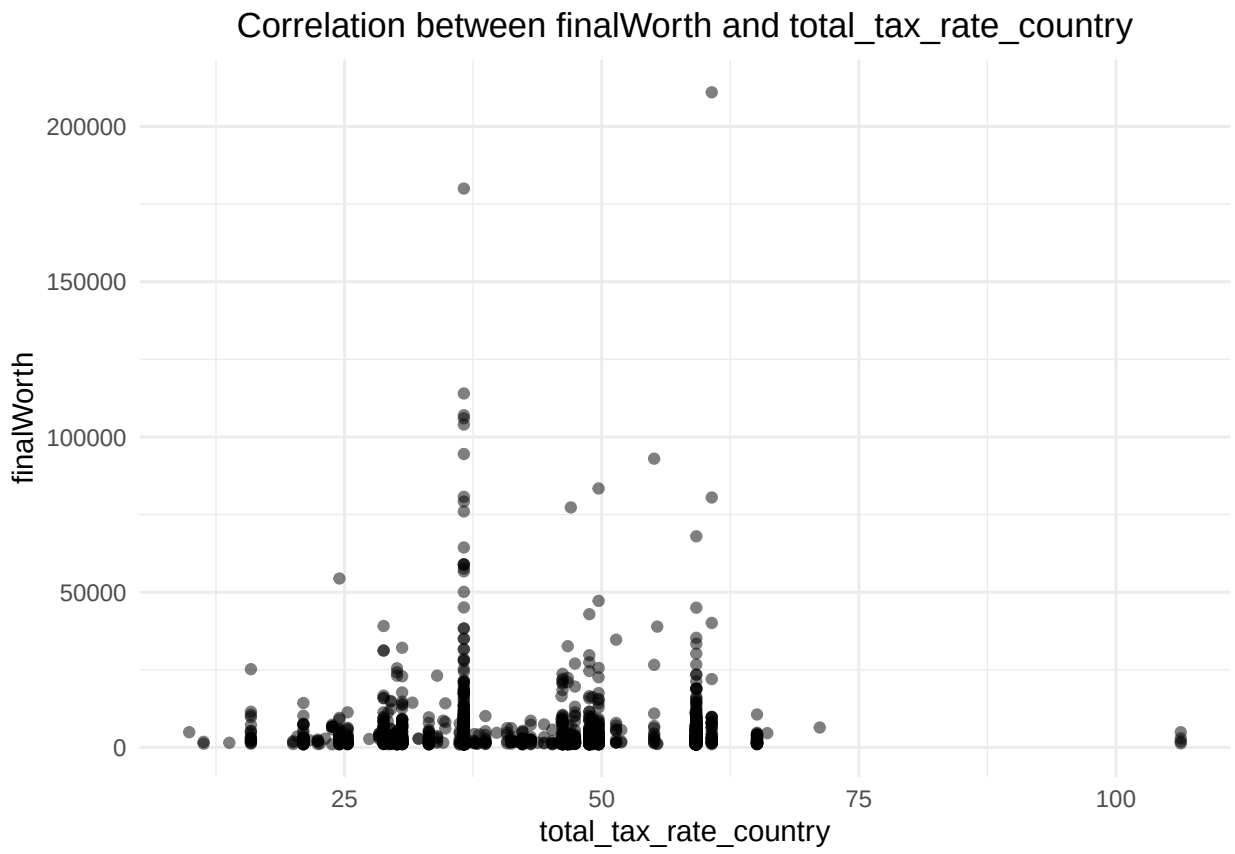
```
## Warning: Removed 164 rows containing missing values (geom_point).
```




```
## Warning: Removed 184 rows containing missing values (geom_point).
```




```
## Warning: Removed 182 rows containing missing values (geom_point).
```



Analysis:

a) Final worth and GDP: A correlation of 0.038 means that there is a very weak positive correlation between final worth and GDP. This means that countries with higher GDPs tend to have more billionaires.

b) Final worth and CPI: A correlation of -0.043 means that there is a very weak negative correlation between final worth and CPI. This means that countries with higher inflation tend to have fewer billionaires.

c) Final worth and tax rate: A correlation of -0.036 means that there is a very weak negative correlation between final worth and tax rate. This means that countries with higher tax rates tend to have fewer billionaires.

