Date	2 oct 2025
Team ID	LTVIP2025TMIDS67798
Project Name	RAINFALL PREDICTION USING MACHINE LEARNING
Maximum Marks	2 Marks

Data Quality Report Template

Project Name: Rainfall Prediction

Team: Lakshmi Sravya Savaram
Mohammad Shouqat Azeez
N Gokul Chowdary
Nallabotula Vijaya Karthik

Date: 13/10/2025

Dataset Source: Dataset Source: Kaggle - Long-term Climatic Data for Cities in Asia
URL: https://www.kaggle.com/datasets/mohammadrahdanmofrad/long-term-climatic-data-for-cities-in-asia

Data Period: [Specify the period covered by your dataset, e.g., 2000–2023]

Geographical Scope: Multiple cities across Asia

Features: Date, Temperature, Humidity, Wind Speed, Rainfall, Weather Conditions

Description: This dataset contains historical climate data from various cities in Asia, including daily measurements of temperature, humidity, wind speed, and rainfall, suitable for predictive modeling.

Dataset Overview

Attribute	Description	Data Type	Example Values	Missing Values (%)
Date	Date of observation	datetime	2025-10-01	0%
Temperature	Daily avg temperature (°C)	float	28.5	2%
Humidity	Daily avg humidity (%)	float	78	1%
Wind Speed	Wind speed (km/h)	float	12	0%
Rainfall	Rainfall amount (mm)	float	10.5	5%
Weather	Weather description	categorical	Rainy, Sunny	0%

Missing Values Analysis

- Summary of missing data by column.
- Visual representation (optional): heatmap or bar chart.
- Handling strategy: e.g., mean/median imputation for numerical, mode or forward fill for categorical.

Feature	Missing Count	Missing Percentage	Handling Method
Temperature	10	2%	Fill with mean
Humidity	5	1%	Fill with median
Rainfall	25	5%	Fill with zero or interpolate

Duplicate Records

- Total duplicate rows: [Number]
- Action taken: [Removed/Kept]

Outlier Analysis

- Identify outliers using: ○ Z-score method ○

  IQR method

- Outliers detected per column:

Feature	Outlier Count	Handling Method
Temperature	3	Capped at max/min
Rainfall	7	Winsorization

Statistical Summary

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Temperature	500	28.6	3.2	22	26	28	31	35
Humidity	500	75.4	10.2	50	68	76	82	98
Rainfall	500	12.1	15.3	0	0	8	18	90

Data Consistency & Integrity Checks

- Check for inconsistent values in categorical fields (e.g., Weather column: “Rainy”, “rainy”, “sunny” → normalized to “Rainy”, “Sunny”).
- Validate date sequences for continuity (no missing days).
- Check for negative or impossible values in numeric columns.

Feature Correlation

- Correlation matrix to assess relationships among features:
  - High correlation may indicate multicollinearity. ○

    Example: Rainfall vs Humidity (correlation = 0.68).

Feature 1	Feature 2	Correlation
Temperature	Humidity	-0.32
Humidity	Rainfall	0.68

Wind Speed Rainfall 0.12

Data Quality Issues Summary

Issue Type	Description	Impact	Resolution
Missing Values	Rainfall missing in 5% records	Medium	Fill using interpolation
Outliers	Extreme rainfall values	High	Winsorization
Duplicates	2 duplicate rows found	Low	Removed
Inconsistencies	Weather column inconsistent	Medium	Standardized labels

---

Conclusion

- Dataset is now clean and ready for feature engineering and model building.
- Notes: Ensure continuous monitoring for incoming data quality in real-time forecasting.