

Ekram_Hassen

December 16, 2024

0.1 Billionaires Statistics Dataset (2023)

Exploring the Global Landscape of Success

Who wouldn't want to be a billionaire? The idea of immense wealth, influence, and the ability to shape the world is a dream shared by many but achieved by very few. Billionaires occupy a unique position in society, wielding power that extends far beyond their bank accounts. From driving technological innovation to influencing global policies, their decisions ripple across industries, politics, and even everyday life. These individuals are more than just wealthy—they are architects of the future.

This project dives into the fascinating lives of billionaires to uncover the patterns and factors that contribute to their success. What common threads run through their stories? How do age, industry, education, and other demographics shape their rise to the top? By answering these questions, we can better understand the mechanisms behind extraordinary wealth and its broader societal impact.

For this analysis, I've selected the Billionaires Statistics Dataset (2023) from Kaggle, a rich collection of data detailing the net worth, industries, and demographics of the world's billionaires. Our goal is to build a predictive model that estimates the net worth (`finalWorth`) of these individuals based on explanatory variables such as industry, country GDP, education, and more. This exploration not only sheds light on the secrets to immense wealth but also offers insights into global economic and social dynamics.

To deepen our understanding, we will develop and test various predictive models to determine which one best captures the complexity of the data. By comparing the performance of these models, we aim to identify the most effective approach for explaining and predicting billionaire wealth.

Through this analysis, we will address questions such as:

What are the key factors that influence a billionaire's net worth? How does the country of origin or industry type impact wealth accumulation? Are self-made billionaires fundamentally different from those who inherit their wealth? Which modeling techniques provide the best fit for this data? This analysis isn't just about numbers; it's about unraveling the story of power, opportunity, and innovation that underpins the lives of the world's wealthiest individuals while leveraging advanced modeling to bring these insights to life.

Key Features

`Rank`: The ranking of the billionaire in terms of wealth. `finalWorth`: The final net worth of the billionaire in U.S. dollars.

`category`: The category or industry in which the billionaire's business operates.

personName: The full name of the billionaire.

age: The age of the billionaire.

country: The country in which the billionaire resides.

city: The city in which the billionaire resides.

source: The source of the billionaire's wealth.

industries: The industries associated with the billionaire's business interests.

countryOfCitizenship: The country of citizenship of the billionaire.

organization: The name of the organization or company associated with the billionaire.

selfMade: Indicates whether the billionaire is self-made (True/False). status: "D" represents self-made billionaires (Founders/Entrepreneurs) and "U" indicates inherited or unearned wealth.

gender: The gender of the billionaire.

birthDate: The birthdate of the billionaire.

lastName: The last name of the billionaire.

firstName: The first name of the billionaire.

title: The title or honorific of the billionaire.

date: The date of data collection.

state: The state in which the billionaire resides.

residenceStateRegion: The region or state of residence of the billionaire.

birthYear: The birth year of the billionaire.

birthMonth: The birth month of the billionaire.

birthDay: The birth day of the billionaire.

cpi_country: Consumer Price Index (CPI) for the billionaire's country.

cpi_change_country: CPI change for the billionaire's country.

gdp_country: Gross Domestic Product (GDP) for the billionaire's country.

gross_tertiary_education_enrollment: Enrollment in tertiary education in the billionaire's country.

gross_primary_education_enrollment_country: Enrollment in primary education in the billionaire's country.

life_expectancy_country: Life expectancy in the billionaire's country.

tax_revenue_country_country: Tax revenue in the billionaire's country.

total_tax_rate_country: Total tax rate in the billionaire's country.

population_country: Population of the billionaire's country.

latitude_country: Latitude coordinate of the billionaire's country.

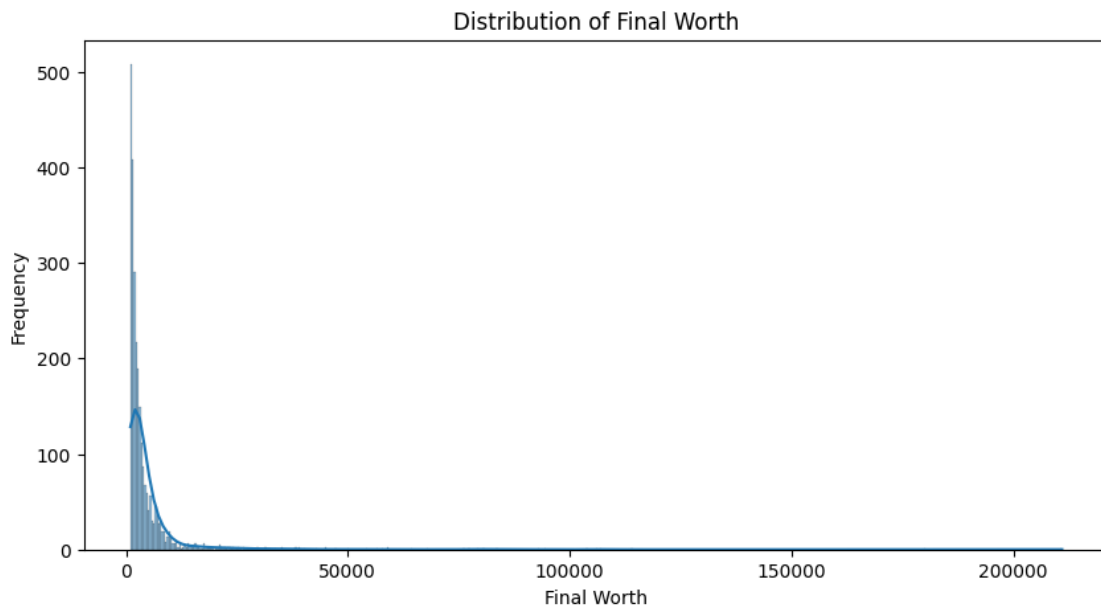
longitude_country: Longitude coordinate of the billionaire's country.

