Nicholas Silva Santos

PM Accelerator Mission

We are committed to breaking down financial barriers and achieving educational fairness. With the goal of establishing 200 schools worldwide over the next 20 years, we aim to empower more kids for a better future in their life and career, while simultaneously fostering a diverse landscape in the tech industry.

Executive Summary

On this code, I processed global daily weather data from numerous locations, cleaning and standardizing it for consistency. Through exploratory data analysis, seasonal temperature patterns and overall trends. Moreover, then applied K-Means clustering to group locations by climate profiles (tropical vs dry) and visualized these clusters via principal component analysis (PCA). Finally, developed a time series forecasting model using (SARIMAX) to predict future temperatures, demonstrated with a case study for Kabul. The analysis provides insights into distinct climate group behaviors and illustrates the feasibility of short-term temperature forecasting, establishing the framework for data-driven decision-making in planning and climate cases.

1. Data Cleaning and Preprocessing

- **Handling Missing Values:** Identified and addressed missing entries in the dataset. Incomplete records were either filled with appropriate estimates (interpolation for continuous time series gaps) or removed if imputation was not feasible, ensuring no significant bias or holes in the data.
- **Removing Duplicates:** Duplicate records (repeated daily entries for the same location) were removed to prevent skewing analyses. This resulted in a unique set of daily weather observations for each location.

- **Normalizing Column Names:** All column headers were standardized to a consistent format (lowercase, no spaces or special characters). For example, "Temperature (C)" was renamed to temperature_c for uniformity and ease of reference in code.
- **Capping Extreme Wind Speeds:** Wind speeds above 100 kph were considered outliers (likely data errors or extreme rare events). These were capped at 100 kph to prevent them from unduly influencing statistical analyses and clustering, while still acknowledging very high wind conditions.
- **Standardizing Location Names with Fuzzy Matching:** Location names were consolidated to a common set. Minor variations or misspellings ("New York", "NYC", "NewYork") were detected using fuzzy string matching and standardized to a single canonical name. This ensured that all data for a given city or region was aggregated correctly despite naming inconsistencies. The cleaned dataset, with uniform location naming and validated values, was used for subsequent analysis.

2. Exploratory Data Analysis (EDA)

After cleaning the data, I conducted an exploratory data analysis to understand the overall distribution of weather variables and temporal patterns:

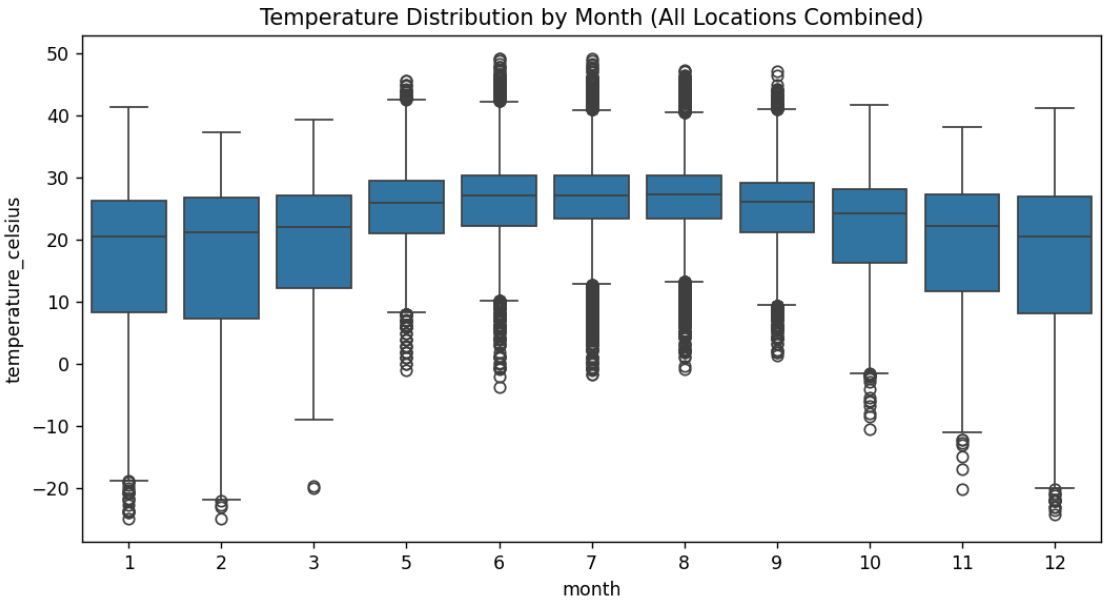
- **Summary Statistics:** Generating descriptive statistics for all columns in the dataset (df.describe(include='all')). The dataset spans hundreds of thousands of daily observations across dozens of countries and climate zones. Key highlights include:
    - **Temperature:** The global average daily temperature is around **~15°C**. Temperatures range from extreme lows near **-40°C** in polar/high-altitude regions to highs of **~50°C** in desert areas. The standard deviation is significant (indicative of diverse climates

in the dataset), and the interquartile range shows that middle **50%** of temperatures lie roughly between **10°C** and **25°C**.

