Report on Environmental Time Series Data Analysis and Predictive Modeling

This report details an in-depth analysis of environmental monitoring data, concentrating particularly on measurements from air quality and temperature prediction. The project involved activities, including data cleaning and exploratory data analysis, feature engineering, advanced time series analysis, outlier detection, and the development of a reliable model for temperature prediction. The overall objective of the project was to provide insights into patterns of air quality and identify unusual events while building a reliable, predictive model for temperature forecasting.

1. Overview of Data and Preprocessing:

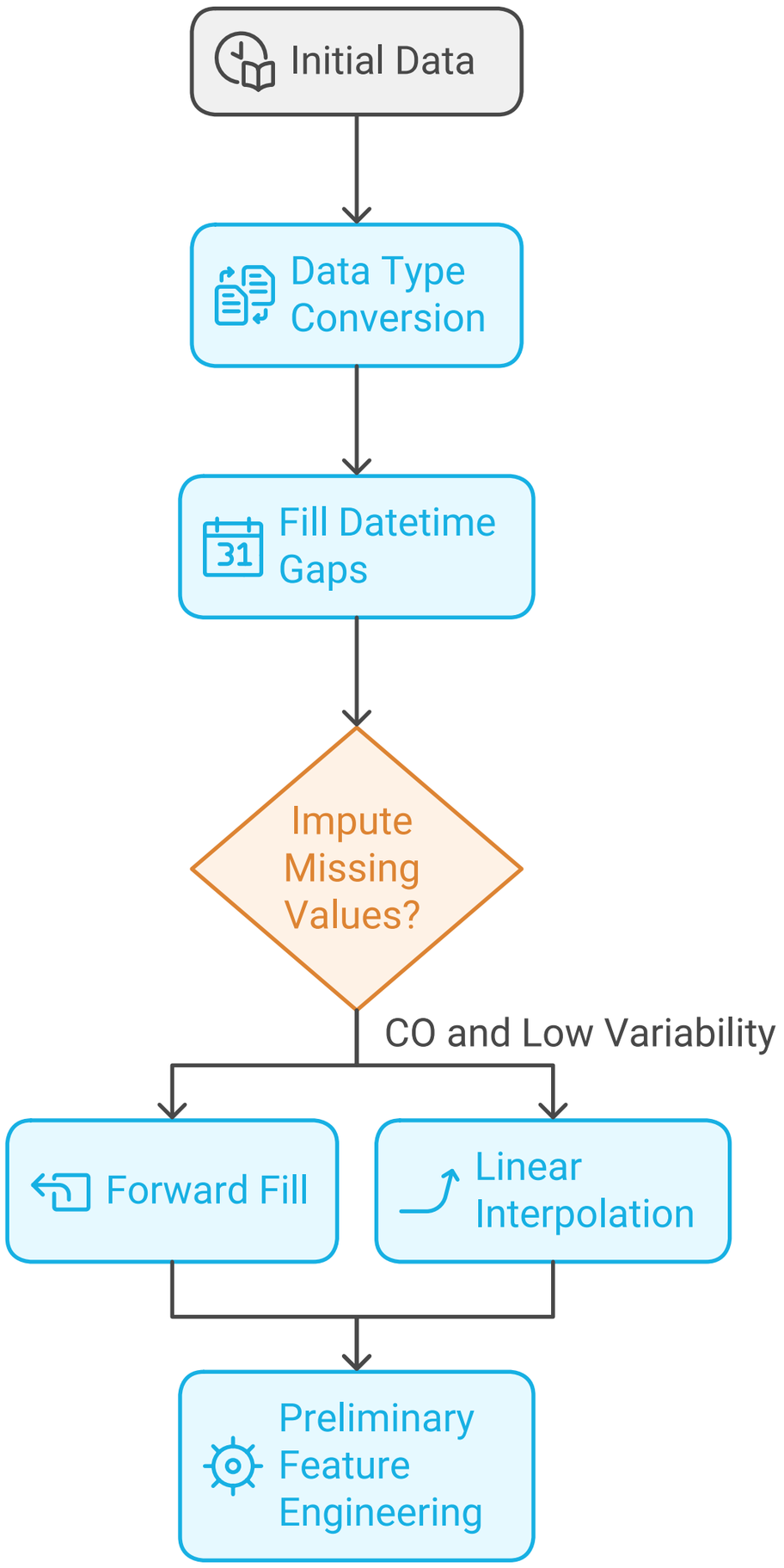
The initial data consisted of the hourly measurements of air quality in the years between 2017 and 2020 with SO2, O3, CO, temperature, particulate matter, and other relevant measurements. Under the preprocessing of the data, the following operations were performed:

Data Type Conversion: The measurement of pollutant concentrations and datetime information were converted to appropriate numeric and datetime formats. Invalid values were also tolerated gracefully and converted to NaN.

