Aly 6015 Project Analyzing Forbes World's Billionaires List 2022

Final Project: Final Project: Initial Analysis Paper

08.05.2022

Group Members:

1- Negin Dehghanian

2- Mohammad Hossein Movahedi

1 Introduction

In the present study, we will investigate the Forbes World's Billionaires Lists of 2022 (Prasert Kanawattanachai, 2022) to see any potential data patterns. The dataset contains 2,658 records relating to billionaires around the world. Table 1 presents a summary of the dataset's features. There are different aspects of a wealthy person in each dataset row. So, in each record, we can find out the billionaire's name, age, collective worth in a million dollars, year of ranking, months of ranking, the industry in which the billionaires made their money, the primary source of money, for example, any specific product and brand, associated country and state and city where the worth was collected, country of citizenship, company name, are they self-made billionaires or not, gender, birth date, their titles, their philanthropy scores (where 5 represents the most generous givers), their residential area, the number of siblings they have, and a brief biography of themselves.

1.2 Research Questions

In this study, we want to answer questions such as:

- Are there any specific patterns in data to distinguish between self-made and non-self-made billionaires?
- Is there a relationship between age, the field of work, and total net worth?
- Is there like and surprising factor and combination of factors (variables) that can help us determine who gets to the list.

To answer the above research questions, we need to explore different dataset columns and extract valuable insights. Since the dataset shows multiple categories for a single person, we could group billionaires according to these categories and analyze the classes accordingly.

2 Methodology: Data preparation

2.1 Data cleaning/ Deleting the about, bio, year, and month columns

Because bio and about are text columns, they are out of the scope of this course. Thus, we wish to remove them from the dataset. But first, we do some text mining, hoping to find some valuable results. Our first step was to combine the two columns into one, then lowercase the words, remove white spaces, quotes, punctuation, and stop words. A Word cloud was then created with colors. According to Figures 1 and 2, company, founded, and chairman are the most frequent words in these two columns.

In addition, we deleted the year and month columns since they did not add any value to the analysis.

2.2 Data cleaning/changing the types of variables

According to table 1, the Birthdate column is categorical data, but it must be converted to date for use with date data. Additionally, the majority of the variables are categorical; however, to use some of them in developing models, we need to convert them to factors. Therefore, we converted category types, self-made, gender, philanthropy Score, and siblings into factors.

2.3 Data cleaning/ creating new features

To make the country and citizenship columns more useful, we grouped them into new columns called country-region and citizenship-region.

2.5 Data cleaning/ missing values

We replaced many of the cells in the dataset with null since they were blank. According to Table 2, several values in the columns are missing: age, birthdate, philanthropy Score, state, city, organization, gender, title, residenceMsa, and numberOfSiblings. We can work on age's missing value because the birthday and age have the same meaning. Because the number of missing Values in the philanthropy Score and numberOfSiblings variables is considerably more significant than the amount of existing data, dealing with null data is challenging. In this scenario, we chose to leave these columns with null values. Data distribution was developed to

replace null values with relevant values (Figure 3) to deal with missing values in the age column. Because the age distribution by gender in Figure 3 is symmetric, we substituted the missing age data with 64.36992, the mean (5 Most Important Data Pre-Processing Techniques - Impute Missing Data - Part II » DevSkrol, 2022). We replaced missing values in categorical variables with "None" or the data mode based on feature nature.

2.6 Data cleaning/ outliers

Dealing with outliers is a very delicate task, and the first step in dealing with them is recognizing them. Our entire dataset is an outlier as only a tiny fraction of people have the chance to become billionaires .and also our dataset contains financial data emphasizing outliers as they are significant parts of our data. This dilemma that every individual data near the edges of our dataset are significantly important put us in the position that we decided to acknowledge the existence of outliers in our dataset but do nothing about them as they are of utmost importance.

2.7 Data cleaning/transforming the final worth to a normal distribution

A glance at our leading target column Figure 4 shows this data doesn't have a normal distribution. Therefore, to deal with this data, we need first to find its distribution and use transforming technics to change it to normal distribution.

First of all, before anything we based on our initial guest, we used the log transform to normalize the data a little bit. Then, we used the Cullen and Frey graph (Delignette-Muller and Dutang, 2015) to decide which distribution fit best with our data. According to Figure 5, our data distribution is either beta, gamma, or Weibull. Therefore, we investigate these distributions according to Figure 6 based on the Q-Q plot and P-P plot (and AIC and BIC of fit tests confirm that) the gamma distribution best fits our data. Therefore, we calculated the parameters of gamma distribution the shape = 98.61073 and rate = 12.40862. Then after this, we used a simple gamma to normal transformation to finally transform our data to normally distributed data in Figure 7.

