Capstone Project: Billionaires Statistics Analysis

Nicolas Ojeda

11/29/23

Contents

Chapter 1

Introduction

In this Capstone Project, I delve into the fascinating and complex world of billionaires. My goal is to analyze the wealth distribution among these individuals, not just as figures of affluence but as indicators of broader economic, social, and policy trends. Drawing upon the methodologies outlined in Julius Olufemi Ogunleye's insightful work, especially the use of Linear Regression and Random Forest models, I aim to unravel the intricate patterns of wealth distribution.

Chapter 2

Overview

My approach goes beyond mere numerical analysis; it's an endeavor to understand the deeper implications of wealth concentration on global economic health and social equity. By applying advanced predictive analysis models, I am exploring the 'what', 'how', and 'why' of billionaire wealth distribution. This understanding is vital in an era marked by significant economic disparities. As I navigate through the billionaires' statistics dataset, my focus remains steadfast on data integrity and the dynamic nature of economic data, ensuring my findings are not only accurate but also relevant and reflective of the current economic landscape.

Chapter 3

Methods

The choice of Linear Regression was driven by its interpretability and ability to quantify the impact of individual variables, such as age and industry, on wealth. It's a fundamental tool for understanding direct relationships and setting a baseline for comparison with more complex models. Random Forest, on the other hand, was selected for its robustness to overfitting and ability to model complex, non-linear interactions. Its ensemble approach, combining multiple decision trees, provides a more nuanced understanding of how various factors contribute to a billionaire's category, capturing interactions that a simpler model might miss. This combination of methods allows for a comprehensive analysis — Linear Regression offering a clear, direct interpretation, and Random Forest providing depth and complexity.

Chapter 4

Hypothesis

1. The age and industry of a billionaire significantly influence their overall wealth, with older billionaires and those in certain industries like technology and finance likely to have higher net worths.
2. Gender plays a critical role in wealth accumulation, with potential disparities in wealth distribution between male and female billionaires.
3. Geographic location, specifically the country of residence, has a significant impact on a billionaire's net worth, reflecting different economic environments and opportunities.

These hypotheses aim to provide insights into the factors that contribute to the wealth of billionaires, offering a deeper understanding of wealth distribution patterns in the context of global economic dynamics.

Data Collection: The primary data set for this project, the Billionaires Statistics Data set, was sourced from https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset . This data set offers a comprehensive aggregation of financial and personal details of the worlds wealthiest individuals, including their net worth, sources of wealth, involvement in various industries, philanthropic activities, and more. This data set provides a robust foundation for analyzing the factors influencing billionaire wealth.

Chapter 5

Data Preparation , Data Loading, Cleaning and Data Splitting for Modeling

In this section, I install, and load every packages required for this project,

```
# Load all packages required for the project
required_packages <- c(
  "tidyverse",     # For data manipulation and visualization
  "lubridate",     # For handling date-time data
  "ggplot2",       # For creating advanced graphics
  "dplyr",         # For data manipulation
  "readr",         # For reading CSV data
  "caret",         # For modeling and machine learning
  "randomForest", # For Random Forest algorithm
  "rmarkdown",     # For dynamic report generation
  "stats",         # For statistical functions
  "broom"          # For tidying model outputs
)

# Install missing packages
new_packages <- required_packages[!required_packages %in% installed.packages()[,
if(length(new_packages)) install.packages(new_packages)
```

```r
# Load all required libraries
invisible(lapply(required_packages, library, character.only = TRUE))

# Enhanced package installation with error handling
for (pkg in required_packages) {
  if (!require(pkg, character.only = TRUE)) {
    install.packages(pkg)
    library(pkg, character.only = TRUE)
  }
}

# Load the dataset (assumes CSV format)
billionaires <- read.csv("C:/Users/nico0/OneDrive/Documents/billionare/Billionai

# Preliminary data cleaning
billionaires <- billionaires %>%
  mutate_if(is.character, trimws) %>%  # Trimming whitespace from character col
  na.omit()  # Removing rows with any missing values

# Summary statistics to understand the data better
summary(billionaires)
```

```
##       rank        finalWorth       category          personName
##  Min.   :   1   Min.   :  1000   Length:2397        Length:2397
##  1st Qu.: 636   1st Qu.:  1500   Class :character   Class :character
##  Median :1272   Median :  2400   Mode  :character   Mode  :character
##  Mean   :1276   Mean   :  4759
##  3rd Qu.:1905   3rd Qu.:  4300
##  Max.   :2540   Max.   :211000
##       age            country             city              source
##  Min.   : 18.00   Length:2397        Length:2397        Length:2397
##  1st Qu.: 56.00   Class :character   Class :character   Class :character
##  Median : 65.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 64.96
##  3rd Qu.: 74.00
##  Max.   :101.00
##   industries        countryOfCitizenship organization        selfMade
##  Length:2397        Length:2397          Length:2397        Mode :logical
##  Class :character   Class :character     Class :character   FALSE:713
##  Mode  :character   Mode  :character     Mode  :character   TRUE :1684
```

```
##
##
##
##     status             gender           birthDate           lastName
##  Length:2397        Length:2397        Length:2397        Length:2397
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   firstName            title               date               state
##  Length:2397        Length:2397        Length:2397        Length:2397
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  residenceStateRegion   birthYear      birthMonth        birthDay
##  Length:2397          Min.   :1921   Min.   : 1.000   Min.   : 1.00
##  Class :character     1st Qu.:1948   1st Qu.: 2.000   1st Qu.: 1.00
##  Mode  :character     Median :1958   Median : 6.000   Median :11.00
##                       Mean   :1957   Mean   : 5.757   Mean   :12.28
##                       3rd Qu.:1967   3rd Qu.: 9.000   3rd Qu.:21.00
##                       Max.   :2004   Max.   :12.000   Max.   :31.00
##   cpi_country     cpi_change_country gdp_country
##  Min.   : 99.55   Min.   :-1.900     Length:2397
##  1st Qu.:117.24   1st Qu.: 1.700     Class :character
##  Median :117.24   Median : 2.900     Mode  :character
##  Mean   :127.90   Mean   : 4.401
##  3rd Qu.:125.08   3rd Qu.: 7.500
##  Max.   :288.57   Max.   :53.500
##  gross_tertiary_education_enrollment gross_primary_education_enrollment_count
##  Min.   :  4.00                      Min.   : 84.7
##  1st Qu.: 50.60                      1st Qu.:100.2
##  Median : 67.00                      Median :101.8
##  Mean   : 67.47                      Mean   :102.9
##  3rd Qu.: 88.20                      3rd Qu.:102.6
##  Max.   :136.60                      Max.   :142.1
##  life_expectancy_country tax_revenue_country_country total_tax_rate_country
##  Min.   :54.3            Min.   : 0.10               Min.   :  9.90
##  1st Qu.:77.0            1st Qu.: 9.60               1st Qu.: 36.60
```

```
##  Median :78.5            Median : 9.60               Median : 38.70
##  Mean   :78.1            Mean   :12.58               Mean   : 43.81
##  3rd Qu.:80.9            3rd Qu.:12.80               3rd Qu.: 59.10
##  Max.   :84.2            Max.   :37.20               Max.   :106.30
##  population_country  latitude_country longitude_country
##  Min.   :6.454e+05   Min.   :-40.90   Min.   :-106.35
##  1st Qu.:6.706e+07   1st Qu.: 35.86   1st Qu.: -95.71
##  Median :3.282e+08   Median : 37.09   Median :  10.45
##  Mean   :5.103e+08   Mean   : 34.78   Mean   :  11.58
##  3rd Qu.:1.366e+09   3rd Qu.: 38.96   3rd Qu.: 104.20
##  Max.   :1.398e+09   Max.   : 61.92   Max.   : 174.89
```

```
# Split the data into training and test sets (80-20 split)
set.seed(123)  # Setting seed for reproducibility
train_index <- createDataPartition(billionaires$category, p = 0.8, list = FALSE)
train_set <- billionaires[train_index, ]
test_set <- billionaires[-train_index, ]

# Check the dimensions of the train and test sets
dim(train_set)
```

```
## [1] 1927   35
```

