**Air Quality Index Prediction**

**-Annamariya Tharayil**

Table of Contents

1. Introduction 3

2. Objective 4

3. Workflow 4

a) Data Gathering 4

b) Data Cleansing 4

c) Data Exploration 5

d) Model Building 5

e) Data Visualization 6

4. Analysis 10

5. Conclusion 12

6. Future Works 12

7. References 12

Table of Figures

Figure 1: Air Quality Index Health Chart 3

Figure 2: AQI for 2016 6

Figure 3 AQI for 2017 6

Figure 4 AQI for 2018(until 30 November) 7

Figure 5 AQI distribution chart for 2016 7

Figure 6 AQI for 2017 8

Figure 7 AQI for 2018 8

Figure 8 Project Workflow 9

Figure 9 Prediction output 10

Figure 10 Categorization of predicted values 11

# Introduction

The Air Quality Index (AQI) is a simple, color-coded, unit-less index that is an effective way to communicate air pollution concentrations to the general public.

The increase in AQI is an indication of higher levels of pollution and could lead to adverse health effects. (San Salvador 2012). An air quality index ranges from 0 to 500, with 500 being the worst AQI value.

Below table shows the degree of health concerns as the AQI increase. (Contributors n.d.)

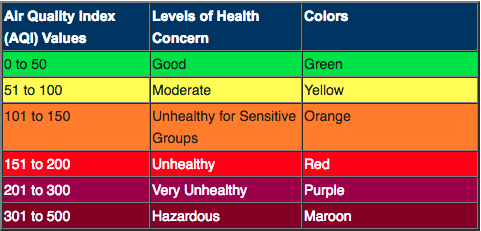


Figure 1: Air Quality Index Health Chart

Changing AQI can be due to change in temperature, climate, increase in emission of air pollutants by industries and increase in the green house gases. The quality of air deteriorates due to presence of air pollutants which includes Ozone, Particulate Matter i.e. PM2.5 and PM10, Carbon Monoxide, Lead, Sulphur dioxide and Nitrogen dioxide. Long-term implications of breathing polluted air can be development of diseases such as asthma, bronchitis, some form of cancer, shorten life span accelerated aging of lungs etc.. People who fall in sensitive group experience higher impacts at lower air pollution and have to be careful as the AQI values increases from 100. (Sacramento Air Quality Management n.d.)

United States Environmental Protection Agency is responsible for monitoring the levels of air pollutants for each hour.

# Objective

The objective of the project is to analysis the AQI data collected from January 2016 –November 2018 for Chicago area and forecast the AQI index for next 10 days based on the historic data collected.

# Workflow

The entire project is divided into 5 phases i.e. Data Gathering, Data Cleansing, Data Exploration, Model Building and Visualization.

## Data Gathering

For this project, historic data from January 2016 – November 2018 was collected and stored in to Azure Data Lake. To collect data, multiple API calls were to be made to airnowapi.org. The data retrieved after the making the API calls is a comma separated file which is stored in “/Test/AQIDATA” folder. In data gathering phase, data is partitioned based on the date on which data was retrieved. The folder structure would look like “/Test/AQIDATA/aqidate=20160101” while the file name will be “000000\_0”. After November 1, 2018 the data was gathered by scheduling the script using Logic App. The Logic App runs a recurrent job which is triggered at 12:00 A.M UTC. The script in the App fetches hourly concentration level of all the pollutants for previous day for fixed location. For this project Chicago area was only considered.

## Data Cleansing

Data is present in Azure Data Lake for use. In azure Data lake information about Latitude, Longitude, Parameter Name, Parameter value, Date Observed, Category number is included. A hive table called “calaqi” is created on the top of the data present in Azure Data Lake. While creating the table all the columns were considered to be string, this ensures that none of the data is lost. Data in Azure Data Lake is partitioned based on the date on which the logic App was run, the same partitions would be followed in the Hive table too. Data in the data lake contains junk characters; in order to clean the junk characters a new hive table was created called calaqi\_values with proper datatypes. Data is inserted from calaqi table to calaqi\_values through required type casting. In the calaqi\_values table apart from the above-mentioned columns Category name is also added. The value for the column is calculated using the category value present in the calaqi table.