Datetime Gap Filling: Missing values for time stamps were carefully located and completed with uniform 1-hour jumps relative to the previous good value, to ensure that continuity of data and integrity of time series data were not lost.

Missing Value Imputation: Missing values in pollutant concentrations were dealt with using particular methods: forward fill for CO and an anonymous variable apparently with low variability, and linear interpolation for particulate matter and O3, the two variables most likely to exhibit smooter transitions.

Preliminary Feature Engineering: Simple time-varying features (year, month, hour) were derived from the datetime field for preliminary inspection and visualization.



2. EDA and Quality Check

A rigorous EDA was performed to understand the nature and pattern of the data and look for potential anomalies:

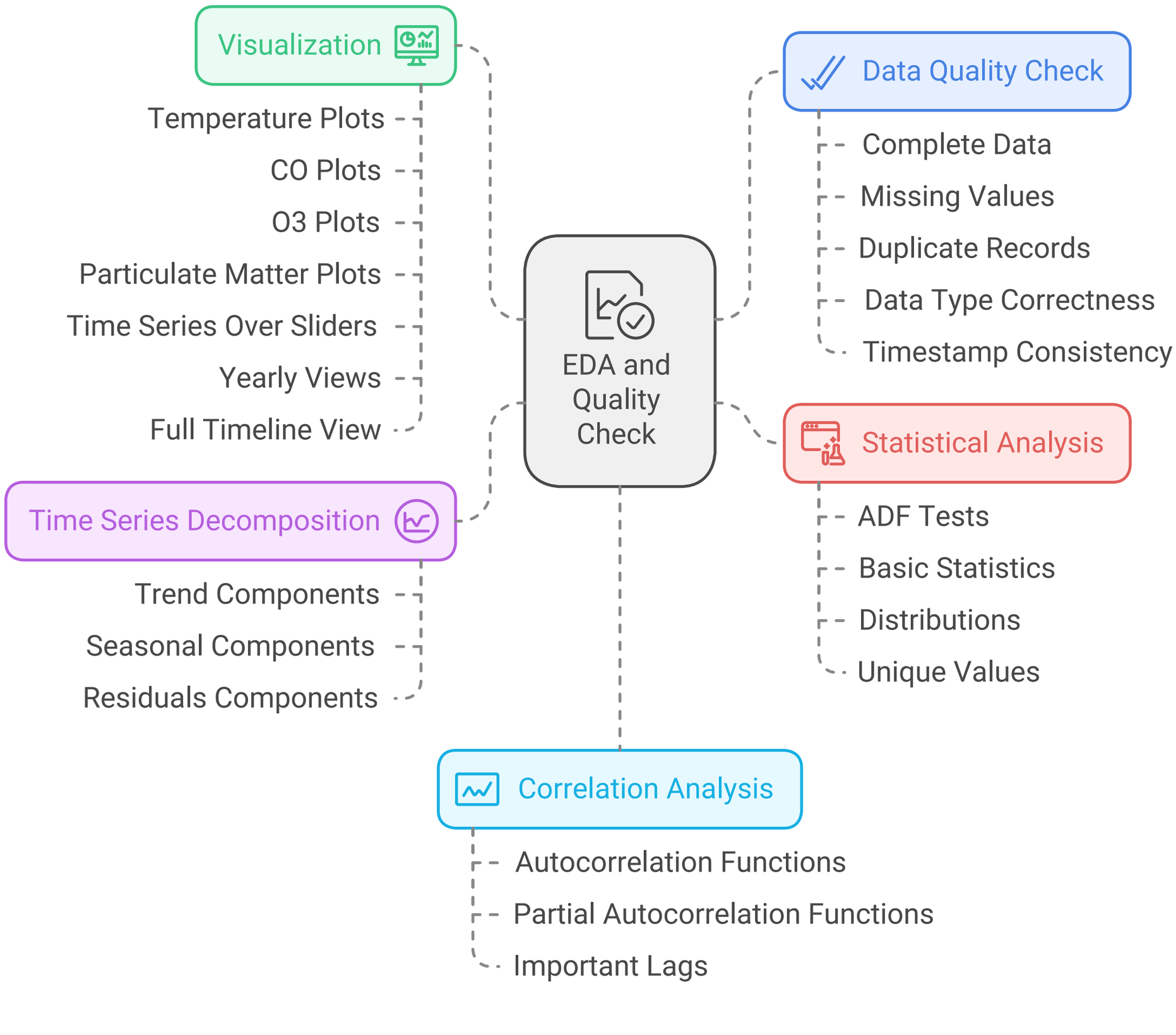
Data Quality Check: All checks were done to determine complete data, missing values, duplicate records, and more important data type correctness. Careful checks were made on timestamp consistency and intervals.

