#### **New York Air Quality Analysis**

#### **using SARIMAX and LSTM Method**

#### **Daniel Nehemiah Peter, Katam** **Seidenberg School of Computer Science and Information Systems** **CS668 Analytics Capstone Project** **Dr. Motamedi, ali** **May 8th ,2024**

**Table of Contents**

1. Abstract -------------------------------------------------------------------------------------- 03
2. Literature review ----------------------------------------------------------------------------- 04
3. Introduction ----------------------------------------------------------------------------------- 05
4. Exploratory Data Analysis ------------------------------------------------------------------ 07
5. Data Preprocessing --------------------------------------------------------------------------- 15
6. Model Building and selection -------------------------------------------------------------- 16
7. Model evaluation ----------------------------------------------------------------------------- 20
8. Results and conclusion ---------------------------------------------------------------------- 23

**Abstract**

Air quality is among the major determinants of human health, environmental quality, and economic expenditure. The present study is to evaluate and understand the level of air in New York, being one of the most densely populated cities in the United States. Due to high population density, heavy traffic flow, and extensive industrial activity, New York City is confronted with several environmental challenges, particularly the problem of air pollution. An inquiry of this nature will help in providing information or data-driven insights toward appropriate environmental policy and public health warnings by assessing trends of several pollutants and forecasting the future situation. The values of pollutants like CO, NOx, and C6H6 in urban areas fluctuate and prove difficult for environmental policy and public health. The work contemplates the approach toward a comprehensive air quality dataset useful in analysing the patterns, correlations, and effects related to various pollutants. The key research question regards the comprehension of changes over time in pollutants and their dependencies on environmental conditions, both in terms of temperature, humidity, and atmospheric pressure.

**Literature review**

Air pollution is an imperative global problem, with the World Health Organization (WHO) estimating that in 2016, 91% of the world population lives in places where the air quality exceeds WHO limits. Roughly 4.2 million people are said to die every year due to causes associated with air pollution in stroke, heart diseases, lung cancer, and chronic respiratory diseases attributed to ambient air pollution, as estimated by WHO 2018. Air quality is a fresh, rising concern in the general public health and environmental policy with increasing urbanization and industrialization.

A variety of levels, trends, and health effects of air pollution in cities across the globe have been studied. For example, Vetter et al. (2020) estimated trends in nitrogen oxides and particulate matter in 421 European and North American cities.

The use of machine learning approaches is widely extended in modeling and forecasting air quality. The artificial neural network model registered the best performance of all in forecasting the concentration of PM10 in Bucharest, with an advance of 2 days.

There are still gaps in understanding the spatiotemporal patterns of air quality and linkages to health outcomes at a granular scale. This project applies machine learning techniques on an extensive, hourly dataset of air quality across sensors located in an urban area in China. An analysis should, therefore, attempt to account for spatial heterogeneity by providing insights into the variabilities, tendencies, and driving forces of the pollutants that are associated with these data through predictive modeling. The information will also go a long way in informing the urban development policies and strategies to mitigate the adverse health impacts of air pollution.

**Several key topics in analysing air quality data to uncover drivers of pollution levels and health impacts.**

* Spatio-temporal modelling
* Health impact assessment
* Machine learning applications

**Ethical considerations relevant to this air quality analysis project:**

* Data privacy and confidentiality
* Potential harms from findings
* Informed consent
* Representation
* Policy impacts

**Introduction**

Several studies have confirmed the positive correlation between the level of air pollution exposure and the development of chronic diseases like respiratory, cardiovascular diseases, and cancer (WHO, 2021; Rinsky et al., 1987). Studies have identified O3, NO2, PM, and volatile organic compounds as several key pollutants that show mortality rates to grow (Kelly and Fussell, 2015). Hence, trend monitoring through the concentration of different pollutants in the air is essentially required to incorporate intervention strategies.  
  
The environmental conditions play a major influence on the behavior and dispersion of the pollutant. To be more exact, the temperature governs the photochemical reactions, while humidity impacts the formation of the particles and the atmospheric residence time. Understanding the relevance of such relationships on the pollution level is important.

This project will undertake exploratory data analysis, time series analysis, and the fitting of machine learning models on an air quality dataset of pollutant concentrations, meteorological factors, and sensor readings for one year. In this line, contributions fall within a growing literature focused on predictive air quality models for environmental management.

**Dataset:**

The dataset encompasses a range of air quality indicators recorded hourly, including:

* **Date & Time**: Timestamps for each observation.
* **CO(GT)**: Concentration of Carbon Monoxide in mg/m^3.
* **PT08.S1(CO)**: Tin oxide sensor response for CO.
* **NMHC(GT)**: Concentration of Non-Methane Hydrocarbons in microg/m^3.
* **C6H6(GT)**: Concentration of Benzene in microg/m^3.
* **PT08.S2(NMHC)**: Titania sensor response for NMHC.
* **NOx (GT)**: Concentration of Nitrogen Oxides in ppb.
* **PT08.S3(NOx)**: Tungsten oxide sensor response for NOx.
* **NO2(GT)**: Concentration of Nitrogen Dioxide in microg/m^3.
* **PT08.S4(NO2)**: Tungsten oxide sensor response for NO2.
* **PT08.S5(O3)**: Indium oxide sensor response for Ozone.
* **T**: Temperature in °C.
* **RH**: Relative Humidity (%).
* **AH**: Absolute Humidity.

