Abstract

Introduction

Methodology

Results and Discussion

-Descriptive analysis

-Linear analysis

-Logistic analysis

Conclusion

Introduction:

With a simple research online, it is simple to discover the amount of billionaires in the world: based on 2022 statistics, we can see that 59,4 million people have more than one billion dollars in their bank, and this is equal to the 0,9% of the world population. At first impact, this amount is greater than expected, and this leads to some questions that probably every individual has asked himself once in his life, that are: how did those people accumulate such an amount of money? Do they have particular requisites? Or is it just the fruit of the received inheritance?

The goal of our analysis is to answer those questions… and many others, conducting a statistical analysis using multiple variables that may influence the worth of a person, and programming tools to facilitate our work.

Given the huge amount of rich people in the world, we decided to conduct our research on a particular category of those: the Billionaires. From our findings, we saw that nowadays there are around 3000 people owning this amount. Eventually, we will discuss the particularities of our dataset in the next paragraph.

Our dataset:

Let's start the discussion presenting our dataset: Our initial dataset presents 3000 observations ordinated by a rank; and each one is referred to a different billionaire (or family) in the world. The variables included are:

* finalWorth (our initial dependent variable): it summarizes the amount of money (worth) of the billionaire or of the family.
* personName: indicates the name of the billionaire or of the family.
* age: indicates the age of the billionaire or of the main exponent of the family.
* country: indicates the country from where the billionaire was born.
* city: indicate the city where the billionaire was living when the dataset was created.
* source: indicates that activity runned by the billionaire that made him possible to increase its worth.
* industries: summarizes the industries in which the billionaires activities are, or from which their worth comes (ex: investments).
* selfMade: it’s a boolean variable indicating whether the billionaire became rich with his own hands, or inherited the most of its worth. A value equal to 1 refers to a self-made billionaire, while a value equal to ‘0 refers to a non self-made billionaire.
* gender: indicates if the billionaire is male or female
* title: indicates the role of the billionaire in its company or its general position
* birthYear: the year in which the billionaire was born
* birthMonth: the month in which the billionaire was born
* birthDay: the day in which the billionaire was born
* cpi\_country: the consumer price index of the country from where the billionaire comes from. This measure is universally recognized as a good proxy for the inflation
* gdp\_country: the gross domestic product of the country from where the billionaire comes from. In a certain way, we can estimate how “rich” a billionaire country is through this variable.
* gross\_primary\_education\_enrollment\_country: this variable was taken from the World Bank Databank (<https://data.worldbank.org/indicator/SE.PRM.ENRR>), and it’s a good approximation for the level of instruction in the billionaire’s country.
* life\_expectancy\_country: it refers to the average age of death in the billionaire's country. Also this statistic was taken from the World Bank Databank.
* total\_tax\_rate\_country: this variable indicates the tax rate that is applied on the billionaires in the country in which they live.
* population\_country: here we have the amount of people currently living in the billionaire’s country

In addition to our “initial” dataset, we created another one to conduct further studies. The reasons that pushed us to create the upgraded dataset, and the way (in terms of practices and modifications) we used to obtain it, are going to be analyzed later on. For now, we will just show the “new” variables that we are going to use in the further studies. Those are:

* being\_graduated: this is another boolean variable. After numerous researches online, we manage to add manually whether all the considered billionaires have graduated from University, or not. We thought that it was a good proxy to describe their background culture, and if this factor can influence in some ways the final worth (especially interesting for the billionaires that are self-made) or other factors, like the probability of having a happy marriage.
* DIVORCED\_YES\_NO: this is a particularly interesting variable. It was a great work for us to check online if all the billionaires have ever divorced once in their life or not. a value equal to 1 suggests that their marriage was not happy; a value equal to 0 indicates instead a prosperous one (till now).
* religion: here we can see in which entity the Billionaire believes. We thought it was interesting to check if this kind of culture-background information affects those people’s worth.
* human\_right\_respect: for this and the following variable we used two different approaches; here, we asked to an AI tool (chat gpt 3.5 - <https://chat.openai.com/>) to rank the billionaires countries in three different levels, based on “the human right respect - perceived - in that area”. We are aware of the fact that this approach is not 100% statistically correct, but still, we are very interested to explain our dependent variables not only with the information derived from the person itself, but also from his background, trying to address aspect like “the context in which he was raised” in our computation.
* democracy\_level: here comes the other approach; we found that the famous journal “The Economist”, classifies every year all the countries in the world depending on the level of democracy perceived. The “democracy index” (<https://en.wikipedia.org/wiki/The_Economist_Democracy_Index>) is nowadays widely used, so we thought it was a good idea to classify the countries through this variable.
* n\_children: our last added variable indicates the number of children that the billionaire has overall.

