Bachelor's Degree in Information Technologies for Sciences

Atmospheric Data Insights

Arnoldo Fernando Chue Sánchez

Course: Descriptive and Inferential Statistics

2024-2

Introduction

“Atmospheric contamination is a serious concern and may be invisible to humans as it begins to build and the concentrations of contaminants may be so gradual that it goes unnoticed. Some atmospheric contaminants may continue to build up and become visible, such as smog or the “brown haze” over cities, but the atmosphere is vitally important to the maintenance of life on Earth.” [1] I started with this description of the enormous problem posed by atmospheric contamination because it illustrates why atmospheric science is one of the most critical areas in Earth sciences.

Specifically, the National Autonomous University of Mexico (UNAM) is the largest institution conducting research on air quality and meteorology in the country. In fact, the University has many research lines and a research institute focused on atmospheric sciences. Therefore, having useful information about air pollutants is crucial. “Air pollutants comprise primary and secondary air pollutants. Primary air pollutants are emitted directly from sources. They include, but are not limited to, particulate matter (PM), sulphur dioxide (SO2), nitric oxides (NOx), hydrocarbon (HC), volatile organic compounds (VOCs), carbon monoxide (CO), and ammonia (NH3). Secondary air pollutants are produced by the chemical reactions of two or more primary pollutants or by reactions with normal atmospheric constituents. Examples of secondary air pollutants are ground level ozone, formaldehyde, smog, and acid mist.” [2]

Therefore, UNAM has created a University Network of Atmospheric Observatories (RUOA). The mission of this project is “to promote research and teaching in atmospheric sciences in the country. Through interdisciplinary and institutional cooperation, the project aims to provide relevant and reliable atmospheric information to study emerging problems and provide solutions to the challenges facing the planet in terms of air pollution, climate change, water resources, and food security, among others.” [3]

One of the stations in this network is located on the UNAM campus in Morelia. This station has been monitoring the air quality and meteorology of the city for almost ten years. During those years, different research groups have developed sensors to obtain the most reliable atmospheric data. However, since May 2023, we have witnessed one of the most interesting climate behaviours in the city, country, and planet.

In this line of thought, we can frame the present work: atmospheric sciences are of vital importance both for the present and the future of our planet. UNAM has already made notable advancements in this area. Particularly on the Morelia campus in Michoacán, sensors and expert groups in atmospheric sciences are already in place. However, there hasn't been much focus on the statistical analysis of the data they possess. In fact, the data cleaning and transformation process alone represent a significant advancement for the university community on campus.

Therefore, the main objectives of this project are to uncover new useful and high-quality information from the observatory data, as well as to statistically validate what research groups with interests in this area have been studying.

The primary achievement of this project is that the findings and analyses provide a comprehensive overview of pollution and meteorological behaviour over the last year (May 2023 – April 2024).

Throughout the document, one can find the official objectives of the work, data sources, as well as the cleaning process that had to be undertaken. Descriptions of the statistical methodologies employed, their development, the results, and their discussion are also included.

Objectives

- From the separate datasets available, create a unified and clean dataset containing both meteorological variables and air quality data.
- Conduct an exploratory data analysis to describe the current landscape of air quality in the city of Morelia, Michoacán.
- Determine which meteorological factors are most correlated with pollutant concentrations. Specifically, analyze how meteorological variables behave when ozone and PM2.5 concentrations vary.

Materials and methodologies

Let's start with the data source: the University Network of Atmospheric Observatories (RUOA). As mentioned in the introduction, UNAM has air monitoring stations in several states of the Republic. In Michoacán, the station is located on the rooftop of Building F at the National School of Higher Studies Unit Morelia. This observatory provides monthly datasets for both meteorology and air quality: https://www.ruoa.unam.mx/index.php?page=estaciones&id=9

These datasets have records every minute since 2015. Therefore, the first task was to select those corresponding to the last year (up until the publication date of this work): May 2023 to April 2024.

Once selected, they had to be concatenated separately to form a meteorological dataset for the last year and another for air quality. With both datasets created, the next step was to merge them using the timestamp of each record. In other words, by taking the timestamp of each instance from both records, instances with a common timestamp were merged to form a single record containing meteorological and air quality data. (In SQL terms, an inner join was performed using the date of the instances).

This was the dataset initially used to begin the work. The samples were initially taken per minute (later on, we will see that we changed the temporality of instances to full days).

Therefore, the variables we initially had were:

- Time
- O3
- O3_flag:
- SO2
- SO2_flag
- NO2
- NO2_flag
- NO

- NO_flag
- CO
- CO_flag
- PM10
- PM10_flag
- PM2.5
- PM2.5_flag
- Temp_Avg
- RH_Avg
- WSpeed_Avg
- WSpeed_Max
- WDir_Avg
- WDir_SD
- Rain_Tot
- Press_Avg
- Rad_Avg

Variable information:

- Time: This variable denotes the date the sample was taken. It's in the format: yyyy-mm-dd HH:MM:SS UTC-6

The air quality variables are:

- O3: ozone. Measured in parts per billion (ppb).
- SO2: sulfur dioxide. Measured in parts per billion (ppb).
- NO2: nitrogen dioxide. Measured in parts per billion (ppb).
- NO: nitrogen oxide. Measured in parts per billion (ppb).
- CO: carbon monoxide. Measured in parts per million (ppm).
- PM10: atmospheric suspended particles with an aerodynamic diameter equal to or less than 10 µm. Measured in µm/m^3.
- PM2.5: atmospheric suspended particles with an aerodynamic diameter equal to or less than 2.5 µm. Measured in µm/m^3.

Each air quality variable is associated with a flag variable indicating the operation status of the sensor taking the sample: O3_flag, SO2_flag, NO2_flag, NO_flag, CO_flag, PM10_flag, PM2.5_flag.

These flag variables can have the following values:

- OK: reliable data
- BDL: below detection limit
- OR: out of range
- OS: out of service

And finally, we have the meteorological variables:

- Temp_Avg: Average temperature. Measured in °C.
- RH_Avg: Average relative humidity. Measured in %.
- WSpeed_Avg: Average wind speed. Measured in m/s.

- WSpeed_Max: Maximum wind speed. Measured in m/s.
- WDir_Avg: Average wind direction. Measured in degrees.
- WDir_SD: Standard deviation of wind direction. Measured in degrees.
- Rain_Tot: Total amount of rainfall during that time period. Measured in mm.
- Press_Avg: Average atmospheric pressure. Measured in hPa.
- Rad_Avg: Average solar radiation. Measured in W/m^2.

With these variables explained, we can now start discussing the data cleaning process.

Data cleaning

Firstly, we had to remove the SO2 variables because there were no measurements of this molecule throughout the year. Since there were no records, it was best to delete it from the dataset.

Next, instances with flags indicating they were out of service were deleted to maintain data integrity rather than filling them with other values.

Regarding handling air quality variables outside the range or below the detection limit, if they were above the span limit, they were marked as null. If they were below the detection limit, values between that limit and its negative value were replaced with half the detection limit value. Anything outside of this range was marked as null.

The limits for each variable and the cleaning criteria were provided by the International Laboratory for Environmental Electronic Devices (LAIDEA).

Pollutant	Units	Analytical method	Detection limit	Span limit
O3	ppb	UV absorption photometry	0.03 ppb	500 ppb
CO	ppm	Non-dispersive infrared	0.04 ppm	25 ppm
NO2, NO	pbp	Chemiluminescence	0.4 ppb	0.3, 0.5 ppm (NO2, NO)
PM2.5	µg/m^3	Separation by diameter and quantification by beta radiation attenuation	4 µg/m^3	400 µg/m^3
PM10	µg/m^3			

Once these cleaning criteria were applied, we moved on to the meteorological variables. The following criteria were also provided by LAIDEA:

- Temperature (Temp_Avg)
  A z-value filter for temperature is used: values between -3 and 3 standard deviations from the mean value are left in the column, and the rest are replaced by null. The standardized anomaly (z, also known as z-score) is a measure of distance, in standard deviation units, between a data value and its mean. Commonly temperature data have a higher standard deviation around 3°C during

winter. Therefore, having -3 or +3 times that standard deviation should be extremely uncommon, and thus, such data points must be cataloged as outliers.

- Wind speed (WSpeed_AVG, WSpeed_Max)
  A simple filter is used for the averaged (WSpeed_AVG) and maximum (WSpeed_Max) wind speeds. Values from 0 to the 99th quantile are left in the column, the rest are then considered as non-typical values with few statistical representativeness. Thus, are replaced by null.
- Wind direction (WDir_Avg)
  Wind direction values between 0.001 and 359.999 are left in the column, the rest are replaced by null.
- Precipitation (Rain_Tot)
  Precipitation values greater than or equal to 0 are left in the column, null values replace the rest.
- Presion (Press_Avg)
  For the Press_Avg column, a filter of z-values for pressure is used: values within -3 to 3 standard deviations from the average value are left in the column, and the rest are set by null.
- Radiation (Rad_Avg)
  A simple filter is used for the Rad_Avg column: values greater than 0.001 are left in the column, and the rest are replaced by null.
- Relative humidity (RH_Avg)
  For column RH_Avg a simple filter is used: values from 1 to 99 are left in the column, and null replaces the rest.

At this point, data cleaning would be completed for research groups using this data. However, (solely for the purposes of this work), we continued cleaning to facilitate statistical analysis.

Firstly, the time attribute was modified. Instead of having a long format in a single column, it was separated into several columns: year, month, day, hour, and minute. This makes it easier to understand and to group data in the future. In fact, separating the date allows for checking data integrity criteria over time for those prioritizing it.

While marking and leaving some data as null is crucial for certain tasks, for our statistical analysis, it's necessary to establish a policy for handling null data, which was designed by the author of this work and is presented as follows:

- Samples with more than 30% of their attributes as null will be eliminated.
- For all variables except solar radiation, we will fill null data with the median. This is to preserve the skewness of each variable (as filling with the mean could alter it).
- In the case of solar radiation, we only have 52.52% of non-null data. Therefore, filling with any value wouldn't be representative. Consequently, when working with this variable, we will do so only with non-null samples.

With this dataset containing samples per minute, we began our work (the results are presented in the following sections). However, due to certain results, a dataset handling samples per day had to be created. This change in temporality was made to reduce data noise and improve the quality of analysis.

Therefore, instances were grouped by year, month, and day. When performing this "group by," decisions had to be made on how to handle the other variables. In this case, the variables selected for work were O3, NO2, NO, CO, PM2.5, PM10, Temp_Avg, RH_Avg, WSpeed_Avg, WDir_Avg, Press_Avg, and Rad_Avg. For all these variables, their data was averaged upon grouping to assign it to the day's data. Additionally, the Rain_Tot variable was included, but since it represents rainfall quantity, instead of averaging the instances of the day, they were summed to provide an objective daily result. In fact, this grouping resolved the issue of null data. Only one day had null solar radiation, which was substituted with its median. Days without records were not included in the dataset, resulting in a total sampling of 315 instances for the year.

Moving on to the methodologies applied in the following sections of the work:

- For descriptive statistics:
    - We used measures of central tendency and dispersion, as well as Fisher's skewness. These are useful for initial data descriptions.
    - Graphically, we will present histograms, boxplots, density plots, scatter plots, dispersion diagrams, correlograms, and regression diagrams.
    - For data grouping, we primarily used qualitative variables from the date. However, some quantitative variables were converted into qualitative ones by establishing ranges. These grouping criteria will be mentioned in their respective experiments.
- Normality tests:
    - Since we have a large number of samples (regardless of the data temporality we work with), the indicated test is the Kolmogorov-Smirnov test.
    - In cases where the data is not normal, Box-Cox transformations will be performed.
    - If there is still no normality after that, non-parametric statistical tests will be conducted.
- Non-parametric inferences:
    - For more than two samples, the Kruskal-Wallis test will be used.
    - If a treatment appears different, Wilcoxon paired tests will be used.
- Parametric inferences:
    - For homoscedasticity of variances, the Snedecor's F test (for two samples) and Levene's test (for three or more samples) were utilized.
    - For means comparison tests:
        - We used t-tests (for two samples).
        - Analysis of Variance (ANOVAs) (for at least three samples).
            - Within ANOVAs, we used post hoc tests when at least one treatment is different. Specifically, we used the Tukey test for its balance.

- Bivariate inferences:
    - Once normality has been verified using the tests described earlier, due to the large number of variables, a correlogram will be used to identify statistically significant correlations (99%) using the Pearson coefficient.
    - After this, linear regression analyses can be performed with their respective graph (95% confidence interval).
    - In this case, we can move on to working with ANOVAs of these variables to obtain more useful information.

Development

Firstly, we need to revisit the information from the datasets we are going to use. Let's remember that we have two data frequencies: air samples per minute and per day. For the per-minute data, we have 443,509 samples, and for the per-day data, we have 315 samples. Regarding the per-minute dataset, it includes the variables that have already been described in the entire dataset. However, for the per-day dataset, we only included the variables O3, NO2, NO, CO, PM2.5, PM10, Temp_Avg, RH_Avg, WSpeed_Avg, WDir_Avg, Press_Avg, Rad_Avg, and Rain_Tot.

For each hypothesis test or specific analysis, the variables to be used will be specified at that moment.

It's worth mentioning that priority was given to the analysis of ozone (O3) and PM2.5 variables. This is because they are of the highest interest to the LAIDEA research group.

Regarding data visualization, as mentioned in the previous section, density plots, histograms, box plots, and whisker plots were used (at least for the descriptive part). For correlations, initially, a correlogram was created, followed by scatter plots, linear regression plots, and even dot plots. The latter were especially used to lead into analysis of variance.

Finally, let's talk about our inferences. Firstly, the confidence level used for almost all tests in this project (except for significance in correlations) was 95%. In the case I mentioned of finding the most statistically significant correlations, a confidence level of 99% was used to be stricter and focus on the most significant correlations.

Regarding our hypothesis tests, they depend on each test we conducted:

- Normality tests. Let's remember that due to the large amount of data we have, we used Kolmogorov-Smirnov. Therefore, the hypothesis test is formulated as follows:
    - H0: The distribution of the random variable is normal.
    - H1: The distribution of the random variable is not normal.

  In this test, our test statistic is Dc.

  Decision rule:
    - $D_C < D_{n,\alpha}$ accept H0
    - $D_C > D_{n,\alpha}$ reject H0
- Homoscedasticity tests for variances:
    - For Snedecor's F test (two samples):
        - H0: The variances are equal.

            - H1: One of the variances is greater than the other (this is adjusted according to the variables we are working with).

        In this case, our test statistic is Fc.

        Decision rule:

            - Fc < $F_{\alpha,df1,df2}$ accept H0
            - Fc > $F_{\alpha,df1,df2}$ reject H0

        - Levene's test (three or more samples):
            - H0: All variances are equal.
            - H1: At least one of the variances is different.

        For this case, the test statistic Fc is also used, and the decision rule is the same as for the two-sample test.

    - Non-parametric tests:
        - For more than two samples: Kruskal Wallis:
            - H0: All samples come from the same population.
            - H1: At least one sample comes from a different population.

        Here, an H test statistic is used, which has a distribution similar to the chi-square distribution. Therefore, the decision rule is as follows:

            - H < χ^2$_{gl,\alpha}$ accept H0
            - H > χ^2$_{gl,\alpha}$ reject H0

        - For two samples: Wilcoxson paired
            - H0: All samples are equals
            - H1: At least one sample is different

        Here, TC statistic is used and the decision rule is:

            - Tc > Tt accept H0
            - Tc < Tt reject H0

- Test of mean comparision
    - For two samples: t student
        - Two tails
            - H0: mean1 = mean2
            - H1: mean1 != mean 2
        - One tail
            - H0: mean1 = mean2
            - H1: mean1 < or > mean2

    Statistic: t student. Decisión rule

        - tc < tt accept H0
        - tc > tt reject H0

    - Analysis of Variance ANOVAS (oneway)
        - H0: □1= □2= … = □k= μ
        - H1: at least one □i≠□k

    Statistic: Fc. Decision rule

        - Fc < $F_{\alpha,df1,df2}$ accept H0
        - Fc > $F_{\alpha,df1,df2}$ reject H0

    - Post Hoc test (if at least one sample is different in an ANOVA): Tukey

        - In this case it is used the statistic W. If the difference of a particular sample is bigger than the Wc, there are significant differences on that sample.
- Bivariate inferences:
    - In this case it is used the Pearson index for correlationship. Specifically for the correlogram. In this part instead of using the correlations with 95% of significant, it is used the 99% level of confidence.

Results

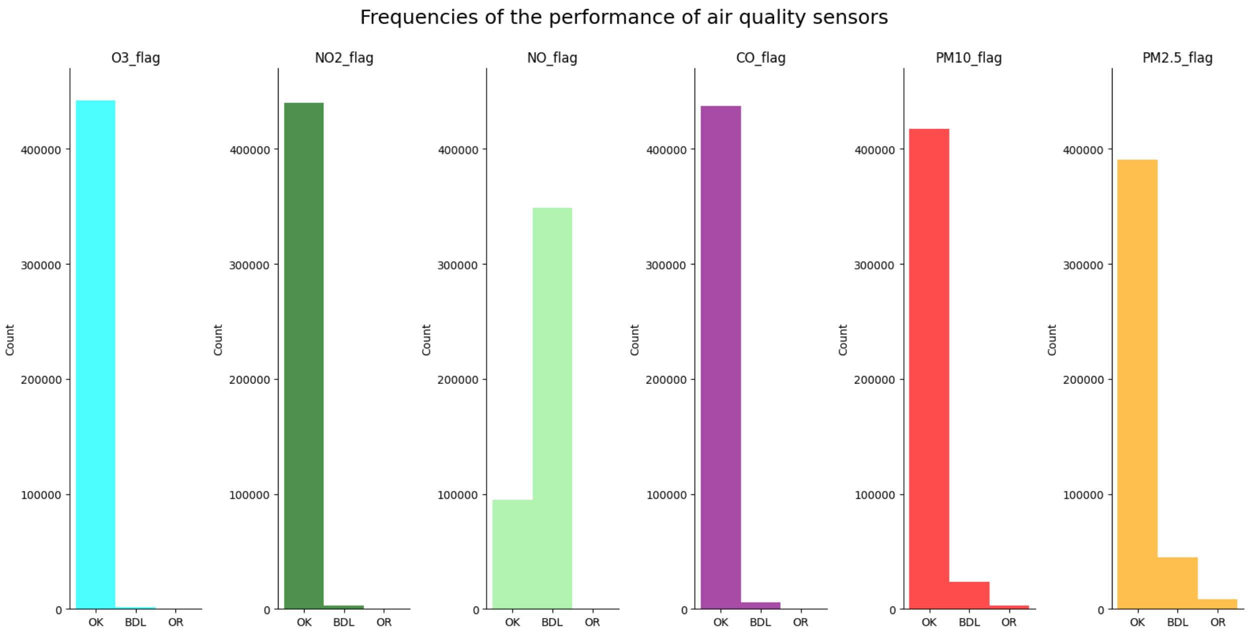
In this section, we will only present the procedures along with their results. The interpretation and analysis will be covered in the next section.

Initially, I intended to work solely with the dataset of samples per minute. The analyses that yielded successful results with this dataset are as follows:

1. Firstly, we calculated the absolute frequencies of the flag variables. Let's remember that at this point, we no longer have instances where the flags indicated out-of-service conditions because we removed them from the dataset.

	O3_flag	**NO2_flag**	**NO_flag**	**CO_flag**	**PM10_flag**	**PM2.5_flag**
OK	442190	440116	348595	437378	417350	390297
BDL	1302	3085	94710	6022	23273	44973
OR	16	307	203	108	2885	8238

Histograms were created to better illustrate this table, showing the frequencies of sensor operation.


Frequencies of the performance of air quality sensors

O3_flag

NO2_flag

NO_flag

CO_flag

PM10_flag

PM2.5_flag

We can also see the categories of OK and BDL (the most frequent ones) along with their relative frequencies.

	O3_flag	NO2_flag	NO_flag	CO_flag	PM10_flag	PM2.5_flag
OK	99.7%	99.24%	21.35%	98.62%	94.1%	88.0%
BDL	0.29%	0.7%	78.6%	1.36%	5.25%	10.14%

2. As part of this analysis of sensor performance, a count of the samples obtained in each month was also conducted. The absolute frequencies for each month are presented in this table:

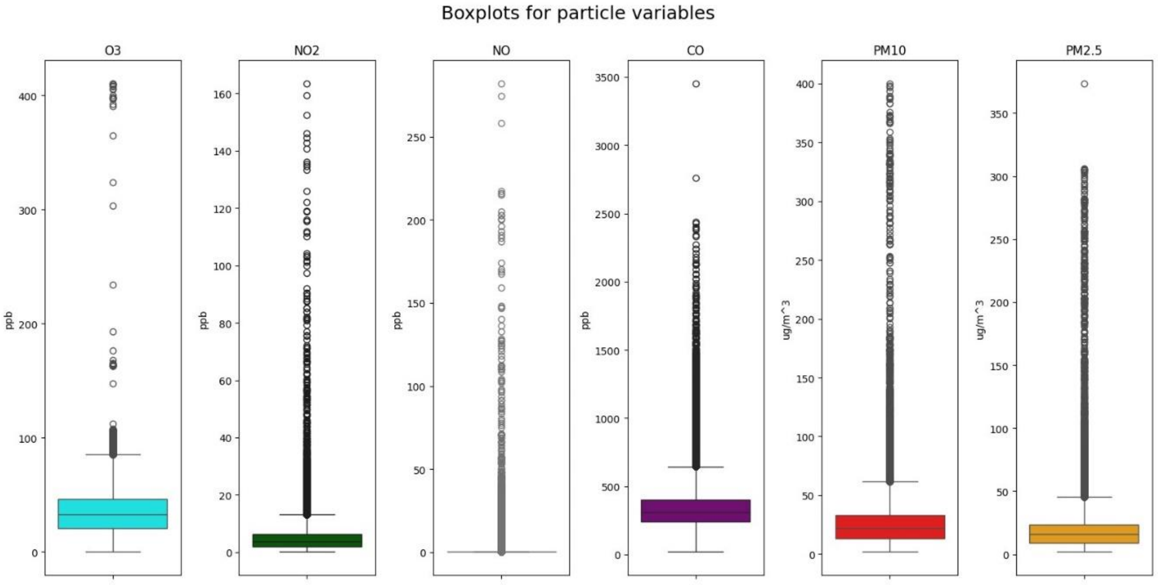
January	40176
February	36600
March	18961
April	40498
May	29463
June	40500
July	39538
August	41555
September	40498
October	41700
November	30900
December	43119

3. Using the air quality variables (O3, NO2, NO, CO, PM10, and PM2.5), we obtained basic data for their description: mean, median, first quartile, third quartile, standard deviation, minimum, maximum, and skewness.