3 Methodology: Data analysis

3-1 Descriptive statistical tables

Once the data have been prepared and quality ensured, the next step is to analyze the data. Firstly, some descriptive statistical tables have been provided in this step to illustrate the overall and detailed state of the data. In table 3, you can view a statistical summary of the data, while in table 4, you can compare self-made and country-region. Table 5 indicates the average age and final worth depending on the self-made.

3-2 Descriptive statistical plots and charts

We produced several plots and charts to understand our data further. Some of the graphs we generated during the data cleansing process (Figures 1- 4). Figures 8 and 9 exhibit box plots of billionaires' ages based on their self-made status and gender type. Figure 10 also depicts a boxplot displaying the distribution of billionaires' final wealth based on their business category. Figure 11 represents the relationship between the final worth and age of billionaires by taking self-made status into account. Figure 12 shows the frequency of data depending on the category of billionaires' businesses.

4 Results and findings: Data interpretation

4-1 Interpreting tables

We first looked at the summary statistics table for all data, as shown in Table 3.

According to this data, the average final worth of billionaires is approximately 4,803 million dollars, while the maximum value is 219000, significantly higher than the average. This is yet another demonstration of the skewness of the final worth distribution.

Furthermore, we can see that the majority of billionaires are men, self-made, and Asian.

According to Table 4, most billionaires from Asia, Africa, Europe, the United States/Canada, and Australia/New Zealand/Oceania are self-made, but billionaires from Latin America/the Caribbean are mostly not. According to Table 5, the average age of female billionaires is around 62, and their ultimate wealth is approximately 4718 million dollars. Compared to women, billionaires' average age is 64, and their average net worth is around 4796 million dollars. Because we did not remove records, the number of observations after cleaning is the same as before.

4-2 Interpreting charts and plots

Some plots have been created to understand the impact of attributes on recognizing self-made billionaires and discovering any special pattern in data. According to Figure 8, both distributions are essentially symmetric, but the average age of self-made billionaires is lower than that of non-self-made billionaires. Perhaps because, as a result of modern technologies, the founders and owners of certain technology-based enterprises are young individuals. We can see that the average age of un-self-made billionaires is lower than the median.

Figure 9 shows the age distribution by gender. This graph indicates that the data distributions for billionaires' ages based on gender are almost symmetric.

Figure 10 presents the distribution of billionaires' net worth based on their business type. We can see that most of the distribution is skewed to the right, implying that a few people have a lot

of wealth in each category. Also, we can notice many outliers. However, there is a substantial variation between the min and max of data based on the nature of the data.

Figure 11 shows a scatter plot demonstrating the impact of self-made on the age and normalized format of billionaires' final worth. We have discovered no linear link between age and wealth.

Furthermore, there is no clear pattern for classifying data based on whether it is self-made. Figure 12 depicts the distribution of several categories of billionaires' enterprises.

The top three industries producing billionaires are finance and investment, manufacturing, and technology.

5 Discussion and conclusion

The questions we defined at the beginning of this study could not be answered clearly.

- In answer to the question: Does data show a distinct pattern between self-made and non-self-made billionaires? Most billionaires are self-made, including those from Asia, Africa, Europe, the United States/Canada, and Australia/New Zealand/Oceania. Self-made billionaires typically have a younger average age than non-self-made billionaires. Due to modern technologies, many technology-based enterprises are founded by and owned by young individuals.
- Research has shown that wealth is not linearly related to age. In addition, there is no clear pattern for classifying data based on whether it was created by the individual or not. After reviewing the above, it is apparent that being a self-made billionaire does not follow a specific pattern, at least based on the data.
- To answer the question: Is there a correlation between age, the field of work, and total net worth? According to statistics, most of the billionaires are men and Asian. We know that women's average age and wealth are lower than their male counterparts. Financial

and investment, manufacturing, and technology are the top three industries introducing billionaires.

- To answer the question: Are there a combination of factors (variables) that will help us determine who gets on the list? So far, we did not find any surprising factors.

The specific characteristics of this dataset likely made it harder to discover patterns in this data. The billionaire's final worth distribution is significantly right-skewed and Gamma. We transformed this distribution to normal, but our knowledge of handling this data is insufficient. Another reason is that most variables in this dataset are categorical. Since we are used to working with numerical data, identifying patterns in this variable is difficult. In addition, the dataset contains only billionaires' information, so we lack access to other people's data to create different categories. Using historical data from prior billionaires' lists can also be a great way to bring diversity to data. Perhaps adding more features, such as a billionaire's education or IQ score, can be regarded as an underlying factor in classifying them as self-made or not.

our next step is to use regression modeling to create a prediction model for the normalized log of final worth and try to see which factors and variables can help us predict the final worth. we will also use GLM to try to determine the gender and self-made factor based on other factors and final worth if we succeed the insight we will get can give us a better understanding of the billionaire's world also we have the birthday of billionaires which might be helpful in creating a time series with independent data and do some analysis on it.

