a.a. 2022/2023

Project report for the

Data Management course

RICHEST PEOPLE IN THE WORLD AND GDP

Belotti Alessandro	896985
Castagna Alessandro	898057
Salvatori Ilaria	898009

Index

Roles:

Alessandro Belotti: data acquisition, data integration, data quality

Alessandro Castagna: data acquisition, data storage, research questions

Ilaria Salvatori: data acquisition, data quality, research questions, documentation

1. Introduction

The objective of this project is to consider whether the presence of billionaires on the territory leads to an increase in the wealth of the country (in terms of GDP per capita) in which they are residents or, more specifically, of the Region or Province. For this purpose, we have chosen to use the data provided online by Forbes magazine regarding the ranking of billionaires in the world. Here, in relation to each person, we could find information such as: the growth of assets from 2014 to 2021 (since the data relating to the GDP of each country are available until that year), the residence, age, gender, type of industry in which they operate. As for the evolution of the heritage, continuing the discussion will notice that some personalities have climbed the ranking only from a more recent date, while others have long been part of it between "highs" and "lows".

Therefore, the heritage of the billionaire is observed with respect to territorial wealth both in terms of state, both in terms of province or region. In particular, in the United States were considered both the total GDP of the country and the GDP of each Member State; similarly, in Europe were considered the wealth produced by individual states and also the regions and provinces that make them up. As regards the territorial division of Europe into smaller units than the States, the nomenclature of territorial units statistics (NUTS) has been taken as a reference, which divides the territory of the European Union for statistical purposes. In particular, there are three levels of NUTS, which divide the territory differently according to the population:

- NUTS 1: according to this nomenclature, the territories covered are 97 and refer to areas such as the Federal States of Germany, the Regions of Belgium, Denmark, Sweden, etc. At this level Italy is divided into supra-regional areas: North-West, North-East, Central, South, Islands. The population here is between 3 and 7 million.
- NUTS 2: in this case 270 territories are considered, including the Italian regions as we know them, the autonomous communities of Spain, the French regions, etc. Here the subdivision is based on a population ranging from 3 million to 800,000. This is the nomenclature used for GDP considerations of European regions.
- NUTS 3: here are considered the smallest territories, such as the Italian Provinces and similar territorial subdivisions of other member countries. The reference population varies between 150,000 and 800,000.

We decided to focus the analysis on Europe and the United States because it was considered interesting to compare the two protagonists of the Western political bloc, which, although sharing international understandings, present great cultural and economic differences. Moreover, most of the billionaires extracted from the Forbes ranking are American or European, so it was considered interesting and more important to conduct an analysis that observed these two giants. Moreover, the choice of observing individual American states and individual European regions at NUTS 2 level is explained by the intention to carry out a more extensive investigation, which would not have been possible if we had focused on the national level.

Moreover, the observation of GDP in smaller areas than in the Member States makes it possible to observe more clearly whether the development of the assets of the billionaire goes hand in hand with the development of the GDP of the Region or of the American State in which he resides (certainly, this trend is less observable taking as a reference the GDP of a whole state, such as the United States) and whether this can therefore be read in terms of the positive or negative

contribution of the individual's assets to the GDP of the region of residence.Un'ulteriore precisazione va fatta riguardo la scelta del periodo di analisi:

- 2021 was chosen as the final year of the analysis because at the time of data collection were not yet available the results relating to GDP 2022 (released only at the end of May 2023) but only forecasts and we preferred to keep our analysis adherent to certain data.
- It was decided to start the analysis in 2014 because the billionaires currently listed are in the ranking from a relatively short time and therefore it would not have made sense to recover the trend of their assets in previous years, since, in fact, irrelevant. Moreover, this choice is compatible with the desire to analyse GDP values starting from where the economic crisis was left behind.

The research questions we want to answer are:

- Do countries with more billionaires have higher average GDP per capita?

- Who are the billionaires with the highest correlation between their worth trend and their Region/State GDP per capita trend?

- What is the percentage of women billionaires in the top 10 richest (with the highest average GDP per capita) countries?

1. Data Acquisition and Data Cleaning

In order to pursue our purpose, we have used data collected both from the Forbes platform relating to the subjects of our interest, and from sites that allow us to obtain information related to the GDP Countries and individual Regions. In particular, to follow we report the 7 datasets and the relative source from which it has been possible to draw, let alone the modality of acquisition.

2.1. Sources

1. **Information about individual billionaires** → true.json
   Through the APIs made available by the Forbes API section, it was possible to make free calls from the https://www.forbes.com website: in particular, the information of our interest was entered in the call url. The result of this search has been inserted in a dictionary, obtaining an output as follows:

```
[{'rank': 1,
  'listUri': 'rtb',
  'finalWorth': 214258.803,
  'personName': 'Bernard Arnault & family',
  'city': 'Paris',
  'source': 'LVMH',
  'industries': ['Fashion & Retail'],
  'countryOfCitizenship': 'France',
  'imageExists': True,
  'gender': 'M',
  'birthDate': -657244800000,
  'lastName': 'Arnault',
  'wealthList': False,
  'estWorthPrev': 214206.544307,
  'familyList': False,
  'squareImage': '//specials-images.forbesimg.com/
cropX2=4000&cropY1=1209&cropY2=5212',
  'bioSuppress': False},
```


2. **Development of the assets of billionaires** → all_worth_trend_bill.json
   From the Forbes website, https://www.forbes.com, individual billionaires were scrapped to record their annual asset developments from 2014 to 2021. This result has also been included in a dictionary, where the name of every billionaire has been associated with the performance of its assets. From this phase it emerged that some billionaires entered the ranking after 2014, or came out before 2021, so not all report the same time range.


```
Elon Musk-->{2014: 8.4, 2015: 12.0, 2016: 10.7, 2017: 13.9, 2018: 19.9, 2019: 22.3, 2020: 24.6, 2021: 151.0}
Bernard Arnault & family-->{2014: 33.5, 2015: 37.2, 2016: 34.0, 2017: 41.5, 2018: 72.0, 2019: 76.0, 2020: 76.0, 2021: 150.0}
Jeff Bezos-->{2014: 32.0, 2015: 34.8, 2016: 45.2, 2017: 72.8, 2018: 112.0, 2019: 131.0, 2020: 113.0, 2021: 177.0}
Larry Ellison-->{2014: 48.0, 2015: 54.3, 2016: 43.6, 2017: 52.2, 2018: 58.5, 2019: 62.5, 2020: 59.0, 2021: 93.0}
Warren Buffett-->{2014: 58.2, 2015: 72.7, 2016: 60.8, 2017: 75.6, 2018: 84.0, 2019: 82.5, 2020: 67.5, 2021: 96.0}
Bill Gates-->{2014: 76.0, 2015: 79.2, 2016: 75.0, 2017: 86.0, 2018: 90.0, 2019: 96.5, 2020: 98.0, 2021: 124.0}
```


