



Money Matters: Investigating the
Relationship between Economic
Indicators and Billionaire Presence using
Machine Learning

Kacper Fila
Eryk Czop
Jakub Banaś

1. Introduction

The driving force behind this project is a curiosity about the complex links between a country's economic well-being and the presence of billionaires. By analyzing key economic indicators, we hope to discover patterns that can help us understand the conditions that foster the rise of billionaire fortunes. These insights can benefit policymakers, economists, and analysts by providing tools for better forecasting and decision-making.

The main goal of this project is to take a closer look at the relationship between selected economic indicators and the prevalence of billionaires in a specific country. Using statistical analysis and data-driven methods, we want to uncover trends, correlations, and potential factors that contribute to extreme wealth concentration. Hypothesis we aim to prove: „Using economic indicators we can predict number of billionaires in particular country.”

This report is organized to provide a thorough exploration of the chosen economic indicators and their impact on the billionaire landscape. After this introduction, we'll dive into the methods we used, where our data came from, and why we picked these specific indicators. Following that, we'll present our findings and analysis, discuss what it all means, and talk about the limitations of our study. Finally, we'll wrap up with a summary of our key discoveries.

2. Dataset description and related work

Our decision leaned towards the Billionaires Statistics Dataset (2023) . For the past few weeks we have worked on it, doing data analysis and implementing machine learning and data science methods.



Pic. 1. Dataset logotype.

“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.” – we can read on source page of this dataset. Here is a breakdown of the dataset 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.

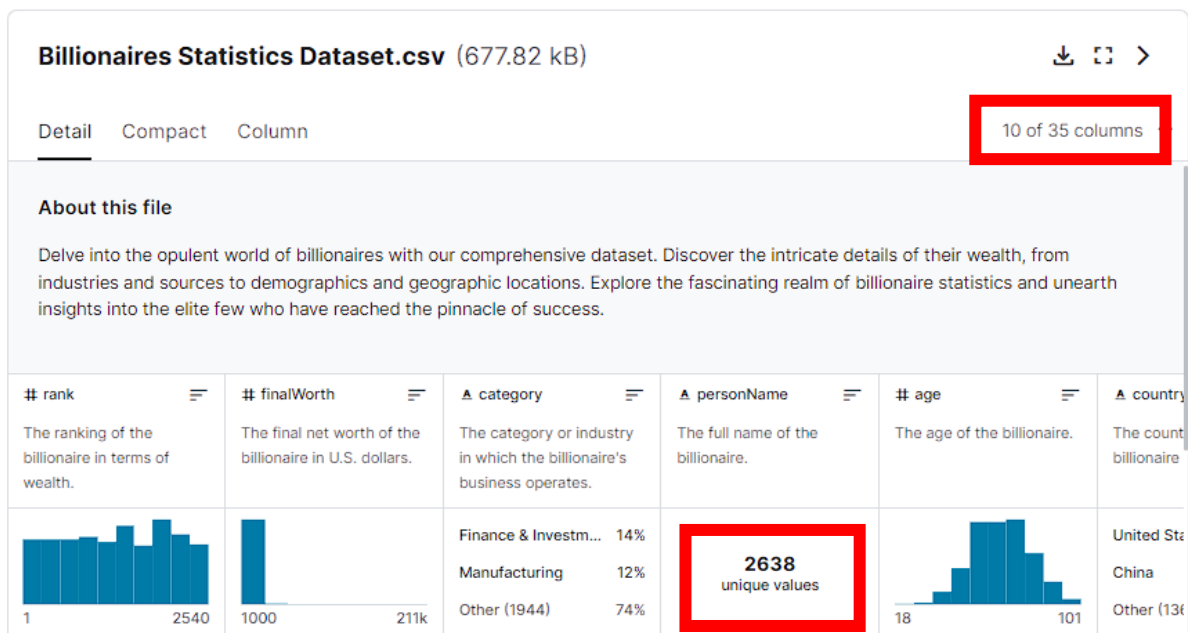
total_tax_rate_country: Total tax rate in the billionaire's country.

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.

This dataset contains 35 features and 2638 rows – number of all billionaires to the day that data was gathered and put into dataset.



Pic. 2. Dataset record.

People have been using this dataset in a well-rounded way—creating visuals for a clear overview, exploring the data for deeper insights, and cleaning it up to ensure accuracy. This multifaceted approach enhances the overall reliability and usefulness of the dataset for various analytical purposes. At the day we did our research there was 26 users that contributed their work on dataset to the main page of this repository.