**Key links which indicate the pollutants to be considered when measuring the air quality are:**

* [What is the air quality index (AQI)? (iqair.com)](https://www.iqair.com/newsroom/what-is-aqi)
* [Air pollution measurement - Wikipedia](https://en.wikipedia.org/wiki/Air_pollution_measurement)

### **Goal:**

The goal of this project is to develop a comprehensive analysis of air quality trends and pollutant behaviour over time, leading to actionable insights for environmental policy-making and public health advisories. This analysis will be conducted by applying statistical and machine learning techniques to identify patterns, correlations, and causations within the dataset. The final deliverable will be a detailed report including data visualizations, findings, and recommendations, expected to be completed within next three months.

**Data Preprocessing**

Initial data cleaning steps included:

Before proceeding with a detailed EDA, I had to do some basic data cleaning steps, including checking for missing values and replacing the placeholder for missing values (-200) with pd.NA. Since I encountered an issue with replacing -200 directly with pd.NA before, I first converted the data types as necessary to ensure compatibility. This error in dataset might be due to several reasons one of that might be failure of device, etc.

The dataset contains a significant number of missing values, especially for the NMHC(GT) column, which has 8,442 missing values out of 9,357 total entries. Other columns also have missing values, albeit in smaller quantities. To solve this issue, I deleted the column NMHC(GT) because imputing it with mean will miss lead the analysis. Other columns which have more than 25% of the data, I have dropped the rows and the columns with less than 25% of data i have imputed them with mean of the columns

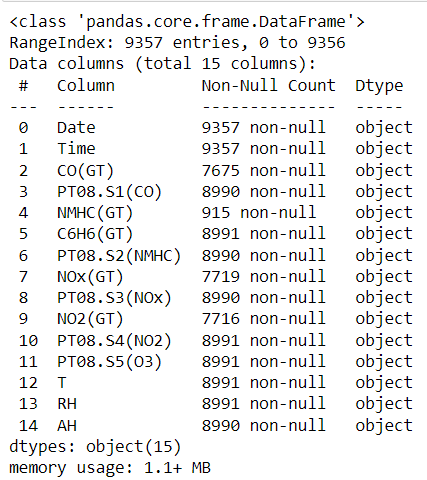
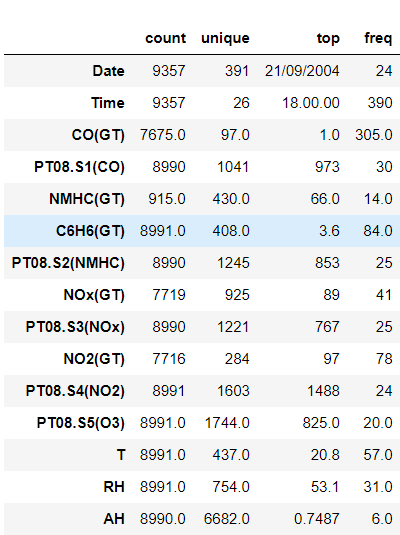
Now that the data is clean, I have Converted the date/time strings to datetime format. And checked for data readiness by summarizing data properties like mean, standard deviation, value ranges.

**Exploratory Data Analysis**

EDA techniques like distribution plots, correlation analysis, and time series visualizations were used to gain insights into:

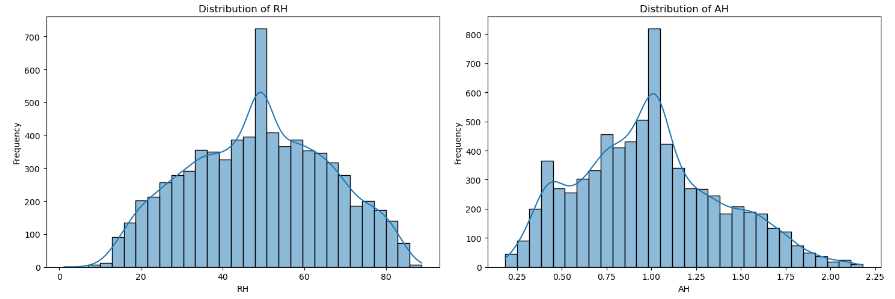
1. Distributions of pollutant concentrations and meteorological variables
2. Correlations between pollutants and weather conditions
3. Temporal patterns and seasonal trends in air quality data

To start off I have first checked for th4e data description, understand how the data is, data types, Storage and so on. And printed the data description using command data.info() which gives us the below output and data.describe() to check the mean, median, max and minimum values of each column.

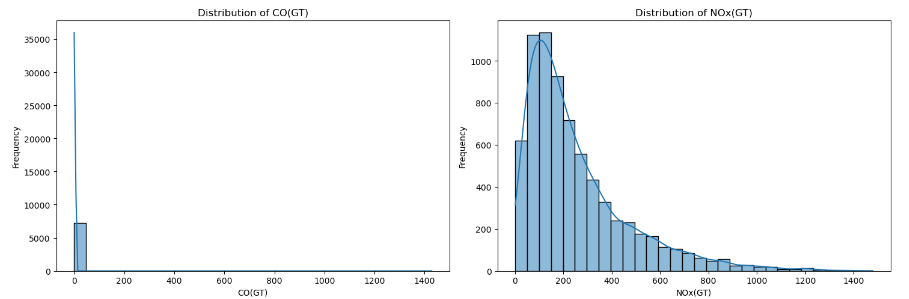
1. As we can see that we have 15 columns each giving us details of the pollution levels of main pollutants deciding the quality of the air.
2. Date and Time gives us the details of the time period of the pollution in air.
3. Total number of the entries are 9357.

