Big Data Analysis Project- Assignment 1 Part A

Impact of Climate Variability
on
Urban Air Pollution:
A Big Data Analysis of Sydney (2015–2025)

Utsav Punia : a1956304

The University of Adelaide

4533_COMP_SCI_7209 : Big Data Project

Table of Contents

1. Problem Description

Urban air pollution is an increasing concern in Australian cities, particularly in Sydney, where a growing population, rising vehicle emissions, and expanding urban infrastructure contribute to elevated levels of pollutants. Climate variability, including rising temperatures, changes in humidity, altered wind patterns, and more frequent extreme weather events, is suspected to intensify these pollution levels, but its specific influence in the Australian urban context remains underexplored (Bureau of Meteorology, 2024a; OECD, 2024). According to the OECD, exposure to air pollutants such as PM2.5 is one of the leading environmental risks to human health worldwide, making this a significant area of concern in developed urban centres.

This project aims to investigate how such climate factors contribute to urban air pollution over the last decade, using Sydney as a representative case study. Sydney was chosen due to its comprehensive monitoring infrastructure and publicly accessible air quality and weather data. The project focuses on understanding the complex interplay between climate variability and pollutant concentrations (such as PM2.5, NO_2, and O_3) using big data techniques. By combining large-scale historical datasets on air quality and meteorological conditions, the study seeks to provide insights that can aid in early warning systems, policy decisions, and urban environmental planning (NSW Department of Climate Change, Energy, the Environment and Water, 2023).

To ensure focus and feasibility, the original plan to include other cities such as Melbourne and Adelaide was reconsidered due to the absence of uniform and accessible APIs or long-format datasets for these regions. Thus, the scope was narrowed to Sydney, where structured hourly records spanning over 10 years are available for both pollutants and meteorological variables. This refined approach strengthens data reliability while still addressing a societally relevant question.

2. Initial Problem Formulation and Sub-Questions

Main Research Question:

❝How has climate variability influenced air pollution levels in Sydney over the past decade?❞

In accordance with the research question, following **sub-questions** are explored:

- **What are the key meteorological factors** (such as temperature, humidity, and wind) that contribute to variations in the concentrations of pollutants like PM2.5 and NO_2 in Sydney?
- **Are there identifiable trends in urban air quality that align with recurring climate patterns,** such as seasonal cycles or extreme weather events?
- **How do urbanization and climatic variability jointly influence average pollution level**s across different regions of Sydney?
- **Can predictive models be developed using historical meteorological and air quality data** to anticipate pollution spikes?

3. Aim and Objectives

3.1. Aim of the project

To analyze ten years of meteorological and air pollution data in Sydney to explore how climate variability affects urban air quality, using big data techniques and initial exploratory analysis. The goal is to derive insights that can support environmental monitoring, public health policy, and potential predictive modelling in later stages of the project.

3.2. Objectives

- To examine large-scale, hourly weather and pollution data collected across key monitoring sites in Sydney from 2015 to 2025.
- To identify which climate variables (e.g., temperature, humidity, wind speed) show the strongest associations with pollutant levels such as PM2.5, NO_2, and O_3.
- To detect seasonal patterns and event-based trends (e.g., during heatwaves or cold fronts) in pollution data.
- To assess whether the available data is suitable for building predictive models in future stages.
- To demonstrate how big data methods can be applied to address urban sustainability

and air quality management challenges.

4. Dataset Description

4.1. Overview of Datasets

This project integrates two primary datasets obtained through API::

- **Air Quality Monitoring Data**, which includes hourly pollutant concentration levels recorded at monitoring stations across New South Wales (NSW), including the Sydney region.
- **Meteorological Data**, which contains hourly environmental variables such as temperature, humidity, wind speed, wind direction, and sigma theta.

Together, these datasets provide a comprehensive basis for analyzing the relationship between climate variability and pollution trends over time.

4.2. Data Source and Structure

The datasets were accessed using the **NSW Government's Air Quality API**, which provides structured environmental observations from its ambient monitoring network. A custom POST request was designed in Python to collect data across selected parameters and monitoring stations. The retrieved data, initially in JSON format, was then normalized and exported to CSV for further analysis.

Source and Documentation: The datasets were accessed using the NSW Government's Air Quality API (NSW Government, 2024a), which provides structured environmental observations. Full documentation is available in the API User Guide (NSW Government, 2024b).

