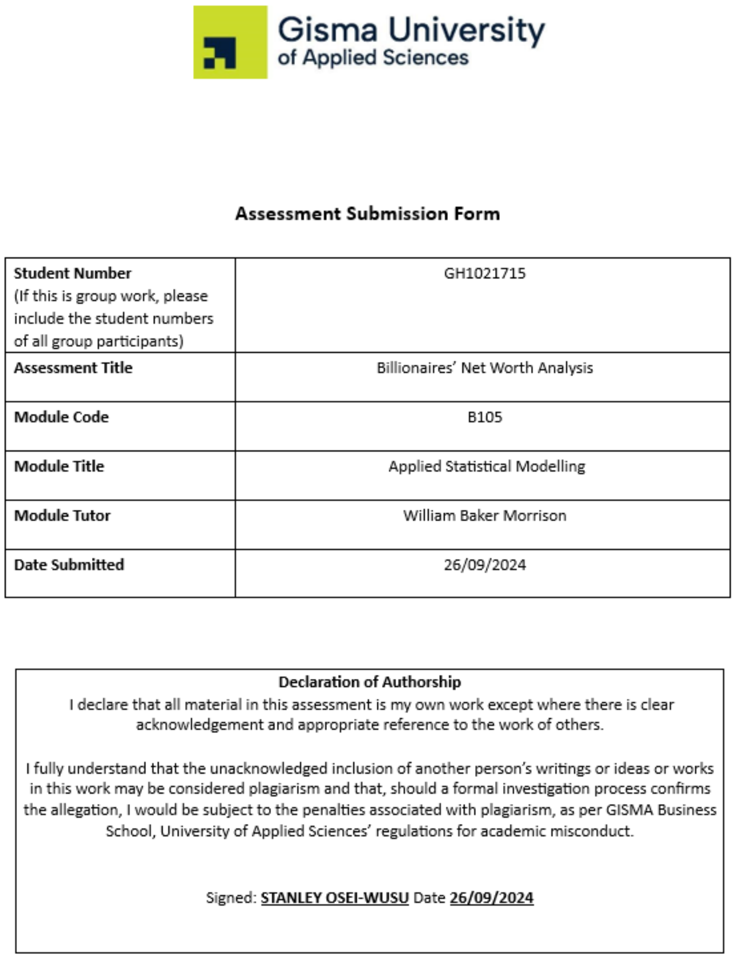


Assessment Submission Form

Student Number (If this is group work, please include the student numbers of all group participants)	GH1021715
Assessment Title	Billionaires' Net Worth Analysis
Module Code	B105
Module Title	Applied Statistical Modelling
Module Tutor	William Baker Morrison
Date Submitted	26/09/2024

Declaration of Authorship

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

I fully understand that the unacknowledged inclusion of another person's writings or ideas or works in this work may be considered plagiarism and that, should a formal investigation process confirms the allegation, I would be subject to the penalties associated with plagiarism, as per GISMA Business School, University of Applied Sciences' regulations for academic misconduct.

Signed: **STANLEY OSEI-WUSU** Date **26/09/2024**

Billionaires' Net Worth Analysis

Stanley Osei-Wusu (GH1021715)

September 26, 2024

Contents

1 Introduction

This report presents a comprehensive statistical analysis of the Billionaires' dataset, which contains information about 2,640 billionaires, including their net worth, age, gender, country of residence, and industry. The goal of this analysis is to answer key business questions, test specific hypotheses, and interpret statistical findings to provide actionable insights for stakeholders.

1.1 Key Business Questions

1. What are the key factors that determine the net worth of billionaires?
2. How do industry and age affect a billionaire's net worth?
3. Are there significant differences in net worth based on gender?
4. Can we predict whether a billionaire belongs to the technology industry based on their age, gender, and net worth?

2 Hypotheses

1. **H0 (Null Hypothesis)**: There is no significant relationship between age, industry, or gender and the net worth of billionaires.
2. **H1 (Alternative Hypothesis)**: There is a significant relationship between age, industry, or gender and the net worth of billionaires.

A second hypothesis for the logistic regression is:

- **H0**: A billionaire's membership in the technology industry is not significantly predicted by their age, gender, or net worth.
- **H1**: A billionaire's membership in the technology industry can be significantly predicted by their age, gender, and net worth.

3 Data Preparation

The data was sourced from the *Billionaires Statistics Dataset* on Kaggle, containing 2,640 records. Several steps were performed to ensure the data was clean and ready for analysis.

3.1 Data Cleaning

- Missing values in the `age` column were imputed using the median age.
- The `finalWorth` column was cleaned by removing non-numeric characters.
- Categorical variables such as `gender`, `industry`, and `country` were converted to factors.
- Rows with missing or zero values in critical variables such as `finalWorth` were removed.

3.2 Sampling

To ensure a balanced analysis, the full dataset was used without sampling, as the size of the dataset (2,640 records) was manageable for the analysis.

