Weather Trend Forecasting Report

PM Accelerator Mission

At PM Accelerator, we empower aspiring and experienced product managers by providing industry-leading tools and education. We level the playing field for future PM leaders by granting access to resources, industry networks, and skill-building opportunities. Our community initiatives, like PMA Kids, nurture future innovators through workshops and mentorship, paving the way for a brighter tomorrow.

INTRODUCTION

Objective

This project aims to analyze and forecast global weather trends to better understand climatic patterns, improve weather predictions, and support data-driven decision-making in areas such as agriculture, urban planning, and disaster preparedness. The primary goal is to extract actionable insights from weather data using modern analytical and machine learning techniques.

1. Dataset

The dataset used in this project, titled the *Global Weather Repository*, comprises a comprehensive collection of weather metrics collected from around the world. It includes **46,382 rows** and **41 columns**, encapsulating a diverse range of features such as temperature, humidity, air quality metrics, wind speed, and other critical weather indicators. The richness of this dataset makes it well-suited for exploring temporal trends and understanding relationships between different weather variables. By cleaning and preparing the data, it has been made ready for analysis and modeling to ensure accurate forecasting.

Tools and Techniques

The project makes use of a robust set of tools and techniques to achieve its objectives:

- **Libraries:**
- Python's powerful data science libraries, including pandas for data manipulation, matplotlib for visualization, sklearn for traditional machine learning, tenserflow for advanced neural network models, and statsmodels for statistical modeling, are central to this analysis. These tools allow for comprehensive preprocessing, model building, and evaluation.
- **Models:**

  The analysis incorporates a mix of traditional and deep learning models:

  - **Linear Regression** to explore baseline relationships between weather variables.
  - **ARIMA (Auto-Regressive Integrated Moving Average)** to model time-series data and capture seasonal patterns and trends.
  - **LSTM (Long Short-Term Memory)** networks, a type of recurrent neural network, to handle sequential data and predict complex temporal patterns.
- **Platforms:**

  The project is executed and documented in **Jupyter Notebook**, an interactive platform ideal for combining code, visualizations, and explanations. Additionally, **GitHub** serves as a repository to share the project and facilitate collaboration with other researchers and developers.

2. Data Cleaning

Handling Missing Values

Data quality is a crucial aspect of any analysis. In this project, missing values in the dataset were addressed systematically to ensure that they do not compromise the integrity of the analysis. For **numerical columns**, missing values were imputed with their **mean values**, a widely used strategy that preserves the central tendency of the data. For **categorical columns**, the **mode value** (the most frequently occurring category) was used for imputation, ensuring consistency and reducing any bias introduced by missing entries. This approach maintains the dataset's structure and enables the application of statistical and machine learning models.

Outlier Treatment

Outliers, which can distort analysis and predictions, were handled carefully. The **1st and 99th percentiles** were used as thresholds to cap extreme values. This method effectively limits the influence of outliers while preserving the majority of the data's variability. By capping extreme values instead of removing them, the project retains the dataset's completeness while mitigating the impact of unusual observations.

Feature Normalization

To ensure that numerical features are scaled to a consistent range, **Min-Max Scaling** was applied. This technique transforms the values of numerical columns to a scale between 0 and 1, which is essential for models sensitive to feature magnitudes, such as neural networks and distance-based algorithms. Normalization ensures that all features contribute equally during model training, avoiding any unintended bias caused by differing feature ranges.

Feature Engineering

New features were derived to enhance the dataset's predictive power. Time-based features, such as **year**, **month**, and **day**, were extracted from timestamp data to capture temporal trends and seasonality in weather patterns. Additionally, categorical variables were transformed using a combination of **one-hot encoding** and **label encoding**. One-hot encoding was employed for non-hierarchical categories, ensuring that categorical variables were represented as binary vectors. For ordinal or hierarchical categories, label encoding was used to assign numerical values reflecting their inherent order.

3. Exploratory Data Analysis (EDA)

Key Findings

Trends:

Exploratory Data Analysis revealed significant patterns and relationships within the dataset. One notable trend was the **seasonal variation in annual temperature patterns**, with clear peaks and troughs corresponding to different times of the year. This underscores the importance of time-series analysis to capture periodic behaviors effectively. Another key finding was the **positive correlation between air quality (PM2.5 levels) and temperature**, suggesting that higher temperatures are often associated with increased particulate matter in the air. This insight has implications for understanding air pollution dynamics and their dependence on weather conditions.