6 Tables and figures

Table 1. Data description summary.

	Feature Title	Feature Description	Feature Type	Number of Missing Values
1	Rank	The position in the Forbes list	Integer	0
2	PersonName	The name and surname of the Billionaires	Character	0
3	Age	The current age of the billionaire	Numerical	86
4	FinalWorth	The final worth in millions $	Numerical	0
5	Year	Year of ranking	Integer	0
6	Month	Months of ranking	Integer	0

	Feature Title	Feature Description	Feature Type	Number of Missing Values
7	Category	The sector of the business, or what segment of the economy they fit into	Character	0
8	Source	The source of wealth	Character	0
9	Country	Associated country which worth has been collected	Character	0
10	State	Associated state which worth has been collected	Character	0
11	City	Associated city which worth has been collected	Character	0
12	CountryOfCitizenship	The name of the country that this billionaire has citizenship with.	Character	0
13	Organization	The name of the company	Character	0

	Feature Title	Feature Description	Feature Type	Number of Missing Values
141	SelfMade	If they are self-made man or not	Character	0
15	Gender	Representing their gender	Character	0
16	BirthDate	Their date of birth	Character	0
171	Title	The billionaire's relationship to the company	Character	0
18	PhilanthropyScore	Their philanthropy scores	Numerical	2272
19	ResidenceMsa	The city in which they reside	Character	0
20	NumberOfSiblings	Number of their siblings	Numerical	2541
21	Bio	Their short biography	Character	0
22	About	Their short biography	Character	0

Table 2. Missing value.

countryOfCitizenship	country	rank	personName	age
0	0	0	0	81
finalWorth	category	source	state	city
0	0	0	1822	24
organization	selfMade	gender	birthDate	title
2243	0	16	93	2176
philanthropyScore	residenceMsa	numberOfSiblings	country_region	Citizenship_region
2174	1930	2445	0	0

Table 3. Descriptive statistic summary.

```
countryOfCitizenship   country             rank       personName            age           finalWorth
Length:2568          Length:2568        Min.   :   1   Length:2568        Min.   : 25.00   Min.   :  1000
Class :character     Class :character   1st Qu.: 665   Class :character   1st Qu.: 55.00   1st Qu.:  1500
Mode  :character     Mode  :character   Median :1292   Mode  :character   Median : 64.00   Median :  2400
                                        Mean   :1301                      Mean   : 64.37   Mean   :  4803
                                        3rd Qu.:1929                      3rd Qu.: 74.00   3rd Qu.:  4300
                                        Max.   :2578                      Max.   :100.00   Max.   :219000
                                                                          NA's   :81
                category      source             state              city           organization
Finance & Investments:374   Length:2568        Length:2568        Length:2568        Length:2568
Manufacturing        :329   Class :character   Class :character   Class :character   Class :character
Technology           :326   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Fashion & Retail     :236
Healthcare           :213
Food & Beverage      :197
(Other)              :893
 selfMade     gender       birthDate            title           philanthropyScore residenceMsa
False: 736       :   0   Min.   :1921-09-11   Length:2568        1   : 174        Length:2568
True :1832   F   : 295   1st Qu.:1947-10-27   Class :character   2   : 135        Class :character
             M   :2257   Median :1957-05-10   Mode  :character   3   :  60        Mode  :character
             NA's:  16   Mean   :1957-05-16                      4   :  16
                         3rd Qu.:1966-07-03                      5   :   9
                         Max.   :1996-07-23                      NA's:2174
                         NA's   :93
numberOfSiblings                      country_region                      Citizenship_region  finalWorth1
2      :  42     Africa                      :  17   Africa                      :  19     Min.   : 6.908
3      :  32     Asia                        :1080   Asia                        :1063     1st Qu.: 7.313
1      :  26     Australia/New Zealand/Oceania:  44   Australia/New Zealand/Oceania:  48     Median : 7.783
4      :  14     Europe                      : 549   Europe                      : 547     Mean   : 7.947
5      :   2     Latin America/Caribbean     :  87   Latin America/Caribbean     :  97     3rd Qu.: 8.366
(Other):   7     USA/Canada                  : 791   USA/Canada                  : 794     Max.   :12.297
NA's   :2445
 NorLogWorth
Min.   :-0.2661
1st Qu.:-0.1117
Median : 0.0603
Mean   : 0.1090
3rd Qu.: 0.2643
Max.   : 1.4411
```

Table 4. Self-made vs country-region table

	Africa	Asia	Australia/New Zealand/Oceania	Europe	Latin America/Caribbean	USA/Canada
False	8	223	13	231	50	211
True	9	857	31	318	37	580