1 Explanatory Data Analysis

1.1 Final Worth Analysis

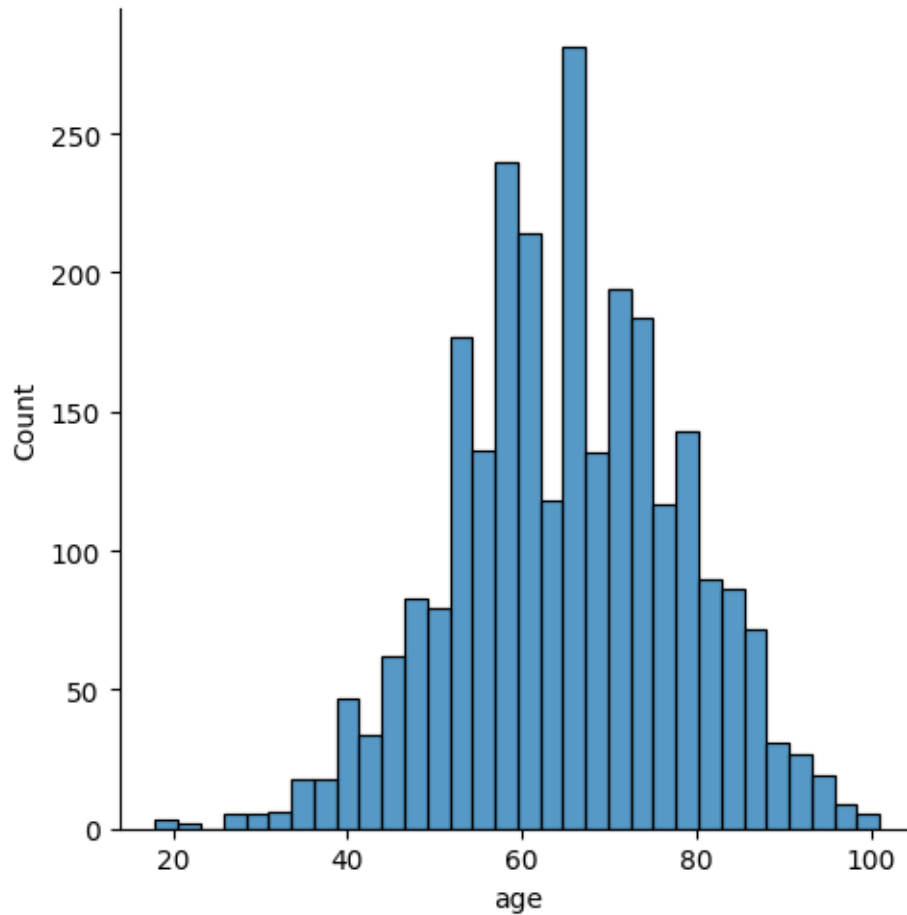
```
count      2640.000000
mean       4623.787879
std        9834.240939
min        1000.000000
25%        1500.000000
50%        2300.000000
75%        4200.000000
max        211000.000000
Name: finalWorth, dtype: float64
```

The data reveals a significant variation in the wealth of billionaires. The average net worth is \$4.62 billion, but the median is \$2.3 billion, indicating that a few extremely wealthy individuals skew the average upward. The highest net worth is \$211 billion, highlighting the considerable wealth disparity. Most billionaires fall within the interquartile range, with net worths between \$1.5 billion and \$4.2 billion, representing the middle 50% of the data. These figures suggest that the distribution is highly uneven, requiring careful handling of extreme values during analysis and prediction.



1.2 Age Analysis

```
<seaborn.axisgrid.FacetGrid at 0x79cf4272ee00>
```



```
age_group
20-30      9
31-40     56
41-50    216
51-60    632
61-70    748
71-80    595
81+     382
Name: count, dtype: int64
```

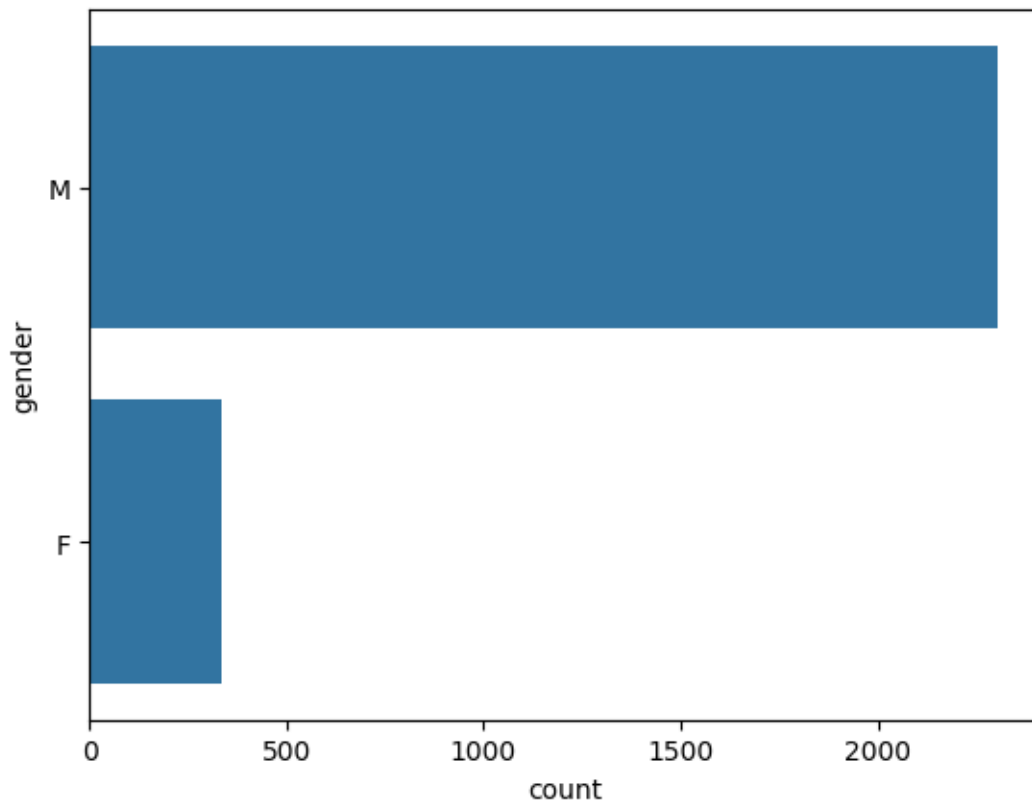
The majority of billionaires fall within the 51-70 age range, with 632 individuals aged 51-60 and 748 aged 61-70, accounting for a significant proportion of the dataset. This suggests that middle-aged to early retirement years are peak periods for accumulating substantial wealth.

Younger billionaires are relatively rare, with only 9 individuals in the 20-30 age range and 56 in the 31-40 range, indicating that becoming a billionaire at a young age is uncommon. On the other hand, the 71-80 group remains substantial with 595 individuals, and 382 billionaires are 81 years or older, showing that a significant number of individuals maintain their billionaire status well into older age.

