Weather Forecasting

Objective

To analyze global weather data and build forecasting models to predict temperature trends, using traditional time-series methods.

Dataset Overview

Name:	GlobalWeatherRepository.csv
Total Records:	62558
Total Columns:	41

	Column Name	Data Type
0	country	object
1	location_name	object
2	latitude	float64
3	longitude	float64
4	timezone	object
5	last_updated_epoch	int64
6	last_updated	object
7	temperature_celsius	float64
8	temperature_fahrenheit	float64
9	condition_text	object
10	wind_mph	float64
11	wind_kph	float64
12	wind_degree	int64
13	wind_direction	object
14	pressure_mb	float64
15	pressure_in	float64
16	precip_mm	float64
17	precip_in	float64
18	humidity	int64
19	cloud	int64
20	feels_like_celsius	float64
21	feels_like_fahrenheit	float64

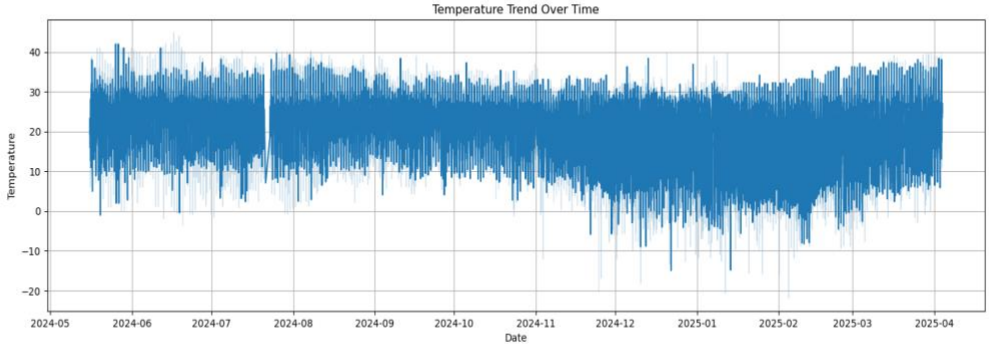
	Column Name	Data Type
22	visibility_km	float64
23	visibility_miles	float64
24	uv_index	float64
25	gust_mph	float64
26	gust_kph	float64
27	air_quality_Carbon_Monoxide	float64
28	air_quality_Ozone	float64
29	air_quality_Nitrogen_dioxide	float64
30	air_quality_Sulphur_dioxide	float64
31	air_quality_PM2.5	float64
32	air_quality_PM10	float64
33	air_quality_us-epa-index	int64
34	air_quality_gb-defra-index	int64
35	sunrise	object
36	sunset	object
37	moonrise	object
38	moonset	object
39	moon_phase	object
40	moon_illumination	int64

