British Journal of Environmental Sciences 12(2),49-58, 2024

Print ISSN: 2055-0219(Print)

Online ISSN: 2055-0227(online)

Website: https://www.eajournals.org/

Publication of the European Centre for Research Training and Development–UK

Applying an Ordinary Least Squares (OLS) Regression Model On Processed Air Quality and Environment Data

Ilma Lili
Department of Computer Sciences, University of Tirana, Albania
Email: ilma.lili@fshn.edu.al

Anxhela Kosta
Department of Computer Sciences, University of Tirana, Albania
Email: Anxhela Kosta@fshn.edu.al

doi: https://doi.org/10.37745/bjes.2013/vol12n24958 Published April 27, 2024

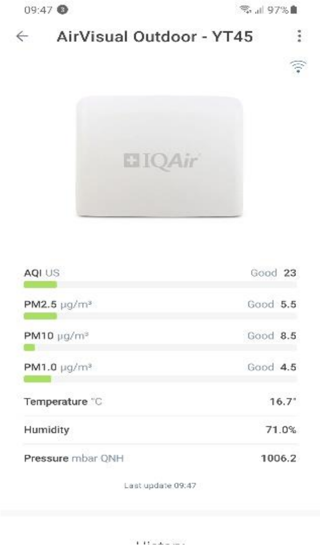
Citation: Lili I., and Kosta A. (2024) Applying an Ordinary Least Squares (OLS) Regression Model On Processed Air Quality and Environment Data, *British Journal of Environmental Sciences* 12(2),49-58

Abstract: *This scientific research is primarily based on real-time data collected on air quality. A comprehensive and extensive study was initially conducted to explore the key factors contributing to air pollution. Other relevant information will encompass additional components such as PM10, PM1, and weather-related factors like temperature, humidity, and air pressure. In order to provide a more reliable but at the same time qualitative information it was essential to examine the anomalies and issues revealed by the gathered data. After carefully identifying and correcting anomalies in the dataset, various statistical analyses have been conducted and results have been presented in both tabular and visual formats. These data are based on fundamental inquiries directly tied to the significance of air quality. After interpreting the statistics, a regression model such as OSL was used always including data that do not have multicollinearity. Based on the findings, it appears that this model is not suitable for forecasting PM2.5 levels because of a significant association between PM10 and PM1*

KEYWORDS: regression model, displot function, data, model, AQI, PM 2.5

INTRODUCTION

There are several elements in the air that affect the quality of the air. An important element is also PM2.5 with a diameter less or equal 2.5 micrometers where, in terms of health, they manage to penetrate deep into the lungs and even into the blood, causing serious health problems such as respiratory infections, heart problems or even strokes.[1] Meanwhile, it is thought that PM2.5 particles also play a role in climate change influencing global temperatures by reflecting solar rays or even absorbing them [2]. The research holds special importance due to the data that, when processed, influences the identification of significant trend patterns in scientific research. Additionally, the cleaning process is essential for ensuring accurate statistical presentation. The study begins by utilizing standardized equipment, such as the Air Visual Outdoor - YT45 monitoring device, which provides real-time information on 7 environmental parameters: PM2.5, PM10, PM1, AQI, temperature, humidity and barometric pressure [3]. It incorporates two independently calibrated laser sensor modules to ensure precise and dependable air quality data. These sensor modules can be easily interchanged for recalibration. The significance lies not only in the accuracy and real-time nature of the data, but also in preserving the historical early measurements. The mobile application can be effortlessly installed on a mobile device, allowing real-time data access with just a network connection. Figure 1 depicts the initial interface accessed through this application, while Figure 2 illustrates the online-accessible dashboard. The information has been downloaded from the dashboard for the annual periods 2022, 2023, 2024. The data is in CSV format, known for its universality and compatibility with Python libraries such as Pandas which has built-in support for CSV files. This simple format can handle a wide range of data types from numeric to strings, making it versatile for various data storage needs.

