Billionaires Trends Analysis

INFX 502 Semester Project

by

Marjan Pahlevani

ULID: C00566222

Department: Informatics

Fall 2024

Table of content

1. Dataset

1.1 Description

The dataset I chose is called the Billionaires Statistics Dataset and contains essential information about the wealthiest people in the world. This dataset has a large number of records equal to 2,640 and 35 columns(features), which can allow various analyses to be performed. It is a good source of information in the examination of the patterns of wealth and population and the impacts of characteristics such as life expectancy and gross domestic product on wealth.

This dataset collected data from various reliable sources such as reports, banks and other financial institutions, and government reports, among others, through various sources such as Bloomberg, Forbes and World Bank.

I selected this dataset because it includes numerical and categorical features, which provide strong preliminary knowledge about the distribution of patterns. Its applicability in a broad spectrum of analysis makes it suitable for conducting correlation analysis, clustering, and regression linear modeling, as the following section on results will show regarding the insight gained about global wealth distribution.

1.2 Dataset source

I collected the dataset from Kaggle, and it's openly available to the public under the MIT License for research purposes only. The dataset is expected to be updated monthly, but the latest update was 10 months ago.

link of the dataset: https://www.kaggle.com/datasets/endofnight17j03/billionaires-statistics-dataset

1.3 Loading the Dataset

I first loaded the dataset into the R environment for analysis and checked the structure of the dataset to ensure that it was properly imported and to inspect its columns and data types.:

```
> Billionaires <- read.csv("c:/User/Mapa/OneDrive/Billionaires_Statistics_Dataset.csv")
```

I used is.data.frame() function in order to make sure that the data is imported in a proper data frame format in R.

```
> is.data.frame(Billionaires)
[1] TRUE
>
```

I ran the head() function to look at the first six rows of the data.

```
> head(Billionaires)
  rank finalWorth             category               personName age       country   city             source         industries countryOfCitizenship
1    1     211000     Fashion & Retail Bernard Arnault & family  74        France  Paris               LVMH   Fashion & Retail               France
2    2     180000           Automotive                Elon Musk  51 United States Austin      Tesla, SpaceX         Automotive        United States
3    3     114000           Technology               Jeff Bezos  59 United States Medina             Amazon         Technology        United States
4    4     107000           Technology            Larry Ellison  78 United States  Lanai             Oracle         Technology        United States
5    5     106000 Finance & Investments          Warren Buffett  92 United States  Omaha Berkshire Hathaway Finance & Investments        United States
6    6     104000           Technology               Bill Gates  67 United States Medina          Microsoft         Technology        United States
                 organization selfMade status gender       birthDate lastName firstName                title          date      state residenceStateRegion birthYear
1 LVMH Moët Hennessy Louis Vuitton    FALSE      U      M  3/5/1949 0:00  Arnault   Bernard       Chairman and CEO 4/4/2023 5:01                                   1949
2                        Tesla     TRUE      D      M 6/28/1971 0:00     Musk      Elon                    CEO 4/4/2023 5:01      Texas                South      1971
3                       Amazon     TRUE      D      M 1/12/1964 0:00    Bezos      Jeff Chairman and Founder 4/4/2023 5:01 Washington                 West      1964
4                       Oracle     TRUE      U      M 8/17/1944 0:00  Ellison     Larry        CTO and Founder 4/4/2023 5:01     Hawaii                 West      1944
5   Berkshire Hathaway Inc. (Cl A)     TRUE      D      M 8/30/1930 0:00  Buffett    Warren                    CEO 4/4/2023 5:01   Nebraska              Midwest      1930
6  Bill & Melinda Gates Foundation     TRUE      D      M 10/28/1955 0:00    Gates      Bill                 Cochair 4/4/2023 5:01 Washington                 West      1955
  birthMonth birthDay cpi_country cpi_change_country         gdp_country gross_tertiary_education_enrollment gross_primary_education_enrollment_country
1          3        5      110.05                1.1  $2,715,518,274,227                                65.6                                      102.5
2          6       28      117.24                7.5 $21,427,700,000,000                                88.2                                      101.8
3          1       12      117.24                7.5 $21,427,700,000,000                                88.2                                      101.8
4          8       17      117.24                7.5 $21,427,700,000,000                                88.2                                      101.8
5          8       30      117.24                7.5 $21,427,700,000,000                                88.2                                      101.8
6         10       28      117.24                7.5 $21,427,700,000,000                                88.2                                      101.8
  life_expectancy_country tax_revenue_country_country total_tax_rate_country population_country latitude_country longitude_country
1                    82.5                        24.2                   60.7           67059887         46.22764          2.213749
2                    78.5                         9.6                   36.6          328239523         37.09024        -95.712891
3                    78.5                         9.6                   36.6          328239523         37.09024        -95.712891
4                    78.5                         9.6                   36.6          328239523         37.09024        -95.712891
5                    78.5                         9.6                   36.6          328239523         37.09024        -95.712891
6                    78.5                         9.6                   36.6          328239523         37.09024        -95.712891
> 
```

Figure 1 First six rows of the Billionaires dataset

I also used tail() function to look at the bottom six rows.

```
> tail(Billionaires)
     rank finalWorth        category               personName age       country       city                         source        industries countryOfCitizenship organization
2635 2540       1000      Healthcare  Yi Xianzhong & family  63         China  Guangzhou                Pharmaceuticals        Healthcare                China
2636 2540       1000      Healthcare                Yu Rong  51         China   Shanghai                 Health clinics        Healthcare                China
2637 2540       1000 Food & Beverage Richard Yuengling, Jr.  80 United States Pottsville                           Beer Food & Beverage        United States
2638 2540       1000   Manufacturing          Zhang Gongyun  60         China      Gaomi Tyre manufacturing machinery     Manufacturing                China
2639 2540       1000     Real Estate Zhang Guiping & family  71         China    Nanjing                    Real estate       Real Estate                China
2640 2540       1000     Diversified            Inigo Zobel  66   Philippines     Makati                    Diversified       Diversified          Philippines
     selfMade status gender       birthDate  lastName firstName title          date        state residenceStateRegion birthYear birthMonth birthDay cpi_country
2635     TRUE      D      M   5/1/1959 0:00        Yi Xianzhong       4/4/2023 5:01                                    1959          5        1      125.08
2636     TRUE      D      M 12/14/1971 0:00        Yu      Rong       4/4/2023 5:01                                    1971         12       14      125.08
2637    FALSE      E      M  3/10/1943 0:00 Yuengling   Richard       4/4/2023 5:01 Pennsylvania           Northeast      1943          3       10      117.24
2638     TRUE      R      M 12/18/1962 0:00     Zhang   Gongyun       4/4/2023 5:01                                    1962         12       18      125.08
2639     TRUE      D      M  8/21/1951 0:00     Zhang   Guiping       4/4/2023 5:01                                    1951          8       21      125.08
2640    FALSE      R      M  11/1/1956 0:00     Zobel     Inigo       4/4/2023 5:01                                    1956         11        1      129.61
     cpi_change_country         gdp_country gross_tertiary_education_enrollment gross_primary_education_enrollment_country life_expectancy_country
2635                2.9 $19,910,000,000,000                                50.6                                      100.2                    77.0
2636                2.9 $19,910,000,000,000                                50.6                                      100.2                    77.0
2637                7.5 $21,427,700,000,000                                88.2                                      101.8                    78.5
2638                2.9 $19,910,000,000,000                                50.6                                      100.2                    77.0
2639                2.9 $19,910,000,000,000                                50.6                                      100.2                    77.0
2640                2.5    $376,795,508,680                                35.5                                      107.5                    71.1
     tax_revenue_country_country total_tax_rate_country population_country latitude_country longitude_country
2635                         9.4                   59.2         1397715000         35.86166         104.19540
2636                         9.4                   59.2         1397715000         35.86166         104.19540
2637                         9.6                   36.6          328239523         37.09024         -95.71289
2638                         9.4                   59.2         1397715000         35.86166         104.19540
2639                         9.4                   59.2         1397715000         35.86166         104.19540
2640                        14.0                   43.1          108116615         12.87972         121.77402
> 
```

Figure 2 Last six rows of the Billionaires dataset

1.4 Cleaning data

I moved into cleaning the dataset to ensure it was prepared for analysis. Therefore, I performed a feature inspection of missing values, wrong data type, duplication, and outliers. The following actions are the procedures that I took to clean the dataset, as well as the R commands that I perform each of these.

1.4.1 Inspecting the Dataset

First of all, I have started to check the structure of the Billionaires dataset. The majority of the categorical variables were stored as characters (chr), which I needed to convert to factors for proper analysis

```
> Billionaires <- read.csv("Billionaires_Statistics_Dataset.csv")
> str(Billionaires)
'data.frame':	2640 obs. of  35 variables:
 $ rank                                      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ finalWorth                                : int  211000 180000 114000 107000 106000 104000 94500 93000 83400 80700 ...
 $ category                                  : chr  "Fashion & Retail" "Automotive" "Technology" "Technology" ...
 $ personName                                : chr  "Bernard Arnault & family" "Elon Musk" "Jeff Bezos" "Larry Ellison" ...
 $ age                                       : int  74 51 59 78 92 67 81 83 65 67 ...
 $ country                                   : chr  "France" "United States" "United States" "United States" ...
 $ city                                      : chr  "Paris" "Austin" "Medina" "Lanai" ...
 $ source                                    : chr  "LVMH" "Tesla, SpaceX" "Amazon" "Oracle" ...
 $ industries                                : chr  "Fashion & Retail" "Automotive" "Technology" "Technology" ...
 $ countryOfCitizenship                      : chr  "France" "United States" "United States" "United States" ...
 $ organization                              : chr  "LVMH Moët Hennessy Louis Vuitton" "Tesla" "Amazon" "Oracle" ...
 $ selfMade                                  : logi  FALSE TRUE TRUE TRUE TRUE TRUE ...
 $ status                                    : chr  "U" "D" "D" "U" ...
 $ gender                                    : chr  "M" "M" "M" "M" ...
 $ birthDate                                 : chr  "3/5/1949 0:00" "6/28/1971 0:00" "1/12/1964 0:00" "8/17/1944 0:00" ...
 $ lastName                                  : chr  "Arnault" "Musk" "Bezos" "Ellison" ...
 $ firstName                                 : chr  "Bernard" "Elon" "Jeff" "Larry" ...
 $ title                                     : chr  "Chairman and CEO" "CEO" "Chairman and Founder" "CTO and Founder" ...
 $ date                                      : chr  "4/4/2023 5:01" "4/4/2023 5:01" "4/4/2023 5:01" "4/4/2023 5:01" ...
 $ state                                     : chr  "" "Texas" "Washington" "Hawaii" ...
 $ residenceStateRegion                      : chr  "" "South" "West" "West" ...
 $ birthYear                                 : int  1949 1971 1964 1944 1930 1955 1942 1940 1957 1956 ...
 $ birthMonth                                : int  3 6 1 8 8 10 2 1 4 3 ...
 $ birthDay                                  : int  5 28 12 17 30 28 14 28 19 24 ...
 $ cpi_country                               : num  110 117 117 117 117 ...
 $ cpi_change_country                        : num  1.1 7.5 7.5 7.5 7.5 7.5 7.5 3.6 7.7 7.5 ...
 $ gdp_country                               : chr  "$2,715,518,274,227 " "$21,427,700,000,000 " "$21,427,700,000,000 " "$21,427,700,000,000 " ...
 $ gross_tertiary_education_enrollment       : num  65.6 88.2 88.2 88.2 88.2 88.2 88.2 40.2 28.1 88.2 ...
 $ gross_primary_education_enrollment_country: num  102 102 102 102 102 ...
 $ life_expectancy_country                   : num  82.5 78.5 78.5 78.5 78.5 78.5 78.5 75 69.4 78.5 ...
 $ tax_revenue_country_country               : num  24.2 9.6 9.6 9.6 9.6 9.6 9.6 13.1 11.2 9.6 ...
 $ total_tax_rate_country                    : num  60.7 36.6 36.6 36.6 36.6 36.6 36.6 55.1 49.7 36.6 ...
 $ population_country                        : int  67059887 328239523 328239523 328239523 328239523 328239523 126014024 1366417754 328239523 ...
 $ latitude_country                          : num  46.2 37.1 37.1 37.1 37.1 ...
 $ longitude_country                         : num  2.21 -95.71 -95.71 -95.71 -95.71 ...
```

Figure 3 Initial structure

I used is.na() to check for missing values that helped me identify which columns needed cleaning and which had missing values.

```
> colSums(is.na(Billionaires))
                              rank                                        finalWorth                                          category
                                 0                                                 0                                                 0
                        personName                                               age                                           country
                                 0                                                65                                                 0
                              city                                            source                                        industries
                                 0                                                 0                                                 0
               countryOfCitizenship                                      organization                                          selfMade
                                 0                                                 0                                                 0
                            status                                            gender                                         birthDate
                                 0                                                 0                                                 0
                          lastName                                         firstName                                             title
                                 0                                                 0                                                 0
                              date                                             state                               residenceStateRegion
                                 0                                                 0                                                 0
                         birthYear                                        birthMonth                                          birthDay
                                76                                                76                                                76
                       cpi_country                                cpi_change_country                                       gdp_country
                               184                                               184                                                 0
  gross_tertiary_education_enrollment gross_primary_education_enrollment_country                           life_expectancy_country
                               182                                               181                                               182
          tax_revenue_country_country                             total_tax_rate_country                                population_country
                               183                                               182                                               164
                  latitude_country                                 longitude_country
                               164                                               164
> |
```

Figure 4 Missing values

The Billionaires dataset has several columns that have missing values according to the colSums(is.na(Billionaires)) output above. In order to deal with the missing values, I applied techniques based on the type of data used in the respective columns. For numerical variables, “age”, “birthyear”, and “population_country”, I replaced missing values with the median of the whole records. Median imputation is less sensitive to outliers and well suits the case when a median of the missing values is a reasonable estimate.

For categorical variables, such as category and country, I replaced missing values with the most common value (mode).

I decided to remove rows with missing values for columns that contain more than 30% of missing values like “gross_primary_education_enrollment_country”, “gross_tertiary_education_enrollment” and “cpi_country”.

```
> columnscheck <- c("cpi_country", "cpi_change_country",
+                   "gross_tertiary_education_enrollment",
+                   "gross_primary_education_enrollment_country",
+                   "tax_revenue_country_country",
+                   "total_tax_rate_country")
> Billionaires <- Billionaires[complete.cases(Billionaires[, columnscheck]), ]
> # Replace missing values in numerical columns with the median
> Billionaires$age[is.na(Billionaires$age)] <- median(Billionaires$age, na.rm = TRUE)
> Billionaires$birthYear[is.na(Billionaires$birthYear)] <- median(Billionaires$birthYear, na.rm = TRUE)
> Billionaires$population_country[is.na(Billionaires$population_country)] <- median(Billionaires$population_country, na.rm = TRUE)
> Billionaires$latitude_country[is.na(Billionaires$latitude_country)] <- median(Billionaires$latitude_country, na.rm = TRUE)
> Billionaires$longitude_country[is.na(Billionaires$longitude_country)] <- median(Billionaires$longitude_country, na.rm = TRUE)
> # Replace missing values in categorical columns with the mode
> Billionaires$category[is.na(Billionaires$category)] <- names(sort(table(Billionaires$category), decreasing = TRUE))[1]
> Billionaires$country[is.na(Billionaires$country)] <- names(sort(table(Billionaires$country), decreasing = TRUE))[1]
>
```

I verified that all the gaps in the data were filled by re-running the colSums(is.na(Billionaires)) function.

```
> colSums(is.na(Billionaires))
                                rank                              finalWorth                                category
                                   0                                       0                                       0
                          personName                                     age                                 country
                                   0                                       0                                       0
                                city                                  source                              industries
                                   0                                       0                                       0
                  countryOfCitizenship                            organization                                selfMade
                                   0                                       0                                       0
                              status                                  gender                               birthDate
                                   0                                       0                                       0
                            lastName                               firstName                                   title
                                   0                                       0                                       0
                                date                                   state                     residenceStateRegion
                                   0                                       0                                       0
                           birthYear                              birthMonth                                birthDay
                                   0                                       0                                       0
                         cpi_country                      cpi_change_country                             gdp_country
                                   0                                       0                                       0
gross_tertiary_education_enrollment gross_primary_education_enrollment_country                 life_expectancy_country
                                   0                                       0                                       0
           tax_revenue_country_country                      total_tax_rate_country                      population_country
                                   0                                       0                                       0
                    latitude_country                       longitude_country
                                   0                                       0
> |
```