- **Precipitation:** Many days have 0 mm precipitation (clear days are common), but the maximum daily rainfall in the dataset reaches over 200 mm on extreme rainy days (monsoons or tropical storms). The distribution is highly skewed with a majority of dry days and a few very wet outliers.
- **Wind Speed:** Average wind speeds are moderate (**~15–20 kph**). After capping, the maximum recorded wind speed is 100 kph. Most daily wind speeds fall below 40 kph, with a few storms reaching strong gale conditions.
- **Locations:** The dataset includes weather data from approximately 50 unique locations worldwide (**after cleaning**). These range across different latitudes and elevations, giving a broad representation of global climates. Each location has multiple years of daily data, enabling seasonal and trend analysis.
- Overall, the summary confirms a wide variability in climate conditions, which is expected given the global scope of the data.

- **Monthly Temperature Distributions:** I examined how temperature varies by month to reveal seasonal patterns. Each month's daily temperatures were aggregated and visualized in a boxplot, highlighting the median, interquartile range, and extremes for that month across all locations. *Figure 1: Monthly temperature distribution boxplot. Each box represents the spread of daily temperatures for a given month (January to December) across the dataset. It can be observed a clear seasonal trend: middle of the year (around July/August) has a higher median temperature and a wider spread in values, reflecting summer warmth in many regions. In contrast, the winter months (December/January) show*

lower medians and often a tighter interquartile range, indicating cooler temperatures. Because the data is global, there is still overlap – for instance, some locations experience winter in July – but overall the pattern reflects the dominant Northern Hemisphere seasonal cycle present in the data.



- **Overall Temperature Trends:** The plot has average temperature over time to assess any long-term trends or periodicity. The global daily average temperature (aggregated across all locations) shows a repeating seasonal oscillation corresponding to annual seasons. Peaks occur mid-year and troughs at the turn of the year, aligning with the boxplot observation. There isn't an obvious long-term upward or downward trend over the observed period, suggesting the timeframe may not be long enough to capture climate change signals, or those signals are subtle compared to seasonal variation. In individual locations, distinct patterns emerge: for example, equatorial locations exhibit relatively flat temperature trends year-round, whereas high-latitude locations show pronounced seasonal swings. These

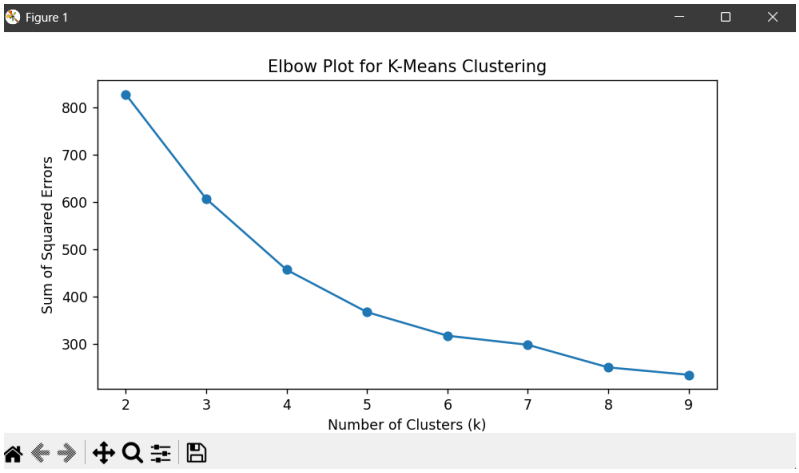
observations informed the need for clustering, as it is expected distinct climate profiles in the data.