## Data Exploration

For data exploration Hive queries were used. On analysis it was found that the data for Ozone and PM2.5 was available in the table. It is assumed that the selected areas did not have other pollutant concentrate in the air. The data extracted fell in one among the four categories i.e. Good, Moderate, Unhealthy for sensitive group and Unhealthy. Categories were calculated based on the AQI value of the pollutants. The AQI value is taken on the hourly basis; hence to find the AQI value for the entire day, highest level of concentrate value for the day is considered and stored in HDFS. Now that we have concentration level for each parameter highest among those parameters is taken into considered as AQI value for the day. Creating an analysis chart in Zeppelin gives an insight that Air Quality index is higher in months between April to August.

## Model Building

A time series is a sequence of data points ordered by time. Time series analysis is a methodology for extracting useful and meaningful information from these data points (Saad 2017). Spark includes spark-ts package, which includes various models that helps in Time series analysis. Time series models such as ARIMA, EWMA, GARCH and HoltWinters are included in the package. For the purpose of this project ARIMA model will be considered. The jar file “sparkts-0.4.0-jar-with-dependencies.jar” needs to be [downloaded](https://repo1.maven.org/maven2/com/cloudera/sparkts/sparkts/0.4.0/) from Maven repository. The git page of [Sandy Ryza](https://github.com/sryza) gives an example of using ARIMA model for single series prediction for stocks (Ryza 2016). The same project was considered as a baseline to create AQI predictor. The calculated AQI is fetched using spark as Resilient Distributed Dataset (RDD). The RDD is divided as training set and testing set. Training set consists of 85% of the entire RDD. The training set is then converted to dense vector so that we can use the data to be fit in the ARIMA model. To fit the data into ARIMA model we require 4 parameters i.e. p, d, q, vector. For this project, p value was considered as 100, while d and q value was considered to be 0. The model gives AIC value as AIC=7256.779310803692 for the predicted values. The root mean square Error for the predicted value and actual value is 14.488397790679226.

## Data Visualization

For Data Visualization, Apache Zeppelin was made to use. A Zeppelin notebook allows us to create interactive dashboard using the data present in HDFS. For the purpose of this project jdbc interpreter was made to use to create dashboards. The AQI index for all the days from 2016 to 2018 is as shown below.

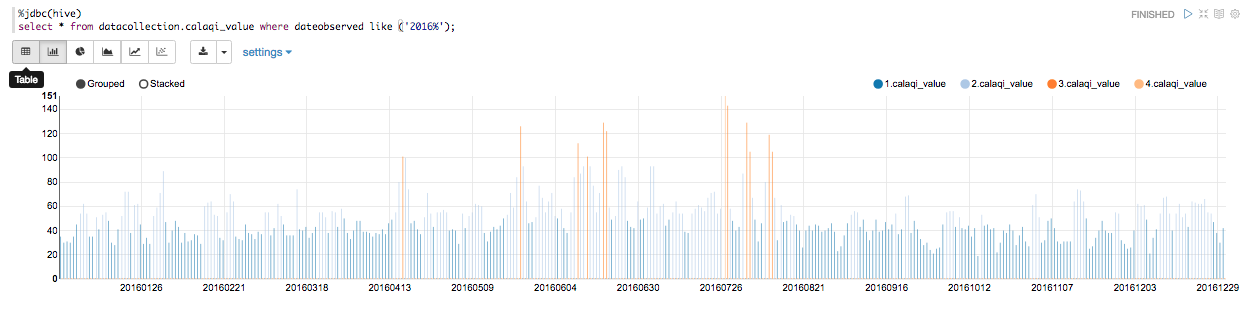


Figure : AQI for 2016

July 27, 2016 had the highest AQI value which was 151, followed by July 28, 2016 with 148 as the AQI.

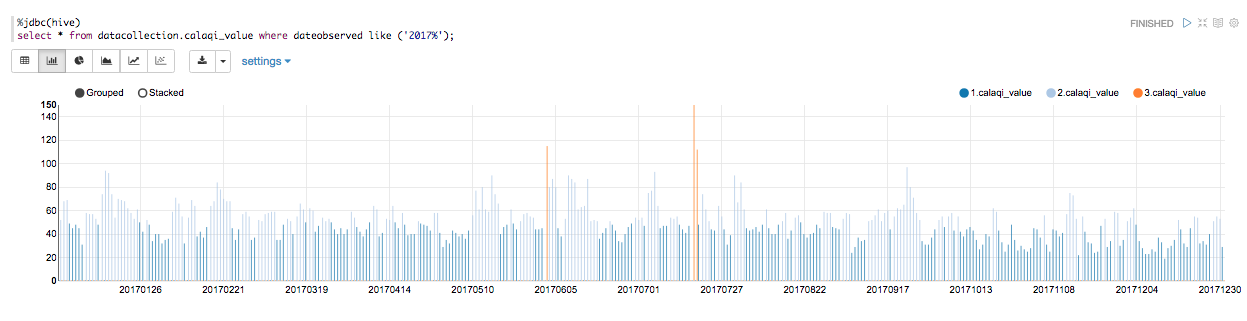


Figure AQI for 2017

July 18, 2017 had the highest AQI value which was 150, followed by June 06, 2017 with 115 as the AQI.