Source: <https://www.kaggle.com/datasets/nelgiriyeewithana/billionaires-statistics-dataset>

3. Exploratory data analysis

So, EDA is our starting point — our way of getting to know the dataset, spotting interesting bits, and deciding where to dig deeper. It's like the first chapter in our data exploration journey. We take a close look at things like averages and spreads to understand where the data tends to gather and how it spreads out. Imagine it as using charts and graphs to paint a picture of how the numbers dance together. We're not just crunching numbers; we're creating visual snapshots that help us see relationships and potential interesting spots in the data.

3.1. Data preparation and cleaning

In this phase of preparing and cleaning our data, we're focusing on simplifying things by narrowing down the number of features to keep the attributes that bring valuable insights to the table. Instead of dealing with everything, we want to keep only the most important ones. The goal of feature selection is to make our analysis more straightforward, efficient, and focused. By reducing the number of features, we not only save computational resources but also make sure that we're working with information that really matters for our analysis and also to prevent feature leakage.

Rank		PersonName	Country	GDPperCapita	TotalTaxRate	CPI
0	1	Bernard Arnault & family	France	40493.928572	60.7	110.05
1	2	Elon Musk	United States	65280.682241	36.6	117.24
2	3	Jeff Bezos	United States	65280.682241	36.6	117.24
3	4	Larry Ellison	United States	65280.682241	36.6	117.24
4	5	Warren Buffett	United States	65280.682241	36.6	117.24
...
2635	2540	Yu Rong	China	14244.677921	59.2	125.08
2636	2540	Richard Yuengling, Jr.	United States	65280.682241	36.6	117.24
2637	2540	Zhang Gongyun	China	14244.677921	59.2	125.08
2638	2540	Zhang Guiping & family	China	14244.677921	59.2	125.08
2639	2540	Inigo Zobel	Philippines	3485.084218	43.1	129.61
2640 rows × 6 columns						

Pic. 3. New data frame consisting of necessary features.

After above operations, we had to ensure that our dataset did not contain NaN (Not a Number) values. It is essential for maintaining data integrity and ensuring accurate and reliable analyses. As you can see below, our number of rows decreased by 184.

	Rank	PersonName	Country	GDPperCapita	TotalTaxRate	CPI
0	1	Bernard Arnault & family	France	40493.928572	60.7	110.05
1	2	Elon Musk	United States	65280.682241	36.6	117.24
2	3	Jeff Bezos	United States	65280.682241	36.6	117.24
3	4	Larry Ellison	United States	65280.682241	36.6	117.24
4	5	Warren Buffett	United States	65280.682241	36.6	117.24
...
2635	2540	Yu Rong	China	14244.677921	59.2	125.08
2636	2540	Richard Yuengling, Jr.	United States	65280.682241	36.6	117.24
2637	2540	Zhang Gongyun	China	14244.677921	59.2	125.08
2638	2540	Zhang Guiping & family	China	14244.677921	59.2	125.08
2639	2540	Inigo Zobel	Philippines	3485.084218	43.1	129.61
2456 rows × 6 columns						

Pic. 4. Cleaned data.

We also checked for duplicates, but there was none. Last operation of acquiring the dataset that we wanted and needed was to correlate number of billionaires with each country – in other words finding **target value**.

PersonName	
Country	
Algeria	1
Argentina	4
Armenia	1
Australia	43
Austria	11
...	...
United Arab Emirates	17
United Kingdom	82
United States	754
Uruguay	1
Vietnam	6
64 rows × 1 columns	

Pic. 5. Series of countries correlated with number of billionaires.

Lastly, we assigned each of economic indicators (GDP per Capita, CPI, Total Tax Rate) to particular countries. This gave us our final dataset that we worked with.

	Country	GDPperCapita	CPI	TotalTaxRate	AmountOfBillionaires
604	Algeria	3948.343279	151.36	66.1	1
554	Argentina	10006.148974	232.75	106.3	4
2375	Armenia	4622.733493	129.18	22.6	1
51	Australia	54049.828812	119.80	47.4	43
36	Austria	50277.275087	118.06	51.4	11

Pic. 6. Final dataset.

