Investigating the association between air pollutant concentrations and temperature changes

1. Introduction:

Motivation:

Air pollution is a pressing global issue with wide-ranging implications for human health, environmental well-being, and climate change. Among the various air pollutants, black carbon(BC) stands out as a significant contributor to the complex web of air quality and climate concerns especially global warming we face. Commonly known as soot, black carbon is emitted by diesel-fuelled vehicles, industrial processes, residential fireplaces, and woodstoves. Wildfires are its largest natural source.[1] Particles have a major impact on how much sunlight reaches and heats our planet. When the amount and distribution of the radiation that heats the entire world changes, this also changes the complex airflow patterns that give us what we normally call "weather". The effects of particles on radiation balance are largest in their source regions, i.e. the most populous and industrialized regions, such as Europe airborne particles have an indirect effect on the climate because the climate impact of clouds depends in part on how particles change the way clouds reflect light. [2] The analysis of air pollution measurements contributes to showing important trends or changes in atmospheric composition. This data analysis report allows us to benchmark the impacts of emissions and increasing temperature. Hence, among all entire air pollutant concentrations, I choose black carbon and DMPS Particle concentrations for this study.

Question:

This report analyses the overall and seasonal trends of equivalent Black Carbon with Aethalometer Measurements and Atmospheric Particles-DMPS concentrations in Ispra, Italy. Based on a comparison between concentrations and temperature changes, It will be determined how the impact of air pollutant's concentration correlated with modification of temperature over 5 years in Ispra.

2. Used Data:

I have used particulate matter data of equivalent Black Carbon Aethalometer, DMPS Measurements, and air temperature data collected at two official monitoring stations namely Ispra Atmosphere Biosphere Climate Integrated monitoring Station of the JRC and Milano, Malpensa weather station, respectively. These data sources have specific environmental and Atmospheric-related interpretations which means the Black Carbon column shows the measurements of equivalent black carbon [ng/m^3]- EBC (880 nm) by Aethalometer, DMPS_Total column interprets measurements of particle number concentration [cm^3]. Particle number concentration was measured with a DMPS - Differential mobility particle spectrometer from 10 nm to 800 nm between year of 2018 to 2022 with the European Commission reuse License. Finally, the temperature values in Celsius are the tavg column representing average temperature, tmin column representing minimum temperature, and tmax column representing maximum temperature between 2018 to 2022 with the Creative Commons Attribution-Non-commercial 4.0 International Public License(CC BY-NC 4.0). To comply with the data licenses the reference of data sources are:

1. Atmospheric Particles-DMPS Particle Concentration 2022, 2021, 2020, 2019,2018 ,Metadata_DMPS_Particle_Concentration
2. Atmospheric Particles-Equivalent Black Carbon Aethalometer 2022, 2021, 2020, 2019,2018 ,Metadata_Equiv_BlackCarbon
3. Meteostat weather service: The Meteostat bulk data interface provides access to full data dumps of individual weather stations. Metadata URL: Metadata weather data source, Data URL: Weather data source, subset full dataset , subset map dataset

As it demonstrates in the following, the transformation output of my pipeline are three tables with non-null value column or invalid data and duplicate data. There is a pipeline shell script in `.project/pipeline.sh` that creates a new SQLight data file from all the data sources as an standalone file. The following results demonstrate my final data information which shows all data types and all columns, nonvalue count, sample data, counts of records and so on.

In [ ]:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 230411 entries, 0 to 230410
Data columns (total 2 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   Date        230411 non-null  datetime64[ns]
 1   DMPS_Total  230411 non-null  float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 3.5 MB
```

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 245474 entries, 0 to 245473
Data columns (total 2 columns):
 #   Column        Non-Null Count   Dtype         
---  ------        --------------   -----         
 0   Date          245474 non-null  datetime64[ns]
 1   Black Carbon  245474 non-null  int64         
dtypes: datetime64[ns](1), int64(1)
memory usage: 3.7 MB
```

