Exploring Billionaire Wealth Dynamics

By: Arpa Banik, Emily Lucia, Janhavi Maniar, Chen Zhang, Jiabei Zhao

Abstract

The analysis of the Billionaires Statistics Dataset serves a crucial purpose in unraveling the intricate web of global economic dynamics. By delving into the distribution of billionaire wealth in the US across diverse industries and regions, this study provides essential insights into billionaire trends and highlights sectors and regions where wealth concentration is most prominent. Furthermore, the examination of socio-economic factors such as inherited wealth contributes to a better understanding of the impact of societal conditions on billionaire wealth. This dataset encompasses information on global billionaires, offering statistics on their businesses, industries, and demographics. It yields valuable insights into the worldwide distribution of wealth, the various business sectors billionaires are involved in, and the demographics of this affluent group. In this analysis, our objective is to explore the factors influencing billionaire wealth. By examining the dataset, we will delve into the distribution of wealth across various industries and regions and evaluate the influence of socioeconomic factors on billionaire wealth. Our ultimate goal is to offer valuable insights into billionaire wealth dynamics, highlight geo-economic disparities, and identify potential areas for additional research. There have not been many similar studies on this topic, in part due to the dataset being compiled from various sources. Additionally, any studies that were completed focused primarily on different data visualization tools, such as charts and maps, without conducting statistical tests. Our methodology involves using advanced regression model building to reveal relationships among variables in the dataset. The study's achievements lie in its ability to shed light on unexplored facets of billionaire wealth dynamics. Through analysis, it unveils trends, patterns, and disparities, offering valuable insights for researchers and the public. This research contributes to the ongoing discourse on economic inequality and provides a foundation for future studies in the field. Ultimately, this study serves as a vital resource for understanding the complexities of billionaire wealth and fostering further research in the pursuit of a more equitable global economic landscape.

Index Terms

Billionaire, regression analysis, wealth, industries, regions, Unites States

Introduction

Under the current global economic situation, it is crucial to understand the dynamics of billionaires' wealth. In fact, billionaires play an important role in shaping economic policy, influencing market trends, and contributing to social development. Our analysis focuses on the distribution of U.S. billionaire wealth by industry, region, age, and self-made. Through the study, the aim is to contribute to the current debate on economic inequality and to identify the main

factors involved in achieving a more equitable global economic landscape. Grand Canyon University (2021) has researched a variety of profiles of wealthy Americans, providing major insights into the educational achievements of many wealthy Americans, the prevalence of starting from nothing, and their contributions to philanthropy. Unlike previous studies centered on data visualization tools, our research methodology utilizes statistical tests to attempt to examine billionaire wealth dynamics from a unique perspective. Specifically, linear regression allows for a comprehensive analysis of how multiple factors work together to influence the subject of the study. The interplay of individual impressionistic factors in billionaire wealth dynamics is further explored.

The Data Set

The data set for this regression analysis is called “Billionaires Statistics Dataset (2023)”, an exhaustive collection of data on billionaires around the world. The dataset is sourced from Kaggle.com, a well-known data science community and platform. This data source was chosen because it provides comprehensive coverage and in-depth information on each billionaire's wealth, demographics, and geographic data.

We consider this Dataset to be a "Cross-sectional dataset". From a point-in-time perspective, our Dataset reflects the wealth of billionaires and other relevant information at a specific point in time in 2023. From the perspective of sample diversity, we can see that the dataset contains different billionaires from different countries and industries. Cross-sectional datasets typically focus on the characteristics of different samples at the same point in time, rather than changes in a single sample over time.

The dataset includes 2,640 records with 35 different attributes, including information such as each billionaire's ranking, final wealth, category, personal details (such as age, country, and city), and industries, the final wealth is measured in Millions. In this dataset we focused on data from 754 billionaires in the United States. To ensure the accuracy and relevance of the analysis, we carefully processed and selected the data set, focusing on the following variables: Age, self-made wealth, Industry, Region, and Final Worth. Through a combined analysis of these variables, our research aims to uncover the key factors that influence the wealth of billionaires in the U.S. and explore how these factors play out across different individuals and environments. In addition, we pay special attention to the contrast between self-created wealth and inherited wealth, as well as the wealth composition of billionaires in different industries and regions. Through this research, we hope to provide a comprehensive perspective for understanding the wealth dynamics of billionaires in the U.S.

During data processing, we found missing values in some columns of the dataset, such as "age", "country" and "city" fields. Since we only focused on billionaires in the U.S., there are fewer missing values in these fields, and we chose to simply remove those missing values. Since the dataset includes billionaires from different countries and industries, it shows a high degree of diversity. Therefore, as mentioned in the introduction, we conduct a subset analysis for the United States and narrow the research object to the top five industries, which are "Technology", "Finance & Investments", "Fashion & Retail", "Technology". "Food & Beverage" and "Real Estate."

By delving into the data of these top U.S. industry billionaires, we aim to reveal the major factors influencing the wealth of this group, providing a comprehensive perspective to understand the wealth dynamics of American billionaires.

