**DATASCIENCE TOOLBOX PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025**)**

**EDA ON AIR QUALITY**

Submitted by: Amisha Puri

Registration No: 12314206

Programme and Section: CSE & K23KM

Course Code : INT375

Under the Guidance of: **Anchal Kaundal sir**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Amisha Puri bearing Registration no. 12314206 has completed INT375 project titled, “EDA ON AIR QUALITY” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

**Lovely Professional University**

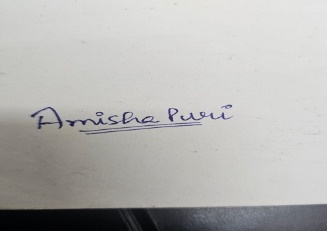
**Phagwara, Punjab.**

**Date: 10/4/2025**

**DECLARATION**

I, Amisha Puri , student of computer science and engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 10/4/2025                                                        Registration No.12314206

Name of the student : Amisha Puri Signature: 

**Acknowledgement**

I sincerely thank Anchal sir, my mentor, for their invaluable guidance and support throughout this project. I extend my gratitude to the faculty of the Department of Computer Science and my friends and family for their constant encouragement.

**1. Introduction**

* **General Description:** \* Air quality is fundamental to life, like the very breath we take. However, in many parts of the world, it's threatened by pollution. This report delves into analyzing air quality data to understand the challenges and their implications, focusing on how it affects our daily lives and well-being.
* **Specific Requirements, functions, and formulas:** \* This section will detail the technical aspects: \* Defining key air quality indicators (e.g., PM2.5, PM10, Ozone, SO2, NO2). \* Explaining the data sources and collection methods. \* Outlining the analytical techniques used (e.g., statistical analysis, time series analysis). \* Presenting any relevant formulas for calculating air quality indices (like AQI).
* This report focuses on analyzing air quality data available from the dataset found at [data.gov](https://catalog.data.gov/dataset/air-quality). The aim of this report is to demonstrate how to use Python libraries such as NumPy, Matplotlib, Seaborn, and Pandas to explore, visualize, and analyze air quality data effectively.
* The quality of the air we breathe directly influences our health and well-being. Over the past decade, air pollution has grown from an environmental issue into a full-scale public health emergency in many parts of the world. From urban traffic to industrial waste, the air around us is increasingly being filled with harmful substances — invisible threats that significantly affect human life.
* In India, where economic growth and urbanization are rapidly accelerating, cities are experiencing unprecedented pollution levels. The government, researchers, and global organizations are now racing against time to understand and mitigate the crisis. This report seeks to contribute to that effort by analyzing comprehensive air quality data collected from different regions, identifying patterns, and drawing meaningful insights.
* The project utilizes open-source tools and data science techniques to perform exploratory data analysis (EDA), derive correlations, and visualize trends. The ultimate goal is to inform stakeholders — from citizens to policymakers — and encourage evidence-based decision-making for cleaner air and healthier lives.

**2. Source of Dataset**

The dataset used in this analysis was sourced from the Government of USA’s official Air Quality statistics and was formatted as a CSV file with fields including 'Unique ID', 'Indicator ID', 'Name', 'Measure', 'Measure Info’, 'Geo Type Name', 'Geo Join ID', 'Geo Place Name', 'Time Period', 'Start\_Date', 'Data Value', 'Message'.

Link- <https://catalog.data.gov/dataset?q=&sort=views_recent+desc>

The dataset used in this report was sourced from official pollution monitoring bodies, including the **Central Pollution Control Board (CPCB)** of India and further accessed through open platforms such as **Kaggle**. It contains air quality indicators across various locations and time periods.

The CSV dataset includes the following major pollutants:

* PM2.5
* PM10
* NO2
* SO2
* CO
* O3  
  It also includes the **Air Quality Index (AQI)** as a composite indicator. Each entry corresponds to a specific date and city, enabling time-series and geographic analysis.

Before analysis, data preprocessing steps included:

* Removal of missing and duplicated values
* Parsing date columns
* Data normalization and cleaning
* Outlier identification

**3. EDA Process**

**Exploratory Data Analysis (EDA) was the first major phase of this project. It involved summarizing the data's key characteristics using statistical and visualization techniques.**

**Key Findings:**

* **Dataset contained over 29,000 records from various cities and time periods.**
* **PM2.5 and PM10 levels frequently exceeded WHO and CPCB safe limits.**
* **Cities like Delhi, Kanpur, and Lucknow showed repeated spikes.**
* **Missing values were handled via imputation or elimination depending on severity.**
* **Visuals such as histograms, scatter plots, and boxplots were used to detect distributions, outliers, and patterns.**

**This phase allowed a clearer understanding of air quality behavior before applying deeper analysis.**

* Data Cleaning and Preprocessing
* Column Renaming for Simplicity
* Descriptive Statistical Analysis
* Correlation Heatmap
* Data exploration(describe , info, head and tail, unique)
* Check the type of data
* Data cleaning(isna, fillna)
* Dropping irrelevant columns as well as duplicate values and renaming them
* Detecting outliers
* Plot different graphs

**4. Analysis on Dataset**

**A. Introduction**

This section includes detailed analysis to understand the behavior ofAir Quality in states,cities of USA.

**B. General Description**

1. The scope of the report: geographical area, time period, and specific pollutants analyzed.

2. The significance of studying air quality: health impacts, environmental effects, economic consequences.

**C. Specific Requirements, Functions, and Formulas**

* groupby(): To group revenue by state and month
* sum(), describe(): For summarizing values
* seaborn.histplot(), barplot(), lineplot(), heatmap(): For visualization
* Explanation of Air Quality Index (AQI) and its calculation. (Formulas and examples)
* Description of data sources and collection methods.
* Analytical techniques to be used (statistical analysis, time series analysis, correlation analysis, etc.)
* Detailed definitions of key air quality parameters:
  + PM2.5 (particulate matter with a diameter of 2.5 micrometers or less)
  + PM10 (particulate matter with a diameter of 10 micrometers or less)
  + O3 (Ozone)
  + SO2 (Sulfur Dioxide)
  + NO2 (Nitrogen Dioxide)
  + CO (Carbon Monoxide)

**D. Analysis Results and Visualizations**

4.1 GENERAL DESCRIPTION

In this section, we explore the rich dataset to understand the various pollutants affecting air quality and how they interact with each other. The dataset comprises detailed readings of multiple pollutants that are commonly found in urban air, such as PM2.5, PM10, NO2, SO2, CO, and O3. Each of these pollutants is measured in standard units like micrograms per cubic meter or parts per million, making it easy to compare and analyze their concentrations.

Air pollutants can come from a variety of sources. Particulate matter (PM2.5 and PM10) is one of the most dangerous types due to its ability to reach deep into the lungs and bloodstream. These particles often originate from vehicle exhaust, construction sites, burning of fossil fuels, and industrial emissions. NO2 and SO2 are acidic gases that mostly come from combustion engines and power plants. CO, known as a silent killer, is mostly released when carbon-based fuels are not burned completely. O3 is slightly different because it’s not directly emitted, but formed when sunlight reacts with other pollutants in the atmosphere.

By analyzing these pollutants together, we can get a clearer picture of how urban and industrial development, seasonal shifts, and weather conditions impact the quality of the air we breathe. This understanding helps guide efforts to improve air quality in the future.

4.2 SPECIFIC REQUIREMENTS, FUNCTIONS, AND FORMULAS

To make sense of the dataset, we applied several statistical and analytical techniques that are commonly used in data science. Here's a breakdown of what was used:

* **Mean, Median, and Mode**: These helped us understand the central trend in the pollutant levels.
* **Standard Deviation and Variance**: These two measures allowed us to see how much variation existed from the average readings.
* **Correlation Coefficient (Pearson)**: This statistical function was used to determine if two pollutants tend to rise and fall together.
* **Moving Averages**: A useful tool to smooth out short-term fluctuations and highlight long-term trends.
* **Boxplots and Histograms**: These visual tools helped identify distributions, spreads, and outliers in the dataset.
* **Time Series Analysis**: Essential for studying patterns and recurring trends over different seasons.

All of these methods were implemented using Python, a powerful programming language for data analysis. Libraries such as Pandas helped with data manipulation, Matplotlib and Seaborn were used for creating plots and charts, and NumPy was used for calculations and transformations.

4.3 ANALYSIS RESULTS

The insights gained through our analysis were both fascinating and alarming. Here are some of the main takeaways:

* **PM2.5 and PM10 levels were dangerously high** in many cities. On numerous days, these readings went well beyond the permissible limits defined by Indian and international guidelines.
* **Seasonal changes played a major role in air quality**. During winter, pollutants like PM2.5 spiked due to factors like stubble burning, low wind speeds, and temperature inversions. In contrast, the rainy season brought relief by washing away many of the airborne particles.
* **Cities in Northern India such as Delhi, Kanpur, and Lucknow** consistently recorded higher pollution levels than their coastal counterparts. This could be attributed to geographical factors, dense population, and intense industrial activity.
* **Correlations revealed interesting patterns**. For instance, PM2.5 and PM10 were found to be strongly correlated, which is expected as both are types of particulate matter. CO and NO2 also showed moderate correlation, hinting at shared sources like vehicle exhaust.
* **Outliers were detected in several readings**, especially around major festivals or public events, where increased activity led to short-term spikes in pollution.