3. **Residence State of each American billionaire** → American_bill_byProvince.json
   Also from the Forbes website https://www.forbes.com that allowed us to extract the information of our interest through the API, we were able to extract and insert in a dictionary the information about the specific American state in which the billionaire resides. This operation was useful and necessary for the subsequent integration of the information concerning the GDP produced by each individual US State.


```
Elon Musk,Texas
Jeff Bezos,Washington
Larry Ellison,Hawaii
Bill Gates,Washington
Warren Buffett,Nebraska
Steve Ballmer,Washington
```


4. **Residence province of each European billionaire** → European_bill_ byProvince.json
   To obtain this information, a scraping operation was carried out from the Geonames site: https://www.geonames.org/search.html?q=galicia&country=. In particular, in the search bar of the site was inserted the name of the European city of residence of the billionaire and in this way was extracted the Province of residence.

```
Bernard Arnault & family, Île-de-France Paris
Francoise Bettencourt Meyers & family, Île-de-France Paris
Amancio Ortega, Galicia A Coruña
Dieter Schwarz, Baden-Wurttemberg Regierungsbezirk Stuttgart
FranÃ§ois Pinault & family, Île-de-France Paris
Giovanni Ferrero, Brussels Capital Bruxelles-Capitale
```

5. **European GDP per capita dividev by European Province** → european_GDP_procapite.csv
   To obtain this information you have downloaded from: https://ec.europa.eu/eurostat/databrowser/view/nama_10r_2gvagr/default/table?lang=en

	TIME	2014	2015	2016	2017	2018	2019	2020	2021
1	Région de Bruxelles-Capitale/Brussels Hoofdste...	57,800	59,500	59,400	60,600	61,600	63,400	60,900	66,200
2	Prov. Antwerpen	37,600	39,100	39,900	40,900	41,900	43,400	43,000	46,900
3	Prov. Limburg (BE)	26,600	26,900	27,400	28,300	29,100	29,700	29,200	32,300
4	Prov. Oost-Vlaanderen	29,100	30,400	31,000	32,000	32,600	33,500	33,000	36,300
5	Prov. Vlaams-Brabant	34,500	35,800	36,700	37,600	38,700	39,900	36,500	39,800
...	...	...	...	...	...	...	...	...	...

6. **American GDP dovided by State** → US_states_GDP.csv
   To obtain this information you have downloaded from: https://apps.bea.gov/itable/?ReqID=70&step=1&acrdn=1#eyJhcHBpZCI6NzAsInN0ZXBzIjpbMSwyNCwyOSwyNSwzMV0sImRhdGEiOltbIlRhYmxlSWQiLCI2MDAiXSxbIkNsYXNzaWZpY2F0aW9uIiwiTm9uLUluZHVzdHJ5Il0sWyJNYWpvcl9BcmVhIiwiMCJdXX0=

	GeoName	2014	2015	2016	2017	2018	2019	2020	2021
0	United States	16932051.0	17390295.0	17680274.0	18076651.0	18609078.0	19036052.0	18509143.0	19609812.0
1	Alabama	189886.3	191335.2	194283.8	196974.9	200372.6	203432.7	199880.8	209979.3
2	Alaska	54188.2	54740.8	54246.6	54278.7	53327.0	53433.8	50705.2	50869.4
3	Arizona	276948.9	282577.0	291275.2	303606.1	314827.5	325395.3	327178.0	347656.0
4	Arkansas	111734.5	112351.0	112798.1	113850.2	115885.2	117126.2	117268.2	123347.3
5	California	2256054.7	2357452.9	2427894.6	2538204.0	2644061.2	2729225.8	2667220.9	2874730.8
6	Colorado	298655.3	312409.7	318953.4	329913.3	342733.2	358438.5	353345.2	373763.3

7. **American population** → Population_us.csv
   From the site https://www.census.gov/en.html was taken the value of the population present in each American state to 2022. This data is fundamental because the data concerning the GDP of the American States are not per capita and, to make them such, it is necessary to relate the GDP of each State to the population present in that State. The desire to have the per capita figure of GDP is explained by the fact that the GDP of the European regions has a given per capita and therefore, in this way, it has been possible to make a coherent comparison in terms of observed values.

8. **GDP per capita for each Country in the world** → world_GDP.csv
   Data was obtained from the site https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.ZG
   Here only the United States and the States of Europe will be filtered and maintained.

	Country Name	2014	2015	2016	2017	2018	2019	2020	2021
0	Aruba	36846.848290	37343.912960	37583.840320	38865.188200	41679.238190	42501.641550	34971.009910	42698.359870
1	Africa Eastern and Southern	3470.550675	3498.125396	3591.099809	3635.564622	3724.868177	3777.972368	3621.058106	3839.470058
2	Afghanistan	2110.829568	2128.125938	2023.834656	2096.093111	2109.929296	2167.704111	2076.138380	1665.805842
3	Africa Western and Central	4143.177054	4075.994459	3998.951876	4045.303263	4160.540965	4264.731035	4174.504565	4409.450961
4	Angola	8123.048065	7274.090475	7027.146634	7216.061373	7042.923829	6881.076241	6362.636076	6491.125578
...	...	...	...	...	...	...	...	...	...

9. **Countries and continents** →Countries and Continents.csv
   Data relating to countries and continents around the world were found at the following link: https://www.kaggle.com/. A data cleaning operation was required in this dataset, since American states were grouped under the heading "United States of America"That is the name of a state, not a continent. However, in order to make a matching that we will see later, it was necessary that these countries were under the North American continent indicated by the acronym NA; this correction was made manually in python.