Key characteristics of the retrieved data:

- **Data Format:** Data retrieved in JSON format Converted to CSV.
- **Temporal coverage:** Hourly data from June 1 ,2015 to June 1,2025.
- **Spatial coverage**: Data obtained from five monitoring sites distributed across urban regions of Sydney, covering eastern, western, and northwestern zones.
- **Fields included:**

Field Name	Description
Site_Id	Unique identifier for the monitoring site
Date	Date of observation
Hour	Hour of observation (1 to 24)

HourDescription	Time range for the hour (e.g., "12 am - 1 am")
Value	Measured value of the parameter
AirQualityCategory	Category assigned based on pollutant thresholds
DeterminingPollutant	Pollutant used to determine air quality category
Parameter.ParameterCode	Code representing the measured parameter (e.g., PM2.5
Parameter.ParameterDescription	Full name of the parameter
Parameter.Units	Abbreviation of the unit of measurement
Parameter.UnitsDescription	Full description of the unit of measurement
Parameter.Category	Type of parameter (e.g., Averages, Exceedences)
Parameter.SubCategory	Subtype within category (e.g., Hourly)
Parameter.SubCategory	Frequency of measurement (e.g., Hourly average)

- **Selected Parameters**

  The variables extracted for analysis include both pollutants and weather-related measurements which were collected under Parameter.ParameterCode :

Table A: Air Pollutants Monitored in the Study

Parameter	**Description**
PM10	Particulate Matter ≤10 micrometres in diameter, affects respiratory health
PM2.5	Fine Particulate Matter ≤2.5 micrometres, penetrates deeper into lungs and bloodstream
NO2	Nitrogen Dioxide, a traffic-related pollutant harmful to lungs

CO	Carbon Monoxide, a gas that reduces oxygen delivery in the body
OZONE	Ground-level Ozone, formed by chemical reactions involving NO_2 and sunlight; irritates airways

Table B: Meteorological Variables Used in the Analysis

Parameter	**Description**
TEMP	Air temperature (°C)
HUMID	Relative humidity (%)
WSP	Wind speed (m/s)
WDR	Wind speed (m/s)
SD1	Sigma Theta – Standard deviation of wind direction (°)

- **Monitoring Sites Selected**

  The following five monitoring sites in Sydney were selected for this study:

  Table C: Selected Monitoring Sites in Sydney

Site Name	**Site ID**	**Region**
Rozelle	39	Sydney East
Chullora	222	Sydney East
Parramatta North	919	Sydney North-west
Campbelltown West	2560	Sydney South-west
Liverpool	107	Sydney South-west

These stations were chosen because they span across geographically diverse urban regions of Sydney including eastern, western, and southwestern corridors offering a balanced spatial distribution. More importantly, these locations report a wide range of both meteorological and pollutant variables (including PM2.5, NO_2, CO, and wind metrics) consistently and reliably, making them ideal for comprehensive temporal and spatial analysis. Site metadata and data availability were confirmed via the **NSW Air Quality API** and its official documentation (NSW DCCEEW, 2023).

4.3. Suitability for Big Data Analysis

The datasets used meet key Big Data criteria:

- **Volume**: Over ten years of hourly data from multiple sites result in hundreds of thousands of records.
- **Variety**: The data includes structured tabular formats from APIs and CSVs, covering multiple pollutant types and meteorological variables.
- **Velocity**: The sources used support frequent updates or continuous data availability, which can be leveraged for real-time forecasting or alert systems in future applications.
- **Veracity**: Sourced from a verified government system, the data is considered reliable for analysis and modelling.

5. Initial Data Processing and Pitfalls

5.1. Overview

The dataset was obtained from the NSW Government's Air Quality API and processed following the structure outlined in the official API documentation, referencing the code templates shared by the department (NSW Department of Climate Change, Energy, the Environment and Water, 2023). The objective at this stage was to convert raw API responses into a time-aligned, analysis-ready dataset through a series of well-defined preprocessing steps, including data extraction, normalization, timestamp construction, transformation, and forward-filling of missing values.

5.2. Timestamp Construction and Formatting

The raw API response separated date and hour into two fields: `Date` (e.g., 01-06-2017) and `HourDescription` (e.g., 12 am – 1 am). A regular expression was applied to extract the starting hour string (e.g., 12 am) and convert it into 24-hour format . The resulting timestamp was created by combining the parsed date and hour into a unified datetime value such as:

2017-06-01 00:00:00

This allowed for precise time indexing of hourly observations. Any rows where this transformation failed (e.g., due to malformed or missing hour strings) were removed.

5.3. Data Transformation and Pivoting

After timestamp creation, the dataset—originally in long format—was reshaped into wide format using a pivot operation. For instance, what was previously multiple rows for the same time (each representing a different parameter) became a single row per site and timestamp.

Example(after pivoting):

