Geoffrey Leigh

Prof. Jaume

Capstone: DS670

May 12, 2017

Mid-term

Abstract:

The aim of my project is to do time series forecasting of pollution based on the station data. The forecasting involves predicting each pollutants value for a given time and location. The Geo-spatial data and Temporal data is made to full use by applying Exponential or ARIMA models. The method will also reveal trend and seasonality associated with each pollutant such as Carbon monoxide, ozone, Sulphur dioxide, nitrous oxide. Thus, the industry use case will be to forecast the pollutant value based on location so we can find when there will be increase in ozone or carbon monoxide with respect to time and location. Once, we know the forecasted value of the pollutant, appropriate steps can be taken to decrease the pollution and maintaining the safety limit throughout the year. The main advantage of this algorithm is that we can isolate which pollutant value will rise and take precautionary steps to tackle that pollutant without worrying about the others.

**Contribution**:

My contribution will be to predict the pollution levels from any of the 449 stations for the following week. The response will be binary to keep it simple and it will state if the levels are above (unsafe) or bellow (safe). The time will be broken in three periods morning, midday and Night. It will be in 8 hour intervals. In theory, the goal is to have a high prediction rate for the following week using times series analysis. This algorithm could be useful in predicting hotspots of pollution and places to avoid during times of the day where there are high unsafe levels.

**State of the Art**:

* [Air pollution and health in urban areas.](http://europepmc.org/abstract/med/9654797?europe_pmc_abs_relarticles_review_redirect=/abstract/MED/10939084)  
  Schwela D   
  Reviews on environmental health [2000 Jan-Jun;15(1-2):13-42]

Air contamination information show attributes like long range relationships and multi-fractal scaling that can be misused to execute a vitality effective, versatile spatial inspecting method for contamination sensor hubs. In this work, we exhibit a) results from de-slanted change investigation to demonstrate the nearness of non-straight flow in genuine contamination datasets assembled from trials did in Cyprus, b) a novel Multi-scale Nearest Neighbors based Adaptive Spatial Sampling (MNNASS) method that decides the consistency and thusly the directional impacts between information from various sensor hubs, and c) execution examination of the calculation as far as vitality reserve funds and estimation exactness.

* [Use of an index to reflect the aggregate burden of long-term exposure to criteria air pollutants in the United States.](http://europepmc.org/abstract/med/9654797?europe_pmc_abs_relarticles_review_redirect=/abstract/MED/11834467)  
  Kyle AD, Woodruff TJ, Buffler PA, Davis DL   
  Environmental health perspectives [2002 Feb;110 Suppl 1:95-102]

The improvement of models to evaluate air contamination exposures inside urban communities for task to subjects in wellbeing contemplates has been recognized as a need range for future research. This paper surveys models for evaluating intraurban introduction under six classes, including: (i) nearness based appraisals, (ii) factual addition, (iii) arrive utilize relapse models, (iv) line scattering models, (v) coordinated discharge meteorological models, and (vi) half and half models consolidating individual or family unit presentation checking with one of the first strategies. They advanced this audit of the displaying methods and results with connected cases from Hamilton, Canada. Also, we subjectively assess the models in light of key criteria imperative to wellbeing impacts evaluation explore. Mixture models seem appropriate to beating the issue of accomplishing populace delegate tests while understanding the part of presentation variety at the individual level. Remote detecting and activity–space investigation will supplement refinements in prior techniques, and with expected advances, the field of introduction appraisal may lessen logical instabilities that now obstruct arrangement intercession went for ensuring general wellbeing.

* [Review of air pollution and health impacts in Malaysia.](http://europepmc.org/abstract/med/9654797?europe_pmc_abs_relarticles_review_redirect=/abstract/MED/12854685)  
  Afroz R, Hassan MN, Ibrahim NA   
  Environmental research [2003 Jun;92(2):71-7]

Bit based neural systems are famous machine learning approach with numerous fruitful applications. Regularization systems speak to their exceptional subclass with strong hypothetical foundation and an assortment of learning conceivable outcomes. In this paper, they concentrate on single and multi-piece units, specifically, we depict the design of an item unit arrange, and portray a developmental learning calculation for setting its parameters including diverse parts from a word reference, and ideal split of contributions to individual items. The approach is tried on genuine information from alignment of air-contamination sensor systems, and the execution is contrasted with a few diverse relapse devices.

