**Data Science Tech Assessment – Rohit Lahori**

PM Accelerator mission:

**1. Introduction**

This report presents an analysis of the Global Weather Repository dataset. The aim is to uncover trends and patterns in global weather data. We perform the following steps:

* **Data Cleaning & Preprocessing:** Remove outliers, convert date fields, and normalize select variables.
* **Exploratory Data Analysis (EDA):** Explore the relationships between variables, assess correlations, and visualize time series data.
* **Panel Regression Modeling:** Build a fixed effects panel regression model to predict normalized temperature, accounting for country-specific effects.
* **Model Evaluation:** Evaluate model performance using common metrics and visualize the in-sample predictions.
* **Advanced EDA:** Implement anomaly detection to identify and analyze outliers.
* **Forecasting with Multiple Models:** Build and compare multiple forecasting models and create an ensemble of models to improve forecast accuracy.
* **Unique Analyses:** We analyze air quality and its correlation with weather parameters and explore how weather conditions differ across countries and continents.

**2. Data Cleaning and Preprocessing**

**2.1 Missing values and sorting**

We check the dataset for missing values and find out the features don’t have any missing values. The data was sorted by the last\_updated timestamp. The last\_updated column was converted to a datetime format to facilitate time series analysis

**2.2 Outlier removal and normalization**

The IQR (Interquartile Range) method was used to remove outliers for key variables (e.g., air\_quality\_Carbon\_Monoxide and wind\_mph), ensuring that extreme values do not skew the analysis.

Using a MinMaxScaler, temperature (in Celsius) and precipitation (in mm) were normalized. New columns (temp\_c\_scaled and precip\_mm\_scaled) were created for further analysis.

**3. Exploratory Data Analysis (EDA)**

**3.1 Correlation Analysis**

* A correlation matrix was computed for the numeric variables to identify relationships among them.
* The correlation matrix was visualized using a heatmap, helping to quickly spot strong and weak correlations between weather attributes.

A screenshot of a graph

AI-generated content may be incorrect.

We can see from the heatmap gust and wind speed are highly correlated. This makes intuitive sense—when wind speeds are high, gusts also tend to be stronger.

Temperature (°C) has a negative correlation with both humidity and pressure\_mb. In general, higher temperatures are often associated with lower humidity (especially in dry climates) and slightly lower pressure.

Temperature and uv\_index also have moderately strong positive correlation along with humidity and cloud as well.

**3.2 Time Series and Scatter Plots**

* **Time Series Analysis:**  
  For a sample country (e.g., the United States of America), the time series plots for temperature and precipitation were created. These plots reveal temporal trends and seasonal variations.
* **Scatter Plot:**  
  A scatter plot of temperature versus precipitation helps to visualize the relationship between these two variables.

A graph with blue dots

AI-generated content may be incorrect.

A graph of a number of patients

AI-generated content may be incorrect.

A graph of blue dots

AI-generated content may be incorrect.

**4. Panel Regression Modeling**

**4.1 Data Preparation for Panel Analysis**

* A panel dataset was created by selecting the relevant variables (including temp\_c\_scaled, precip\_mm\_scaled, humidity, wind\_mph, and pressure\_mb) and setting a multi-index with country and last\_updated.

**4.2 Building the Fixed Effects Model**

* A fixed effects panel regression (PanelOLS) was used to model normalized temperature as a function of weather predictors, accounting for country-specific effects.

**4.3 Model Evaluation**

* **In-Sample Predictions:**  
  The model's fitted values were compared to the actual normalized temperature values.
* **Performance Metrics:**  
  Common evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were calculated to quantify the model’s performance.

Evaluation Metrics for the PanelOLS Model:

Mean Squared Error (MSE): 0.0134

Mean Absolute Error (MAE): 0.0939

Root Mean Squared Error (RMSE): 0.1157

**4.4 Visualization of Model Predictions**

* For the sample country, the actual versus fitted normalized temperature is plotted over time. This visualization helps assess how well the model captures the temporal patterns.

A graph of blue and orange lines

AI-generated content may be incorrect.

**5. Advanced EDA and Anomaly Detection**

**5.1 Anomaly Detection**

Using the **IsolationForest** algorithm, we detect outliers in key weather features such as temperature, humidity, wind speed, and pressure. This helps flag unusual observations that may indicate data errors or extreme weather events.

**Code Overview:**

* **Feature Selection:**  
  The model uses temperature\_celsius, humidity, wind\_mph, and pressure\_mb as input features.
* **Anomaly Flagging:**  
  Observations labeled as -1 by IsolationForest are considered anomalies. We then visualize these anomalies using a boxplot for temperature.

A graph of a temperature

AI-generated content may be incorrect.

**6. Forecasting with Multiple Models**

We build three forecasting models on a subset of the data (United States of America) and then create an ensemble to improve forecast accuracy.

**6.1 Data Preparation**

* Convert the last\_updated column to datetime.
* Focus on data for the United States, setting the index as the timestamp.
* Handle missing values by time interpolation and ensuring a regular frequency.

**6.2 Forecasting Models**

**Model 1: ARIMA**  
An ARIMA model (order (5,1,0)) is fitted on the training data to forecast temperature.

**Model 2: SARIMAX**  
A seasonal ARIMAX model is used with seasonal order (1,1,1,12) to account for possible seasonality in the data.

**Model 3: Random Forest Regressor**  
A regression-based approach is applied by creating lag features. A Random Forest Regressor forecasts temperature using past values.

The models are evaluated using Mean Absolute Error (MAE):

ARIMA MAE: 0.95

SARIMAX MAE: 1.25

Random Forest MAE: 1.92

**6.3 Ensemble Forecasting**

An ensemble forecast is created by computing weights inversely proportional to the MAE of each model. These weights are then used to compute a weighted average forecast.

Weighted Ensemble MAE: 1.01

**7. Unique Analyses**

**7.1 Air Quality and Weather Correlation**

We analyze the relationship between air quality indicators (e.g., CO, Ozone, PM2.5, etc.) and weather parameters (e.g., temperature, humidity, wind speed). A correlation matrix is computed and visualized as a heatmap.

A screen shot of a chart

AI-generated content may be incorrect.

**7.2 Weather Differences Across Countries and Continents**

We explore how average weather conditions differ across major countries and continents. For instance:

* **By Country:**  
  Average temperature, humidity, and precipitation are computed and visualized for the top 10 countries by data volume.

A graph of a temperature

AI-generated content may be incorrect.

A graph of a number of people

AI-generated content may be incorrect.

A graph of a number of blue bars

AI-generated content may be incorrect.

* **By Continent:**  
  By mapping countries to continents, we compute and visualize average weather conditions by continent.

A graph of a temperature

AI-generated content may be incorrect.

A graph of a number of blue rectangular bars

AI-generated content may be incorrect.

A graph of a number of blue rectangular bars

AI-generated content may be incorrect.