4 Exploratory Data Analysis

4.1 Descriptive Statistics

The dataset contains a wide range of billionaire net worths, with the minimum being $1 billion and the maximum reaching $211 billion. The mean net worth is approximately $4.6 billion, and the distribution is heavily skewed toward lower values.

```
Summary statistics:
Min. Net Worth: 1 billion
Max. Net Worth: 211 billion
Median Net Worth: 2.3 billion
Mean Net Worth: 4.6 billion
Age Range: 18 - 101 years
```

4.2 Distribution of Net Worth

The net worth distribution is right-skewed, with the majority of billionaires having a net worth below $10 billion, while a small number hold extremely high values.

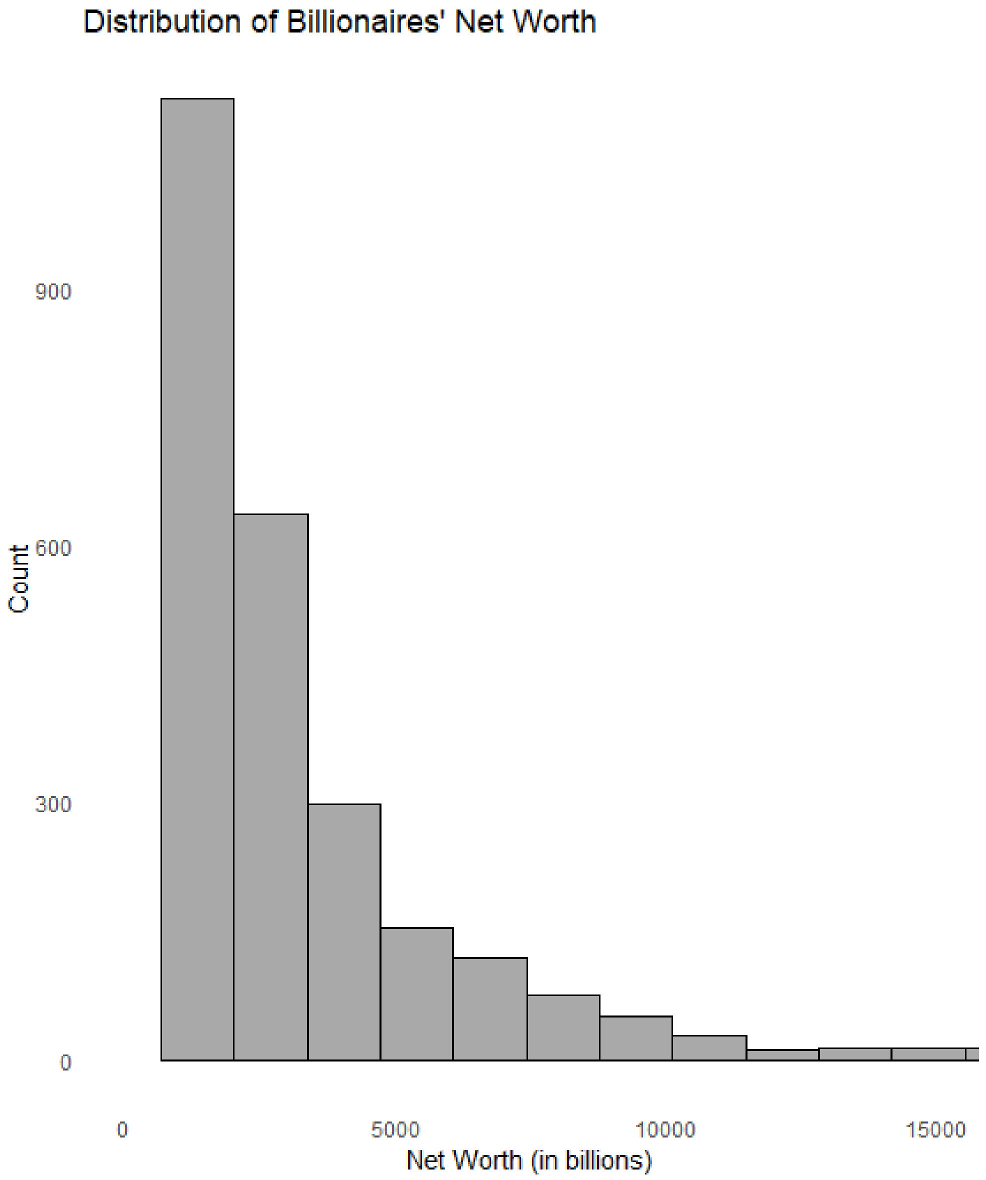


Figure 1: Distribution of Billionaires' Net Worth

4.3 Age vs Net Worth

A scatter plot of age vs. net worth reveals a weak positive correlation, with older individuals generally having more wealth.

