Air Quality Data Retrieval and Processing Report

Shiwen Xu(520569045)

April 19, 2024

1 Dataset Description

Data Source

The air quality monitoring in NSW [2] are openly accessible and provides API to stream data. The data, protected under the CC BY 4.0 license, allows user to copy, remix, transform, and redistribute for any purpose with indication of changed made and appropriate credit given. All 3 datasets (OBSERVATION, PARAMETER, SITE) are in form of texts and retrieved from semi-structure (JSON) data format. SITE and PARAMETER datasets remain stable, whereas the OBSERVATION data are updated every hour. The Department of Planning, Industry, and Environment (DPIE) of the NSW Government claims that recent data may be undergone only preliminary quality checks and could be subject to adjustments during the validation process to address technical issues, whereas source (OBSERVATION) data for older time periods are reliable and consistent. There is no personal or sensitive information involved so no extra cleaning step is not needed.

Retrieve and Preprocess Data

The New South Wales air quality and meteorology data is retrieved from the Azure Cloud Data Warehouse, under the Department of Climate Change, Energy, the Environment and Water, using an Application Programming Interface (API) data service. This API service can be accessed and streamed either as current real-time hourly data or downloaded as historical data. The retrieving process can be considered as a form of on-demand ingestion as the system allows users to retrieve specific data from the service only when needed. The SITE and PARAMETER data are retrieved all at once using the HTTP GET requests, whereas historical OBSERVATION data are retrieved by querying specific data by adding a request body to the HTTP POST request.
We retrieved the ozone concentration level of Sydney adjacent regions from past 5 years as our historical observation data. The depletion of the thin (stratospheric) ozone layer over Australia is widely recognized for its role in allowing more Ultraviolet (UV) radiation to reach the Earth's surface [3], which consequently increases the risk of skin cancer. Therefore, high ozone concentration in stratospheric level can be beneficial to human. On the contrary, high ground ozone level can irritate human eyes, affect lung functions, and worsen asthma [5]. Nonetheless, less public attention is paid to the detrimental effect of ground-level ozone. Therefore, among all parameters, we aim to examine the average monthly ground-level ozone data of Sydney adjacent regions from past 5 years. All data are loaded in JSON format, converted to pandas dataframe, and loaded in Comma-Separated Values(CSV) files for visualization. On top of that, tables are created and data are stored in PostgreSQL database for automatic generation of Entity Relationship Diagram(ERD).

2 Data Exploration Description

- **Missing and Duplicates** The SITE Dataset contains 17 missing values for both longitude and latitude column. Compared to other regions where only one site location is missing, region Upper Hunter-Muswellbrook has 2 sites location missing. The PARAMETER dataset has no value missing. The BBSERVATION data have 409 missing values. Values in column AirQualityCategory and DeterminingPollutant are all missing. Hence, we simply drop these two columns. All missing values are filled with `np.NaN` in pandas dataframe and filled with -999 to store in database. No duplicates are founded.

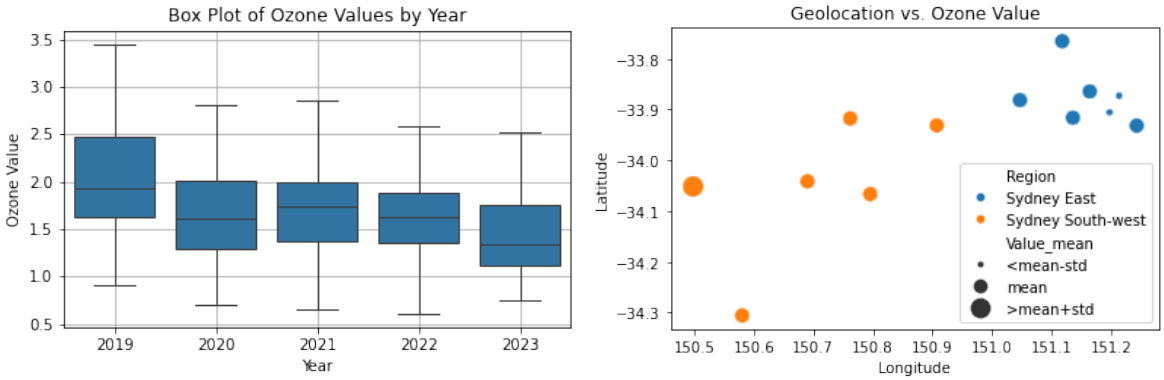
- **Data Format** Data types are checked using `.dtypes` function. In the SITE dataset, numeric data types are correct, and only categorical data types are converted. In the PARAMETER dataset, data types are converted to category and string type correspondingly. In the OBSERVATION dataset, date attribute are converted to datetime64 type using `.to_datetime` function. The rest are converted to category and string type correspondingly.

