AKINLEYE R. AFEEZ

MPS, Analytics

ALY 6000:

Introduction to Analytics

Dr. Mohsen Soltanifar

Date: Oct. 23rd 2023

INTRODUCTION

This project focuses on the analysis of Billionaire Statistics dataset chosen from the Kaggle.com website. This dataset contains statistics on the world's billionaires, including information about their businesses, industries, and personal details.

It provides insights into the wealth distribution, business sectors, and demographics of billionaires worldwide.

DATA CLEANSING

To get started, I installed the following packages into the R script to work with.
Janitor: A package for data cleaning and data frame tidying.

Naniar: A package that works with missing values in R.
Dplyr: A package for data manipulation and data wrangling in R.
Ggplot: Enables that plot charts in R

Then we started by working with:

Missing values (NAs)

Managing NAs. Correcting data types.

Manipulating strings

Reorganizing the data.

DATA CLEANSING Contd.

Replace missing values in the "Age" column with the mean age: Calculation was done to find the mean age of the age of billionaires excluding the missing values using the code:

- mean_age <- mean(b_data$age, na.rm = TRUE)print(mean_age)

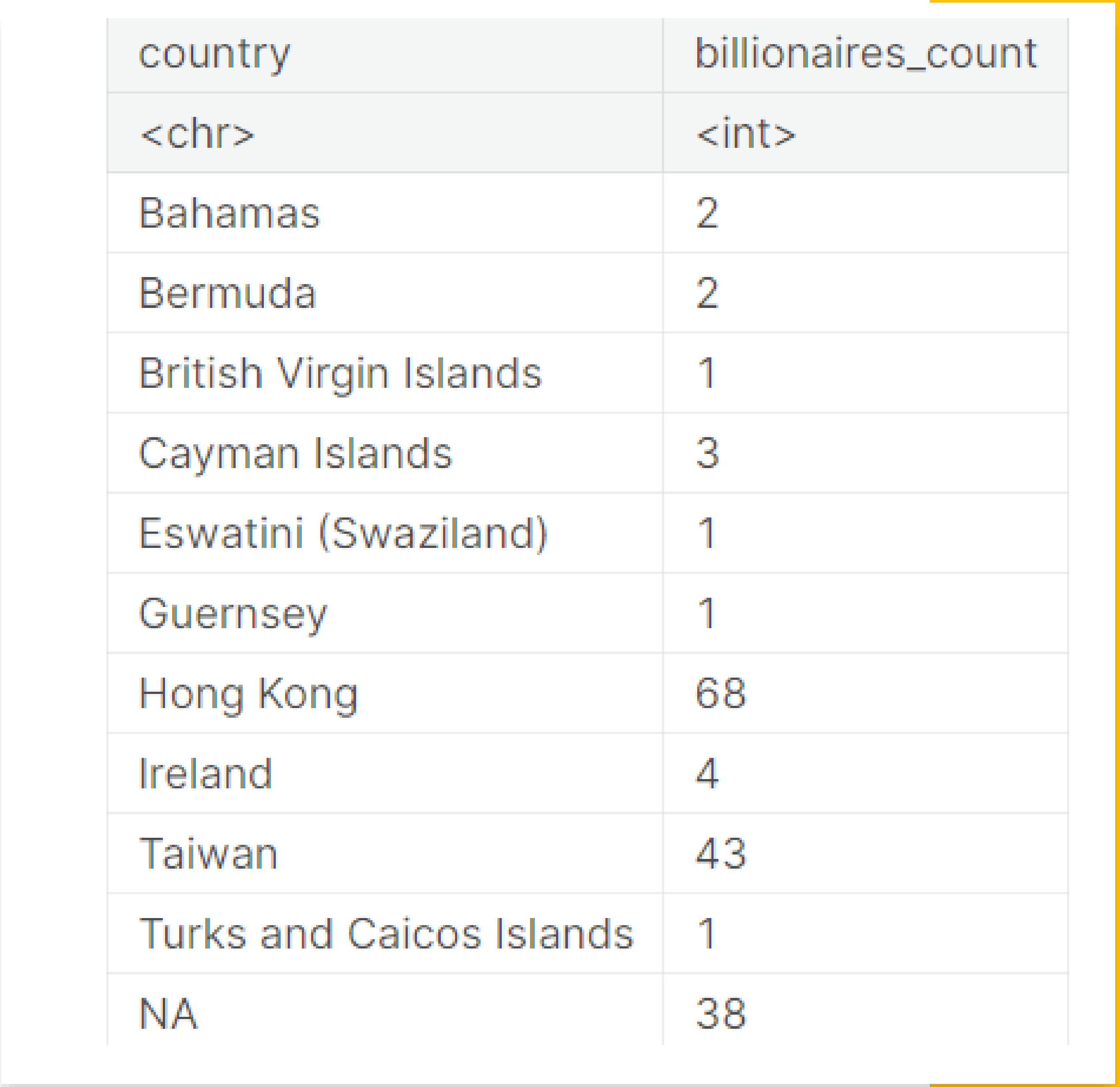
Then, replace the missing values with the mean age:

- b_data$age[is.na(b_data$age)] <- mean_age

Following actions was also made to standardize the dataset:

Remove Commas and Dollar Signs in the "gdp_country" country and convert to numeric type

Summarizing the count of billionaires in each country with missing GDP data:



country	billionaires_count
<chr>	<int>
Bahamas	2
Bermuda	2
British Virgin Islands	1
Cayman Islands	3
Eswatini (Swaziland)	1
Guernsey	1
Hong Kong	68
Ireland	4
Taiwan	43
Turks and Caicos Islands	1
NA	38

The figure shows the count and grouped of billionaires by country of origin with 38 billionaires whose origins are missing.

DATA CLEANSING Contd.

- We ignore these billionaires without origin and fill in the missing values with an external GDP file that accompanied our dataset.
- We fill in the missing values for known countries and will ignore 38 missing values that their countries are unknown.

```
b_data$gdp_country <- ifelse(is.na(b_data$gdp_country),
external_gdp_data$gdp_country[match(b_data$country,
external_gdp_data$country)],b_data$gdp_country)
```

RESEARCH QUESTIONS: 1

What is the distribution of the age of billionaires in 2023? Are there any notable age trends among the world's billionaires?

I approached this question by calculating the measure of central tendencies of our data.

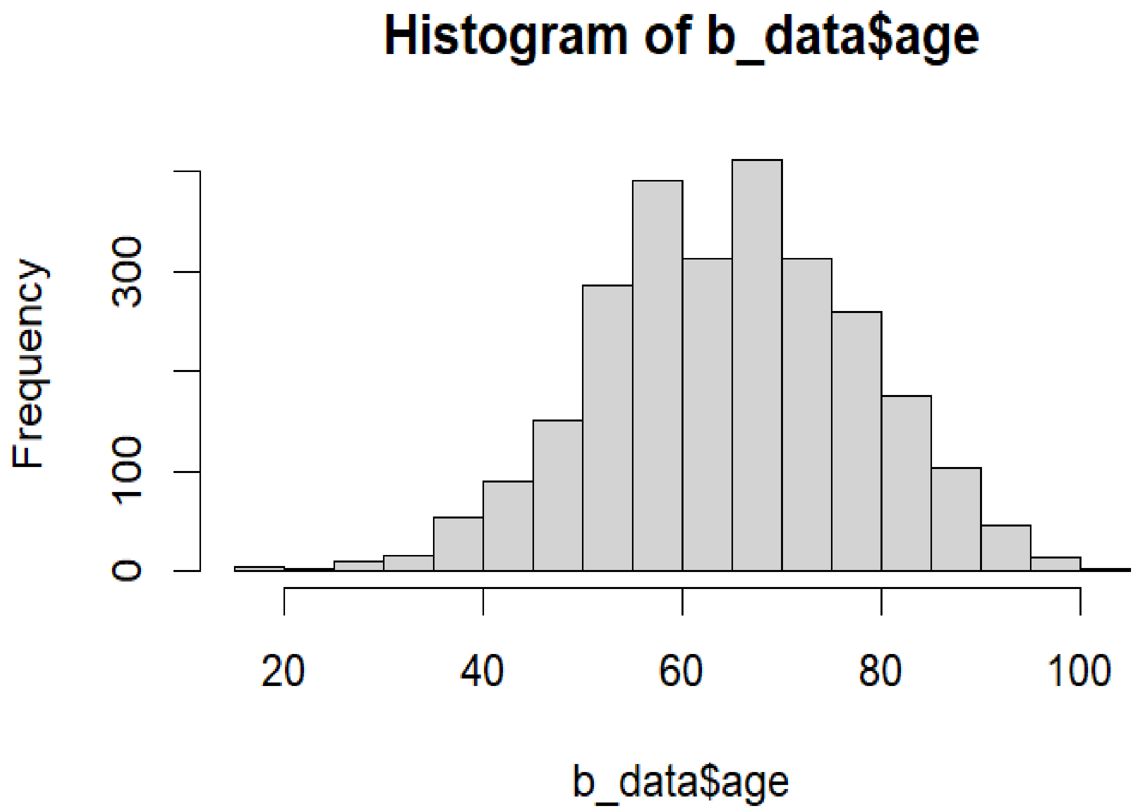
mean(b_data$age)

median(b_data$age)

sd(b_data$age)

IQR(b_data$age) - (InterQuartile Range)

Mean:	65.14019
Median:	65.14019
SD:	13.0938
IQR:	18



The median age, which is also approximately 65.14 years, is very close to the mean age. This indicates that the age distribution of billionaires is relatively symmetric. Some billionaires are notably younger or older than the average. Most billionaires tend to be in their mid-60s.