Table 5. Age vs final worth based on gender type table

Group.1	age	finalWorth
F	62.85766	4718.644
M	64.56185	4796.167

Figure 1. Word cloud based on bio and about features

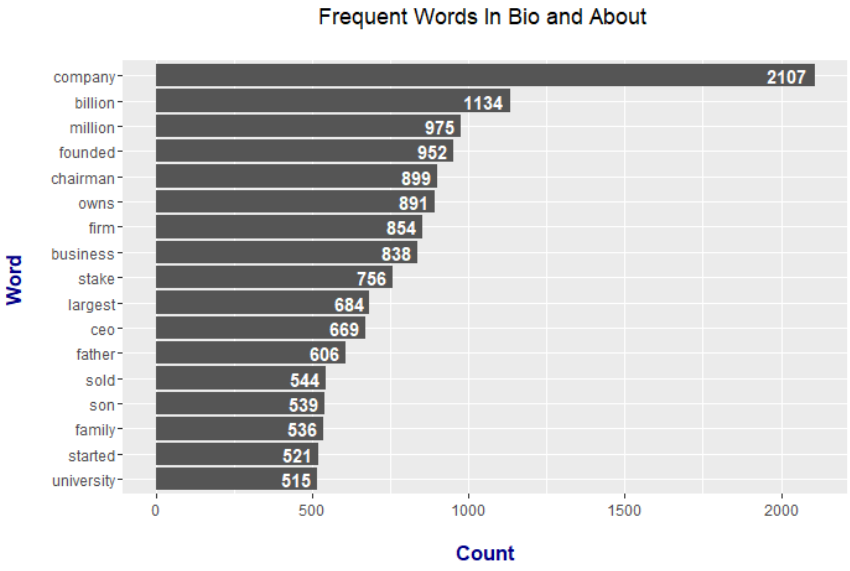

Figure 2. Most frequent words in Bio and About columns

Figure 3. Histogram for age based on gender

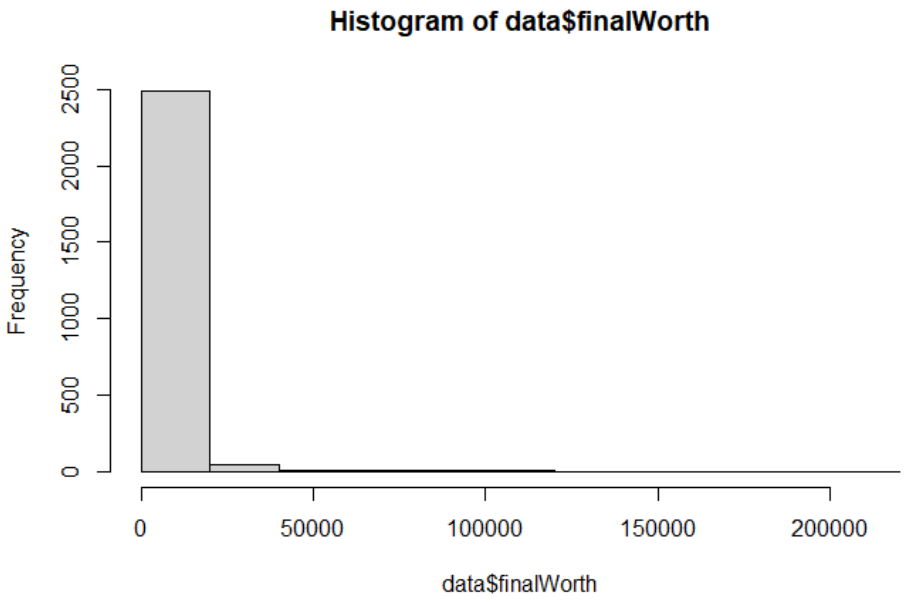

