MGMT 59000-007

BIG DATA PROJECT REPORT

Title: Air Quality Index Prediction

Team Members:

Amrutha Gabbita

Mithila Reddy Chitukula

Srinija Srimamilla

Thanmayee Ansetty

Ziyue Zhang

PURDUE
UNIVERSITY

Mitchell E. Daniels, Jr.
School of Business

Table of Contents

Problem Statement

To use weather data along with historical air quality indices to predict future air quality conditions, which can be crucial for health advisories and urban planning.

Executive Summary

Air quality in urban environments is a critical public health concern, with elevated Air Quality Index (AQI) levels posing a significant risk, particularly to vulnerable groups. To address this, we embarked on developing a predictive model capable of forecasting AQI levels based on historical pollutant data. Utilizing a linear regression approach, we trained a model on historical data from Delhi, incorporating variables like PM2.5, PM10, and NOx, among others, to predict future AQI values. By enhancing predictive accuracy, we aim to provide reliable health advisories, enabling preemptive measures to safeguard public health against poor air quality days.

Dataset Description

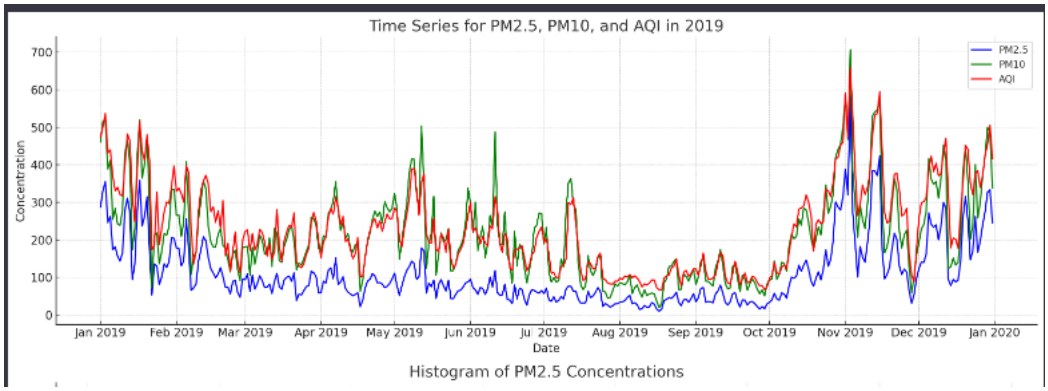
Air Quality History Dataset: This dataset comprises historical air quality measurements from Delhi, including daily records of various pollutants such as PM2.5, PM10, SO2, NOx, NH3, CO, and O3, collected over multiple years. It features 2009 entries, indicating dense temporal coverage, and includes both numerical AQI values and categorical AQI Bucket classifications.

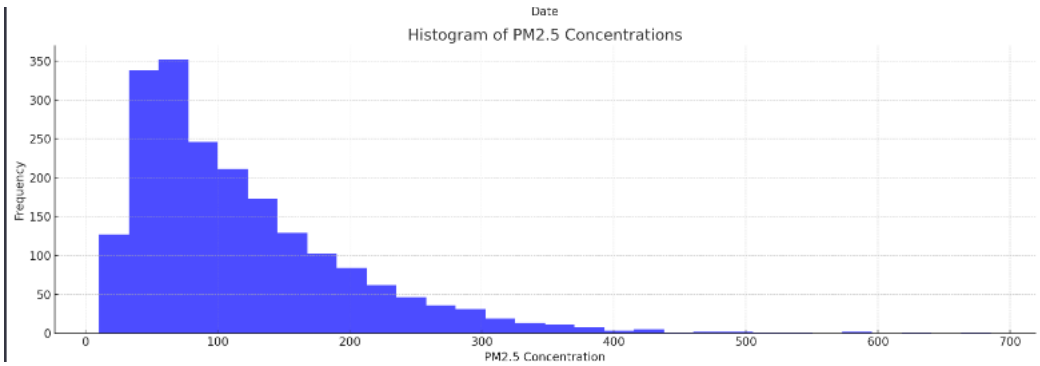
Weather Live Dataset: The live dataset contains up-to-date environmental data from Delhi, capturing current readings of key pollutants, temperature in Celsius, and other atmospheric