Figure 5 Fixed Missing values

1.4.2 Converting chr Variables to Factors

Many categorical variables, such as category, gender, and country, were stored as character data. I converted those variables to factors because this resulted in better memory efficiency and compatibility with statistical tools in R.

```
> Billionaires$category <- as.factor(Billionaires$category)
> is.factor(Billionaires$category)
[1] TRUE
> Billionaires$country <- as.factor(Billionaires$country)
> is.factor(Billionaires$country)
[1] TRUE
> Billionaires$city <- as.factor(Billionaires$city)
> is.factor(Billionaires$city)
[1] TRUE
> Billionaires$source <- as.factor(Billionaires$source)
> is.factor(Billionaires$source)
[1] TRUE
> Billionaires$industries <- as.factor(Billionaires$industries)
> is.factor(Billionaires$industries)
[1] TRUE
> Billionaires$countryOfCitizenship <- as.factor(Billionaires$countryOfCitizenship)
> is.factor(Billionaires$countryOfCitizenship)
[1] TRUE
> Billionaires$organization <- as.factor(Billionaires$organization)
> is.factor(Billionaires$organization)
[1] TRUE
> Billionaires$selfMade <- as.factor(Billionaires$selfMade)
> is.factor(Billionaires$selfMade)
[1] TRUE
> Billionaires$status <- as.factor(Billionaires$status)
> is.factor(Billionaires$status)
[1] TRUE
> Billionaires$gender <- as.factor(Billionaires$gender)
> Billionaires$state <- as.factor(Billionaires$state)
> is.factor(Billionaires$state)
[1] TRUE
> Billionaires$residenceStateRegion <- as.factor(Billionaires$residenceStateRegion)
> is.factor(Billionaires$residenceStateRegion)
[1] TRUE
> Billionaires$title <- as.factor(Billionaires$title)
> is.factor(Billionaires$title)
[1] TRUE
```

```
> str(Billionaires)
'data.frame':	2456 obs. of  35 variables:
 $ rank                                      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ finalWorth                                : int  211000 180000 114000 107000 106000 104000 94500 93000 83400 80700 ...
 $ category                                  : Factor w/ 18 levels "Automotive","Construction & Engineering",..: 5 1 17 17 6 17 12 18 3 17 ...
 $ personName                                : chr  "Bernard Arnault & family" "Elon Musk" "Jeff Bezos" "Larry Ellison" ...
 $ age                                       : int  74 51 59 78 92 67 81 83 65 67 ...
 $ country                                   : Factor w/ 64 levels "Algeria","Argentina",..: 19 62 62 62 62 62 62 34 24 62 ...
 $ city                                      : Factor w/ 720 levels "","A Coruña",..: 483 28 394 325 466 394 443 402 423 255 ...
 $ source                                    : Factor w/ 877 levels "3D printing",..: 462 796 29 579 81 504 98 787 223 504 ...
 $ industries                                : Factor w/ 18 levels "Automotive","Construction & Engineering",..: 5 1 17 17 6 17 12 18 3 17 ...
 $ countryOfCitizenship                      : Factor w/ 71 levels "Algeria","Argentina",..: 21 68 68 68 68 68 68 38 28 68 ...
 $ organization                              : Factor w/ 290 levels "","ABC Supply",..: 159 255 8 189 29 32 35 9 208 155 ...
 $ selfMade                                  : Factor w/ 2 levels "FALSE","TRUE": 1 2 2 2 2 2 2 2 1 2 ...
 $ status                                    : Factor w/ 6 levels "D","E","N","R",..: 6 1 1 6 1 1 6 6 1 1 ...
 $ gender                                    : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ birthDate                                 : chr  "3/5/1949 0:00" "6/28/1971 0:00" "1/12/1964 0:00" "8/17/1944 0:00" ...
 $ lastName                                  : chr  "Arnault" "Musk" "Bezos" "Ellison" ...
 $ firstName                                 : chr  "Bernard" "Elon" "Jeff" "Larry" ...
 $ title                                     : Factor w/ 97 levels "","Advisor","Athlete",..: 18 5 21 52 5 36 5 77 68 85 ...
 $ date                                      : chr  "4/4/2023 5:01" "4/4/2023 5:01" "4/4/2023 5:01" "4/4/2023 5:01" ...
 $ state                                     : Factor w/ 46 levels "","Alabama","Arizona",..: 1 40 44 10 26 44 30 1 1 44 ...
 $ residenceStateRegion                      : Factor w/ 6 levels "","Midwest","Northeast",..: 1 4 6 6 2 6 3 1 1 6 ...
 $ birthYear                                 : int  1949 1971 1964 1944 1930 1955 1942 1940 1957 1956 ...
 $ birthMonth                                : int  3 6 1 8 8 10 2 1 4 3 ...
 $ birthDay                                  : int  5 28 12 17 30 28 14 28 19 24 ...
 $ cpi_country                               : num  110 117 117 117 117 ...
 $ cpi_change_country                        : num  1.1 7.5 7.5 7.5 7.5 7.5 7.5 3.6 7.7 7.5 ...
 $ gdp_country                               : chr  "$2,715,518,274,227 " "$21,427,700,000,000 " "$21,427,700,000,000 " "$21,427,700,000,000 " ...
 $ gross_tertiary_education_enrollment       : num  65.6 88.2 88.2 88.2 88.2 88.2 88.2 40.2 28.1 88.2 ...
 $ gross_primary_education_enrollment_country: num  102 102 102 102 102 ...
 $ life_expectancy_country                   : num  82.5 78.5 78.5 78.5 78.5 78.5 78.5 75 69.4 78.5 ...
 $ tax_revenue_country_country               : num  24.2 9.6 9.6 9.6 9.6 9.6 9.6 13.1 11.2 9.6 ...
 $ total_tax_rate_country                    : num  60.7 36.6 36.6 36.6 36.6 36.6 36.6 55.1 49.7 36.6 ...
 $ population_country                        : num  6.71e+07 3.28e+08 3.28e+08 3.28e+08 3.28e+08 ...
 $ latitude_country                          : num  46.2 37.1 37.1 37.1 37.1 ...
 $ longitude_country                         : num  2.21 -95.71 -95.71 -95.71 -95.71 ...
> 
```

Figure 6 Converted factor variables

1.4.3 Converting character Variables to numeric

It seems the gdp_country column is stored as a character (chr) instead of numeric because it likely includes non-numeric characters such as commas and Dollar signs (e.g., "$21,427,740,000,000"). To fix this, I cleaned the column by removing such formatting and converting it to numeric. Here's how I handled this:

```
> Billionaires$gdp_country <- gsub("[^0-9.-]", "", Billionaires$gdp_country)
> Billionaires$gdp_country <- as.numeric(Billionaires$gdp_country)
> is.numeric(Billionaires$gdp_country)
[1] TRUE
> 
```

I also converted the date variable format to date from char.

```
> Billionaires$date <- as.Date(Billionaires$date, format = "%m/%d/%Y")
```

```
> str(Billionaires)
'data.frame':	2456 obs. of  35 variables:
 $ rank                                    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ finalWorth                              : int  211000 180000 114000 107000 106000 104000 94500 93000 83400 80700 ...
 $ category                                : Factor w/ 18 levels "Automotive","Construction & Engineering",..: 5 1 17 17 6 17 12 18 3 17 ...
 $ personName                              : chr  "Bernard Arnault & family" "Elon Musk" "Jeff Bezos" "Larry Ellison" ...
 $ age                                     : int  74 51 59 78 92 67 81 83 65 67 ...
 $ country                                 : Factor w/ 64 levels "Algeria","Argentina",..: 19 62 62 62 62 62 62 34 24 62 ...
 $ city                                    : Factor w/ 720 levels "","A Coruña",..: 483 28 394 325 466 394 443 402 423 255 ...
 $ source                                  : Factor w/ 877 levels "3D printing",..: 462 796 29 579 81 504 98 787 223 504 ...
 $ industries                              : Factor w/ 18 levels "Automotive","Construction & Engineering",..: 5 1 17 17 6 17 12 18 3 17 ...
 $ countryOfCitizenship                    : Factor w/ 71 levels "Algeria","Argentina",..: 21 68 68 68 68 68 68 38 28 68 ...
 $ organization                            : Factor w/ 290 levels "","ABC Supply",..: 159 255 8 189 29 32 35 9 208 155 ...
 $ selfMade                                : Factor w/ 2 levels "FALSE","TRUE": 1 2 2 2 2 2 2 2 1 2 ...
 $ status                                  : Factor w/ 6 levels "D","E","N","R",..: 6 1 1 6 1 1 6 6 1 1 ...
 $ gender                                  : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ birthDate                               : chr  "3/5/1949 0:00" "6/28/1971 0:00" "1/12/1964 0:00" "8/17/1944 0:00" ...
 $ lastName                                : chr  "Arnault" "Musk" "Bezos" "Ellison" ...
 $ firstName                               : chr  "Bernard" "Elon" "Jeff" "Larry" ...
 $ title                                   : Factor w/ 97 levels "","Advisor","Athlete",..: 18 5 21 52 5 36 5 77 68 85 ...
 $ date                                    : Date, format: "2023-04-04" "2023-04-04" "2023-04-04" "2023-04-04" ...
 $ state                                   : Factor w/ 46 levels "","Alabama","Arizona",..: 1 40 44 10 26 44 30 1 1 44 ...
 $ residenceStateRegion                    : Factor w/ 6 levels "","Midwest","Northeast",..: 1 4 6 6 2 6 3 1 1 6 ...
 $ birthYear                               : int  1949 1971 1964 1944 1930 1955 1942 1940 1957 1956 ...
 $ birthMonth                              : int  3 6 1 8 8 10 2 1 4 3 ...
 $ birthDay                                : int  5 28 12 17 30 28 14 28 19 24 ...
 $ cpi_country                             : num  110 117 117 117 117 ...
 $ cpi_change_country                      : num  1.1 7.5 7.5 7.5 7.5 7.5 7.5 3.6 7.7 7.5 ...
 $ gdp_country                             : num  2.72e+12 2.14e+13 2.14e+13 2.14e+13 2.14e+13 ...
 $ gross_tertiary_education_enrollment     : num  65.6 88.2 88.2 88.2 88.2 88.2 88.2 40.2 28.1 88.2 ...
 $ gross_primary_education_enrollment_country: num  102 102 102 102 102 ...
 $ life_expectancy_country                 : num  82.5 78.5 78.5 78.5 78.5 78.5 78.5 75 69.4 78.5 ...
 $ tax_revenue_country_country             : num  24.2 9.6 9.6 9.6 9.6 9.6 9.6 13.1 11.2 9.6 ...
 $ total_tax_rate_country                  : num  60.7 36.6 36.6 36.6 36.6 36.6 36.6 55.1 49.7 36.6 ...
 $ population_country                      : num  6.71e+07 3.28e+08 3.28e+08 3.28e+08 3.28e+08 ...
 $ latitude_country                        : num  46.2 37.1 37.1 37.1 37.1 ...
 $ longitude_country                       : num  2.21 -95.71 -95.71 -95.71 -95.71 ...
> |
```

Figure 7 Fixed structure of Billionaires dataset

After cleaning, the dataset has 2456 records and 35 variables in the correct format. Green, blue, and yellow color shows converted variables.

1.5 Variable Table

The following table lists and describes each of the 35 variables (columns) in the dataset:

Table 1 Variable Table

Variable name	Description	Mode
Rank	Rank of a person in terms of wealth.	Integer
finalWorth	Net worth of the individual.	Numeric
category	Classification (automotive, engineering, etc.).	Factor with 18 level
personName	Full name of the individual.	Character
age	Age of the individual.	Numeric
country	Country of residence of the individual.	Factor with 64 levels
city	City of residence of the individual.	Factor
source	Source of wealth.	Factor
industries	Industries associated with the individual.	Factor with 18 levels
countryOfCitizenship	Country of citizenship.	Factor with 71 levels
organization	Organization associated with the individual.	Factor
selfMade	Indicates whether wealth is self-made or inherited.	Factor with 2 levels
status	Professional status (e.g., CEO, Founder).	Factor with 6 levels
gender	Gender of the individual.	Factor with 2 levels
birthDate	Date of birth of the individual.	Integer
lastName	Last name of the individual.	Character
firstName	First name of the individual.	Character
title	Title of individual (e.g., advisor, athlete).	Factor
date	Date of data entry.	Date
state	State or region of residence of the individual.	Factor with 46 level
residenceStateRegion	Detailed state/region of residence.	Factor with 6 level
birthYear	Year of birth of the individual.	Integer
birthMonth	Month of birth of the individual.	Integer
birthDay	Day of birth of the individual.	Integer
cpi_country	Consumer Price Index (CPI) for the individual's country.	Numeric
cpi_change_country	Change in CPI for the individual's country.	Numeric
gdp_country	Gross Domestic Product (GDP) of the individual country.	Numeric
gross_tertiary_education_enrollment	Tertiary education enrollment rate in the country.	Numeric
gross_primary_education_enrollment_country	Primary education enrollment rate in country.	Numeric
life_expectancy_country	Life expectancy in the individual's country.	Numeric
tax_revenue_country	Tax revenue in the individual's country.	Numeric
total_tax_rate_country	Total tax rate in the individual's country.	Numeric
population_country	Population of the individual's country.	Numeric
latitude_country	Latitude of the individual's country.	Numeric
longitude_country	Longitude of the individual's country.	Numeric

1.6 Expectations

In this project, I expect to identify useful patterns or trends concerning various aspects that influence the wealth and demography of billionaires based on the Billionaires Statistics Dataset. Given the numerical and categorical data, I will seek to describe how these characteristics work and what the results will be on global wealth distribution.

I anticipate the correlation analysis to have high values for numerical variables such as "age", "finalWorth" and "gdp_country". For instance, I expect to know if the billionaires in their fifties have more net worth than billionaires in their thirties or if the young billionaires major in technological firms. I also predict that countries with higher GDPs will have more billionaires with higher net worth.

Trend analysis by geographical location will probably reveal that countries that have the most billionaires are the most industrially developed countries, like the United States and China. I expect that industries like technology and finance will be in these regions because these are the most developed economic regions. I expect that such origins as technology and finance will be more typical for self-made billionaires, while such spheres as inheritance will be more characteristic of fashion and retail. Bar plots and contingency tables will be used to analyze the distribution of sources of wealth by industry and location, respectively, while the Chai-square test will be used to establish whether two categorical variables, Self-Made and industries, bear any relationship or not.

I also anticipate finding some outliers in the data set, such as very rich billionaires or young people with large fortunes. These are the outliers that will be detected by constructing boxplots and histograms. Likewise, I expect the distributions of variables such as final worth and age to be positively skewed, given that wealth is more or less bounded at the high end.

I expect to use clustering techniques to develop a list of billionaires using similar features. I expect to find groups of billionaires according to the industry and their asset value, as well as according to country efficiency indicators such as GDP and population. These clusters will raise patterns of wealth accumulation and explain how similar characteristics bring people together.

2. Analysis

In this section, I examined the Billionaires Statistics Dataset with the help of basic methods of visualization and statistics. The analysis concerns searching for relationships between variables, identifying patterns, and knowledge about distributions of wealth and factors influencing it.

2.1 Scatterplots and Correlation Matrices (numeric variables)

To establish the relation between the numeric variables, I used correlation coefficients and displayed them by correlation matrix and scatterplot. The age variable, final Worth, and gdp_country were investigated in the current analysis.

```
> Bil_num <- data.frame(Billionaires[ ,c("finalWorth", "age", "population_country", "gdp_country")])
> cor(Bil_num)
                    finalWorth         age population_country gdp_country
finalWorth          1.00000000  0.06227230        -0.05356773  0.03746733
age                 0.06227230  1.00000000        -0.16539155 -0.06718313
population_country -0.05356773 -0.16539155         1.00000000  0.44672519
gdp_country         0.03746733 -0.06718313         0.44672519  1.00000000
```

Figure 8 Correlation Matrices

As seen in the correlation matrix most of these variables have either very low or no correlation at all. For example, the association between finalWorth and age is equal to 0.062, which means that their wealth practically does not depend on their age. Equally, finalWorth does not have any correlation with one country's GDP as it is correlated at 0.037. But a moderate positive relationship of 0.447 between population_country and gdp_country means that those countries with big population have high GDP.

```
> plot(Bil_num, main = "Scatterplot Matrix of Numerical Variables", pch = 19, col = "Blue")
>
```

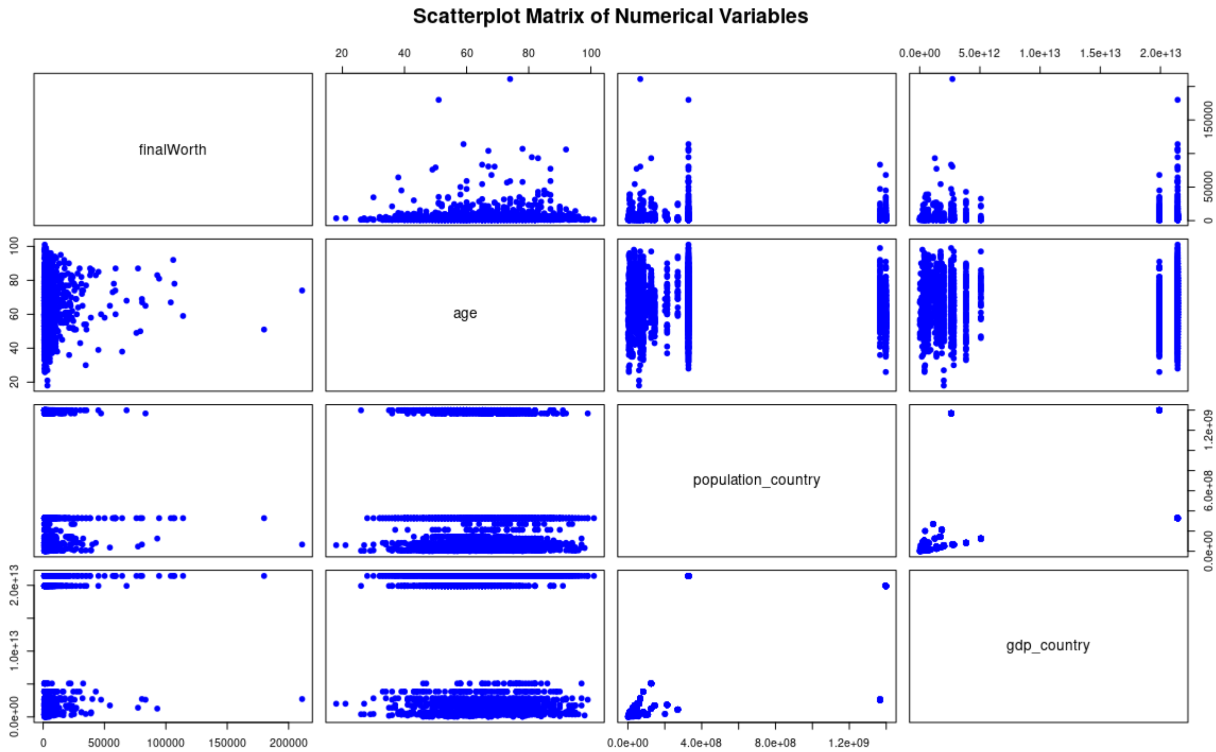

Figure 9 Scatterplot Matrix of Numeric Variables

The scatterplot of finalWorth and age and the scatterplot of finalWorth and population_country do not indicate any trend. However, as for the variables of the plot of population_country and gdp_country, there is a visible tendency of increase that corresponds to their moderate positive relation, which means that the higher the size of the population, the higher the GDP. These scatterplots clearly show the separation of most of the variables and the higher correlation between population and economic output. It is also important to note that the above results match the correlation matrix to give an overview of the data fields.

