Investigating the association between air pollutant's concentration and temperature changes

June 6, 2024

1.*Introdunction* :

Motivation :

Air pollution has long been recognized as a threat to human health and the environment. Although emissions of most primary pollutants have declined in Europe, North America, and Japan from the 1990s until the present, air pollution is still a serious problem in many places. It is currently considered to be one of the most critical environmental issues in the world. My report and my key questions are about how the impact of air pollutant concentration is related to weather modification especially temperature changes over 5 years in Ispra, Italy.

2.*DataSources* :

in the first research, I found two datasets related to equivalent black carbon and particular matter concentration in 2021. However, these datasets are simple and clean so there is no need to do specific data Transformation methods such as normalizing and removing duplicate rows on them. As soon as I realized they did not meet my requirements, I decided to expand my timeline into 5 different years from 2018 to 2022.

Licenses :

I chose these datasets to have less Nan value in comparison to others also the license is an open data license as well as the publisher is a valid publisher as I am assured of their data quality and data accuracy. (Joint Research Center and Meteostat service). Equivalent black carbon and particular matter concentrations are based on the European Commission reuse notice License and Meteostat data is provided under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Public License (CC BY-NC 4.0). As this report is academic and educational (non-commercial) and I also provide a link to the source datasets, thus I reference the main provider. Moreover, I do not alter any data and any modifications for the purpose of data engineering have been indicated. Therefore, I ensure that this usage of data from the Joint Research Center and Meteostat service is legally followed and fulfills their obligations.

Datasource Description:

In this project, all datasets are structured data in CSV format which has the following data quality dimensions: their accuracy and consistency is not sufficient as their integer range and data type format is not correct in some columns. e.g. black carbon field. All data that I need for analyzing is uncompleted such as missing values in Black carbon measurement. Since all these data are historical, the age of data is appropriate so they are Timeliness. they are relevant and related to the main question. Overall, There are 13 datasets in 3 main groups:

Datasource1: 5 datasets for measurements of particle matter number concentration in Ispra, Italy, each of them is for one specific year(2018-2022). I demonstrate links for the year of 2021 as a sample to make it short.

These datasets contain Atmospheric Particle numbers which have been calculated based on two methods of DMPS_Total and CPC_Total Particle Concentration in a specific Date and time. Metadata URL: DMPS_Total Data URL: DMPS_Particle_Concentration_2021

Datasource2: 5 datasets for measurements of equivalent black carbon concentration in Ispra, Italy, each of them is for one specific year(2018-2022). These datasets contain Atmospheric Particles-Equivalent Black Carbon which has been calculated based on Aethalometer at a specific Date and time. Metadata URL: Metadata_Equiv_BlackCarbon Data URL: Equiv_BlackCarbon_AETH_2021

Datasource3: finally, one dataset is for weather changes from 1973 to 2024. For using this data source, I should find the nearest weather station to Ispra based on latitude and longitude, It is Milano, Malpensa weather station. This dataset contains The date string in the format of YYYY-MM-DD, The average air temperature in °C, The maximum air the temperature in °C, The daily precipitation total, The maximum snow depth in mm, The average wind direction in degrees, The average wind speed in km/h, The peak wind gust in km/h, The average sea-level air pressure in hPa', The daily sunshine total in minutes. Metadata URL: Metadata weather data source Data URL: Weather data source

In addition, there are side data sources essential for understanding the concept of the main dataset. These documents typically help mappings to column names to their corresponding meanings in the primary in Datasource.

