COMP7507 Final Report

Designing an Interactive Dashboard for Citizen Science and Expert Analysis: Exploring Trends and Relationships Between Air Pollutants and Weather Variables in Hong Kong

Group 15

SRIVASTAVA Dhruv (3035667792)

NG Tsz Chiu (3036198176)

CHEUNG Ngai Yan Irene (3036409367)

LEUNG Ho Kong (3036409862)

ZHANG Longyi (3036411956)

Table of Contents

1. Introduction

1.1. Motivation

Air pollution remains a pressing environmental and public health concern in urban regions such as Hong Kong (HK). Despite growing awareness of its adverse effects, public engagement with air quality data and protective behaviors remains limited. According to a survey by the World Green Organisation (WGO), over 75% of respondents in HK recognize air pollution as a significant issue yet approximately 80% of them reported taking no personal measures to safeguard their health against its impacts [1]. This disconnect between awareness and action highlights a critical need for tools that not only inform but also empower individuals to engage with environmental data in meaningful ways.

The development of an interactive dashboard serves as a response to this gap. By making air quality and weather data accessible and interpretable to both general citizens and expert users, the dashboard aims to foster greater public participation, promote environmental literacy, and support data-driven decision-making. Through intuitive design and interactive features, the tool seeks to bridge the divide between information availability and user engagement, ultimately contributing to more informed and proactive responses to air pollution in Hong Kong.

1.2. Background

Weather parameters may affect concentrations of different air pollutants in the atmosphere [2]. Analysing the general situation of the pollutant concentrations and weather conditions at different locations in HK and drawing conclusions based on the data can be challenging given the availability of large volumes of data from various sources. For HK, weather and pollution data visualisations such as the City Dashboard and Air Quality Health Index (AQHI) offers real-time updates, high-level pollution summary and forecasts for weather conditions for the general public. However, an integrated application to display the complex relationships between individual pollutants and different weather patterns, which can be of use to domain experts in addition to the public, remains lacking.

1.3. Objectives

The primary objectives of this project are twofold. First, to develop an interactive dashboard that effectively serves both general users and Subject Matter Experts (SMEs), enabling intuitive exploration and analysis of air quality and weather data. The dashboard is designed to accommodate varying levels of data literacy, offering simple visualizations for public users and advanced analytical tools for experts.

Second, the project aims to foster public interest in environmental issues and promote citizen science, which encourages public engagement in environmental and health-related issues by providing accessible tools for data exploration and awareness [3]. By promoting user participation through constructivist interaction with the data, the dashboard aims to empower individuals to actively engage with air quality information and assume a more involved role in environmental monitoring and awareness.

1.4. Highlights

Overview Tools for Pollutants and Weather Variables
Users have access to dedicated tools that provide a clear and interactive summary of individual air pollutants (e.g., NO_2, O_3) and weather features (temperature, humidity, and wind speed).

Interactive Visualizations for Correlation Exploration
The dashboards offer specialized visualizations that enable users to investigate and identify potential correlations between different pollutants, or between pollutants and weather variables. These tools help users visually detect patterns, associations, and possible causal links among the measured entities.

Forecast Dashboard for All Features
A forecasting component is included, allowing users to view and analyze predictions for all monitored pollutants and weather parameters. This dashboard employs historical data to generate forward-looking insights, aiding in trend anticipation and informed decision-making.

2. User-Centered Dashboard Design

2.1. Target Users

In developing the Tableau dashboard, we considered the distinct needs of two primary user groups: general citizens and expert users. General users are assumed to have limited technical expertise and are primarily interested in accessible insights, such as air pollutant levels in their district or seasonal weather patterns. To support this, the dashboard includes intuitive temporal and district-level filters, enabling users to easily query variables of interest without being overwhelmed by complexity.

SMEs, including environmental researchers, data analysts, and policy professionals, require more advanced functionality. Their use cases often involve in-depth analysis, hypothesis testing, and decision-making support. Accordingly, the dashboard is designed to accommodate granular data exploration, multi-variable comparisons, and the investigation of relationships between air pollutants and meteorological variables. Features such as correlation plots, trend analysis tools, and potential modeling components are considered to enhance analytical depth and support expert-level inquiry.

By tailoring interaction designs to these two user profiles, the dashboard aims to balance simplicity and sophistication, ensuring accessibility for the general public while maintaining analytical rigor for expert users.

2.2. Selection of Visualization Tools

We have chosen Tableau as the tool for visualization, as it enables dynamic filtering and drill-downs, making it easy to explore relationships between pollutants and weather conditions across time and district areas. Tableau may provide more polished interactions than Google Sheet/Excel via map visualizations, and it is more user-friendly than Python or web tools, requiring no coding. It is particularly useful for the 2 groups of users we have in mind, due to the intuitive interactions that can be implemented.

3. Datasets

3.1. Data Sources

The data used in this study were obtained from the Hong Kong Observatory (HKO) and the Environmental Protection Department (EPD), both of which are recognized and authoritative government sources. From a data science perspective, sourcing data from such reputable institutions ensures a high level of data integrity, reliability, and transparency. This foundational quality is essential for conducting robust analyses, drawing valid conclusions, and supporting evidence-based decision-making.

This approach ensures that the insights and conclusions derived from the dashboard are both accurate and meaningful. By relying on high-quality, official data sources, the analytical outputs maintain a high level of credibility, enabling users to make informed decisions and interpretations with confidence.

Our final dataset was compiled from different data sources listed below. These datasets contained geospatial and timeseries data.

3.1.1. Daily Temperature – HKO

Link: https://data.gov.hk/en-data/dataset/hk-hko-rss-daily-temperature-info-hko
Description: This dataset provides the daily temperature (in degree celsius) measurements from multiple observation stations across HK. It includes comprehensive daily temperature metrics (minimum daily temperature, maximum daily temperature and mean daily temperature) and a data completeness indicator.

3.1.2. Daily Total Rainfall -- HKO

Link: https://data.gov.hk/en-data/dataset/hk-hko-rss-daily-total-rainfall
Description: This dataset provides the daily total rainfall in millimeters (mm) recorded at various weather monitoring stations across HK. Each data point includes the year, month, day, rainfall amount, and a data completeness indicator.

3.1.3. Daily Mean Wind Speed -- HKO

Link: https://data.gov.hk/en-data/dataset/hk-hko-rss-daily-mean-wind-speed
Description: This dataset records the daily average wind speed at different locations in kilometers per hour (km/hr). Each data point includes the year, month, day, wind speed, and a data completeness indicator.

