# **IBU 512 Advanced Topics in Database Systems**

Dataset Results

**Topic:** Exploratory Data Analysis on the Billionaire Dataset

**Dataset:** [Billionaires Statistics Dataset (2023)](https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset/)

**Code:** [Google Collab Python Notebook](https://drive.google.com/file/d/1o09uO-ZSZpkg7Inwi1TfAo3mjC94LLCd/view?usp=sharing)

## Dataset Explanation

The dataset used for this analysis is the Billionaire Statistics Dataset, which provides detailed information about global billionaires, their demographics, industries, wealth sources, and geographic distribution.

This dataset includes various attributes related to billionaires worldwide, including:

* Net worth: The total wealth of each billionaire.
* Industries: The sector in which the billionaire has accumulated wealth.
* Geographical data: Countries and regions where billionaires are concentrated.
* Demographics: Age, gender, and citizenship of billionaires.
* Source of wealth: Whether the billionaire's wealth was self-made or inherited.
* Economic indicators: GDP, tax revenue, and population of countries where billionaires reside.

### ***1.1. Dataset Attributes***

The dataset originally contained a wide range of columns (35 columns), as detailed in the table below:

| **Column Name** | **Description** |
| --- | --- |
| **rank** | The ranking of the billionaire in the dataset based on net worth. |
| **finalWorth** | The billionaire’s total net worth (in billions of USD). |
| **category** | The broad classification of the billionaire’s industry. |
| **personName** | The full name of the billionaire. |
| **age** | The age of the billionaire at the time of data collection. |
| **country** | The country where the billionaire primarily resides. |
| **city** | The city where the billionaire primarily resides. |
| **source** | The main company or asset responsible for the billionaire’s wealth (e.g., Amazon, Tesla). |
| **industries** | The specific industry in which the billionaire has accumulated wealth (e.g., Technology, Retail). |
| **countryOfCitizenship** | The country where the billionaire holds official citizenship. |
| **organization** | The main organization or business associated with the billionaire. |
| **selfMade** | Whether the billionaire is self-made (True) or inherited (False). |
| **status** | Current status of the billionaire (e.g., Active, Retired, Deceased). |
| **gender** | The gender of the billionaire (Male/Female). |
| **birthDate** | The full birthdate of the billionaire (YYYY-MM-DD format). |
| **lastName** | The billionaire’s last name (if applicable). |
| **firstName** | The billionaire’s first name (if applicable). |
| **title** | Any official title or designation held by the billionaire (e.g., CEO, Founder). |
| **date** | The date when the billionaire’s wealth was last updated. |
| **state** | The state where the billionaire resides (if available). |
| **residenceStateRegion** | The region or province where the billionaire lives. |
| **birthYear** | The year of birth of the billionaire. |
| **birthMonth** | The month of birth of the billionaire. |
| **birthDay** | The day of birth of the billionaire. |
| **cpi\_country** | The Consumer Price Index (CPI) of the billionaire’s country (indicates inflation level). |
| **cpi\_change\_country** | The percentage change in CPI, showing inflation trends. |
| **gdp\_country** | The Gross Domestic Product (GDP) of the billionaire’s country. |
| **gross\_tertiary\_education\_enrollment** | The percentage of the population enrolled in higher education in the billionaire’s country. |
| **gross\_primary\_education\_enrollment\_country** | The percentage of the population enrolled in primary education. |
| **life\_expectancy\_country** | The average life expectancy in the billionaire’s country. |
| **tax\_revenue\_country\_country** | The tax revenue collected as a percentage of GDP. |
| **total\_tax\_rate\_country** | The total tax rate in the billionaire’s country. |
| **population\_country** | The total population of the billionaire’s country. |
| **latitude\_country** | The latitude coordinate of the billionaire’s country. |
| **longitude\_country** | The longitude coordinate of the billionaire’s country. |

### 1.2. Dataset Selection

Although the dataset initially contained 35 attributes, not all were relevant for analysis. 15 key columns were selected that provided the most valuable insights for the analysis.

The selected attributes are:

1. rank
2. personName
3. finalWorth
4. age
5. gender
6. industries
7. countryOfCitizenship
8. selfMade
9. gdp\_country
10. tax\_revenue\_country
11. total\_tax\_rate\_country
12. population\_country
13. source
14. organization
15. city

### ***1.3. Dataset Size and Reduction***

* The dataset originally contained 2640 rows.
* Certain records with excessive missing values were removed to ensure data quality and accuracy.
* After preprocessing, the final dataset contained 2457 rows.