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1826 entries, 0 to 1825
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype         
---  ------  --------------  -----         
 0   Date    1826 non-null   datetime64[ns]
 1   tavg    1826 non-null   float64       
 2   tmin    1826 non-null   float64       
 3   tmax    1826 non-null   float64       
dtypes: datetime64[ns](1), float64(3)
memory usage: 57.2 KB
```

	Date	tavg	tmin	tmax
0	2018-01-01	3.2	2.6	4.2
1	2018-01-02	4.4	-1.0	13.0
2	2018-01-03	4.0	0.0	8.6

3. Analysis:

Initial Inspection:

At the first step of analysing, I take a look at some statistical point of view for columns, that it makes me to be familiar with my data. To illustrate an initial inference needs to install some dependencies. Thus, I have Installed Visualization Library matplotlib, seaborn, and a requirements text file in my project repository. As it has shown, I utilize the describe function which is used to get a descriptive statistics summary of the numerical columns in the dataset. This includes mean, count, std deviation, percentiles, and min-max values of all the columns.

In [ ]:

Out[ ]:

	Date	Black Carbon
count	245474	245474.000000
mean	2020-06-11 14:33:15.837522176	1347.680960
min	2018-01-01 00:05:00	-39259.000000
25%	2019-03-31 13:17:30	369.000000
50%	2020-06-11 11:10:00	728.000000
75%	2021-08-30 11:02:30	1592.000000
max	2022-11-18 09:05:00	52795.000000
std	NaN	1678.343174

In [ ]:

```
DMPS_df.describe()
```

```
Out[ ]:
                               Date      DMPS_Total
count                        230411   230411.000000
 mean  2020-07-16 20:29:34.095416064    5876.551670
  min           2018-01-01 00:04:14    -3535.039706
  25%    2019-04-25 23:53:06.500000    3361.173864
  50%           2020-07-15 01:19:21    4900.345953
  75%           2021-10-12 23:54:12    7340.669689
  max           2022-12-31 23:52:49  214203.718795
  std                           NaN    4060.013642
```