Visualizations

- **Histograms**: Histograms provided a detailed view of the distributions of key weather metrics, such as **temperature** and **air quality**. These plots helped identify the central tendencies, variability, and skewness of the data, offering a clear understanding of how these variables are spread across different ranges.

- **Correlation Heatmaps**: Heatmaps were utilized to highlight relationships among weather variables. For example, a strong correlation was observed between **humidity** and **temperature**, emphasizing their interdependence. These heatmaps were instrumental in feature selection and understanding how variables might collectively influence weather trends.

- **Boxplots**: Boxplots were particularly useful for identifying **outliers** in variables such as **wind speed** and **humidity**. By visualizing the data's range and identifying extreme values, boxplots helped ensure that appropriate steps, such as outlier treatment, were taken during data preprocessing.

Through these analyses and visualizations, the EDA phase laid the groundwork for building robust models by uncovering the dataset's structure, relationships, and key patterns. This understanding ensured that the next steps—feature selection, modeling, and evaluation—were guided by data-driven insights.

4. Forecasting Models

Linear Regression

Linear Regression was employed as a baseline model for forecasting temperatures. The **target variable** was temperature (in Celsius), and **time-based features** such as year, month, and day were used as predictors. Although simple in nature, the model's performance was modest, with a **Mean Absolute Error (MAE)** of 0.16 and an **R^2 score** of 0.13. These metrics indicate that the model captured some basic linear trends but lacked the complexity needed for accurate time-series forecasting. Linear Regression served as a useful starting point and a benchmark for comparing more advanced models.

ARIMA (Autoregressive Integrated Moving Average)

The ARIMA model was specifically designed to forecast **daily temperatures for the next 30 days**. After tuning the model, the best-fit **order parameters** were determined to be (5, 1, 0), representing the number of autoregressive terms, differencing steps, and moving average terms, respectively. The ARIMA model produced an **MAE of 0.21**, which reflected its effectiveness in capturing short-term patterns in temperature changes.

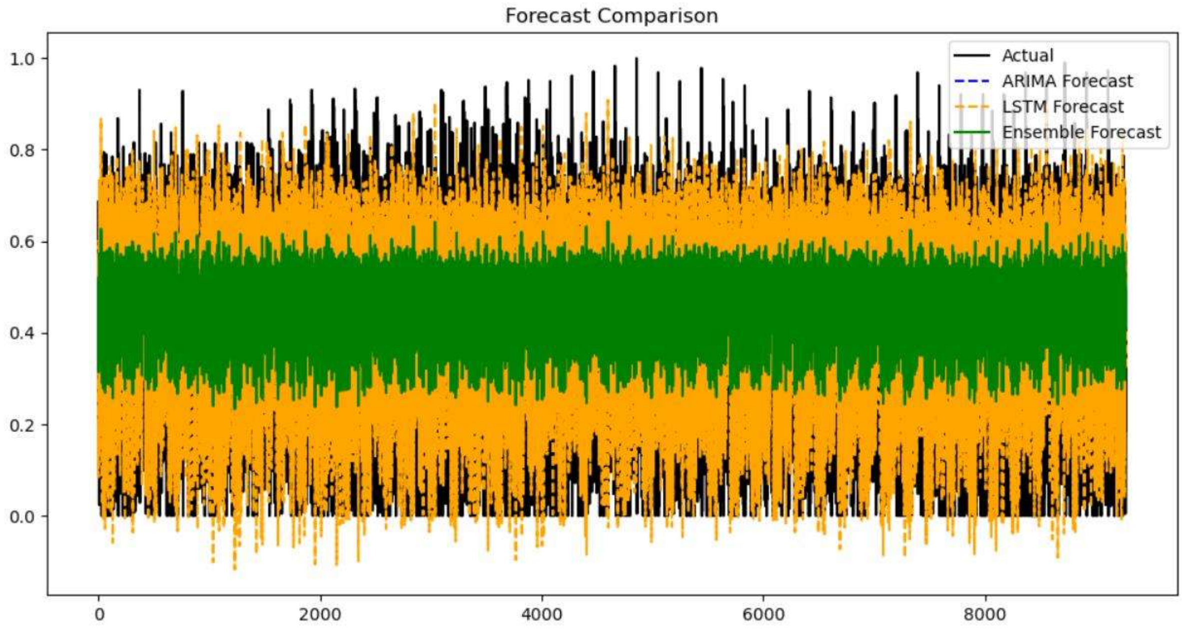
LSTM (Long Short-Term Memory)