1.3 Gender analysis

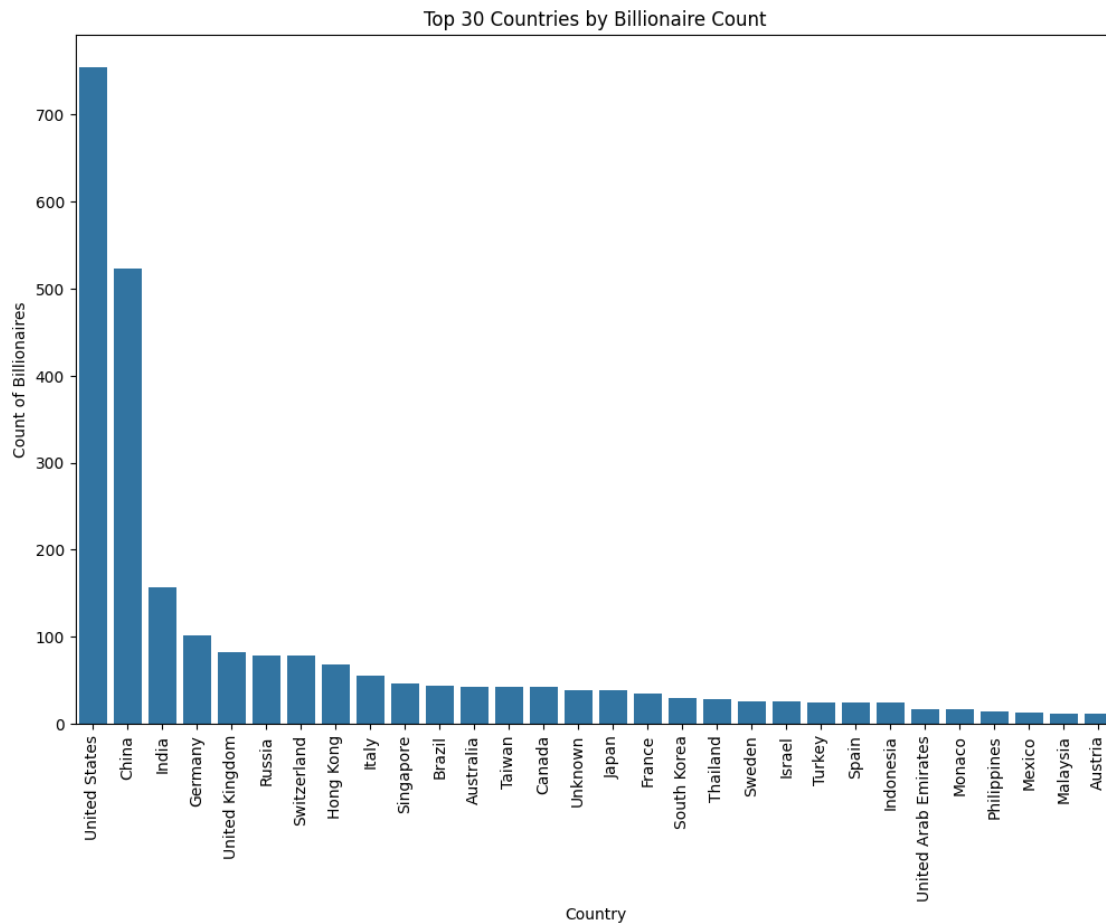
```
gender
M    2303
F     337
Name: count, dtype: int64
```

```
<Axes: xlabel='count', ylabel='gender'>
```



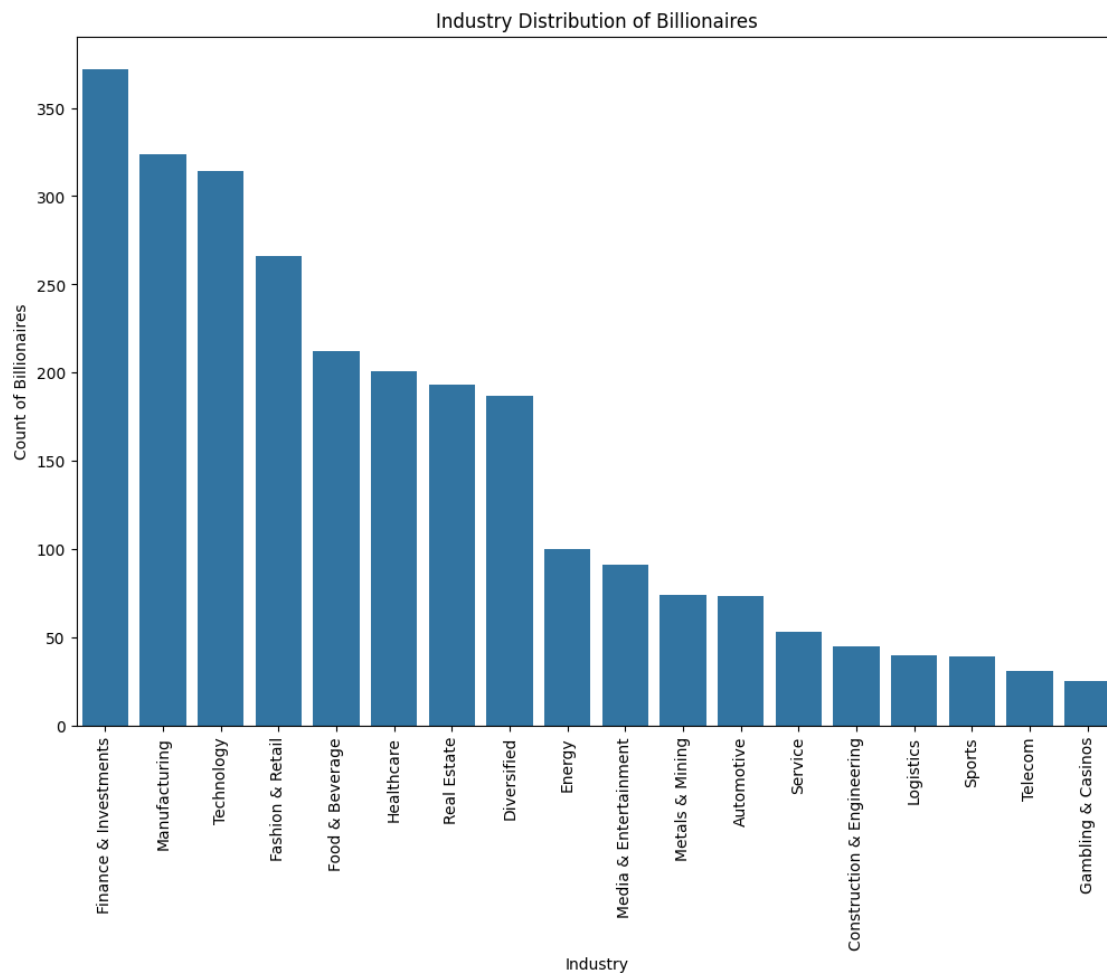
This indicates that men overwhelmingly dominate the billionaire demographic, accounting for approximately 87% of the total, while women represent just 13%. This disparity could be attributed to historical, cultural, and systemic factors that have limited women's access to opportunities for wealth creation, particularly in high-revenue industries and leadership roles.

