Introduction

- ✓ This presentation focuses on the **analysis and modeling** of air quality data in **Beijing**.
- ✓ The analysis utilizes air quality data from 12 monitoring stations in Beijing, covering the period from March 1, 2013, to February 28, 2017
- ✓ The study utilizes data from 12 monitoring stations over a four-year period to predict PM2.5 concentrations.
- ✓ It’s objective is to develop a software application for predicting air quality, focusing on PM2.5, based on meteorological and environmental factors.

Steps Involved:

- ✓ Data collection & preprocessing
- ✓ Exploratory Data Analysis (EDA)
- ✓ Machine learning model development
- ✓ Application deployment for real-time predictions

Data Handling

- ✓ **Data Source**:
  - ✓ 12 monitoring stations in Beijing (2013-2017), 420,768 hourly records.
- ✓ **Challenges**:
  - ✓ Having to precheck the data if they are consistent across all file
  - ✓ Missing values (e.g., wind direction 'wd' and pollutants).
  - ✓ Data integration from multiple CSV files.
- ✓ **Solution**:
  - ✓ Use **pandas** for data cleaning (imputation, forward fill) and merging data from all stations into a unified DataFrame.
  - ✓ For analysis, **seaborn and matplotlib** was used for plotting graphs which summarizes information about the data
  - ✓ **Scikit-learn** was used for machine learning as it contains modules necessary for this with additional ones like xgboost.

Data Information

- ✓ It includes pollutant concentrations, weather, and station details.
- ✓ Hourly records for each station is merged into a DataFrame for analysis.
- ✓ Each station had 18 columns with 35064 records which totalled to 420,768.
- ✓ Each record represents measurements taken at a specific time and location in the city..
- ✓ All columns are numerical (consist of float and integer) except for wind direction (wd) and station. Here are the columns
  - ✓ **Pollutants: PM2.5, PM10, NO2, SO2, CO, O3**
  - ✓ **Meteorological features: TEMP, PRES, RAIN, WSPM**
  - ✓ **Time-based information: Year, Month, Day, Hour**

	No	year	month	day	hour	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	wd	WSPM	station
0	1	2013	3	1	0	9.0	9.0	3.0	17.0	300.0	89.0	-0.5	1024.5	-21.4	0.0	NNW	5.7	Dongsi
1	2	2013	3	1	1	4.0	4.0	3.0	16.0	300.0	88.0	-0.7	1025.1	-22.1	0.0	NW	3.9	Dongsi
2	3	2013	3	1	2	7.0	7.0	NaN	17.0	300.0	60.0	-1.2	1025.3	-24.6	0.0	NNW	5.3	Dongsi
3	4	2013	3	1	3	3.0	3.0	5.0	18.0	NaN	NaN	-1.4	1026.2	-25.5	0.0	N	4.9	Dongsi
4	5	2013	3	1	4	3.0	3.0	7.0	NaN	200.0	84.0	-1.9	1027.1	-24.5	0.0	NNW	3.2	Dongsi

Data Exploration and Precheck-up

Data Pre-Checkup

- ✓ **Key Insights from Missing Data:**
  - ✓ The pollutants (PM2.5, PM10, SO2, NO2, CO, O3) have significant missing data, especially CO.
  - ✓ Wind direction (wd) has 18% missing data, which could impact wind-related analysis.
  - ✓ Other columns like TEMP, PRES, DEWP, RAIN, and WSPM have relatively few missing entries.
  - ✓ Year, month, day, hour does not have any missing values (0 missing)
- ✓ **Checking data types:**
  - ✓ All columns are numerical (consist of float and integer) except for wind direction (wd) and station
- ✓ **Generating basic statistics to summarize the data**

Exploration

Statistical Summary

- ✓ **Pollutants i.e** PM2.5, PM10, SO2, NO2, CO, and O3 show significant variation, with CO having the highest mean of 1230.77 and maximum of 10,000. PM2.5 and PM10 have large ranges, with values up to 999.
- ✓ **Meteorological Variables:** Temperature ranges from -19.9°C to 41.6°C, with an average of 13.54°C. Pressure shows minor variation, with values between 982.4 and 1042.8 hPa.
- ✓ **Wind and Rain:** Wind speed averages 1.73, and rain shows a very low mean of 0.06, indicating limited precipitation in the dataset.