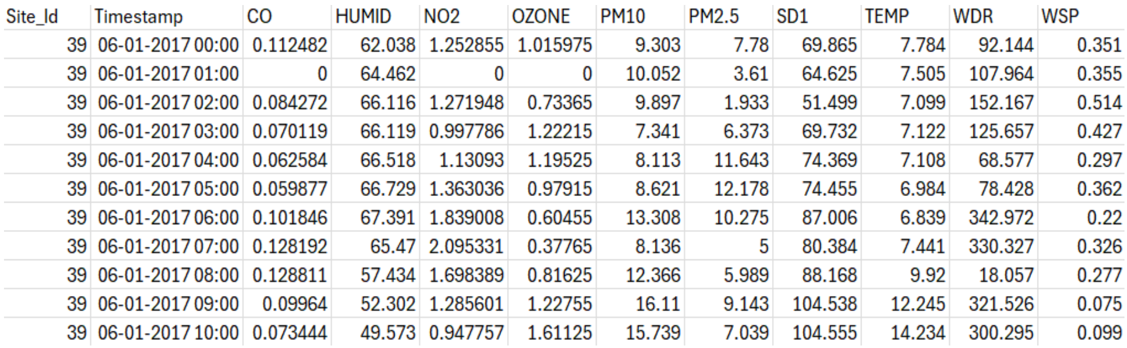
Site_Id	**Timestamp**	**CO**	**HUMID**	**NO2**	**OZONE**	**PM10**
39	06-01-2017 00:00	0.112482	62.038	1.252855	1.015975	9.303

5.4. Missing Value Handling

Missing parameter values were handled using forward-fill (fillna(method='ffill')), which propagates the last observed value forward until a new measurement is recorded. This method is particularly appropriate for environmental sensor datasets, where temporary gaps in data collection are common but values tend to evolve gradually. Forward-filling ensures time-series continuity without introducing artificial zero values or statistical distortions.

5.5. Output Format and Analysis Readiness

The following data format is achieve dafter data acquisition and first stage preprocessing :



Site_Id	Timestamp	CO	HUMID	NO2	OZONE	PM10	PM2.5	SD1	TEMP	WDR	WSP
39	06-01-2017 00:00	0.112482	62.038	1.252855	1.015975	9.303	7.78	69.865	7.784	92.144	0.351
39	06-01-2017 01:00	0	64.462	0	0	10.052	3.61	64.625	7.505	107.964	0.355
39	06-01-2017 02:00	0.084272	66.116	1.271948	0.73365	9.897	1.933	51.499	7.099	152.167	0.514
39	06-01-2017 03:00	0.070119	66.119	0.997786	1.22215	7.341	6.373	69.732	7.122	125.657	0.427
39	06-01-2017 04:00	0.062584	66.518	1.13093	1.19525	8.113	11.643	74.369	7.108	68.577	0.297
39	06-01-2017 05:00	0.059877	66.729	1.363036	0.97915	8.621	12.178	74.455	6.984	78.428	0.362
39	06-01-2017 06:00	0.101846	67.391	1.839008	0.60455	13.308	10.275	87.006	6.839	342.972	0.22
39	06-01-2017 07:00	0.128192	65.47	2.095331	0.37765	8.136	5	80.384	7.441	330.327	0.326
39	06-01-2017 08:00	0.128811	57.434	1.698389	0.81625	12.366	5.989	88.168	9.92	18.057	0.277
39	06-01-2017 09:00	0.09964	52.302	1.285601	1.22755	16.11	9.143	104.538	12.245	321.526	0.075
39	06-01-2017 10:00	0.073444	49.573	0.947757	1.61125	15.739	7.039	104.555	14.234	300.295	0.099

Fig a: Sample Output after Initial Data Processing

This format is suitable for the following analyses:

- Correlation analysis between pollutants and meteorological factors
- Time-series trend identification
- Feature engineering for potential predictive modelling tasks

5.6. Identified Limitations and Future Considerations

- Temporal Scope: The current data covers only June 2017. Scaling to the full 10-year range will require batching, automation, and careful handling of API rate limits and storage constraints.
- **Geographic Scope Limitation:**
  The project was originally intended to include a comparative analysis across multiple Australian cities, such as Melbourne and Adelaide. However, during the data collection phase, it became evident that long-term, high-resolution hourly data for pollutants and meteorological variables was either unavailable or inconsistent outside New South Wales. As a result, the study was scoped specifically to Sydney, where reliable and extensive historical data could be obtained via the NSW Air Quality API, which is also mentioned in the Problem Description.
- **Occasional Missing Parameters:**
  Although the selected sites report all required variables, temporary sensor outages result in occasional missing values. Forward-filling was used to maintain continuity, but these gaps may affect accuracy during sensitive modelling.
- **Dropped Records from Timestamp Errors:**
  A small number of rows were dropped due to parsing issues in HourDescription. These were excluded to ensure the integrity of datetime alignment.

6. Refined Problem and Plan

6.1. Refined Problems and Support Questions

Main Research Question:

❝How has climate variability influenced air pollution levels in Sydney over the past decade?❞

Supporting Questions:

- **What are the key meteorological factors** (such as temperature, humidity, and wind) that contribute to variations in the concentrations of pollutants like PM2.5 and NO_2 in Sydney?
- **Are there identifiable trends in urban air quality that align with recurring climate patterns**, such as seasonal cycles or extreme weather events?
- **How do urbanization and climatic variability combining influence** average pollution levels across different regions of Sydney?
- **Can predictive models be developed using historical meteorological and air quality data to anticipate pollution spikes?**

The research and support questions remain unchanged, as the data obtained supports a robust analysis within the Sydney region.

