Documentation on: - Data Cleaning and Solution Planning

Team Name: - Data Dynamos

Problem Statement: - Weather Data Analysis and Prediction

Problem Statement Description

Use weather datasets to predict different weather conditions for specific region. This can help in planning for agricultural, travel needs or etc.

Data Cleaning Steps

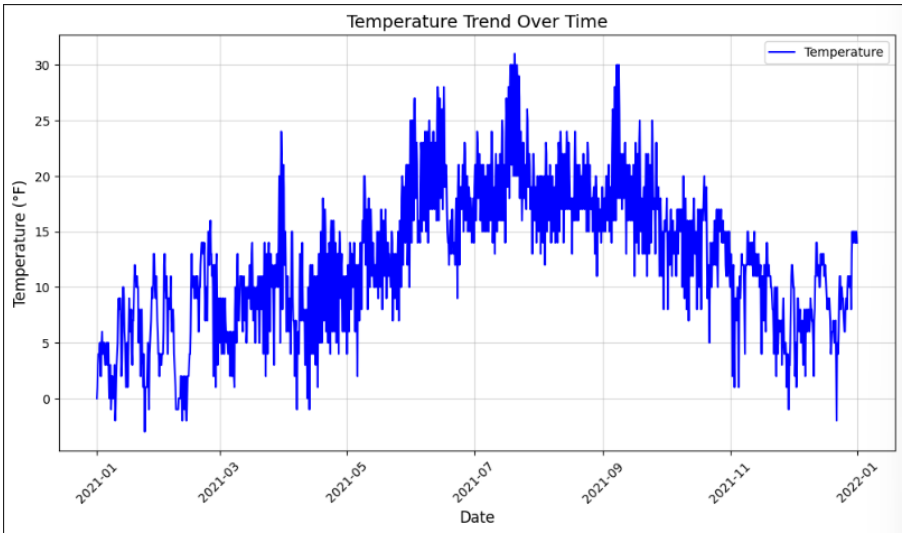
1. **Importing Libraries:**
    - **Pandas** – Used for reading, writing, and manipulating tabular data efficiently. It helps in handling missing values, filtering, and transforming datasets.
    - **NumPy** – Essential for numerical computations, handling large datasets, and working with arrays and missing values.
    - **Matplotlib & Seaborn** – Used for data visualization, allowing for plotting graphs, histograms, and trend analysis to better understand the dataset.
    - **Scikit-learn (sklearn)** – Provides machine learning tools for regression, classification, and data preprocessing, essential for building predictive models.
    - **Statsmodels** – Used for statistical analysis and time series modeling, helping to analyze trends and seasonal patterns in the dataset.
2. **Initial Data Assessment:**
    - Inspecting for missing values and duplicates in the dataset.
    - Verifying and adjusting data types (e.g., converting timestamps to datetime).
3. **Missing Values Handling:**
    - For numerical features (temp, baro): We will use mean or median imputation because these methods provide a balanced replacement for missing values without significantly distorting the distribution.
    - For categorical features (desc, Day_of_Week): Using mode imputation because it preserves the most common category in the dataset, maintaining the categorical distribution and avoiding introducing any bias.
4. **Feature Engineering:**
    - Converting date_id to a readable date format and extracting temporal features (e.g., hour, week).
    - Combining Day, Month, Year into a single Date column.

5. **Outlier Detection:**
    - Identifying outliers in continuous variables (temp, baro, wind) using boxplots or statistical methods as it helps detect these anomalies visually or quantitatively.
    - Replacing outliers with bounds based on domain knowledge. This ensures that extreme values do not disproportionately influence regression coefficients, distance metrics, or optimization algorithms.
6. **Scaling and Normalization:**
    - Normalizing numerical features using techniques like Min-Max Scaling or Standardization to ensure models performs optimally.