3.2. Data splitting

In the data splitting stage, we partition our dataset into three subsets: training, validation, and test sets, with a ratio of 60/20/20. This division is a preferred option because it allows us to train our model on a substantial portion of the data (60%), fine-tune its performance using the validation set (20%), and ultimately assess its generalization on unseen data using the test set (20%). This balance helps prevent overfitting, where the model becomes too tailored to the training data, and ensures that its performance is robust and applicable to new, unseen information.

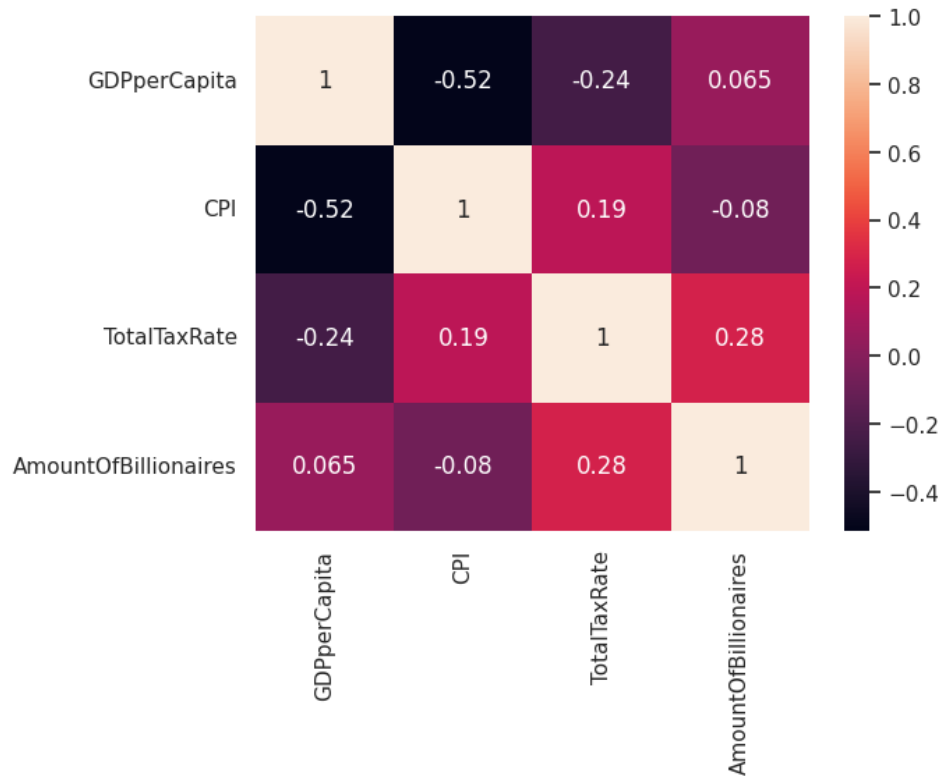
```
[ ] # Define features and target variable
X = df_country[['GDPperCapita', 'CPI', 'TotalTaxRate']]
y = df_country['AmountOfBillionaires']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=20)
X_train, X_val, y_train, y_val = train_test_split(X_train, y_train, test_size = 0.25, random_state=20)
```

Pic. 7. The best language of programming in the world in action.