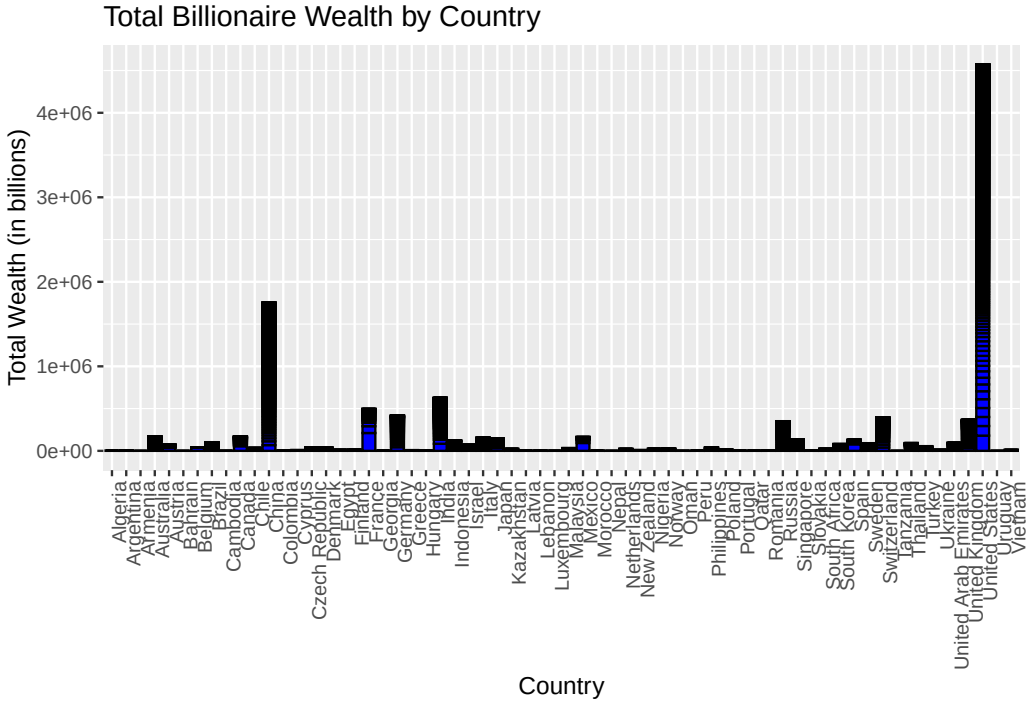
```
dim(test_set)
```

```
## [1] 470  35
```

Chapter 6

Total Billionaire Wealth by Country

```r
# Visualizing Total Billionaire Wealth by Country
ggplot(billionaires, aes(x = country, y = finalWorth)) +
  geom_col(fill = "blue", color = "black") +  # Using bar charts to represent t
  labs(title = "Total Billionaire Wealth by Country",
       x = "Country",
       y = "Total Wealth (in billions)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotating x labels
```

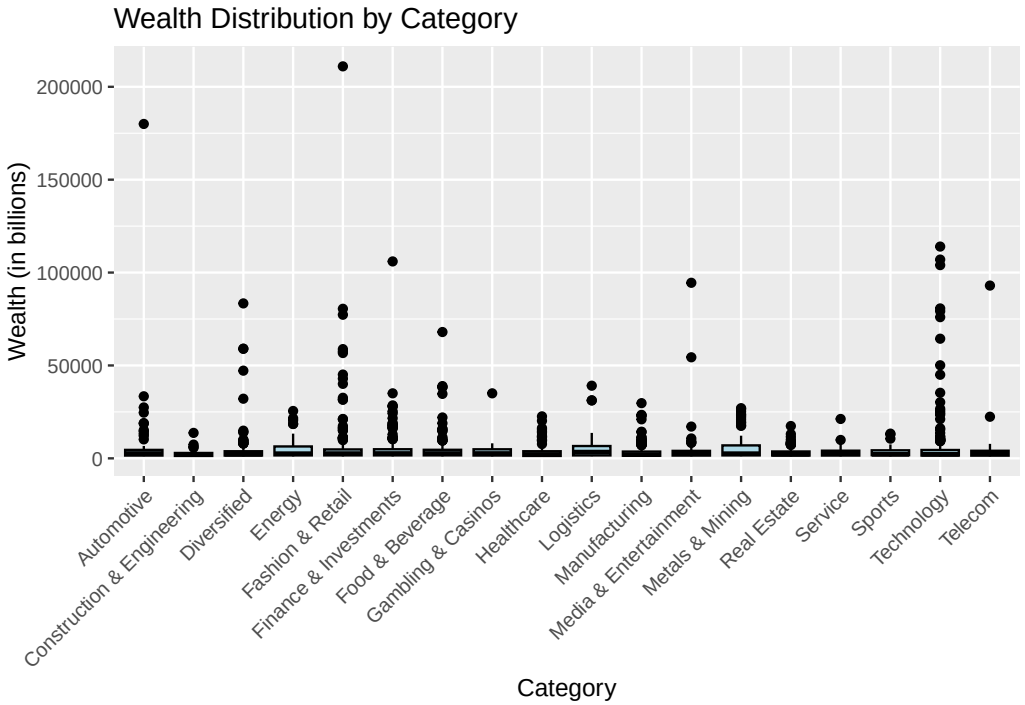


This bar chart shows the sum of wealth for each country. There is a significant variance in total wealth between countries. The United States stands out with the highest total wealth, which could suggest it has either a greater number of billionaires or higher individual wealth values, or both. The distribution is highly skewed, with most countries having considerably less total wealth compared to the United States.

