DAI Assignment-1 (23114023 – Darsh Jain)

1. Data Preprocessing

1.1 Data Loading

- The dataset is loaded from weatherAUS.csv using pandas.read_csv(), with the Date column parsed as a datetime index.

1.2 Handling Missing Values

- The code prints the number of missing values for each column.
- Imputation techniques, including interpolation, are applied.
- Some variables undergo direct removal if they have too many missing values.

1.3 Data Summary

- The dataset's shape is printed to understand the number of rows and columns.
- data.describe() provides statistical summaries (mean, min, max, percentiles).
- data.info() helps identify categorical and numerical variables.

---

2. Exploratory Data Analysis (EDA)

2.1 Outlier Detection and Removal

- The Interquartile Range (IQR) method is used to identify and remove outliers.
- The code initially removes outliers but later replaces them with the lower or upper bound instead.

2.2 Frequency Distribution

- Histograms and bar plots visualize the frequency distributions of categorical and numerical variables.

2.3 Correlation Analysis

- A **heatmap** visualizes the correlation matrix of numerical features.
- Highly correlated variables are identified.

3. Feature Engineering: Use of Trigonometric Transformations

3.1 Sine and Cosine Transformations

- The notebook applies **sine** and **cosine** transformations to wind direction features:
  - WindGustDirCos, WindGustDirSin
  - WindDir9amCos, WindDir9amSin
  - WindDir3pmCos, WindDir3pmSin

- This transformation helps preserve the circular nature of wind direction data while making it suitable for numerical analysis.

---

4. Multivariate Analysis

4.1 Pair Plots

- sns.pairplot() is used to explore relationships between multiple numerical variables.

4.2 Scatter Plots (with Multiple Variables)

- A scatter plot is created with:
  - **X-axis:** Temp3pm
  - **Y-axis:** Humidity3pm
  - **Hue:** Location
  - **Style:** WindGustDirCos
  - **Size:** Pressure3pm

4.3 Grouped Comparisons

- Box plots and violin plots are used to compare numerical variables across different categorical values.

---

5. Visualization Techniques

5.1 Heatmaps

- Correlation among numerical variables is visualized using a **heatmap**.

5.2 Box Plots and Bar Plots

- Box plots visualize distributions and detect outliers.
- Bar plots compare numerical variables against categorical labels.

---

6. Conclusion and Insights

- The dataset underwent significant preprocessing, including outlier handling and imputation.
- Exploratory analysis revealed relationships between different weather variables.
- Various visualization techniques helped uncover patterns and trends.
- Multivariate techniques such as scatter plots, pair plots, and heatmaps provided deep insights into the interactions between multiple features.