**Detailed Report on Global Weather Data Analysis**

**1. Introduction**

The analysis revolves around a comprehensive global weather dataset with 24,442 entries collected from various countries and regions, including weather metrics such as temperature, wind speed, precipitation, and air quality indices. The dataset also contains geographic coordinates, time zones, and information on celestial phenomena such as sunrise, sunset, and moon phases. This report documents the data cleaning and preprocessing steps, exploratory data analysis (EDA), forecasting models, and feature importance analysis. Additionally, spatial and time series decomposition analyses are presented to provide insights into global weather patterns and forecast temperature trends.

**2. Data Overview**

The dataset consists of 41 columns, each representing a unique weather or air quality attribute, including:

* **Country** and **Location Name**: Geographical identifiers for the region.
* **Latitude** and **Longitude**: Geographic coordinates for spatial analysis.
* **Temperature**: Recorded in both Celsius and Fahrenheit.
* **Wind Data**: Including wind speed in both kilometers per hour (kph) and miles per hour (mph).
* **Precipitation**: Recorded in millimeters and inches.
* **Air Quality Indices**: Including PM2.5, PM10, Carbon Monoxide, Nitrogen Dioxide, and Ozone levels.
* **Celestial Events**: Sunrise, sunset, moon phase, and moon illumination data.

**3. Data Cleaning and Preprocessing**

**3.1 Missing Values**

Upon analyzing the dataset, there were no missing values in the primary weather attributes. However, placeholder values like -9999 were found in air quality columns such as Carbon Monoxide and Nitrogen Dioxide, which were replaced with NaN to handle them appropriately during further analysis.

**3.2 Outlier Handling**

Outliers were detected in certain air quality columns:

* **Air Quality (PM2.5, Carbon Monoxide)**: Some regions recorded unusually high values, and extreme outliers were replaced or treated using standard deviation limits.
* **Temperature**: Some anomalies were observed, such as very low or high temperatures that didn’t align with the geographical region or season.

**3.3 Feature Scaling and Encoding**

To improve the performance of machine learning models:

* **Normalization**: Continuous variables like wind speed and air quality measures were scaled using MinMaxScaler to bring them into a comparable range.
* **One-Hot Encoding**: Categorical variables such as weather conditions (e.g., ‘Clear’, ‘Partly Cloudy’), wind direction, and moon phases were converted into numerical format using one-hot encoding, enabling models to process these features effectively.

**3.4 Feature Selection**

To reduce redundancy, columns like temperature\_fahrenheit and wind\_mph were dropped as they duplicated information already present in Celsius and kph formats.

**4. Exploratory Data Analysis (EDA)**

**4.1 Temperature Distribution**

* **Average Temperature**: 26.33°C, with a range from -3.7°C to 49.2°C. Most locations exhibited temperatures between 20°C and 35°C, indicating a global dataset spanning various climate zones.
* **Temperature Spread**: A standard deviation of 7.03°C indicates variability across different regions, with tropical and desert regions showing significantly higher temperatures than temperate or polar regions.

A graph of a temperature

Description automatically generated

**4.2 Precipitation Distribution**

* **Average Precipitation**: 0.16 mm, indicating that the majority of locations had little or no rainfall. This could imply that the dataset includes data from dry seasons or desert climates where rainfall is infrequent.
* **Outliers**: Some regions showed significantly higher precipitation levels, up to 27.82 mm, which could represent monsoon or storm-affected areas.

A graph of a number of precipitation

Description automatically generated

**4.3 Correlation Analysis**

A heatmap was created to explore relationships between the key weather variables:

* **Temperature and Humidity**: A moderate negative correlation (-0.48) was observed, suggesting that warmer temperatures are generally associated with lower humidity.
* **Temperature and Atmospheric Pressure**: A strong negative correlation (-0.48) indicated that higher temperatures are often linked with low-pressure systems, potentially explaining hot, dry conditions.
* **Temperature and UV Index**: A positive correlation (0.54) showed that higher temperatures tend to be associated with stronger UV radiation, possibly due to clearer skies.

A screenshot of a graph

Description automatically generated

**5. Time Series Analysis**

**5.1 Daily Averages**

The dataset was resampled by day to calculate daily averages for temperature, precipitation, and other weather metrics:

* **Temperature Trends**: The daily average temperature hovered between 24°C and 28°C over time, showing slight fluctuations but generally remaining consistent across most regions.
* **Precipitation Trends**: Precipitation levels were minimal in most regions, with daily averages indicating mostly dry conditions except for occasional spikes in some locations.

A graph showing the temperature of the day

Description automatically generated

A graph showing a number of days and months

Description automatically generated

A graph of blue dots

Description automatically generated

**5.2 Correlation Heatmap**

Correlation analysis between weather features showed interesting relationships:

* **Humidity**: Negatively correlated with temperature (-0.48), as expected.
* **UV Index**: Positively correlated with temperature (0.54), indicating that regions with more sunshine and heat tend to have higher UV levels.

