Billionaires Statistics Dataset (2023) - Data Curation Project

Nency Patel , Darshan Patel, Ajay Babu Popuri
Group-26

---

1. Introduction

Data Curation is the process of maintaining data to ensure that it is accurate and useful for its intended purpose. Data curation works with various steps to ensure that data is properly maintained for use and reuse.

2. Understanding Dataset

Dataset is downloaded from Kaggle. The world's billionaires are profiled in this dataset, which also includes personal and business-related information. It offers information on the global distribution of wealth, industry sectors, and characteristics of billionaires.

2.1 Potential UseCase of Dataset

- Wealth distribution analysis: Explore the distribution of billionaires' wealth across different industries, countries, and regions.
- Demographic analysis: Investigate the age, gender, and birthplace demographics of billionaires.
- Self-made vs. inherited wealth: Analyze the proportion of self-made billionaires and those who inherited their wealth.
- Economic indicators: Study correlations between billionaire wealth and economic indicators such as GDP, CPI, and tax rates.
- Geospatial analysis: Visualize the geographical distribution of billionaires and their wealth on a map.
- Trends over time: Track changes in billionaire demographics and wealth over the years.

2.2 Brief description of Dataset

This dataset contains statistics on the world's billionaires, including information about their businesses, industries, and personal details. It provides insights into the wealth distribution, business sectors, and demographics of billionaires worldwide.

ATTRIBUTE NAME	DATATYPE	DESCRIPTION
RANK	NUMBER	The ranking of the billionaire in terms of wealth.

FINALWORTH	NUMBER	The final net worth of the billionaire in U.S. dollars.
CATEGORY	STRING	The category or industry in which the billionaire's business operates.
PERSONNAME	STRING	The full name of the billionaire.
AGE	NUMBER	The age of the billionaire.
COUNTRY	STRING	The country in which the billionaire resides.
CITY	STRING	The city in which the billionaire resides.
SOURCE	STRING	The source of the billionaire's wealth.
INDUSTRIES	STRING	The industries associated with the billionaire's business interests.
COUNTRYOFCITIZENSHIP	STRING	The country of citizenship of the billionaire.
ORGANIZATION	STRING	The name of the organization or company associated with the billionaire.
SELFMADE	STRING	Indicates whether the billionaire is self-made (True/False).
STATUS	STRING	: "D" represents self-made billionaires (Founders/Entrepreneurs) and "U" indicates inherited or unearned wealth.
GENDER	STRING	The gender of the billionaire.
BIRTHDATE	NUMBER	The birthdate of the billionaire.
LASTNAME	STRING	The last name of the billionaire.
FIRSTNAME	STRING	The first name of the billionaire.
TITLE	STRING	The title or honorific of the billionaire.
DATE	NUMBER	The date of data collection.
STATE	STRING	The state in which the billionaire resides.
RESIDENCESTATEREGION	STRING	The region or state of residence of the billionaire.
BIRTHYEAR	NUMBER	The birth year of the billionaire.
BIRTHMONTH	NUMBER	The birth month of the billionaire.
BIRTHDAY	NUMBER	The birth day of the billionaire.
CPI_COUNTRY	FLOAT	Consumer Price Index (CPI) for the billionaire's country.

CPI_CHANGE_COUNTRY	FLOAT	CPI change for the billionaire's country.
GDP_COUNTRY	NUMBER	Gross Domestic Product (GDP) for the billionaire's country.
GROSS_TERTIARY_EDUCATION_ENROLLMENT	FLOAT	Enrollment in tertiary education in the billionaire's country.
GROSS_PRIMARY_EDUCATION_ENROLLMENT_COUNTRY	FLOAT	Enrollment in primary education in the billionaire's country.
LIFE_EXPECTANCY_COUNTRY	FLOAT	Life expectancy in the billionaire's country.
TAX_REVENUE_COUNTRY_COUNTRY	FLOAT	Tax revenue in the billionaire's country.
TOTAL_TAX_RATE_COUNTRY	FLOAT	Total tax rate in the billionaire's country.
POPULATION_COUNTRY	NUMBER	Population of the billionaire's country.
LATITUDE_COUNTRY	FLOAT	Latitude coordinate of the billionaire's country.
LONGITUDE_COUNTRY	FLOAT	Longitude coordinate of the billionaire's country.