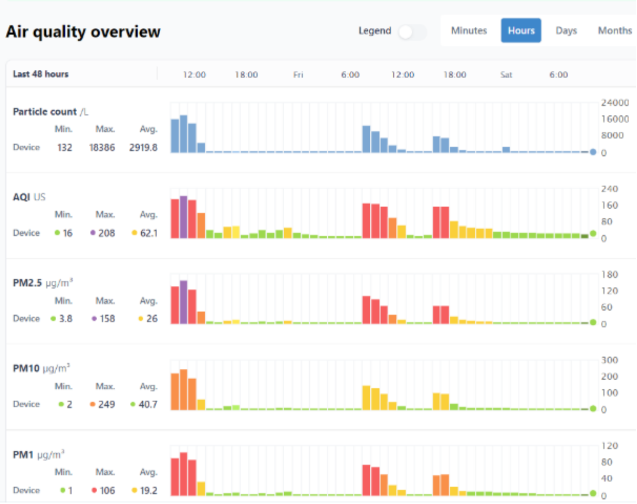


AirVisual Outdoor - YT45

AQI US	Good	23
PM2.5 µg/m³	Good	5.5
PM10 µg/m³	Good	8.5
PM1.0 µg/m³	Good	4.5
Temperature °C		16.7°
Humidity		71.0%
Pressure mbar QNH		1006.2

Last update 09:47

Figure 1. Mobile app for air visualizer parameters



Air quality overview

Last 48 hours

	Min.	Max.	Avg.
Particle count /L			
Device	132	18386	2919.8
AQI US			
Device	● 16	● 208	● 62.1
PM2.5 µg/m³			
Device	● 3.8	● 158	● 26
PM10 µg/m³			
Device	● 2	● 249	● 40.7
PM1 µg/m³			
Device	● 1	● 106	● 19.2

Figure 2. Dashboar air quality overview

The dataset used is in CSV format and contains 9436 records with 23 columns. The information received was gross and normally needed adjustments and precisions to narrow it down to reach the analysis that will be done for air quality parameters, mainly PM2.5. Given that the purpose of this research is precisely data analysis, the use of Jupyter notebook version 6.5.4 with the Python 3.11.3 programming language was considered as a tool. The libraries used are: Panda which is known for the analysis of mainly tabular data. It has tools that enable cleaning, transformation or even manipulation of data. It is mainly worked with dataframe objects that resemble excel spreadsheets. During the work, the lack of data was found as a result of external factors, and with the help of this library, the missing information was properly filled. It also integrates with Matplotlib by providing methods to. simple for graphical presentation of various statistics. The NumPy library that helps works mainly for scientific calculations mainly in supporting matrices with a rich collection of high level mathematical functions who's operates on these arrays. matplotlib.pyplot is another used library that is versatile offering a wide range of plot types, from histogram and bar charts to scatter plots and line graphs. Normally, the attribute of this library is its customization and integration with Panda and NumPy. For the Seaborn library, it was chosen to use the distplot function which combines Matplotlib functionality with improved features for visualizing the distribution of univariate set of observations. When sns.set() is called, it sets the aesthetic parameters for the plots to Seaborn's default style. In statsmodels.formula.api, the formula interface is used to define the relationship between a response variable and one or more explanatory variables. Considering that the last step of this study is to find a suitable regression model.