Figure 4. Histogram for final worth

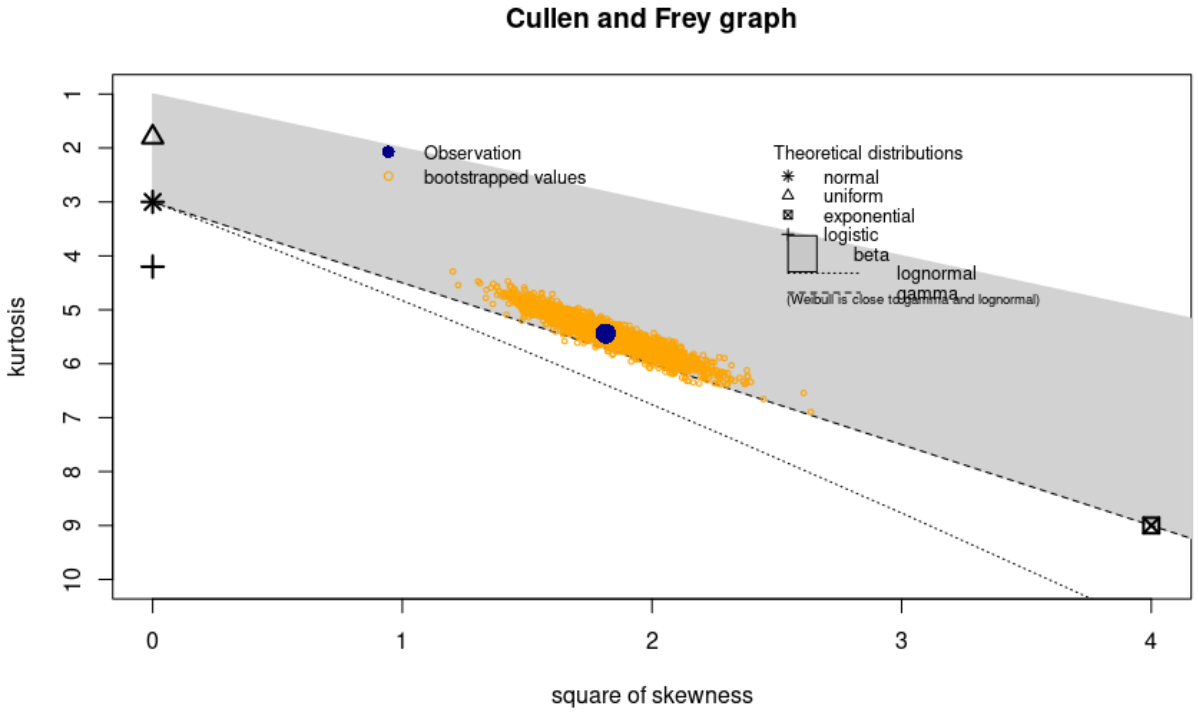

Figure 5 Cullen and Frey graph

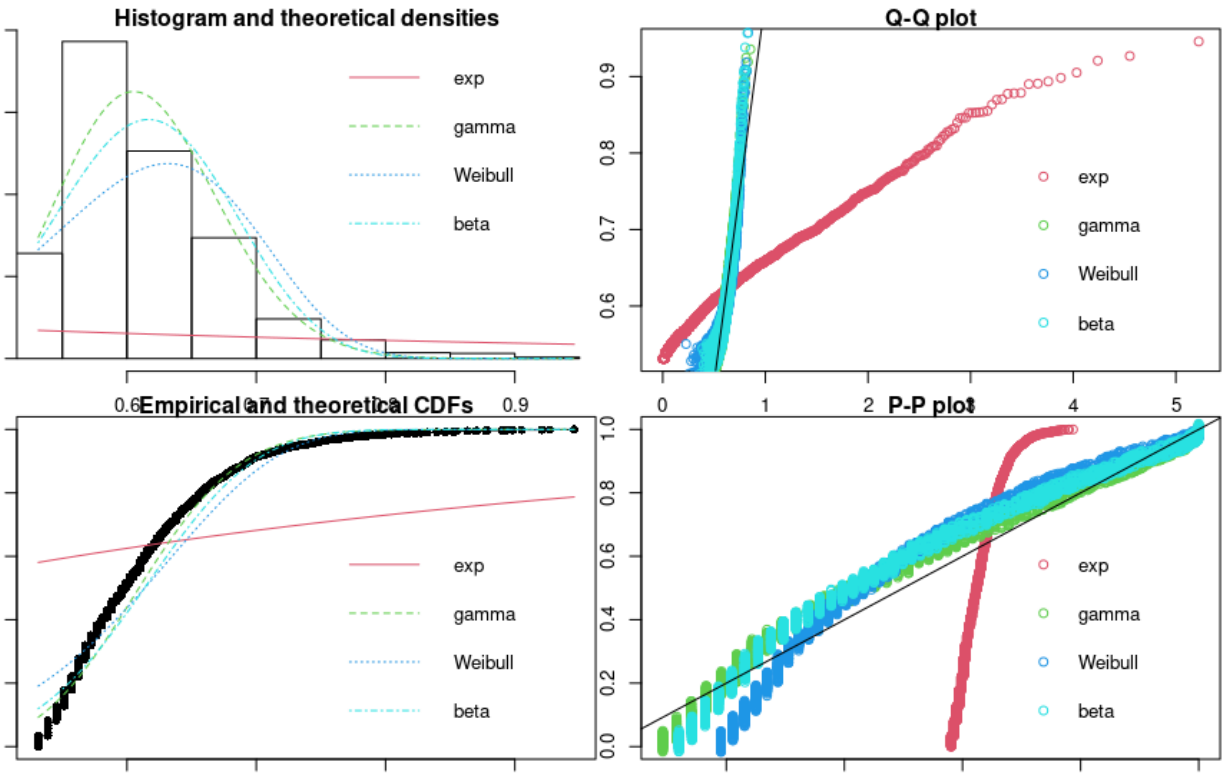

figure 6 descriptive plots for fitting

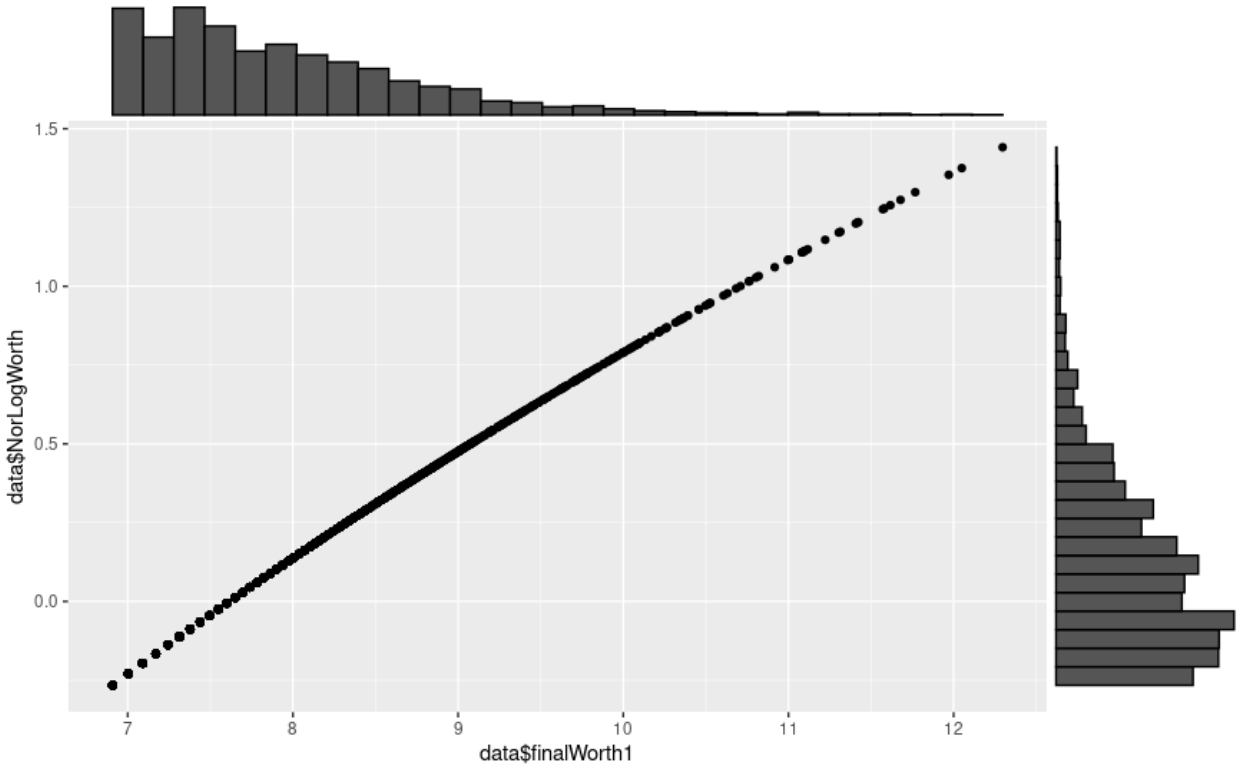

Figure 7 transformation to normal

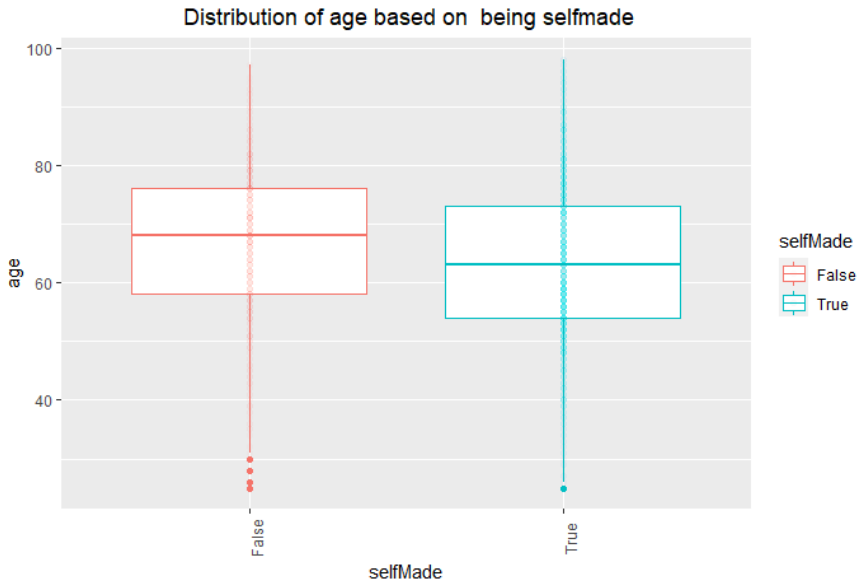

