Weather Trend Forecasting

Howard Mach

PM Accelerator Mission

By making industry-leading tools and education available to individuals from all backgrounds, **we level the playing field for future PM leaders.** This is the PM Accelerator motto, as we grant aspiring and experienced PMs what they need most – Access. We introduce you to industry leaders, **surround you with the right PM ecosystem**, and discover the new world of AI product management skills.

Objective

Accurately predicting the perceived temperature accounting for factors like humidity, wind speed, and atmospheric pressure is essential for enhancing real-world weather insights and improving public comfort advisories. Unlike raw temperature measurements, the "feels like" metric reflects the actual thermal sensation experienced by individuals, making it especially valuable for outdoor planning, health advisories, and resource management. By building a forecasting model that integrates both meteorological data and time-derived features from chronologically sorted observations, we aim to deliver timely, context-aware predictions that better capture human temperature perception and offer a more practical lens for weather-related decision-making.

Methodology Overview

Preprocessing

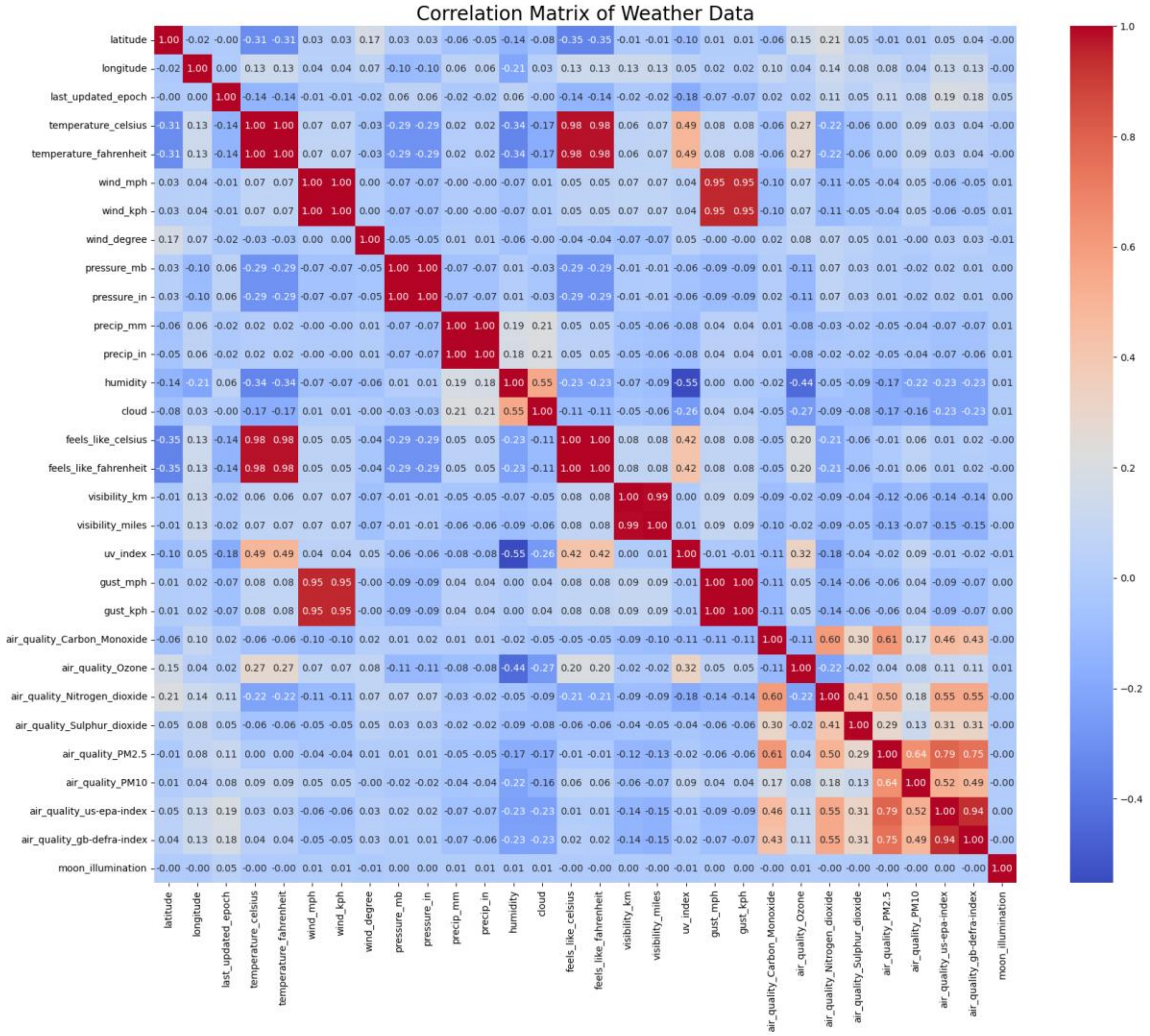
Before model development, the dataset was carefully examined for missing values to ensure data completeness and reliability. This initial quality assessment enabled informed decisions around data cleaning and helped prevent issues during training and evaluation.