Note: *The change in problem description by eliminating the other cities was handled during the data acquisition phase, and the research question presented in the Initial Problem Formulation was framed after this major refinement was made.*

6.2. Next Steps for Analysis

With a cleaned and time-aligned dataset in place, the next phase of the project involves exploring the relationships between climate conditions and air pollution levels across different parts of Sydney. The analysis will focus on both statistical trends and spatial variations to better understand how meteorological changes shape urban air quality.

6.2.1. Investigating Relationships Between Pollutants and Weather Conditions

To identify which climate factors influence pollutant levels, we will begin with correlation analysis. Pearson correlation coefficients will be computed between each pollutant (PM2.5, PM10, NO_2, CO, O_3) and the selected meteorological variables (temperature, humidity, wind speed, wind direction, sigma theta). A heatmap of these correlations will help visualise where strong associations exist, such as between high temperatures and ozone levels, or low wind speeds and PM2.5 accumulation.

6.2.2. Temporal Trend Analysis

To uncover how pollution behaves over time, we will analyse hourly and daily variations across the dataset. Line plots will be used to show short-term fluctuations, while boxplots grouped by hour of day and day of week will help identify recurring patterns.

6.2.3. Comparing Pollution Across Different Locations

By comparing pollutant levels across the five monitoring sites, we aim to understand how air quality varies geographically within Sydney. Site-wise boxplots will highlight differences in average concentrations and outliers, while multi-panel time plots will allow side-by-side visual comparisons of trends at each location.

6.2.4. Exploring Pollution Under Specific Weather Conditions

To better understand how extreme weather affects pollution, the data will be grouped by climate thresholds,e,g, high temperature days versus cooler ones, or windy versus still conditions. By comparing pollutant levels across these groups, we can examine whether certain meteorological states contribute to pollution build-up or dispersion.

6.2.5. Understanding Distributions and Preparing for Modelling

Histograms and density plots will be used to examine the distribution of each variable. This step will help identify skewed data, possible outliers, or the need for transformations before modelling. Features such as rolling averages or lagged pollutant values may also be generated for the predictive modelling phase in future scope.

6.3. Backup Question and Data Source

If long-term data retrieval becomes infeasible due to persistent gaps, inconsistent parameter reporting, or rate limitations in the NSW Air Quality API, a narrower, event-based research question has been prepared:

❝ ***What are the typical pollution patterns during high-temperature events in Sydney, and how do they vary across different urban locations?*** ❞

This question shifts the focus from long-term trends to short-term climate events. It allows for meaningful analysis using targeted time windows, such as periods of extreme heat, where pollutant behaviour is known to deviate from normal patterns. These windows can be selected using historical climate summaries provided by the Bureau of Meteorology(Bureau of Meteorology (2024b)).

The same NSW Air Quality API can be used to extract relevant air quality and meteorological data for those specific dates. This approach is more manageable in terms of data volume and does not require full coverage over a decade. It remains consistent with the overall aim of exploring the relationship between climate variability and air quality in Sydney.

This backup plan ensures that the project remains viable even if full historical datasets cannot be obtained or prepared on time.

\

7. References

1. Bureau of Meteorology (2024a). *Climate trends and extreme weather in Australia*. Available at: https://www.bom.gov.au/climate/change/ [Accessed 12 June 2025].
2. **Bureau of Meteorology (2024b).** *Climate history.* Available at: http://www.bom.gov.au/climate/history/ [Accessed 12 June 2025].
3. NSW Department of Climate Change, Energy, the Environment and Water (2023). *Air Quality Monitoring Network Overview*. Available at: https://www.environment.nsw.gov.au/ [Accessed 12 June 2025].
4. NSW Government (2024a). *Air Quality API*. Available at: https://www.airquality.nsw.gov.au/air-quality-data-services/air-quality-api [Accessed 12 June 2025].
5. NSW Government (2024b). *Air Quality API – Application Programming Interface User Guide*. Available at: https://www.environment.nsw.gov.au/sites/default/files/air-quality-application-programming-interface-user-guide-210346.pdf [Accessed 12 June 2025].
6. OECD (2024). *Air pollution*. Available at: https://www.oecd.org/en/topics/sub-issues/air-pollution.html [Accessed 12 June 2025].