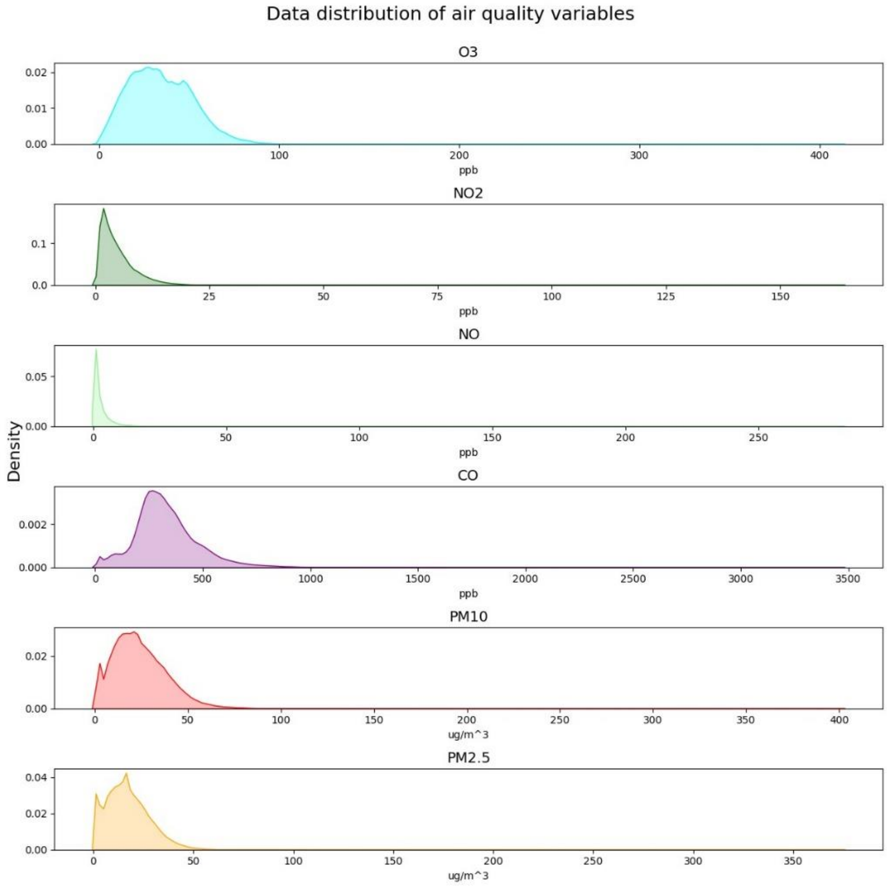
	O3	NO2	NO	CO	PM10	PM2.5
Mean	34.1781	4.8262	0.8407	0.3305	24.3758	17.4330
Median	32.57	3.72	0.2	0.31	22.18	15.99
First quartile	20.61	1.98	0.2	0.24	13.55	9.17
Third quartile	46.57	6.47	0.2	0.4	32.92	23.83
Standar deviation	17.6207	4.0499	2.8866	0.1537	15.6372	11.99
Minimum	0.015	0.2	0.2	0.02	2.0	2.0
Maximum	410.2	163.27	281.94	3.45	399.64	373.28
Fisher asymetry	0.79	3.86	27.31	1.35	2.85	3.50

Since we obtained such large skewness values, it was suspected that there were issues with the data. However, work continued nonetheless until the switch to daily data.

To accompany these measures of central tendency, dispersion, and skewness (and to further investigate the high skewness), box plots were created first.



As we can see, there is a huge amount of outlier data in each of the particle variables. To further analyze these highly atypical distributions, density plots were created to visualize the data distribution.

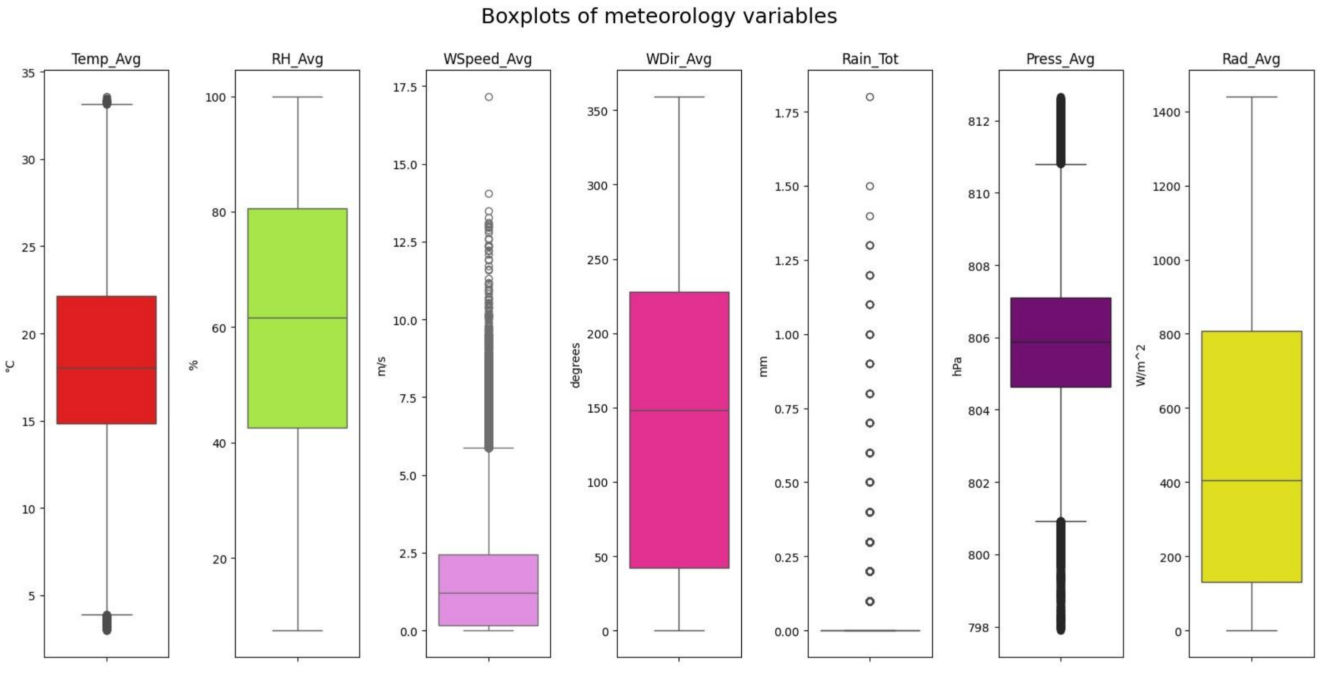


At this point, we could go directly to normality tests to demonstrate that these distributions are not, and will not be, normal. However, to complete this initial data description, we carried out the same procedure, but now with the meteorological variables.

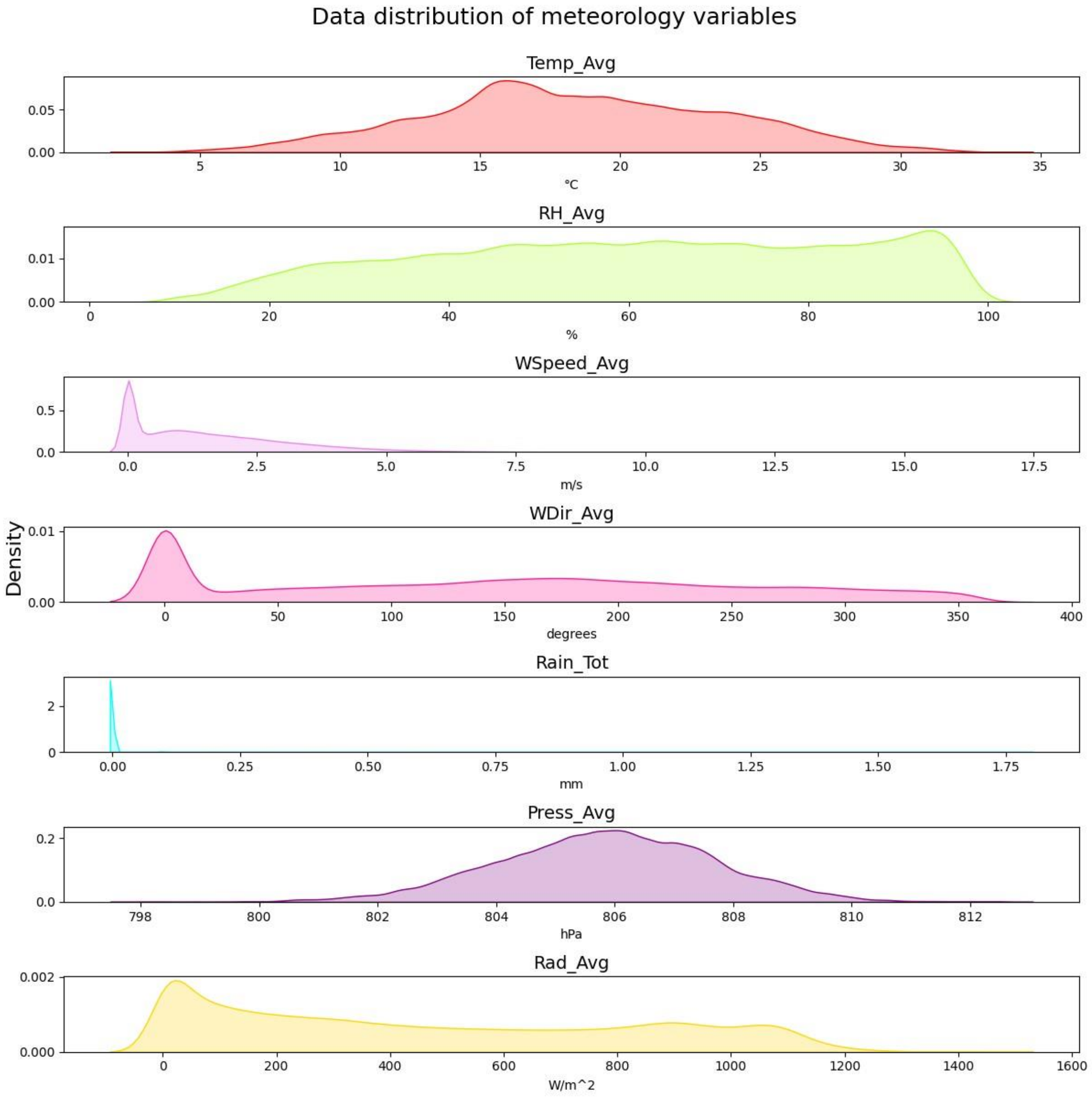
4. First, we present a descriptive table.

	Temp_Avg	RH_Avg	WSpeed_Avg	WSpeed_Max	WDir_Avg	WDir_SD	Rain_Tot	Press_Avg	Rad_Avg
Mean	18.34	60.65	1.5642	2.056	144.94	9.55	0.0010	805.83	471.90
Median	18.06	61.71	1.217	1.618	147.8	7.68	0.0	805.87	404.5
First quartile	14.85	42.58	0.172	0.451	42.73	0.174	0.0	804.62	130.1
Third quartile	22.17	80.6	2.454	3.164	228.6	14.25	0.0	807.09	807.0
Standar deviation	5.26	23.01	1.523	1.926	108.05	10.21	0.0190	1.84	366.84
Minimum	2.966	7.40	0.0	0.0	0.0	0.0	0.0	797.89	0.046
Maximum	33.57	99.99	17.15	18.78	360	103.4	1.8	812.65	1439.0
Fisher asymetry	0.04	-0.17	1.10	1.07	0.13	2.00	36.21	-0.08	nan

Note that due to the null values for solar radiation, its skewness could not be calculated (nan = not a number)
If we observed the box plots

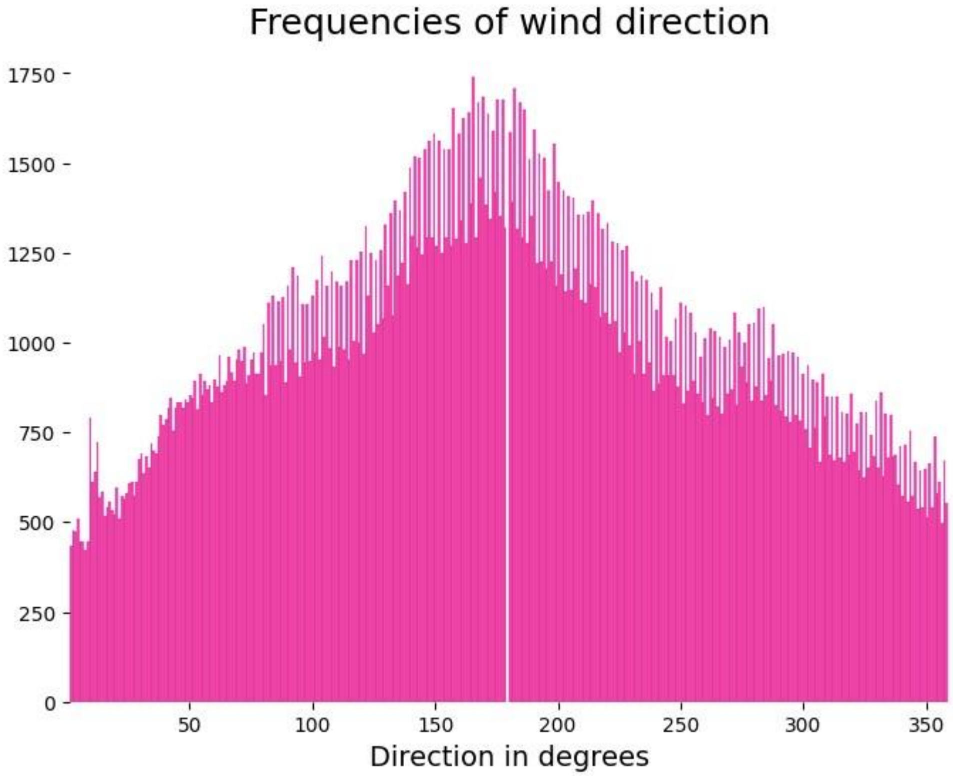


Unlike the air quality variables, here we can notice some distributions without many null values. However, the rainfall variable has only outlier data. Therefore, it is advisable to create density plots.

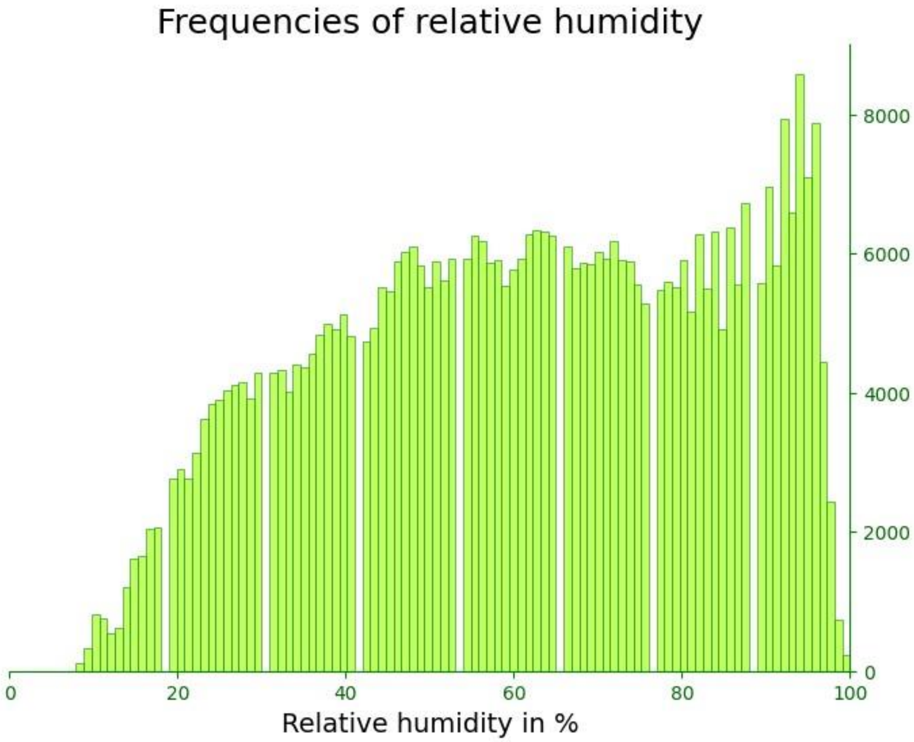


After this initial data description, more visualizations of specific variables were conducted.

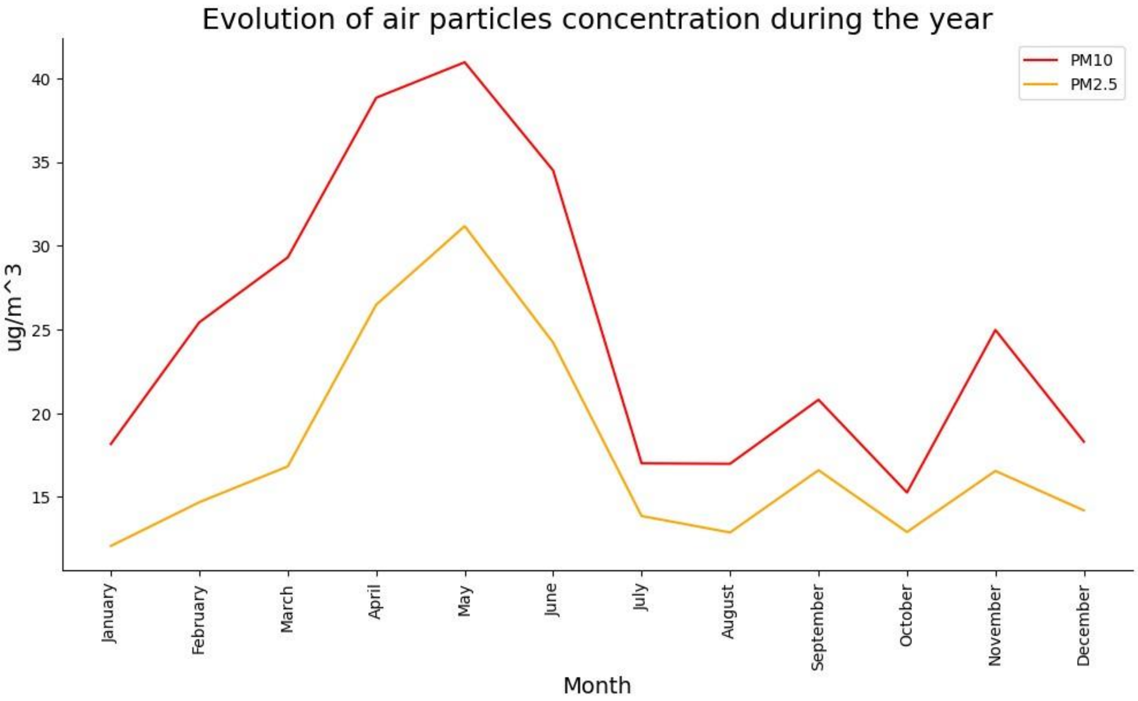
5. For the wind direction frequencies, they were rounded to fit exactly within the 360 degrees in which the wind can blow:



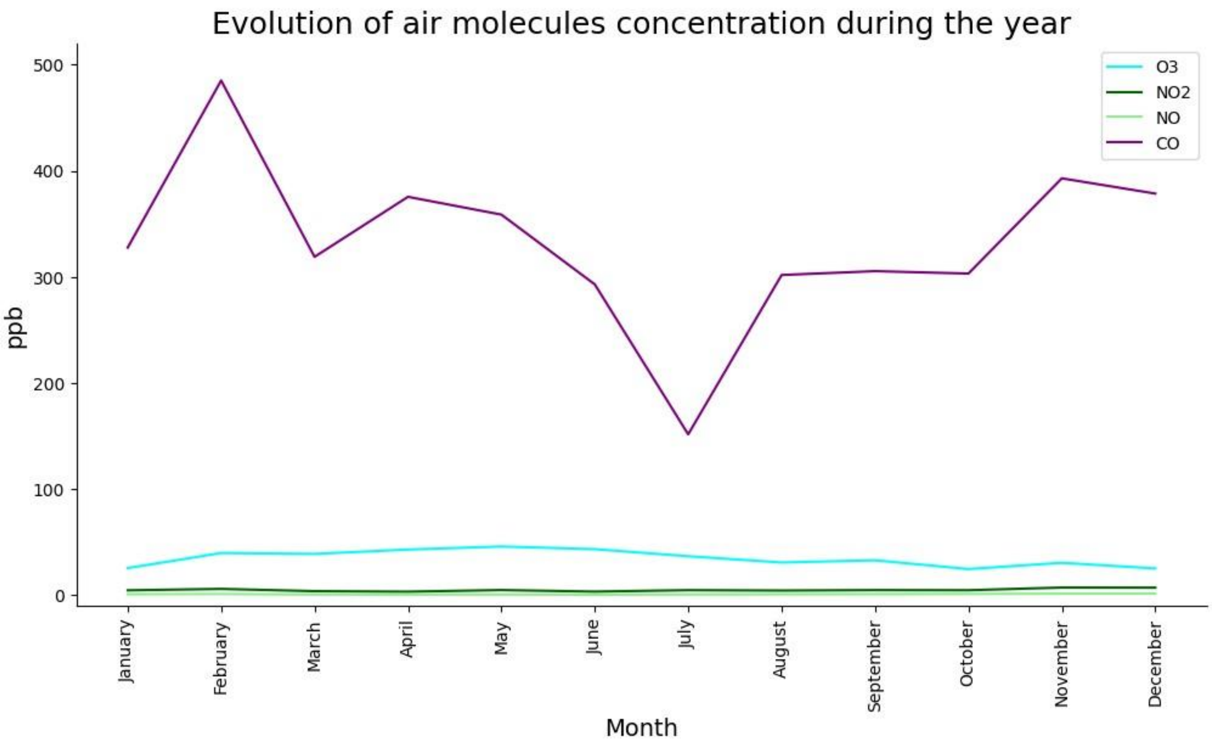
6. Similarly, with the relative humidity (RH) variable, it was rounded to its respective percentage to create a histogram of humidity frequencies.