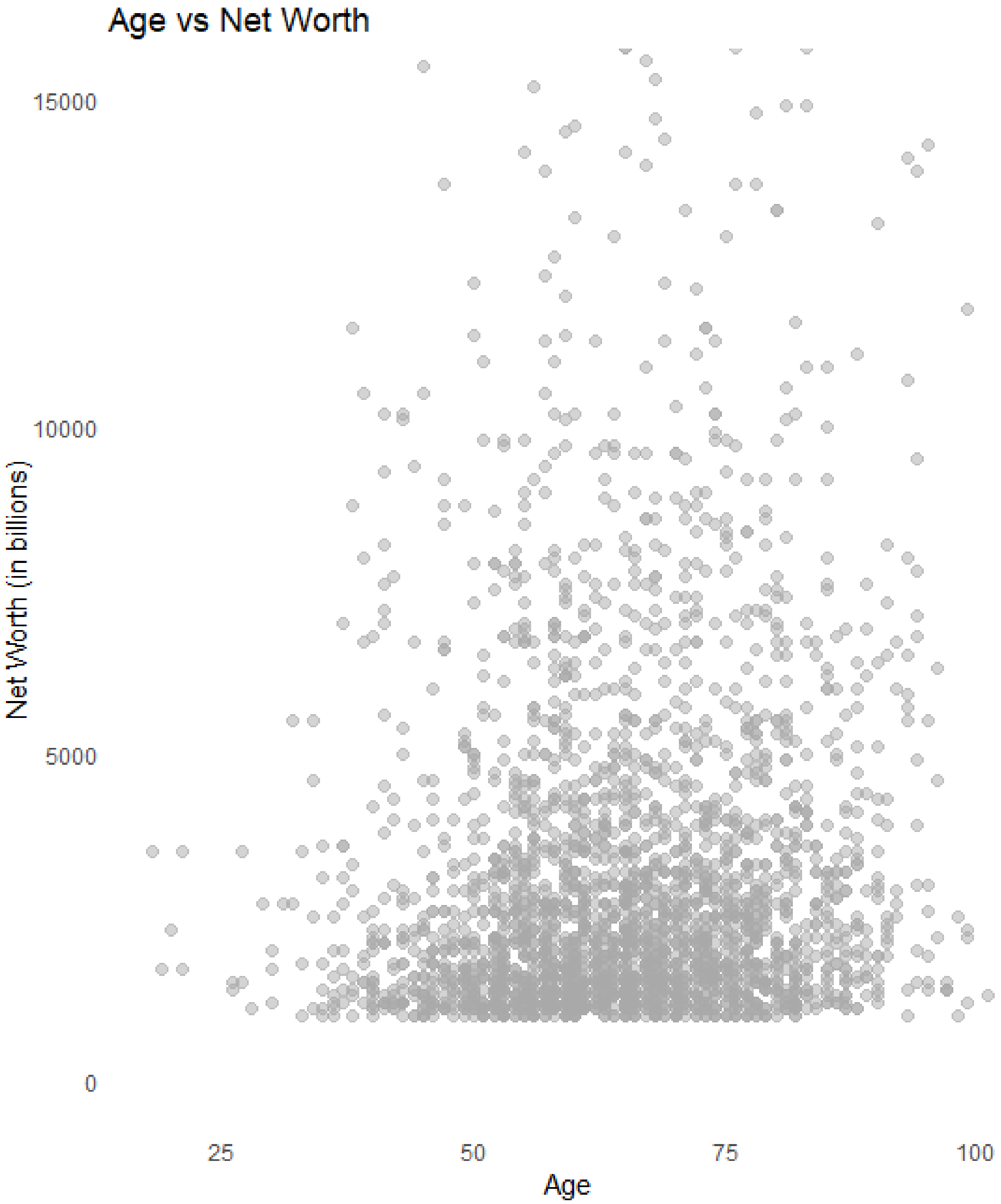


Figure 2: Age vs Net Worth

4.4 Net Worth by Industry

The boxplot below shows the distribution of net worth across different industries, highlighting that sectors like Energy and Finance tend to have a broader range of wealth.