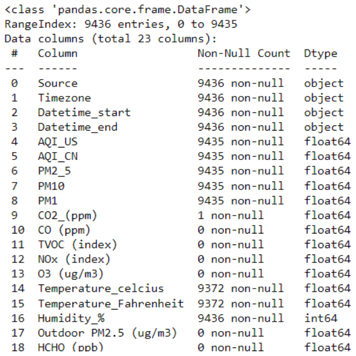
METHODOLOGY AND TECHNIQUES

The dataset is in CSV format and comprises 9436 records with 23 columns. The extensive information received needed refinement and clarification to focus on analyzing air quality parameters, particularly PM2.5.The study seeks to conduct data analysis, and as a result, Jupyter notebook version 6.5.4 was selected along with Python 3.11.3 as the programming language of choice for this task. The following libraries are used, starting with the Pandas library, known for analyzing mainly tabular data. It provides tools to clean, transform, and manipulate data. It primarily works with data frame objects that resemble Excel spreadsheets. In the course of the study, it was observed that the absence of data was attributable to external circumstances. However, with the assistance of the library, the deficient information was appropriately supplemented. Additionally, it facilitates integration with Matplotlib by offering methods that enable the straightforward graphical representation of diverse statistical data. The NumPy library primarily facilitates scientific calculations by providing a comprehensive set of high-level mathematical functions that act on matrices. The matplotlib.pyplot package is a versatile tool that provides a diverse selection of plot types, including histograms, bar charts, scatter plots, and line graphs. Typically, this library's feature set includes customization and integration with NumPy and Panda. The selection of the distplot function for the Seaborn library was based on its ability to integrate Matplotlib capability with enhanced features for the purpose of showing the distribution of a univariate set of observations. Plot aesthetic parameters are set to Seaborn's default style when sns.set() is executed.In statsmodels.formula.api, the formula interface is used to define the relationship between a response variable and one or more explanatory variables considering the fact that the last step of this study is to identify if OLS regression model is a fitting model for air quality elements.

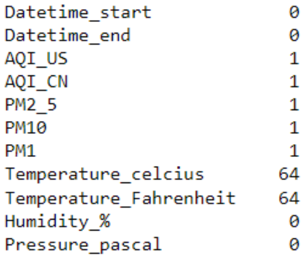
DATA ANALYSES AND FINDINGS

After defining the base libraries, the CSV file was read and converted into a DataFrame using pd.read_csv().

An analysis revealed that 57% of the columns contain data in over 95% of cases, while 9% (2 columns) have only 19% records with data and 34% are entirely empty. It was determined that the columns without data are not necessary for fitting the regression model, so they were dropped using data.drop() functionality.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9436 entries, 0 to 9435
Data columns (total 23 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Source                   9436 non-null   object
 1   Timezone                 9436 non-null   object
 2   Datetime_start           9436 non-null   object
 3   Datetime_end             9436 non-null   object
 4   AQI_US                   9435 non-null   float64
 5   AQI_CN                   9435 non-null   float64
 6   PM2_5                    9435 non-null   float64
 7   PM10                     9435 non-null   float64
 8   PM1                      9435 non-null   float64
 9   CO2_(ppm)                1 non-null      float64
 10  CO (ppm)                 0 non-null      float64
 11  TVOC (index)             0 non-null      float64
 12  NOx (index)              0 non-null      float64
 13  O3 (ug/m3)               0 non-null      float64
 14  Temperature_celcius      9372 non-null   float64
 15  Temperature_Fahrenheit   9372 non-null   float64
 16  Humidity_%               9436 non-null   int64
 17  Outdoor PM2.5 (ug/m3)    0 non-null      float64
 18  HCHO (ppb)               0 non-null      float64
```

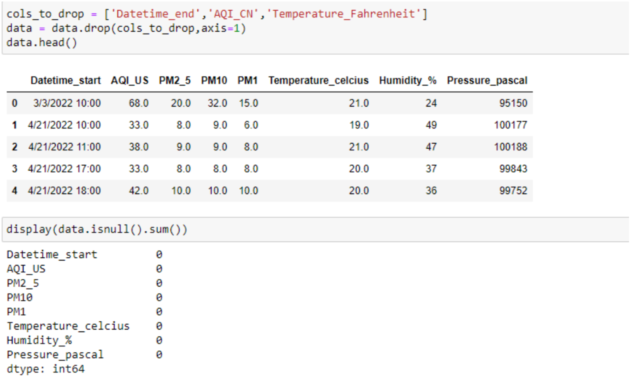


```
Datetime_start             0
Datetime_end               0
AQI_US                     1
AQI_CN                     1
PM2_5                      1
PM10                       1
PM1                        1
Temperature_celcius       64
Temperature_Fahrenheit    64
Humidity_%                 0
Pressure_pascal            0
```

Figure 3-Before Data Manipulation

Figure 4-Identification of NULL cases

Additionally, any duplications were checked through data.duplicated().any() function. Instead NULL values were evaluated using data.isnull().sum() function. The mean imputation technique[4] was then utilized to replace NULL values in specific columns such as AQI_US, PM2_5, PM10, PM1, Temperature_celcius, Humidity_% and Pressure_pascal. This approach is typically beneficial for machine learning algorithms.