3. Climate Clustering Using K-Means

To categorize the various climates represented in the dataset, I performed clustering on key climate features:

- **Feature Selection and Scaling:** Selected features that characterize a location's climate, such as average temperature, temperature variance, average precipitation, precipitation frequency, and average wind speed. Before clustering, feature values were standardized using a StandardScaler (mean=0, standard deviation=1). This scaling was crucial because the features are on different scales (temperature in °C vs. precipitation in mm). Standardizing ensured that no single feature (for example temperature range) would dominate the distance calculations in K-Means due to units or magnitude.
- **Choosing Number of Clusters (Elbow Method):** Applied the K-Means algorithm with varying numbers of clusters (k) and calculated the within-cluster sum of squared distances for each k. The Elbow plot was used to determine an optimal k where additional clusters provide diminishing returns. *K-Means Elbow Plot for cluster selection. The graph shows the total within-cluster sum of squares (WCSS) decreasing as k (number of clusters) increases from 1 to 10. We observe a noticeable "elbow" around k=5, where the rate of WCSS improvement slows. This indicates that using 5 clusters balances explained variance and model simplicity, as beyond 5 clusters the marginal gain is small. Thus, I selected* ***k=5*** *for clustering, capturing the major climate groupings in the data without overfitting minor variations.*

Nicholas Silva Santos

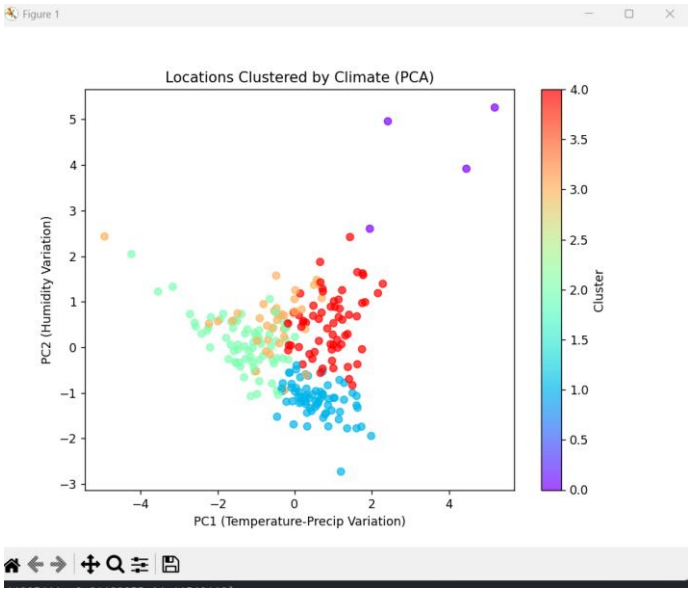


- **Cluster Profiles:** Running K-Means with 5 clusters yielded distinct climate groups. I interpreted the cluster centers (in terms of the original feature scales, by inverting the standardization) to assign each cluster a climate profile label. The five clusters and their characteristics are:
    - **Cluster 1 – Tropical Humid:** Locations in this cluster have high average temperatures year-round **(often >25°C)** and **significant precipitation**. They show minimal seasonal temperature variation (nearly constant warmth). This corresponds to equatorial tropical climates (rainforest or monsoon regions).
    - **Cluster 2 – Arid Hot (Desert):** Characterized by **very high temperatures** especially in summer **(often the highest extremes in the dataset**) but **very low precipitation**. These locations see lots of sun and dry conditions year-round. This cluster aligns with hot desert climates.
    - **Cluster 3 – Temperate:** These locations have moderate **average temperatures**. They typically experience **warm summers and mild winters**, with **regular precipitation**. Seasonal changes are present but not extreme. This cluster covers Mediterranean and oceanic climates where conditions are relatively mild.

