WEATHER DATA ANALYSIS & TEMPERATURE PREDICTION ML PROJECT

Presented by : Mariam Ghanim

CHALLENGE:

Accurate temperature prediction is crucial for various applications but remains complex due to multiple interacting weather variables.

Why This Matters

Agriculture: Crop planning and frost protection

Energy Management: Heating/cooling demand forecasting

Transportation: Safety planning and route optimization

Daily Life: Personal planning and comfort decisions

Emergency Services: Preparation for extreme weather conditions

Our Goal:

Build a robust, accurate ML model that can predict temperature using readily available weather parameters, and deploy it in an accessible web application.

Key Components

1. **Data Cleaning**
2. **Exploratory Data Analysis**
3. **Feature Engineering & Encoding**
4. **Modeling, Tuning & Results**

Data Cleaning Steps

1 Duplicate Removal

```
Number of duplicate rows: 24
Removed 24 duplicate rows

Shape after removing duplicates: (96429, 12)
```

2 Handling Missing Values

```
Missing Precip Type: 0.54% of data

Since missing values are less than 1%, we can safely drop them.
New shape after dropping missing values: (95912, 12)
```

3 Convert Formatted Date from string to datetime

```
Before conversion:
Data type: object
Sample value: 2006-04-01 00:00:00.000 +0200

================================================
After conversion:
Data type: datetime64[ns, UTC]
Sample value: 2006-03-31 22:00:00+00:00
```

4 Dropping loud cover as it has only one value (only zero)

```
New shape of cleaned dataframe:

(95912, 11)
```

5 Convert categorical columns to 'category' type for Less Memory Usage

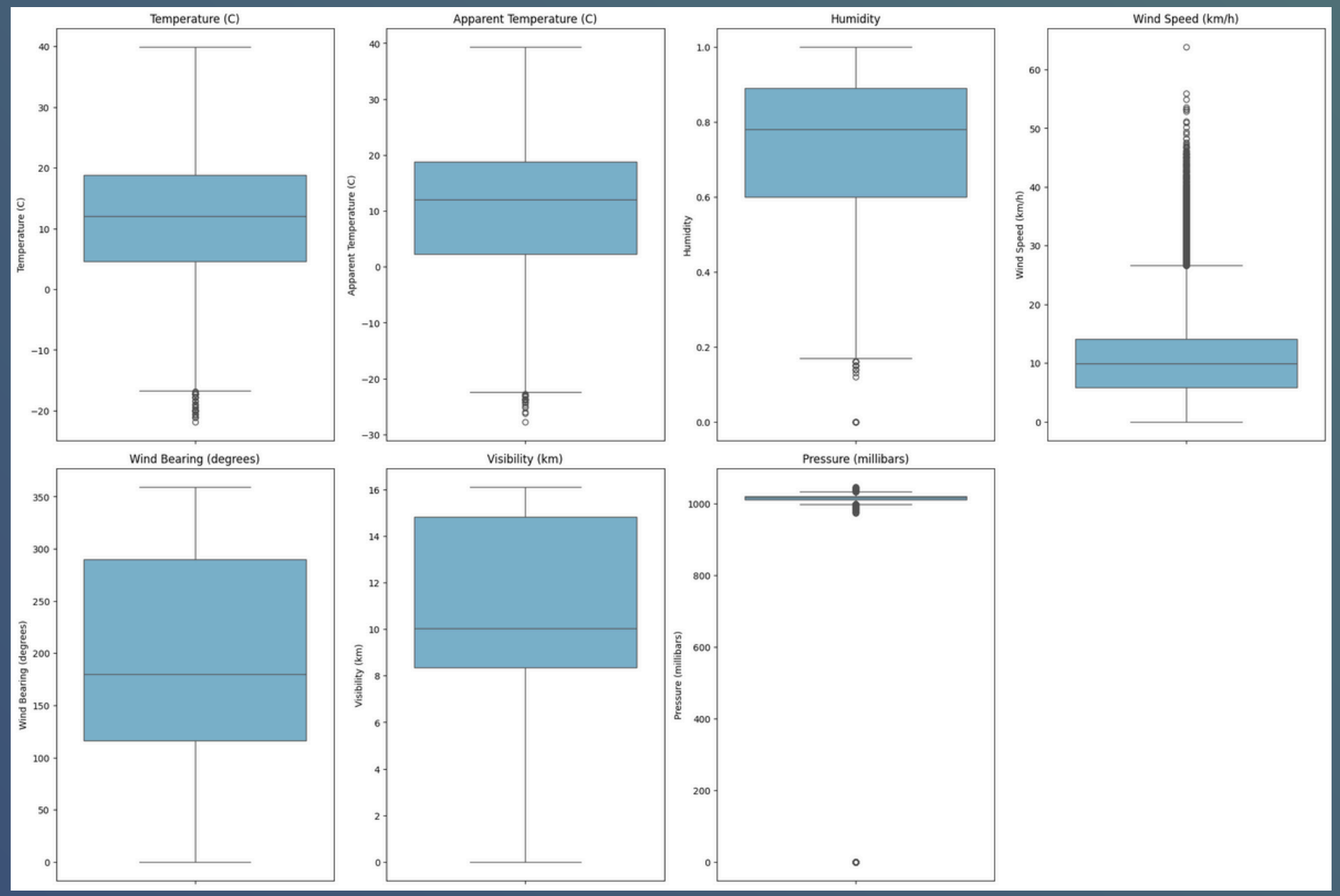
```
Converting categorical columns to 'category' t
Summary: category
Precip Type: category
Daily Summary: category
```