RESEARCH QUESTIONS: 2

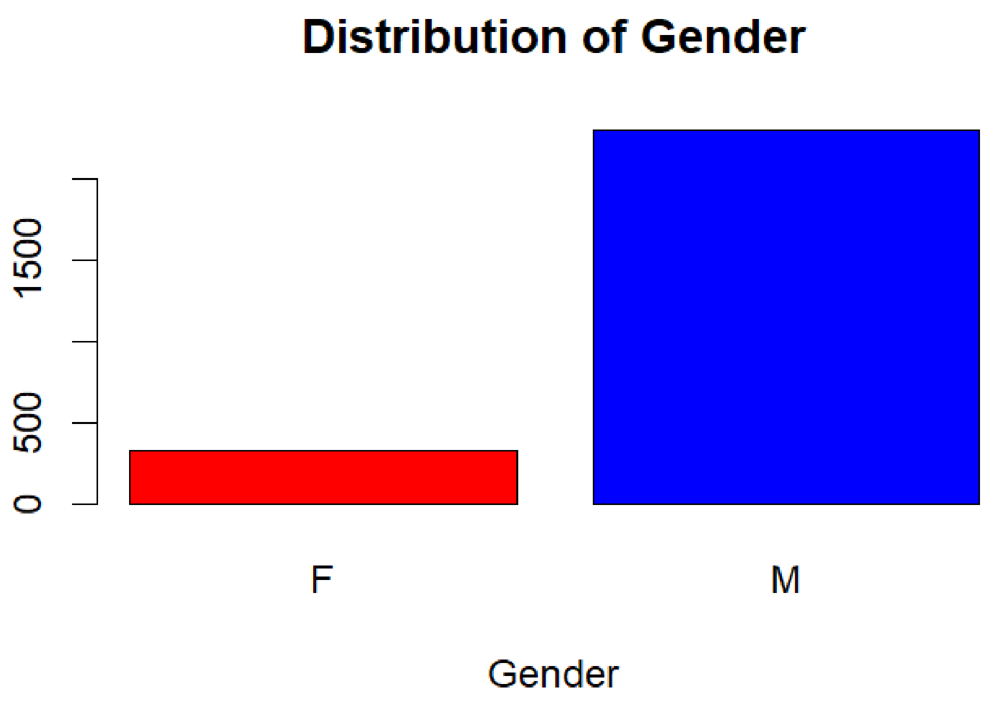
What is the distribution of the gender of billionaires in 2023? What is the gender makeup of the billionaire population?

I approached this question by calculating the frequency table of the Gender variable.

```
billionaire_gender<-b_data$gender
gender_frequency<-table(gender)
print(gender_frequency)
```

Billionaires Gender	
Female	Male
337	2303

% of Gender	
Female	12.76%
337	87.23%



Approximately 12.76% of billionaires are female, while around 87.23% are male. This means that the majority of billionaires in 2023 are male, with female makes up a smaller proportion of the billionaire population

RESEARCH QUESTIONS: 3

Is there a linear relationship between a billionaire's age at the time of data collection and their final net worth? In other words, does age correlate with wealth for billionaires in 2023?

We looked at the relationship between Age and Final Worth using the Scatter plot chart.