3.1.4. Air Quality Monitoring Data – EPD

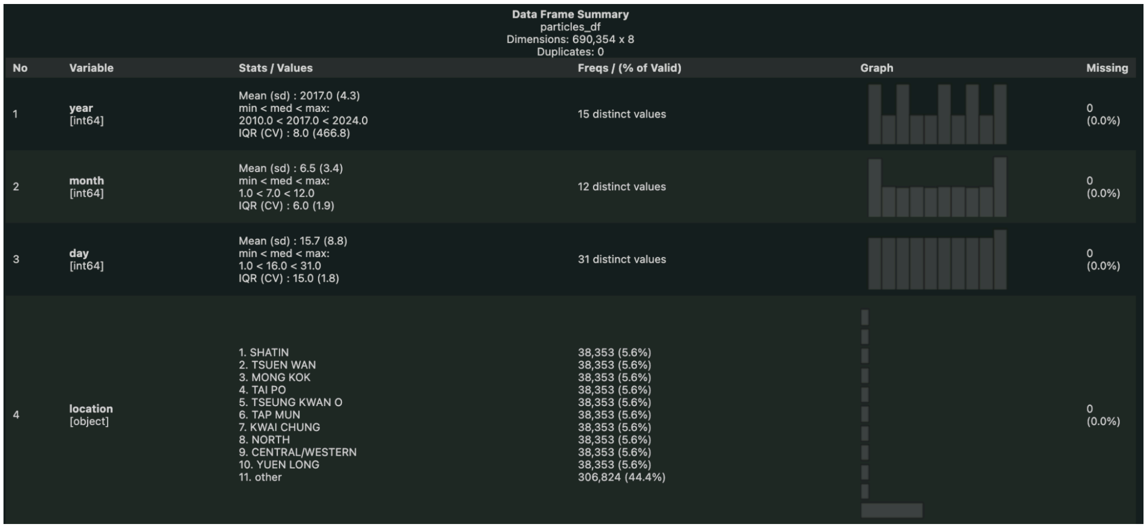
Link: https://cd.epic.epd.gov.hk/EPICDI/air/station/?lang=en
Description: This dataset provides air pollutant concentration data from HK's official Air Quality Monitoring Network. The daily average concentration values for each pollutant are collected in micrograms per meter cube ($\mu g/m^3$), except Carbon Monoxide (CO), which is collected in $10\mu g/m^3$.

3.2. Preprocessing

In this project, we used Python as the primary programming language for data pre-processing, utilizing the Jupyter Notebook environment for interactive data analysis and visualization. Key libraries included matplotlib for graphical representation, and dfSummary for comprehensive data summarization. The preprocessing phase was essential to ensure data integrity, consistency, and readiness for subsequent modeling and analysis.

3.2.1. EDA

To gain an initial understanding of the dataset, we conducted an Exploratory Data Analysis (EDA) using the dfSummary tool. This library provides a detailed statistical overview of each variable, including measures such as mean, minimum, maximum, interquartile range, standard deviation and percentage of missing variables (a screenshot from Jupyter notebook is provided in figure 1 showing the statistical overview). This allows us to gain visibility on the overall data distribution and data completeness. Additionally, it generates histograms that visually depict the distribution of each variable, facilitating the identification of skewness, outliers, and potential data quality issues.



Data Frame Summary
particles_df
Dimensions: 690,354 x 8
Duplicates: 0

No	Variable	Stats / Values	Freqs / (% of Valid)	Graph	Missing
1	**year** [int64]	Mean (sd) : 2017.0 (4.3) min < med < max: 2010.0 < 2017.0 < 2024.0 IQR (CV) : 8.0 (466.8)	15 distinct values		0 (0.0%)
2	**month** [int64]	Mean (sd) : 6.5 (3.4) min < med < max: 1.0 < 7.0 < 12.0 IQR (CV) : 6.0 (1.9)	12 distinct values		0 (0.0%)
3	**day** [int64]	Mean (sd) : 15.7 (8.8) min < med < max: 1.0 < 16.0 < 31.0 IQR (CV) : 15.0 (1.8)	31 distinct values		0 (0.0%)
4	**location** [object]	1. SHATIN 2. TSUEN WAN 3. MONG KOK 4. TAI PO 5. TSEUNG KWAN O 6. TAP MUN 7. KWAI CHUNG 8. NORTH 9. CENTRAL/WESTERN 10. YUEN LONG 11. other	38,353 (5.6%) 38,353 (5.6%) 38,353 (5.6%) 38,353 (5.6%) 38,353 (5.6%) 38,353 (5.6%) 38,353 (5.6%) 38,353 (5.6%) 38,353 (5.6%) 38,353 (5.6%) 306,824 (44.4%)		0 (0.0%)

Figure 1. dfSummary Tool in Use during EDA.

The use of Jupyter Notebook allowed for dynamic interaction with the data, enabling iterative exploration and immediate visualization. This step was crucial for uncovering underlying patterns, assessing variable relationships, and guiding subsequent data cleaning and transformation procedures.

3.2.2. Data Cleaning

The data cleaning process involved several key steps to enhance the quality and comparability of the dataset:

3.2.2.1. Normalization of Pollutant Units

To ensure fair and meaningful comparisons across different pollutant measurements, we standardized all pollutant concentration values to consistent units. This normalization process mitigates discrepancies arising from heterogeneous measurement scales and facilitates accurate cross-variable and cross-location analyses.

3.2.2.2. District Labelling

To enrich the dataset with geospatial context, we incorporated latitude and longitude coordinates corresponding to the 18 administrative districts of HK. These coordinates were sourced from Google Maps and appended to the existing CSV files using python. This enhancement enables spatial analysis and visualization, such as mapping pollution levels across districts.

3.3. Data Combination Strategy

We collected data from the sources listed in section 3.1 with the target period of January 2010 to December 2024 to allow for comprehensive analyses of the data. The data was delivered in different CSV files in different formats (for columns). We built an internal format for the different variables and cleaned and combined the data using Python (in Jupyter notebooks) into a single Excel file to input into Tableau. The data was further formatted after discovering input requirements in Tableau when building specific visualizations.

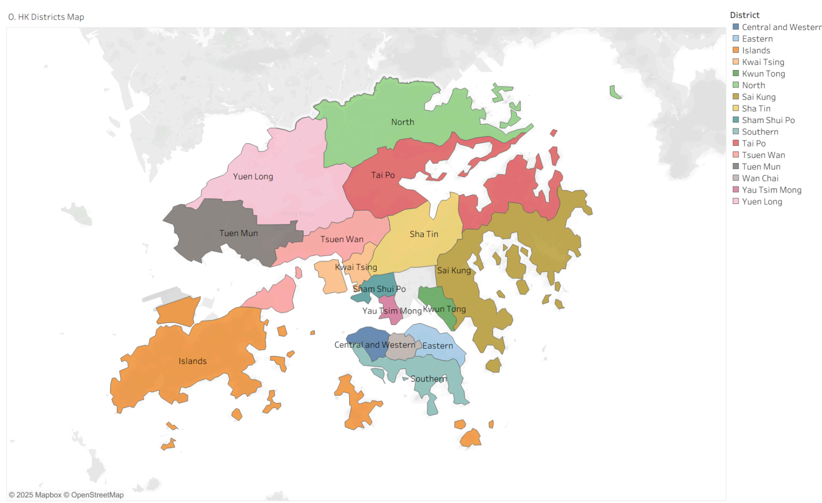


Figure 2. Map of HK districts, available on the Overview page in Tableau to help non-expert users find their district.

