Project\_CIS603

Saikiran Jelloji

2023-05-07

## Summary

The project has included more of visualization to understand the distribution of main air pollutants, sulphur and carbon, in urban areas. The project has considered different cities and main air pollutants for each city. The project has included a multiple linear regression model that can be used for predicting the occurence of air pollution from the main air pollutants like the sulphur and carbon. The project has illustrated that sulphur is the leading air pollutant in the urban areas. The p-value of the linear regression model has illustrated that the SO2.Mean, CO.Mean, and NO2.Mean are not statistically significant to our model.

## Introduction

Air pollution has become a serious challenge in urban areas. This is because air pollution is associated with diverse health issues like cancer, a deadly disease, or heart disease. Due to the high number of individuals, industries, and traffic, urban areas have been identified as the most vulnerable to air pollution. Therefore, my project tries to identify the factors that are highly causing air pollution in urban areas and take the right actions. Therefore, the importance of this project provides the advantage of identifying the main causes of air pollution and also forecast the levels of air pollution in urban areas. With the identification and forecasting, it will be helpful to create awareness among people on taking precautions that would be vital to avoid being exposed to air pollutants. In addition, the project is important to the environmental regulatory boards in implementing regulations that might help reduce air pollution in urban areas.

## Literature Review

There have been research papers that have shown interest in the predictions of the levels of air pollution by the use of techniques associated with statistical models. The research paper by Mustafaraj et al. (2011) has implemented a technique that used both auto regressive and neural network models to predict air pollution levels. Another study by Qin & Gong (2022) incorporated the use of machine learning methods for the predictions of Carbon dioxide levels in China.

## Research Question

The research question for the project is:

1.What is the data analysis method or technique applicable in R for predicting the levels of air pollution?

## Data

The data used in the project has been obtained from sites that store data associated with environmental factors. The data contains the nitrogen and carbon dioxide levels variables that are measured in the air and information on the temperature, humidity, or wind speed.

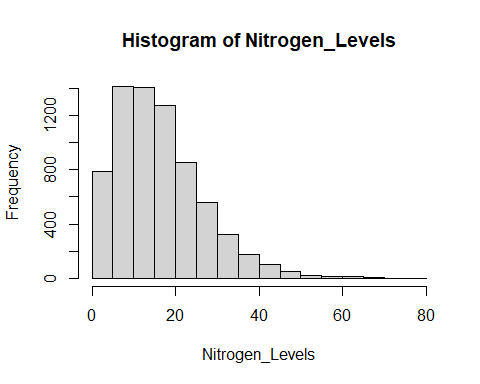
## X State.Code County.Code Site.Num  
## 1 118048 6 73 1  
## 2 510823 42 7 14  
## 3 470000 6 73 1  
## 4 265886 42 101 4  
## 5 108040 6 37 4002  
## 6 504742 36 81 124  
## Address State County  
## 1 80 E. 'J' ST., CHULA VISTA California San Diego  
## 2 EIGHT STREET AND RIVER ALLEY Pennsylvania Beaver  
## 3 80 E. 'J' ST., CHULA VISTA California San Diego  
## 4 1501 E. LYCOMING AVE. Pennsylvania Philadelphia  
## 5 3648 N. LONG BEACH BLVD., LONG BEACH California Los Angeles  
## 6 Queens College 65-30 Kissena Blvd Parking Lot#6 New York Queens  
## City Date.Local NO2.Units NO2.Mean NO2.1st.Max.Value  
## 1 Chula Vista 17-Aug-01 Parts per billion 8.347826 16  
## 2 Beaver Falls 27-Oct-05 Parts per billion 16.166667 23  
## 3 Chula Vista 21-Oct-05 Parts per billion 15.695652 26  
## 4 Philadelphia 06-Dec-02 Parts per billion 24.333333 50  
## 5 Long Beach 07-Aug-01 Parts per billion 9.826087 15  
## 6 New York 08-Aug-05 Parts per billion 27.375000 49  
## NO2.1st.Max.Hour NO2.AQI O3.Units O3.Mean O3.1st.Max.Value  
## 1 6 15 Parts per million 0.037333 0.048  
## 2 19 22 Parts per million 0.006792 0.015  
## 3 9 25 Parts per million 0.042208 0.053  
## 4 6 47 Parts per million 0.036708 0.053  
## 5 5 14 Parts per million 0.024167 0.033  
## 6 22 46 Parts per million 0.021958 0.037  
## O3.1st.Max.Hour O3.AQI SO2.Units SO2.Mean SO2.1st.Max.Value  
## 1 9 41 Parts per billion 2.885714 3.3  
## 2 8 13 Parts per billion 4.062500 5.0  
## 3 12 45 Parts per billion 3.086957 4.0  
## 4 9 45 Parts per billion 7.237500 16.6  
## 5 9 28 Parts per billion 0.652174 5.0  
## 6 11 31 Parts per billion 4.475000 9.6  
## SO2.1st.Max.Hour SO2.AQI CO.Units CO.Mean CO.1st.Max.Value  
## 1 11 NA Parts per million 0.479167 0.6  
## 2 11 NA Parts per million 0.412500 0.5  
## 3 14 6 Parts per million 0.637500 0.7  
## 4 8 NA Parts per million 0.508333 0.6  
## 5 9 7 Parts per million 0.273913 0.3  
## 6 8 NA Parts per million 0.600000 0.8  
## CO.1st.Max.Hour CO.AQI Respiratory.Diseases Respiratory.disease.Dummy  
## 1 12 7 1 Yes  
## 2 21 6 1 Yes  
## 3 7 8 1 Yes  
## 4 9 7 1 Yes  
## 5 0 NA 1 Yes  
## 6 22 NA 1 Yes