8. Appendix

8.1. Site Metadata Endpoint

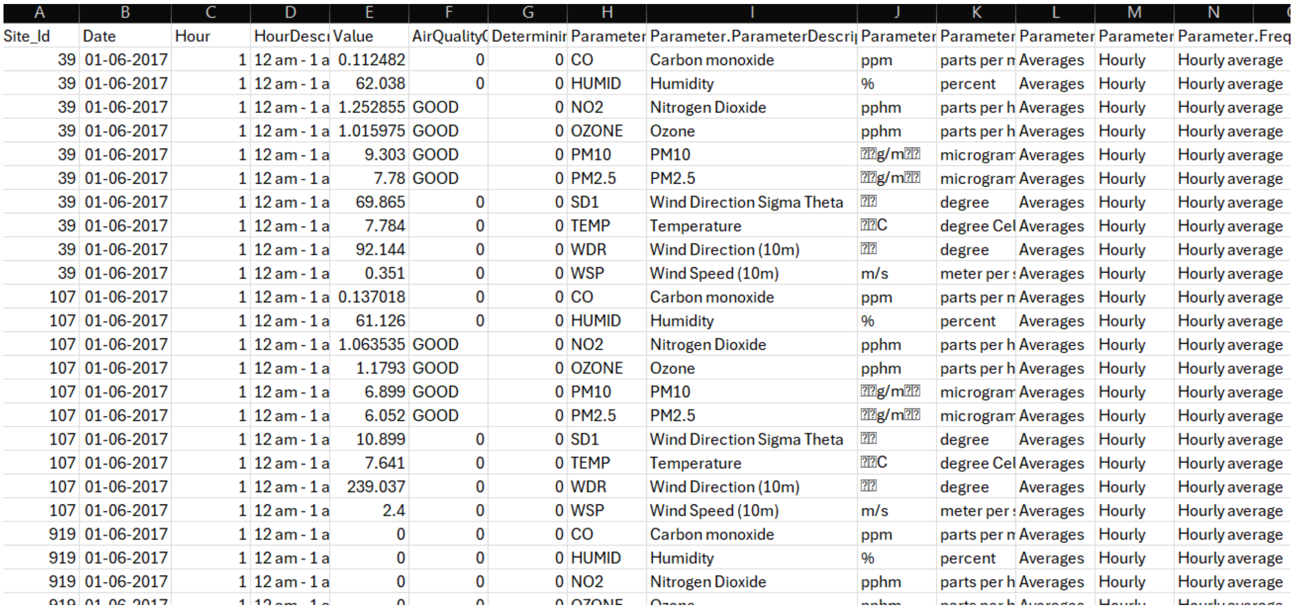
URL: https://data.airquality.nsw.gov.au/api/Data/get_SiteDetails

The correct site ids were taken from this endpoint.

8.2. Parameter Metadata Endpoint

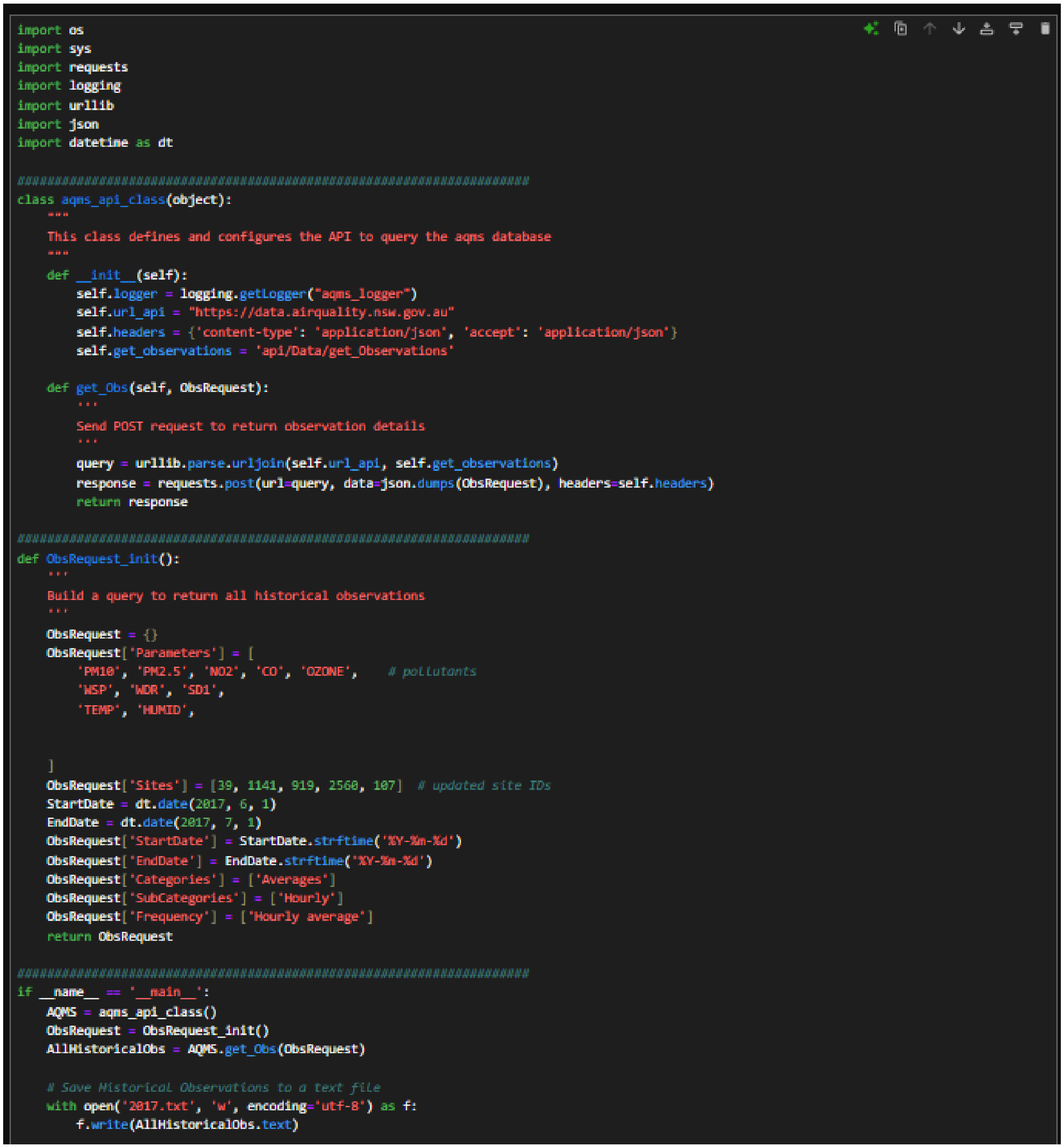
This endpoint provided the correct parameters that are required to be passed in ths OBSRequest function while calling the api for data acquisition.