The timestamp column, *last_updated*, was then converted to a proper datetime format, which allowed the dataset to be sorted chronologically. This step ensured that temporal consistency was maintained, especially for time-aware feature extraction and forecasting tasks.

Feature Selection

To identify the most informative predictors for forecasting the perceived temperature (*feels_like_celsius*), a strategic feature selection process was undertaken that combined domain expertise with data-driven analysis. The target variable represents perceived temperature, which is influenced by both weather conditions and temporal dynamics. Therefore, a mix of meteorological and time-derived features were considered.

A correlation heatmap was generated to assess the relationships between numerical variables within the dataset, serving as a visual tool for uncovering patterns of linear association. This method was selected for its clarity and efficiency in highlighting both strongly related features and potential issues like multicollinearity.

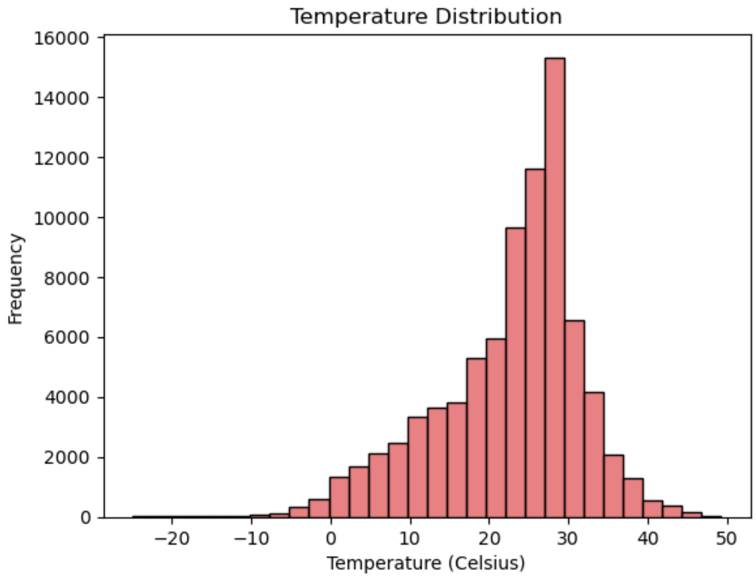


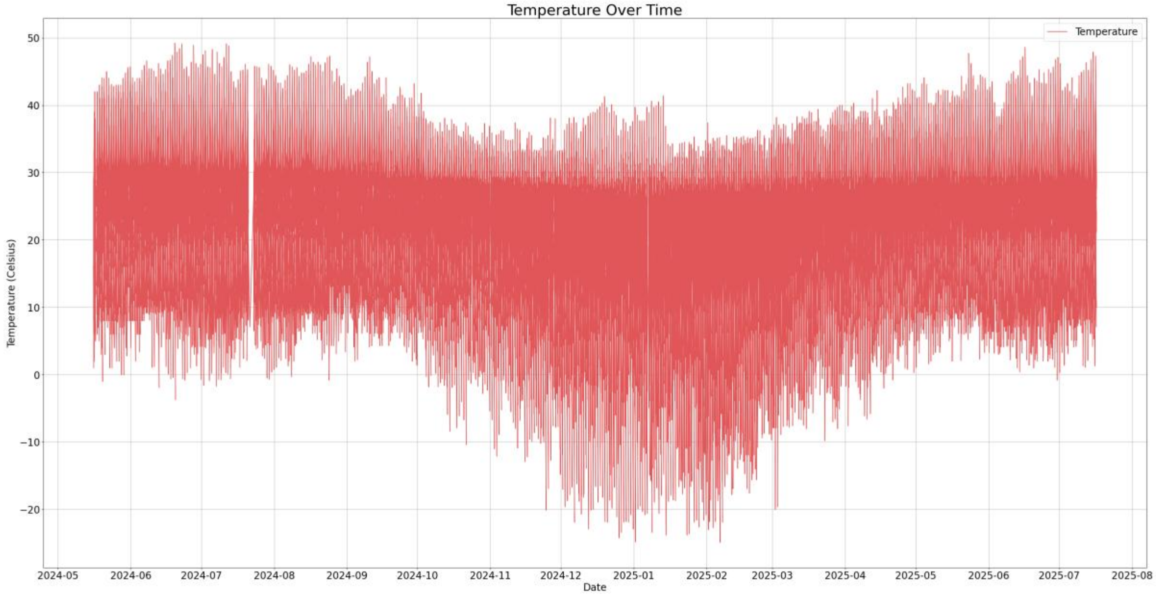
Based on correlation insights from the heatmap, the following predictors were chosen:

Predictor	**Rationale**
temperature_celsius	Strong correlation with target
humidity	Influences perceived temperature
wind_kph	Affects air movement & perception
pressure_mb	Related to broader weather patterns
precip_mm	Indicates moisture and cooling
cloud	Impacts sunlight and thermal feel

Predictor	**Rationale**
hour_sin & hour_cos	Capture cyclical daily weather patterns
dayofyear	Seasonal variation

