“I Wanna Be a Billionaire”

BIA-500 Group 6 Capstone Project

Team Members:

Shweta Sanjay Churi

Estella Clevenger

Jyoti Rajesh Khare

Daniel Salib

Pamela Yong

Table of Contents

I. Project Definition

a. Project Overview

The "I Wanna Be A Billionaire" project is driven by the desire to assist people and businesses in realizing their wealth-related objectives as well as the goal of financial empowerment. Many people dream of becoming billionaires, but few achieve this goal. This project aims to uncover insights and patterns among billionaires that could reveal how they attained their wealth. The Data set we utilized is from Kaggle and additional input data is from Google searches for missing information from data set obtained from Kaggle.

b. Problem statement

The richest 1% owns almost half of the world's wealth and accounts for almost two-thirds of the world's new wealth. As Percy Bysshe Shelley once said, "The rich get richer, and the poor get poorer." This research project tackles the perplexing task of becoming wealthy and successful financially in the current economic climate. In addition to analyzing the societal ramifications of wealth distribution, it seeks to empower people and businesses in their pursuit of financial goals by offering data-driven insights, solutions, and financial literacy. Ultimately, it hopes to contribute to a fairer and more knowledgeable economic environment. To test the null hypothesis of: "I will not become a billionaire in this lifetime" vs alternative hypothesis of, "I will become a billionaire in this lifetime."

c. Metrics

Some of the metrics that we used to measure results were demographic metrics such as the age distribution of billionaires and the gender distribution among billionaires. For wealth and GDP metrics we used a comparison of total wealth of billionaires compared to the GDP among countries. We also used a distribution of wealth by industry (tech, finance, retail, etc...),

and the most common and frequent source of wealth. For net worth metrics, we used the total net worth distribution and the top 10 countries with billionaires.

II. Analysis

a. Data Exploration

The data set lists the number of billionaires in the world. At the time of the data download, the number stands at 2640 but it changes daily. While the data set has 38 columns with different types of useful and meaningful information, it's not complete. For example, many organizations were missing, it's either unknown or there are too many companies to list. There are also certain abnormalities, an example, Taiwan and Hong Kong are listed as countries but both GDP and population were left blank; we manually added it to the raw data through searches via Google. Currently, we will be categorizing them as countries as it serves our visualizations better than leaving their data blank or grouping them with China. Certain people's age could not be found and hence we could only put N/A. For the age, while some of the names listed "& family" it seems to only refer to the oldest person alive holding part of the wealth, so while the family might be on the list, as individuals, a person might or might not be a billionaire.

b. Data visualization

i.

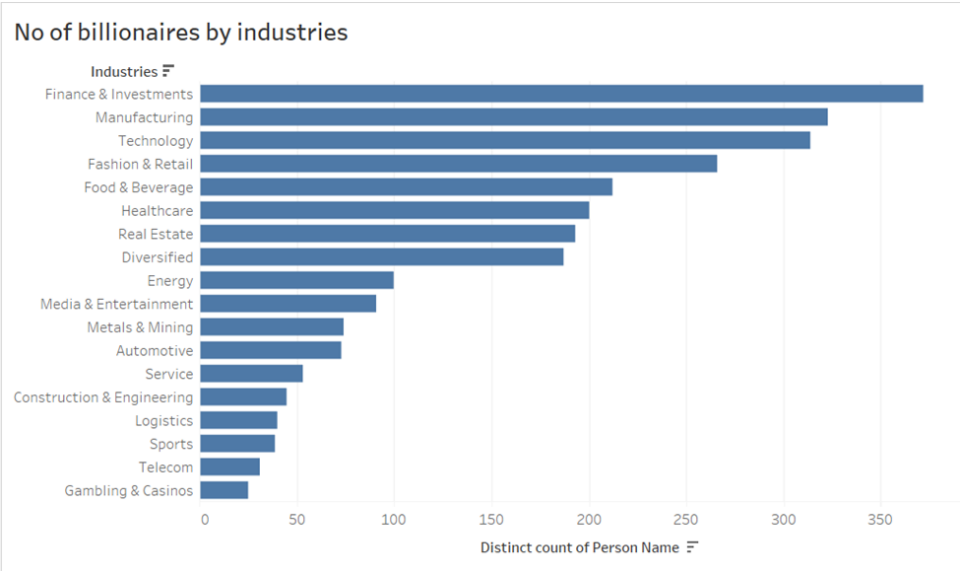
Excel Data Visualizations:

Excel Link: BIA 500 - CAPSTONE PROJECT DATA SET (WORKING COPY).xlsm
(Note: Please open in Desktop version of Excel, if not, the maps will not load. Also, this file is only accessible by @stevens.edu users.)