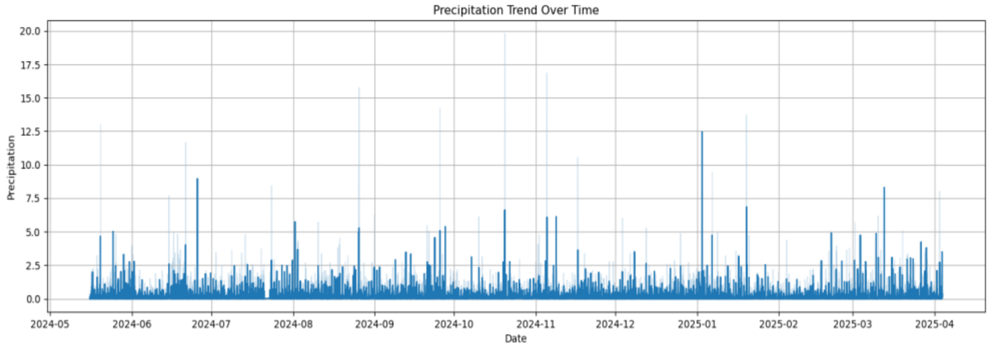
Data Cleaning & Preprocessing

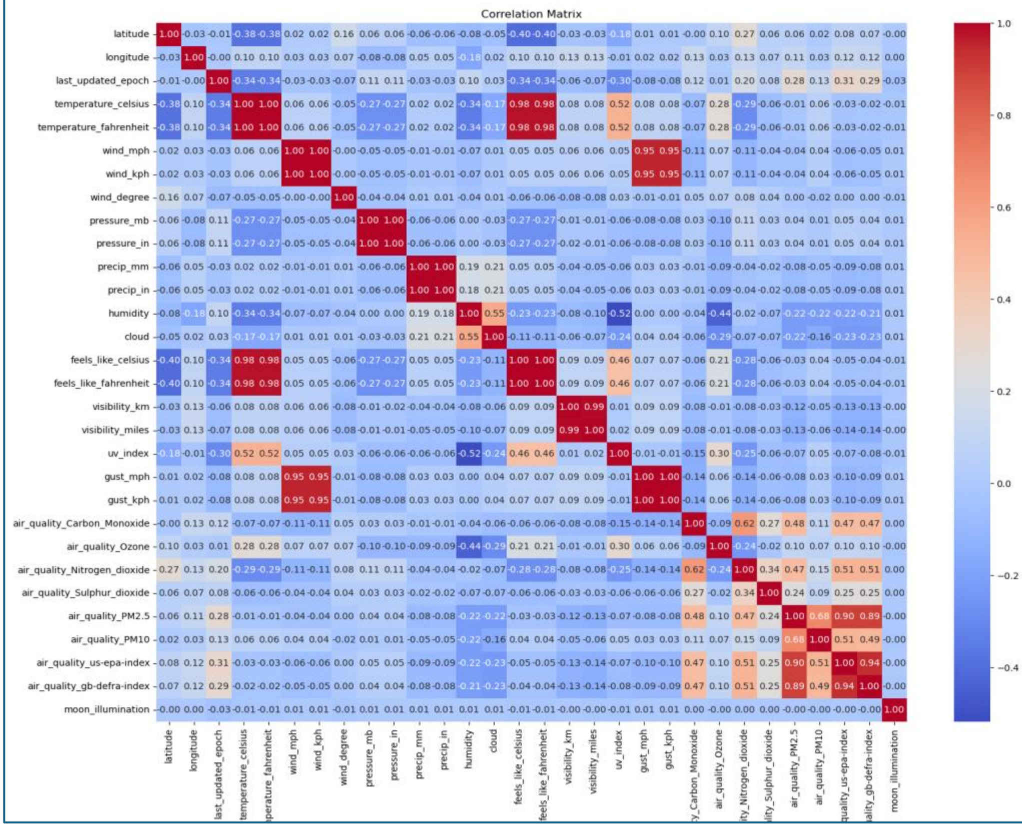
- No missing values are present in the dataset provided.
- The ‘IsolationForest’ algorithm is used to detect and remove outliers from numerical columns in the dataset.
- It assigns an "outlier" label (-1) to anomalous rows based on a contamination rate of 1% and filters them out, leaving only normal data points (1).
- The "outlier" column is then dropped to clean up the dataset.

Exploratory Data Analysis

- Analyzed correlations between various features like temperature, humidity, wind, precipitation, pressure etc.
- Visualized trends in temperature and precipitation over time
- Visualized weather differences across countries


Temperature Trend Over Time
Temperature
40
30
20
10
0
−10
−20
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
2025-03
2025-04
Date


Precipitation Trend Over Time
Precipitation
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
2025-03
2025-04
Date


Correlation Matrix
latitude
longitude
last_updated_epoch
temperature_celsius
temperature_fahrenheit
wind_mph
wind_kph
wind_degree
pressure_mb
pressure_in
precip_mm
precip_in
humidity
cloud
feels_like_celsius
feels_like_fahrenheit
visibility_km
visibility_miles
uv_index
gust_mph
gust_kph
air_quality_Carbon_Monoxide
air_quality_Ozone
air_quality_Nitrogen_dioxide
air_quality_Sulphur_dioxide
air_quality_PM2.5
air_quality_PM10
air_quality_us-epa-index
air_quality_gb-defra-index
moon_illumination
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Forecasting with Multiple Models

Evaluation:

Model	MAE	MSE	RMSE	R^2 Score
ARIMA	6.9863	73.4687	8.5714	-0.0367
Facebook Prophet	6.5294	75.1736	8.6703	-0.0607
SARIMAX	9.9575	164.5829	12.8290	N/A
XGBoost Regressor	8.7602	96.4141	9.8191	-0.3604
Fine-Tuned XGBoost	0.0861	0.1770	0.4207	0.9983

Inference:

1. **ARIMA**:
   - Moderate performance with MAE of 6.9863 and RMSE of 8.5714.
   - Negative R^2 (-0.0367) indicates the model does not explain the variance in the data well, suggesting it struggles with capturing complex temporal patterns.