The data collection stations for the different weather and pollutant variables were found to be at different locations. In order to match the data for analyses, we decided to aggregate the data by district. This also had the advantage of allowing us to create choropleths in Tableau for the different variables which may help in communicating their spatial distribution. Some districts had more than one data collection station and the variable values for such districts are computed by averaging the values of all the stations. This is not a perfect approach as district classification can be arbitrary and may include different natural and man-made features affecting the variable values that may hide relevant signals or bring non-relevant signals forward. One example is the Islands District, as shown in Figure 2, containing different islands which may have their own independent signals, due to their isolation from each other, but are aggregated under the same district.

4. Visualisations

As mentioned previously, we used Tableau to visualize the data we collected as described in section 3. We chose Tableau as it has good support for different data visualizations and analytics. We also designed a colour palette for the variables to maintain consistency across the dashboards as defined in Figure 3 to reduce the potential cognitive load on the user when mapping the different colour values to the same variable in different dashboards. We chose low luminance colours with sufficient contrast for showing the pollutant variables and high luminance colours for the weather variables.

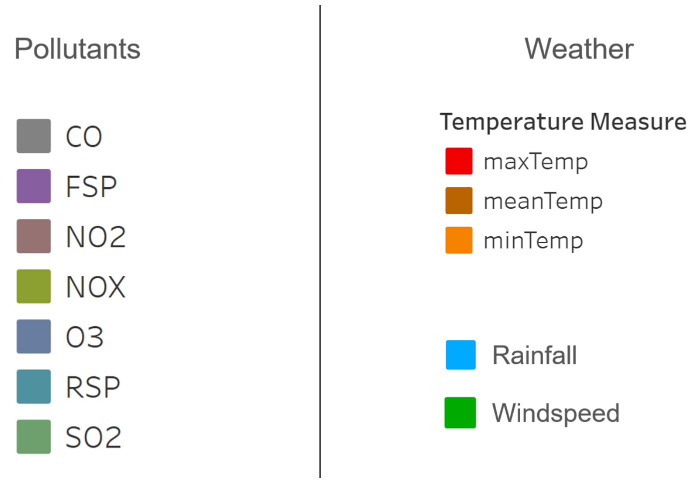


Figure 3. Colour palette for the variables. maxTemp, meanTemp, and minTemp refer to the daily maximum, mean and minimum temperature measurements respectively. The pollutant acronyms are consistent with visualizations on Tableau.

We built 9 dashboards in Tableau to support our users. An **Overview** dashboard to provide instructions and other information such as the acronyms used. Dashboards **D1 (Pollutants Overview)**, **D2 (Temperature Overview)**, **D3 (Rainfall Overview)** and **D4 (Wind Speed Overview)** allow for exploration of the pollutant and weather variables individually. These are effective in helping concerned citizens who may be interested in learning simple facts about these variables. Dashboards **D5 (Pollutant Vs Weather Correlation)**, **D6 (Pollutant Vs Pollutant Correlation)**, and **D7 (Weather Vs Weather Correlation)** contain visualizations that help in analysing the correlation of the variables amongst themselves. These are catered towards SMEs who would want to perform deeper analysis of the data and observe the relationships of the variables. Together **D1-7** may help an SME to perform analysis according to their preferred style, using a *bottom-up* approach, exploring individual variables first then their relationships, or a *top-down* approach, exploring the relationships first, then the individual variable. **D8 (Forecast)** provides forecasts for the variable values. This can help the users understand the potential future trends in the variable values for future planning. We use the built-in features of Tableau for the predictions of the values. Sections 5.1, 5.2 and 5.3 discuss the visualizations available in detail. Section 5.1 focuses on dashboards **D1-4**, which provide tools to explore the variables individually, section 5.2 discusses the visualizations available in dashboards **D5-D7**, section 5.3 discusses the visualizations made available with forecasting in **D8**.

4.1. Dashboards for Individual Variable Exploration

We provide dashboards **D1-4** to explore the variables individually. They all have similar visualizations available: choropleths, line graphs and box plots. **D1.V1-7** choropleths show the spatial distribution of the average daily pollutant concentration, which may help in identifying districts with high or low pollutant concentrations. Similar to **D1.V1-7**, we built **D2.V1**, **D3.V1** and **D4.V1** to show the spatial distribution of the daily mean temperature, rainfall and wind speed respectively. **D1.V8** is a line graph showing the historic daily concentration values averaged over a month. We included this visualization to allow the user to explore any patterns in the historic data and identify any long-term trends. We chose to aggregate the data by month here as we found daily or weekly aggregation was not suitable as it was cluttering the space, making pattern discovery difficult. Aggregation by month offered the right balance in data resolution and clutter on the screen. It also allows the user to select what pollutants should show up in the graph. Similarly, **D2.V2**, **D3.V2**, and **D4.V2** line graphs also show the historic trend of the temperature measurements, rainfall and wind speed respectively. **D1.V9** is a box plot that shows the distribution of the daily concentration values grouped by month of the chosen pollutant. This helps the user to understand the distribution of the concentration and its variability over a month by observing the median and interquartile range, the outliers and seasonal differences in pollutant concentrations. Similarly, **D2.V3**, **D3.V3**, and **D4.V3** box plots show the distribution of the weather features. The dashboards also allow filtering the data by choosing the time period of analysis and selecting the district on the map to filter the data for that specific district on the entire dashboard. These interactions help the user in utilizing the same visualizations for a specific time period and district to perform the analysis. For the visualizations in **D1-4** we deliberately chose to display the different visualization types in isolation (one at a time on the screen). This fragmented approach provides more screen space to allow the user to explore minute patterns in the data, which could be missed if the visualizations were given a smaller space on the screen. The drawback of the approach is that the user needs to *scroll* on the dashboard and perform the filtering actions away from the visualization, especially for the line and the box plots, negatively impacting the user experience.

4.2. Dashboards for Variable Relationship Exploration

We built dashboards **D5-7** to allow the users to explore the relationships amongst all the variables. **D5** has visualizations to help the users to explore the relationships between pollutant variables and weather variables. Visualizations in **D6** help in exploring the relationships amongst the pollutant variables, while those in **D7** help in exploring the relationships amongst the weather variables. We make use of scatter matrices annotated with trend lines to allow the users to explore the correlations between the variables under different data filter settings. **D5.V1**, **D6.V1**, and **D7.V1** are the scatter matrix plots available on the different dashboards. **D5.V1** and **D6.V1** have filters on the Y-axis to allow selection of specific pollutants in the plots to make good use of the screen space by allowing the user to select the relevant pollutants for comparison. All scatter matrix plots have controls to filter data based on time period and district to allow analyses at a fine level. Parallel Coordinate Plots are effective in exploring multi-variate data, but Tableau does not have native support for them. We have provided an incomplete implementation of a PC Plot for pollutants in **D6.V2** and **D6.V3** to still allow for some exploration. **D6.V2** shows a static PC Plot while **D6.V3** allows selection of two pollutants to be put side-by-side for comparison. The order of the pollutants on the Y-axis is always consistent in both visualizations, which allows **D6.V2** to serve as context for **D6.V3** when picking specific pollutants as shown in Figure 4. **D6.V1** and the PC plot pair **D6.V2** and **D6.V3** have isolated controls to allow fine configuration for analysis.