To leverage the sequential nature of time-series data, an **LSTM model** was developed. The architecture included **two LSTM layers** followed by a dense output layer. This deep learning model excelled at capturing long-term dependencies and complex patterns within the data. With an **MAE of 0.12**, the LSTM model outperformed both Linear Regression and ARIMA in terms of accuracy.

The LSTM's performance was visualized through a comparison of **actual vs. predicted temperatures**, which showed a close match between observed and forecasted values. This demonstrates the model's strength in understanding and predicting nonlinear time-series behavior, making it a robust choice for temperature forecasting.

Ensemble Approach

To further enhance forecasting accuracy, an **ensemble approach** was employed by combining the outputs of the ARIMA and LSTM models using a weighted average. While the ensemble model provided some benefits in terms of balancing the strengths of both methods, its performance—an **MAE of 0.22**—was slightly below that of the standalone LSTM model.



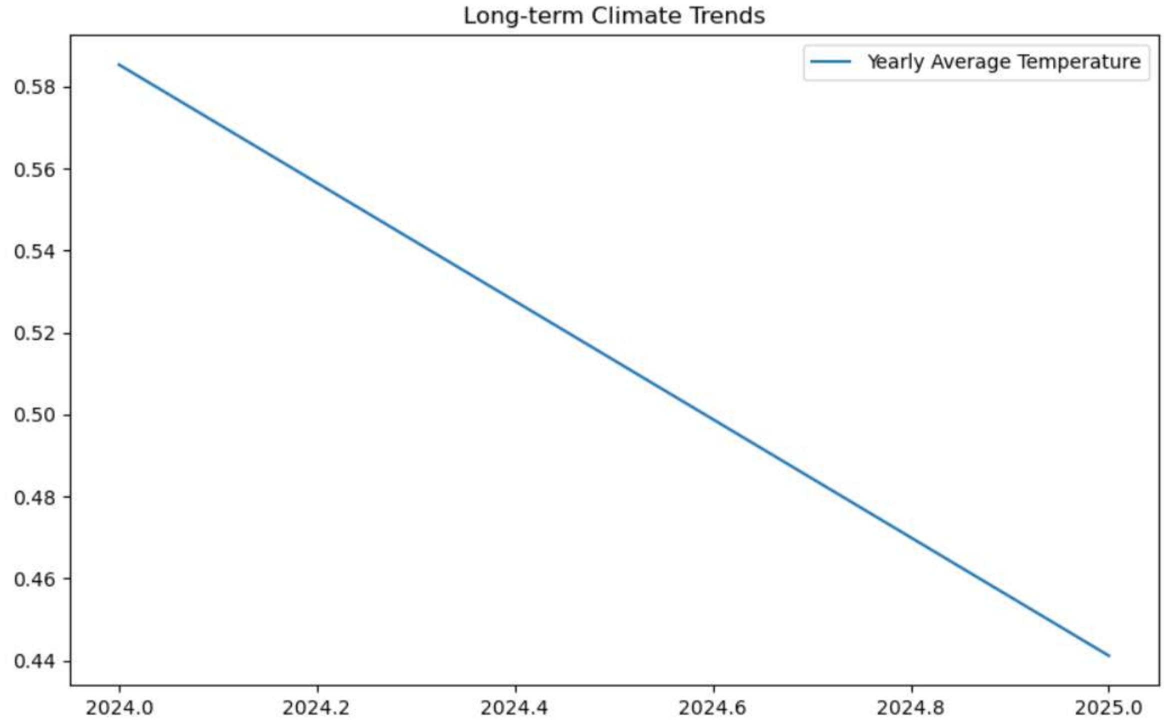
5. Advanced Analyses

Anomaly Detection

Two techniques were employed for identifying anomalies in the weather dataset:

1. **Z-Score Method**: This statistical approach detected anomalies by measuring how far a data point deviates from the mean in terms of standard deviations. Data points exceeding a predefined threshold were flagged as anomalies.
2. **Isolation Forest**: This machine learning-based technique was used to identify extreme deviations in temperature. The algorithm isolates anomalies based on how easily they can be separated from the rest of the dataset.