8.3. Raw Data from API (before first-stage data pre-processing):



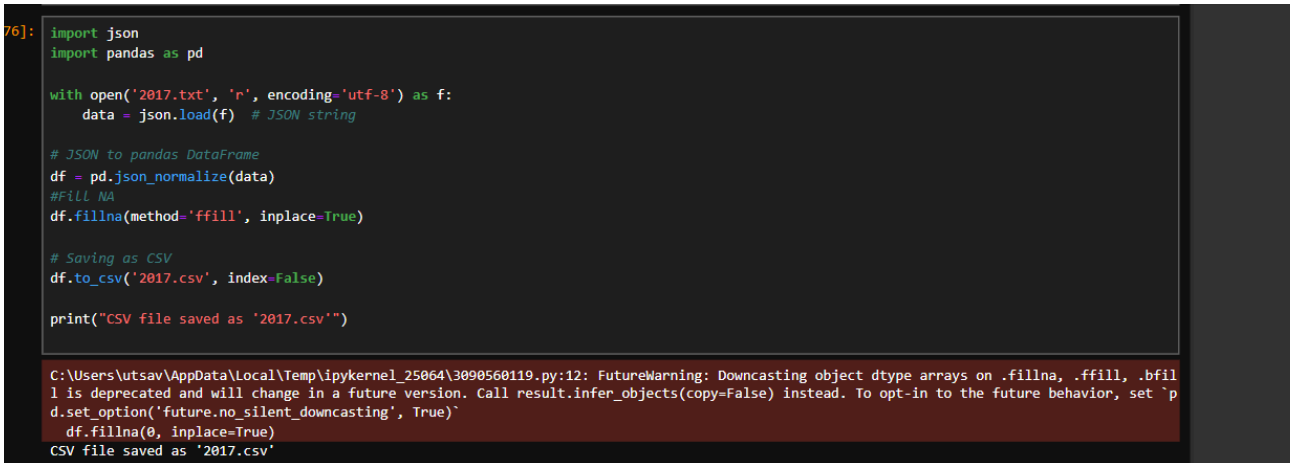
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Site_Id	Date	Hour	HourDesc	Value	AirQuality(	Determinir	Parameter	Parameter.ParameterDescri	Parameter	Parameter	Parameter	Parameter	Parameter.Freq
39	01-06-2017	1	12 am - 1 a	0.112482	0	0	CO	Carbon monoxide	ppm	parts per n	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	62.038	0	0	HUMID	Humidity	%	percent	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	1.252855	GOOD	0	NO2	Nitrogen Dioxide	pphm	parts per h	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	1.015975	GOOD	0	OZONE	Ozone	pphm	parts per h	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	9.303	GOOD	0	PM10	PM10	⍰⍰g/m⍰⍰	microgram	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	7.78	GOOD	0	PM2.5	PM2.5	⍰⍰g/m⍰⍰	microgram	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	69.865	0	0	SD1	Wind Direction Sigma Theta	⍰⍰	degree	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	7.784	0	0	TEMP	Temperature	⍰⍰C	degree Cel	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	92.144	0	0	WDR	Wind Direction (10m)	⍰⍰	degree	Averages	Hourly	Hourly average
39	01-06-2017	1	12 am - 1 a	0.351	0	0	WSP	Wind Speed (10m)	m/s	meter per s	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	0.137018	0	0	CO	Carbon monoxide	ppm	parts per n	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	61.126	0	0	HUMID	Humidity	%	percent	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	1.063535	GOOD	0	NO2	Nitrogen Dioxide	pphm	parts per h	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	1.1793	GOOD	0	OZONE	Ozone	pphm	parts per h	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	6.899	GOOD	0	PM10	PM10	⍰⍰g/m⍰⍰	microgram	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	6.052	GOOD	0	PM2.5	PM2.5	⍰⍰g/m⍰⍰	microgram	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	10.899	0	0	SD1	Wind Direction Sigma Theta	⍰⍰	degree	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	7.641	0	0	TEMP	Temperature	⍰⍰C	degree Cel	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	239.037	0	0	WDR	Wind Direction (10m)	⍰⍰	degree	Averages	Hourly	Hourly average
107	01-06-2017	1	12 am - 1 a	2.4	0	0	WSP	Wind Speed (10m)	m/s	meter per s	Averages	Hourly	Hourly average
919	01-06-2017	1	12 am - 1 a	0	0	0	CO	Carbon monoxide	ppm	parts per n	Averages	Hourly	Hourly average
919	01-06-2017	1	12 am - 1 a	0	0	0	HUMID	Humidity	%	percent	Averages	Hourly	Hourly average
919	01-06-2017	1	12 am - 1 a	0	0	0	NO2	Nitrogen Dioxide	pphm	parts per h	Averages	Hourly	Hourly average
919	01-06-2017	1	12 am - 1 a	0	0	0	OZONE	Ozone	pphm	parts per h	Averages	Hourly	Hourly average