2.1.2 Correlation of Education Enrollment and Final Worth

```
> correlation <- cor( Billionaires$gross_tertiary_education_enrollment,
+                     Billionaires$finalWorth, use = "complete.obs" )
> print(paste("Correlation coefficient:", correlation))
[1] "Correlation coefficient: 0.0677547502004039"
> 
```

In order to analyze the correlation between education enrollment and billionaire wealth, I calculated the Pearson coefficient. This makes the findings show that there is a very low but positive relationship between tertiary education enrollment and a billionaire's final worth, with a correlation coefficient of 0.0678 and means that the correlation is positive in nature, but the intensity of the correlation is rather weak. In functional terms, it means that even when a country has high enrollment in higher education, the flow to the billionaires' wealth is as negligible. However, other factors might contribute more to the determination of the wealth levels. This scatterplot shows the correlation between gross tertiary education enrollment and final worth. The greater part of the dots focuses on the lower values of both variables, which means that the majority of people originated from countries with less gross tertiary education enrollment and, thus, possess less final worth.

```
> ggplot(Billionaires, aes(x = gross_tertiary_education_enrollment, y = finalWorth)) +
+   geom_point()
> |
```

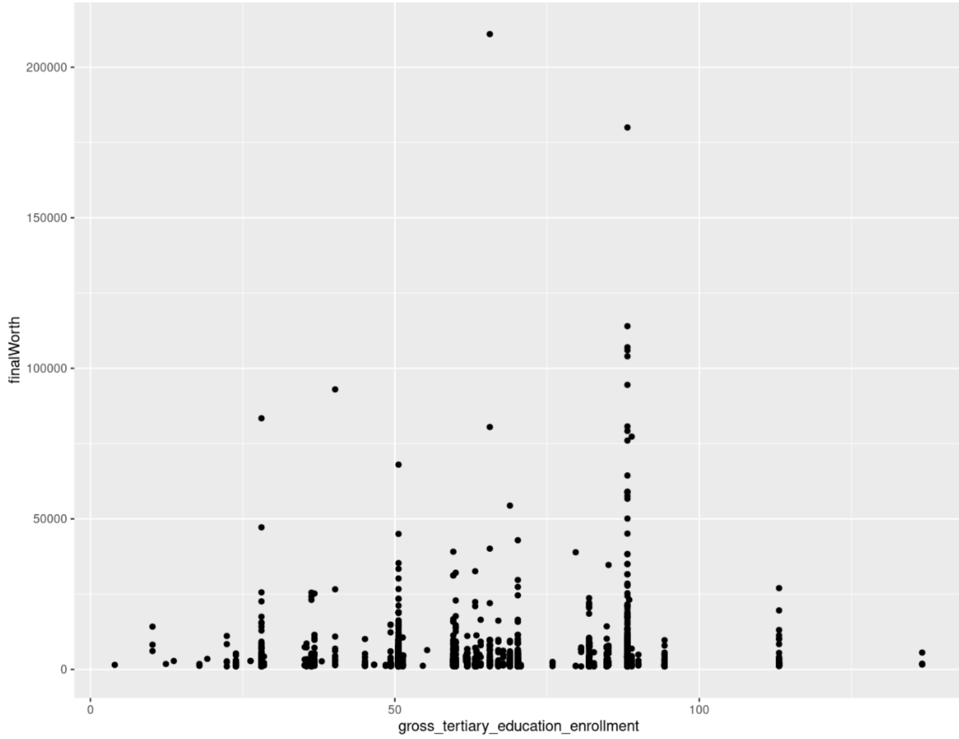

Figure 10 The correlation between education enrollment and final wealth

This scatterplot shows the correlation between gross tertiary education enrollment and final worth. The greater part of the dots focuses on the lower values of both variables, which means that the majority of people originated from countries with less gross tertiary education enrollment and, thus, possess less final worth.

2.1.3 Scatterplot of Age Vs Net Worth

Below is the scatterplot of Age vs Net Worth, which shows the relationship between the age and final worth of billionaires. The plot shows a wide distribution with no clear linear trend, and it means that age alone is not a strong predictor of net worth. There are some outliers at very high net worth values, particularly for the age group between 70 and 100 years old. Across all age groups, the majority of data points bunch at a lower net worth level and indicates that factors other than age can influence the net worth of a billionaire.

```
> library(ggplot2)
> ggplot(Billionaires, aes(x = age, y = finalWorth)) +
+   geom_point(alpha = 0.6) +
+   labs(title = "Scatterplot of Age vs Net Worth", x = "Age", y = "Net Worth (in billions)")
> 
```

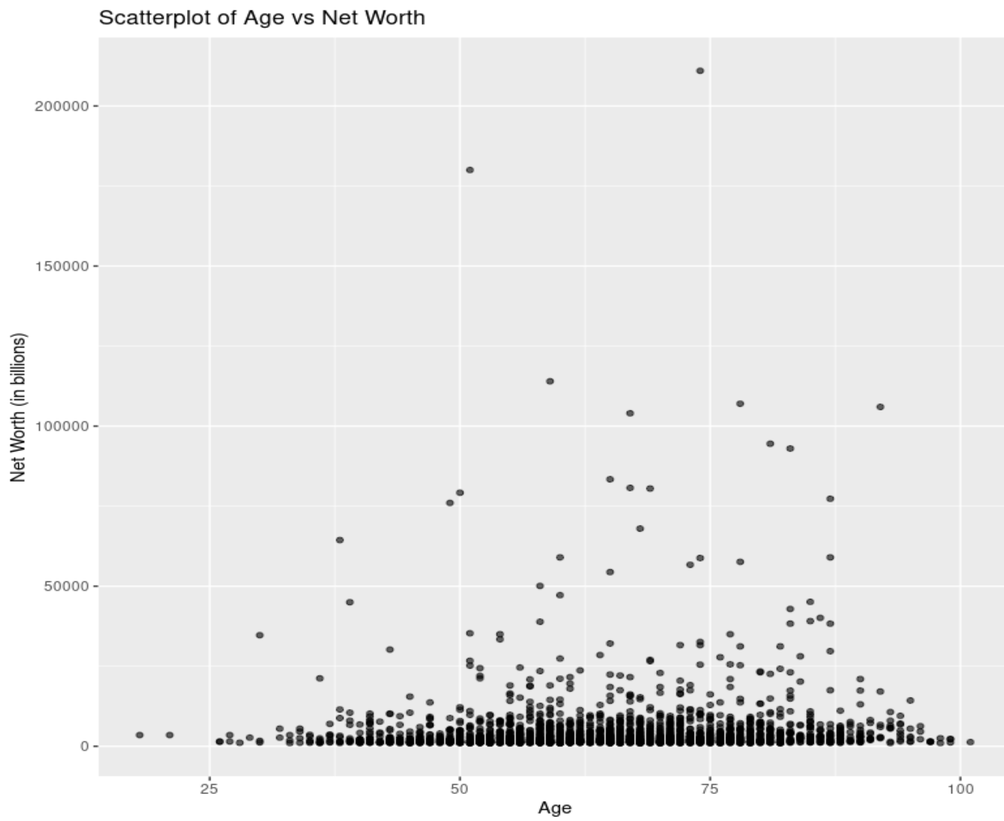

Figure 11 Scatterplot of Age Vs Net Worth

2.2 Contingency Tables and Heatmaps (categorical Variable)

In this step, I analyzed the relationships between categorical variables. I used contingency tables and visualized these relationships with a heatmap. I explored how industries and self-made are related.

```
> table_industries_selfMade <- table(Billionaires$industries, Billionaires$selfMade)
> print(table_industries_selfMade)

                           FALSE TRUE
  Automotive                  33   37
  Construction & Engineering  19   23
  Diversified                101   78
  Energy                      25   71
  Fashion & Retail           101  149
  Finance & Investments       78  265
  Food & Beverage             95  105
  Gambling & Casinos           4   18
  Healthcare                  52  143
  Logistics                    6   27
  Manufacturing               80  216
  Media & Entertainment       24   62
  Metals & Mining             19   52
  Real Estate                 46  114
  Service                     19   29
  Sports                      14   24
  Technology                  22  277
  Telecom                      4   24
>
```

As it is shown in the table, some industries are just hot spots for self-made billionaires. In Technology, there are 277 self-made billionaires versus only 22 who inherited the money that is a huge difference and really puts into perspective how innovation can lead to immense personal fortunes. Finance & Investments tells a similar story and shows that savvy financial moves can build wealth from the ground up. Healthcare also shows a strong trend toward self-made wealth, with 143 self-made billionaires versus 52 inherited ones, while Energy has 71 versus 25. These patterns show how different industries offer varying opportunities for building wealth, whether you're striking out on your own or building on a family legacy.

```
> library(ggplot2)
> Ct <- table(Billionaires$industries, Billionaires$selfMade)
> heatmap_data <- as.data.frame(Ct)
> colnames(heatmap_data) <- c("Industry", "SelfMade", "Count")
> ggplot(data = heatmap_data, aes(x = SelfMade, y = Industry, fill = Count)) +
+   geom_tile()
>
```

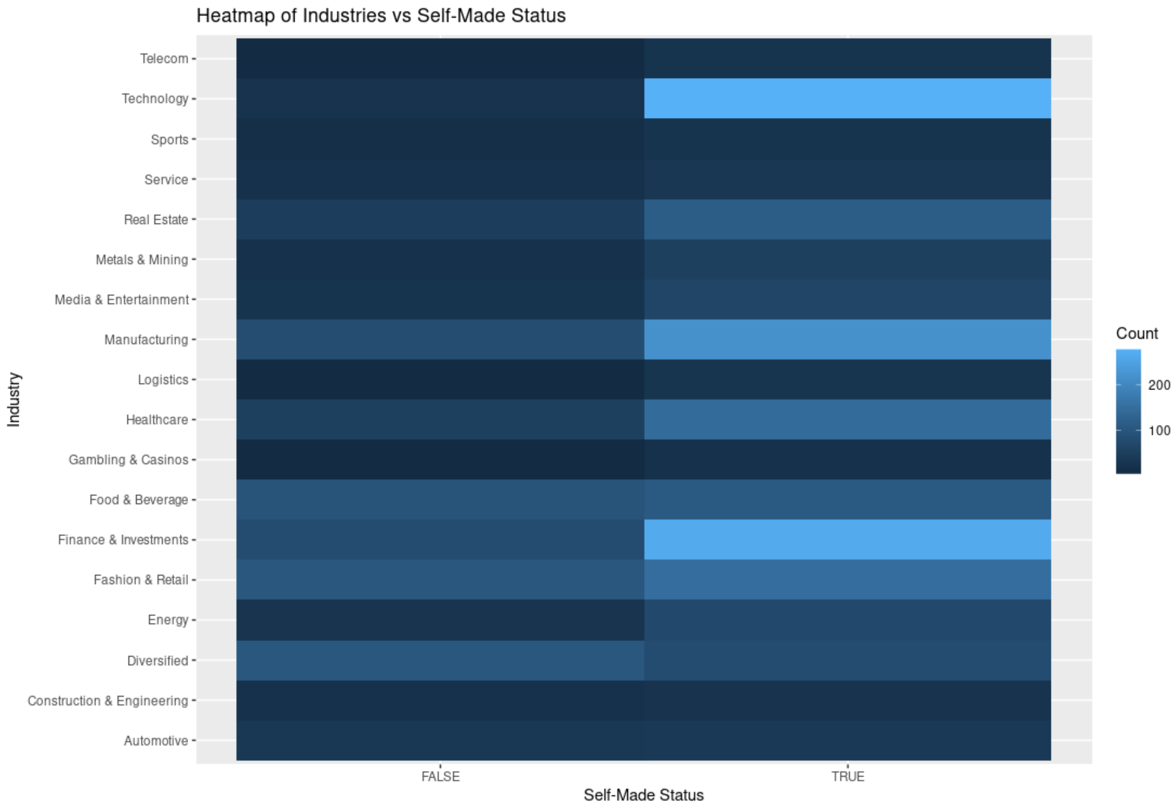

Figure 12 Heatmap of Industries Vs Self-Made Status

The heatmap depicts industries and the self-made status of billionaires. Every cell reflects the number of billionaires belonging to that particular Industry and self-made status (TRUE for self-made, FALSE for inherited wealth). The darker blue represents the lower frequency, and the lighter blue represents the higher number. Technology, Finance & Investments have the highest proportion of self-made billionaires (TRUE). Fashion & Retail and Healthcare, along with Manufacturing, also own a good number of people who are self-made. Real estate has a relatively small number of people who made it on their own.

2.3 Pie Charts

To better understand the global distribution of the key categorical variables, I developed pie charts that shed light on the industries, the self-made variable, and the gender. These charts gave a broad view of how the data is spread in the different categories. The pie chart on industries showed that the majority of billionaires are in the Finance & Investment, Technology, Fashion & Retail, and Manufacturing industries, and all these four industries have large sample populations.

```
> pie(table(Billionaires$industries), main = "Distribution of Industries")
>
```

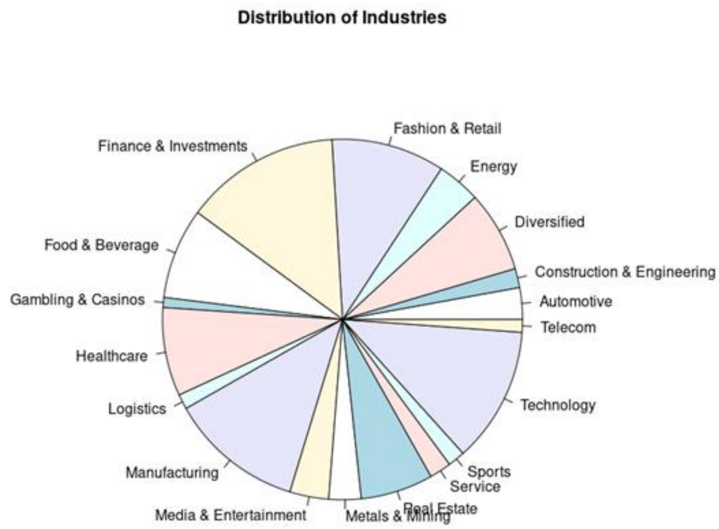

Figure 13 Distribution of Industries

Similarly, the pie chart for self-made highlighted that the majority of billionaires' money was earned self-made.

```
> pie(table(Billionaires$selfMade), main = "Self-Made vs Inherited")
```

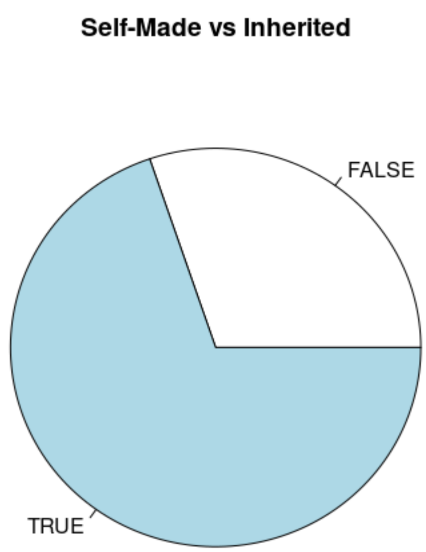

Figure 14 Self-made Pie chart

The pie chart for gender revealed that the majority of billionaires are male.

```
> pie(table(Billionaires$gender), main = "Gender Distribution")
```

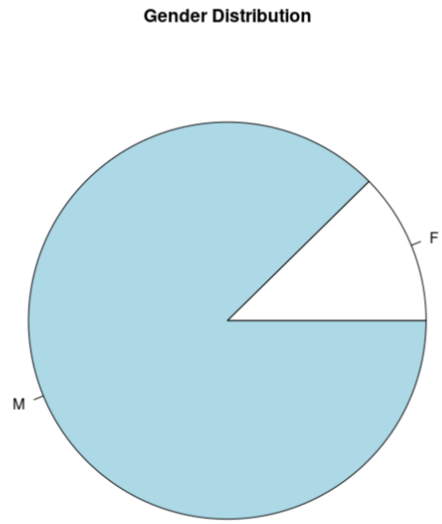

Figure 15 Gender Distribution

2.4 Histograms and density plots (Numeric variables)

With the aim of checking the distribution of the numeric variables in the dataset, I used the histograms and the density plots. The histogram and Density plot helped in learning how the different quantitative variables, such as net worth, age, GDP, or population, are spread, especially when they are grouped by other qualitative variables.

2.4.1 Histograms and density plots

- Final Worth

```
> ggplot(data = Billionaires, aes(x = finalWorth)) +
+   geom_histogram(aes(y = after_stat(density)), bins = num_bins, color = "black", fill = "skyblue", alpha = 0.6) +
+   geom_density(color = "darkgreen", linewidth = 1) +
+   scale_x_log10() +
+   labs(
+     title = "Distribution of Final Worth  with Density",
+     x = "Log of Final Worth (in billions)",
+     y = "Density"
+   )
> |
```

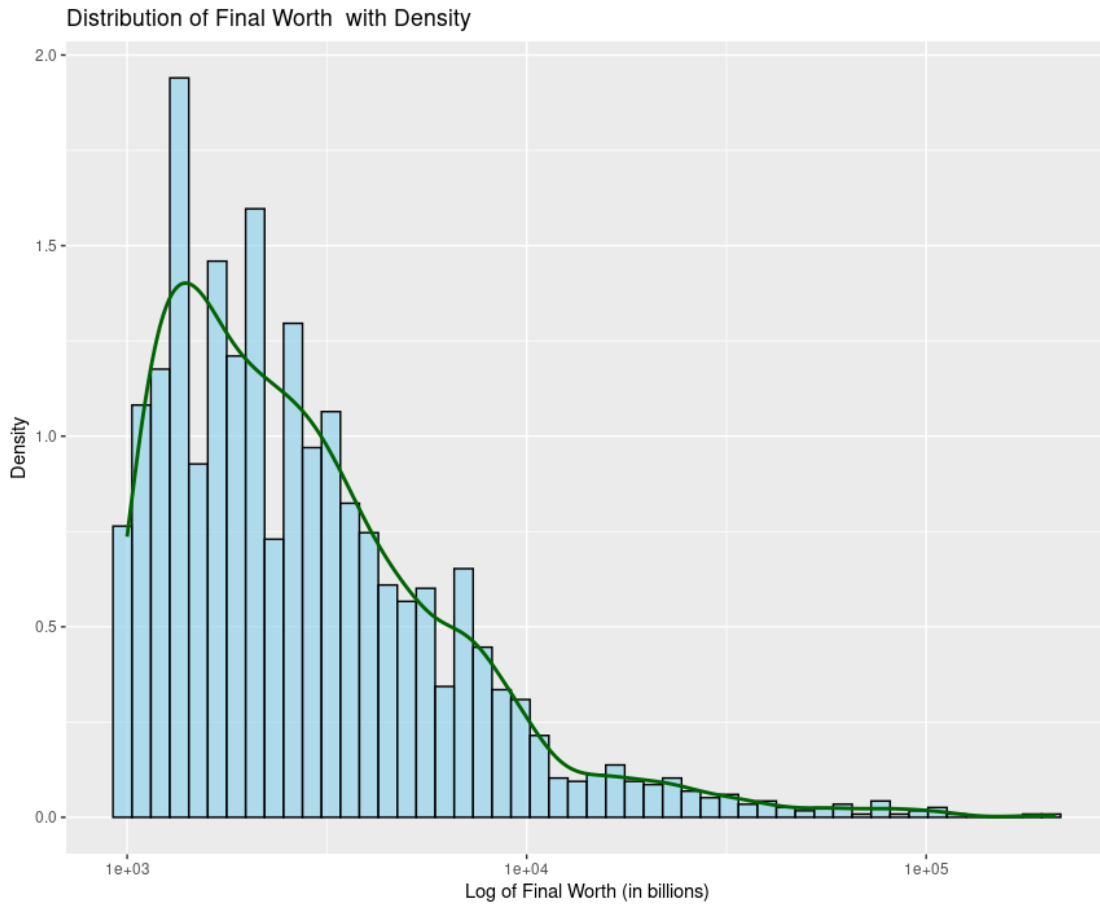

Figure 16 The histogram and density of Final Worth

I adjusted this visualization by applying log scale to the x-axis, which served as the more sensible measure of the billionaire wealth. This transformation reduced the variation in the net worth values and enhanced the patterns within the data. The histogram now clearly indicates that most of the billionaires reside at the beginnings of the shaded area of the graph, and the density plot is almost identical to it. The representational scale is logarithmic and reduces the effect of some super-rich individuals with extreme value for net worth and portfolios. As we have seen from the above visualization, it refines the point that wealth is skewed with most people falling in the lower wealth categories; thus, including it in the analysis is beneficial.