Theory and Methods

Regression analysis is a statistical tool that allows for the examination of relationships between two or more variables. It serves two primary purposes: to predict the value of a dependent variable based on knowledge of independent variables or to describe the effect of independent variables on a dependent variable. The general model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \in$$

Here, y is the dependent variable that is being solved or examined. The intercept is represented by β_0, and the coefficients of the independent variables are the β_1 through β_k values. The values of the actual variables are represented by x_1 through x_k. In our final mode, the dependent variable is the final worth of the billionaires, and the independent variables are their age, their involvement in the Technology industry, their involvement in the Fashion & Retail industry, the billionaire's residence in the West, and an interaction between involvement in the Technology industry and residence in the West. The final term in the general model, $\in$, is the error term. It acknowledges and constitutes the margin of error that is present between the theoretical model and the actual observed results. The model provides a broad understanding of patterns and behaviors of the data. However, the fitted model is better suited for more precise and accurate predictions because it is specifically fitted to the dataset. This helps minimize the difference between the predicted and actual results. The fitted model is depicted as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$

The pieces of this model represent similar values as the general model, but for a specific observation. So, $\hat{y}_i$ is the value of the dependent variable for the ith observation. $\hat{\beta}_0$ is still the intercept and β_1 through β_k are the coefficients for the independent variables ($x_{1i}, x_{2i}, \ldots x_{ki}$). The β values will remain the same between the two models. The x_{ki} values are the values of the independent variables for the ith observation.

Before utilizing the model, the data must meet the four assumptions of regression analysis: normality, linearity, constant variance, and independence. The normality assumption states that the population must have a normal distribution to avoid inaccuracy due to skewedness. The linearity clause requires the independent and dependent variables to have a linear relationship, meaning that change in one variable will lead to a proportional change in the other. The

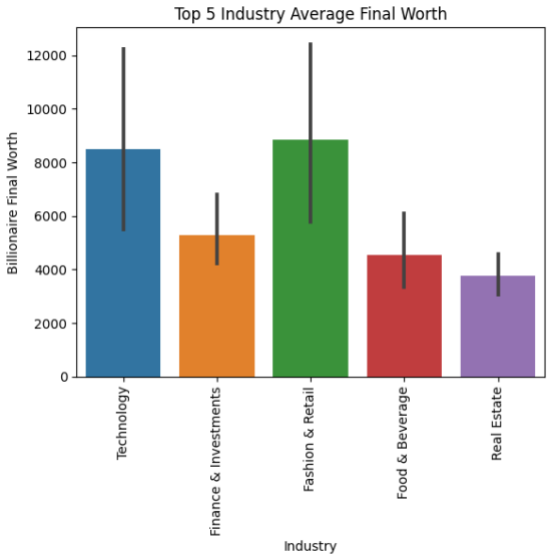
assumption of constant variance, also known as homoscedasticity, outlines that the spread of residuals should be the same at each level of the predicted response to discourage unbiased estimators of parameters. Lastly, the independence assumption simply states that the observations should be independent of each other. In our Data Analysis section, we will discuss how our data met these requirements.

Data Analyses

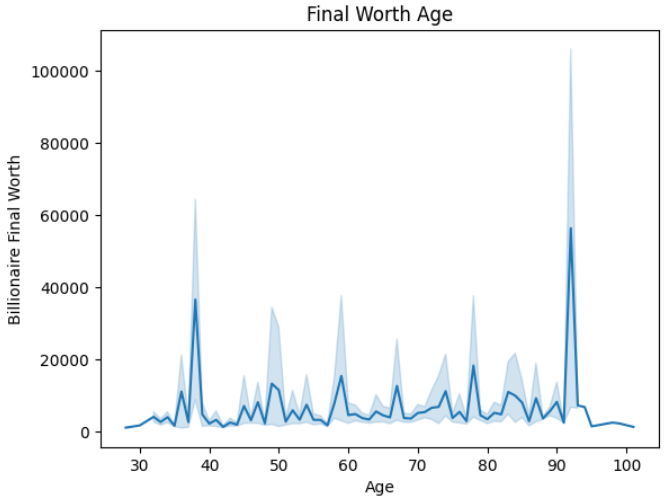
Exploratory Data Analysis (EDA)

To analyze the data, we can start with graphs and charts to understand the causal relationship between the variables. The following is the analysis of the charts that we found meaningful：

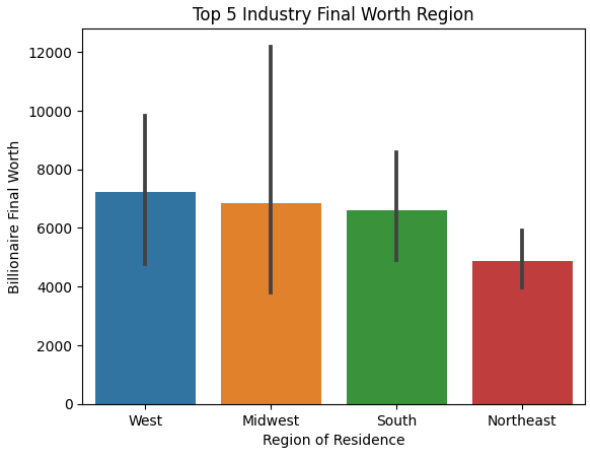
From the bar plot of top 5 industries and final worth, we can see that Fashion & Retail and Technology are the dominant industries that create the most final value. Followed by Finance & Investment, Food & Beverage and Real Estate.