2.2. Data cleaning and pre-enrichment

In this step, both the data cleaning and the pre-enrichment phases were carried out simultaneously. It was considered appropriate to speak of pre-enrichment because here we tried to find the points of contact between the different datasets so that we can perform with agility the real phase of integration and enrichment later on Mongo DB.
The phases of what was defined pre-enrichment were as follows:

1) Since our analysis focuses on European and American billionaires, it has been necessary to attribute to each subject belonging to a certain continent. For this purpose, starting from the ***true.json*** file, through a scraping procedure, the *city of residence* of each billionaire were extracted, in order to insert them in the search bar of the Geonames site and thus obtain the State of residence. The information about the continent has been added thanks to a dataset called Countries and Continents, which for each state of the world (including the states of residence of billionaires) associates the Continent of origin (the association was simple because the strings of the name of the State taken from Geonames and the name of the State present in Countries_and_Contiens.csv are equal).
   In this way, the information from the true.json has been enriched with the specifications highlighted in green:

{'uri': 'bernard-arnault', 'rank': 1, 'listUri': 'rtb', 'finalWorth': 238188
.732, 'personName': 'Bernard Arnault & family', 'city': 'Paris', 'source': '
LVMH', 'industries': ['Fashion & Retail'], 'countryOfCitizenship': 'France',
'imageExists': True, 'gender': 'M', 'birthDate': -657244800000, 'lastName':
'Arnault', 'wealthList': False, 'estWorthPrev': 237579.9297, 'familyList': F
alse, 'squareImage': '//specials-images.forbesimg.com/imageserve/5dc05518ca4
25400079c659f/416x416.jpg?background=000000&cropX1=0&cropX2=4000&cropY1=1209
&cropY2=5212', 'bioSuppress': False, 'continent': 'EU', 'residence_state': '
France'}

2) Since our aim is also to observe the wealth of the individual billionaire relative to the wealth (in terms of GDP) of the place where he resides, it was necessary to integrate the dataset ***european_bill_byProvince.json*** with the GDP values of each European region contained in the dataset ***european_GDP_procapite.csv***. To this end, the Regions written in their original language in *European_bill_byProvince.json* have been translated into English to bring them into line with those of *european_GDP_procapite.csv*; Furthermore, the incorrect characters have been replaced with spaces and slash symbols have been removed.
   At this point, through the similarity function, we proceeded with the matching between the names of the provinces of the two datasets. Some matching did not work and, since it was sporadic cases, we proceeded with the manual imputation in python.
   The similarity cosine is a measure of similarity between two vectors in a vector space and represents a particularly widespread obstruction to compare the similarity between texts. It is calculated by measuring the angle between the two vectors in the context of the vector space. The smaller the angle, the greater the similarity between vectors; in particular, it varies between -1 and 1, where 1 indicates perfect similarity and -1 indicates total dissimilarity.
   At the end of the procedure, the new data set, ***prov_matched_bill.json***, replaced the former european_bill_byProvince.json. It contains the name of the billionaire and the region of residence, written in a manner consistent with what is present in the dataset of European GDP; In this way, it will be possible to combine information on European billionaires and data on the wealth of each region in which they live.

3) The ***world_gdp.csv*** file shows the performance of GDP for each country in the world from 2014 to 2021; of our interest are of course the GDP of the USA and the GDP of the European Continent (in which we also find Russian billionaires belonging to the Moscow Region which, geographically, is part of Europe). The GDP values reported in the dataset have been rounded to two decimal places in order to make them more readable, obtaining the following result:

	Unnamed: 0	Country Name	2014	2015	2016	2017	2018	2019	2020	2021
0	0	Aruba	36846.85	37343.91	37583.84	38865.19	41679.24	42501.64	34971.01	42698.36
1	1	Africa Eastern and Southern	3470.55	3498.13	3591.10	3635.56	3724.87	3777.97	3621.06	3839.47
2	2	Afghanistan	2110.83	2128.13	2023.83	2096.09	2109.93	2167.70	2076.14	1665.81
3	3	Africa Western and Central	4143.18	4075.99	3998.95	4045.30	4160.54	4264.73	4174.50	4409.45
4	4	Angola	8123.05	7274.09	7027.15	7216.06	7042.92	6881.08	6362.64	6491.13

In addition, several states have reported missing values, which have been remedied in two different ways depending on the situation:

- Countries reporting all missing values for GDP from 2014 to 2021 were eliminated;
- In countries with at least one GDP value, a mean replacement has been performed.

Thanks to the association between the State in this dataset and the State of residence of each billionaire present in true.json, it has been possible to further enrich the true.json with information on the evolution of wealth of the State in which the billionaire resides.

4) In ***US_states_GDP.csv*** dataset the present values were replaced by the GDP per capita of each State, thanks to the ratio between the wealth of the State indicated in the dataset and the number of population present in the specific State. The transformation of values into per capita has been carried out in order to easily compare the values of GDP in European regions, which, in fact, are expressed as GDP per capita.
Finally, the values originally represented in thousands of dollars, were written respecting the extended figure and a rounding to two decimal places was established.
Here is the difference in the values before (as reported in the data acquisition phase) and after:

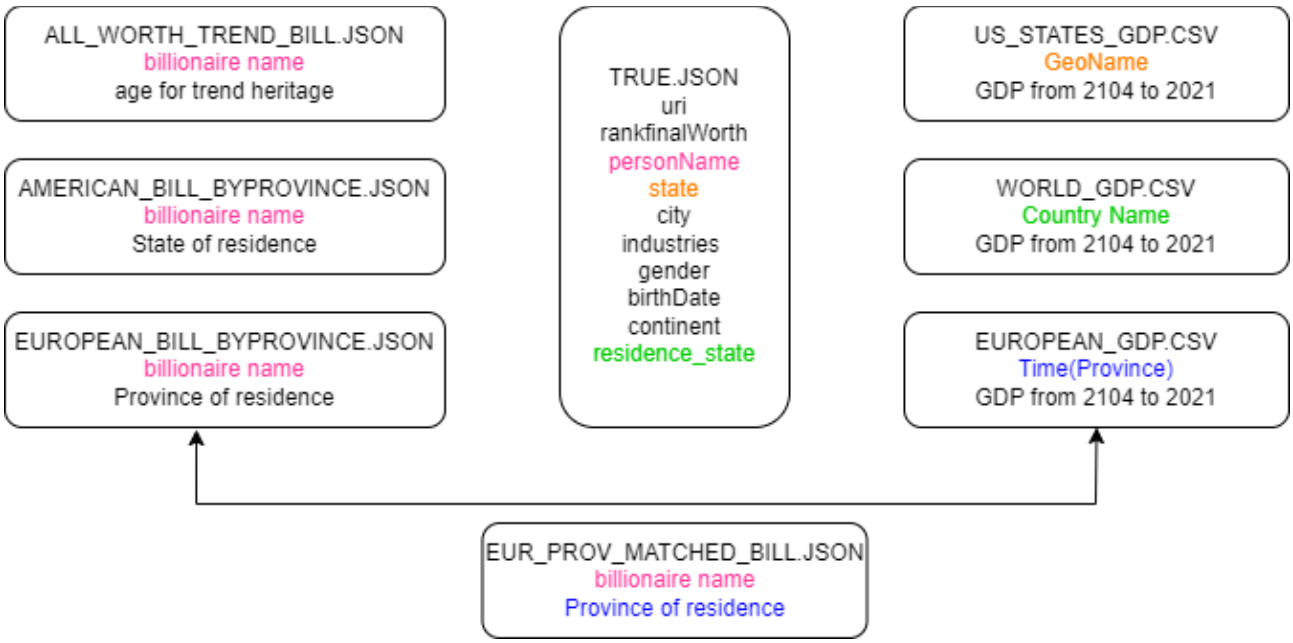
	GeoName	2014	2015	2016	2017	2018	2019	2020	2021
0	United States	16932051.0	17390295.0	17680274.0	18076651.0	18609078.0	19036052.0	18509143.0	19609812.0
1	Alabama	189886.3	191335.2	194283.8	196974.9	200372.6	203432.7	199880.8	209979.3
2	Alaska	54188.2	54740.8	54246.6	54278.7	53327.0	53433.8	50705.2	50869.4
3	Arizona	276948.9	282577.0	291275.2	303606.1	314827.5	325395.3	327178.0	347656.0
4	Arkansas	111734.5	112351.0	112798.1	113850.2	115885.2	117126.2	117268.2	123347.3
5	California	2256054.7	2357452.9	2427894.6	2538204.0	2644061.2	2729225.8	2667220.9	2874730.8
6	Colorado	298655.3	312409.7	318953.4	329913.3	342733.2	358438.5	353345.2	373763.3