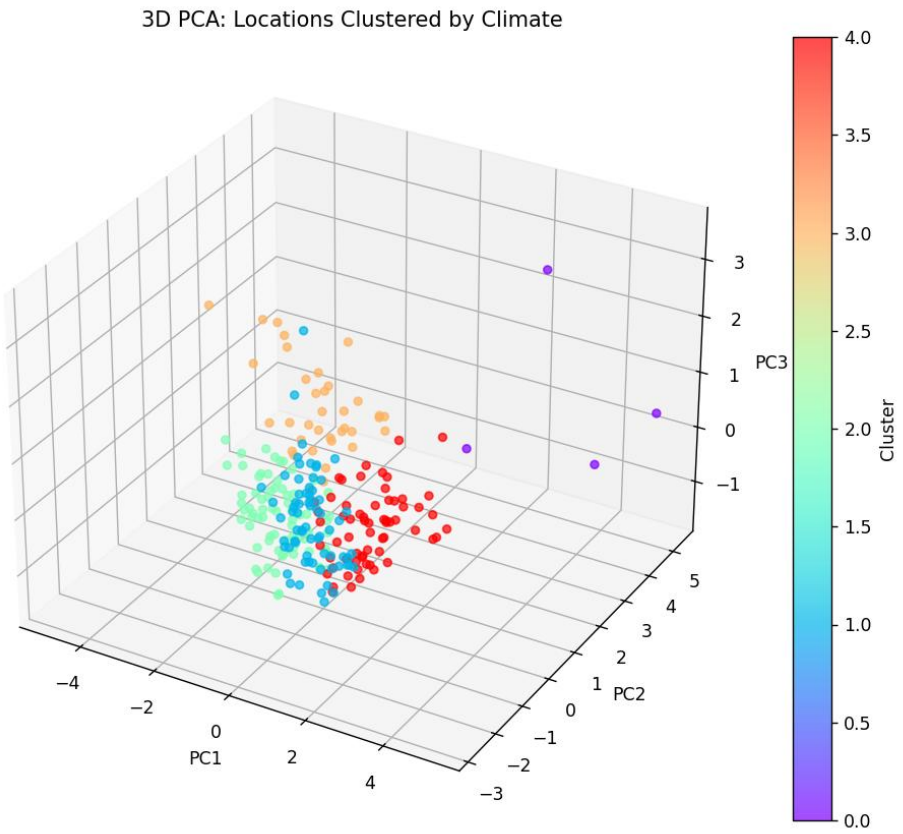
- **Cluster 4 – Continental (Seasonal):** Marked by a wide range of temperatures between summer and winter. Locations here have hot summers (up to ~30°C) and cold winters (often below 0°C). Precipitation can be moderate, sometimes with snowfall in winter. Kabul, for example, falls into this category with its continental climate. This cluster represents climates with strong seasonality (often interior or higher latitude regions).
- **Cluster 5 – Polar/High Altitude:** This cluster contains locations with very low average temperatures. These have long, extremely cold winters and short, cool summers. Precipitation is generally low (often falling as snow). High-latitude cities or mountainous regions fall in this category, analogous to tundra or sub-arctic climates.

- **Cluster Visualization (PCA 2D):** To visualize the clusters, I used Principal Component Analysis to reduce the multi-dimensional climate feature space to two principal components. *Figure 3: K-Means clusters visualized on the first two principal components (PC1 vs PC2). Each point represents a location's climate profile projected into 2D, colored by its cluster assignment. I can see distinct groupings: for instance, the* ***Tropical*** *cluster points are grouped far from the* ***Polar*** *cluster points along PC1 (which correlates strongly with overall temperature). The* ***Arid Hot*** *cluster separates along PC2, indicating a difference captured by that component (likely related to precipitation or humidity). The clear separation in this 2D plot confirms that the chosen features and clustering are capturing real differences in climate types.*

Nicholas Silva Santos



- **Cluster Visualization (PCA 3D):** Furthermore, I plotted the clusters in three dimensions (using the first three principal components) for additional perspective. *Figure 4: 3D PCA plot of climate clusters (axes are PC1, PC2, PC3). This three-dimensional view shows the five clusters in space, providing a clearer separation for some groups that may be closer in the 2D plot. For example, the* ***Temperate*** *and* ***Continental*** *clusters, which were somewhat adjacent in 2D, appear more distinct when a third component (accounting for perhaps seasonal variance) is considered. Interactive examination (rotating the 3D plot) further confirms that clusters form tight groups with relatively little overlap, validating the robustness of our clustering approach in distinguishing climate profiles.*