```
cols_to_drop = ['Datetime_end','AQI_CN','Temperature_Fahrenheit']
data = data.drop(cols_to_drop,axis=1)
data.head()
```

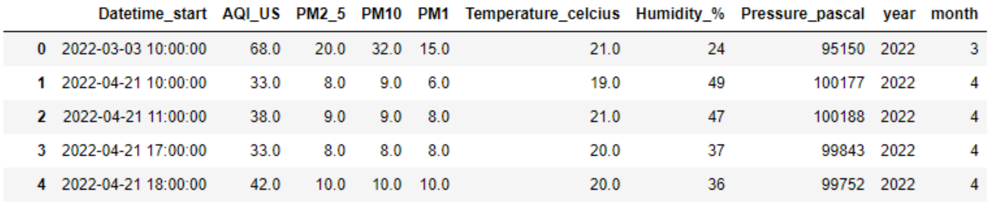
	Datetime_start	AQI_US	PM2_5	PM10	PM1	Temperature_celcius	Humidity_%	Pressure_pascal
0	3/3/2022 10:00	68.0	20.0	32.0	15.0	21.0	24	95150
1	4/21/2022 10:00	33.0	8.0	9.0	6.0	19.0	49	100177
2	4/21/2022 11:00	38.0	9.0	9.0	8.0	21.0	47	100188
3	4/21/2022 17:00	33.0	8.0	8.0	8.0	20.0	37	99843
4	4/21/2022 18:00	42.0	10.0	10.0	10.0	20.0	36	99752