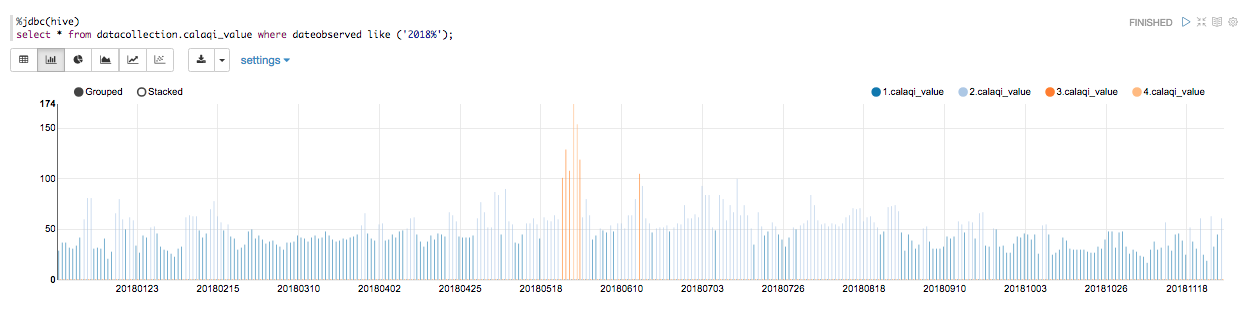


Figure AQI for 2018(until 30 November)

Data for 2018 suggests, we have bunch of days in May having AQI greater than 2.

