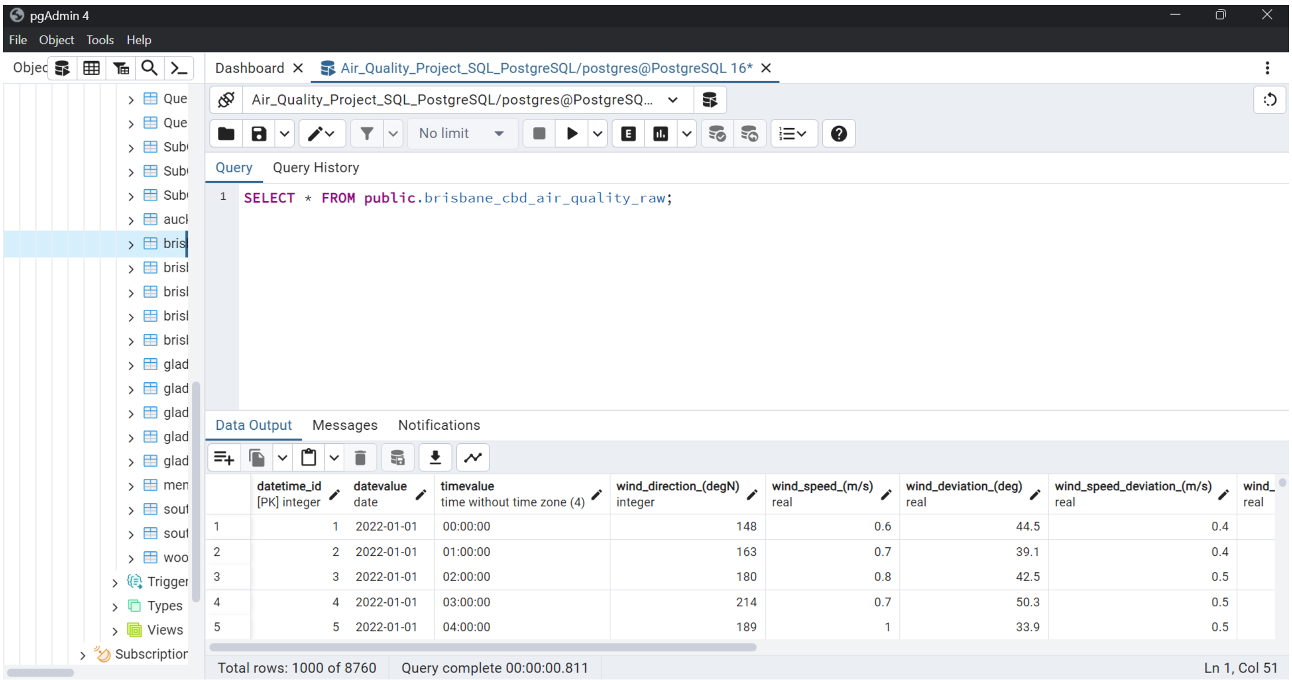
Project 1 – Weather Data Comparison

INTRODUCTION

Comparing Brisbane and Gladstone weather data is essential to discover potential inferences in both location's local pollutants, industry impacts and finding undiscovered variables crucial to fully understanding both regions. Therefore, an analysis was taken on both regions using multiple BOM (GOV.Q, 2023) weather station datasets, including Brisbane's CBD, South and Wooloongabba areas with Gladstone's Auckland Point, Memorial Park and South. Composite datasets were built using reasonable substitute data and back-filling or removing columns if data was insufficiently filled. Afterwards, analytics queries were performed with results visualised on this report and briefly discussed.

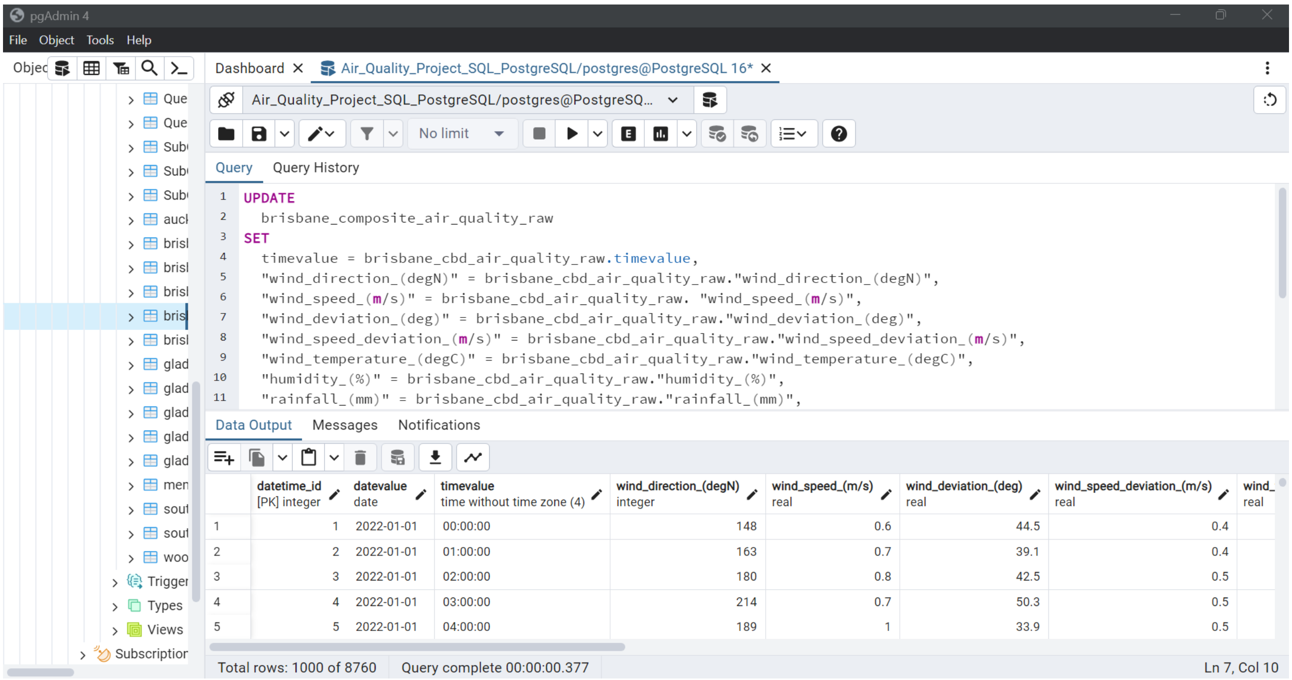
DATA SOURCING/CLEANING

Data was sourced from Australia's 2022 meteorological records (GOV.Q, 2023), with 6 datasets downloaded as structured .csv data, registered on-site as "3-star" quality for data reading. Initial data was hourly records of 2022 weather metrics and sized from 8760x11 to 8760x19 (rows x columns). These datasets required formatting ID columns for easier SQL queries and renaming columns headers to fit a standard camel-case format. Several "Null" values were present in most datasets, so data filling composite datasets was required to generate a usable datasets for analysing multiple weather metrics. Data was then substituted into a PostgreSQL database for warehousing and general management, stored as integer, date, time, and real data for 5 decimal places of accuracy. Below shows the starting dataset results, considering Brisbane CBD data:

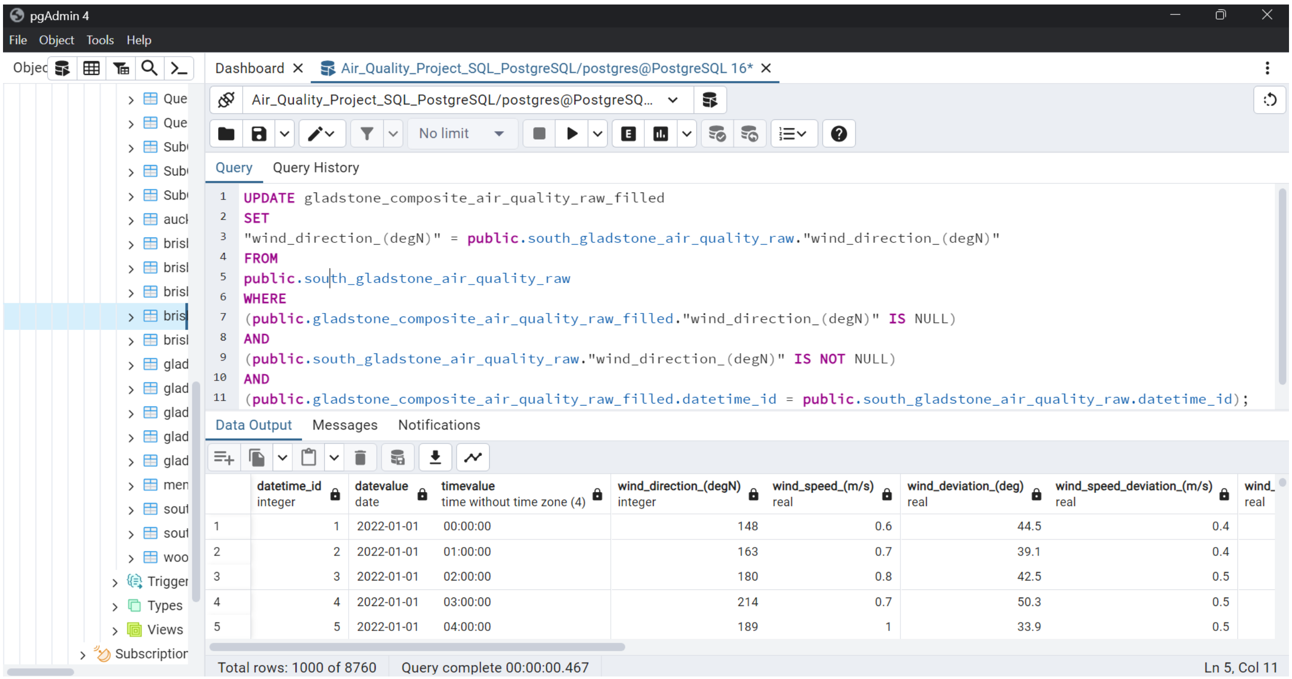


Cleaning data composed of a few core steps, but largely consisted of data cleaning using substitutes of geographically close weather stations to create a single 'composite' dataset, as well as back-filling data for persistently null steps given a tolerance of <500 nulls, or ~5.7% unfilled data. The cause of data incompleteness was unknown, however, in several instances it was either damage to selective measuring apparatus or maintenance work, so the nulls were likely correlated data. From this, data filling was implemented with substitutes and forward-filled data. Implementing substitute data required ranking locations by proximity to the desired location, Brisbane and Gladstone's central districts, so ordered by CBD > South Brisbane > Woolongabba and Memorial > Auckland >

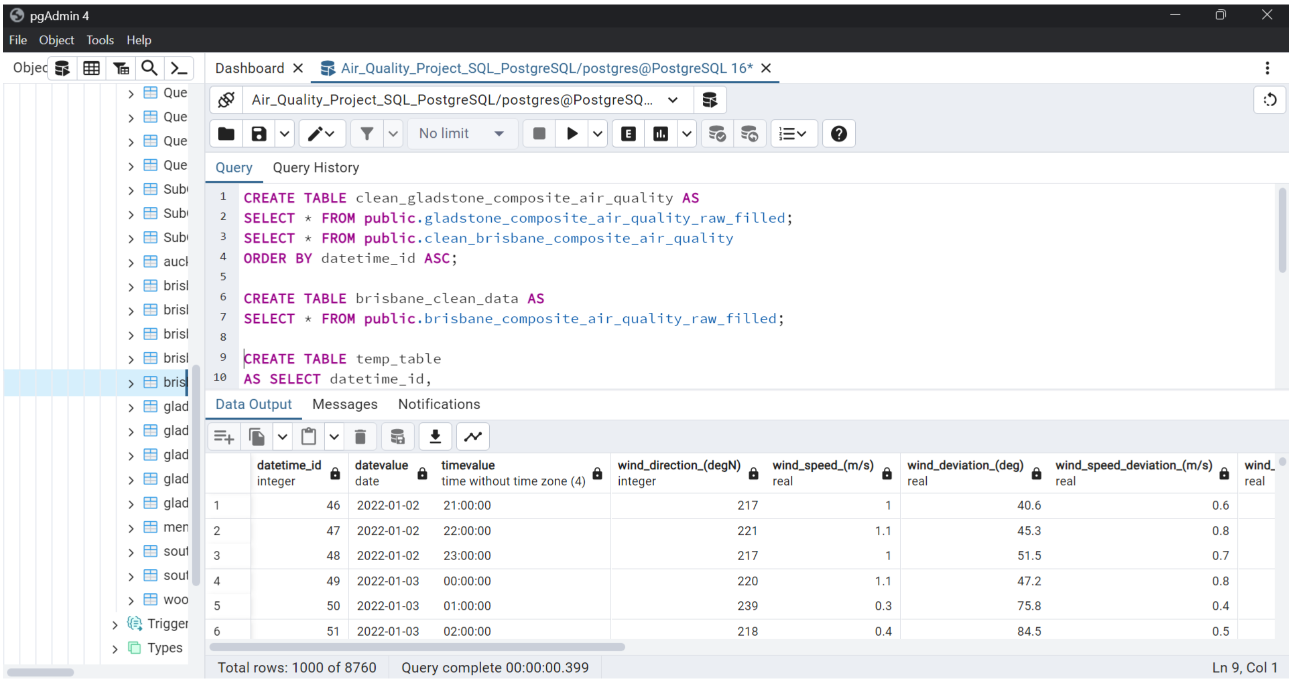
South Gladstone. The implemented substitution strategy involved first generating a composite dataset, and filling it with the priority data (Memorial Park or Brisbane CBD, shown below):