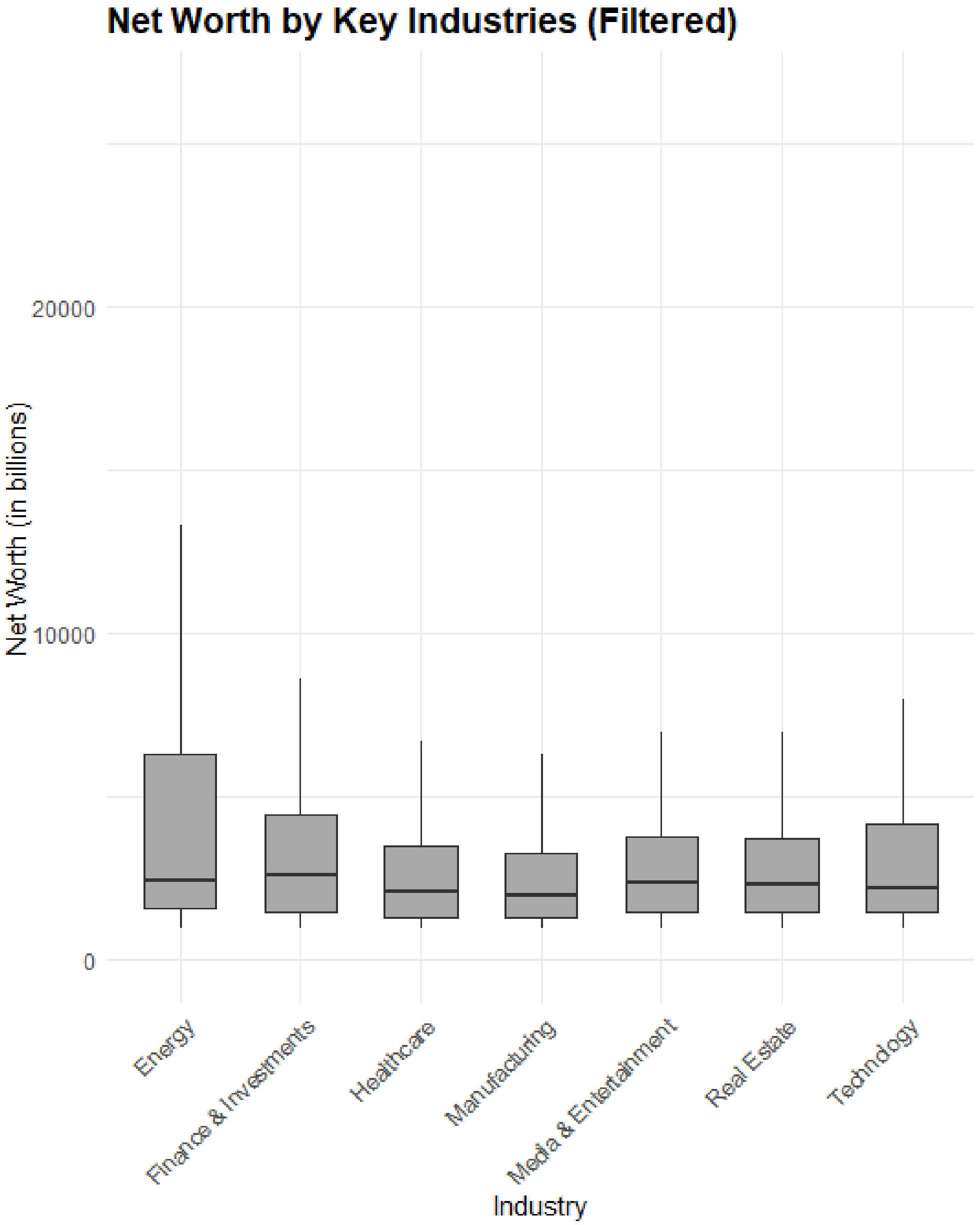


Figure 3: Net Worth by Key Industries (Filtered)

4.5 Average Net Worth by Gender

The bar chart indicates that male billionaires have a slightly higher average net worth compared to their female counterparts, but the difference is not statistically significant.