Visualizations used

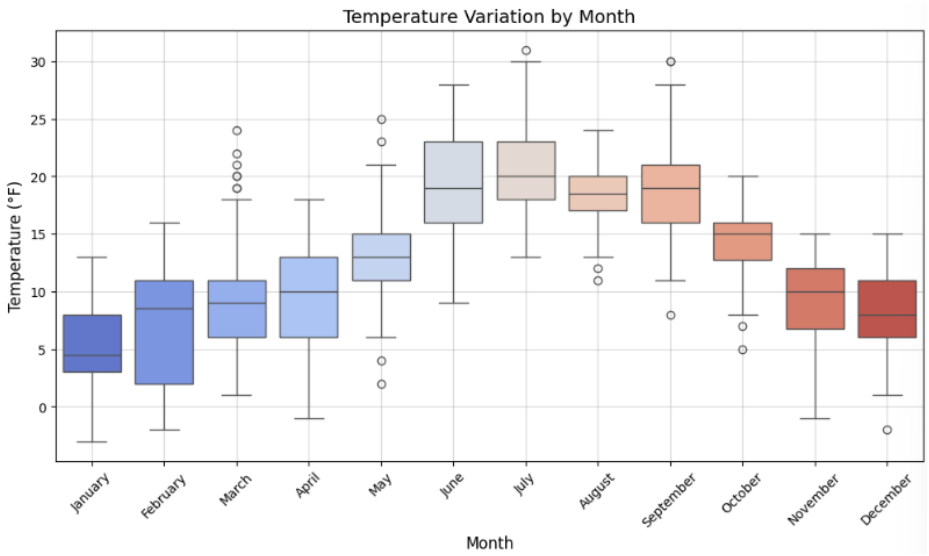
1. **Line Plot: Temperature Trends Over Time:**

    This line graph displays temperature trends over time, showcasing daily fluctuations and seasonal patterns. It provides a clear visualization of how temperature changes over the course of a year.

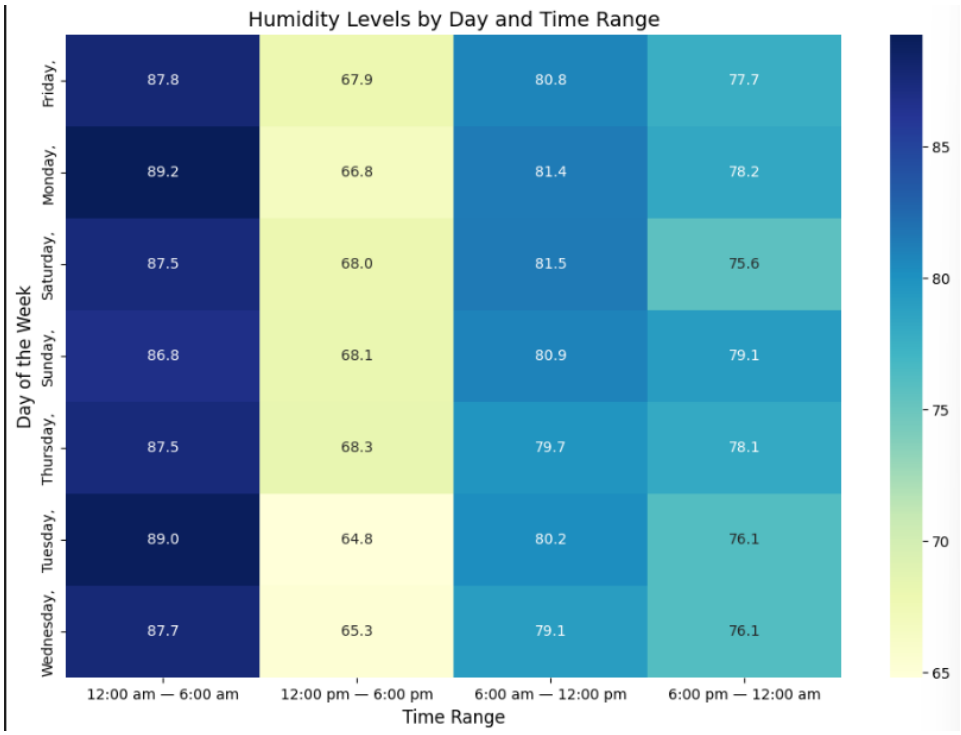


2. **Box Plot: Temperature Variation by Month:**

Box plot showing monthly temperature variations. It indicates seasonal trends, with temperatures rising in summer (June–August) and dropping in winter (December–February)