**Limitation**

They concentrated too much on single and multi-piece units, with neural network using the parameters in a way it is difficult to interpret the usages of variables in the model. That is, the model acted as a black box where the variables used are unknown significance.

* [The potential impacts of climate variability and change on air pollution-related health effects in the United States.](http://europepmc.org/abstract/med/9654797?europe_pmc_abs_relarticles_review_redirect=/abstract/MED/11359687)  
  Bernard SM, Samet JM, Grambsch A, Ebi KL, Romieu I   
  Environmental health perspectives [2001 May;109 Suppl 2:199-209]

We have built up a displaying framework for anticipating the activity volumes, emanations from stationary and vehicular sources, and barometrical scattering of contamination in a urban region. A friend paper addresses show improvement and its applications. This paper portrays an examination of the anticipated NOx and NO2 fixations with the consequences of a urban air quality checking system. We played out a factual investigation concerning the assention of the anticipated and measured hourly time arrangement of fixations, at four observing stations in the Helsinki metropolitan range in 1993. The anticipated and measured NO2 focuses concurred well at all the stations considered. The understanding of model expectations and estimations for NOx and NO2 was better for the two rural checking stations, contrasted and the two urban stations, situated in downtown Helsinki.

* [Health effects of outdoor air pollution.](http://europepmc.org/abstract/med/9654797?europe_pmc_abs_relarticles_review_redirect=/abstract/MED/8172037)  
  Neher JO, Koenig JQ   
  American family physician [1994 May 1;49(6):1397-404, 1407-8]

Albeit some accord has risen among the logical and administrative groups that the urban encompassing climatic blend of ignition related toxins is a determinant of populace wellbeing, the relative harmfulness of the concoction and physical segments of this perplexing blend stays vague. Day by day death rates and simultaneous information on sizefractionated particulate mass and vaporous contaminations were gotten in eight of Canada's biggest urban areas from 1986 to 1996 comprehensive with a specific end goal to inspect the relative danger of the segments of the blend of surrounding air poisons to which Canadians are uncovered. Positive and measurably huge affiliations were seen between day by day varieties in both gas-and particulate-stage contamination and day by day changes in death rates. The relationship between air contamination and mortality couldn't be clarified by temporalvariation in either death rates or climate elements. Fine particulate mass (under 2.5 μm in normal aerometric measurement) was a more grounded indicator of mortality than coarse mass (in the vicinity of 2.5 and 10 μm). Measure fractionated particulate mass clarified 28% of the aggregate wellbeing impact of the blend, with the rest of the impacts represented by the gasses. Forty-seven essential focuses were gotten for the fine and coarse division utilizing nondestructive x-beam fluorescence procedures. Sulfate focuses were gotten by particle chromatography. Sulfate particle, iron, nickel, and zinc from the fine portion were most unequivocally connected with mortality. The aggregate impact of these four segments was more noteworthy than that for fine mass alone, proposing that the attributes of the perplexing substance blend in the fine division perhaps a superior indicator of mortality than mass alone. However,the variety in the impacts of the constituents of the fine division between urban communities was more noteworthy than the variety in the mass impact, inferring that there are extra lethal segments of fine particulate matter not analyzed in this review whose fixations and impacts change between areas. One of these segments, carbon, speaks to a large portion of the mass of fine particulate matter. We prescribe that estimations of essential and organiccarbon be embraced in Canadian urban situations to look at their potential impacts on human wellbeing.

* [[Asthma and household chemical pollutants (with the exception of tobacco)].](http://europepmc.org/abstract/med/9654797?europe_pmc_abs_relarticles_review_redirect=/abstract/MED/9551510)  
  Krieger P, de Blay F, Pauli G, Kopferschmitt MC   
  Revue des maladies respiratoires [1998 Feb;15(1):11-24]

