Project 1

For this assignment, student groups can choose any of the three project topics outlined. It is perfectly acceptable if multiple groups choose the same project, as each group will be selecting their own database, variables, and making key choices throughout the analysis. As a result, even if you choose the same project as another group, I expect your reports to be unique and reflect your individual approach. In addition to submitting a PDF of your report, you must also include any Python code you developed for your analysis, as all analysis must be conducted using Python. Please submit your project through Moodle no later than **14 October 2024**, and ensure both the report and code are included in your submission (put the code in a single file).

1) Air Quality Analysis in a Major City

Requirements:

1. **Data Collection**: Each group must find an open-source time series dataset on air quality for a major city. The dataset must include measurements for at least three air quality parameters (e.g., PM2.5, PM10, SOx, NOx, CO, etc.) from at least two monitoring stations within the city.
2. **Data Cleaning**: If needed, the collected dataset must be cleaned to handle outliers, and ensure proper formatting for analysis. Groups must document the cleaning process in detail.
3. **Summary Statistics**: Calculate and present summary statistics (e.g., mean, median, standard deviation, variance, minimum, and maximum etc., [No need to cover this list, choose your own set of statistics]) for each parameter at both stations. Provide a table summarizing these statistics for easy comparison.
4. **Time Series Plots**: Generate time series plots for each parameter at both stations. These plots should clearly show how the air quality changes over time, and be properly labeled.
5. **Histograms**: Create histograms for each air quality parameter to show the distribution of values across the time period for both stations. Use these to identify any skewness, outliers, or common value ranges.
6. **Correlation Analysis**:
    - Perform correlation analysis between different air quality parameters (e.g., PM2.5 vs. NOx) within the same station. Calculate correlation coefficients and provide statistical evidence to support any significant correlations.

- Perform correlation analysis between the two stations for each parameter (e.g., PM2.5 at Station A vs. PM2.5 at Station B). Identify and interpret any significant differences or similarities.

7. **Trend Analysis**: Select one parameter at one station (e.g., PM2.5 at Station A) and perform a trend analysis. Identify whether the parameter shows a significant increasing or decreasing trend over time. Use statistical tests or visual trend lines to support your conclusions.
8. **Reporting**: Submit a report that includes:
    - The summary statistics table.
    - Time series plots and histograms for each parameter.
    - Results and interpretation of the correlation analyses.
    - Trend analysis results for the selected parameter and station.
    - A discussion of the key findings, including any insights on air quality trends in the chosen city and potential implications for public health or environmental policy.

2) Climate Data Analysis Using ERA5-Land Dataset

Requirements:

1. **Data Collection**: Each group must download a subset of the **ERA5-Land dataset** from a reliable source (e.g., Copernicus Climate Data Store) for a specific region or country. The dataset should cover at least five years of data and include at least three meteorological parameters (e.g., temperature, precipitation, wind speed, soil moisture, surface radiation).
2. **Data Cleaning**: Groups must clean the dataset by handling missing values, ensuring consistency in units, and formatting the data appropriately for analysis. The cleaning process must be thoroughly documented.
3. **Summary Statistics**: Calculate and present summary statistics (mean, median, standard deviation, variance, minimum, and maximum [No need to cover this list, choose your own set of statistics]) for each parameter across the entire time period. Display these statistics in a summary table for clear comparison.
4. **Time Series Plots**: Generate time series plots for each meteorological parameter over the selected region and time period. The plots should show trends or patterns in the data over time, with clear labels and annotations.

5. **Histograms**: Create histograms for each meteorological parameter to show the distribution of values during the time period. Use these to identify any skewness, outliers, or common value ranges.
6. **Correlation Analysis**:
    - Perform correlation analysis between different meteorological parameters (e.g., temperature vs. precipitation). Calculate the correlation coefficients and provide statistical evidence for any significant relationships.
    - Interpret the correlations between parameters, explaining possible physical or environmental reasons behind any strong correlations or lack thereof.
7. **Trend Analysis**: Select one parameter (e.g., temperature or precipitation) and perform a trend analysis for the selected region. Identify whether the parameter shows a significant increasing or decreasing trend over the five-year period. Use statistical methods (e.g., linear regression, moving averages) to quantify the trend and provide evidence to support your conclusions.
8. **Reporting**: Submit a report that includes:
    - The summary statistics table for each parameter.
    - Time series plots and histograms for each meteorological parameter.
    - Results and interpretation of the correlation analysis.
    - Trend analysis results for the selected parameter.
    - A discussion of the key findings, focusing on how meteorological conditions in the region may have evolved over time and potential implications for climate change or regional environmental management.

3) Analysis of Species Distribution Changes Over Time

Requirements:

1. **Data Collection**: Each group must find an open-access biodiversity dataset containing species occurrence data over time. Possible data sources include the *Global Biodiversity Information Facility (GBIF)*, iNaturalist, or other local/regional biodiversity databases. The dataset should cover at least 10 years of species occurrences for at least two species within a specific geographic region.

2. **Data Cleaning**: Clean the dataset by removing any records with missing geographic or temporal data and ensure all species occurrence points are formatted correctly for analysis. Document the data cleaning process thoroughly, including any assumptions made during data preprocessing.
3. **Summary Statistics**: Calculate and present summary statistics (e.g., total number of occurrences, number of unique locations, species richness, and changes in occurrence frequency over time [No need to cover this list, choose your own set of statistics]). Present these statistics in a clear and concise summary table.
4. **Time Series Plots**: Create time series plots showing the number of occurrences for each species over time. These plots should help visualize how the distribution or frequency of occurrences for each species has changed over the selected time period.
5. **Correlation Analysis**:
    - Perform a correlation analysis between the occurrences of the two selected species over the time period. Calculate correlation coefficients and determine whether there is any significant relationship between the occurrences of the two species. Provide statistical evidence to support any conclusions.
    - Interpret whether the two species show positive or negative correlation in their occurrence patterns and discuss potential biological or ecological reasons behind the relationship.
6. **Trend Analysis**: Choose one species and perform a detailed trend analysis to assess how its occurrence frequency has evolved over the selected time period. Investigate whether the species shows significant increases or decreases in occurrence. Support your conclusions with visual or statistical evidence.
7. **Reporting**: Submit a report that includes:
    - The summary statistics table on species occurrences.
    - Time series plots showing changes in occurrence frequency for both species.
    - Results and interpretation of the correlation analysis between the two species' occurrences.
    - Trend analysis results for the selected species.
    - A discussion of the key findings, including potential biological or ecological factors influencing species distribution changes and their conservation implications.