**Assignment 15**

Final Teleprompter Script

DS-670

Neha Kumari

**Contribution of Competitor Article**

The competitor article I used had worked on the same dataset that is, the Smart City, CityPulse on air pollution. The paper described the K-means clustering to group the data and identify the healthiest location in the city of Aarhus in Denmark. The primary pollutant used in the study was ozone.

The ultimate goal will be to build an Air Polluting monitoring System in smart cities that can provide real time analytics. These contributions will help in building a smart city where we can in know the air pollution levels in advance. This will help in taking preventive measures as well as in future urban development planning. With Air monitoring systems the general public can estimate the exposure to the pollution they will have on a day, spreading awareness about air pollution and its harmful effects on human health. As we know Smart city is all about Internet of things, contributions would make the city much more planned and connected. The results from one contribution can be used for other smart city project like Road traffic congestion. There are many research projects going on in this direction. One is in China, Beijing where the pollution levels have reached at alarming levels. IBM, China has started a project to study the effect of pollutants under different conditions.

**Description of your contribution**

Like the competitor article, I also started with k- means clustering and grouped the data to find healthy and unhealthy areas. I extended the research with time series forecasting also using ARIMA model so that we could estimate the pollution levels for the next day. ARIMA models are considered to give good results based on the parameters we choose that is the AR order (p), differencing order (d) and the moving average (q). I used various combinations of these parameters to get the best working model. This would give an estimate of the pollution exposure to the people living and commuting in that place. The accuracy of the forecasting largely depends on the accuracy of the data. We worked on the data generated by the sensors, the accuracy of the data depends how accurately the sensors captured the data and the location of the sensors placed. The sensor technology is comparatively new and not very developed.

The pollutant I used for this study was ozone, as it is considered the most important air pollutant in air pollution levels.

**Data Source and content**

The data set was from the smart city pulse project <http://iot.ee.surrey.ac.uk:8080/datasets.html#pollution>.

The data had total 449 files. Each file is the collection of the air pollutant levels of a particular location at 5 minutes interval. This data is collected from Aarhus city in Denmark.

The data had the following fields- ozone, particulate matter, carbon monoxide, sulfur dioxide, nitrogen dioxide, latitude, longitude and timestamp. Each file has around 17k records and in total the data has around 7500000 records. The duration for which the data is collected is from August 1st 2014 to October 1st 2014. The data is collected and generated by sensors. We are assuming that the data shows the correct levels of the pollutants. The pollution levels are measured in AQI (Air Quality Index).

Some of the harmful effects of these pollutants on human health are discussed below:

* Ozone: There are two types of ozone – healthy and unhealthy. Ozone occurring naturally in the atmosphere is called the good ozone, whereas ozone at the ground level (at troposphere) is considered as bad ozone. Ozone is created due to chemical reactions of the gases emitted by vehicles. Also under sunshine and without sunshine the reactions differ. It causes health issues among adults and children like respiratory diseases, congestion, and chest pain throat irritation.
* Particulate matter: Particulate pollution can cause heart diseases, ling cancer and asthma attacks. PM usually comes in two size ranges – PM10 and PM 2.5. Diesel is the main source of PM in vehicle emissions; wildfires smoke is another cause of increased PM levels in the environment.
* Carbon monoxide: The burning of vehicular fuels produces it. Breathing carbon monoxide can cause headaches, nausea, dizziness, vomiting. It is said that high CO levels causes slow poisoning in the humans when exposed for a long time because it reduces the amount of oxygen reaches to human tissues.
* Sulfur Dioxide: The main source of Sulfur dioxide in the atmosphere is fossil fuel burning and industrial pollution. It can combine with other particles in the environment and can cause haze. Haze reduces the visibility. It also is harmful for old monuments and stones, as it stains and damages them. On human health it can cause wheeze, cough, chest-tightness. It can aggravate the respiratory problems.
* Nitrogen dioxide: It is also responsible for lung diseases in humans particularly children and older adults. It is highly reactive gas. When combined with water, oxygen and other chemicals it causes acid rain.

**Your Method**

K-means clustering:

* We used K-means clustering which is an unsupervised learning algorithm.
* We started with an initial cluster size of 4 and continued till 10.
* We used ozone levels as the deciding factor in classifying the areas.
* At first we took the data with all five pollutants and applied k-means, where k changed from 4 through 10.
* At k=10 we got best results where we found the minimum levels of ozone at some latitude and longitude
* The other approach we used was using only ozone in the data set
* We applied k-means again from k=4 through 10, to get the results.
* This time our results improved as we got lower levels of ozone in our cluster, which we classified as the healthiest area in the city.