Our Approach:

After having shown the object of our analysis, we want to clarify the methodology used and the goal that we want to reach.

Even though our reports appear quite straightforward till now, our main aim is probably not something you would expect. Before finding the dataset, we were asking ourself if datasets found online are really reliable to run statistical analysis. In the Business Analytics class we learnt that sometimes it is even difficult to read “real” datasets with programming tools; a lot of adjustments have to be done, quite often we can run into missing values, and sometimes the given variable are “not enough” to give a statistical significance answer to the problem analyzed.

Our aim is to do a comparison: in the first part, we will run a linear regression, in order to see how all the mentioned independent variables affect the finalWorth of the employee. In this section we will first customize the original dataset using a feature engineer approach (eliminate at least most of the missing values and run the data on excel, so that it is at least possible to obtain some results). After a brief descriptive analysis on the original dataset it will be possible for us to understand the relationship between the variables, and we will be ready to answer our first question.

In the second part, we will use the “improved dataset”. This was obtained by:

* reducing the number of observations from 3000 to 300.
* eliminating every missing (NaN) value in the dataset. We did research online to find all the factors related to every billionaire. We believe that this will make our analysis more accurate and more meaningful, rather than replacing all the values with “the mean of the observations”.
* adding the already mentioned “new variables”. Those will help us to explain a bigger portion of the dependent variable.
* Fixing the errors: it was common that all billionaires with a feature in common (ex. same Nation), had different values in other related variables. In the new dataset it was possible for us to manually fix those errors.

We are aware that the reduction from 3000 to 300 observations will have a big impact on our investigation. At the same time, we predict that all the adjustments in term of new variables, fix of the errors and replacement of NaN values, will more than compensate this reduction, and help us the have at least some feedbacks.

We want to be clear on our purpose: even though we don’t expect a high level of significance from our study, we know that our computations will help us to understand “in broad terms” the relationship between the variables. Given this, we want to give a suggestion in case of further studies - in order to get more remarkable results, we recommend to increase the number of observations and use more variables related especially to the singles billionaire (and not only about his country), which for us student was impossible to find and add by hand.

In the last part of our survey, we will check for explanations changing the dependent variable. One approach is related to the use of the logistic regression tool, in order to see how all the variables predict the probability for a billionaire to get divorced or not. More discussions will be done in the last paragraph.

A clarification: the dataset was “read '' using excel, while all the computations and statistical analysis are done with the coding language “Python”. We used several anaconda’s and jupiter libraries in order to import statistical or mathematical functions, necessary for our operations.

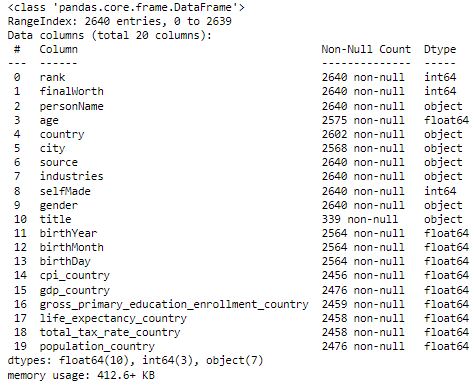
Our path:

Let’s start our investigation. We will divide the work in different segments:

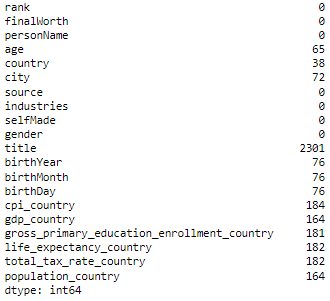
1. Feature Engineering on the original dataset
2. Descriptive Analysis on the original dataset
3. Linear Regression on the original dataset, using finalWorth as dependent variable
4. Feature Engineering on the new dataset
5. Descriptive analysis new dataset
6. Linear Regression on the new dataset, using finalWorth as dependent variable
7. Logistic Regression on the new dataset, with DIVORCED\_YES\_NO as dependent variable

**Feature Engineering on the original dataset:**

After having read the dataset, we had two major problems: on one side, some of our variables were considered by Python with a different “type” compared to the correct one that we were expecting. The variables cpi\_country and gdp\_country were seen as objects, probably because the dataset was presented with commas instead of dots and vice versa. Furthermore, the boolean variable selfMade was not read correctly (of course, before changings also its type was object). We converted the boolean values in “1” for self-made and “0” for non-self-made, using a Python function.

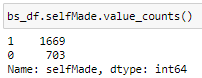


After having solved those problems, we moved to the other side: practically, our original dataset was full of missing values for almost every variable! As mentioned, our goal was to “customize” the dataset only in the second part of the project, in order to show the differences in the results using the two datasets. So, here we modified some of the missing values (mostly by eliminating the observation, sometimes using a Python function to substitute the mean of the variable in the place of the NaN value). İt was “just the necessary” to have the result in the regression.

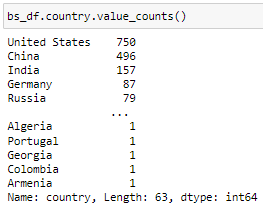
(original dataset missing values)

**Descriptive Analysis on the original dataset:**

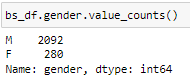
In this part our goal was to show the relationship between the original variables and our initial dependent variable on finalWorth. We know that through the tool of linear regression it is possible to see how an aggregate of independent variables affects the dependent one. With descriptive analysis, our goal was instead to see the relationship one by one, of some variables on finalWorth, visualizing with graphs the results. We will show our results with small comments in order to highlight the most significant insight of the relationship.



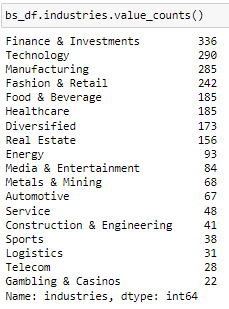
In our dataset there are more self-made billionaires than non-self-made. More than double!



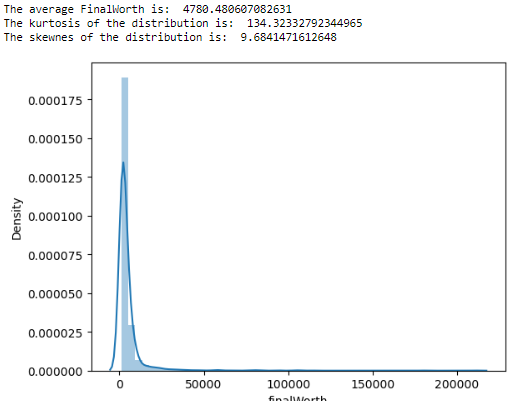
The United States and China are the most represented countries.



The number of female billionaires is way less than the mans.



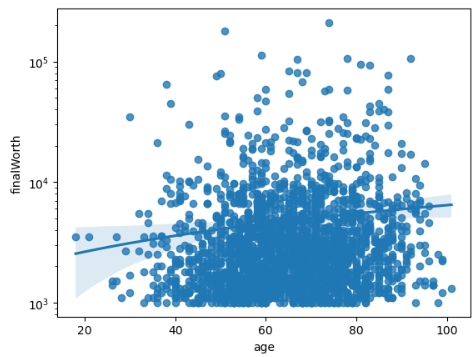
Here we can see the representations in all the industries from the billionaires.