1. Country – Number of billionaires in a specific country. Broken down into ranges from 0-100, 101-200, and so on and so forth.
2. Country Top 10 – Only the top 10 listed countries with the most number of billionaires, all the rest are then marked as others.
3. US States – Map representation of total wealth in only the United States by states. Top five are California, New York, Texas, Florida and Washington in that order. There are many reasons to consider, such as Global HQ's location, Technology-Silicon Valley, Opportunities, Networking, State Incentives and Taxes.
4. Family Wealth vs Individual Wealth – Broken down by % of the 2640 who are wealthy families or individuals then comparing the monetary value and percentage.
5. Self-Made vs Inherited – Broken down by % of the 2640 who are self-made or inherited the money and to see the difference in monetary value.
6. Gender – Male vs Female Billionaires, which surprisingly, the % comparison between count and wealth are the same.

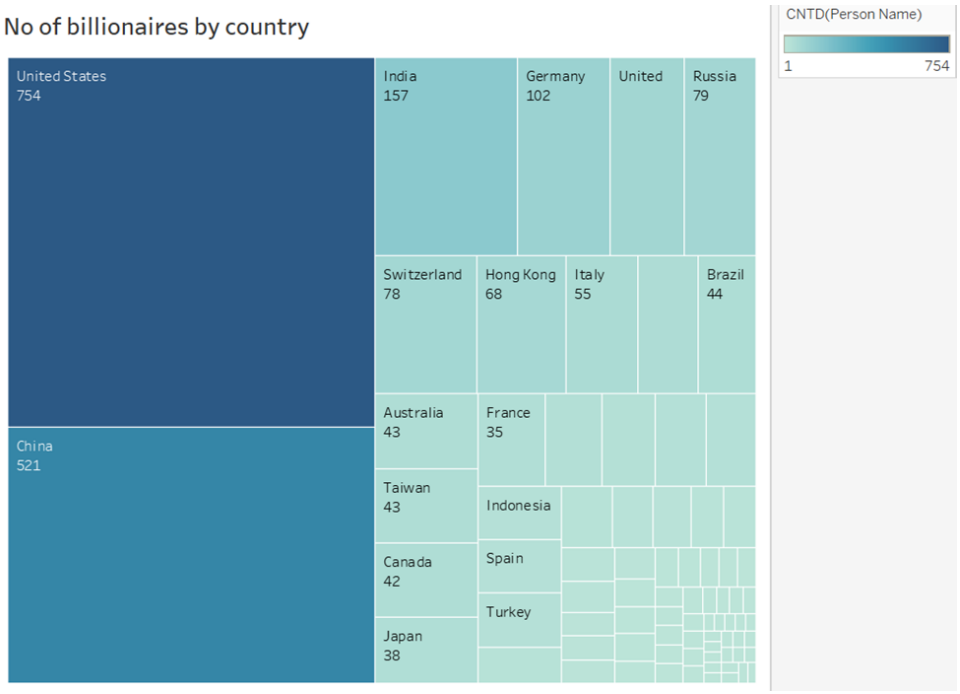
7. Age – What age range does the billionaire fall into. There is a bit of a margin of error as there were 59 entries with N/A as the age and data could not be found online. Overall, most range from ages 41-90 which is a large range, most prevalent is 51-70.
8. Country GDP – The thought was to see if there was a correlation between the GDP of a country compared to the wealth of the billionaires but there does not seem to be a pattern that could be seen.
9. Country Population vs Wealth – Due to Outliers, the data did not show a direct correlation.
10. Country Population with Top 3 Outliers Removed – The thought was to isolate the top 3 outliers and see if then there is a correlation between a country's population and wealth, but the scatter chart showed that the numbers were all over the place. A higher population doesn't mean a country has more wealth and a lower population doesn't mean a country is less wealthy, even taking out the top 3 outliers.
11. Industry – Which industry has the most total wealth. Which surprisingly showed Fashion & Retail coming in second and Construction & Engineering coming in last.
12.

Tableau Data Visualizations:

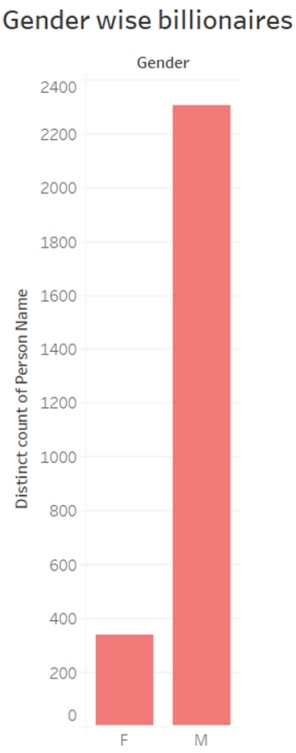


This graph shows the number of billionaires by industry. We can see that finance and investments has the highest number of billionaires followed by manufacturing and technology.