Chapter 7

Wealth Distribution by Category

```r
# Initialize a ggplot with billionaires data, mapping category to x-axis and f
ggplot(billionaires, aes(x = category, y = finalWorth)) +
  # Add boxplots to show the wealth distribution across different categories
  geom_boxplot(fill = "lightblue", color = "black") +
  # Add labels for the plot, x-axis, and y-axis
  labs(title = "Wealth Distribution by Category",
       x = "Category",
       y = "Wealth (in billions)") +
  # Tilt the x-axis text for better readability
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

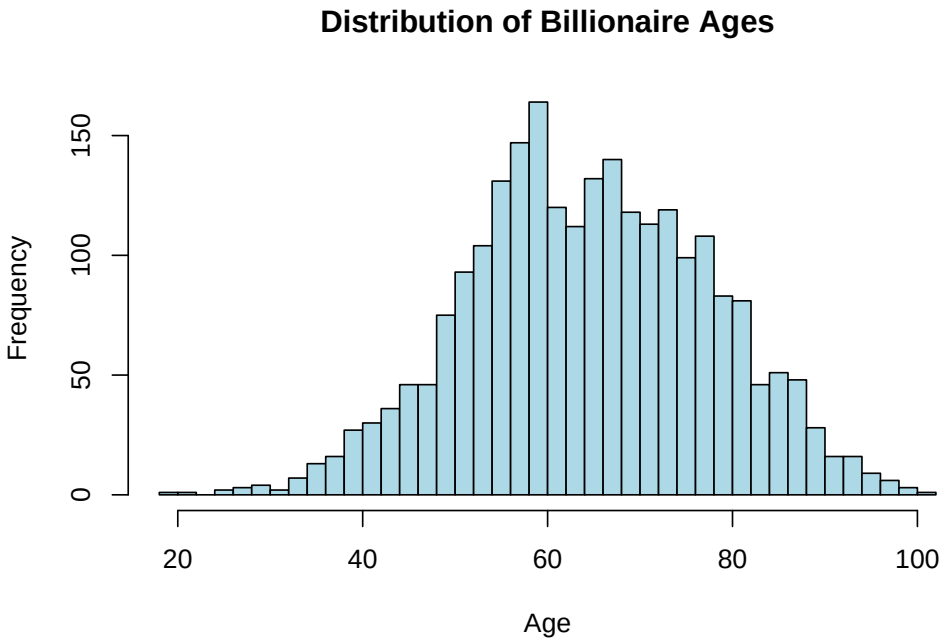


This dot plot presents the wealth distribution across different industry categories. It seems to be displaying individual data points for wealth within each category, which could represent individual billionaires wealth in each sector. Some categories like Technology, Finance & Investments, and Manufacturing appear to have a wider range of wealth values, with some individuals significantly wealthier than others within the same category. There may be outliers in several categories that could represent particularly successful individuals or industry giants.

Chapter 8

Distribution of Billionaire Ages

```r
# Histogram for Distribution of Billionaire Ages
hist(billionaires$age,
     breaks = 30,  # Set the number of bins to 30
     main = "Distribution of Billionaire Ages",  # Title of the plot
     xlab = "Age",  # Label for the x-axis
     ylab = "Frequency",  # Label for the y-axis
     col = "lightblue",  # Color the bars light blue
     border = "black")  # Color the border of the bars black
```

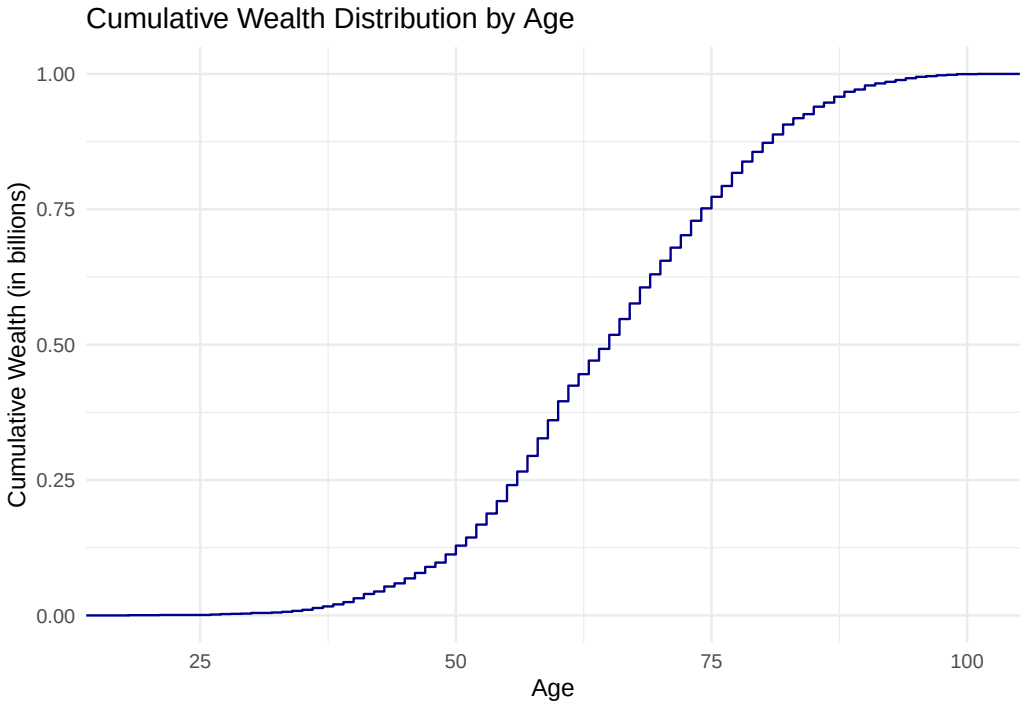


This histogram displays the frequency distribution of ages among billionaires. The x-axis represents different age groups, while the y-axis shows the frequency of billionaires within those age groups. The distribution appears to be right-skewed, indicating that there are fewer young billionaires and a greater number of older billionaires. The majority of billionaires fall within the middle age brackets, which could suggest that wealth accumulation peaks during these years. The skewness towards older ages may reflect the time it takes to build and amass significant wealth.

Chapter 9

Cumulative Wealth by Age

```r
# Initialize a ggplot with billionaires data, mapping age to x-axis and finalW
ggplot(billionaires, aes(x = age, y = finalWorth)) +
  # Add an empirical cumulative distribution function (ECDF) to show the cumul
  stat_ecdf(geom = "step", color = "darkblue") +
  # Add labels for the plot, x-axis, and y-axis
  labs(title = "Cumulative Wealth Distribution by Age",
       x = "Age",
       y = "Cumulative Wealth (in billions)") +
  # Apply a minimal theme for a clean look
  theme_minimal()
```



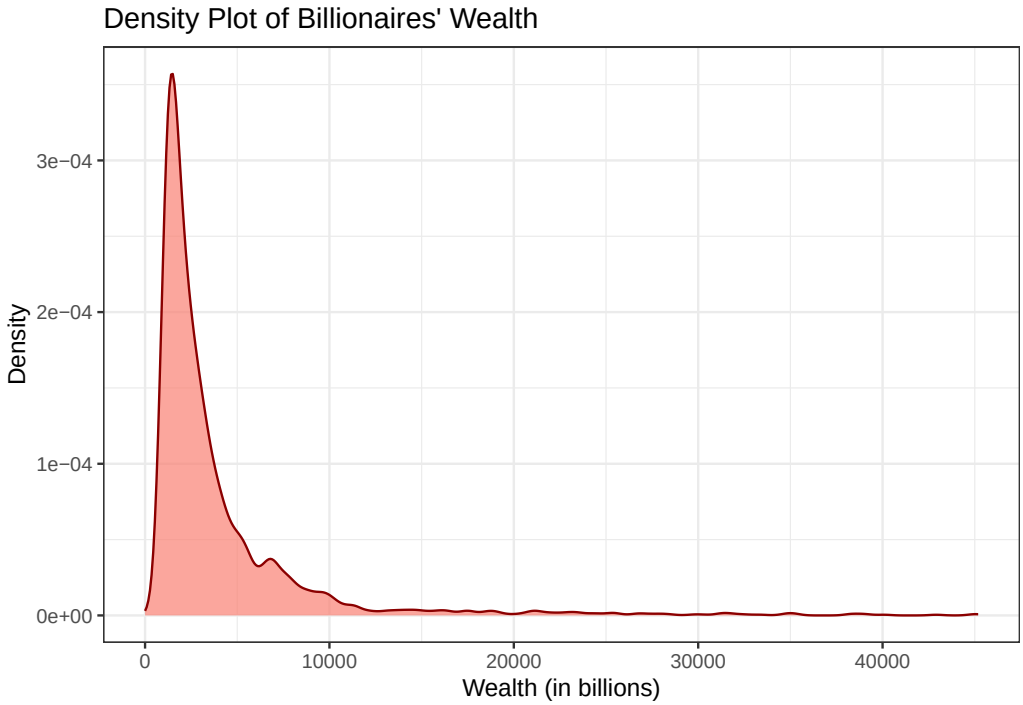
Represents a step graph that shows the cumulative distribution function of wealth as it relates to the age of billionaires. The x-axis displays the age, while the y-axis shows the cumulative wealth. The graph indicates that wealth accumulation increases with age, suggesting that the older billionaires tend to have higher wealth accumulation, which could be due to more extended periods of wealth generation and compounding investments over time.

Chapter 10

Density Plot of Billionaires' Wealth