Overall, the weak correlations between final worth and GDP, CPI, and tax rate suggest that there is no clear relationship between these variables

Geospatial Analysis

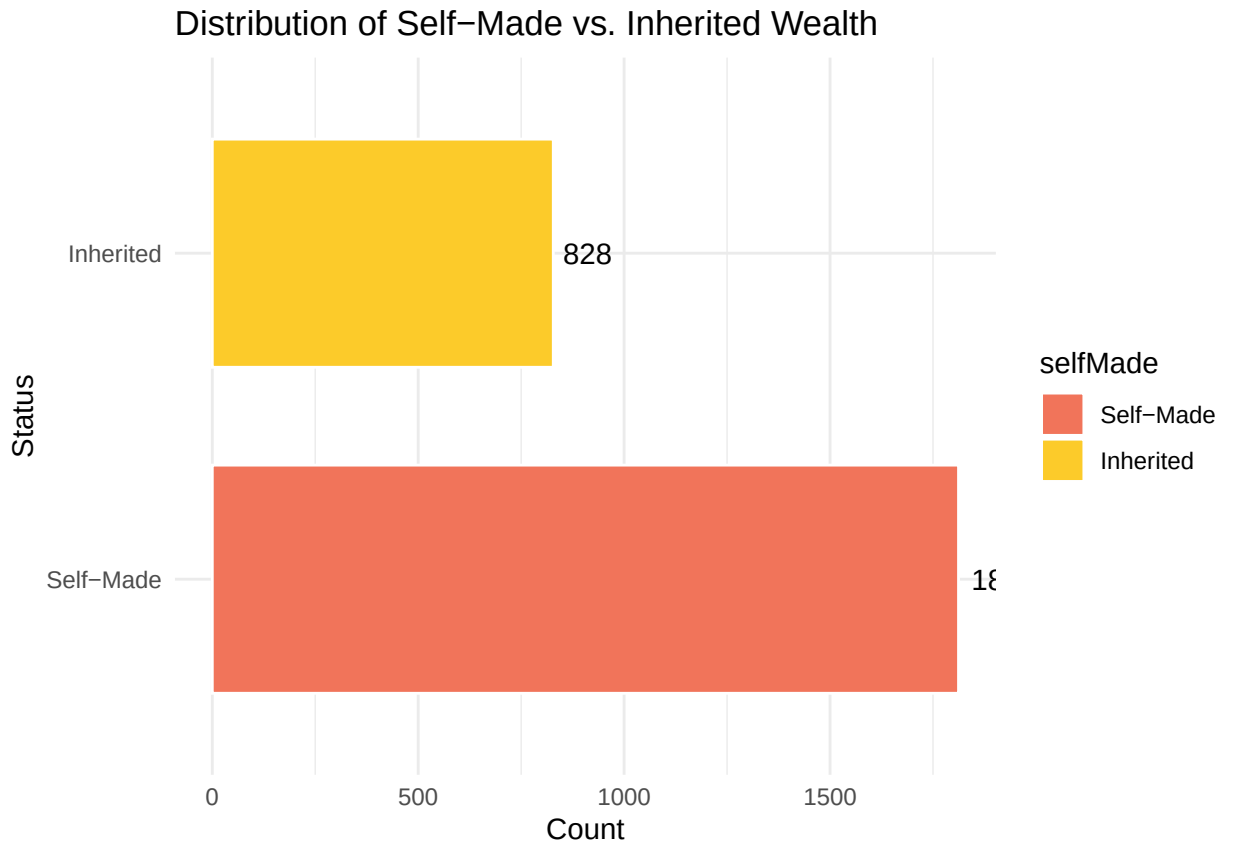
```
library(leaflet)

# Geospatial distribution of billionaires
map <- leaflet(df ) %>%
  addTiles() %>%
  addMarkers(~longitude_country, ~latitude_country, popup = ~paste(personName, "<br>Net Worth: $", final
```

```
## Warning in validateCoords(lng, lat, funcName): Data contains 164 rows with
## either missing or invalid lat/lon values and will be ignored
```

```
# Displaying the map
map
```

Self-made vs. Inherited Wealth

```
total_self_made <- table(df$selfMade)
mapping <- c("Self-Made" = TRUE, "Inherited" = FALSE)
total_self_made_df <- as.data.frame(total_self_made)
names(total_self_made_df) <- c("selfMade", "count")
total_self_made_df$selfMade <- factor(total_self_made_df$selfMade, levels = c(TRUE, FALSE), labels = c(


ggplot(total_self_made_df, aes(x = selfMade, y = count, fill = selfMade)) +
  geom_bar(stat = "identity", width = 0.7, color = "white") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Distribution of Self-Made vs. Inherited Wealth", x = "Status", y = "Count") +
  theme(legend.position = "right") +
  geom_text(aes(label = count), hjust = -0.2) +
  scale_fill_manual(values = c("Self-Made" = "#F1745A", "Inherited" = "#FCCB2A"))
```



Top 5 billionaires in the world

```
top_5_billionaires <- df[order(df$finalWorth, decreasing = TRUE), c("personName", "finalWorth")][1:5, ]
ggplot(top_5_billionaires, aes(x = personName, y = finalWorth)) +
  geom_bar(stat = "identity", fill = "#F1745A") +
  theme_minimal() +
  labs(title = "Top 5 Billionaires and Their Final Worth",
       x = "Billionaire",
       y = "Final Worth (in billions USD)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none")
```