```
In [ ]: temp_df.describe()
```

```
Out[ ]:
                      Date         tavg         tmin         tmax
count                 1826  1826.000000  1826.000000  1826.000000
 mean  2020-07-01 12:00:00    13.893538     8.580778    19.253122
  min  2018-01-01 00:00:00    -3.800000   -10.600000     0.000000
  25%  2019-04-02 06:00:00     6.700000     1.225000    12.000000
  50%  2020-07-01 12:00:00    13.550000     9.000000    19.000000
  75%  2021-09-30 18:00:00    21.500000    16.000000    27.000000
  max  2022-12-31 00:00:00    30.600000    25.500000    37.000000
  std                  NaN     8.138400     7.967887     8.495913
```

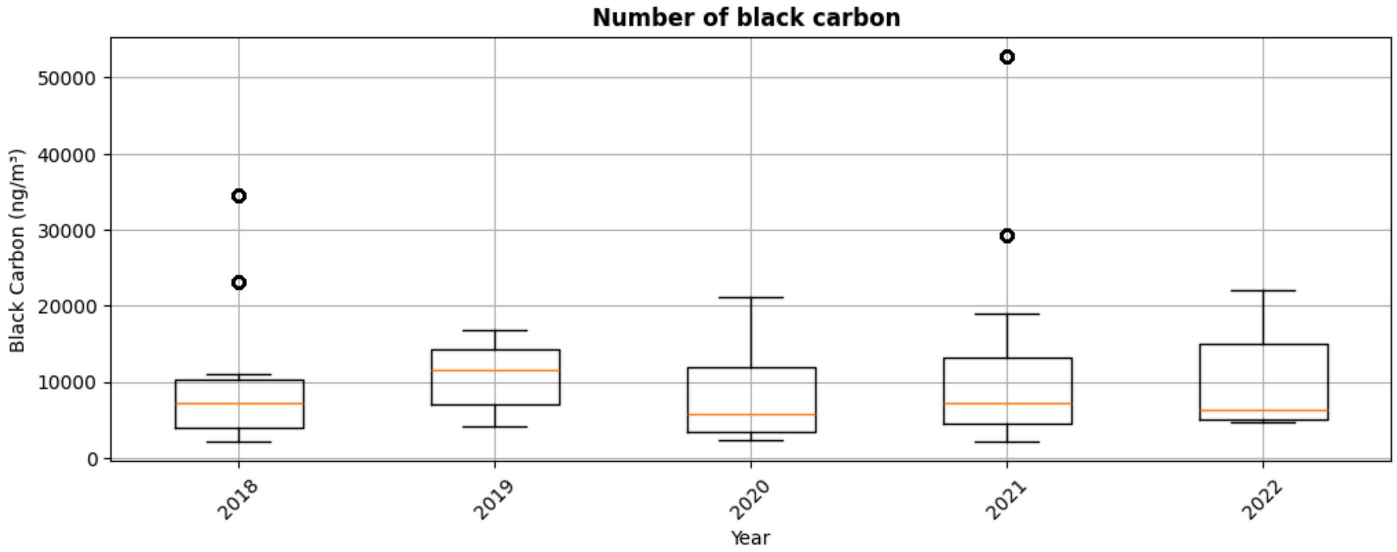
Method and Result:

Upon finalizing the ETL procedures, to gain benefits from analyzing valuable information, I utilize some built-in functions and data visualization and time series analysis methods in this regard. Analysing data points collected or recorded at specific time intervals to identify trends and seasonal patterns is my procedure. Subsequently, the result will be indicated correlations so that it provides insights to look at source apportionment time series data and make comparisons to indicators over the timeline.

Trend analysis of atmospheric concentrations:

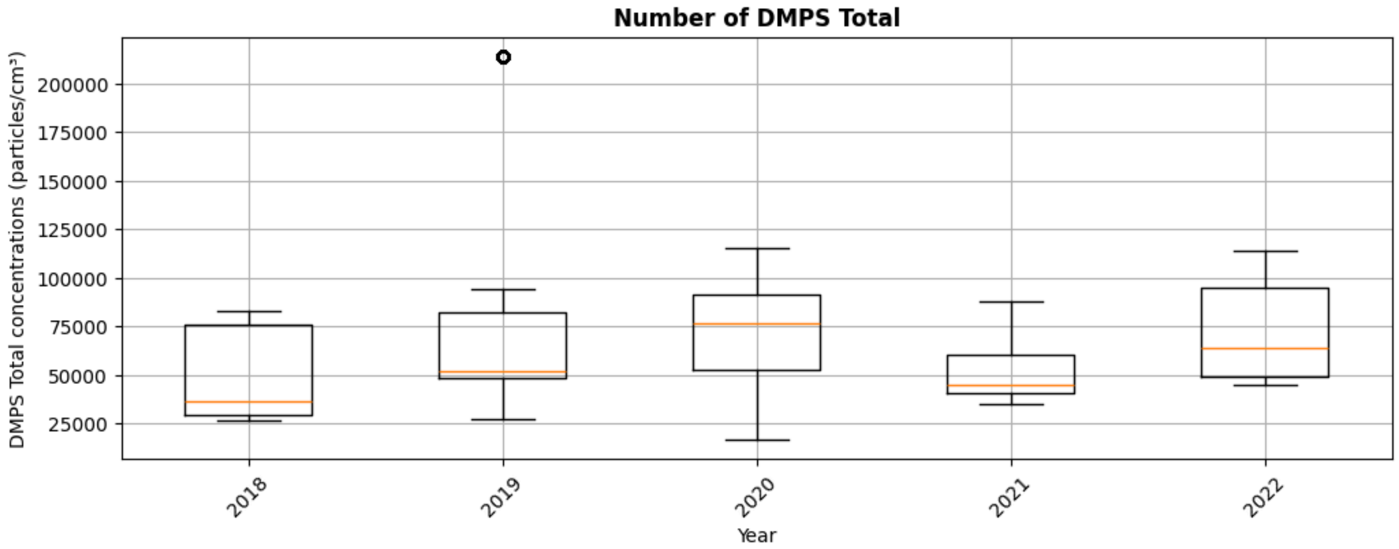
I start with visualizing concentrations data with the aid of a boxplot for annual Maximum Black Carbon(ng/m^3) concentrations: As it has shown, Black Carbon emission follows variant distribution over the years as well as data are not symmetrically distributed: Median: the median line for the years 2020 and 2022 are closer to Q1. Thus, it indicates that more than half of the year Black Carbon emissions were below the median. IQR: As we are going forward from 2018 to 2022, the interquartile range(IQR) is getting wider, so it suggests variability in annual Black Carbon emission but the difference is not significant. Whiskers: In 2018, there is a short whisker that implies consistent Black Carbon emission levels within the 1.5 * IQR range. However, it changes through the following years. Outliers: In 2021 and 2018, there are outliers above the upper whisker, that it might represent unusual Black Carbon emissions. Investigating these years can reveal if they were due to specific events or other factors such as errors in sensor systems or miscalculating with Aethalometer. This is an example of my limitations and uncertainty in measuring data.