3. Identify Issue

- Using some simple Python code, identify the various data types your data includes as the first step in getting to know it.
- Pandas library from Python is useful tool where We will be able to extract insights from a big dataset by breaking it up into smaller, more manageable chunks with the help of these tools.
- To identify inconsistent, redundant, and missing data, data visualization has been done on the dataset via different charts and tables.

4. Challenges

The major challenges or potential issues faced in the provided code:

Handling Geolocation Operations: The code uses the geopy library to fetch missing country values based on latitude and longitude. However, this can be resource-intensive or may face issues due to API limitations, network availability, or rate limits from the geolocation service provider.
It might lead to slower execution or potential failures due to internet connectivity issues or if the service doesn't respond within a certain timeframe.

Data Imputation based on Grouping: The imputation strategy based on the mode or mean within each country/state grouping might not always be the best approach, especially if the dataset lacks diversity within certain groups. This can lead to biased imputations.

Handling Missing Data: The code uses several methods like mode, mean, and fillna() for imputing missing values. However, the appropriateness of these methods depends on the nature of the data. It's essential to understand the data distribution before choosing an imputation strategy.

Assumptions and Hard-Coded Values: There are assumptions made regarding using the mode or mean for imputation when certain groups might not have a mode or have an unreliable mode. This can introduce bias or inaccuracies in the dataset.

Potential Data Integrity Issues: The code doesn't perform data validation or error handling explicitly, which might lead to issues if the fetched geolocation data is incorrect or if there are discrepancies between the location data and the provided country/state information.

Data Conversion and Cleaning: The code performs operations like converting object data types to float and string replacements for columns like 'gdp_country'. However, it assumes consistent formatting and may break if the data structure deviates from the assumed format.

These challenges highlight potential areas for improvement and caution in the code implementation for handling missing values, geolocation-based data fetching, and data cleaning processes.

5. Problems

5.1 Null Values in dataset

ATTRIBUTE NAME	NULL VALUES
RANK	0
FINALWORTH	0
CATEGORY	0
PERSONNAME	0
AGE	65
COUNTRY	38
CITY	72
SOURCE	0
INDUSTRIES	0
COUNTRYOFCITIZENSHIP	0
ORGANIZATION	2315
SELFMADE	0
STATUS	0
GENDER	0
BIRTHDATE	76
LASTNAME	0
FIRSTNAME	3

TITLE	2301
DATE	0
STATE	1887
RESIDENCESTATEREGION	1893
BIRTHYEAR	76
BIRTHMONTH	76
BIRTHDAY	76
CPI_COUNTRY	184
CPI_CHANGE_COUNTRY	184
GDP_COUNTRY	164
GROSS_TERTIARY_EDUCATION_ENROLLMENT	182
GROSS_PRIMARY_EDUCATION_ENROLLMENT_COUNTRY	181
LIFE_EXPECTANCY_COUNTRY	182
TAX_REVENUE_COUNTRY_COUNTRY	183
TOTAL_TAX_RATE_COUNTRY	182
POPULATION_COUNTRY	164
LATITUDE_COUNTRY	164
LONGITUDE_COUNTRY	164

5.2 Remove Unnecessary data in column

In the BIRTHDATE Column , the formate is ‘MM/DD/YYYY 0:00’ where time string is the same 0:00 for whole dataset which is not accurate. Remove Unnecessary time from the BIRTHDATE column.

5.3 Column Datatype are irrelevant

GDP_COUNTRY column has object data type because the values start with $ so, will convert it to float

5.4 Redundant Data

Some column has other column with same meaning but name is different.