```
display(data.isnull().sum())
```

```
Datetime_start         0
AQI_US                 0
PM2_5                  0
PM10                   0
PM1                    0
Temperature_celcius    0
Humidity_%             0
Pressure_pascal        0
dtype: int64
```

Figure 5-After Data Manipulation

Bearing in mind that some statistics will be applied to these data, it is necessary to add some columns with separate information on the year, month, day, hour, etc. The Datetime_start column is considered as an object and it is necessary to convert it to the date format to extract the detailed columns that serve us in the statistical part.



	Datetime_start	AQI_US	PM2_5	PM10	PM1	Temperature_celcius	Humidity_%	Pressure_pascal	year	month
0	2022-03-03 10:00:00	68.0	20.0	32.0	15.0	21.0	24	95150	2022	3
1	2022-04-21 10:00:00	33.0	8.0	9.0	6.0	19.0	49	100177	2022	4
2	2022-04-21 11:00:00	38.0	9.0	9.0	8.0	21.0	47	100188	2022	4
3	2022-04-21 17:00:00	33.0	8.0	8.0	8.0	20.0	37	99843	2022	4
4	2022-04-21 18:00:00	42.0	10.0	10.0	10.0	20.0	36	99752	2022	4

Figure 6-Data with the new columns

STATISTICS REPORT

The primary source of inquiries for many statistical data typically comes from a variety of interests [5]. Using questions helps us interact with the data and its internals more effectively. Finding the air concentration density of PM2.5, PM10, and PM 1 is one of the questions involved. This question is summarized in the graphs below.

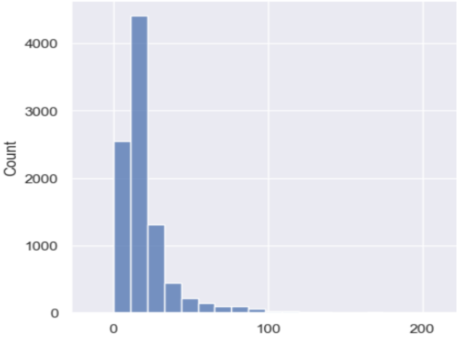