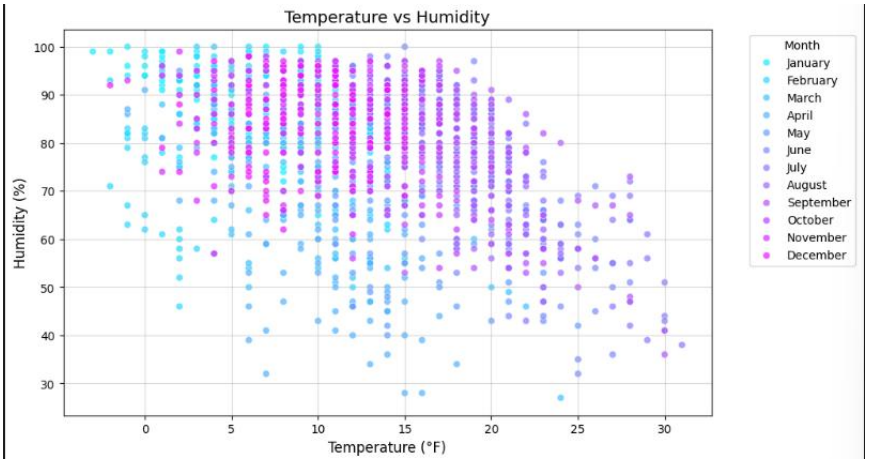


3. **Heatmap: Humidity by Day and Time:**
   This heatmap illustrates average humidity levels across different days of the week and time ranges.It offers a structured overview of humidity patterns based on day and time.