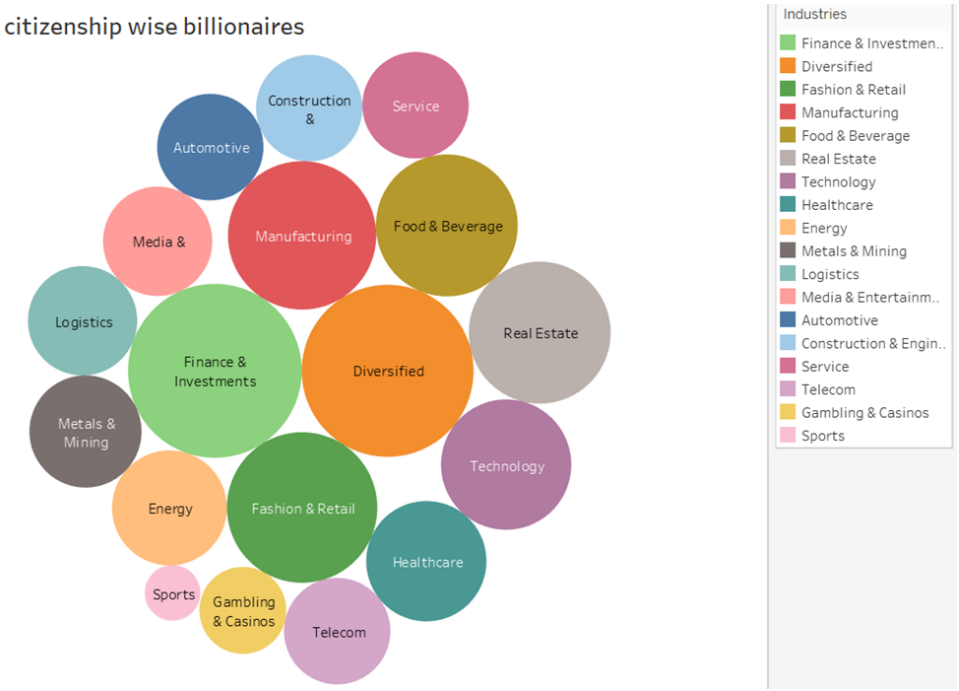
From the graph we can see that sports, telecom and casinos have the lowest number of billionaires.



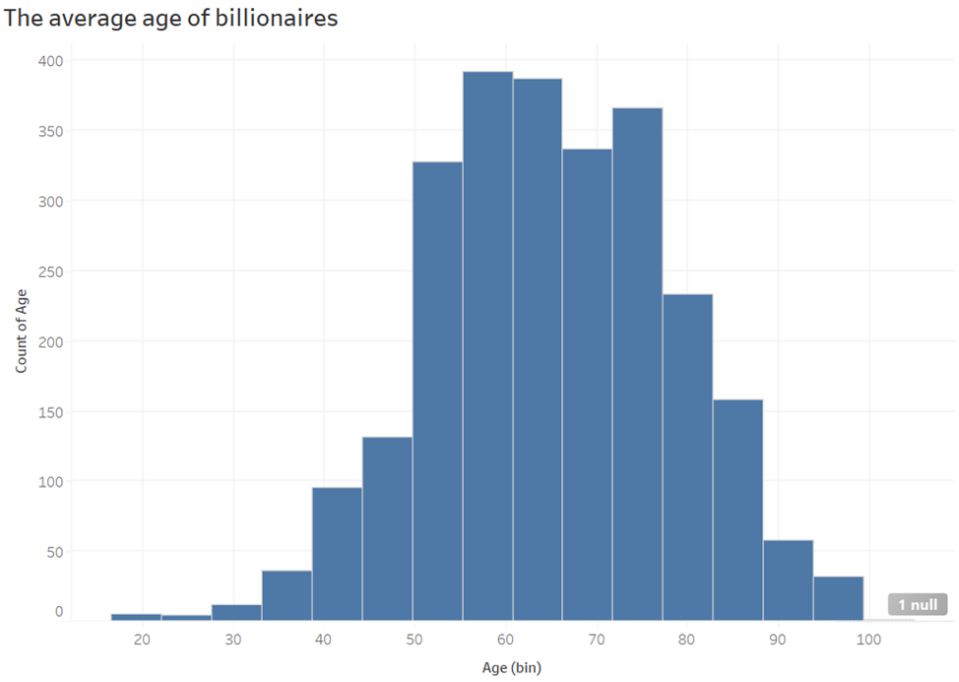
From the above graph we can state that the United States has the highest number of billionaires from the dataset. Then next we have China and India wherein there are many billionaires. When you hover over the data you can see that the Uzbekistan has the least and the only billionaire.



