# Assignment – 8
# Midterm Proposal

## Contribution:

Research paper titled “predict air pollution in Brasov city with regression models” is concentrated on what are the parameters that influence the results of air pollution in Brasov city. In this study they followed some methodologies to predict the pollution in urban air pollution. Also explained about how vehicle transportation is important and what are the minimum requirements used for these kind of transportations. By this how our civilization in advancing and improving technological advancements. On the other hand these advancements are also have negative impacts on the world we live in. These negative impacts will result in the environment changes, health concerned issues and our world, earth we live in will day by day will become unsuitable for humans. These are serious issues every nation and every individual should consider. Our future generation will depend on the actions we take now. So considering seriousness of the environment, this study helps in understand distribution and level of air pollution in small area compared to the world Brasov city. They tried different regression models to come up with better prediction of the air pollution. In my research and I am going to consider these methods to understand how it is distributed and predicted and I will implement those on our dataset. I also consider the parameters like correlation between area and pollution, population and pollution. Because we cannot say that these part of city have high pollution based on parameters, there could be more vehicles running in these areas at some particular time interval and which results in these kind of increase in air pollution. So population should also be a main factor in deciding and predicting the air pollution.

Another research paper titled “Modelling the impact of road traffic on air pollution in urban environment case study: A new overpass in the city of Craiova” represents the work of authors on correlation between average speed of vehicles and volume of emissions pollution evaluations. They also considered the parameters and factors like infrastructure and area of travel for better understanding. For example in highways, average speed is comparatively more than the average speed of vehicles in normal roads these factors are also important in higher understanding. They also stated that by implementing an overpass in the Craiova city, it can reduce the emissions of gases like CO, C6H6, NO2 and O3. They performed their research on Craiova city dataset. The abstract mainly started with stating population is being increased day by day and due to increase in population there is a need to increase in transportation systems and technical advancements. Transportation is the main source of air pollution now-a-days. There will be effect on humans, animals, infrastructure, landscapes and plants due to air pollution. In this paper they performed the correlation between average speed and gas emissions. As speed is also a considerable factor for gas emissions and in turn results in air pollution.

In “Traffic Air Pollution and Mortality Rate Advancement Periods” they directly calculating the relation between pollution and mortality rate. Air pollution is directly related with mortality rates. The effect of air pollution in respect to different reasons for death in a population is of general wellbeing significance and has not been introduced. In this study, increase in rate time periods is related with air pollution exposures were evaluated. They performed the tests at a centre in Hamilton, Ontario, Canada, in the vicinity of 1985 and 1999 for pulmonary testing. Cox regression model was used to model mortality from all normal causes from 1992 to 2001 in connection to various organs in body.

As discussed previous air pollution will not only make the world unsafe and unsuitable for future generations to live it also have huge effect on human health. Not only humans, even animals and plants. Due to damage caused by air pollution on human health, it reduces the mortality rate of the humans. In our research we are not going into mortality rate and how humans are being affected by the air pollution. We are going to perform some regression models to see how the data is distributed. This distribution of the data is related to dataset of Brasov city. Where there are locations and values for the different emissions of gases.

Other similar articles provided in the reference are following the similar approach and finding the predictions of mortality rates, how it is effected to one's health. These research also considered different factors varying from simple highways to as complicated as emission of gases with respect to speed. These all factors play major role in air pollution. In our research we consider the location, time and different gases as we have access to this data only and perform different regression models and test which one is best suited and gives more accurate results in predicting the pollution for next hour or next day. This paper will also provide with the areas which are highly polluted. This can be used in various field to make the world better place by implementing some measures reduce the pollution or showing the public what are the dangerous places to live or travel regarding air pollution.

## State of the Art:

Pollution is when natural resources like Air, Water, Land and other parts of environments starts to become unsuitable or unsafe to use. Air pollution is one of the most dangerous pollutions humans are facing right now. Air pollution may result in various harmful diseases and mortality rate will also fall due to this. There are many factors which results in air pollution, humans are also one of the factors for causing air pollution and important factor. Mostly air pollution is caused due to motor vehicles and many recent electronic devices. There are many ways to calculate the air pollution, sensors can be used to understand the percentage of different parameters which cause the air pollution.