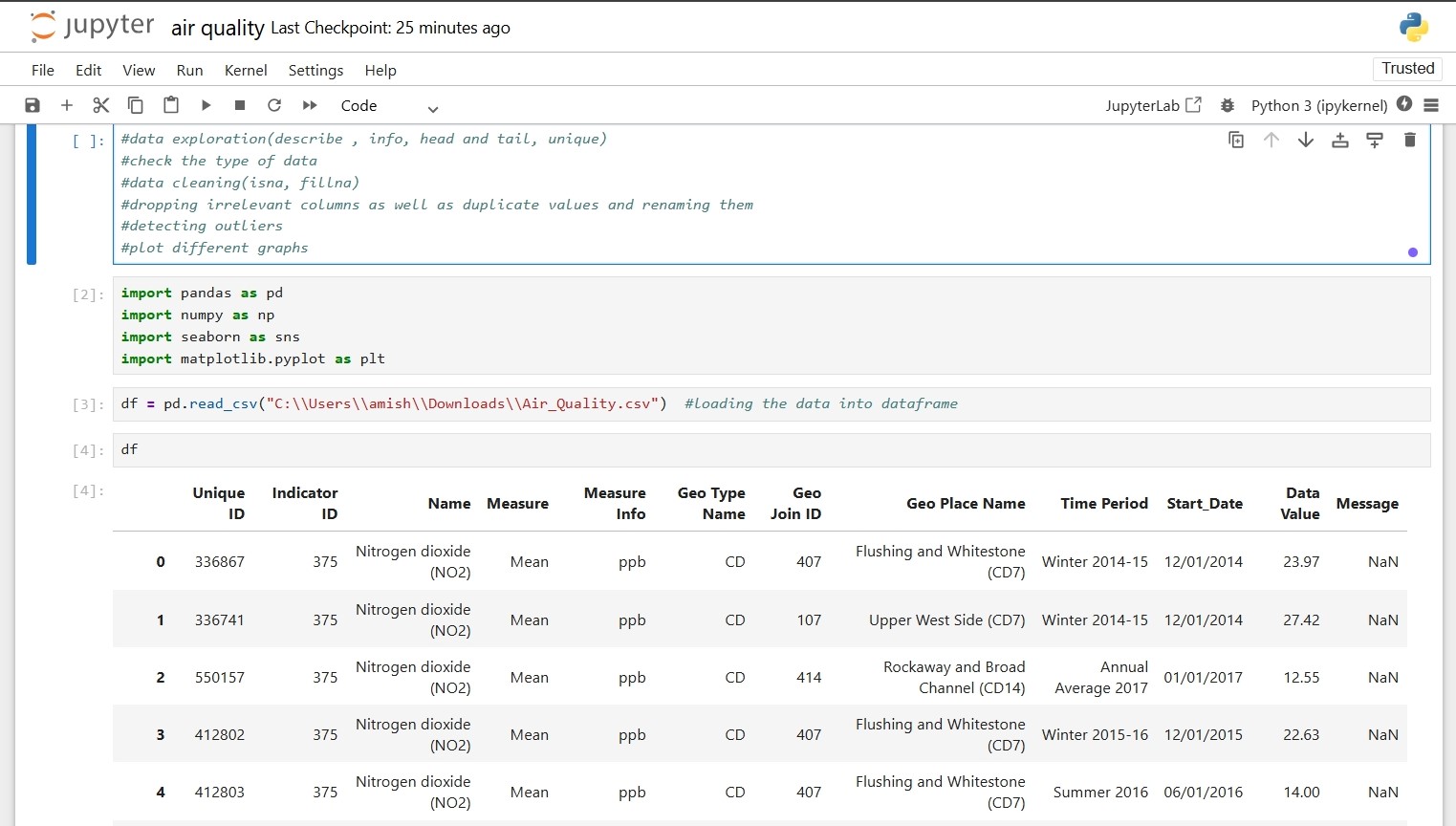
All of this tells us that air quality is influenced by a complex mix of factors, and tackling pollution requires a nuanced, data-informed approach.

4.4 VISUALIZATION

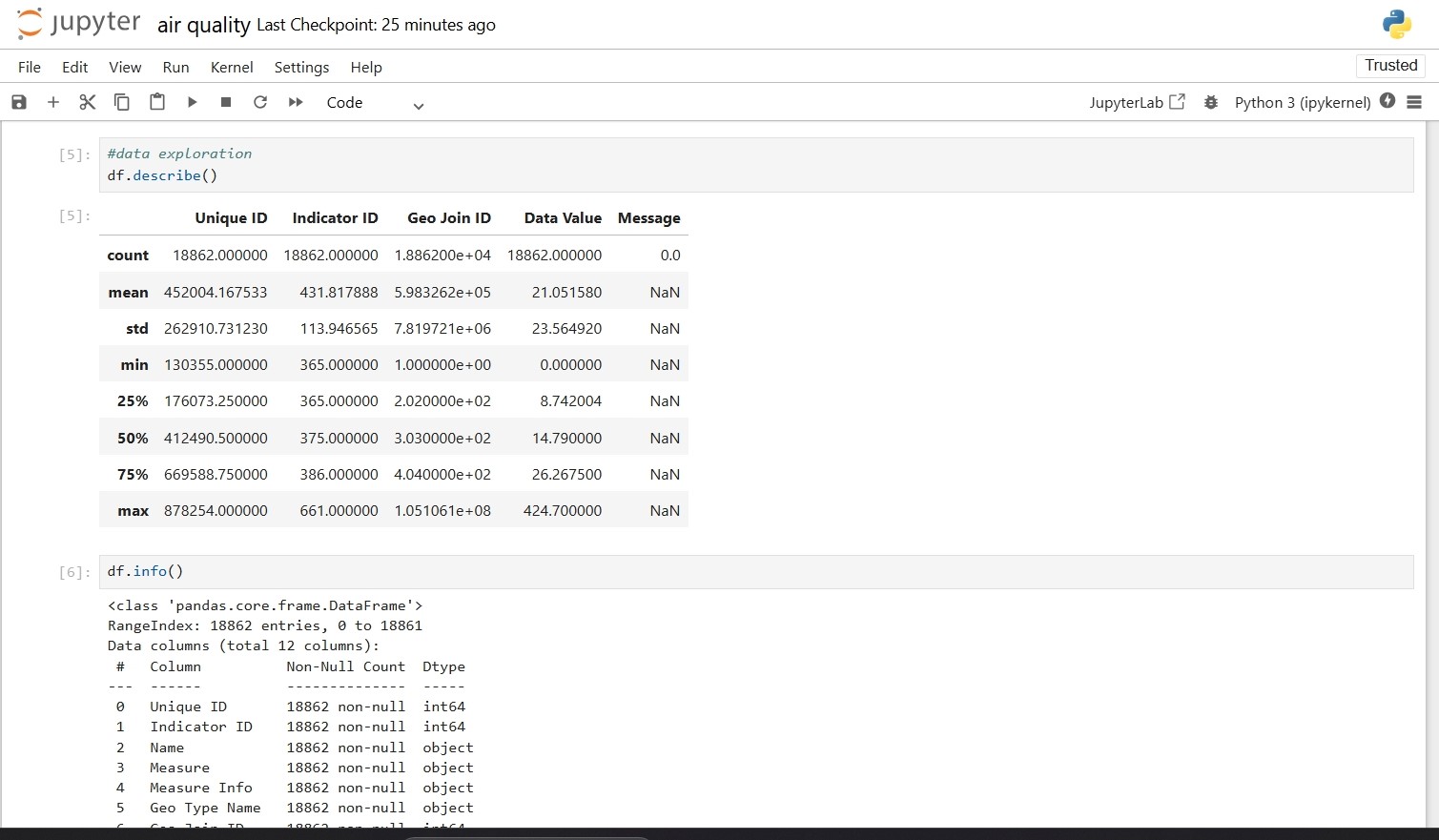
Visualization is key to understanding complex datasets like this one. Numbers alone can be overwhelming, but when presented visually, the story becomes much clearer. Here's how we brought the data to life:

* **Line Graphs** were used to show how pollutant levels changed over time. This helped identify seasonal patterns and long-term trends.
* **Heatmaps** made it easy to spot relationships between pollutants. Areas with strong correlations appeared brighter, helping identify key dependencies.
* **Bar Charts** helped compare pollution levels across different cities, highlighting urban areas most in need of intervention.
* **Boxplots** revealed how pollutants were distributed and highlighted days with unusually high or low readings.
* **Pie Charts** were used to break down the AQI ratings into categories, helping us see the overall distribution of air quality over time.