```
In [ ]:
```



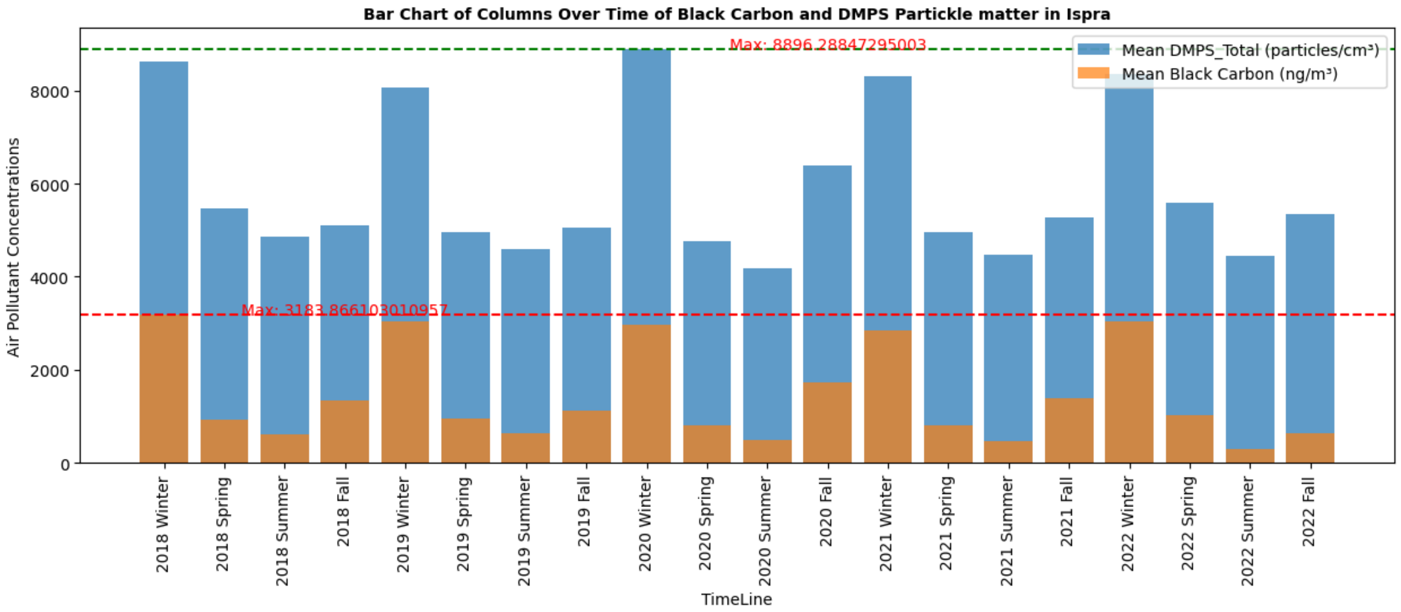
The following plot displays the box-whisker plots of the annual Maximum total of DMPS Particle concentrations(particles/cm³). A clear variation could be observed in the annual Maximum total of DMPS Particle concentrations, especially in 2021. Thus, data are not symmetrically distributed: Median: the median line for the years of 2019 and 2021 is closer to Q1. Thus, it indicates that for more than half of the year, DMPS Particle Concentration was below the median. IQR: except for 2021, the interquartile range(IQR) of the other years are similar to each other and wider than IQR in 2021, so it suggests variability in annual DMPS Particle concentrations against 2021. Whiskers: In 2018, there is a short whisker that implies consistent DMPS Particle concentration levels within the 1.5 * IQR range. However, it changes through the following years. Outliers: In 2019, there are outliers above the upper whisker, that might represent unusual DMPS Particle concentrations. Investigating this year can reveal if it was due to specific events or other factors such as errors in sensor systems or miscalculating with a DMPS measurement spectrometer. This is an example of my limitations and uncertainty in measuring data.

```
In [ ]:
```



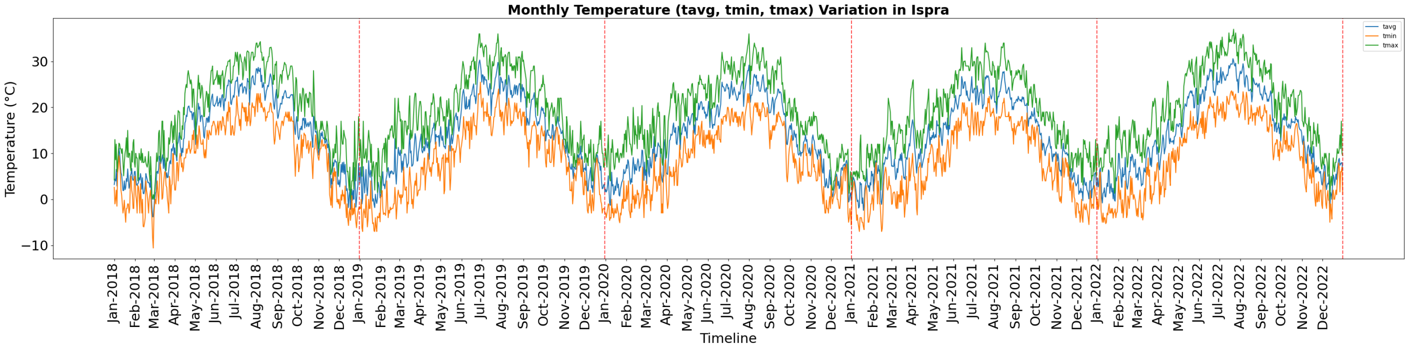
The below bar chart displays the mean value of air pollutant concentrations in seasons from 2018 to 2022 illustrating in winter, there are the maximum emissions for whether black carbon or DMPS Particle concentrations and also minimum pollution occurs in summer. In general, the trends of air pollutant concentrations are winter, spring or fall, and summer respectively from high to low. The maximum of arithmetic mean occurs in winter 2020 for DMPS Particle concentrations and in winter 2018 for black carbon. In this way, by comparing trends in consecutive seasons, the effect of air pollutant concentrations is revealed in a specific year.