Figure 7-Density of recorded concentration PM2.5 in the air

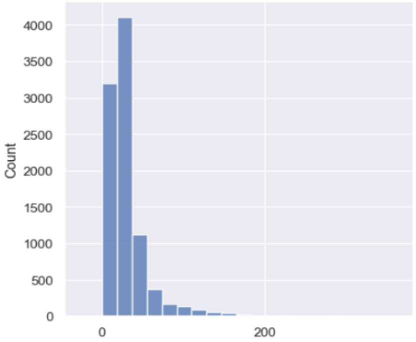


Figure 8-Density of recorded concentration of PM 10 in the air

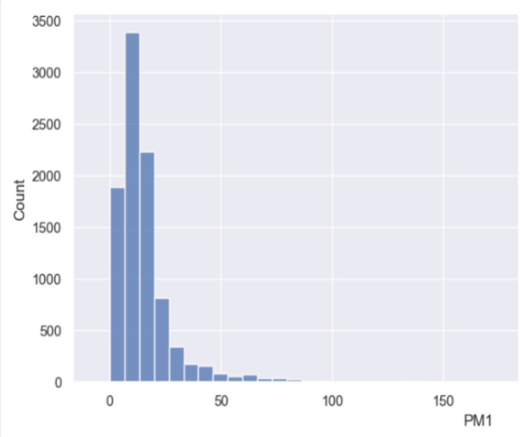


Figure 9-Density of recorded concentration PM 2.5 in the air

What pattern may be seen on a weekly basis for the hourly measurement of PM2.5 concentration in the air? This one means find the daily average of PM2.5 contained in the air in any given hour.

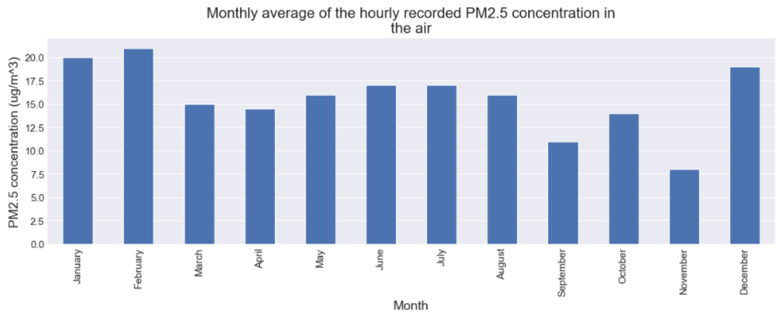
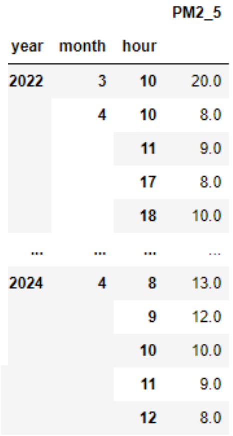


Figure 10-Monthly average of the hourly recorded for PM2.5



			PM2_5
year	month	hour	
2022	3	10	20.0
	4	10	8.0
		11	9.0
		17	8.0
		18	10.0
...	...	...	...
2024	4	8	13.0
		9	12.0
		10	10.0
		11	9.0
		12	8.0

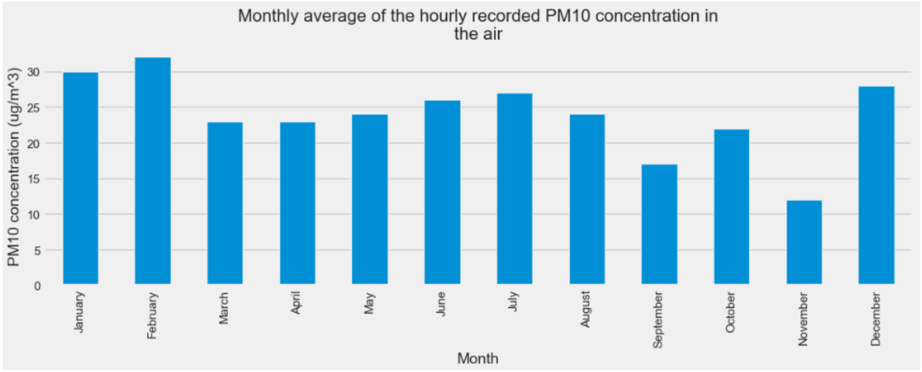


Figure 11-Monthly average of the hourly recorded for PM 10

At what time of the day do we expect the amount of AQI_US concentration in the air to be high

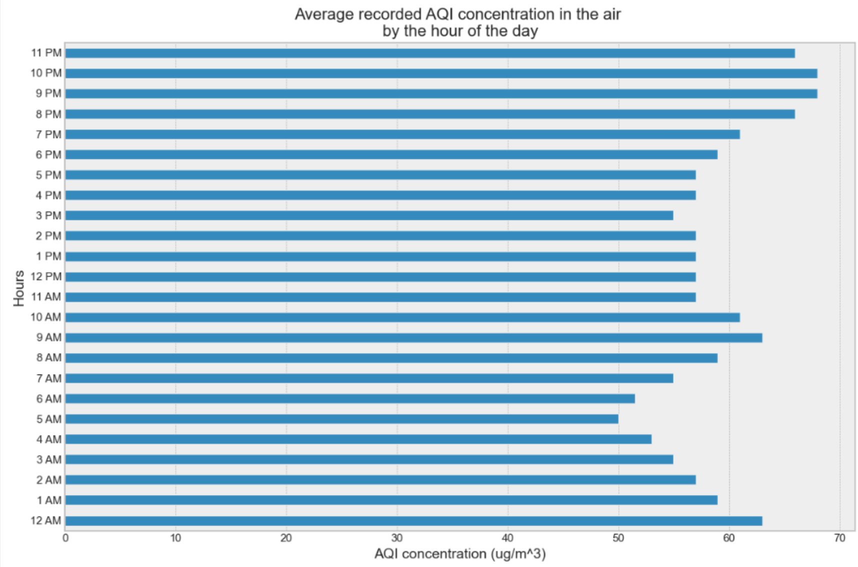


Figure 12 -Average recorded AQI concentration in the air by the hour of the day

The chart shows that the peak of AQI values occurs during the late evening hours, with an average value reaching approximately 70, resulting in a Moderate air quality rating. The Air Quality Index generally spans from normal values (0-50) to moderate values (51-100), where potential issues may be observed in sensitive groups of people. Problematic AQI values are those equal or greater than 101.