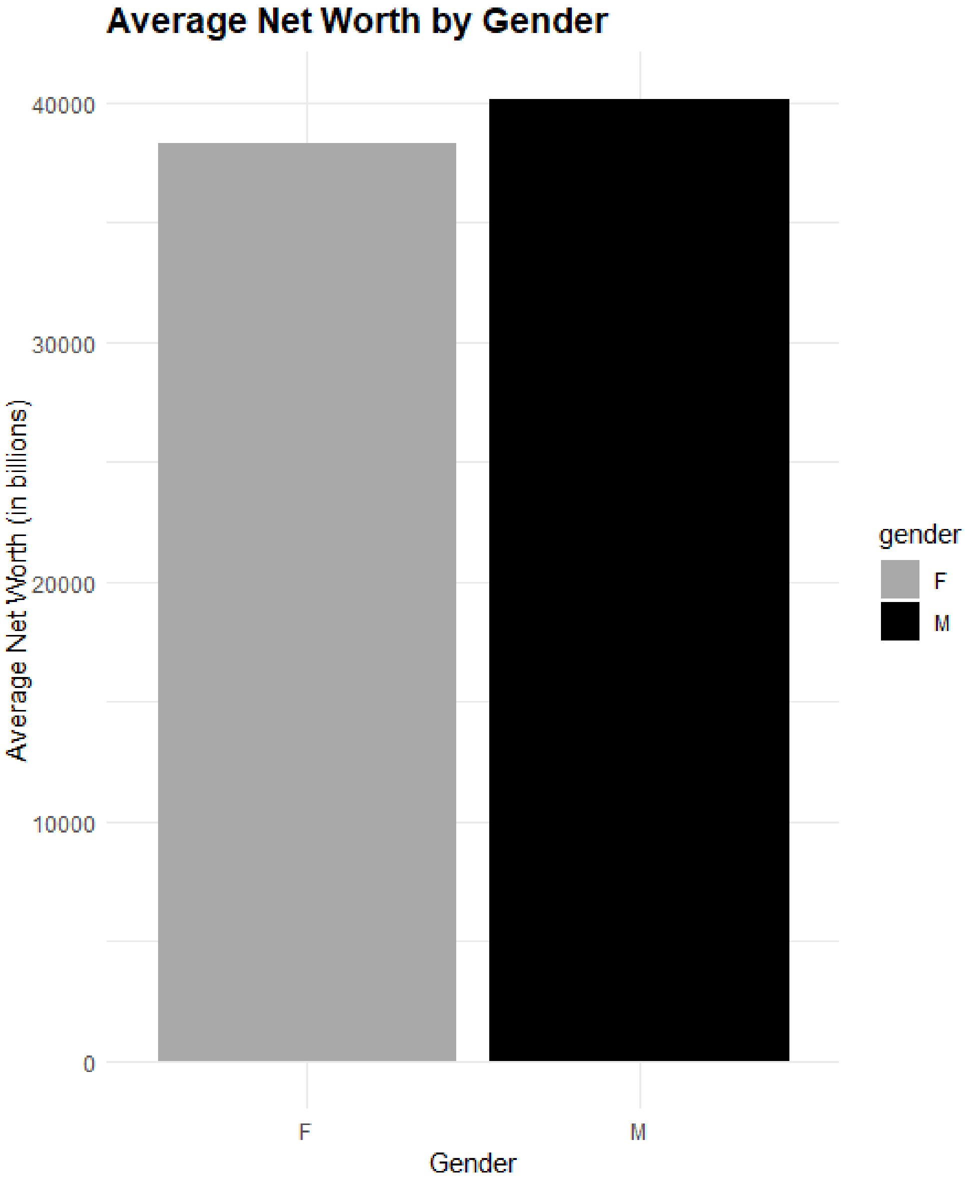


Figure 4: Average Net Worth by Gender

5 Inferential Statistics

5.1 Hypothesis Testing: T-test for Gender and Net Worth

A Welch Two Sample t-test was conducted to compare the mean net worth between male and female billionaires.

```
Welch Two Sample t-test
t = 0.77808, df = 422.13, p-value = 0.437
Mean in group F = 4023.952, Mean in group M = 3803.423
```

Since the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating no significant difference in net worth based on gender.

5.2 ANOVA for Industry

An ANOVA test was performed to determine if there are significant differences in net worth across industries.

```
ANOVA results:
F-value: 3.316, p-value: 4.87e-06
```

The p-value is significant, suggesting that industry has a statistically significant effect on net worth.

5.3 Logistic Regression

A logistic regression model was used to predict whether a billionaire belongs to the technology industry based on age, gender, and net worth.

```
Logistic Regression Results:
Age: p < 0.001, Gender: p = 0.027, Net Worth: p = 0.366 (not significant)
```

The results indicate that age (p ¡ 0.001) and gender (p = 0.027) are statistically significant predictors of technology industry membership, while net worth (p = 0.366) is not a significant predictor.

5.4 ROC Curve

5.4 ROC Curve

The ROC curve for the logistic regression model shows an AUC of approximately 0.7, indicating a moderately good fit for predicting membership in the technology industry.

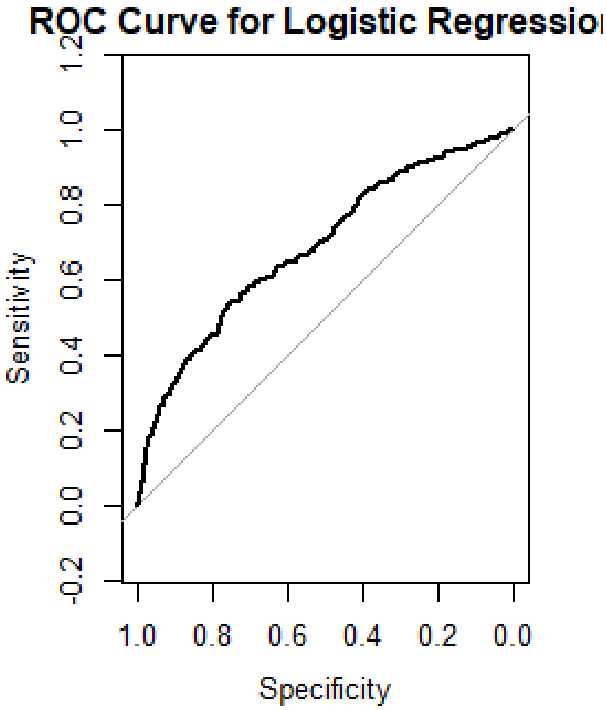


Figure 5: ROC Curve for Logistic Regression

6 Conclusion

The analysis provided several key insights:

- Age and industry significantly influence a billionaire's net worth.
- There is no significant difference in net worth between male and female billionaires.
- Logistic regression models indicate that age and gender are predictive of whether a billionaire belongs to the technology industry, but net worth is not.

6.1 Limitations

The analysis was limited by the lack of more granular data on billionaire behaviors or financial strategies. Additionally, there may be unobserved variables that impact wealth accumulation not included in this dataset.