A screenshot of a graph

Description automatically generated

**6. Forecasting Models**

**6.1 ARIMA Model**

An Auto-Regressive Integrated Moving Average (ARIMA) model was built to forecast temperature trends:

* **Model Parameters**: The ARIMA model was configured with parameters (p=5, d=1, q=0) after tuning based on the dataset.
* **RMSE (Root Mean Squared Error)**: The ARIMA model achieved an RMSE of 0.97, indicating a reasonable level of accuracy in forecasting temperature trends.

**6.2 Exponential Smoothing Model (ETS)**

The ETS model was applied to forecast temperature trends, but it performed worse than the ARIMA model:

* **RMSE**: The ETS model yielded a higher RMSE of 1.35, suggesting that it was less effective for this particular dataset due to the absence of strong seasonal patterns.

**6.3 Simple Moving Average (SMA)**

The SMA model, which uses simple averaging, performed best with an RMSE of 0.99. This suggests that temperature fluctuations were relatively stable, making the SMA a reliable predictor for this dataset.

**6.4 Ensemble Model**

A combined ensemble model using predictions from ARIMA, ETS, and SMA was built:

* **RMSE**: The ensemble model yielded an RMSE of 1.14, which was better than ETS but did not outperform SMA alone.

A graph showing the different types of data

Description automatically generated with medium confidence

**7. Feature Importance Analysis**

**7.1 Random Forest Model**

A Random Forest Regressor was trained to evaluate the importance of different features in predicting temperature:

* **Top Features**:
  + **Humidity**: Scored the highest with an importance value of 0.30.
  + **Pressure (mb)**: Followed closely with an importance value of 0.25, indicating its influence on temperature.
  + **Ozone Levels**: Had moderate importance, with a score of 0.12, highlighting its connection to warmer temperatures.
* **Least Important Features**: Wind speed and Sulfur Dioxide had minimal importance in predicting temperature, as their effects were negligible compared to other weather features.

A graph with different colored bars

Description automatically generated

**7.2 Permutation Importance**

A permutation importance analysis was performed to further confirm the results:

* **Key Observations**: Humidity and pressure remained the top features influencing temperature predictions, while pollutants like PM10 and Nitrogen Dioxide had moderate importance.

A graph with different colored bars

Description automatically generated

**8. Geospatial Analysis**

**8.1 Spatial Map of Temperature**

Using geographic coordinates, a scatter map was created to visualize temperature distributions globally:

* **Hottest Regions**: The Middle East, especially Kuwait and Iraq, consistently recorded the highest temperatures, often exceeding 49°C.
* **Coldest Regions**: Southern Hemisphere countries like Australia and Chile displayed the lowest temperatures, particularly during their winter months.A map of the world with different colored circles

  Description automatically generated

**8.2 Country-wise Temperature Comparison**

A bar chart was generated to compare the average temperatures across countries:

* **Top 5 Hottest Countries**: Kuwait (49.2°C) and Iraq (49.1°C) had the highest average temperatures, confirming the extreme desert climate of these regions.
* **Top 5 Coldest Countries**: Australia recorded the lowest temperatures, with values dropping to -3.7°C during its winter season.

**9. Time Series Decomposition**

To better understand the underlying patterns in temperature data, the time series was decomposed into its trend, seasonal, and residual components:

* **Trend**: The trend component showed a gradual increase in temperature from May to July, followed by a slight decline toward the end of September.
* **Seasonality**: The seasonal component exhibited small but regular fluctuations, possibly linked to daily temperature cycles or local weather patterns.
* **Residuals**: Residuals remained close to zero, except for a few outliers, indicating that most of the variability in temperature was explained by the trend and seasonality components.

**10. Air Quality and Temperature Correlation**

An in-depth correlation analysis was performed between temperature and air quality measures:

* **Ozone and Temperature**: A moderate positive correlation (0.35) suggested that Ozone levels tend to rise in warmer regions.
* **Pollutant Interrelationships**: High intercorrelation was observed between pollutants such as Carbon Monoxide, Nitrogen Dioxide, and PM2.5, indicating that these pollutants often increase together.

A screenshot of a graph

Description automatically generated

**11. Conclusions**

* **Global Temperature Patterns**: The dataset revealed significant variability in global temperatures, with regions like Kuwait and Iraq experiencing extreme heat, while Australia and Chile faced colder conditions.
* **Best Forecasting Model**: The Simple Moving Average (SMA) model outperformed others, indicating that temperature trends in the dataset were relatively stable and predictable with simple models.
* **Air Quality Insights**: Ozone levels showed the strongest correlation with temperature, while other pollutants like Carbon Monoxide and Nitrogen Dioxide had minimal direct relationships with temperature.

**12. Future Work**

* **Model Refinement**: Further exploration of seasonal components in the ETS model or more advanced ARIMA parameter tuning could improve forecasting accuracy.
* **Feature Expansion**: Incorporating additional features like solar radiation or atmospheric moisture could enhance predictive models, particularly for regions with more complex weather patterns.
* **Long-term Climate Analysis**: Given the seasonal trends and geographical diversity in the dataset, a deeper longitudinal analysis of climate change patterns could provide valuable insights into global warming and environmental shifts.