Exploratory Data Analysis

1 Univariate analysis

I. Checking for outliers through boxplot

Observation: Outliers are present in Temperature, Apparent Temperature, Humidity, and Wind Speed.

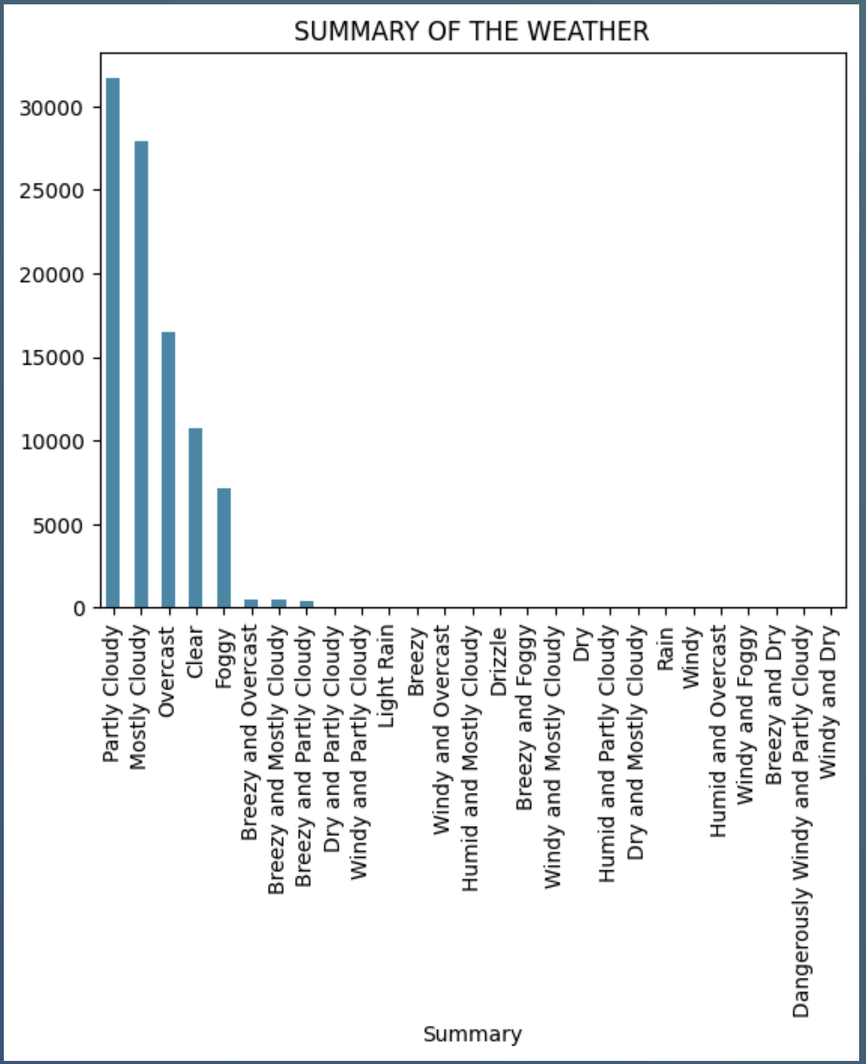


Exploratory Data Analysis

1 Univariate analysis

II. Explore Summary Column

Observation: In the given dataset, most of the days are partly cloudy, followed by mostly cloudy, overcast, and foggy

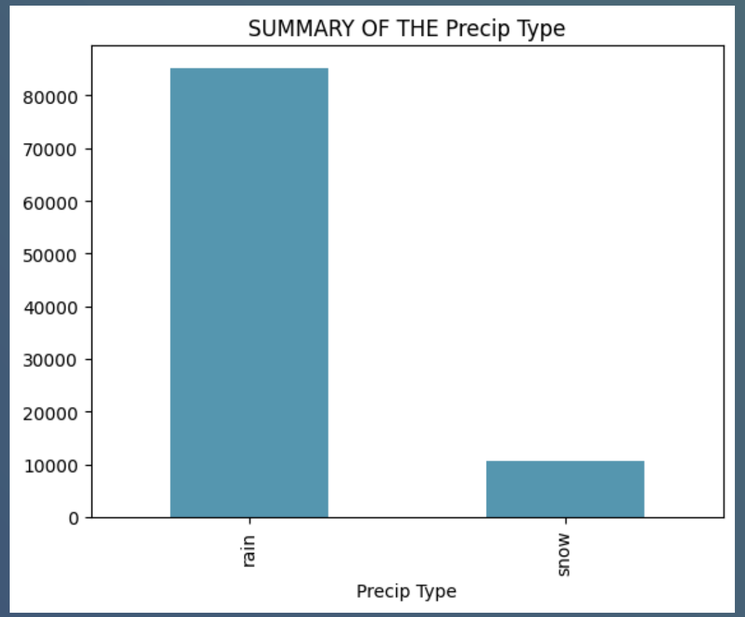


Exploratory Data Analysis

1 Univariate analysis

III. Checking precipitation type

Observation: The given dataset has only two kinds of precipitation: rain and snow. Among these, most of the days experienced rain.



III. Note on 'Daily Summary' Column

- The 'Daily Summary' column is more like a description of the day.
- Proper analysis would require NLP techniques.
- Therefore, this column is dropped from the dataframe.

```
Daily Summary
Mostly cloudy throughout the day.          20020
Partly cloudy throughout the day.           9930
Partly cloudy until night.                  6169
Partly cloudy starting in the morning.      5177
Foggy in the morning.                       4201
                                            ...
Rain until afternoon.                         17
Rain until morning.                           12
Light rain in the morning.                    11
Drizzle starting in the evening.               9
Light rain overnight.                          3
Name: count, Length: 214, dtype: int64
```