- BIRTHDATE column has date while another columns BIRTHDAY , BIRTHYEAR and BIRTHMONTH are splitting columns from birthdate column.
- PERSONNAME column has enough to satisfy FIRSTNAME and LASTNAME column’s value. Without losing any information remove these two column.

5.5 Get missing location in column from longitude and latitude.

- COUNTRY column have missing values which can be derived from LOGTITUDE_COUNTRY and LATITUDE_COUNTRY column using Geopy.

- Imputing values of other columns using the specific country as reference, as these values remain same for that specific country.

6. Tools and Methods

- Vizier environment to perform all the necessary operations.
- Geopy is a Python client for several popular geocoding web services.
- Pandas is an open-source Python library that provides data structures and data analysis tools for working with structured data, primarily tabular data.
- Numpy is an open source Python library consisting of multidimensional array objects and a collection of routines for processing those arrays.
- Matplotlib, Seaborn and Plotply for data tranformation.

7. Data cleaning

7.1 Eliminate the redundant column

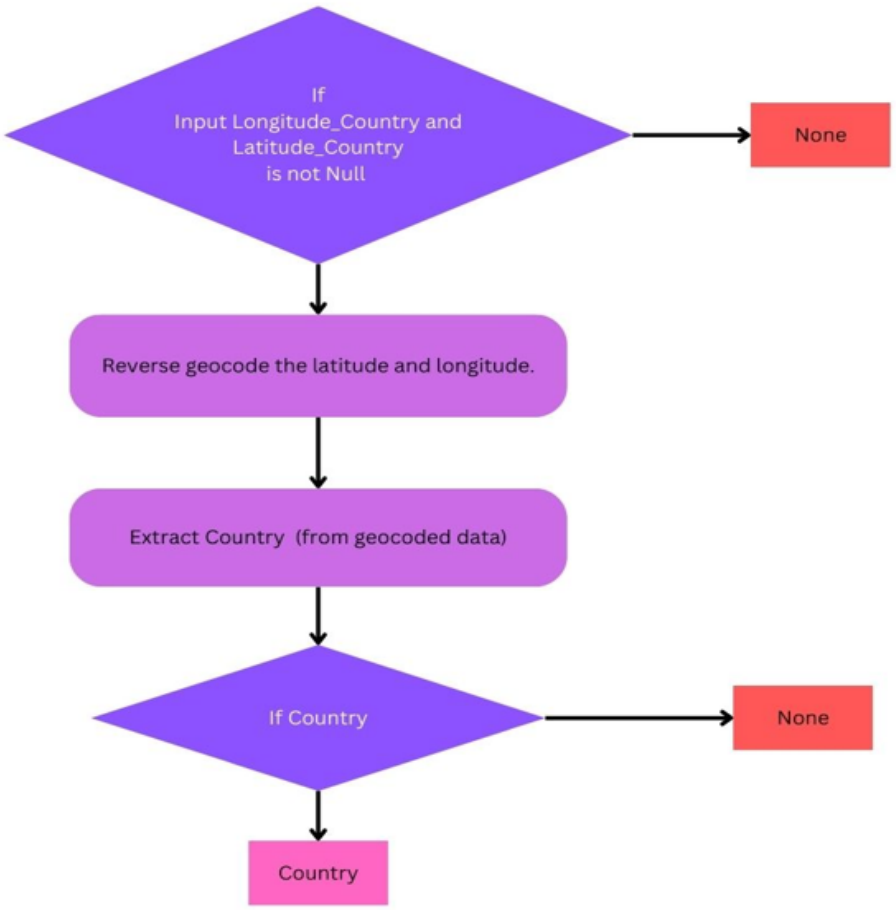
Removing redundant column from dataset such as BIRTHYEAR, BIRTHDAY, BIRTHMONTH , LASTNAME and FIRSTNAME.

7.2 Remove Unnecessary data in column

Removing time string from BirthDate column and date column using pandas from python.