Data URL: subset full dataset and subset map dataset

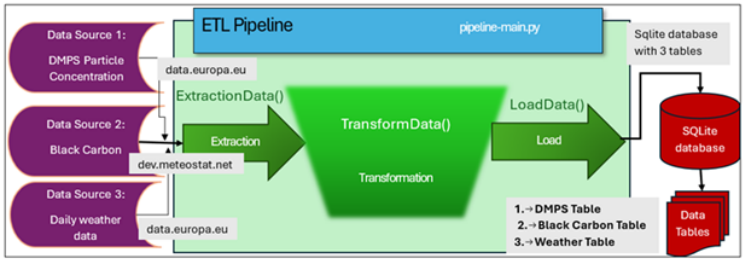
3.*DataPipeline* :

Method :

There is a pipline shell script in .project/pipeline.sh that creates a new SQLite database from all of these data sources as a standalone file through the ETL process. The requirements are just SQLqlit in Python packages.

PipelineStructure :

As illustrated in the following diagram, the project follows a structured ETL (Extract, Transform, Load) pipeline approach. All these modules are implemented as functions.



Extraction :

In this phase, All data sources are obtained from various online sources and implemented as an Extraction function. However, due to differentiation in Date Time, The datasets come from varied sources, accurately, they are in the same areas. Therefore, DPMS and Black Carbon have been appended based on date time respectively. The dataset weather is a GZ file.
So it is downloaded and unzipped afterward, it is saved as a CSV file.

***Transformation* :**

I separated my transformation process into two parts which worked based on the dataset's URLs. Some of the transformations are shared such as removing duplicated data, counting and removing Nan value due to most of them are Nan and not useful, renaming all columns in date time format to a unique name like Date that I can merge and manipulate them in the analyzing phase to plot them conveniently. Thus, I put these steps in the main body of the transformation function in order to apply them to the entire data source. Continuously, check the validation of data manually and randomly. Converting Black carbon data type to integer values as there is no float number for none of them. Converting data type txt to timestamp for data fields, since it provides the calculation of built-in data time functions in the next level. Afterward, specifically for the Weather dataset, due to the lack of headers, I should find its header from metadata information to add it to the data, and after that, I can remove useless columns. Finally, I rectified the scope duration with the help of filtering. Subsequently, these steps perform data frame cleaning and generate a new, refined version.

***Loads* :**

The result is to generate the final dataset stored in an SQLite database named "AtmosphericAndTemperatureAnalytics.sqlite" including three tables.

***ResultandLimitations* :**

The final result illustrates the outline of the SQL query from three tables of SQLite database, I utilized SQLite in view of the fact that it is lightweight. There are some missing data for the last month of 2022.

	Date	DMPS_Total
0	2018-01-01 00:04:14	12629.411394
1	2018-01-01 00:16:01	12779.947706
2	2018-01-01 00:27:47	13376.037479
3	2018-01-01 00:39:33	12633.864300
4	2018-01-01 00:51:20	12351.211675

	Date	Black Carbon
0	2018-01-01 00:05:00	6728
1	2018-01-01 00:15:00	6768
2	2018-01-01 00:45:00	7304
3	2018-01-01 00:55:00	7021
4	2018-01-01 01:05:00	6973

	Date	tavg	tmin	tmax
0	2018-01-01 00:00:00	3.2	2.6	4.2
1	2018-01-02 00:00:00	4.4	-1.0	13.0
2	2018-01-03 00:00:00	4.0	0.0	8.6
3	2018-01-04 00:00:00	6.5	2.0	12.0
4	2018-01-05 00:00:00	2.9	-1.5	6.0

for Black carbon measurement. For this issue, I searched more and looked forward to other institutes to provide data but it was not helpful as my project is related to a specific location, therefore it is hard to find data. Generally, data is for big cities or regions. I assume that the measurements have been done precisely because it is provided by valid publisher but always there is a miscalculation.

In conclusion, due to enhancements in data quality, the following data quality dimensions are expressed:

1. Their accuracy and consistency are sufficient as their integer range and data type format are correct. e.g.the temperature data reflects the real data and it is correct in the format aspect.
2. All data that I need for analyzing is getting completed and they contain all the necessary information.
3. Since all these data are historical, the age of the data is appropriate so they are Timeliness.
4. They are relevant and related to the main question.