Next, an update function was implemented with sufficient conditions to overwrite NULL values without either replacing original data or adding in extra NULLs. The result is shown below, again following Brisbane's data as an example:



Forward-filling was implemented by generating a dummy table with the forward-filled values, then the composite datasets were updated using the previously-used update function. The results are below:

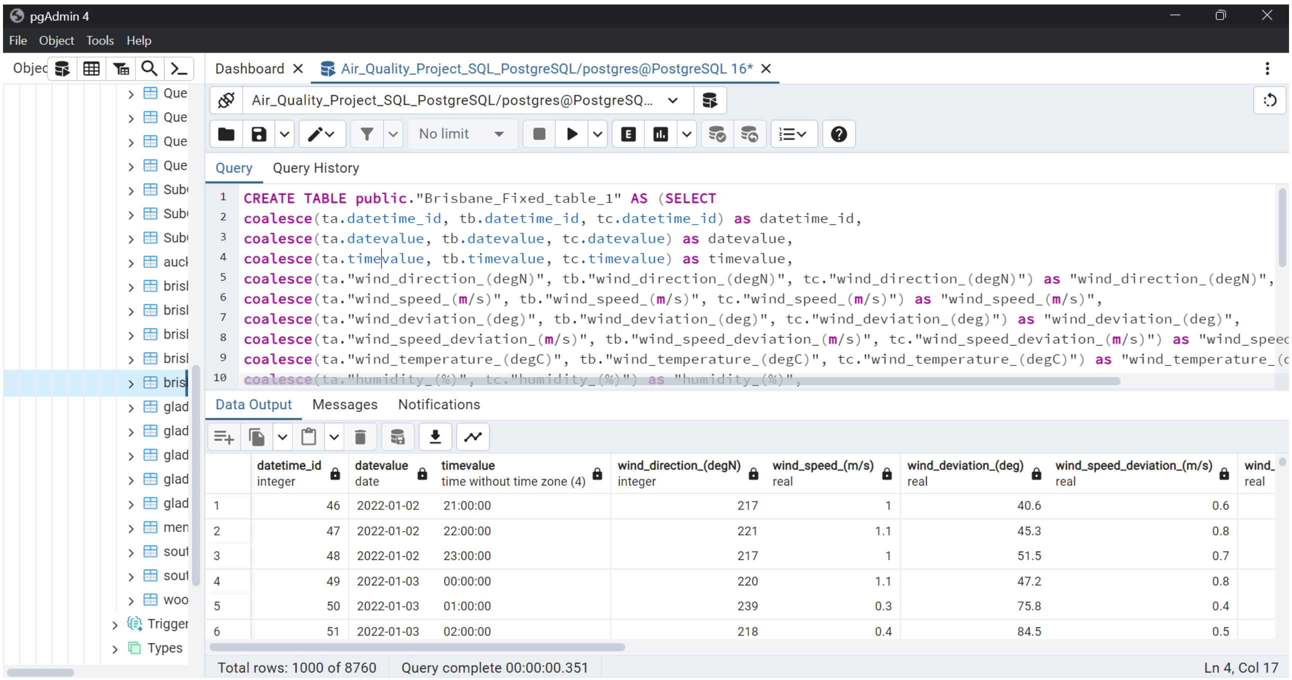


[Note that images were taken post processing, so SQL procedure times are faster.]

Overall, Memorial Park and Wooloongabba data contributed new metrics to the composite datasets, whereas both South region districts largely acted as substitute data for missing data. Forward filling occurred afterwards, filling in missing values, as well as dropping data metrics with large missing values. Therefore, the composite Brisbane dataset was 8760 by 18 and Gladstone 8760 by 21, dropping some Gladstone Chemical variables due to unfilled data.

During data cleaning and querying, additional procedures were needed. These included fixing overwritten data, fixing timestamp timezones and reformatting data, fixed in initial cleaning and analysis. Additionally, datasets were also renamed throughout the project for clarity.

After review, a shortened version of the original substitution process was implemented using postgreSQL's Coalesce function to apply substitution order, combined with a similar forward-filling process to clean the entire dataset in less time. Below shows an extract of the code, with appropriate output, for Brisbane:



To simplify viewing query analysis, a uniform timestamp as "Date-Time" was generated using TO_x and concatenate functions to combine given date-time and hour functions into one unified timestamp function. As the default timezone is not Brisbane, it was preset during implementation, creating a timestamp format of "2022-01-01 01:00:00+10" for query generation. Code showing the timestamp generation is below:

Query Query History

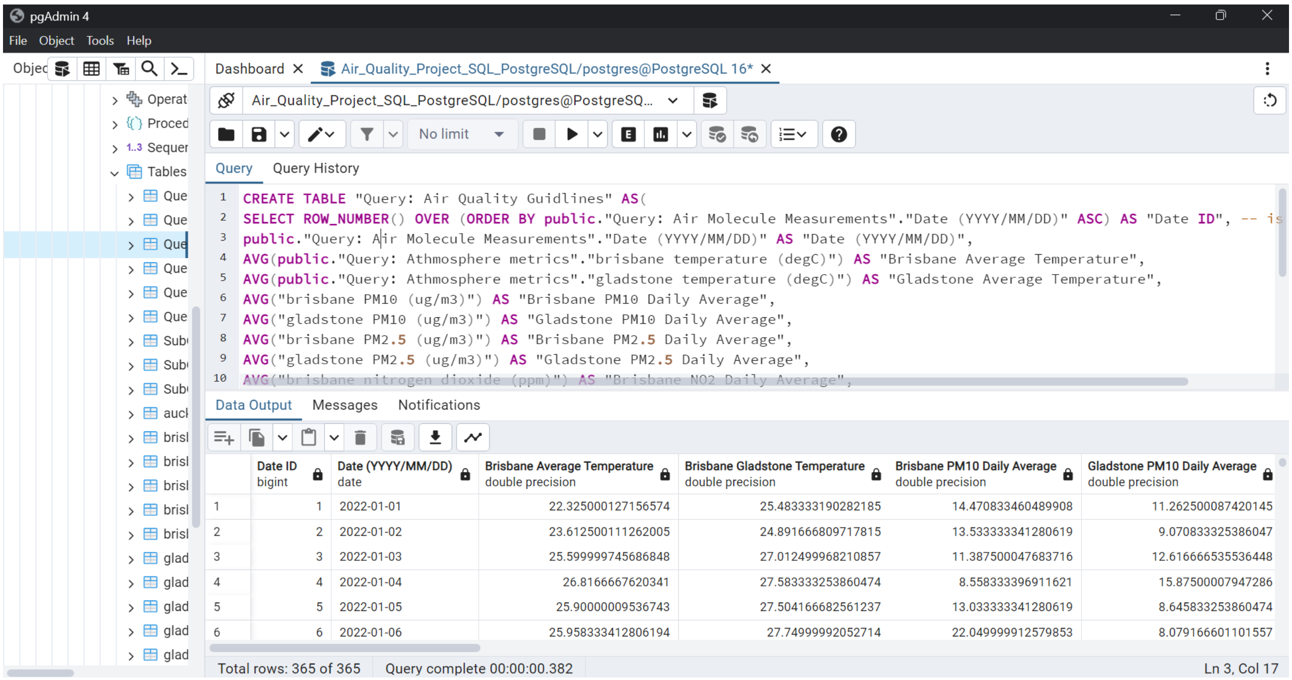
```
TO_TIMESTAMP( to_char("Date (YYYY/MM/DD)"::date, 'YYYY-MM-DD')||' '||
              to_char("Hour (24:00)"::time without time zone, 'HH24:MI'),
              'YYYY-MM-DD HH24:MI TZH')::timestamptz AS "Date-Time"
```

After these processes, analysis was performed using SQL querying and Tableau visualisations to identify important aspects of the collected data and queries.

