Project Report

# GitHub URL

[Your projects (github.com)](https://github.com/Crowmium1?tab=projects)

# Abstract

This project is taking weather data from an API and cleaning that data into a dataframe. Other pollution data is downloaded from an online API into an excel file. This is loaded into Python and combined with the weather data. Once this data is fully cleaned to suit the purposes of the study, this data is imported to another Python file for plotting variables against each other. Some summary statistics are also performed on this dataframe.

# Introduction

This project feels as though it would be an easy first project for analyzing data and getting the hang out the Pandas, Matplotlib and other small libraries I will be using it for.

# Dataset

I selected an API from the website ([Interactive Weather API Explorer - WeatherAPI.com](https://www.weatherapi.com/api-explorer.aspx)) as a source of data for weather variables over a 14 period for the first 14 days of April. I then selected another dataset from ([Air Quality Historical Data Platform (aqicn.org)](https://aqicn.org/data-platform/register/)), where I found a free dataset for the Galway region which was downloadable in csv. format.

# Implementation Process

* There was some trial and error involved in finding out how to get the API to connect and put that information into a dataframe. Then I had to figure what the dataset contained, so that I could take the correct information.
* I experimented with hourly data first, but decided on selecting the daily data instead. I needed to figure out how to pull all the different days with nested information for each one into one dataframe.
* Once I was able to do this, my next problem was that I only had one set of values in the dataframe. I needed to select a range a dates to pull from and iterate these using the enumerate method. This method solved the problem I was having somewhat with the indexing not working properly. The other way I tackled this problem was by resetting and renaming the index and removing unwanted indexes.
* I checked for connection to the internet as a condition to input and remove information. Else an error message was printed. I removed unwanted rows and columns.
* I imported the AirQuality.csv. Read it through Pandas library and selected the desired columns using iloc.
* I combined both dataframes once both were cleaned. I removed more columns, altered max rows and columns and converted the date to days, so that they will read better on a graph.
* The first plot I made was avg. temperature vs. air pollutants over time. I included some summary statistics, such as the mean and standard deviation on the graph.
* The second plot I made was a series of 4 subplots all with time in days on the x-axes an pollutants on one of the y-axes. Another weather variable was present on the opposite y-axis for comparison.
* A legend is included, along with different colours for the multiple lines on the line plots.

# Results

# A screenshot of a computer Description automatically generated with medium confidence **Fig.1 – Nested Dataframe**

A screenshot of a computer screen

Description automatically generated with medium confidence  
**Fig.2 – DataFrame 1 successfully cleaned and indexed**

A screenshot of a computer

Description automatically generated with medium confidence**Fig.3 – DataFrame 2 imported**

A screenshot of a computer

Description automatically generated  
**Fig.4 – DataFrame 2 Column headers**

A picture containing text, diagram, line, plot

Description automatically generated**Fig.5 – Showing the effect of average temperature vs. pollutant volumes over time.**

A picture containing diagram, text, line, plot

Description automatically generated**Fig.6 – Showing the effect of Weather Variable vs. Pollutant Volumes over time.**

# Insights

There were some summary statistics analysis done on the first graph. It is evident that the concentrations of PM2.5 are significantly higher than both NO3 and PM10 levels at any point during these 14 days.  
 Some more general insights I can pull from these graphs is as the temperature rises and falls in tandem with all of the different pollutants more closely than any of the other variables being measured here.  
 It would be interesting to see some evaluation metrics being done on these graphs, such as the root mean error and the root mean squared error. The simplest thing which can be done here may be finding the (y = m \* x + b) to model the data. This is a form of linear regression where the point closest between points is found by finding the slope and intercept between points.

# References

# [Interactive Weather API Explorer - WeatherAPI.com](https://www.weatherapi.com/api-explorer.aspx) – Weather Dataset for Galway City

* [Air Quality Historical Data Platform (aqicn.org)](https://aqicn.org/data-platform/register/) – Briarhill Galway Pollution Dataset