- Age

The histogram of age distribution is nearly normal, and most billionaires are in their 60s and 70s. The result shows that the process of building up wealth takes years, and more often, people garner their billionaire status in their later years. Also, density plot is align with it.

```
> ggplot(data = Billionaires, aes(x = age, y = after_stat(density))) +
+   geom_histogram(bins = round(sqrt(nrow(Billionaires))), fill = "skyblue", color = "black", alpha = 0.6) +
+   geom_density(color = "darkgreen", linewidth = 1) +
+   labs(title = "Histogram of Age with Density Overlay", x = "Age", y = "Density")
> |
```

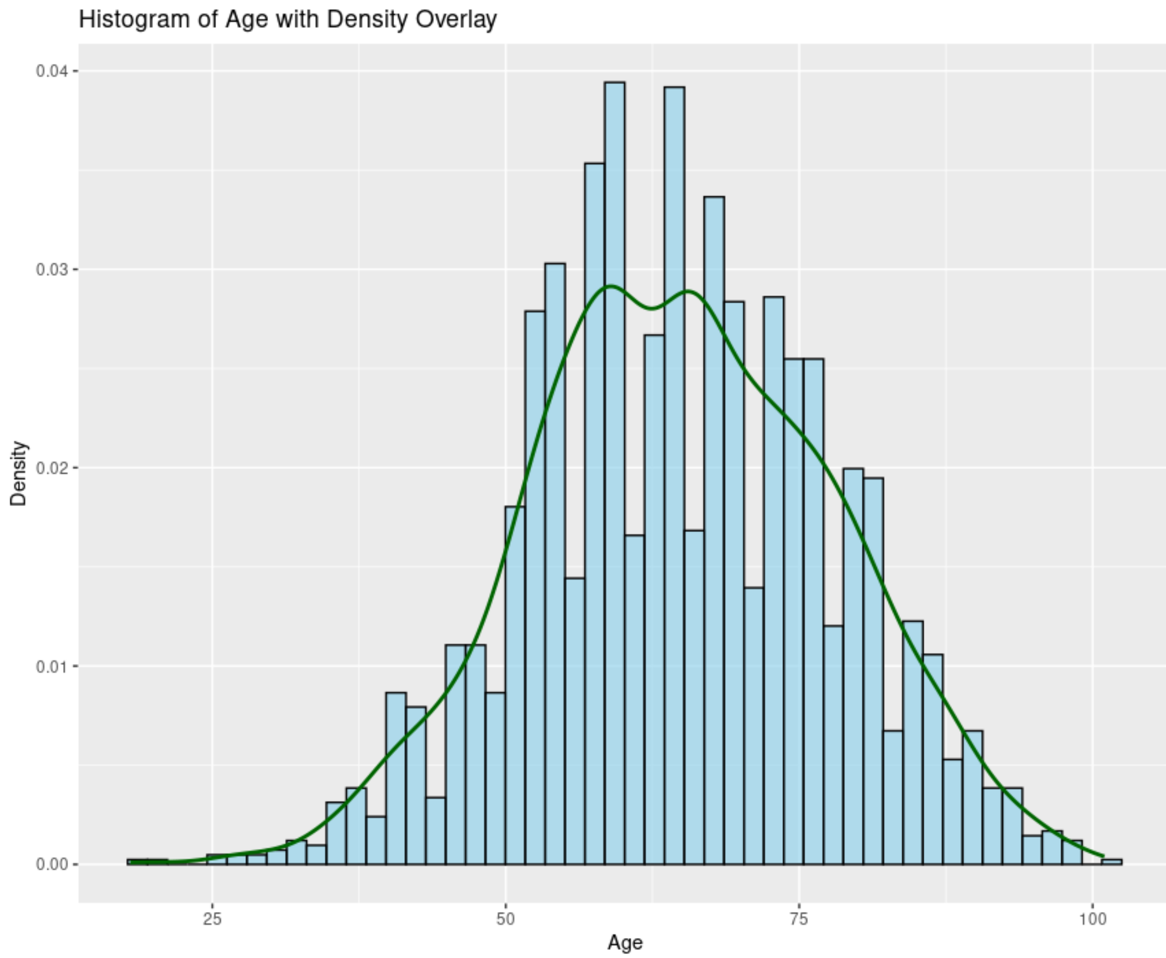

Figure 17 Histogram and Density plot of Age

- Gross Domestic Product (GDP)

I plotted the histogram of GDP by country with density curve. The GDP histogram has a positively skewed distribution, indicating that most of the countries incorporated GDP values only in the lower region of the scale. However, a few countries display higher GDP values than the rest, as depicted by the bars on the right of Figure 17.

```
> ggplot(data = Billionaires, aes(x = as.numeric(gdp_country))) +
+   geom_histogram(aes(y = after_stat(density)), bins = floor(sqrt(nrow(Billionaires))),
+                  fill = "skyblue", color = "black") +
+   geom_density(color = "darkgreen", linewidth = 1) +
+   labs(
+     title = "Histogram of GDP by Country with Density Overlay",
+     x = "GDP (in billions)",
+     y = "Density"
+   )
> |
```

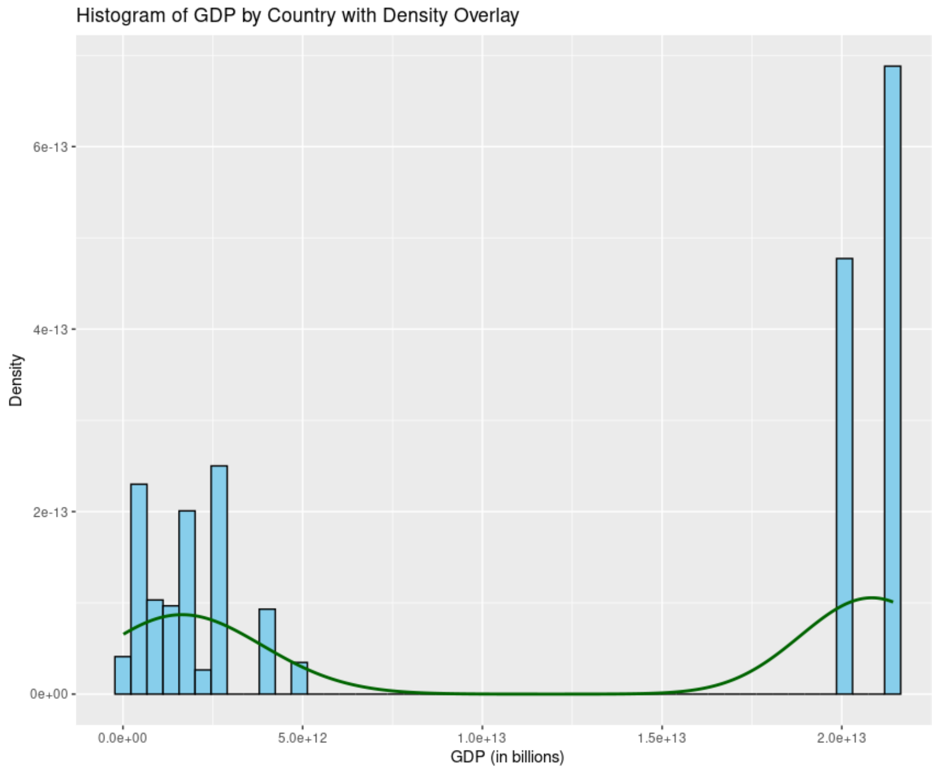

Figure 18 Histogram and Density plot of GDP by Country

- Population Country of billionaires

The histogram with density plot for the population by country of billionaire persons shows an interesting distribution. Most billionaires are from countries with relatively smaller populations, which is evident from the high density on the side of the plot. A sharp peak in the middle is countries with a population of 500 million, likely populous countries with many billionaires. On the extreme right, there is a smaller number of billionaires, but they are from highly populated countries, more than one billion, and this implies that population could determine the billionaire population across the world.

```
> ggplot(data = Billionaires, aes(x = population_country)) +
+   geom_histogram(bins = floor(sqrt(nrow(Billionaires))), fill = "skyblue", color = "black", aes(y =
after_stat(density))) +
+   geom_density(color = "darkgreen", linewidth = 1) +
+   labs(title = "Histogram of Population by Country with Density Overlay", x = "Population (in billi
ons)", y = "Density")
>
```

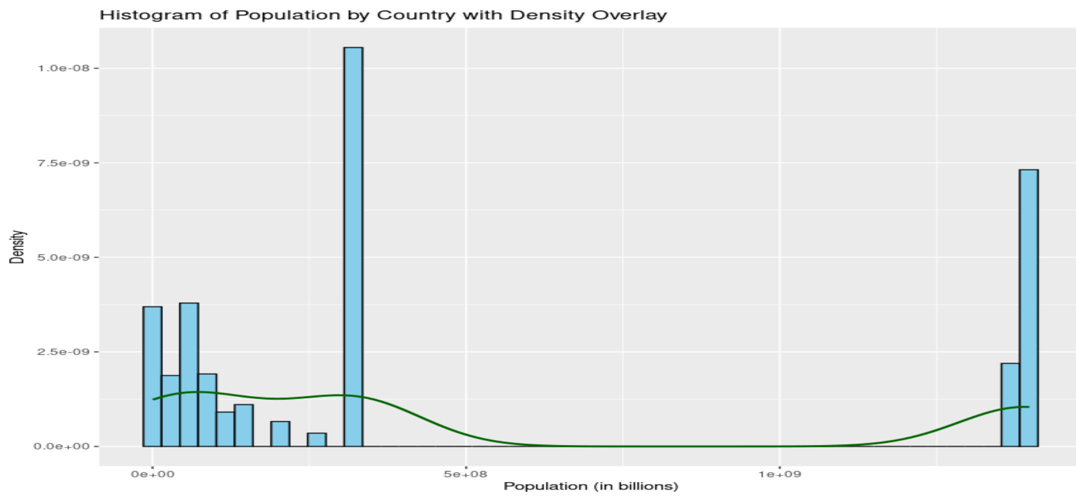

Figure 19 Histogram of Population by Country

- Histogram and density plot of Life expectancy

The distribution of life expectancy is normal; most of the countries have values between 70-80 years.

```
> ggplot(Billionaires, aes(x = life_expectancy_country)) +
+   geom_histogram(binwidth = 2, fill = "skyblue", color = "black", alpha = 0.7, aes(y =
after_stat(density))) +
+   geom_density(color = "darkgreen", linewidth = 1) +
+   labs(title = "Histogram of Life Expectancy by Country with Density Overlay",
+        x = "Life Expectancy (Years)",
+        y = "Density")
```

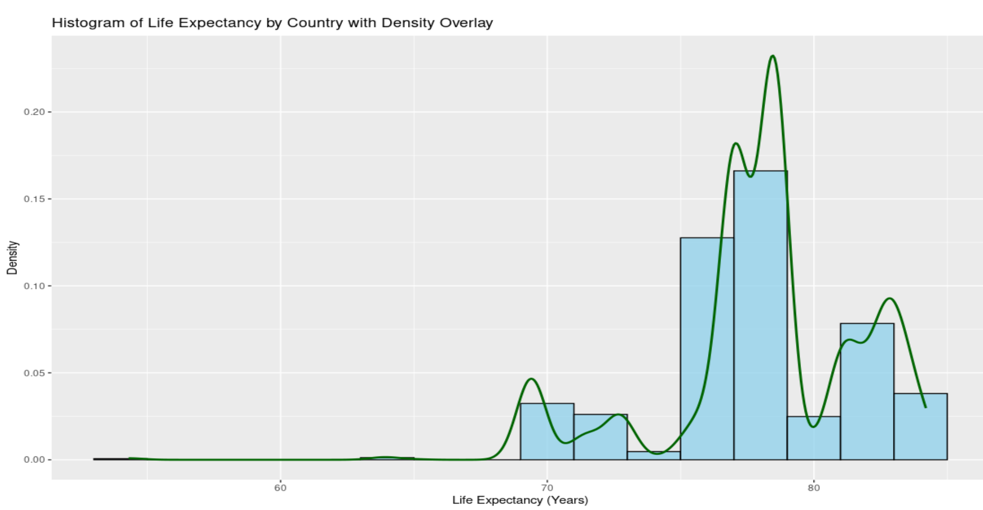

Figure 20 Histogram of Life Expectancy by Country

2.4.2 Grouped Density Plots

- Final Worth by Self-Made Status

By grouping final worth by self-made status, the density plot shows that self-made billionaires exhibit a broader range of wealth distribution. In contrast, those with inherited wealth are concentrated in a narrower band.

```
> ggplot(data = Billionaires, aes(x = finalWorth, fill = selfMade)) +
+   geom_density(alpha = 0.3) +
+   labs(title = "Density Plot of Final Worth by Self-Made Status", x = "Final Worth (in billions)",
y = "Density")
>
```

The density plot of final worth by self-made status reveals that most billionaires, whether self-made or not, have lower net worth, with wealth sharply declining as net worth increases. I believe distributions overlap slightly, but both groups follow a similar pattern, emphasizing that extreme wealth is rare in both categories.

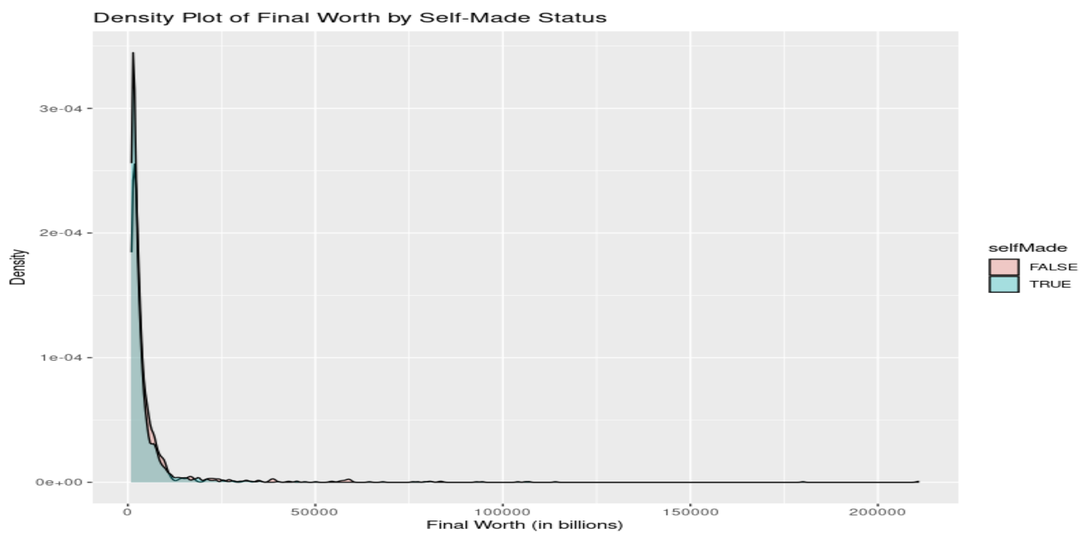

Figure 21 Density Plot of Final Worth by Self-Made Status

- Age by Gender

The density plot illustrates differences in how male and female billionaires' ages are distributed. Male billionaires tend to have a more even spread across ages, while female billionaires show a peak in mid-life.

```
> ggplot(data = Billionaires, aes(x = age, fill = gender)) +geom_density(alpha = 0.3) +
+ labs(title = "Density Plot of Age by Gender", x = "Age", y = "Density")
>
```

Figure 22 Density Plot of Age by Gender

2.5 Bar plots (categorical variables)

Bar graphs are a great way to visually analyze the relationships between two categorical variables or summarize the distribution of one categorical variable. Below are various bar graphs using the Billionaires dataset and related interpretation and code.

2.5.1 Category

Bar plots help to understand which industries are most important for the formation of billionaire wealth based on their distribution by categories such as “Technology," “Finance,” "Retail,” etc.

```
> ggplot(Billionaires, aes(x = category)) + geom_bar(fill = "skyblue", color = "black") +
+       labs(title = "Barplot of Categories", x = "Category", y = "Count") +
+       theme(axis.text.x = element_text(angle = 45, hjust = 1))
> 
```

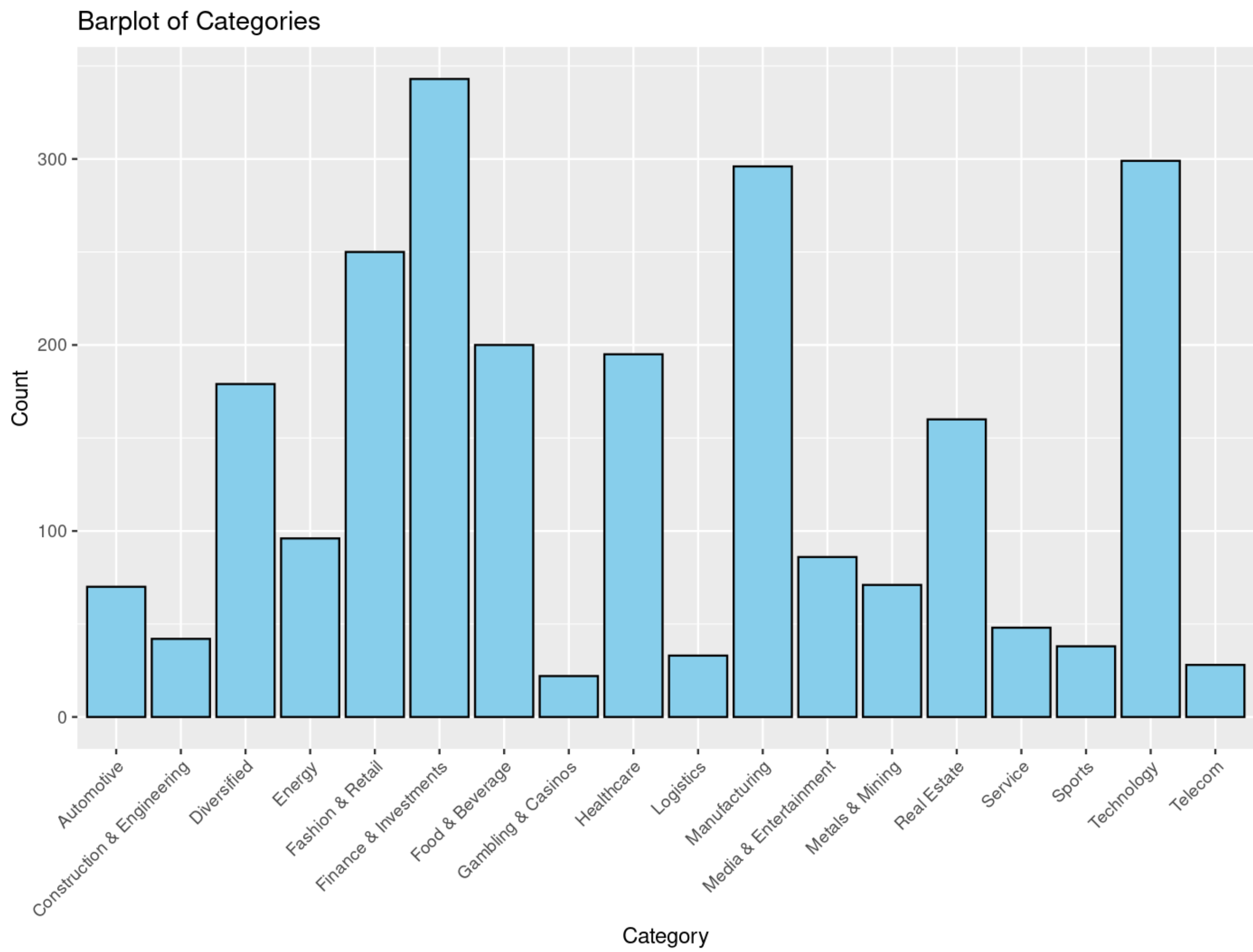

Figure 23 Bar plot of Categories

2.5.2 Gender

The gender distribution shows how many of the billionaires are male and how many are female.

```
> ggplot(Billionaires, aes(x = gender)) +
+ geom_bar(fill = "pink", color = "black") +
+ labs(title = "Barplot of Gender", x = "Gender", y = "Count")
```