1.4 Country Analysis



The majority of billionaires originate from the United States, followed by China and India. This concentration emphasizes the global economic dominance of these countries and their ability to create environments conducive to wealth generation. Their strong economies, technological innovation, and entrepreneurial ecosystems contribute significantly to this wealth concentration, showcasing their critical role in shaping global financial landscapes.

1.5 Industry Analysis



Number of Billionaires by Industry:

industries

Finance & Investments	372
Manufacturing	324
Technology	314
Fashion & Retail	266
Food & Beverage	212
Healthcare	201
Real Estate	193
Diversified	187
Energy	100
Media & Entertainment	91
Metals & Mining	74
Automotive	73
Service	53
Construction & Engineering	45

Logistics	40
Sports	39
Telecom	31
Gambling & Casinos	25

Name: count, dtype: int64

Average Net Worth per Industry (in USD):

industries	
Automotive	7195.890411
Telecom	6564.516129
Fashion & Retail	6386.466165
Metals & Mining	6037.837838
Logistics	5987.500000
Technology	5980.573248
Diversified	4840.641711
Gambling & Casinos	4820.000000
Media & Entertainment	4697.802198
Energy	4535.000000
Food & Beverage	4515.094340
Finance & Investments	4314.784946
Sports	3448.717949
Real Estate	3406.217617
Service	3271.698113
Healthcare	3200.000000
Manufacturing	3145.061728
Construction & Engineering	2633.333333

Name: finalWorth, dtype: float64

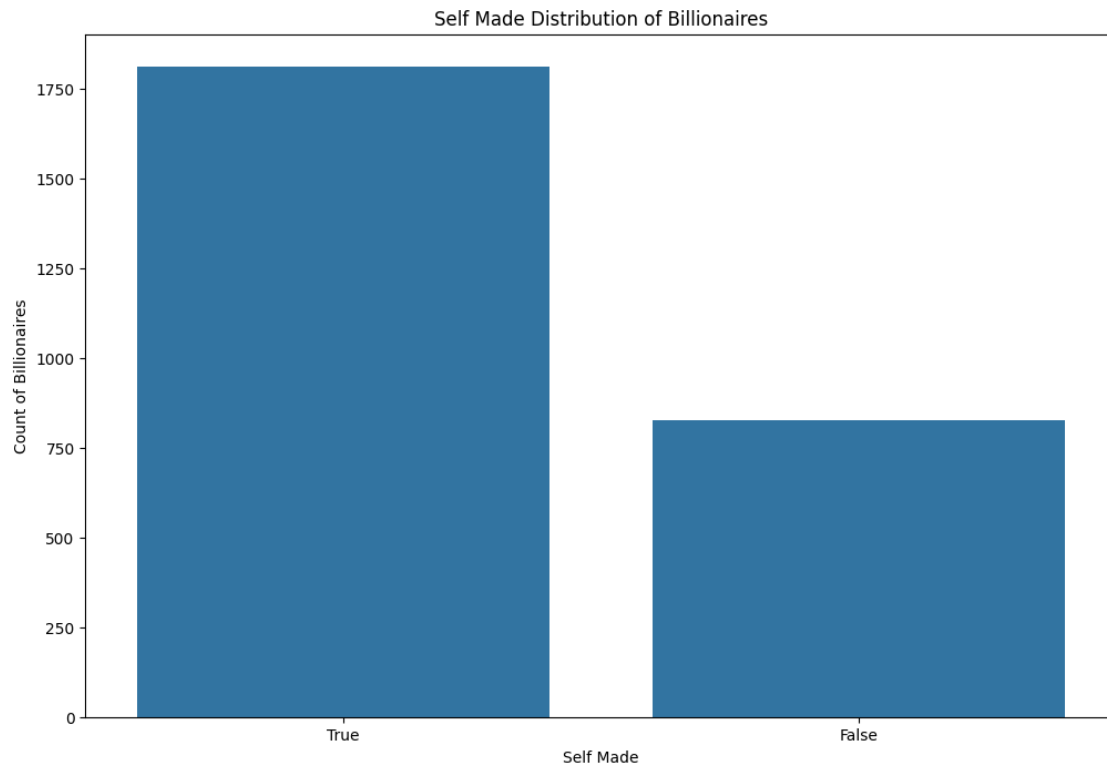
Finance & Investments has the most billionaires (372), followed by Manufacturing (324) and Technology (314), showing that these sectors are big worldwide wealth creators.

In terms of average net worth, Automotive leads with around \$7.2 billion, followed by Telecom (\$6.6 billion) and Fashion & Retail (\$6.4 billion). These industries, while having fewer billionaires than sectors such as finance and manufacturing, have much higher average wealth. This implies that a smaller number of people in these industries have a higher stake.

This implies that a smaller number of people in these industries own a bigger percentage of the wealth, showing a high concentration of wealth among specific businesses or individuals.