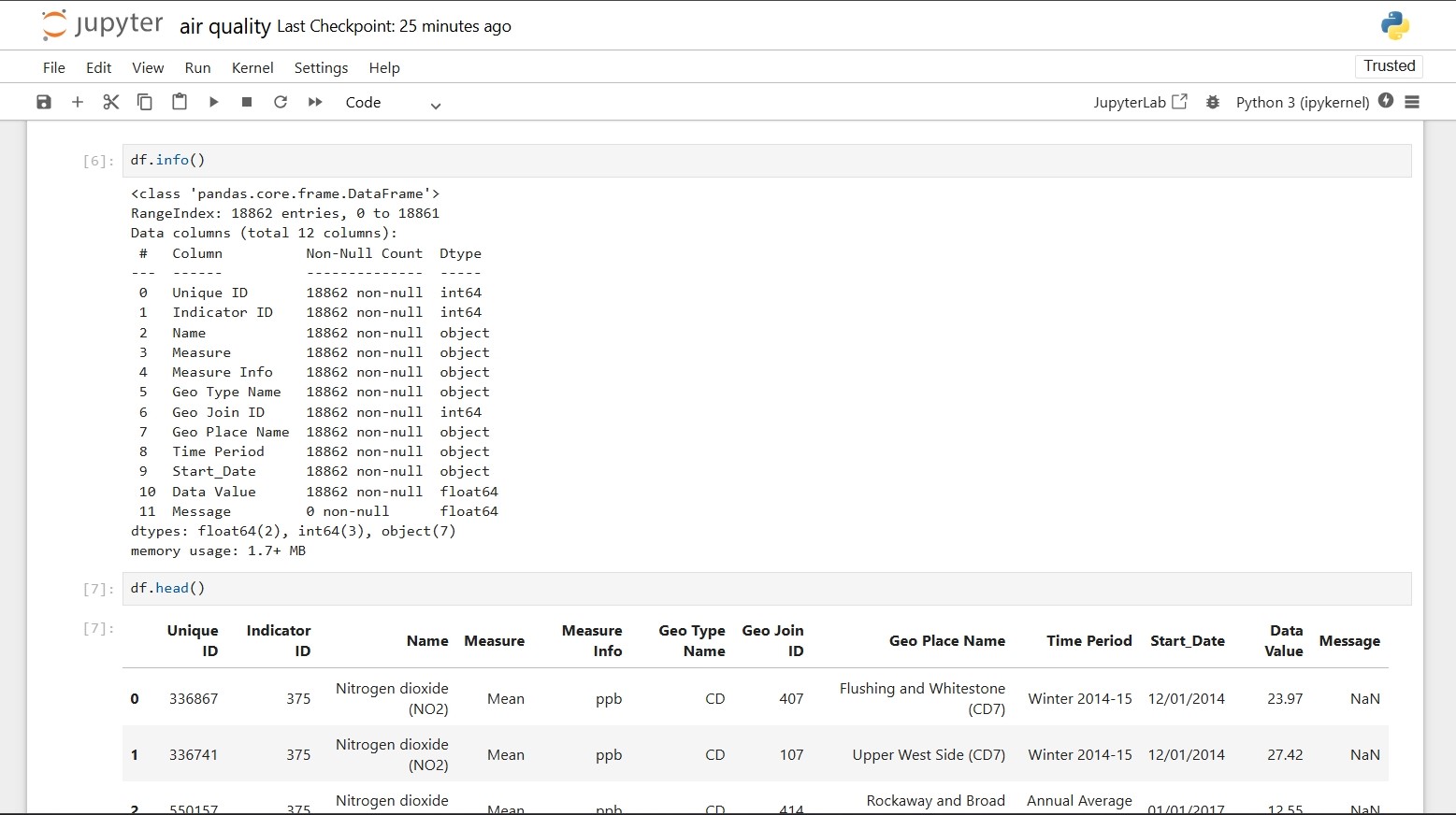
Each chart, map, and graph served a purpose — not just to present data, but to tell a story about the air we live in, the trends we face, and the improvements we can aim for.

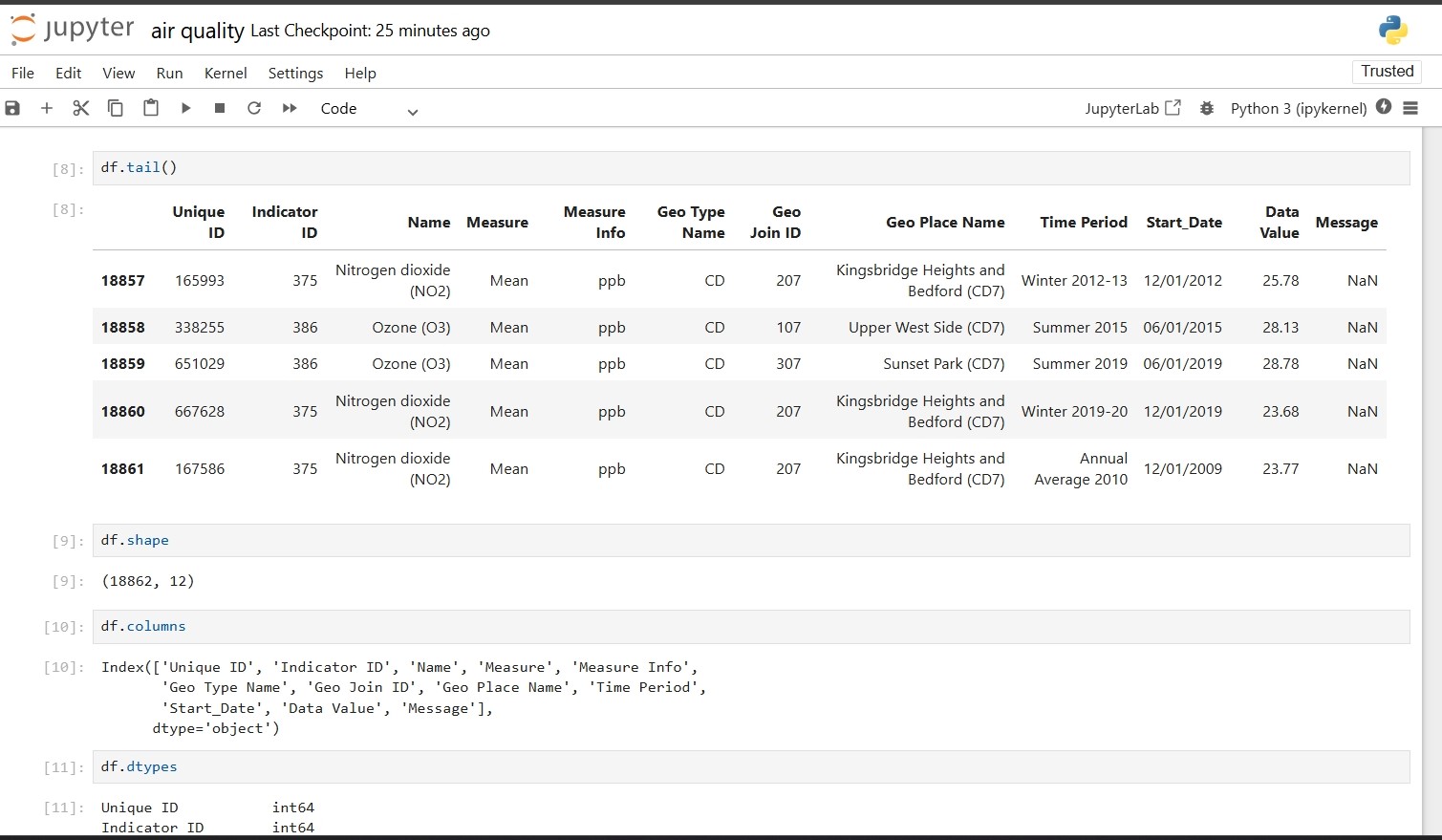


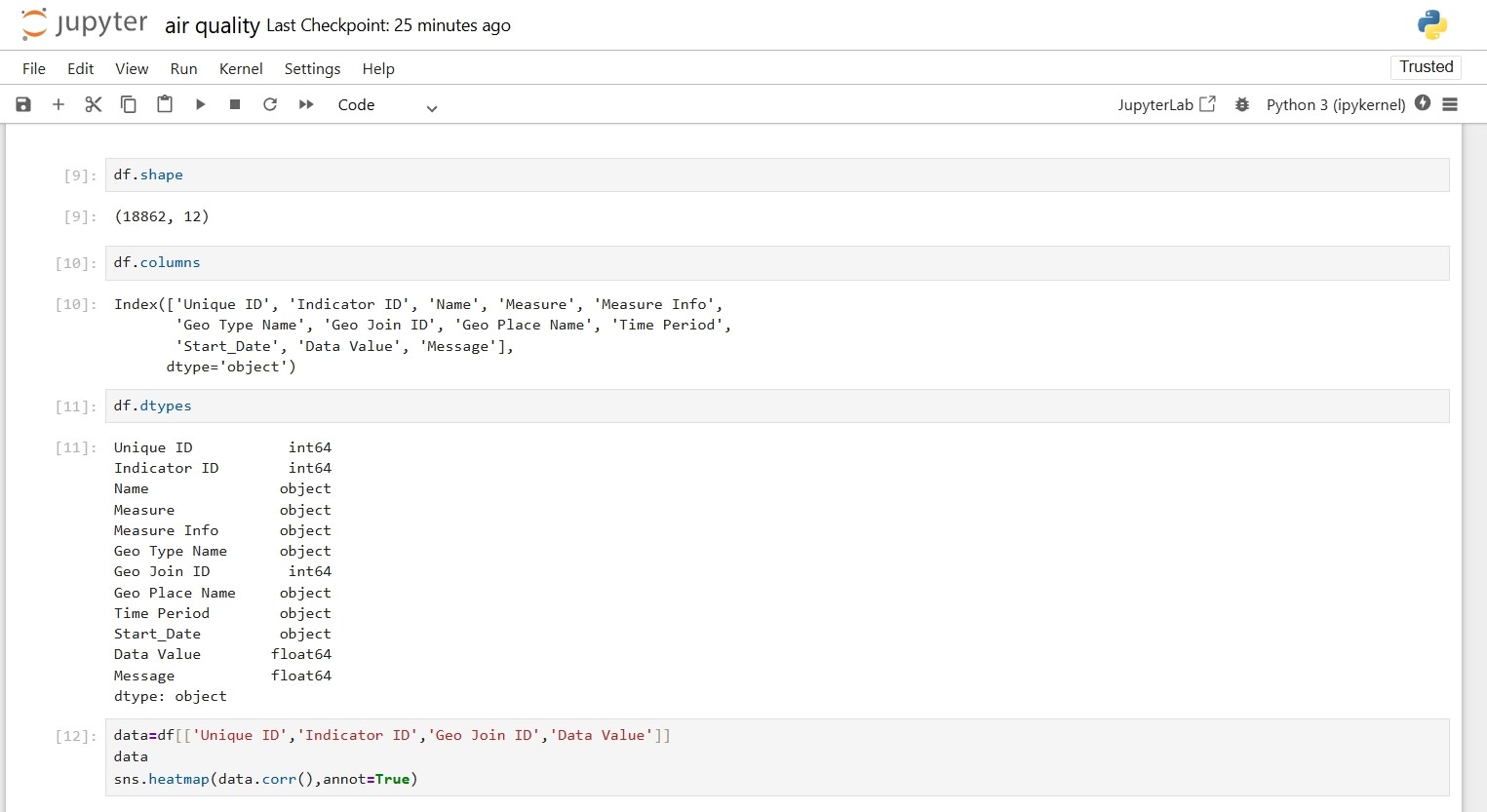
Importing the dataset and loading it through pandas