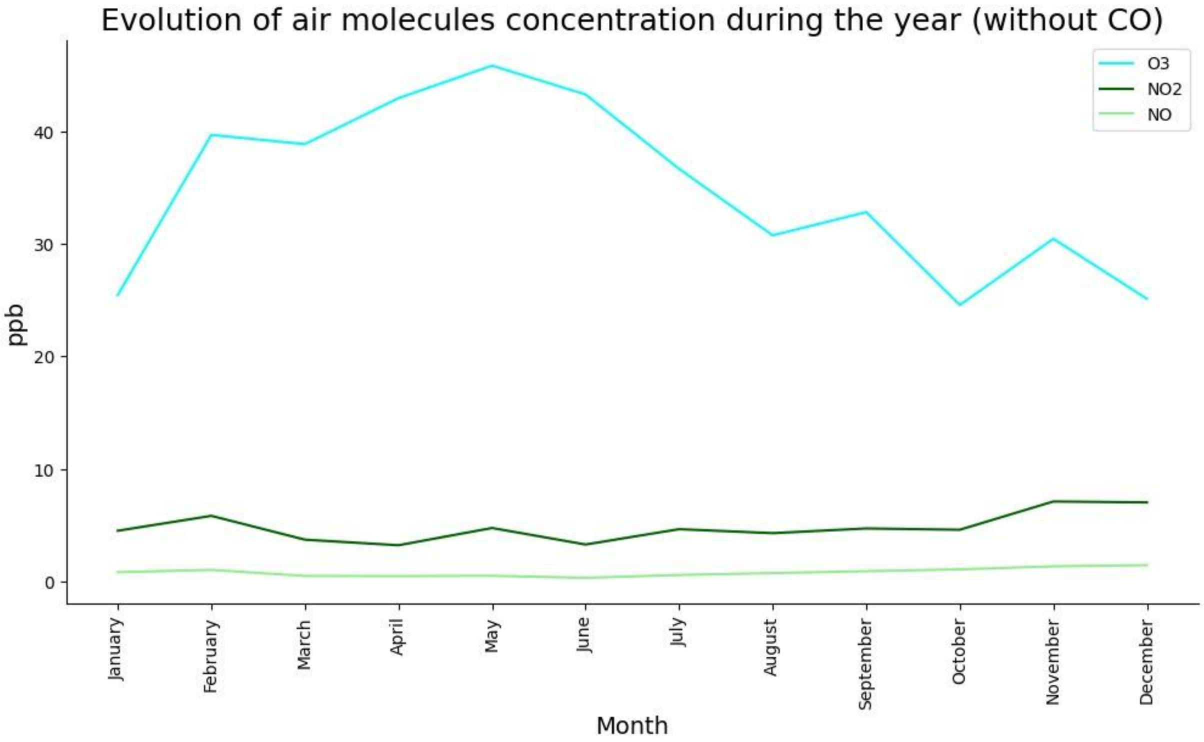
7. Moving on to the air quality variables, they were grouped by their respective month. This was done to better visualize the data throughout the year. Let's remember that due to the units, CO had to be converted to parts per billion to match the other molecules. Additionally, PM10 and PM2.5 were graphed separately due to their units.

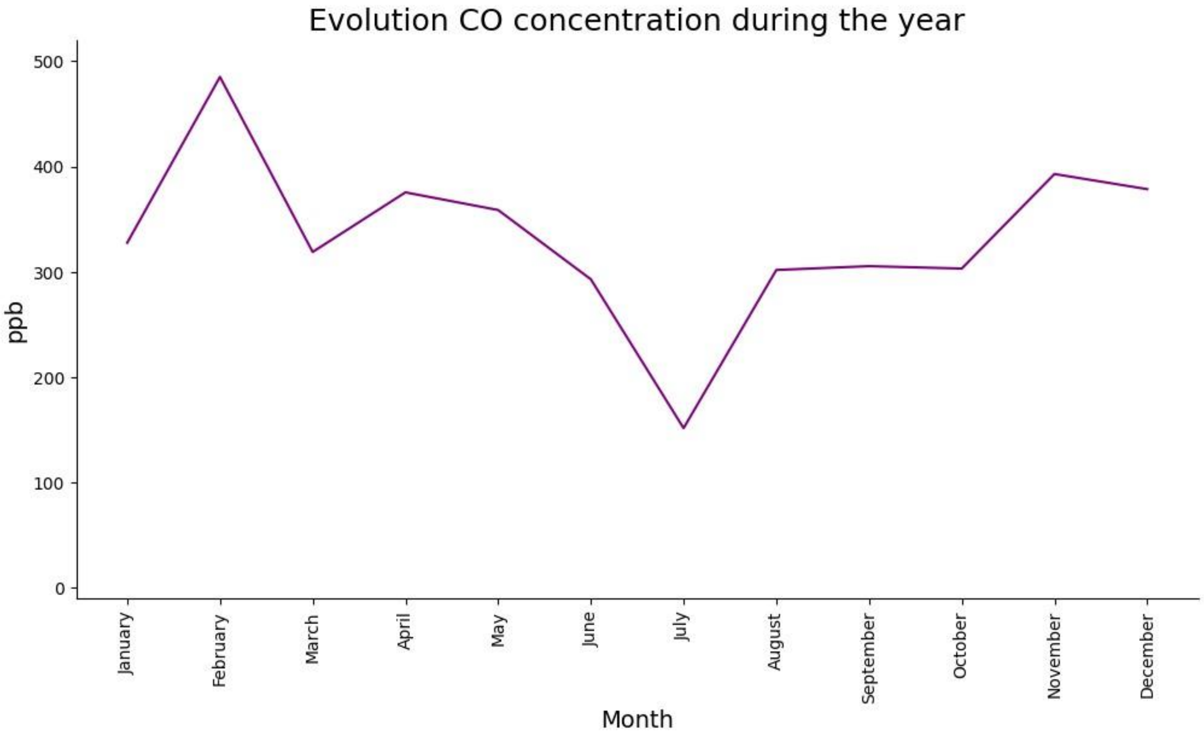


Now, with the molecules:

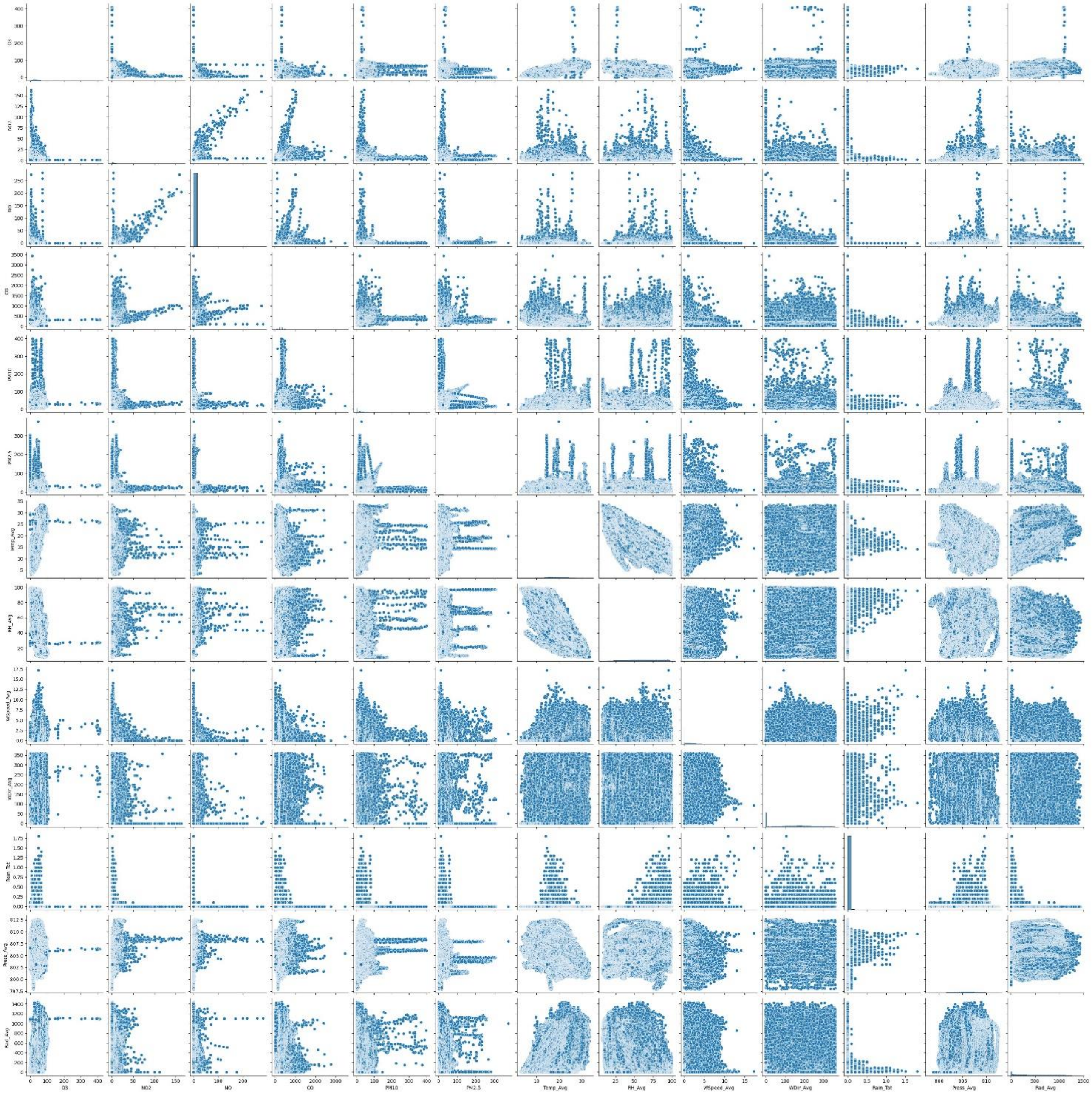


By separating CO into a separate graph, we can improve the visualization scale for the other variables, given the quantities of CO.:



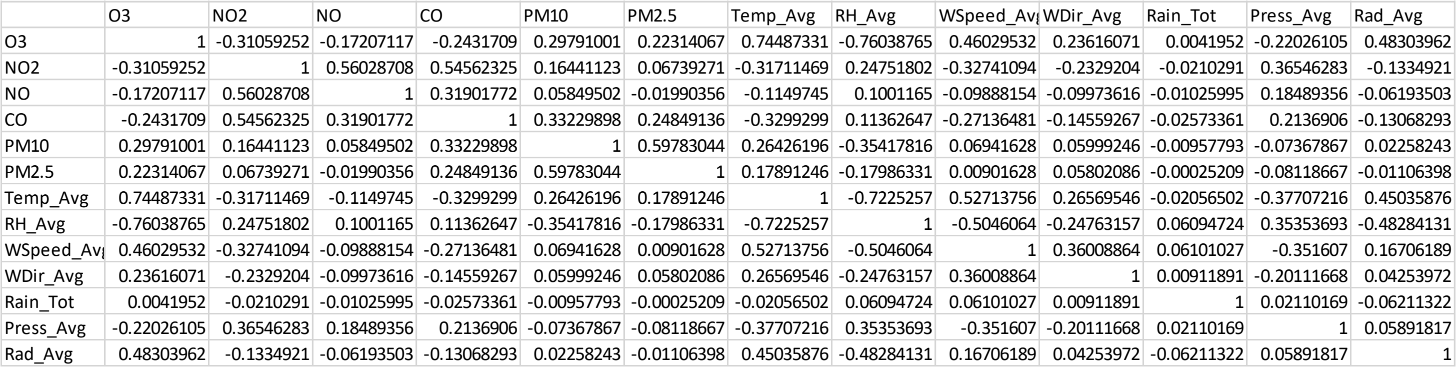


8. Continuing with the analysis, the goal was to continue creating scatter plots to identify possible correlations. Therefore, only the following variables were selected: O3, NO2, NO, CO, PM10, PM2.5, Temp_Avg, RH_Avg, WSpeed_Avg, WDir_Avg, Rain_Tot, Press_Avg, Rad_Avg.
   Thus, creating a scatter plot of all variables against each other resulted in the following:



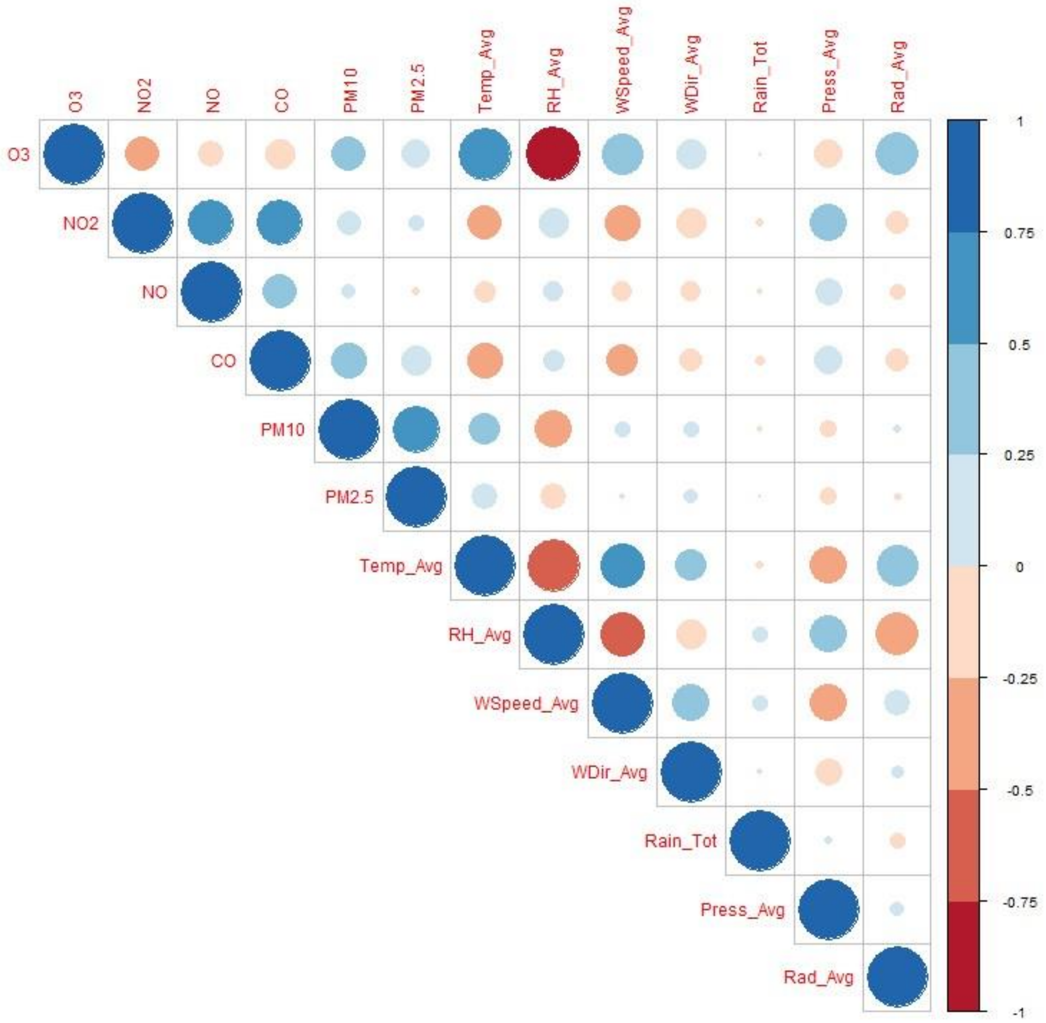
9. Based on the previous result, before considering creating regression plots and models, it was better to analyze the correlations.

First, we calculated the Pearson correlation coefficient for each pair from the previous scatter plot:

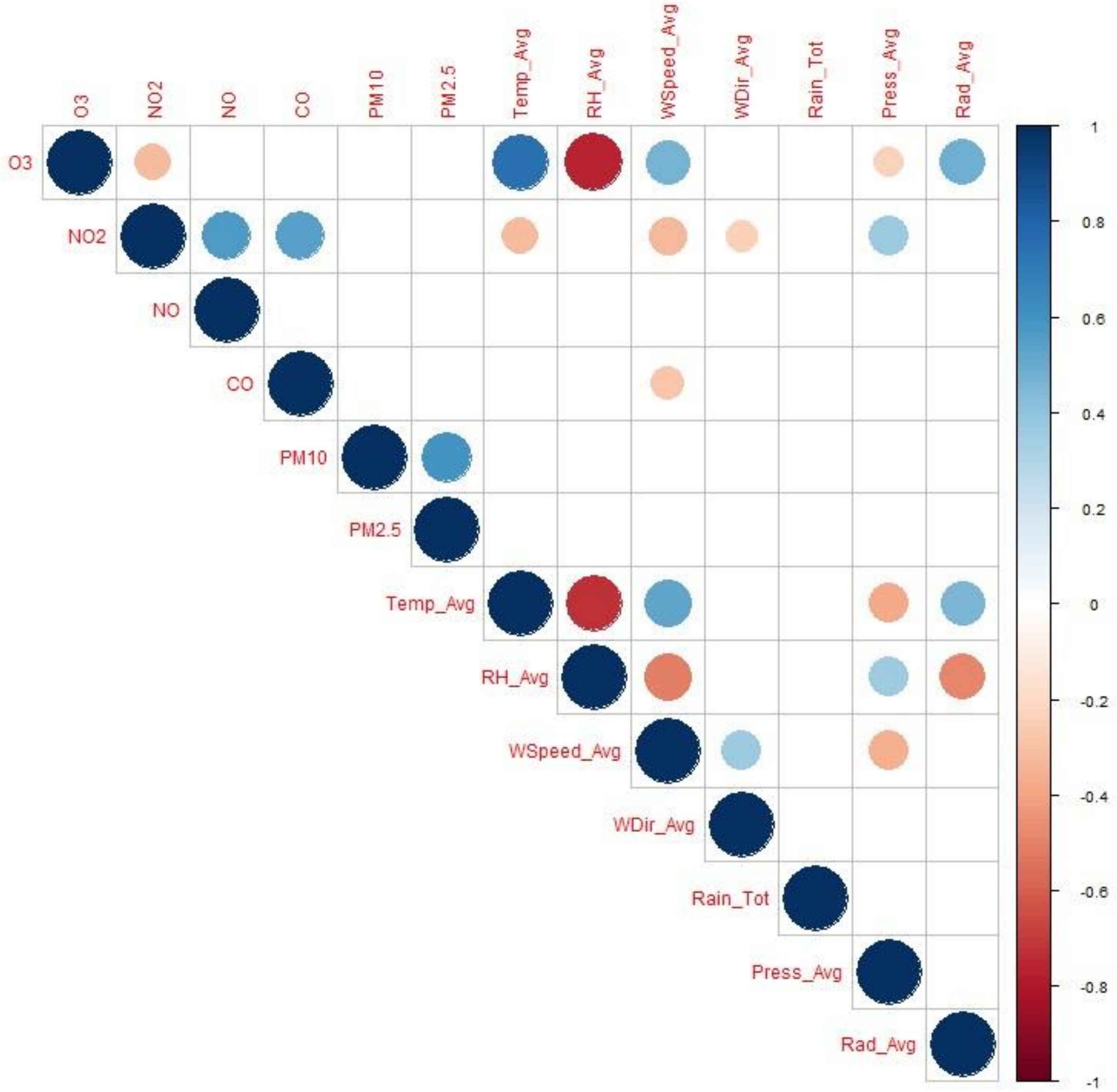


	O3	NO2	NO	CO	PM10	PM2.5	Temp_Avg	RH_Avg	WSpeed_Av	WDir_Avg	Rain_Tot	Press_Avg	Rad_Avg
O3	1	-0.31059252	-0.17207117	-0.2431709	0.29791001	0.22314067	0.74487331	-0.76038765	0.46029532	0.23616071	0.0041952	-0.22026105	0.48303962
NO2	-0.31059252	1	0.56028708	0.54562325	0.16441123	0.06739271	-0.31711469	0.24751802	-0.32741094	-0.2329204	-0.0210291	0.36546283	-0.1334921
NO	-0.17207117	0.56028708	1	0.31901772	0.05849502	-0.01990356	-0.1149745	0.1001165	-0.09888154	-0.09973616	-0.01025995	0.18489356	-0.06193503
CO	-0.2431709	0.54562325	0.31901772	1	0.33229898	0.24849136	-0.3299299	0.11362647	-0.27136481	-0.14559267	-0.02573361	0.2136906	-0.13068293
PM10	0.29791001	0.16441123	0.05849502	0.33229898	1	0.59783044	0.26426196	-0.35417816	0.06941628	0.05999246	-0.00957793	-0.07367867	0.02258243
PM2.5	0.22314067	0.06739271	-0.01990356	0.24849136	0.59783044	1	0.17891246	-0.17986331	0.00901628	0.05802086	-0.00025209	-0.08118667	-0.01106398
Temp_Avg	0.74487331	-0.31711469	-0.1149745	-0.3299299	0.26426196	0.17891246	1	-0.7225257	0.52713756	0.26569546	-0.02056502	-0.37707216	0.45035876
RH_Avg	-0.76038765	0.24751802	0.1001165	0.11362647	-0.35417816	-0.17986331	-0.7225257	1	-0.5046064	-0.24763157	0.06094724	0.35353693	-0.48284131
WSpeed_Av	0.46029532	-0.32741094	-0.09888154	-0.27136481	0.06941628	0.00901628	0.52713756	-0.5046064	1	0.36008864	0.06101027	-0.351607	0.16706189
WDir_Avg	0.23616071	-0.2329204	-0.09973616	-0.14559267	0.05999246	0.05802086	0.26569546	-0.24763157	0.36008864	1	0.00911891	-0.20111668	0.04253972
Rain_Tot	0.0041952	-0.0210291	-0.01025995	-0.02573361	-0.00957793	-0.00025209	-0.02056502	0.06094724	0.06101027	0.00911891	1	0.02110169	-0.06211322
Press_Avg	-0.22026105	0.36546283	0.18489356	0.2136906	-0.07367867	-0.08118667	-0.37707216	0.35353693	-0.351607	-0.20111668	0.02110169	1	0.05891817
Rad_Avg	0.48303962	-0.1334921	-0.06193503	-0.13068293	0.02258243	-0.01106398	0.45035876	-0.48284131	0.16706189	0.04253972	-0.06211322	0.05891817	1