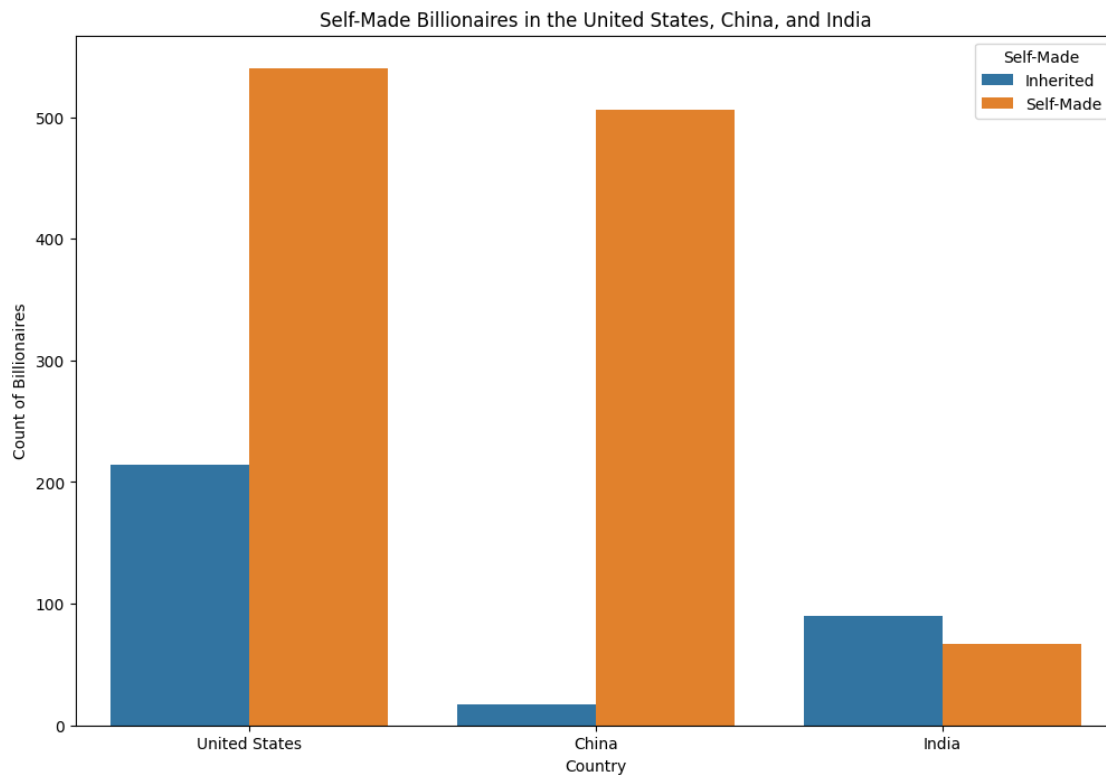
Healthcare (\$3.2 billion), Manufacturing (\$3.1 billion), and Construction & Engineering (\$2.6 billion) have lower average net worths, which could imply that these industries have more billionaires with more evenly distributed wealth, or that their market structures result in less wealth concentration at the top.

1.6 Self Made Analysis



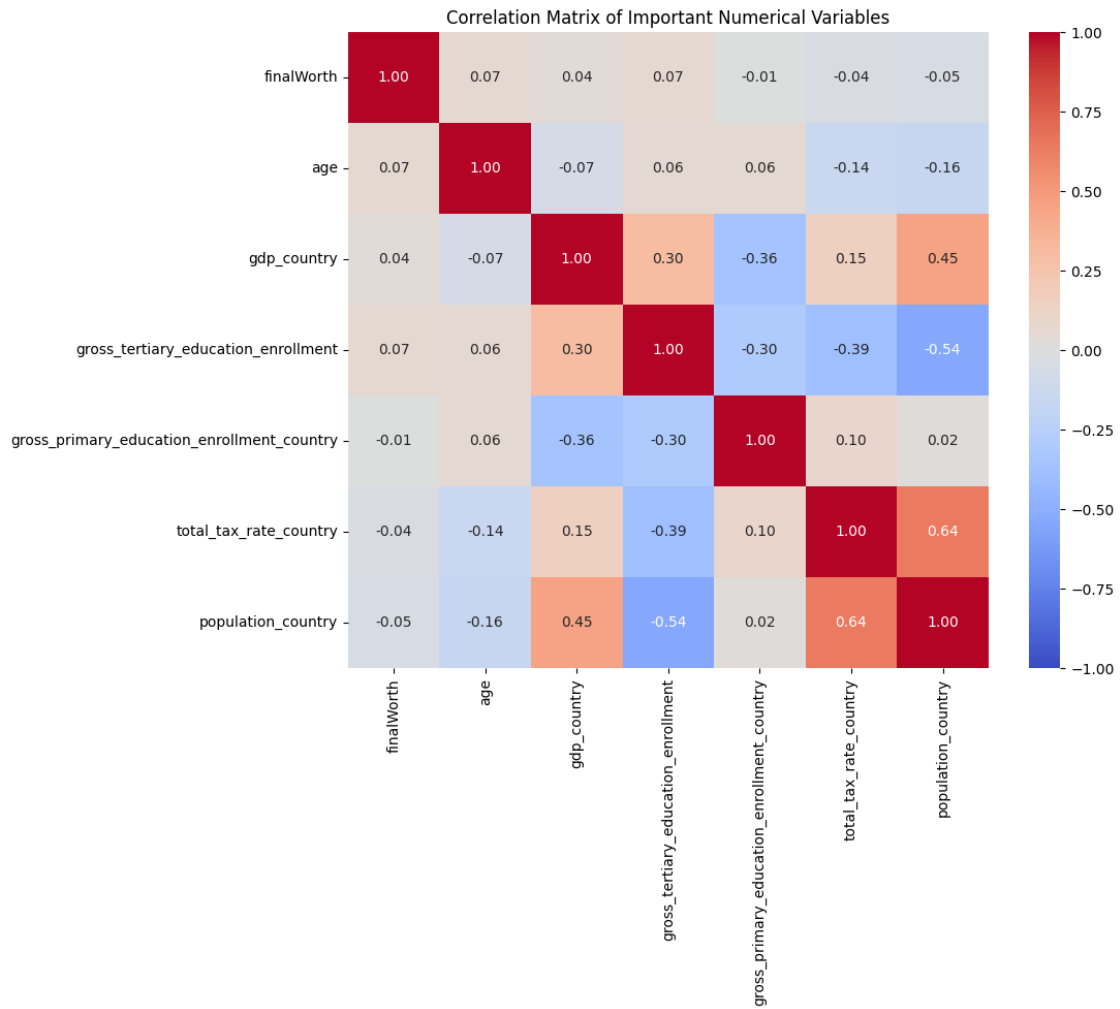
The majority of billionaires are self-made, having accumulated their wealth through their own efforts and ventures. They built their fortunes independently, often starting from humble beginnings and achieving success through innovation, entrepreneurship, and strategic investments.