Exploratory Data Analysis

2 Bivariate analysis

I. Summary vs Precipitation Type

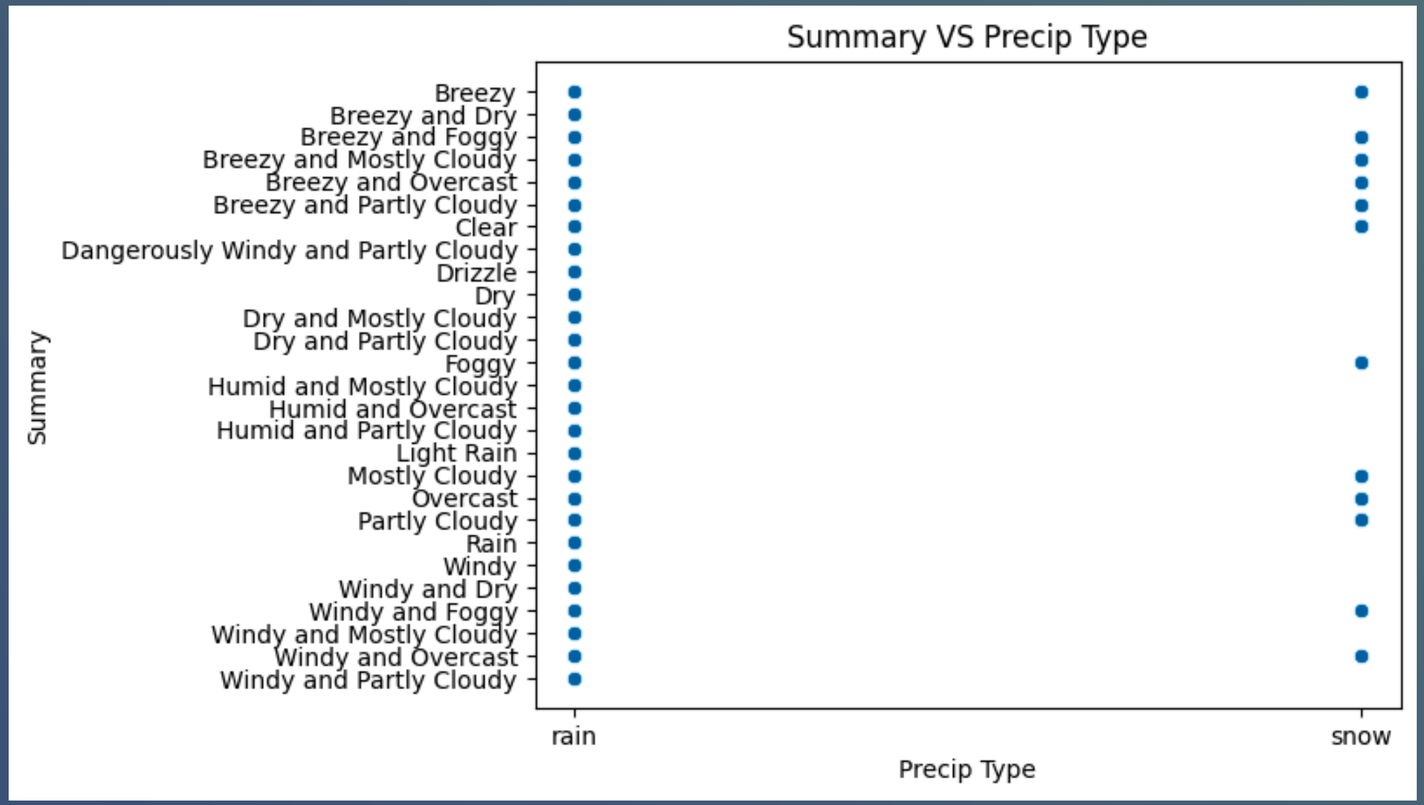
Observation:

Rain:

- Can occur under all types of weather in the dataset.

Snow:

- Only observed under specific conditions: cloudy, foggy, or windy days.



Exploratory Data Analysis

2 Bivariate analysis

II. Snow vs Rain

Observation:

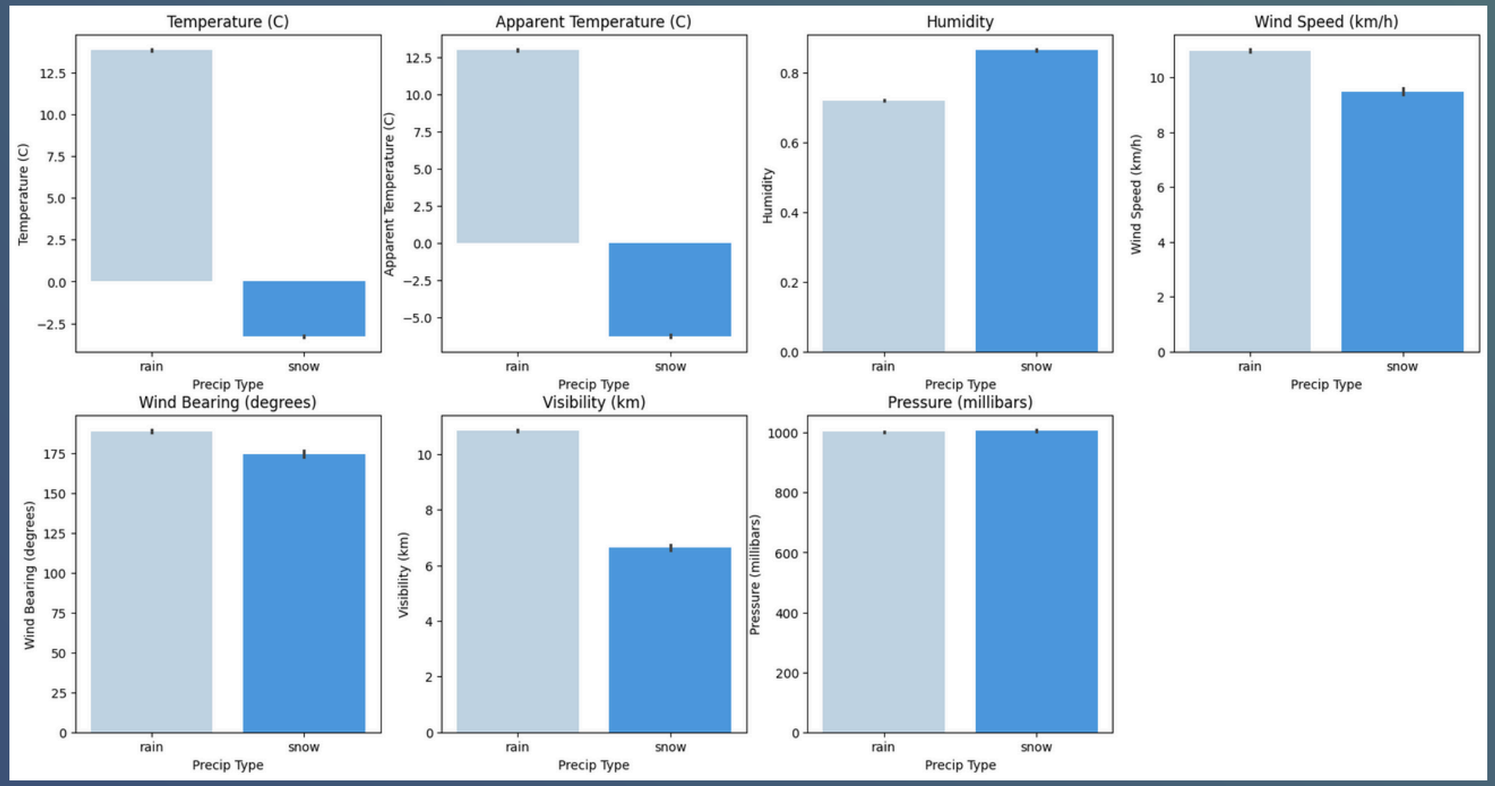
- Temperature & Apparent Temperature:
- Lower on snowy days compared to rainy days.

- Average Humidity:
- Higher on snowy days.

- Average Wind Speed:
- Lower on snowy days compared to rainy days.

- Average Visibility:
- Lower on snowy days compared to rainy days.

- Average Pressure:
- No significant difference between snowy and rainy days.



Feature Engineering and Encoding Steps

Feature Engineering

DateTime Feature

Created Year, Month, Day, Hour, WeekOfYear

Temporal Feature Derivation

Derived Season and Time of Day from date and time.

Feature Categorization

Categorized key variables (Temperature, Wind Speed, Humidity, Pressure, Visibility).