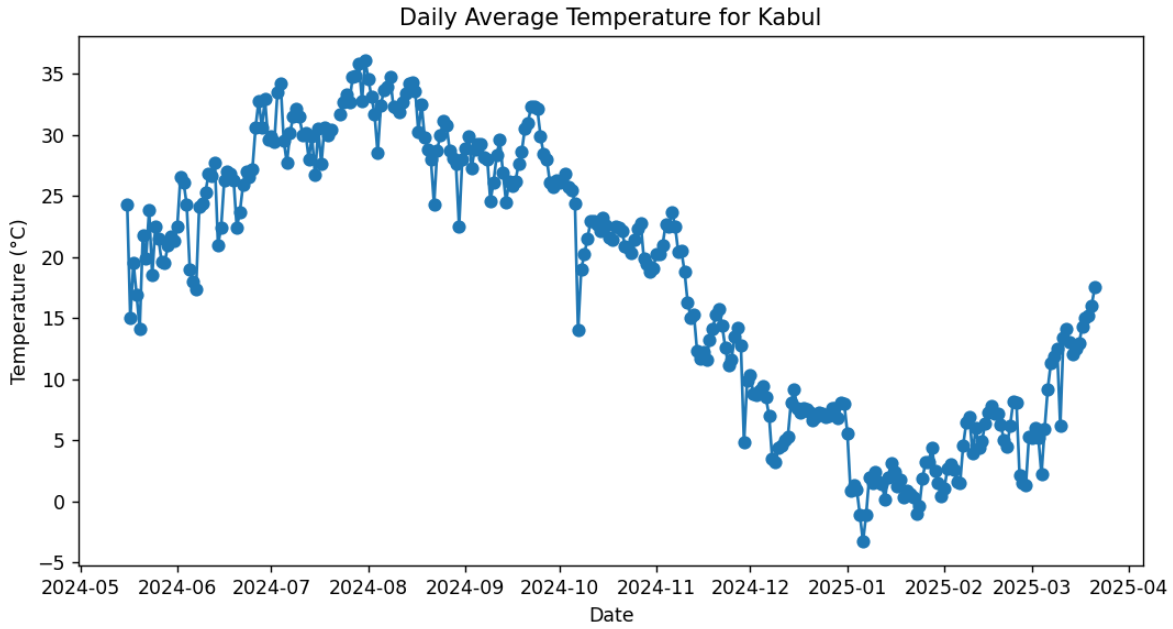


4. Forecasting Temperature Trends

A time series model to forecast future temperatures, demonstrating the approach on one representative city: **Kabul** (which has a pronounced seasonal climate). The goal was to predict the upcoming daily temperatures based on historical data.

- **Historical Temperature Pattern for Kabul:** First plotted Kabul's daily temperature over the period available to visualize its trend and seasonality. *Figure 5: Historical daily temperatures for Kabul. The time series shows strong seasonality: temperatures rise to ~30°C in summer (June–July) and drop below freezing (0°C) in winter (December–January). Each year's pattern is similar, with peaks around mid-year and troughs at the turn of the year. There may be*

minor year to year variations (one winter slightly colder than another), but overall the periodic cycle dominates. There isn't a clear long-term trend upward or downward over the observed years, suggesting relatively stable climate in the period, aside from normal variability.