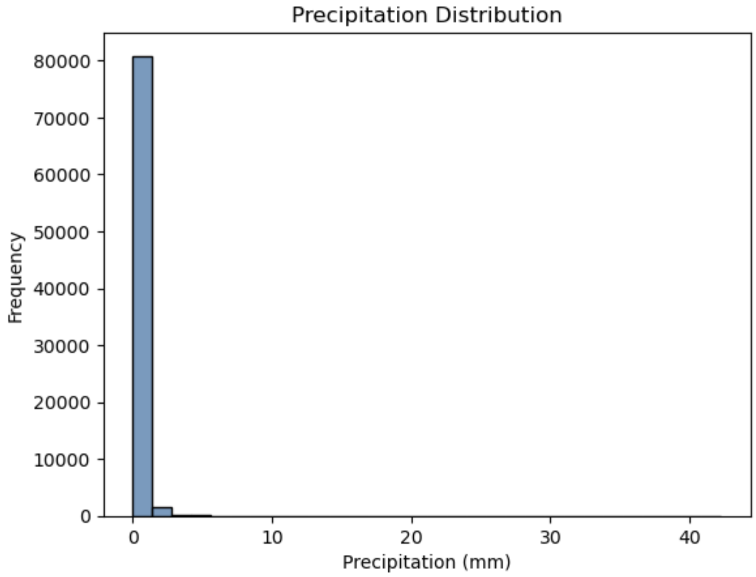
Building on this curated feature set, exploratory data analysis (EDA) visualizations were generated to further assess the distribution, temporal trends, and inter-variable relationships of each predictor—offering deeper insights into their behavior and relevance prior to model development. To further illustrate these insights, the following visualizations examine each predictor individually—revealing patterns, anomalies, and seasonality that inform both the modeling strategy and the interpretation of forecast results.

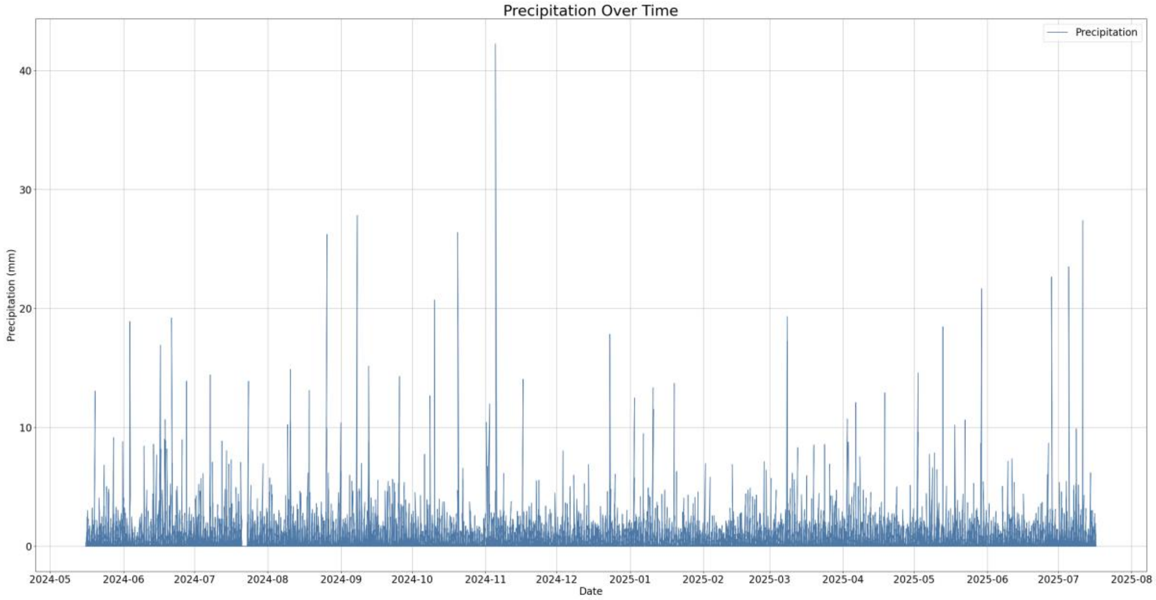




The first graph, a histogram of temperature readings in Celsius, shows a distribution clustered primarily between **10°C and 40°C**, with a clear peak around **30°C**—suggesting that moderate to warm conditions dominate the dataset. There's a slight right skew, indicating relatively fewer high-temperature occurrences beyond 40°C and minimal cold extremes below 0°C.

