Project Iteration 1 Report

Muhammad Qasim 21L-5246

Syed Usman Ali 21L-5405

Section: 7B

1. Introduction

This project involves weather data analysis with the goal of predicting precipitation (rainfall). The objective is to preprocess the dataset to make it suitable for predictive modeling. The data wrangling phase is essential for identifying patterns and trends in weather conditions, which are critical for making accurate predictions regarding precipitation (rainfall). The dataset is sourced from NASA POWER, includes weather conditions recorded over time with the goal of predicting precipitation (rainfall) and contains various weather parameters such as temperature, humidity, and wind speed.

2. Dataset Overview

The dataset contains 1094 rows and 11 columns. The features include temperature, humidity, wind speed, and precipitation. The columns are as follows:

- **YEAR**: Year of observation
- **MO**: Month of observation
- **DY**: Day of observation
- **Temperature at 2 Meters (°C)**
- **Dew/Frost Point at 2 Meters (°C)**
- **Temperature at 2 Meters Maximum (C)**
- **Temperature at 2 Meters Minimum ©**
- **Specific Humidity at 2 Meters (g/kg)**
- **Relative Humidity at 2 Meters (%)**
- **Precipitation Corrected (mm/day)**

- **Wind Speed at 10 Meters (m/s)**

This dataset provides weather-related conditions which can help in predicting precipitation and understanding weather patterns.

3. Data Loading and Exploration

The data was loaded into the environment using pandas' read_csv() function. The head() and tail() methods were used to inspect the first and last few records of the dataset. Additionally, describe() and info() functions were used to summarize the dataset, providing information on the data types and summary statistics.

- **df.info():**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1094 entries, 0 to 1093
Data columns (total 11 columns):
 #   Column                                 Non-Null Count  Dtype  
---  ------                                 --------------  -----  
 0   YEAR                                   1094 non-null   int64  
 1   MO                                     1094 non-null   int64  
 2   DY                                     1094 non-null   int64  
 3   Temperature at 2 Meters ©              1094 non-null   float64
 4   Dew/Frost Point at 2 Meters (C)        1094 non-null   float64
 5   Temperature at 2 Meters Maximum (C)    1094 non-null   float64
 6   Temperature at 2 Meters Minimum ©      1094 non-null   float64
 7   Specific Humidity at 2 Meters (g/kg)   1094 non-null   float64
 8   Relative Humidity at 2 Meters (%)      1094 non-null   float64
 9   Precipitation Corrected (mm/day)       1094 non-null   float64
 10  Wind Speed at 10 Meters                1094 non-null   float64
dtypes: float64(8), int64(3)
memory usage: 94.1 KB
```

- **df.describe():**

	YEAR	MO	DY	Temperature at 2 Meters ©	Dew/Frost Point at 2 Meters (C)	Temperature at 2 Meters Maximum (C)	Temperature at 2 Meters Minimum ©	Specific Humidity at 2 Meters (g/kg)	Relative Humidity at 2 Meters (%)	Precipitation Corrected (mm/day)	Wind Speed at 10 Meters
count	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000
mean	2021.999086	6.521024	15.706581	24.815777	12.788940	31.724680	18.856243	10.728995	53.420064	2.590137	2.306892
std	0.816683	3.447026	8.792141	8.070966	8.224657	7.615615	8.061764	5.761845	18.171100	7.764667	0.793325
min	2021.000000	1.000000	1.000000	7.540000	-5.410000	13.830000	0.320000	2.560000	9.500000	0.000000	0.570000
25%	2021.000000	4.000000	8.000000	17.562500	5.900000	25.650000	11.922500	5.920000	40.440000	0.000000	1.730000
50%	2022.000000	7.000000	16.000000	26.280000	11.535000	32.655000	19.355000	8.670000	54.155000	0.000000	2.170000
75%	2023.000000	10.000000	23.000000	31.050000	20.940000	37.047500	25.975000	15.975000	67.690000	1.197500	2.760000
max	2023.000000	12.000000	31.000000	40.560000	27.190000	47.820000	33.510000	23.250000	95.060000	96.170000	7.020000

- **df.head():**

	YEAR	MO	DY	Temperature at 2 Meters ©	Dew/Frost Point at 2 Meters (C)	Temperature at 2 Meters Maximum (C)	Temperature at 2 Meters Minimum ©	Specific Humidity at 2 Meters (g/kg)	Relative Humidity at 2 Meters (%)	Precipitation Corrected (mm/day)	Wind Speed at 10 Meters
0	2021	1	1	11.09	-1.05	18.75	5.82	3.54	45.00	0.00	1.43
1	2021	1	2	10.99	1.71	18.87	5.91	4.52	56.12	0.41	1.82
2	2021	1	3	12.78	10.72	17.98	8.79	8.18	87.69	1.54	2.55
3	2021	1	4	14.45	13.18	18.49	11.69	9.58	92.12	3.36	2.38
4	2021	1	5	14.19	13.42	17.01	11.78	9.70	95.06	32.72	3.84

- **df.tail():**

	YEAR	MO	DY	Temperature at 2 Meters ©	Dew/Frost Point at 2 Meters (C)	Temperature at 2 Meters Maximum (C)	Temperature at 2 Meters Minimum ©	Specific Humidity at 2 Meters (g/kg)	Relative Humidity at 2 Meters (%)	Precipitation Corrected (mm/day)	Wind Speed at 10 Meters
1089	2023	12	26	15.95	4.68	23.08	9.92	5.37	48.56	0.0	1.18
1090	2023	12	27	14.60	4.86	21.62	9.49	5.43	53.31	0.0	1.45
1091	2023	12	28	13.03	4.43	21.74	7.23	5.25	58.12	0.0	2.27
1092	2023	12	29	13.35	4.33	20.53	7.79	5.25	56.56	0.0	1.25
1093	2023	12	30	12.95	3.88	19.33	8.33	5.07	55.50	0.0	1.09

4. Data Cleaning

- **Missing Data**: There were no missing values in the dataset, as confirmed using df.isna().sum().

```
df.isna().sum()
```

	0
YEAR	0
MO	0
DY	0
Temperature at 2 Meters ©	0
Dew/Frost Point at 2 Meters (C)	0
Temperature at 2 Meters Maximum (C)	0
Temperature at 2 Meters Minimum ©	0
Specific Humidity at 2 Meters (g/kg)	0
Relative Humidity at 2 Meters (%)	0
Precipitation Corrected (mm/day)	0
Wind Speed at 10 Meters	0

- **Duplicate Removal**: No duplicate records were found in the dataset as the shape remains same before and after dropping duplicates.
- **Data Type Conversion**: The YEAR, MO, and DY columns were combined into a single Date column for time-based analysis if needed to examine the pattern over some time period.
- **Outlier Detection**: As initially all columns' data types were numerical so their skewness is checked first. After checking the skewness, if it is 0 then outliers are detected and removed using Z-score distribution otherwise IQR method is

used. For the time being, outliers are found in the target variable (Precipitation Corrected) and Wind Speed at 10 Meters column. Outliers are removed for the Wind Speed column but for Precipitation further analysis e.g. Model Performance etc. will determine whether to remove these outliers as it is not being removed now as they represent valid, meaningful points i.e. rare but possible.

5. Data Transformation

- **Scaling**: Standardization (Z-score normalization) was applied to the numerical features. This technique was chosen because the dataset contains variables with different scales, and standardization ensures that the model treats all features consistently. In addition to that, standardization is also not affected by the outliers