Forecasting using ARIMA time series model:

* Before applying ARIMA model, we first separated the timestamp field in the data set into two different columns namely – date and time.
* The time series model rely heavily on the past values to forecast the future values.
* Then we simply plotted the data to understand our data well. The data did not show any patterns like trend, seasonality or cycles.
* Then we converted our data to time series object and plotted the ACF and PACF graphs
* The ACF graph was decaying slowly indicating that the series was non-stationary. ARIMA models work for stationary series only, so we converted our non-stationary series we included the differencing factor of 1.
* We used auto arima function to give us the best model
* The model we got was ARIMA(1,1,0).
* We then plotted the residuals plot of the model for ACF and PACF graphs. This time the graph indicated stationary series.
* To see if the model was a good fit or not we use the Akaike information criteria (AIC) and Baysian information criteria (BIC). The lower these values are the better the model is. The values were low for this model. These values indicates how much information is lost, if this model is chosen.
* Seasonality is either annually or quarterly and is detected for data that gets affected by the seasons, like rain, stock prices or number of tourists in some place. Some data shows a trend of overall increase or decrease. For example sales of houses over the years would show an overall increase. Some data has cycles, like increase and then decrease.
* While the ACF and PACF graphs visually determine if the series is stationary or not. It also helps in choosing the right order for ARIMA. There is ADF test (augmented Dickey-Fuller), which can also test statistically if the series is stationary, or not. The null hypothesis is that the series is not stationary while the alternate hypothesis is that the series is stationary
* The AR model also called the auto regressive model means using the past values in the regression. If we use only one value it means the AR order is 1,AR (1) or ARIMA (1,0,0) if two pasts values are used then it is an AR(2) model or equivalently ARIMA(2,0,0). A moving average MA(q), combines the past errors, q determines the number of error terms of the past models to include in the current model. The d value represents the degree of differencing (I (d)). It simply means the number of subtracting the past values from current. Usually the value of d doesn’t go beyond 2.
* We compared the AIC and BIC values with other ARIMA models with different combinations of parameters p, d and q. We found the minimum AIC and BIC values at order (1,1,0).
* After fitting the model, we also plotted the residuals and it did look like white noise as needed.
* At last we used this model for forecasting the ozone levels for the next day.
* Main packages used were: ggmap,maps,ggplot2 mainly for visualizing the cluster data on maps. Sqldf package allowes us to write sql like queries in R. We used it to group our data by longitude and latitude.

Readr and dplyr are data manipulating packages and helped us to combine our 449 files into a single data frame. Forecast was used for forecasting time series., Auto.Arima function and arima function.

**Quantitative Results 1**

Results for k means clustering:

K=10

longitude latitude avg\_ozone particulate

1 10.17052 56.17031 100.98485 123.39569

2 10.18366 56.16245 115.32841 124.68005

3 10.17410 56.15099 114.33345 89.32272

4 10.18570 56.16588 111.94914 102.53156

5 10.18032 56.17061 81.92253 121.50843

6 10.18209 56.17875 137.23991 119.00751

7 10.15168 56.16368 94.99104 101.03512

8 10.17870 56.16995 127.04528 103.73866

9 10.18328 56.17163 118.22879 112.93678

10 10.17883 56.15726 106.61785 112.47661

With K = 10, for cluster 5, the ozone concentration is minimum at longitude 10.18032 and latitude 56.17061.This approach gave us minimum ozone levels as 81.92253 and maximum ozone levels: as 136.89854

**Quantitative Results 2**

We used different parameters for ARIMA using arima function but finally we

Used, ARIMA (1,1,0) model for forecasting ozone level for next day. This model gave us the minimum AIC and BIC values. We did not include moving average in our model, as there was no trend in our model, so no smoothing was needed. Auto.arima function is used to get best ARIMA model confirmed by ACF, PACF plots.

The values we obtained were: standard error 0.0075, AIC 90854.27 and log likelihood value -45425.13

**Discussion: Comparison with your Competitor**

The Results we got of k means were different from the competitor articles. We both used 10 clusters. The competitor got the minimum level of ozone concentration as 89.66607 AQI at longitude = 10.18129 and latitude 56.16788. The maximum concentration levels it got was 129.12724.

We also used 10 clusters and got the K = 10 and at longitude 10.18032, latitude 56.17061 got the minimum levels 81.92253

We got the Ozone maximum levels 137.23991 at longitude 10.18209 and latitude 56.17875. The difference in the results might be due to the fact that k-means initialization varies on every run and it only calculates the local minima which depends on it’s initialization.

For time series ARIMA(1,1,0) gave us the best results, this was confirmed by the diagnostic plot ACF and PACF. The AIC was minimum for this model, which was 90854.27. We used this model to predict the ozone levels for next 24 hours.

**Performance on Big data**

We used zeppelin for this implementation. Zeppelin provides an interactive notebook that supports multiple programming languages. We used Pyspark and R for this project. One of the greatest advantages of zeppelin we found was it is fast. Loading the data took less than 30 seconds even though the number of records was in lakhs. Also zeppelin can be deployed directly to clusters without re-writing the code again. Most data operations took less than 30 seconds, these operations included data loading, data visualization graphs, applying k-means algorithms and time series forecasting using ARIMA model.

**Conclusion**

The takeaways from this research project are:

* K-means gave us the most environmentally clean area are at longitude 10.18032 and latitude 56.17061 where ozone level was, 81.92253 Air Quality Index whereas the unhealthiest are was found at longitude 10.18209 and latitude 56.17875 with ozone levels 137.23991.
* We used time series ARIMA model to predict the next day ozone levels with a confidence of 95%.
* According to the table in the presentation that provides the AQI levels and it’s meaning, we conclude that the healthy area we found had levels around 81, which is moderate on AQI. The unhealthiest are we found had the levels 136.89854, which may affect people in the sensitive group and not the general public.
* The overall air quality index of the Aarhus city is good and healthy for the general public living.