A Data Analysis of Covid19 Air Pollution Implications Through HDFS and Spark

Project Design Document

Team Members

Michael Anthony | Sydney Bookstaver | Teddy Li | Alyson Chen | Vahe Tascian

Project Start Date

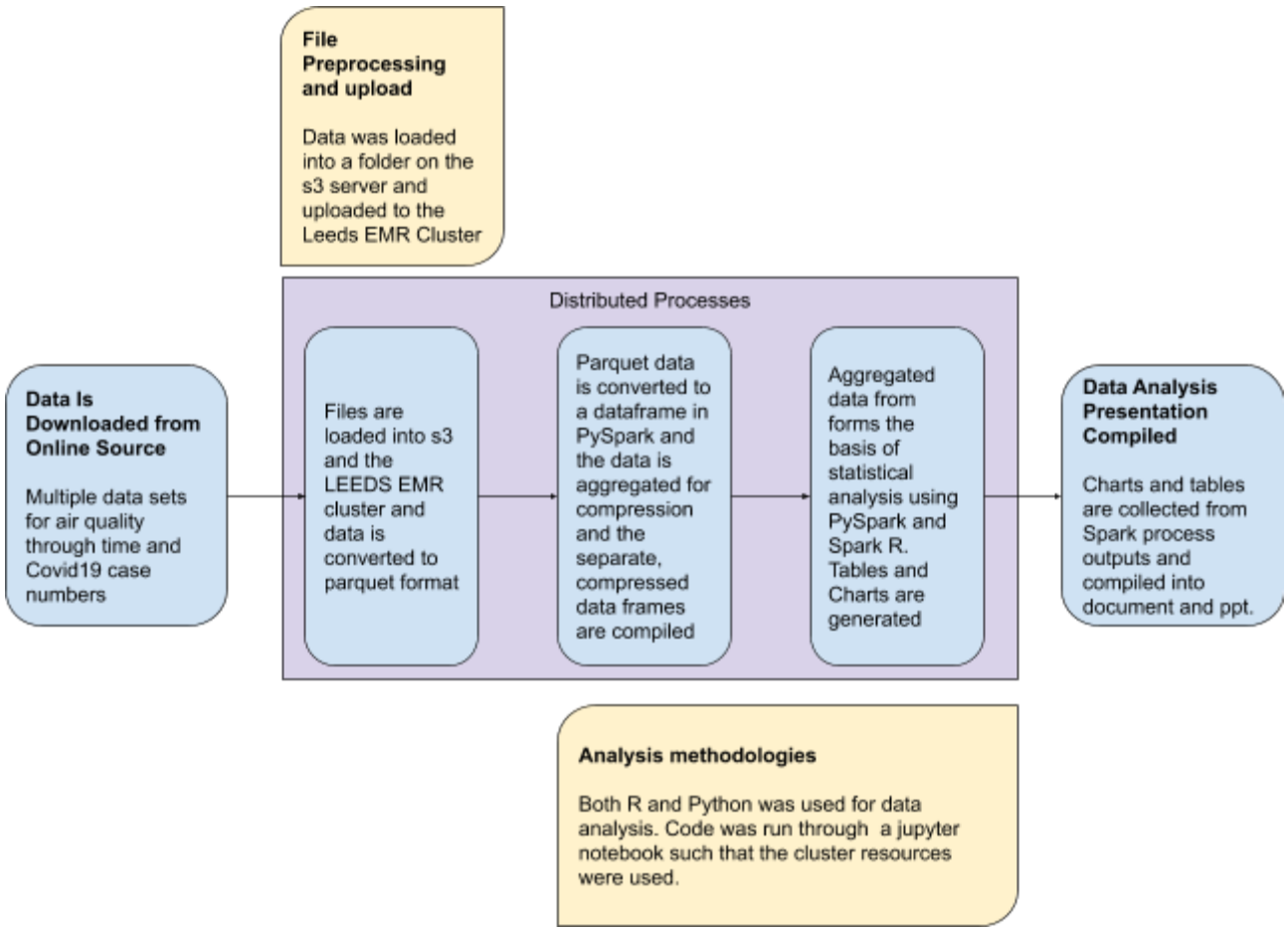
April 12th, 2020

Submission Details

Professor Peigang Zhang
MSBX-5420 | Unstructured and Distributed Data Modeling

Project Process Diagram

A visual representation of the data manipulation process chronicling the transition from data gathering through to distributed storage, distributed computation, and results.