Air pollution will have different parameters to consider. Air pollution can be measured using Air quality Index (AQI) metric. This is measured using nearly 449 observation points in total. If AQI increases it also increase chances of health related issues. Sensors can be used for measuring parameters. For the sake of dataset they used one sensor for each traffic sensors available in the area. These sensors will give information and values like Carbon Monoxide levels, Nitrogen dioxide level, Sulphur Dioxide level, Particulate Matter and ozone index level as specified accordingly in Air Pollution Index from wiki. Sensor measures the values by initially assigning it to a value from 25 to 100 according what it is measuring and how dense it is for example Carbon Monoxide. Next for every 5 minutes previous value will be added by a random number from 1 to 10 if its value is below 20. Same way if the value is above 210, a random number from 1 to 10 is subtracted from the previous value. Else a random value from -5 to +5 is added to last value. It is followed this way because the values would not fall in low and high suddenly and keep the values more realistic and confine them in bounds.

The dataset is from Citypulse website, where it provides with the air pollution dataset of Brasov in Romania. There are 449 different locations where data is collected. Dataset consists of 449 different csv files, each file for each location. Each csv file have parameters like location (latitude and longitude's), timestamp (it is recorded for every 5 minutes), ozone level, particulate matter level, carbon monoxide level, sulphur dioxide and nitrogen dioxide. Dataset is from the dates August 1 2014 to September 30 2014, 2 months of data which is recorded every 5 minutes. Using this dataset we can find some pattern in the air pollution and help in bringing that down. There are some ideas or

implementations which we follow through in this paper for better understanding and conclude with some insight.

From the dataset we can find the locations and their respective gas emissions. From these factors we can find some patterns like correlation between time and pollution. We can also perform time-series calculations and can predict the pollution on that particular location at this particular interval of time. From this analysis we can mark the highly polluted areas and may avoid frequent visits in these related areas. These can be calculated by the dataset we have, there are 449 locations and each location have one separate file. Each file has around 18000 records with location and emission values. So each location can be taken at a time and sum all the values. These can be categorized into different time series like calculate for monthly, daily and hourly. From this we can find a pattern and may predict for the next day or hour. These can be very helpful in understanding the pollution distribution and help people in avoiding those areas. We can use Brasov city data and use the location data and find the values of the pollution and show case the highly polluted areas in the city by heatmap. Using these heatmap we can find the highly polluted areas so that public can avoid those in case of travelling. This can also help in taking measures for reducing the pollution in those areas, may be by planting more trees or making vehicles to take alternate route. Some ideas of show casing the polluted areas are by making specific areas green where it is not polluted and red where it is highly polluted.

Using different regression models and predicting the next day's or next hour's pollution in that particular area. We can implement all these with different tools available like tableau, R, SQL. As this is not a huge data we can simple load it into R and run regression models. If it is huge data set and even SQL cannot handle the performance, we can go for big data technologies like Hadoop, MapReduce and Spark. If size of dataset is reasonable enough to run on normal databases, it is better to use in it as it is faster and economical for the performance.

Some areas are hugely populated, in these areas pollution is high automatically. When there is huge population, they need more transportation systems. Transportation systems may include different modes of transportations. Even emissions from flights are causing damage to ozone layer. This in turn will reflect on living organisms on earth like cancer. Pollution will have huge impact on health of living organisms. Humans are the main factors for these air, water and other pollutions. Air pollution in mainly caused due to emissions from the motor vehicles or any related equipment. Due to this emission of poisonous gases into the air, they undergo some chemical reactions with the atmosphere and result in bad environment. This environmental change is in turn reflect on human health and effect badly on future generations.

Pollution is relatively proportional to the population of the area. For example less densely populated areas have less transportation so this result in less gas emissions and which results in less pollution. So correlation between pollution and population is also considered in study. This makes huge difference in understanding and finding the patterns. If we did not consider the population and calculated the pollution, we cannot rely on that results and it is not a good model to consider for further calculations or predicting the pollution. One limitation for this approach is that we need to have access to the population dataset and we need to have location data too for that population. We have the pollution dataset and locations based on sensors. But we do not have sensors for calculating the population. Even if we had, it would not be same locations as the sensor data. The population and pollution locations will operate on different areas. So this approach cannot be accurate enough, though we can find the city population and can calculate. But this will be for whole and cannot be relied on it.