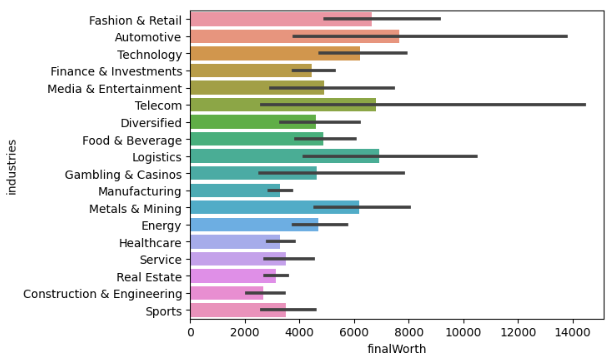
This graph shows the distribution of final worth. It clearly doesn’t follow a normal distribution, due to high kurtosis and right skew.

A high kurtosis indicates that most of the observations fall close to the mean of the distribution. In our case, this is caused by some “very very rich billionaires”, that make the distribution have “a long tail”.

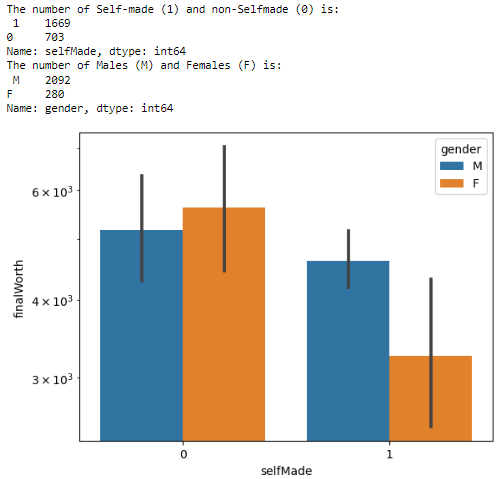
A right skew, indicates instead that the observation on the right side of the mean (and so, related to “richier” billionaires, affect the distribution more than the one on the left side. We interpreted this as if “there is more money over the mean”.



This graph shows a positive linear relation between the finalWorth and the age of billionaires: the more the billionaires are old, the more they are on average richer.

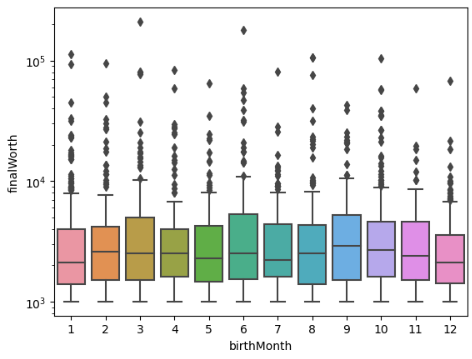


The industries that make billionaires richer are Automotive and Logistics, while the least profitable are Manufacturing and Constructions and Engineering

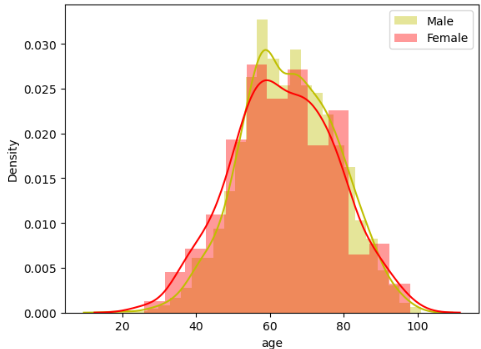


This graph shows an interesting and unexpected result: on one hand, the self-made men are on average wealthier than self-made women. On the other hand, non-self-made women are on average richer than non-self-made men.

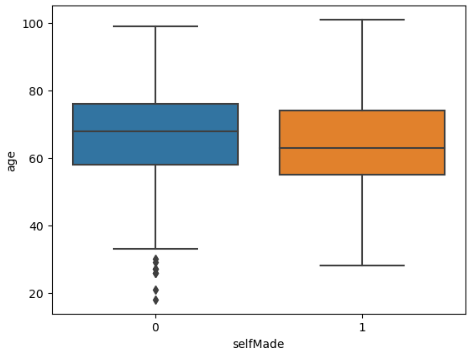
Probably this result derives from the fact that many women increased their wealth due to billionaire divorces