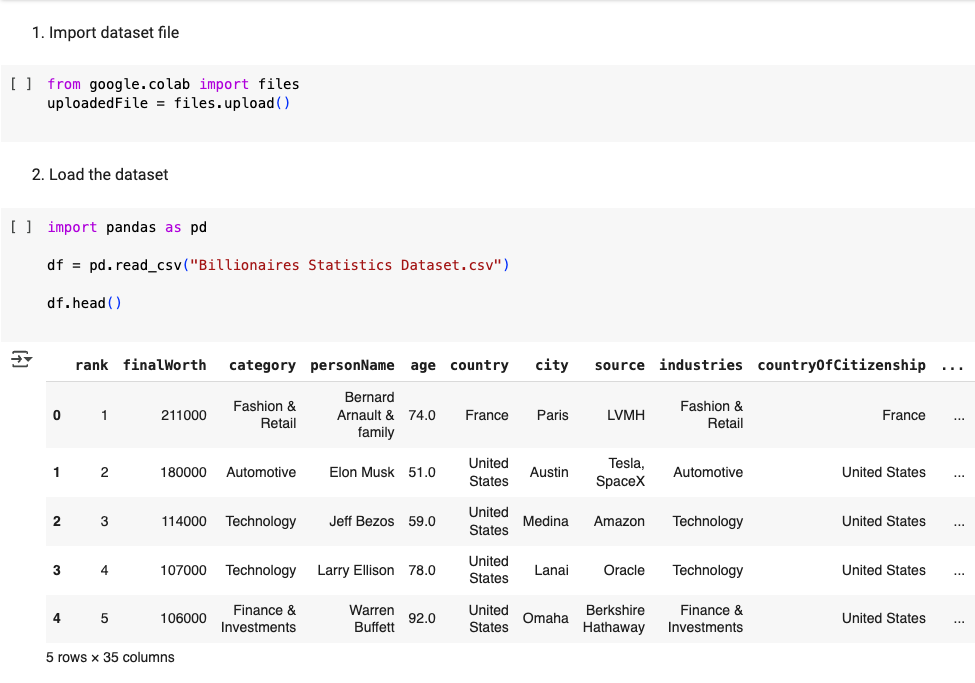
## Dataset Preprocessing and Cleaning

Before conducting Exploratory Data Analysis (EDA), several preprocessing steps were applied to clean the dataset and ensure its quality. These steps included:

* Handling missing values
* Dropping unnecessary columns
* Transforming data types
* Fetching missing values using APIs
* Removing duplicates and final dataset adjustments

### 2.1. Data Loading

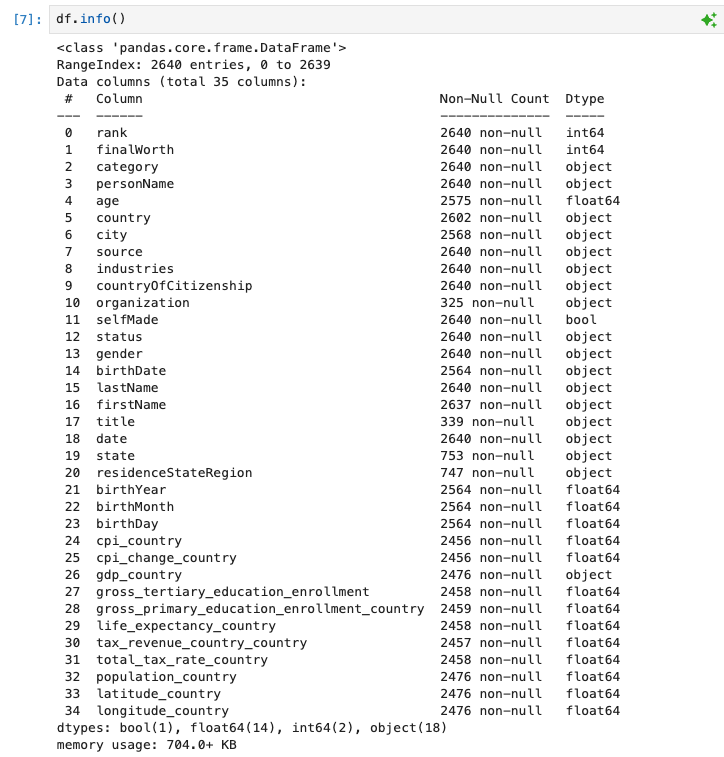
The dataset was first imported into Python using pandas library. The first few rows were displayed using .head() to understand its structure, as shown in the Figure 1.

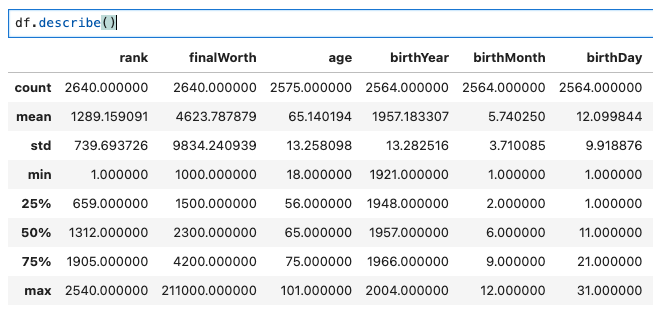


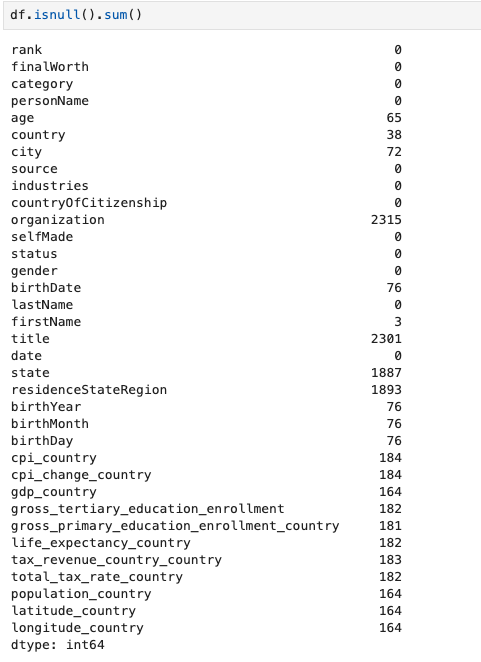
##### Figure 1: Dataset import and presenting

### 2.2. Initial Data Inspection

To check column names, missing values, and data types, commands as shown in figure 2 were used. This helped to identify unnecessary columns and missing values.