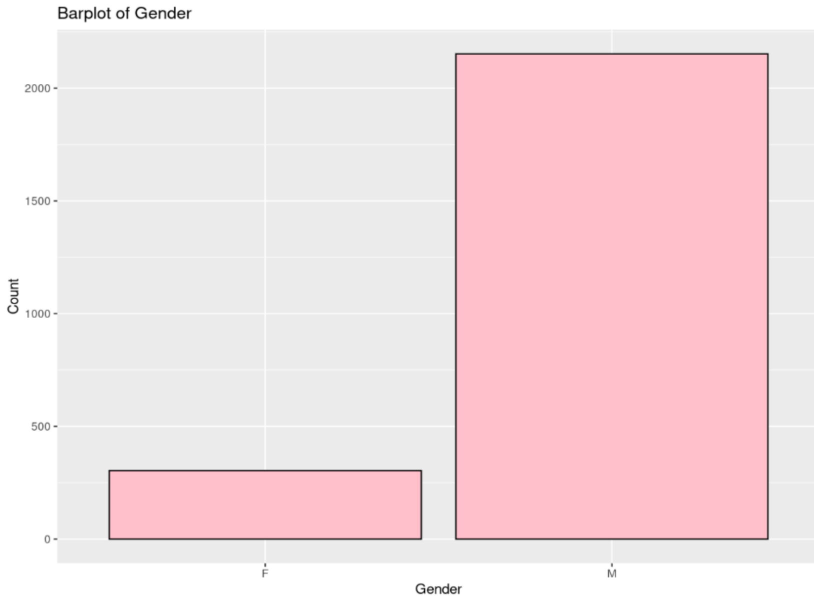

Figure 24 Bar Plot of Gender

2.5.3 Country

The number of billionaires by the country can reveal the situation with economic opportunities and the market in these countries.

```
> ggplot(Billionaires, aes(x = country)) +
+   geom_bar(fill = "lightgreen", color = "black") +
+   labs(title = "Barplot of Countries", x = "Country", y = "Count") +
+   theme(axis.text.x = element_text(angle = 90, hjust = 1))
>
```

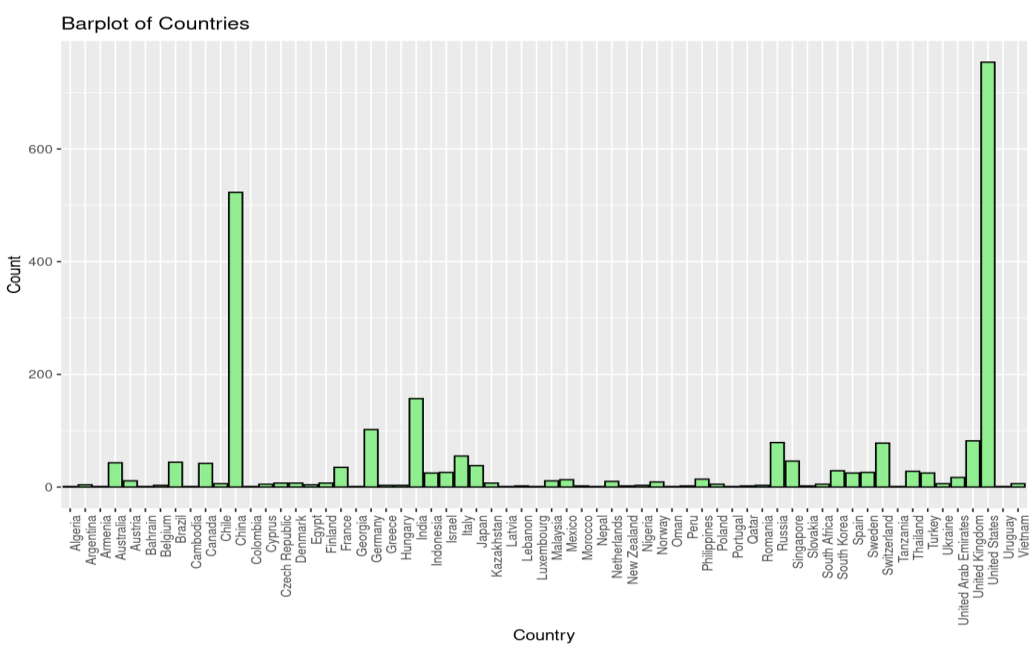

Figure 25 Bar Plot of Countries

2.5.4 Self-Made vs Inherited

The bar plot represents the number of billionaires who have accumulated their money through their endeavors instead of inheriting it.

```
> ggplot(Billionaires, aes(x = selfMade)) +
+    geom_bar(fill = "skyblue", color = "black") +
+   labs(title = "Barplot of Self-Made vs Inherited", x = "Self-Made", y = "Count")
>
```

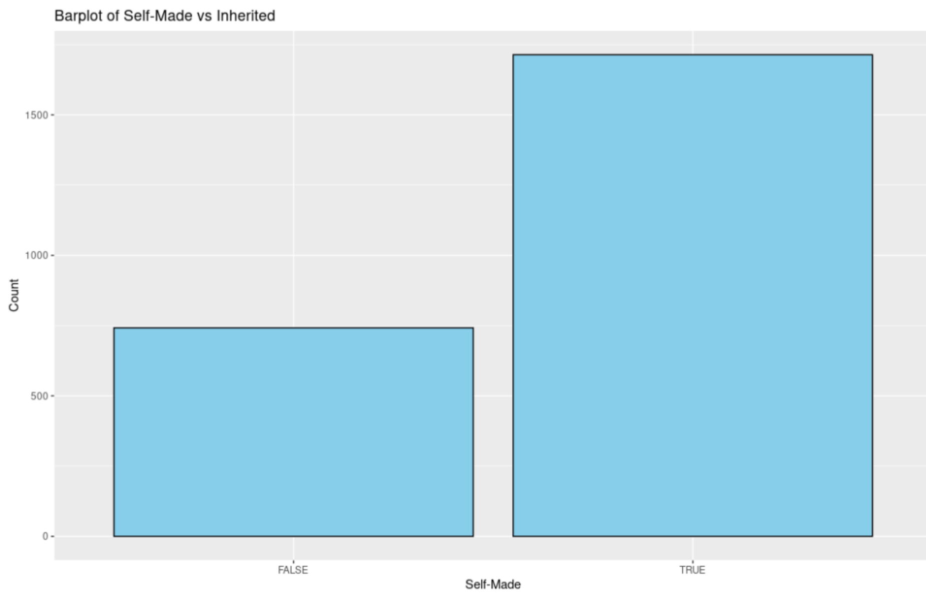

Figure 26 Bar Plot of Self-Made Vs Inherited

2.5.5 Industries vs. Country

This bar graph examines the number of billionaires by industry in various countries. The bar graph labeled as industries vs. countries represents the global billionaire's distribution across the industries. Technology and finance are leading industries in the United States and China and fashion and retail are popular in France and Italy. The picture shows that several critical sectors of the world economy are concentrated in several countries, indicating the international division of labor. For instance, the local market leaders in technology are the United States, while the market leaders in fashion and retail are France. This distribution reflects the relative economic capabilities and leadership in different countries' industries.

```
> ggplot(data = Billionaires, aes(x = industries, fill = industries)) +
+   geom_bar() +
+   facet_wrap(~ country, scales = "free_y") +
+   labs(title = "Industries by Country",
+        x = "Industry", y = "Count") +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none") +
+   theme_minimal()
```

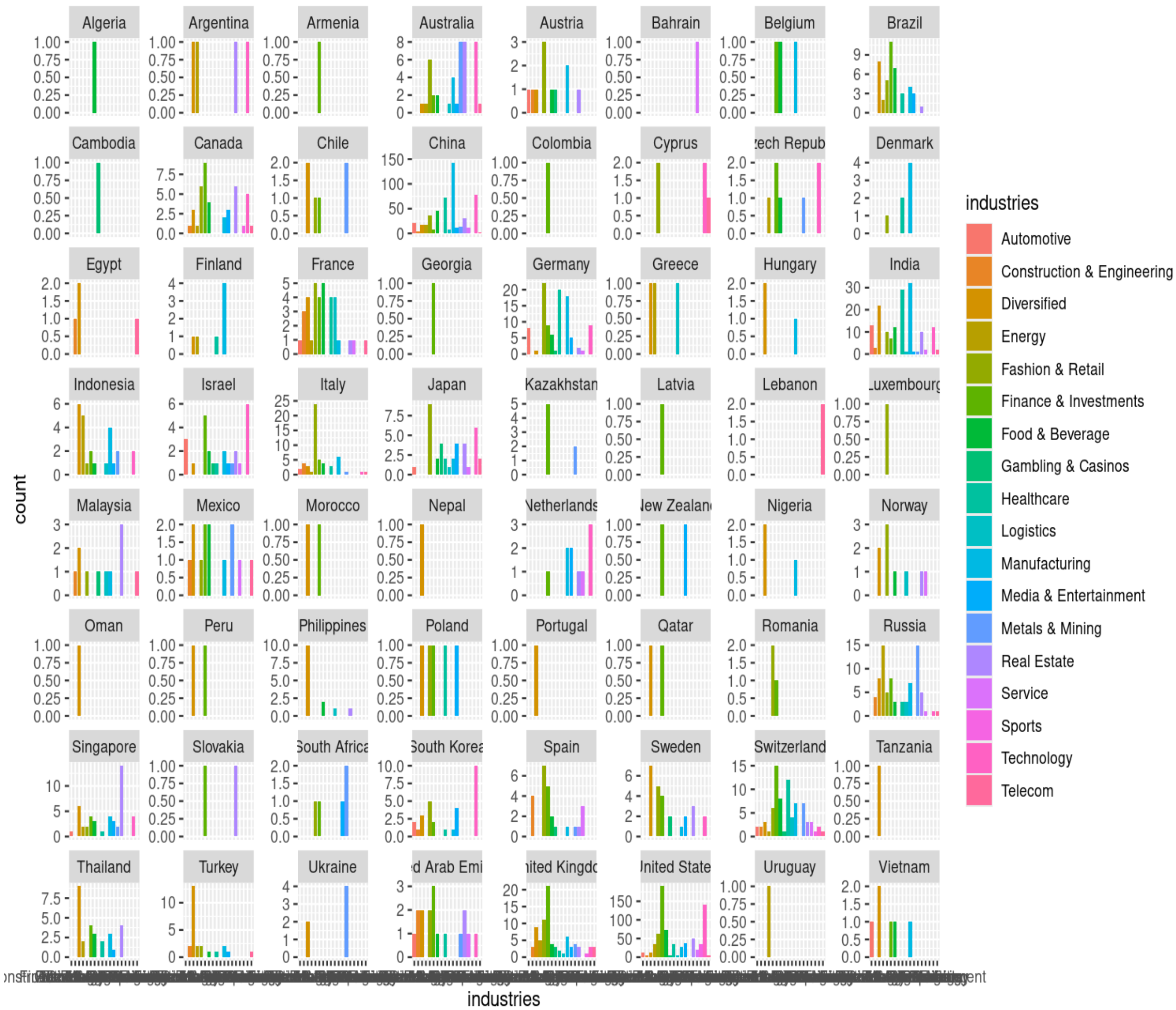

Figure 27 Industries Vs. Countries

2.5.6 Gender across Industries

The bar graph shows the number of males and females in various industries.

```
> bardata3 <- table(Billionaires$gender, Billionaires$industries)
> bardf3 <- as.data.frame(bardata3)
> colnames(bardf3) <- c("Gender", "Industry", "Count")
> ggplot(data = bardf3, aes(x = Industry, y = Count, fill = Gender)) +
+   geom_bar(stat = "identity", position = "dodge") +
+   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
+   labs(
+     title = "Gender Across Industries",
+     x = "Industry",
+     y = "Count"
+   )
> 
```

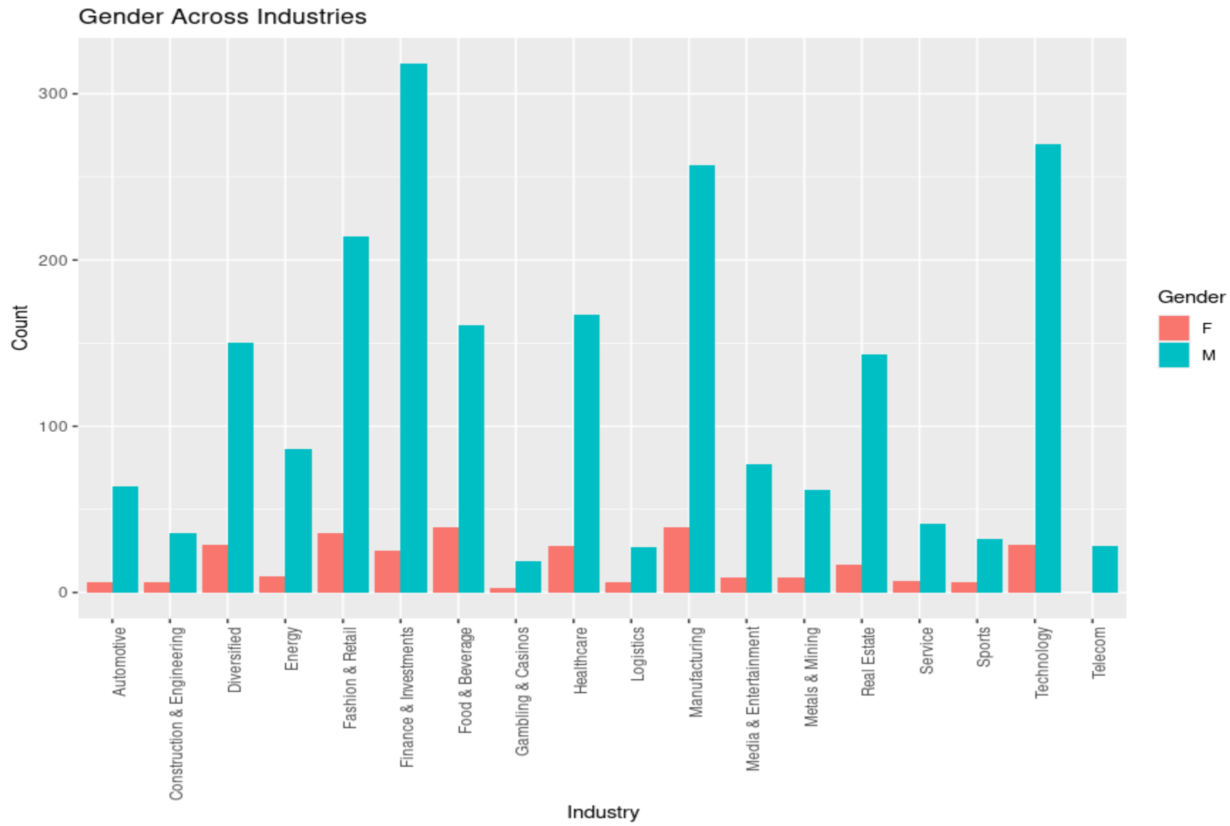

Figure 28 Bar plots of Gender Across Industries

From the bar plot, males are dominant in all industries, and they work in all the high-domain sectors such as "Finance & Investments," "Manufacturing," and "Technology." Females are visible in almost all industries, although fewer in number compared to males.

2.5.7 Country Vs Self-Made Status

The bar graph refers to the distribution of billionaires by Country and whether or not they self-made their wealth (True or False). The United States has the highest number of billionaires, many of which have accumulated their wealth. Likewise, by the value, China also has many billionaires, most of whom self-made their wealth, which signifies the Country's venture creation.