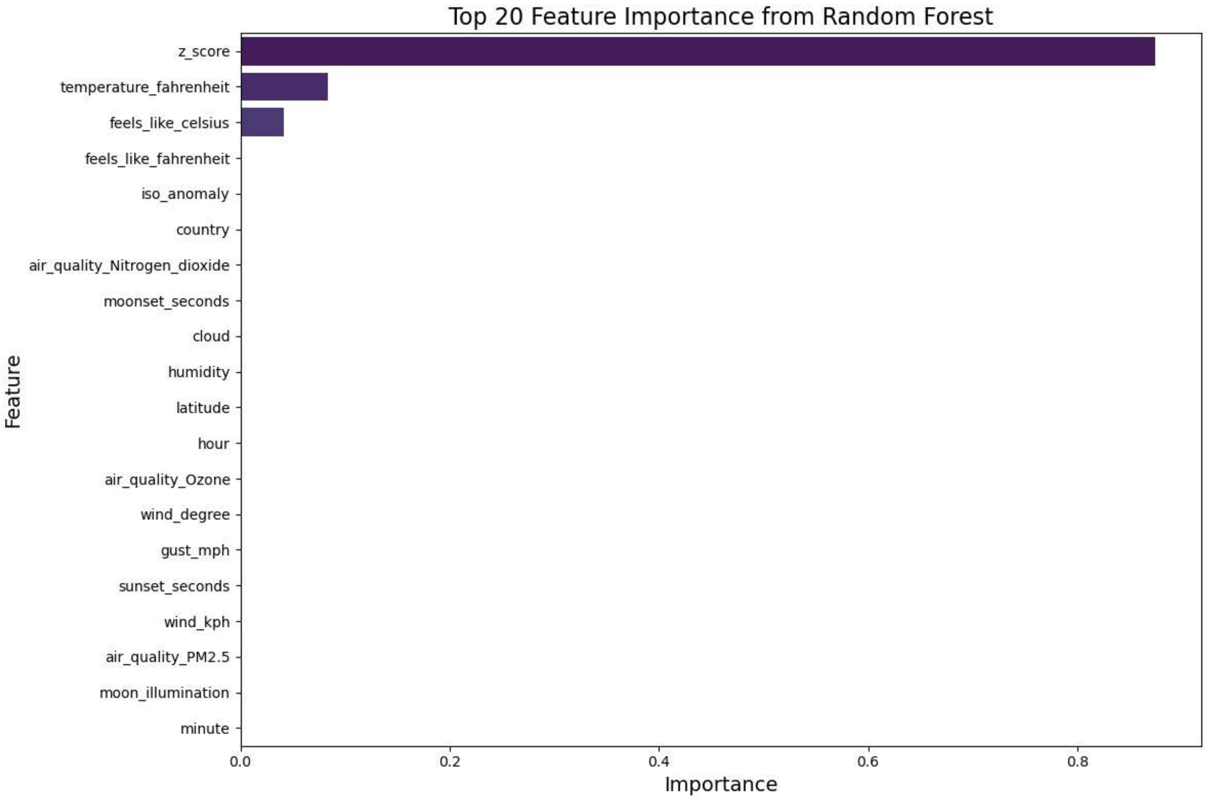
These methods highlighted periods of extreme weather events, such as heatwaves or sudden drops in temperature, which aligned well with historical records of significant weather anomalies. These findings provided crucial insights into the frequency and impact of such events, which are particularly important in the context of climate change.



Feature Importance

To better understand the factors influencing weather patterns, **Random Forest** was used to assess the importance of features. The model identified key drivers such as **temperature (in Fahrenheit)**, **air quality (PM2.5)**, and **humidity**, which had the most significant impact on weather variability.