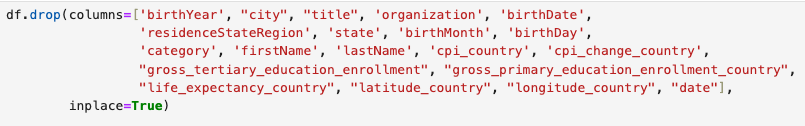




##### Figure 2: Initial Dataset Inspection

### 2.3. Dropping Unnecessary Columns

Since many columns were either redundant or had excessive missing values, 20 columns that were unnecessary for the analysis were removed.



##### Figure 3:Removing Unnecessary Columns

### 2.4. Handling Missing Values

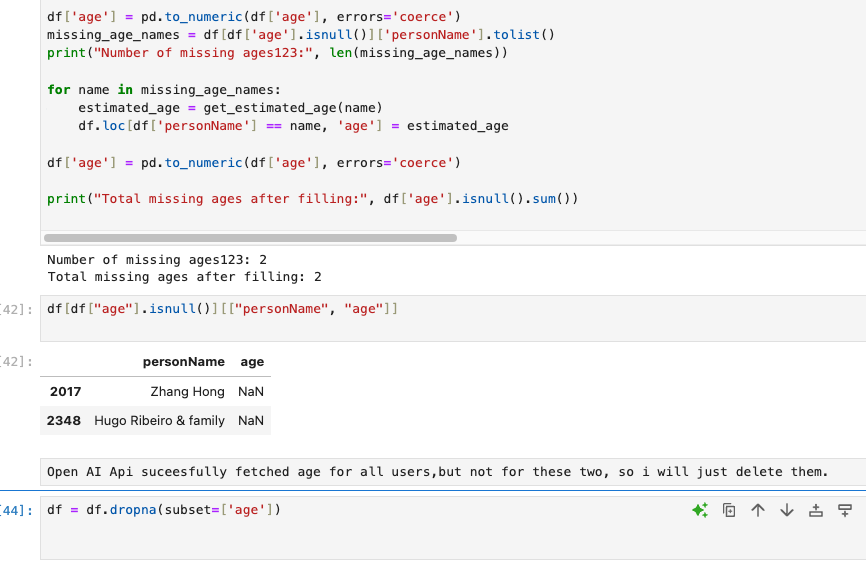
Some rows with missing values in key columns (like country) were removed since there was not a lot of them.



##### Figure 4: Handling Missing Values

### 2.5. Filling Missing Age and Economic Indicator Values using APIs

Some billionaire ages were missing (53 total). To fill them, I used the OpenAI API to estimate the ages. The API successfully retrieved ages for most billionaires, but those rows were removed for two individuals where it failed. Then some country-level economic indicators (GDP, tax rates, population) were missing. World Bank API was used to fetch those missing values. Rows were removed where API fetch failed (remaining missing values).

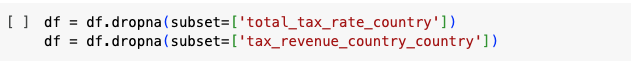


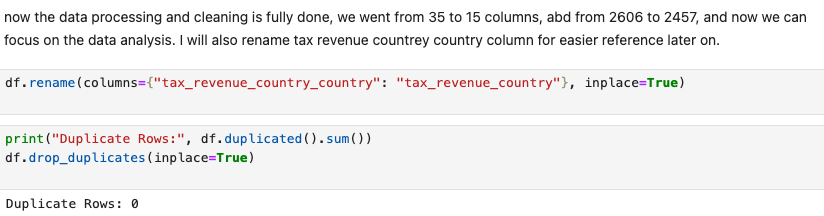
#### 

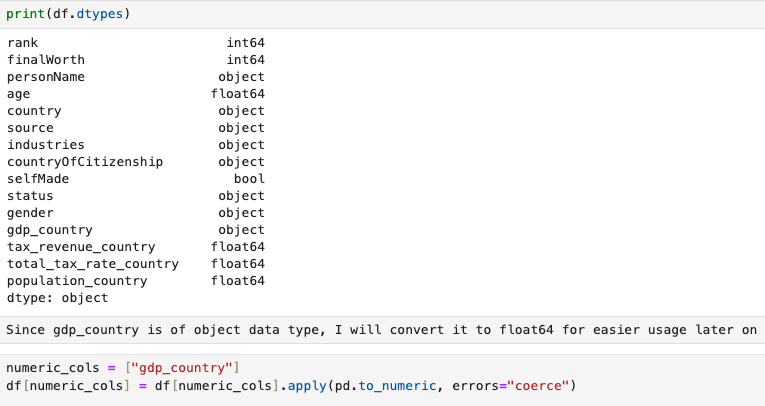
##### Figure 5: Filling Missing Values Using APIs

### 2.6. Final Cleaning and Data Type Fixes

A small amount of rows were removed that were left empty after the API fetch failed. Additionally, the confusing column name was changed to be easily understandable. Duplicate rows were also removed. Numeric data was converted to proper data types.







##### *Figure 6: Final Clean Up*

## Exploratory Data Analysis (EDA)

This data analysis plays a crucial role in understanding global wealth distribution. This section will investigate key trends and patterns in billionaire demographics, industries, geographic distribution, and economic correlations.