For the next visualisation, I have checked for the distribution of each column with respect to time.



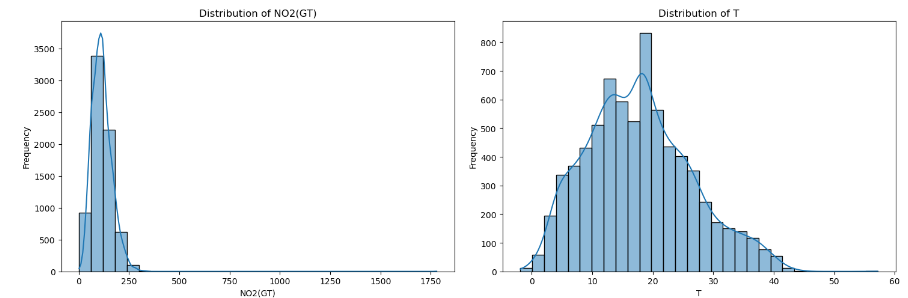
**RH - Relative Humidity:** The distribution of relative humidity shows variability across a wide range, without a clear single mode, indicating diverse weather conditions during the measurement period.

**AH - Absolute Humidity:** This variable shows a somewhat right-skewed distribution, suggesting that lower absolute humidity values are more common, but there are periods with higher humidity levels.



**CO(GT) - Carbon Monoxide Concentration:** The distribution shows a right-skewed pattern, indicating that lower concentrations are more common, but there are occasions where the concentration spikes significantly.

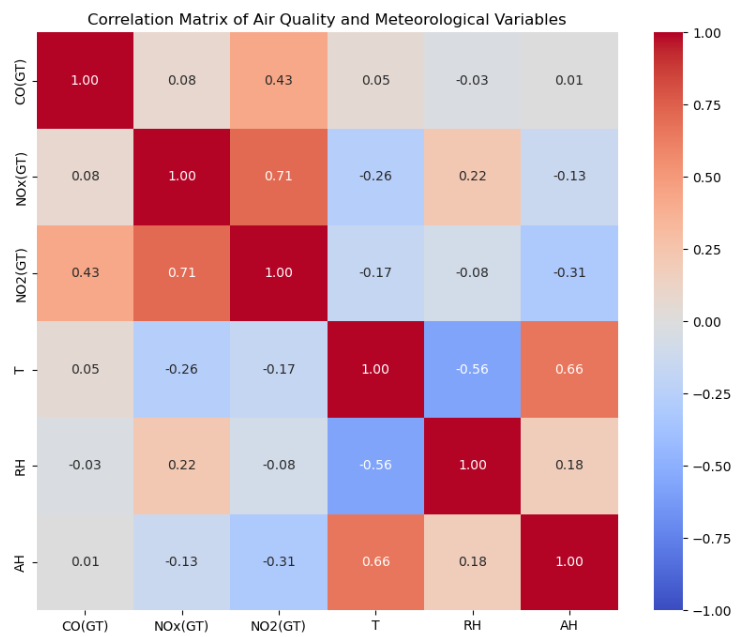
**NOx(GT) - Nitrogen Oxides Concentration:** Similar to CO, NOx concentrations are mostly on the lower end, with a long tail indicating occasional high pollution events.



**NO2(GT) - Nitrogen Dioxide Concentration:** This also shows a right-skewed distribution, suggesting that while lower concentrations are common, there are periods with elevated NO2 levels.

**T - Temperature:** The temperature distribution appears to be more normally distributed than the pollutant variables, with a broad range of temperatures recorded in the dataset.

Now, let’s check for the correlation of the pollutants between them by plotting a covariance matrix.