Visualization: Plots of key variables (temperature, CO, O3, particulate matter) were created using time series over sliders zooming and by different views and selections through year 1, 2, 3 views and the full timeline view. This was all quite easy to visually inspect the trend and pattern from the data.

Statistical Analysis: Augmented Dickey-Fuller tests to check stationarity as its stationarity is extremely crucial while analyzing time series. Basic statistics with distributions have been checked, as well as unique values.

Time Series Decomposition: The time series was decomposed into trend, seasonal, and residuals components using seasonal decomposition in order to understand the underlying pattern.

Correlation Analysis: Through the computation of autocorrelation functions as well as partial autocorrelation functions (ACF and PACF), the relationships between the time series were better identified along with important lags.



3. Time Series Analysis and Outlier Detection:

Advanced techniques have applied to further explore the time series and pick out anything interesting in the data:

Lag Analysis: Lag plots were made at various time scales (1-hour, 3-hour, 24-hour, 8640-hour) to identify short-term and long-term correlations. These plots will indicate the amount of influence past values have on the future measurements.

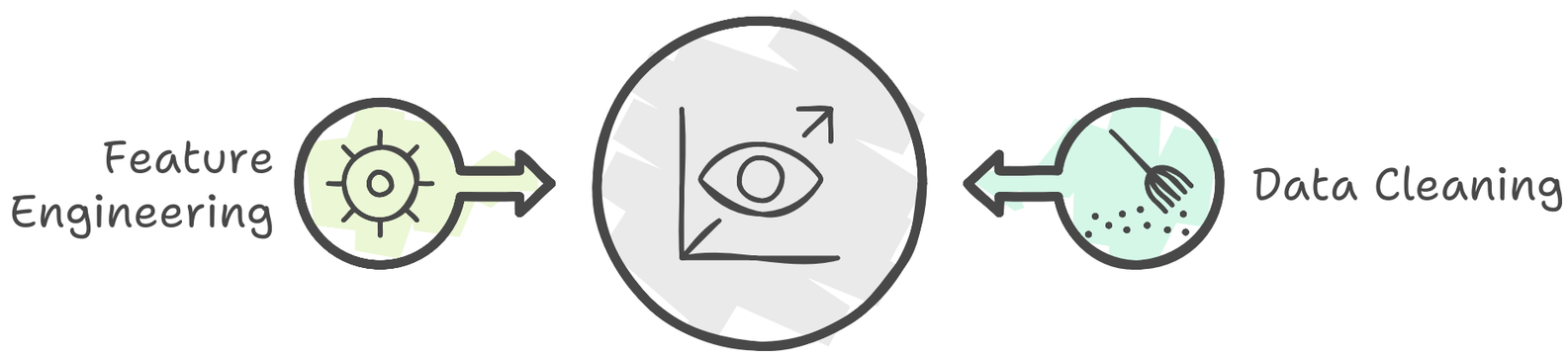
Detection of Outliers: Percentile-based outlier detection was performed over all numeric columns by calculating 1st and 99th percentiles. This will enable the separation of unusual environmental events or probable measurement errors. A separate dataset of outlier rows is created, which can be used for further investigation.