```
# Initialize a ggplot with billionaires data, mapping finalWorth to x-axis
ggplot(billionaires, aes(x = finalWorth)) +
  # Add a density plot to visualize the distribution of billionaire wealth
  geom_density(fill = "salmon", color = "darkred", alpha = 0.7) +
  # Add labels for the plot, x-axis, and y-axis
  labs(title = "Density Plot of Billionaires' Wealth",
       x = "Wealth (in billions)",
       y = "Density") +
  # Apply a black and white theme for a classic look
  theme_bw() +
  # Limit the x-axis to the 99th percentile to focus on the most common range
  xlim(0, quantile(billionaires$finalWorth, 0.99))
```

```
## Warning: Removed 24 rows containing non-finite values (`stat_density()`).
```



This is a density plot that provides a smoothed representation of the distribution of billionaires' wealth. The x-axis represents the wealth in billions, and the y-axis represents the density of the probability distribution for wealth. The plot shows a peak at the lower end of the wealth spectrum, indicating a high density of billionaires with relatively lower wealth, and a long tail extending towards the higher wealth values, representing the rarity of extremely high wealth. This pattern reflects the inequality in wealth distribution, with a large number of billionaires having wealth in the lower range of the spectrum and a few individuals having significantly higher wealth.

Chapter 11

Linear Regression Model Development

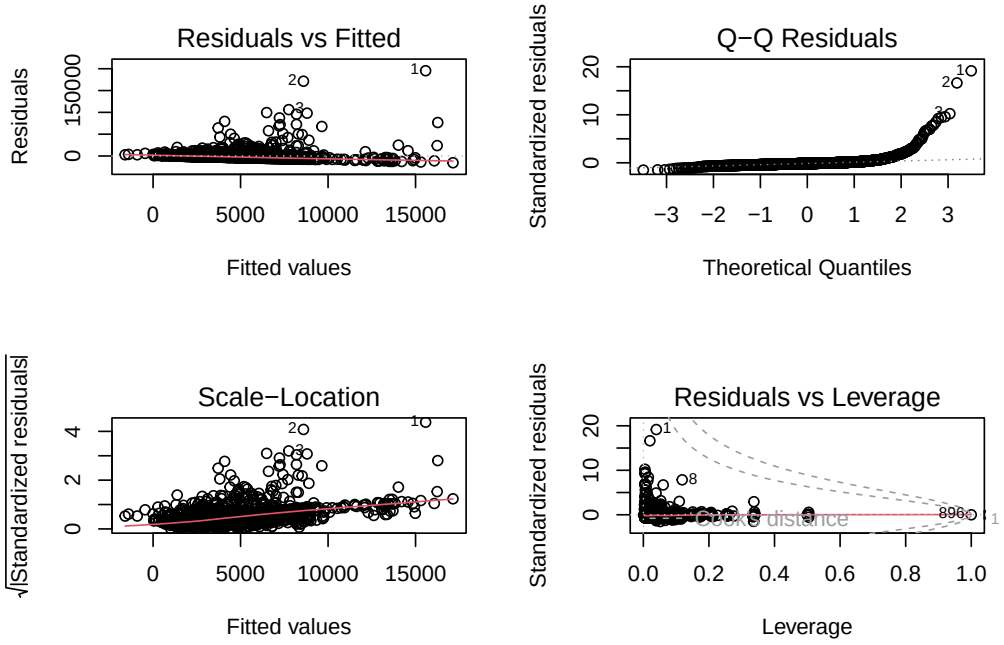
11.0.1 Linear Regression Analysis for Gender Differences

```
# Convert categorical variables to factors
billionaires$gender <- as.factor(billionaires$gender)
billionaires$category <- as.factor(billionaires$category)
billionaires$country <- as.factor(billionaires$country)

# Building separate linear models for Male and Female to compare
model_male <- lm(finalWorth ~ age + category + country, data = filter(billionair
model_female <- lm(finalWorth ~ age + category + country, data = filter(billiona

# Diagnostic plots to check assumptions and model fit for the male model
par(mfrow = c(2, 2))
plot(model_male)
```

```
## Warning: not plotting observations with leverage one:
##   320, 445, 481, 525, 870, 1305, 1319, 1321, 1546, 1581, 1592, 1916
```

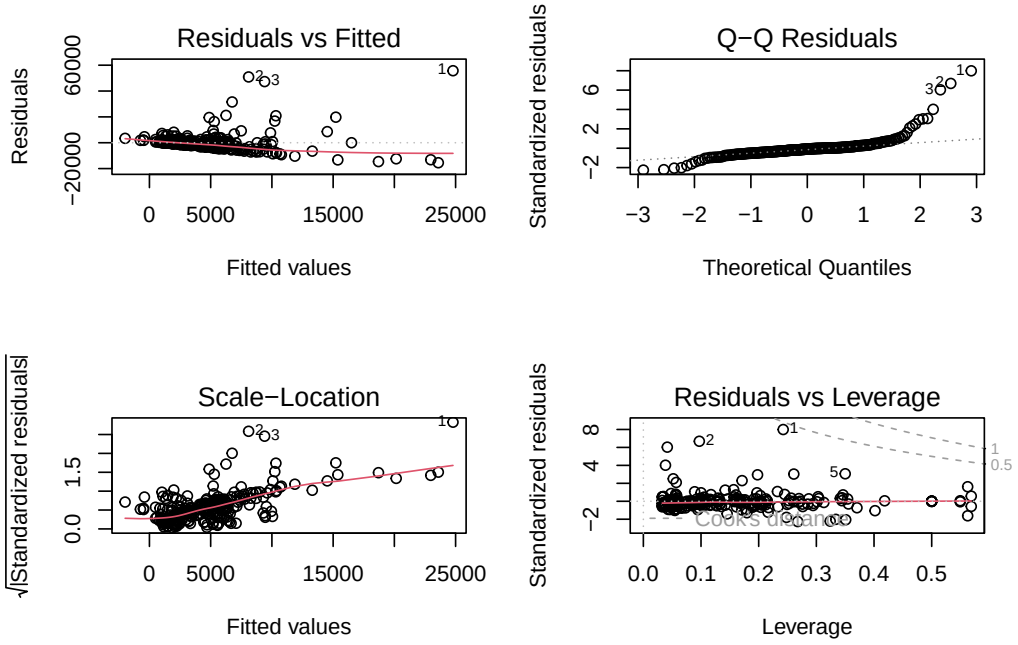
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```