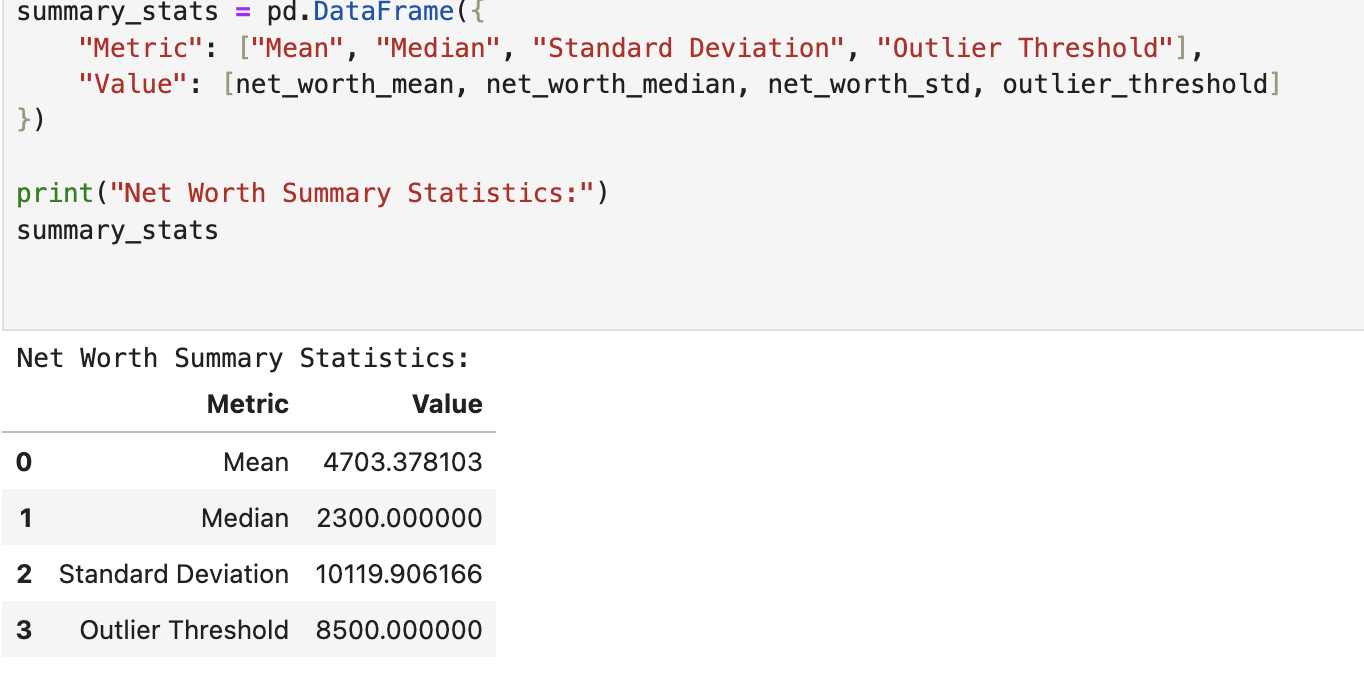
### 3.1. Net Worth Analysis

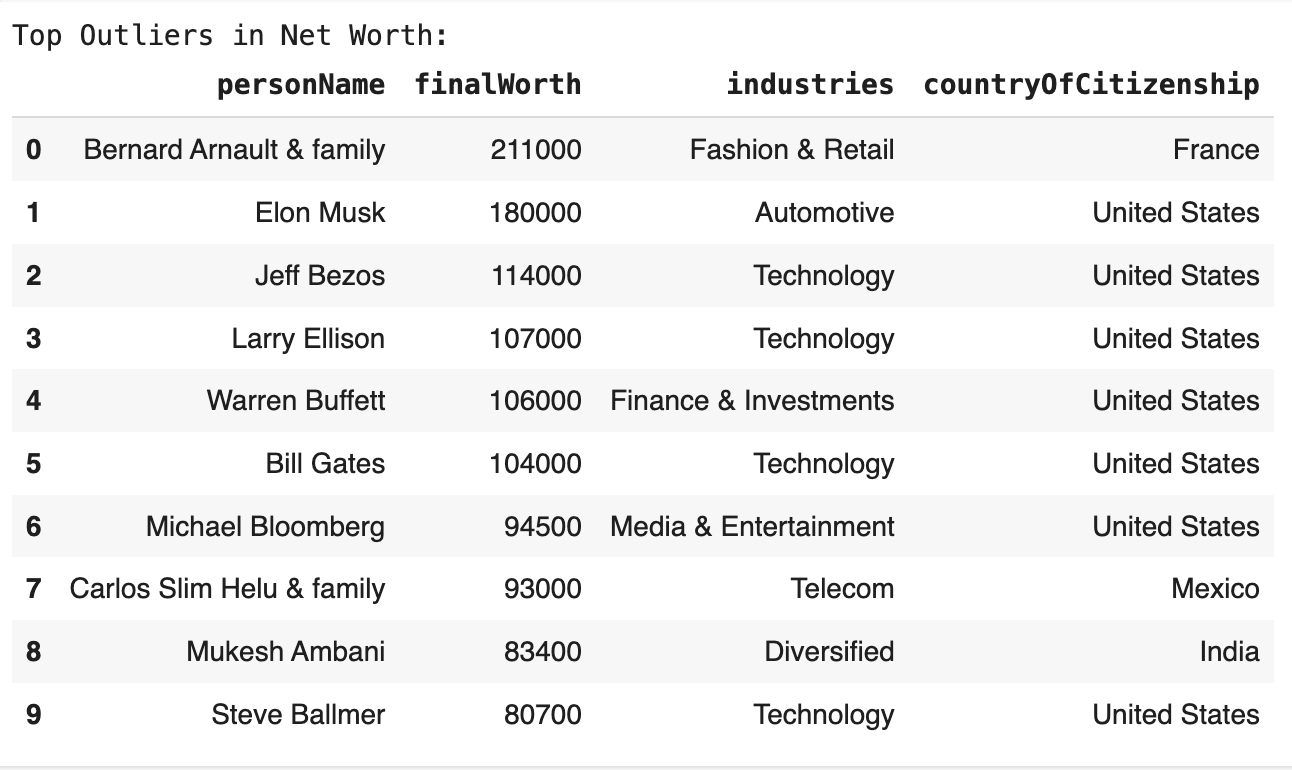
In this section, the distribution of billionaire net worth was analyzed, outliers were identified, and it was examined which industries and countries have the highest total billionaire wealth.