The below pie-chart gives us an idea of distribution of categories AQI

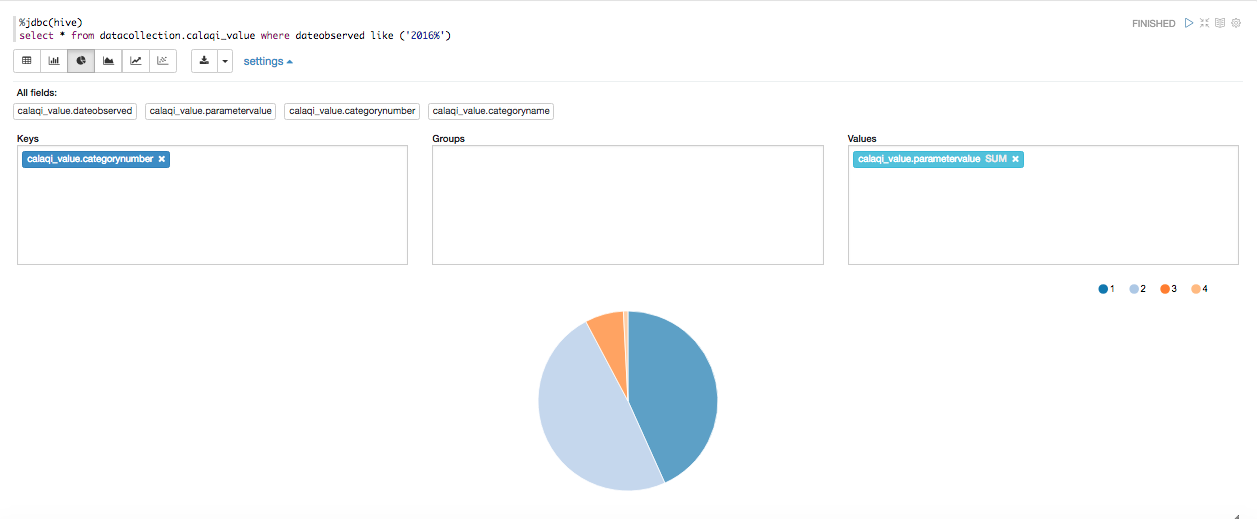


Figure AQI distribution chart for 2016

The graph gives out following information

43% of the year had AQI as category 1

49% of the year had AQI as category 2

7% of the year had AQI as category 3

1% of the year had AQI as category 4

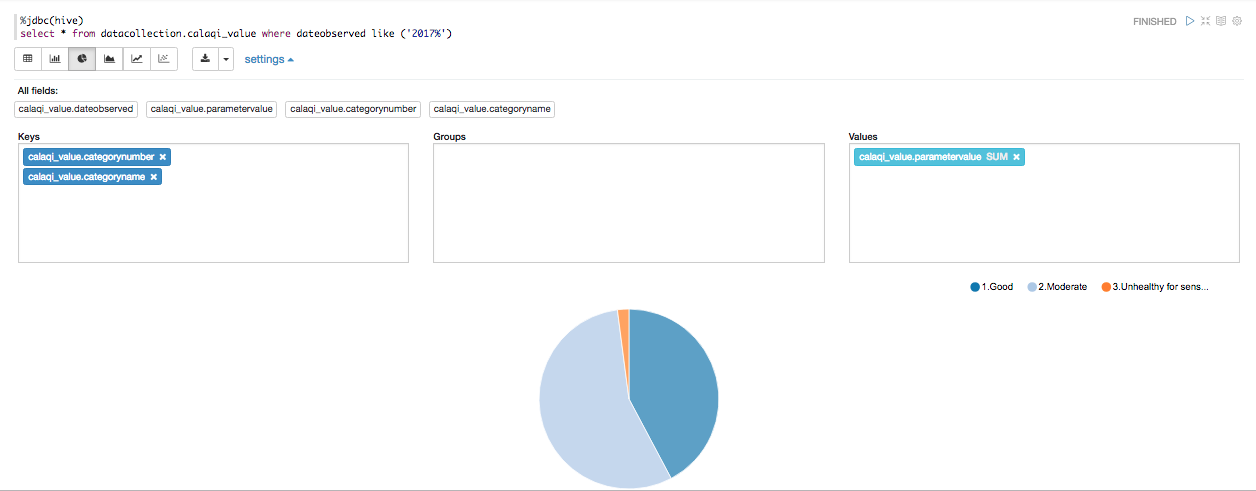


Figure AQI for 2017

The graph gives out following information

42% of the year had AQI as category 1

56% of the year had AQI as category 2

2% of the year had AQI as category 3

0% of the year had AQI as category 4

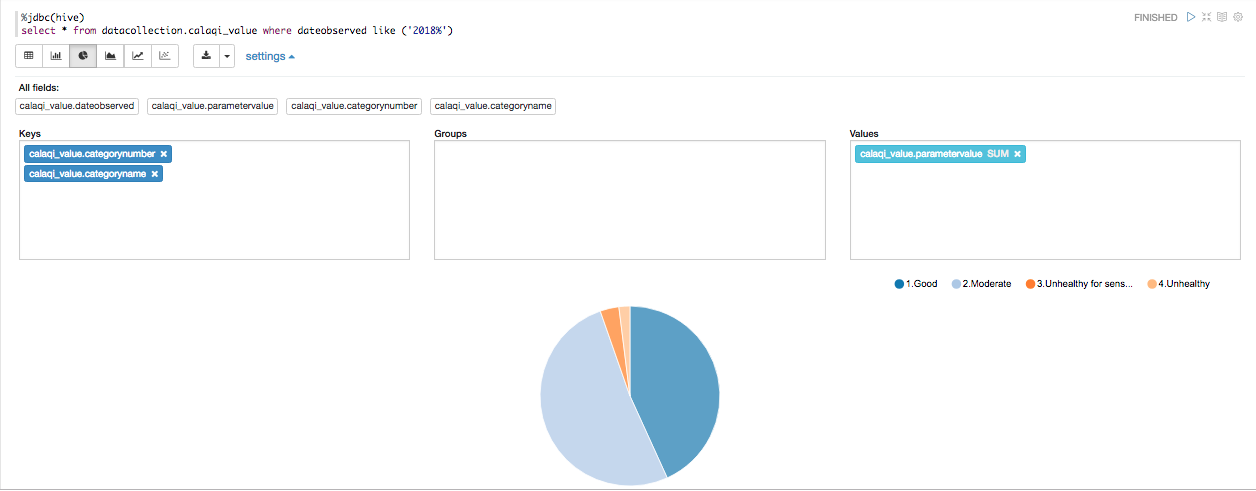


Figure AQI for 2018

The graph gives out following information

43% of the year had AQI as category 1

51% of the year had AQI as category 2

3% of the year had AQI as category 3

2% of the year had AQI as category 4

The following image depicts the overall flow of the Project

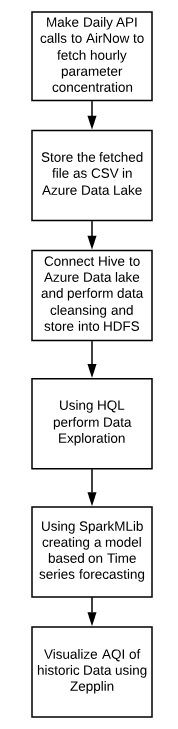


Figure Project Workflow

# Analysis

To run the ARIMA model, the data was converted to a Dense Vector. The predicted values and the expected values are as follows. The error square is also calculated which is as shown below.