4. City-Specific Analysis :

Focused analysis was conducted on data. This depicts how the pipeline could be applied to city-specific data.

Data Cleaning and Export: Missing values were handled through ffill and interpolation. Cleaned data were also exported to a new file named "test_modi_c1.csv".

Feature Engineering (Advanced): The following features are engineered: time-based features - day, hour, month, etc. Pollution indices: basic, ratios, weighted; categorical features, time of day, weekend, season, etc. A rich feature set is important to good predictive modeling.



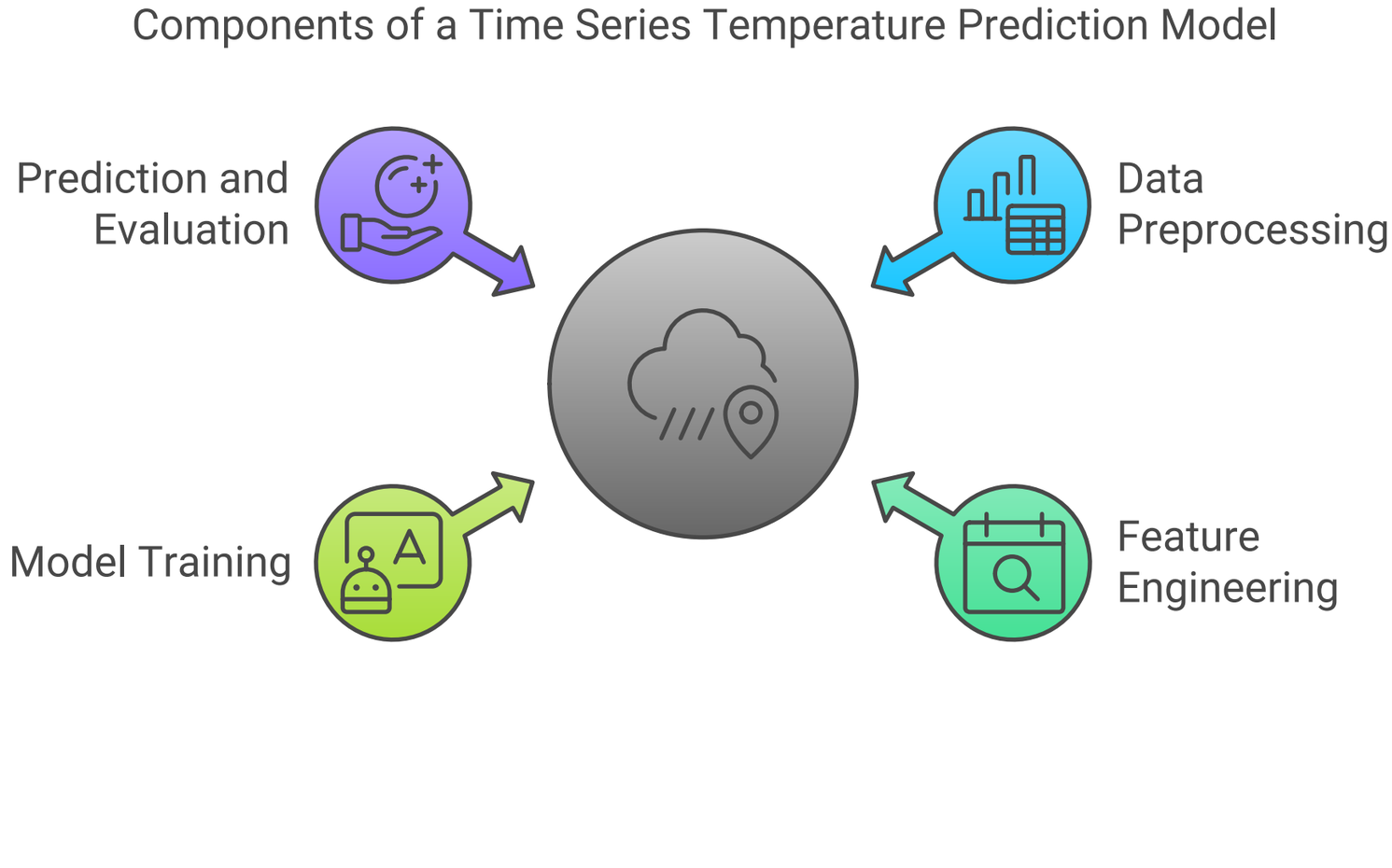
5. Time Series Temperature Prediction Model:

Preprocessing: RobustScaler for numeric features and OneHotEncoder for categorical features will ensure that data is robust and machine learning models are compatible.