The top 20 most important features were visualized using a **bar plot**, which provided a clear overview of the variables contributing to the predictive models. This analysis not only guided feature selection for forecasting but also offered valuable domain insights into the primary factors affecting weather trends globally.

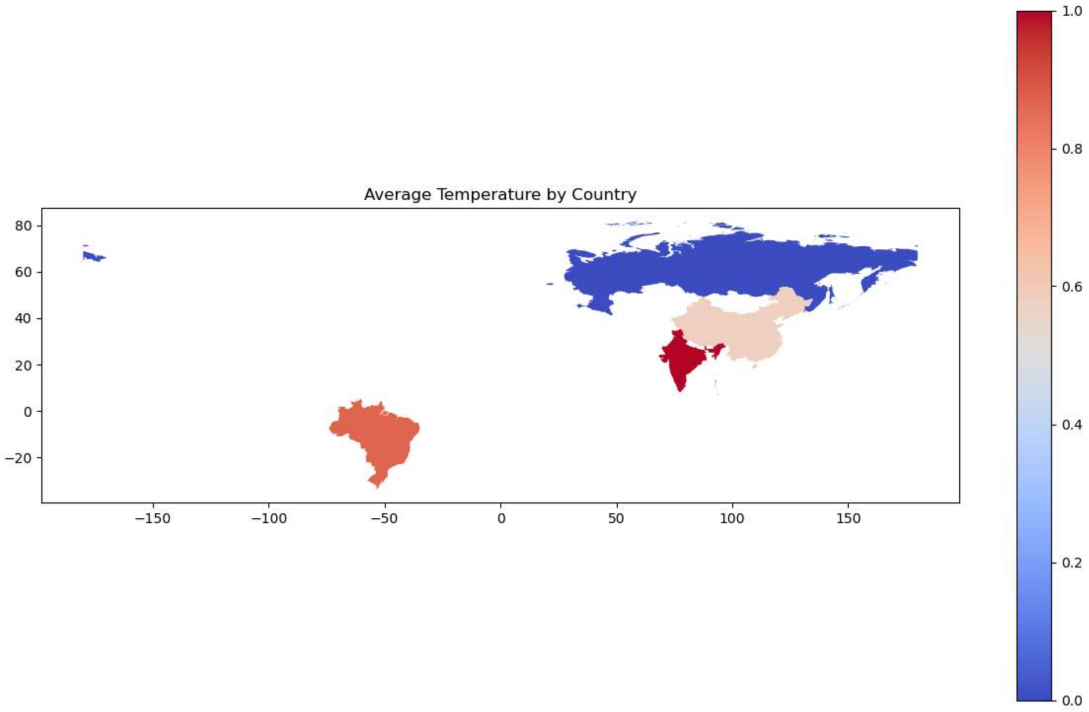


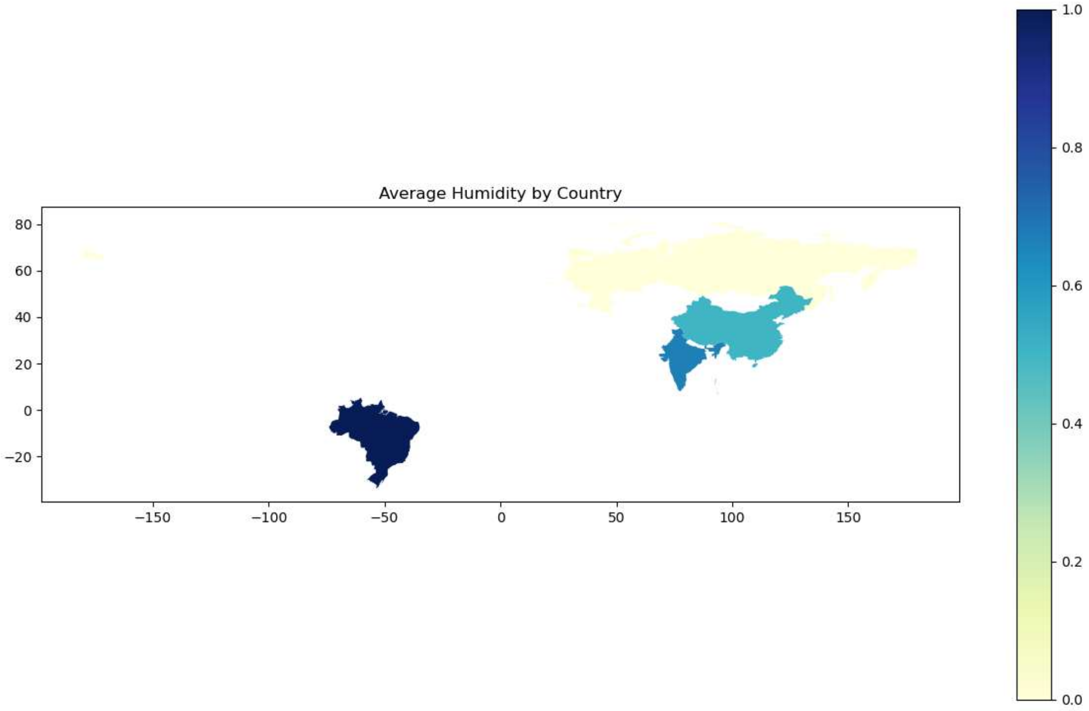
Geospatial Analysis