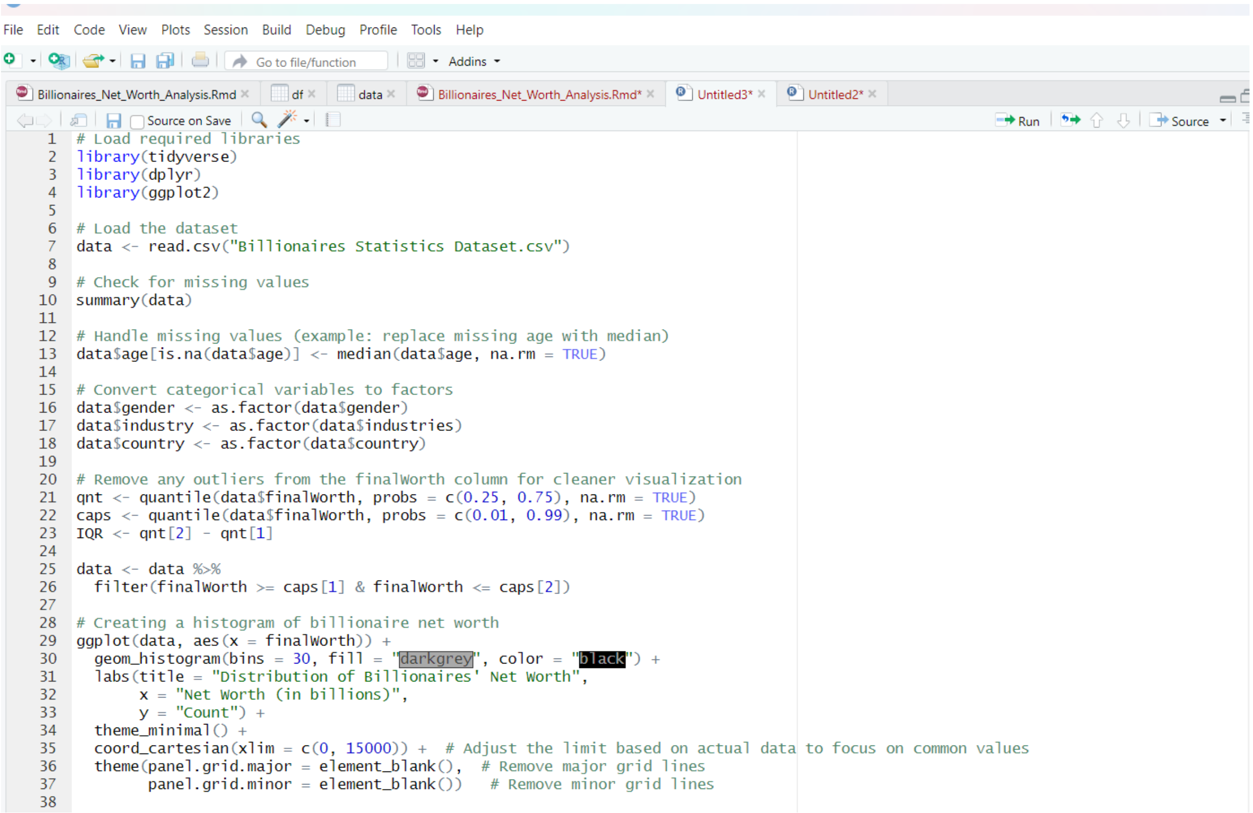
6.2 Future Work

Future analyses could explore non-linear models and include additional variables such as investment behavior, philanthropy, and inheritance patterns, which may provide deeper insights into wealth accumulation among billionaires.