#### 3.1.1. Summary Statistics

To get an overview of billionaire's net worth, certain calculations were done.

* Mean (Average Net Worth) – the average billionaire wealth.
* Median Net Worth – the middle billionaire net worth, less influenced by extreme values.
* Standard Deviation – a measure of how much billionaire net worth varies.
* Outliers Using IQR (Interquartile Range) Method – extreme billionaire net worth values.



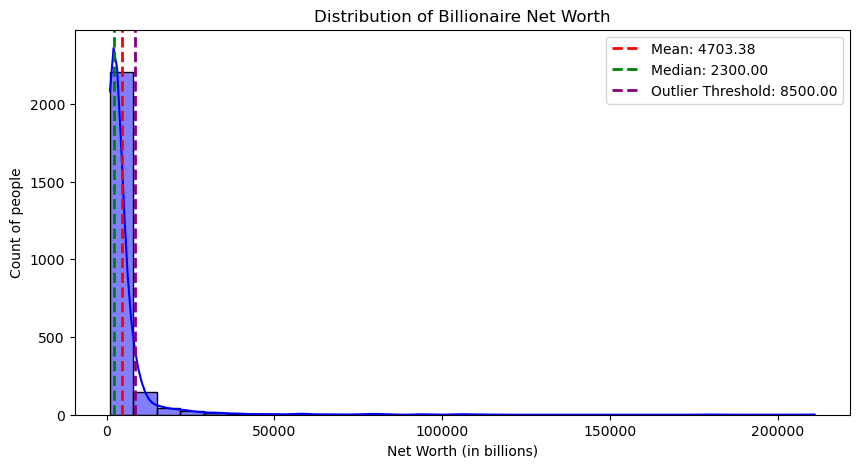


##### Figure 7: Billionaire Statistics and Outliers

As a result, the mean and median billionaire wealth was calculated. Most billionaires have lower net worth, however, outlier calculation has shown that there are a couple of billionaires who are more rich than others.

#### 3.1.2. Visualize Net Worth Distribution

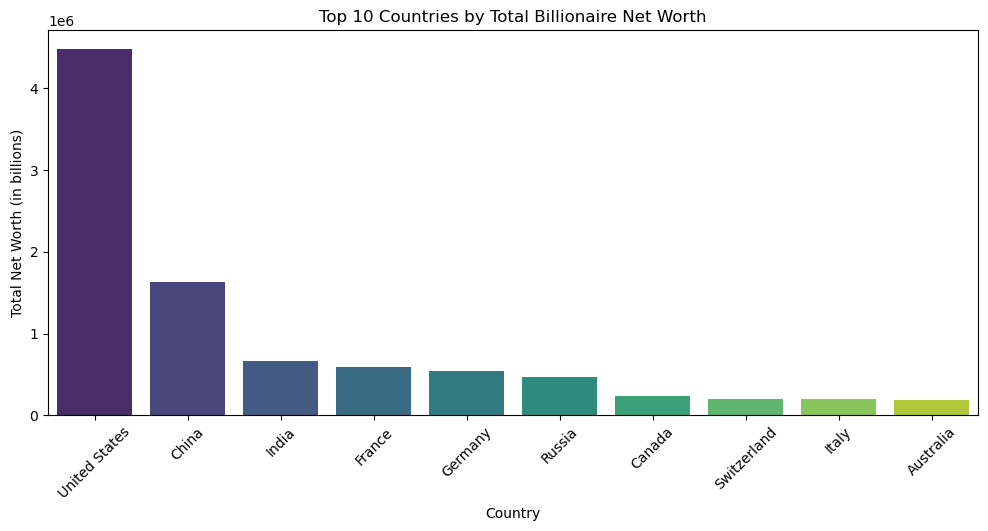
A histogram was plotted to visualize how billionaire wealth is distributed. The distribution is highly skewed, meaning a few billionaires hold extreme wealth, while the majority have significantly lower net worth. Vertical lines indicate the mean, median, and outlier threshold.



##### Figure 8: Distribution of Billionaire Net Worth

**3.1.3. Countries with the Highest Total Net Worth**

To explore wealth concentration globally, billionaire net worth was aggregated by country and the top 10 countries with the highest billionaire wealth were identified. The United States and China lead in total billionaire net worth. Emerging economies show significant billionaire presence but have a lower total wealth. A bar chart was created to compare the top 10 wealthiest countries by billionaire net worth.