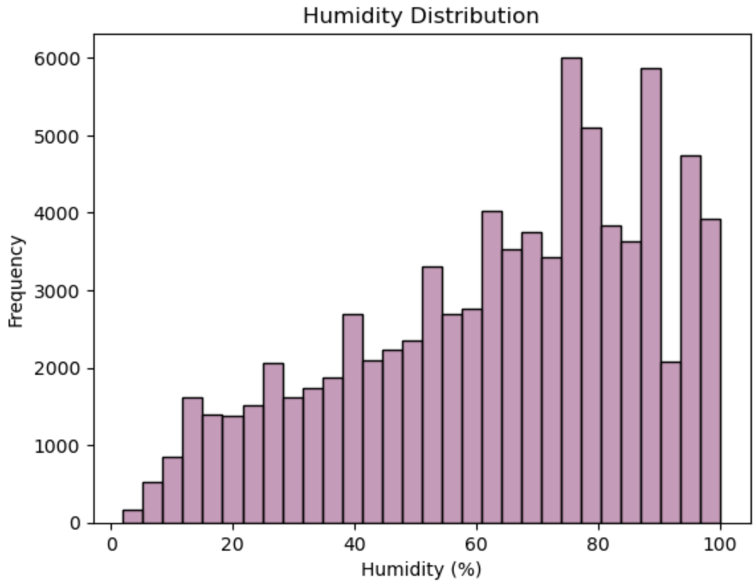
The second graph presents a **temperature time series** from **June 2022 to August 2025**, displaying clear **seasonal fluctuations**. Temperatures rise sharply during summer months, often exceeding 40°C, and dip below freezing during winter, reflecting expected cyclical climate patterns. This trend underscores the importance of incorporating **seasonal and cyclical time features** into the forecasting model to effectively capture temporal variability in temperature behavior

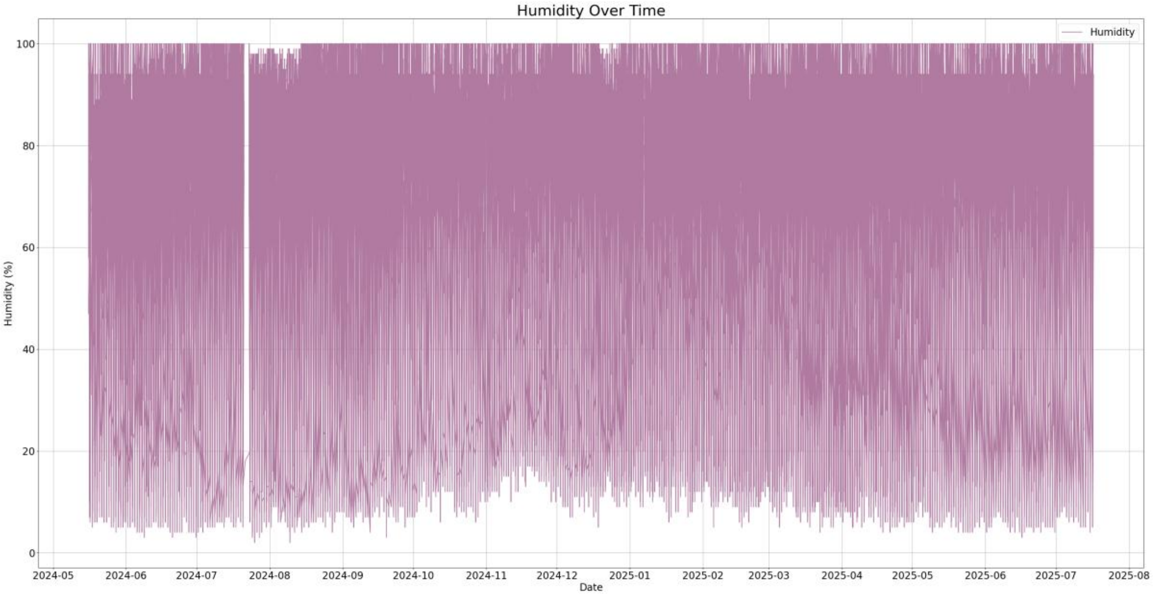




The first graph, titled **Precipitation Distribution**, shows a sharply right-skewed histogram where most of the observations fall at or near **0 mm**. This suggests that most recorded days experienced little to no precipitation, with relatively rare instances of moderate to heavy rainfall reaching above **40 mm**.

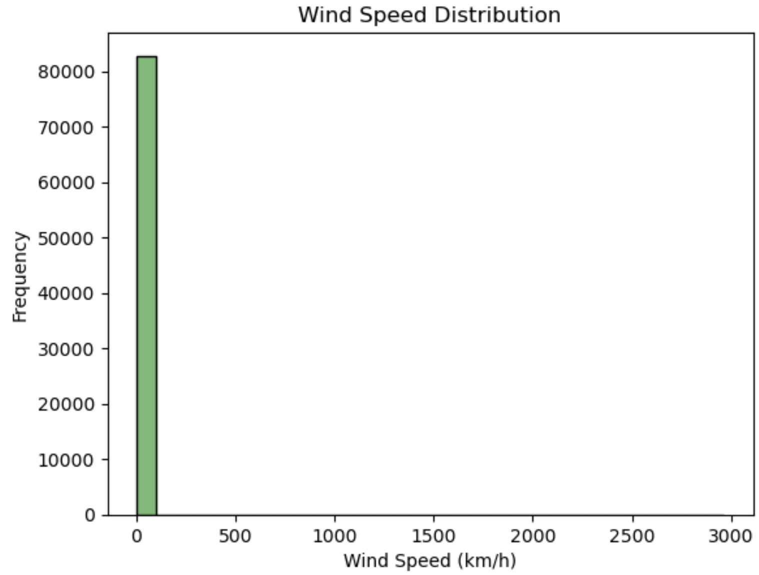
The second graph, **Precipitation Over Time**, reveals sporadic but significant spikes in rainfall between **May 2024 and August 2025**, with intermittent periods of dryness in between. These peaks highlight weather events that may correspond to seasonal patterns or storm occurrences, making this feature valuable for understanding the timing and intensity of rainfall when forecasting perceived temperature or broader weather conditions.