Figure 8. Box plot for age based on being self-made

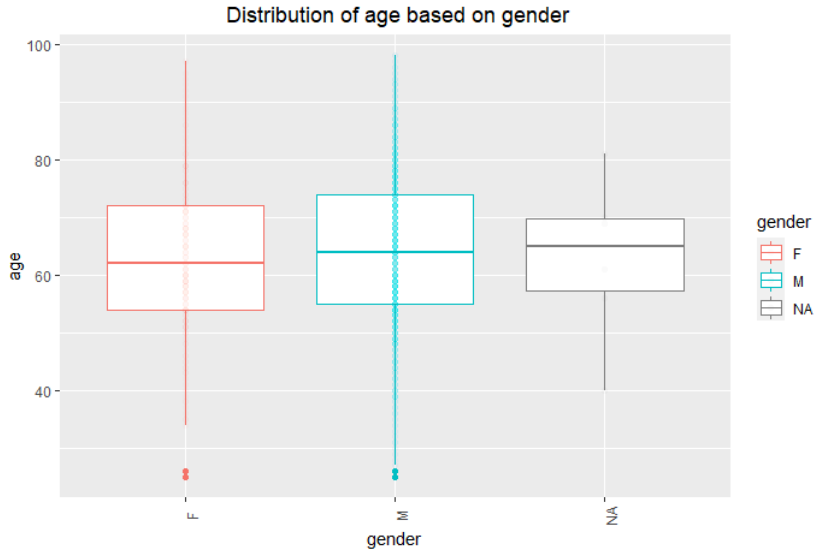

Figure9. Box plot for age based on gender

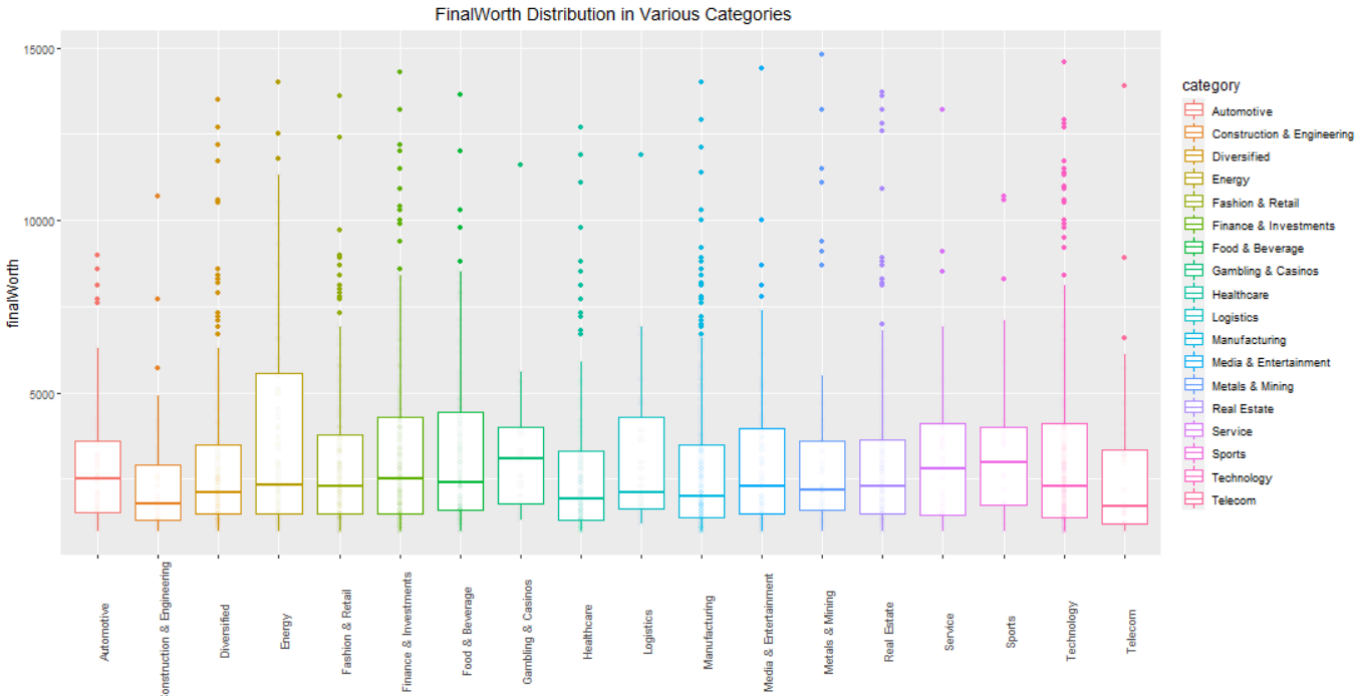

Figure 10. Box plot for final worth distribution based on category

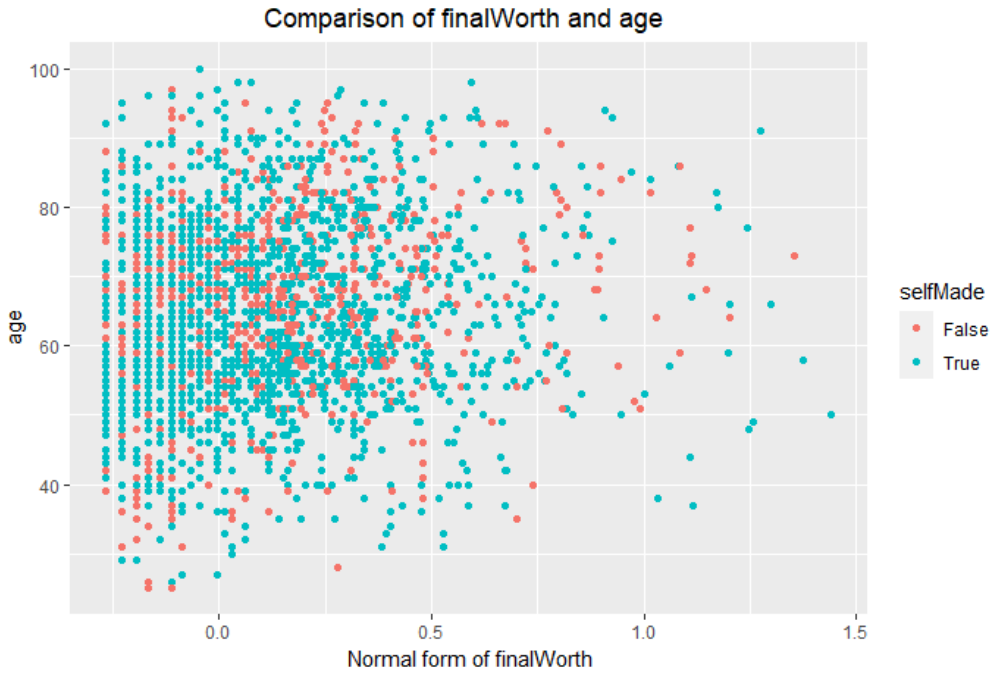

Figure 11. Scatter plot for normalized format of final worth vs. age