As a component of the EU-subsidized SAVIAH extend, a relapse based technique for mapping activity related air contamination was created inside a GIS situation. Mapping was done for NO2 in Amsterdam, Huddersfield and Prague. In every middle, studies of NO2, as a marker for activity related contamination, were led utilizing uninvolved dissemination tubes, uncovered for four 2-week time frames. A GIS was additionally settled, containing information on observed air contamination levels, street arrange, activity volume, arrive cover, height and other, privately decided, highlights. Information from 80 of the observing locales were then used to build a relapse condition, on the premise of indicator ecological factors, and the subsequent condition used to guide air contamination over the review zone. The precision of the guide was then evaluated by contrasting anticipated contamination levels and observed levels at a scope of free reference locales. Comes about demonstrated that the guide created to a great degree great expectations of observed contamination levels, both for individual reviews and for the mean yearly focus, with r2 0.79-0.87 crosswise over 8-10 reference focuses, however the precision of forecasts for individual study periods was more factor. In Huddersfield and Amsterdam, additionally observing likewise demonstrated that the contamination delineate solid assessments of NO2 fixations in the next year

* [Exercise and outdoor ambient air pollution.](http://europepmc.org/abstract/med/9654797?europe_pmc_abs_relarticles_review_redirect=/abstract/MED/11477012)  
  Carlisle AJ, Sharp NC   
  British journal of sports medicine [2001 Aug;35(4):214-22]

The reason for the present paper is to propose a methodical way to deal with the relapse examinations that are key to this sort of research. We contend that the outcomes may rely on upon various specially appointed components of the investigation, including which meteorological factors to alter for, and the way in which diverse slacked estimations of particulate matter are joined into a solitary presentation measure". We likewise analyze the topic of whether the impacts are direct or nonlinear, with specific consideration regarding the likelihood of a limit impact", i.e. that signicant impacts happen just over some edge.

These focuses are outlined with an informational index from Birmingham, Alabama, referred to by Schwartz (1993) and since broadly re-broke down. For this informational index, we and that the outcomes are touchy to whether moistness is incorporated alongside temperature as a meteorological variable, and to the denition of the introduction measure. We likewise and proof of a limit impact, with the best increment in mortality occuring over 50 g/m3, which is the long haul normal level allowed by the current NAAQS. In this manner, on the premise of this informational collection, the requirement for a more tightly NAAQS is not set up.

Despite the fact that this specific investigation is focussed just on one informational index, the issues it raises are average around there of research. We don't debate that there is a sensible level of proof connecting climatic particulate matter with unfavorable wellbeing results even inside the levels allowed by current directions. Nonetheless, the impression has been made by a portion of the distributed writing that such affiliations are overwhelmingly upheld by epidemiological research. Our perspective is that the measurable investigations permit diverse elucidations, and that the case for more tightly controls can't be construct exclusively in light of investigations of this nature.

**Limitation**

The more extensive ramifications for particulate matter and wellbeing is that in a regular informational collection, there are many issues that should be considered before a decision of a causal impact can be drawn. Rough examinations that don't consider conceivable option elucidations of the information are of restricted an incentive with regards to an open verbal confrontation over contamination controls

* [Pulmonary effects of air pollution.](http://europepmc.org/abstract/med/9654797?europe_pmc_abs_relarticles_review_redirect=/abstract/MED/4898543)  
  Wolkonsky PM   
  Archives of environmental health [1969 Oct;19(4):586-92]

Investigations of air contamination and human wellbeing have developed from graphic investigations of the early marvels of extensive increments in unfriendly wellbeing impacts taking after outrageous air contamination scenes to time-arrangement investigations in light of the utilization of advanced relapse models. Truth be told, progressed measurable techniques are important to address the difficulties inalienable in the location of a generally little contamination hazard within the sight of potential confounders. This paper surveys the history, techniques, and discoveries of the time-arrangement thinks about assessing wellbeing dangers related with here and now introduction to particulate matter (PM), however a significant part of the dialog is pertinent to epidemiological investigations of air contamination when all is said in done. We audit the basic part of epidemiological reviews in setting administrative models and the historical backdrop of PM the study of disease transmission and time-arrangement investigation.

We additionally compress late time-arrangement comes about and finish up with an examination of ebb and flow and future headings of time-arrangement investigation of particulates, including research on mortality relocation and the determination of results from accomplice and time-arrangement ponders.