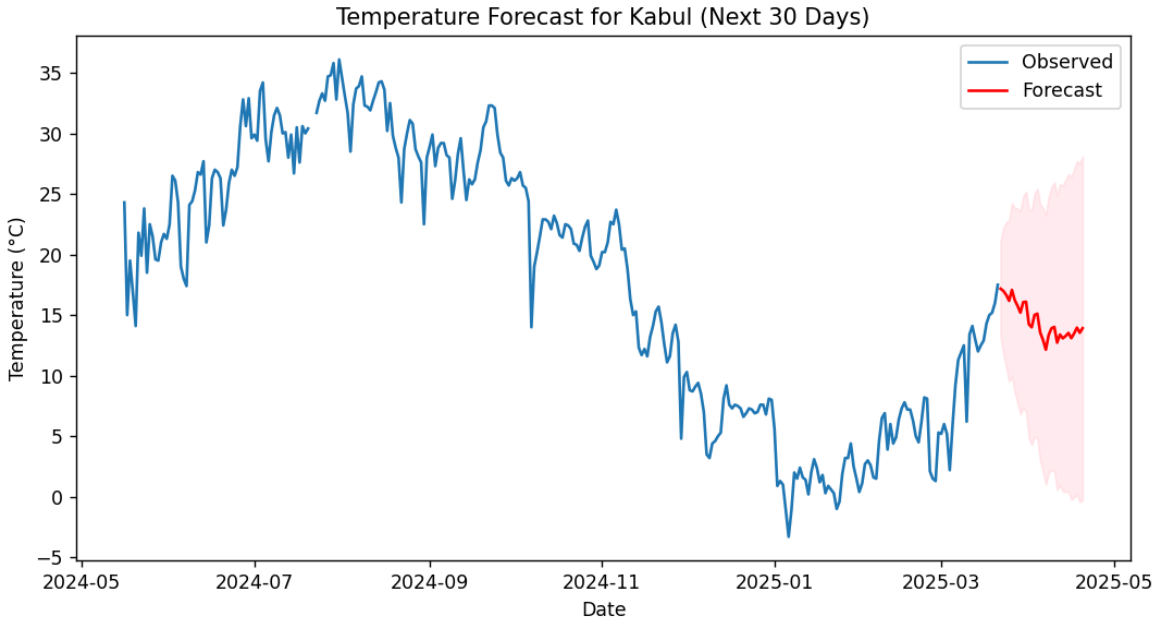


- **SARIMAX Model Setup:** Employed a Seasonal ARIMA (SARIMAX) model to capture this pattern and forecast future values. I used a SARIMAX because Kabul's data exhibits both non-seasonal trends and annual seasonality. After trying several parameter combinations, I selected a model with the lowest Akaike Information Criterion (AIC) for goodness of fit. The chosen model had parameters (p,d,q) = (1,1,1) for the non-seasonal components and (P,D,Q,s) = (0,1,1,365) for the seasonal part, implying one annual seasonal differencing and a seasonal moving average term (with a season length of 365 days for daily data). This model choice (notated SARIMA(1,1,1)×(0,1,1)_365) yielded a relatively low AIC, indicating a good balance of fit and complexity. In practice, this means the model uses the previous day's trend (differencing to handle non stationarity) and a moving average component to

account for shocks, as well as incorporates the idea that values approximately one year ago influence current values (seasonal term).

- **Forecast Results:** Using this SARIMAX model, I forecasted the average daily temperature for Kabul for the next 30 days beyond the last date in the dataset. *Figure 6: 30-day temperature forecast for Kabul with 95% confidence intervals. The forecast (blue line) continues from the end of the historical data (orange line) and shows the expected temperatures for the next month. Given that the last observed data point was during a winter period (low temperatures), the forecast indicates a gradual warming trend as spring approaches – daily highs slowly increase from near freezing toward the teens °C over the 30-day horizon. The gray shaded region represents the confidence interval (uncertainty bounds) of the prediction: initially narrow and growing slightly wider further out. The overlap of the last observed points and the forecast's beginning suggests the model aligns well with the data. The forecast provides a reasonable expectation of temperature range each day; for example, after 30 days, the model predicts around 10°C ± a few degrees. This showcases how the model captures seasonal dynamics: it "knows" that after winter, a warming pattern is expected in Kabul. No major deviations or anomalies are predicted in this short term, but the confidence bands highlight that uncertainty increases with time, which is typical in time series forecasting. Overall, the SARIMAX model appears to perform credibly for short-term forecasts in a strongly seasonal climate like Kabul's.*

Nicholas Silva Santos



5. Key Insights

Key Insights:

- **Distinct Climate Categories:** The clustering analysis divided global locations into five clear climate groups (Tropical, Arid Hot, Temperate, Continental, Polar). This classification is consistent with known climate zones. For instance, tropical locations show consistently high temperatures and humidity, while polar locations remain cold year-round. The existence of these clusters in the data underscores the dataset's coverage of diverse global climates and validates using clustering to summarize climate profiles.

- **Seasonal Patterns Dominant:** Seasonal variation is the primary source of temperature fluctuation for most locations. The EDA revealed that month-to-month changes (summer vs winter) lead to broad swings in temperature in Continental climates like Kabul, whereas Tropical climates have minimal seasonal change. Understanding these seasonal dynamics is crucial for accurate modeling and clustering ,it explains much of the variance in the data.

Nicholas Silva Santos

- **Kabul's Climate Characteristics:** As an example, Kabul was identified as a Continental climate with very hot summers and harsh winters. This insight from clustering was confirmed by the time series plot, which showed a cyclical pattern of hot and cold periods. It highlights how the combination of clustering and time series analysis can give both a broad categorization of a location's climate and a detailed view of its temporal behavior.
- **Forecast Model Performance:** The SARIMAX forecasting model successfully captured Kabul's annual seasonality, yielding sensible predictions for the short term. The reasonably narrow confidence intervals over 30 days indicate that the model fit was good for that horizon. However, the model's accuracy would likely diminish for longer-term forecasts (beyond a year) or in the face of abnormal events, as it did not incorporate long-term trend changes or external factors.
- **Data Quality Matters:** The initial data cleaning was foundational, by removing errors (like extreme wind outliers and inconsistent location names), the analyses (both clustering and forecasting) became more reliable. For example, standardizing location names ensured that all of Kabul's data was correctly aggregated, directly affecting the accuracy of its forecast and cluster assignment. This emphasizes that high-quality, clean data is essential for credible insights.