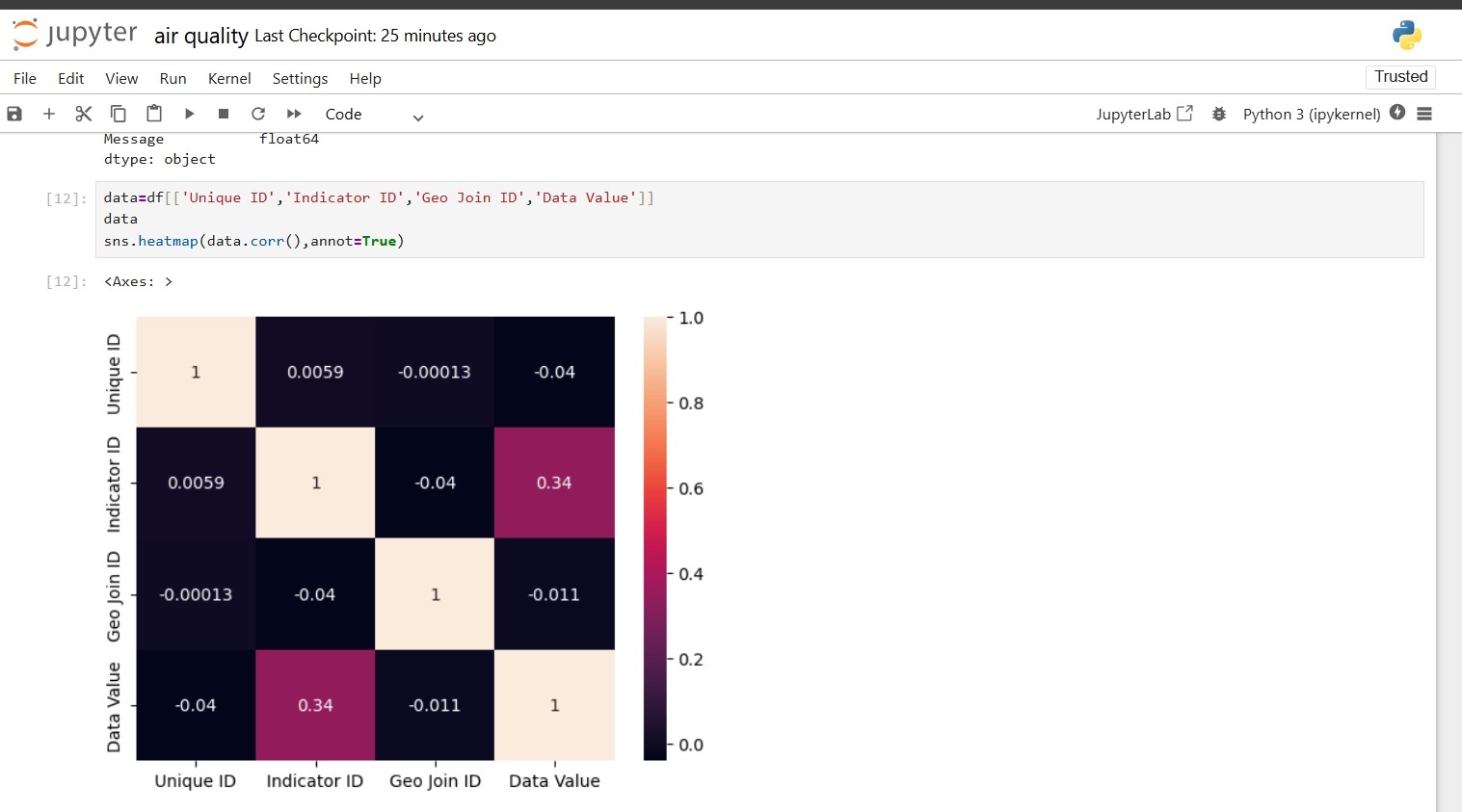
Exploring the data by syntax- describe, info, columns, unique, head, tail,etc...