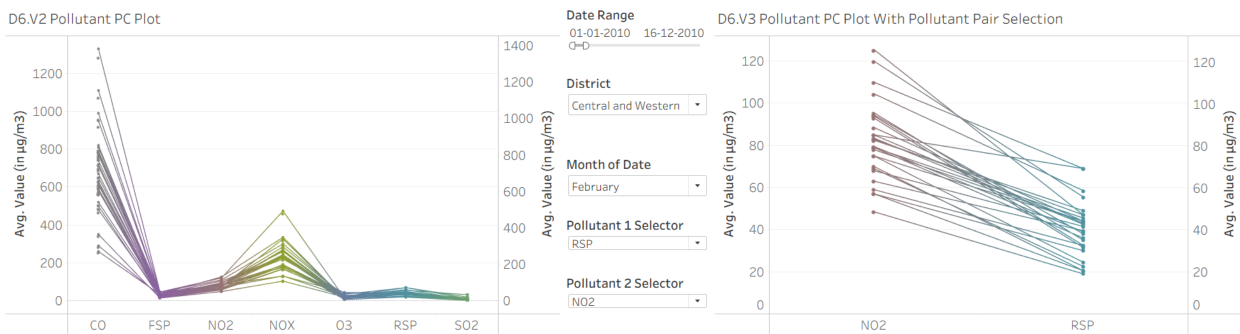


Figure 4. Visualizations D6.V2 and D6.V3 have consistent ordering of pollutant appearance on the Y-axis. NO2 is always on the left hand side of RSP in both plots. This property allows D6.V2 to serve as context for D6.V3.

4.3. Forecast

The visualizations in **D8** use the forecasting feature of Tableau to help the users to understand potential future trends in the variable values to make future plans. We show the historic trends in the variable values in the line graphs and then show the estimated values. The filtering controls allow selecting specific districts to forecast district-specific values. Visualizations **D8.V1** and **D8.V2** provide selection of exact variables to be plotted to make good use of the screen space and focus the analyses.

5. User Analyses

Our dashboard is designed to serve two types of users: concerned citizens and SMEs. This section demonstrates how each user group can benefit from different features in the dashboard through real-world use scenarios.

5.1. Concerned Citizens

5.1.1. Planning a short trip

A resident with asthma is planning a short summer trip and wants to minimize exposure to air pollution. Using the Pollutants Overview (D1), they examine historical pollutant data — particularly ozone (O_3) and FSP (Fine Suspended Particulates) — across different districts and months.

The data reveals that districts in the eastern part of HK, such as Tai Po and Sai Kung, tend to have higher ozone concentrations, particularly during the autumn months. In contrast, western districts generally report lower average O_3 (shown in figure 5) levels. For FSP, the pattern is opposite: western districts show higher concentrations compared to eastern ones. In terms of seasonality, ozone levels are lower in June and July, while October shows the highest concentrations. Conversely, FSP peaks in January and is lowest in June, aligning with winter pollution buildup and better summer dispersion.

D1.V1.5 Average Daily O3 Concentration

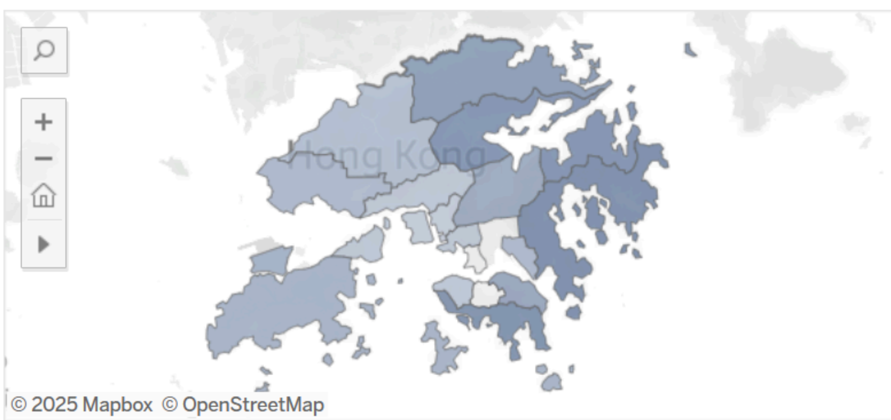


Figure 5. Illustrates the spatial distribution of average daily from 2010-2024 ozone concentrations across HK, highlighting higher levels in eastern districts like Tai Po and Sai Kung and lower levels in the western regions.

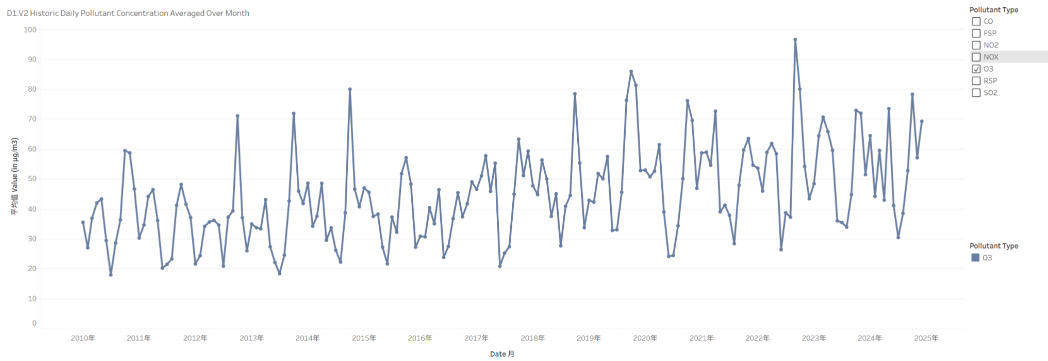


Figure 6. Presents the historic monthly average of daily ozone concentrations from 2010 to 2025, highlighting a clear seasonal trend with recurring peaks in October and dips in June and July.

Based on this insight, the resident chooses to avoid high-O_3 eastern districts like Sai Kung in October, and instead plans a trip to a western district in July, where both ozone and FSP levels have been lower in the past and are expected to be low as predicted via forecast — making the trip safer and more comfortable for their respiratory health.

5.1.2. Avoid Carbon monoxide(CO) Exposure

A resident is concerned about exposure to Carbon Monoxide(CO), especially due to its impact on heart and brain function during long periods of low-level exposure.

Using the Pollutants Overview (D1), they explore historical CO concentration data across districts and by month. They discover that Tuen Mun, Yuen Long, and Wan Chai tend to have higher average CO levels, likely due to dense traffic and street canyon effects.

Seasonally, the data shows that CO concentrations peak in January, likely due to winter inversion layers and increased vehicle use, while May and July show consistently lower CO levels across most districts.