7.3 Missing Location retrieve by Geopy

Using Geopy library , Getting missing Country from longitude_country and latitude_country.
Below image shows that how geopy library work for getting country from the inputs of longitude and latitude of country. If both latitude and longitude are not null, it initializes a geolocator using the Nominatim service and performs reverse geocoding to get an address based on the given latitude and longitude.It extracts the 'address' field from the geocoded location and then extracts the 'country' field from the address. The 'country' information is assigned to the variable country. If 'country' is not available in the address, it assigns None to Country.



For state missing value using geopy steps would like this:

Define a function (get_state_from_coordinates) that takes latitude and longitude as input and uses Geopy to retrieve the state information.
Load and Handle Missing Latitude and Longitude:Load your dataset using pandas. Assuming the dataset has columns named 'latitude' and 'longitude', handle missing values using linear interpolation.
Iterate through the rows of the dataframe where the 'state' column has missing values.
Call the get_state_from_coordinates function to retrieve the state information based on the available latitude and longitude.
Update the 'state' column in the dataframe with the retrieved state information.

7.4 Null Values

For missing age add mean value of age.
Imputing missing values in the 'city' column based on the 'country' column using the mode of the city within each specific country means filling in missing 'city' values by using the most frequently occurring city within the same 'country

-: Identify rows where the 'city' column has missing values.
-: Group the data by the 'country' column.
-: For each group (each country), find the mode (most frequently occurring value) of the 'city' column.
-: Replace missing 'city' values in that group with the mode of the 'city' column for that specific country.
-: Repeat for All Groups:
    -Repeat this process for each unique country in the dataset.

Do same step as state for residenceStateRegion finding missing values as well as for gross_tertiary_education_enrollment, gross_primary_education_enrollment_country , life_expectancy_country,tax_revenue_country_country,total_tax_rate_country,population_country,latitude_country,longitude_country,cpi_country,cpi_change_country.

Birthdate, Title will remain as it is because the values cannot be imputed.

7.5 Change Datatype

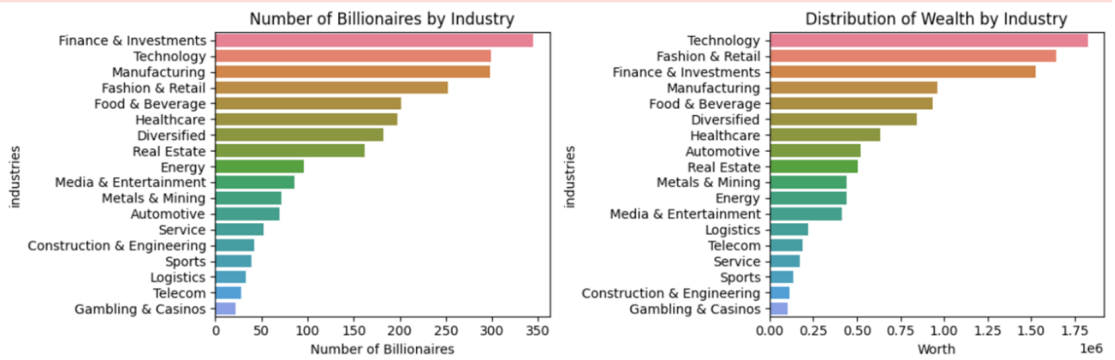
Gdp_country has $ sign and it become object datatype so remove it and convert it into float that can be used into data extraction in future.

Birthdate has Object datatype convert it to Date formate.

