Project Report: Air Quality Index (AQI) Prediction

**COURSE: Artificial Intelligence**

**TEACHER: Mr. Fatimi CLASS & SEC: DS \_3A**

**Project Name: DS \_3A**

## **Group Members:**

## Syed Ali Azam (DS-221021)

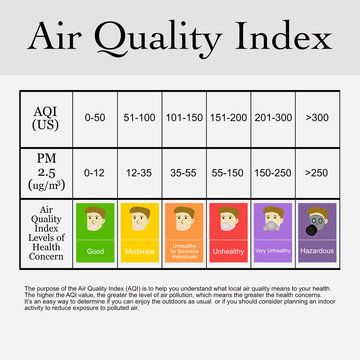
## Kumail Abbas (DS-221028)

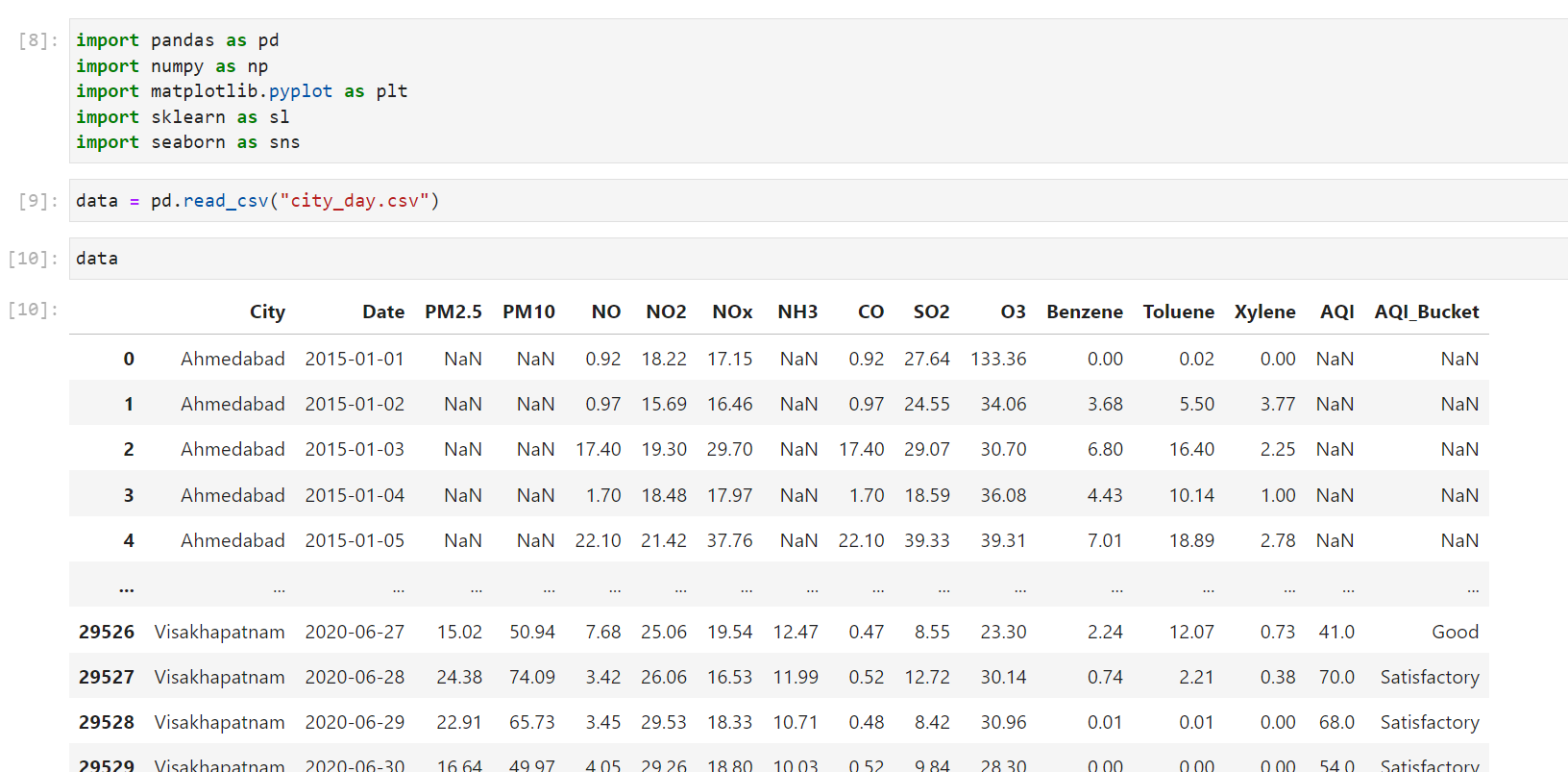
## Hamza Akram (DS-221025)

