**Michael Christopher – s224830467**

**CAPRI - Correlation of Air Pollution and Respiratory Illness**

*What have been done?*

1. **Finalised Data Collection for Air Quality Index**

Completed the acquisition of AQI datasets across multiple regions, ensuring coverage for core pollutants (CO, NO₂, O₃, PM10, PM2.5).

*So what?* This establishes a solid foundation for analysis, ensuring our models are supported by comprehensive and reliable inputs.

Data from <https://discover.data.vic.gov.au/dataset/epa-air-watch-all-sites-air-quality-hourly-averages-yearly>

A screenshot of a computer

AI-generated content may be incorrect.

Data from [Air Pollution in Victoria: Real-time Air Quality Index Visual Map](https://aqicn.org/map/victoria/)

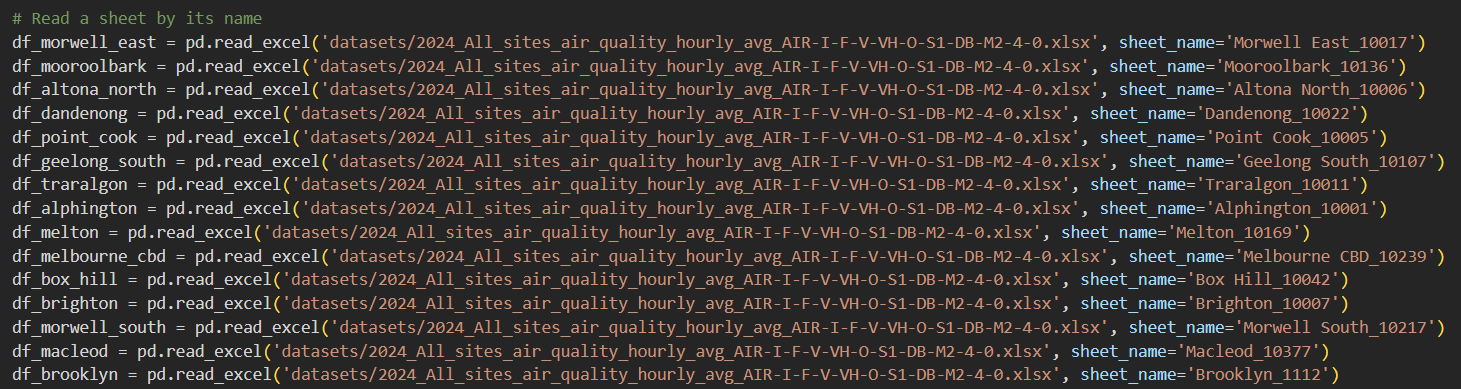
A screenshot of a table

AI-generated content may be incorrect.

1. **Developed Data Cleaning & Preprocessing Script**

Implemented Python scripts to handle missing values, align timestamps, and standardise pollutant units across datasets.

*So what?* Automated preprocessing streamlines future updates, reduces error risk, and ensures consistent quality in downstream modelling.



In this case, I use a formula in <https://document.airnow.gov/technical-assistance-document-for-the-reporting-of-daily-air-quailty.pdf> to calculate the AQI

A paper with text and images

AI-generated content may be incorrect.

A table with numbers and letters

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

1. **Tested API Call with WAQI (World Air Quality Index) API**

Successfully wrote and executed test code to fetch real-time AQI data, validating the feasibility of API-based integration.

*So what?* This capability enables dynamic data ingestion, supporting future extensions like interactive dashboards or predictive monitoring.

A screen shot of a computer program

AI-generated content may be incorrect.

1. **Air Quality Prediction Models**

Began predictive modelling using multiple machine learning approaches (e.g., Linear Regression, Random Forest, Gradient Boosting, etc.) to forecast AQI values, aiming to identify the model with the best predictive performance.

A computer screen shot of a program code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Next actions?**

1. Refine preprocessing pipeline for full dataset integration with health data (AIHW).
2. Begin exploratory data analysis (EDA) to identify patterns and validate dataset reliability.
3. Prototype initial visualisations (e.g., pollutant trends, heatmaps) for technical and non-technical audiences.
4. Work more on the machine learning modelling so it is more in depth and meaningful.