**DSC 540 Project Milestone 1**

**Logan Quandt**

**Data Sources:**

Flat File:

<https://www.kaggle.com/datasets/zgrcemta/world-gdpgdp-gdp-per-capita-and-annual-growths>

This link contains multiple CSV files with information pertaining to World economic data gathered by World Bank (GDP, GDP Growth, GDP per capita, GDP per Capita growth, GDP PPP and GDP PPP growth) broken down by country for the years 1960-2020.

Website with table:

[Population by Country (2022) - Worldometer (worldometers.info)](https://www.worldometers.info/world-population/population-by-country/)

This website contains a table for the world population down by country. Beyond providing the total population the website provides fields such as yearly change, net change, fertility rate, migrants and more.

API:

<https://datahelp.imf.org/knowledgebase/articles/1968408-how-to-use-the-api-python-and-r>

This API connects to the International Monetary Funds data allowing me to pull macro and micro economic indicators by country, region or set of countries from multiple different databases. An example may be pulling in net import data for the United Kingdom.

**Relationships:**

All three data sources are connected by the common theme of being broken down into country level identifiers such as country name or country code which can be used as an ID after some cleaning.

CSV Files – Country Name and Country Code

Website – Country (or Dependency)

API – REF\_AREA (Ex: GB = United Kingdom)

**Action Plan:**

To meet the five milestones, the first step of my action plan is to merge the five different datasets found on the Kaggle link. I will join them all based off the country name, and this will expand the amount of data readily available in a single dataset. I will also go through the dataset and look for and handle missing or null values. For example, some countries do not have any data on their GDP in these CSV files so I can likely remove those rows. There may also be outliers in the data such as very small countries which may not be significant to my analysis.

Another challenge I will have to overcome is I will have to format the different country level identifiers to make them easily joinable as the ID specifically the API which has a ref area code which may differ from the country code. I also will likely have to clean the country names as abbreviations or alternate spellings may cause duplicates which I will have to remove. After solving for these issues, I can replace the headers to simplify the coding process and make what we are looking at clearer.

My belief is that this data will give me insights into country level economic indicators and give me a better understanding of how the world economy is both growing and perhaps shifting in terms of economic power. If the 2020 data is accurate it can perhaps give insights into the impact of COVID-19 on world GDP. Based on these datasets, someone could further build on it to research multiple different things such as trends that may cause GDP growth. As all the data is open to the public and does not contain any personal information, I do not believe there will be any ethical considerations that I will encounter.