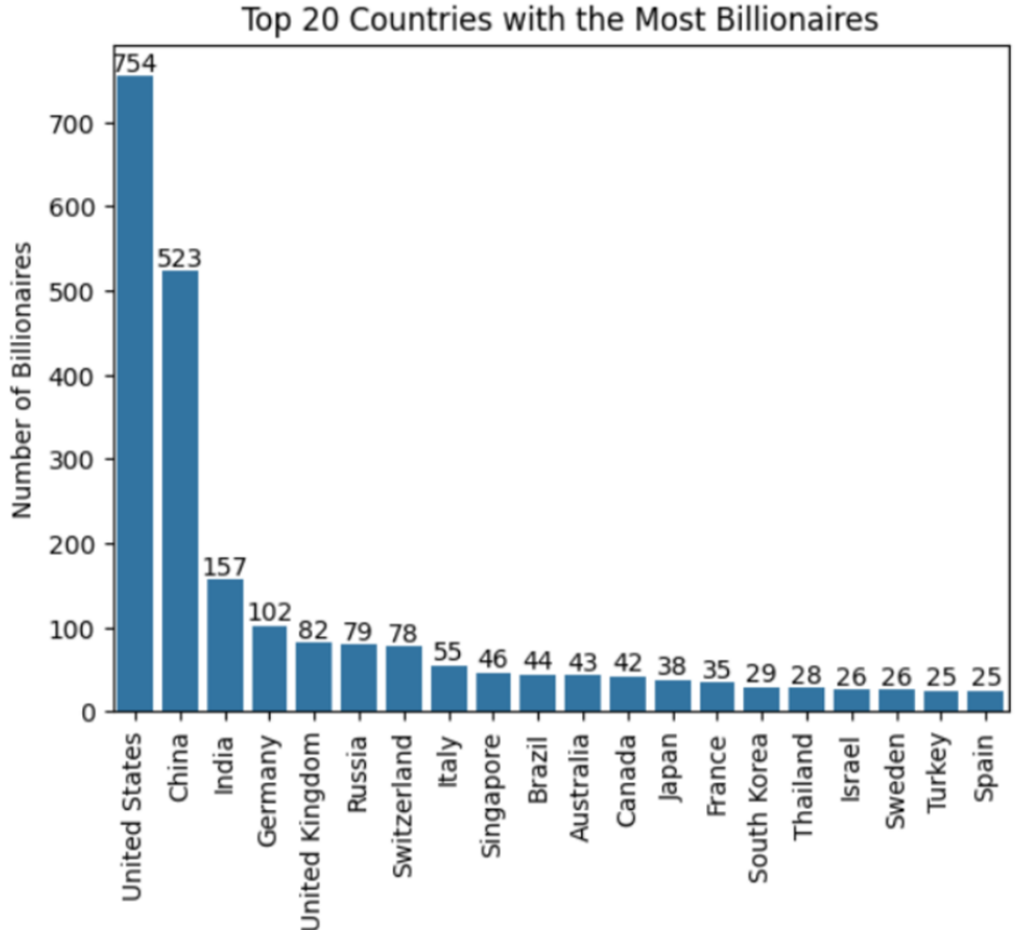
8. Data Transformation

Data transformation using plots involves visualizing data in different ways to gain insights, identify patterns, and make informed decisions. Plots play a crucial role in transforming raw data into meaningful visual representations.

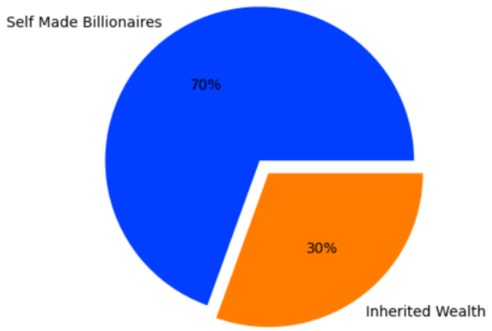
Here we have plotted graphs for number of billionaires by Industry and distribution of distribution of wealth by industry.From the first graph, most number of billionaires came from the Finance and Investments industry where as it ranks 3rd in industry with most wealth. Technology industry stands top with the most wealth. We can also observe that the top five industries with the most number of billionaires are also in the top five industries with the most wealth, in no particular order.