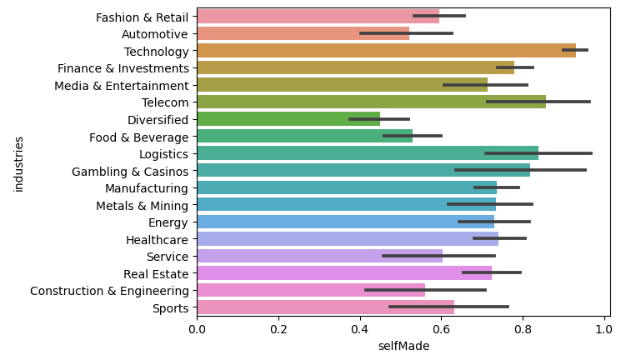
A bad news for people that believe in the horoscope: unfortunately it seems that on average the month of birth doesn’t affect the finalWorth (of course, we were expecting this result)



The age distribution of billionaires is very close to normal and has a similar pattern for both men and women.



This graph shows instead that non-self-made billionaires are usually older than self-made ones.

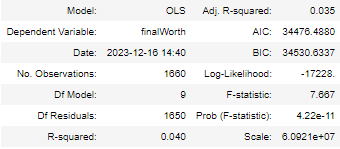


This last bar plot indicates that most of the self-made billionaires come from the Technology industry; we were expecting this result.

**Linear Regression on the original dataset, using finalWorth as dependent variable:**

In this part we will conduct a linear regression. The first step is dropping from the list of the independent variables, all the ones that are insignificant for the regression. those are: personName, city, source, birthYear, birthDay, title and rank. Secondly, we created dummies for the object variables.

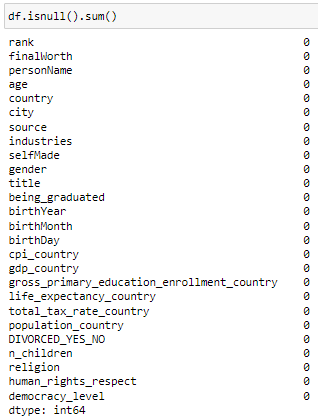
After this step we can eliminate the correlated variables, that are only “life\_expectancy\_country” and “country\_China”, and run the regression (dropping all the variables with a p-value > 0.05). Here the results:





The R-squared is really low: it means that overall, the independent variable does not explain much of the dependent one. We probably should include more independent variables or change the object of our analysis. We also note another strange result: even though the p-value for the variable gdp\_country is close to 0, it seems that its coefficient is 0 (a p-value lower than 0.05 suggest that with that level of significance, we can statistically say that that variable is affecting the dependent one - and so that the coefficient is statistically different from 0).

We can interpret the results as follows: being older (the more the billionaire has a higher age), coming from France (respect to the benchmark country), coming from Fashion & Retail, Metal & Mining. Technology or Diversified industries (compared to the benchmark industry), increase the finalWorth. While being self-made and a higher tax rate, decrease the finalWorth.

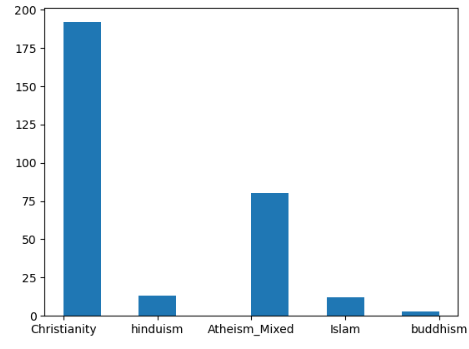
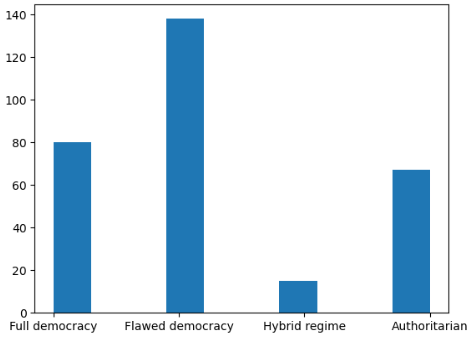
**Feature Engineering on the new dataset:**