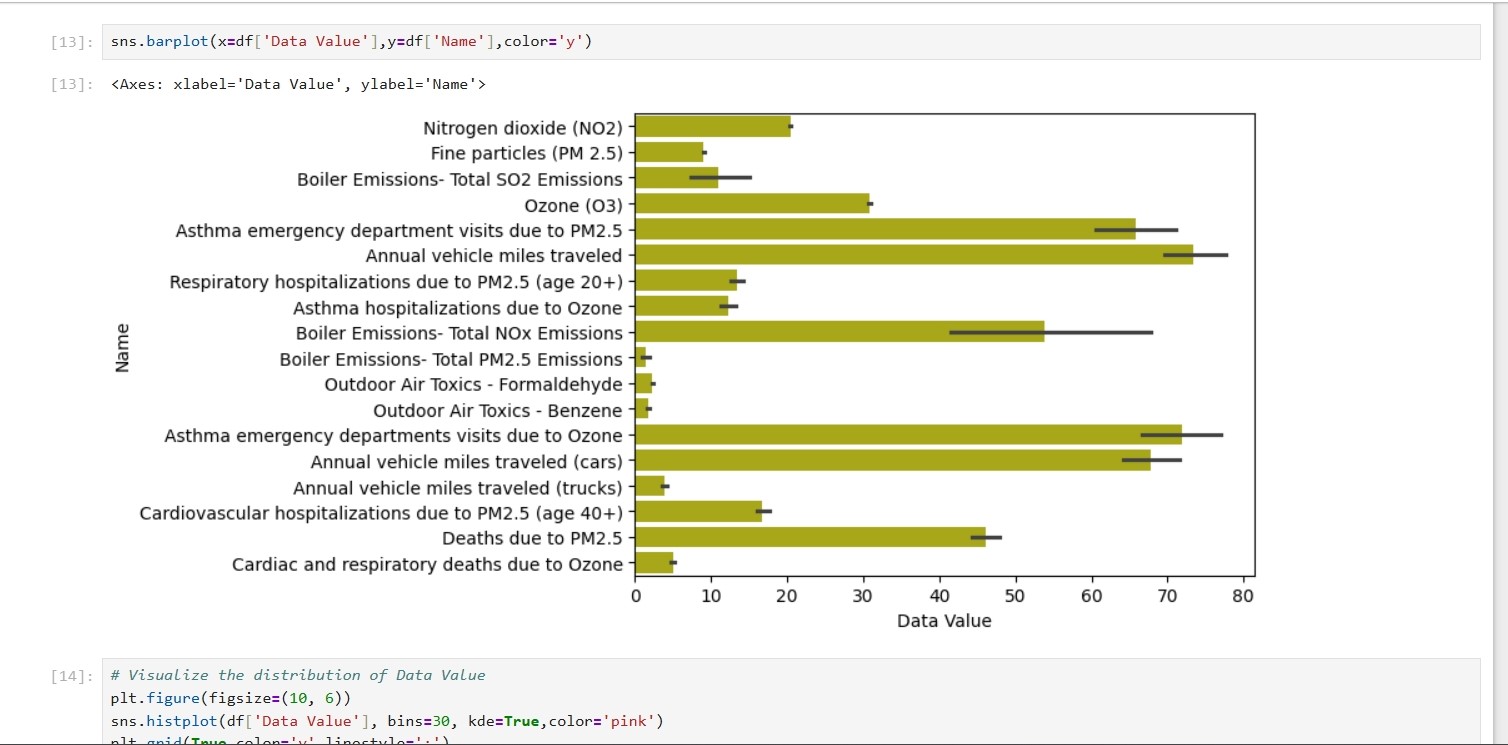




Now using heatmaps to detect relation between Unique ID, Geo Join ID, Data Value

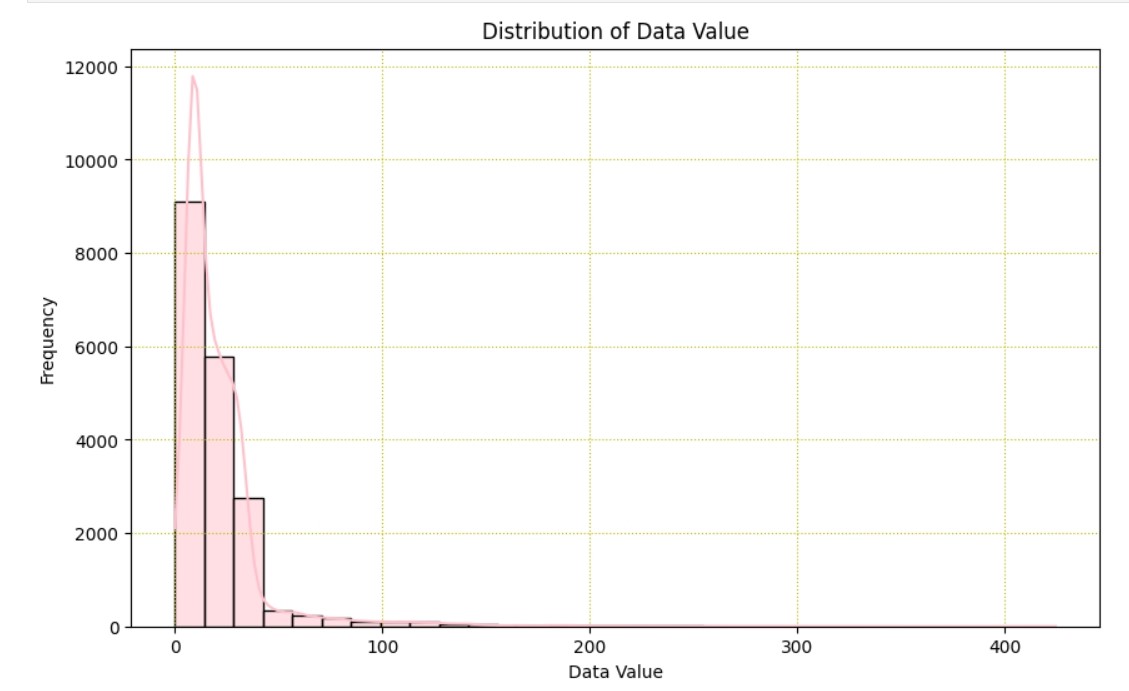


Plotting a barplot between Data Values and Name (Geo Names)

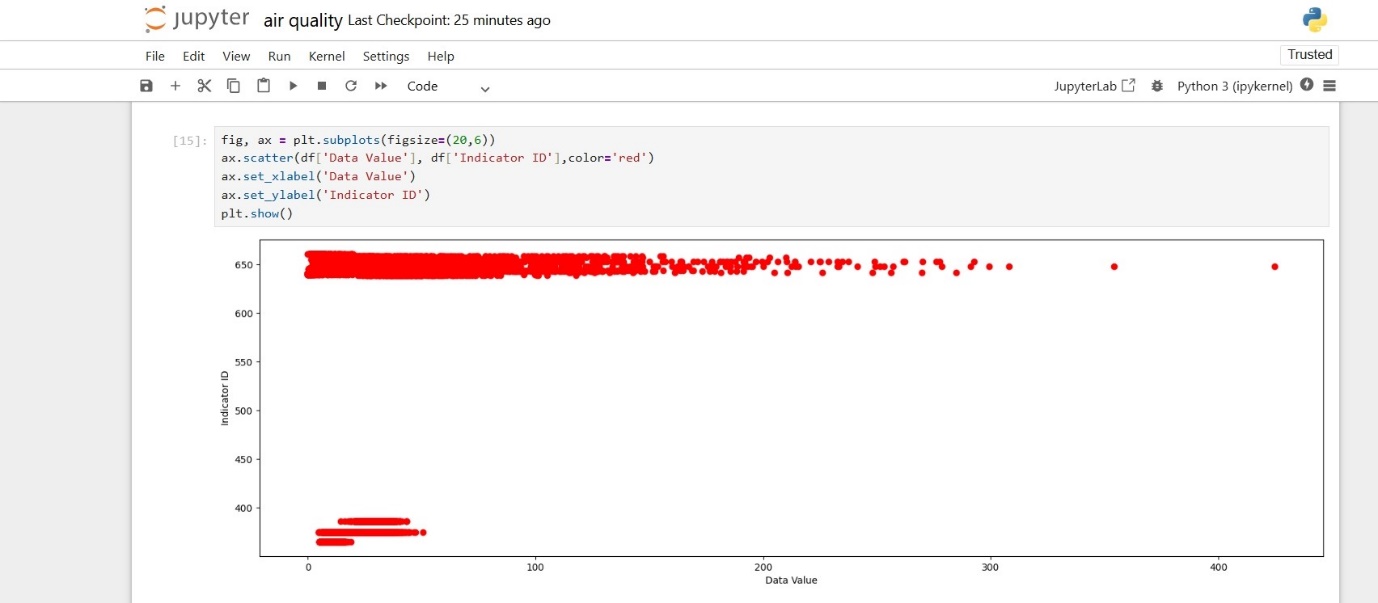


Now visulaizing the data by plotting a histogram with the help of matplot between Data Values and Frequency