Visualizing data distributions

- ✓ In order to understand the trends in depth, analyzing relationships between variables was done using tabular and graphical methods to observe trends

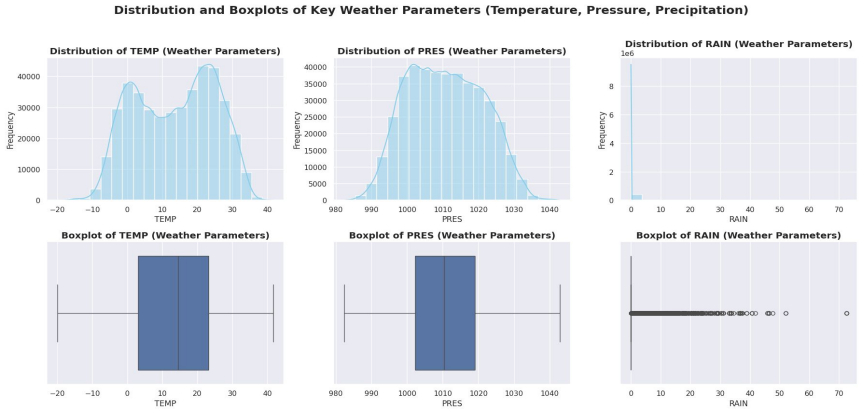
Data Preprocessing

- ✓ **Missing Values**:
  - ✓ Numerical data imputed with the mean.
  - ✓ Wind direction (`wd`) filled using forward fill.
- ✓ **Feature Engineering**:
  - ✓ Combined Year, Month, and Day into a single `datetime` column.
  - ✓ Dropped the 'No' column (record ID) for simplicity.

Data Analysis and EDA

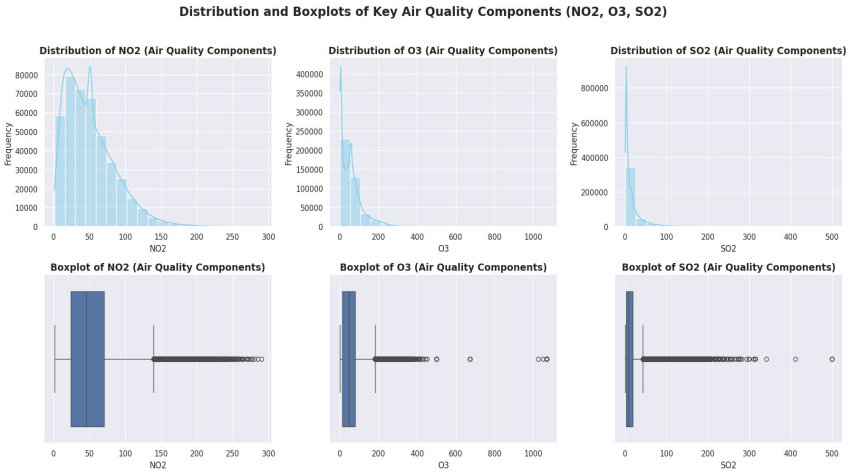
Key FIndings on Distributions

- ✓ The temperature distribution is likely bimodal, which is also near-normal, centred around the mean with no outliers.
- ✓ The pressure distribution is near symmetrical, with small changes around a mean value of approximately 1010. No outliers are identified.
- ✓ Rainfall distribution is very right-skewed, showing multiple outliers due to a large deviation in values. Most data points show little or no rainfall, while a smaller number of days have heavy rainfall.