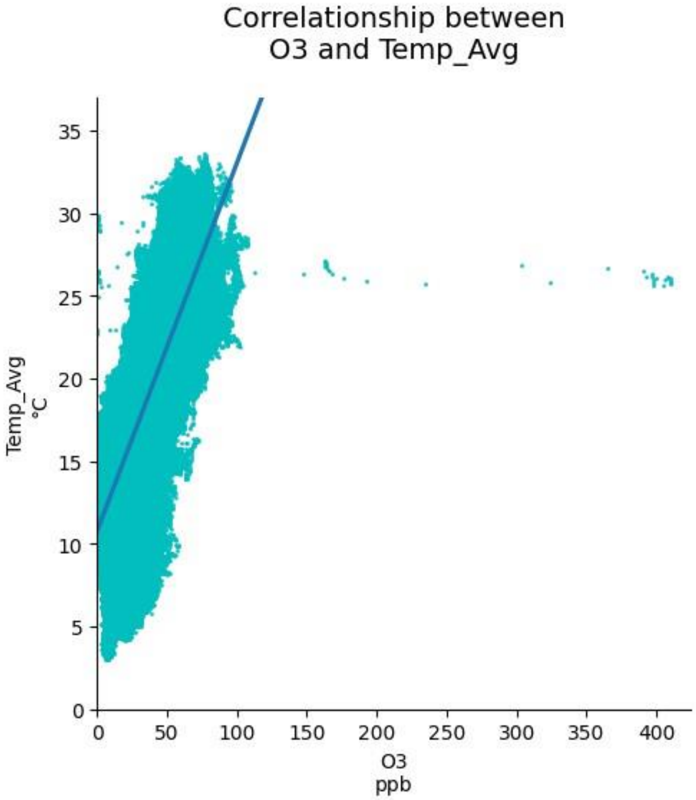
With these values, the following correlogram was obtained:

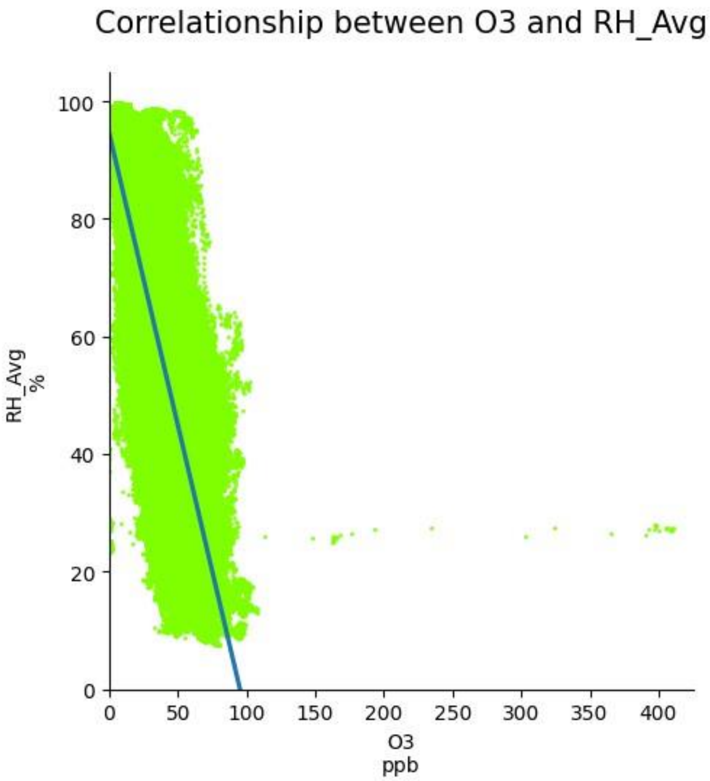


However, to obtain more significant results, we calculated the correlations that are statistically significant at a confidence level of 99%.



10. Once these more significant relationships were obtained, we chose to plot those with a higher coefficient of determination (r^2): O3 vs Temp_Avg and O3 vs RH_Avg. In addition to the scatter plot, we will draw a regression line that better fits the data:





11. After all these procedures, it is necessary to test the normality of the data before making inferences. Therefore, the same variables used in the correlogram are subject to normality testing. (O3, NO2, NO, CO, PM10, PM2.5, Temp_Avg, RH_Avg, WSpeed_Avg, WDir_Avg, Rain_Tot, Press_Avg) Let's remember that we should use the Kolmogorov-Smirnov test due to the amount of data.

Variable	**P-value**
O3	2.2e-16
NO2	2.2e-16
NO	2.2e-16
CO	2.2e-16
PM10	2.2e-16
PM2.5	2.2e-16
Temp_Avg	2.2e-16
RH_Avg	2.2e-16
WSpeed_Avg	2.2e-16
WDir_Avg	2.2e-16
Rain_Tot	2.2e-16
Press_Avg	2.2e-16

All variables don’t have normality. So, it is needed to do boxcox transformations according to their lambdas

Variable	Box-Cox Transformation
O3	$\sqrt{x}$
NO2	log(x)
NO	$\frac{1}{x^2}$
CO	$\sqrt{x}$
PM10	$\sqrt{x}$
PM2.5	$\sqrt{x}$
Temp_Avg	x
RH_Avg	x
WSpeed_Avg	x
WDir_Avg	$\sqrt{x}$
Rain_Tot	$\frac{1}{x^2}$
Press_Avg	x^2

Doing the normality test again, but with the box cox transformation:

Variable	P-value
O3	2.2e-16
NO2	2.2e-16
NO	2.2e-16
CO	2.2e-16
PM10	2.2e-16
PM2.5	2.2e-16
Temp_Avg	2.2e-16
RH_Avg	2.2e-16
WSpeed_Avg	2.2e-16
WDir_Avg	2.2e-16
Rain_Tot	2.2e-16
Press_Avg	2.2e-16

Although the all the values for Dc decreased, it was not enough to achieve normality.

In this case what comes next are the non-parametric tests. But, all these results are very unusual. So, the approach here (instead of doing non-parametric test with low reliable data) is to focus on a bigger temporality. The idea behind this is that in low temporalities (as the 1 minute per sample) all the samples generate much noise.

That is the reason why the dataset of samples per day was created. So let's repeat the same process of descriptive analysis, but now for this new dataset.

12. First at all, now we just have the year, month, and day of each sample. Besides, from now on we will work just with these variables: O3, NO2, NO, CO, PM10, PM2.5, Temp_Avg, RH_Avg, WSpeed_Avg, WDir_Avg, Rain_Tot, Press_Avg, Rad_Avg. Remember that this new dataset have only 315 samples (50 days were not available).

First, the descriptive tables:

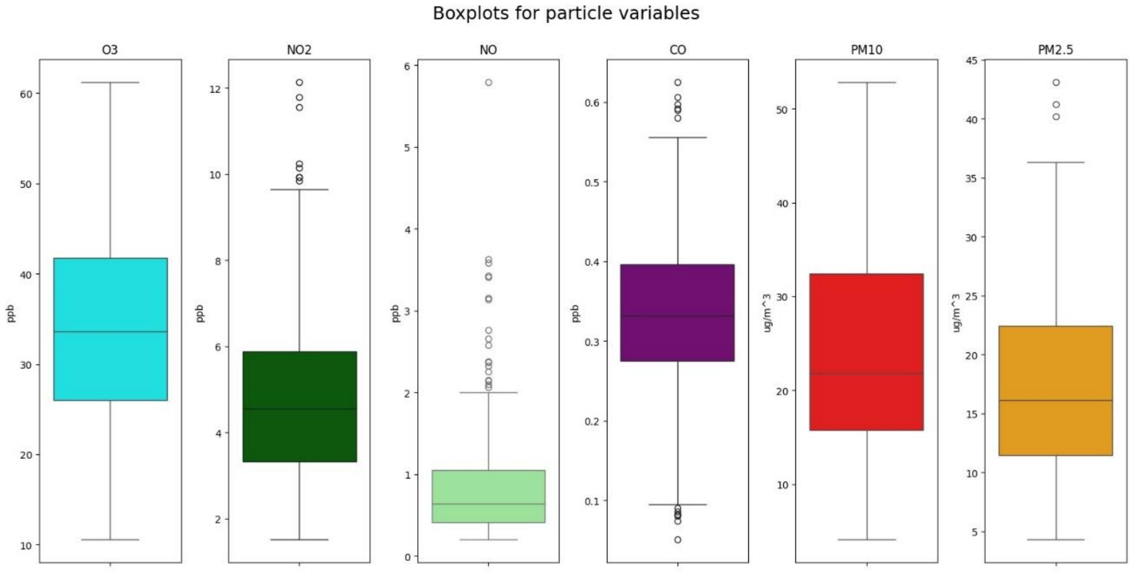
Air quality:

	O3	**NO2**	**NO**	**CO**	**PM10**	**PM2.5**
Mean	34.0474	4.8367	0.8381	0.3310	24.5149	17.4772
Median	33.6012	4.5474	0.6380	0.3313	21.8194	16.1388
Std	10.0502	2.0002	0.6780	0.1037	10.9295	7.8591
Min	10.5729	1.5085	0.2000	0.0507	4.0972	4.3052
25%	26.0263	3.3140	0.4069	0.2753	15.7859	11.4616
75%	41.7517	5.8749	1.0464	0.3962	32.4311	22.4167
Max	61.1739	12.1345	5.7879	0.6247	52.8176	43.0924
Fisher asymetry	0.07	0.91	2.68	-0.11	0.46	0.69

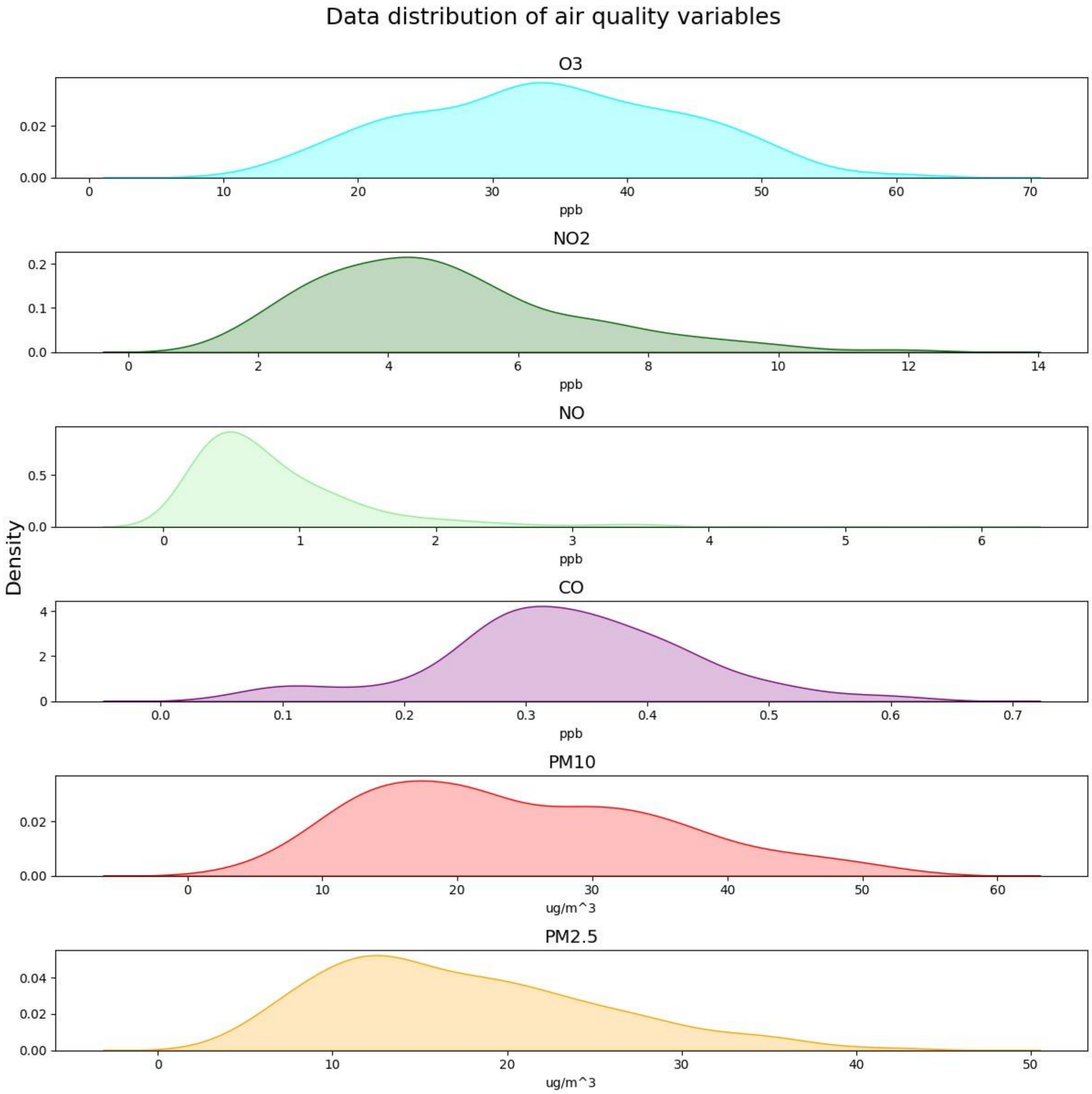
Meteorological variables:

	Temp_Avg	**RH_Avg**	**WSpeed_Avg**	**WDir_Avg**	**Press_Avg**	**Rad_Avg**	**Rain_Tot**
Mean	18.2947	60.7989	1.5489	144.6036	805.8282	468.1455	1.4555
Median	18.5791	64.6393	1.4913	146.5271	805.7542	483.8610	0.00
Std	2.9760	16.0522	0.5521	31.6714	1.4189	121.9885	4.1247
Min	10.0623	23.3850	0.3424	63.3813	800.7429	0.0460	0.00
25%	16.3998	46.3544	1.1636	124.4876	804.9208	392.4523	0.00
75%	19.9609	74.1703	1.9225	165.3201	806.7331	567.4862	0.3000
Max	24.6258	92.7811	3.4809	272.0374	810.3584	660.3957	41.7000
Fisher asymetry	-0.11	-0.38	0.64	0.16	0.05	-0.75	5.02

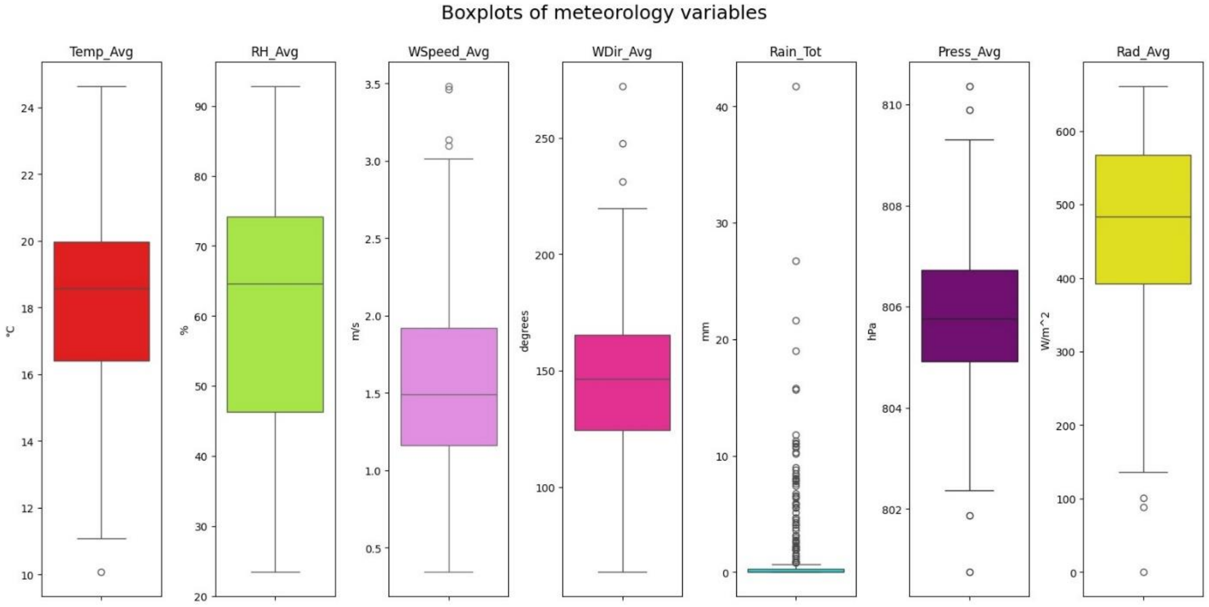
Box plots of air quality variables:



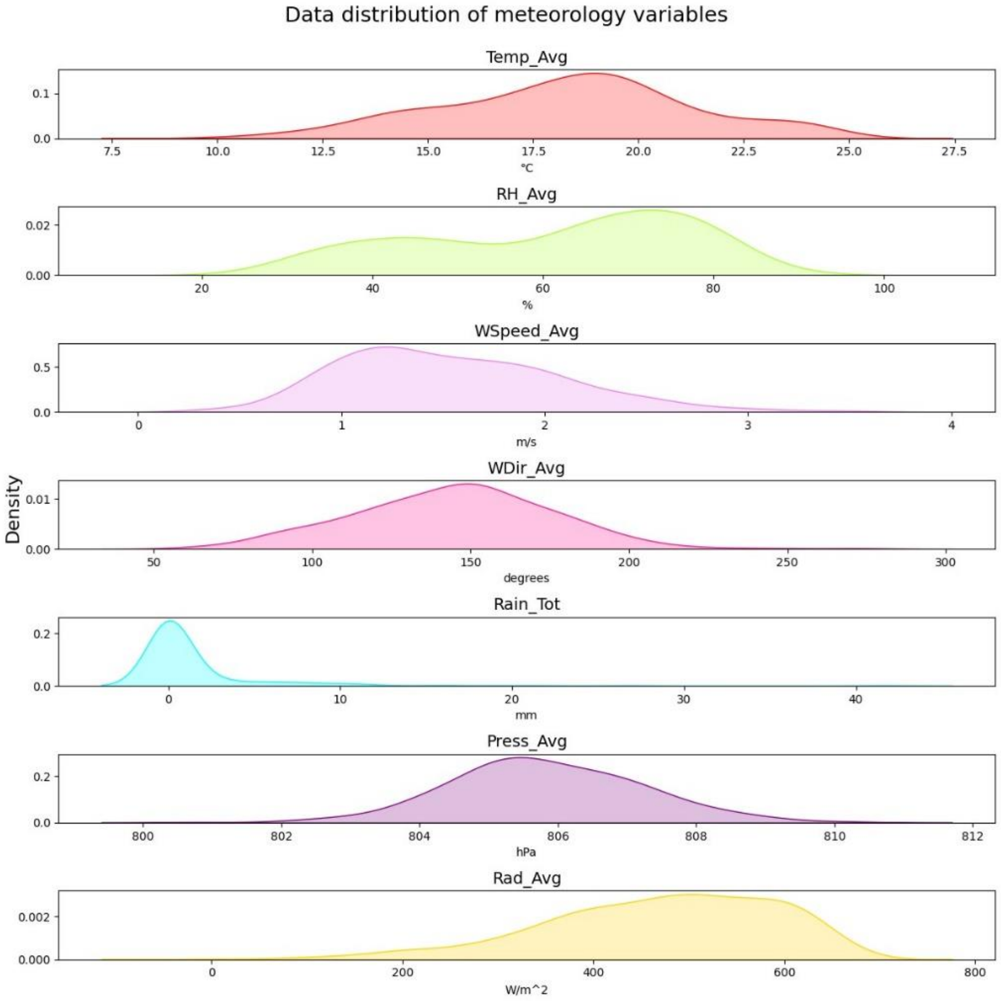
Density distribution plots:



Box plots of meteorological variables:



Density distribution plots:



13. After describing these new data, it's important to subject them to a normality test. This is especially crucial since it's precisely the reason we chose to work with this higher temporal resolution. Kolmogorov Smirnov test will be used again because 315 samples are still too many for Shapiro test

Variable	**P-value**
O3	0.60
NO2	0.009
NO	1.21e-8
CO	0.11
PM10	0.0041
PM2.5	0.0061
Temp_Avg	0.59
RH_Avg	0.001
WSpeed_Avg	0.04
WDir_Avg	0.82
Rain_Tot	2.2e-16
Press_Avg	0.83
Rad_Avg	0.13

NO2, NO, PM10, PM2.5, RH_Avg, WSpeed_Avg, and Rain_Tot need to have a box cox tranformacion

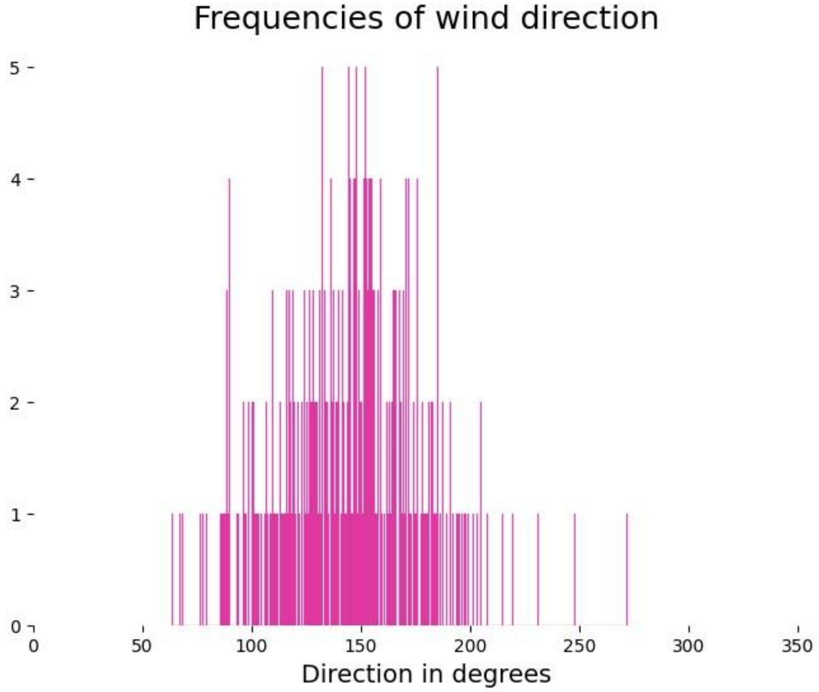
Variable	**Box-Cox Transformation**
NO2	log(x)
NO	log(x)
PM10	$\sqrt{x}$
PM2.5	log(x)
RH_Avg	x^2
WSpeed_Avg	$\sqrt{x}$
Rain_Tot	$\sqrt{x}$

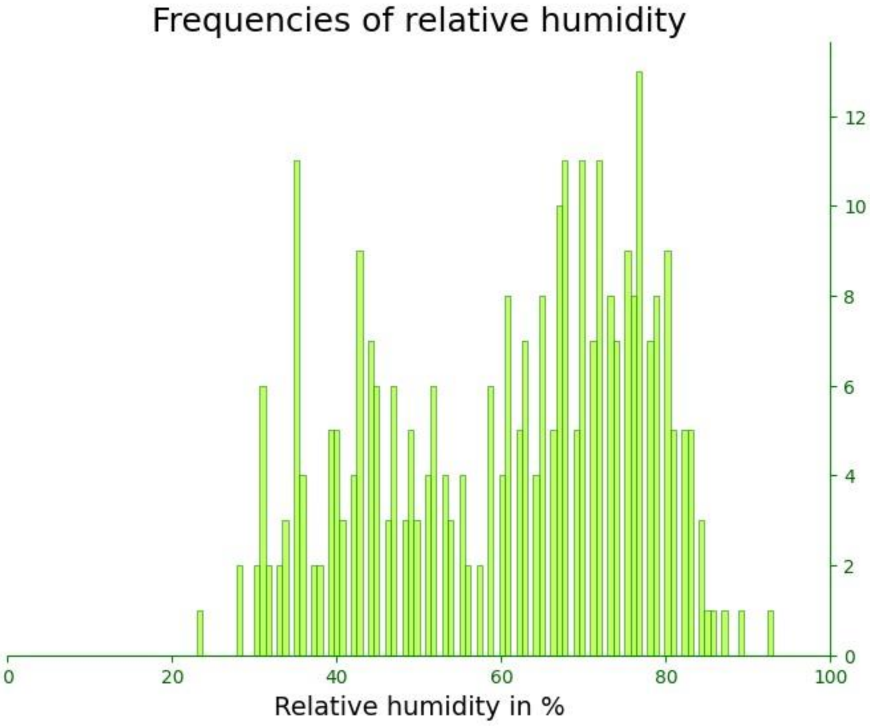
Doing again the normality test for the transformed variables.