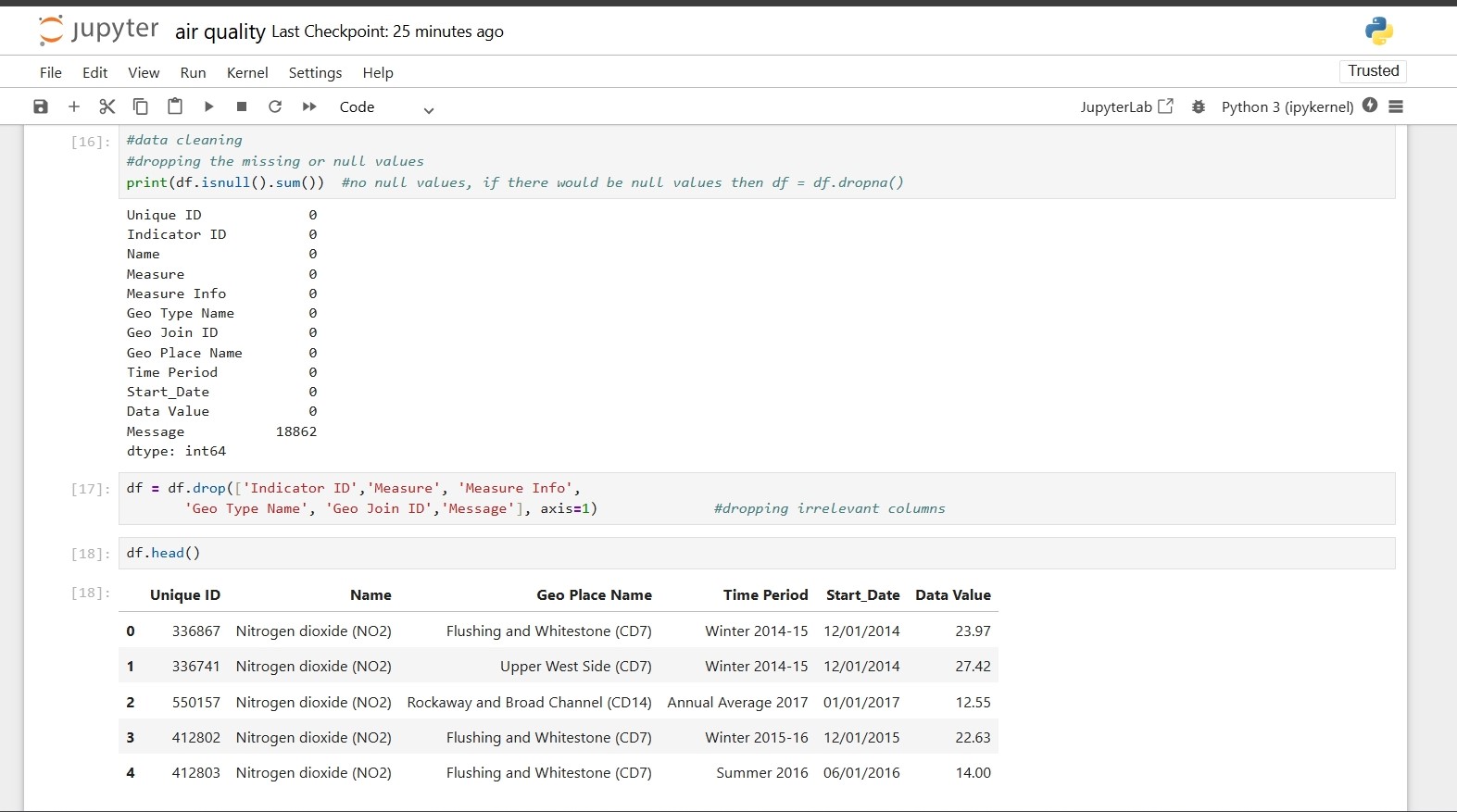




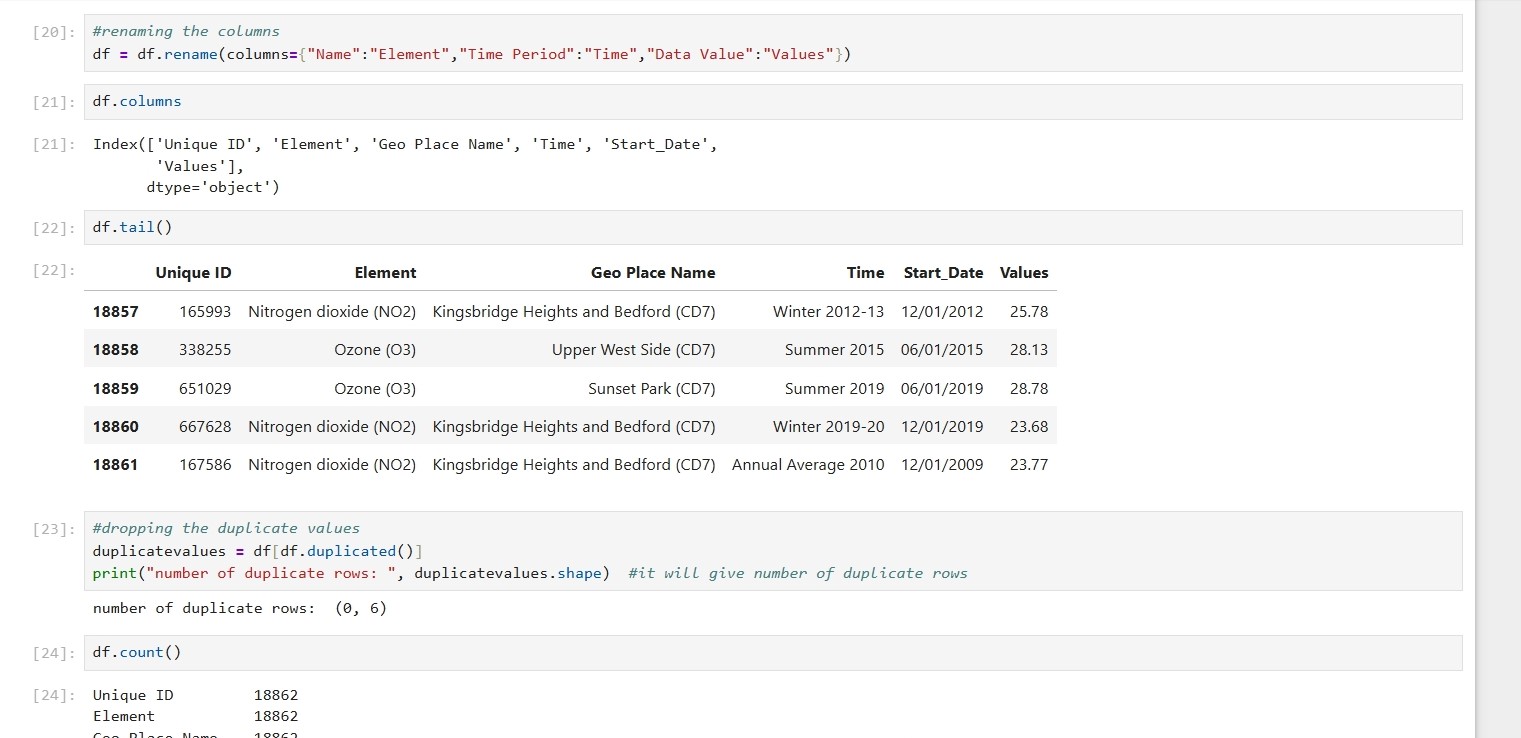
Now using scatterplot to visualize the data between Data Values and Indicator ID