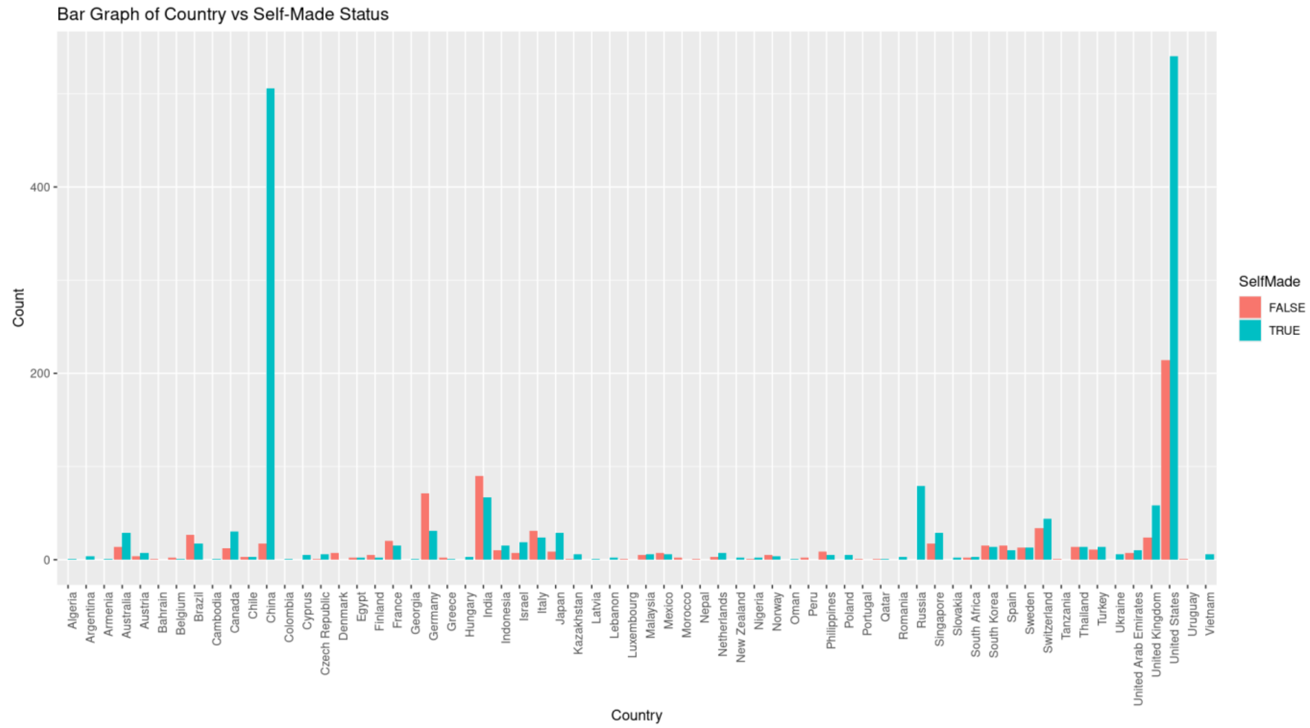

Figure 29 Bar Graph of Country Vs. Self-made Status

```
> bardata <- table(Billionaires$country, Billionaires$selfMade)
> bardf <- as.data.frame(bardata)
> colnames(bardf) <- c("Country", "SelfMade", "Count")
> ggplot(data = bardf, aes(x = Country, y = Count, fill = SelfMade)) +
+   geom_bar(stat = "identity", position = "dodge") +
+   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
+   labs(
+     title = "Bar Graph of Country vs Self-Made Status",
+     x = "Country",
+     y = "Count"
+   )
```

2.5.8 Title Distribution Across Billionaires

I chose a bar graph so that I would be able to examine the frequency of different titles among billionaires. The bar plot also provides the frequency of each of the title, which assists me in determining which leadership role or professional title is more frequent.

```
> filtereddata <- Billionaires[Billionaires$title != "", ]
> ggplot(filtereddata, aes(x = title, fill = title)) +
+   geom_bar() +
+   labs(title = "Distribution of Titles Among Billionaires",
+        x = "Title", y = "Count") +
+   theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 10),
+         legend.position = "none")
>
```

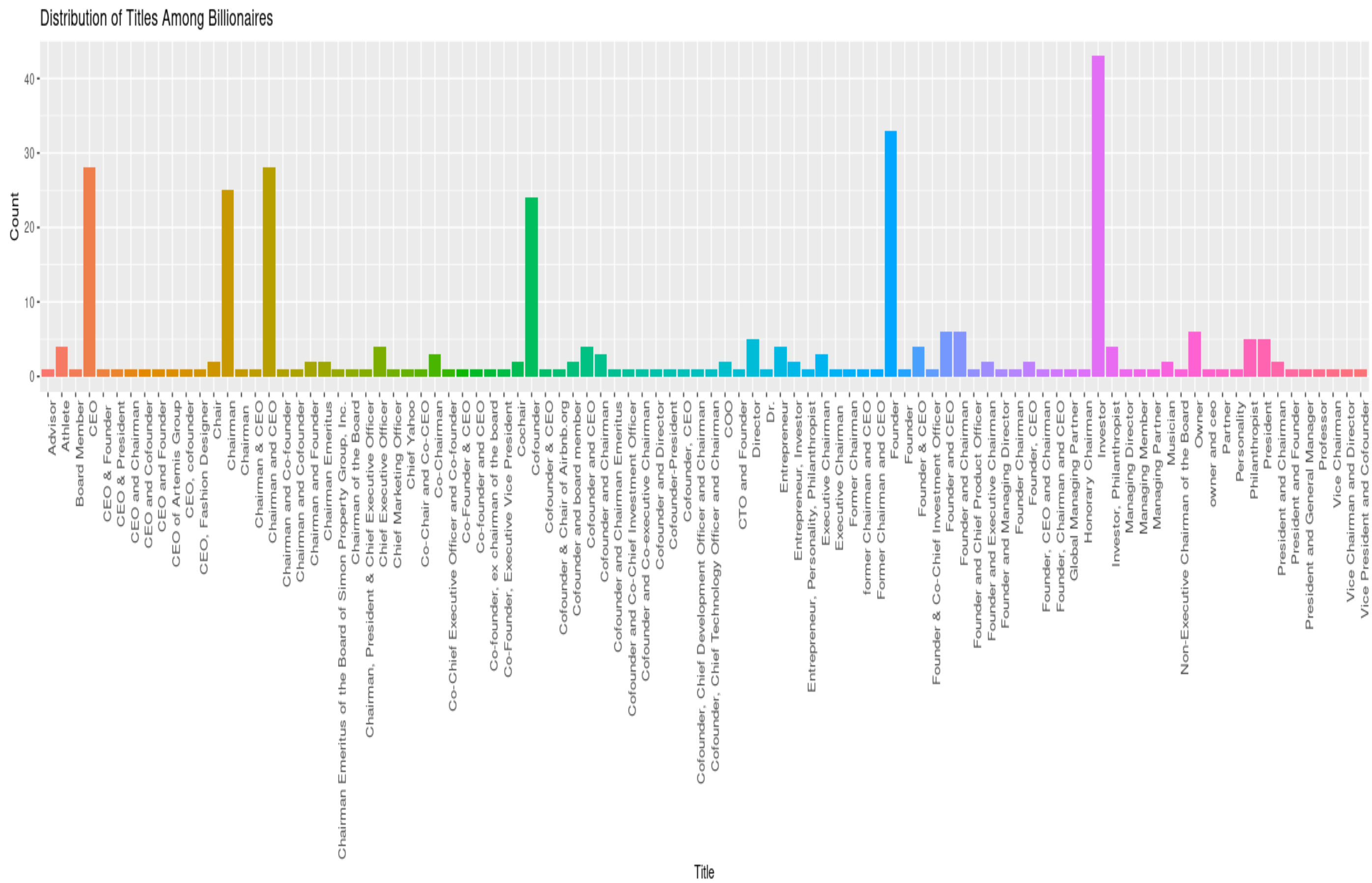

Figure 30 Distribution of Titles Among Billionaires

The bar graph above shows the frequency distribution of the titles of billionaires, excluding any null or missing values. As can be seen from the chart, some titles overrepresent the dataset. The most frequently used position is the CEO, with numbers many times higher than those of other positions. Because CEOs are the most dominant profession among billionaires.

Many of the titles in the billionaire's dataset are Founder, Chairman, and Managing Partner, which also shows their significance in startups and companies.

2.6 Chi-Square Test of Independence

I did a chi-square test to ensure that industries and self-made status have a dependency with the result indicating a $p < 0.05$. The result shows a strong relationship between these two variables, meaning that the type of wealth (self-made or inherited) depends on the industrial category. These results corroborate the visible trends in the bar graphs and the heatmap; for instance, while Technology and Finance self-made billionaires' percentages are high, Real Estate's is low.

```
> conttable <- table(Billionaires$industries, Billionaires$selfMade)
> chisqtest <- chisq.test(conttable)
> chisqtest

        Pearson's Chi-squared test

data:  conttable
X-squared = 210.25, df = 17, p-value < 2.2e-16

>
```

2.7 Summary of statistics (Numeric variables)

```
> numericsummary <- summary(Billionaires[, sapply(Billionaires, is.numeric)])
> print(numericsummary)
      rank        finalWorth          age          birthYear      birthMonth        birthDay      cpi_country
 Min.   :   1   Min.   :  1000   Min.   : 18.00   Min.   :1921   Min.   : 1.000   Min.   : 1.00   Min.   : 99.55
 1st Qu.: 636   1st Qu.:  1500   1st Qu.: 56.00   1st Qu.:1948   1st Qu.: 2.000   1st Qu.: 1.00   1st Qu.:117.24
 Median :1312   Median :  2300   Median : 65.00   Median :1958   Median : 6.000   Median :11.00   Median :117.24
 Mean   :1285   Mean   :  4699   Mean   : 64.91   Mean   :1957   Mean   : 5.757   Mean   :12.28   Mean   :127.76
 3rd Qu.:1905   3rd Qu.:  4300   3rd Qu.: 74.00   3rd Qu.:1966   3rd Qu.: 9.000   3rd Qu.:21.00   3rd Qu.:125.08
 Max.   :2540   Max.   :211000   Max.   :101.00   Max.   :2004   Max.   :12.000   Max.   :31.00   Max.   :288.57

 cpi_change_country  gdp_country         gross_tertiary_education_enrollment gross_primary_education_enrollment_country
 Min.   :-1.900     Min.   :1.367e+10   Min.   :  4.00                     Min.   : 84.7
 1st Qu.: 1.700     1st Qu.:1.736e+12   1st Qu.: 50.60                     1st Qu.:100.2
 Median : 2.900     Median :1.991e+13   Median : 65.60                     Median :101.8
 Mean   : 4.364     Mean   :1.168e+13   Mean   : 67.26                     Mean   :102.9
 3rd Qu.: 7.500     3rd Qu.:2.143e+13   3rd Qu.: 88.20                     3rd Qu.:102.6
 Max.   :53.500     Max.   :2.143e+13   Max.   :136.60                     Max.   :142.1

 life_expectancy_country tax_revenue_country_country total_tax_rate_country population_country  latitude_country longitude_country
 Min.   :54.30           Min.   : 0.10               Min.   :  9.90         Min.   :6.454e+05   Min.   :-40.90   Min.   :-106.35
 1st Qu.:77.00           1st Qu.: 9.60               1st Qu.: 36.60         1st Qu.:6.706e+07   1st Qu.: 35.86   1st Qu.: -95.71
 Median :78.50           Median : 9.60               Median : 41.20         Median :3.282e+08   Median : 37.09   Median :  12.57
 Mean   :78.12           Mean   :12.55               Mean   : 43.98         Mean   :5.143e+08   Mean   : 34.83   Mean   :  12.60
 3rd Qu.:80.90           3rd Qu.:12.80               3rd Qu.: 59.10         3rd Qu.:1.366e+09   3rd Qu.: 38.96   3rd Qu.: 104.20
 Max.   :84.20           Max.   :37.20               Max.   :106.30         Max.   :1.398e+09   Max.   : 61.92   Max.   : 174.89

>
```

The summary statistics show facts regarding wealth, education, life expectancy, and the country of billionaires. The final Worth range is about 1 billion to 211,000 billion for a median of 2,290 billion, proving that they are not spread evenly across the population. Billionaires are between 18 and 100 years of age, and the median and average age is 65. The gross tertiary education enrollment stands at 65. 67%, with the highest figure of 136.60 %, implying that access to higher education remains a major issue in many countries of the world. The life expectancy ranges from 54.30 to 84.20, and the mean age expectancy is 78.52 years, indicating differences in health enrolment and Development. These numbers provide an understanding of the various components of the socioeconomic and demographical character of billionaires’ wealth and global inequality.

```
> summary(Billionaires$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   56.00   65.00   64.91   74.00  101.00 
> 
```

For example, Summary statistics of age variable show that the minimum age is 18 and the maximum age of billionaires is 101.00.

I also look at the summary of all variables of the dataset to find more insight to continue

```
> summary(Billionaires)
      rank        finalWorth                        category     personName             age                    country          city                   source                       industries      countryOfCitizenship
 Min.   :   1   Min.   :  1000   Finance & Investments:343   Length:2456        Min.   : 18.00   United States :754   New York:  99   Real estate    : 122   Finance & Investments:343   United States:733
 1st Qu.: 636   1st Qu.:  1500   Technology           :299   Class :character   1st Qu.: 56.00   China         :523   Beijing :  68   Diversified    :  88   Technology           :299   China        :484
 Median :1312   Median :  2300   Manufacturing        :296   Mode  :character   Median : 65.00   India         :157   Shanghai:  64   Investments    :  87   Manufacturing        :296   India        :165
 Mean   :1285   Mean   :  4699   Fashion & Retail     :250                      Mean   : 64.91   Germany       :102   London  :  61   Pharmaceuticals:  80   Fashion & Retail     :250   Germany      :115
 3rd Qu.:1905   3rd Qu.:  4300   Food & Beverage      :200                      3rd Qu.: 74.00   United Kingdom: 82   Moscow  :  60   Software       :  63   Food & Beverage      :200   Russia       :102
 Max.   :2540   Max.   :211000   Healthcare           :195                      Max.   :101.00   Russia        : 79   Mumbai  :  56   Hedge funds    :  41   Healthcare           :195   Canada       : 60
                                 (Other)              :873                                       (Other)       :759   (Other) :2048   (Other)        :1975   (Other)              :873   (Other)      :797
                          organization   selfMade                     status     gender    birthDate           lastName           firstName                       title          date                           state
                                 :2136   FALSE: 742   D                  :1136   F: 304   Length:2456        Length:2456        Length:2456                           :2123   Min.   :2023-04-04              :1703
 Meta Platforms                  :   4   TRUE :1714   E                  : 251   M:2152   Class :character   Class :character   Class :character   Investor       :  43   1st Qu.:2023-04-04   California: 178
 Gap Inc.                        :   3                N                  : 127            Mode  :character   Mode  :character   Mode  :character   Founder        :  33   Median :2023-04-04   New York  : 128
 Airbnb, Inc.                    :   2                R                  :  61                                                                     CEO            :  28   Mean   :2023-04-04   Florida   :  94
 Alimentation Couche Tard Inc. Cl A:   2              Split Family Fortune:  68                                                                    Chairman and CEO:  28  3rd Qu.:2023-04-04   Texas     :  70
 Alphabet                        :   2                U                  : 813                                                                     Chairman       :  25   Max.   :2023-04-04   Illinois  :  24
 (Other)                         : 307                                                                                                             (Other)        : 176                        (Other)   : 259
     residenceStateRegion   birthYear      birthMonth        birthDay       cpi_country     cpi_change_country  gdp_country          gross_tertiary_education_enrollment gross_primary_education_enrollment_country
                 :1709   Min.   :1921   Min.   : 1.000   Min.   : 1.00   Min.   : 99.55   Min.   :-1.900   Min.   :1.367e+10   Min.   :  4.00                      Min.   : 84.7
 Midwest         :  67   1st Qu.:1948   1st Qu.: 2.000   1st Qu.: 1.00   1st Qu.:117.24   1st Qu.: 1.700   1st Qu.:1.736e+12   1st Qu.: 50.60                      1st Qu.:100.2
 Northeast       : 190   Median :1958   Median : 6.000   Median :11.00   Median :117.24   Median : 2.900   Median :1.991e+13   Median : 65.60                      Median :101.8
 South           : 241   Mean   :1957   Mean   : 5.757   Mean   :12.28   Mean   :127.76   Mean   : 4.364   Mean   :1.168e+13   Mean   : 67.26                      Mean   :102.9
 U.S. Territories:   1   3rd Qu.:1966   3rd Qu.: 9.000   3rd Qu.:21.00   3rd Qu.:125.08   3rd Qu.: 7.500   3rd Qu.:2.143e+13   3rd Qu.: 88.20                      3rd Qu.:102.6
 West            : 248   Max.   :2004   Max.   :12.000   Max.   :31.00   Max.   :288.57   Max.   :53.500   Max.   :2.143e+13   Max.   :136.60                      Max.   :142.1

 life_expectancy_country tax_revenue_country_country total_tax_rate_country population_country  latitude_country longitude_country
 Min.   :54.30           Min.   : 0.10               Min.   :  9.90         Min.   :6.454e+05   Min.   :-40.90   Min.   :-106.35
 1st Qu.:77.00           1st Qu.: 9.60               1st Qu.: 36.60         1st Qu.:6.706e+07   1st Qu.: 35.86   1st Qu.: -95.71
 Median :78.50           Median : 9.60               Median : 41.20         Median :3.282e+08   Median : 37.09   Median :  12.57
 Mean   :78.12           Mean   :12.55               Mean   : 43.98         Mean   :5.143e+08   Mean   : 34.83   Mean   :  12.60
 3rd Qu.:80.90           3rd Qu.:12.80               3rd Qu.: 59.10         3rd Qu.:1.366e+09   3rd Qu.: 38.96   3rd Qu.: 104.20
 Max.   :84.20           Max.   :37.20               Max.   :106.30         Max.   :1.398e+09   Max.   : 61.92   Max.   : 174.89

> 
```

2.8. Outliers/box plots

The main objective of this section is to detect and analyze outliers in data, applying box plots for two different types of variables – categorical and numeric. The boxplot is helpful for presenting numeric data and their distribution across categories.

2.8.1 Final Worth/ Industry

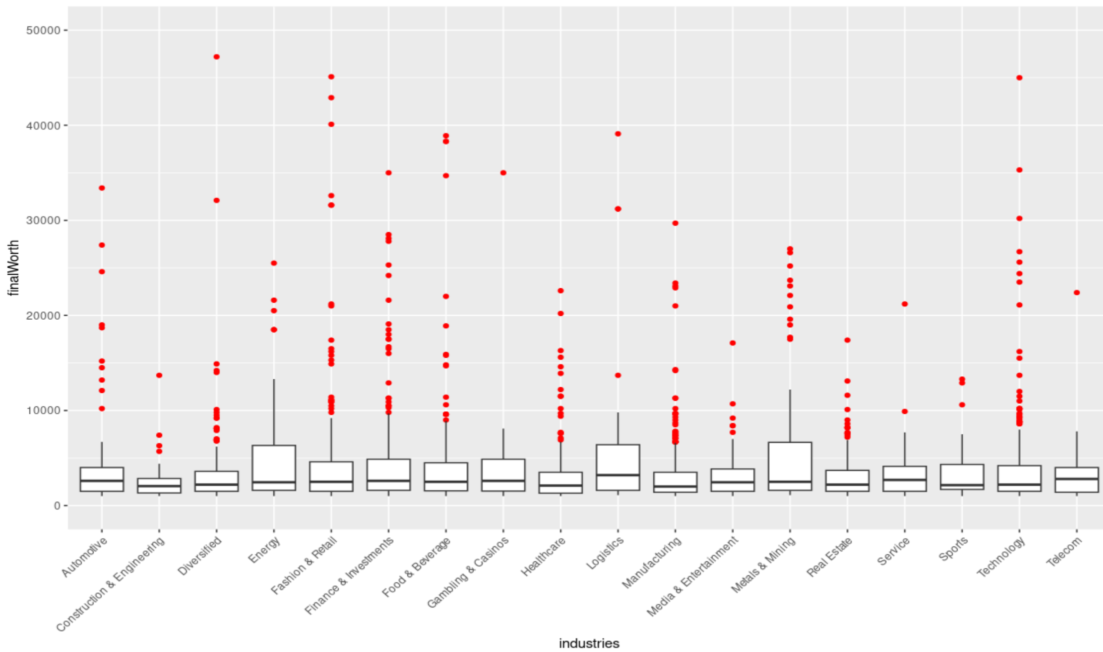

Figure 31 Box Plot of Final Worth/ Industry

```
> ggplot(data = Billionaires, aes(x = industries, y = finalWorth)) +
+   geom_boxplot(outlier.color = "red") +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
>
```

The box plot shows the distribution of the final worth by the industries. Almost all industries have a tiny IQR, meaning that most people in each industry fall within the same wealth bracket. However, there are many such outliers, as marked in red, where specific representatives of certain sectors, for some reason, have much more assets than others in their field. The upper boundary at the y-axis assists in focusing on variability within the primary data, making central tendencies and spread easy to detect while revealing the existence of extremely wealthy people.