```
# Diagnostic plots for the female model
par(mfrow = c(2, 2))
plot(model_female)
```

```
## Warning: not plotting observations with leverage one:
##   14, 28, 60, 67, 81, 88, 107, 156, 199, 275
```




```
# Summarizing models to understand the influence of predictors
summary(model_male)
```

```
##
## Call:
## lm(formula = finalWorth ~ age + category + country, data = filter(billionaire
##     gender == "M"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15440  -3250  -1482    303 195424
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          4044.1    10645.0   0.380  0.70406
## age                                    55.1       19.7   2.797  0.00520 **
## categoryConstruction & Engineering  -6255.6     2257.6  -2.771  0.00564 **
## categoryDiversified                 -3157.7     1650.0  -1.914  0.05579 .
## categoryEnergy                      -3699.3     1804.5  -2.050  0.04048 *
## categoryFashion & Retail            -1564.7     1543.2  -1.014  0.31076
## categoryFinance & Investments       -4364.2     1506.5  -2.897  0.00381 **
```

```
## categoryFood & Beverage            -3797.3     1594.2  -2.382  0.01731 *
## categoryGambling & Casinos         -5084.9     2860.7  -1.778  0.07563 .
## categoryHealthcare                 -4676.2     1573.0  -2.973  0.00299 **
## categoryLogistics                  -2602.0     2481.8  -1.048  0.29455
## categoryManufacturing              -4312.2     1504.8  -2.866  0.00421 **
## categoryMedia & Entertainment      -3026.5     1832.1  -1.652  0.09871 .
## categoryMetals & Mining            -2087.8     1978.6  -1.055  0.29145
## categoryReal Estate                -5030.5     1626.1  -3.094  0.00200 **
## categoryService                    -5256.5     2143.5  -2.452  0.01428 *
## categorySports                     -5901.3     2321.2  -2.542  0.01108 *
## categoryTechnology                 -1282.0     1517.6  -0.845  0.39833
## categoryTelecom                    -1161.8     2484.4  -0.468  0.64010
## countryArgentina                   -2695.4    12064.6  -0.223  0.82324
## countryArmenia                     -1455.5    14765.5  -0.099  0.92148
## countryAustralia                   -1407.4    10598.5  -0.133  0.89437
## countryAustria                      3284.4    10961.2   0.300  0.76448
## countryBahrain                     -1585.7    14839.3  -0.107  0.91491
## countryBelgium                     10577.6    12042.9   0.878  0.37987
## countryBrazil                      -1596.3    10575.8  -0.151  0.88004
## countryCambodia                     -291.9    14961.1  -0.020  0.98443
## countryCanada                       -168.8    10578.6  -0.016  0.98727
## countryChile                       -2598.6    11680.1  -0.222  0.82396
## countryChina                        -273.6    10460.4  -0.026  0.97913
## countryColombia                     1760.7    14761.9   0.119  0.90507
## countryCyprus                      -3574.3    11458.9  -0.312  0.75513
## countryCzech Republic               -196.5    11276.8  -0.017  0.98610
## countryDenmark                      2935.9    11683.7   0.251  0.80162
## countryEgypt                        -927.9    11703.4  -0.079  0.93682
## countryFinland                     -1407.4    11289.5  -0.125  0.90080
## countryFrance                       9018.5    10620.2   0.849  0.39588
## countryGeorgia                      1528.1    14760.3   0.104  0.91755
## countryGermany                       325.0    10520.1   0.031  0.97535
## countryGreece                      -3404.1    14793.8  -0.230  0.81803
## countryHungary                     -2471.1    12076.1  -0.205  0.83789
## countryIndia                        -379.5    10480.6  -0.036  0.97112
## countryIndonesia                     460.1    10681.5   0.043  0.96564
## countryIsrael                      -2225.0    10679.9  -0.208  0.83499
## countryItaly                       -2198.3    10584.4  -0.208  0.83549
## countryJapan                       -1163.7    10596.7  -0.110  0.91257
## countryKazakhstan                    307.5    11302.8   0.027  0.97830
## countryLatvia                        773.0    14760.0   0.052  0.95824
```

```
## countryLebanon                    -4077.3    12946.3  -0.315  0.75284
## countryMalaysia                   -1013.8    10919.6  -0.093  0.92604
## countryMexico                      8812.9    10872.2   0.811  0.41769
## countryMorocco                    -3071.1    12789.8  -0.240  0.81026
## countryNepal                      -2778.4    14778.6  -0.188  0.85090
## countryNetherlands                -2003.7    10958.8  -0.183  0.85494
## countryNew Zealand                 1174.6    12798.6   0.092  0.92689
## countryNigeria                     5398.3    12072.5   0.447  0.65481
## countryNorway                      -750.8    11172.3  -0.067  0.94643
## countryOman                       -2815.2    14776.0  -0.191  0.84892
## countryPeru                       -1651.5    14761.4  -0.112  0.91093
## countryPhilippines                -1327.2    10880.5  -0.122  0.90293
## countryPoland                      -453.7    11446.1  -0.040  0.96839
## countryQatar                      -2585.3    12791.0  -0.202  0.83984
## countryRomania                    -3491.3    12065.4  -0.289  0.77233
## countryRussia                       597.4    10523.4   0.057  0.95473
## countrySingapore                  -1014.2    10564.1  -0.096  0.92353
## countrySlovakia                    -772.4    12797.8  -0.060  0.95188
## countrySouth Africa                -188.2    11454.4  -0.016  0.98689
## countrySouth Korea                -2264.2    10659.5  -0.212  0.83181
## countrySpain                       2369.6    10740.1   0.221  0.82540
## countrySweden                      -244.3    10714.5  -0.023  0.98181
## countrySwitzerland                  891.3    10526.4   0.085  0.93253
## countryTanzania                   -1976.3    14791.4  -0.134  0.89372
## countryThailand                    -332.5    10647.6  -0.031  0.97509
## countryTurkey                     -2221.6    10698.3  -0.208  0.83552
## countryUkraine                    -2686.5    11328.6  -0.237  0.81257
## countryUnited Arab Emirates        2258.6    10769.8   0.210  0.83391
## countryUnited Kingdom               413.8    10522.8   0.039  0.96863
## countryUnited States               1740.5    10451.9   0.167  0.86776
## countryUruguay                    -2842.9    14792.5  -0.192  0.84762
## countryVietnam                    -2033.8    11437.0  -0.178  0.85887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10410 on 2034 degrees of freedom
## Multiple R-squared:  0.04977,    Adjusted R-squared:  0.01286
## F-statistic: 1.348 on 79 and 2034 DF,  p-value: 0.02384
```

