**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.

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