Unnamed: 0	GeoName	2014	2015	2016	2017	2018	2019	2020	2021
0	United States	50995.31	52375.43	53248.78	54442.57	56046.11	57332.06	55745.13	59060.09
1	Alabama	37602.39	37889.31	38473.21	39006.12	39678.95	40284.93	39581.56	41581.33
2	Alaska	73807.58	74560.26	73887.13	73930.85	72634.58	72780.05	69063.53	69287.18
3	Arizona	38121.62	38896.32	40093.62	41790.95	43335.56	44790.20	45035.59	47854.35
4	Arkansas	36898.94	37102.53	37250.18	37597.63	38269.66	38679.49	38726.38	40733.93
5	California	57636.24	60226.69	62026.29	64844.41	67548.78	69724.51	68140.45	73441.78
6	Colorado	51392.19	53759.03	54885.06	56771.03	58977.06	61679.60	60803.16	64316.67

5) The ***european_GDP_procapite.csv*** file reported values in euro and, in order to make them uniform with US dollar GDP, the currency was converted to the dollar. Moreover, as in the case of the GDP of the US States, the values expressed in thousands of dollars have been reported respecting the extended figure and maintaining here also a rounding equal to two decimal places. The result obtained is as follows:

	TIME	2014	2015	2016	2017	2018	2019	2020	2021
0	Région de Bruxelles-Capitale/Brussels Hoofdste...	62447.12	64283.80	64175.76	65472.24	66552.64	68497.36	65796.36	71522.48
1	Prov. Antwerpen	40623.04	42243.64	43107.96	44188.36	45268.76	46889.36	46457.20	50670.76
2	Prov. Limburg (BE)	28738.64	29062.76	29602.96	30575.32	31439.64	32087.88	31547.68	34896.92
3	Prov. Oost-Vlaanderen	31439.64	32844.16	33492.40	34572.80	35221.04	36193.40	35653.20	39218.52
4	Prov. Vlaams-Brabant	37273.80	38678.32	39650.68	40623.04	41811.48	43107.96	39434.60	42999.92
...	...	...	...	...	...	...	...	...	...

Below is a graph to make clear about the matching made:



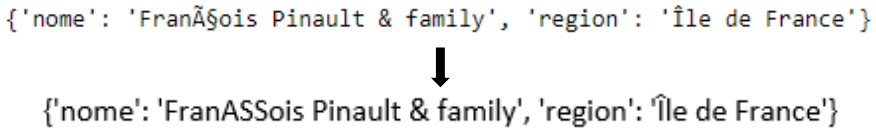
3. Data Storage

Regarding data storage, the json format was chosen for several reasons:

- The data about the billionaires was extracted through the Forbes API, which gave us json files.
- There are different null values in the annual performance of each billionaire's assets, so a relationship approach would be limited;
- The relational approach is to be excluded also because it would not allow to have a certain level of flexibility scheme such as to allow additions of data concerning the patrimony in the following years.

At this stage we chose to use the Mongo DB Compass tool to work on the data we have available. In fact, on Mongo DB Compass you can upload both json and csv files, without having to write lines

of code; in particular, the uploaded csv files are automatically transformed into json, thus homogenizing the format of all our datasets.
In order to import files to Mongo DB Compass, it was necessary to make a transformation of the schema file to all json files, labeling each considered attribute; for example, the expression {Elon musk : Texas} was transformed into {name: Elon Musk, Region: Texas}.
Thanks to these transformations, it was possible to call directly the reference attribute, such as the name or the State to which it belonged; this step was fundamental for the subsequent data integration and enrichment.
In particular, the following datasets have been modified using the unidecode function to translate names containing special characters; through this function, the names have not been translated in a grammatically correct way but in order to allow the subsequent matching.



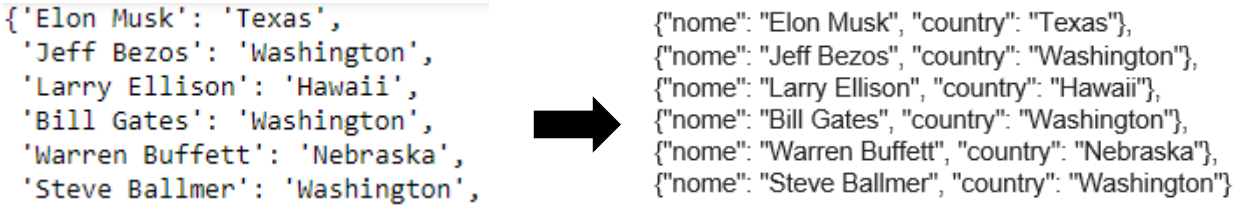
1. **Eur_prov_matched_bill.json**: here the labels "name" and "region" have been added to the list values, resulting in the following result:

{'nome': 'Bernard Arnault & family', 'region': 'Île de France'}
{'nome': 'Francoise Bettencourt Meyers & family', 'region': 'Île de France'}
{'nome': 'Amancio Ortega', 'region': 'Galicia'}
{'nome': 'Dieter Schwarz', 'region': 'Stuttgart'}
{'nome': 'FranASSois Pinault & family', 'region': 'Île de France'}
{'nome': 'Giovanni Ferrero', 'region': 'Région de Bruxelles-Capitale/Brussels Hoofdstedelijk Gewest'}
{'nome': 'Mark Mateschitz', 'region': 'Salzburg'}

2. **all_worth_trend_bill.json**: here again a list of dictionaries was created.


```
{'nome': 'Elon Musk', 2014: 8.4, 2015: 12.0, 2016: 10.7, 2017: 13.9, 2018: 19.9, 2019: 22.3, 2020: 24.6, 2021: 151.0}
{'nome': 'Bernard Arnault & family', 2014: 33.5, 2015: 37.2, 2016: 34.0, 2017: 41.5, 2018: 72.0, 2019: 76.0, 2020: 76.0, 202
1: 150.0}
{'nome': 'Jeff Bezos', 2014: 32.0, 2015: 34.8, 2016: 45.2, 2017: 72.8, 2018: 112.0, 2019: 131.0, 2020: 113.0, 2021: 177.0}
{'nome': 'Larry Ellison', 2014: 48.0, 2015: 54.3, 2016: 43.6, 2017: 52.2, 2018: 58.5, 2019: 62.5, 2020: 59.0, 2021: 93.0}
{'nome': 'Warren Buffett', 2014: 58.2, 2015: 72.7, 2016: 60.8, 2017: 75.6, 2018: 84.0, 2019: 82.5, 2020: 67.5, 2021: 96.0}
{'nome': 'Bill Gates', 2014: 76.0, 2015: 79.2, 2016: 75.0, 2017: 86.0, 2018: 90.0, 2019: 96.5, 2020: 98.0, 2021: 124.0}
{'nome': 'Larry Page', 2014: 32.3, 2015: 29.7, 2016: 35.2, 2017: 40.7, 2018: 48.8, 2019: 50.8, 2020: 50.9, 2021: 91.5}
```


3. **american_bill_byProvince.json**: the names of the billionaires and their Member State have been included in a list of dictionaries. Here it is possible to appreciate the difference regarding how much reported in the data acquisition:



4. **true.json**: This file appears as before, with name correction containing special characters only.

5. **US_states_GDP.json**: the previous csv file has been transformed into json because reading the file on Mongo DB the last column of the dataset was not read; Therefore, to avoid problems of further processing we preferred to directly transform the file in json format, which allowed us to read all the information correctly.

The files thus obtained have been placed on Mongo DB through a process that starts from the creation of an account on Mongo DB Atlas, where it was possible to create the database containing all the datasets of our interest. The choice to make this step on Atlas is explained by the fact that here you can make the connection to Mongo DB Compass.

4. Data integration and enrichment

The datasets described above were merged on the Mongo DB Compass. In a single pipline merging led to the creation of two definitive datasets:

- **dataset about European regions GDP**
  Here the **true.json** was merged with the **eur_prov_matched.json** through *personName* nel true.json and *name* in eur_prov_matched.json; afterwards, **eur_prov_matched** was merged with the dataset **european_GDP_procapite** on the basis of *region* in eur_prov_matched and *time* in european_GDP_procapite. This result was then combined with the dataset **all_worth_trend_bill.json** through *name* of eur_prov_matched and *name* di all_worth_trend_bill.json.
  The final output of this process is as follows: to each billionaire are associated the characteristics present in the true.json, the evolution of the patrimony and the performance of the GDP of the Province or Region in which it resides.
  Below is the code in Mongo db:

```
[
 {
  $project: {
   uri: 1,
   rank: 1,
   finalWorth: 1,
   personName: 1,
```

```
    city: 1,
    source: 1,
    industries: 1,
    countryOfCitizenship: 1,
    gender: 1,
    continent: 1,
    residence_state: 1,
   },
  },
  {
   $lookup: {
    from: "eur_prov_matched_bill",
    localField: "personName",
    foreignField: "nome",
    as: "result",
   },
  },
  {
   $match: {
    continent: "EU",
   },
  },
  {
   $unwind: "$result",
  },
  {
   $lookup: {
    from: "european_GDP_procapite",
    localField: "result.region",
    foreignField: "TIME",
    as: "state/region_GDP",
   },
  },
  {
   $lookup: {
    from: "world_GDP",
    localField: "residence_state",
    foreignField: "Country Name",
    as: "state_GDP",
   },
  },
  {
   $lookup: {
    from: "all_bill",
```

```
    localField: "personName",
    foreignField: "nome",
    as: "bill_worth",
   },
  },
]
```

- **dataset sui GDP dei singoli Stati americani**

  Here the **true.json** was merged with **US_states_GDP** through *State* in true.json and *GeoName* in US_states_GDP. Afterwards, null values (ie non-US billionaires) were excluded by asking to select only billionaires having as Continent NA. This has been combined with the dataset **all_worth_trend_bill.json** through *personName* del true.json and *name* di all_worth_trend_bill.json.

  Finally, we obtained the following aggregate information: the US billionaire and his specifications from the true.json, the performance of his assets and the GDP performance of the US state in which he resides.

  Below is the code in Mongo db:

```
[
 {
  $lookup: {
   from: "US_states_GDP",
   localField: "state",
   foreignField: "Country Name",
   as: "state/region_GDP",
  },
 },
 {
  $match: {
   continent: "NA",
  },
 },
 {
  $lookup: {
   from: "all_bill",
   localField: "personName",
   foreignField: "nome",
   as: "bill_worth",
  },
 },
 {
  $lookup: {
   from: "world_GDP",
   localField: "residence_state",
```

