DAI Assignment

Name: Vedant Airon **Enrollment No:** 23116106

---

1. Dataset Overview

This report presents an analysis of a dataset containing weather-related data. The primary objectives include data cleaning, understanding variable distributions, identifying relationships, and performing multivariate analysis to extract meaningful insights.

- **Shape**: 145460 data entries, 23 features

---

2. Data Cleaning

- **Missing Values:** Several columns contained missing values, with some having more than 40% missing data.
    - **Over 40%:** Sunshine, Evaporation, Cloud3pm, Cloud9am -> **Dropped**
    - **Categorical:** Fixed via **mode** (11 Features)
    - **Numerical:** Fixed via **mean** (6 Features)
- **Duplicates:** Checked for and handled potential duplicate entries
    - No duplicates were found.
- **Outliers:** Several columns showed negligible outliers except 2 columns.
    - Boxplot, Histogram with KDE, and Q-Q plots along with skewness, kurtosis, and normality tests were performed to analyse each outlier contribution and to handle accordingly.
    - **Rainfall:** 20% outliers -> Fixed partially via Log Transformation
    - **WindGustSpeed:** 4% outliers -> Winsorization was applied
    - Rest 10% showed 0-1% outliers -> Removed
- **Data Types:** Some categorical variables required formatting adjustments.

- **Date**: Formatted into **YYYY-MM-DD** format.
- **Location:** Switched from CamelCase to TitleCase.
- **Wind Directions:** Uppercased and extra space removed.

---

3. EDA: Exploratory Data Analysis

- **Univariate Analysis:**
  - **Numerical:** Statistical Analysis including mean, median, mode, std dev, skewness, kurtosis etc along Histogram and Boxplots were used to understand the data distribution.
    - **Skewness:**
      - **Rainfall:** Highly positive skewed (2.12%)
      - **Humiditiy9am:** Slightly negatively skewed (-0.35%)
      - **Rest:** Almost symmetric distribution
    - **Kurtosis:**
      - **Rainfall**: Slightly high -> **leptokurtic distribution** (3.88%)
      - **Rest:** (<1%) negative -> **platykurtic distribution**
    - **Histogram:**
      - Temperature Variables**:** Fairly normal distribution
      - **Rainfall:** Highly right-skewed
      - **WindGustSpeed:** Uniform and multimodel
      - **WindSpeed9am & Windspeed3pm:** normal distribution but with some skewness
      - **Humidity9am & Humidity3pm**: slight negative skew
      - **Pressure9am & Pressure3pm**: **negative kurtosis** and appear quite symmetric
    - **Boxplot**:
      - **MinTemp & MaxTemp:** relatively symmetric with some outliers.
      - **Rainfall**: Highly skewed distribution with outliers.
      - **WindGustSpeed:** fairly spread-out distribution with some prominent outliers.
      - **WindSpeed9am & WindSpeed3pm:** some skewness but have fewer outliers.
      - **Humidity9am & Humidity3pm:** few outliers on the lower end
      - **Pressure9am & Pressure3pm:** Nearly normal distributions with minor outliers.
      - **Temp9am & Temp3pm:** near-normal distribution.
  - **Categorical:** Frequency distribution, bar plots and pie charts were used to study the data efficiently.

        - **Location:** distribution appears relatively uniform.
        - **WindGustDir:** dominant wind gust direction appears to be **'W' (West)**
        - **WindDir9am:** dominant wind direction at 9am appears to be **'N' (North)**
        - **WindDir3pm:** dominant wind direction at 3pm appears to be **'SE' (Soth East).**
        - **RainToday & RainTomorrow: highly imbalanced** with majority being **'NO'**

- **Bivariate Analysis and Multivariate:** Relationship between Variables
    - **Numerical-Numerical** was deeply analysed using **Correlation Matrix** and **Scatter Plots** highlighting strong positive-negative correlation and patterns with non-linear dependencies respectively.
    - **Numerical-Categorical** was studied using Box Plots, Violin Plots and Bar Plots showing key distribution of numerical data across each categorical data making it easier to highlight trends across groups.
    - **Categorical-Categorical** analysis involved Stacked Barplots along with Crosstab Heatmaps to identify dominant categories and potential dependencies.
    - **Pairplots and Heatmaps** were extremely useful for studying multiple features simultaneously and complex and detailed analysis.

---

4. Findings & Interpretations

- Variables like temperature and humidity showed a strong correlation.
- Wind direction and rainfall had noticeable categorical influences.
- Missing data in critical columns required careful imputation or exclusion strategies.
- Outlier handling improved data integrity and consistency.

---

5. Conclusion

The analysis provided valuable insights into weather trends and relationships between various atmospheric factors with Rainfall and WindGustDir showing significant dominance. Multiple locations were studied with similar amount of data to deeply understand how minor factors could play crucial role in weather changes. Future work could involve predictive modelling using machine learning for deeper insights.