Project Tools Summary

Tool	Project Use Case
S3	The initial data upload location - s3 was used as an intermediate storage location before data was pushed to the EMR cluster.
GitHub	Used to collaborate, organize and share code and project documents.
PySpark	Installed and utilized pySpark to upload data and create aggregations with SQLcontext to run several select statements. Used pySpark tools to create a new data frame with only relevant columns.
Spark R	Installed sparklyr to connect to spark locally. Attempted to connect with spark locally and through the cluster; however, we ran into some problems with java installations and plan to revisit this in order to implement spark with R code.
AWS	We plan to use aws to store data and obtain enough computing power to run a time series model on the entire dataset.

Data dictionary

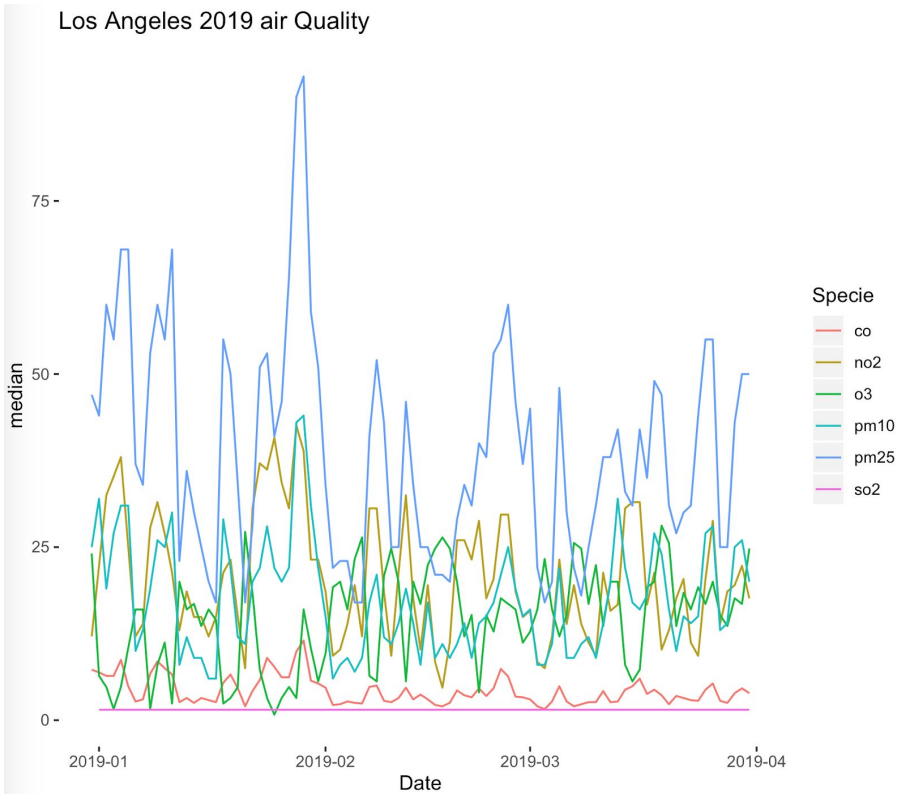
Date: Date the data measurement was collected
Country: Country location of air quality measurement
City: City of the air quality measurement
Specie: type of measurement collected (dew, so2, humidity, wind speed, pressure, wind-quest, o3, temperature, wind gust, co, pm10, pm25, no2, precipitation, wd, aqi, pol, uvi, mepaqi, pm1, neph)
Count
Min
Max
Median
Variance

Data Cleaning Operations

- Removed beginning columns of each air quality dataset
- Removed unwanted rows where species was related to weather conditions such as wind speed, temperature, pressure ect.
- Formated date column to be recognized as a date and ordered in chronological order