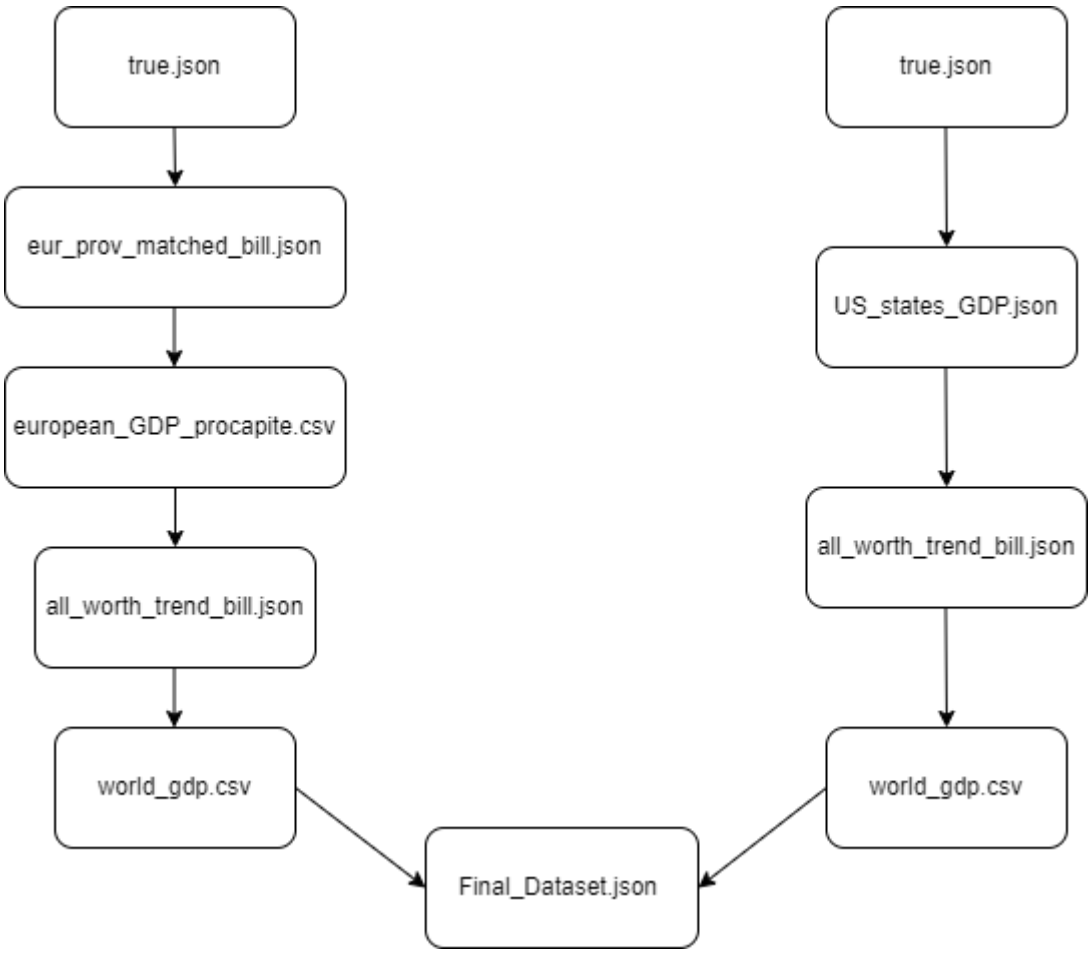
```
      foreignField: "Country Name",
      as: "state_GDP",
    },
  },
  {
    $project: {
      uri: 1,
      rank: 1,
      finalWorth: 1,
      personName: 1,
      city: 1,
      source: 1,
      industries: 1,
      countryOfCitizenship: 1,
      gender: 1,
      continent: 1,
      residence_state: 1,
      "state/region_GDP": 1,
      bill_worth: 1,
      state_GDP: 1,
    },
  },
```

To create both datasets, an aggregation pipeline has been produced, that is, a series of operations performed sequentially on the documents of a collection to obtain the desired results.
About the same functions that Mongo DB offers for aggregation have been used:

- $lookup: Join between two collections based on a common field.
- $match: Filters documents based on certain criteria.
- $project: projects the fields specified in the final result.

In both datasets aggregations are made starting from the true.json file and always joining with variables present in this file, except for a join made with the result of the join operation between the true json and the eur_prov_matchedbill.
For this purpose, the $unwind function has been added that allows you to "unroll" a field of an array document and generate a separate document for each element of the array, thus allowing aggregation with a variable not present in the original dataset.

Below is an explanatory diagram of the steps taken:



The data of the American pipeline have been linked with the data obtained from the pipeline of the Europeans, thus obtaining the Final_Dataset.json containing first all the data related to the American billionaires and then those Europeans.
Therefore, the Final_Dataset.json file will contain all the aggregated information for each individual billionaire: specifications present in the true.json (previously enriched with information on Country of residence, Continent and State/Region of residence)the GDP of the State/Region in which it resides, the development of its assets and the GDP of the country of residence.
For example, Alicia Walton will report, in addition to her specifications in true.json, data on the performance of the GDP of Texas (the US state in which she resides), the performance of her assets and the US GDP.
This information in the Final_dataset.json will appear as follows:


```
[{
 "_id": {
  "$oid": "646e26dd1f261df8170518ea"
 },
 "uri": "alice-walton",
 "rank": 19,
```

```
"finalWorth": 61730.879,
"personName": "Alice Walton",
"city": "Fort Worth",
"source": "Walmart",
"industries": [
 "Fashion & Retail"
],
"countryOfCitizenship": "United States",
"gender": "F",
"continent": "NA",
"residence_state": "United States",
"state/region_GDP": [
 {
  "2014": 51748.17,
  "2015": 54328.94,
  "2016": 54804.35,
  "2017": 56140.63,
  "2018": 58469.91,
  "2019": 60211.43,
  "2020": 59121.42,
  "2021": 61405.05,
  "_id": {
   "$oid": "647212d11f261df817052a03"
  },
  "Country Name": "Texas"
 }
],
"bill_worth": [
 {
  "2014": 34.3,
  "2015": 39.4,
  "2016": 32.3,
  "2017": 33.8,
  "2018": 46,
  "2019": 44.4,
  "2020": 54.4,
  "2021": 61.8,
  "_id": {
   "$oid": "646e1f271f261df817050540"
  },
  "nome": "Alice Walton"
 }
],
"state_GDP": [
```

```
   {
    "2014": 55123.84979,
    "2015": 56762.72945,
    "2016": 57866.74493,
    "2017": 59914.7778,
    "2018": 62805.25376,
    "2019": 65094.79943,
    "2020": 63027.67953,
    "2021": 69287.53659,
    "_id": {
     "$oid": "6470ce941f261df817052818"
    },
    "Country Name": "United States"
   }
  ]
},
```


Finally, it should be remembered that, among the European billionaires, there are also the Russians, who, as mentioned above, belong to the Moscow Region, which is considered European because it is part of the Russian Federation State, for which we have the GDP (whereas they are not among the European regions because they are considered only the regions belonging to the European Union, of which Moscow is not part).