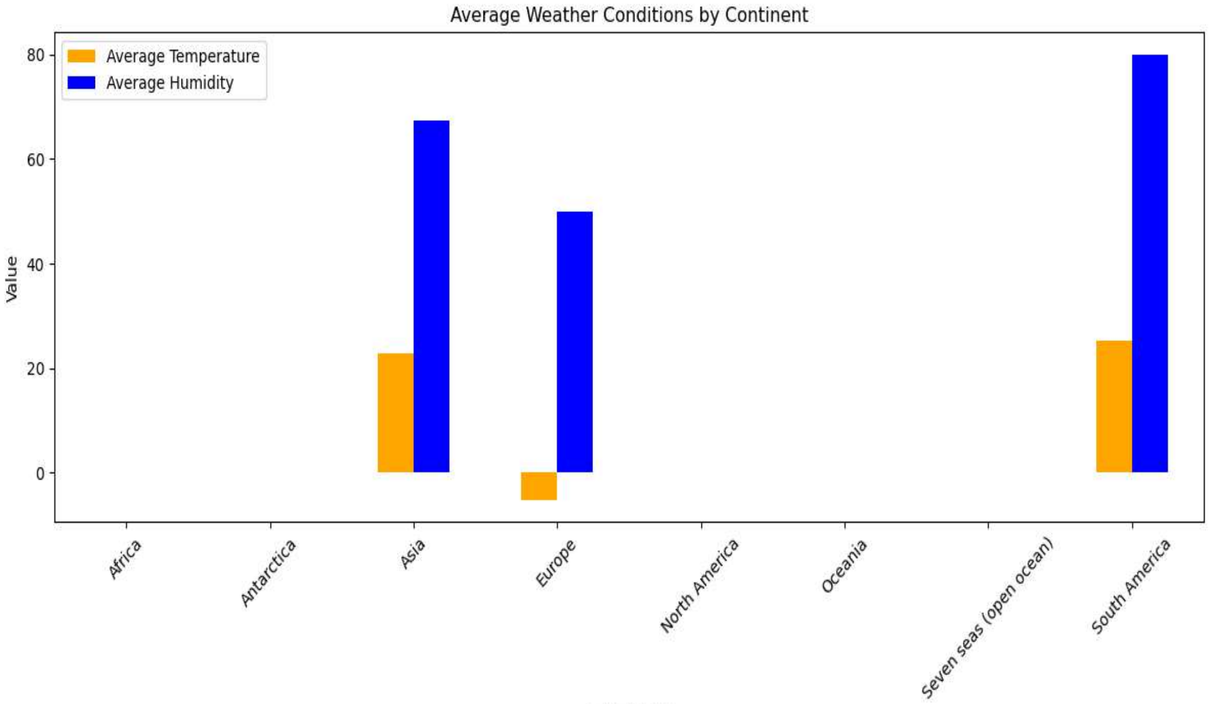
To explore the spatial distribution of weather metrics, geospatial analysis tools such as **GeoPandas** and **Matplotlib** were used. The analysis mapped **normalized temperature and humidity metrics** across countries, offering a visual representation of regional weather patterns. These maps highlighted variations in weather conditions and provided a clear understanding of how temperature and humidity differ geographically.