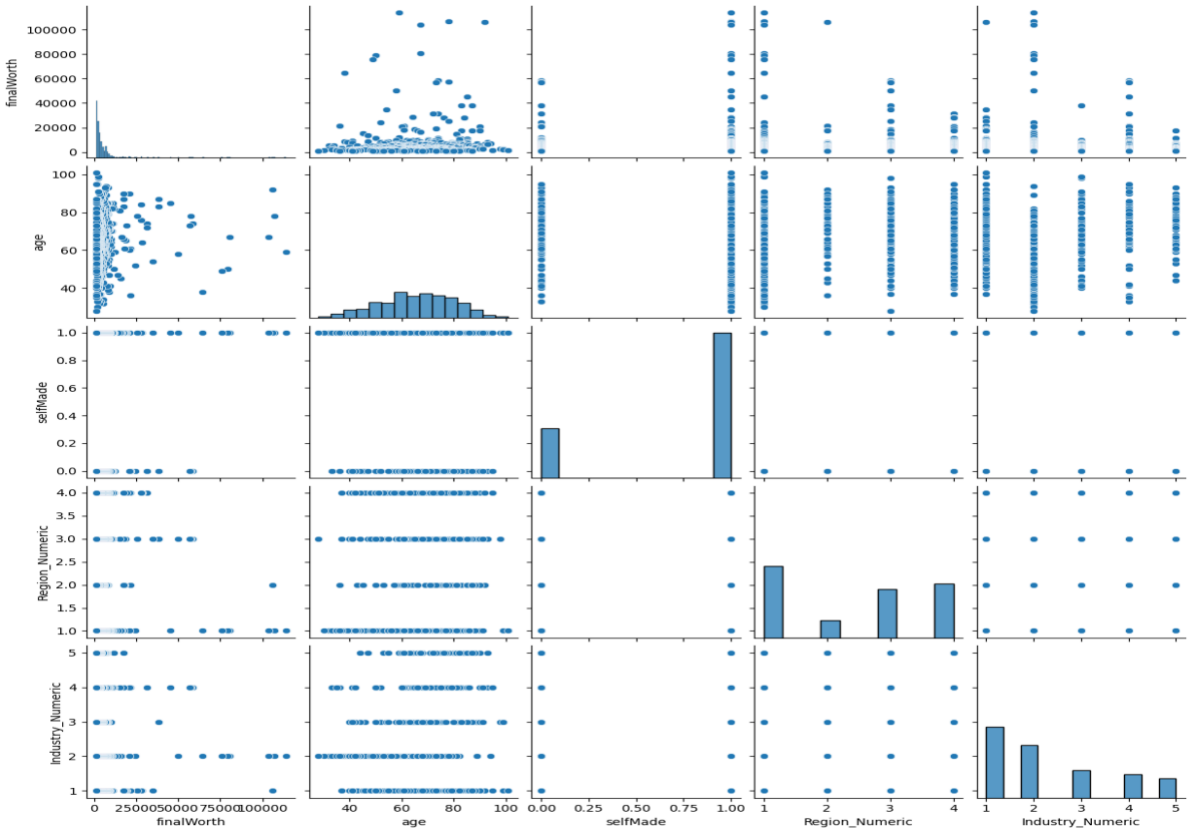
The line chart shows the relationship between the age of billionaires and their final worth. There are two distinct peaks in the graph. One of the larger peaks occurs around the age of 90. This suggests that although the age range of billionaires is large, the higher age may correspond to a higher final worth. This may reflect the cumulative effect of wealth growth over time, or it may reflect the fact that a very small number of billionaires have exceptionally high stature at that age. Fluctuations in the lower age range (30-40 years old) may indicate the dynamic nature of the wealth of young billionaires.



The next bar chart compares the final values of top five major industries in four different regions of the United States: the West, the Midwest, the South, and the Northeast. The West leads the way in terms of final value, followed by the Midwest and the South, leaving the Northeast with the lowest value of the four regions.



Additional causality maps we generated can be found in the Appendix.