Variable	**P-values**
NO2	0.84
NO	0.76
PM10	0.28
PM2.5	0.33
RH_Avg	0.009
WSpeed_Avg	0.28
Rain_Tot	2.2e-16

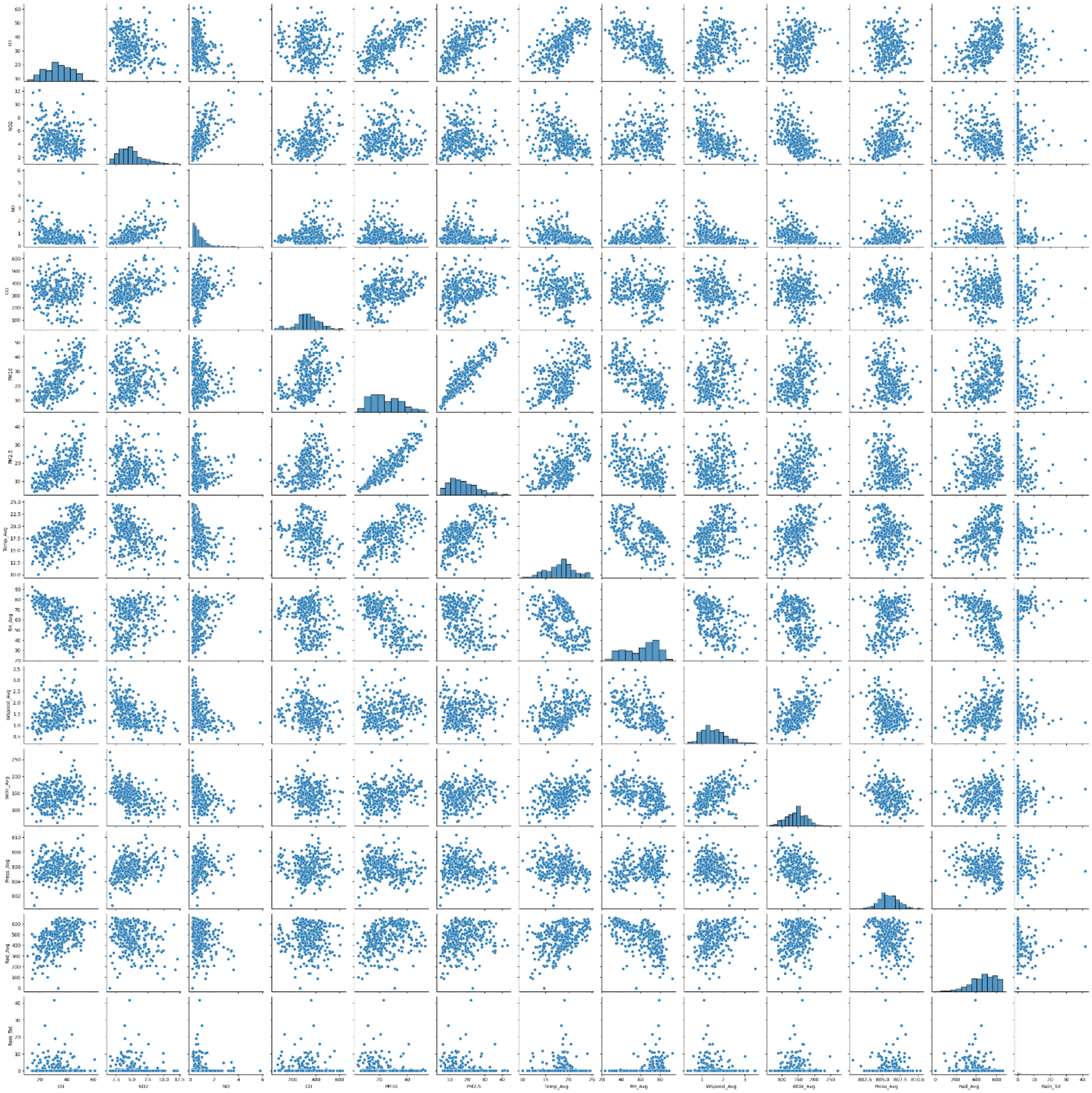
Therefore, the only variables that did not achieve normality were: RH_Avg and Rain_Tot. About Rain_Tot it was impossible: it has too many outliers. However, RH_Avg can be explained due to its bimodal distribution that we can see in the density plot.

14. At this point, to demonstrate how the data changes with temporal resolution, histograms of the variables WDir_Avg and RH_Avg are presented. These histograms were created in previous points. Additionally, for the RH_Avg variable, the bimodality that led to non-normality is also shown.

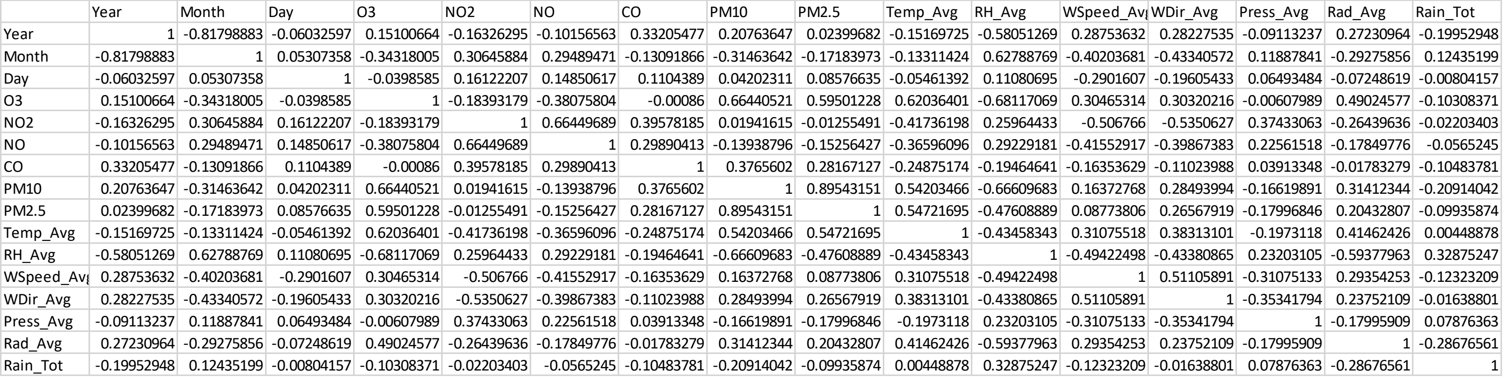




15. Once data normality has been tested, we can proceed with the same correlation procedure using this dataset.
    It stars with the dispersion plot of all the pairs of the variables:

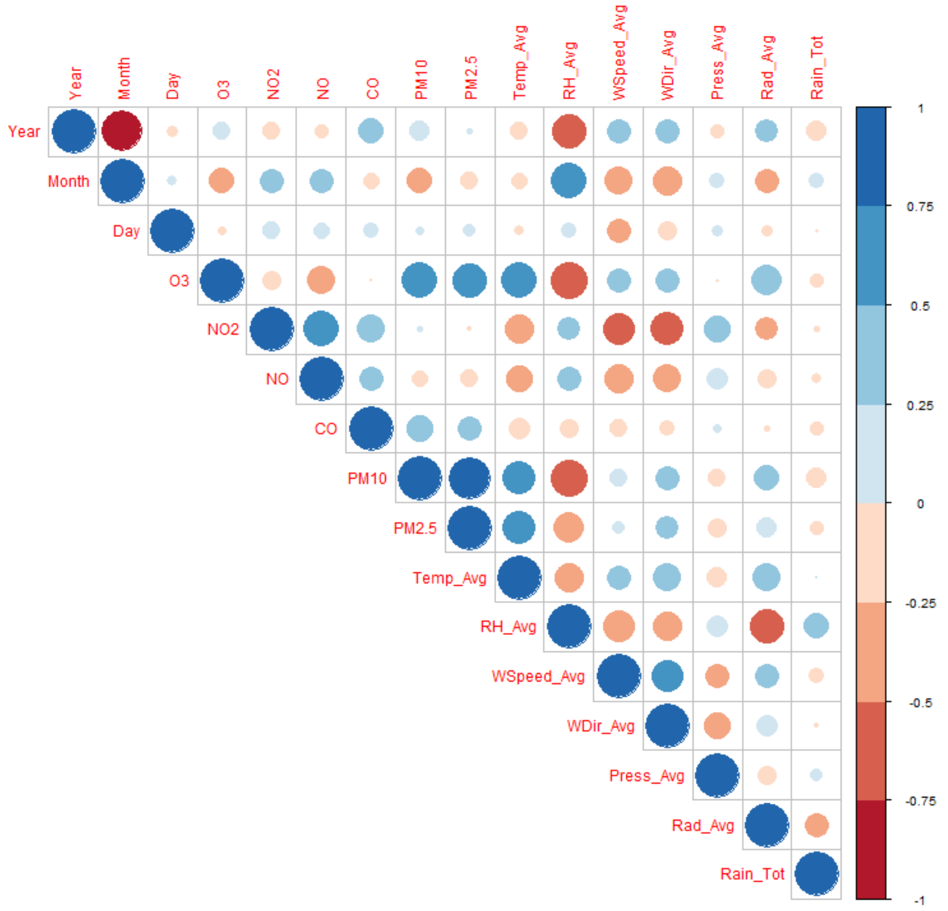


Moving forward, we present the Pearson correlation coefficients:

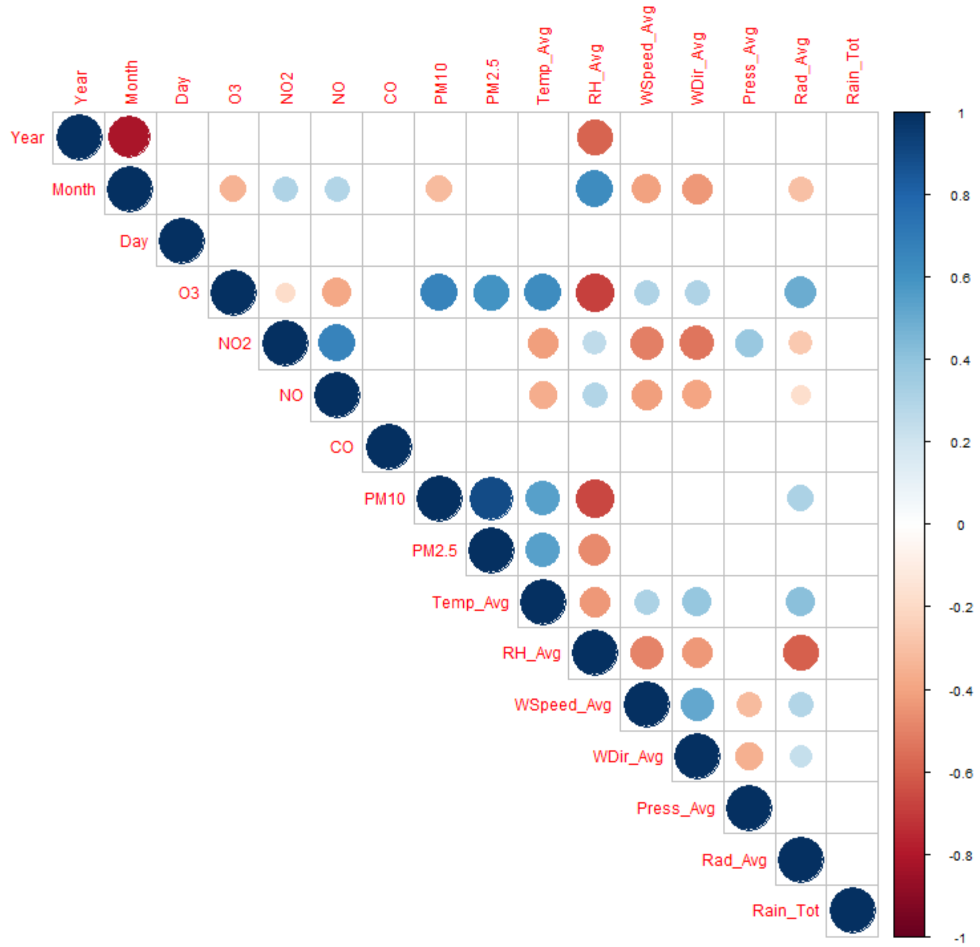


	Year	Month	Day	O3	NO2	NO	CO	PM10	PM2.5	Temp_Avg	RH_Avg	WSpeed_Av	WDir_Avg	Press_Avg	Rad_Avg	Rain_Tot
Year	1	-0.81798883	-0.06032597	0.15100664	-0.16326295	-0.10156563	0.33205477	0.20763647	0.02399682	-0.15169725	-0.58051269	0.28753632	0.28227535	-0.09113237	0.27230964	-0.19952948
Month	-0.81798883	1	0.05307358	-0.34318005	0.30645884	0.29489471	-0.13091866	-0.31463642	-0.17183973	-0.13311424	0.62788769	-0.40203681	-0.43340572	0.11887841	-0.29275856	0.12435199
Day	-0.06032597	0.05307358	1	-0.0398585	0.16122207	0.14850617	0.1104389	0.04202311	0.08576635	-0.05461392	0.11080695	-0.2901607	-0.19605433	0.06493484	-0.07248619	-0.00804157
O3	0.15100664	-0.34318005	-0.0398585	1	-0.18393179	-0.38075804	-0.00086	0.66440521	0.59501228	0.62036401	-0.68117069	0.30465314	0.30320216	-0.00607989	0.49024577	-0.10308371
NO2	-0.16326295	0.30645884	0.16122207	-0.18393179	1	0.66449689	0.39578185	0.01941615	-0.01255491	-0.41736198	0.25964433	-0.506766	-0.5350627	0.37433063	-0.26439636	-0.02203403
NO	-0.10156563	0.29489471	0.14850617	-0.38075804	0.66449689	1	0.29890413	-0.13938796	-0.15256427	-0.36596096	0.29229181	-0.41552917	-0.39867383	0.22561518	-0.17849776	-0.0565245
CO	0.33205477	-0.13091866	0.1104389	-0.00086	0.39578185	0.29890413	1	0.3765602	0.28167127	-0.24875174	-0.19464641	-0.16353629	-0.11023988	0.03913348	-0.01783279	-0.10483781
PM10	0.20763647	-0.31463642	0.04202311	0.66440521	0.01941615	-0.13938796	0.3765602	1	0.89543151	0.54203466	-0.66609683	0.16372768	0.28493994	-0.16619891	0.31412344	-0.20914042
PM2.5	0.02399682	-0.17183973	0.08576635	0.59501228	-0.01255491	-0.15256427	0.28167127	0.89543151	1	0.54721695	-0.47608889	0.08773806	0.26567919	-0.17996846	0.20432807	-0.09935874
Temp_Avg	-0.15169725	-0.13311424	-0.05461392	0.62036401	-0.41736198	-0.36596096	-0.24875174	0.54203466	0.54721695	1	-0.43458343	0.31075518	0.38313101	-0.1973118	0.41462426	0.00448878
RH_Avg	-0.58051269	0.62788769	0.11080695	-0.68117069	0.25964433	0.29229181	-0.19464641	-0.66609683	-0.47608889	-0.43458343	1	-0.49422498	-0.43380865	0.23203105	-0.59377963	0.32875247
WSpeed_Av	0.28753632	-0.40203681	-0.2901607	0.30465314	-0.506766	-0.41552917	-0.16353629	0.16372768	0.08773806	0.31075518	-0.49422498	1	0.51105891	-0.31075133	0.29354253	-0.12323209
WDir_Avg	0.28227535	-0.43340572	-0.19605433	0.30320216	-0.5350627	-0.39867383	-0.11023988	0.28493994	0.26567919	0.38313101	-0.43380865	0.51105891	1	-0.35341794	0.23752109	-0.01638801
Press_Avg	-0.09113237	0.11887841	0.06493484	-0.00607989	0.37433063	0.22561518	0.03913348	-0.16619891	-0.17996846	-0.1973118	0.23203105	-0.31075133	-0.35341794	1	-0.17995909	0.07876363
Rad_Avg	0.27230964	-0.29275856	-0.07248619	0.49024577	-0.26439636	-0.17849776	-0.01783279	0.31412344	0.20432807	0.41462426	-0.59377963	0.29354253	0.23752109	-0.17995909	1	-0.28676561
Rain_Tot	-0.19952948	0.12435199	-0.00804157	-0.10308371	-0.02203403	-0.0565245	-0.10483781	-0.20914042	-0.09935874	0.00448878	0.32875247	-0.12323209	-0.01638801	0.07876363	-0.28676561	1

Plotting the correlogram:

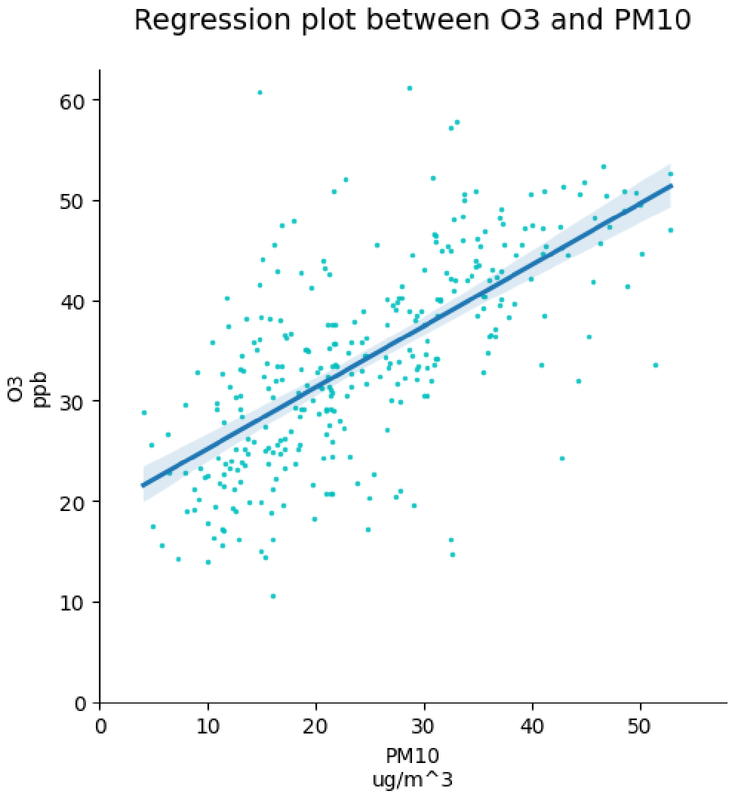


Due to it has too many correlations, it’s necessary to create the correlogram with the most significant correlations (with 99% of confidence).