1.7 Self-Made Billionaires in the United States, China, and India



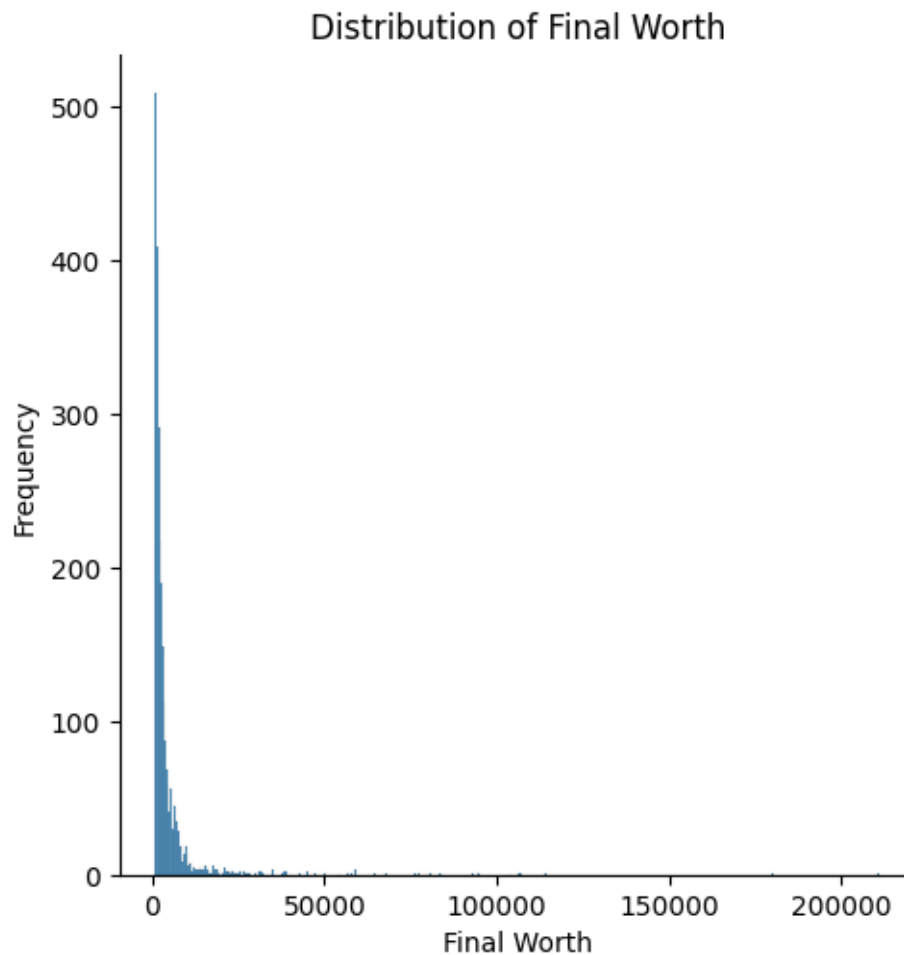
In the United States and China, the majority of billionaires are self-made. However, in India, the majority of billionaires inherit their wealth. In China, the proportion of self-made billionaires is significantly higher compared to inherited wealth, whereas in the United States, there is a notable number of billionaires who inherited their wealth. The contrast is striking, as in China, inherited wealth represents a much smaller portion of the billionaire population.

1.8 Correlation matrix

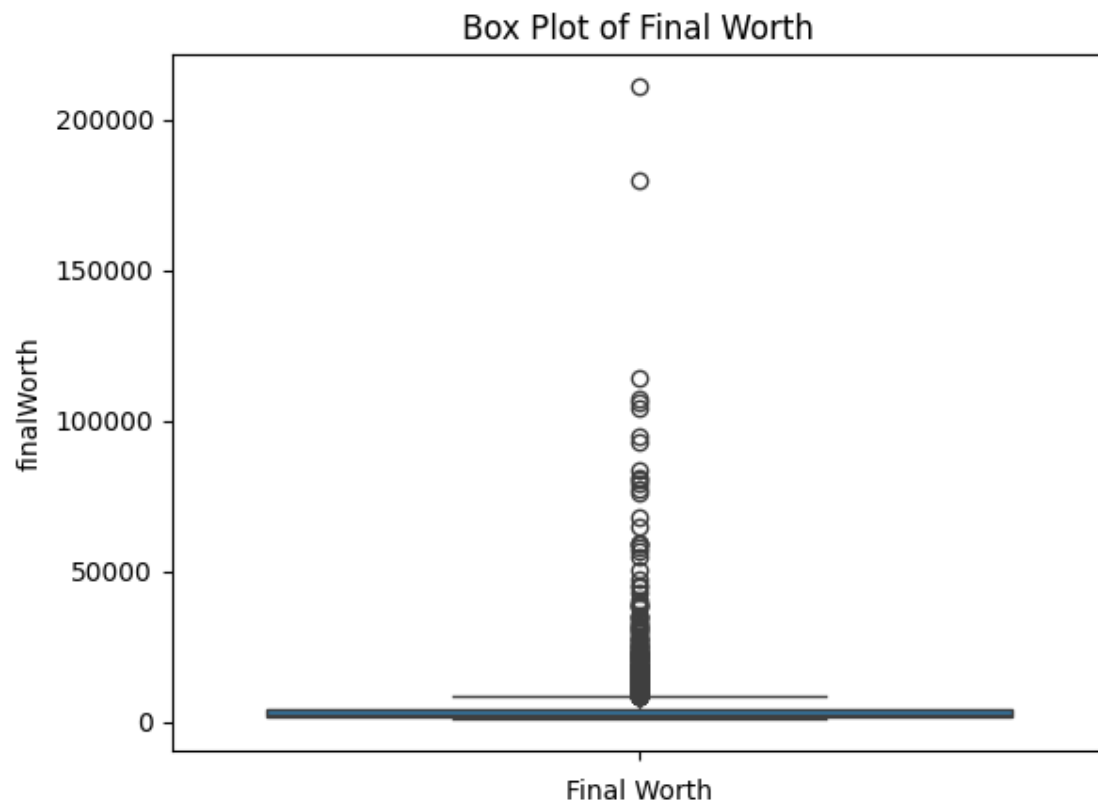


There is no significant correlation between final worth and the other variables based on the linear relationships captured by the correlation analysis. However, this does not mean that there is no connection between the variables; it simply indicates that the relationships may not be linear. In the next section, we will explore non-linear models, which may uncover more complex interactions that linear methods cannot capture.