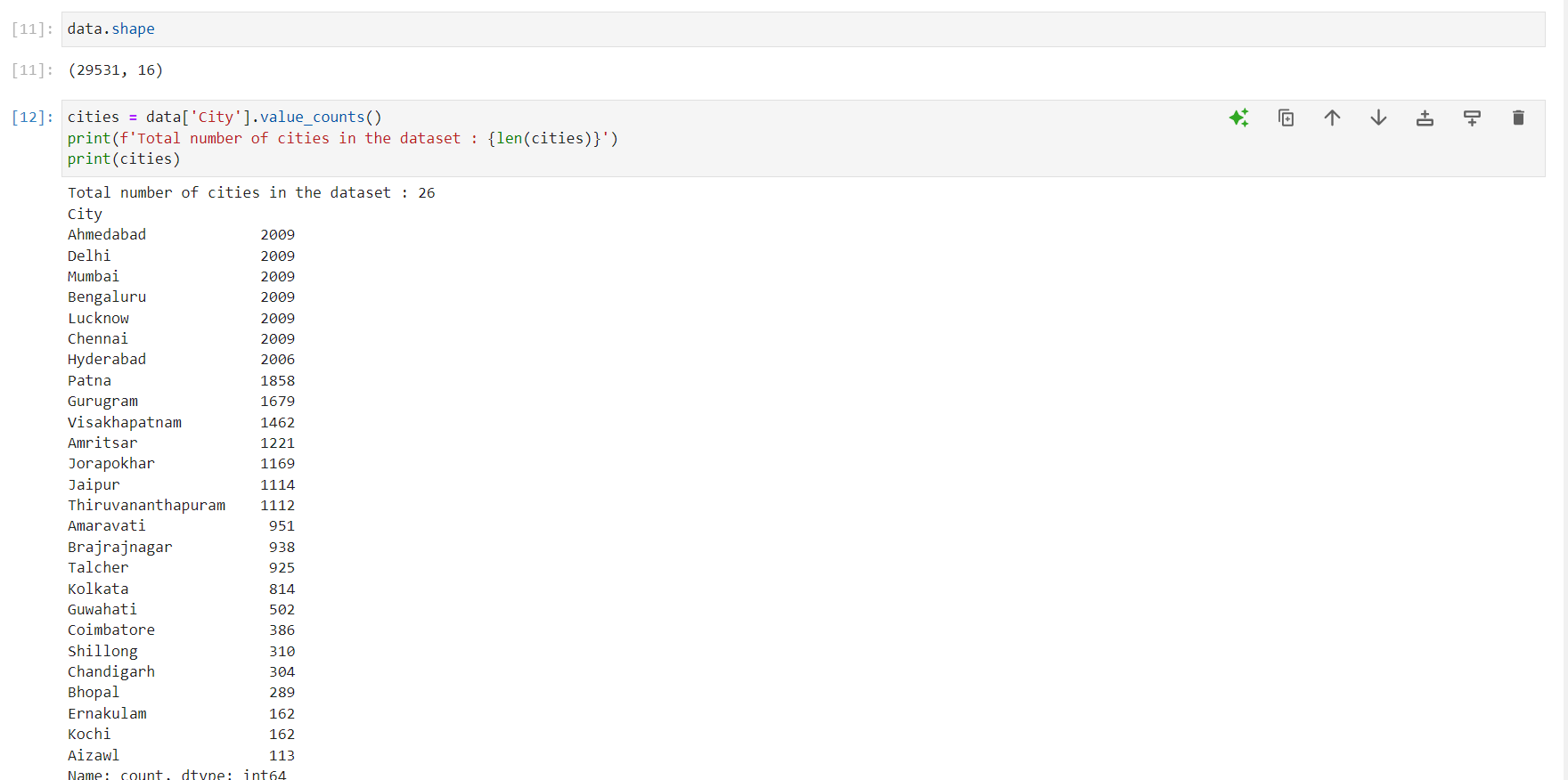
## Zaheer Ahmed (DS-221009)