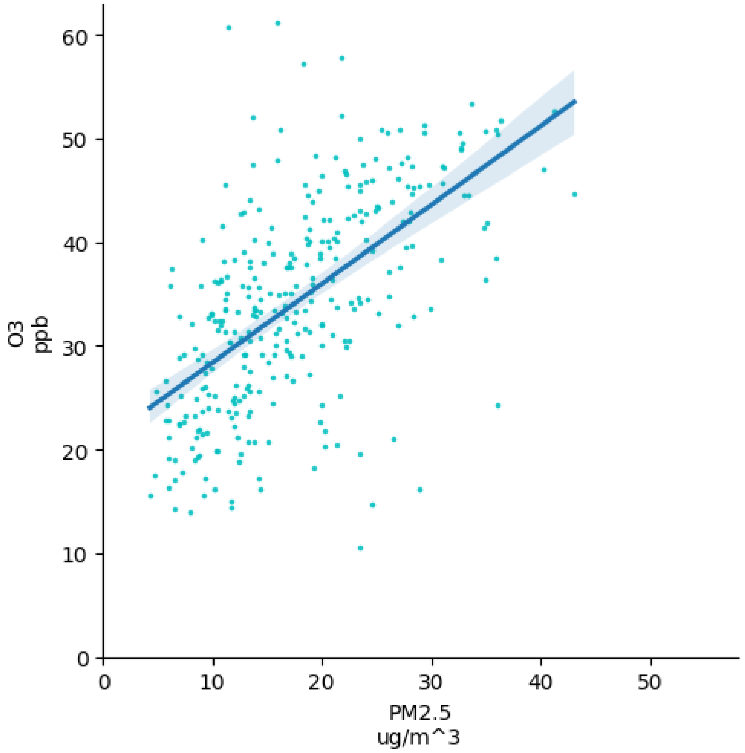


16. Based on the last correlogram, we will create scatter plots for our regression models with a confidence level of 95% for the following pairs of variables (especially because they are the most important for our analysis).
    - O3 – PM10
    - O3 – PM2.5
    - O3 – Temp_Avg
    - O3 – RH_Avg *strong negative correlation
    - O3 – Rad_Avg
    - NO2 – NO
    - NO2 – WSpeed_Avg
    - NO2 – Wdir_Avg
    - NO – WSpeed_Avg
    - NO – Wdir_Avg
    - PM10 – PM2.5 *strong positive correlation
    - PM10 – Temp_Avg
    - PM10 – RH_Avg *strong negative correlation

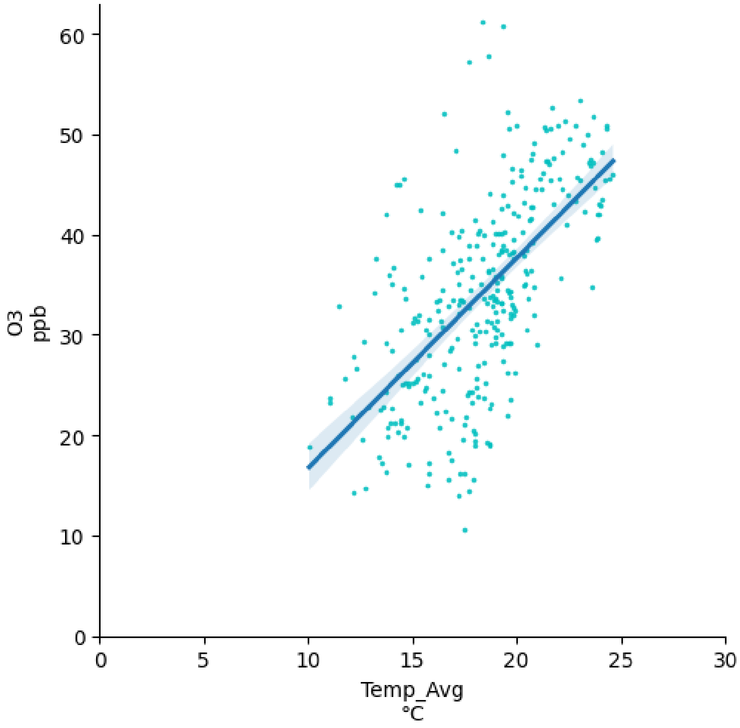
- PM2.5 – Temp_Avg
- PM2.5 – RH_Avg

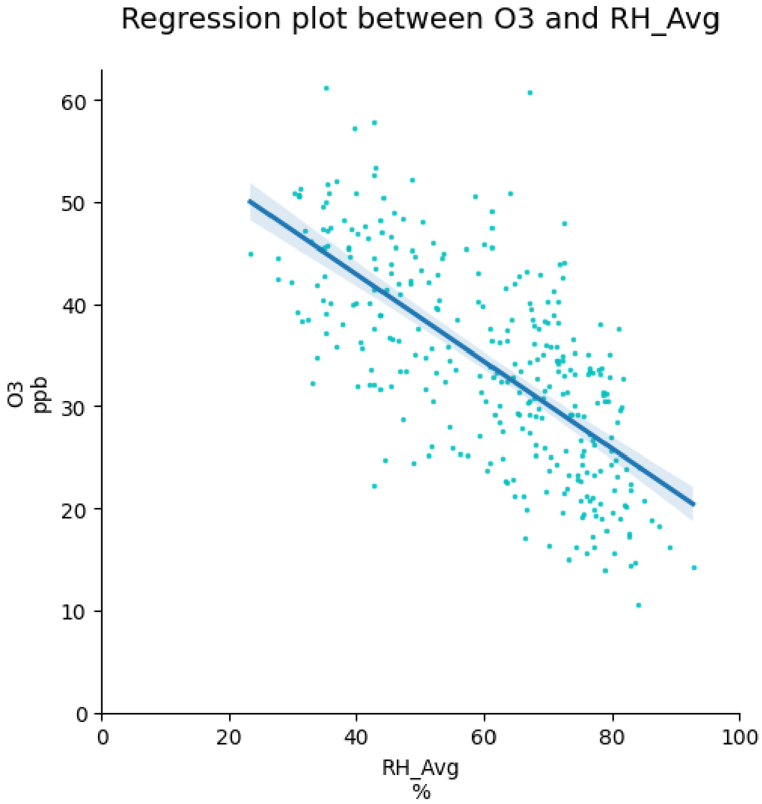


Regression plot between O3 and PM2.5

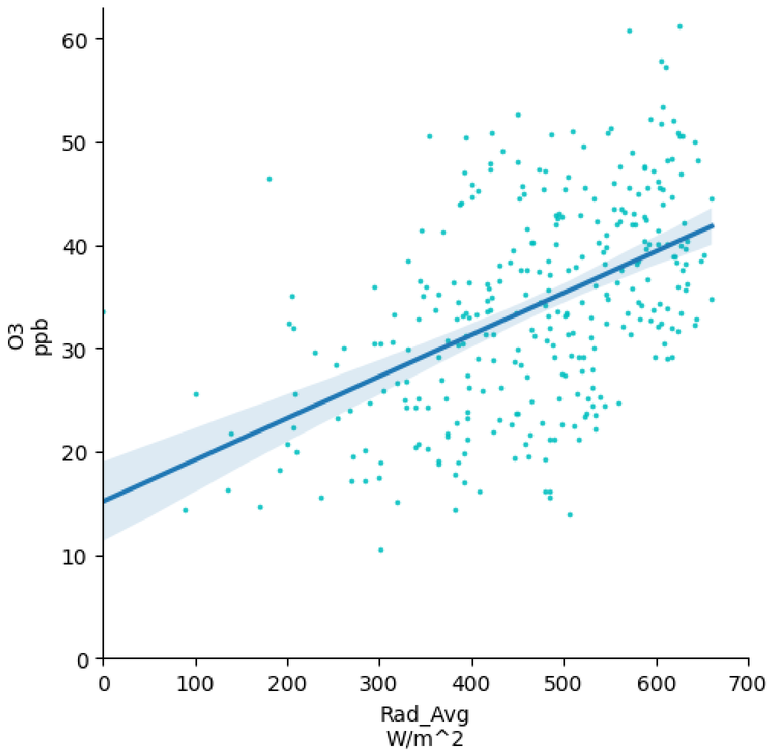


Regression plot between O3 and Temp_Avg

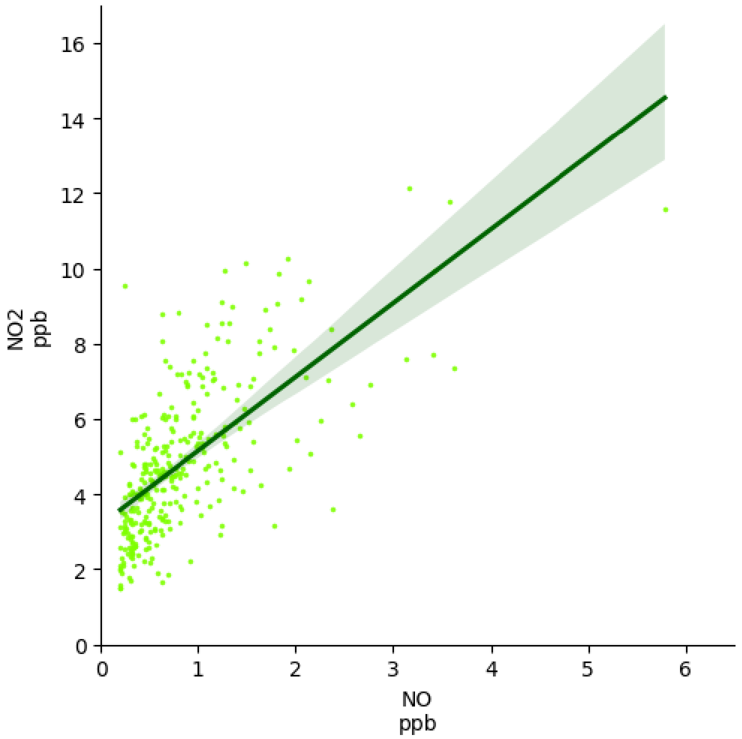


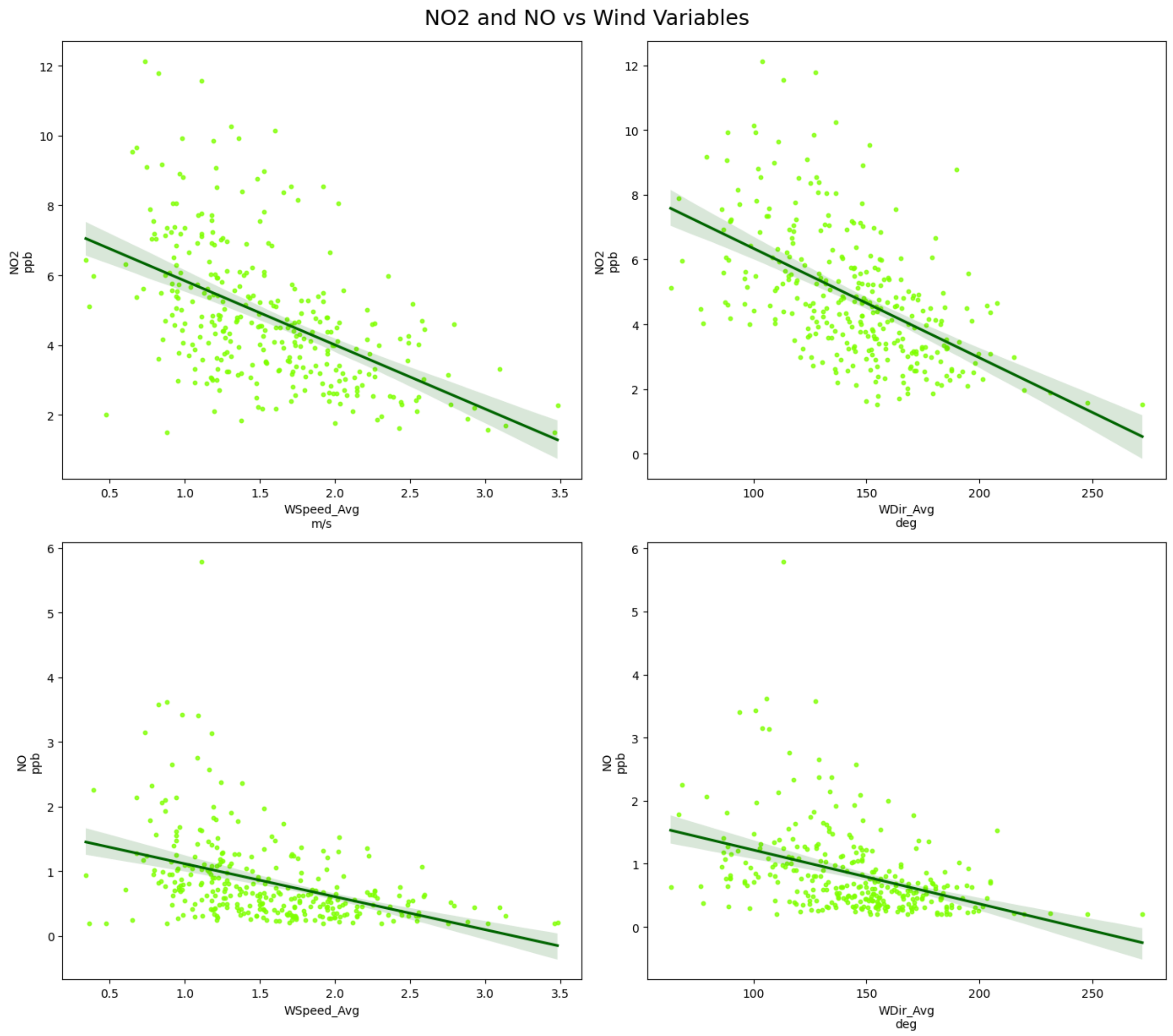

Regression plot between O3 and RH_Avg
O3
ppb
0
10
20
30
40
50
60
0
20
40
60
80
100
RH_Avg
%

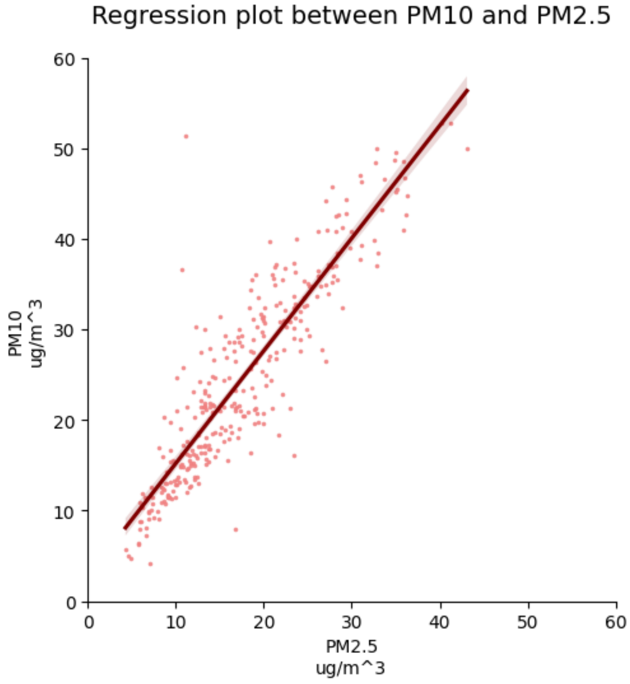
Regression plot between O3 and Rad_Avg

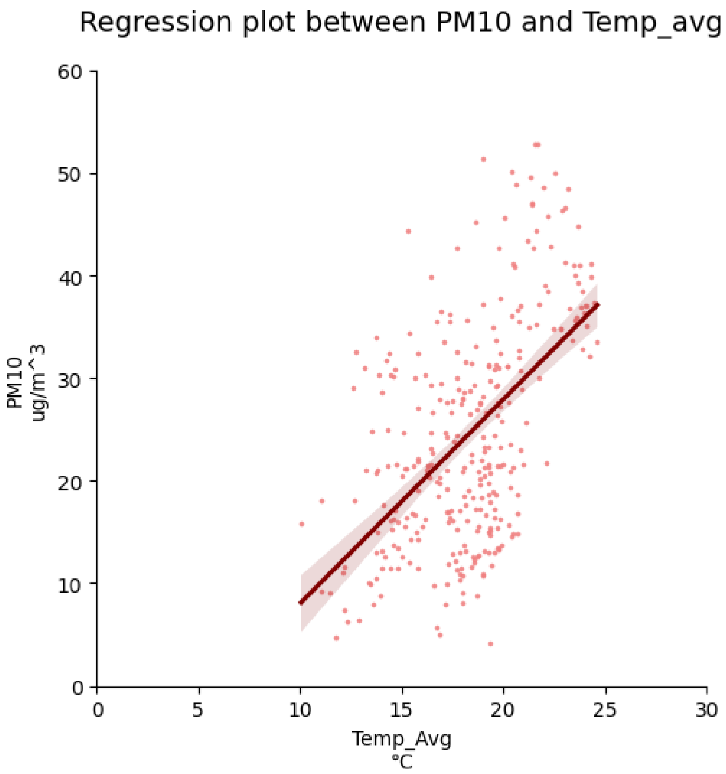


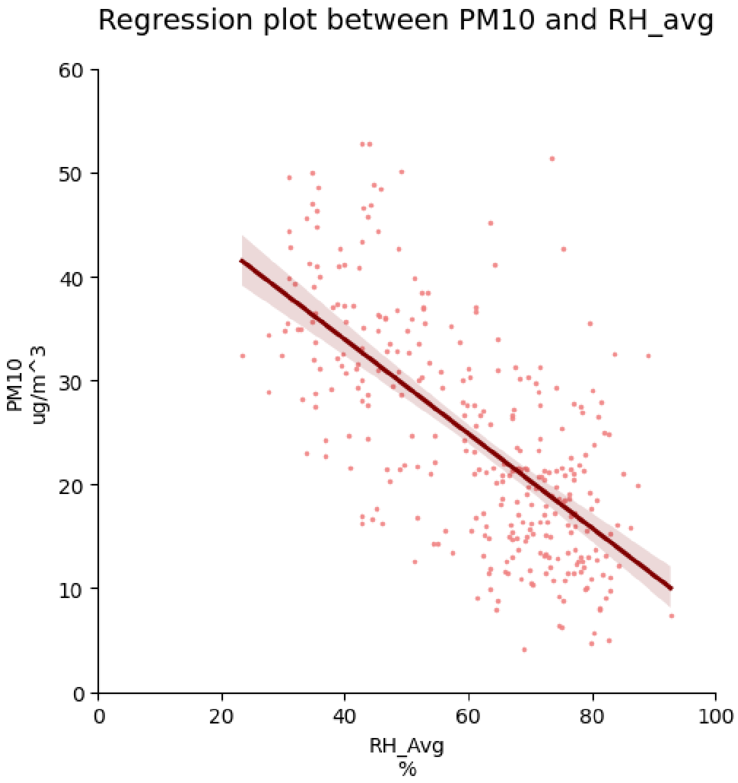
Regression plot between NO2 and NO



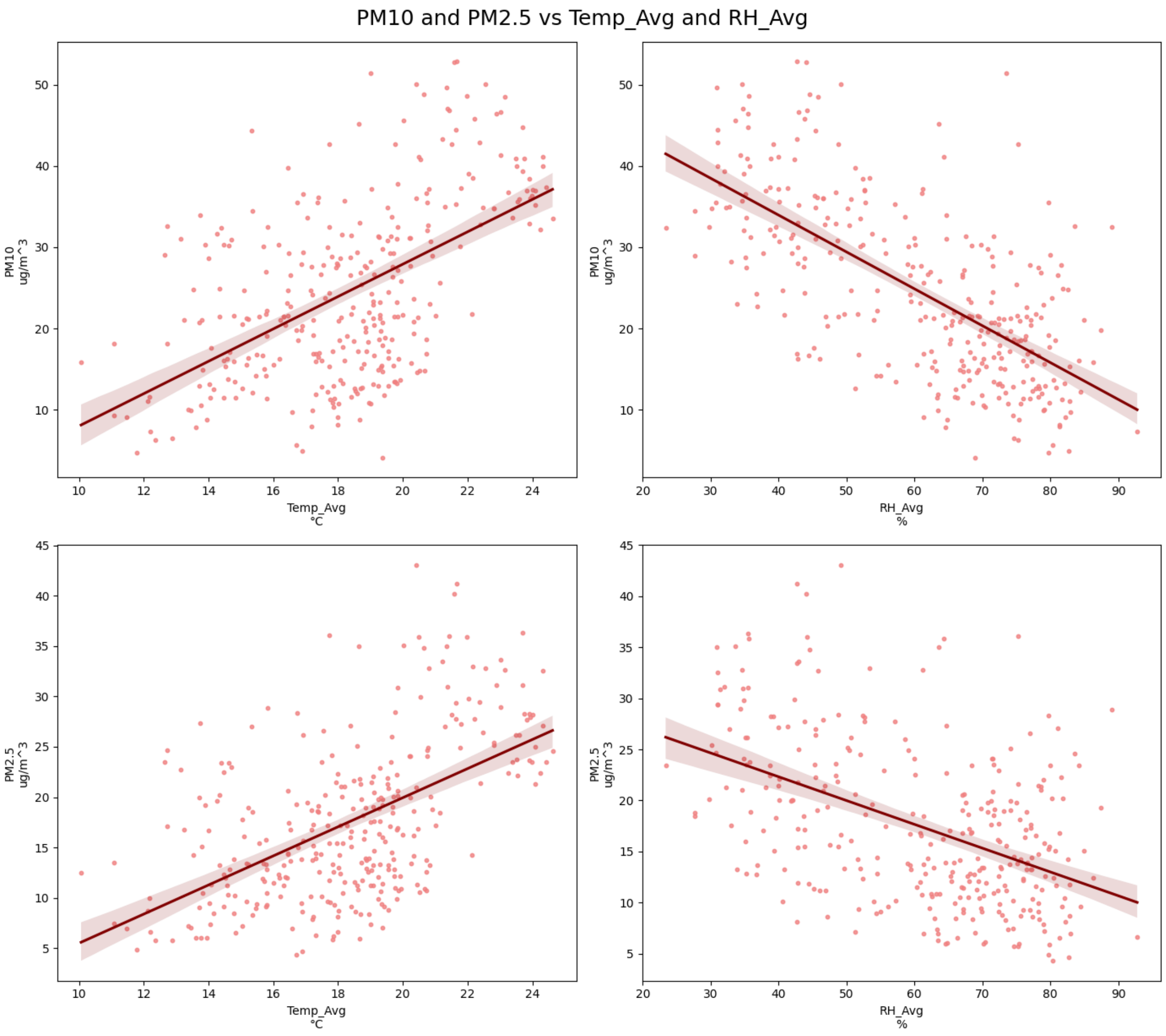

NO2 and NO vs Wind Variables


Regression plot between PM10 and PM2.5


Regression plot between PM10 and Temp_avg
PM10
ug/m^3
0
10
20
30
40
50
60
0
5
10
15
20
25
30
Temp_Avg
°C


Regression plot between PM10 and RH_avg
PM10
ug/m^3
0
10
20
30
40
50
60
0
20
40
60
80
100
RH_Avg
%

PM10 and PM2.5 vs Temp_Avg and RH_Avg

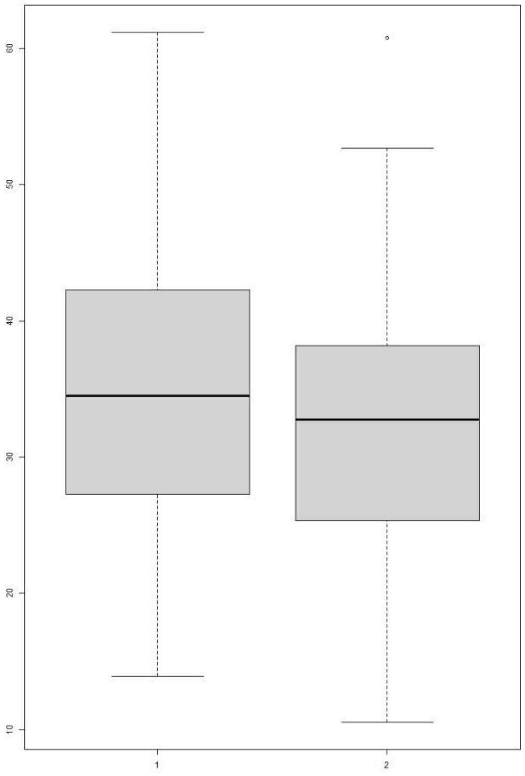


Let's now move on to our final experiments: the inferences.