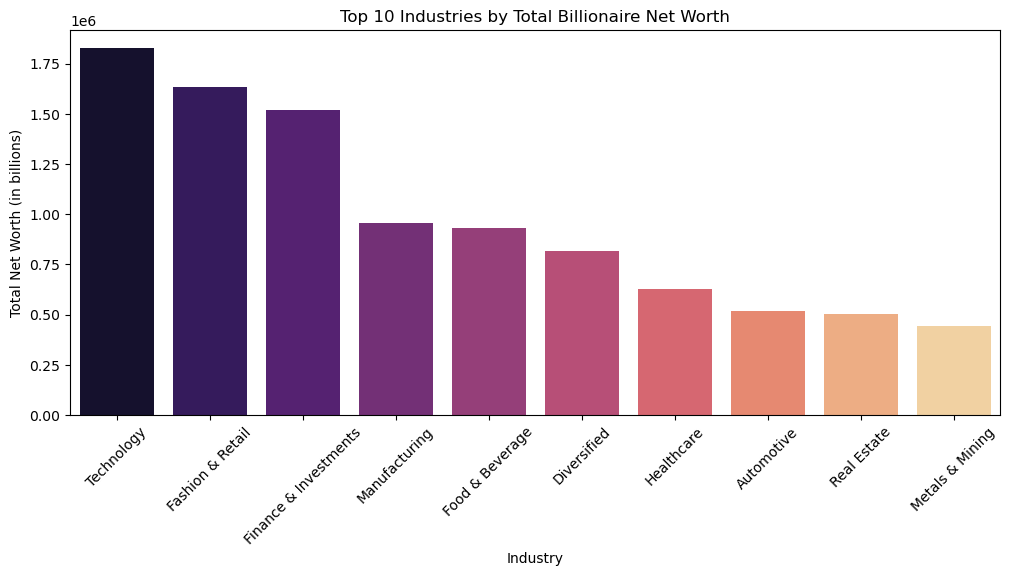
##### Figure 9: Distribution of Billionaires per Country

#### 3.1.4. Industries with the Most Billionaire Wealth

Billionaire wealth was analyzed by industry, identifying the top 10 most lucrative sectors. Technology, Finance, and Retail dominate billionaire wealth. Certain industries have fewer billionaires but extremely high net worth per individual. A bar chart was created to showcase which industries hold the most billionaire wealth. Additionally, I calculated the average billionaire net worth per industry to see if some industries consistently produce wealthier billionaires.

**

##### Figure 10: Top 10 Industries with Average Net Worth

****

##### Figure 11: Top 10 Industries by Total Billionaire Net Worth

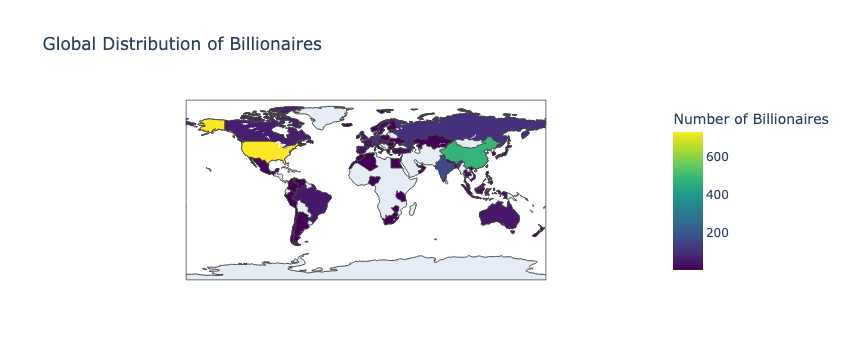
### 3.2. Geographical Distribution

The geographical analysis of billionaires and their net worth has provided valuable insights into how wealth is distributed across the globe. The first choropleth map highlights the number of billionaires in different countries, while the second map visualizes the total billionaire wealth by country.

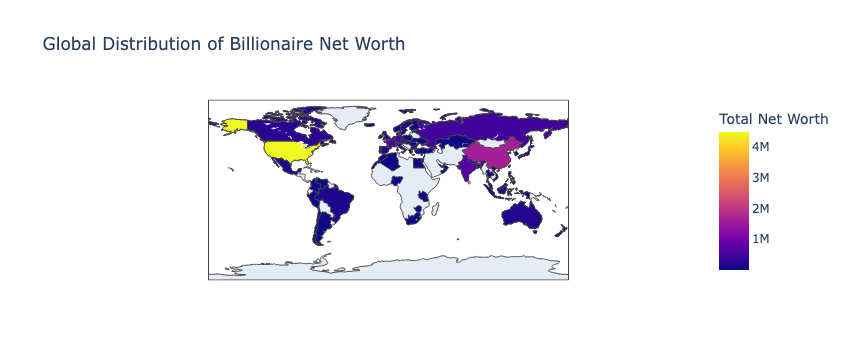
Key findings from this analysis are:

* The United States and China lead in billionaire count, indicating that these two countries are home to the highest number of billionaires.
* The United States dominates in total billionaire wealth, far surpassing other nations, reflecting its economic strength and wealth concentration.
* Other countries with significant billionaire populations include India, Germany, and Russia, showing that billionaire wealth is not only concentrated in Western economies but also in emerging markets.
* Countries with lower billionaire counts often have significantly lower total net worth, suggesting that wealth is highly concentrated in a few dominant economies.

These findings reinforce the idea that economic powerhouses like the U.S. and China play a crucial role in global wealth accumulation, while other nations with fewer billionaires exhibit a more evenly distributed economic structure.