```
In [ ]:
```



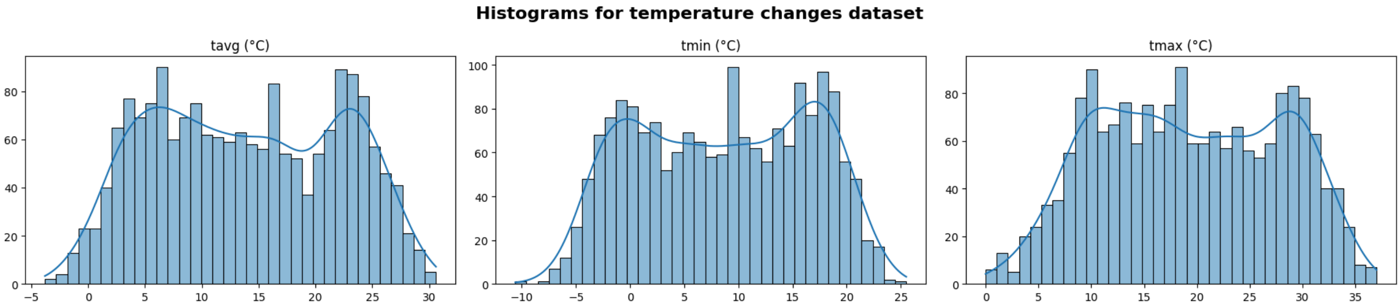
For the last data source, The temperature variations in Ispra show moderate changes over the five-year period, highlighting the region's distinct climatic characteristics. The data, which includes average temperature (tavg), minimum temperature (tmin), and maximum temperature (tmax), displays clear seasonal patterns. Typically, Ispra experiences higher temperatures during the summer months (June to October) and cooler conditions in the winter (December to February). By analyzing these temperature trends over several years, I can infer climatic shifts in the region, understand its seasonal characteristics, and identify any emerging patterns that might have an impact. In general, The trend of temperature variations has nearly the same fluctuation over these 5 years except for March 2018.

In [ ]:



Afterwards, I utilize Histograms of the weather table. These histograms provide a visual representation of the frequency distribution, helping to identify common ranges and general patterns in temperature changes. Histograms are created for each of the columns and it shows the distribution of values for columns such as tavg, tmin, tmax. The number of bins is set to 35, and kernel density estimation (KDE) is enabled to visualize the underlying distribution. The data is time-related and It suggests that there are two distinct periods within the data range that have higher frequencies of temperature occurrences. So, one peak could represent summer temperatures and the other winter temperatures. First Peak is around -1°C in tmin column, that it corresponds to the winter temperatures. The second Peak is around 30°C in tmax column which corresponds to the summer temperatures.

In [ ]:



Question: Investigating the Correlation Between Temperature Change and Air Pollutant Concentrations in Ispra, Italy:

I now have a comprehensive view of the individual outputs from the ETL pipeline. To address this question, The next step is to manipulate this data to uncover the correlations between them. The project's data is time series, with values continuously changing over time. Therefore, I convert the Date column in these three tables to the "%Y-%M-%D" format to create a consistent column for merging all tables. This merging is facilitated by using the groupby function with the max() operation. Below, I present the information on the merged data.

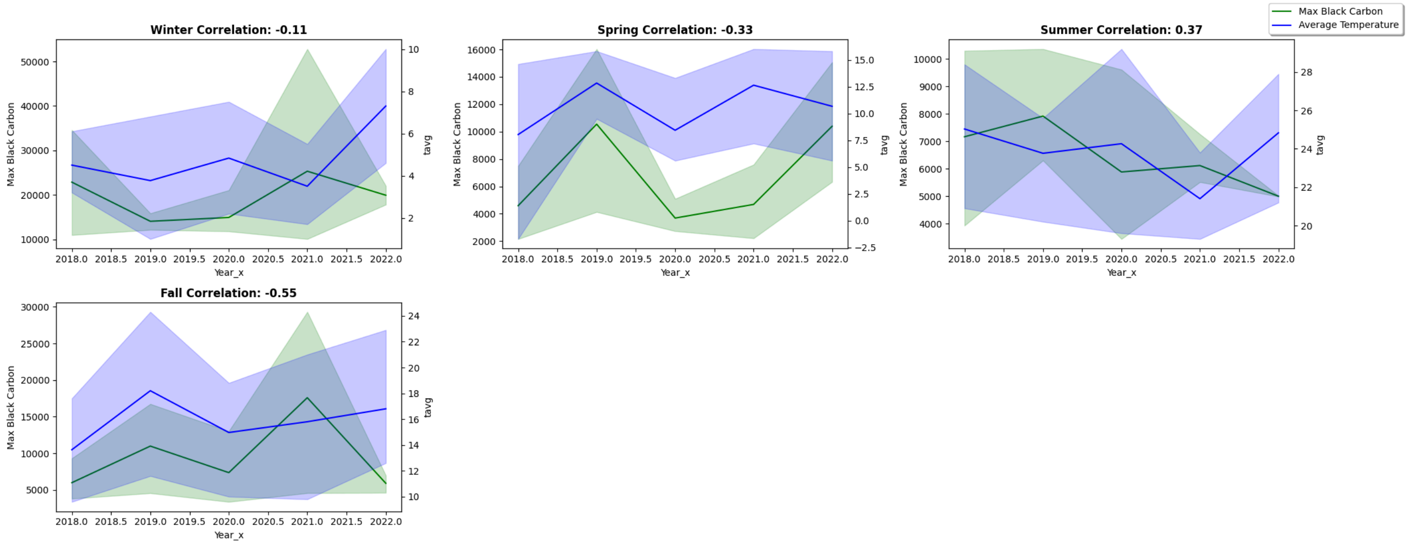
In [ ]:

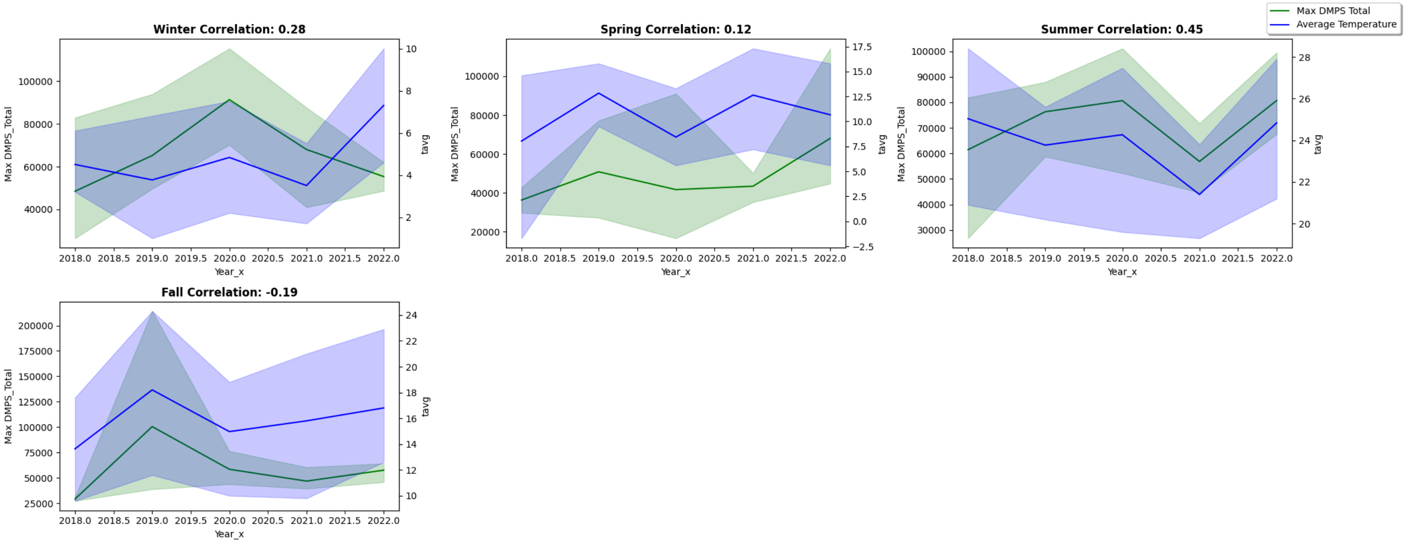
	Year_x	Month_x	Max Black Carbon	Join Date	Season	Max DMPS_Total	Date	tavg	tmin	tmax	Month_y	Year_y
0	2018	1	10997	2018-01-01	Winter	26522.440646	2018-01-01	3.2	2.6	4.2	1	2018
1	2018	2	23051	2018-02-01	Winter	82883.127327	2018-02-01	6.1	5.0	7.0	2	2018
2	2018	3	7495	2018-03-01	Spring	29714.849725	2018-03-01	-1.7	-3.6	0.0	3	2018

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59 entries, 0 to 58
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Year_x            59 non-null     int32
 1   Month_x           59 non-null     int32
 2   Max Black Carbon  59 non-null     int64
 3   Join Date         59 non-null     object
 4   Season            59 non-null     object
 5   Max DMPS_Total    59 non-null     float64
 6   Date              59 non-null     datetime64[ns]
 7   tavg              59 non-null     float64
 8   tmin              59 non-null     float64
 9   tmax              59 non-null     float64
 10  Month_y           59 non-null     int32
 11  Year_y            59 non-null     int32
dtypes: datetime64[ns](1), float64(4), int32(4), int64(1), object(2)
memory usage: 4.7+ KB
<Figure size 1400x400 with 0 Axes>
```