A. Ozone-Rain

Here, we will only work with the O3 variable. Rainfall will be our grouping variable: we have samples of ozone concentration for days without rain and samples of concentration for days with rain.

If we create box plots for both sets of samples (no rain left and rain right) we can see that there is a minimum difference. So, let’s do a t-test



a) Skewness: Both skewness are in a normal range. In fact, no rainfall samples have almost symetry distribution. skewness = 0.01

b) Degrees of freedom. No rain = 224. Rain = 91

c) Normality test: No rain – p value = 0.55. Rain – p value = 0.98. Both are normal

d) Variance homogeneity. Fc = 1.11 Ft = 1.35. Fc < Ft. H0 accepted. Variances have homogeneity

e) T-student test.

H0: mean(no_rain) <= mean(rain)

H1: mean(no_rain) > mean(rain)

p-value = 0.02

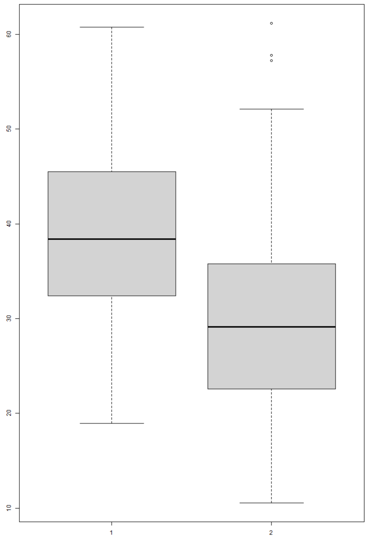
Tc = 2.05

Tt = 1.64

Due to Tc > Tt alternative hypothesis accepted. Ozone concentration is bigger during the days when there is not rainfall

B. Ozone – Season

In this case, we divided the ozone concentration samples into two groups: those belonging to the warm seasons (spring and summer) and the cold seasons (autumn and winter). Here, the box plot is arranged with the warm seasons on the left and the cold seasons on the right.



a) Skewness. Both skewness are in a normal range

b) Degrees of freedom. Hot season: 158. Cold season: 155

c) Normality test. Hot season – p value = 0.28. Cold season – p value = 0.46 Both are normal

d) Variance homogeneity. Fc = 1.17 Ft = 1.78 H0 accepted. Variances have homogeneity

e) T-student test.

H0: mean(hot season) <= mean(cold season)

H1: mean(hot season) > mean(cold season)

p-value = 2.2e-16

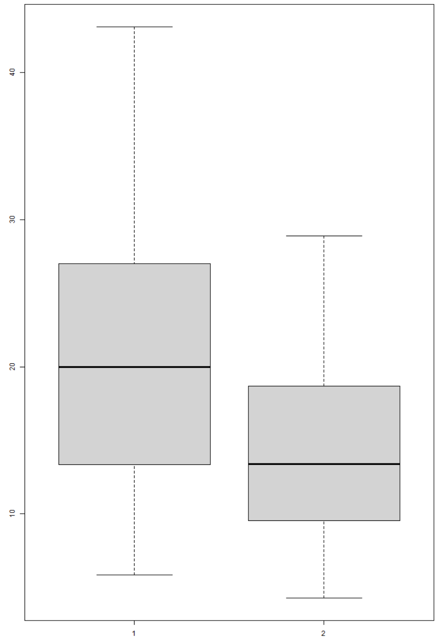
Tc = 8.68

Tt = 1.64

Due to Tc > Tt alternative hypothesis accepted. Ozone concentration is bigger during the summer and spring

C. PM2.5 – Season

In this case, we divided the PM2.5 concentration samples into two groups: those belonging to the warm seasons (spring and summer) and the cold seasons (autumn and winter). Here, the box plot is arranged with the warm seasons on the left and the cold seasons on the right.



a) Skewness. Both skewness are in a normal range. In fact, both have very similar values, so they have almost the same skewness

b) Degrees of freedom. Hot season: 158. Cold season: 155

c) Normality test. Hot season – p value = 0.23. Cold season – p value = 0.30 Both are normal

d) Variance homogeneity. Fc = 2.31 Ft = 1.30 H0 rejected. Var(hot_season) > var(cold_season)

e) T-student test.

H0: mean(hot season) <= mean(cold season)

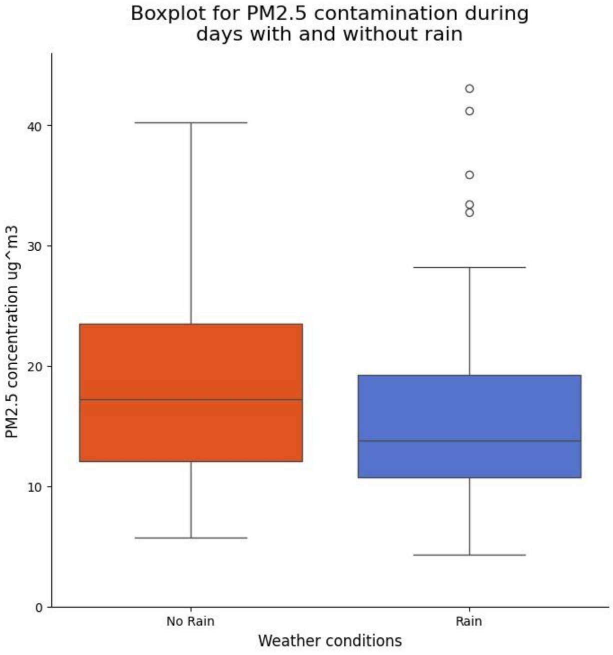
H1: mean(hot season) > mean(cold season)

p-value = 4.93e-14

Tc = 7.84

Tt = 1.65

Due to Tc > Tt alternative hypothesis accepted. PM2.5 concentration is bigger during the summer and spring



D. PM2.5 – Rain

In this case, we divided the PM2.5 concentration samples into two groups: those belonging to the days with rain and those with no rain. Here, the box plot is arranged with no rain samples on the left and samples with rain on the right.

a) Skewness. Both skewness are in a normal range. The samples of the days with rain have a bigger skewness: 1.18

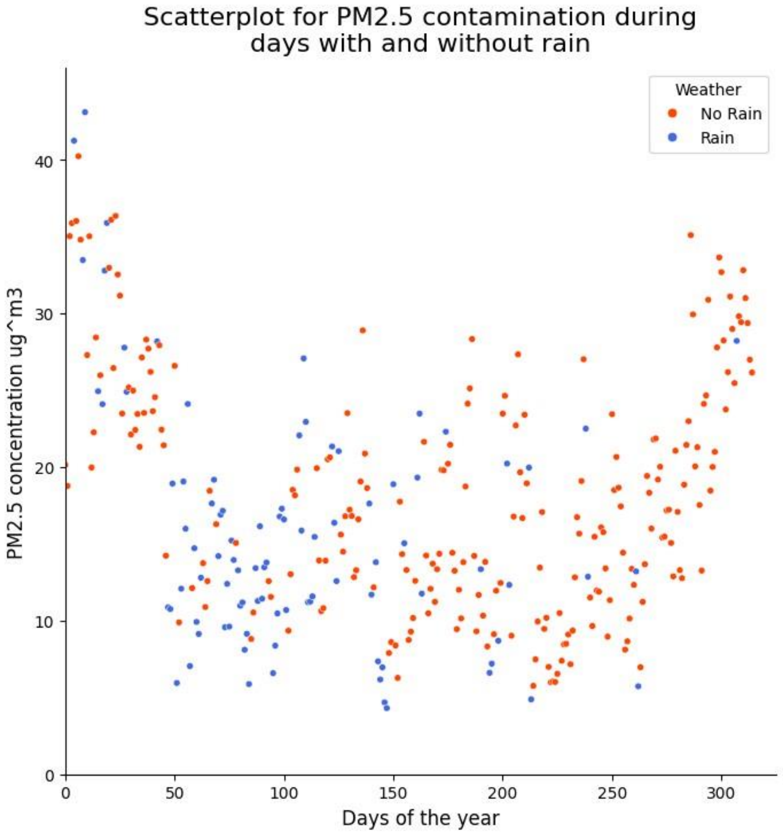
b) Degrees of freedom. No rain: 224. Rain: 91

c) Normality test: No rain – p value = 0.03. Rain – p value = 0.17. Only one is normal

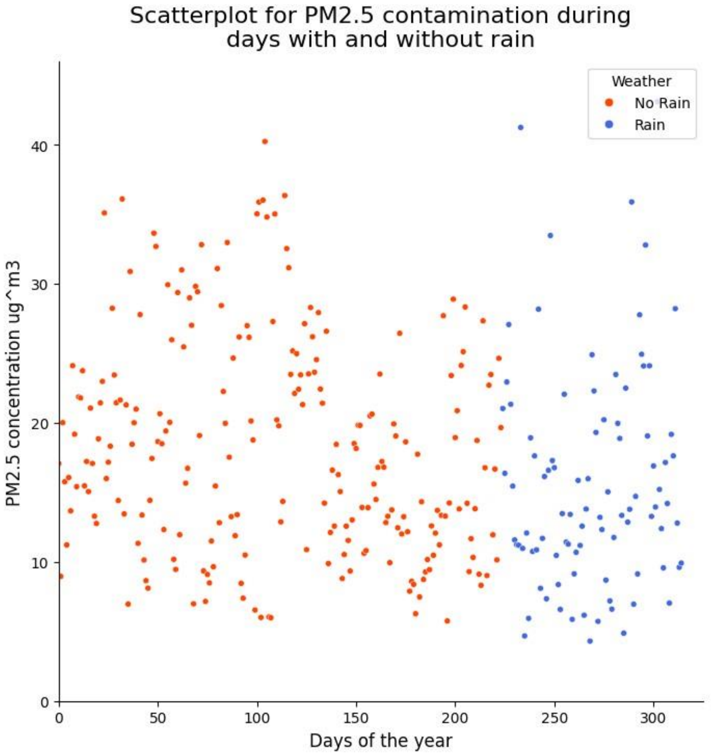
Due to one has normal distribution and the other must be transformed with box cox, it is obvious that with the transformation the variances will be different.
Variance homogeneity
H0: var(no_rain) = var(rain)
H1: var(no_rain) < var(rain)
Fc = 76.90
Ft = 1.35
Fc > Ft -> H0 rejected. Variances are different
So, the t – test would be:
H0: mean(transform_no_rain) <= mean(rain)
H1: mean(transform_no_rain) > mean(rain)
P value = 1
Tc = -13.78
Tt = 1.66
H0 accepted: There is no enough evidence to demostrate that concentration of PM2.5 is bigger during the days with no rainfall

However, if we repeat the process with both variables transformed with box cox, we get that the variances are the same:
H0: var(no_rain) = var(rain)
H1: var(no_rain) < var(rain)
Fc = 1.11
Ft = 1.35
Fc < Ft -> H0 accepted. Variances equal
And also if we don't do the box cox transformation:
H0: var(no_rain) = var(rain)
H1: var(no_rain) < var(rain)
Fc = 1.05
Ft = 1.35
Fc < Ft -> H0 accepted. Variances are equal
In both cases we can do the t-test getting a tc = 2.87 for the samples transformed and a tc = 2.62 for the samples without transformation. So, in both cases we can reject H0 and affirm that PM2.5 concentration during the days when there is no rainfall is bigger.

The question is which result is correct? Even using scatter plot for visualize raw data:



We can sort the samples trying to improve the plot:



Since the response isn't clear this way, we can move on to the analysis of variance to find a definitive answer regarding the variability of different amounts of rainfall on ozone concentration.

1. ANOVA. PM2.5 – Rain

   For this analysis of variance, we want to determine if there is a difference in PM2.5 concentration with respect to different rainfall treatments. These treatments were defined as follows:

   - Rain_Tot = 0: No Rain
   - Rain_Tot < 10: Rain
   - Rain_Tot >= 10 Heavy Rain

   ANOVA results:

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Weather	2	420	210.05	3.454	0.0328
Residuals	312	18975	60.82		

   Then, we can reject the null hypothesis: not all the rain samples have the same impact on the PM2.5 concentration

   Before, doing a Tukey test for finding the rain categories that actually have impact on the PM2.5 concentration, we hace to check the normality of the residuals and the homogeneity of the variances.

   p value = 0.02 There is not normality

   So, let's do a Kruskal-Wallis test instead of the ANOVA that we did:

   - chi-squared = 8.4563
   - df = 2
   - p-value = 0.0145

   In this case, we also can reject the null hypothesis: not all the rain samples have the same impact on the PM2.5 concentration.

   Finding with pairwise Wilcoxon test the difference between treatments.

	Heavy Rain	No rain
No rain	0.750	-
Rain	1.000	0.016

   The pair: No Rain - Rain is the one that has a significant difference

2. ANOVA PM2.5 – Temp_Avg

   For this analysis of variance, we want to determine if there is a difference in PM2.5 concentration with respect to different temperature treatments. These treatments were defined as follows:

   - Low Temperature <= 15
   - Medium Temperature <= 20
   - High Temperature > 20

ANOVA results:

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Temperature	2	7005	3502	88.2	<2e-16
Residuals	312	12390	40		

We can reject the null hypothesis (with high level confidence): not all the temperature categories have the same impact on the PM2.5 concentration
Before, doing a Tukey test for finding the temperature categories that actually have impact on the PM2.5 concentration, we have to check the normality of the residuals and the homogeneity of their variances.
p-value = 0.27: There is normality on the residuals

Variance homogeneity Levene test

	Df	F value	Pr(>F)
Group	2	3.6794	0.0263
	312		

p-value = 0.02 -> Variances are different

So let's do an ANOVA Welch test because variances are different

- F = 67.918
- Num df = 2
- Denom df = 107.97
- P value < 2.2e-16

We can reject the null hypothesis: not all the temperature categories have the same impact on the PM2.5 concentration

Tukey test for checking which temperature category has a different effect on the PM2.5 concentration

	Diff	Lwr	Upr	P adj
Low Tem – High Tem	-12.5317	-15.2369	-9.8266	0.0000
Medium Tem – High Tem	-10.3179	-12.3166	-8.3191	0.0000
Medium Tem – Low Tem	2.2138	-0.1664	4.5942	0.0744

diff: mean difference between the two categories according to the PM2.5 concentration
lwr - upr confidence interval
p adj. Significance.

Only Medium and Low Temperature don't have significant differences

3. ANOVA Ozone – Temp_Avg

For this analysis of variance, we want to determine if there is a difference in O3 concentration with respect to different temperature treatments. These treatments were defined as follows:

- Low Temperature <= 15
- Medium Temperature <= 20
- High Temperature > 20

ANOVA results:

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Temperature	2	11240	5620	85.63	<2e-16
Residuals	312	20477	66		

We can reject the null hypothesis (with high level confidence): not all the temperature categories have the same impact on the O3 concentration

Before, doing a Tukey test for finding the temperature categories that actually have impact on the O3 concentration, we have to check the normality of the residuals and the homogeneity of their variances.
p-value = 0.37: There is normality on the residuals

Variance homogeneity Levene test

	Df	F value	Pr(>F)
Group	2	5.7441	0.0035
	312		

p-value = 0.003 -> Variances are different

So let's do an ANOVA Welch test because variances are different

- F = 131.71
- Num df = 2
- Denom df = 124.92
- P value < 2.2e-16

We can reject the null hypothesis: not all the temperature categories have the same impact on the O3 concentration

Tukey test for checking which temperature category has a different effect on the O3 concentration

	Diff	Lwr	Upr	P adj
Low Tem – High Tem	-17.5656	-21.0433	-14.0878	0.0000
Medium Tem – High Tem	-11.8544	-14.4239	-9.2848	0.0000
Medium Tem – Low Tem	5.7112	2.6511	8.7713	4.51e-05

In this case all the temperature categories have a different effect on the ozone concentration. The bigger the temperature, the more concentration of ozone we have

4. ANOVA Ozone – RH

For this analysis of variance, we want to determine if there is a difference in O3 concentration with respect to different humidity treatments. These treatments were defined as follows:

- Humidity 1 < 33
- Humidity 2 < 43
- Humidity 3 < 53
- Humidity 4 < 63
- Humidity 5 < 73
- Humidity 6 < 83
- Humidity 7 >= 83

ANOVA results:

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
RelativeHumidity	2	15451	2575.2	48.76	<2e-16
Residuals	308	16265	52.8		

We can reject the null hypothesis (with high level confidence): not all the relative humidity categories have the same impact on the O3 concentration

Before, doing a Tukey test for finding the RH categories that actually have impact on the O3 concentration, we have to check the normality of the residuals and the homogeneity of their variances.

p-value = 0.93: There is normality on the residuals

Variance homogeneity Levene test

	Df	F value	Pr(>F)
Group	6	0.9769	0.4409
	308		

p-value = 0.44 -> Variances are equal

Tukey test for checking which relative humidity category has a different effect on the ozone concentration

	Diff	Lwr	Upr	P adj
Humidity 2-Humidity 1	-0.9240	-7.3039	5.4558	0.9995
Humidity 3-Humidity 1	-5.5160	-11.9289	0.8968	0.1445
Humidity 4-Humidity 1	-9.6965	-16.2733	-3.1197	0.0003

Humidity 5-Humidity 1	-11.6634	-17.7320	-5.5949	0.0000
Humidity 6-Humidity 1	-19.0294	-25.0979	-12.9609	0.0000
Humidity 7-Humidity 1	-27.4703	-36.9127	-18.0278	0.0000
Humidity 3-Humidity 2	-4.5920	-9.0422	-0.1418	0.0381
Humidity 4-Humidity 2	-8.7724	-13.4557	-4.0892	0.0000
Humidity 5-Humidity 2	-10.7394	-14.6772	-6.8016	0.0000
Humidity 6-Humidity 2	-18.1054	-22.0431	-14.1676	0.0000
Humidity 7-Humidity 2	-26.5462	-34.7827	-18.3098	0.0000
Humidity 4-Humidity 3	-4.1804	-8.9085	0.5475	0.1223
Humidity 5-Humidity 3	-6.1474	-10.1383	-2.1564	0.0001
Humidity 6-Humidity 3	-13.5133	-17.5043	-9.5224	0.0000
Humidity 7-Humidity 3	-21.9542	-30.2162	-13.6922	0.0000
Humidity 5-Humidity 4	-1.9669	-6.2162	2.2823	0.8154
Humidity 6-Humidity 4	-9.3329	-13.5822	-5.0836	0.0000
Humidity 7-Humidity 4	-17.7737	-26.1636	-9.3839	0.0000
Humidity 6-Humidity 5	-7.3659	-10.7761	-3.9557	0.0000
Humidity 7-Humidity 5	-15.8068	-23.8045	-7.8091	0.0000
Humidity 7-Humidity 6	-8.4408	-16.4385	-0.4431	0.0309

In close categories there are not significant changes. Specifically, in low percentages of rh, there are no significant changes on the ozone concentration. But, the more rh we have, the more ozone concreation we get.

5. ANOVA Ozone – Rad_Avg
   For this analysis of variance, we want to determine if there is a difference in O3 concentration with respect to different solar radiation treatments. These treatments were defined as follows:

- Radiation 1 <= 300
- Radiation 2 < =500
- Radiation 3 > 500

ANOVA results:

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Radiation	2	6120	3060	37.3	3e-15
Residuals	312	25597	82		

We can reject the null hypothesis (with high level confidence): not all the solar radiation categories have the same impact on the O3 concentration

Before, doing a Tukey test for finding the radiation categories that actually have impact on the O3 concentration, we have to check the normality of the residuals and the homogeneity of their variances.
p-value = 0.6299: There is normality on the residuals

Variance homogeneity Levene test

	Df	F value	Pr(>F)
Group	2	1.7462	0.1761
	312		

p-value = 0.17 -> Variances are equal

Tukey test for checking which radiation category has a different effect on the ozone concentration

	Diff	Lwr	Upr	P adj
Radiation 2 - Radiation 1	7.1279	2.7914	11.4644	0.0003
Radiation 3 - Radiation 1	13.8761	9.5242	18.2280	0.0000
Radiation 2 - Radiation 1	6.7482	4.2250	9.2713	0.0000

All the categories presents different impact on the ozone concentration. Specially, low levels of radiation decrease the ozone concentration

Discussion

To analyze and discuss the results from the previous section, they will be presented in the order they were shown.