# Data:

Data we use for this Capstone project is pollution data from the link http://iot.ee.surrey.ac.uk:8080/datasets.html#pollution. The time frame of the data of 2 years which is from June 2012 to June 2014. This data is survey data of UK, which have RSS feeds from municipality department. Data is downloaded from the citypulse website. This website provides with various semantical datasets which can be used for various purpose. This data is collected from various partners of CityPulse EU and this data is relevant resource for smart city data. There are so many papers which of them cited this website.

This website basically collects the data from the surrey county council webcasts at http://surreycc.public-i.tv/core/portal/home, this have the webcast data of previous and upcoming events. We can access recent ones from the link http://surreycc.public-i.tv/core/portal/home#webcast and previous ones can be accessed from the link http://connect.surreycc.public-i.tv/site/webcasts.php. The timeline from the data which usually provided by the surrey website shows the agenda of the live webcast and recent ones. In the same way it may also refer to index points from the previous ones. There are some additional tabs too which used to access the additional features like slides resources, speaker profile and chats. There is also sharing and embedded buttons which can be used to integrate in the personal or our own websites, blogs or any forms. When we open the link and look for library listings we find all the events going on at present, resent and previous ones. Which also includes with time and date of events. It also includes with small description about this event and most importantly venue of the event where it is going to happen. This helps in lot of people to find their interesting events and follow them.

## Data Load:

We are using Zeppelin to build our project and run the test on the dataset. Zeppelin is notebook which is web based. This is used and will provide interactive data analytic functions which can be used by data analysts. We can use SQL, Scala, R, Python and many number of languages in with this. This runs on top of spark. This zeppelin is provided with data integration, data discovery, and data analytics and data visualization. Zeppelin provides in-built spark integration which help is fast access and fast processing of the data and perform analytics without any difficulties and faster than any other. Even data visualization is done with ease basic chart are provided with data retrieval so it is easy to check. Due to spark the performance is very high.

Zeppelin installation consist of steps like download the binary files from the link https://zeppelin.apache.org/download.html. The current latest version is 0.6.2. After downloading, extract the zip file and save it in a folder. We can also build this with source. To install zeppelin, need to run this command from terminal /bin/install-interpreter.sh –all, which installs all the interpreters available. After installation, need to go to http://localhost:8080 where zeppelin web interface is provide and we can perform the actions. We can learn using the tutorials provided in the web application and seek help for any assistance.

There are many interpreters available to work on like Python, Spark, Scala, R, Sql and so on. These list of interpreters are listed in conf/interpreters-list document.

# Method:

We used Apache Zeppelin to run different methods in our project. Apache Zeppelin enables interactive data analytics on a web based notebook. Apache Zeppelin supports different languages like Python, Scala, R, SQL and more. The main advantage of using Apache Zeppelin is, it has built in Apache Spark integration. In many other notebook or tools we need to build separate module or import the libraries

required. Some of the features Apache Zeppelin provides with spark integration are Automatic injection of data using SparkContext and SQLContext, load libraries or jar files from local file system or maven repositories, display the job progress and job cancelation. Normally Apache Zeppelin comes with whole package required for data analysis. For example, matplotlib library is mostly used for the data visualization and data analysis in python, so latest version of Apache Zeppelin comes with integrated Matplotlib library, we do not have to import it separately and also leverage the Spark performance speed by pyspark Interpreter which fast and reliable.

In this project we are going to use python in Zeppelin. In Apache Zeppelin we can write python scripts by using %python tag before writing python scripts. But we need to use spark integrated python too as for now we have only limited dataset and in future if we are going to build large scale project based on this, we might need to leverage spark parallel processing and high volume processing. For using python in spark we need to use **%pyspark** and this will change the interpreter into python on spark. Apache Zeppelin provides all these inbuilt so we do not have to reconfigure any setting to get this functionalities. More over Apache Zeppelin is open source which gives huge advantage to programmers and application builders and contribute in the development of tools which are helpful

First we need to import the data into the spark data file system. For this we can use the SQLContext function where we need to specify the directory of all the files. This will consider all the files in the directory and read them. Now that data is in the sqlcontext object we can use sql commands too to see the output and work on functions. This is very good approach because we need to configure so many things, change settings, paths, connections before so that we can import, validate, make changes and export where as in Apache Zeppelin we have all the functionalities, dataset and we can use which comfortable programming language we are comfortable with to build the project and run successfully.