Model-Specific Feature Engineering: This would include lagged features by hour, day, and week, rolling statistics, and environmental indices such as stability in the weather. This can capture time dependencies and more intricate relationships between variables.

Model Training and Hyperparameter Optimization: We used robust evaluation of the model with time series cross-validation and Optuna for automated hyperparameter tuning. The metrics to be optimized were MSE, MAE, R2 as well as direction accuracy.

Prediction and Evaluation: This model gives predictions along with confidence intervals using a bootstrap approach. Standard metrics as well as direction accuracy form evaluation. The importance analysis of features delivers insight about how the model behaves.



6. Key Insights and Future Directions:

Patterns in the data were visible and diverse cleaning methodologies used for different pollutants by the analysis. Time series analysis shows dependencies both short term as well as long term which play a central role in establishing dynamics of air quality. All outliers were identified to be tracked for further study.

Multicollinearity Analysis: Highly collinear features should be rectified to get robust modelling. To achieve this purpose, VIF analysis and matrices of correlation were both carried out using Pearson and Spearman. These will be used to select and correct the highly related features to avoid multicollinearity that could pose a problem for valid, stable coefficients of the model.

Granger Causality: The Granger causality tests were conducted to see if past values of one variable can predict future values of the other. It establishes causal relationships - or the absence of it between environment factors and temperature, and also helps detect delayed effects through multiple lags.

Cross-Correlation Analysis: In cross-correlation analysis, the two time series can be used to calculate the relationship at various lags. The lag proves ideal when the correlation is strongest, suggesting whether cause and effect have occurred at some specified time lag. Thus, visualization of cross-correlation patterns helps understand these complex relationships.

The TimeSeriesOutlierDetector class implemented a suite of outlier detection methods: STL Decomposition, Rolling Z-score, Rolling Percentiles, Isolation Forest, and Median Absolute Deviation (MAD). This diversity of methods allows for the identification of outliers with a much richer and comprehensive approach than offered through different kinds of anomalies.