1. This first point was heavily focused on sensor operation. As we could see in both the histograms and tables, the vast majority provided more than 90% of the data under the OK flag, indicating reliable data measurements. However, for PM2.5 and NO, we did have a significant percentage of measurements below the detection limit. Especially with the NO sensor, we do have an issue, as more than 75% of the data was flagged as BDL. Based on this analysis, it would be advisable to verify the operation of this sensor: it's possible that

pollution levels for this molecule are simply low, or alternatively, it may be necessary to improve the sensor's sensitivity to enhance measurements.

2. This analysis was quite straightforward, but the results were interesting. Working with measurements per month, one would expect that months with 31 days should have the same number of samples (as well as months with 30 days among themselves). Additionally, one would expect February to have fewer samples. However, here we can note several things:
    - March is the month with the least instances by far: almost less than half compared to the others.
    - Almost all months have a different number of instances. In fact, the only ones that have the same number are April and September.

These modifications could be due to the removal of some instances due to null data or when merging the meteorology and air quality datasets (during the data cleaning). However, it would be advisable to review the functioning of the minute-by-minute sensor measurements.

3. This point is extremely important for our analysis as it constitutes the first description of the air quality data, at least for the dataset per minute.

Based on the initial table with measures of central tendency, dispersion, and skewness:

- The first thing we noticed is that all variables have positive skewness. We can justify this with density plots and boxplots: most data points are clustered within a certain range, but there are a significant number of extremely large outliers that contribute to the positive skewness of the variables. This was partly expected because pollution levels may behave similarly for most of the year, but during certain peak events, they can spike significantly (even after data cleaning).
- Justifying these highly positively skewed distributions, we have extremely high Fisher skewness values, reaching up to 27.31. Additionally, we can observe that in all variables, the mean is greater than the median.
- Another common pattern in the variables (also caused by these particular distributions) is a very high standard deviation. In several cases, the standard deviation is greater than the mean itself, and in the lesser cases, it is close to half of the mean value. This can be explained by the presence of outliers.

Going into more detail about the outliers, the boxplots clearly show both the large number of outliers and their considerable magnitudes. In the visualization, the boxes are very small or virtually nonexistent, indicating that the interquartile range practically does not change.

For this reason, density plots were created for each variable afterward. With such extreme data, it was important to observe the distribution. We can see in these plots that the peaks of the distributions are heavily skewed to the left, while the outliers extend the plots towards the right.

Furthermore, the topology of the distribution peaks shows that most are not smooth. This indicates that there is a value that repeats frequently, generating the peak, but the data is not normally distributed.

In fact, considering these observations, it was expected that normality would not be achieved.

An interesting hypothesis to explain these distributions would be to verify the high peaks of pollution and contextualize them according to date and meteorology.

4. Continuing with the same procedure, but now for the meteorological variables. First, if we observe the table:
   - Here we can see that most variables have more common asymmetries. Only in the case where we get a NaN because remember that for the average radiation, we left many outliers.
   - However, here we can also see the highest asymmetry in the entire dataset: 36.21 in the amount of rainfall.
   - Regarding the other variables, we can see some relatively high standard deviations, but for the most part, they are within normal ranges.

If we continue with the boxplots, we can justify what was described in the tables: most of the plots remain in order and with few outliers except for WSpeed_Avg and Rain_Tot.

For wind speed, it shows a similar behavior to the air quality variables: in its interquartile range, it stores variables that do not change much in magnitude. However, very high outliers cause the distribution to change. In this case, the outliers could be samples where there were very high wind gusts that caused the values to rise.

Now, regarding rainfall, we have a very particular diagram: the interquartile range is at 0 and everything above is considered an outlier. The initial interpretation might be sad and concerning: the atypical aspect is that it is raining (thus justifying the droughts we have experienced in the last year).

However, it is important to remember that we are working with samples per minute. It is most normal for it not to be raining every minute.

It is precisely due to these results that the decision was made to create the dataset with daily temporal granularity, to truly analyze what happens with the weather and air quality in a more appropriate context and with less noise.

We see this reflected in the density plots for the meteorological variables per minute. Rain_Tot shows a distribution that practically looks like a peak at 0.0. The same goes for wind speed and direction (in fact, we will analyze this point in the next analysis). It was not necessary to conduct the Kolmogorov-Smirnov tests to demonstrate that we would not find normality.

5. Based on the previous results, it was decided to create a histogram for the wind direction frequencies. If we round the values to integers and plot the 360 degrees, we see a histogram that appears to be normal. This is especially true if we consider 360 instead of 0 (because otherwise we would be adding an extra

data point). What would be interesting here is to complement this analysis with geographical data from the atmospheric observatory to justify the regions where the wind is strongest.

6. A similar procedure was carried out for the relative humidity data. Since it is in percentage units, it can be discretized into 100 intervals. However, here we did notice an increase in the distribution of the data. This suggests that despite the lack of rain, humidity levels did not decrease completely
7. Due to the issues with the temporal nature of the samples, in order to graph the air quality variable concentrations throughout the year, monthly averages were calculated to create a visualization that truly tells the story of how the data evolved (which in data visualization is known as storytelling). This was the approach taken to create the graphs in this section.

   Regarding PM10 and PM2.5, we observe an almost synchronized behavior (which we will see later with their paired scatter plot). During the hot and dry periods (from March to June), we see higher pollution levels. Specifically, the peaks of higher pollution are found in months without rain. Decreases are observed in July and at the beginning and end of the year (during summer and winter rains). These factors led us to generate the hypothesis tests we conducted at the end of the results section.

   Cases of NO2 and NO show that they remain constant. This would indicate that they are pollutants that do not change significantly over time, but rather remain stable.

   However, for O3, we notice a behavior similar to that of PM: during the hottest seasons, with more sunlight and less rain, ozone concentrations increase. The decrease is caused by the temperature drop along with the low rainfall of the year.

   Finally, we have CO. For this pollutant, we see that the peak is in February, not at the beginning of summer like all the other variables. However, its behavior shows that it remains relatively stable over time. Only in July is there a considerable decrease and the maximum we mentioned in February. This would suggest a different behavior from the rest of the variables that would be worth explaining.
8. As mentioned earlier, the goal of this section was to find variables with a strong correlation. However, in terms of visualization, this graph did not prove to be a powerful tool (especially due to the number of variables). That's why it was better to perform the calculations in the next section.
9. At this point, we began with a table to directly search for the highest Pearson coefficients (or lowest for negative correlations). While some coefficients were found with ozone values above or below 0.70, there were others very close to 0. This could be explained by the subsequent finding that the variables were not normally distributed. Therefore, the Pearson coefficient was not suitable for use. However, what was decided at the time to filter out statistically significant correlations was to raise the confidence level to 99%. With this, it was observed

that in these cases, O3 had a positive correlation with Temp_Avg and a negative correlation with RH_Avg. This was important with the minute-level data. However, with the change in temporal resolution, we will see later on that there were more correlations that helped us achieve the project's objective (linking an air quality variable with a meteorological variable).

10. At this point, the goal was to continue from the previous step with the most pronounced correlations in this dataset that would be useful for our study: O3 with Temp_Avg (positive) and O3 with RH_Avg (negative).

    Upon creating the visualization, we see that we have so many data points that even if we reduce the thickness of the points, they still appear as a large cloud of data. This is not necessarily a bad thing, but for the purpose of observing regressions (by plotting the linear regression line), we see that it is very difficult for them to fit the model. This suggests that no matter how pronounced a relationship is between a meteorological factor and an air quality factor, we may not be able to see it linearly.

    Also, note that there are data points that deviate significantly not only from the regression line but also from the overall data cloud. These data points are precisely the outliers in the molecule concentration (in this case, O3).

    With these results, we can hypothesize that for O3, the meteorological factors that most affect it are temperature and relative humidity.

11. We then moved on to the normality tests. The truth is, they were done merely as a formality, but after all the analysis we've conducted up to this point, it was expected that normality would not be achieved even after performing Box-Cox transformations with their corresponding lambda values.

    The interesting aspect here is to note that the p-value never changed from 2.2e-16. This means that H0, which assumes normality, was decisively rejected. The risk of committing a Type I error was very low. However, an important observation is that although the Box-Cox transformations did reduce the values of Dc, it was not enough.

    This can be explained by the large amount of data we have in the minute-level dataset: they generate a lot of noise, and with such a large dataset, one would expect that they would easily exhibit normality (however, since they have distributions with very large skewness, they naturally do not exhibit normality)

12. Due to all of the above, it was decided to continue with the daily time series dataset. Continuing with the minute-level data was not a good idea since the information they would provide would not be as reliable.

    Therefore, this point was used to perform descriptive statistics on this new dataset. Firstly, looking at the tables, we can see that with this change in time granularity, neither the mean nor the median were significantly altered. However, the outliers were adjusted to the point that most of the skewness values entered a very low range. This is a great success in data treatment because we did not modify the interquartile range or the measures of central tendency, but we did manage to deal with the outliers that were affecting the data.

As mentioned, the skewness values went from highly skewed values to almost symmetrical in some cases. However, there are still some variables that present outliers, especially rainfall.

Lastly, regarding the tables, another justification that we managed to reduce noise in the data can be seen in the reductions in standard deviation in all variables.

Moving on to the boxplots, the visualization improved significantly without the outliers. Now we can observe the interquartile range of almost all variables to better understand their usual behavior. However, even though there was a considerable improvement, there are variables that still show very high extreme values.

In the case of NO, we can see from the boxplot that the concentration data tends to remain very low. However, it seems that on some days of the year, the concentration increased more than usual.

The interpretation that again provides strong results is that of rainfall. Even when summing up daily rainfall, what is still atypical is for it to rain. Demonstrating statistically that based on the past year, rainy days are atypical cases.

Finally, looking at the density plots, starting with the air quality variables, we notice that some like O3 or CO may directly reach normality. However, cases like NO2 or NO still seem to retain a positive skewness. So, in the normality tests, we will likely need to perform Box-Cox transformations on them.

Now, with the meteorological variables, we see that several of them can reach normality. However, I am intrigued by the fact that for RH_Avg, we had not seen with the boxplot that we may have a bimodal distribution. This could be due to the change in time granularity; however, this variable may still pose problems with normality.

13. When conducting normality tests again, but now for this daily time series dataset, we see what we were observing in the density plots. Some variables turned out to be normally distributed directly, but others had to be transformed.

    However, the important results worth mentioning are that the variables that remained non-normal were RH_Avg and Rain_Tot. Regarding rainfall, this was expected: most of the year does not experience rain, and the days with heavy rainfall created a distribution with a strong positive skewness.

    As for relative humidity, the bimodal nature makes it very difficult to transform it to achieve normality. Perhaps with a more sophisticated function than those suggested with the lambdas of the Box-Cox transformations, normality could be achieved. However, for working with it, it would be advisable to perform non-parametric tests or conduct subsampling to create separate subsets that exhibit normality.

14. Just to observe the change in distribution when modifying the temporal granularity, histograms were created for the wind direction and relative humidity variables. For WDir_Avg, we see the same result as in the minute dataset but less dense due to the reduction in data quantity.

    On the other hand, the issue of bimodality in relative humidity is much clearer in the histogram. What's interesting is that it's not an outlier but rather two local maximum regions in the distribution. We could interpret this as humidity oscillating between two extreme values throughout the year, meaning relative humidity became more extreme over the past year.
15. Reaching this point, we repeated the procedure to look for relationships between pairs of correlated variables. We are particularly interested in those between air quality and meteorological variables.

    Upon reviewing the first scatter plot of all the variables of interest, we can see that we no longer have as many scattered data points. Already at this stage, we can see data that seems to fit a linear regression model. This, combined with the normality we now have, allows us to calculate Pearson coefficients.

    Looking at the table of these coefficients, we see once again that we have some values very close to zero. However, a case like PM was expected to achieve a correlation close to 0.90, especially after observing its synchronization over time.

    Now, moving on to the correlogram, we see that there are indeed many correlations, but again, we apply the filter to keep only those that are statistically significant with a 99% confidence level.

    Specifically, if we also focus on those of interest (contaminant-meteorology pairs) and some other correlated contaminants, those are the ones we made graphs and regression models for.

    Note that the highest/lowest correlations were with the RH_Avg variable.
16. Regarding ozone, we see from the graphs that it is correlated with PM pollutants. In fact, although the data clouds go beyond the confidence interval of the regression, it is evident that they do form a linear relationship. When observing the ozone relationship with other meteorological variables, we find that cases involving temperature and relative humidity fit the regression model very well. This was expected because humidity tends to lower pollution levels, while temperature elevation is correlated with factors that contribute to ozone generation. As for solar radiation, it seems not to fit the model, which is somewhat contradictory because it was expected to have a stronger correlation since solar radiation also contributes to low-level ozone formation.

    Moving on to nitrogen-related variables (NO2 and NO), we observe that the correlation between them is not as clear, but there is a correlation between each of them and the same meteorological factors. In other words, NO2 and NO separately are affected by the same meteorological factors, but they do not interact in a way that affects each other.

Finally, we have PM10 and PM2.5. Due to their similarity, these two show the strongest correlation, which is evident from how closely they align with the regression line. This was expected given their behavior throughout the year and the fact that they are the same particle but of different sizes. Therefore, in this case, the difference in sizes does not make a significant impact.

If we do the same as with nitrogen-related variables and compare them with their most marked meteorological variables, we see that both PM10 and PM2.5 are affected almost equally by temperature and humidity. The only notable difference is that PM2.5 is not as affected by relative humidity as PM10 (in this case, the smaller size makes it more resistant to humidity). However, both are affected almost equally by meteorological factors, and there is a correlation between the concentrations of each particle.

Let's now move on to the inferences. From this point on, we will focus on the air quality variables that are of greater interest to the research groups on campus: ozone and PM2.5.

A. Ozone – Rain

Given the lack of normality, both the calculation of the Pearson coefficient (which assumes normality to be calculated) and the plotting were not clear in determining if rain affects ozone concentration. Therefore, this t-student test was conducted to verify if there are significant differences between ozone samples on rainy days and non-rainy days. Specifically, we expect the concentration to be higher on non-rainy days than on rainy days (thus, we can conduct a one-tailed test). The calculations supported this intuition, which was helpful because although the box plot already showed some differences, the t-test made it clearer. Therefore, indeed, rain does affect ozone concentration.

B. Ozone – Season

These tests for warm and non-warm seasons focus on separating the samples into the two semesters when the university is attended. In other words, here we ask, are there significant differences in ozone concentration between the spring-summer semester and the autumn-winter semester? The box and whisker plots already gave us the answer from the start (which would later be confirmed with the Ozone and Temperature ANOVA): ozone concentration is indeed higher in the summer semester than in the autumn semester. Therefore, it would be interesting to raise awareness among the university community about the causes and consequences of air pollution and how it affects us, especially in the spring-summer semesters.

C. PM2.5 – Season

Repeating the same subsampling procedure with the seasons, but now with PM2.5, we see very similar results to those with ozone. However, here they are slightly less conclusive. This means that the difference in PM2.5 pollution between the summer semesters is indeed greater than in the winter semester. However, it is not as pronounced as with ozone. This implies that meteorological factors do not affect PM2.5 as much; it is more of a human factor.

D. PM2.5 – Rain

Following our last inference with the t-student test before moving on to ANOVA, the plan was to repeat the same procedure with PM2.5 on rainy and non-rainy days, similar to what was done with ozone. However, a problem arose: one of the samples did not exhibit normality. Initially, it might be thought that using the Box-Cox transformation would suffice because normality was achieved. But now we face a scaling problem:

If we work under the assumption of normality, the non-rainy samples must be transformed, so they are no longer in the same units and scales as the rainy day samples. By doing this, it was expected that the variances would be different, and there would not be significant evidence to say that PM2.5 concentration is higher on non-rainy days. These conclusions are reached if we respect the assumptions of normality. However, I consider these assumptions a minor issue compared to comparing a transformed sample with one that is not.

That's why the procedure was repeated with both transformed and non-transformed samples. In both cases, the same conclusion was reached: the null hypothesis is rejected—PM2.5 concentration is indeed higher on non-rainy days than on rainy days.

But now comes the dilemma: which option to choose? If we sacrifice scales for normality, we must accept H0, but if we sacrifice normality for uniform scales, we must reject H0. In decision-making in these cases, it's best to refer to the graphs: to obtain a visualization that helps us decide better.

Looking at the boxplots, we can start to deduce the cause of the problem: there are outliers on some rainy days. That is, in the samples of rainy days, there were atypical cases where PM2.5 concentration increased.

This is reinforced by the scatter plots presented (both ordered and unordered): the data points for rainy days are below, but the outliers prevent this inference from being made mathematically.

For all these reasons, it was decided to conduct an ANOVA for this PM2.5 relationship with rain. The only difference is that for rain, we will have 3 treatments: no rain, rain, and extreme rain. With this test, we can demonstrate that PM2.5 concentration is indeed affected by rain.

E. ANOVA PM2.5 – Rain

As mentioned in the previous point, this ANOVA was conducted to find a formal answer to the hypothesis about rain and PM2.5 concentration.

As expected, normality was not found in the residuals of these samples even with the 3 treatments. However, Kruskal-Wallis tests were used to verify if there were differences between concentrations in different amounts of rain, and indeed, these differences were observed.

With the Wilcoxon tests, the differences were further examined, and the results are interesting: to decrease PM2.5 concentration, rain is needed. It's not necessary for it to rain heavily; in fact, the differences with heavy rain were not very significant.

These results, although unexpected, are quite good because it shows that a lot of rain is not needed to reduce concentrations.

F. ANOVA PM2.5 – Temp_Avg

The final inference regarding PM2.5 is related to temperature. We had already demonstrated that PM2.5 does change its concentration throughout the seasons of the year. However, the question here is whether temperature is one of those factors that modifies the concentration.

In this case, it was indeed the case: temperature does change the concentration; however, the differences in temperature are only noticeable with very high temperatures; there is not much difference between moderate and low temperatures. This may indicate that when it is very hot, the concentration of PM2.5 can indeed rise, but only at thresholds above 20°C.

G. ANOVA Ozone – Temp_Avg

Let's repeat this same procedure with 3 different temperature treatments for ozone. The results here show that ozone concentration is indeed sensitive to temperature changes. This is because all 3 pairs of treatments showed significant differences. With this, we can demonstrate that unlike PM2.5, ozone is indeed affected by temperature changes, even within ranges of 5°C among them. This was expected when looking at the scatter plot of O3 with Temp_Avg and seeing that the regression line had a relatively large positive slope compared to other factors.

H. ANOVA Ozone – RH_Avg

Now we move on to an ANOVA with relative humidity. Remember that it's one of the factors that has a fairly strong correlation with ozone. However, here, more detailed treatments were proposed, meaning that 7 treatments for humidity were conducted to detect the real threshold of change to notice significant differences in ozone concentration.

In this context, once the ANOVA and post hoc tests were completed, we noticed that among close treatments (those with not much variation in relative humidity between them), there were no significant differences. Strong differences begin to be noticeable when there are differences of at least 3 treatments.

This indicates that ozone is not so sensitive to relative humidity: there must be a really significant change (more than 20% humidity) for the ozone concentration to rise or fall.

I. ANOVA Ozone – Rad_Avg

A final analysis of variance: ozone concentration with different treatments of solar radiation. In theory, solar radiation is one of the factors that most affects ozone concentration. However, throughout this work, it has not been shown to be a key factor. That's why this ANOVA with 3 treatments of solar radiation is being conducted.

The results were very statistically significant: there are differences in the treatments. In fact, there are significant differences between all the treatments. Moreover, the difference in ozone concentration between each treatment is approximately 6.5 ppb. In other words, for every increase of 200 W/m^2 in solar radiation, the O3 concentration will rise by 6.5 ppb.

Conclusions

Working with atmospheric data is practically synonymous with working with big data. Fortunately, today we have these large databases, but as a result, it is necessary to have a deep understanding of statistics in order to fully leverage all the information they can provide us.

If we truly want to extract all the information available in this dataset, this work constitutes only the first steps: creating the dataset, cleaning it, providing an initial description of the data, and making the first inferences. The cleaning and validation of data are crucial because without them, the statistical analysis loses all validity. This part was done very well throughout the project; however, for future work, it is necessary that anyone wishing to work on the project already has preprocessed data ready to start working with.

On the other hand, the general description that was achieved was a significant step forward in outlining the current landscape. However, the next step to be taken is to start making more detailed descriptions by selecting only some variables, and possibly exploring other temporalities.

Regarding the inferences, what is presented here constitutes a good mathematical demonstration of several assumptions that have been previously proposed by research groups interested in the data. However, the most significant outcome of this is the emergence of new research questions. Many of the latest inferences regarding air quality variation can be further explored with the aim of showing the community what is really happening with the air we breathe.

In terms of the collective effort made during the course, this work fortunately allowed the application of each and every topic studied. We practically have a case study that can be used by anyone wishing to apply their knowledge in descriptive and inferential statistics. This, coupled with a solid foundation in data visualization, makes the report more accessible for those interested in starting the topic.

If there were one area to improve (and just to mention one because the course was simply perfect), I personally would seek to apply a first approach to multivariable analysis techniques. This is because we have more than 20 variables in the dataset. It would be useful and interesting to be able to use many variables simultaneously rather than just two at a time.

Finally, there is one last point, which is the data itself. None of this would be possible without the samples we can work with. Therefore, it is essential to raise awareness about the importance of collecting this data, both for being able to perform these statistical works and for starting to deploy machine learning models in the near future.

References

[1] Cook, J., Farmer, G. T. (2013). Climate Change Science: A Modern Synthesis. Volume 1 - The Physical Climate. Springer.

[2] Tan, Z. (2014). Air Pollution and Greenhouse Gases. From Basic Concepts to Engineering Applications for Air Emission Control. Springer.

[3] UNAM. Red Universitaria de Observatorios Atmosféricos. RUOA. Last accessed on April 18th, 2024 from https://www.ruoa.unam.mx/index.php?page=home

All the datasets can be found in:
https://www.ruoa.unam.mx/index.php?page=estaciones&id=9

All the calculations, procedures, and the work in general for this project are available in the repository: https://github.com/arnoldtics/Atmospheric_Data_Insights