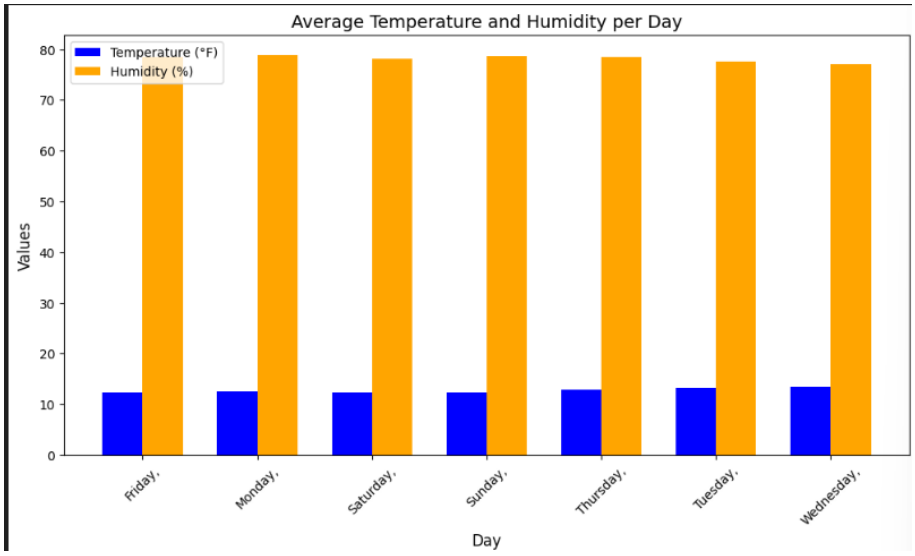


4. **Scatter Plot: Temperature vs. Humidity:**

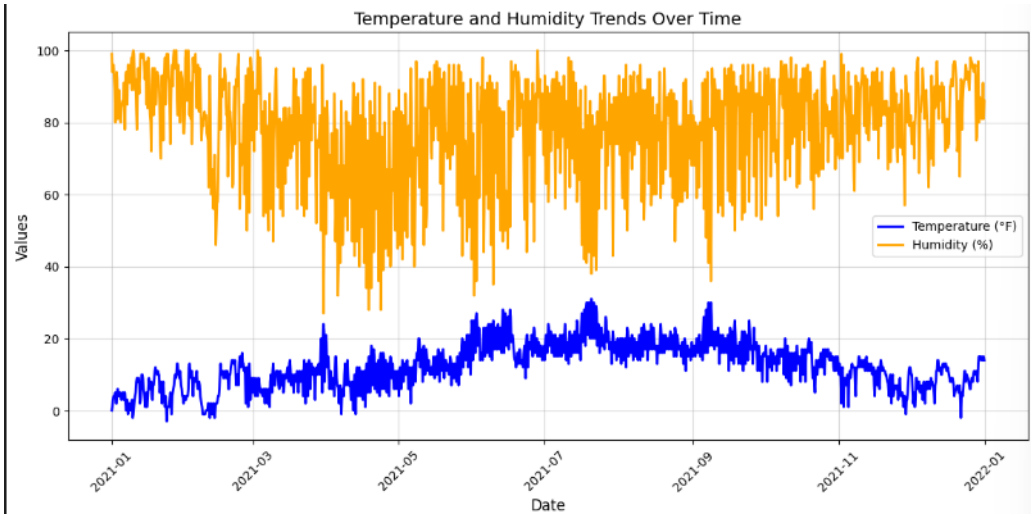
Scatter plot with points in blue and purple. It suggests clustered data with possible correlations or categorical groupings.



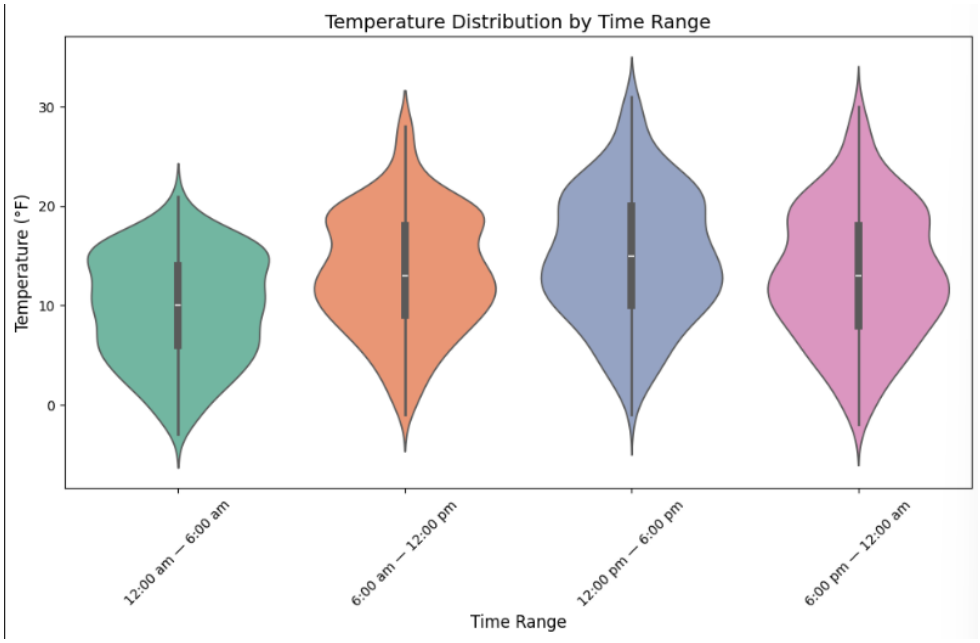
5. **Grouped Bar Chart: Average Temperature and Humidity per Day:**
   This bar chart compares the average temperature and humidity for each day of the week.It enables a quick comparison of temperature and humidity averages across different days.