- **Data Values** Categorical values for each column are checked whether needs to be standardized by using the `.unique` function to get all distinct values. All values are from defined categories for all three datasets. Numeric values are examined through `.describe` function to get the overview of the value distributions. The minimum and maximum longitude is 139.7370 and 153.2935 degree respectively. The minimum and maximum latitude is -36.9078 and -28.8321 degree respectively. The average and maximum ozone concentration for Sydney adjacent regions from past 5 years is 1.6883 and 3.4465 parts per hundred billion (pphm) respectively. Both are comparatively lower than the national air quality standards [1], which set maximum ozone as 0.08 ppm (8 pphm).

- **Transformation** No further transformation is applied on the SITE and PARAMETER dataset. Since the parameter attribute in OBSERVATION dataset is loaded as dictionary type, this column is unfolded into ParameterCode and ParameterDescription columns using the `.json_normalize` function. ParameterDescription column is dropped to reduce redundancy.

3 Data Visualization Description

Three types of visualization, including box plot, scatter plot, and line chart, to explore the ground-level ozone distribution and changing patterns with respect to time and space variations. Specifically, the retrieved OBSERVATION data contains monthly ozone level for the past five years (from January 1, 2019, to December 31, 2023) from Sydney East, Sydney South-west regions.
The box plot, as shown in Figure 1 (a), shows the distribution of ozone parts per hundred million(pphm) value along with its central tendency and variability for each year. Year 2019 has a long upper whiskers, indicating that there is a significant portion of observed ozone value higher the median. The minimum, median, and maximum values of each box plot indicate a trend of decreasing ozone levels over the past five years. In the future, number of exceedances data can be incorporated to better assist the identification of trends in ozone pollution levels over times.
The scatter plot, as shown in Figure 1 (b), explores whether the average ground-ozone value is associated with geo-locations of each site. The ozone values for each site are averaged over past five years and categorized into 3 groups, which are less than $mean - std$, between $mean - std <$



(a) (b)

Figur 1: Box plot figure (a) on the left displays distribution of ozone concentration for each past year. Scatter plot figure (b) on the right displays the relationship between the geo-location and the averaged ozone value for each site.

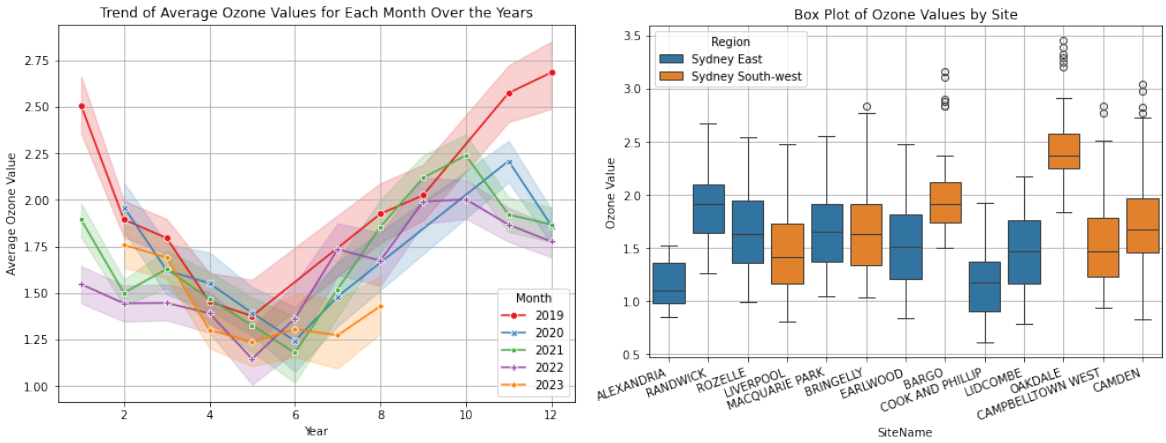
$x < mean + std$, and larger than $mean + std$. The average ozone value for 10 out of 13 sites lies in second interval. Two sites in Sydney East region has least average ground-ozone values, whereas 1 site in Sydney South west has largest average. More sites can be included in the future to explore the correlation between ground-level ozone and specific geographic locations.

The line chart, as shown in Figure 2 (a), displays trend of average ozone value over each month, and different color represents different years. Winter season (from May to Aug) has the lowest average ozone value, whereas the Summer season(from Nov to Feb) has the highest average. This line chart is consistent with reports on trends in ozone for weather conditions, which indicates that ozone is more readily formed on warm, sunny days when the air is stagnant [4]. Given this initial observation, it would be beneficial to include temperature data in our analysis to explore the correlation between ozone levels and temperature more comprehensively.