Key FIndings on Distributions for NO2, O3 and SO2

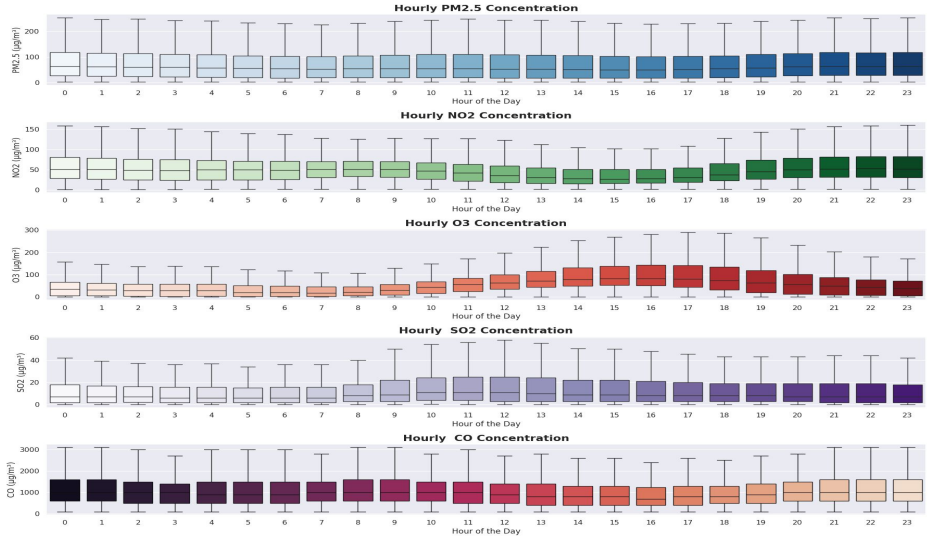
- ✓ The NO2 distribution is right-skewed, showing higher concentrations during winter and lower levels during warmer months, with fewer outliers.
- ✓ The Ozone distribution is right-skewed with a very large variation, hence more outliers.
- ✓ SO2 is right-skewed with a majority of days showing low concentrations, but some outliers indicating pollution spikes.



Trend Analysis

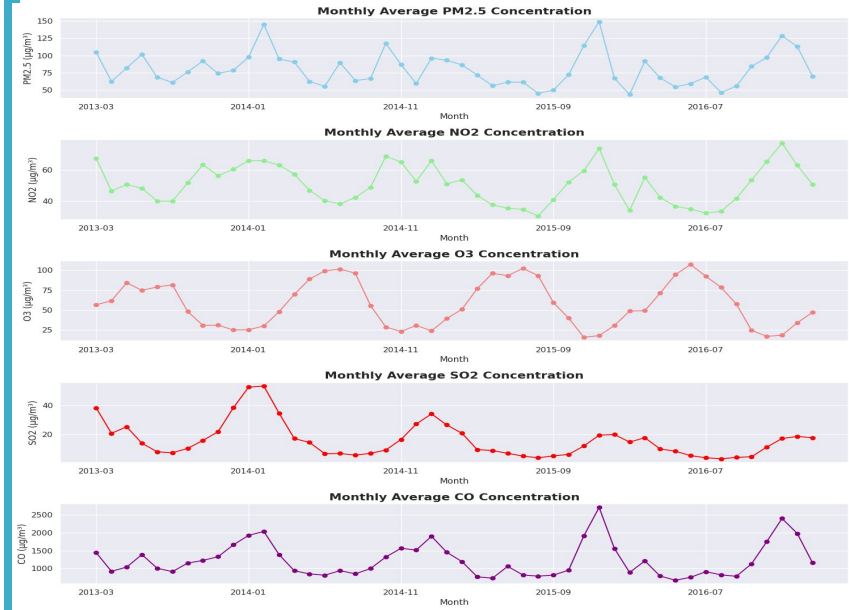
Monthly Trend Analysis.