DATA ANALYSIS

Implementing a data analysis involved several queries using joins and categorical data. This included Air Molecule Measurements, Air Quality Data, Guidelines, Atmosphere Metrics, Wind Speed Deviations and Categorical Weather metrics.

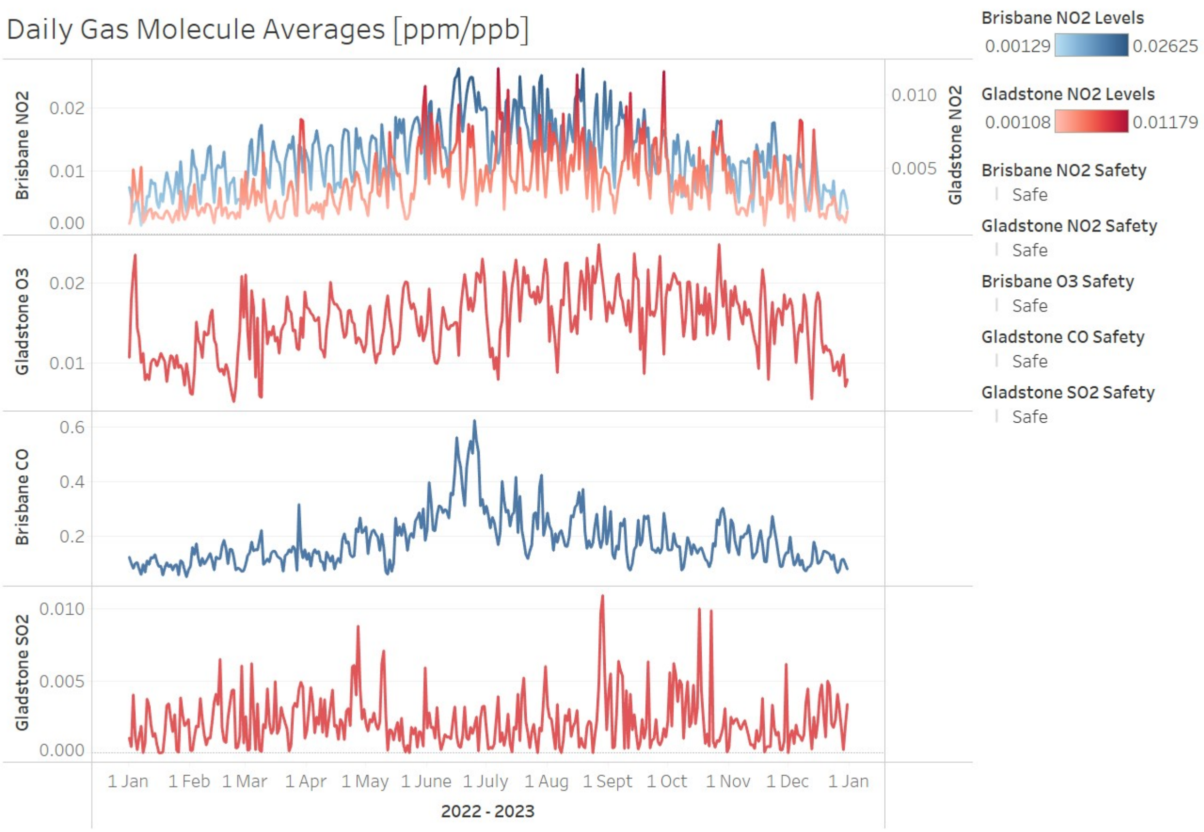
Query implementation required several functions, including the creation of several combined query datasets using postgresSQL. Datasets included a query collecting composite weather metrics (temperatures, air pressures, etc) as well as air smog (PM10, PM2.5, etc) and molecule pollutants (carbon monoxide, sulphur dioxide, etc) shortened to their chemical labels (ie, CO, SO2, etc). These were relabelled for clarity and generated as separate datasets for a clean template to export as a .csv for use in Tableau. Expansions were also made to these initial datasets, expanding with sub-queries and injected categorical data to expand on data analysis. Additions included astrological and meteorological seasons, World Health Organisation air pollutant measurements and calculated columns such as wind speed percentiles and air metrics conversions, explained further in the data analysis section. An example query is shown below, covering the WHO guidelines:



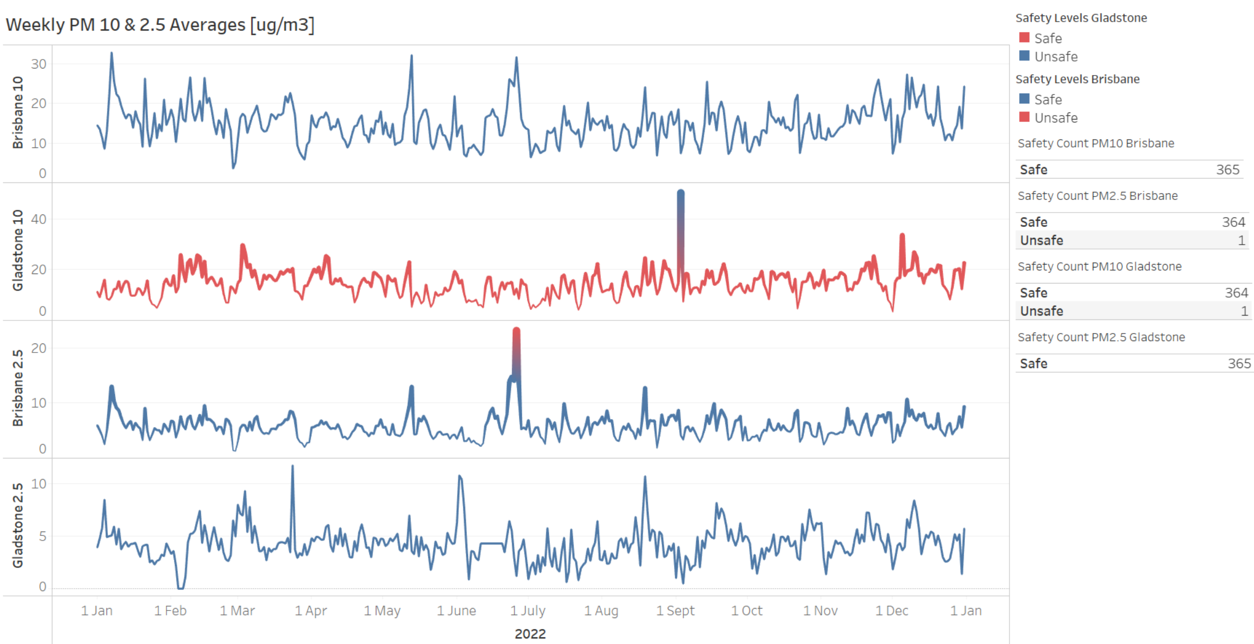
RESULTS AND VISUALISATIONS

Overall, the results of this projects were interesting. Generally, computations were at most 1.186 seconds, so generally fast for SQL to implement. The Brisbane composite weather stations were better filled as nulls were lower, and overlap led to only 3 time periods across all data columns needing forward-fill results compared to Gladstone's, where up to ~16.639% of data was missing from formaldehyde. From this, eliminated rows included: Benzene, Toulene and Formaldehyde from Gladstone data, whereas all Brisbane data was kept. Although some data elimination was implemented, the remaining data was easily sufficient for data analysis. Additionally, the forward-filled results had relatively little effect on final visualisation results for all datasets, bar a few Gladstone metrics. Note that data on Tableau often had visualisation issues at large row numbers, such as the 8760 row count of the completed graphs, so much of the data had to be averaged to weekly or daily data for both clarity and performance.

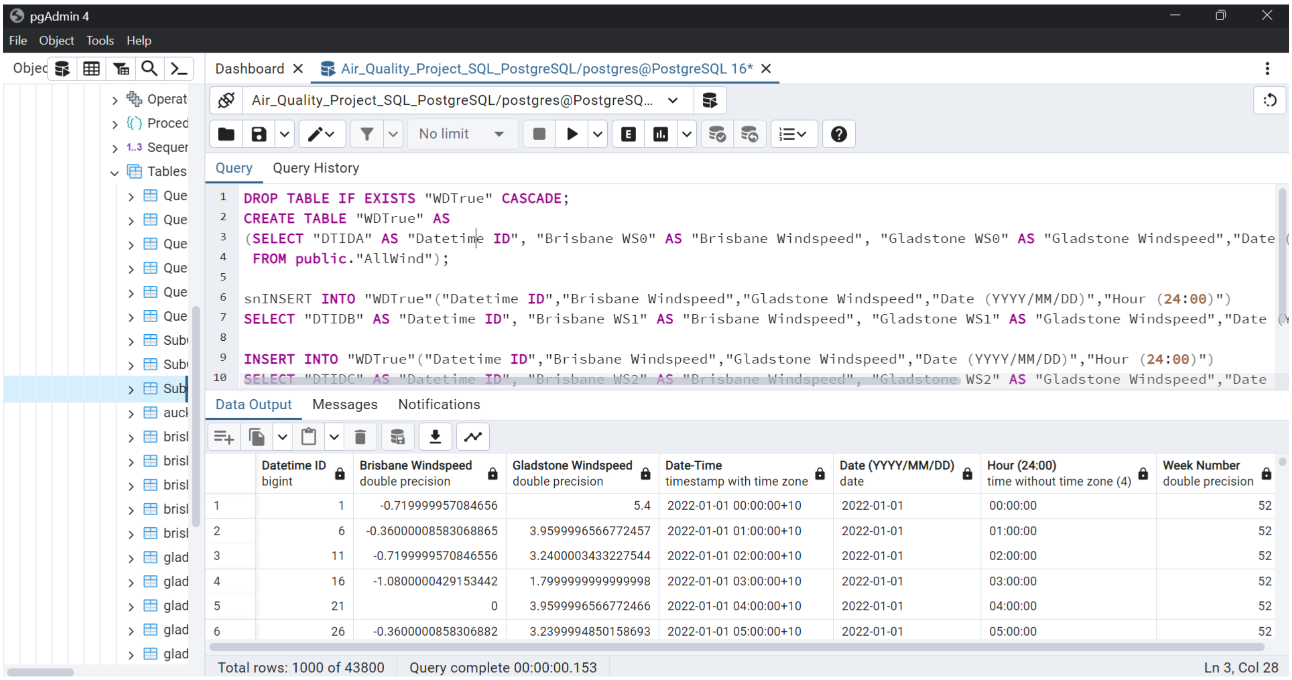
The initial analysis and visualisation involved collecting gas molecule data and comparing it to WHO guidelines for safe levels. These measurements (WHO, 2024) were either used or converted to form using a simplified ppb/ppm to ug/m3 converter, exploiting the ideal gas law to account for temperature ranges. It was found that for specific gas compounds, no unsafe levels were detected at either Brisbane or Gladstone, indicating safe management of gaseous waste from industrial zones. This was visualised as weekly data seen below, utilising Tableau's multi-parameter modelling to show nitrogen dioxide levels as one component, and safety categorically shown by uniform line thickness:



Secondly, smog particulates (PM10, 2.5 and VRPs) showed unsafe levels at Gladstone's PM10 and Brisbane's PM2.5 for the respective days 3rd September and 26th June. The outcome is interesting as there is no correlation with graphical data, indicating an external source for these spikes which could merit further investigation. A visualisation is shown below, indicating a clear Brisbane and Gladstone spike, shown by both a thicker line and inverted colour scheme. Notably, the spike in emissions may exaggerate the outlier, so a side table displays the number of unsafe days for each parameter, shown below:



The raw Wind Speed data was reformatted to work with the implementation of Tableau's visualisations. To visualise percentile ranges for wind speeds, a range of data not provided by the raw data is required, however it includes the standard wind speed deviation. Therefore, an indirect solution includes using the standard deviation to generate range values for each hour at ± 1 or 2 deviations. Implementing this solution required SQL to generate and merge these values, which was done using a sub-query "WDTrue", relabelled "SubQuery: Wind Speed Ranges", which was then used for Tableau visualisations, shown below:



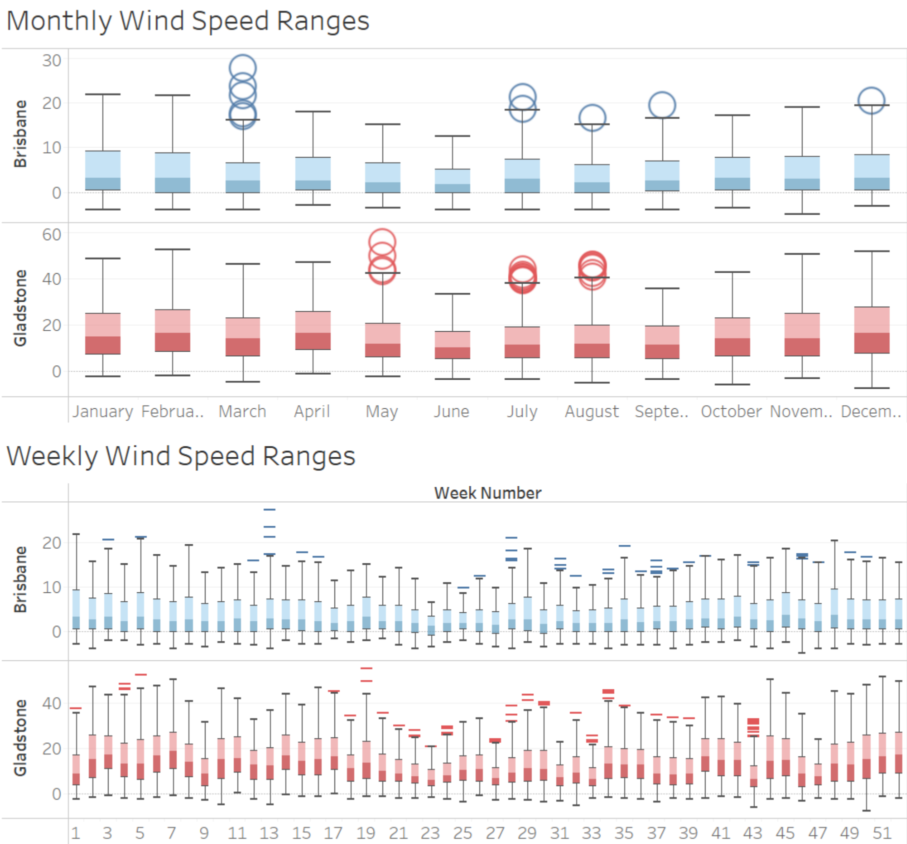
Using this generated data, box-whisker plots were created to show the average wind speed and its ranges given a reported direction (note that negative results mean the wind blew in the opposite direction). Additionally, from generated distributions, some results were outside the expected range of monthly/weekly plots, so were shown in the dashboard below:

Brisbane and Gladstone
Wind Ranges (km/h)

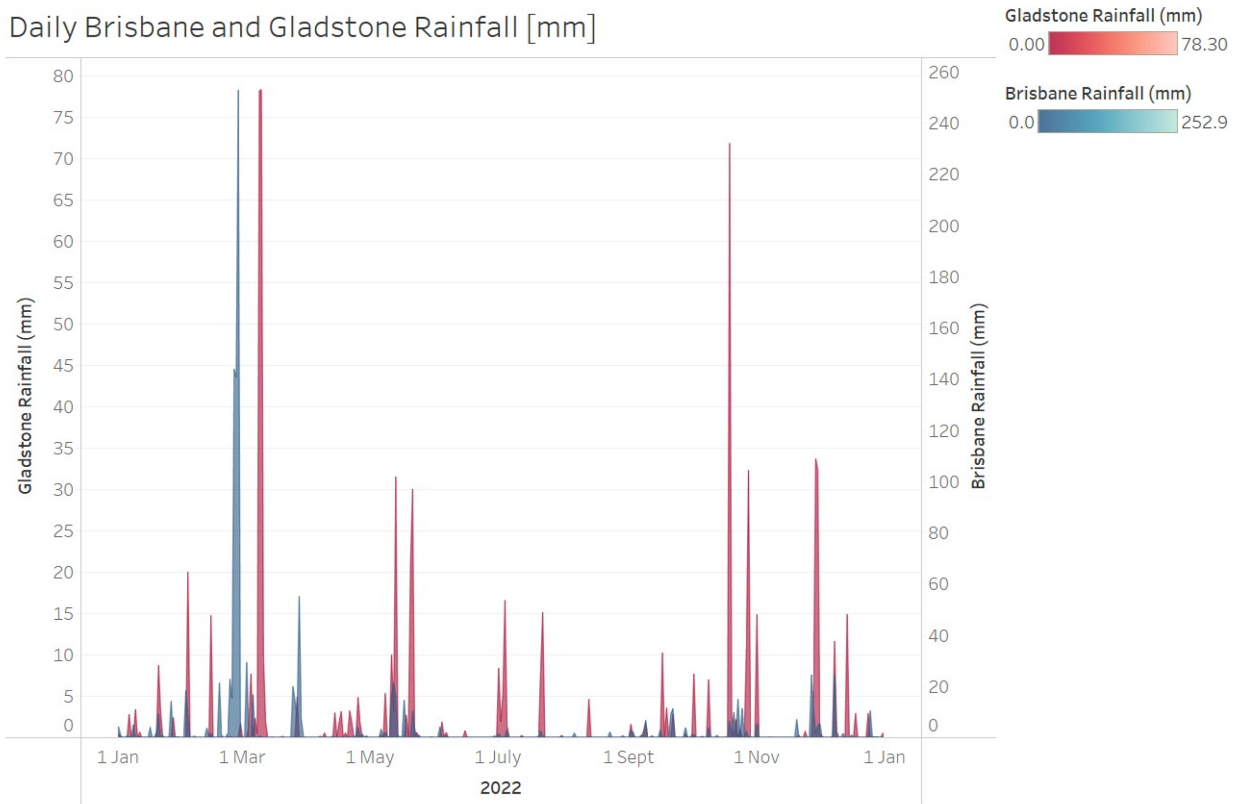
Annual Statistics

Avg. Brisbane Windspeed	2.72
Max. Brisbane Windspeed	27.72
Std. dev. of Brisbane Windspeed	3.51
Avg. Gladstone Windspeed	12.06
Max. Gladstone Windspeed	55.44
Std. dev. of Gladstone Windspeed	8.91