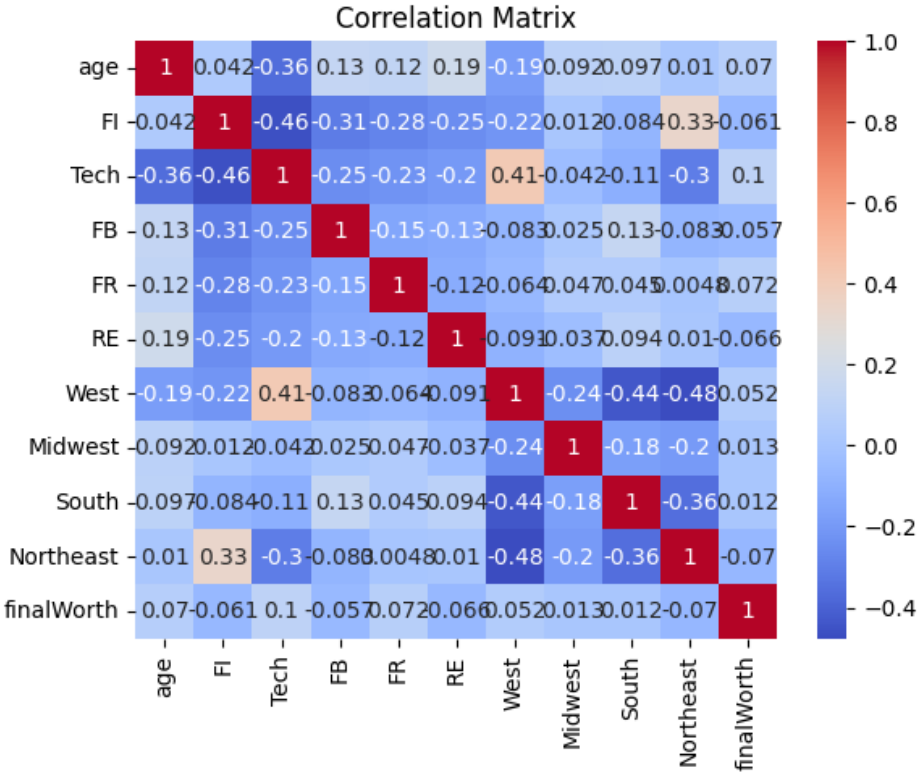


To gain an understanding of the connections between our variables and visually assess patterns and potential clusters we utilized Scatter Cloud Matrices. This technique allows us to explore and comprehend the pairwise interactions among our variables; age, self-made status, industry categories (Finance & Investments Technology, Food & Beverage, Fashion & Retail Real Estate), and residence regions (West, Midwest, South, Northeast).

Contrary to our expectations our analysis reveals that there is no association between age and final worth among billionaires. The scatter cloud matrix and correlation analysis suggest that there is no trend in this regard. This challenges the held belief of a correlation between age and financial success. This unexpected finding prompts us to reevaluate the role of age in predicting the worth of billionaires in our dataset. Similar to what we observed with age, our analysis indicates that there is no relationship between self-made status and final worth. Although self-made billionaires make up a portion of our dataset, their financial success does not follow a pattern. The scatter cloud matrix doesn't provide a path indicating that it's important to explore the factors that influence someone’s overall worth, beyond just being self-made. When we looked at the scatter cloud matrices for regions where people live, we discovered patterns. Specifically,

the West region had clusters of individuals with worth suggesting that where someone lives may have an impact on their overall worth.

The scatter cloud matrices also hinted at interactions between industry categories and residence regions. To capture these relationships and improve our model's ability to explain things we included interaction terms in model iterations. We also identified outliers and influential data points by examining the scatter plots. This information helped us make decisions about how to handle these observations during the modeling process.



During our investigation of the dataset, we thoroughly examined the pairwise correlation matrix to measure the connections between variables and most importantly their relationship with individuals' final worth.

We noticed a correlation between age and final worth suggesting that on average older individuals tend to have a higher net worth. The Finance & Investments (FI) sector demonstrates a correlation with worth indicating that individuals in this industry may have a lower net worth on average. However, Technology (Tech) exhibits a correlation implying that being involved in the technology sector may positively impact an individual's final worth. The correlation with residence regions varies significantly emphasizing the significance of location when assessing one's worth. For example, residing in the West region shows a correlation with a higher net worth.

These insights derived from correlations helped us select features and construct interaction terms for our regression analysis. Variables showing correlations, with worth were

given more priority in our model to improve predictive accuracy. In the following sections we will explore the analysis of regression demonstrating how well the model performs and its capacity to capture the relationships observed in the correlation matrix.

Assumptions

An important factor in our analysis of regression is to consider and validate assumptions. The reliability of our results depends on these assumptions being true. Below we will discuss each assumption. Explain how our analysis aligns with them.

1. Linearity- The linearity assumption suggests that the relationship between the variables and the dependent variable is linear. In our analysis we have used regression models ensuring that this assumption is met. We have used the variables in their form without any transformations, which maintains the linearity assumption.

2. Independence of Residuals- The independence of residuals assumes that there are no correlations among the errors. To test for autocorrelation, in residuals we have examined them using the Durbin Watson statistic. Although the Durbin Watson value falls below the range indicating autocorrelation further investigation is required. To address this concern incorporating time series analysis or additional variables can account for any dependencies.

3. Homoscedasticity- Homoscedasticity implies that the variance of residuals remains constant across all levels of variables. Our residual analysis and scatter plots indicate variance. While there are no signs of heteroscedasticity it may be worth exploring standard errors in future iterations to account for potential heteroscedasticity.

4. Normality of Residuals- The assumption of normality in residuals implies that the errors follow a distribution. Our analysis of the histogram suggests a departure from normality as indicated by skewness and kurtosis. However, it's important to note that for sample sizes the central limit theorem supports the reliability of our estimates. Furthermore, hypothesis testing is generally robust when normality assumptions are violated in datasets.

Model Fitting

Brute force variable selection is a method used in regression analysis that involves exploring all combinations of independent variables to find the subset that best fits the model. It entails evaluating the performance of each combination based on criteria, such as adjusted R-squared F-statistic or other relevant metrics. Despite being computationally intensive this approach ensures a search for predictor variables.