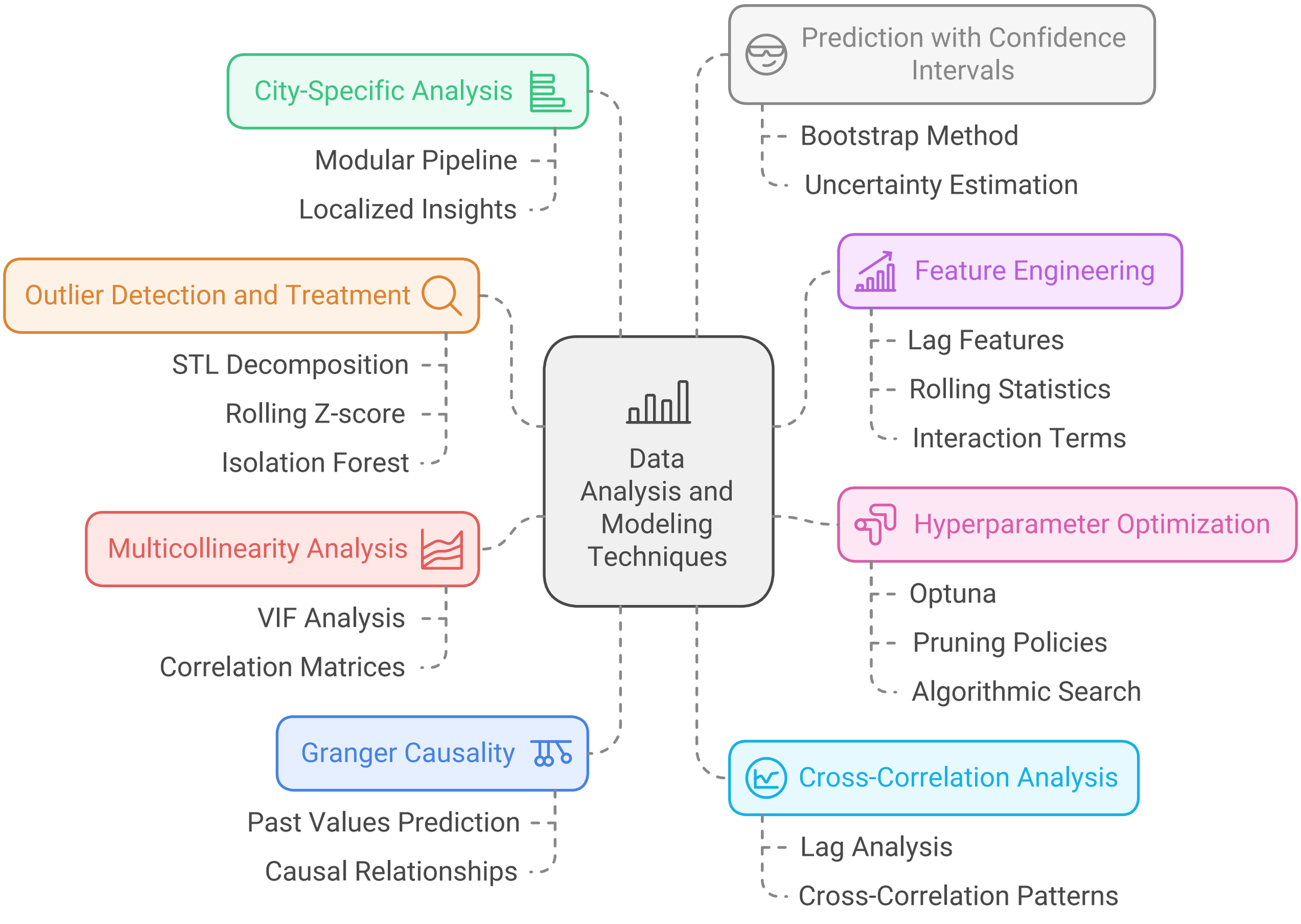
Outlier Treatment Strategies: The identified outliers were handled using appropriate techniques, for instance, interpolation and rolling median replacement. Visualization of original vs. cleaned data enables evaluation of the results of outlier removal. Statistical summaries for detected outliers provide information about the nature and extent of anomalies.

City-Specific Analysis Pipeline: The pipeline constructed for data vividly demonstrates the modular and scalable nature of this approach. This demonstrates how one can apply the analysis to cities or specific regions for localized insights and intervention.

Time Series Temperature Predictor – Feature Engineering Details: The feature engineering carried out for the predictor is impressive. Lag features are created at varied scales: hourly, daily, and weekly. Rolling statistics provide context for recent trends. Most environmental indices, like weather stability indices, summarize a number of factors with one representative index. Interaction terms capture the manner in which different variables exhibit synergistic effects. The differences by time features track the rates of change in environmental factors. These complex feature sets are at the core of the model's predictive power.

Hyperparameter optimization with Optuna In the project, the strength lies in the fact that it uses Optuna to optimize hyperparameters. Along with pruning policies, using Optuna's algorithmic search strategies assures full exploration and optimization of the hyperparameter space, which leads to a more accurate model. Most important of all, studying various metrics ensures the balance of their optimization strategy.

Prediction with Confidence Intervals: Providing confidence intervals along with predictions gives an expression of uncertainty quite important for realistic applications, and the bootstrap-based method provides a good approach to the problem of estimating such intervals.


Data Analysis and Modeling Techniques
City-Specific Analysis
Modular Pipeline
Localized Insights
Prediction with Confidence Intervals
Bootstrap Method
Uncertainty Estimation
Outlier Detection and Treatment
STL Decomposition
Rolling Z-score
Isolation Forest
Feature Engineering
Lag Features
Rolling Statistics
Interaction Terms
Multicollinearity Analysis
VIF Analysis
Correlation Matrices
Hyperparameter Optimization
Optuna
Pruning Policies
Algorithmic Search
Granger Causality
Past Values Prediction
Causal Relationships
Cross-Correlation Analysis
Lag Analysis
Cross-Correlation Patterns