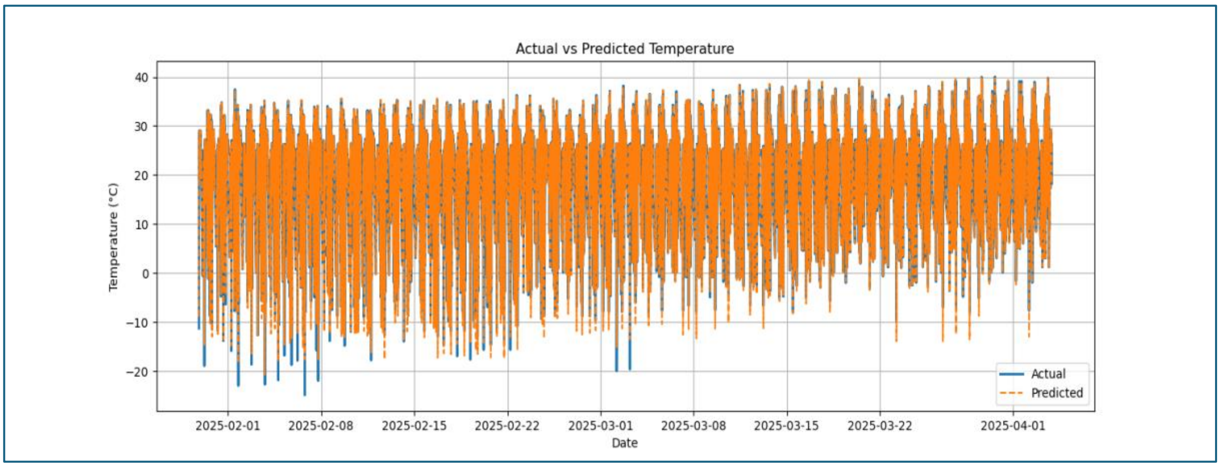
2. **Facebook Prophet**:
   - Slightly better MAE (6.5294) compared to ARIMA, but RMSE (8.6703) is slightly higher.
   - Negative R^2 (-0.0607) implies Prophet's seasonality handling is insufficient for this dataset, possibly due to noise or irregular patterns.

3. **SARIMAX**:
   - Poor performance with high MAE (9.9575) and RMSE (12.8290).
   - R^2 is not applicable, likely due to issues in model fitting or data resolution (e.g., monthly aggregation).

4. **XGBoost Regressor**:
   - MAE (8.7602) and RMSE (9.8191) indicate suboptimal predictions compared to ARIMA and Prophet.
   - R^2 (-0.3604) suggests overfitting or poor feature engineering in the initial implementation.

5. **Fine-Tuned XGBoost**:
   - Exceptional improvement with MAE (0.0861), RMSE (0.4207), and near-perfect R^2 (0.9983).
   - Indicates effective hyperparameter tuning and feature engineering, making it the best-performing model by far.

Fine-Tuned XGBoost Model



- Temporal features and lag features are created to capture trends and dependencies in the data.
- The target and feature variables are separated, and the dataset is split into training and testing sets.
- The XGBoost Model is trained using hyperparameter tuning with randomized search to optimize its performance on predicting temperature.
- ‘RandomizedSearchCV’ performs hyperparameter tuning by sampling random combinations from the hyperparameter space defined in params.
- The hyperparameters that will be tuned during the ‘RandomizedSearchCV’ are:
  - **n_estimators**: Number of boosting rounds or trees.
  - **max_depth**: Maximum depth of each tree.
  - **learning_rate**: The step size for each iteration.
  - **subsample**: Fraction of samples used for training each tree (used for regularization).
  - **colsample_bytree**: Fraction of features used for training each tree.
- The best model is used to make predictions on the test data.