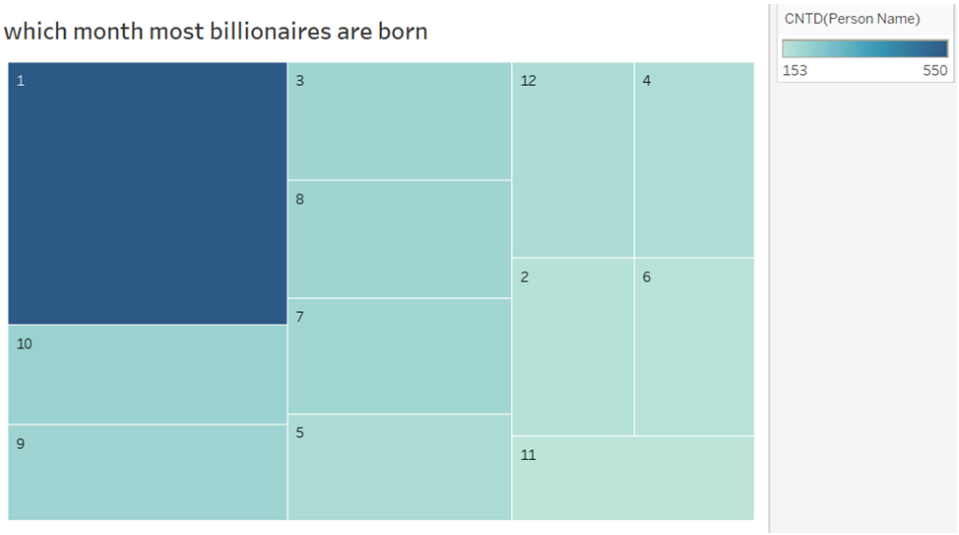
From this graph we can clearly see that there are more number of male billionaires than females.



Here we have classified the various industries based on citizenship. Here different color identifies different sector of the industry.

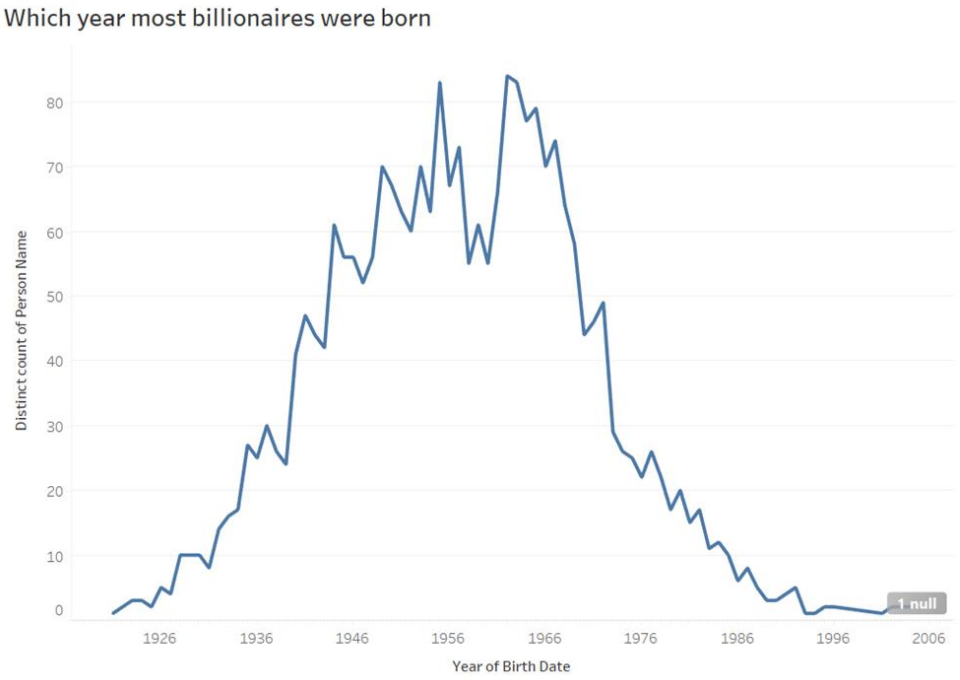


From this graph we can clearly see that the average age of billionaires falls between 50 to 60 years. And the age bracket of 17 to 22 have least billionaires compared to all.

