Climate Data Analysis: Exploratory Data Analysis & Regression

1. Introduction

This report presents an end-to-end exploratory data analysis (EDA) and regression modeling process applied to a climate dataset. The objective is to understand data distributions, relationships between climate variables, validate statistical assumptions, and build an interpretable regression model.

2. Dataset Description

The dataset contains 520 observations with climate-related variables including temperature, precipitation, humidity, wind speed, atmospheric pressure, CO■ concentration, region, and month. The data is synthetic but statistically realistic and suitable for analytical demonstration.

3. Data Quality Assessment

Initial inspection showed no missing values or data type inconsistencies. Summary statistics revealed reasonable ranges across all variables, indicating a clean and reliable dataset.

4. Exploratory Data Analysis

Univariate analysis showed that temperature, humidity, and pressure follow approximately normal distributions, while precipitation exhibited significant right skewness. Visualization techniques such as histograms, boxplots, and heatmaps were used to understand data behavior.

5. Normality Testing

The Shapiro–Wilk test was applied to numeric variables. Results confirmed normality for most features except precipitation, which required transformation before modeling.

6. Feature Transformation

A logarithmic transformation was applied to precipitation to reduce skewness and improve regression model performance.

7. Relationship Analysis

Correlation analysis showed a negative relationship between temperature and pressure, and a positive relationship between humidity and precipitation. No severe multicollinearity was detected.

8. Regional Influence

Grouping by region revealed that tropical regions have higher humidity and precipitation, arid regions exhibit lower rainfall and higher variability, while temperate regions remain relatively stable.

9. Regression Modeling

A linear regression model was built to predict temperature using selected climate features. Categorical variables were encoded, and the dataset was split into training and testing sets.

10. Model Evaluation

Model performance was evaluated using R^2 and RMSE. The model demonstrated reasonable explanatory power with interpretable coefficients, making it suitable for analytical insights.

11. Interpretation of Results

Atmospheric pressure showed a negative influence on temperature, while region and humidity had significant effects. CO■ concentration showed weak short-term influence on temperature.

12. Conclusion

This analysis demonstrates a complete EDA and regression workflow. The dataset is well-prepared for further modeling such as predictive analytics, clustering, or time-series analysis.

13. Future Work

Future improvements include advanced regression techniques, residual diagnostics, PCA, clustering, and integration of real-world climate datasets.