Check for if there are any missing values in the data.

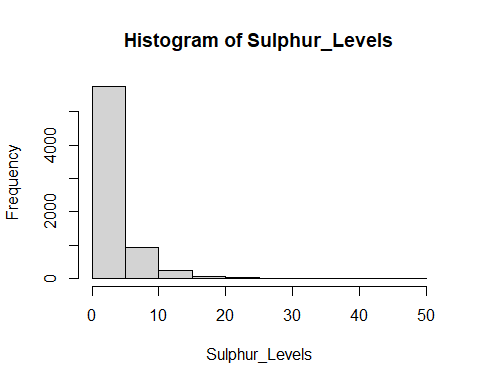
## X State.Code County.Code   
## 0 0 0   
## Site.Num Address State   
## 0 0 0   
## County City Date.Local   
## 0 0 0   
## NO2.Units NO2.Mean NO2.1st.Max.Value   
## 0 0 0   
## NO2.1st.Max.Hour NO2.AQI O3.Units   
## 0 0 0   
## O3.Mean O3.1st.Max.Value O3.1st.Max.Hour   
## 0 0 0   
## O3.AQI SO2.Units SO2.Mean   
## 0 0 0   
## SO2.1st.Max.Value SO2.1st.Max.Hour SO2.AQI   
## 0 0 3498   
## CO.Units CO.Mean CO.1st.Max.Value   
## 0 0 0   
## CO.1st.Max.Hour CO.AQI Respiratory.Diseases   
## 0 3578 0   
## Respiratory.disease.Dummy   
## 0