Average recorded air concentration elemenets by the hour of the day?

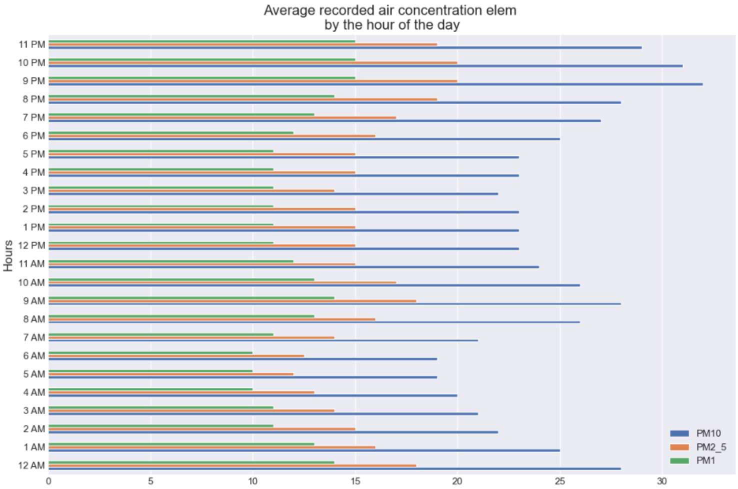
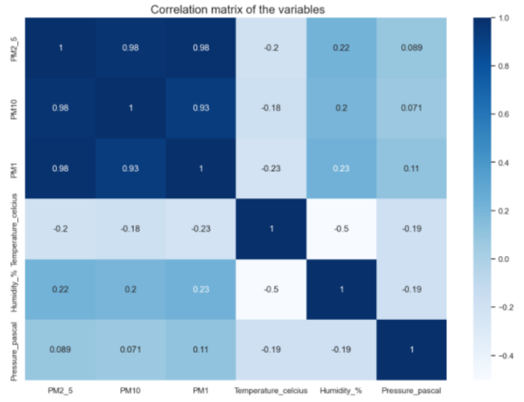


Figure 13-Average recorded air concentration elem by the hour of the day

How to visualize the correlation between the features of the data understanding if they affect the amount of PM2.5 concentration in the air?

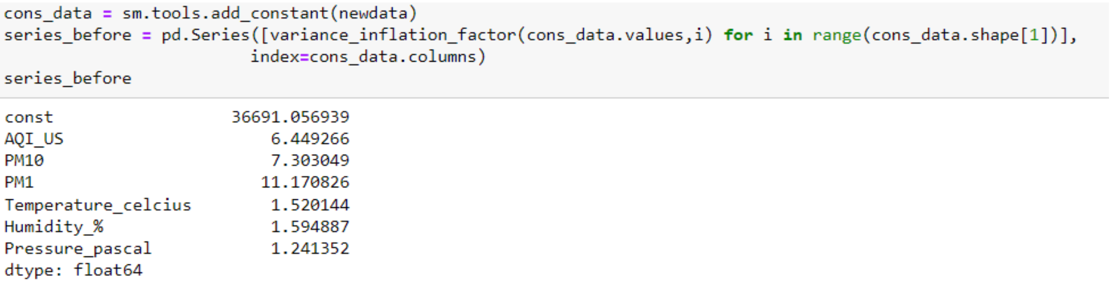


	PM2_5	PM10	PM1	Temperature_celcius	Humidity_%	Pressure_pascal
PM2_5	1	0.98	0.98	-0.2	0.22	0.089
PM10	0.98	1	0.93	-0.18	0.2	0.071
PM1	0.98	0.93	1	-0.23	0.23	0.11
Temperature_celcius	-0.2	-0.18	-0.23	1	-0.5	-0.19
Humidity_%	0.22	0.2	0.23	-0.5	1	-0.19
Pressure_pascal	0.089	0.071	0.11	-0.19	-0.19	1

Figure 14-Correlation matrix of the variables

EVALUATING AND FITTING OLS REGRESSIVE MODEL

To implement the OLS regression model, we need to address multicollinearity among variables and then fit a regression model using statsmodels. This involves using the variance inflation factor function from the statsmodels.stats.outliers_influence library. Before applying the variance inflation factor function, it's important to ensure that the dataframe contains only numerical data, has no missing values, and includes an intercept value for the data. If your DataFrame cons_data does not already include a column of ones (the intercept), this can cause an error which is added with statmodels's add_constant function.

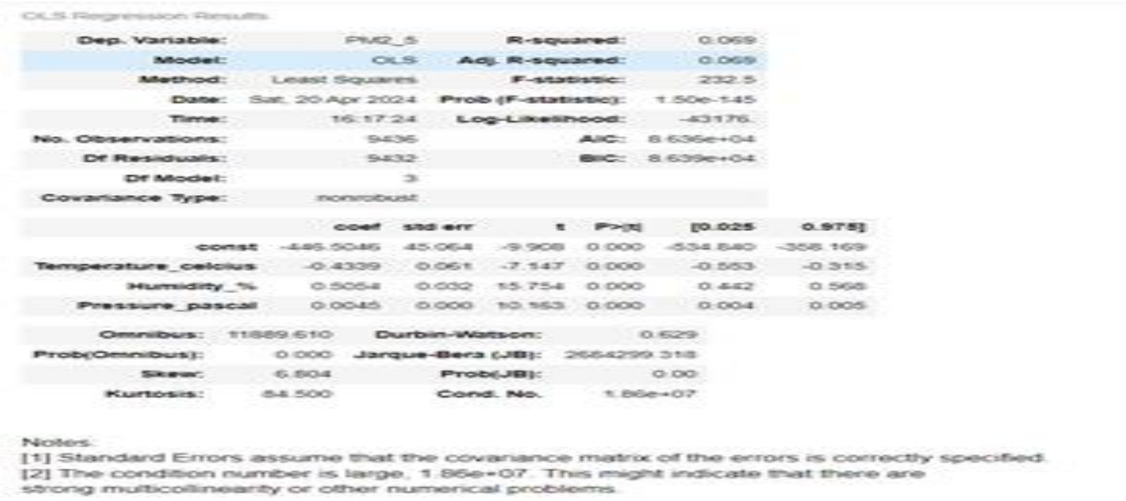