2.8.2 Age by Self-Made Status

The box plot below presents the age of billionaires categorized according to whether they are self-made. However, the median age of self-made billionaires is slightly lower than inherited billionaires, so self-made people are relatively young. Both groups have similar ranges and low-end outliers for self-made billionaires, showing that a few people in this category are very young. The box plot further reveals that these billionaires' self-made status and age distribution, similar to that reported overall, suggest that they are younger.

```
> ggplot(data = Billionaires, aes(x = selfMade, y = age)) +
+     geom_boxplot(outlier.color = "red") +
+     labs(title = "Box Plot of Age by Self-Made Status", x = "Self-Made Status", y = "Age")
>
```

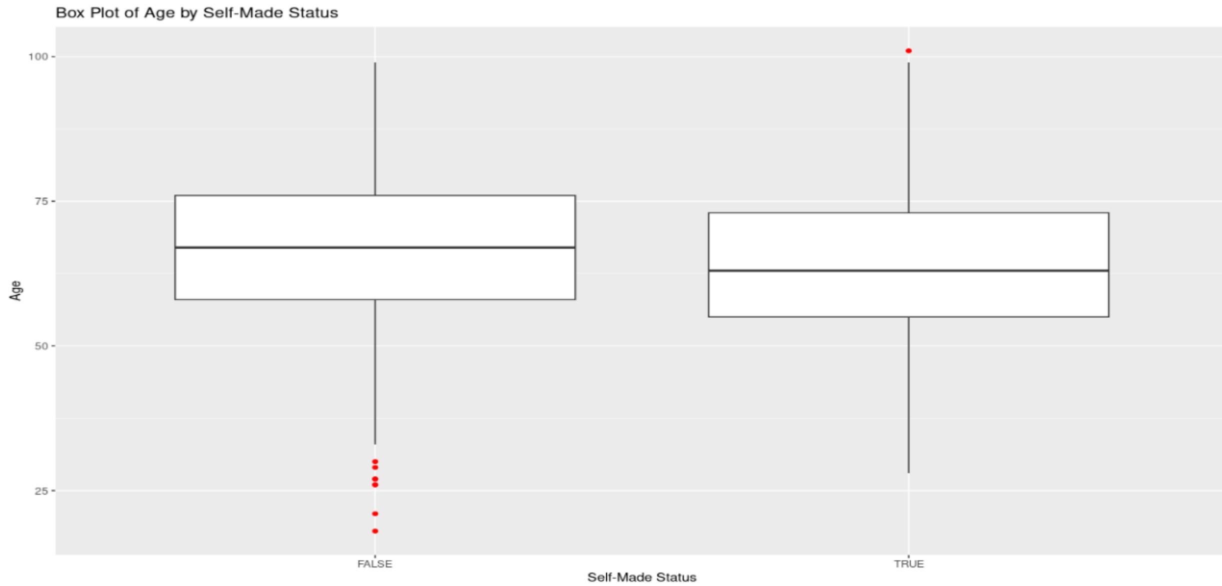

Figure 32 Box Plot of Age by Self-Made Status

2.8.3 Education Enrollment vs. Final Worth

The box plot displays the final wealth of billionaires categorized by their education levels: low, medium, high, and very high. The results suggest that while the average billionaire's wealth is not heavily affected by their educational attainment, billionaires can emerge from any academic background.

```
> ggplot(Billionaires, aes(x = education_category, y = finalWorth)) +
+   geom_boxplot(outlier.color = "red", outlier.size = 1) +
+   scale_y_log10() +
+
+   labs(
+     title = "Distribution of Final Worth by Education Category",
+     x = "Education Category",
+     y = "Log of Final Worth (in billions)"
+   )

>
```

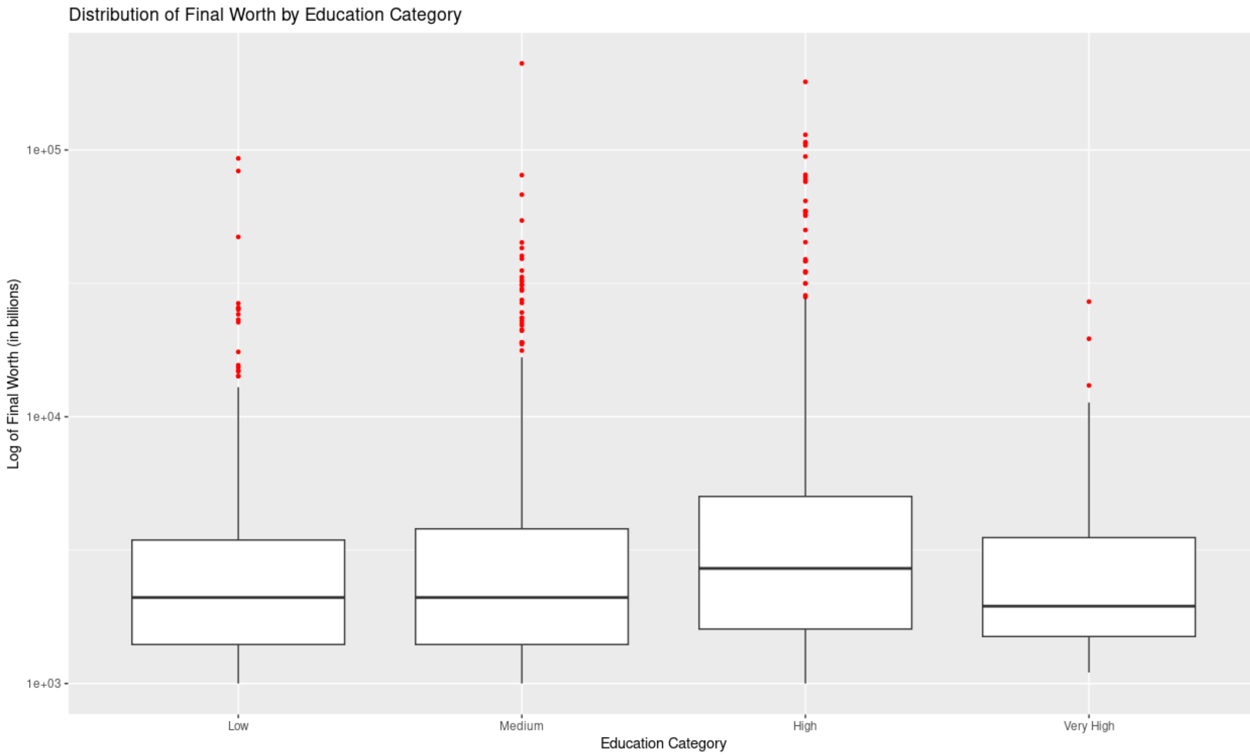

Figure 33 Distribution of Final Worth by Education Category

2.9 World Map Showing Billionaires by Country

To see the number of billionaires in each country, I created a gradient-colored world map. Dark blue tones represent the countries with more billionaires, including China, Russia, and the United States, and lighter tones represent the countries with smaller amounts of billionaires.

This map helps to show where the world's wealth is concentrated and where economic activities are most concentrated or least concentrated depending on the color contrast chosen.

```
> library(dplyr)
> library(ggplot2)
> library(rnaturalearth)
> library(rnaturalearthdata)
> library(sf)
> billionaire_counts <- Billionaires %>%
+     group_by(country) %>%
+     summarise(billionaire_count = n(), .groups = 'drop')
> billionaire_counts$country <- gsub("United States", "United States of America", billionaire_counts$country)
> world_map <- ne_countries(scale = "medium", returnclass = "sf") %>%
+     left_join(billionaire_counts, by = c("name" = "country"))
> world_map$billionaire_count[is.na(world_map$billionaire_count)] <- 0
> ggplot(data = world_map) +
+     geom_sf(aes(fill = billionaire_count), color = "black", size = 0.1) +
+     scale_fill_gradient(low = "lightblue", high = "darkblue", na.value = "gray90",
+                         name = "Billionaire Count") +
+     labs(title = "World Map Showing Billionaires by Country") +
+     theme_minimal() +
+     theme(legend.position = "bottom")
>
```

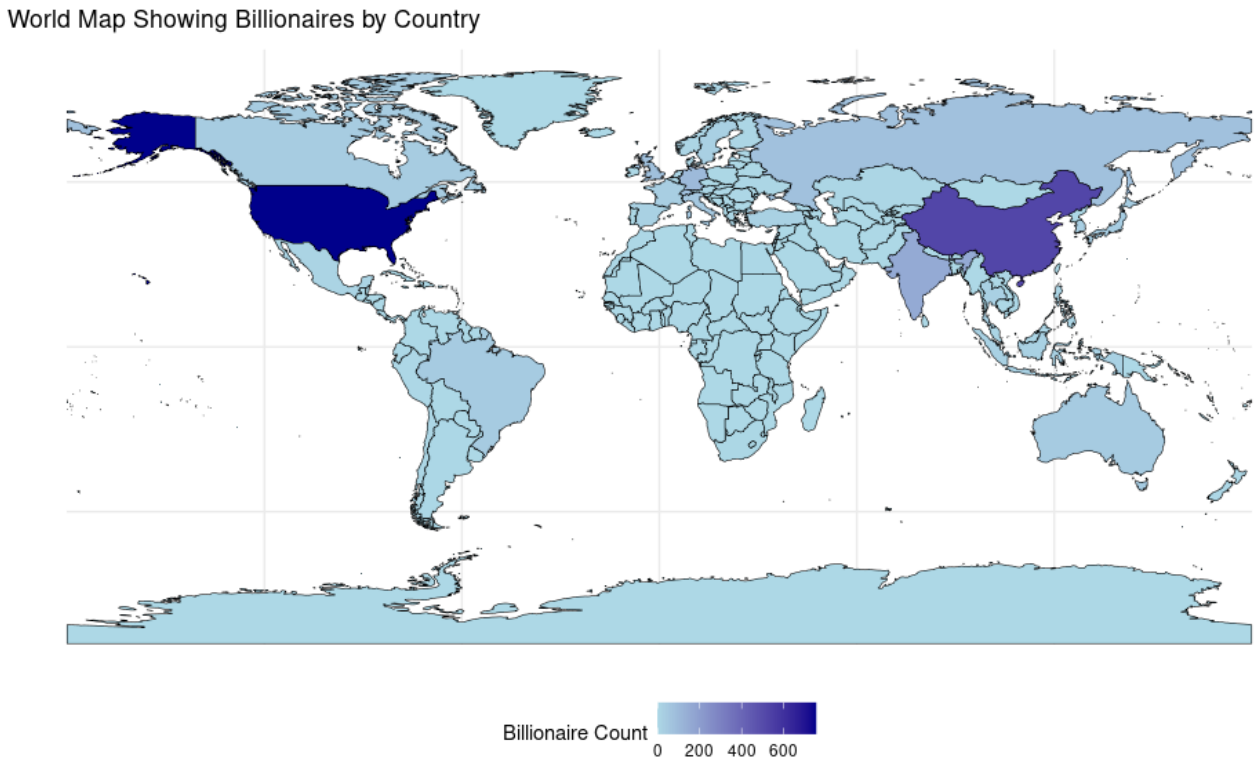

Figure 34 World Map Showing Billionaires by Country

2.10 K-mean Clustering

I conducted a cluster analysis to uncover helpful patterns among billionaires using all numeric variables in the dataset. First, a new dataset was named Billionaires.Cluster that contains only the variables required for the analysis. Subsequently, the scale function was used to normalize the data in to make sure that the different variables make a similar contribution in the clustering process.

```
> library(ggplot2)
> library(cluster)
> Billionaires.Cluster <- data.frame(
+   Billionaires$finalWorth,
+   Billionaires$age,
+   Billionaires$gdp_country,
+   Billionaires$population_country,
+   Billionaires$life_expectancy_country
+ )
> Billionaires.Cluster.Scale <- scale(Billionaires.Cluster)
> Billionaires.Cluster.Scale.DF <- as.data.frame(Billionaires.Cluster.Scale)
```

To determine the number of clusters, the within-group sum of squares (WSS) approach is employed. For the determination of the best value of WSS for K-means clustering, a graph of WSS as a function of the number of clusters k was plotted and the 'elbow' method was used. In the elbow plot, I was realized that the number of clusters that would be efficient for the dataset is 4

```
> wss <- function(k) { kmeans(Billionaires.Cluster.Scale, k, nstart =
> k.values <- 1:15
> wss_values <- sapply(k.values, wss)
> # Plot WSS to find the elbow
> plot(k.values, wss_values, type = "b", pch = 19, frame = FALSE,
+      xlab = "Number of Clusters K",
+      ylab = "Total Within-Clusters Sum of Squares")
```

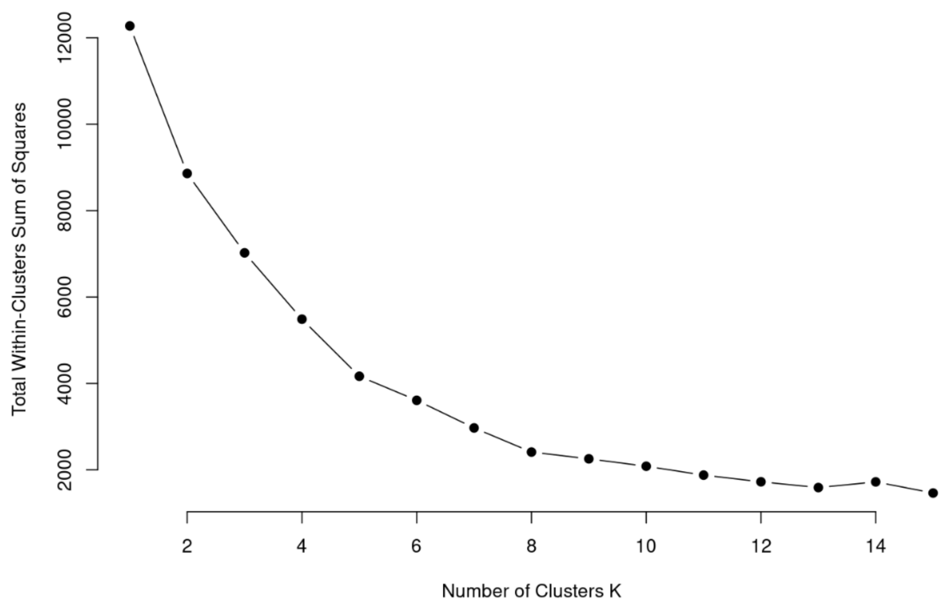

Figure 35 Optimal Number of Cluster

When the number of clusters is defined as 4, k-means clustering was conducted. Thus, the algorithm used in the study divided the data into four categories based on the similarity of the values in the numeric variables. The output gave the sizes of the clusters as 803, 24, 370, 1259 respectively to help explain significant groups within the analysis.

```
> Billionaires.kmeans <- kmeans(Billionaires.Cluster.Scale, centers = 4, nstart = 25)
> Billionaires$Cluster <- Billionaires.kmeans$cluster
> print(Billionaires.kmeans)

> print(Billionaires.kmeans)
K-means clustering with 4 clusters of sizes 803, 24, 370, 1259

Cluster means:
  Billionaires.finalWorth Billionaires.age Billionaires.gdp_country
1             -0.07278018       0.08708725               -1.0371505
2              8.01589017       0.32140806                0.5073039
3             -0.09597972       0.07538701               -1.0295459
4             -0.07817823      -0.08382688                0.9543992
  Billionaires.population_country Billionaires.life_expectancy_country
1                      -0.8446179                           0.93608125
```

```
2                       -0.2746717                            0.12446541
3                        0.2720988                           -1.81562130
4                        0.4639744                           -0.06583045

Clustering vector:
   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 4 1 1 1 1 4 4 4 4 4 1 4
1 1 4 4 1
  [43] 1 4 1 4 4 4 1 1 4 3 3 3 4 1 1 4 3 3 4 4 1 3 1 3 1 3 1 4 3 4 4 4 4 1 3 3 4
1 4 4 4 4
  [85] 4 4 4 3 4 1 4 4 3 4 4 4 1 1 1 4 1 4 1 4 1 1 3 4 3 4 1 4 1 1 4 4 1 3 3 1 4
1 4 4 4 4
 [127] 4 1 3 4 1 1 4 4 4 4 4 1 1 1 4 4 4 1 1 1 1 3 4 4 4 3 4 4 3 3 4 1 4 1 4 4 4
3 1 1 4 1
 [169] 1 4 4 4 1 1 1 1 1 1 3 1 4 1 4 4 4 4 4 4 4 4 1 1 1 4 4 1 3 1 1 4 3 4 4 4 1
1 1 4 1 1
 [211] 1 3 4 4 4 4 4 1 4 1 4 1 3 1 1 3 4 4 4 4 4 3 1 4 4 1 3 3 4 3 4 4 1 1 4 1 1
4 3 4 4 4
 [253] 1 4 4 4 1 4 4 3 1 4 4 4 4 1 1 4 4 4 4 4 1 4 3 1 4 1 1 4 4 4 4 4 4 4 1 4 1
4 1 4 3 4
 [295] 3 3 1 3 4 4 4 4 4 1 1 4 4 1 4 4 1 1 4 4 1 1 4 3 4 3 3 4 4 4 3 4 4 4 4 4 1
4 4 1 1 4
 [337] 4 4 4 4 4 4 3 4 4 4 1 1 1 1 1 4 4 1 1 4 4 4 4 4 1 4 4 1 4 4 4 4 4 1 4 4 4
1 1 1 4 4
 [379] 4 1 1 4 1 4 1 1 4 1 3 1 1 4 4 4 1 3 4 1 1 3 1 4 1 4 3 1 1 4 4 4 4 1 4 4 4
4 1 4 3 4
 [421] 1 1 1 4 4 4 1 4 4 3 4 4 4 4 1 1 4 4 4 4 4 4 1 4 1 4 1 1 4 4 4 1 4 1 4 1 3
4 1 4 1 1
 [463] 1 1 1 1 4 4 1 4 4 3 4 4 4 4 4 4 1 1 4 4 4 4 1 1 4 1 4 4 4 4 3 4 1 4 1 4 4
4 4 4 4 1
 [505] 1 1 4 3 1 4 1 4 4 1 1 1 1 1 4 4 3 4 1 4 4 3 3 1 1 4 4 3 4 4 1 1 1 4 4 4 4
4 4 4 3 4
 [547] 1 4 1 3 1 1 4 4 4 4 4 4 4 4 4 3 4 4 1 4 4 4 1 4 4 1 4 4 3 1 4 4 1 3 1 4 1
3 4 4 4 4
 [589] 4 4 1 1 1 3 4 4 4 3 1 4 1 4 4 4 4 1 3 3 3 4 4 3 4 4 4 4 4 1 3 4 4 4 4 3 4
1 4 4 4 4
 [631] 4 4 4 4 3 3 4 1 1 1 4 4 4 1 1 4 4 3 4 3 3 4 1 4 1 4 4 4 4 4 4 3 4 4 1 4 4
1 4 3 3 4
 [673] 4 1 4 1 1 3 4 1 4 4 1 4 1 1 1 4 4 1 4 3 3 3 4 1 1 1 4 4 3 4 4 4 3 1 4 4 4
4 4 4 4 4
 [715] 3 4 1 4 4 4 4 4 4 4 1 4 4 4 4 4 4 3 1 1 4 1 4 4 4 4 1 4 3 3 1 4 1 1 1 1 4
4 3 4 4 1
 [757] 1 1 1 1 4 1 4 1 4 4 4 4 4 3 4 1 4 4 1 1 1 1 1 1 4 4 3 4 4 1 1 4 1 3 1 1 3
4 4 1 4 4
 [799] 4 1 4 1 3 3 3 1 1 1 1 4 4 4 4 4 1 4 4 4 3 1 1 4 4 4 4 1 4 4 4 4 3 1 4 1 1
1 4 1 1 1
 [841] 4 4 3 4 1 1 1 4 4 4 4 1 1 1 3 1 4 1 3 1 4 4 4 1 3 3 3 1 4 4 1 4 4 3 3 1 4
4 1 4 4 1
 [883] 4 1 4 4 4 3 1 1 4 4 4 4 4 4 1 4 4 3 3 1 1 3 4 4 1 3 1 4 4 3 1 3 4 4 1 4 3
1 1 3 4 1
 [925] 1 1 4 4 4 1 4 1 4 4 4 1 4 4 4 4 1 3 1 4 3 1 1 4 3 1 1 4 3 4 1 4 1 4 1 4 1
1 1 4 4 1
 [967] 1 4 4 1 4 1 4 1 1 4 4 1 1 4 3 1 1 4 4 4 4 1 1 1 3 3 1 1 3 3 4 4 4 1
```

```
 [ reached getOption("max.print") -- omitted 1456 entries ]

Within cluster sum of squares by cluster:
[1] 1385.8262  410.3672 1012.6081 2678.0553
 (between_SS / total_SS =  55.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

The cluster centers helped to get the understanding of the subjects within each cluster. For example, Cluster 2 embodied individuals with higher net assets in contrast to Cluster 3 that embodied individuals in countries with lower life span.

Last of all, I generated a cluster plot for the analysis of the results that I have got. The clusplot function was then applied to present the data in two dimensions which illustrated where the k-means had clustered the data.