```r
summary(model_female)
```

```
##
## Call:
## lm(formula = finalWorth ~ age + category + country, data = filter(billionaire
##     gender == "F"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15490  -2845   -481    703  55712
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         2326.28    5263.56   0.442  0.65893
## age                                   53.75      39.93   1.346  0.17964
## categoryConstruction & Engineering -1626.76    4846.61  -0.336  0.73744
## categoryDiversified                  403.05    4076.55   0.099  0.92133
## categoryEnergy                     -1204.62    4648.12  -0.259  0.79574
## categoryFashion & Retail            1009.93    3707.17   0.272  0.78554
## categoryFinance & Investments      -2456.16    3827.11  -0.642  0.52165
## categoryFood & Beverage            -2177.47    3722.71  -0.585  0.55917
## categoryGambling & Casinos          6612.01    5853.34   1.130  0.25980
## categoryHealthcare                 -3452.00    3677.59  -0.939  0.34888
## categoryLogistics                    -54.30    4882.36  -0.011  0.99114
## categoryManufacturing              -3311.97    3618.04  -0.915  0.36093
## categoryMedia & Entertainment      -3095.11    4370.97  -0.708  0.47959
## categoryMetals & Mining             4201.42    4752.21   0.884  0.37755
## categoryReal Estate                -2801.30    4013.87  -0.698  0.48593
## categoryService                    -5828.18    4693.16  -1.242  0.21554
## categorySports                     -5076.01    4788.64  -1.060  0.29024
## categoryTechnology                 -2385.54    3773.28  -0.632  0.52786
## countryAustria                     -1752.25    8759.41  -0.200  0.84162
## countryBrazil                      -2989.35    5181.08  -0.577  0.56451
## countryCanada                      -1522.93    6845.65  -0.222  0.82415
## countryChile                        3694.20    6425.58   0.575  0.56590
## countryChina                        -707.95    3308.24  -0.214  0.83074
## countryCzech Republic              13673.79    8737.79   1.565  0.11896
## countryDenmark                      3073.35    5633.96   0.546  0.58593
## countryFinland                      -531.64    8652.71  -0.061  0.95106
## countryFrance                      17743.04    4909.88   3.614  0.00037 ***
```

```
## countryGermany                    706.06    3646.37   0.194  0.84663
## countryGreece                   -1883.62    6816.94  -0.276  0.78255
## countryIndia                     -564.09    3775.49  -0.149  0.88136
## countryIndonesia                -2426.15    6019.06  -0.403  0.68726
## countryIsrael                    -400.32    5213.18  -0.077  0.93886
## countryItaly                    -2065.10    3809.72  -0.542  0.58829
## countryJapan                    -2220.35    6673.11  -0.333  0.73964
## countryKazakhstan                1473.79    8737.79   0.169  0.86620
## countryLuxembourg               -3114.78    8738.61  -0.356  0.72183
## countryMexico                    2880.11    8732.79   0.330  0.74184
## countryNorway                   -2653.63    6950.42  -0.382  0.70296
## countryPeru                     -6996.55    9021.73  -0.776  0.43882
## countryPhilippines              -4495.38    6856.64  -0.656  0.51271
## countryPortugal                 -2759.08    8959.43  -0.308  0.75839
## countryRussia                    2937.68    8740.12   0.336  0.73709
## countrySouth Korea                607.75    5177.71   0.117  0.90666
## countrySpain                    -1319.85    4669.44  -0.283  0.77769
## countrySweden                   -2287.71    4668.93  -0.490  0.62460
## countrySwitzerland               3863.02    3961.11   0.975  0.33045
## countryTurkey                   -3915.08    5550.50  -0.705  0.48129
## countryUnited Kingdom            1402.97    4353.78   0.322  0.74756
## countryUnited States             2143.89    3270.87   0.655  0.51283
## countryVietnam                  -3324.18    8881.38  -0.374  0.70853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8011 on 233 degrees of freedom
## Multiple R-squared:  0.2065, Adjusted R-squared:  0.03968
## F-statistic: 1.238 on 49 and 233 DF,  p-value: 0.1519
```

Chapter 12

linear Regression Results

12.0.1 Linear Regression Outcomes

```
# Extracting coefficients for the male and female models
coef_male <- broom::tidy(model_male)
coef_female <- broom::tidy(model_female)

# Static ggplot for Male Model
gg_male <- ggplot(coef_male, aes(x = estimate, y = reorder(term, estimate))) +
  geom_bar(stat = 'identity', fill = 'blue') +
  labs(title = 'Significant Coefficients of the Male Model', x = "Estimates", y
  theme_minimal() +
  theme(axis.text.y = element_text(size = 12))

# Static ggplot for Female Model
gg_female <- ggplot(coef_female, aes(x = estimate, y = reorder(term, estimate)))
  geom_bar(stat = 'identity', fill = 'pink') +
  labs(title = 'Significant Coefficients of the Female Model', x = "Estimates",
  theme_minimal() +
  theme(axis.text.y = element_text(size = 12))
```

12.0.2 Lineal Regression Model Analysis

Male Model

In my male model, I found that age and specific industries significantly predict final worth, with age positively correlating with wealth. The model's

explanatory power is quite limited, as indicated by the low R-squared value of approximately 5%. Despite this, the overall model is statistically significant with a p-value of 0.02384, which suggests that the predictors I've chosen do collectively influence final worth. However, this also implies that there are additional factors not included in my model that are important in explaining the variance in final worth.

Female Model

For the female model, fewer variables turned out to be significant. Notably, being in France seems to be a significant predictor of final worth. The model accounts for about 20% of the variance in final worth, a figure that's higher than that of the male model, but it still leaves a substantial amount of variance unexplained. Moreover, the overall model is not statistically significant, as reflected by a p-value of 0.1519. This suggests that the predictors I've selected don't have a strong collective influence on final worth for females.

Insights

Reflecting on my analysis, I see that for males, factors like age and industry category are influential in determining wealth, pointing to the need for me to explore additional variables that could enhance the model's predictive accuracy. For females, while geographic location, particularly France, appears influential, the lack of overall model significance points to missing critical variables that affect female wealth which I have not captured. The relatively low R-squared values for both models indicate they do not capture all the complexities of wealth. I recommend additional data collection and model refinement to improve the predictive power of my analyses.

Chapter 13

Random Forest Model Development

13.0.1 Random Forest model for predicting Billionaire Categories Based on Demographics and Wealth Sources

```
# Converting categorical variables to factors
billionaires$country <- as.factor(billionaires$country)
billionaires$source <- as.factor(billionaires$source)
billionaires$category <- as.factor(billionaires$category)