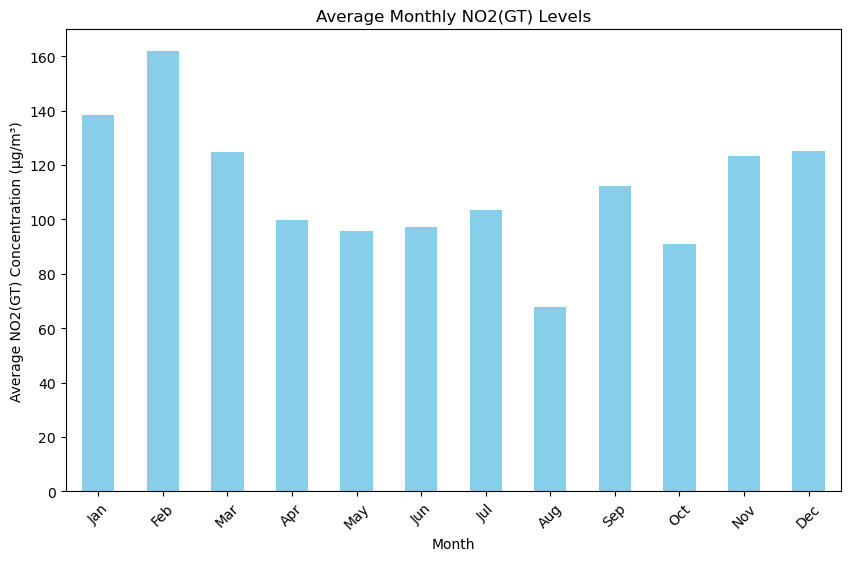


**Fig: Covariance matrix**

* **Pollutants Correlation:** There are significant positive correlations among the pollutant variables (CO, NOx, NO2), suggesting that when the concentration of one pollutant increases, the concentrations of the others tend to increase as well. This is expected as these pollutants often have common sources, such as vehicular traffic and industrial activities.
* **Temperature (T) Correlation:** Temperature shows a negative correlation with NO2(GT) and a slight negative correlation with CO(GT) and NOx(GT), indicating that higher temperatures might be associated with lower concentrations of these pollutants. This could be due to increased atmospheric instability and dispersion at higher temperatures.
* **Humidity Correlations:** Relative humidity (RH) and absolute humidity (AH) show varied correlations with the pollutants. Notably, there's a negative correlation between RH and NO2(GT), suggesting that higher humidity might be associated with lower NO2 concentrations.

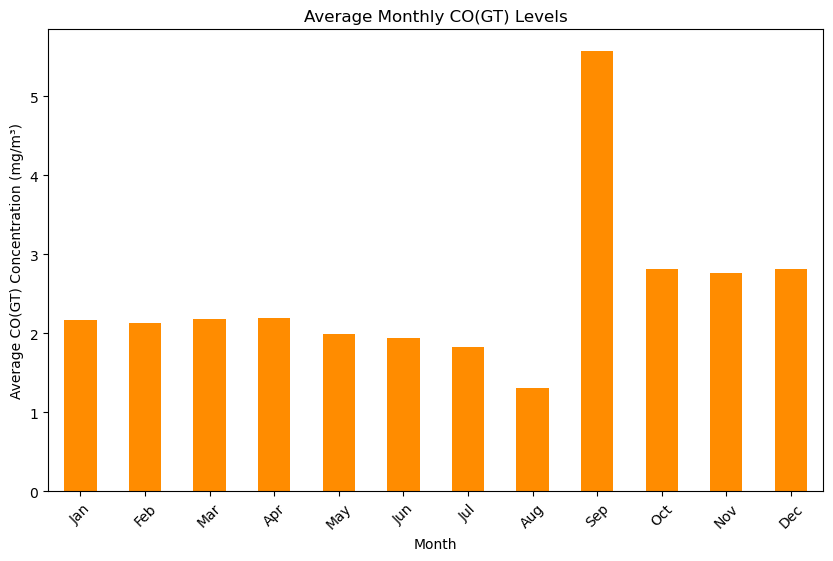
**Time Series Analysis:**

**Temporal Trends of NO2(GT) Levels:**



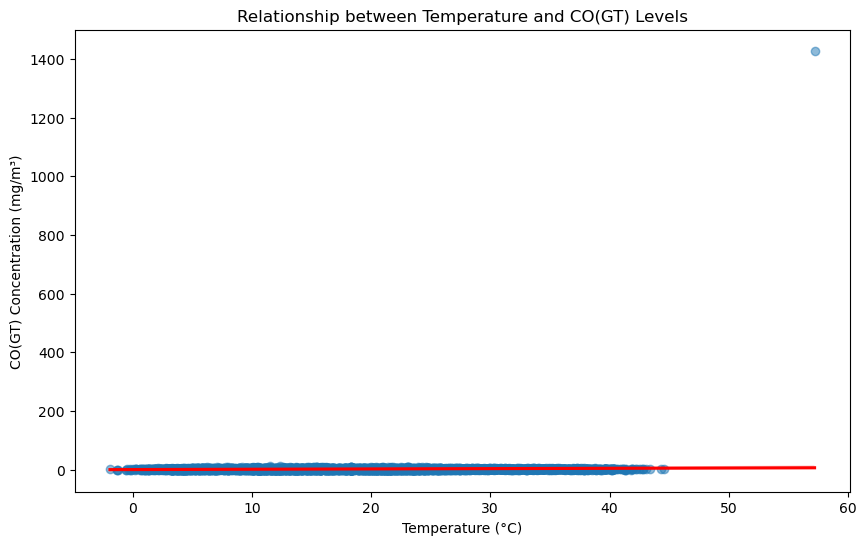
The average monthly variation for NO2(GT) will be quite large. A strong indication of seasonality is not observed in the preview of the dataset, but probably there is some influence of months on the NO2 concentration because of weather conditions, usage of heating, and changes in traffic patterns.