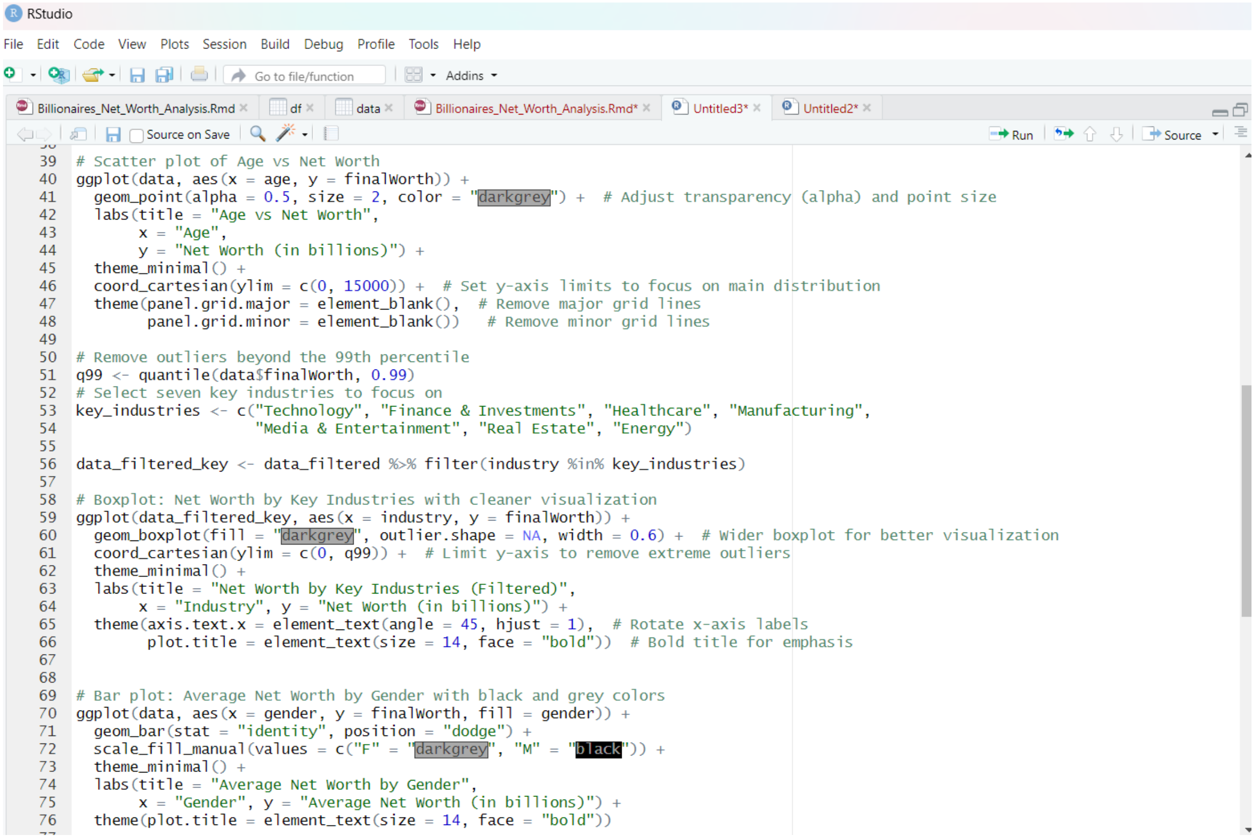
7 GitHub Repository

The full R code and scripts for this analysis are available on GitHub at the following link: https://github.com/stanleymay20/Bi

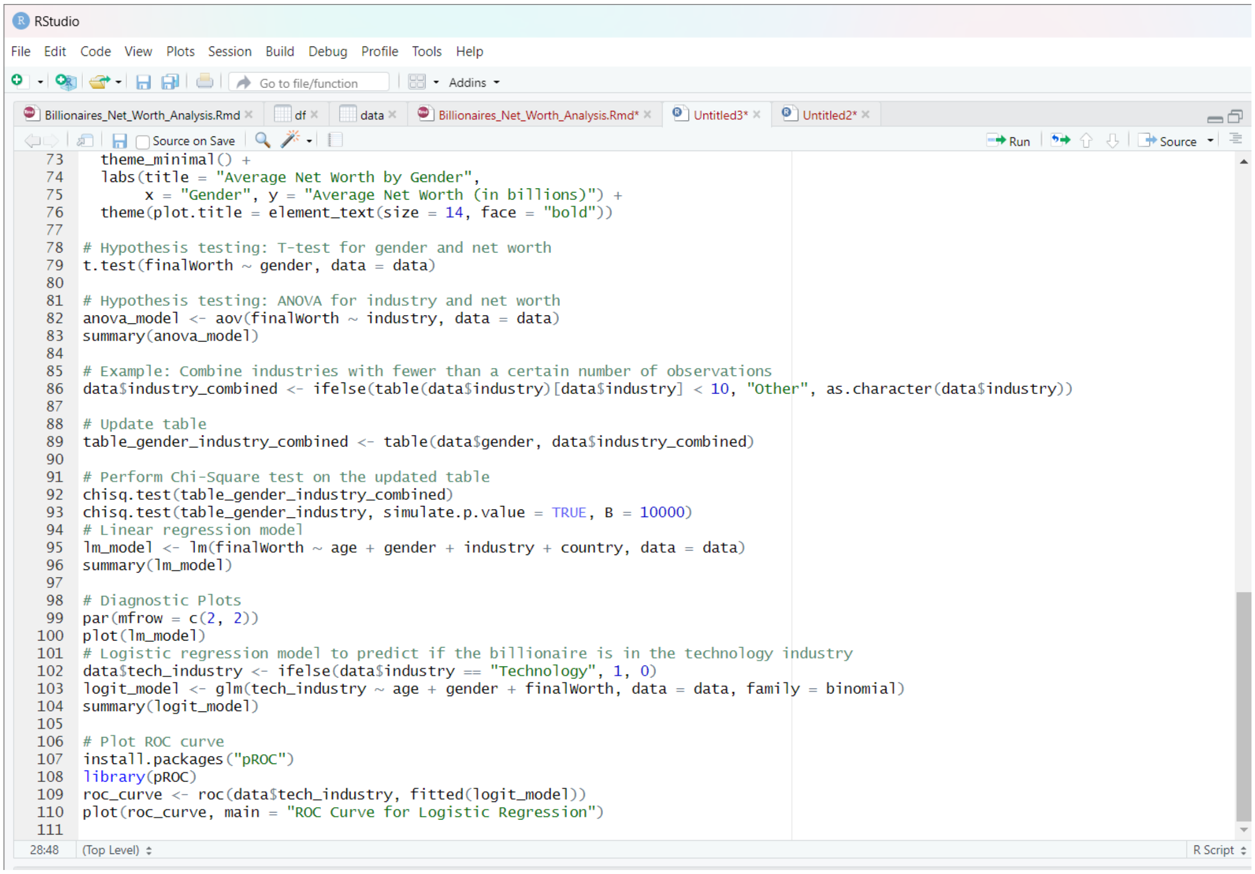
8 Screenshots of R coding



```r
# Load required libraries
library(tidyverse)
library(dplyr)
library(ggplot2)

# Load the dataset
data <- read.csv("Billionaires Statistics Dataset.csv")

# Check for missing values
summary(data)

# Handle missing values (example: replace missing age with median)
data$age[is.na(data$age)] <- median(data$age, na.rm = TRUE)

# Convert categorical variables to factors
data$gender <- as.factor(data$gender)
data$industry <- as.factor(data$industries)
data$country <- as.factor(data$country)

# Remove any outliers from the finalWorth column for cleaner visualization
qnt <- quantile(data$finalWorth, probs = c(0.25, 0.75), na.rm = TRUE)
caps <- quantile(data$finalWorth, probs = c(0.01, 0.99), na.rm = TRUE)
IQR <- qnt[2] - qnt[1]

data <- data %>%
  filter(finalWorth >= caps[1] & finalWorth <= caps[2])

# Creating a histogram of billionaire net worth
ggplot(data, aes(x = finalWorth)) +
  geom_histogram(bins = 30, fill = "darkgrey", color = "black") +
  labs(title = "Distribution of Billionaires' Net Worth",
       x = "Net Worth (in billions)",
       y = "Count") +
  theme_minimal() +
  coord_cartesian(xlim = c(0, 15000)) +  # Adjust the limit based on actual data to focus on common values
  theme(panel.grid.major = element_blank(),  # Remove major grid lines
        panel.grid.minor = element_blank())   # Remove minor grid lines