- ✓ Higher pollution levels for PM2.5, CO, and NO2 are observed during peak hours (7 AM to 9 AM and 5 PM to 7 PM).
- ✓ SO2 levels are elevated from 7 AM to about 6 PM.
- ✓ Ozone (O3) has higher levels between 11 AM and 7 PM, peaking in the afternoon (1 PM to 6 PM).



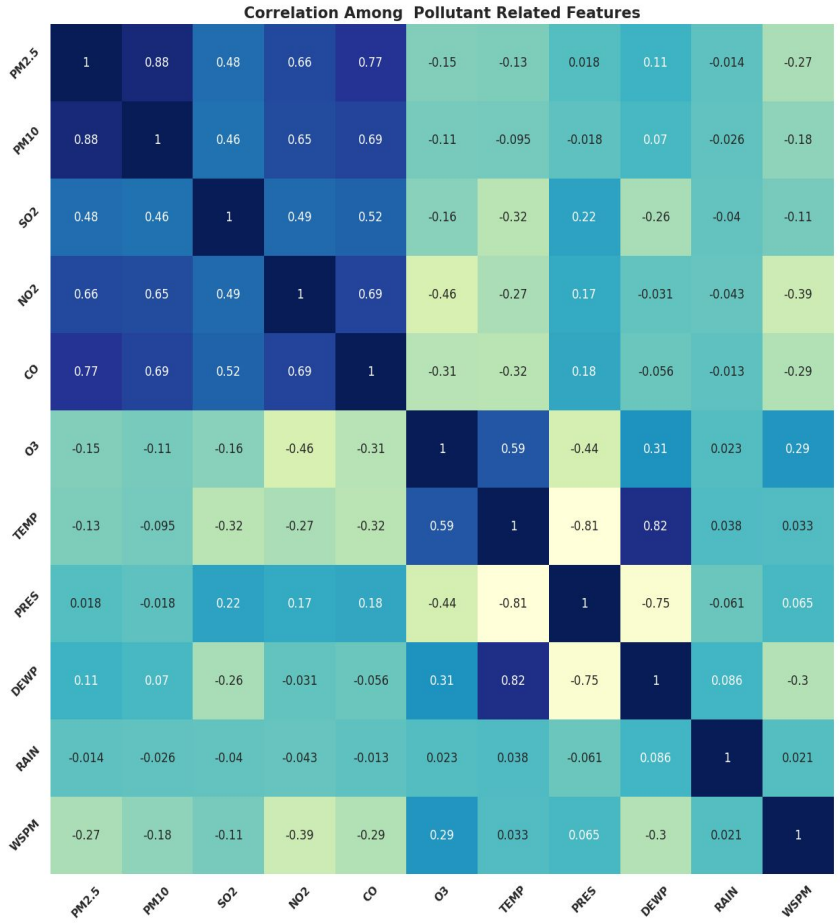
Monthly Trend Analysis.

- ✓ PM2.5, SO2, and NO2 concentrations increase during winter (November to February).
- ✓ Ozone (O3) shows higher levels in the warmer months (May to August).
- ✓ CO follows a similar trend to NO2



Correlation Analysis

- ✓ PM2.5 has strong positive correlations with PM10 of 0.88, CO of 0.77, and NO2 of 0.66.
- ✓ Moderate correlation with SO2 of 0.48.
- ✓ Weak correlations with wind speed of 0.27 and O3 of 0.15.
- ✓ Very weak correlations with temperature (0.13), dew point of 0.11, pressure of 0.02, and rain of 0.01.
- ✓ Ozone (O3) behaves differently, showing negative correlations with most primary pollutants, especially NO2
- ✓ Temperature and dew point have a strong relationship.
- ✓ Rainfall and wind speed show weak to moderate impacts on air quality.