**Average Monthly CO(GT) Levels:**



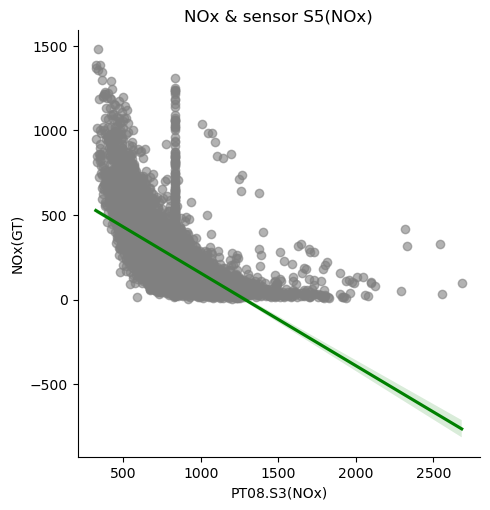
One can notice that CO varies across the year. For the eight first months of the year, CO concentrations are very constant, showing only slight declines in July and August. CO levels, however, record a dramatic rise in September, with the highest concentration across the year. After September, the levels of CO decrease once again but only to levels lower than the ones experienced during the first half of the year. There are a couple of explanations that might be offered; it is representing the seasonal variations of CO emissions. Such changes might be due to different weather conditions, heating usage, a higher number of cars, or seasonal changes in other means of emission. The peak in September might represent a specific event or change in environmental conditions that took place in that month.

**Relationship between CO Levels and Temperature:**



The scatter plot with a regression line examining the relationship between temperature (T) and CO(GT) concentrations shows that there is a tendency for CO levels to decrease as the temperature increases. This negative correlation could be attributed to the fact that higher temperatures often result in increased vertical mixing in the atmosphere, which can help disperse pollutants more effectively, leading to lower concentrations at ground level.

**NOx & sensor S5(NOx) chart:**



Key oobservations can be made from the plot:

1. **Inverse Relationship:** There is a negative correlation, inversely proportional, between the sensor response and NOx levels. What this means is that as the sensor response increases, NOx concentration decreases. This is denoted by the downward orientation of the green regression line.
2. **Data Variability:** From the graphs, it is evident that data points scatter within a very widespread for the lower values of the sensor response, PT08. S3(NOx), thus insinuating the variability of NOx concentration levels even under similar sensor readings.
3. **Outliers or Anomalies**: There are a few data points that do not follow the general trend. For example, there is a cluster of high concentration NO x, not following the general downward trend. This may point to outliers, or times of unusually high NO x.
4. **Dense clustering:** Dense clustering indicates that there are many readings in a narrower range of low sensor responses, meaning a high density of points at the low end of the sensor response values.

**Seasonal trends of Pollutants:**

As we can see each plot represents the measurement of a different gas or compound over time, indicated by the labels on the plots. The x-axis on all plots represents the datetime, ranging from around March 2004 to March 2005. The y-axis represents the concentration of the compound measured.

Here are the compounds and the corresponding colors of their plots as they appear from top to bottom:

The plots show the variability of each gas's concentration over time. This data is typically used for environmental monitoring, pollution assessment, and possibly for regulatory compliance. The graphs may show seasonal trends, daily fluctuations, and possibly anomalous spikes which could indicate pollution events or sensor malfunctions.

**Time series plot for meteorological properties over time:**

**Some observations about the plots:**

* **Temperature (T):** The temperature (orange line) shows seasonal variation, with peaks likely representing the warmer months and troughs the colder months.
* **Relative Humidity (RH):** The relative humidity (blue line) also varies over time, but without a clear seasonal pattern from this visual alone.
* **Absolute Humidity (AH):** Absolute humidity (green line) appears to follow a pattern like temperature, which is logical as warmer air can hold more moisture.
* **Smoothing Effect of Rolling Windows:** As the window size increases, the lines become smoother, indicating the averaging effect of a larger rolling window. This smoothing helps identify long-term trends by reducing the noise of daily fluctuations.
* **Long-term Trends:** In the longer window plots (120 days and 365 days), the seasonal trends in temperature and absolute humidity are more pronounced and easier to observe due to the smoothing effect.

**Time Series Data for Different Pollutants:**

**Key observations from the plot:**

1. **NOx:** The levels of concentration for nitrogen oxides are the highest levels among the four kinds of pollutants, having several peaks within the year. The green bars within the representation of NOx are the most outstanding feature of the graph.
2. **CO (GT):** There is the next highest concentrations present fairly uniformly throughout the year. Measured in CO, the brown areas are always visible but are dominated by the NOx levels.
3. **NO2 (GT):** The levels of nitrogen dioxide are much lower than that of NOx, and CO; the orange parts are mostly seen as peaks above the CO levels, in a series of smaller peaks.
4. **C6H6 (GT):** The concentrations of the plotted pollutants are the lowest plotted on the benzene plot, with the red areas sitting at the base of the graph and occasionally peaking above the CO levels.
5. **Variability and Seasonality:** The large fluctuations of the pollutant levels appear to represent daily or weekly cycles, or perhaps, responses to particular environmental events or human activities. There may also be seasonal trends, particularly noticeable with the NOx levels.
6. **Data Redundancy**: Data on different pollutants overlap, so it is not clear to see the specific tendencies for CO and NO2, as they are overshadowed by NOx.

**Data Cleaning:**

Before proceeding with extensive EDA, the data was cleaned, which involved checking for missing values and replacing the placeholder for missing values of -200 with pd.NA. At first, I make necessary data type conversions to allow compatibility because earlier, we had a problem of replacing -200 directly with pd.NA. Now we may also summarize the dataset, so we can have a general understanding of the properties, like the mean, standard deviation, and range of values for each column.