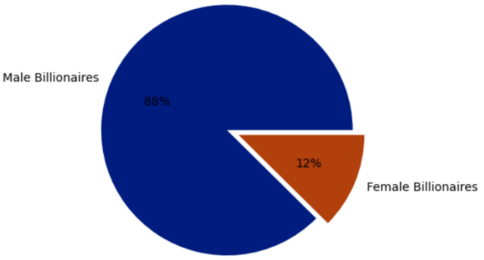


In this we plotted a graph graph for top 20 countries holding billionaires. United states, China and India are in top 3 positions of the list. And also we can observe that there is a huge difference between these positions like from 1st to 2nd it is around 230 and from 2nd to 3rd, it is around 360.

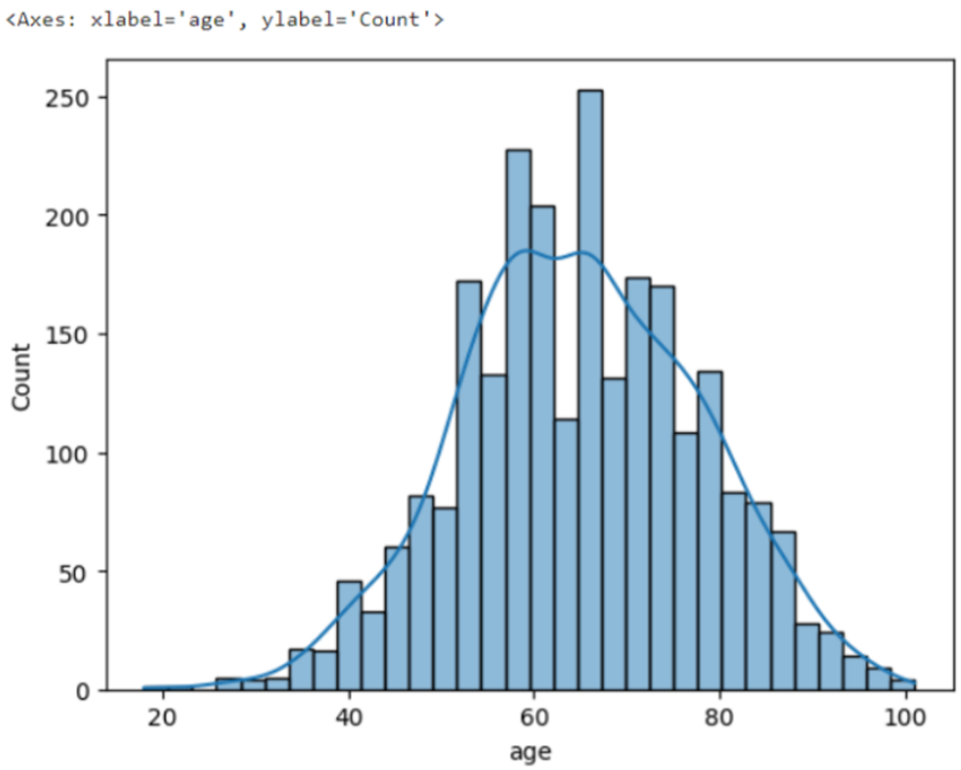


This pie chart indicates that the self made billionaires are with a huge percentage of 70% whereas it 30% who inherited the wealth.

