**Work\_Log**

**Progress Update - Data Analysis for First Question**

**Work Completed (as till 5:00 PM, 11-02-25): by Shriniwas**

* **Documentation & Structuring:**
  + Appended the questions for both sub-questions (Basic Analysis and Machine Learning) on Google Docs.
  + Reviewed and structured the file system to ensure clarity on where all components are stored and where teammates should append their work.
* **Dataset Preparation:**
  + Uploaded the required datasets in the **Raw Dataset** folder under **Dataset**.
  + Cleaned four datasets for basic analysis and saved them in the **Cleaned Dataset** folder:
    - *GDP*
    - *Affected by Disaster vs. GDP*
    - *Full Disaster Dataset*
    - *Income Group*
* **Code & Data Merging:**
  + All Python code for data cleaning is available in:  
    **notebooks/Data\_cleaning/Q1\_data\_cleaning.ipynb**
  + Merged the datasets **Affected by Disaster vs. GDP** and **Full Disaster Dataset** using the primary key **ISO Code** and **Year**.
  + The merging was necessary due to missing values in the *Affected by Disaster vs. GDP* dataset. The missing values will be derived from the merged dataset.

**Next Steps (Remaining Work):**

1. **Handling Missing Values:** Identify dependent attributes to fill in missing values in the merged dataset.
2. **GDP Dataset Transformation:** The GDP dataset has **yearly data in a horizontal format**, whereas other datasets have it in a **vertical format**. Convert the GDP dataset into a vertical format.
3. **Final Merging:** After fixing the GDP dataset structure, merge it with the current merged dataset (**Affected by Disaster vs. GDP + Full Disaster Dataset**).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Country | ISO | Year | Income Classification | Total Affected | Total Deaths | No. Injured | GDP per capita (PPP) | Total Damage ('000 US$) | Magnitude | Region | Affected per 100,000 |
| Afghanistan | AFG | 1987 | Low-income | 457.38 | 252 | 1182 | 2,160.86 | 15,000,000 | 7.5 | Developing | 457.38 |
| Honduras | HND | 1990 | Low-income | 1000 | 100 | 500 | 3,271.40 | 10,000,000 | 7.2 | Developing | 500 |
| Cote d'Ivoire | CIV | 2000 | Developing | 1200 | 50 | 200 | 4,085.20 | 8,000,000 | 6.8 | Developing | 600 |

**Ideal Final Dataset Structure:**

**Analysis to be Conducted After Dataset Completion:**

1. **Summary Statistics:** Calculate mean, median, and standard deviation of GDP changes for each income classification.
2. **Correlation Analysis:** Identify relationships between *Total Affected, Magnitude of Disasters,* and *GDP per capita*.
3. **T-tests/ANOVA:** Compare GDP losses between developed and developing countries to check for significant differences.
4. **Regression Analysis:** Predict how disaster severity (e.g., *Magnitude, Total Affected*) impacts GDP decline.

Work Completed as on 16/2/25

1. Merged the disaster, GDP, income classification of each country, Co2 emissions, Funds allocated and temperature in a single dataset ( E-Coders\_SAS\_Curiosity\_Cup\datasets\merging\_datasets\final\_merged\_dataset.csv" )
2. Problem Faced : During the dataset merge operation, a **Cartesian product issue** arose due to a **many-to-many relationship**, leading to unintended data duplication and an inflated dataset. To address this, we **aggregated** the total funds allocated to each country for a given year using a **groupby operation with summation**. This ensured that the dataset contained **unique country-year combinations**, preventing redundant records and maintaining data integrity before performing the merge.

Disaster dataset :

classified the events according to disaster category, and identified disaster categories that have a direct relation with climate change.

Now, disaster start month, day and year was given along with end year, month and day, which was used to calculate duration of disaster.

( problem faced in this : start month, day and year was given in the dataset but there was missing values in start Month and start day, which was imputed to have 1 value i.e 1 in start month and 1 in start day, so if month and day are missing it because 1st of Jan, If End day was missing, it was imputed with value as ([start day]+1)mod30 and if end month was missing, it was imputer with ([start month]+1)mod12. Now these columns didn’t have any missing values. Now We convert these columns to start\_date and end\_date where the values are concatenated and stored as date objects )

Now, GDP and Disaster database was merged according to Year and Country, new columns added were GDP\_start\_year and GDP\_end\_year and saved and appended to disaster dataset

Now, a database with income classification of each country and different years was collected i.e identified the income classification of different countries at different time like Income classification of India is low-middle income in 1984. Now according to [organization], only high income countries are classified as developed countries and others as under developed countries, so this column was also added to the dataset and merged with the disaster dataset. This would help us to identify pattern of occurrence of disasters in developed as well as under developed countries.

Co2 per capita data was from 1990, disaster dataset was 1960 so, we used Co2 emission per country dataset whose values were from 1960. We used these values and a different population dataset to figure our per capita Co2 emission from 1960 itself. Again for missing values in Co2 per country, we used Co2 emission per capita multiplied by population of that time to calculate the Co2 emission per country. Now, these columns didn’t have any missing values and were merged the disaster dataset.

Now, total funds allocated to a country to prevent natural disasters was added to the disaster dataset. This would enable us to understand if the funds are helping the country reduce the impact of disaster on their economy and people. Aid provided to a country in a single financial year through various different programmes was added as the funds dataset had different programmes through which a country received aids in the same financial year, if merged like this, it would result in cartesian product and the final dataset would have multi fold of data, Hence the funds allocated in a FY was summed. This resulted in a dataset with total funds allocated to a country in a single financial year. This data was merged with the disaster dataset giving us our final dataset which is now to be used.

But we observed that there were still missing values in latitude and longitude the the place of occurance of disaster. This was overcome by using LocationIQ’s ForwardGeocoding API which gave us the latitude and longitude of the countries. This data which was fetched was added in the disaster dataset wherever there was missing values.

To address remaining missing values following steps were taken:

ISO\_code derived from country in income classification dataset. If ISO\_code not present Country was derived from ISO\_code

Null values in Location, Magnitude Scale, in disaster dataset was changed to Unknown

Disaster Dataset : Aid contribution, magnitude, Total Death, no Injured, Affected Number, Total Affected, Reconstruction Cost, Reconstruction cost adj, Insured damage, Insured damage Adj, Total Damage and total damage adj total funds, GEF grant, Cofinancing null values changed to 0

**Append your Work after this if any :\**