Model Building

The goal of model building is to predict air quality, specifically **PM2.5 concentration**, based on available features. The following regression models are employed:

- ✓ **Linear Regression**
- ✓ **Random Forest Regressor**
- ✓ **XGBoost**
- ✓ **Gradient Boosting Regressor**

The trained models will then be evaluated using the following metrics;

- ✓ **Mean Absolute Error (MAE)**
- ✓ **Mean Squared Error (MSE)**
- ✓ **Mean Squared Error (MSE)**
- ✓ **R^2 (Coefficient of Determination)**

Beside above, residual plots are used for visual model assessment.

Model Performance

- ✓ **XGBoost** and **Random Forest** outperform other models with the lowest MAE, MSE, and RMSE, and the highest R^2 of 0.94.
- ✓ **Gradient Boosting** offers strong performance but slightly lags behind XGBoost and Random Forest in accuracy.
- ✓ **Linear Regression** has the highest error metrics, indicating weaker predictive performance compared to the other models.

Gradient Boosting for Hyperparameter Tuning:

- ✓ It was selected as it seemed to have higher chances of performance improvement across all metrics due to its flexibility and it is a light weight boosting model
- ✓ **Balances complexity** and performance, making it a good choice for optimization.

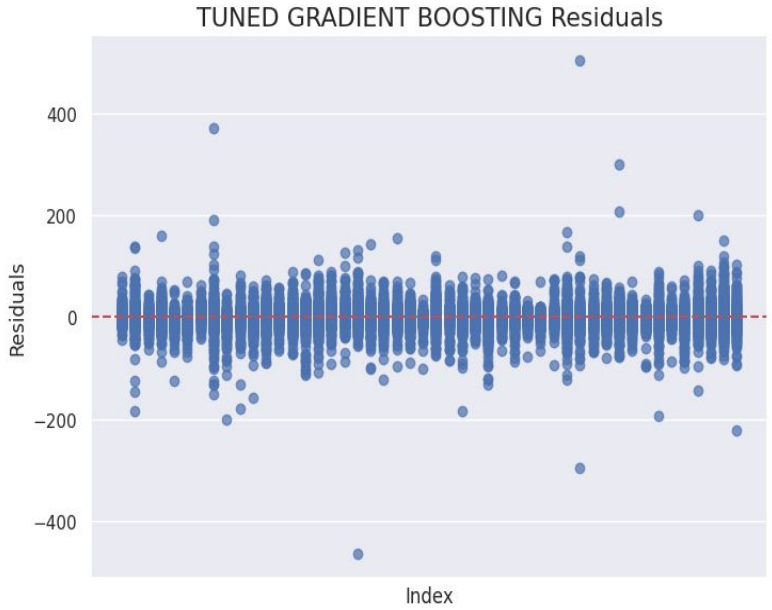
Model	MAE	MSE	RMSE	R^2
Linear Regression	20.48	1015.19	31.86	0.84
Random Forest	12.36	403.25	20.08	0.94
Gradient Boosting	15.68	604.44	24.59	0.91
XGBoost	12.04	365.90	19.13	0.94

Hyperparameter Tuning

- ✓ **GridSearchCV was used together with the Gradient Boosting Regressor** to perform hyperparameter tuning using GridSearchCV.
- ✓ This process aims to find the optimal combination of hyperparameters to minimize prediction error and enhance model performance. Various parameters like learning rate, depth, estimators etc are used.

Tuned Model Results:

- ✓ **MAE**: 12.46 (Improved from 12.36 in Random Forest)
- ✓ **MSE**: 419.16 (Lower than 604.44 in Gradient Boosting)
- ✓ **RMSE**: 20.47 (Better than 24.59 in Gradient Boosting)
- ✓ **R^2**: 0.93 (Improved from 0.91 in Gradient Boosting)
- ✓ **Residual plot** on the left confirms the model's enhanced fit and reduced errors.
- ✓ **Tuned model** was saved for use in the application.