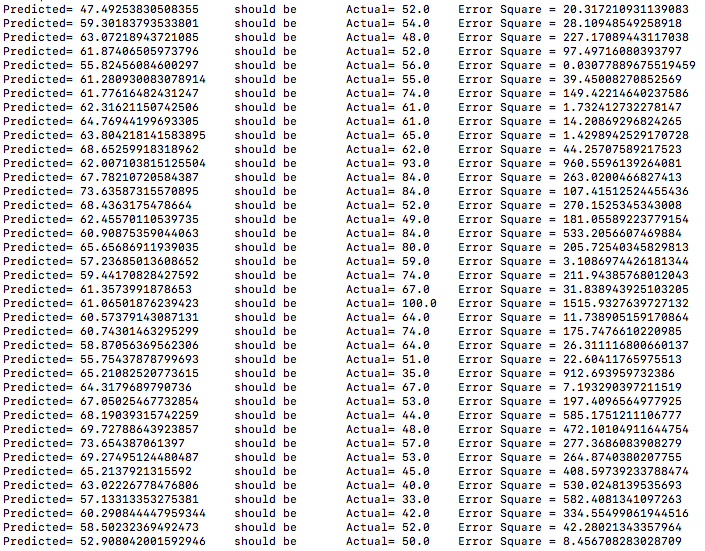


Figure Prediction output

The predicted values were also categorized according to the AQI value which as shown below

First column is the date for which prediction is made.

Second column is the predicted value.

The last column is the categorization of the predicted value.

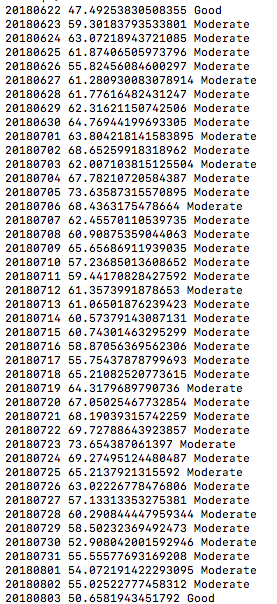


Figure Categorization of predicted values

The total misclassification percent is 4.

Since the predicted values fall in either good or moderate category, accuracy of only good and moderate was considered.

The Accuracy rate for category Good is 72%

The Accuracy rate for category Moderate is 73%.

# Conclusion

In this project time series analysis and forecasting is used to analyze and forecast the AQI value in Chicago. Time is considered as the only parameter to build the ARIMA model, which is a popular method in forecasting. The total error rate for misclassification is found to be 4%, which is considered to be negligible. The Accuracy rate for both the categories lies between 70-75%. Through this project people can be aware of the future AQI for the area that they stay and take necessary precautions if the AQI category turns out to be Unhealthy.

# Future Works

In the current project time was the only parameter to calculate the air quality. Air quality is affected by other factors such as Wind speed, traffic, temperature etc. Model could be built by taking into consideration the above-mentioned parameters.

# References

**Contributors, Wikipedia. *Air Quality Index****.* https://en.wikipedia.org/wiki/Air\_quality\_index (accessed November 8, 2018).

**Saad, Mohamed. *Badrit.* June 29, 2017.** https://badrit.com/blog/2017/5/29/time-series-analysis-using-spark#.W\_tkAxNKi1s (accessed November 18, 2018).

**Sacramento Air Quality Management. *Health Effects****.* http://www.sparetheair.com/health.cfm (accessed November 2018, 27).

**San Salvador, El Salvador. *United States Environmental Protection Agency.*** April 16, 2012. https://www.epa.gov/sites/production/files/2014-05/documents/zell-aqi.pdf (accessed Nov 27, 2018).

**Ryza, Sandy. *Git Hub.* February 27, 2016.** https://github.com/sryza/spark-ts-examples/tree/master/jvm/src/main/scala/com/cloudera/tsexamples (accessed November 20, 2018).