**Limitation**

Two of the crucial research concerns are the need to determine what properties of PM are associated with greater risk for adverse health effects and to better understand the implications of using ambient indicators as a surrogate for exposure. Other critical research needs are further characterization of the effects on the general population and susceptible subpopulations, emissions sources, deposition and fate of PM in the respiratory tract, combined effects of PM and other pollutants, and biological mechanisms. An important assumption in many epidemiological studies of air pollution is a consistent, harmful effect from PM, as defined by size, without attention to the heterogeneous nature of PM. This heterogeneity hampers comparison across studies for different locations, because differences in risk estimates may be due to a difference in PM characteristics

**Methods**:

Linear Model: Linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

ARIMA Model: ARIMA models are the most general class of models for forecasting a time series which can be made to be “stationary” by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging or deflating (if necessary). A random variable that is a time series is stationary if its statistical properties are all constant over time.  *A stationary series has no trend, its variations around its mean have a constant amplitude, and it wiggles in a consistent fashion*, i.e., its short-term random time patterns always look the same in a statistical sense.  The latter condition means that its *autocorrelations* (correlations with its own prior deviations from the mean) remain constant over time, or equivalently, that its *power spectrum* remains constant over time.  A random variable of this form can be viewed (as usual) as a combination of signal and noise, and the signal (if one is apparent) could be a pattern of fast or slow mean reversion, or sinusoidal oscillation, or rapid alternation in sign, and it could also have a seasonal component.  An ARIMA model can be viewed as a “filter” that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts.

 A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. Time series *analysis* comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series *forecasting* is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time

The data consists of the following columns

1. Ozone
2. particullate\_matter
3. carbon\_monoxide
4. sulfure\_dioxide
5. nitrogen\_dioxide
6. longitude
7. latitude
8. timestamp

**Latitude and Longitude :**

The Latitude and Longitude variables show the place where the readings have been recorded. This helps us pin point places where there is less pollution or more points and concentrate more on these places looking for reasons.

**Timestamp** :

The timestamp helps us to know when was the reading recorded. This variable is useful in giving insights about the behavior of pollution over the years or forecast how it will be in the future.

**Ozone** :

Ozone is a particle of three oxygen molecules bound together (O3). It is flimsy and exceptionally receptive. Ozone is utilized as a detergent, an aerating specialist, and a cleansing operator for air and drinking water. At low fixations, it is dangerous. Despite the fact that ozone contamination is shaped primarily in urban and rural territories, it winds up in rustic ranges too, conveyed by winning winds or coming about because of autos and trucks that go into country zones. Noteworthy levels of ozone contamination can be identified in provincial regions to the extent 250 miles (402 kilometers) downwind from urban modern zones.

**Carbon Monoxide:**

CO is a dismal, scentless gas shaped by the fragmented response of air with fuel. CO contamination happens fundamentally from emanations delivered by fossil fuel–powered motors, including engine vehicles and non-street motors and vehicles, (for example, development hardware and vessels). Larger amounts of CO for the most part happen in ranges with substantial movement blockage. Different wellsprings of CO emanations incorporate mechanical procedures, (for example, handling of metals and concoction fabricating), private wood blazing, and common sources, for example, woodland fires. Woodstoves, gas stoves, tobacco smoke, and unvented gas and lamp fuel space radiators are indoor wellsprings of CO. The most elevated amounts of CO ordinarily happen amid the colder months of the year when reversal conditions (when the air contamination gets to be distinctly caught close to the ground underneath a layer of warm air) are more incessant.

CO can bring about hurtful wellbeing impacts by diminishing oxygen conveyance to the body's organs and tissues. Introduction to lower levels of CO is most genuine for the individuals who experience the ill effects of coronary illness, and can bring about trunk torment, lessen the capacity to work out, or—with rehashed exposures—may add to other cardiovascular impacts.

Indeed, even solid individuals can be influenced by elevated amounts of CO. Individuals who inhale large amounts of CO can create vision issues, lessened capacity to work or learn, decreased manual aptitude, and trouble performing complex assignments. At abnormal states, CO is noxious and can bring about death.