5. Data quality

Data quality was carried out on all datasets imported into Mongo DB and on the final dataset; in particular, the following parameters were analysed:

- Accuracy
- Completeness
- Timeliness and Currency
- Consistency

5.1 Accuracy

Through this parameter it is possible to determine how much the reported data reflect the reality. Since this is economic data acquired by official platforms, most of our data respects the accuracy parameter, except in two datasets where changes have been made to the values present in the original datasets

Dataset	Length	Accuracy
European_GDP_procapite.csv	307	The currency transformation has been computed with the current EUR/USD change value (no change value for each year)
European_prov_matched_bill.json	493	63 Russian billionaires won't be matched with the European GDP dataset because in this dataset Russia was not take into account (here were considered UE Countries)
US_states_GDP.csv	52	The population dataset used for computing the per capita GDP is referred only to 2022

5.2 Completeness

Through this parameter we can observe how much the phenomenon is represented inside the dataset.

Dataset	Length	Missing values	% of missing values
world_GDP.csv	264	-	0%
US_states_GDP.csv	52	-	0%
European_GDP_procapite.csv	307	-	0%
true.json	2624	133	5%
European_bill_byProvince.json	493	-	0%
American_bill_byProvince.json	753	-	0%
All_worth_trend_bill.json	1089	59	5.4%
European_prov_matched_bill.json	493	180	36.5%
finalGDP_world_bill.json	1228	155	12,6%*

.* these are only missing valuesrelated to European Regions that we do not have the value of GDP, such as Russia, Switzerland, Monaco, etc..

In the **true.json** dataset 133 missing values emerged on python as follows:

```
### COUNTING HOW MANY BILLIONAIRES HAVE NO RESIDENCE CONTINENT

cont = 0

for el in category_list:
    try:
        if el['continent'] == '':
            cont += 1
    except KeyError:
        continue
print(f"{cont} null values of continent")
```

```
133 null values of continent
```

These null values are relative to the Continent residence of each billionaire, since among the information in the true.json there was not the city of residence of 133 billionaires, without which it was not possible to find neither the state nor the continent of residence.

In the dataset **All_worth_trend_bill.json** json note that the 59 missing values are given by the sum of:

- 17 of 441 european billionaires; therefore, the percentage of missing values in the performance of the patrimonies of European billionaires is about 3.8%..
- 42 of 648 american billionaires; therefore, the percentage of missing values in the asset performance of American billionaires is about 6.5%.

These values refer to those billionaires on the list, for whom, however, no asset performance is reported between 2014 and 2021.

These values emerged on python by entering a code that allowed us to count how many billionaires do not show a trend in the period under study:

```
### COUNTING HOW MANY BILLIONAIRES HAVE NO WORTH TREND IN THE PERIOD UNDER STUDY

cont_eu = 0
cont_us = 0
cont_null_eu = 0
cont_null_us = 0

for el in worth_trend:
    if el in american_bill_byProvince:
        cont_us += 1
        if len(worth_trend[el]) == 0:
            cont_null_us += 1
    if el in european_bill_byProvince:
        cont_eu += 1
        if len(worth_trend[el]) == 0:
            cont_null_eu += 1
print(f"{cont_eu} vs {cont_null_eu} : {cont_eu - cont_null_eu} european scraped")
print(f"{cont_us} vs {cont_null_us} : {cont_us - cont_null_us} us scraped")
```

```
441 vs 17 : 424 european scraped
648 vs 42 : 606 us scraped
```

In the dataset **European_prov_matched_bill.json**, the 180 missing values include:

- 63 russian billionaires: these values have been "lost" in the matching between the datasets *european_bill_byProvince.json* and *European_GDP_procapite.csv*, since in the former the

European continent is considered and then Moscow is taken into account; on the contrary, the second considers data referring to the European Union, of which Moscow is not part. These values are all imputed to the Moscow Province, which will be considered as belonging to the European continent in the world_GDP, even if they will appear only in the analysis concerning the States and not the Regions. In this way, to associate the value of Russia's GDP to the State of the Russian billionaires, it was necessary to replace the word "Russia" on the true.json in "Russian Federation" to make it compatible with the wording of world_GDP.

- 117 values are missing, some for one reason, others for another:
  - 98 missing values because the name of the province taken by Geonames is absent in the GDP dataset of the European Regions (European_GDP_procapite.csv); this is because the geographical area returned by Geonames is different from the European Region considered in the dataset on European GDP (e.g. Swiss cantons).
  - 19 missing values because the name of the Region taken by Geonames is different from the name of the Region in the dataset European_GDP_procapite.csv. For example, on Geonames, looking for "Valencia", we are given back the province "Valencia Valencia", while in the dataset European_GDP_procapite.csv it is reported as "Comunitat Valenciana". These values were imputed manually, reducing the percentage of missingu values to 32.6%.

In the **finalGDP_world_bill.json**: 155 missing values missing values concern the only European billionaires for whom GDP has been associated for the State but not for the Region of residence.

5.3 Temporal quality

The temporal quality is measured through two parameters: timeliness and currency. In particular, timeliness measures how data is up to date with respect to a particular moment, while currency measures how quickly the data is updated compared to the corresponding phenomenon in the real world.

Dataset	Temporal quality
European_GDP_procapite.csv	The most up-to-date value of GDP refers to 2021, not 2022
US_states_GDP.csv	The most up-to-date value of GDP refers to 2021, not 2022
All_worth_trend_bill.json	Some billionaires have no worth trend or partial trend in the period 2014/2021
finalGDP_world_bill.json	The most up-to-date value of GDP refers to 2021, not 2022. Some billionaires have no worth trend or partial trend in the period 2014/2021

5.4 Consistency

This parameter makes it possible to verify that the data represent all aspects of the reality of interest.

Dataset	Consistency
world_GDP.csv	Rounded values at 2 decimal numbers
US_states_GDP.csv	It wasn't per capita, so we compute this value by dividing it by population
European_GDP_procapite.csv	The values are in euro, we change the currency in USD, and from thousands to unitary values
true.json	For further enrichment we have changed the semantic value of these States of residence: Russia and Slovakia
European_bill_byProvince.json	The European provincies are not written in the same format of the European GDP dataset

6. Answer to reserach questions

From the final_dataset.json in which all our information was aggregated, it was possible to answer research questions.

1) Do countries with more billionaires have higher average GDP per capita?