```



```r
# Scatter plot of Age vs Net Worth
ggplot(data, aes(x = age, y = finalWorth)) +
  geom_point(alpha = 0.5, size = 2, color = "darkgrey") +  # Adjust transparency (alpha) and point size
  labs(title = "Age vs Net Worth",
       x = "Age",
       y = "Net Worth (in billions)") +
  theme_minimal() +
  coord_cartesian(ylim = c(0, 15000)) +  # Set y-axis limits to focus on main distribution
  theme(panel.grid.major = element_blank(),  # Remove major grid lines
        panel.grid.minor = element_blank())  # Remove minor grid lines

# Remove outliers beyond the 99th percentile
q99 <- quantile(data$finalWorth, 0.99)
# Select seven key industries to focus on
key_industries <- c("Technology", "Finance & Investments", "Healthcare", "Manufacturing",
                    "Media & Entertainment", "Real Estate", "Energy")

data_filtered_key <- data_filtered %>% filter(industry %in% key_industries)

# Boxplot: Net Worth by Key Industries with cleaner visualization
ggplot(data_filtered_key, aes(x = industry, y = finalWorth)) +
  geom_boxplot(fill = "darkgrey", outlier.shape = NA, width = 0.6) +  # Wider boxplot for better visualization
  coord_cartesian(ylim = c(0, q99)) +  # Limit y-axis to remove extreme outliers
  theme_minimal() +
  labs(title = "Net Worth by Key Industries (Filtered)",
       x = "Industry", y = "Net Worth (in billions)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels
        plot.title = element_text(size = 14, face = "bold"))  # Bold title for emphasis


# Bar plot: Average Net Worth by Gender with black and grey colors
ggplot(data, aes(x = gender, y = finalWorth, fill = gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("F" = "darkgrey", "M" = "black")) +
  theme_minimal() +
  labs(title = "Average Net Worth by Gender",
       x = "Gender", y = "Average Net Worth (in billions)") +
  theme(plot.title = element_text(size = 14, face = "bold"))
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Billionaires_Net_Worth_Analysis.Rmd
df
data
Billionaires_Net_Worth_Analysis.Rmd*
Untitled3*
Untitled2*
Source on Save
Run
Source
28:48
(Top Level)
R Script

```r
  theme_minimal() +
  labs(title = "Average Net Worth by Gender",
       x = "Gender", y = "Average Net Worth (in billions)") +
  theme(plot.title = element_text(size = 14, face = "bold"))

# Hypothesis testing: T-test for gender and net worth
t.test(finalWorth ~ gender, data = data)

# Hypothesis testing: ANOVA for industry and net worth
anova_model <- aov(finalWorth ~ industry, data = data)
summary(anova_model)

# Example: Combine industries with fewer than a certain number of observations
data$industry_combined <- ifelse(table(data$industry)[data$industry] < 10, "Other", as.character(data$industry))

# Update table
table_gender_industry_combined <- table(data$gender, data$industry_combined)

# Perform Chi-Square test on the updated table
chisq.test(table_gender_industry_combined)
chisq.test(table_gender_industry, simulate.p.value = TRUE, B = 10000)
# Linear regression model
lm_model <- lm(finalWorth ~ age + gender + industry + country, data = data)
summary(lm_model)

# Diagnostic Plots
par(mfrow = c(2, 2))
plot(lm_model)
# Logistic regression model to predict if the billionaire is in the technology industry
data$tech_industry <- ifelse(data$industry == "Technology", 1, 0)
logit_model <- glm(tech_industry ~ age + gender + finalWorth, data = data, family = binomial)
summary(logit_model)

# Plot ROC curve
install.packages("pROC")
library(pROC)
roc_curve <- roc(data$tech_industry, fitted(logit_model))
plot(roc_curve, main = "ROC Curve for Logistic Regression")

```