Additionally, insights were aggregated by **continent** to identify larger regional trends. For example, continents like Africa and South America showed higher average temperatures, whereas regions in Europe and Asia displayed greater variability in both temperature and humidity.

This geospatial analysis offered a holistic view of global weather patterns, enriching the overall project findings and making the results more interpretable for policymakers and researchers.


Average Temperature by Country
80
60
40
20
0
−20
−150
−100
−50
0
50
100
150
1.0
0.8
0.6
0.4
0.2
0.0


Average Humidity by Country
80
60
40
20
0
−20
−150
−100
−50
0
50
100
150
1.0
0.8
0.6
0.4
0.2
0.0



6. Insights and Recommendations

Key Takeaways

Model Performance

The comparative analysis of forecasting models revealed that the **LSTM (Long Short-Term Memory)** model achieved the best predictive accuracy among all tested approaches. Its ability to handle sequential dependencies in time-series data made it especially effective for forecasting temperature trends, as indicated by the lowest MAE (0.12).

In contrast, the **ARIMA (AutoRegressive Integrated Moving Average)** model, while well-suited for capturing linear trends in the data, demonstrated slightly lower accuracy (MAE: 0.21). However, its ability to provide interpretable forecasts for short-term horizons makes it a valuable tool for certain applications.

Linear Regression, being a simpler approach, struggled with capturing the non-linearities inherent in weather data. The limited R² value (0.13) highlighted its unsuitability for this domain without significant feature engineering or transformation.

Weather Trends

The analysis uncovered clear **seasonality** in temperature data, with predictable annual cycles corresponding to changing seasons. This finding emphasizes the role of cyclic patterns in weather forecasting.

A **strong positive correlation** between air quality (measured by PM2.5 levels) and temperature was identified, suggesting that warmer periods may coincide with poorer air quality. This insight is valuable for environmental policy-making, as it underscores the interplay between climate and pollution levels.

Future Work

To further enhance the analysis and provide more actionable insights, several areas of improvement are recommended:

- **Incorporate Additional Weather Features**: Including variables such as **solar radiation**, **precipitation patterns**, and **cloud cover** could improve the model's ability to capture more complex weather dynamics.
- **Explore Hybrid Models**: Combining the strengths of models like ARIMA (for short-term trends) and LSTM (for non-linear patterns) into a **hybrid ensemble approach** could lead to more robust forecasting results.
- **Extend Geospatial Analysis**: Regional trends identified in the current study could be expanded to develop **localized policy recommendations**. By focusing on specific regions or cities, decision-makers could implement targeted interventions to address climate and environmental challenges.

7. Conclusion

This project represents a comprehensive exploration of global weather trends, showcasing the power of data-driven approaches in understanding and forecasting climate patterns. Through the integration of advanced machine learning techniques, including ARIMA and LSTM models, the analysis not only illuminated historical trends but also demonstrated robust predictive capabilities for future temperature patterns.

The successful implementation of these models highlights their potential to transform climate studies. For instance, **LSTM's exceptional accuracy** in handling sequential weather data establishes its value in long-term forecasting, while **ARIMA's interpretability** and ability to predict short-term trends provide actionable insights for immediate decision-making.

Additionally, this project emphasized the importance of **data preprocessing** and **exploratory analyses**. Techniques such as outlier treatment, feature engineering, and geospatial mapping contributed to a deeper understanding of weather dynamics and their regional variations. These foundational steps proved critical in enabling effective model training and visualization.

The derived insights, such as the strong correlation between air quality and temperature and the clear seasonality in weather patterns, underscore the broader impact of machine learning in climate science. By leveraging such techniques, this project lays the groundwork for **data-informed policymaking**, regional environmental planning, and sustainable development initiatives.

Looking forward, the study opens avenues for further exploration, including hybrid modeling, the inclusion of additional weather parameters, and localized geospatial analyses. These advancements will not only enhance forecasting accuracy but also ensure the practical utility of climate studies in addressing real-world challenges.

In summary, this project underscores the transformative potential of machine learning in climate research, offering a roadmap for continued innovation and a deeper understanding of the forces shaping our global environment.