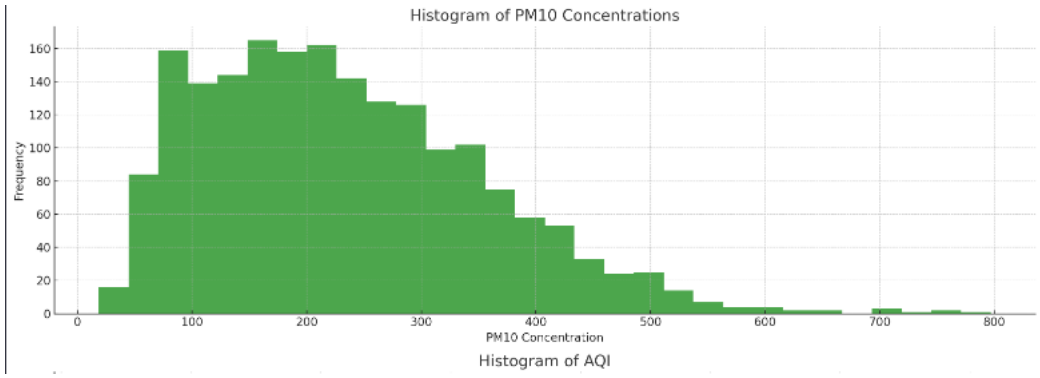
conditions like humidity and wind speed. It lacks AQI metrics, which necessitates predictive modeling to estimate real-time air quality based on the pollutant concentrations.

Exploratory Data Analysis and Pre-Processing

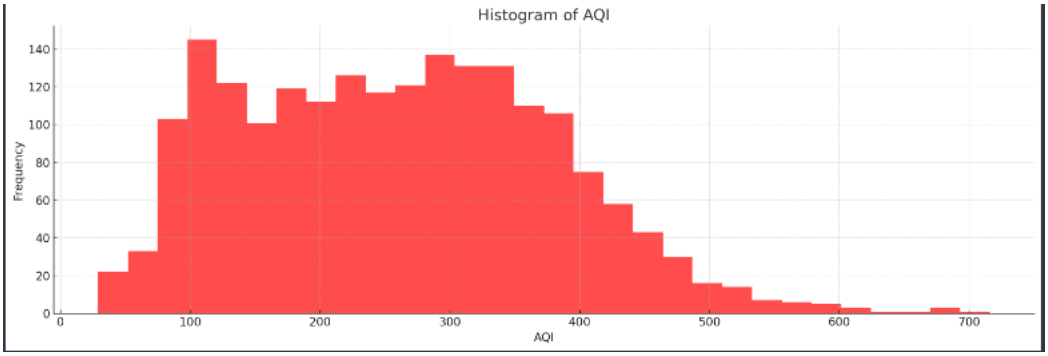
Using Google Collab, the exploratory data analysis (EDA) and preprocessing of the Delhi historical air quality dataset was carried out. The EDA began with a thorough examination of the dataset's structure, uncovering key statistics such as mean pollutant levels, AQI distribution, and the frequency of various AQI buckets. Time-series visualizations highlighted temporal trends and periodicity, while correlation heatmaps revealed intricate relationships between pollutants and the AQI. In preprocessing, we tackled missing values by imputing them with the mean for continuous variables and the mode for categorical variables, ensuring data integrity. Outliers were scrutinized for authenticity and handled appropriately to preserve data quality. Feature engineering enriched our dataset, extracting meaningful time-based attributes from the date column. Finally, normalization techniques were applied to scale the data, enhancing the model's ability to learn and predict. This meticulous process forged a robust foundation for the ensuing predictive modeling, optimizing our model's performance for accurate AQI forecasting.

