Aman Samria	20110014
Nilanshi Patel	20110122
Priyanshu Sakhare	20110148
Riya Dhantoliya	20110168
Sankarshan Kulkarni	20110184
Shaandili Vajpai	20110186
Vikash Vishnoi	20110226

MA 202
Pre-Mid Semester Project

Calibrating Readings of Low-Cost Instruments to Obtain Accurate Concentrations of PM 2.5 in the Atmosphere

---

Contents:

Problem statement:

In this project, our main motive is to find the correction factor for the low-cost instrument for converting it into a high-cost device that measures the pollution levels of the gas molecules in the atmosphere. Finding an accurate measure of the pollution levels in the atmosphere is an essential component of any city development experiment. But wanting high accuracy data means having a high precision giving instrument, resulting in a high-cost mechanism. Therefore, we want to lower the cost and so wish to have a low-cost device that can perform the same job without compromising the quality of the readings obtained. Consequently, we are finding a regression model that will allow us to convert low-cost instrument readings into tasks that resemble those recorded by a high-cost instrument (which will henceforth serve as our reference instrument). Solving this problem opens up a door to cost-effectiveness, high accuracy data plotting, and regression modeling. Similar methods to measure air quality trends in cities have been used before.

We used two different devices to measure particle concentrations: one is a high-cost instrument, and the other is a low-cost instrument. The high-cost tools can measure particle concentrations briefly, resulting in more accurate findings. Compared to a high-cost mechanism, a low-cost device measures particle concentration over a more extended period.

Regression analysis is a collection of procedures used in statistical modeling to determine the connection between dependent and independent variables. The dependent variable is almost always a result, whereas the independent variable is almost always a prediction. The most frequent type of regression analysis is linear regression, in which a set of mathematical criteria is used to identify the line that best fits the data.

Method:

Step 1: Read file with data set: import numpy library to interpret and analyze data values, pandas to extract data, matplotlib.pyplot to read the data set file.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Step 2: Extract data into 3 different arrays: low-cost instrument concentration readings, reference instrument concentration readings, and merged array of the times at which the readings of the respective instruments.

```
import copy
Low_cost_data = copy.copy(dataset)[:268]
Low_cost_data.drop('Reference Instrument', inplace=True, axis=1)
Low_cost_data.drop('Time_2', inplace=True, axis=1)

High_cost_data = copy.copy(dataset)
High_cost_data.drop('Low Cost Instrument', inplace=True, axis=1)
High_cost_data.drop('Time_1', inplace=True, axis=1)

time = copy.copy(Low_cost_data["Time_1"])
FILTER_HIGH_COST_DATA = copy.copy(High_cost_data[High_cost_data.Time_2.isin(time)])

Low_cost_data.rename(columns={'Time_1':'Time'}, inplace = True)
High_cost_data.rename(columns={'Time_2':'Time'}, inplace = True)
Low_cost_data['Low Cost Instrument'] = Low_cost_data['Low Cost Instrument'].astype(int)
```

Step 3: Merge the concentration value arrays. Plot the concentration values vs the time values. Also plot the reference instrument readings vs the low-cost instrument readings. By observing the general trend of values from the graph, we decide to use the linear regression model to find a regression from low-cost instrument readings to reference instrument readings.