As already mentioned, we created the new dataset by

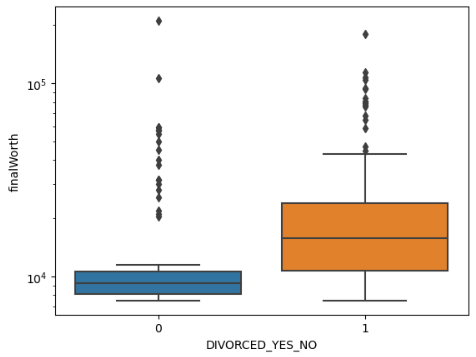
* reducing the number of observations from 3000 to 300.
* eliminating every missing (NaN) value in the dataset.
* adding new variables.
* Fixing the errors.

**Descriptive analysis new dataset:**

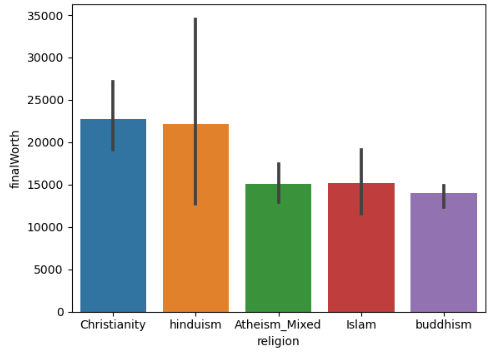
Here our attention is concentrated on the relationship between the new variables. Again, we are going to show the results with a brief comment.



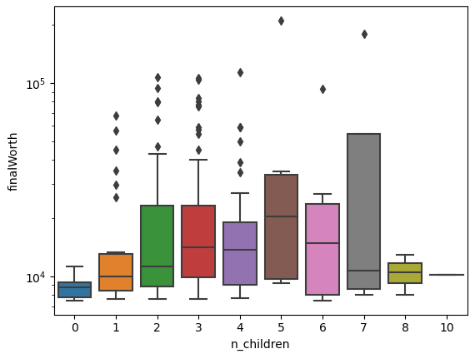
Through the bar plots we can see that Christianity is the most popular religion and that most of the billionaires come from countries in which the level of democracy is flawed.



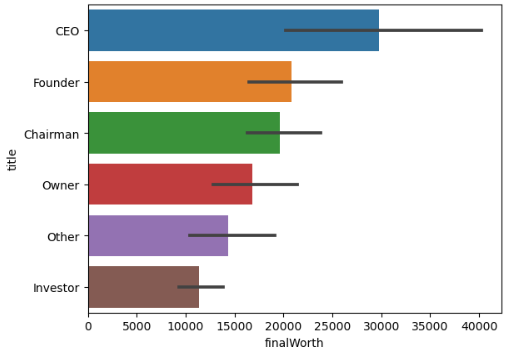
Divorced billionaires have on average a higher final worth. It is interesting to see that the gap between the 25% and 75% percentile is way bigger in the divorced case. We also noted a big amount of outliers.



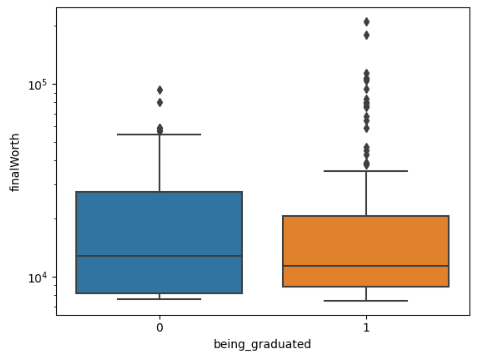
Christinanity and Hinduism religions lead to higher final worth. This result seems strange to use and might be affected by the huge reduction of observations.



This graph shows an interesting relation between the number of children and the final worth. Despite all the outliers, we can see a positive trend, since the more the number of children increases, the more, on average, the final worth increases. There were very few billionaires with more than 5 children, so we feel free not to consider those ones in our analysis. We will check through linear regression if our interpretation is correct.