8.4. Code Screenshot 1:



```python
import os
import sys
import requests
import logging
import urllib
import json
import datetime as dt

######################################################################
class aqms_api_class(object):
    """
    This class defines and configures the API to query the aqms database
    """
    def __init__(self):
        self.logger = logging.getLogger("aqms_logger")
        self.url_api = "https://data.airquality.nsw.gov.au"
        self.headers = {'content-type': 'application/json', 'accept': 'application/json'}
        self.get_observations = 'api/Data/get_Observations'

    def get_Obs(self, ObsRequest):
        '''
        Send POST request to return observation details
        '''
        query = urllib.parse.urljoin(self.url_api, self.get_observations)
        response = requests.post(url=query, data=json.dumps(ObsRequest), headers=self.headers)
        return response

######################################################################
def ObsRequest_init():
    '''
    Build a query to return all historical observations
    '''
    ObsRequest = {}
    ObsRequest['Parameters'] = [
        'PM10', 'PM2.5', 'NO2', 'CO', 'OZONE',    # pollutants
        'WSP', 'WDR', 'SD1',
        'TEMP', 'HUMID',


    ]
    ObsRequest['Sites'] = [39, 1141, 919, 2560, 107]  # updated site IDs
    StartDate = dt.date(2017, 6, 1)
    EndDate = dt.date(2017, 7, 1)
    ObsRequest['StartDate'] = StartDate.strftime('%Y-%m-%d')
    ObsRequest['EndDate'] = EndDate.strftime('%Y-%m-%d')
    ObsRequest['Categories'] = ['Averages']
    ObsRequest['SubCategories'] = ['Hourly']
    ObsRequest['Frequency'] = ['Hourly average']
    return ObsRequest

######################################################################
if __name__ == '__main__':
    AQMS = aqms_api_class()
    ObsRequest = ObsRequest_init()
    AllHistoricalObs = AQMS.get_Obs(ObsRequest)

    # Save Historical Observations to a text file
    with open('2017.txt', 'w', encoding='utf-8') as f:
        f.write(AllHistoricalObs.text)
```