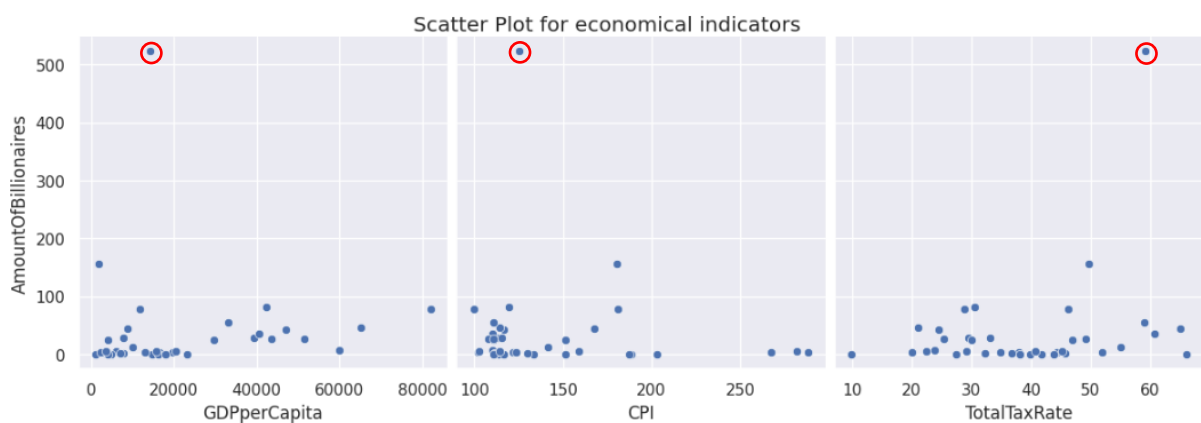
3.3. Data visualization

Our examination of the correlation matrix brought insights into the strength and direction of relationships between variables. Surprisingly, the correlation matrix indicated a lack of significant correlation between the number of billionaires and the chosen economic indicators. This finding challenges initial expectations of a clear connection and emphasizes the importance of visualizations in uncovering nuances and outliers that may not be apparent from raw data alone.



Pic. 8. Correlation matrix.

Despite of fact, that correlation matrix above might suggest that economic indicators are not the best suit for our task, we did further data analysis and decided to try to prove our hypothesis. As you can see below, the scatter plot looks much better than the correlation matrix and reveals a linear relationship that we will leverage in subsequent analysis.



Pic. 9. Each of economic indicators vs. amount of billionaires.

One more important observation resulting from visual analysis is that our dataset contains few outliers, where country of USA is the biggest one standing out – marked with red circle on the above picture.

4. Implementation and configuration of models

In this section, we implement regression models to extract statistical insights from our dataset, starting with Linear Regression. We also used:

- weighted linear regression,
- regression using ANN.

4.1. Linear regression

This method allows us to represent the strength and direction of the linear associations, providing a statistical foundation for understanding how changes in economic indicators correspond to variations in the number of billionaires. Linear Regression offers a powerful tool for predictive analysis and helps us identify which economic factors are most influential in shaping the prevalence of billionaires. Through this statistical approach, we seek to gain a deeper understanding of the quantitative impact of economic indicators on extreme wealth concentration within our dataset.




Pic. 10. Visualization of linear regression on scatterplots.

4.2. Weighted linear regression

To improve our results, we decided to implement weighted linear regression. The reason for doing that is simple – we wanted to find out if reducing or increasing influence of outliers (like mentioned previously USA). In conclusion, reducing influence of outliers gave better outcome. What's our reasoning – we wanted to assign higher weights (close to 1) to countries that had small number of billionaires, and are majority in our case. Analogically – lower weights for countries with higher amount of billionaires – to lower their impact on our regression line.

```
0.994258
1.000000
1.000000
0.998086
0.850718
0.998086
0.999043
0.996172
0.977033
0.999043
0.974163
0.500478
0.973206
0.967464
0.977033
1.000000
0.958852
0.922488
0.960766
1.000000
0.988517
0.976077
0.998086
1.000000
0.996172
0.997129
0.956938
0.926316
0.925359
1.000000
1.000000
1.000000
0.948325
0.999043
0.998086
0.996172
0.976077
0.995215
AmountOfBillionaires, dtype: float64
```



Pic. 11. Assigned weights (Red arrow – outlier).

4.3. Regression using MLP

As our last model we decided to implement MLP in hope of improving our results. To ensure the best possible parameters, we used grid search algorithm with cross validation. Also the reasons for using only 1- and 2-layers MLP's is that:

- our data is distributed in linear manner,
- dataset consists of low number of samples.

Having that in mind, we decided that such configuration is enough and justifiable. Below we present best parameters found by grid search algorithm:

```
[ ] param = {
    'hidden_layer_sizes':[(x) for x in range(1, 11)],
    'activation': ['identity','logistic','tanh','relu'],
    'random_state':[10],
    'max_iter':[4000],
    'solver': ['lbfgs','sgd','adam']
}
reg = MLPRegressor()
grid_search = GridSearchCV(reg, param, cv=5, scoring='neg_mean_absolute_error')
grid_search.fit(X_train2, y_train2)
print(f"BEST PARAMETERS: {grid_search.best_params_}") # to get the best parameters
print(f"BEST ESTIMATOR: {grid_search.best_estimator_}") # to get the best estimator
print(f"ALL RESULTS: {grid_search.cv_results_}") # to get all results

BEST PARAMETERS: {'activation': 'tanh', 'hidden_layer_sizes': 1, 'max_iter': 4000, 'random_state': 10, 'solver': 'adam'}
BEST ESTIMATOR: MLPRegressor(activation='tanh', hidden_layer_sizes=1, max_iter=4000,
random_state=10)
```