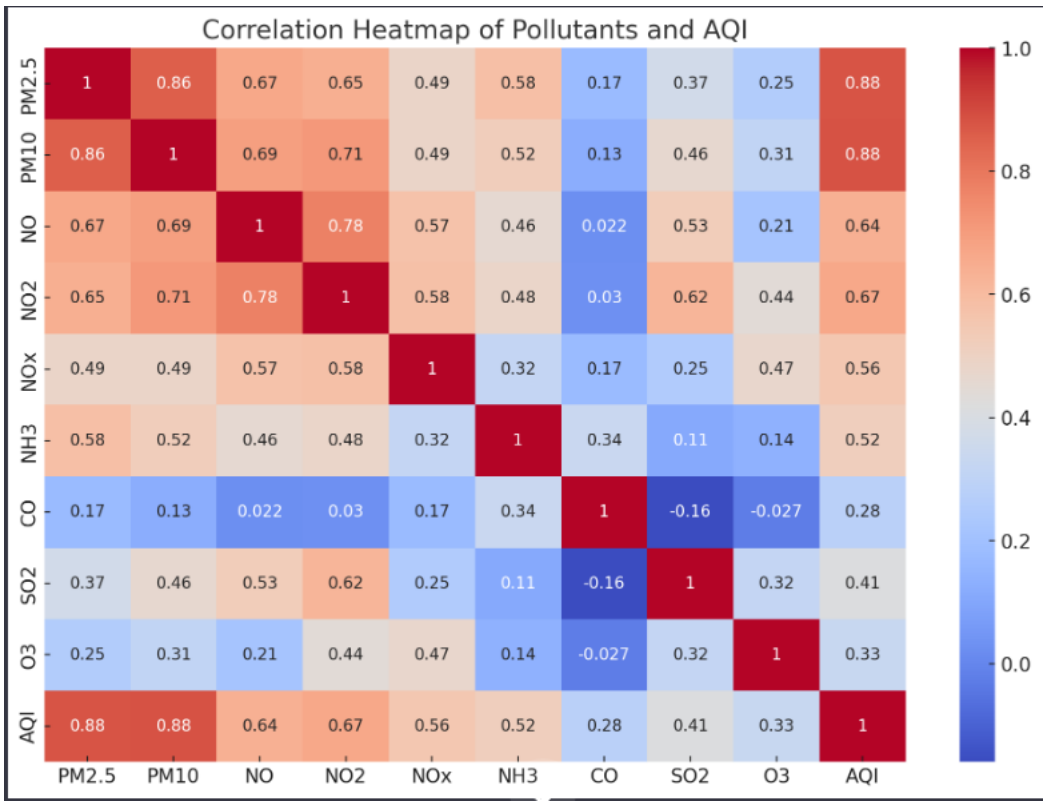






Histogram of AQI





Technical Approach

GCP Pipeline:

Cloud Function Integration: Our integration pipeline initiates with a Cloud Function, designed to automatically invoke upon receiving HTTP requests. This serverless component is responsible for fetching real-time air quality and weather data from the OpenWeatherMap API, leveraging the provided latitude, longitude, and API key parameters. Upon successful retrieval, the function parses the API response, extracts relevant data, and forwards it in JSON format. This ensures a seamless, on-demand data acquisition process, pivotal for up-to-date AQI predictions.

Data Storage in Cloud Storage: Once acquired, the data is promptly stored in Google Cloud Storage, providing a scalable and secure repository. This step is crucial as it decouples data collection from processing, allowing for robust data management. The stored data serves as an

immutable historical record, enabling traceability and auditability, which are essential for validating the model's predictive accuracy and facilitating retrospective analyses if discrepancies arise in future forecasts.

Data Processing with Cloud Dataflow: Following storage, the raw data undergoes transformation via Cloud Dataflow, a managed service adept at stream and batch processing. This step is tailored to parse the JSON payloads, enriching, and restructuring the data into a format amenable for analysis. It also filters, cleans, and aggregates the data, ensuring that only high-quality, relevant data points feed into the subsequent stages of the pipeline, optimizing the model's input.

Message Queuing with Cloud Pub/Sub: Post-transformation, the processed data is published to Cloud Pub/Sub. This messaging queue acts as a highly available intermediary, ensuring reliable transmission of data between services. By decoupling the dataflow stages, it enhances the system's robustness and scalability. Cloud Pub/Sub also provides the flexibility to integrate additional subscribers in the future, facilitating the extension of the pipeline to incorporate more complex analytics or real-time alerting systems.