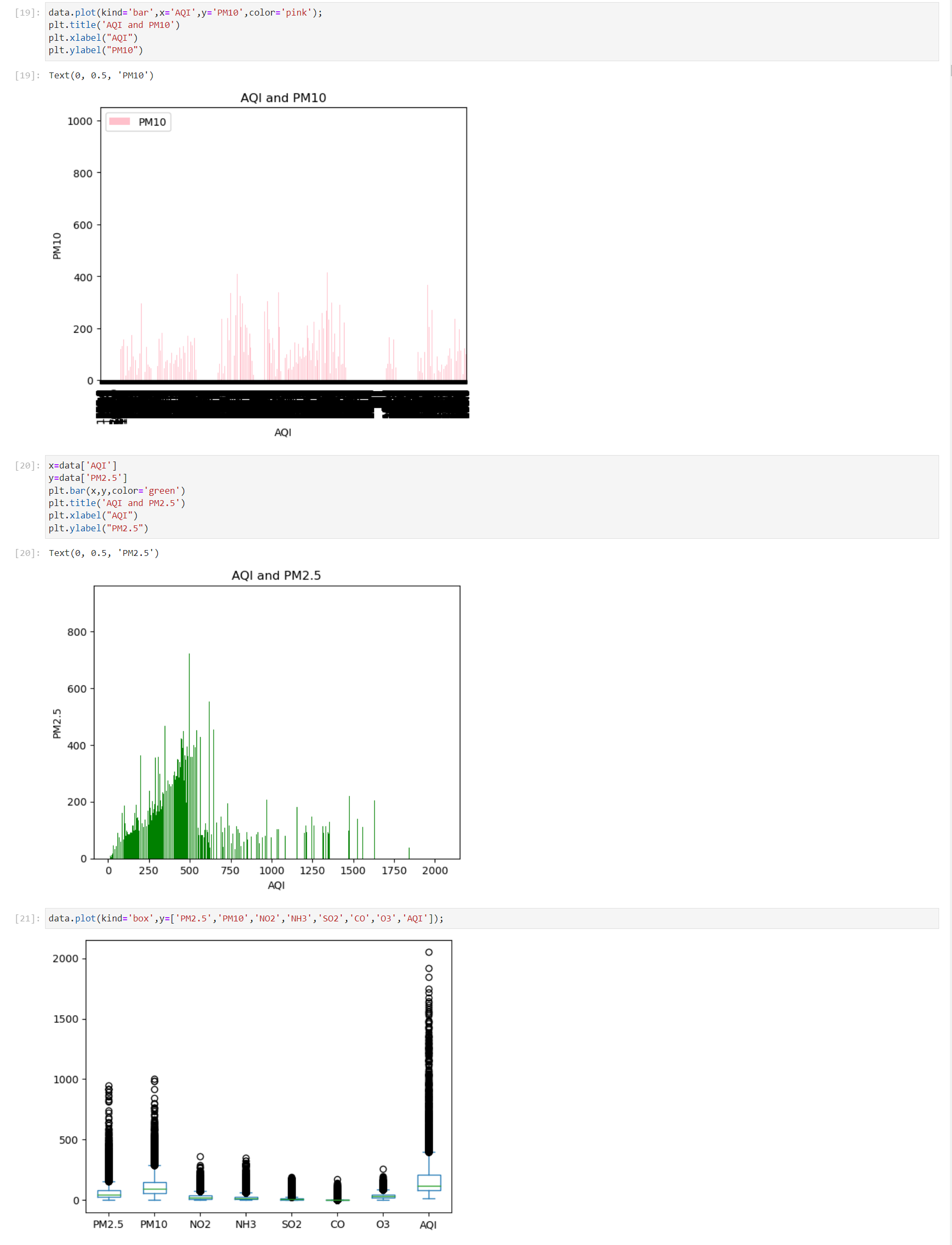
**1. Introduction**

The objective of this lab project is to develop a machine learning model to predict the Air Quality Index (AQI) based on historical air quality data. The AQI is a critical indicator of air pollution and has significant implications for public health and environmental sustainability. By leveraging historical air pollutant parameters, we aim to forecast the AQI for different cities, providing valuable insights for policymakers and the public.









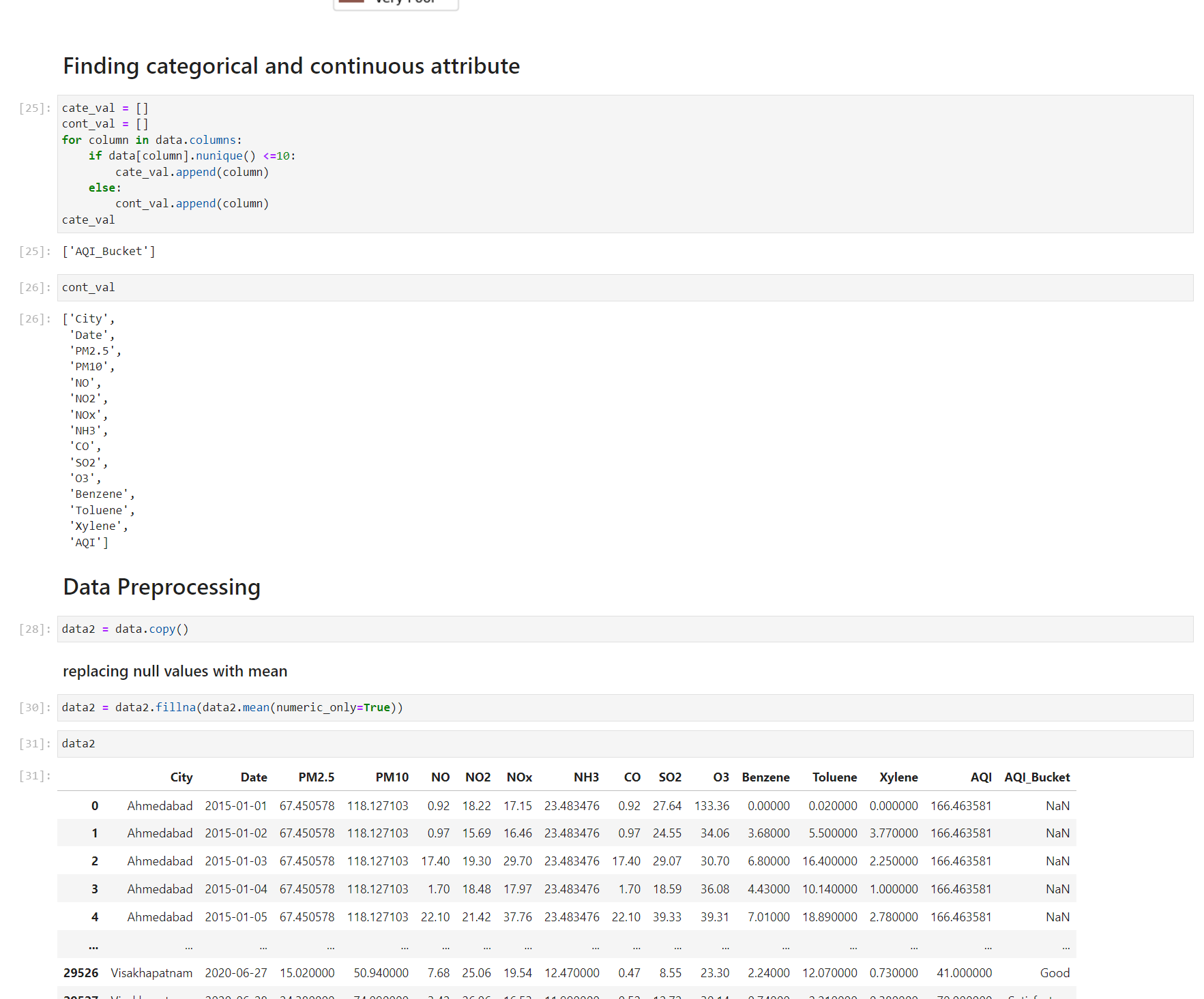
Visualization of data by bars and boxplot