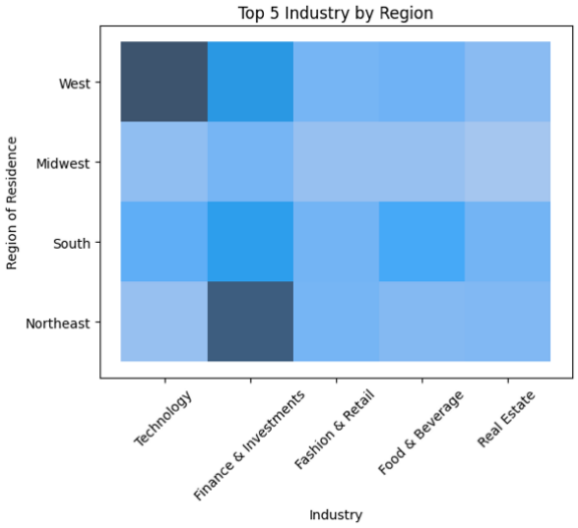
Our initial analysis using force included considering variables like 'age' 'FI' (Finance & Investments) 'Tech' (Technology) and 'FR' (Fashion & Retail). The outcome of this selection process resulted in a model, with an adjusted R value of 0.027 and an associated F statistic value of 4.620 indicating an improvement compared to a null model. However, this model had limited

power overall as shown by the R squared value of 0.035. The variables 'age,' 'Tech,' 'FR' and 'West' were included in this model with 'Tech' having a positive impact on final worth.

Even though individual variables showed significance we didn't consider this model as the choice due to its relatively low explanatory power and minimal improvement compared to the null model. The adjusted R squared, which considers the number of variables in the model, also fell below our desired threshold.

When considering the model, we recognized the need for refinement beyond brute force selection. We explored interactions and other relevant variables to enhance the model's ability to explain. The iterative process of developing the model and incorporating insights played a role in selecting a final model that captured the intricate relationship, between variables and billionaires' final worth. In sections we delve into how our analysis progressed from brute force selection to achieving performance and addressing key assumptions.

After understanding the variables that the brute force method finds significant, it is important to understand what this method may be missing in the model, such as different interactions and variable significance. We can see from the brute force model output that Finance and Investment was not a significant predictor as the p-value was 0.261. All remaining variables from the brute force model, age, technology, and fashion and retail, were left in the model. The reason for introducing West into the model was due to an important interaction found in our initial stages. This important interaction to consider in the dataset is the industries relations to different regions in the United States. We can see the different relationships with this heatmap:



We can see from this graph that the majority of those in the technology industry tend to be from the Western region and those in the finance and investment industry tend to be from the Northeast. These connections make sense as the West is known for Silicon Valley in California being a technology center and Wall Street in the Northeast being a trading center. Using this information on interactions, we will test them being input to the brute force model.

For the level two complexity of the model, we introduced the variable West making our list of predictor variables age, technology, fashion and retail, and west. The output of this model gave a lower R^2 adjusted value, 0.025, compared to the brute force which was 0.027. We look towards the R^2 adjusted value compared to the traditional R^2 value because it brings into consideration the number of variables whereas the traditional R^2 will always increase with added variables.

With this information, the next level of complexity was to add in the interaction between technology and the West making our list of predictor variables age, technology, fashion and retail, West, and technology interaction with the West. This output gave a more desirable R^2 adjusted value of 0.043. This added interaction, however, did cause our variables West and technology to become insignificant, but we are okay with this as they are included in the interaction which is significant. This shows that the interaction between technology and the West is important when fitting our model and should be included.

Level four complexity involved introducing back in the finance and investments variables along with adding in the Northeast variable as well. This was because there was a high connection between these two variables seen in the heatmap that is important to test out. After adding these variables to our model and running the test, we see that it did increase our R^2 adjusted value slightly to 0.044 compared to level three complexity of 0.043. This, however, came at a cost because neither variable, the Northeast nor finance and investments were found to be significant in predicting final worth with p-values being 0.238 and 0.18 respectively.

Since this new model did increase our R^2 adjusted value, level five complexity was introduced which added an interaction variable between the Northeast and the finance and investments industry. Adding this term, however, did hurt our model as our R^2 adjusted value decreased to 0.042 which is below level three and level four complexities. We also continue to see finance and investments, the Northeast, and the interaction between them to be insignificant. This caused a reconsideration of keeping the variables Northeast, finance and investments, and the interaction between the two in the model.

The final decision for the model was using the level three complexity model. This model included the variables age, technology, fashion and retail, West, and interaction between technology and the West. This gave us the final fitted line model:

Final Worth = -2471.83 + 115.51(age) - 260.5(Tech) + 3881.8(FR) - 2321.4(West) + 9007.4247(Tech & West)

This model reveals the predicted final worth of a billionaire given their age, if they are in the technology industry or the fashion and retail industry, and if they are in the Western region of the United States. We cannot interpret the intercept of the fitted line as it would be impossible to include someone of zero years of age and for their final worth to be -$2.471 billion as this would be extrapolation. In terms of the age variable, our model tells us that as the billionaire increases one year in age, their wealth will increase by $115.51 million. It also shows how if you are in the fashion and retail industry, your wealth increases by $3.881 billion. In terms of the interaction between technology and the West, we see that if you are in the technology industry but not in the

West, your wealth decreases by $260.5 million. Inversely, if you are in the West but not in the technology industry, your wealth decreases by $2.321 billion. If you are in both the technology industry and the West, however, your final wealth increases by $9.007 billion.