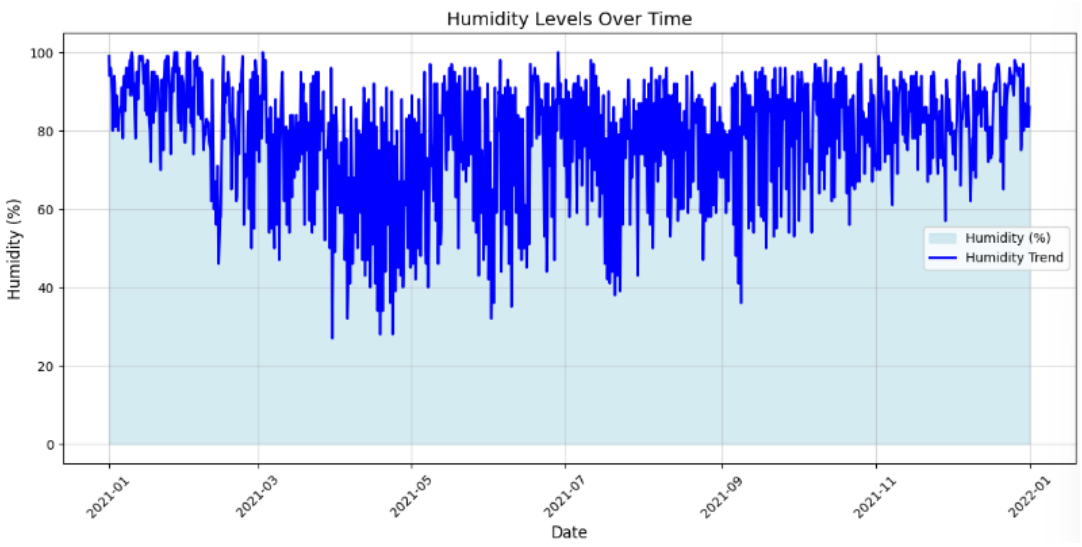
6. **Line Plot: Compare Temperature and Humidity Over Time:**
   This line graph plots both temperature and humidity trends over time. It Provides a side-by-side comparison of how temperature and humidity change over time.



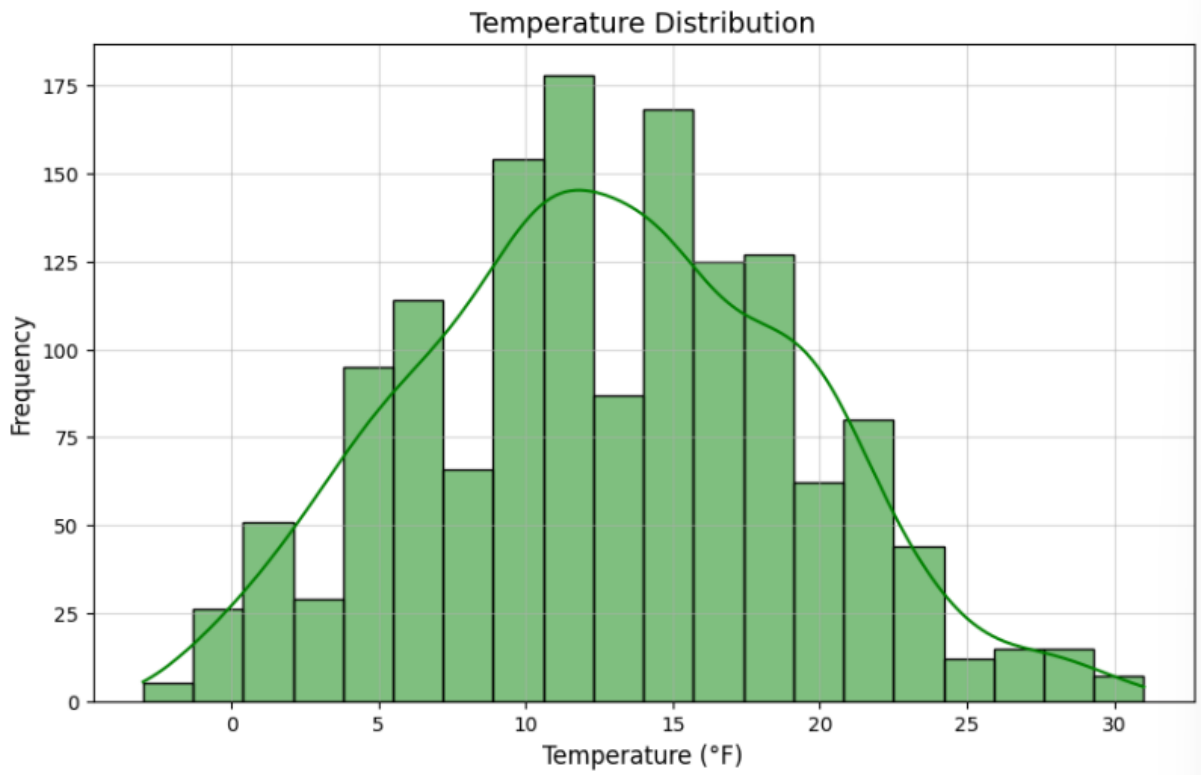
7. **Violin Plot: Temperature Distribution by Time Range:**
   This violin plot visualizes the distribution of temperatures across different time ranges.It shows how the spread of temperature data varies depending on the time of day.