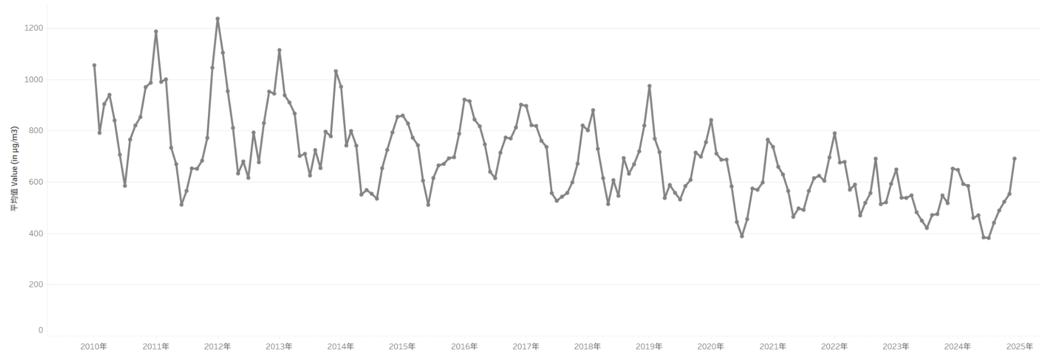


Figure 7. Presents the historic monthly average of daily CO concentrations from 2010 to 2025, highlighting a clear seasonal trend with recurring peaks in January and dips in July.

With this insight, the resident chooses to:

- Avoid commuting through high-CO areas like Wan Chai when possible
- Take indoor MTR transfers instead of roadside bus terminals in January

5.2. Subject Matter Experts

SMEs such as environmental scientists, public health researchers, and policy analysts often require deeper insights into pollutant behavior, multi-variable relationships, and long-term trends to inform policy or health recommendations. Our dashboard supports such expert use through detailed correlation tools, pollutant-pollutant interactions, and forecasting features.

5.2.1. Identifying High-Risk Zones for Traffic-Linked Pollution

A transportation policy advisor is tasked with evaluating how vehicular traffic may contribute to localized pollution hotspots.

By examining Carbon Monoxide (CO) and Nitrogen Dioxide (NO_2) using Pollutants Overview (D1), they identify that Wan Chai, Tuen Mun, and Yuen Long have consistently higher CO levels. Using D6 (Pollutant Correlation), the advisor observes a strong positive correlation between CO and NO_2, further confirming vehicular emissions as a shared source. Cross-referencing with monthly trend visualizations, they find peak CO levels in January, suggesting effects from winter inversion layers and increased traffic volume. The advisor uses this analysis to recommend targeted traffic restrictions and roadside emission controls during winter in affected districts.

5.2.2. Forecast-Based Policy Planning

A climate policy expert explores how forecasted trends of pollutants may guide resource allocation.

In the Forecast Dashboard (D8), they focus on PM2.5 (FSP) trends across several high-risk districts. They observe that Sham Shui Po and Kwai Tsing are projected to have persistently high FSP levels over the coming months.

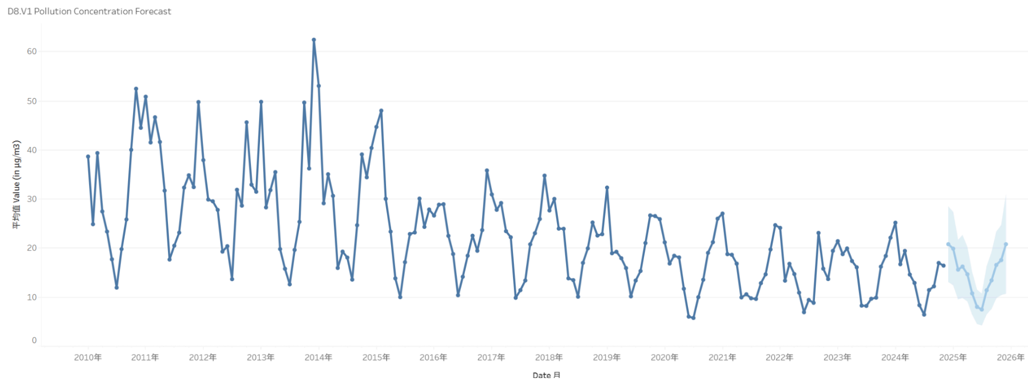


Figure 8. Displays the forecasted monthly average concentrations of PM2.5 (FSP), showing a seasonal trend with expected declines in mid-2025 followed by a gradual rebound — offering insight for preemptive policy actions in high-risk districts.

The expert uses this forecast to recommend proactive measures such as issuing early public health advisories, equipping schools and elderly centers in high-risk districts with temporary indoor air filtration systems, and launching targeted citizen awareness campaigns to reduce outdoor exposure during peak pollution periods.

5.2.3. Distinctive Geospatial Distribution of Ozone Relative to Other Air Pollutants

A public health researcher has undertaken a detailed analysis of the geospatial distribution of ozone in comparison to other air pollutants, utilizing map-based visualizations and statistical evaluations of long-term and seasonal ozone data from multiple monitoring stations (D1.V1.5). This approach aims to identify emerging urban hotspots and shifts in spatial distribution patterns, thereby elucidating the underlying factors influencing these trends and informing the development of targeted mitigation strategies.

Analysis of the ozone concentration map (D1.V1.5), particularly when examining individual districts, reveals that elevated ozone levels in Hong Kong are predominantly found in the eastern and northern districts, such as Sai Kung, Tai Po, and North District. Conversely, geospatial mapping of other air pollutants (D1.V1.1-1.7) highlights higher concentrations in the western districts and areas such as Yau Tsim Mong.

Furthermore, when applying a temporal filter (D1.V1.5) to the ozone concentration data, the analysis demonstrates distinct shifts in spatial patterns over time. Between 2010 and 2015, elevated ozone concentrations were primarily located in the western and northern parts of Hong Kong. Since 2016, however, there has been a pronounced shift, with southern and eastern districts experiencing increased ozone levels.

The insights generated through this expert analysis are essential for enhancing air quality management policies. Specifically, they provide evidence to support the prioritization of districts for intervention and the implementation of geographically targeted measures to effectively reduce ozone exposure for both urban and rural communities in Hong Kong.

6. Discussion

6.1. Limitations

6.1.1. Sequential Visual Layout UX

The current dashboard design requires users to navigate through individual visualizations sequentially via dropdown menus, manually selecting each tab for analysis. This segmented approach hinders simultaneous comparison across multiple data dimensions and limits the potential for integrated, holistic exploration. For both expert and general users, this structure may reduce analytical efficiency and cognitive coherence. Implementing a unified dashboard layout—where multiple visual components are presented in a single, interactive view—would enhance contextual understanding and support a more fluid narrative flow.

6.1.2. Non-Intuitive Filtering Mechanisms

Although the dashboard incorporates regional and temporal filtering capabilities, the current map-based selection method may not be intuitive for all users. Specifically, it requires users to either hover over map regions to reveal district names or rely on their ability to recognize district shapes based on geographic outlines. This design increases cognitive load, particularly for general users unfamiliar with HK's district geography, potentially hindering efficient interaction and data exploration.