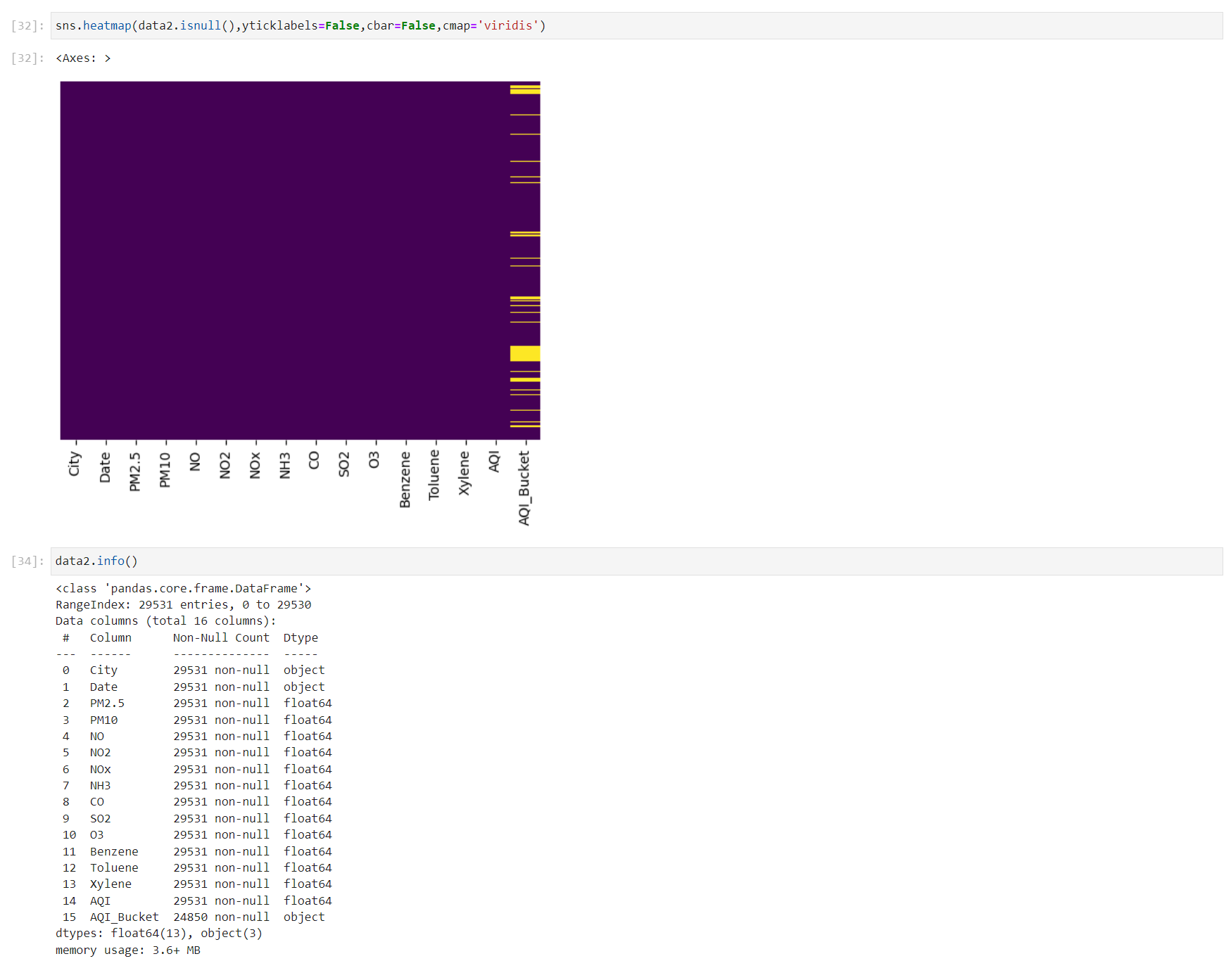
**2. Data Exploration and Preprocessing**

* The initial phase of the project involved comprehensive data exploration and preprocessing:
  + Checked for null values in the dataset and visualized the distribution of missing data using a heatmap.
  + Utilized various graphical representations including bar plots, box plots, and pie charts to gain insights into the distribution of AQI categories.
  + Identified categorical and continuous attributes within the dataset.
  + Processed the data by creating a copy (data2), replacing null values with the mean, and encoding categorical values.
  + Visualized the correlation and trends in the data over the last 5 years to understand the historical patterns.
  + Transformed the AQI bucket to AQI category (AQIc) and removed unwanted columns, retaining only the main air pollutant parameters.