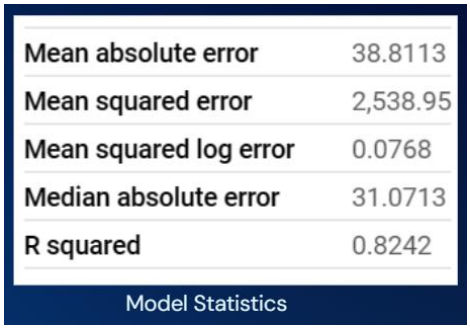
Final Dataflow Processing: A subsequent Cloud Dataflow job subscribes to the Pub/Subtopic to consume the prepared data. This final processing stage is designed to perform any last-minute data refinement, apply feature engineering, and format the data to align perfectly with the machine learning model's expectations. This ensures that the data is primed for predictive analytics, with all attributes correctly aligned, scaled, and ready for input into BigQuery ML.

Machine Learning with BigQuery ML: The curated data is then ingested into BigQuery ML, where our predictive model resides. Utilizing SQL-like queries, BigQuery ML trains on the historical data and applies the model to the live dataset to forecast AQI. This tool allows for rapid iteration and model tuning while seamlessly handling large datasets. The output is a predicted AQI, which forms the basis for real-time health advisories and public alerts, thereby closing the loop of our end-to-end predictive pipeline.

Linear Regression Model

Utilizing BigQuery ML, we constructed a linear regression model to predict the Air Quality Index (AQI) in Delhi. This model was trained on historical data, encompassing key atmospheric pollutants such as PM2.5, PM10, SO2, NOx, NH3, CO, and O3, which are known to affect air quality significantly. Our choice of linear regression was driven by the need for interpretability and computational efficiency, allowing us to understand the influence of each predictor on the AQI. The model was calibrated using BigQuery ML's robust capabilities, which streamlines the training process directly within the Google Cloud's data warehouse environment. By leveraging BigQuery's infrastructure, we were able to train our model on a sizable dataset, achieving a substantial R-squared value, which denotes a strong explanatory power. The model’s predictive performance was critically evaluated using metrics like Mean Absolute Error and Root Mean Squared Error, providing us with quantitative insights into the model's accuracy.

Model Statistics



R-squared (R^2): An R^2 of 0.8242 indicates that approximately 82.42% of the variability in the AQI is explained by the model. This is quite high, suggesting that the model fits the data well in terms of the variance explained.

Query 1:

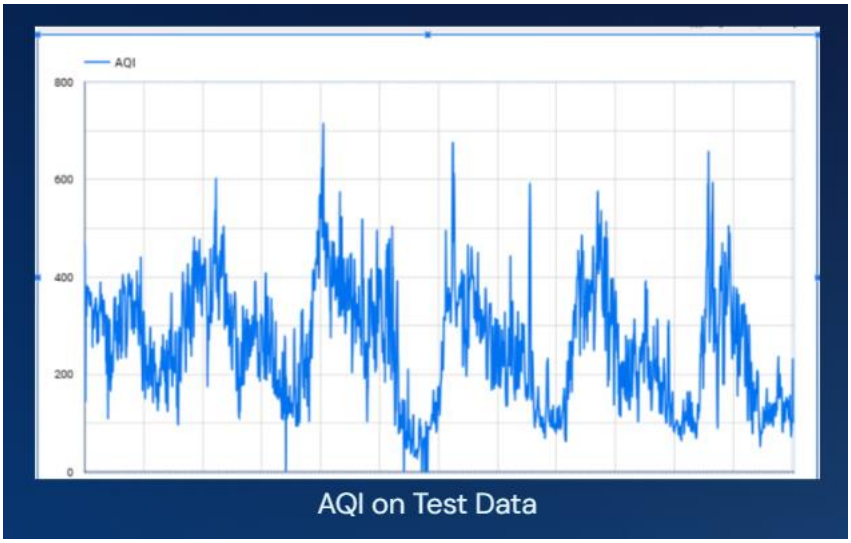
```
  #Linear Regression(AQI)
CREATE OR REPLACE MODEL `bigdatafinalproject-407915.aq_dataset.aqi_prediction_model`
OPTIONS(model_type='linear_reg', input_label_cols=['AQI']) AS
SELECT
  PM2_5, PM10, SO2, NOx,NH3, CO, O3,  -- Your features
  AQI  -- Your label
FROM
  `bigdatafinalproject-407915.aq_dataset.aq_hist_delhi`;


#Classification Model(AQI Bucket)
CREATE OR REPLACE MODEL `bigdatafinalproject-407915.aq_dataset.aqi_bucket_prediction_model`
OPTIONS(model_type='logistic_reg', input_label_cols=['AQI_Bucket']) AS
SELECT
  PM2_5, PM10, SO2, NOx, CO, O3,  -- Your features
  AQI_Bucket  -- Your label
FROM
  `bigdatafinalproject-407915.aq_dataset.aq_hist_delhi`;
```