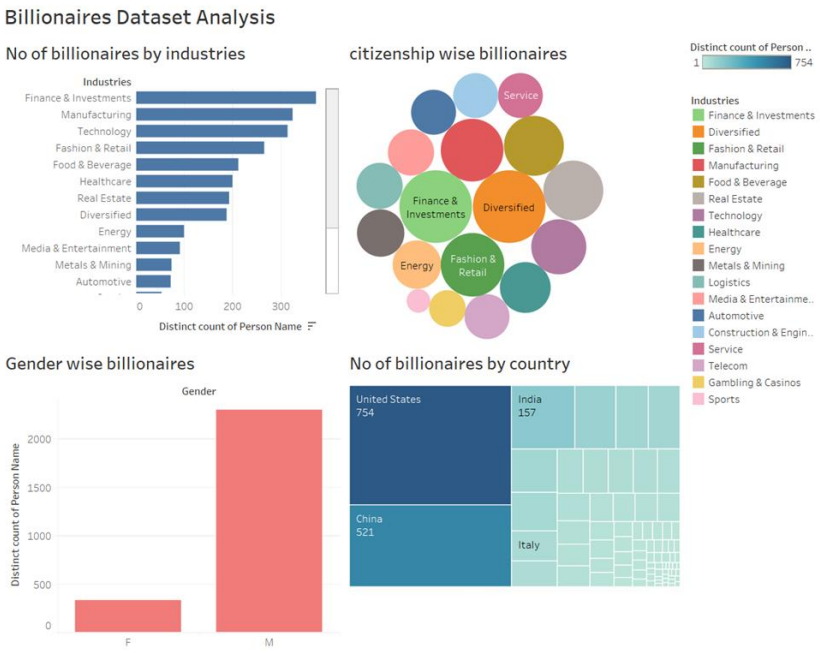


From this data we can say that January is the month most billionaires are born followed by October and September.

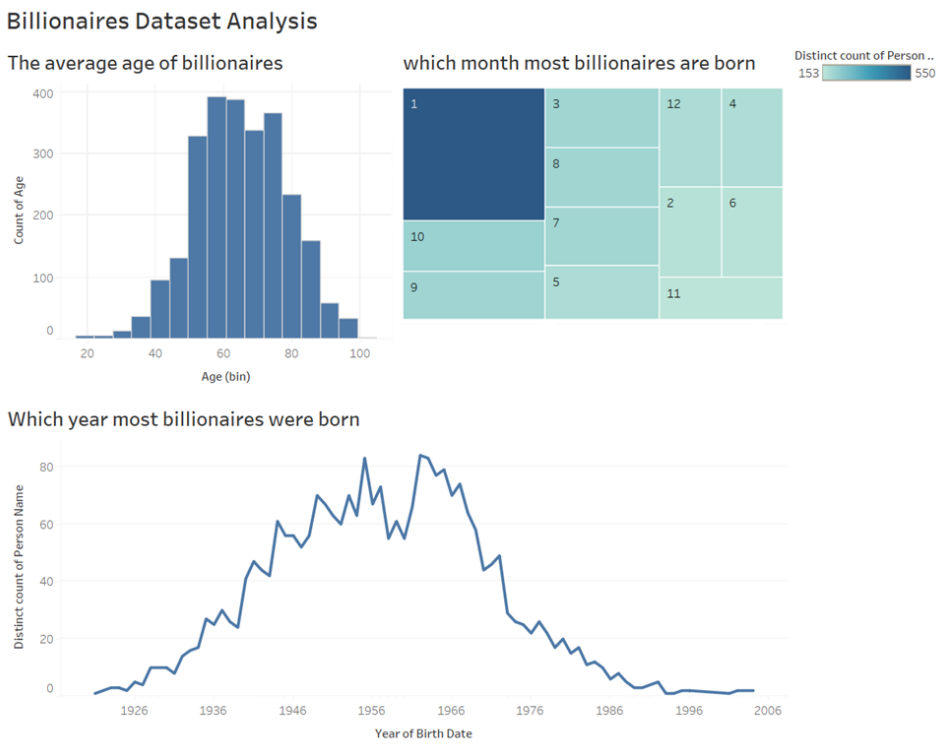
November, which is the 11th we can say that this is the month the least billionaires are born.



From the above graph we can say that the most billionaires were born during 1962 and 1955 as these years have the highest peak value. We can see a sudden increase after 1936 and sudden drop after 1966.