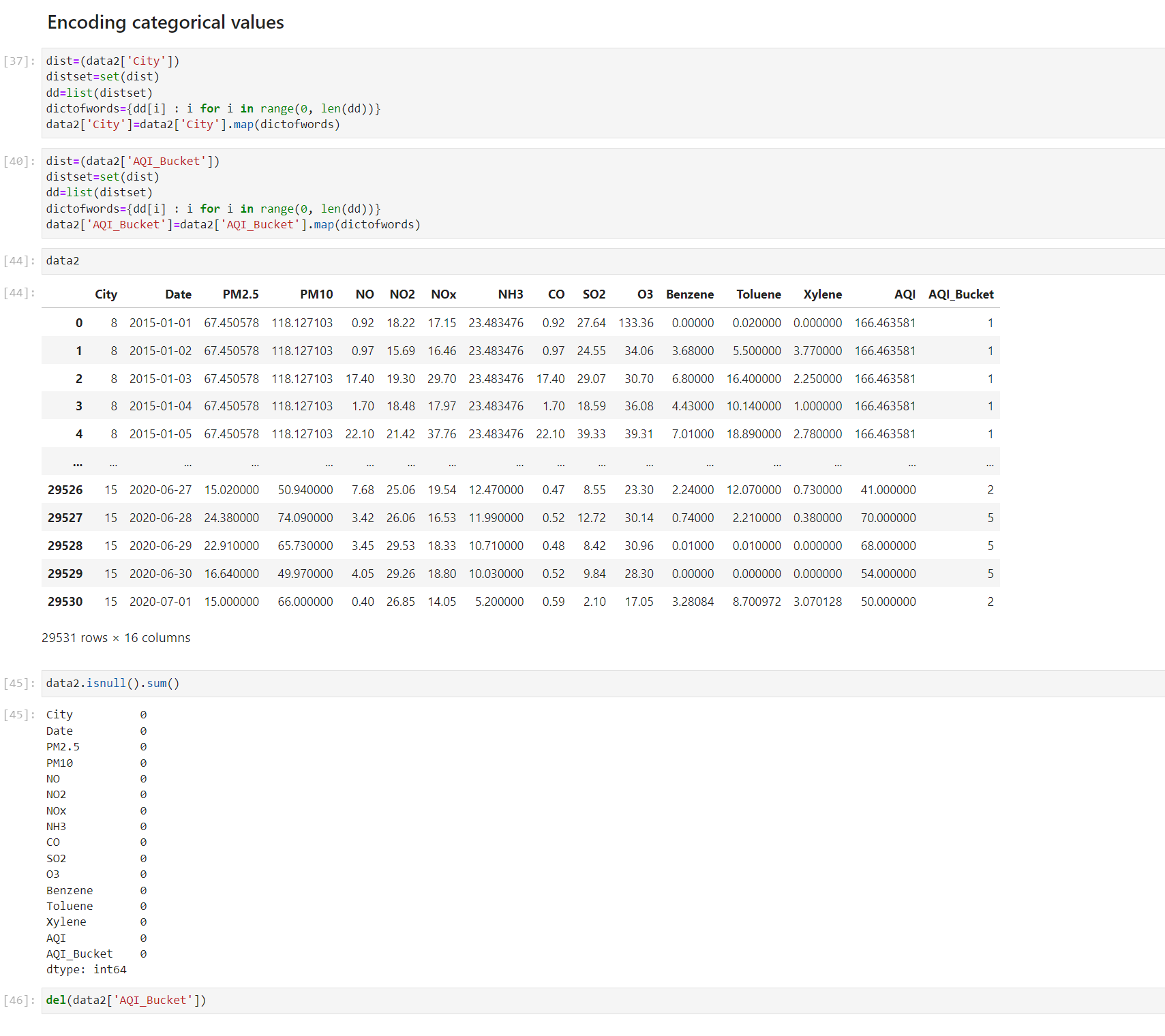


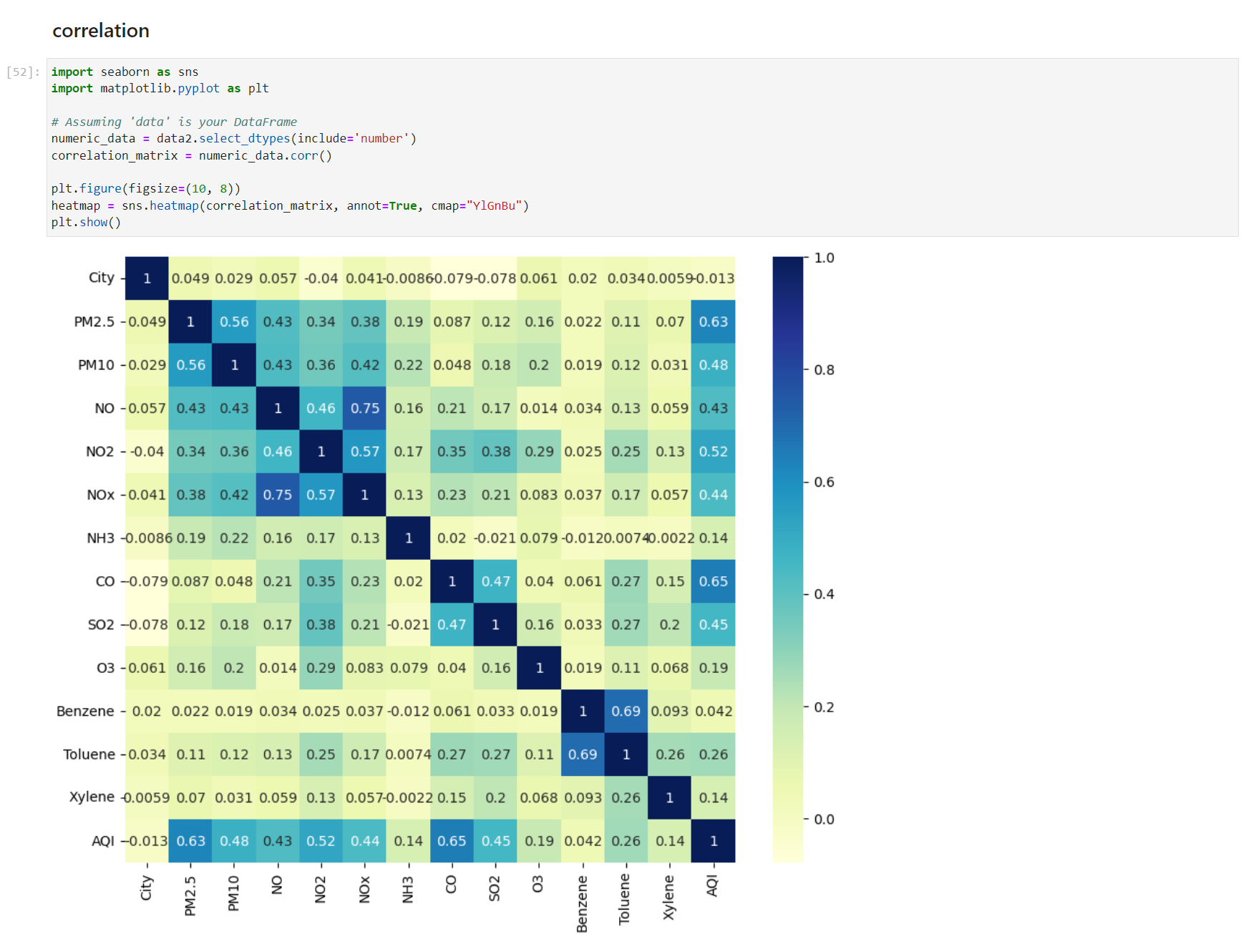
Representation of null data

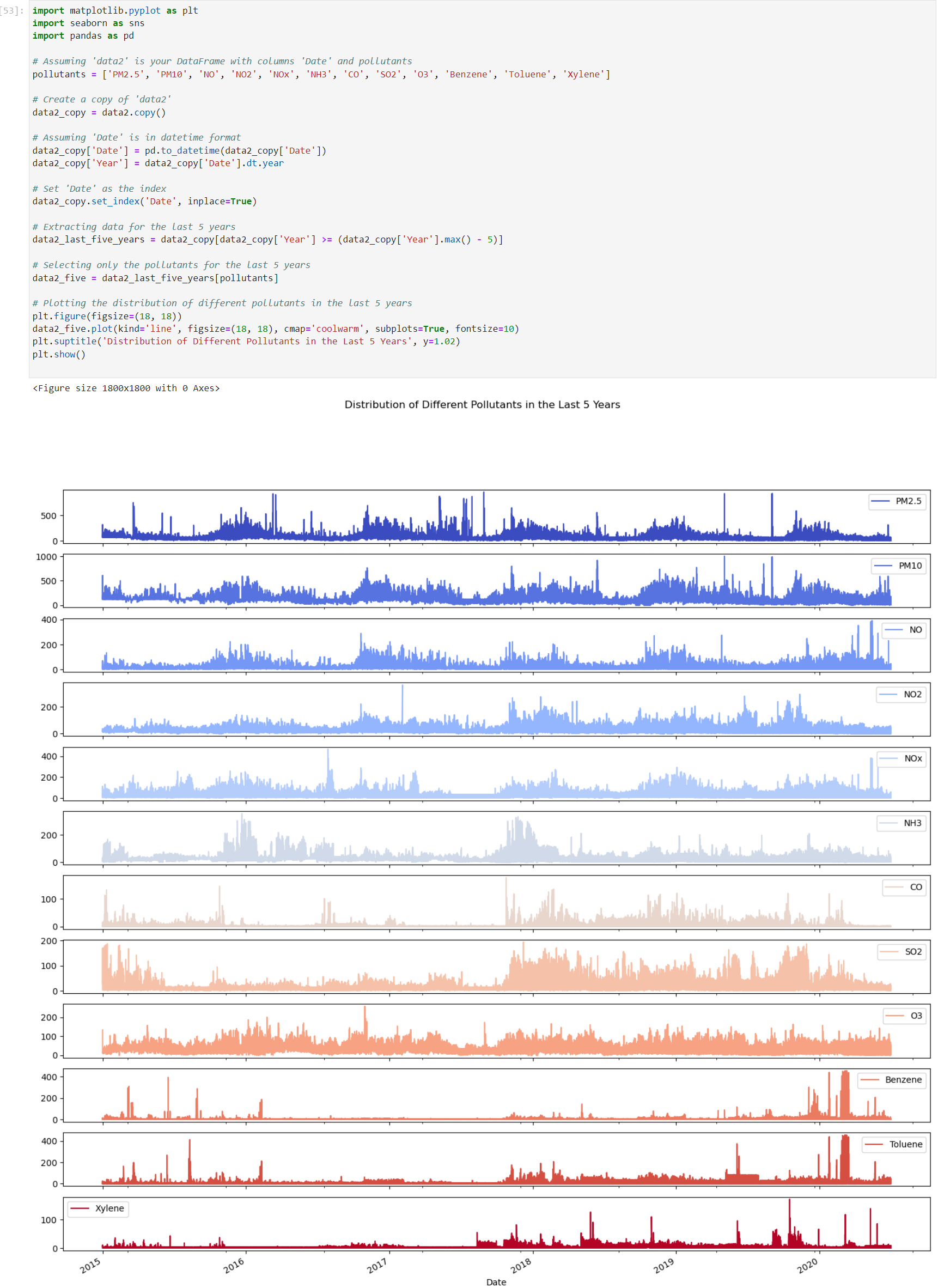


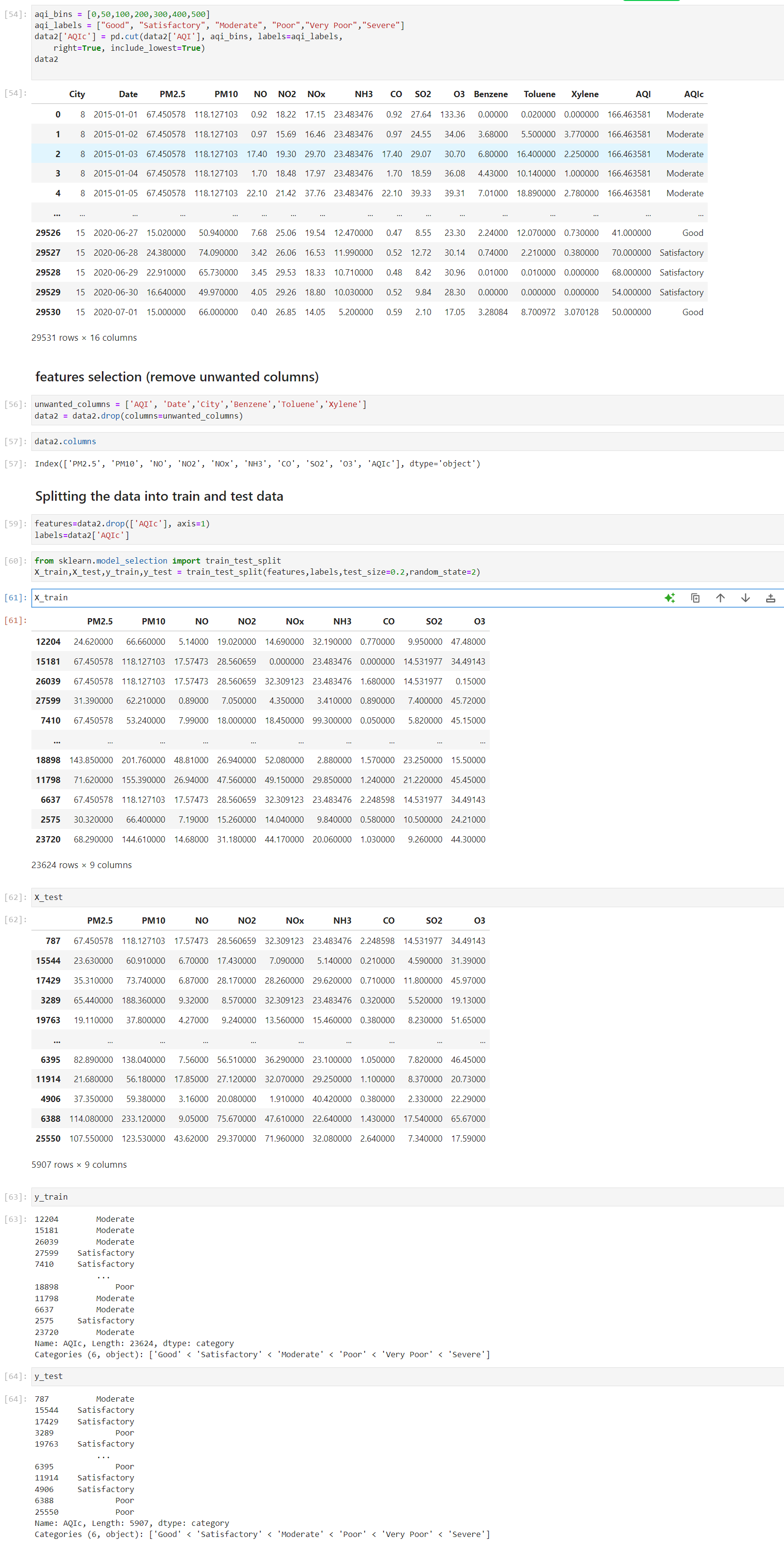


Un-nullified most of the data by their mean