## Visualizations

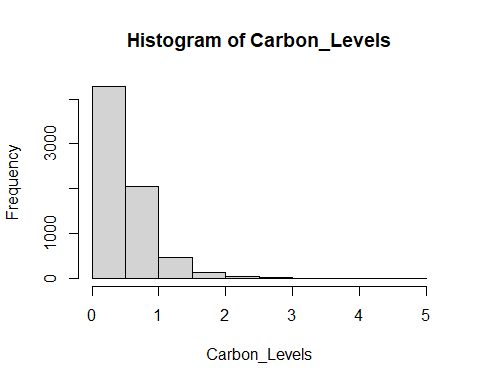
## Histogram for Mean Nitrogen Levels



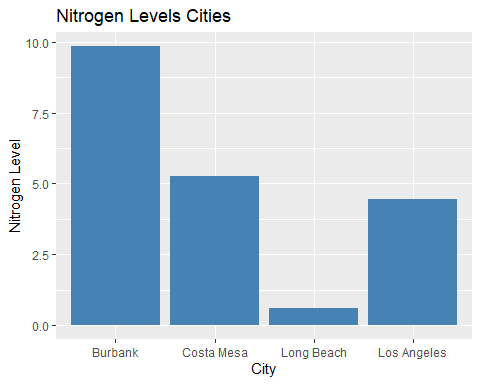
## Histogram for Mean Suphur Levels



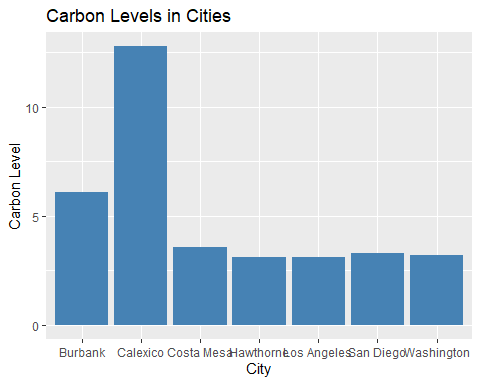
## Histogram for Carbon Levels



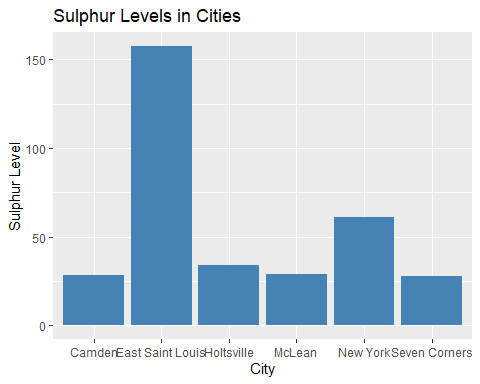
## Barchart fot Nitrogen Levels in Cities



## Barchart fot Carbon Levels in Cities



## Barchart fot Sulphur Levels in Cities

 ## Analysis

Split Dataset to 80% training set and 20% testing set.

## data\_splitted  
## 1 2   
## 5593 1407

The training set has 5593 observations and testing set has 1407 observations. ## Build the predicting Linear Regression model

##   
## Call:  
## lm(formula = Respiratory.Diseases ~ SO2.Mean + CO.Mean + NO2.Mean,   
## data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.12678 -0.09599 -0.09119 -0.08682 0.92624   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0868051 0.0073529 11.806 <2e-16 \*\*\*  
## SO2.Mean 0.0006991 0.0011552 0.605 0.545   
## CO.Mean -0.0154271 0.0130169 -1.185 0.236   
## NO2.Mean 0.0007094 0.0005185 1.368 0.171   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2902 on 5589 degrees of freedom  
## Multiple R-squared: 0.0005507, Adjusted R-squared: 1.42e-05   
## F-statistic: 1.026 on 3 and 5589 DF, p-value: 0.3796

The equation of the linear regression model for predicting the probability of respiratory diseases through the three main air pollutants, sulphur, carbon, and nitrogen can be formulated as: Predicted= 0.086+0.0006991\* SO2.Mean- 0.01542\* CO.Mean+0.007094*NO2.Mean. For example: Predicted= 0.086+0.0006991* 10- 0.01542\* 20+0.007094\*15 If answer is: 1 represents yes there is respiratory disease for example lung cancer 0 represents no respiratory disease .

## Conclusion

The project has been successfully performed and it has been clear on the main air pollutants in urban areas and in different cities. Sulphur is main air pollutant in urban areas. Also the linear regression model have shown that the predictor variables are not statistically significant to our model (Maverick, 2023).

## References

Maverick. (2023). Learn to Predict Using Linear Regression in R With Ease (Updated 2023). Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/12/predicting-using-linear-regression-in-r/> Mustafaraj, G., Lowry, G., & Chen, J. (2011). Prediction of room temperature and relative humidity by autoregressive linear and nonlinear neural network models for an open office. Energy and Buildings, 43(6), 1452-1460. Qin, J., & Gong, N. (2022). The estimation of the carbon dioxide emission and driving factors in China based on machine learning methods. Sustainable Production and Consumption, 33, 218-229.