Assignment 1: Comprehensive Analysis of Global Weather Data

Available: Jan 08, 2025 **Deadline:** Jan 25, 2025

Overview

Welcome to your first assignment for **FRE 521D**! This assignment will require you to analyze a global weather dataset and apply various data analysis, engineering, and visualization techniques. By the end of this assignment, you will have gained experience in cleaning, processing, and modeling complex datasets, as well as presenting your findings in a structured format.

Learning Objectives

By completing this assignment, you will:

- Explore and understand the structure of a real-world dataset.
- Apply key data cleaning and preprocessing techniques.
- Perform exploratory data analysis (EDA) to uncover patterns.
- Engineer new features to enhance the dataset.
- Build predictive models using machine learning algorithms.
- Effectively communicate your findings using visualizations and written reports.

Dataset

The dataset contains weather-related data points from various global locations. It includes both **categorical** and **numerical** features, which you will explore and process. Key features in the dataset are:

- **Location Details:** `country`, `location_name`, `latitude`, `longitude`, and `timezone`.
- **Weather Parameters:** `temperature_celsius`, `humidity`, `wind_kph`, `condition_text`, `pressure_mb`.
- **Derived Conditions:** `feels_like_celsius`, `air_quality_PM2.5`, `moon_phase`.
- **Temporal Data:** `last_updated`, `sunrise`, and `sunset`.

Step 1: Understanding the Dataset

Start by getting familiar with the dataset:

1. **Data Loading:**
   - Read the dataset into a Pandas DataFrame. Ensure proper handling of file formats (CSV or Excel).
   - Print the first 5 rows of the dataset using `.head()` and identify its structure.
2. **Initial Exploration:**
   - Identify the types of features (numerical vs. categorical).
   - Use `.info()` and `.describe()` to understand the data types, ranges, and missing values.
3. **Dataset Summary:**
   - Summarize the features in a markdown cell or your report:
     - Which columns are categorical?
     - Which columns are numerical?
     - Are there any temporal or geospatial columns?

Step 2: Data Cleaning

Prepare the dataset for analysis:

1. **Handle Missing Data:**
   - Identify missing values using `.isnull().sum()`.
   - For numerical columns (e.g., `temperature_celsius`, `humidity`), fill missing values with the **mean** or **median**.
   - For categorical columns (e.g., `condition_text`), replace missing values with the **mode**.
2. **Remove Duplicates:**
   - Check for duplicate rows using `.duplicated()` and drop them if necessary.
3. **Date and Time Conversion:**
   - Convert `last_updated` to a proper datetime format using `pd.to_datetime`.
   - Extract new features such as:
     - `hour_of_day`: Extract the hour of the day from `last_updated`.
     - `is_daytime`: Determine whether the timestamp corresponds to daytime or nighttime based on `sunrise` and `sunset`.
4. **Verify Data Integrity:**
   - Check for any anomalies in ranges. For example:
     - `temperature_celsius` should not exceed 70°C or drop below -100°C.
     - `humidity` should be between 0 and 100%.

Step 3: Exploratory Data Analysis (EDA)

Analyze the data to gain meaningful insights:

1. **Summary Statistics:**
   - Generate statistical summaries for all numeric columns using `.describe()`.
   - Highlight key insights, such as:
     - The average global temperature.
     - Variations in air quality indicators (`air_quality_PM2.5` and `air_quality_Carbon_Monox`
2. **Visualizations:**
   - **Univariate Analysis:**
     - Create histograms for columns like `temperature_celsius`, `humidity`, and `wind_kph`.
     - Plot a bar chart for the distribution of `condition_text`.
   - **Multivariate Analysis:**
     - Plot a correlation heatmap for numerical features.
     - Create scatter plots for:
       * `temperature_celsius` vs. `humidity`
       * `temperature_celsius` vs. `feels_like_celsius`
   - **Location Analysis:**
     - Use `latitude` and `longitude` to map temperatures on a world map.

Step 4: Model Building and Application

After cleaning and exploring the dataset, you will build predictive models to gain insights into the data and make meaningful predictions. Below are the tasks and guiding questions to help structure this section:

1. **Objective Definition:**
   - Modeling questions (you need to answer only one of them):
     - Can we predict the `temperature_celsius` based on features like `humidity`, `wind_kph`, and `pressure_mb`?
     - How accurately can we classify weather conditions (`condition_text`) using numerical features?
2. **Train-Test Split:**
   - Split the dataset into training and testing sets.
   - Ensure a balanced distribution of classes if working with classification.
3. **Model Selection:**
   - Use appropriate models based on the prediction goals:

  - For regression: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor.
  - For classification: Logistic Regression, Decision Trees, Support Vector Machine (SVM).
- Compare models based on their performance metrics.

4. **Evaluation Metrics:**
   - Use appropriate metrics to evaluate the models:
     - Regression: Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R^2).
     - Classification: Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.
   - Provide visualizations such as residual plots for regression or ROC curves for classification.
5. **Model Interpretation:**
   - Analyze the importance of features contributing to the model's predictions.
   - Answer questions like:
     - Which features are the most influential in predicting `temperature_celsius`?
     - How well does the model generalize to unseen data?

Submission Guidelines

Your submission should include:

1. **Jupyter Notebook:**
   - Ensure the code is clean, well-commented, and runs without errors.
2. **Report (PDF):**
   - Include the following sections:
     - Introduction and objectives.
     - Data cleaning and preprocessing.
     - EDA and key findings.
     - Feature engineering.
     - Model building and evaluation.
     - Conclusion and recommendations.

Evaluation Criteria

Component	**Points**
Data Cleaning	15
EDA	20
Feature Engineering	15
Visualization	20
Predictive Modeling	25
Report and Documentation	15
Total	**100**