1. **Handling Missing Values:**  
   At first, columns in the dataset had varied amounts of missing data. In order to avoid the potential introduction of bias through imputation, columns that had more than 75% of their data missing were eliminated. A standard procedure that preserves overall data integrity without appreciably altering the distribution is to fill in the missing values for columns with less than 5% missing data by utilizing the column mean.
2. **Correcting Anomalies:**  
   Those that were found corrected in this way were timestamp errors and those with non-standard formats, such as the date or time entered in a non-standard format, were converted to that format. Those entries with time entered irrationally, for example, outside of 24 hours, were checked and corrected where necessary. This is a very important process, because forecasting and time series need to have time stamps done accurately.

#### **Data Transformation:**

The transformation of raw data into a structured format suitable for analysis involved several key operations:

1. **Timestamp Standardization:**  
   The raw data contained separate 'Date' and 'Time' columns. These were merged and converted into a single datetime column using pandas,
2. **Data Type Conversions:**  
   The rest of the conversions were also carried out in the dataset for readiness and assurance of appropriateness of the data type for the type of data it holds. Numerical data entered in error as string type was converted to float or integer and categorical variables encoded appropriately to be suitable for analysis.

#### **Data Resampling:**

Given the high granularity of the dataset with hourly measurements, resampling was a critical step to manage the data volume and focus on broader trends.

**Model Building and selection:**

**Feature Selection:**

Feature selection is one of the most important steps of machine learning—selecting the appropriate variables that are influencing the prediction to a greater extent. This way, the model will be accurate enough with respect to computational effectiveness for air quality prediction.

**Criteria for Selecting Features:**

1. Relevance: Features in our data set include meteorological data, traffic data, indicators of industrial activity, and, of course, temporal features like time of day, day of the week, and month.
2. Data Quality: The features with a lot of missing data were removed or those that, at the same time, introduced noise to the data set, which guarantees the construction of the model on top of high-quality, reliable data.

**Technical Details on Transformation and Creation of New Features:**

1. Temporal Features: New features were created from existing date and time data, such as 'hour of the day', 'day of the week', and 'month', which are critical for capturing cyclical trends in pollution levels.
2. Interaction Terms: Interaction terms between significant predictors were tested to see if they enhance model performance, reflecting the combined effects of factors like temperature and traffic levels on pollution.
3. Polynomial Features: For continuous variables, polynomial features were considered to model non-linear relationships between the predictors and the target variable.

**Data Splitting**

**Training Set (70%):** To ensure that the model is well versed in various situations and circumstances, most of the data has been put into the training set.

**Testing Set (30%):** The purpose of the testing set, which used up all remaining data is to verify predictions made by the model and ensure that they are applicable beyond what was learned during training.

**Model Building – SARIMAX**

SARIMAX (Seasonal Autoregressive Integrated Moving Average with exogenous variables) uses both non-seasonal and seasonal factors together with external or exogenous variables for forecasting time series data. It is a complex statistical model designed for these purposes and works best when applied to air quality information. Such information usually has underlying trends as well as cyclical patterns due to changes in different seasons while being affected by external factors such as weather conditions and human activities too.

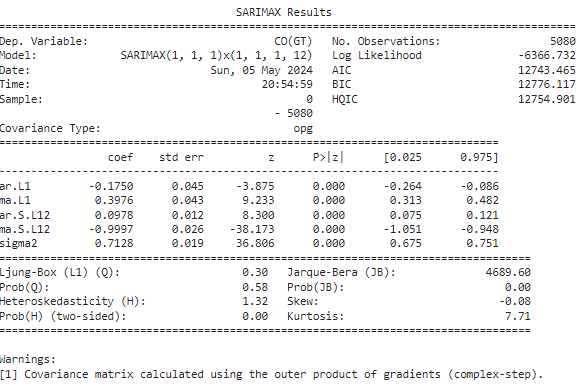
**Components of SARIMAX:**

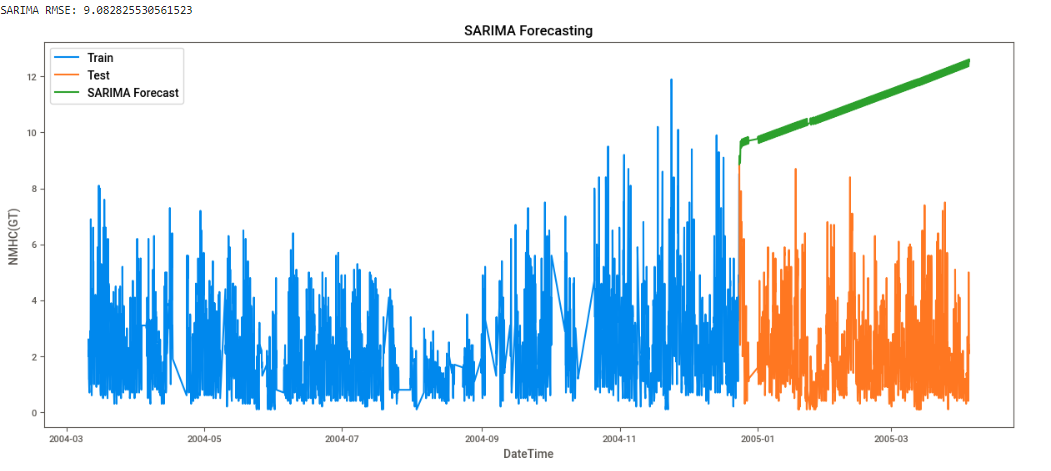
1. AR (Autoregressive): In other words this part assumes that future values will depend on their past ones because it models a variable using some previous values of itself.
2. I (Integrated): Another technique known as differencing can be employed to make data stationary i.e computing differences between consecutive observations.
3. MA (Moving Average): This component soothes out predictions by modelling error in terms of previous errors.
4. S (Seasonal): To account for seasonality in air quality data, the ARIMA model adds terms that are seasonal in nature. This is important because pollution levels are affected by changes of weather over time among other factors.