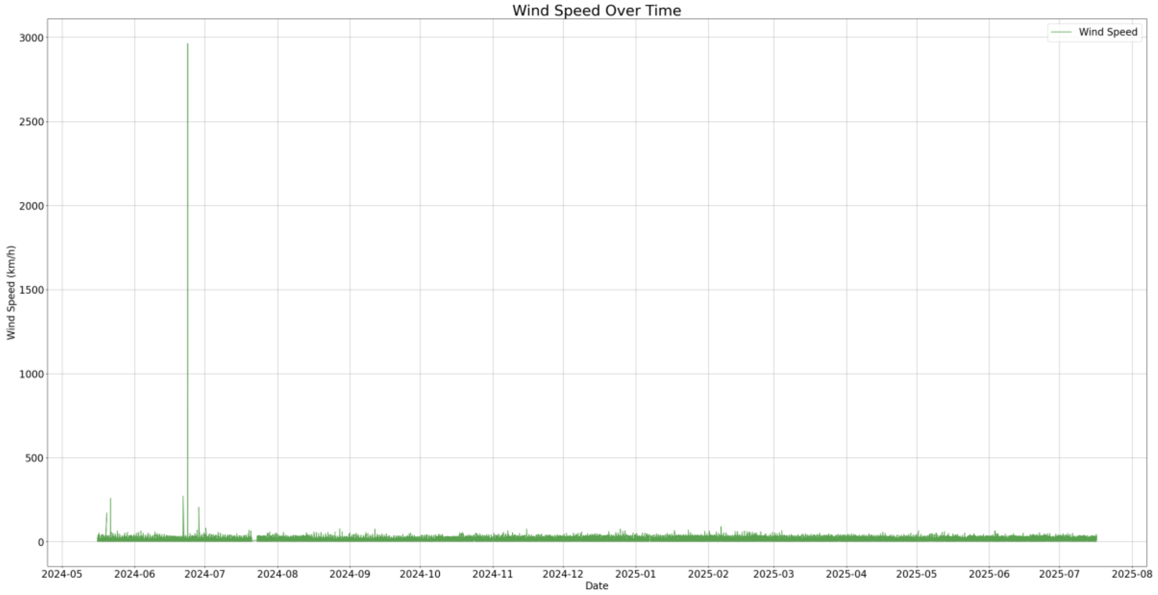




The first graph, **Humidity Distribution**, displays a clear concentration of humidity values between **60% and 100%**, with noticeable peaks around **80% and 90%**. This suggests that the dataset frequently captures relatively humid conditions, which may amplify the perceived temperature and impact comfort levels, making it a meaningful predictor in the model.

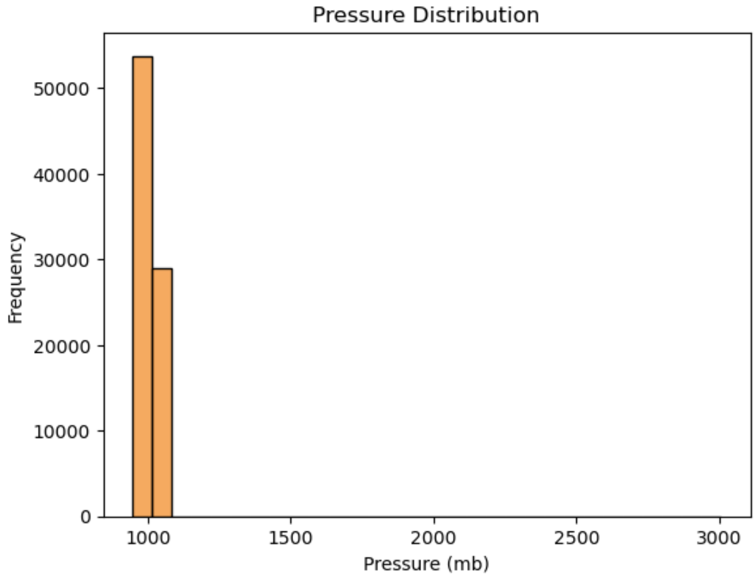
The second graph, **Humidity Over Time**, shows considerable variation with humidity often swinging between very low and very high values. These fluctuations appear consistently throughout the timeline from **May 2024 to August 2025**, with a visible gap in the data around **July 2024**. This kind of temporal volatility highlights the importance of time-aware features when modeling how humidity influences perceived temperature.