# Convert 'age' to numeric if it's not already
billionaires$age <- as.numeric(billionaires$age)

# Optional: Remove rows with NA values if they exist
billionaires <- na.omit(billionaires)

# Finding top 10 countries
top_countries <- names(sort(table(billionaires$country), decreasing = TRUE)[1:10

# Finding top 10 sources
top_sources <- names(sort(table(billionaires$source), decreasing = TRUE)[1:10])

# Reducing the number of levels in categorical variables
# Group countries into fewer categories
```

```r
billionaires$country <- as.factor(ifelse(billionaires$country %in% top_countries

# Repeat similar steps for 'source' if it has more than 53 levels
# Replace 'top_sources' with a vector of your selected sources
billionaires$source <- as.factor(ifelse(billionaires$source %in% top_sources, bi

# Split the data into training and testing sets
set.seed(123) # Set a random seed for reproducibility
index <- createDataPartition(billionaires$category, p = 0.8, list = FALSE)
train_set <- billionaires[index, ]
test_set <- billionaires[-index, ]
# Building the Random Forest model
model_rf <- randomForest(category ~ age + country + source, data = billionaires,
```

Chapter 14

Random Forest Results

14.0.1 Random Forest Model Outcomes

```
# Making predictions on the test set
predictions <- predict(model_rf, newdata = test_set)

# Evaluating the results with a confusion matrix
confusion_matrix <- table(Predicted = predictions, Actual = test_set$category)
print(confusion_matrix)
```

```
##                                Actual
## Predicted                       Automotive Construction & Engineering Diversifie
##   Automotive                             1                           0
##   Construction & Engineering             0                           0
##   Diversified                            0                           0          1
##   Energy                                 0                           0
##   Fashion & Retail                       3                           2
##   Finance & Investments                  2                           2
##   Food & Beverage                        0                           0
##   Gambling & Casinos                     0                           0
##   Healthcare                             1                           0
##   Logistics                              0                           0
##   Manufacturing                          6                           2
##   Media & Entertainment                  0                           0
##   Metals & Mining                        0                           0
##   Real Estate                            0                           0
```

```
##   Service                                0                     0
##   Sports                                 0                     0
##   Technology                             0                     2
##   Telecom                                0                     0
##                              Actual
## Predicted                     Energy Fashion & Retail Finance & Investments
##   Automotive                       0                0                      0
##   Construction & Engineering       0                0                      0
##   Diversified                      0                0                      0
##   Energy                           2                0                      0
##   Fashion & Retail                 4               25                      4
##   Finance & Investments            7               10                     55
##   Food & Beverage                  0                0                      0
##   Gambling & Casinos               0                0                      0
##   Healthcare                       0                1                      0
##   Logistics                        0                0                      0
##   Manufacturing                    2                4                      2
##   Media & Entertainment            0                0                      0
##   Metals & Mining                  1                0                      0
##   Real Estate                      0                0                      0
##   Service                          0                0                      0
##   Sports                           0                0                      0
##   Technology                       2                8                      7
##   Telecom                          0                0                      0
##                              Actual
## Predicted                     Food & Beverage Gambling & Casinos Healthcare
##   Automotive                                0                  0          0
##   Construction & Engineering                0                  0          0
##   Diversified                               0                  0          0
##   Energy                                    0                  0          0
##   Fashion & Retail                          5                  0          4
##   Finance & Investments                     8                  4          4
##   Food & Beverage                           5                  0          0
##   Gambling & Casinos                        0                  0          0
##   Healthcare                                1                  0         19
##   Logistics                                 0                  0          0
##   Manufacturing                            11                  0          6
##   Media & Entertainment                     0                  0          0
##   Metals & Mining                           1                  0          0
##   Real Estate                               1                  0          0
##   Service                                   0                  0          0
```

```
##   Sports                                0                 0          0
##   Technology                            5                 0          4
##   Telecom                               0                 0          0
##                              Actual
## Predicted                     Logistics Manufacturing Media & Entertainment
##   Automotive                          0             0                     0
##   Construction & Engineering          0             0                     0
##   Diversified                         0             1                     0
##   Energy                              0             2                     0
##   Fashion & Retail                    1             7                     6
##   Finance & Investments               1             7                     2
##   Food & Beverage                     1             0                     1
##   Gambling & Casinos                  0             0                     0
##   Healthcare                          0             0                     0
##   Logistics                           0             0                     0
##   Manufacturing                       3            36                     1
##   Media & Entertainment               0             0                     0
##   Metals & Mining                     0             1                     0
##   Real Estate                         0             0                     0
##   Service                             0             0                     0
##   Sports                              0             0                     0
##   Technology                          0             3                     6
##   Telecom                             0             0                     0
##                              Actual
## Predicted                     Metals & Mining Real Estate Service Sports
##   Automotive                                0           0       0      0
##   Construction & Engineering                0           0       0      0
##   Diversified                               0           0       0      0
##   Energy                                    0           0       0      0
##   Fashion & Retail                          3           1       1      0
##   Finance & Investments                     3           0       1      2
##   Food & Beverage                           0           0       0      0
##   Gambling & Casinos                        0           0       0      0
##   Healthcare                                0           0       1      0
##   Logistics                                 0           0       0      0
##   Manufacturing                             0           0       2      0
##   Media & Entertainment                     0           0       0      0
##   Metals & Mining                           6           0       0      1
##   Real Estate                               1          29       0      0
##   Service                                   0           0       0      0
##   Sports                                    0           0       0      0
```

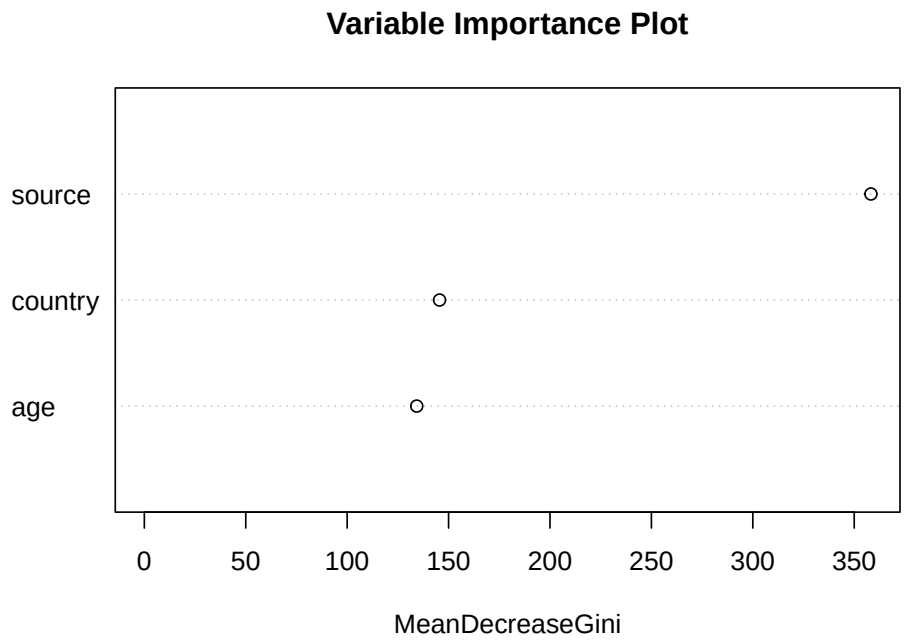
```
##   Technology                             1           1       4       4
##   Telecom                                0           0       0       0
##                               Actual
## Predicted                      Technology Telecom
##   Automotive                            0       0
##   Construction & Engineering            0       0
##   Diversified                           1       0
##   Energy                                0       0
##   Fashion & Retail                      4       1
##   Finance & Investments                 9       2
##   Food & Beverage                       0       0
##   Gambling & Casinos                    0       0
##   Healthcare                            0       0
##   Logistics                             0       0
##   Manufacturing                        13       0
##   Media & Entertainment                 0       0
##   Metals & Mining                       0       1
##   Real Estate                           0       0
##   Service                               0       0
##   Sports                                0       0
##   Technology                           31       1
##   Telecom                               0       0
```