To improve usability and reduce pre-attentive processing demands, more accessible filtering techniques—such as dropdown menus with clearly labeled district names or annotated maps displaying district labels—should be considered. These enhancements would support a more inclusive and user-friendly interface, accommodating both expert and non-expert users.

6.2. Future Developments

6.2.1. Streamlined User Flow

To enhance usability, especially for a diverse audience that includes both general and expert users, a more streamlined and purpose-driven layout is recommended. A unified dashboard interface can guide users intuitively toward relevant visualizations based on their analytical goals or domain expertise.

For general users, simplified views with guided narratives and minimal interaction complexity can reduce cognitive load. Expert users, on the other hand, may benefit from access to advanced filtering, multi-dimensional analysis, and customizable views.

A potential improvement in design is to adopt a higher-level filtering system that guides users toward relevant visualizations and controls based on their analytical goals via user mode selection:

- User Mode Selection: Introduce a toggle or parameter control (e.g., “Basic Mode” vs. “Expert Mode”) at the top of the dashboard.
- Basic Mode: Displays simplified filters (e.g., dropdowns for district and year) and high-level summaries.
- Expert Mode: Reveals advanced filters (e.g., correlation explorations, multi-variable selectors) and detailed analytical views.

6.2.2. Statistical Insights and Expert Collaboration

The current forecasting functionality within the dashboard relies on Tableau’s built-in forecasting tool, which provides a convenient but generic approach to time-series prediction. However, this method has not been backtested for accuracy, nor has the underlying model been fine-tuned to reflect domain-specific characteristics of air pollutant and meteorological data. As such, the reliability of the forecasted values remains uncertain.

To enhance the credibility and analytical robustness of the dashboard, future development should include collaboration with environmental scientists and statistical experts. These specialists can assist in validating the forecasting methodology, calibrating model parameters, and identifying appropriate statistical techniques tailored to the nature of the data. Furthermore, integrating more interactive and transparent ways to communicate forecast accuracy—such as confidence intervals, residual plots, or model comparison dashboards—would improve user trust and engagement. This collaborative approach would ensure that the dashboard not only presents data but also supports informed decision-making through statistically sound insights.

6.2.3. Enhancing Engagement Through Web-Based Interactivity

Initially, the dashboard design considered incorporating pollution complaint statistics [4] to enrich the analysis. However, due to data incompleteness identified during exploratory data analysis, this feature was excluded from the current version. To address this limitation and foster greater user engagement, one proposed enhancement is the integration of real-time citizen input. For example, users could submit textual feedback or observations related to air quality, which could then be visualized dynamically using a word cloud embedded within the dashboard.

This participatory feature would not only increase interactivity but also support sentiment analysis, offering qualitative insights into public perception and environmental concerns. Implementing such functionality would require transitioning from Tableau to more flexible, web-based platforms such as D3.js or Streamlit, which support real-time data input, dynamic visualizations, and enhanced customization. This migration would enable a more responsive and collaborative user experience, aligning with the principles of citizen science and open environmental data.

References

[1] World Green Organisation. (2018, August 31). Hong Kong citizens lack knowledge on air pollution. https://thewgo.org/website/eng/news/air-pollution-and-health/

[2] E. Zender-Świercz, B. Galiszewska, M. Telejko, and M. Starzomska, “The effect of temperature and humidity of air on the concentration of particulate matter - PM2.5 and PM10,” Atmospheric Research, vol. 312, p. 107733, Oct. 2024, doi: https://doi.org/10.1016/j.atmosres.2024.107733.

[3] Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. Frontiers in Public Health, 8, 14. https://doi.org/10.3389/fpubh.2020.00014

[4] Common Spatial Data Infrastructure. (n.d.). Air pollutant concentration by monitoring station. https://portal.csdi.gov.hk/geoportal/?datasetId=epd_rcd_1629267205235_67879

Individual Contributions

COMP7507 Individual Contributions

LEUNG Ho Kong (3036409862)

Dataset Proposal and Evaluation:

Proposed and assessed datasets from HKO and EPD based on relevance, temporal granularity, and spatial coverage. Evaluated data completeness, consistency, and potential for integration to support meaningful visualizations and correlation analyses.

User Use Case Evaluation:

Participated in identifying core user personas (general citizens and SMEs) and aligning dashboard features with their goals. Contributed to mapping typical user tasks to appropriate visualizations and interaction flows for effective engagement and insight extraction.

Scenario/Dashboard Testing

Tested multiple dashboards for usability, filter responsiveness, and data accuracy. Simulated realistic use cases (e.g., trip planning, pollution risk assessment) to validate dashboard functionality and provide feedback for layout and interaction improvements.

COMP7507 Individual Contributions

Cheung Ngai Yan Irene (3036409367)

Part 1: Details of Tasks Completed

1. Dataset Research

I contributed by helping to research searching for relevant datasets on the Hong Kong Open Data platform and also other government platforms such as the EPD. This step was crucial for identifying sources that could provide meaningful insights for the project. At this point, the scope of our project was still broad as we considered data relating to citizens' lifestyles, e.g. EV purchase habits can also be considered as a metric.

2. Data Pre-processing, Cleaning, and Exploratory Data Analysis (EDA)

After collecting the datasets, I normalized district names into a consistent format and appended the corresponding district center coordinates to each row. I performed exploratory data analysis using the dfsummary library in Python to better understand the data structure and quality. During this process, I realized that some datasets were incomplete for certain time periods, which could potentially affect the overall data quality.

3. Visualization Design, Prototyping and Testing

I proposed several ideas for visualizations, such as correlation plots and word clouds. To choose appropriate color schemes, I used online tools like ColorKit. Initially, I suggested earthy tones for representing pollutants (to maintain similarity) and use high luminance and contrasting colors for other weather variables. However, after further discussion, our team pointed out that the low luminance colors were too similar and could cause confusion. Eventually, we settled on using low-luminance colors for pollutants and high-luminance colors for weather variables to improve clarity and contrast. I had also contributed by doing some prototyping of the dashboard, i.e. plotting pollutants on a map, and using the in-built forecast function on Tableau to predict pollutant levels. I also helped to write some in-dashboard overview description which would help the users navigate the dashboard, with the rationale that it would help different user groups find the tabs/visualizations that they would need.

Colors

1. Pollutant colors (same color palette ‘earthly tones’-> low luminance colors) -> Unify design by using analogy
2. Rainfall, Temperature, Windspeed (high luminance colors shows contrast from Pollutant colors) -> shows contrast from pollutant colors

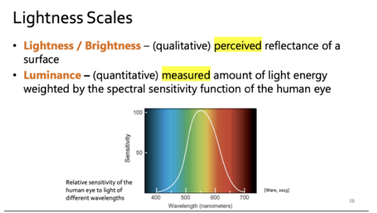


Figure 9.

Overview page

Welcome to the Pollutant & Weather Dashboard!
Explore air quality in your district or Hong Kong in general and see how it connects with weather patterns like temperature, rainfall, and windspeed.