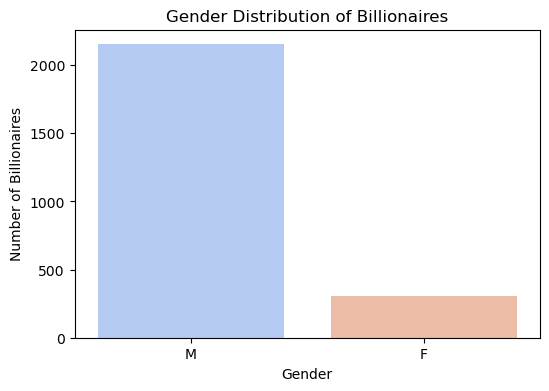
##### Figure 12: Global Distribution of Billionaires



##### Figure 13: Global Distribution of Billionaire Net Worth

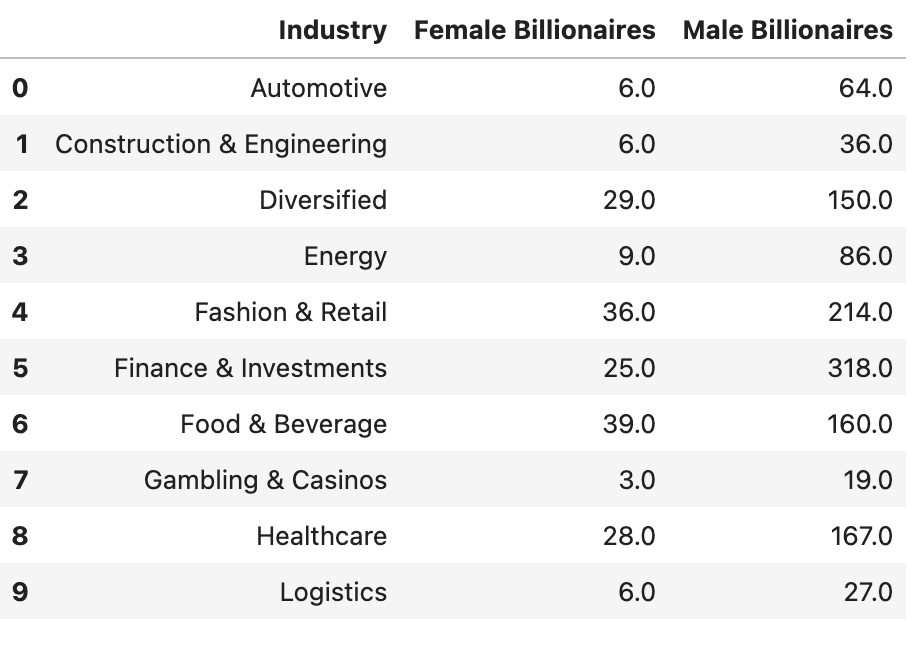
### 3.3. Gender Analysis

The gender distribution analysis reveals a significant disparity between male and female billionaires. The bar chart indicates that the vast majority of billionaires are male, with female billionaires constituting a small fraction of the overall billionaire population. This suggests a gender imbalance in global wealth accumulation, which could be attributed to historical socio-economic factors, industry structures, and access to business opportunities.



##### Figure 14: Gender Distribution of Billionaires

Furthermore, the industry-wise distribution highlights that male billionaires dominate across almost all sectors. However, certain industries, such as Fashion & Retail and Food & Beverage, have a relatively higher proportion of female billionaires compared to others like Automotive, Energy, and Finance & Investments, which are heavily male-dominated. This suggests that certain industries offer more opportunities for female entrepreneurs.

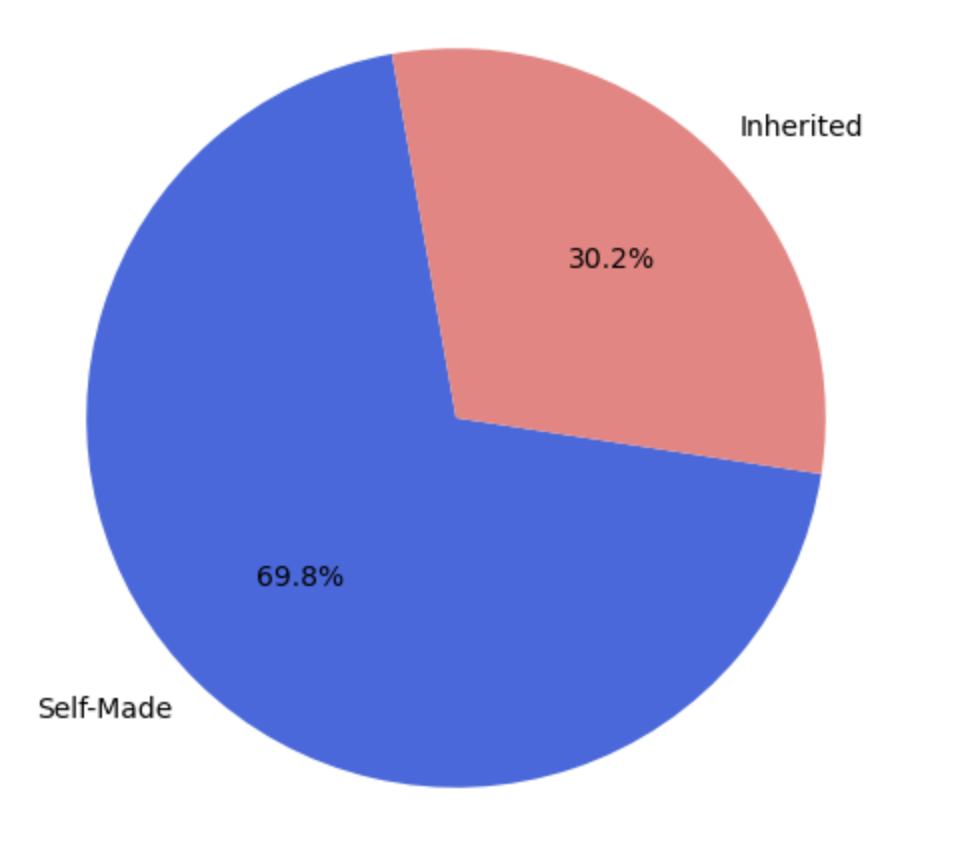


##### Figure 15: Rate of Billionaires per Gendre in Top 10 Industries

Overall, the analysis confirms that wealth accumulation at the billionaire level is still highly skewed towards men, and while some industries show a slightly more balanced distribution, the overall gender gap remains substantial.

