Billionaires Dataset Project Proposal

- Jonathan Ibarra and Boone Losche

Why we chose it:

For this project, we decided to go with a dataset of billionaires which includes many features such as their country's tax rate, their age, the industry they are in, their education level, etc. We chose this project because we were deeply interested in finding any correlations between an individual's circumstances/attributes and their immense wealth. The end ideal goal is to create a model that can fairly accurately predict one's accumulated wealth (given they are a billionaire.) Going into this project we are aware that there are hundreds of factors that can influence an individual's wealth, not all of which are captured by this dataset. However, given this dataset's focus on the billionaire's country's conditions, such as tax rate, CPI, life expectancy, GDP, etc, we are interested in seeing the relationships that exist between the features and one's ability to accumulate unthinkable amounts of wealth.

Our Approach:

To start we will be cleaning up the data and pruning/editing some features. The Residence_State and Residence_Region features have a majority of NULL entries, this could very well skew the results thus requiring us to remove them. We also will not be entering personal information such as Full_Name given the fact that someone's name is not a reliable tool to predict one's wealth.

We will then begin testing to figure out the best algorithms and input variables to maximizing the accuracy of predicting. Luckily with over two thousand observations in the dataset and thus have lots of data to train and test on. One big step we will work out will be how to handle a blend of quantitative and categorical features within our dataset. With preliminary research we have seen three possible approaches to begin our search with:

- Linear regression (some categorical variables though may throw this off)
- Random Forest
- Neural Net

Finally, once we decide on the best algorithm and input values we will start the jupyter notebook and make sure to have plenty of tables and graphs to show our project off and compare it against a baseline.

Steps:

1. We gather our information necessary for the project (e.g. model, data, and understandings of content) **Week 1-2**
2. We begin to write the jupyter notebook **Week 3-4**
    a. A clear introduction to the topic we worked on, including a detailed description of the dataset
    b. A clear description of what we did and how we did it, with results that are clearly presented in tables and graphs
    c. Discussion of our results and what we learned from our exploration
    d. Finally, polish our notebook to be well organized and well written.
3. Submit.

Team Responsibilities:

Our team consists of two people. We have previously worked together for a semester-long project in CS 414 and have developed a very solid system for fairly and evenly distributing work. Our main goal as a team is to learn as much as possible; With this in mind, we aim to both be engaged in every step of the process together. We will try and augment our learnings from SCRUM and use them to create weekly check-ins where we discuss what we have worked on and ask for assistance with any roadblocks. We both will be responsible for large decisions but will split up the larger goals into smaller pieces we can work on independently till our next conversation.

For the first part of the project we will split up the research and testing phase, Boone will take cleaning the data and Random Forest, and Jonathan will take Linear Regression and Neural Net. From there we will begin the bulk of the project once the algorithm is decided. We will fairly split up the writing and coding, making sure we both get enough experience to maximize our respective experience on the subjects. Overall we are very excited to break into this project, especially having the chance to learn about new algorithms like Random Forest and Neural Nets.

The Dataset:

https://www.kaggle.com/datasets/javiersab/billionaires-dataset-cleaned

Features: position, wealth, industry, full_name, age, country_of_residence, city_of_residence, source,citizenship, gender, birth_date,last_name, first_name,residence_state, residence_region, birth_year,birth_month, birth_day, cpi_country, cpi_change_country, gdp_country, g_tertiary_ed_enroll, g_primary_ed_enroll, life_expectancy, tax_revenue, tax_rate, country_pop, country_lat, country_long, Wcontinent