Getting Started:
Start with D1 for an overview of pollutant levels.
Use D2–D4 to get an overview of weather conditions.
Interact with the maps to pick districts and explore other controls available to support your analyses.

Go Deeper:
Check D5–D7 to explore how pollution and weather interact amongst themselves.
Visit D8 for forecasts based on past data.

Pollutants Covered:

CO = Carbon Monoxide
FSP = Fine Suspended Particulates
NO2 = Nitrogen Dioxide
NOX = Nitrogen Oxides
O3 = Ozone
RSP = Respirable Suspended Particulates
SO2 = Sulphur Dioxide

General Instructions:

1. Clicking the district on the maps in D1-D4 filters the data in the entire dashboard to only contain the data for the selected district to allow district-specific analyses.
2. Other controls for filtering the data in the different visualizations are available on the dashboards.
3. The controls of D6.V1 and D6.V2 are independent of each other.

Figure 10.

4. Limitations and Future Developments

We had discussed some limitations of the current UI/UX with the teaching assistant, particularly the filtering options on the maps. Through internal discussions, we recognized that users might find it difficult to navigate and interpret the map shapes. For the report, I suggested future development ideas such as implementing a user mode toggle (basic vs. expert) to reduce cognitive load. I also proposed migrating to web-based solutions to allow more control, enable advanced features, and support real-time feedback and textual data collection.

Part 2. Results or findings of the tasks

1. Dataset Research

Through the EDA, I found that some of the datasets were incomplete or too high-level to generate meaningful insights about the variables we were interested in. Although these datasets were initially proposed in our project plan, we ultimately decided not to use them due to their limitations.

2. Data Pre-processing

During the data cleaning phase, I encountered confusion regarding the units used to measure pollutants. To ensure fair comparisons in the report, I consulted groupmates and we conducted additional research to understand the units and then normalized them across the dataset.

3. Dashboard Prototype and Testing

While experimenting with the forecast function in our dashboard, I realized that we needed more backtesting to verify the accuracy of the data (which was not done in interest of time). Despite this, we decided to include the feature as it could be particularly useful for expert users.

I also discovered that the dashboard wasn't very intuitive as a collaborative tool. Our team often had to go back and forth, and there was a lack of control over versioning. One of my groupmates attempted to use Git for source control, but it didn’t work well with our setup. In the end, we each experimented with features individually, and one groupmate consolidated everything. While

Tableau was effective for visual representation from a user’s perspective, it wasn’t as developer-friendly. It didn’t support an iterative design process well and made it difficult to revise design choices.

COMP7507 Individual Contributions

Ng Tsz Chiu (3036198176)

1. Dataset Research

1.1. Greenhouse Gas (GHG) Emissions

The project began with an investigation of datasets related to greenhouse gas (GHG) emissions. Official open data sources from Hong Kong, such as the Environmental Protection Department (EPD), provide comprehensive GHG emissions inventories. These datasets detail annual total emissions, sector-specific contributions (including electricity generation, transportation, and waste), along with per capita and carbon intensity figures. The latest inventories cover trends from 1990 to recent years, which facilitate time-series analyses and sector-focused assessments. The primary reason for examining this information was its relevance to general air quality and its possible indirect influence on the Air Quality Health Index (AQHI), which is affected by overall air pollution.

1.2. Carbon Intensity

This dataset enables evaluation of decarbonization efficiency relative to the region's economic output, with annual data available from 2005 onwards. Hong Kong has demonstrated a steady decline in carbon intensity since 2005, reflecting reforms in the energy sector and economic shifts. The objective was to assess whether carbon intensity correlates with short-term air quality measures, although this metric generally serves as a broader, long-term indicator.

1.3. Waste-related Emissions

Data on emissions from the waste sector, including methane from landfills and pollutants from incineration, is also available through the EPD. These statistics illustrate trends and the relative contribution of waste management to Hong Kong's total GHG emissions, which the waste management is particularly burning or decomposition that can release particulates and gases that elevate AQHI values. The motivation for exploring this dataset was to investigate possible connections between waste handling and urban air quality levels.

2. Exploratory Data Analysis

2.1. Greenhouse Gas (GHG) Emissions

The exploratory data analysis (EDA) of GHG emission datasets targeted identifying the main contributing sectors (such as electricity, transport, and waste) and tracking temporal trends. Visualizations were created to illustrate annual emissions, sectoral breakdowns, and geographic patterns, showing varied trends by sector and area. Although these datasets revealed progress in terms of Hong Kong's carbon reduction goals, their annual aggregation limited their suitability for direct comparison with the AQHI, which varies daily or hourly.

2.2. Carbon Intensity

The EDA on carbon intensity datasets investigated efficiency trends over time and contrasted regions with differing emission intensities. Through distribution and temporal analyses, a broad understanding of Hong Kong's decarbonization progress emerged. The consistent downward trend in carbon intensity reflects policy impacts and shifts within sectors. Year-over-year comparisons, including anomalies such as those related to the pandemic's economic effects, offered insights relevant to policymaking. Despite their value for climate-related research, these datasets do not directly correlate with immediate air quality health indices.

2.3. Waste-related Emissions

Analysis of waste emissions data highlighted temporal patterns such as trends over time, fluctuations due to waste policies, and seasonal variations in landfill outputs. More detailed data at the facility or process level provided opportunities for in-depth investigation. However, as with previous datasets, establishing direct associations with the AQHI is challenging because these emissions are aggregated annually, whereas health indices require more granular temporal data.

3. Use Case Scenario based on Final Work

Dashboard Usage:

- Enabled visualization of geospatial distribution of ozone and other air pollutants through map-based displays.
- Allowed statistical analysis of long-term and seasonal ozone data from multiple monitoring stations.
- Provided interactive temporal filtering to examine changes in ozone concentration patterns over different years.
- Supported district-level investigation, highlighting specific areas with elevated pollutant levels.
- Facilitated comparison between ozone patterns and other pollutant distributions for comprehensive spatial understanding.

Scenario Highlighted:

- A public health researcher used the dashboard to identify distinct geographic hotspots of ozone versus other pollutants.
- Found that elevated ozone was mainly in eastern and northern districts, and recently surged in southern districts, while other pollutants concentrated in western districts.
- Temporal analysis showed a spatial shift in ozone hotspots from western/northern districts (2010–2015) to southern/eastern districts (post-2016).
- The findings supported targeted air quality management, enabling prioritization of districts and tailored mitigation strategies to reduce ozone exposure for diverse communities in Hong Kong.

COMP7507 Individual Contributions

Srivastava Dhruv (3035667792)

1. Introduction

This project provides visualization tools to explore the historic pollutant and weather data in Hong Kong and the relationships between these variables. The project was executed in different phases: data collection, data pre-processing, visualization design and implementation, and analysis along with intermittent tasks undertaken to meet course requirements with respect to the project deliverables. My primary contributions were in the data pre-processing, visualization design and implementation phases of the project, and in the preparation of deliverables relating to the final Tableau dashboards, demonstration and the report. Section 2 discusses my contributions in detail.