8. **Area Chart: Humidity Levels Over Time:**
   This line graph displays humidity levels over time, highlighting a general trend.It effectively captures the overall humidity pattern throughout the year.



9. **Histogram: Temperature Distribution:**
   This histogram illustrates the distribution of temperature data, with a fitted curve indicating the probability density function. It provides insight into the frequency of different temperature values.



Proposed Machine Learning Models

1. Time Series Model (Timeseis Model):

Since our dataset contains sequential time-based data, the Timeseis Model is ideal for analyzing sequential time-based data, enabling accurate predictions by capturing trends, seasonality, and recurring patterns. It forecasts future weather conditions, such as temperature, humidity, and barometric pressure, using historical data. Additionally, the model identifies long-term trends and seasonal variations for improved analysis.

2. Regression Model (Linear Regression for Continuous Variables):

Linear regression is a simple model and will be used for predicting continuous variables, when features have linear dependencies. It will predict weather variables like temperature, humidity, and pressure, based on inputs such as wind speed and past conditions, helping to understand weather influences.

3. Classification Model (Support Vector Machines for Categorical Outputs)

Support Vector Machines (SVM) are ideal for high-dimensional data and complex decision boundaries. With categorical weather descriptions in our dataset, SVM is well-suited for classification tasks as it will classify weather conditions (e.g., "Low clouds," "Drizzle") based on factors like temperature, humidity, and wind speed, improving weather prediction and analysis.

Evaluation Metrics

1. For Linear Regression (Regression Metrics):

- **Mean Absolute Error (MAE)**: It will measure the average magnitude of errors in predictions.
- **Root Mean Squared Error (RMSE)**: It will penalize larger errors more heavily than MAE.
- **R-squared (R^2)**: It will represent the proportion of variance explained by the model.

2. For SVM (Classification Metrics):

- **Accuracy**: It will show overall correctness of the model.
- **Precision**: It will show the proportion of true positives among predicted positives.
- **Recall**: It will show the proportion of true positives among actual positives.

- **F1 Score**: It will show the harmonic mean of precision and recall.
- **Confusion Matrix**: It will show provide a breakdown of predicted vs. actual values.

3. Time Series Metrics:

- **Mean Absolute Percentage Error (MAPE)**: It will measure prediction accuracy as a percentage.
- **Mean Squared Error (MSE)**: Average of squared differences between predictions and actual values.
- **Symmetric Mean Absolute Percentage Error (SMAPE)**: A balanced metric for percentage errors.