Encoding

Ordinal Encoding

Applied ordinal encoding for ordered categories.

One-hot encoding

Used one-hot encoding for nominal features (Precip Type, Summary).

Feature Selection & Model Optimization

Feature Selection (Filter Method)

- Applied SelectKBest with f-regression
- Reduced features from 56 → top 12 most impactful
- Improved model efficiency and reduced noise

Temp_Range_Indicator_encoded
Humidity
Comfort_Index_encoded
These features showed the strongest relationship with temperature

Hyperparameter Tuning Strategy

GridSearchCV applied across:

- Linear Regression
- Random Forest
- LightGBM

3-Fold Cross-Validation

- Ensured model stability
- Reduced risk of overfitting

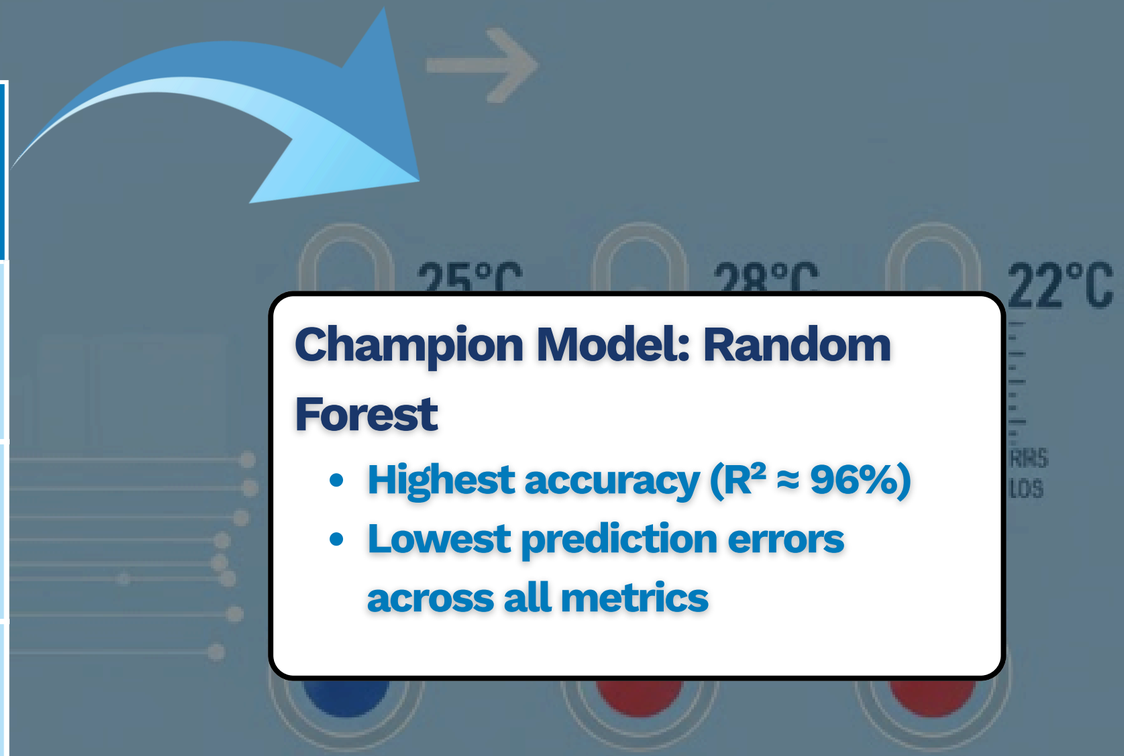
StandardScaler used for feature normalization

Final Results & Model Insights

Objective

Compare model performance and select the best model for deployment

Model	MAE	RMSE	R^2 Score
Random Forest	1.4469	1.9055	**0.9604**
LightGBM	1.6079	2.0219	0.9554
Linear Regression	2.2607	2.7039	0.9203



Champion Model: Random Forest

- **Highest accuracy ($R^2 \approx 96\%$)**
- **Lowest prediction errors across all metrics**