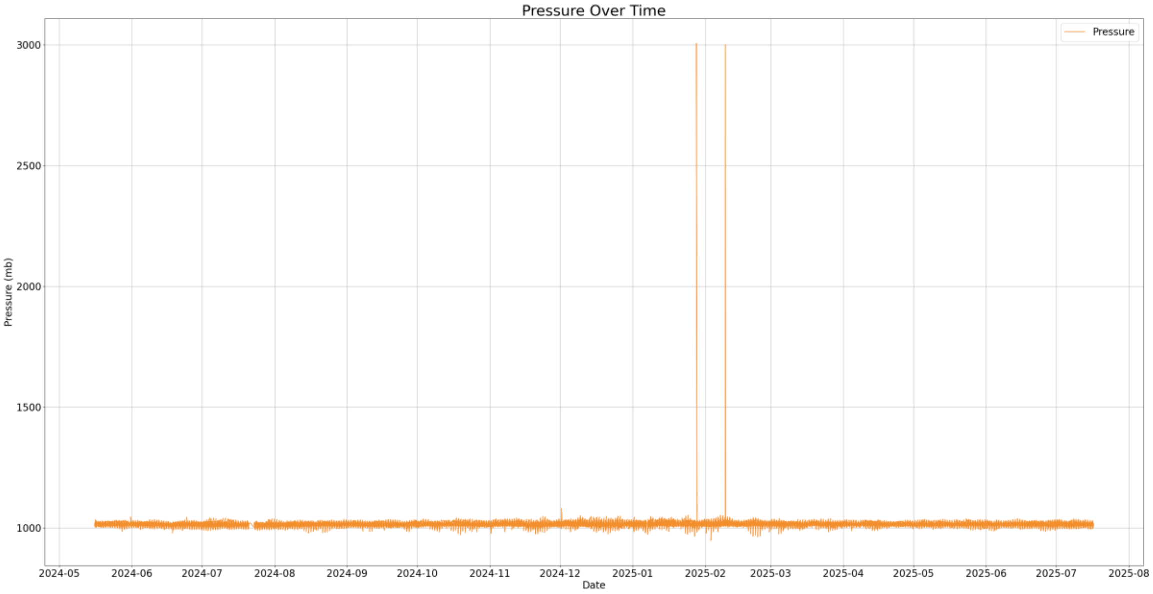




The first graph, **Wind Speed Distribution**, shows an extreme right-skewed histogram where nearly all values cluster around **0 km/h**, with an exceptionally rare spike near **3000 km/h**. This outlier suggests a potential data anomaly or sensor error, as such speeds are not physically plausible under normal atmospheric conditions.

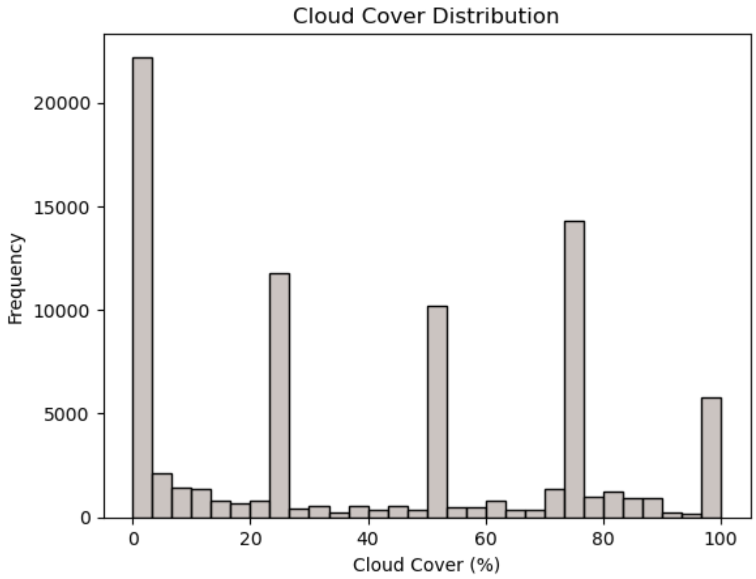
The second graph, **Wind Speed Over Time**, confirms this anomaly visually—there's a dramatic spike around **June 2024**, while the rest of the timeline shows consistently low wind speeds. This reinforces the idea that the dataset may contain erroneous or outlier values that should be investigated and possibly removed before modeling, as they could skew learning and predictions.