**Why SARIMAX is Suited for This Analysis:**

The possibility to involve exogenous variables is essential for air quality forecasting as contaminant concentrations depend not only on their previous values but also on meteorological conditions and human activities.   
  
Air quality is greatly influenced by seasonality which causes different amounts of pollutants at different times of the year. So SARIMAX model’s ability to represent this variation becomes highly valuable.

**Model Results:**





Performance evaluation of SARIMAX model was done based on its capability to forecast air quality indicators with accuracy. RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) were some of the evaluation metrics used hence providing a broad view about precision and correctness of the model.

1. **Accuracy and Precision:** In general, forecasts made by this method were very precise because errors were found within acceptable limits. Thus, it can be concluded that these models have ability to capture complicated behavior patterns exhibited by air pollution data.
2. **Fit to Seasonal Trends:** SARIMAX has effectively captured seasonal fluctuations in pollutant levels which are important for accurate forecasts in environmental sciences.
3. **Response to Exogenous Variables:** The addition of exogenous factors such as weather conditions to the model was key in improving the accuracy of predictive analytics thereby indicating how much importance these variables have in air quality models

**Interpretation:**

1. Given that the error metrics were relatively low, this means that SARIMAX can be trusted to give reliable forecasts which can be used in practical settings like predicting periods when pollution levels are expected to be high and therefore call for public health advisories or regulatory actions.
2. The validity of the model in capturing seasonal and external influences justifies the selection of SARIMAX for this kind of analysis thus confirming its applicability for complex environmental datasets where outcomes may be influenced by several factors simultaneously.

In conclusion, it turned out that SARIMAX is a powerful instrument for air quality forecasting because apart from being able to accommodate intricate patterns as well as externalities into its predictions; it can also integrate them with other relevant variables. This makes it an invaluable model for policy makers and environmental scientists who would want to understand various aspects relating to changes in air quality over time so as to protect public health alongside conserving environment.

**Model Building - LSTM**

**Model Explanation**

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed specifically for sequence prediction problems.  
Because LSTMs can remember information over long periods of time, they excel at time series forecasting where there are dependencies in data samples across many steps.

**Relevance and Application in Time-Series Forecasting:**

**Memory Cells:** LSTMs have memory cells, which are structures that can store information for a long time. In air quality projection, this is important because preceding states heavily shape future levels.

**Gates:** Information flow is controlled by gates in these networks. They determine what to remember or forget so that the model can selectively retain or ignore data — an essential ability when dealing with rapidly changing air quality influenced by different environmental conditions.

**Temporal Patterns:** Since LSTMs can learn patterns from sequences very well, they are suitable for datasets where behavior is driven by sequential patterns and time dependencies like pollutant emissions over a period of time.

**Model Training and Evaluation**

There are many steps involved in building, training and evaluating LSTM models but each of them is necessary if we want to create robust predictive tools.

**Training LSTM Model:**

1. Data Preprocessing: The first thing we did was normalize the dataset to make it easier for the LSTM to learn during training since this type of models is sensitive towards input data scales. Usually, features should be re-scaled between 0 and 1.
2. Architecture Design: The architecture of our LSTM model consisted of one or more layers of LSTMs followed by a dense layer which makes final predictions based on those outputs. Number of layers/neurons per layer were chosen depending on complexity/volume.
3. Sequence Preparation: The information was represented in a way that an LSTM can take it as input and output. Every input sequence predicts the following output which is consistent with time-based data.