```
cons_data = sm.tools.add_constant(newdata)
series_before = pd.Series([variance_inflation_factor(cons_data.values,i) for i in range(cons_data.shape[1])],
                          index=cons_data.columns)
series_before
```

```
const                  36691.056939
AQI_US                     6.449266
PM10                       7.303049
PM1                       11.170826
Temperature_celcius        1.520144
Humidity_%                 1.594887
Pressure_pascal            1.241352
dtype: float64
```

Figure 15-Adding constant on dataframe and VIF

As can be seen from the figure AQI_US , PM1 and PM10 are values greater than 5 which are removed. Then consdata2 is created containing only the independent variables, while the variable y holds the data of PM2.5 and will be used as the predicted value.



```
OLS Regression Results
           Dep. Variable:            PM2_5          R-squared:      0.069
                  Model:              OLS      Adj. R-squared:      0.069
                 Method:    Least Squares          F-statistic:      232.5
                   Date: Sat, 20 Apr 2024  Prob (F-statistic):  1.50e-145
                   Time:         16:17:24      Log-Likelihood:     -43176.
       No. Observations:             9436                 AIC:  8.636e+04
           Df Residuals:             9432                 BIC:  8.639e+04
               Df Model:                3
        Covariance Type:        nonrobust

                          coef   std err        t   P>|t|    [0.025     0.975]
                const -445.5046   45.064   -9.908   0.000  -534.840   -356.169
  Temperature_celcius   -0.4339    0.061   -7.147   0.000    -0.553     -0.315
           Humidity_%    0.5054    0.032   15.754   0.000     0.442      0.568
      Pressure_pascal    0.0045    0.000   10.163   0.000     0.004      0.005

         Omnibus: 11889.610     Durbin-Watson:          0.629
  Prob(Omnibus):      0.000  Jarque-Bera (JB):  2654299.318
            Skew:      6.604         Prob(JB):          0.00
        Kurtosis:     84.500         Cond. No.      1.86e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.86e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 16-OLS regression model results

After analyzing the results [6], it is comprehensible that the OLS model from statsmodels gives us an accuracy of 6% (0.069) which is not satisfactory for prediction. Through this research it was understood that it is necessary to use other models in which the accuracy should be greater than 90%.

CONCLUSION AND FUTURE RESEARCH

Based on the results of this research, it is essential to evaluate the prediction using various regression models such as Linear Regression one of the most common and comprehensive statistical and machine learning algorithm [7]. Decision Tree Regressor fails to deal with linear relationships, Linear Decision Tree Regressor is build to overcome this limitation by combining the linear regressor with the decision tree regressor [8]. Research more about RandomForestRegressor, and GradientBoostingRegressor. Comparing these models will help determine the best method and draw consistent conclusions for future scientific endeavors.

REFERENCES

[1] Yu-Fei Xing, Yue-Hua Xu, Min-Hua Shi, and Yi-Xin Liancorresponding author *The impact of PM2.5 on the human respiratory system* J Thorac Dis. 2016 Jan; 8(1): E69–E74.

[2] Shau-Liang Chen, Sih-Wei Chang, Yen-Jen Chen & Hsuen-Li Chen *Possible warming effect of fine particulate matter in the atmosphere* Commun Earth Environ 2, 208 (2021).

[3] I. Lili, A. Kosta and E. Xhina. "The use of smart devices (IOT) to monitor the air quality: a case study at the Faculty of Natural Sciences ". ceur-ws.org. vol. 3402. pp. 89-86.

[4] ‘Imputation statistics’ (2024). Wikipedia. Available at Url *https://en.wikipedia.org/wiki/Imputation_(statistics)* (Accessed: 21.April.2024).

[5] P. Owusu, "Beijin Air quality (predicting PM2.5)", github.com, (Accessed: 21 Apr. 2024).

[6] Stuti Singh, “Interpretation results of OLS” URL: *https://medium.com/analytics-vidhya/how-to-interpret-result-from-linear-regression-3f7ae7679ef9* published 06.06.2020

[7] Maulud, Dastan & Mohsin Abdulazeez, Adnan. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends. 1. 140-147. 10.38094/jastt1457.

[8] Desai, Nihal & Patel, Vatsal. (2021). Linear Decision Tree Regressor: Decision Tree Regressor Combined with Linear Regressor.