Quality Check

Final Model

Final Worth = -2471.83 + 115.51(age) - 260.5(Tech) + 3881.8(FR) - 2321.4(West) + 9007.4247(Tech & West)

OLS Regression Results

Dep. Variable:	finalWorth	**R-squared:**	0.052
Model:	OLS	**Adj. R-squared:**	0.043
Method:	Least Squares	**F-statistic:**	5.634
Date:	Sun, 03 Dec 2023	**Prob (F-statistic):**	4.55e-05
Time:	20:16:16	**Log-Likelihood:**	-5624.6
No. Observations:	518	**AIC:**	1.126e+04
Df Residuals:	512	**BIC:**	1.129e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2471.8316	2973.499	-0.831	0.406	-8313.592	3369.928
age	115.5101	41.862	2.759	0.006	33.267	197.753
Tech	-260.5003	2121.066	-0.123	0.902	-4427.563	3906.563
FR	3881.8483	1749.135	2.219	0.027	445.484	7318.212
West	-2321.4203	1512.361	-1.535	0.125	-5292.617	649.777
TECH & West	9007.4247	2767.700	3.254	0.001	3569.978	1.44e+04

Omnibus:	568.492	**Durbin-Watson:**	0.107
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	24952.949
Skew:	5.219	**Prob(JB):**	0.00
Kurtosis:	35.360	**Cond. No.**	412.

Brute Force Model

Final Worth = -4009.4267 + 115.16(age) + 115.156(FI) + 5944.27(Tech) + 4781.03(FR)

```
Selected Features (Brute Force): ['age', 'FI', 'Tech', 'FR']
```

OLS Regression Results

Dep. Variable:	finalWorth	**R-squared:**	0.035
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	4.620
Date:	Sun, 03 Dec 2023	**Prob (F-statistic):**	0.00113
Time:	20:01:58	**Log-Likelihood:**	-5629.3
No. Observations:	518	**AIC:**	1.127e+04
Df Residuals:	513	**BIC:**	1.129e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4009.4287	3252.479	-1.233	0.218	-1.04e+04	2380.389
age	115.1559	42.594	2.704	0.007	31.476	198.836
FI	1673.5442	1486.960	1.125	0.261	-1247.736	4594.824
Tech	5944.2747	1684.875	3.528	0.000	2634.171	9254.379
FR	4781.0307	1970.957	2.426	0.016	908.891	8653.170

Omnibus:	578.073	**Durbin-Watson:**	0.075
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	26616.570
Skew:	5.350	**Prob(JB):**	0.00
Kurtosis:	36.447	**Cond. No.**	428.

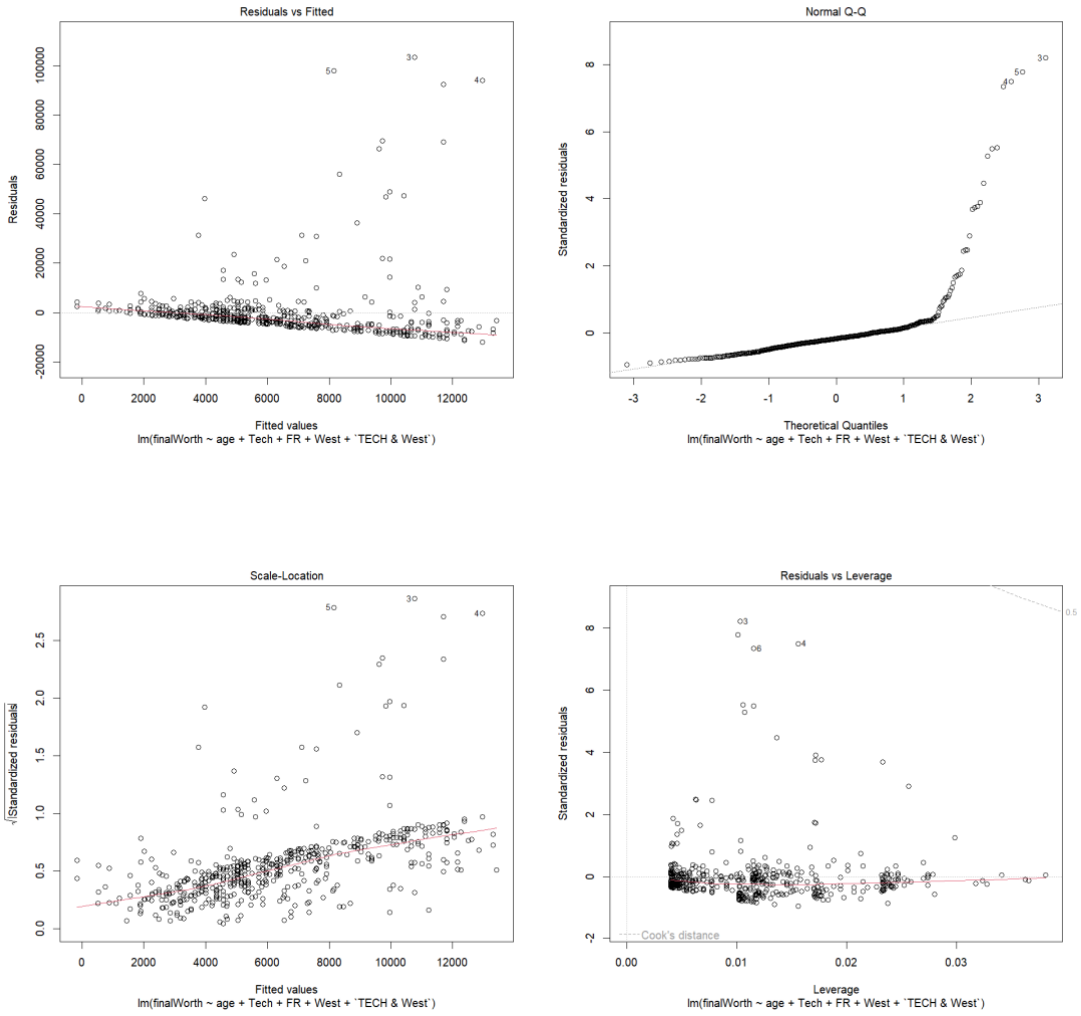
It is important in our analysis to ensure that the fitted model we have is of high quality and does not appear to overfit or extrapolate the data. To ensure our model is a quality predictor of final worth among billionaires, we must analyze many factors of the level three quality against the factors of the brute force, level one complexity, model. The steps we will go through to ensure our model is sufficient are checking the R^2 and adjusted R^2 values, low MSE values, the global F and individual T-tests are passed, residuals are good, there is a low PRESS value, there is no multicollinearity, and the model is overall logical. All these factors combined will show which model is a more accurate predictor.

