Session 2 Exercise: Data Analysis and Visualization with NumPy, pandas, and Matplotlib

In this exercise, you will clean and analyze a dataset using NumPy, pandas, and matplotlib. This will allow you to:

- Perform some data cleaning on the dataset.
- Perform numerical operations with NumPy.
- Load, explore, and filter data using pandas.
- Visualize data trends with matplotlib.

The dataset provided is `city_temperature_with_nan.csv` (https://drive.google.com/file/d/1-RF1x3N2aRyN9qcjZ1iANgcX2zr8BYr8/view?usp=sharing) with the following columns:

- **City**: Name of the city.
- **Date**: Date of the recorded temperature (YYYY-MM-DD).
- **Temperature**: Recorded temperature in degrees Celsius (some missing values).
- **Rainfall (mm)**: Recorded rainfall in millimeters (some missing values).
- **Time Zone**: Time zone city is located.

Submission Instructions:

1. Rename your notebook in the following format: lastname_firstname.ipynb
2. Download your notebook from the dashboard to your local computer.
3. Email your notebook to egarza@tacc.utexas.edu, use subject line: ACSC Session 2 Deliverable

Deadline to Submit

- If you don't plan to attend the training session then you deadline is 10:00p CT, 9:00p MT, 8:00p PT, 7:00p AKT, 5:00p HT Wednesday (10/30/24).
- For all others you can submit by 5:00p CT, 4:00p MT, 3:00p PT, 2:00p AKT, 12:00p HT Monday (11/04/24)

Pre Setup: importing necessary libraries

In [ ]:
```
# import any libraries that you will require to clean and analyze the dataset
```

Part 0: Data Cleaning with pandas

Before starting the analysis, we need to ensure our data is clean and free of inconsistencies. This is an essential step in data analysis to handle missing values and make sure our data is in a suitable format for analysis.

Task:

1. Identify missing values in the dataset.
2. Why do you think those values are missing? Write your hypothesis as a comment in the code.
3. Apply the following methods to handle missing values:
   - Method 1: Fill missing values in `Temperature` with a specific value (e.g., mean or median of the column) for each city.
   - Method 2: Replace missing values in `Rain` with zero.
4. Check for and remove any duplicate rows.

Let's start by identifying missing values in the `Temperature` and `Rainfall (mm)` columns.

In [ ]:
```
# Load the dataset with NaN values
```

In [ ]:
```
# Take a look at a sample of 15 rows from the dataset
```

In [ ]:
```
# Display the number of missing values in each column
```

In [ ]:
```
# Handling missing values
# Option 1: Fill NaN values in Temperature with the mean value of the column
```

In [ ]:
```
# Option 2: Fill NaN values in Rainfall with 0 (assuming no rainfall as a possi
```

In [ ]:
```
# Check for duplicates and remove them
```

In [ ]:
```
# Verify the data after cleaning, you can use a sample of 15 rows
```

Part 1: Numerical Analysis with NumPy

First, we'll use NumPy to perform basic operations on the Temperature data.

Task:

- Extract the Temperature column as a NumPy array.
- Calculate the mean, minimum, maximum, and standard deviation of the temperatures.

In [ ]:
```
# Convert the Temperature column to a NumPy array
```

In [ ]:
```
# Calculate statistics: mean_temp, min_temp, max_temp, std_temp
```

In [ ]:
```
#print out your output
```

Part 2: Data Analysis with pandas

In this section, we'll load and explore the dataset using pandas.

Task:

- Display basic information and statistics on the dataset.
- Calculate the mean temperature and total rainfall for each city.

In [ ]:
```
# Display the first few rows
```

In [ ]:
```
# Display basic info and statistics
```

In [ ]:
```
# Calculate the mean temperature and total rainfall by city
```

In [ ]:
```
# print out your output
```

Part 3: Filtering and Subsetting Data

We will now filter the data to focus on specific records.

Task:

- Filter records where temperature is above 25°C.
- Subset data to only include records from a specific city (e.g., "New York").

In [ ]:
```
# Filter records where temperature is above 25°C
```

In [ ]:
```
# Filter records for a specific city (e.g., New York)
```

Part 4: Data Visualization with Matplotlib

Visualize the data using matplotlib to identify trends and comparisons.

Task:

- Plot the temperature trend over time for a specific city (e.g., "New York").
- Create a bar chart comparing total rainfall across different cities.

In [ ]:
```
# import necessary libraries

# Line plot for temperature trend over time (e.g., for New York)


```

```
# Bar chart for total rainfall by city
```

Summary

In this exercise, you demonstrated how to:

- Perform data cleaning with pandas.
- Conduct numerical analysis with NumPy.
- Explore and filter data with pandas.
- Visualize data trends and comparisons with matplotlib.

Complete each section and ensure the results are well-documented.

Submission Instructions:

1. Rename your notebook in the following format: lastname_firstname.ipynb
2. Download your notebook from the dashboard to your local computer.
3. Email your notebook to egarza@tacc.utexas.edu, use subject line: ACSC Session 2 Deliverable

Deadline to Submit

- If you don't plan to attend the training session then you deadline is 10:00p CT, 9:00p MT, 8:00p PT, 7:00p AKT, 5:00p HT Wednesday (10/30/24).
- For all others you can submit by 5:00p CT, 4:00p MT, 3:00p PT, 2:00p AKT, 12:00p HT Monday (11/04/24)