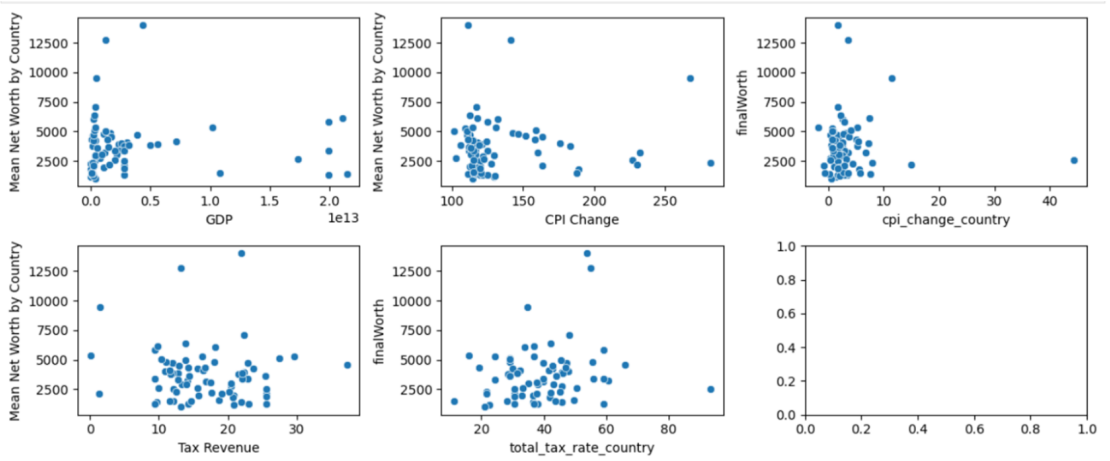




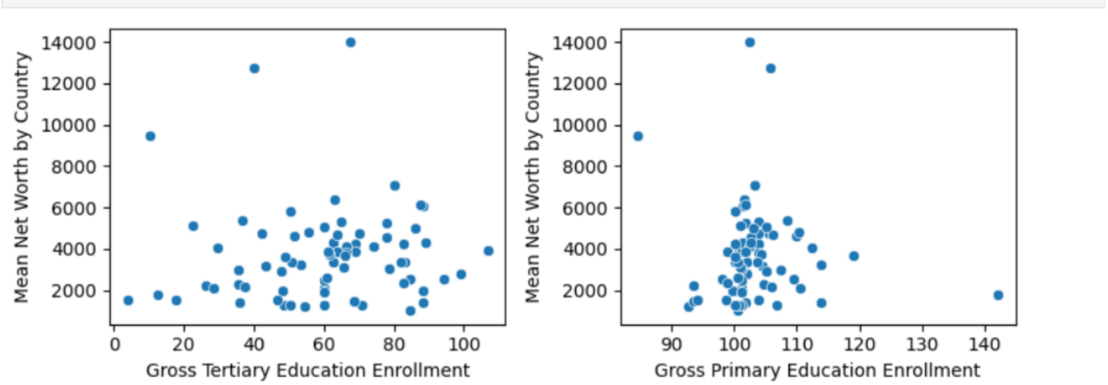
Similarly if we look into the another one, we can clearly see that the percentage of male is dominant over the female.



If we have look at this graph, majority of the billionaires age lies between 50 to 70. and also the count of the young billionaires is very low.
Same here, we plot age against each industry in total wealth vise.



Most number of billionaires came from countries with low GDP. Billionaires who came from countries with GDP tend to be outliers.
Billionaires with the highest net worth came from countries with low GDP.
Most billionaires also came from countries with low Consumer Price Index (CPI).
There are few billionaires who came from countries with high CPI, these billionaires have lower mean Net Worth.
Billionaires with the highest net worth came from countries with low CPI.
A considerable number of billionaires are from countries with tax revenues ranging from 10 to 25 USD. The billionaires with the highest net worth also came from this tax revenue range.
There are also billionaires who came from countries with very low tax revenue.
Most billionaires came from countries with tax rate of 20% to 50%.



The Gross Tertiary Education Enrollment does not reveal a strong correlation to a billionaire's net worth, as shown in the plot where points are scattered on different values.
On the contrary, most billionaire's came from countries with Gross Primary Education Enrollment of 100% or more.

9. Conclusion and Future Work

In summary, After cleaning data improved the quality and our use of visualizations effectively communicated the success of implemented solutions, showcasing concrete examples of improved

data quality. Emphasizing enhanced data reliability, we demonstrated positive outcomes from our curation efforts, instilling confidence in decision-making.

References -

https://geopy.readthedocs.io/en/stable