The box plot, as shown in Figure 2 (b), displays the distribution ground-ozone concentration for each site. The average ozone value for each site used in Figure 1 (b) is sensitive to extreme values and does not convey information about the variability or dispersion of the data. Therefore, this box plot aims to offers detailed illustrations of the distribution. Consistent with Figure 1 (b), the distribution for orange and blue box plots show no clear association between region geo-location and ozone values. Site Oakdale in Sydney South-west has the highest ozone concentration on average with many unusual high ozone values, whereas site Alexandria has the lowest ozone concentration on average with no unusual extreme ozone values occurred in past 5 years. To investigate what causes large different ozone concentrations of these two sites, other atmosphere conditions, such as temperature, carbon/sulphur/nitrogen monoxide, can be included to assist in-depth analysis.

4 Database Description

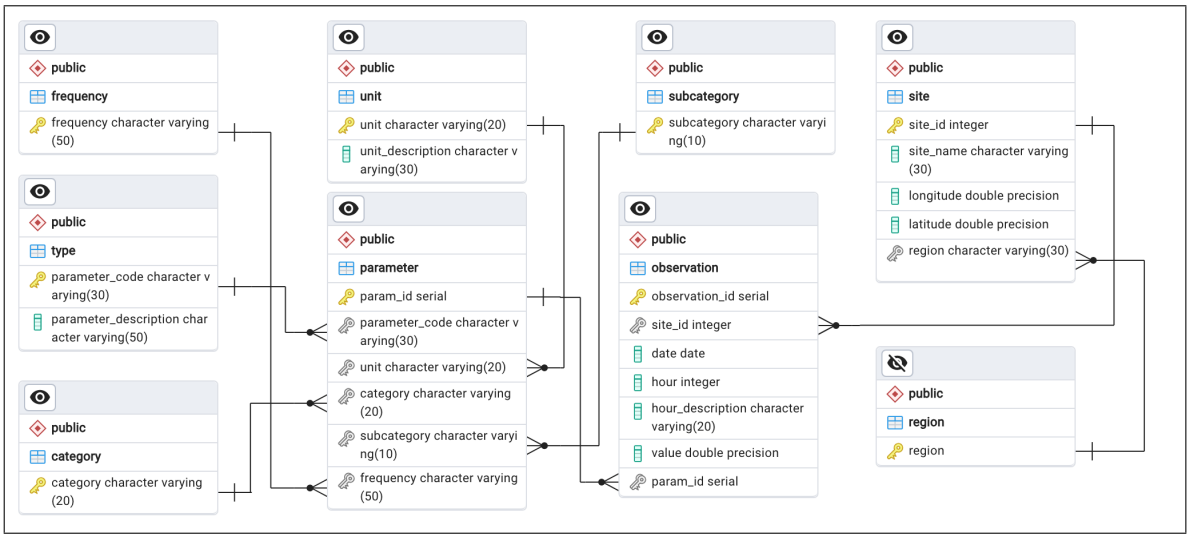
The proposed schema, as shown in Figure 3, includes 9 tables altogether. Two tables are created out of the site dataset. Since each site only belongs to a single region, a region table is created with only 1 column to save the region information. The site table contains every attribute in the original dataset, and the site_id is the primary key with region as foreign key referencing the region table. Frequency, category, subcategory, unit, parameter are separated into 5 tables, and the parameter table references them correspondingly. A surrogate primary key, param_id, is added to the parameter table since there



(a) (b)

Figur 2: Line chart figure (a) on the left: month on x-axis, monthly averaged ozone value on y-axis. Box plots figure (b) on the right: site names on x-axis, ozone value distribution on y-axis

is not a clear combination of attributes that always uniquely identify each row. For example, only ParameterCode is enough to uniquely identify the row when ParameterCode is 'humid', whereas in other cases, ParameterCode, Units, SubCategory, and Frequency need to be combined to identify a unique record. For simplicity, observation table adopts the similar way by creating a surrogate primary key even though compound primary key (site_id, date, and param_id) can identify a unique record. All the lines in the schema diagram denote 1-to-Many relationship. The proposed schema reduces redundancy and improves data integrity by atomizing each entity into its smallest logical components.



Figur 3: Database Schema

References

[1] *National Environment Protection (Ambient Air Quality) Measure Variation*. Accessed on April 19, 2024. 2015. URL: https://www.legislation.gov.au/F2007B01142/2016-02-03/text.

[2] Matthew Riley m. fl. "Air quality monitoring in NSW: From long term trend monitoring to integrated urban services". I: *Air Quality and Climate Change* 54.1 (2020), s. 44–51.

[3] *The ozone layer*. Accessed on April 16, 2024. URL: https://www.dcceew.gov.au/environment/protection/ozone/ozone-science/ozone-layer.

[4] *Trends in Ozone Adjusted for Weather Conditions*. Accessed on April 16, 2024. URL: https://www.epa.gov/air-trends/trends-ozone-adjusted-weather-conditions.

[5] *Types of air pollution*. Access on April 18, 2024. URL: https://www.airquality.nsw.gov.au/types-of-air-pollution.