This is the first dashboard which combines all the 4 worksheets. It gives a detailed overview of the dataset. These dashboards help to give Consolidate key information in one place.



This is the second dashboard which also give a quick view which helps to understand the dataset better and gives us information about the trends in billionaires and gives insights about their average age, month and year they are born.

III. Methodology

a. Data Preprocessing:

Missing Values: Missing values were identified and addressed for columns like GDP, age, and industry. Different techniques like imputation or deletion were employed depending on the nature of the missing values and their impact on the analysis.

Outlier Detection: Outlier detection was performed on all numeric data using techniques like Z-score analysis or Interquartile Range (IQR). Identified outliers were carefully analyzed and either removed from the dataset or treated based on their context.

b. Data Analysis:

Exploratory Data Analysis (EDA): EDA was conducted to understand the distribution of data across various attributes like country, industry, age, etc. This involved visualizations like histograms, pie charts, and box plots to gain insights into the data and identify potential patterns and relationships.

Random Forest Classifier: A Random Forest classifier was implemented to predict the target variable of "self-made" for everyone in the dataset. The model was chosen due to its robustness to outliers and its ability to handle a large number of features.

Model Training and Evaluation: The data was split into training (70%) and testing (30%) sets. The Random Forest model was trained on the training set, and its performance was evaluated on the testing set using metrics like accuracy, precision, recall, and F1-score.

Feature Importance: Feature importance was extracted from the Random Forest model to identify the most significant predictors of self-made wealth. This information provided valuable insights into the factors contributing to becoming a billionaire.

c. Refinement:

Error Analysis: The model's performance was carefully analyzed, and potential sources of error were investigated. This involved examining misclassified data points and identifying areas where the model could be improved.

Parameter Tuning: Based on the error analysis, model parameters were further tuned to enhance its accuracy. In this case, increasing the number of trees to 150 resulted in a performance improvement

IV. Results

a. Model evaluation and validation

To determine the influence of certain characteristics on becoming a billionaire, it would have been ideal to have a dataset that included both billionaires and non-billionaires. Due to the unavailability of such comprehensive data, we hypothesized that constructing a model to predict whether a billionaire is self-made might offer similar insights. By identifying the most influential features in this prediction, we aimed to correlate these findings with our initial visualizations.

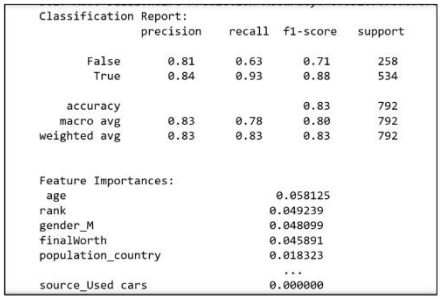
Upon importing the dataset, we managed missing values in numerical columns through median imputation, a method effective at minimizing the influence of outliers, hence maintaining data integrity. The dataset was then refined by removing columns that could lead to overfitting due to non-predictive or overly specific information. Specifically, columns such as 'personName', 'lastName', 'firstName', 'date', 'birthDate', 'birthDay', 'birthMonth', 'birthYear', 'latitude_country', and 'longitude_country' were dropped to bolster the model's generalization capability.

Categorical variables underwent one-hot encoding to eliminate model bias resulting from different variable types or scales. Following this transformation, we conducted a thorough check to replace any remaining NaN values with zero, ensuring our dataset was not missing values after the one-hot encoding.

We then chose the RandomForestClassifier due to its enhanced accuracy in performing predictions across multiple decision trees. The hyperparameter setting of 150 trees was selected through trial and error to balance computational efficiency against model performance and to avoid underfitting and or overfitting. We also permitted trees to grow without depth constraints, allowing the model to capture complex patterns, though we closely monitored for overfitting via cross-validation.

The dataset was then split into a 70/30 ratio for training and testing, to provide a large enough volume of data for training while still retaining a ample amount for model validation. We assessed model performance through accuracy and a detailed classification report, evaluating precision, recall, and F1-score to gain an understanding of the model's classification abilities.

b. Justification

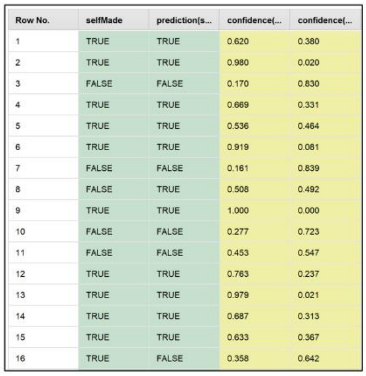


Classification Report:

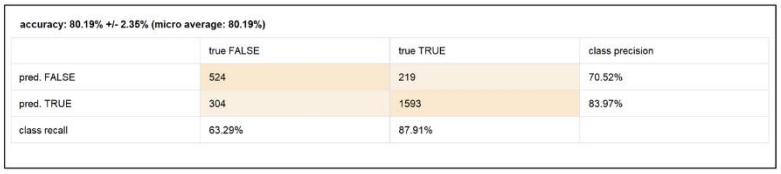
	precision	recall	f1-score	support
False	0.81	0.63	0.71	258
True	0.84	0.93	0.88	534
accuracy			0.83	792
macro avg	0.83	0.78	0.80	792
weighted avg	0.83	0.83	0.83	792

Feature Importances:

age	0.058125
rank	0.049239
gender_M	0.048099
finalWorth	0.045891
population_country	0.018323
...	...
source_Used cars	0.000000



Row No.	selfMade	prediction(s...	confidence(...	confidence(...
1	TRUE	TRUE	0.620	0.380
2	TRUE	TRUE	0.980	0.020
3	FALSE	FALSE	0.170	0.830
4	TRUE	TRUE	0.669	0.331
5	TRUE	TRUE	0.536	0.464
6	TRUE	TRUE	0.919	0.081
7	FALSE	FALSE	0.161	0.839
8	FALSE	TRUE	0.508	0.492
9	TRUE	TRUE	1.000	0.000
10	FALSE	FALSE	0.277	0.723
11	FALSE	FALSE	0.453	0.547
12	TRUE	TRUE	0.763	0.237
13	TRUE	TRUE	0.979	0.021
14	TRUE	TRUE	0.687	0.313
15	TRUE	TRUE	0.633	0.367
16	TRUE	FALSE	0.358	0.642



accuracy: 80.19% +/- 2.35% (micro average: 80.19%)

	true FALSE	true TRUE	class precision
pred. FALSE	524	219	70.52%
pred. TRUE	304	1593	83.97%
class recall	63.29%	87.91%	

The above images display the outputs of two implementations of a RandomForest model designed to predict the 'selfMade' status of billionaires: one executed through Python code and the other via the RapidMiner platform. While the core model remains consistent across both platforms, subtle variations in their results prompt a closer examination of each tool's methodologies and capabilities.

The Python model, with its 81.7% accuracy, illustrates strong predictive performance. This approach benefits from precise control over data preprocessing, including the capability to one-

hot encode categorical variables—a technique not readily available in the unlicensed version of RapidMiner. As a result, the Python model may have leveraged a more nuanced understanding of categorical distinctions, potentially contributing to its slightly higher accuracy.

In contrast, the RapidMiner model, displaying a marginally lower accuracy of 80.19%, employs cross-validation within its workflow—a technique that inherently provides a more robust assessment of model performance by validating the model against multiple subsets of the data. However, without one-hot encoding, RapidMiner may treat categorical variables differently, possibly affecting the model's ability to distinguish between nuanced categorical states.

These slight differences in model accuracy can, therefore, be attributed to the divergent preprocessing capabilities and validation techniques of the two platforms. RapidMiner's use of cross-validation ensures a rigorous evaluation of the model's generalizability, but its limitation in encoding categorical variables might constrain the model's predictive nuances. Meanwhile, the Python implementation, with its complete preprocessing steps, including one-hot encoding, may capture a more detailed representation of the data, despite the potential lack of cross-validation in the presented code.

Ultimately, the differences underscore the importance of consistent data treatment across modeling tools and the impact of software capabilities on model performance. Both models present a reliable analysis of predictive factors for self-made billionaires, with their slight differences highlighting the balance between model complexity and evaluation rigor.

In examining the feature importances from the Python output, 'age', 'rank', 'gender', and 'finalWorth' are highlighted as significant predictors. However, the inclusion of 'rank' and

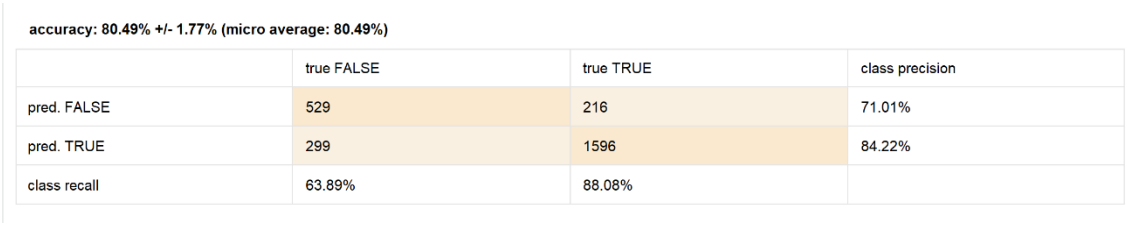
'finalWorth' may lead to potential data leakage since these features could be directly associated with the target variable, potentially leading to overly optimistic performance estimates.