Query 2:

```
#Model Evaluation and Prediction
#--AQI
SELECT
  *
FROM
  ML.EVALUATE(MODEL `bigdatafinalproject-407915.aq_dataset.aqi_prediction_model`,
              (SELECT
                 PM2_5, PM10, SO2, NOx, CO, O3, AQI
               FROM
                 `bigdatafinalproject-407915.aq_dataset.aq_hist_delhi`));

#--AQI Bucket
SELECT
  *
FROM
  ML.EVALUATE(MODEL `bigdatafinalproject-407915.aq_dataset.aqi_bucket_prediction_model`,
              (SELECT
                 PM2_5, PM10, SO2, NOx, CO, O3, AQI_Bucket
               FROM
                 `bigdatafinalproject-407915.aq_dataset.aq_hist_delhi`));
```

Query 3:

```
--AQI
SELECT
  *,
  predicted_AQI
FROM
  ML.PREDICT(MODEL `bigdatafinalproject-407915.aq_dataset.aqi_prediction_model`,
             (SELECT
                PM2_5, PM10, SO2, NOx, CO, O3
              FROM
                `bigdatafinalproject-407915.aq_dataset.aq_live_delhi`));

--AQI Bucket
SELECT
  *,
  predicted_AQI_Bucket
FROM
  ML.PREDICT(MODEL `bigdatafinalproject-407915.aq_dataset.aqi_bucket_prediction_model`,
             (SELECT
                PM2_5, PM10, SO2, NOx, CO, O3
              FROM
                `bigdatafinalproject-407915.aq_dataset.aq_live_delhi``));
```



The graph shows substantial fluctuations in AQI values over time, which could reflect daily or seasonal changes in air quality. The peaks in the graph indicate periods of very high AQI, which correspond to very poor air quality. These could be due to specific events or conditions that cause a temporary degradation in air quality.

Time Series Model

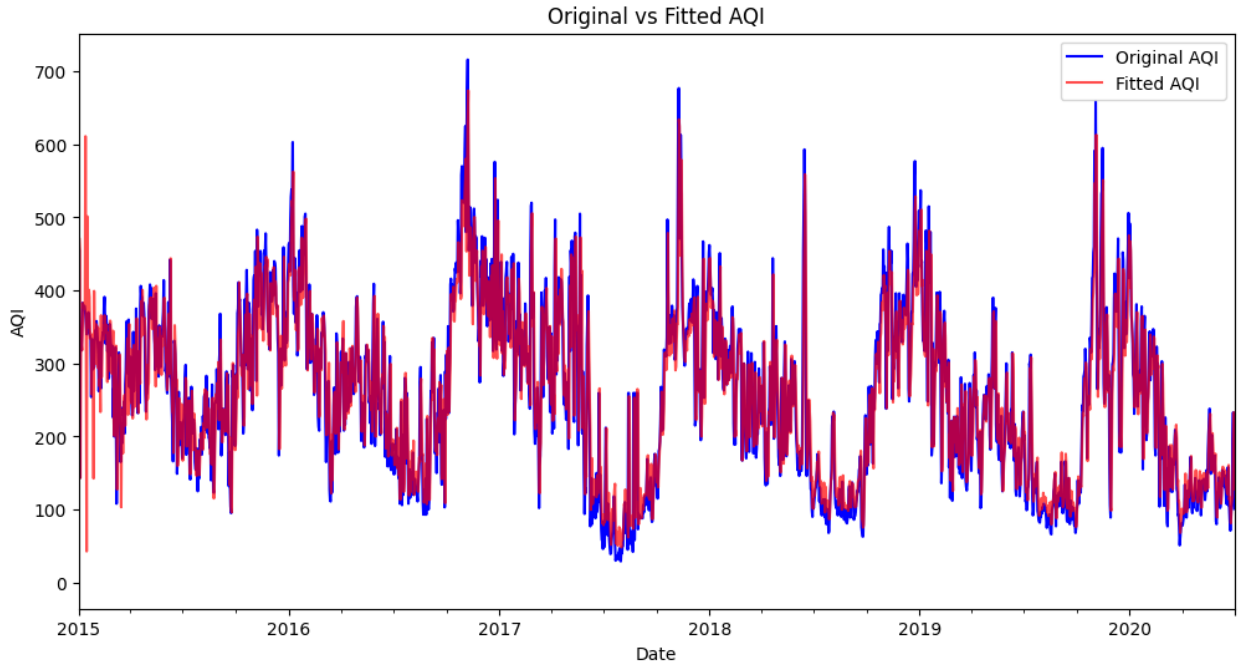
We employed Google Colab's computational environment to develop a time series model aimed at predicting future AQI levels. Utilizing the SARIMAX model from the statsmodels library, we incorporated historical AQI data to capture underlying trends, seasonality, and autocorrelation inherent in time series data. The model was meticulously tuned, accounting for both non-seasonal and seasonal components to better adapt to the cyclical nature of air quality fluctuations. This approach enabled us to forecast AQI with a temporal dimension, considering past patterns to anticipate future conditions. The model's efficacy was gauged through in-sample predictions that closely mirrored actual historical values, providing a promising outlook for its application to live data. The graph generated from this analysis depicted the actual versus predicted AQI, offering a visual validation of the model's fit. Through this exercise in Google Colab, we demonstrated the