```
> clusplot(Billionaires.Cluster.Scale.DF, Billionaires.kmeans$cluster,
+         color = TRUE, shade = TRUE, labels = 2, lines = 0)
```

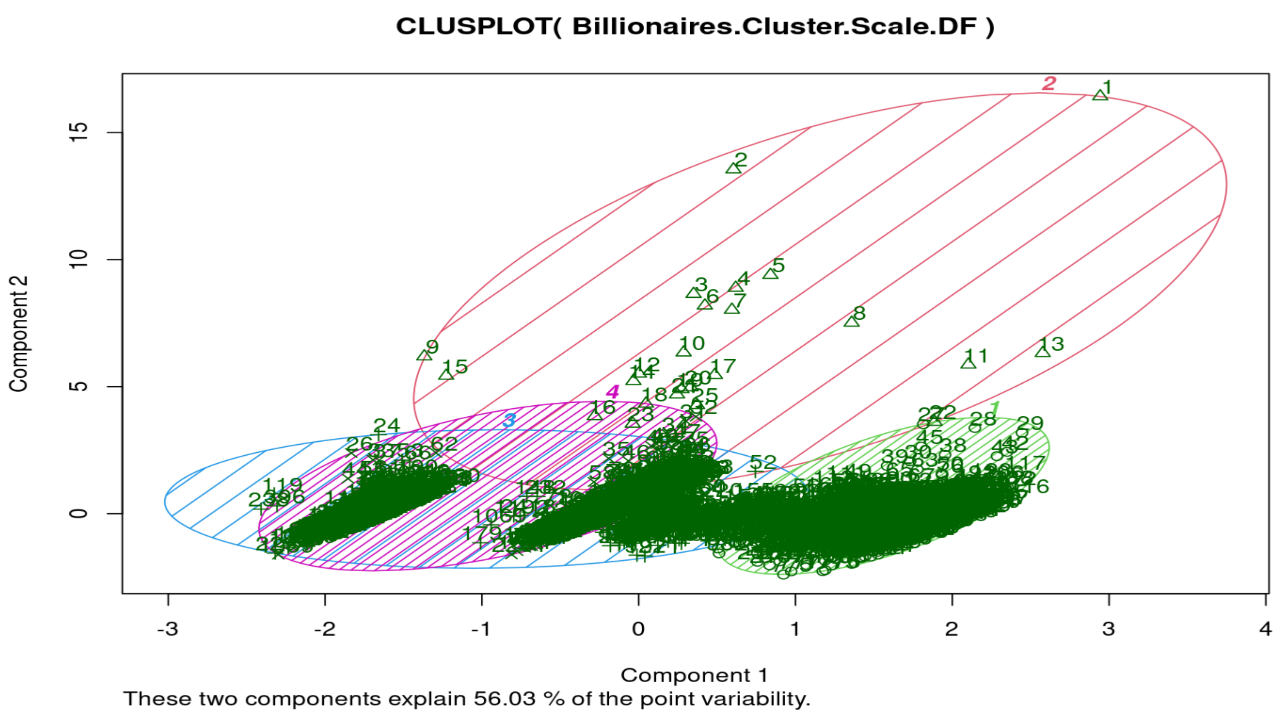

Figure 36 Custers

From the plot analysis, it was easy to observe the four clusters of the population under study. Every cluster corresponded to a segment of the data set; it was useful to understand how billionaires can be divided according to their wealth, age, etc.

```
> head(Billionaires)
  rank finalWorth             category               persnName age       country
city
1    1     211000     Fashion & Retail Bernard Arnault & family  74        France
Paris
2    2     180000           Automotive                Elon Musk  51 United States A
ustin
3    3     114000           Technology               Jeff Bezos  59 United States M
edina
4    4     107000           Technology            Larry Ellison  78 United States
Lanai
5    5     106000 Finance & Investments          Warren Buffett  92 United States
Omaha
6    6     104000           Technology               Bill Gates  67 United States M
edina
              source            industries countryOfCitizenship
1               LVMH      Fashion & Retail               France
2      Tesla, SpaceX            Automotive        United States
3             Amazon            Technology        United States
4             Oracle            Technology        United States
5 Berkshire Hathaway Finance & Investments        United States
6          Microsoft            Technology        United States
                      organization selfMade status gender       birthDate lastName
1 LVMH Moët Hennessy Louis Vuitton    FALSE      U      M   3/5/1949 0:00  Arnault
2                            Tesla     TRUE      D      M  6/28/1971 0:00     Musk
3                           Amazon     TRUE      D      M  1/12/1964 0:00    Bezos
4                           Oracle     TRUE      U      M  8/17/1944 0:00  Ellison
5   Berkshire Hathaway Inc. (Cl A)     TRUE      D      M  8/30/1930 0:00  Buffett
6  Bill & Melinda Gates Foundation     TRUE      D      M 10/28/1955 0:00    Gates
  firstName                title       date      state residenceStateRegion birthYea
r
1   Bernard     Chairman and CEO 2023-04-04                                      194
9
2      Elon                  CEO 2023-04-04      Texas                South      197
1
3      Jeff Chairman and Founder 2023-04-04 Washington                 West      196
4
4     Larry      CTO and Founder 2023-04-04     Hawaii                 West      194
4
5    Warren                  CEO 2023-04-04   Nebraska              Midwest      193
0
6      Bill              Cochair 2023-04-04 Washington                 West      195
5
  cpi_country cpi_change_country  gdp_country gross_tertiary_education_enrollment
1      110.05                1.1 2.715518e+12                                65.6
2      117.24                7.5 2.142770e+13                                88.2
3      117.24                7.5 2.142770e+13                                88.2
4      117.24                7.5 2.142770e+13                                88.2
5      117.24                7.5 2.142770e+13                                88.2
6      117.24                7.5 2.142770e+13                                88.2
  gross_primary_education_enrollment_country life_expectancy_country
1                                      102.5                    82.5
```

```
2                                     101.8                  78.5
3                                     101.8                  78.5
4                                     101.8                  78.5
5                                     101.8                  78.5
6                                     101.8                  78.5
  tax_revenue_country_country total_tax_rate_country population_country latitude_cou
ntry
1                        24.2                   60.7           67059887         46.2
2764
2                         9.6                   36.6          328239523         37.0
9024
3                         9.6                   36.6          328239523         37.0
9024
4                         9.6                   36.6          328239523         37.0
9024
5                         9.6                   36.6          328239523         37.0
9024
6                         9.6                   36.6          328239523         37.0
9024
  longitude_country Cluster
1          2.213749       2
2        -95.712891       2
3        -95.712891       2
4        -95.712891       2
5        -95.712891       2
6        -95.712891       2

>
```

I conducted a clustering analysis using k-means on the billionaires' dataset to explore significant groupings based on numeric attributes: net worth, age, GDP of the countries, population, life expectancy. The output revealed four clusters. Cluster 1 that comprised billionaires with moderately low net worth and age levels and connected this group to countries with small populations but relatively long life spans. Cluster 2, which includes Bernard Arnault and Elon Musk, demonstrated people with extremely high net worth who are slightly older than the world's average age, and most of them are associated with high GDP countries like the United States and France. Cluster 3 was concerned with billionaires from countries with relatively low GDP, young populations, and low life expectancy, which pointed to emerging wealth from developing areas. Lastly, Cluster 4 includes mostly moderately rich billionaires who are relatively young, and the countries they represent are either highly populated or have a high demographic activity.

Thus, the suggestions of the cluster analysis helped sort the billionaires into four groups, which reflect their financial and demographic characteristics

.

2.11 Multiple Regression

I employed the multiple linear regression model, considered final worth as the dependent variable. I selected other numeric variables from the data set to understand the factors affecting the billionaire's wealth. More specifically, this analysis intended to establish how a billionaire's wealth correlates with those demographical, geographic, and economic factors.

Actually, before constructing the model, I conducted a preliminary analysis to distinguish variables that might impact final worth. The examinations of the scatter plots and correlation analysis showed an intense correlation with the variables such as age, gdp_country, and life_expectancy_country.

```
> cor(Billionaires[, c("finalWorth", "age", "gdp_country", "life_expectancy_country")])
                        finalWorth         age gdp_country life_expectancy_country
finalWorth              1.00000000  0.06227230  0.03746733              0.02265391
age                     0.06227230  1.00000000 -0.06718313              0.02049957
gdp_country             0.03746733 -0.06718313  1.00000000             -0.05716769
life_expectancy_country 0.02265391  0.02049957 -0.05716769              1.00000000
> |
```

I standardized all numeric variables to improve comparability and model stability. To handle categorical data, such as industries and countries, I used the fastDummies package to create dummy variables, ensuring they were suitable for analysis.

```
> numeric_variables <- Billionaires[, sapply(Billionaires, is.numeric)]
> numeric_variables_scaled <- scale(numeric_variables)
> library(fastDummies)
> categorical_variables <- Billionaires[, c("industries", "country")]
> categorical_dummies <- dummy_cols(categorical_variables, remove_most_frequent_dummy = TRUE)
> regression_data <- cbind(numeric_variables_scaled, categorical_dummies)
> |
```

I built a multiple linear regression model to predict final worth using age, gdp_country, and life_expectancy_country as predictors. Here is the model equation:

$$finalWorth=\beta0+\beta1\cdot age+\beta2\cdot gdp\_country+\beta3\cdot life\_expectancy\_country+\epsilon$$

```
> model <- lm(finalWorth ~ age + gdp_country + life_expectancy_country, data = Billionaires)
> summary(model)

Call:
lm(formula = finalWorth ~ age + gdp_country + life_expectancy_country,
    data = Billionaires)

Residuals:
   Min     1Q Median     3Q    Max
 -5689  -3114  -2146   -374 205969

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.154e+03  4.410e+03  -0.942  0.34633
age                      5.042e+01  1.573e+01   3.205  0.00137 **
gdp_country              4.572e-11  2.140e-11   2.136  0.03275 *
life_expectancy_country  6.460e+01  5.475e+01   1.180  0.23816
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10090 on 2452 degrees of freedom
Multiple R-squared:  0.006185,	Adjusted R-squared:  0.004969
F-statistic: 5.086 on 3 and 2452 DF,  p-value: 0.001642

>
```

I conducted a residual analysis to make sure that the model was valid. I used the Residuals vs. Fitted plot to confirm linearity and consistent variance in the data.

```
> par(mfrow = c(2, 2))
> plot(model)

>
```

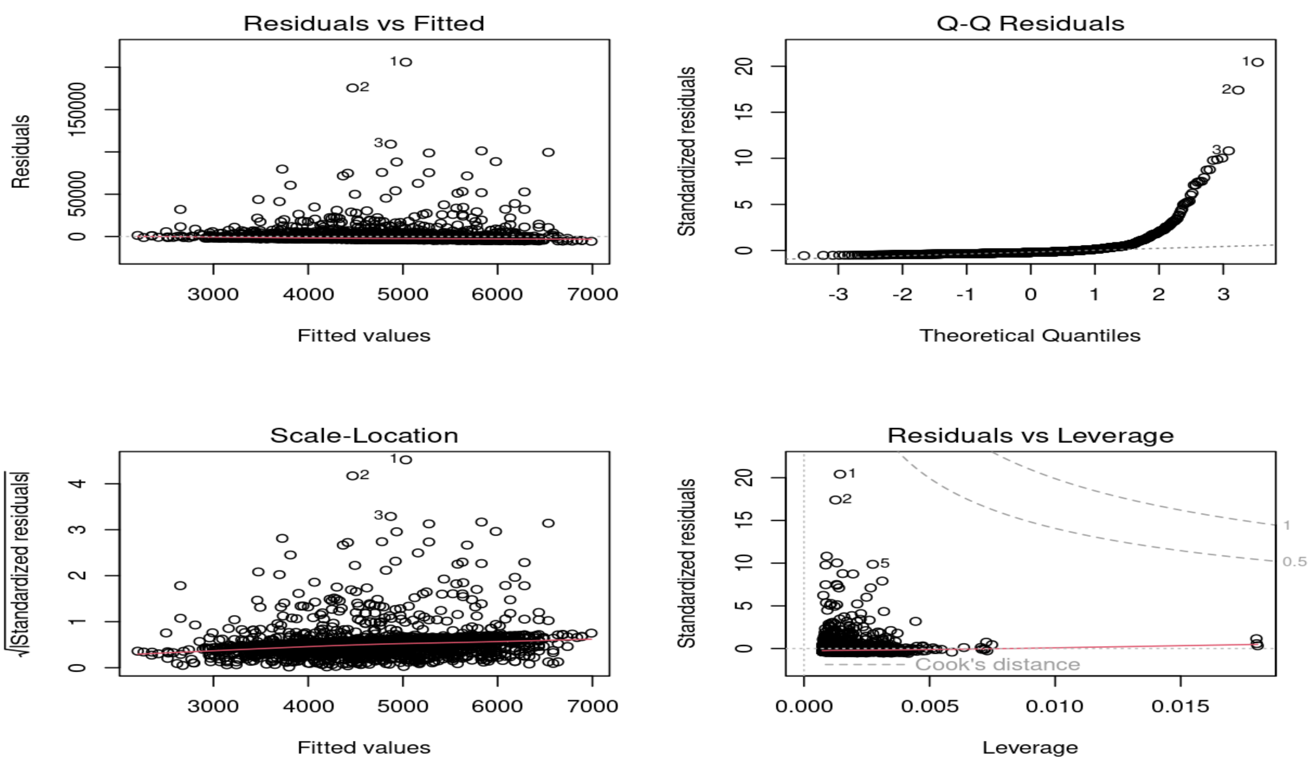

Figure 37 Residual analysis

According to the regression analysis, three variables significantly affected a billionaire's net worth. As expected, the results for GDP per country were positive, which testifies to the fact that billionaires in countries with high GDP usually possess high net worth. Experience was also indicated to be a contributing factor, as a qualitative one, which means that as people grow older, they earn more money. Lastly, life expectancy by country is increasing, meaning that billionaires tend to live in countries with higher living expectancy, probably due to better social living standards. These results suggest a complementary relationship between individual characteristics and national conditions to form billionaire wealth.

```
> library(Metrics)
> set.seed(123)
> training_indices <- sample(1:nrow(Billionaires), 0.7 * nrow(Billionaires))
> training_data <- Billionaires[training_indices, ]
> testing_data <- Billionaires[-training_indices, ]
> model <- lm(finalWorth ~ age + gdp_country + life_expectancy_country, data = training_data)
> predictions <- predict(model, newdata = testing_data)
> mse <- mse(testing_data$finalWorth, predictions)
> print(paste("Mean Squared Error:", mse))
[1] "Mean Squared Error: 65859785.1123105"
>
```

To check the accuracy of this regression model, I divided the data set into a training data set and a testing data set and cross-checked the data; the Mean Squared Error (MSE) equals 65,859,785.11. This value is an initial test showing the model's accuracy in estimating net worth; other evaluations, such as residual analysis or R^2, may also help. The analysis revealed key insights: The role of age is apparent in wealth generation with the help of GDP per country, emphasizing the economic factors that define wealth generation per country while using the life expectancy index as the variable portraying social and health support for wealth sustenance. Evidently, these results show a significant relationship between the elements of individual and economic characteristics of billionaires' wealth.

3. Summary

I studied the Billionaires Statistics Dataset to reveal information about wealth inequality, demographics and the relationship between individual and economic factors that contribute to net worth.

As I expected before and mentioned it in exception section correlation matrices and scatterplots showed weak correlations but moderate positive correlations between population size and GDP, illustrating the power of larger populations to fuel economic growth. K-means clustering of the dataset divided it into four clusters and differentiated the youngest innovators from older traditional wealth builders.

A multiple linear regression model proved the importance of both personal and economic factors and yielded good predictions indicated by a computed Mean Squared Error by using the predictor variables like age, GDP, life expectancy. Visualizations, such as box plots and density plots, showed patterns such as outliers in the Technology and Finance sectors and patterns of wealth accumulation skewed towards older people. Statistical analysis assured the strong association between the industry and self-made status with socio-economic differences in wealth accumulation across sectors.

This analysis gave me valuable information regarding how individual and economic factors shape billionaire wealth and I found that the majority of billionaires earned their money with efforts and work in top industries instead of inherited it.