As can be viewed below, removing the parameters in question ended up benefiting both models.

```
Self-Made Billionaire Prediction Accuracy: 0.8358585858585859
Classification Report:
              precision    recall  f1-score   support

       False       0.80      0.66      0.72       258
        True       0.85      0.92      0.88       534

    accuracy                           0.84       792
   macro avg       0.83      0.79      0.80       792
weighted avg       0.83      0.84      0.83       792
```



accuracy: 80.49% +/- 1.77% (micro average: 80.49%)

	true FALSE	true TRUE	class precision
pred. FALSE	529	216	71.01%
pred. TRUE	299	1596	84.22%
class recall	63.89%	88.08%	

The RapidMiner output offers a confusion matrix that affirms a high-class precision and recall, especially for the 'True' class. The similar performance metrics across both platforms reinforce the reliability of the model's predictions.

The use of two different modeling approaches serves as an additional form of cross-validation, strengthening our confidence in the model's predictive power when we observe consistent results.

V. Conclusion

a. Reflection

Analyzing big data with limited resources revealed the many-sided nature of wealth accumulation and the similar qualities that define self-made billionaires. Our research delved into various measurable aspects such as demographic distributions, industry involvement, and geographical influences. However, the underlying factors driving an individual's journey to immense wealth often went beyond measurable metrics. Elements such as personal innovation, a knack for seizing economic opportunities at opportune moments, and sometimes a stroke of unforeseen luck, play pivotal roles. These absract elements, although crucial, elude concrete measurement and remain as mysterious as the concept of fortune itself. Thus, while our data-driven approach provided insightful correlations and trends, it only scratches the surface of the complexity that is billionaire wealth creation.

b. Improvement

To enhance the quality of our data analysis in the future, we would suggest purchasing complete data sets from reliable sources rather than relying on Kaggle, which often contains incomplete information. This is because redundant data consumes memory that could be better utilized elsewhere. It would also be advisable to set a defined end date for data collection, considering the fluctuating nature of billionaires' net worth due to variable factors like stock prices. For instance, while Bernard Arnault & Family ranked first in our data set, Elon Musk leads as of 11/30/2023.

Moreover, working with smaller data sets could be more beneficial, as managing excessive data can be challenging, especially with the hardware and software limitations we currently face. Even simple tasks in Excel can cause lagging issues.

We also observed that compatibility issues arise across different platforms and versions of the same software. For example, discrepancies in charts, graphs, and maps are evident when transitioning from Excel desktop to the 365 Online version. Issues with theme consistency, color schemes, sizing, and column layouts are common when switching between program versions and licenses. To mitigate this, in the future we would ensure that all participants are utilizing the same programs.

Furthermore, compatibility issues between Mac IOS and Windows created difficulties in sharing work across different user environments. in the future, we could either cross train across platforms to mitigate compatibility issues or prioritize work on a single operating system. Lastly, the diversity in the programs used for script execution underscores the need for a more universally accessible and free application, as not all users have access to paid subscriptions.

VI. References

1. Nidula Elgiriyewithana. (2023, October) "Billionaires Statistics Dataset (2023)." Kaggle. https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset
2. Forbes. (2023, November) "The World's Real Time Billionaires." Forbes. https://www.forbes.com/real-time-billionaires/#52a073483d78
3. Dbpedia. (2023) "About DBPedia." DBPedia. https://www.dbpedia.org/about/
4. Ali Hassan. (2023, October 31) "16 Billionaires Who Live Like Regular People." Yahoo! Finance. https://finance.yahoo.com/news/16-billionaires-live-regular-people-203226243.html
5. Soumya Karlamangla. (2022, March 8) "Why So Many Billionaires Live in California." NYTimes. https://www.nytimes.com/2022/03/08/us/billionaires-california.html
6. OpenAI. (2022, November 30). Assistance with citation and report formatting was provided by ChatGPT
7. Khanyi Mlaba. (2023, January 19). “The Richest 1% Own Almost Half the World’s Wealth & 9 Other Mind-Blowing Facts on Wealth Inequality. Global Citizen.” https://www.globalcitizen.org/en/content/wealth-inequality-oxfam-billionaires-elon-musk/#:~:text=The%20richest%201%25%20own%20almost%20half%20of%20the%20world's%20wealth,99%25%20of%20the%20world's%20population.