**Week 2 – Lab Assignment**

**NOTES:**

1. **Before attempting this lab assignment, make sure you have completed the exercise in week 2 - lab exercise**
2. **Upon completing this exercise, you’ll need to upload your python (jupyter .ipynb file) to canvas before the end of this week.**

**Assignment details:**

* **Rules:**
  + The code must be **fully commented**.
  + **Print all outputs** so results are clearly visible.
  + **Use Pandas and Matplotlib** to complete the tasks.
  + **Do not use NaN values**—filter or replace missing data where necessary.
  + Your final script should run **without errors** when executed.
* **Datasets:** Use one or more of the following (you should have those already from Week 1, if not ask the staff available at the lab):
  + GlobalLandTemperaturesByCountry.csv
  + owid-co2-data.csv
  + owid-covid-data.csv

**📌 Tasks**

**🔹 Part 1️ Data Cleaning & Exploration**

💡 **Objective:** Prepare and clean the dataset before analysis.

**Q1.** Choose **one dataset** (CO₂, COVID, or Temperature) and clean it by:

* Removing **all rows with missing values** for key columns.
* Filtering only **data before 2023** (to avoid incomplete data).
* Keeping only **relevant columns** that will be useful for analysis.

**Q2.** Display basic **summary statistics** using .describe().

* What are the **minimum, maximum, and average values** for key numerical columns?
* Check for **missing values** using .isna().sum() and confirm they were removed.

**🔹 Part 2️ Aggregation & Filtering**

💡 **Objective:** Extract meaningful insights from the dataset.

**Q3.** Identify the **top 5 countries or regions** based on a key metric:

* If working with CO₂ → Select the **highest CO₂-emitting countries in 2021**.
* If working with Temperature → Find the **countries with the highest average temperature** before 2023.
* If working with COVID → Find the **5 worst-affected countries in 2021** based on total cases or deaths.

**Q4.** Find the **average value of a key metric per year** using .groupby().

* If using CO₂, find the **average CO₂ emissions per year**.
* If using Temperature, calculate **yearly average temperatures**.
* If using COVID, find **new cases per year**.
* Display results **as a DataFrame**, showing the **yearly trend**.

**🔹 Part 3️ Advanced Data Processing**

💡 **Objective:** Apply transformations and enhance the dataset.

**Q5.** Add a **new column** that shows **percentage change** in a key metric year-over-year.

* Example: If using CO₂, calculate **percentage change in CO₂ emissions per year**.
* Example: If using Temperature, calculate **percentage change in average temperature per year**.
* Example: If using COVID, calculate **percentage change in total cases per year**.

💡 **Hint:** Use .pct\_change() to compute the percentage difference between years.

**Q6.** Merge this dataset with **another dataset** (if applicable).

* Example: **Merge CO₂ data with temperature data** for the same country.
* Example: **Merge COVID data with population data** (if the dataset contains population figures).
* Ensure that the merged dataset **only includes years with available data in both datasets**.

**🔹 Part 4: Data Visualisation (Q7 - Q9)**

**Q7: Line Chart - Trend Over Time**

💡 **Objective:** Create a **line chart** showing how a key metric has changed over time.

* Select a **country or region** and **plot the trend** of one metric (e.g., CO₂ emissions, temperature, COVID cases).
* The **x-axis should represent the year**, and the **y-axis should represent the chosen metric**.
* Add **labels, a title, and a legend** to make the chart meaningful.

**Example Questions:**

* How have **CO₂ emissions in the UK** changed over the years?
* How has **the global average temperature** fluctuated over time?
* How did **COVID cases in Vietnam** evolve throughout 2021?

📌 **Deliverable:** A line chart showing the selected trend.

**Q8: Bar Chart - Top Contributors**

💡 **Objective:** Use a **bar chart** to compare top contributors for a key metric in a given year.

* Identify the **top 10** contributors (countries or regions) for a chosen metric.
* The **x-axis should represent the country or region**, and the **y-axis should represent the metric value**.
* Use a **single year** to make the comparison meaningful (e.g., CO₂ emissions in 2021).
* Rotate the x-axis labels if necessary for readability.

**Example Questions:**

* Which **10 countries had the highest CO₂ emissions in 2021**?
* Which **10 countries had the highest temperatures before 2023**?
* Which **10 countries had the highest COVID cases in 2021**?

📌 **Deliverable:** A bar chart comparing the top 10 contributors for a metric.

**Q9: Scatter Plot - Relationship Between Two Metrics**

💡 **Objective:** Explore **correlation** between two variables using a **scatter plot**.

* Select **two related metrics** from a dataset.
* The **x-axis should represent one variable**, and the **y-axis should represent the other**.
* Ensure that the dataset **does not contain NaN values** for these metrics.
* **Fit a trend line (optional)** to show possible correlations.

**Example Questions:**

* Does **higher CO₂ emissions correlate with higher temperatures**?
* Does **a country's population size affect total CO₂ emissions**?
* Is there **a link between COVID cases and total deaths per country**?

📌 **Deliverable:** A scatter plot showing the relationship between two metrics.

**✅ Submission Guidelines**

Ensure your Jupyter Notebook **includes all required charts and explanations**.

* Save your notebook as **Week2\_lab\_assignment\_YourID.ipynb**.
* Upload your **.ipynb file** to Canvas before the deadline.