To indicate the relationship between temperature changes and air pollutant concentrations, I will apply both line plots and Pearson's correlation coefficient (r). This method will help determine if there is a significant correlation between temperature variations and the concentrations of black carbon and DMPS atmospheric particles in individual seasons. This thorough analysis is essential for identifying whether certain seasons display distinct patterns or trends that differ from the overall trend. I will use line plots to simultaneously show air pollutant concentrations and temperature changes. In these plots, the x-axis represents the years, while the y-axis labels the target variables (black carbon or DMPS atmospheric particles and the average temperature changes). The line in each plot denotes the mean value for each year, and the shaded area around the line indicates the variability range (upper and lower bounds) for each year, providing insights into annual and seasonal fluctuations.

In [ ]:

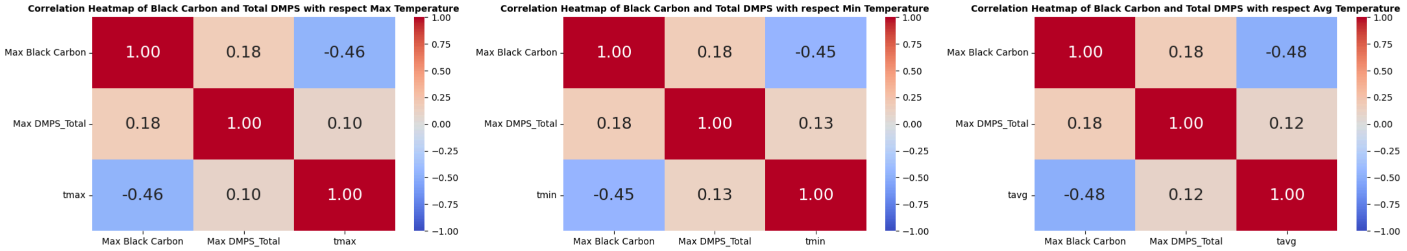




The total number of maximum Black Carbon concentrations in Ispra shows a moderate negative relationship with an average temperature value (-0.55) seasonally over 5 years. This suggests that as Black Carbon concentrations rise, the temperatures descend. In particular, It indicates to Fall season in Ispra. However, the summer season, shows a moderate positive correlation with an average temperature value (0.37) This suggests that as Black Carbon concentrations rise, there is an increase in temperatures. It demonstrates that Black Carbon concentrations has more influence on cooler weather (-0.55) than warmer weather (0.37). The total number of Atmospheric Particles-DMPS Particle Concentrations in Ispra shows a various positive and negative relationship with average temperature values seasonally over 5 years. It exhibits a weak positive correlation in spring (0.12) and a moderate negative relationship (-0.19) in Fall. Moreover, it illustrates a moderate positive relationship in summer and winter, respectively(0.45,0.28). This suggests that In particular, It indicates to Fall season in Ispra, as Atmospheric Particles-DMPS rise, temperatures descend. However, in the other seasons, especially in summer, This suggests that as Atmospheric Particles-DMPS rise, there is an increase in temperatures. It demonstrates that Atmospheric Particles-DMPS has more influence on warmer weather (0.45) than cooler weather (0.28).