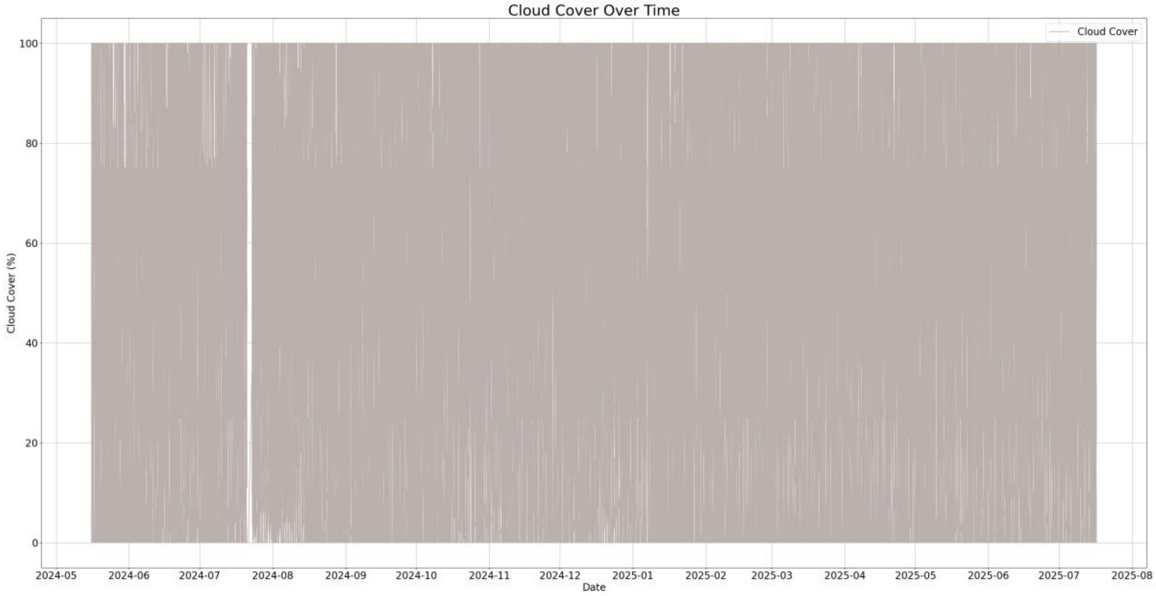




The first graph, **Pressure Distribution**, shows that most atmospheric pressure readings cluster tightly around **1000 mb**, which aligns with typical sea-level pressure conditions. The distribution is narrow and bell-shaped, indicating a stable and consistent range of pressure measurements in the dataset. However, there is a sparse tail extending toward higher values, hinting at a few anomalously high readings.

The second graph, **Pressure Over Time**, reflects this stability visually—with pressure remaining consistently near 1000 mb throughout the observed period from **May 2024 to August 2025**. Notably, there are two prominent spikes in early **2025** where pressure values briefly surge close to **3000 mb**, which suggests either sensor error or outlier events that warrant further investigation before modeling. These anomalies could distort regression outcomes and are good candidates for removal or correction during data cleaning.





The first graph, **Cloud Cover Distribution**, displays a distinct multimodal histogram with prominent peaks at **0%, 20%, 40%, 60%, 80%, and 100%**. This pattern suggests that cloud cover percentages were recorded in discrete intervals, possibly rounded or bucketed during collection. The repeated spacing indicates frequent classification at standard visibility thresholds—useful for modeling categorical or ordinal cloud behavior.

The second graph, **Cloud Cover Over Time**, shows a dense scatter of values between **May 2024 and August 2025**, capturing high variability across seasons and daily weather changes. Cloud cover percentages fluctuate continuously, reflecting realistic shifts between clear skies and full overcast. This variability reinforces its role as a dynamic predictor of perceived temperature and supports the inclusion of cloud data in the forecasting model.

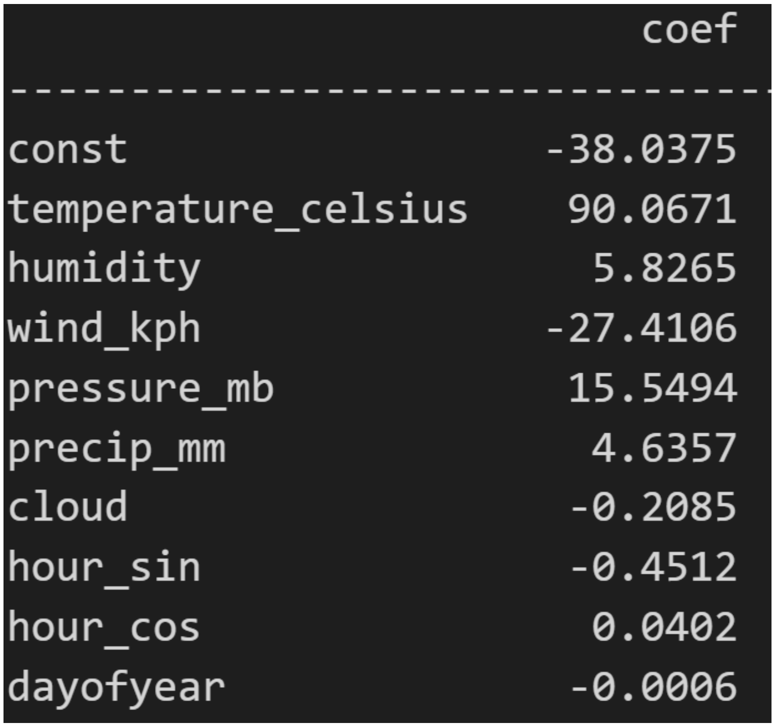
Modeling & Evaluation

For model training and evaluation, the dataset was split chronologically to preserve temporal integrity, allocating **75%** of the data for **training** and the remaining **25%** for **testing**. This approach ensures that the model learns from past conditions and predicts forward-looking scenarios without temporal leakage.

A **linear regression model** was employed to forecast *feels_like_celsius*, leveraging both environmental and time-derived features. This choice offers interpretability and serves as a reliable baseline for understanding linear relationships within the dataset.

Model performance was assessed using three standard metrics: **Mean Absolute Error (MAE)** to measure average prediction error, **Root Mean Squared Error (RMSE)** to penalize larger deviations, and **R² Score** to quantify the proportion of variance explained by the model. These evaluations provide insight into both accuracy and generalization capability, forming the foundation for future model comparison and refinement.