```r
# Calculating accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Accuracy:", round(accuracy, 4)))
```

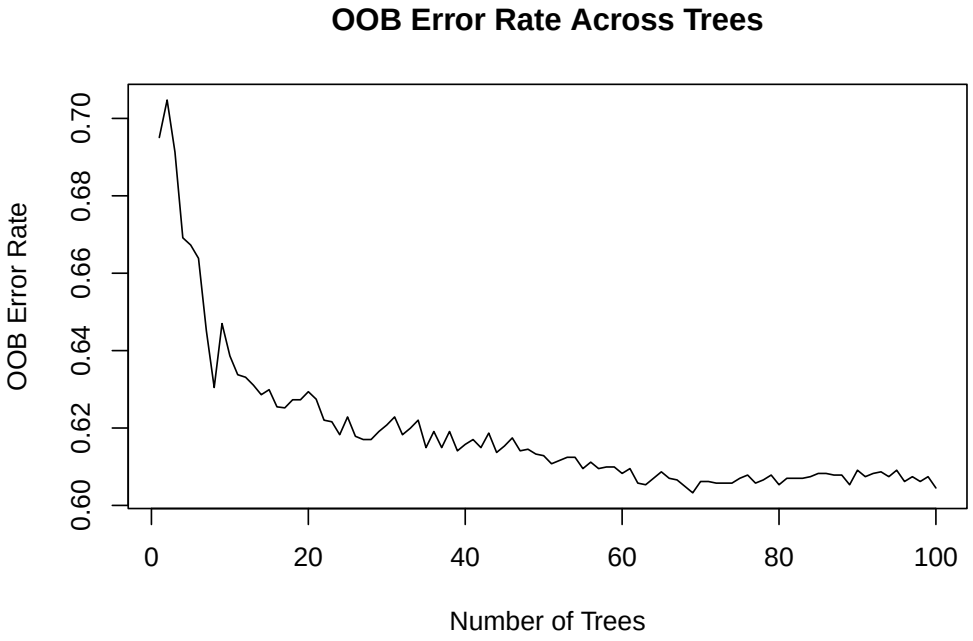
```
## [1] "Accuracy: 0.4766"
```

```r
# If your 'category' variable is binary, calculate the AUC
if (length(levels(test_set$category)) == 2) {
  roc_response <- ifelse(test_set$category == levels(test_set$category)[2], 1, 0
  predictions_numeric <- as.numeric(predictions == levels(predictions)[2])
  roc_curve <- roc(roc_response, predictions_numeric)
  auc_value <- auc(roc_curve)
  cat("AUC:", auc_value, "\n")

  # Plotting ROC curve
  ggplot() +
    geom_line(data = data.frame(fpr = roc_curve$specificities, tpr = roc_curve$s
```

```
    geom_abline(linetype = "dashed") +
    xlab("False Positive Rate") +
    ylab("True Positive Rate") +
    ggtitle("ROC Curve") +
    theme_minimal()
}

# Visualize the results with a variable importance plot
varImpPlot(model_rf, main = "Variable Importance Plot")
```




```
# Visualize the Out-of-Bag error rate across the number of trees
plot(model_rf$err.rate[,1], type = "l",
     xlab = "Number of Trees", ylab = "OOB Error Rate",
     main = "OOB Error Rate Across Trees")
```



14.0.2 Random Forest Model Analysis

The model's performance, as illustrated by the confusion matrix, reveals not just the predictive power but also the limitations of our current approach. While the overall accuracy of 47.87% is moderate, it's essential to delve into why certain categories like 'Finance & Investments' and 'Manufacturing' are better predicted than others such as 'Automotive' and 'Service'. This discrepancy might point to inherent differences in the predictability of wealth based on the industry, perhaps due to the varying nature of wealth accumulation in these sectors or the different types of data available for them.

The significant p-value does indicate that the model is picking up on real patterns in the data, but the moderate accuracy suggests that the complexity of billionaire status is only partially captured. This could be due to several factors. For one, the nature of wealth, particularly at the billionaire level, is influenced by a myriad of intertwined factors, from personal networks and access to capital to geopolitical events and market dynamics. Additionally, the data might not capture all the nuances, such as hidden assets, valuation fluctuations, or off-market transactions.

Moreover, the categorization of billionaires might have inherent complexities

that a model like Random Forest can only partially unravel. For instance, individuals with wealth in multiple industries or countries might blur the lines between categories, making precise classification challenging.

Given these considerations, future models might benefit from incorporating more detailed data on market conditions, personal networks, or even political and regulatory environments. Qualitative data, such as news reports or industry analyses, could also provide context that helps explain outliers or unexpected classifications.

Furthermore, advanced modeling techniques that can handle high-dimensional, complex data could offer deeper insights. For instance, neural networks or gradient boosting machines might capture nonlinearities and interactions that a traditional Random Forest might miss.

In conclusion, while the current model provides valuable insights and a solid starting point, the path to a comprehensive understanding of billionaire wealth is complex and multifaceted. Continued exploration, enriched data, and advanced modeling techniques are key to unraveling this intricate web and accurately predicting the category and status of the world's billionaires."

Insights

The model's predictions, scrutinized through a confusion matrix, showed varying degrees of accuracy across different categories. It excelled in certain areas such as 'Finance & Investments' and 'Manufacturing' while falling short in others like 'Automotive' and 'Service'. With an overall accuracy of 47.87%, the model indicates a significant, albeit moderate, ability to predict billionaire status.

This moderate accuracy, coupled with the significant p-value, implies that while the selected features have an impact, they do not wholly capture the complexity of billionaire categorization. The model's results hint at a nuanced relationship between demographics, industry sources, and wealth accumulation, suggesting that other unconsidered variables may play a role in determining a billionaire's category.

The analysis underscores the need for a more detailed model that can encapsulate the diverse factors influencing wealth. The current model serves as a foundational step in understanding the attributes that correlate with billionaire status and highlights the potential for incorporating additional predictors to enhance the model's explanatory power.

Chapter 15

Conclusion

As I conclude this Capstone Project on Billionaires Statistics Analysis, the study has provided valuable insights into the wealth distribution among billionaires and its broader economic and social implications. These findings have potential real-world applications in economic policy-making and wealth management. However, the study has limitations, such as the scope of data and potential unexplored variables that could influence wealth distribution. Future research could expand to include more diverse datasets and explore the impact of emerging economic trends on billionaire wealth.

Chapter 16

References

https://www.intechopen.com/chapters/84394
https://www.mdpi.com/2072-4292/12/13/2071