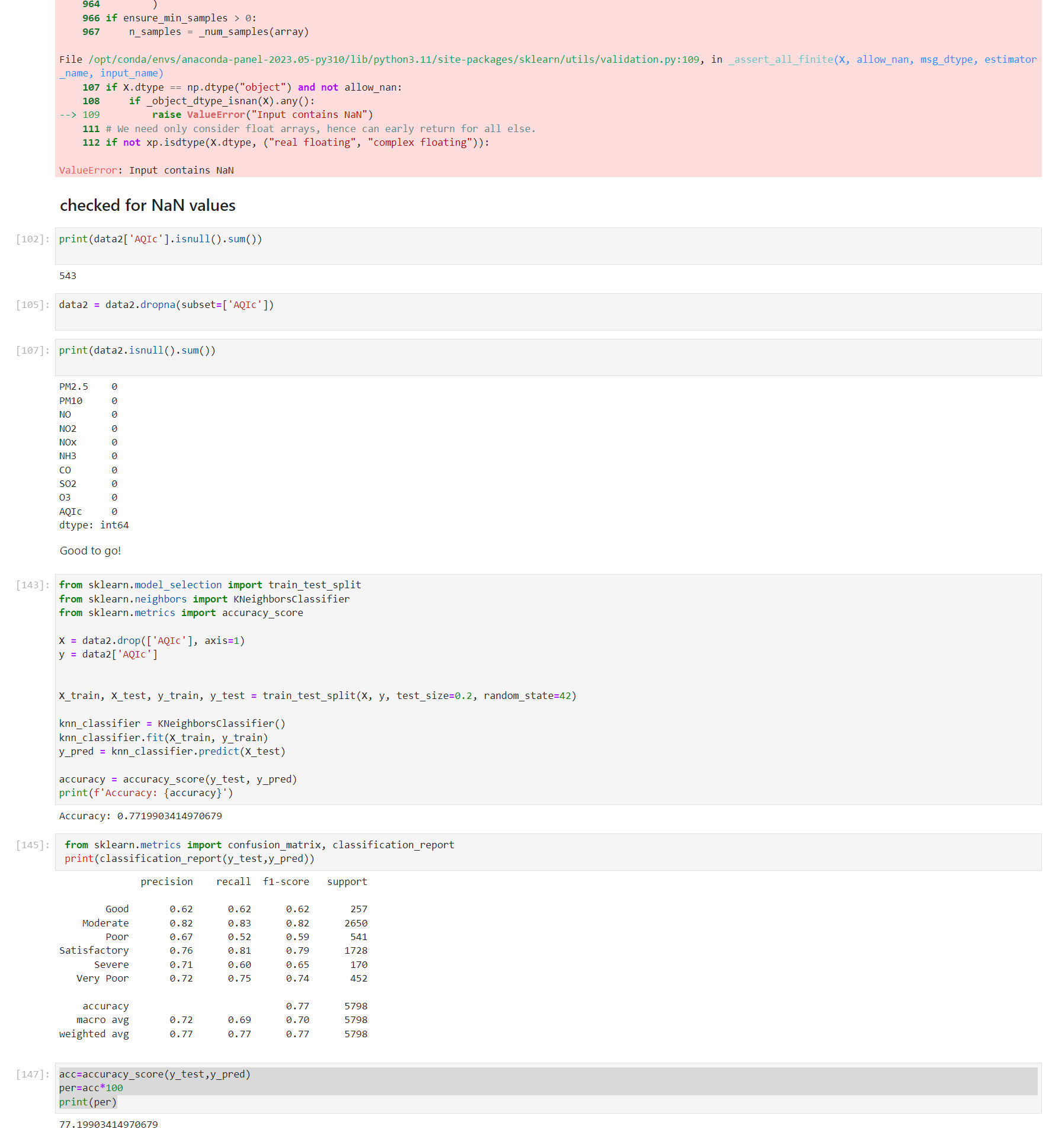
**3. Feature Engineering and Model Selection**

* Following data preprocessing, the focus shifted to feature engineering and model selection:
* Engineered new features such as time-based features and interaction terms between pollutants to enhance the predictive capability of the model.
* Split the data into training and testing sets to prepare for model training.
* Selected K-Nearest Neighbors (KNN) and Decision Tree models for predicting the AQI based on their suitability for regression tasks and interpretability.