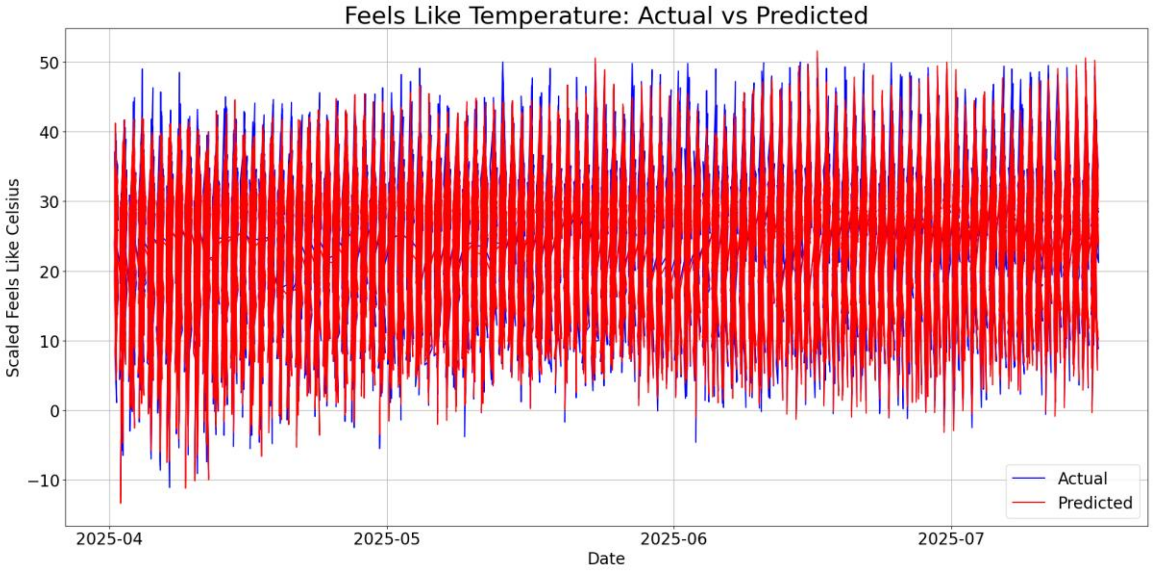
The coefficients of the model are:



```
                          coef
-------------------------------
const                 -38.0375
temperature_celsius    90.0671
humidity                5.8265
wind_kph              -27.4106
pressure_mb            15.5494
precip_mm               4.6357
cloud                  -0.2085
hour_sin               -0.4512
hour_cos                0.0402
dayofyear              -0.0006
```

Results

Metric	Value
MAE	*1.242*
RMSE	*1.733*
R²	*0.964*



The error metrics indicate that the forecasting model performs with impressive accuracy. The **Mean Absolute Error (MAE)** of **1.242°C** suggests that, on average, the model's predictions deviate from actual values by just over one degree—an acceptably small margin for temperature forecasting. The **Root Mean Squared Error (RMSE)**, calculated at **1.733°C**, reflects slightly higher typical deviations, penalizing larger errors more heavily. These figures together suggest that the model captures most variations effectively, with minimal outliers or extreme misses. Most notably, the **Coefficient of Determination (R²)** value of **0.964** demonstrates that the model explains 96.4% of the variance in the perceived temperature, validating its reliability and robustness for real-world applications.

The visualization supports these quantitative results with compelling clarity. The time series plot shows the **actual values (blue)** and **predicted values (red)** closely aligned across the test period, especially from **April to July 2025**. The red prediction curve tracks seasonal temperature rises and dips with consistent precision, rarely diverging from the actual data line. There's no significant lag, overfitting, or flat-line behavior, which confirms that the model appropriately responds to cyclical weather shifts and dynamic input signals.

F-statistic: 3.397e+05

p-value: 0.0

The extremely high **F-statistic** value of **339,700** indicates that the overall regression model has strong explanatory power, meaning the collective set of predictors significantly contributes to forecasting *feels_like_celsius*. The accompanying **p-value of 0.0** confirms this statistical significance, suggesting that there is virtually no likelihood the observed relationship between the predictors and the target variable occurred by chance.

Altogether, the error analysis and visual validation affirm that the model is well-calibrated for short-term temperature perception forecasting. Minor deviations, where present, appear localized and do not distort overall performance. Future improvements may focus on expanding the forecast horizon, validating the model on unseen data from other seasons, or experimenting with nonlinear algorithms like ensemble trees or support vector regressors to further enhance predictive capability.

Key Takeaways

- Time-derived features and normalization significantly improved data consistency.
- Predictor selection based on heatmap insights created a strong baseline model.
- Visualization and metrics confirm reasonable accuracy, with room for tuning or deeper models.

The forecasting model has proven effective in capturing both environmental and temporal patterns that influence perceived temperature. Its success suggests strong potential for expanding into more complex domains, such as integrating air quality indices, UV exposure levels, or wind chill to further enhance real-world usability.

Looking ahead, incorporating additional data streams like satellite imagery or probabilistic weather forecasts could offer richer context, helping the model adapt to more nuanced climate conditions and improving its responsiveness during extreme weather events.