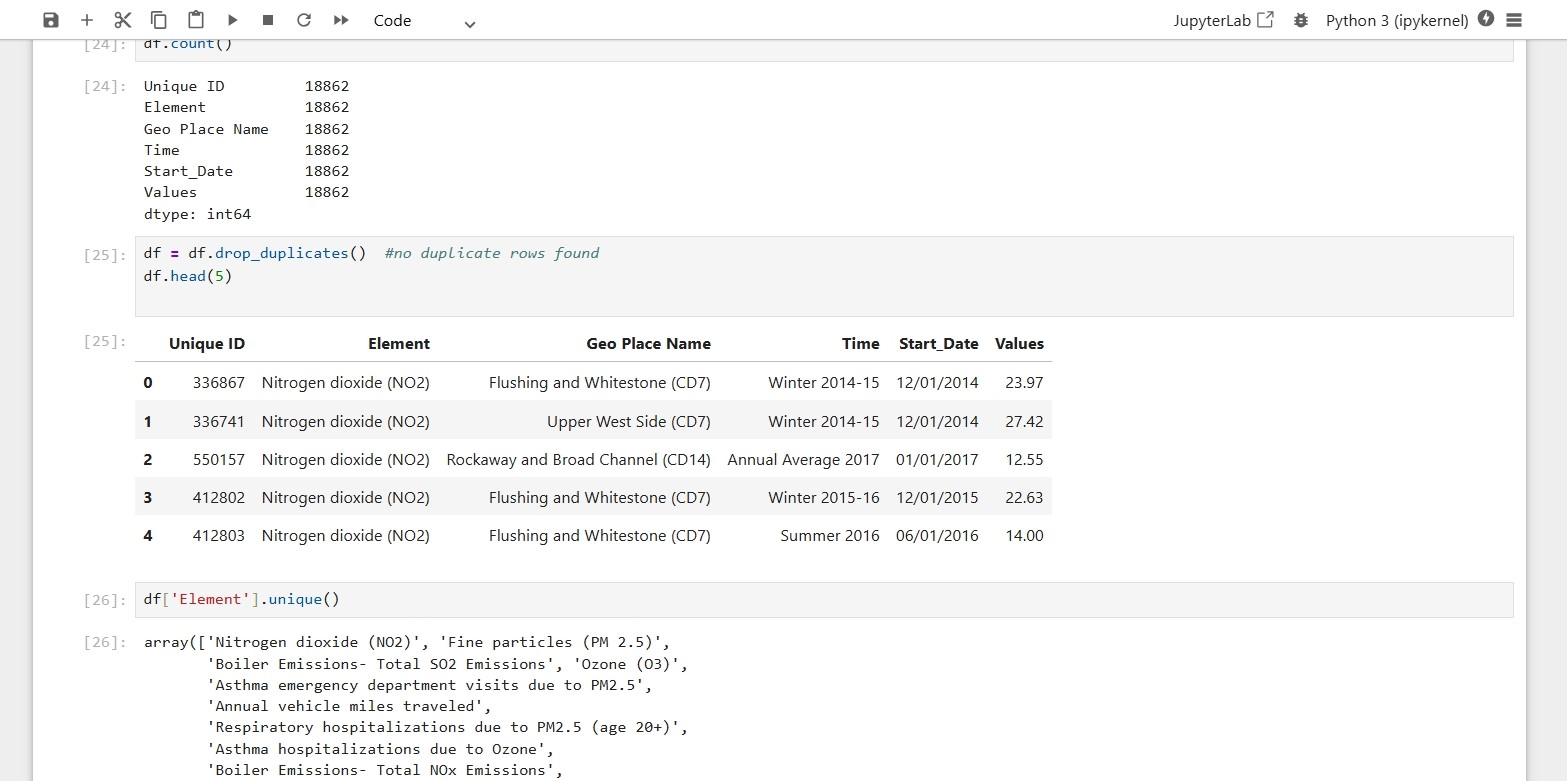


Cleaning the Data:-

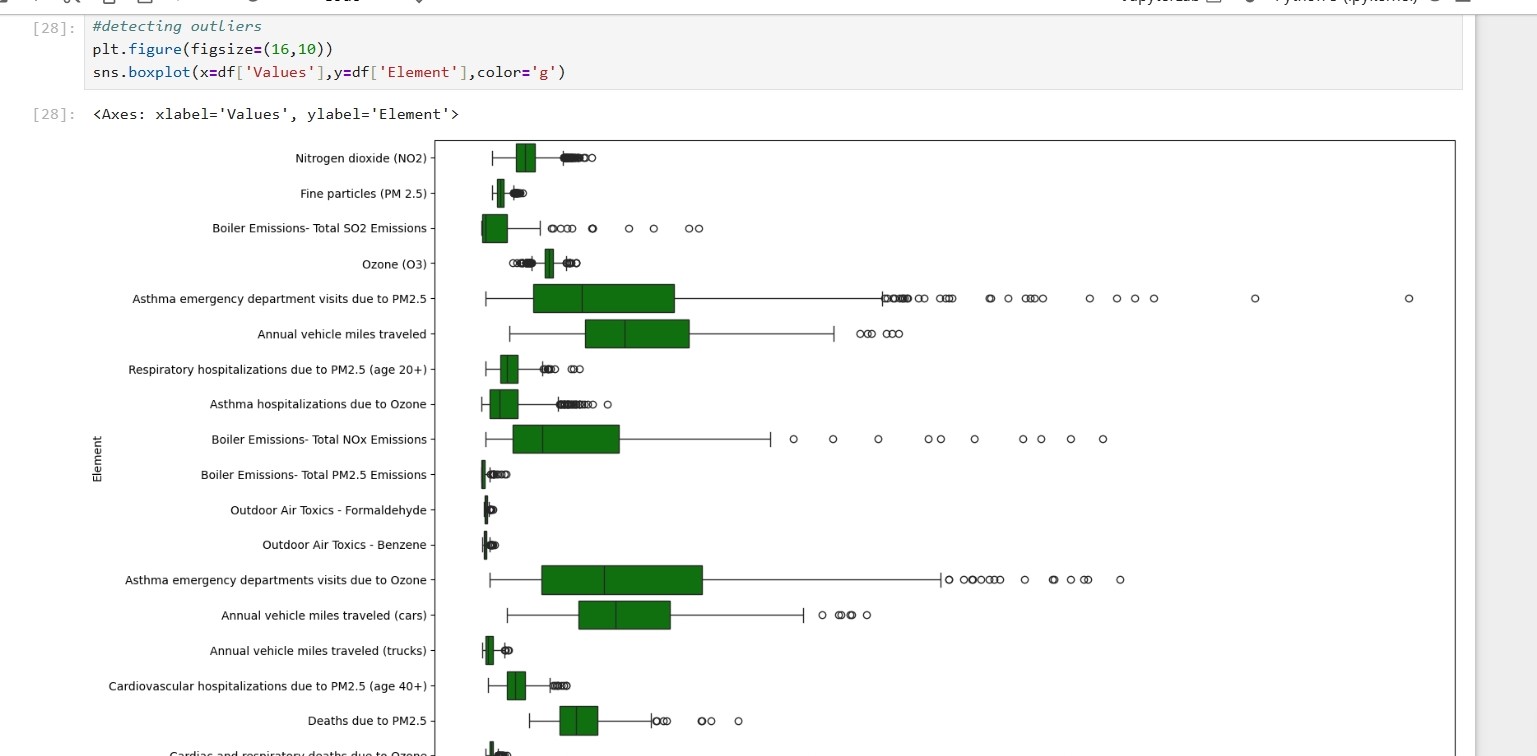


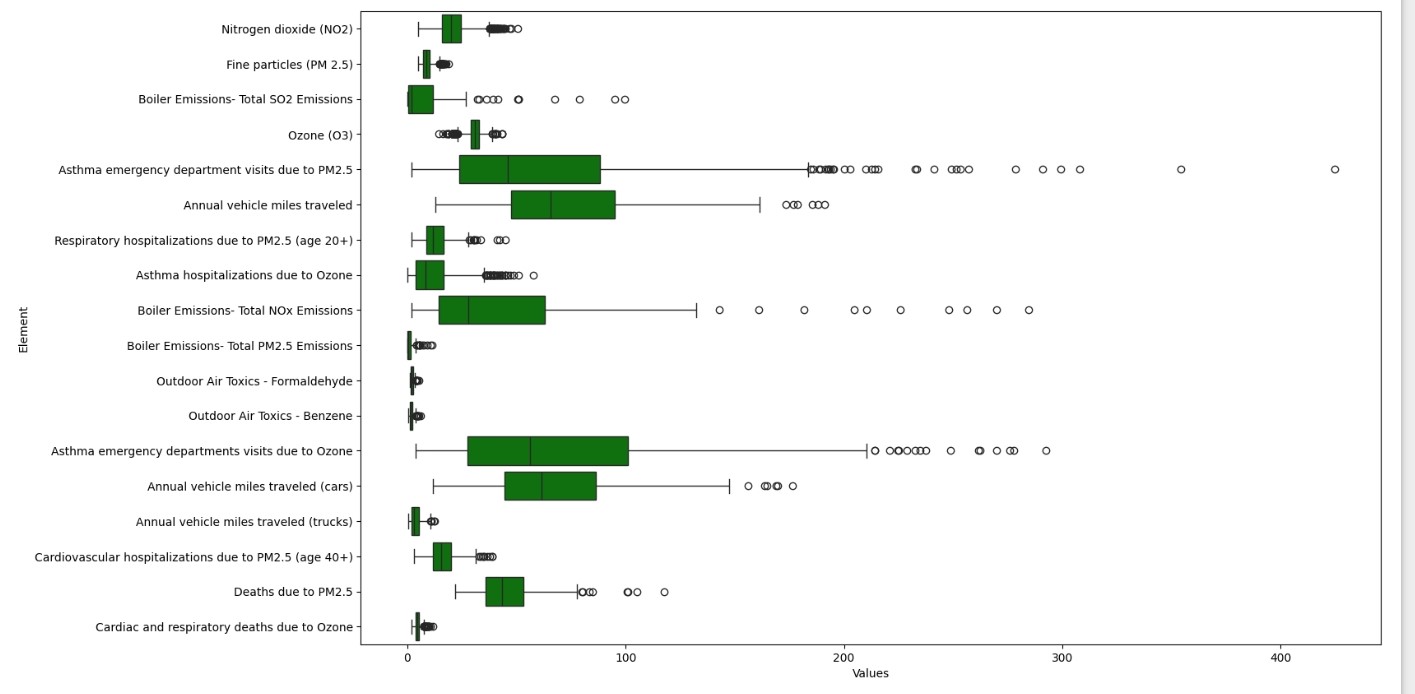
Dropping the data and then renaming the columns of the dataset





