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

Raw Data Sources and Data Quality Report

Project Name: Rainfall Prediction

Team: Lakshmi Sravya Savaram

Mohammad Shouqat Azeez

N Gokul Chowdary

Nallabotula Vijaya Karthik

Date: 13/10/2025

Raw Data Sources

Dataset	Source URL	Data Period	Geographical Scope	Features	Name
Climate Across Asia	Kaggle Link	2000–2023	Multiple Asian cities	Date, Temperature, Humidity, Wind Speed, Rainfall, Weather Conditions	

Description:

This dataset contains historical climate data across Asian regions, including daily measurements of rainfall, temperature, humidity, and wind speed. It is suitable for rainfall prediction modeling using machine learning techniques.

Dataset Overview

Attribute	Description	Data Type	Example Values	Missing Values (%)
Date	Date of observation	datetime	2023-01-01	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%
Attribute	**Description**	**Data Type**	**Example Values**	**Missing Values (%)**
Weather	Weather description	categorical	Rainy, Sunny	0%

Total Records: [Number of rows]
Total Features: [Number of columns]

Missing Values Analysis

Feature	Missing Count	Missing Percentage	Handling Method
Temperature	10	2%	Fill with mean
Humidity	5	1%	Fill with median
Rainfall	25	5%	Fill using interpolation or 0

Notes: Missing values were handled using imputation methods to maintain dataset integrity.

Duplicate Records

Total Duplicate Rows Action Taken

[Number]	Removed duplicates to maintain data quality

Outlier Analysis

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

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Rainfall	500	12.1	15.3	0	0	8	18	90

Data Consistency & Integrity Checks

- Standardized categorical values (e.g., “Rainy”, “Sunny”)
- Validated date sequences for continuity
- Checked for negative or impossible values in numeric columns

---

Feature Correlation

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

- The dataset has been cleaned, validated, and is ready for feature engineering and modeling.
- Continuous monitoring is recommended for future incoming data to maintain quality.