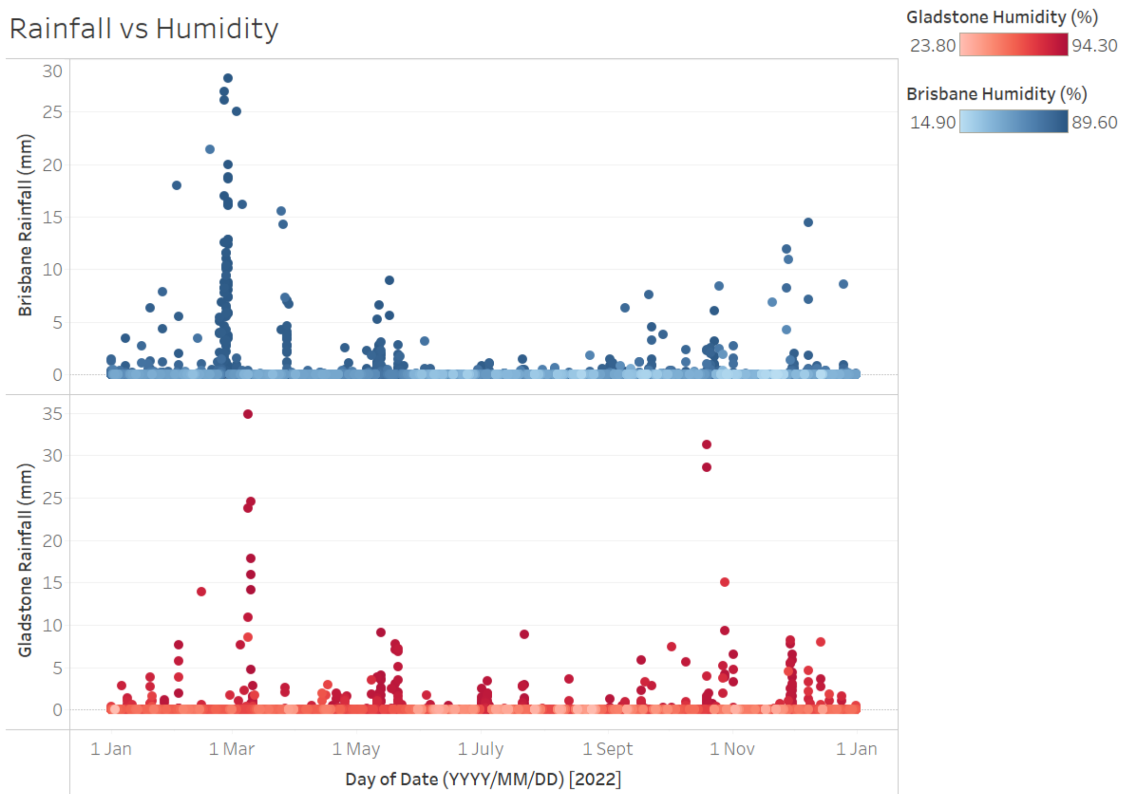
Note: Lines/Circles are 95th+ percentile results



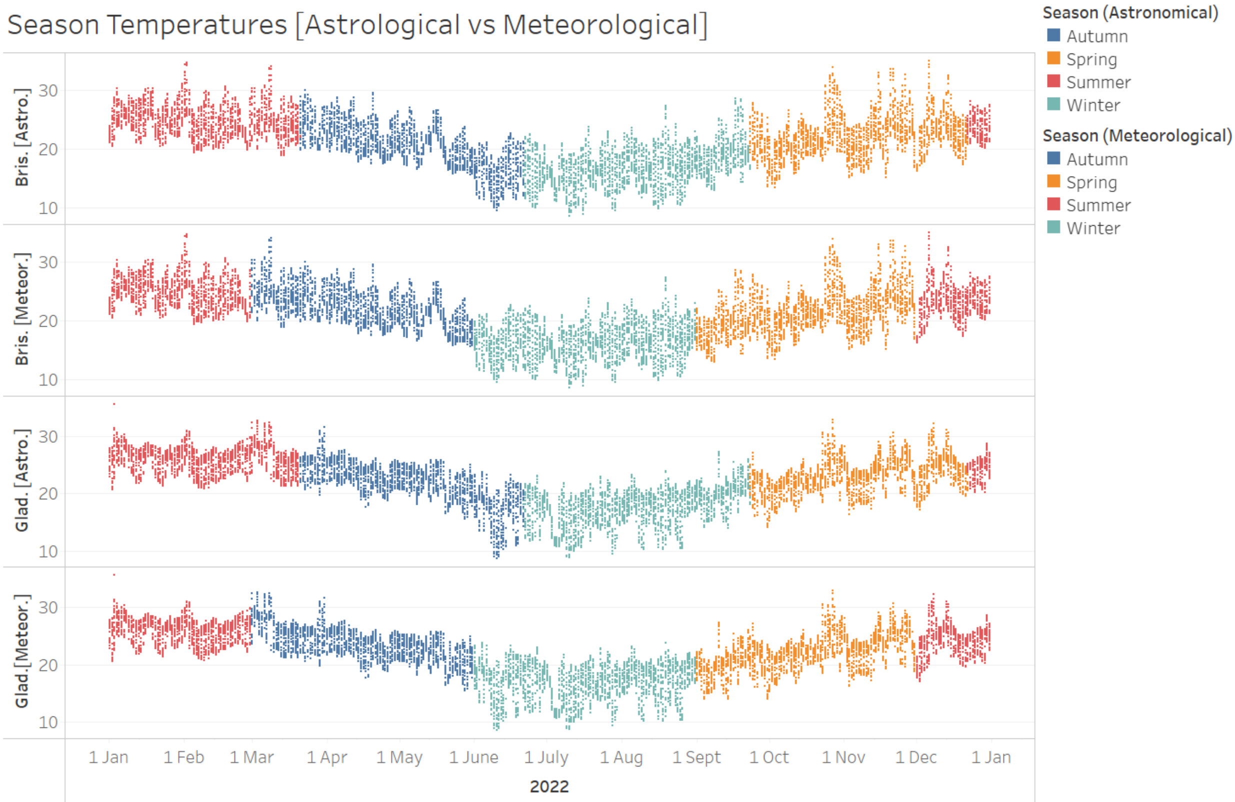
Finally, atmospheric data was analysed with monthly and seasonal categories which were implemented into the project to assess potential correlations. Initial queries included analysing the overall rainfall in the year, shown below. A colour gradient was generated for visual clarity, as due to Tableau normalising the charts, a solid line melded the two parameters.



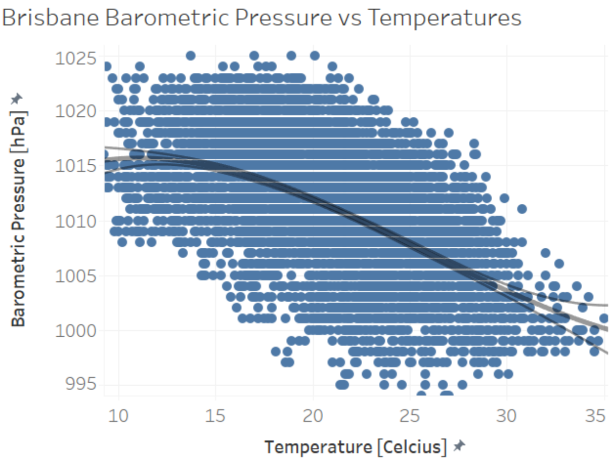
A second analysis compared rainfall to humidity, showing a clear colour gradient between drier and wetter days being strongly correlated, as expected. Below shows a visual graph of these effects:

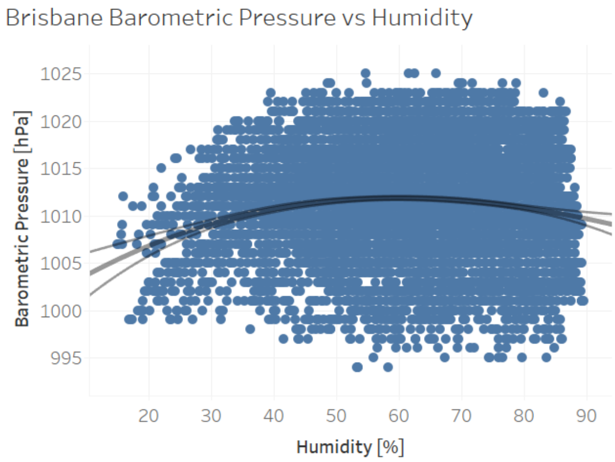


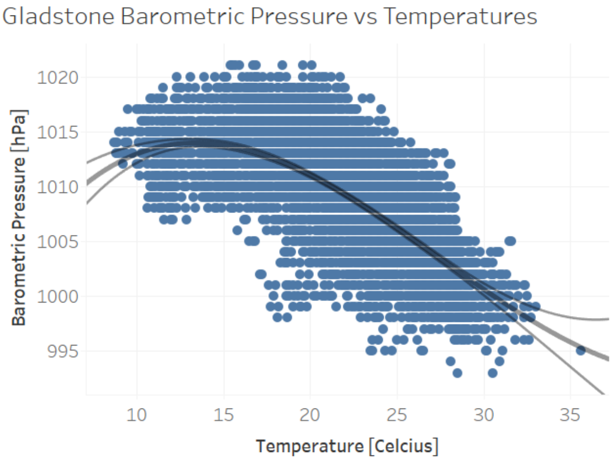
Further analysis was undertaken by gauging temperatures to meteorological and astrological seasons to determine their rough accuracy. Below shows that although the dates are relatively similar, Brisbane data tends to be better represented with astrological data for warmer months (such as the end of summer), and with Gladstone meteorological data for the ends of seasons (such as the start of summer and winter). This was visualised using daily data and coloured by season. The results below clearly show the season offsets and their differences.

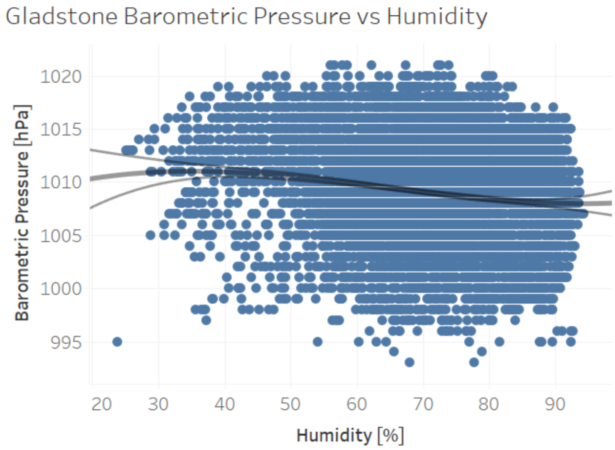


Next, metrics were compared to establish correlations. It was found that barometric pressure is weakly inversely correlated to temperate and seems to have no significant correlation to humidity. Analysis found strong trend lines but at weak (R^2 <0.40) power, indicating that trends could be present, but more complex data is needed for confirmation. Below shows these visualisations with trend-lines in Tableau:



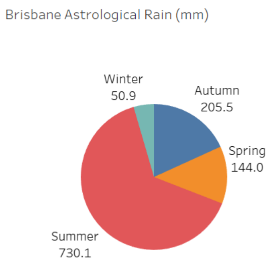


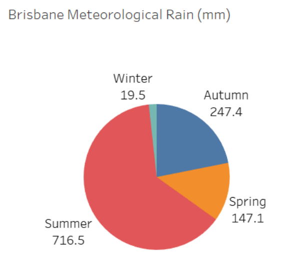




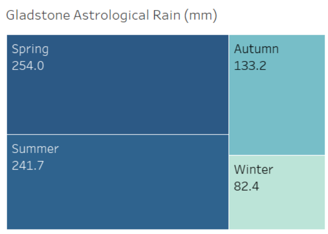
A final visualisation included generating a dashboard, comparing astrological to meteorological data in a more visual and direct way. For distinction, pie charts were used to visualise rainfall for Brisbane, and Tile charts for Gladstone. Overall, compared to astrological data, more rainfall occurred during "Summer" in Brisbane's meteorological rainfall results, whereas "Autumn" had the most rainfall for Gladstone. The dashboard below also clearly shows the differences in temperature:

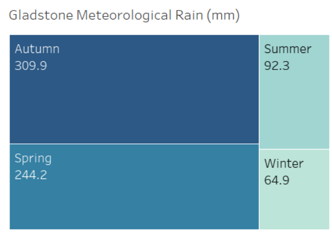
	Season (Meteorological)			
	Autumn	Spring	Summer	Winter
Brisbane Rainfall (mm)	247.4	147.1	716.5	19.5
Gladstone Rainfall (mm)	309.9	244.2	92.3	64.9
Avg. Brisbane Humidity (%)	68.7	58.2	64.6	59.5
Avg. Gladstone Humidity (%)	74.4	71.6	70.5	67.9
Avg. Brisbane Temperature (degC)	22.3	21.2	24.3	16.2
Avg. Gladstone Temperature (degC)	23.6	22.0	25.4	16.9





	Season (Astronomical)			
	Autumn	Spring	Summer	Winter
Brisbane Rainfall (mm)	205.5	144.0	730.1	50.9
Gladstone Rainfall (mm)	133.2	254.0	241.7	82.4
Avg. Brisbane Humidity (%)	66.4	57.9	66.3	60.5
Avg. Gladstone Humidity (%)	71.7	70.5	71.7	70.5
Avg. Brisbane Temperature (degC)	20.3	22.2	24.7	16.9
Avg. Gladstone Temperature (degC)	21.5	23.1	25.8	17.6





CONCLUSION

In summary, it was found that most gas molecule concentrations remained within safe limits according to WHO guidelines bar 2 days in Gladstone PM10 and Brisbane PM2.5 levels. Additionally, a mild inverse correlation was seen between temperature and barometric pressure. Finally, astrological seasons fit Brisbane better compared to meteorological data for Gladstone. These results were visualised in Tableau from querying PostgreSQL managed and cleaned composite datasets. Using these data metrics, I have demonstrated how meteorological data differences between Brisbane and Gladstone can directly infer unique characteristics and pollution indicators in these regions, as well as how SQL can be successfully used to implement multiple data analytics procedures to clean large and complex sets of raw data.

Citations:

WHO. (2021, September 22). What are the WHO Air quality guidelines? World Health Organization. https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines

WHO. (2024). WHO global air quality guidelines [Review of WHO global air quality guidelines]. https://iris.who.int/bitstream/handle/10665/345329/9789240034228-eng.pdf

time and date. (2024). Solstices & Equinoxes for Brisbane (Surrounding 10 Years) [Review of Solstices & Equinoxes for Brisbane (Surrounding 10 Years)]. Www.timeanddate.com. https://www.timeanddate.com/calendar/seasons.html

Calendarr Australia. (2023). Seasons Of The Year In Australia [Review of Seasons Of The Year In Australia]. Www.calendarr.com. https://www.calendarr.com/australia/seasons-of-the-year-in-australia/

GOV, Q. (2023). Air Quality Monitoring - 2022 [Review of Air Quality Monitoring - 2022]. Queensland Government. https://www.data.qld.gov.au/dataset/air-quality-monitoring-2022