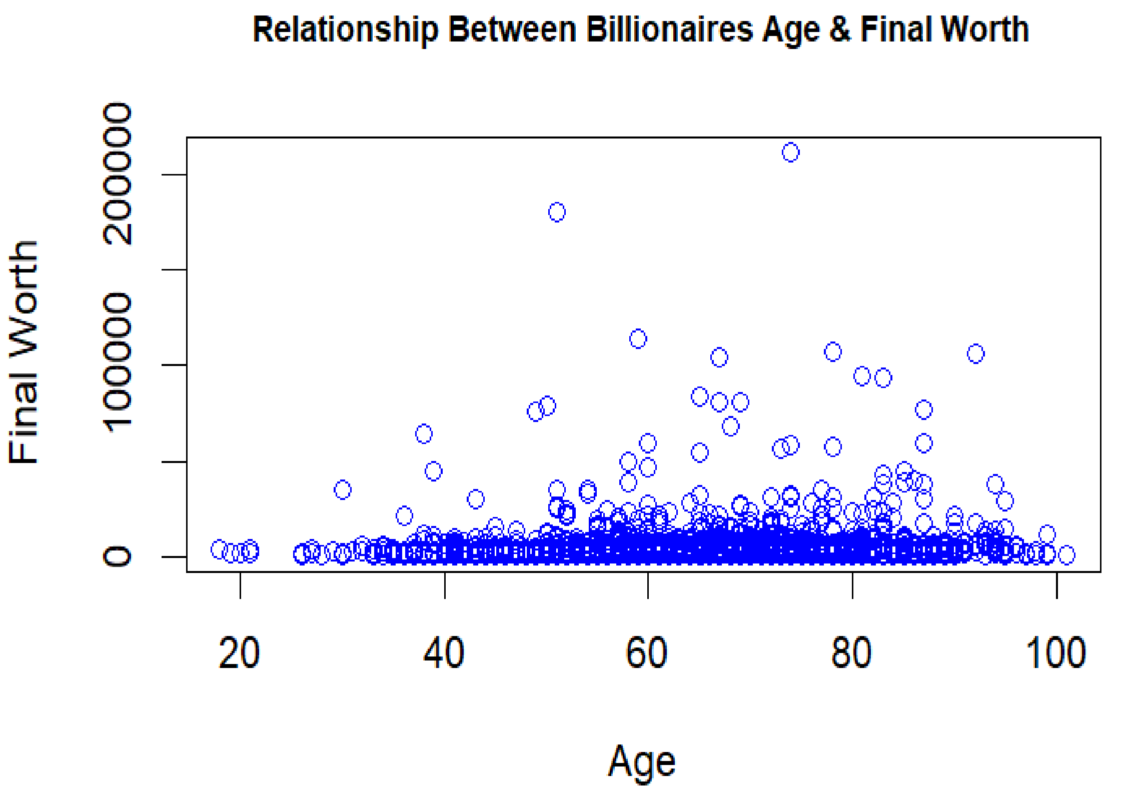
Variables: Age (Quantitative) and Final Worth (Quantitative)

```
plot(b_data$age,b_data$final_worth,
main = "Relationship Between Billionaires Age &
Final Worth",
xlab = "Age",
ylab = "Final Worth",
col = "blue",
pch = 1,
cex.main = 0.8)
```

We run a correlation/regression analysis to ascertain this.

```
correlation<- cor(b_data$final_worth, b_data$age)
print(correlation)
```

Ans: 0.06695056



A correlation of 0.06695056 indicates a very weak, positive relationship between age and final net worth among billionaires in your dataset, and it does not provide much predictive power or explanatory value for final net worth based on age.

CONCLUSION

- The median age, which is also approximately 65.14 years, is very close to the mean age. This indicates that the age distribution of billionaires is relatively symmetric. Some billionaires are notably younger or older than the average. Most billionaires tend to be in their mid-60s.
- Approximately 12.76% of billionaires are female, while around 87.23% are male. This means that the majority of billionaires in 2023 are male, with female makes up a smaller proportion of the billionaire population
- A correlation of 0.06695056 indicates a very weak, positive relationship between age and final net worth among billionaires in your dataset, and it does not provide much predictive power or explanatory value for final net worth based on age.

REFERENCES

Bluman, G. (2018). Elementary Statistics. A step-by-step Approach, 10th Edition, McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. C. Elementary Statistics: A Step by Step Approach - 10th Edition - Solutions and answers | Quizlet

Data Daft. (2020, February 4). dplyr::group_by() | How to use dplyr group by function | R Programming [Video]. YouTube. *https://www.youtube.com/watch?v=5lJX4IgmqRA*

Eugene O. (2021, February 12). How To... Perform Vector Arithmetic in R #14. [Video]. YouTube. (1294) How To... Perform Vector Arithmetic in R #14 – YouTube

Kabacoff, I. (2020). R in Action. 3rd Edition, Manning Publications. https://www.manning.com/books/r-in-actionthird-edition

Kaggle: https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset/

R Programming 101 Group by and Summarise functions in R programming - use the tidyverse package to wrangle your data [Video]. YouTube. (1294) Group by and Summarise functions in R programming - use the tidyverse package to wrangle your data – YouTube

R Programming 101 Bar charts and Histograms using ggplot in R [Video]. YouTube (1294) Bar charts and Histograms using ggplot in R - YouTube

Shobha, K. (2021. Descriptive Statistics R-Software-R Studio [Video]. YouTube. (1294) DESCRIPTIVE STATISTICS R SOFTWARE – YouTube