For insights into all 5 years, I use a heatmap along with Pearson's r correlation coefficient to evaluate the relationships between variables. Pearson's r ranges from -1 to +1, where -1 denotes a perfect negative correlation (indicating one variable decreases as the other increases), and +1 denotes a perfect positive correlation (indicating both variables increase together). Values close to 0 indicate little to no correlation. This method helps identify significant relationships between black carbon or DMPS atmospheric particles and maximum temperature.

In [ ]:



```
Correlations of Max Black Carbon and Max DMPS Total with respect to Temperatures:
Max DMPS_Total      0.096488
Max Black Carbon   -0.461225
Name: tmax, dtype: float64
Max DMPS_Total      0.133284
Max Black Carbon   -0.447234
Name: tmin, dtype: float64
Max DMPS_Total      0.115889
Max Black Carbon   -0.478773
Name: tavg, dtype: float64
```

The following considerations can be noted from the analysis conducted using the heatmap and Pearson's r correlation coefficient. Existing Black Carbon concentrations in Ispra atmosphere seem to have a more pronounced influence on average temperature variations. Hence, I apply the Maximum number of Black Carbon concentrations and the Maximum number of atmospheric particles-DMPS Particle Concentration to indicate more influence. The total number of maximum Black Carbon concentrations in Ispra shows a moderate negative relationship with average temperature value (-0.478773) maximum temperature value (-0.461225) and minimum temperature value (-0.447234), respectively from highest impact to lower impact over 5 years. This suggests that as Black Carbon concentrations rise, the temperatures descend annually. On the other hand, the total number of maximum Atmospheric Particles-DMPS Particle Concentration exhibits a weak positive correlation with average temperature value (0.115889), maximum temperature value (0.096488), and minimum temperature value (0.133284). This suggests a slight increase in temperature over the 5 years. This indicates that atmospheric particles-DMPS concentration has little to no influence on temperature variations.

4. Conclusions:

As mentioned in previous steps, the trends and changes in black carbon and DMPS atmospheric particles individually, have no significant changes over these 5 years, Similarly, temperature changes such that they follow the annual trends. In this situation, I investigated more. So, when I compared them with each other and looking at the data collectively, I found their correlation, which demonstrates that their correlation is more seasonal than annual. I can infer that more changes occur in Fall for black carbon and in summer for air pollutant concentrations. However, black carbon emissions make the temperature to be cooler and air pollutant concentrations make the temperature to be warmer over 5 years annually. Also, it seems that there are some changes in trends of black carbon and DMPS atmospheric particles besides temperature variations in 2020, This lack of correlation suggests that the set of other factors might be more influential in determining temperature variations. In other research [3], DMPS atmospheric particles and Black carbon contribute to global warming by directly absorbing solar energy and releasing it as heat. However, there is not enough evidence to conduct that in Ispra. Over these years do not exhibit a notable correlation with either temperature change. This indicates that year-to-year variations in temperature and changes in black carbon and DMPS atmospheric particles do not have a predictable or consistent relationship. To answer my question, I investigate associations step by step and mention them in detail, briefly, equivalent black carbon and DMPS atmospheric particles have a slight impact on temperature changes. In conclusion, Each step in data engineering methods, from data preparation to visualization and statistical testing, plays a crucial role in the overall analysis process.

Limitations of the Project Report:

There is a notice in source links announcing that The information and links of datasets provided are maintained in distributed and heterogeneous information systems. Although they strive to maintain and keep links and information updated, this may not always be possible because of changes that are not registered (e.g. broken links) and updated in the relevant information systems. Missing data: This comparison and ensuring of the accuracy of the correlation analysis is possible only if sufficient data is available for the period of interest. In many instances, reliable long-term sites for comparison may not be available. Furthermore, any conclusions derived from this comparison depend on the assumption that trends observed at long-term sites accurately represent the true trend and are not significantly influenced by external factors. Geographical limitation: Including more regions in the dataset would have provided a more comprehensive understanding of the correlation between air pollutant concentrations and temperature changes across the entire country of Italy over more years to have more data. Influencing other factors: This study examined the direct correlation between temperature changes and air pollutant concentrations without thoroughly investigating other potential influencing factors, such as additional particulate matter, geographical conditions, and greenhouse gas emissions.

5. reference:

1. Improving the current air quality index with new particulate indicators using a robust statistical approach
2. Air & Environment2023 Swedish Environmental Protection Agency
3. Atmospheric black carbon concentrations Nov 1, 2022