Checking the R^2 and adjusted R^2 values were checked during the model fitting stage and the model chosen had the second-highest adjusted R^2 value, 0.043, throughout our testing. Our final model, level three complexity, was chosen despite not having the highest adjusted R^2 because the level four complexity model had more insignificant variables added. The brute force model, on the other hand, had a much lower adjusted R^2 value, 0.027. This shows that the final model predicted the outcome accurately 4.3% of the time compared to the brute force model which was accurate only 2.7% of the time.

The next factor of importance is ensuring the model has a comparatively low MSE, mean squared error, value. The MSE informs the average squared difference between the observed and predicted values. We look for the model with the lower MSE as it will have lower amounts of error when predicting. When comparing the MSE of the final model versus the brute force, 159,967,522 and 162,583,032 respectively, we see that the final model had a lower MSE value, making it the better prediction model using this factor compared to the brute force model.

The next two tests, the global F-test and individual T-test, were both checked in the model fitting stage like the adjusted R^2 values. The global F-test indicates whether or not there is one good variable present in the model. The F-test for both the final model and brute force model were significant in that their p-values were 0.000 and 0.001 respectively, meaning that both models had at least one significant variable. In more depth with the individual T-tests, we found that the brute force model used finance and investment in their model while having a p-value over our alpha, 0.05. All other variables used in the brute force model, however, were significant in predicting the final worth of a billionaire. In terms of the final model, all variables were found to be significant excluding technology and the West. This was not a surprise, however, because the model includes an interaction between technology and the West which can cause their independent variables to become insignificant. It is important to still include them since they relate to the interaction.

Final Model Residual Plots



When analyzing the residual trends formed from the final model, we see results consistent with a model that is of good quality. Specifically looking at the residuals versus fitted plot, we see the majority of the points spread along the x-axis. This shows that our predicted values are like the actual values in that our model is an accurate predictor. The outliers seen in this plot can be described as the 'mega-rich' that fall even outside the standards of being a billionaire. Next is the Q-Q plot which describes the distribution of the data. In our Q-Q plot, we see that our points all fall on the desired 45° angle line until you reach the end. This is like the last outcome in that this change in pattern is the mega billionaires who fall outside of the normal range.

We see these patterns continue in the Scale-Location plot, where we see many of the points along the desired 45° angle line with some outliers. In the residuals versus leverage plot, we see large groupings along the x-axis with few outliers remaining outside this area. All this information together shows that our data is a relatively normal distribution with some outliers on

the higher side of the final worth. It was in our best interest to keep these outliers in the model, however, because they are important in showing where the most profitable industries are and how this can impact other billionaires that fall close behind them.

The next measurement in checking the quality of the models is understanding their PRESS values. Lower PRESS values indicate better predictive performance in the model. Comparing the final model to the brute force model we see the PRESS values are 83,993,802,376 and 85,146,419,167 respectively. From these numbers we see that the final model has a lower PRESS value comparatively showing it has better predictive performance. Next is looking into multicollinearity which is when there are many strong linear relationships between the independent variables. The measurement used for this analysis is VIF, variance inflation factor. The VIF is approximately how much the variance of the estimated regression coefficients is changed due to multicollinearity. The results for the final model and brute force model are:

Final Model:

	Variable	VIF
0	const	28.630802
1	age	1.160480
2	Tech	2.886072
3	FR	1.058368
4	West	1.723973
5	TECH & West	3.804976

Brute Force Model:

	Variable	VIF
0	const	33.704168
1	age	1.182066
2	Tech	1.791806
3	FI	1.632493
4	FR	1.322212

When analyzing the VIF results for both models, we see that besides the intercept, no other values are affected by multicollinearity. If the values were above 10, then we would have reason to believe multicollinearity was affecting the results of the model.