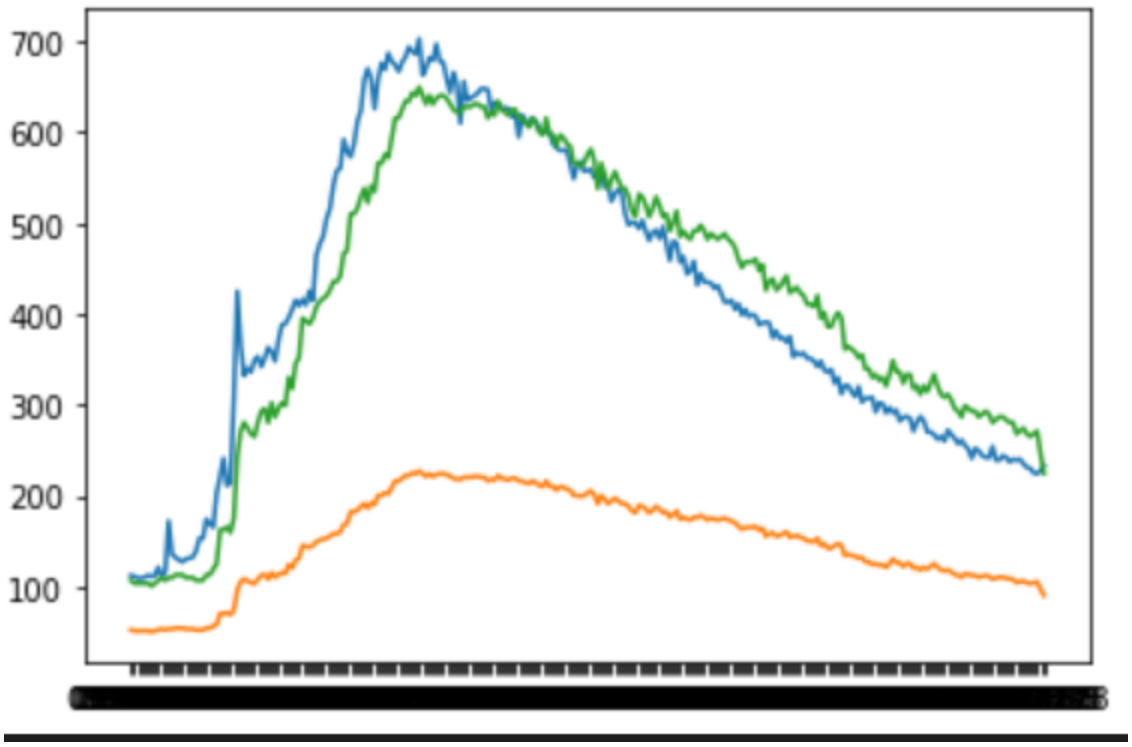

FIGURE 1.

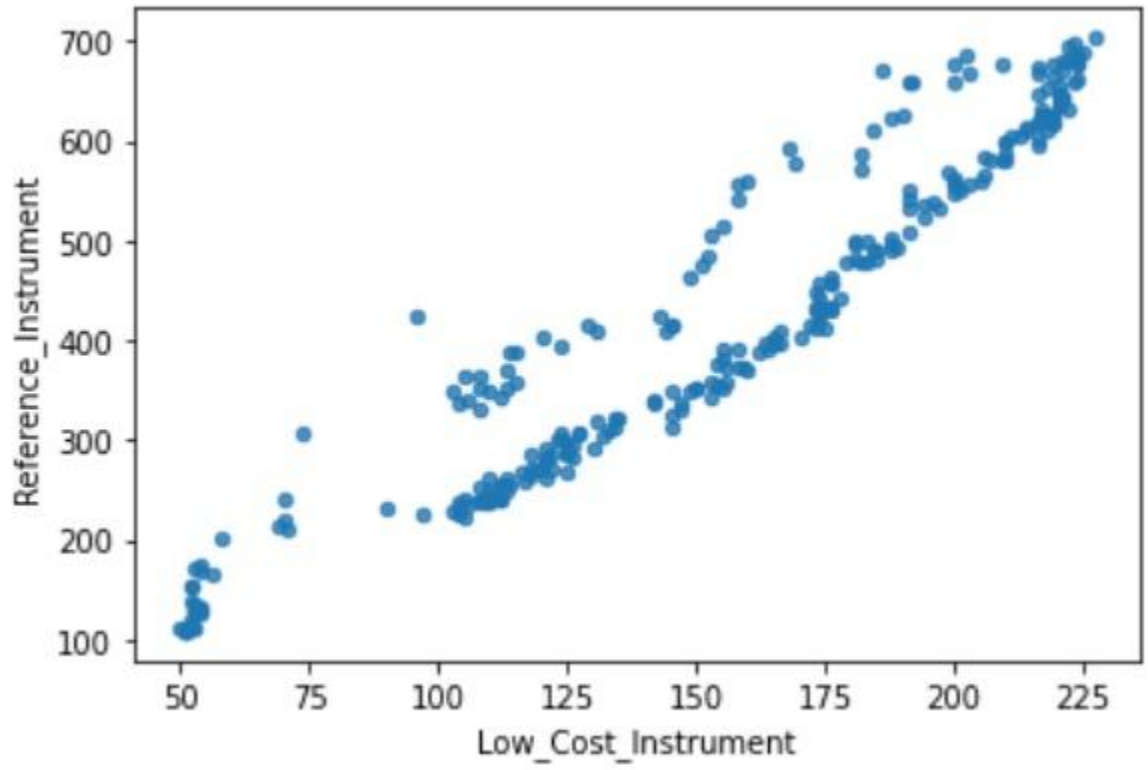

FIGURE 2.