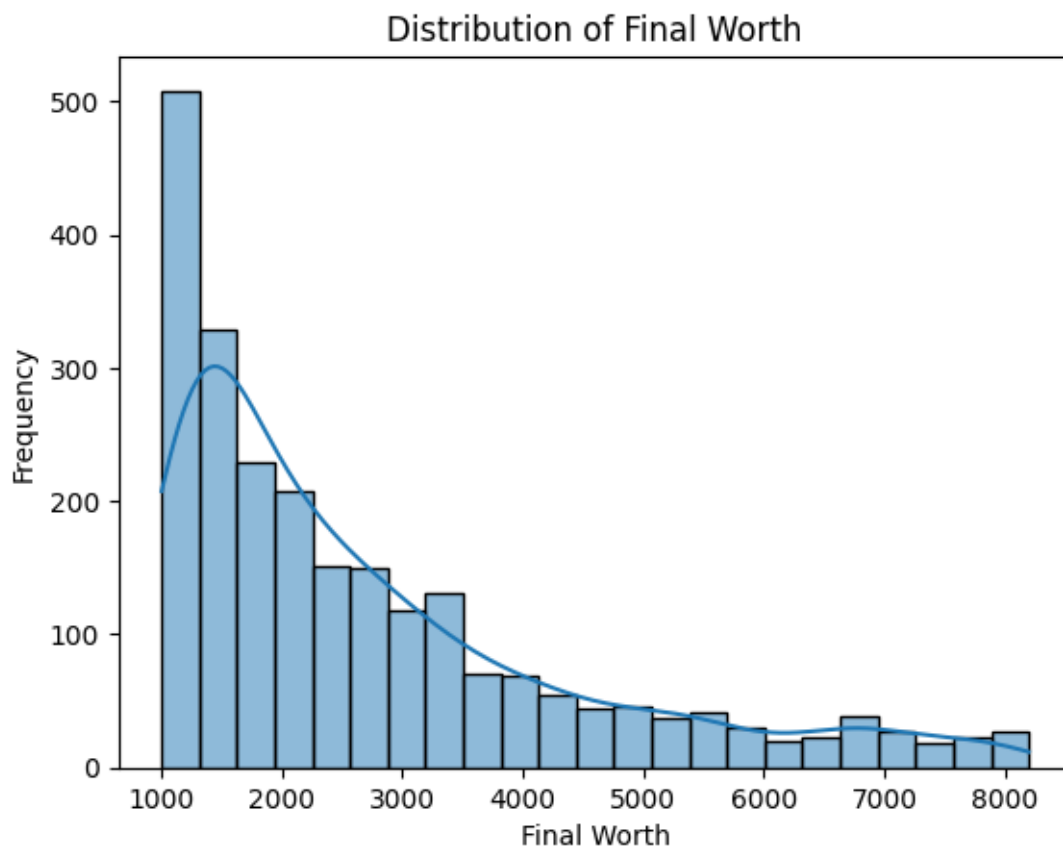
2 Modeling and Interpretation



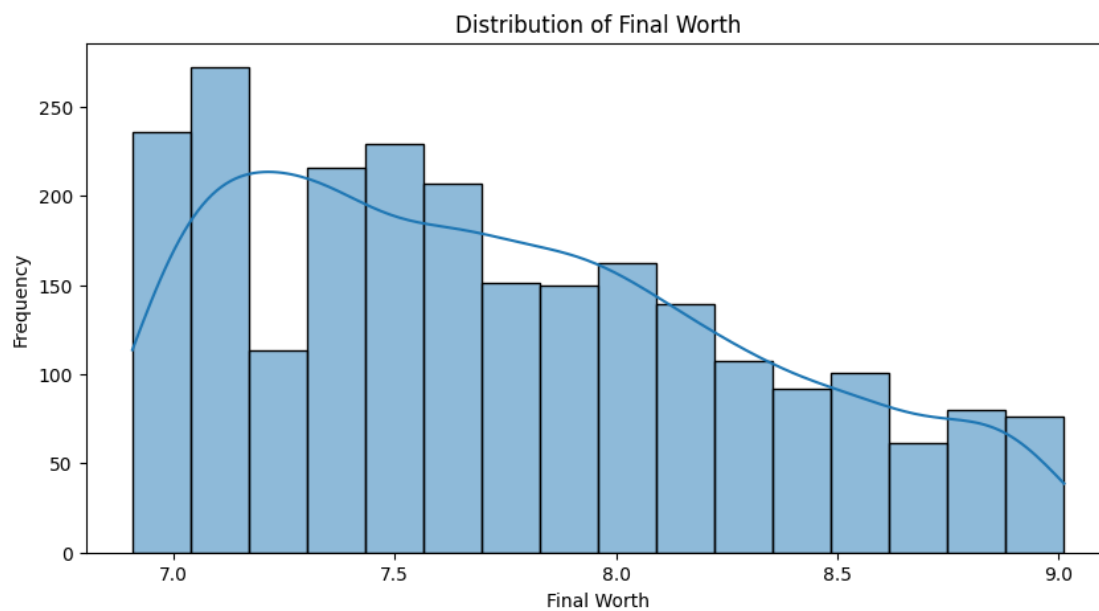
The target variable is highly skewed, which could be problematic for predictive models. As a result, I first examined the boxplot to check for outliers. If any outliers were detected, I proceeded to remove them to improve the model's performance.

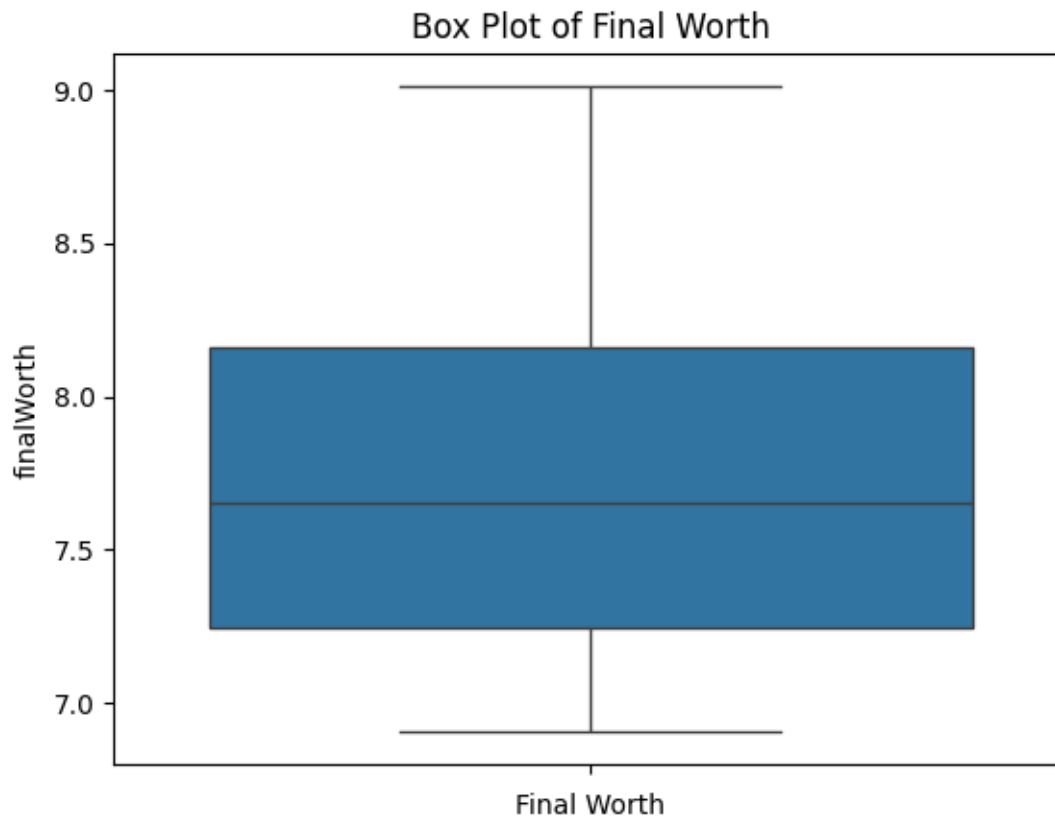


```
count    2392.000000
mean      2729.891304
std       1719.917126
min       1000.000000
25%       1400.000000
50%       2100.000000
75%       3500.000000
max       8200.000000
Name: finalWorth, dtype: float64
```



Then, I applied a logarithmic transformation to the target variable to reduce skewness and improve the model's performance.