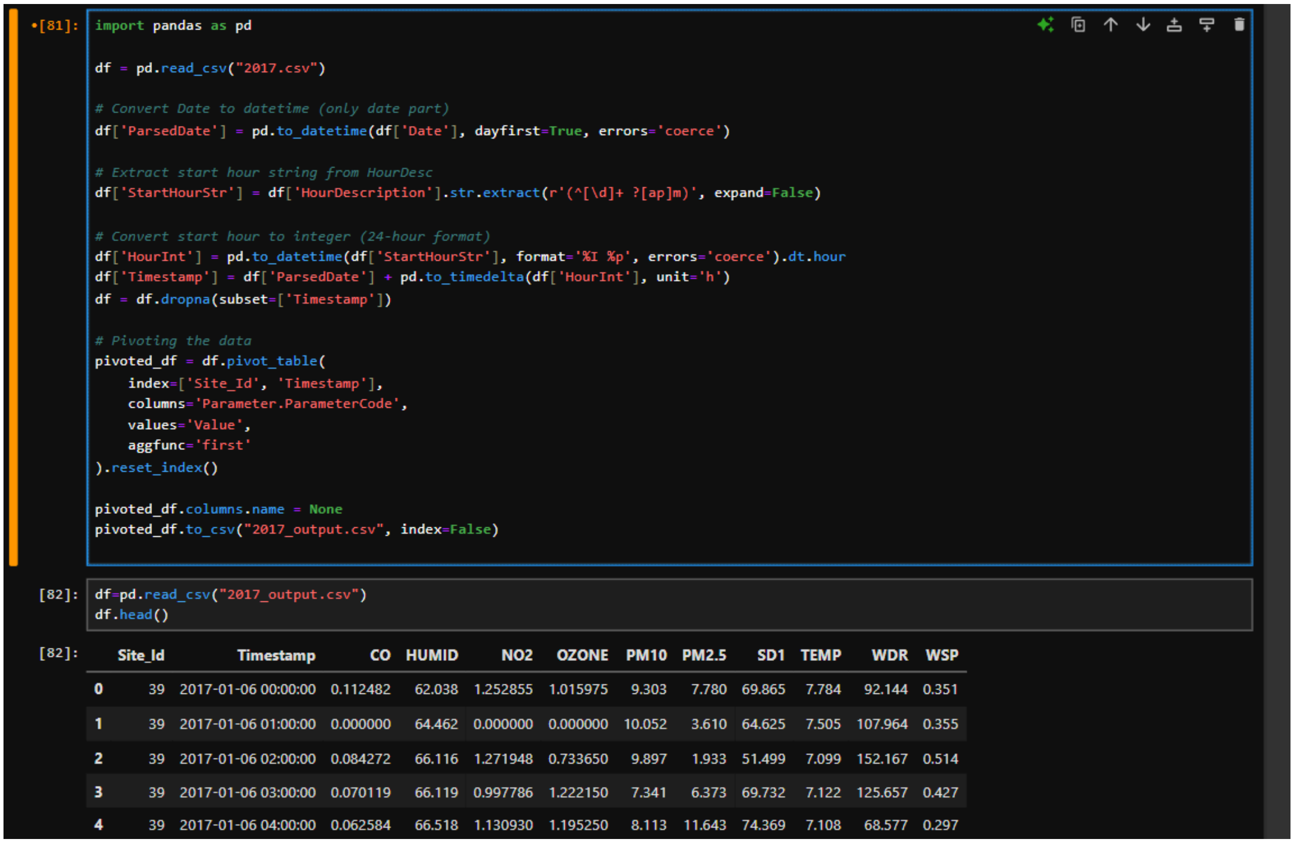
8.5. Code Screenshot 2



```
76]: import json
     import pandas as pd

     with open('2017.txt', 'r', encoding='utf-8') as f:
         data = json.load(f)  # JSON string

     # JSON to pandas DataFrame
     df = pd.json_normalize(data)
     #Fill NA
     df.fillna(method='ffill', inplace=True)

     # Saving as CSV
     df.to_csv('2017.csv', index=False)

     print("CSV file saved as '2017.csv'")
```

```
C:\Users\utsav\AppData\Local\Temp\ipykernel_25064\3090560119.py:12: FutureWarning: Downcasting object dtype arrays on .fillna, .ffill, .bfil
l is deprecated and will change in a future version. Call result.infer_objects(copy=False) instead. To opt-in to the future behavior, set `p
d.set_option('future.no_silent_downcasting', True)`
  df.fillna(0, inplace=True)
CSV file saved as '2017.csv'
```

8.6. Code Screenshot 3



```
•[81]: import pandas as pd

       df = pd.read_csv("2017.csv")

       # Convert Date to datetime (only date part)
       df['ParsedDate'] = pd.to_datetime(df['Date'], dayfirst=True, errors='coerce')

       # Extract start hour string from HourDesc
       df['StartHourStr'] = df['HourDescription'].str.extract(r'(^[\d]+ ?[ap]m)', expand=False)

       # Convert start hour to integer (24-hour format)
       df['HourInt'] = pd.to_datetime(df['StartHourStr'], format='%I %p', errors='coerce').dt.hour
       df['Timestamp'] = df['ParsedDate'] + pd.to_timedelta(df['HourInt'], unit='h')
       df = df.dropna(subset=['Timestamp'])

       # Pivoting the data
       pivoted_df = df.pivot_table(
           index=['Site_Id', 'Timestamp'],
           columns='Parameter.ParameterCode',
           values='Value',
           aggfunc='first'
       ).reset_index()

       pivoted_df.columns.name = None
       pivoted_df.to_csv("2017_output.csv", index=False)
```

```
[82]: df=pd.read_csv("2017_output.csv")
      df.head()
```

[82]:

	Site_Id	Timestamp	CO	HUMID	NO2	OZONE	PM10	PM2.5	SD1	TEMP	WDR	WSP
0	39	2017-01-06 00:00:00	0.112482	62.038	1.252855	1.015975	9.303	7.780	69.865	7.784	92.144	0.351
1	39	2017-01-06 01:00:00	0.000000	64.462	0.000000	0.000000	10.052	3.610	64.625	7.505	107.964	0.355
2	39	2017-01-06 02:00:00	0.084272	66.116	1.271948	0.733650	9.897	1.933	51.499	7.099	152.167	0.514
3	39	2017-01-06 03:00:00	0.070119	66.119	0.997786	1.222150	7.341	6.373	69.732	7.122	125.657	0.427
4	39	2017-01-06 04:00:00	0.062584	66.518	1.130930	1.195250	8.113	11.643	74.369	7.108	68.577	0.297