**Sulphur Dioxide :**

Sulfur dioxide is a gas. It is imperceptible and has a terrible, sharp smell. It responds effortlessly with different substances to frame hurtful mixes, for example, sulfuric corrosive, sulfurous corrosive and sulfate particles. Around 99% of the sulfur dioxide in air originates from human sources. The fundamental wellspring of sulfur dioxide noticeable all around is modern action that procedures materials that contain sulfur, eg the era of power from coal, oil or gas that contains sulfur. Some mineral metals likewise contain sulfur, and sulfur dioxide is discharged when they are handled. What's more, modern exercises that blaze fossil energizes containing sulfur can be essential wellsprings of sulfur dioxide. Sulfur dioxide is likewise present in engine vehicle emanations, as the consequence of fuel ignition.

Sulfur dioxide affects human health when it is breathed in. It irritates the nose, throat, and airways to cause coughing, wheezing, shortness of breath, or a tight feeling around the chest. The effects of sulfur dioxide are felt very quickly and most people would feel the worst symptoms in 10 or 15 minutes after breathing it in. Those most at risk of developing problems if they are exposed to sulfur dioxide are people with asthma or similar conditions.

**Nitrogen Dioxide :**

Nitrogen dioxide is a frightful noticing gas. Some nitrogen dioxide is shaped actually in the air by lightning and some is created by plants, soil and water. Be that as it may, just around 1% of the aggregate sum of nitrogen dioxide found in our urban communities' air is framed thusly. Nitrogen dioxide is an imperative air poison since it adds to the development of photochemical brown haze, which can impacts and affect human wellbeing.

The main effect of breathing in raised levels of nitrogen dioxide is the increased likelihood of respiratory problems. Nitrogen dioxide inflames the lining of the lungs, and it can reduce immunity to lung infections. This can cause problems such as wheezing, coughing, colds, flu and bronchitis.

Increased levels of nitrogen dioxide can have significant impacts on people with asthma because it can cause more frequent and more intense attacks. Children with asthma and older people with heart disease are most at risk.

**References**

Wang, W., & Guo, Y. (2009, October). Air pollution PM2. 5 data analysis in Los Angeles long beach with seasonal ARIMA model. In *Energy and Environment Technology, 2009. ICEET'09. International Conference on* (Vol. 3, pp. 7-10). IEEE.

Kingsy, G. R., Manimegalai, R., Geetha, D. M., Rajathi, S., Usha, K., & Raabiathul, B. N. (2016, November). Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data. In *Region 10 Conference (TENCON), 2016 IEEE* (pp. 1945-1949). IEEE.

Gupta, P., Kandakatla, K., De, S., & Jana, S. (2013, February). Feasibility analysis on integrated recharging and data collection in pollution sensor networks. In *Communications (NCC), 2013 National Conference on* (pp. 1-5). IEEE.

Zhang, Q., Xu, Y., Xu, X., & Lou, Z. (2011, June). Application of Bayesian network in water quality risk analysis and pollution reduction decision making from small data. In *Remote Sensing, Environment and Transportation Engineering (RSETE), 2011 International Conference on* (pp. 84-88). IEEE.

Jain, R., & Shah, H. (2016, October). An anomaly detection in smart cities modeled as wireless sensor network. In *Signal and Information Processing (IConSIP), International Conference on* (pp. 1-5). IEEE.

Ali, S., Tirumala, S. S., & Sarrafzadeh, A. (2014, December). SVM aggregation modelling for spatio-temporal air pollution analysis. In *Multi-Topic Conf*

Ojeda-Magaña, B., Cortina-Januchs, M. G., Barrón-Adame, J. M., Quintanilla-Domínguez, J., Hernandez, W., Vega-Corona, A., ... & Andina, D. (2010, March). Air pollution analysis with a pfcm clustering algorithm applied in a real database of salamanca (mexico). In *Industrial Technology (ICIT), 2010 IEEE International Conference on* (pp. 1297-1302). IEEE.

Baralis, E., Cerquitelli, T., Chiusano, S., Garza, P., & Kavoosifar, M. R. (2016, May). Analyzing air pollution on the urban environment. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016 39th International Convention on* (pp. 1464-1469). IEEE.

Yoon, H. J., Xu, S., & Tourassi, G. (2016, February). Predicting lung cancer incidence from air pollution exposures using shapelet-based time series analysis. In *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on* (pp. 565-568). IEEE.