Pic. 12. 1-layer MLP with best parameters.

```
param = {
    'hidden_layer_sizes':[(x, y) for x in range(1, 11) for y in range(1, 11)],
    'activation': ['identity','logistic','tanh','relu'],
    'random_state':[10],
    'max_iter':[4000],
    'solver': ['lbfgs','sgd','adam']
}
reg = MLPRegressor()
grid_search = GridSearchCV(reg, param, cv=5, scoring='neg_mean_absolute_error')
grid_search.fit(X_train2, y_train2)
print(f"BEST PARAMETERS: {grid_search.best_params_}") # to get the best parameters
print(f"BEST ESTIMATOR: {grid_search.best_estimator_}") # to get the best estimator
print(f"ALL RESULTS: {grid_search.cv_results_}") # to get all results

BEST PARAMETERS: {'activation': 'relu', 'hidden_layer_sizes': (2, 10), 'max_iter': 4000, 'random_state': 10, 'solver': 'adam'}
BEST ESTIMATOR: MLPRegressor(hidden_layer_sizes=(2, 10), max_iter=4000, random_state=10)
```

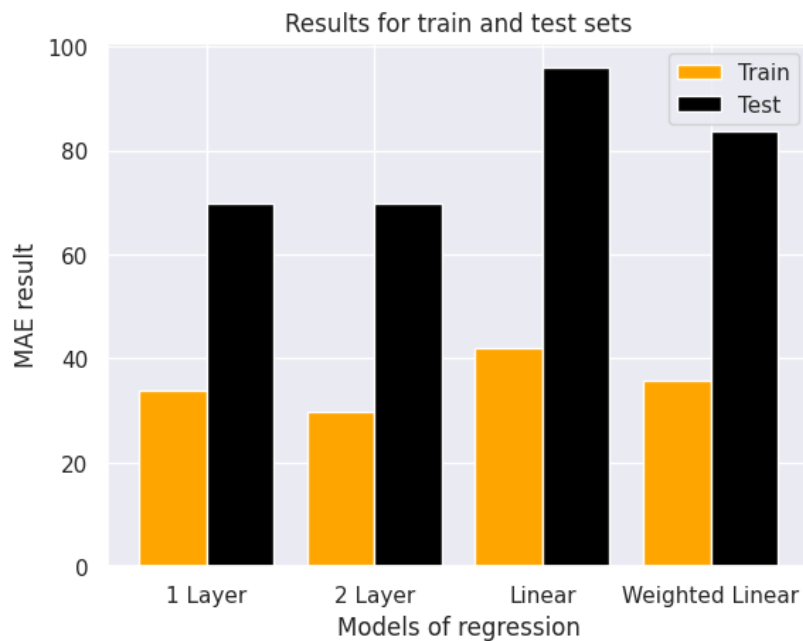
Pic. 13. 2-layer MLP with best parameters.

4.4. Final assessment of regression models

In this section we compare the results and effectiveness of our regression models. To measure our error we used R2 and MAE. We decided to use two measures because we wanted to make sure that we are getting valid results.



Pic. 14. R2 error for every model.



Pic. 15. MAE error for every model.

5. Summary and conclusions

Discovering why our hypothesis didn't hold reveals a few practical challenges. The dataset became smaller after cleaning, which might have impacted our ability to draw solid conclusions. Notably, outliers like the United States and China could be throwing off our results, overshadowing trends in other countries. Our initial choice of features might not be the best predictors of billionaire prevalence, urging us to reconsider. It's crucial to recognize these issues and adjust our approach. Exploring alternative models and additional features could give us a more straightforward and clearer picture of the connections between economic indicators and the number of billionaires. This practical and pragmatic approach ensures a more grounded and understandable analysis in the future work.

6. Acknowledgements

1). Chat GPT for writing report – checking typos, constructing more understandable phrases, help with translation.

2). Contribution of each author of the project:

Hypothesis formulation – E. Czop, J. Banaś, K. Fila

Relevant work research – K. Fila

Exploratory Data analysis – E. Czop, J. Banaś

Model implementation and configuration – E. Czop

Model testing – J. Banaś

Report writing – K. Fila

Project organization – J. Banaś

Project results presentation – E. Czop, J. Banaś, K. Fila