The final factor that goes into understanding the quality of the model is whether the variable coefficients seem reasonable and logical. For this analysis, we will only need to see if the final model is logical as this is the model we have chosen for the most accurate predictions. It makes sense for increasing age to go along with increasing final wealth by $115 million because individuals tend to gain more money over time naturally through business growth. When looking at fashion and retail, we saw previously that those in fashion and retail tended to have higher final worths than other industries. This proves our model to be logical as it increases the billionaire's wealth by $3.881 billion if they are in the fashion and retail industry.

Implementation

At first glance, it does not make sense as to why the model says that if you are in the technology industry or live in the West your final worth would decrease by $260 million or $2.321 billion respectively, unless you consider the interaction. If you live in the West and are not in the technology industry, you are not part of the higher-profit industry of that area, so your

final wealth goes down. Similarly, if you are in the technology industry but you are not in the West, your final worth decreases because you are not in the location where those in the technology industry are more likely to prosper. If you are in the technology industry and live in the West, however, your final wealth will increase by $9.007 billion because you are in the prime location for that industry.

When using our final model to predict a billionaire's worth, it is important to compare two similar individuals and compare the differences between predictions when we change individual factors. In this case, we will compare the differences between two 65-year-old billionaires in the technology industry, with one being from the Northeast and the other being from the West. We maintain the age of 65 between the billionaires to ensure we are analyzing the difference in geographic location.

When utilizing the model on two 65-year-old billionaires in the technology industry from the Northeast and the West, we predict their final worth to be $2.714 billion and $11.461 billion respectively. That is an $8.747 billion difference in final worth between these two billionaires despite the only difference being their location.

We can understand from the model that the interaction of the West and the technology industry is applied to the second individual causing an increase of $9.007 billion. There was also an offsetting decrease, however, from the individual West variable of $260.5 million. These two changes together gave us the overall increase of final worth between two like individuals in the technology industry with the distinguishing factor being location in the Northeast or the West.

Conclusion

By studying the wealth dynamics of billionaires in U.S., we found that the tech industry and the West region of United State have a significant impact on their ultimate wealth. However, there are limitations to this analysis. Because wealth is a dynamic indicator and susceptible to changes over time, our dataset may not fully capture this. In addition, focusing only on U.S. billionaires and their ultimate wealth may not capture the full complexity of global wealth accumulation.

To address these limitations, future research should include longitudinal analyses to track the evolution of billionaires' wealth over time. In addition, expanding the analysis to include international billionaires will provide a more comprehensive understanding of global wealth patterns, providing a more nuanced perspective on global wealth dynamics. In addition, we plan to introduce more factors that affect wealth, such as political and economic policies, and explore the impact of segmented industries. At the same time, it analyzes how social and cultural factors affect the accumulation and distribution of wealth, and deeply understands the problem of wealth inequality.

References

Giriyewithana, N. 2023. Billionaires Statistics Dataset [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset/data

An Analysis of America's billionaires. GCU. (n.d.). https://www.gcu.edu/blog/gcu-experience/facts-about-rich-americans

Appendix

Figure 1 – Final Worth by Gender (EDA)

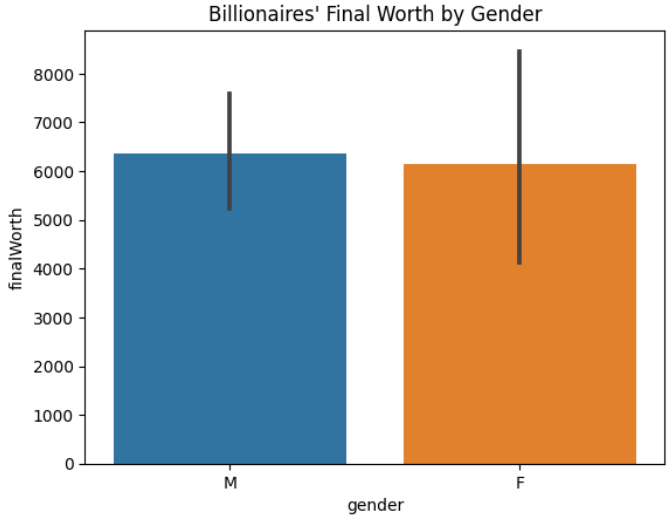


Figure 2 – Final Worth by Self-Made Status (EDA)

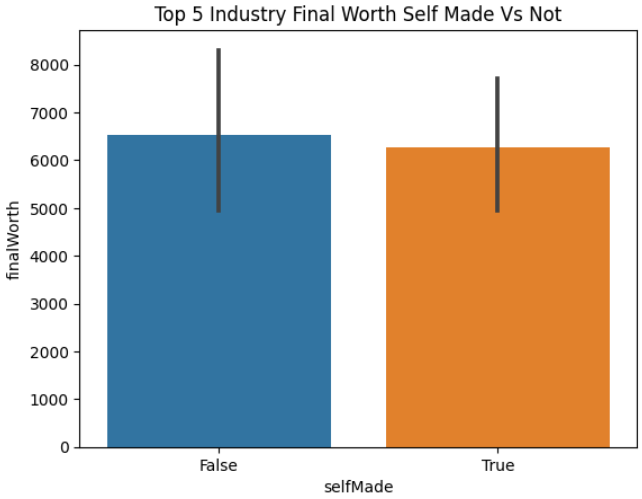


Figure 3 – Self-Made by Industry (EDA)