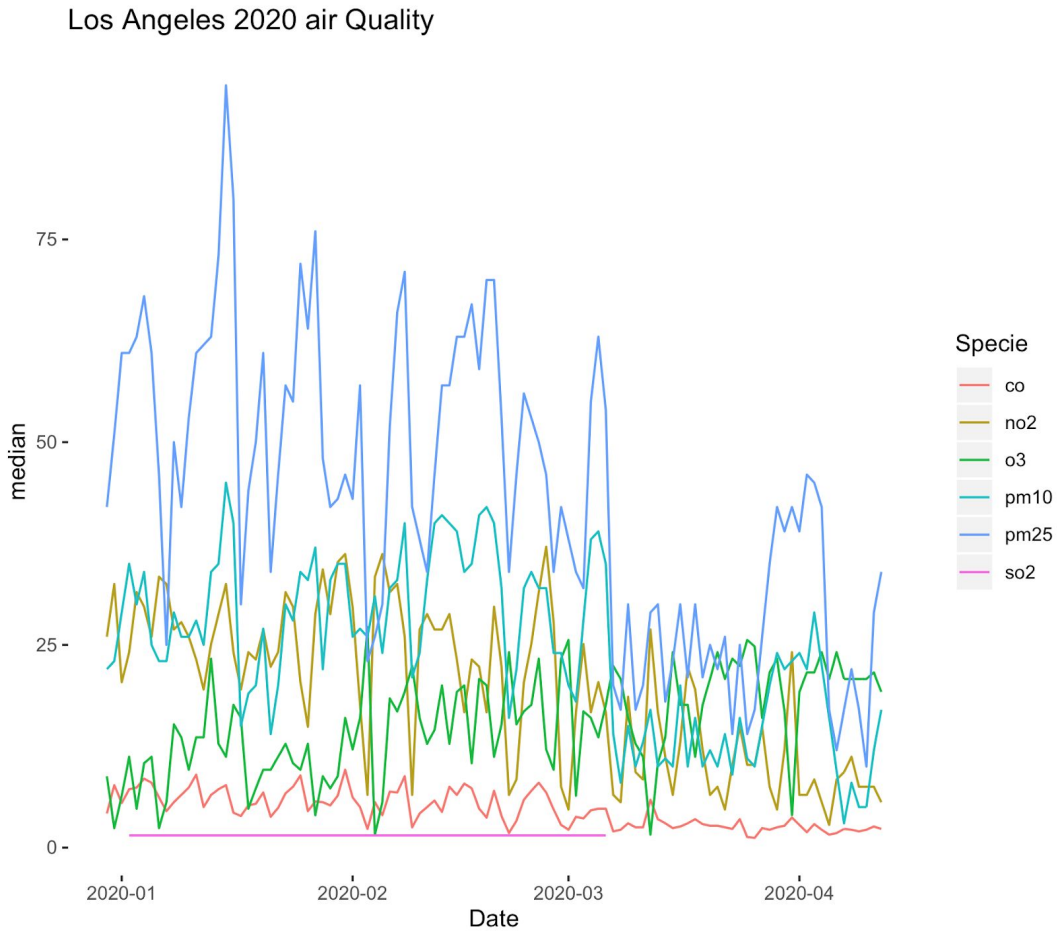
Preliminary Conclusions

Looking at measurements of air quality from 2019 and 2020 we hoped to discover some improvements in air quality as a result of COVID-19 social distancing policies. After loading the data in with pyspark we created some tables using SQL context and select statements to aggregate the data. At first glance, the data does indicate some differences in air quality measurements from last year compared to this year's recent months. To gain a better understanding of the air quality data we did some aggregations using R, to explore which cities had the highest median level of no2 emissions from 2019 and 2020. We found that Jerusalem, Tel Aviv, Soeul and Milan were some of the cities with the highest concentrations of NO_2. After noticing that Seoul was consistently in the top 10 for different measurements of air pollutant concentration we decided to look further into the data in Seoul. The average maximum emissions in Seoul in 2019 was 92.72 parts per million of pollutant concentration, whereas the average maximum emissions for 2020 were 77.90 representing a 23% reduction. We also noticed the difference between emissions in 2019 and 2020 in Los Angeles was substantial. We then created a graph displaying each air quality measurement from January to April in 2019. The graph demonstrates a general consistency among air quality measurements in 2019.



Looking at the same measurements for 2020 we see a decline in the various measurements starting around mid February. Measurements of pm25, pm10 co, and no2 seem to have declined in the last few months compared to where they were in the

previous months and last year.