### 3.4. Source of Wealth

The majority of billionaires are self-made. The pie chart clearly illustrates that 69.8% of billionaires are self-made, while 30.2% inherited their wealth. This suggests that, despite the notion that wealth is passed down, a significant proportion of billionaires have built their fortunes independently.



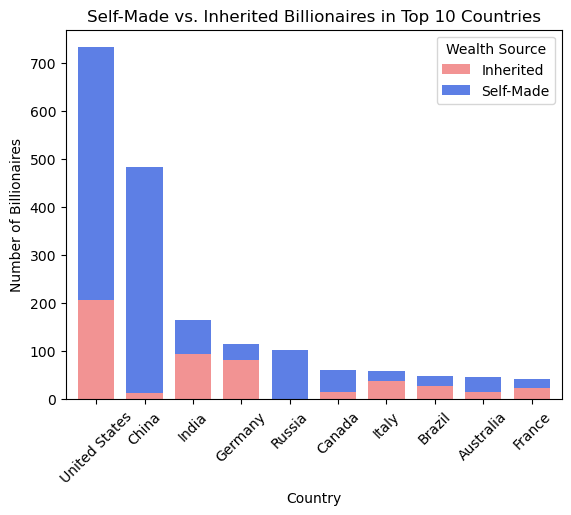
##### Figure 16: Self-made vs inherited billionaire wealth chart

Some industries have a higher concentration of self-made billionaires. Technology, Finance & Investments, and Manufacturing stand out as industries dominated by self-made wealth. In contrast, industries like Fashion & Retail and Real Estate have a larger share of inherited wealth. This is likely due to the nature of these industries, where family-owned businesses are passed down through generations.



##### Figure 17: Self-made vs inherited billionaire per industry

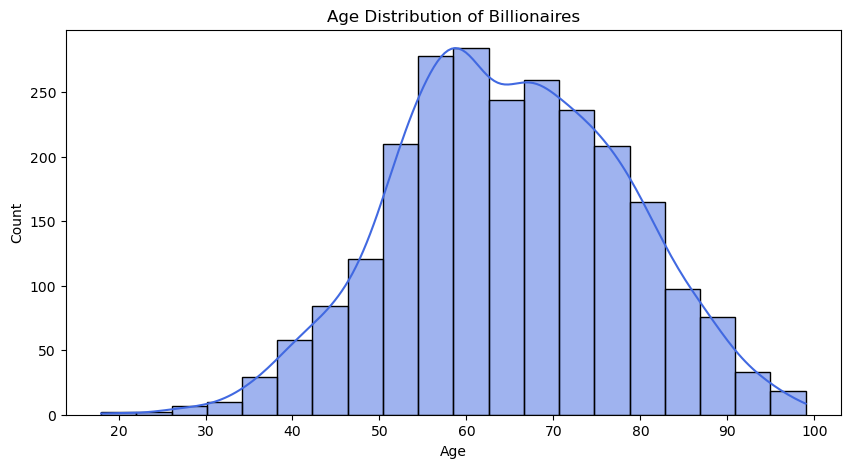
In country-related terms, the United States and China have the highest number of self-made billionaires, reinforcing the entrepreneurial culture in these countries. Germany and Russia have a more balanced mix of self-made and inherited billionaires, reflecting their long-standing industrial and financial legacies. India has a high number of self-made billionaires, indicating a growing startup and entrepreneurial ecosystem.



##### Figure 18: Self-made vs inherited billionaire per country

### 3.5. Age Distribution

The age distribution of billionaires was visualized using a histogram, providing insights into the most common ages among the world's wealthiest individuals. The results indicate that most billionaires fall within the 50-80 age range, with the highest concentration around 60 years old. The distribution is approximately bell-shaped, suggesting a normal distribution where younger and older billionaires are less common. This visualization confirms that the majority of billionaires tend to be older, indicating that wealth accumulation often takes decades to build.



##### Figure 19: Age Distribution Between Billionaires

To further analyze the correlation between age and industry, the average age of billionaires in different industries was calculated alongside their average net worth. The results highlight the following key findings:

* The Automotive, Telecom, and Logistics industries have the highest average age of billionaires, indicating that these sectors might require long-term experience and capital accumulation.
* The Technology industry has the lowest average age (around 56 years old), suggesting that younger billionaires are more likely to emerge in this field. This aligns with real-world trends where tech entrepreneurs (e.g., founders of startups) achieve billionaire status at a younger age compared to those in traditional industries.
* Industries such as Metals & Mining, Energy, and Finance & Investments tend to have older billionaires, reflecting the need for long-term capital investment and market experience.

#### 

##### Figure 20: Correlation between Industry, Average Net Worth and Average Age

This analysis suggests that age plays a crucial role in wealth accumulation, with most billionaires reaching their peak net worth later in life. However, the technology sector stands out as an industry where younger individuals can achieve billionaire status more quickly, likely due to innovation, venture capital funding, and rapid business scaling.

##### 