2. Contributions

This section details my contributions in this project in detail.

2.1 Data Processing: Cleaning

After the data sources were identified, the data was collected into a GitHub repository for version control and Jupyter notebooks were created to process the data to get it into the right shape for inputting into Tableau to build the different visualizations. The data for the same variable was supplied in different CSV files. I contributed in the initial phase of data cleaning by understanding the data format for each dataset, designing and defining an internal data format for each dataset by means of creating Pydantic (https://docs.pydantic.dev/latest/) definitions for the data, and then combining and transforming the raw data to the internal data format. The approach of defining an internal data format was chosen to simplify further data processing tasks as now the data attributes and what possible values they could take will be well known. I also completed some data processing tasks later in the project due to the input requirements of visualizations in Tableau for the data.

2.2 Visualization Design & Implementation

I contributed to the discussions on the design of the dashboards and in deciding the types of visualizations and interactions to be used to allow the users to effectively explore the data. This process involved revisiting the course materials and looking for appropriate types of visualizations to use for the temporal and geospatial data. One of the visualization tools that I pushed to include in the dashboards was the Parallel Coordinate Plot (PC Plot), which can be a great tool to analyse the relationships in multivariate data. Tableau does not have native support for PC Plots, so I implemented a visualization that comes close to a PC Plot, included in dashboard **D6** to analyse relationships amongst the pollutants. We often found ourselves working against the limitations in Tableau Public as it did not offer sufficient flexibility in our specific use case, despite its extensive feature set, highlighting the importance of tool selection. Another limitation that we encountered was collaboration in Tableau Public. It was not possible for different users to edit the same Tableau workbook. It did not have version control as well, making prototyping and experimentation with the Tableau settings on different visualizations inefficient, requiring intermediate *saves* before making any changes and closing the program without saving in case we wanted to revert the settings. I tried to experiment with Tableau Public Desktop to see if we can potentially use Git to combine the changes in the workbook file to allow collaboration and version control, but that attempt failed as Tableau seems to maintain a dependency file for every project (which is most likely a database file) in the host machine which was not allowing the workbook to open in other machines. In the end, we decided that one group member would be assigned to manage the Tableau workbook and group members will take screenshots of the settings of any visualizations that they create on Tableau and send them to that member for applying to the workbook. I undertook this responsibility and managed the project workbook while also implementing the visualizations and the dashboards myself.

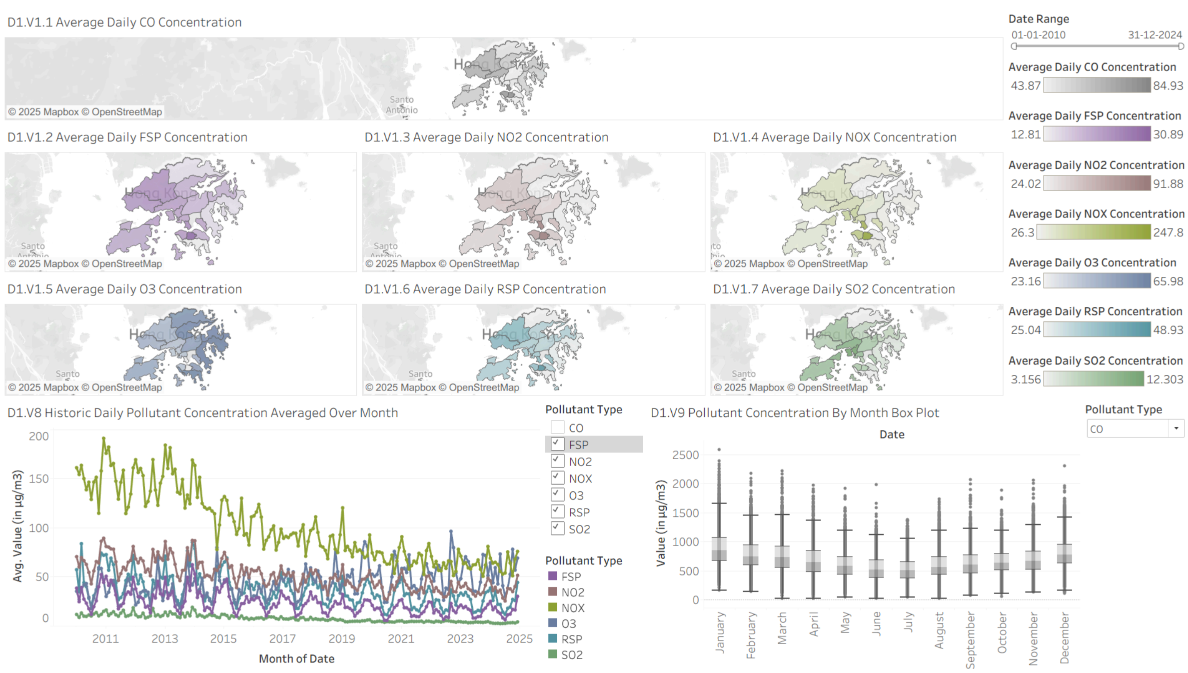


Figure 11. Alternative Rendering of Dashboard D1 by Showing all the Content on a Single 1080p Screen.

The dashboard layout of showing one visualization at a time was a design choice that I advocated for. Due to the scale of the data, the visualizations would have been very cluttered if we tried to arrange all the dashboard visualizations in a single page rendering on a standard 1080p screen, making trend and pattern discovery difficult as seen in figure 11, which shows the alternative choice of displaying all the visualizations in a 1080p screen. The mini-maps (**D1.V1-7**) are extremely small, making understanding the spatial distribution of the pollutants difficult. The line graph (**D1.V8**) is compressed along the horizontal direction, making the trend and pattern discovery difficult. Initially, I wanted there to be some way so that the user can "focus" on a specific visualization, e.g., clicking the visualization on the dashboard makes it larger, so the trends and patterns could be explored well while also allowing the observation of the interactive changes across the dashboard, but I could not find such a feature on Tableau. The current approach dedicates more screen space to each visualization, allowing finer analysis, but has the issue of inability to observe global changes made via interactions, and adding the overhead of scrolling to reach certain settings.

2.3. Deliverables

I made contributions in the preparation of the project deliverables: demonstration and the project report.

COMP7507 Individual Contributions

Zhang Longyi (3036411956)

I further prepared the data structures from our candidate datasets for Tableau visualisations in Pandas. Specifically, it involves creating uniform date objects for time-series visualisations, imputing missing values, and correcting discrepancies in neighbourhood encoding in the original datasets due to the different agencies responsible for collection. This was later used to infer the district of each data point for polished map visualisations.

I created preliminary visualisations for progress reports and future dashboard designs and participated in designing the layout for the particle/weather element dashboards and proposing visualisations on them to fulfill user tasks. Admittedly, I was largely uninvolved with implementation for the Tableau workbook and later stages in general.

I contributed to writing the datasets section of the report as well as editing and proofreading in general.