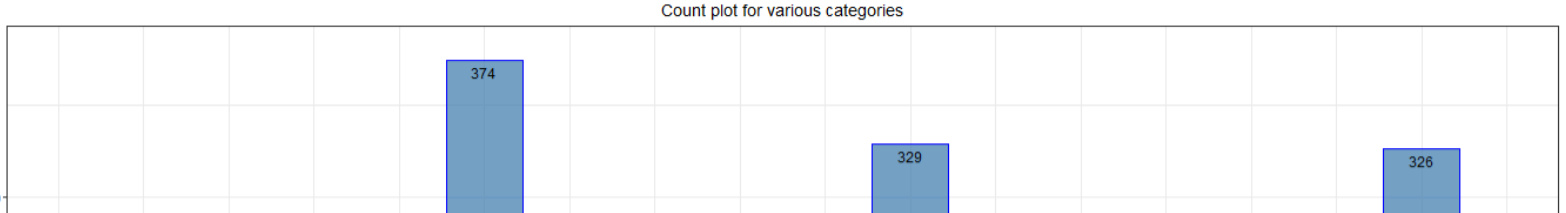

Figure 12. Distribution of billionaires based on category

7 Bibliography

Prasert Kanawattanachai (2022). *Forbes World's Billionaires List 2022*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/prasertk/forbes-worlds-billionaires-list-2022?select=forbes_2022_billionaires.csv [Accessed 23 Apr. 2022].

5 Most important Data Pre-Processing Techniques—Impute missing data—Part II » DevSkrol. (n.d.). Retrieved February 28, 2022, from https://devskrol.com/2022/01/25/imputation-of-missing-data/

Delignette-Muller, M.L. and Dutang, C. (2015). fitdistrplus: AnRPackage for Fitting Distributions. *Journal of Statistical Software*, 64(4). doi:10.18637/jss.v064.i04.