**Data Source**

The “Billionaires Statistics Dataset (2023)” was acquired from Kaggle.com.

The owner and complier of the data set is Nidula Elgiriyewithana, a data scientist from Sri Lanka and a Kaggle Progression System Grand Master. The data was compiled on April 4, 2023, and uploaded to Kaggle in September 2023.

I have chosen this data set because it meets all the requirements as stated in the project brief and shows promise for producing an interesting and enlightening project.

**Data Profile**

**Collection:**

The dataset was compiled from various sources, including Forbes and other financial publications. No further details on the sources was provided, other than this link to country tax rates: <https://idea.usaid.gov/cd/argentina/domestic-revenue> mobilization#tab-revenue- composition (link to Argentina specifically).

The contents of this data set include details on each of the 2,640 billionaires listed as of April 4, 2023. There are fields for basic demographics of each person and data on their worth in US dollars, the industries in which they made their wealth, and whether or not they are self-made. The data set also includes information on the countries in which the people live, such as the life expectancy, consumer price index, GDP, and education enrolment rates for their country. After cleaning the data frame includes 2,639 rows and 27 columns.

**Data Understanding:**

• I used df.describe() to view basic descriptive statistics for the data set. The max tax rate is 106% - this is the tax rate assigned to Argentina. It is explained on the Kaggle discussion for this data set as follows: (source)

https://idea.usaid.gov/cd/argentina/domestic-revenue-mobilization.

“Total tax rate measures the amount of taxes and mandatory contributions payable by businesses after accounting for allowable deductions and exemptions as a share of commercial profits. Taxes withheld (such as personal income tax) or collected and remitted to tax authorities (such as value added taxes, sales taxes or goods and service taxes) are excluded.” This explains how a tax rate can be over 100%

**Limitations:**

There are various limitations to this data set. First, there is no way to verify most of the data sources since limited information was provided by the owner. It appears that the data was collected manually, in which case, there could be transcription errors. Lastly, some of the columns lack units. Three columns in particular are troublesome: status, gross\_tertiary\_education\_enrollment, and gross\_primary\_education\_enrollment\_country. The “status” field contains alphabetic values that are not defined. And the two columns on education enrolment do not contain units and do not appear to be percentages.

**Ethical Considerations:**

There may be billionaires who are not represented in this data set and so there could be a sampling bias. There could also be a reporting bias because this data may be self-reported or collected from public sources, leading to inaccuracies. Additionally, individuals may inflate their successes rather than failures, leading to an overrepresentation of wealth depending on what metrics they include. Also, measurement bias could exist in the metrics used to measure wealth. Things like net worth may not capture the full extent of a billionaire's wealth, leading to underestimations or overestimations. Not knowing the criteria for inclusion in this dataset means we can’t know what metrics were used other than just “final worth.” This is also just a snapshot in time. Year to year, there could be a turnover in the billionaires listed, changing the analysis results. If this data were available year after year, a time-series analysis could be useful.

**Questions to Explore**

➢ Where are most billionaires located?

➢ Is there a country or city with the most?

➢ Do billionaires live longer than the life expectancy for their

country?

➢ Which country has less billionaires ?

➢ Is there a correlation between the primary or tertiary education

enrolment in a country and the number of billionaires in that

country?

➢ Is there a correlation between the gross domestic product of a

country and the number of billionaires in that country?

➢ Is there a correlation between the population of a country and

the number of billionaires in that country?

➢ Are more male or female billionaires self-made vs billionaires by

inheritance?

➢ Are male or female billionaires concentrated in any country or

region of the globe?

➢ Is there is any correlation between revenue generated by the Tax

in the Country and number of billionaires in the country?