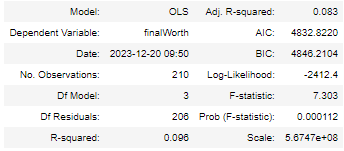
CEO’s are on average richer: this is also an unexpected result that may be influenced by the small amount of observations.

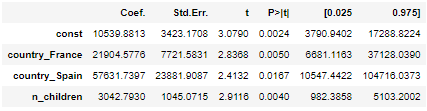


Finally, it seems that on average being graduated does not affect significantly the final worth of the billionaires.

**Linear Regression on the new dataset, using finalWorth as dependent variable:**

The second linear regression is yet to be run. We are expecting of course different results from the previous one due to all the changes applied. after having completed al the necessary adjustment as in the first case (creation of X and Y variables, dropping the same insignificant variables, creating dummies for categorical variables, adding a constant to the X, checking correlation between the independent variables - here the correlated variables are more - splitting in train and test using a test size of 0.3), we run the regression and obtained the following results:





The R-squared value did not change much, meaning that we didn’t manage to improve our analysis through the new dataset. Here we have just three significant variables, that are country\_France, country\_Spain and n\_children, and they all have positive coefficient (of course the variables related to the nations, have to be compared to the “benchmark state”, and the positive coefficient indicate that coming from France or Spain increases the final worth more compared to the benchmark State).

It is really interesting the result for the n\_children variable. Having more children increases the finalWorth, and the same result can be seen in the graph in the descriptive analysis part.

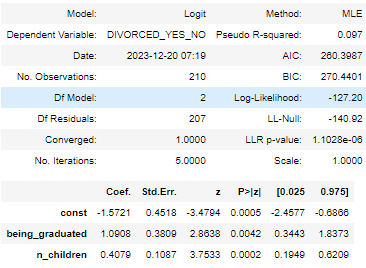
**Logistic Regression on the new dataset, with DIVORCED\_YES\_NO as dependent variable:**

Given the non-significant results, we decided to use our dataset with a different approach. Since  our main goal is to demonstrate if a dataset took online can give reliable results, and till now the answer in our case is no, we wanted to investigate if changing the dependent variable and the tool used to run the analysis, we can improve it (initially, we read that our dataset was meant to run a linear regression on the variable finalWorth). Here we are going to run a logistic regression, using as a dependent variable the new “DIVORCED\_YES\_NO”.

The overall goal is: given a series of information about an entity, place it into the correct category. The information is related to all the variables in our dataset; our categories are: a billionaire that is not divorced; a billionaire that has divorced.

With our analysis we want to predict whether a billionaire is likely to divorce or not given some information (here the dependent variable is categorical).

After all the passages already repeated in the linear regression part, we obtained this output:



It seems that the Pseudo R-squared is very low again, and that the only two significant variables are: being\_graduated and n\_children.

**Conclusion:**

With our results we managed to answer the main question that we asked ourselves. Unfortunately it is very tough to improve a given dataset, especially when the number of observations is really high. Searching and adding new variables, filling all the missing values for every different observation, coping with errors etc… can take a lot of time, energy and resources. Even though we managed just to slightly improve the dataset, we are aware of the possible sequential changes that can increase the significance of the dataset even more, till reasonable results. Those are:

* Incrementing again the number of observations; even considering not only  billionaires in the world (that are around 3000), but also some of the richest millionaires would be beneficial for the interpretations.
* Control if ALL the values in the dataset are correct. This was a task too big for a group of students like us, but, if in the future there will be a possibility to implement a new software that can check whether numbers are reliable or not, it would improve the computations a lot.
* incrementing the number, but especially the “quality” of the variables. As it can be seen, most of the variables added are related to the country from which the billionaire comes, and those ones were not explaining the final worth probability because they are too generic. On the other hand, it resulted that the new specific variables like n\_children or being\_graduated were way more significant. For us it was impossible to add further variables related to the “billionaire itself”, because, as already mentioned, all we did was by hand! For further studies we suggest adding more qualitative variables (especially if it is possible to use a system such as an AI tool to do the work for you, and not doing everything by hand). One suggestion is using the Total amount of money given in charity, even though being sure of the reliability of the information found online may not be the right approach.