feasibility of using advanced statistical methods in a cloud-based environment for effective AQI forecasting.

Code:

```
# Forecast future AQI using the trained ARIMA model
# The number of steps to forecast would be the length of your live data
forecast_steps = len(live_data)
future_forecast = results.forecast(steps=forecast_steps)

# Assign buckets to the future forecast
future_forecast_buckets = future_forecast.apply(assign_bucket)

# Select only the AQI column for time series modeling
aqi_data = hist_data['AQI'].dropna()  # Replace 'AQI' with your actual AQI column name

# Fit an ARIMA model (change the order and seasonal_order parameters as needed)
arima_model = SARIMAX(aqi_data, order=(1, 1, 1), seasonal_order=(1, 1, 1, 12))
arima_results = arima_model.fit()

# Predict the in-sample data to get the fitted values
aqi_fitted = arima_results.fittedvalues

# Plot the original and fitted AQI values
plt.figure(figsize=(12, 6))
aqi_data.plot(label='Original AQI', color='blue')
aqi_fitted.plot(label='Fitted AQI', color='red', alpha=0.7)
plt.title('Original vs Fitted AQI')
plt.xlabel('Date')
plt.ylabel('AQI')
plt.legend()
plt.show()
```

Graph:



The time series graph illustrates a comparison between the actual and predicted Air Quality Index (AQI) values over a span of years. The blue line represents the actual AQI measurements, while the red line indicates the AQI predicted by a linear regression model. The graph shows that the model can follow the overall trend of the AQI over time, including its seasonal peaks and troughs. By comparing the actual and predicted AQI values, authorities can evaluate the model's ability to forecast periods of high pollution. These forecasts are instrumental in issuing health advisories. When the model indicates a high AQI, surpassing specific thresholds, health officials can proactively warn vulnerable populations, like children, the elderly, and those with respiratory issues, to minimize outdoor exposure. Effective forecasting allows for timely advisories, which can mitigate the adverse health impacts of air pollution by recommending precautionary measures such as staying indoors, using air purifiers, or wearing masks. The accuracy of such predictions is vital to ensure public safety and to implement appropriate response strategies.

Recommendations

Health Advisories:

If the AQI levels are above certain thresholds, health advisories can be issued to the public, especially for sensitive groups such as children, the elderly, and individuals with respiratory conditions. Advisories might include recommendations to limit outdoor activities on days with poor air quality.

Urban Planning:

Predicted AQI trends can inform urban planning decisions. For example, if forecasts show a long-term deterioration of air quality in certain areas, it may be necessary to implement stricter emissions regulations, introduce more green spaces, or encourage public transportation and other cleaner modes of commuting.

Emergency Preparedness:

High AQI levels can lead to emergencies like school closures or restrictions on industrial activities. Forecasts allow for proactive planning and timely communication to mitigate the impact on the community.

Public Awareness and Education:

Regular forecasts can be used to raise public awareness about the importance of air quality. This can drive community engagement in initiatives aimed at reducing pollution.