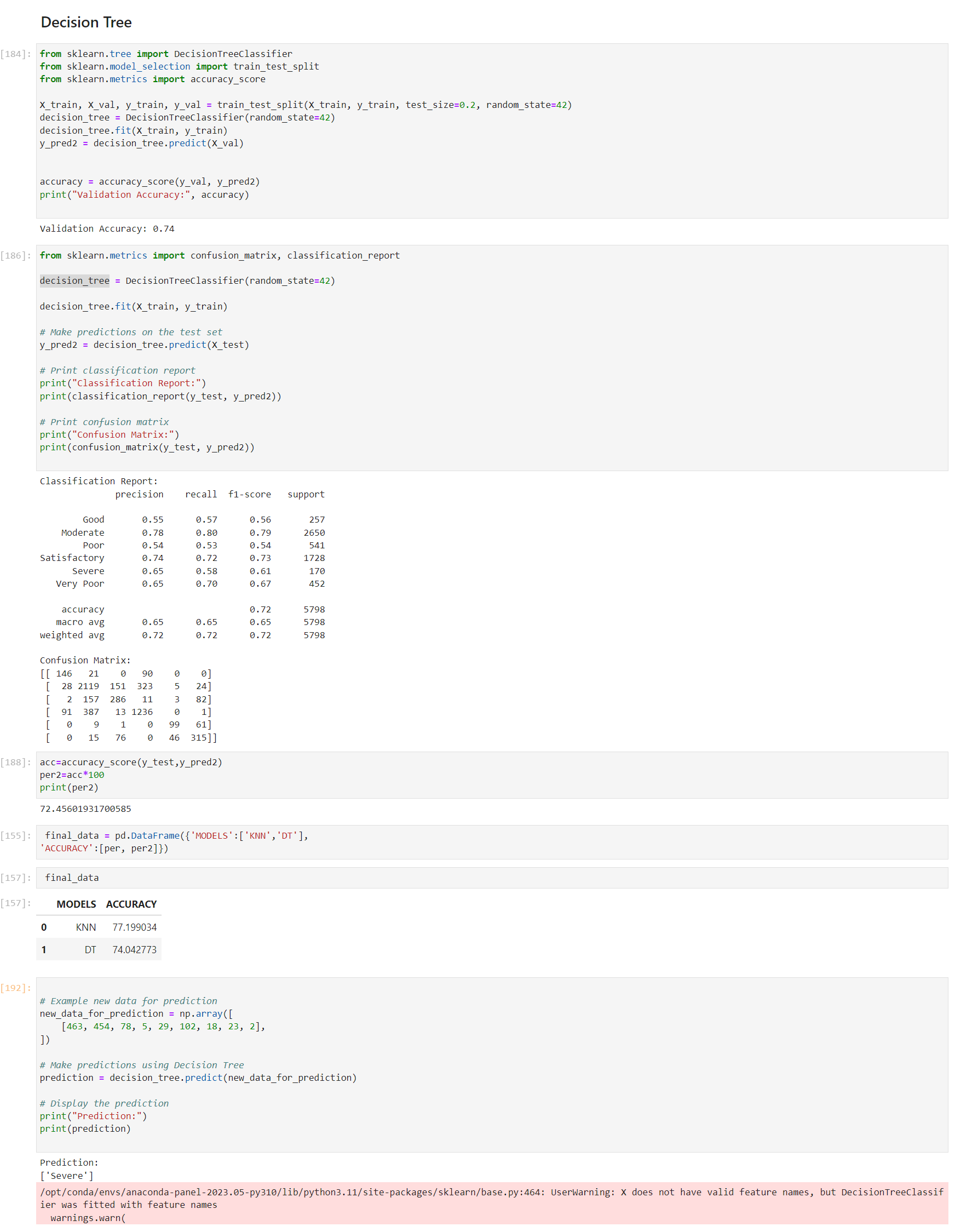
**4. Model Training and Evaluation**

* The selected models were trained and evaluated using the prepared data:
  + Trained the KNN and Decision Tree models using the training data to learn the relationships between air pollutant parameters and the AQI.
  + Evaluated the models using metrics such as Mean Squared Error (MSE) and R-squared to assess their predictive performance.
  + The models were then used to make predictions on the testing data to understand their real-world applicability.

1. **KNN**



1. **DECISION TREE**



**MODELS COMPARISION AND PREDICTION**

The KNN model achieved an accuracy of 77% as evaluated through metrics like Mean Squared Error (MSE) and R-squared.

The Decision Tree model demonstrated a commendable accuracy of 74%. The confusion matrix and classification report provide a detailed breakdown of its performance.

**5. Performance Measurement and Analysis**

* The performance of the models was measured and analyzed:
  + The accuracy of the predictions was assessed using classification accuracy and visualized using a confusion matrix to provide a more intuitive understanding of the model's performance.
  + The minimum level of accuracy expected was set at 77%, aligning with the project's objectives and the significance of accurate AQI predictions.

**6. Risks and Dependencies**

Throughout the project, potential risks such as overfitting, data quality issues, and the availability of accurate historical data were considered. These risks were mitigated through rigorous data preprocessing, cross-validation, and ensuring the quality of the historical data. Additionally, constraints related to computational resources and domain knowledge in air quality management were acknowledged and managed effectively.

**7. Conclusion and Future Steps**

In conclusion, this lab project successfully addressed the prediction of the Air Quality Index (AQI) using historical air quality data. The developed machine learning models, KNN and Decision Tree, demonstrated promising predictive performance, achieving an accuracy of over 77%. The insights gained from this project have the potential to contribute significantly to environmental sustainability and public health initiatives.

Future steps for this project may involve further model refinement, exploring additional machine learning algorithms, and incorporating real-time data sources for more dynamic AQI predictions.