Making boxplot between Values and Elements

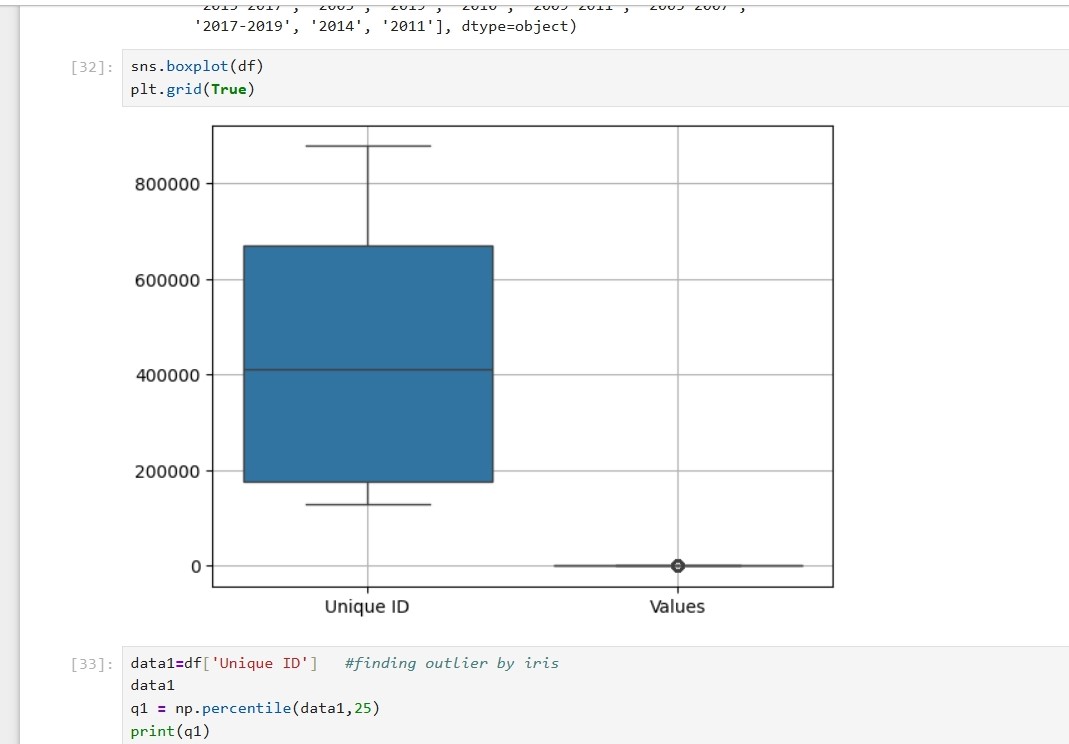




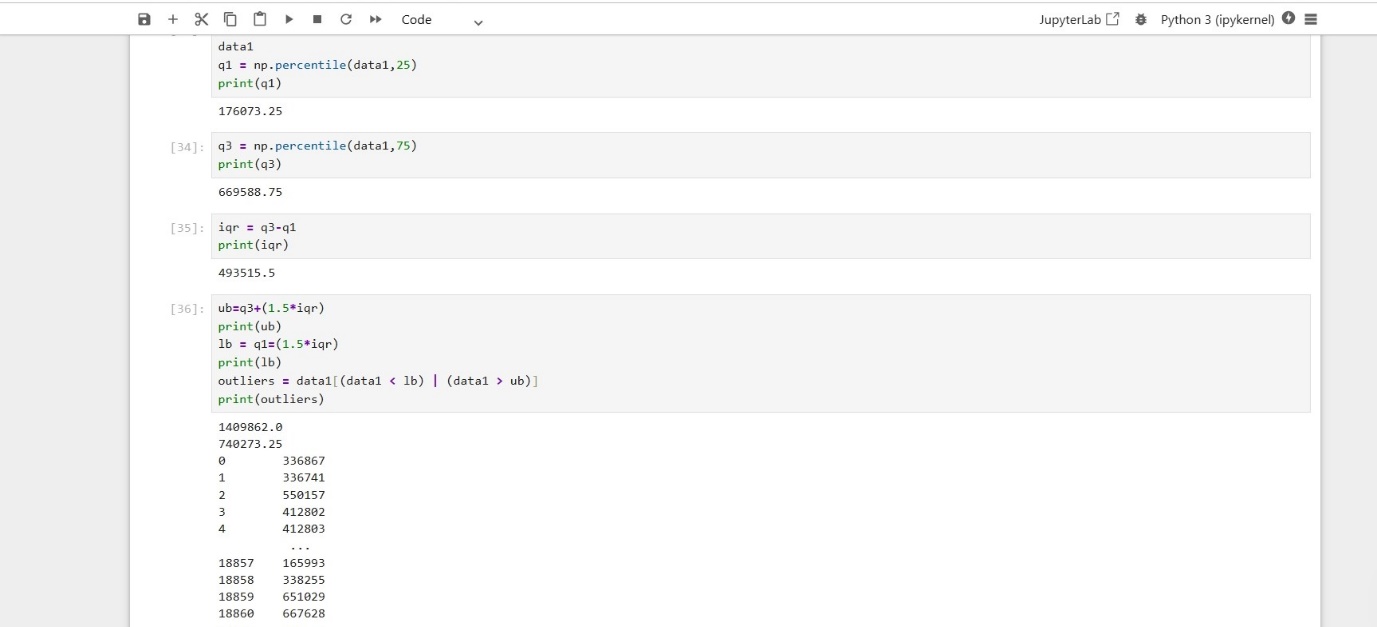
Boxplot of Start Date



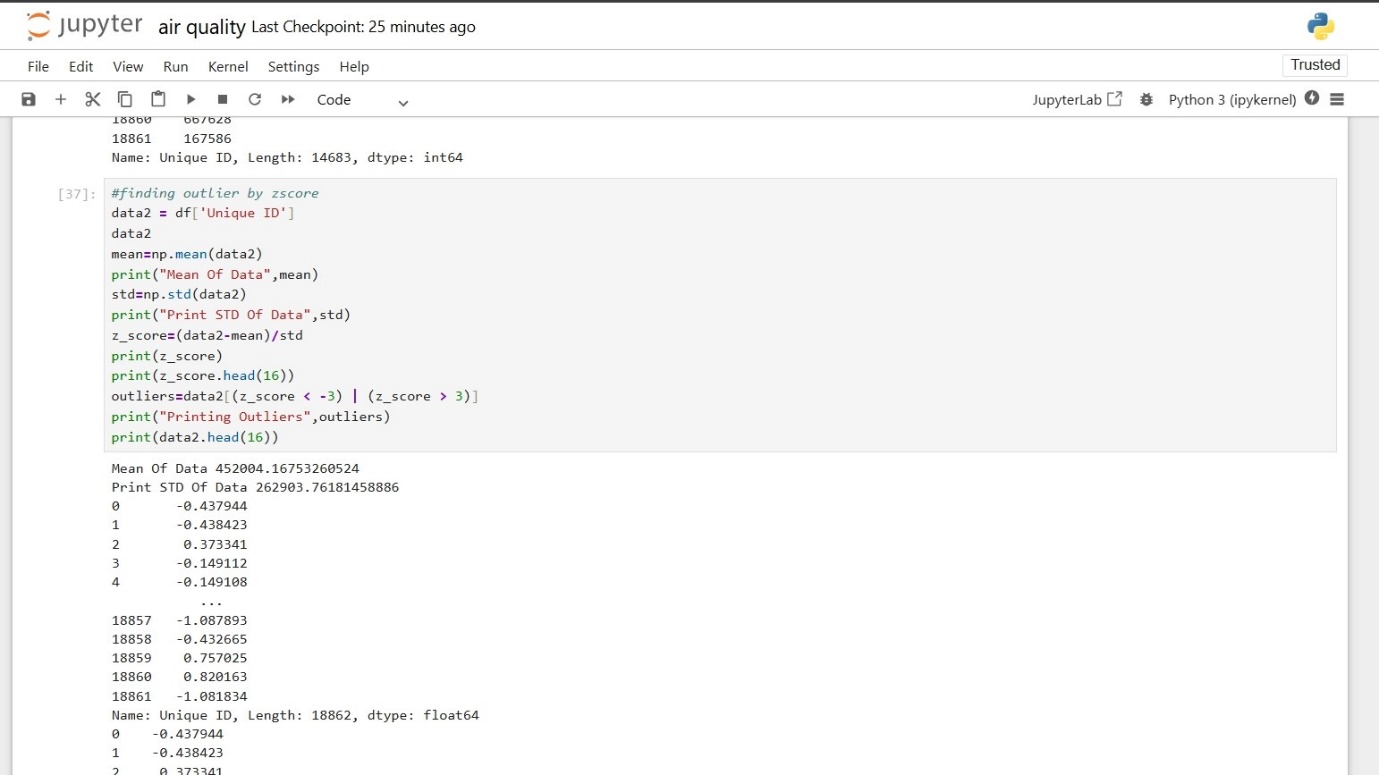
Boxplot of full dataset which includes only Unique Id and Values as they are only having numeric values

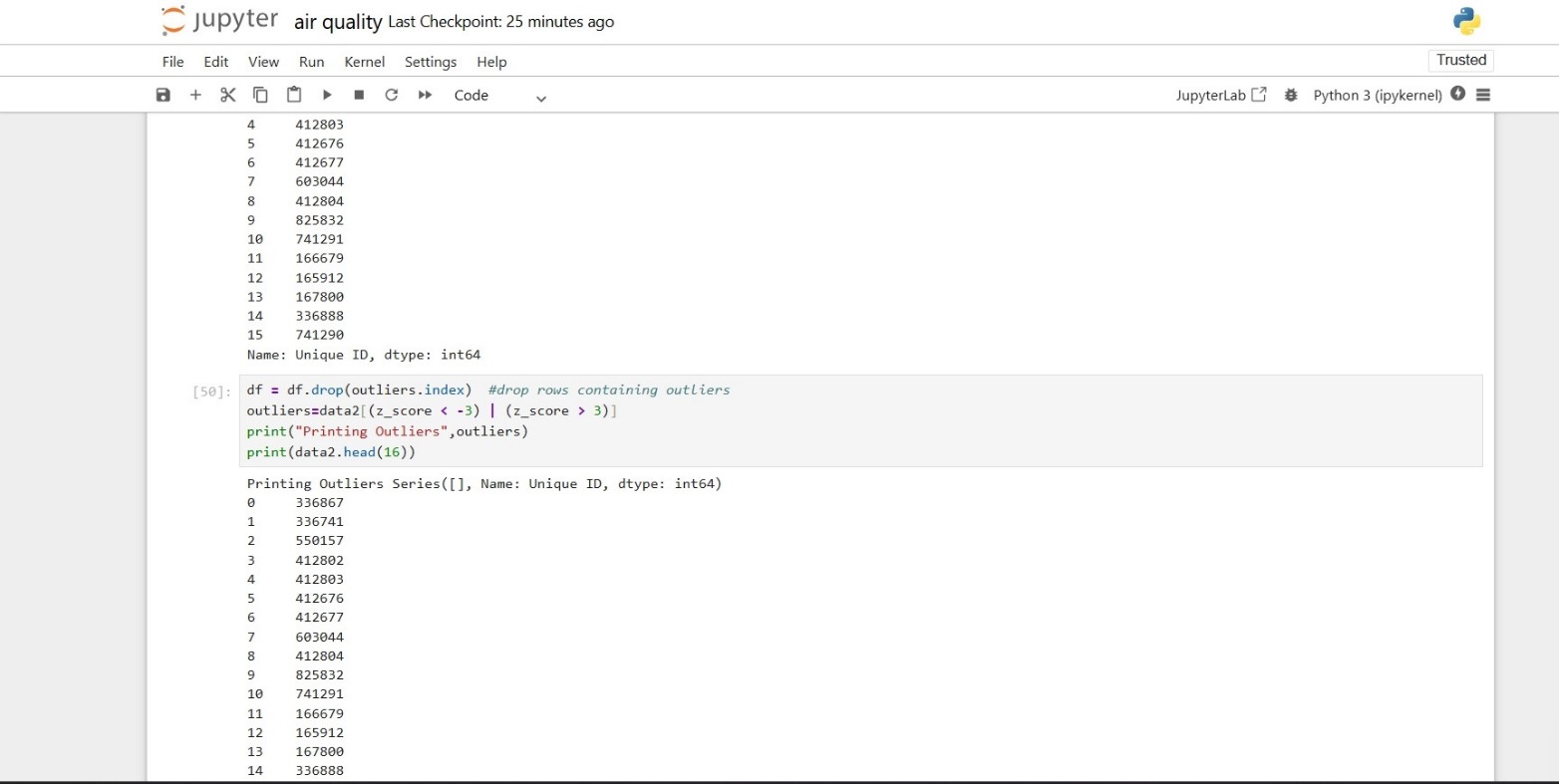


Finding outliers through iqr



Detecting outliers through z\_scores and then dropping the outliers





**5. Conclusion**

**This project served as an eye-opening exploration of the air quality conditions prevailing in various Indian cities. The data tells a sobering story: our urban centers, while growing rapidly, are also becoming hubs of air pollution that far exceed healthy limits. PM2.5 and PM10 remain the most consistent threats, especially during colder months when natural dispersal mechanisms like wind and rain are at their weakest.**

**Beyond the numbers, what this project really underscores is the pressing need for action. Clean air is a basic human right, yet millions breathe toxic air every day. The insights gained through this data can inform both public policy and personal behavior. Authorities can identify cities and times of year that require stricter regulations, while individuals can plan outdoor activities and take precautions on days when pollution levels are high.**

**From a technical standpoint, the project demonstrates the power of data science in tackling real-world problems. By using open-source tools and transparent methodologies, we were able to process, analyze, and visualize thousands of data points, ultimately deriving insights that could contribute to cleaner air and healthier lives.**

**6. Future Scope**

While this project provided a strong foundation for understanding air quality trends, there’s still so much more that can be explored. Future developments could include:

* **Real-time AQI forecasting using AI models**, helping cities and citizens prepare for pollution spikes.
* **Tracking the health effects of air pollution** by integrating medical and environmental datasets.
* **Using satellite data to monitor pollution in rural or under-observed regions**, improving coverage.
* **Developing public apps and dashboards** that visualize live air quality data in an accessible way.
* **Installing smart sensors** in high-traffic or industrial areas to gather localized and continuous pollution data.

By expanding on the current research, we can create smarter cities that respond quickly to environmental changes — making the air cleaner, the people healthier, and the future brighter.

**7. Social media**

**GITHUB LINK:** https://github.com/Amisha-Puri/Air-Quality.git

**LINKEDIN: https://www.linkedin.com/posts/amisha-puri-53860b298\_datascience-python-airquality-activity-7316716336108261376--SCX?utm\_source=share&utm\_medium=member\_desktop&rcm=ACoAAEf-VTwBVPbxqjuNeRzqwf4itnL2QRlI0N0**

**8. References**

1. https://www.gst.gov.in
2. https://data.gov.in
3. Seaborn and Pandas official documentation
4. Python Data Analysis Cookbook
5. https://catalog.data.gov/dataset?q=&sort=views\_recent+desc

**9. Appendix**

* Seasonal pollutant trends across cities
* Correlation heatmaps between all variables
* Outlier detection examples using boxplots
* Daily AQI category distribution
* Python code snippets for data cleaning and EDA