To answer this question, it was first necessary to observe how many billionaires were resident in each country; The data have been arranged in descending order in order to make it clear at first glance which countries have more and fewer billionaires. From this it emerged that the billionaires of our interest reside in a total of 28 countries between the USA and 27 European states; in particular, it is clearly appreciable the great difference between the number of billionaires living in the United States and those who have residence in the second country by number of billionaires (Germany). Finally, the number of billionaires resident in a certain country was associated with the average value of GDP between 2014 and 2021 for that specific state.
Thanks to this approach it has been possible to clearly highlight that it is not necessarily true that in countries with a greater number of billionaires there is also a GDP per capita on average higher.

	N_Billionaire	Country_Name	Mean_GDP
0	751	United States	61235.421410
1	97	Germany	53131.493705
2	93	Russian Federation	27711.724122
3	64	Switzerland	70229.963014
4	56	Italy	41607.775795
5	31	France	45564.990925
6	24	Sweden	52924.849920
7	24	Spain	38466.799735
8	14	Monaco	190606.729464
9	10	Netherlands	56097.806949
10	9	Austria	54848.999427
11	8	Norway	66631.125173
12	7	Denmark	55940.659083
13	7	Finland	48207.960091
14	6	Ukraine	12150.101198
15	5	Poland	31339.643860
16	4	Ireland	80366.790925
17	3	Belgium	51506.217821
18	3	Romania	28148.195549
19	2	Cyprus	38364.745608
20	2	Slovak Republic	31125.482252
21	2	Greece	28769.391272
22	1	Albania	13190.426124
23	1	Portugal	33276.407383
24	1	Luxembourg	116373.078013
25	1	Liechtenstein	169148.765476
26	1	Andorra	41088.504733
27	1	Hungary	30906.060742

Some values have been highlighted in the table: in red those values of the GDP of the States higher than the US GDP but with a lower concentration of billionaires; in blue those low GDP but which do not refer to countries with few billionaires.
In some cases the opposite is observed: for example, the Russian Federation has 93 billionaires and is in third place in the ranking in question, but the average GDP associated with this country is one of the lowest on this list. The above also suggests that wealth is likely to be concentrated in the hands of a few super rich people in Russia. Conversely, countries such as Liechtenstein and Luxemburg, in which only a billionaire appears to reside, are among the countries with the highest average GDP per capita; this suggests that wealth would therefore appear to be equally distributed among the population.

2) Who are the billionaires with the highest correlation between their worth trend and their Region/State GDP per capita trend?

The correlation between the development of the billionaire's assets and the GDP of the European Region/US State in which he resides was calculated with the Pearson index. The values thus obtained were ordered from the highest value of the index to the lowest value.

In this way it has been possible to highlight the States/Regions with the highest value of the index, as it has been evidenced from the output here of continuation:

```
           Billionaire  Pearson correaltion
0            Daniel Ek             0.997998
1        Bert Beveridge             0.988241
2     Herbert Wertheim              0.982886
3           Jeff Bezos             0.981290
4    Michael Bloomberg             0.981247
5         Arturo Moreno             0.979405
6          Laurie Tisch             0.977823
7            Mark Cuban             0.975748
8       Robert F. Smith             0.975006
9  Reinhold Schmieding             0.973664
                  Billionaire  Pearson correaltion
776             Michael Platt            -0.812828
777            Martin Haefner            -0.818783
778           Forrest Preston            -0.819838
779           Ronald Perelman            -0.838403
780  Jaroslav Hascak & family            -0.855138
781              Bernd Freier            -0.863694
782            Ivar Tollefsen            -0.867990
783         Gabriella Meister            -0.881738
784           Heikki Kyostila            -0.887301
785              James France            -0.904857
The most influential billionaire is Daniel Ek, with a correlation value of his worth and the General Domestic Product of his co
untry (Düsseldorf) of 0.997998.
This correlation coefficient is based on the following values:
1) General Domestic Product of Düsseldorf: {'2019': 55856.68, '2020': 55208.44, '2021': 59746.12}.
2) Daniel Ek's worth trend: {'2019': 2.2, '2020': 2, '2021': 4.6} (each values is expres in billions of dollars).
```

This means that: the first 10 billionaires show a trend of assets similar to the GDP of the State/Region in which they reside; the last 10, show one instead of growing assets, the GDP of the State/Region of residence decreases.
At the end of the output are the values related to the evolution of the assets and the GDP of the State/ Region of residence of the billionaire who has the highest Paerson index; Furthermore, it should be noted that only such a result has been reported as an example.
Looking at the figures for Daniel Ek, we can see that the data refer only to the years 2019, 2020, 2021. This happens because we only have the trend of the assets in these three years, probably because previously the subject had not yet climbed the ranking of Forbes. Consequently, in order to calculate the correlation, the GDP values of the Region of residence were observed only for those years.

3) What is the percentage of women billionaires in the top 10 richest (with the highest average GDP per capita) countries?

Initially, countries were ranked from the richest to the richest in terms of GDP per capita. Only the 10 richest countries have been selected and the number of resident billionaires has been compared, and in particular, women have been extracted, thus constituting the column of the number of billionaires in each country considered. At this point, the percentage of women in the total of billionaires was calculated.

	State	Average GDP per capita	Women percentage	N women billionaire	N billionaire
0	Monaco	190606.73	21.43	3	14
1	Liechtenstein	169148.77	0.00	0	1
2	Luxembourg	116373.08	100.00	1	1
3	Ireland	80366.79	0.00	0	4
4	Switzerland	70229.96	17.19	11	64
5	Norway	66631.13	25.00	2	8
6	United States	61235.42	13.18	99	751
7	Netherlands	56097.81	0.00	0	10
8	Denmark	55940.66	42.86	3	7
9	Austria	54849.00	11.11	1	9

In any case, the percentage of billionaire women is considerably lower than the percentage of billionaire men; of course, this does not apply to Luxembourg, where only a billionaire resident, in this case a woman (the same thing but also applies to Liechtenstein).
Highlighted in blue you can see the country with a higher percentage of billionaire women: the highest figure is reported by Denmark, followed by Norway and the Principality of Monaco, which is also the country on average richer.
In red, it was considered interesting to highlight how the US, which has the highest number of resident billionaires, is one of the countries with the lowest percentage of billionaire women in this analysis.
To conclude, this research question leads us to think that even in the world of the super rich there are still discriminating factors that lead women to never reach values, in a broad sense, equal or superior to those of men.