## Operations:

There are four main operations considered in this project and one future implementation or operation for better understanding of the situation. The four main operations are show case location of sensors which collect the data of the pollution and records it. The different factors for the air pollution these sensors collect are ozone, particulate matter, carbon monoxide, sulphur dioxide, nitrogen dioxide, additional to these it also records latitudes, longitudes and time stamp. With this readings, these following operations are implemented:

**Show locations of sensors:**

In the data files we have, there are latitudes and longitudes of reading that took place. Each file is from one different location, so each file have data from one sensor. So using this locations we can show where the sensors are and how polluted those areas are. There are 449 locations in the directory so Brasov city has records of 449 pollution records.

**Heat map showcasing density of pollution:**

Now that we have the data of different locations of pollution records so we can see how the pollution is spread in the city. We have the locations of the sensors in our data, using those latitudes and longitudes and using maps library in python we can showcase the Brasov city. Those coordinate we locate them on the map, if it is highly dense colour it represents the highly polluted area.

**Time series for those locations or busy location**:

We can also plot the time series of a particular location throughout the day, how different factors of pollution distributed along the time. We can build these plots using matplotlib library. We can also

compare different factors how they are related to each other and we can find correlation between each of them. The data we have is the records for every 5 minutes update, so we can have accurate time series plot. By observing the time series and density of pollution we can find the busy routes in the city. This could help in finding the highly polluted areas and people can avoid these routes as it will be harmful to health

**Regression models and predict:**

Predicting can be made using different regression models. Some of the regression models we can use are logistic regression model, multiple regression, box-cox, principle component regression and Ridge regression. Once we run all the models compare them with each other and observing which one is the best model and decide which one to use for further implementations.

**Cross validate Pollution rates and population:**

If we consider only the pollution rates of areas, it is not sufficient because highly populated area will obviously have high transportation rates. With increase in transportation there is high chances of air pollution around this area. We need to find the area wise population of Brasov city and as we already have the pollution rates of those areas, we can cross validate both of them and examine if there is any relation between pollution and population. Theoretically there should be relation. Vehicle usage is directly proportional to the air pollution. Not only vehicles emit harmful gases into the air, even some electrical tools which we use on daily bases will also emit harmful gases like Air Conditioners, refrigerator will also emit some pollution.

## Functions:

Functions are basically used to eliminate the reuse of piece of code. In technical point of view functions are set of code which perform certain operations on the values provided from the parameters passed. This functions can be called as many times as required and perform operations as many times as required without writing the same code repeatedly over and over again. In simple words functions are explained with an example like, sum of two numbers calculation is repeating over and over again in code, this adding two numbers is written into one function. So whenever there is requirement of sum of two numbers, we can directly call this function by passing the two numbers which are needed to add. This will return the output, sum of those two numbers.

Functions are divided into two categories, user defined functions and inbuilt functions. Inbuilt functions are the functions which are written and available with the programming language, this does not need to write separately. For example “print” function in python. Print function is already written in programming language itself, so whenever we call “print” function with any value passing into it, it will print it into the console without any additional coding. Whereas user defined functions are written by programmers for their convenience and they define the rules of that function. For example sum of two numbers function. This function is written manually by asking the user to enter two numbers which are to be added. Inside the function these two numbers which are passed through function are added together by the addition operation which is mentioned inside the function. Then there is a return function which returns the final value after all the operations executed in that function.

For this project we are going to use three main user defined functions readFiles, locationValues, timeValues. ReadFiles function is built mainly to read csv file in the directory and return the list of all the data from csv files. This function takes the directory as input string, this input directory is accessed by the python and gets all the files from that directory. Now that we have all the files from directory (we have 449 csv files in this directory). Each file represents the location of sensor across the city and each file will have five air pollution factors, latitude and longitudes and timestamp for the readings.

So in for loop we will be taking each file at a time and read it line by line. Each line represents one record in csv. So every line consists of 5 values of pollution factors, 2 location values and 1 timestamp. So each line is read and splits using delimiter “,”, now these 8 values are stored in the list. As we are in the loop of all the files, these lists generated by each file is added to another list and this big list consists of all the 449 files. Using all these lists, looping them we can work on our operations using pyspark.