2.1 Baseline

Baseline Mean Squared Error: 0.3208229296883327

2.2 Multiple Regression

```
Pipeline(steps=[('encode',
                  ColumnTransformer(transformers=[('num', StandardScaler(),
                                                  ['age', 'gdp_country',
                                                  'gross_tertiary_education_enrollment',
                                                  'gross_primary_education_enrollment_country',
                                                  'total_tax_rate_country',
                                                  'population_country']),
                                                  ('cat',
                                                  OneHotEncoder(handle_unknown='ignore'),
                                                  ['country', 'industries',
                                                  'selfMade', 'gender'])])),
                  ('regressor', LinearRegression())])
```

Training MSE: 0.28450735150404
 Test MSE: 0.306377843680439

	Feature	Coefficient
23	cat__country_Denmark	1.054455
50	cat__country_Nigeria	0.927925
20	cat__country_Colombia	0.914227
12	cat__country_Belgium	-0.822640
27	cat__country_Georgia	0.796127
..
84	cat__industries_Food & Beverage	0.009382
79	cat__industries_Construction & Engineering	-0.008449
98	cat__gender_F	0.003673
99	cat__gender_M	-0.003673
2	num__gross_tertiary_education_enrollment	-0.001635

[100 rows x 2 columns]

2.3 K-Nearest Neighbors Regression Model

Best Hyperparameters: {'model__n_neighbors': 50}
 Training MSE: 0.29610734432526276
 Test MSE: 0.3079731604666978

2.4 Random Forest

```
GridSearchCV(cv=5,
              estimator=Pipeline(steps=[('encode',
                                         ColumnTransformer(transformers=[('num',
                                                                              StandardScaler(),
                                                                              ['age',
                                                                              'gdp_country',
                                                                              'gross_tertiary_education_enrollment',
                                                                              'gross_primary_education_enrollment_country',
                                                                              'total_tax_rate_country',
                                                                              'population_country']),
                                                                              ('cat',
                                                                              OneHotEncoder(handle_unknown='ignore'),
                                                                              ['country',
                                                                              'industries',
                                                                              'selfMade',
                                                                              'gender'])])),
              ('model', RandomForestRegressor())]),
              param_grid={'model__max_depth': [3, 4, 5],
                          'model__n_estimators': [50, 100, 150]},
              scoring='neg_mean_squared_error')

{'model__max_depth': 4, 'model__n_estimators': 150}
```


Training MSE: 0.290871658418162

Test MSE: 0.28952704805141516

3 Findings

The linear model does well, with a training MSE of 0.2845 and a test MSE of 0.3064, both of which are lower than the baseline MSE of 0.3208, capturing meaningful patterns.

The K-Nearest Neighbors (KNN) model, with the optimal hyperparameter of 50 neighbors, has a training MSE of 0.2961 and a test MSE of 0.3080, both of which are lower than the baseline MSE of 0.3208. While the KNN model's training MSE is somewhat higher than the linear model's, both models show identical performance in terms of test MSE, implying generalization on unseen data.

The Random Forest model, with a training MSE of 0.2909 and a test MSE of 0.2895, performs similarly to the KNN model in terms of test MSE, with both models falling below the baseline MSE of 0.3208. The Random Forest model exhibits low overfitting, with fairly similar training and test MSE values, indicating that it generalizes effectively without overfitting the data.

When comparing the three models, the linear model has the lowest training MSE, while the KNN and Random Forest models show very similar performance in terms of their test MSE. The Random Forest model, being a more complex model, performs almost as well as the linear model, capturing more complex relationships in the data with minimal overfitting. All three models outperform the baseline, with the linear model slightly ahead in terms of training performance, while the Random Forest and KNN models perform competitively on the test data.

	Importance
age	0.028705
gross_primary_education_enrollment_country	0.021025
total_tax_rate_country	0.018651
country	0.012620
gdp_country	0.012492
industries	0.007034
gross_tertiary_education_enrollment	0.003719
selfMade	0.001631
population_country	0.001587
gender	-0.000126

The feature importance values indicate that Age (0.031516) is the most influential feature in predicting the target variable, followed by Total Tax Rate Country (0.022721) and Gross Primary Education Enrollment Country (0.019119), which also contribute meaningfully to the model.

4 Summary

Model Comparison: The linear model is the best performer in terms of training MSE but slightly lags behind Random Forest and KNN on the test data. Both Random Forest and KNN models offer similar test performance, with Random Forest slightly outperforming KNN.

Best Balance of Complexity and Performance: Random Forest strikes the best balance between

complexity and performance, capturing complex relationships while maintaining good generalization without overfitting.

Based on the test MSE and generalization ability, Random Forest emerges as the most well-rounded model for this dataset, though KNN and linear models also perform adequately and may be preferred for simpler, faster implementation in certain contexts. Further hyperparameter tuning and feature engineering could further enhance performance across all models.

5 Next Step/ Improvement

In the next steps, I plan to address several key areas to improve the analysis. First, I aim to update the dataset with the most recent data to reflect current trends, as the existing data is outdated and may not accurately capture the present dynamics. Additionally, I will ensure that all units of measurement are clearly defined and properly documented, as the current dataset lacks clarity on this front. This will enhance the accuracy and transparency of the analysis.

I also intend to expand the dataset by incorporating more relevant features, such as lifestyle-related variables, to deepen the insights. For instance, understanding whether billionaires are first-generation wealth creators, their educational backgrounds, and the industries contributing to their wealth can provide valuable context for the analysis. Exploring these factors will help capture more nuanced relationships within the data.

I plan to use advanced models like XGBoost, SVM, and Neural Networks, along with creating new features, to improve prediction accuracy.

Finally, once the updated data is analyzed, I aim to compare the findings with historical trends to uncover changes in wealth creation patterns, sectoral influences, and other emerging trends. These steps will ensure a comprehensive, accurate, and insightful analysis that better informs decision-making.