Dicken, R. A., Rubby, S. M. F., Naz, S., Khaled, A. A., Rahman, S. A., Rahman, S., & Rahman, R. M. (2015, June). Analysis and classification of respiratory health risks with respect to air pollution levels. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on* (pp. 1-6). IEEE.

Zhu, J., Hu, Y., Qin, C., & Yin, L. Z. (2012, December). Simulation analysis of air pollution dispersion based on interactive virtual geographic environment. In *Information Science and Control Engineering 2012 (ICISCE 2012), IET International Conference on* (pp. 1-4). IET.

Atanassov, E., & Ivanovska, S. (2013, May). Computation and analysis of Sobol coefficients for air pollution concentrations over the territory of Bulgaria. In *Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on* (pp. 234-237). IEEE.

Feizizadeh, B., & Blaschke, T. (2013). Examining urban heat island relations to land use and air pollution: multiple endmember spectral mixture analysis for thermal remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(3), 1749-1756.

Baojun, Y., Yanxia, L., Jun, C., & Jing, X. (2012, June). Air quality and environmental pollution analysis in Qinhuangdao city. In *World Automation Congress (WAC), 2012* (pp. 1-3). IEEE.

Song, J., Liu, X., He, H., Shou, S., & Zhang, J. (2012, June). Synoptic Analysis and Numerical Simulations of an Air Pollution Episode over Ningbo City in Winter. In *Computational Sciences and Optimization (CSO), 2012 Fifth International Joint Conference on* (pp. 892-895). IEEE.

Zhai, Q., Li, L., Yuan, Y., & Guo, Z. (2011, June). Correlation analysis of air pollution and respiratory disease in City A. In *Remote Sensing, Environment and Transportation Engineering (RSETE), 2011 International Conference on*(pp. 197-200). IEEE.

Zhang, X., Zhang, Z. F., & Zhang, R. Q. (2010, June). Analysis on Air Pollution Risk Assessment of Volatiles from Mangliaia aromatica by GC/MS. In *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on* (pp. 1-3). IEEE.

Wen-hui, Z., Hui-li, G., Wen-ji, Z., Zhu, L., & Tang, T. (2009, July). Spatial variation of inhalable particulate matter and its influence factor analysis during the regional air pollution study. In *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009* (Vol. 3, pp. III-148). IEEE.

Cheng-Jian, Z., Xian-Hua, L., Xiang-Yong, S., & Qing-Zhou, L. (2009, May). Analysis on the correlation of atmospheric path radiation and air pollution index. In *Urban Remote Sensing Event, 2009 Joint* (pp. 1-5). IEEE.

Johansson, J., Marthinsson, B. G., & Eng, S. T. (1978). Computer Automation and Error Analysis of a CO2-Laser Long-Path Absorption System for Air Pollution Monitoring. *IEEE Transactions on Instrumentation and Measurement*, *27*(4), 358-363.

*Results of the Project*

The results from my project were to predict the levels of pollution in near the near-term.

The output will be a mean value for that day. Any levels about the mean have a binary response. I would classify any value higher than the mean as unsafe and anything below as safe. As I was dissecting the data, there were strong correlation with Carbon and Sulfates on a daily level. I also had a high F-score with all the independent variables. (See png files on github). I then proceed to use Times series forecast to predict make predictions.

Six Time Series Models were used to best predict the four pollutants. The time series models used were Time Series Regression, and ARIMA Model. Taking the best model with less error rate for each pollutants.

Linear Model: Linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

ARIMA Model: ARIMA models are the most general class of models for forecasting a time series which can be made to be “stationary” by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging or deflating (if necessary). A random variable that is a time series is stationary if its statistical properties are all constant over time.  *A stationary series has no trend, its variations around its mean have a constant amplitude, and it wiggles in a consistent fashion*, i.e., its short-term random time patterns always look the same in a statistical sense.  The latter condition means that its *autocorrelations* (correlations with its own prior deviations from the mean) remain constant over time, or equivalently, that its *power spectrum* remains constant over time.  A random variable of this form can be viewed (as usual) as a combination of signal and noise, and the signal (if one is apparent) could be a pattern of fast or slow mean reversion, or sinusoidal oscillation, or rapid alternation in sign, and it could also have a seasonal component.  An ARIMA model can be viewed as a “filter” that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts.

 A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Time series *analysis* comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series *forecasting* is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time