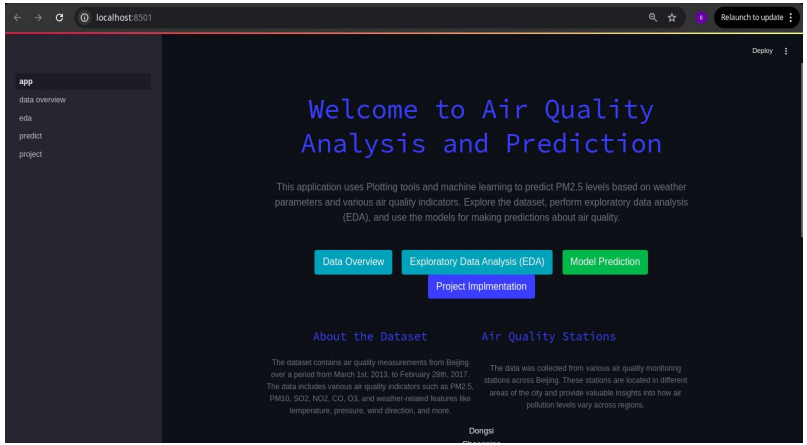
Application Structure and Navigation

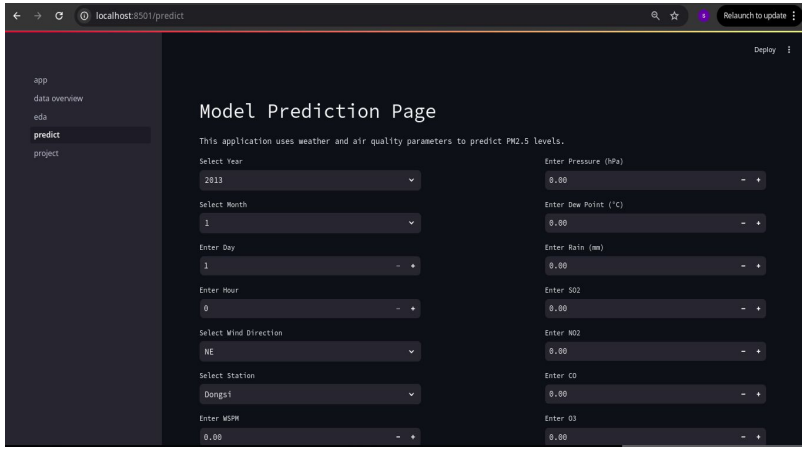
The application was built using Streamlit since it simplifies the development process, allowing for rapid prototyping and deployment. It is a Multi-Page Design where it have four distinct pages, each serving a specific purpose:

- ✓ **Home Page -** Serves as the central hub, providing an overview of the project and navigation links to the other pages.
- ✓ **Data Overview -** Presents a summary of the air quality dataset used in the project, including key statistics and visualizations.
- ✓ **EDA** - Provides interactive tools and visualisations to explore the relationships between air quality parameters, weather conditions, and time-based trends. An additional to see is a whole project code.
- ✓ **Prediction Page** - Allows users to input environmental data and receive PM2.5 predictions generated by the hypertuned Gradient Boosting model.

On the left is sample two pages, others can be seen on demove

SAMPLE PAGES





Critical Reflection: Application Development and Recommendations

Recommendations:

- ✓ **The analysis showed higher PM2.5, CO, and NO2 levels during peak hours (7-9 AM and 5-7 PM), likely due to increased traffic emissions. To address this, implement targeted interventions such as:**
  - ✓ **Promote Public Transportation**
  - ✓ **Alternative Transportation and promote cycling and walking as healthy and eco-friendly alternatives.**
- ✓ **The data shows increased PM2.5, SO2, and NO2 concentrations during colder months. Strategies to combat this include:**
  - ✓ **Promote Cleaner Industrial Practices**
  - ✓ **Promote the use of low-sulphur fuels in industries and transportation to reduce SO2 emissions.**

Q & A

THANK YOU