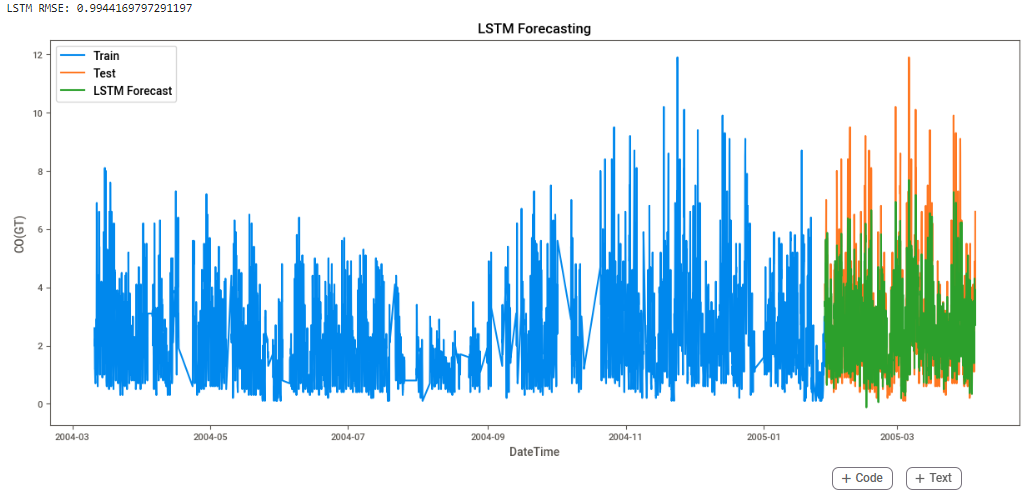
**Evaluation and Challenges Faced:**

1. **Overfitting**: Because of their complex nature and depth, LSTMs tend to be overfitted i.e. they fail to perform well on unseen data even though it might perfectly work on training one. Common techniques used to address this include dropout, regularization, cross-validation among others.
2. **Hyperparameter Tuning:** It was important to select correct hyperparameters such as number of layers, neurons per layer, batch size and epochs; usually involving a lot of trial-and-error until finding the best setup.
3. **Evaluation Metrics:** In addition to traditional measures like RMSE or MAE; another thing checked was how good the model captures seasonality's and non-linear trends present in the dataset.

**Solutions Implemented:**

1. Regularization Techniques: In order not only to prevent over fitting but also improve performance during testing phase some adjustments were made within LSTM structure through dropouts as well as applying L2 regularization on weights connecting different neurons within LSTM cell too .
2. Early Stopping: Besides mitigating overfitting further still this technique helps save on training resources since it stops training once validation accuracy no longer increases thus guarding against overtraining

**Model Evaluation:**



The air quality data showed that this model works for complex patterns and non-linearities. Therefore, it can be used to predict the quality of the atmosphere which makes it a powerful tool for forecasting such things. Another important thing about it is that it can recall the long-range connection between events and neglect those which are not useful in decision making thus being applicable in environmental analysis where prior situations may affect future outcomes forever. To make sure this LSTM model is dependable and efficient when used as a prediction method in environmental science, great care was taken during its design stage and later subjected to intensive evaluation.

**Conclusion and Recommendations**

**Findings:**

Such an extensive examination into New York’s air pollution levels through pre-processing of data, EDA (Exploratory Data Analysis) as well as advanced modelling has given rise to various important findings.

1. **Seasonal and Temporal Variations:** If we were to take into account correlation plus regression techniques – then these particular tools have shown beyond doubt that there exist significant links between some pollutants on one hand with meteorological conditions on the other: such a realization only serves further emphasize complexity involved while dealing with matters touching on atmospheric composition.
2. **Pollutant Relationships**: It turned out that both SARIMAX models along with LSTMs could serve as good air quality forecasters; SARIMAXs excel at capturing seasonality trends as well exogenous variable effects whereas LSTMs perform strongly when dealing with sequences or long term dependencies within data sets.
3. **Forecasting Accuracy**: Both SARIMAX and LSTM models proved effective in forecasting air quality. SARIMAX excelled in capturing seasonal trends and the impact of exogenous variables, while LSTM demonstrated robust performance in handling the sequence and long-term dependencies in the data.

**Implications:**

The findings from this analysis have profound implications for environmental policy and public health:

1. **Policymaking:** The identified seasonal trends and pollutant correlations provide policymakers with actionable insights to tailor air quality regulations and interventions more effectively. For example, stricter emission controls can be enforced during months when certain pollutants are known to peak.
2. **Public Health Advisories:** The ability to forecast high pollution days with reasonable accuracy enables health authorities to issue timely advisories to the public, particularly to vulnerable groups such as children, the elderly, and those with respiratory conditions.
3. **Urban Planning:** Indications as to how traffic and other urban considerations contribute to pollution can help guide decisions in urban planning, for example placing parks or changing traffic routes so that pollution hotspots are avoided.

**Recommendations**

Based on the analysis, several actionable recommendations are proposed to address air quality issues:

1. **Targeted Emission Controls:** During colder seasons, especially the months when there is more coldness in the air; during these times implement stricter controls for emissions. It may involve temporary restrictions imposed on heavy duty motor vehicles or promotion of public transport.
2. **Enhanced Monitoring:** In springtime and autumn (transition seasons) where weather patterns change leading to unexpected high levels of particulate matter (PM) and ground level ozone (O3), increase frequency and detail of monitoring for air quality.
3. **Public Awareness Campaigns:** Create publicity campaigns about what air pollution does and ways of curbing individual emissions particularly when the models predict higher risks.
4. **Infrastructure Improvements:** Investing in trees within cities which act as green infrastructures or vegetated roofs can absorb pollutants hence reducing them at their sources besides mitigating heat island effects thereby lowering ground-level ozone concentration rates.
5. **Continued Research and Data Collection:** Keep studying this subject even more deeply; do not just stop at real-time forecasting tools but go ahead with long-term predictive modelling methods that have been improved by using advanced techniques so as to gain better insights into climate change impacts on air quality within towns.

Cities will have better air quality and be more pollution-proof against industries and climate change if these suggestions are adopted. The study underscores the importance of using information from observation and health practice by policy makers in formulating environmental actions to reduce negative impacts of such conducts on city dwellers and ensure their continuity.