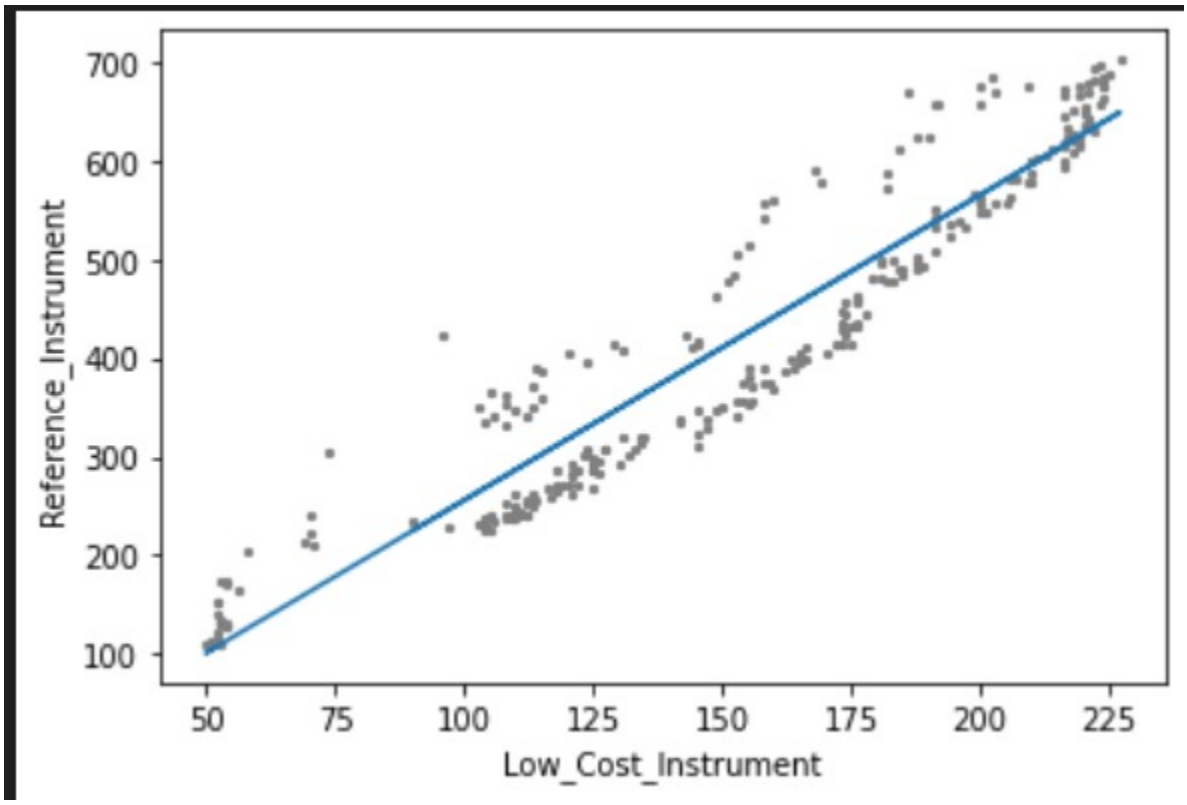

FIGURE 3.

Step 4: Find the correlation between the low-cost instrument readings and the reference instrument readings. This will define the strength of the relationship between the dependent (reference instrument readings) variable and independent (low-cost instrument readings) variable.

correlation_data= Merged_data.corr()

Ascend= correlation_data['Reference Instrument'].sort_values()
Time_axis = Merged_data["Time"]
RefIns = Merged_data["Reference Instrument"]
LowIns = Merged_data["Low Cost Instrument"]

plt.plot(Time_axis, RefIns)
plt.plot(Time_axis, LowIns)

Step 5: We then find the best-fit line (the regression line) on the reference vs low-cost instrument graph. To find the equation of the line, we choose two points on it and use the two-point method. The regression factor is the slope of the line, the square of which describes how much variations in the low-cost instrument readings affect variations in the reference instrument readings.

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(data, labels)
```

Step 6: Find the error in the regression by finding the mean square error. The readings of the instruments are calibrated thus:

```
from sklearn.metrics import mean_squared_error
rand_predictions = model.predict(data)
mse = mean_squared_error(labels, rand_predictions)
rmse = np.sqrt(mse)

t=Time_axis[5].split(':')
x1= int(t[0])*60+int(t[1])*1 +int(t[2])/60
y1 = Predictions[10]
t1=Time_axis[69].split(':')
x2= int(t1[0])*60+int(t1[1])*1 +int(t1[2])/60
y2 = Predictions[100]
m = (y2-y1) / (x2-x1)
c = y1 - m*x1

print("Equation : y = (", m, ")*x + (", c, ")")
```

y = Reference instrument readings
x = Low-cost instrument readings
m (sometimes also written as R) = regression factor
c = intercept of the best-fit line
Ɛ = error (mean square error)

$$y = mx + c + \mathcal{E}$$

By substituting the values that we have obtained, our final regression equation is:

$$y = 8.202x - 4771.08573$$

Coherence of :
Low Cost Instrument: 0.949516
High Cost Instrument: 1
Root Mean Square Error: 52.7929787

Scope for Improvement (Potential Problem Areas):

To improve this model, We can use multiple regression methods in place of the linear regression method, which is used in our model. Because if we have more than two variables(in this case, we have the efficiency of the low-cost instrument and efficiency of a high-cost instrument as dependent and independent variables, respectively), if we had one more extra variable, such as the efficiency of average cost instrument, then it will be difficult for us to find the error factor. Also, in the multiple regression method, the chances of failure are reduced.

Conclusion:

According to the studies about the devices (PM2.5 monitoring devices), the expensive devices measure data more accurately, and the low-cost device measures the data with error. And obtaining expensive instrumentation to monitor air quality can be costly, so we decide to use the linear regression model to find a regression from low-cost instrument readings to reference instrument readings.
Using this regression model, we get approximately the same quality of data as the reference device.

References:

https://www.investopedia.com/terms/r/r-squared.asp#:~:text=R%2DSquared%20is%20a%20statistical,s

Previous studies that have used similar methods to calibrate low-cost instruments:

https://www.sciencedirect.com/science/article/pii/S0269749115000160

https://www.sciencedirect.com/science/article/pii/S0021850221005644