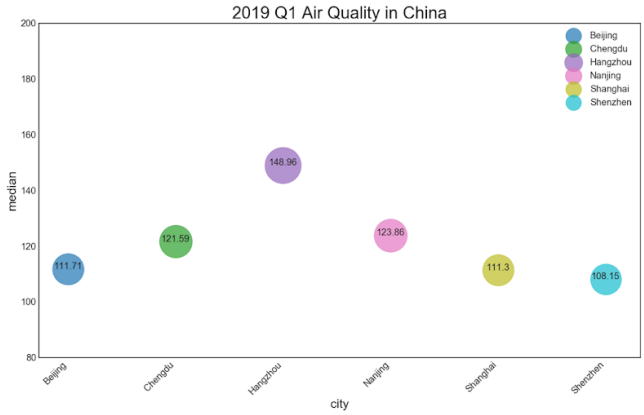
Analyze the air quality between 2019 Q1 and 2020 Q1

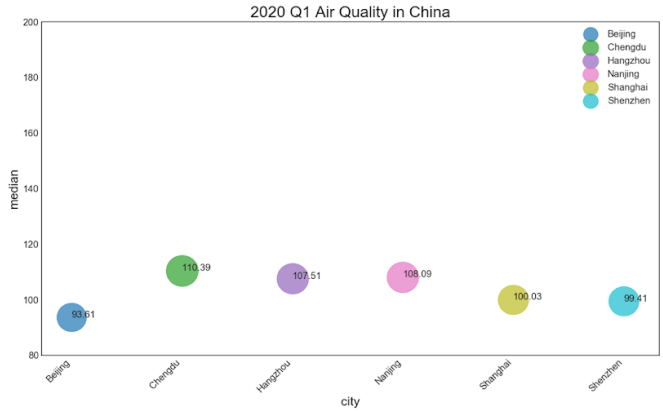
One of our missions for this project is to see how the air quality changes during the COVID-19 pandemic. To get more insightful and practical information about the changes in the air quality, we decided to make a comparison between 2019 Q1 (when COVID-19 hasn't happened) and 2020 Q1 (COVID-19 outbreak start). Moreover, we picked six major cities in the U.S. and China to investigate their air quality changes

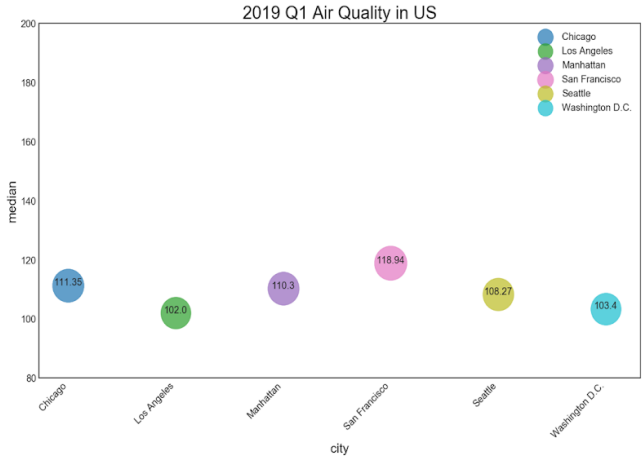
Process

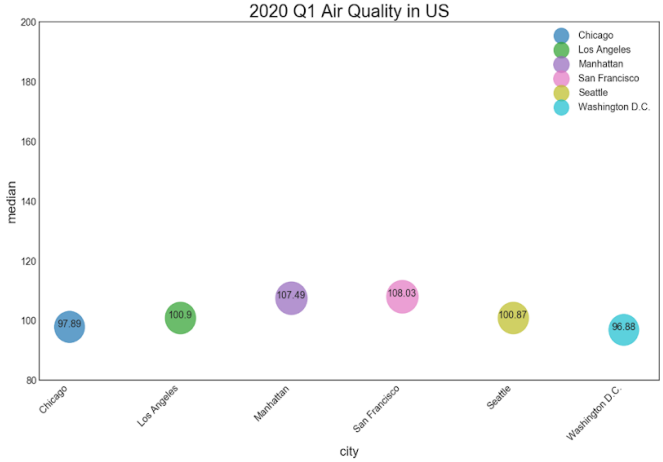
1. Extract six cities in China and in the US 2019 and 2020
2. Group by group by city and median
3. Start to create bubble plot for 2019 Q1 and 2020Q2 air quality in China and US main cities

Results


2019 Q1 Air Quality in China
median
city
Beijing
Chengdu
Hangzhou
Nanjing
Shanghai
Shenzhen
111.71
121.59
148.96
123.86
111.3
108.15


2020 Q1 Air Quality in China
median
city
Beijing
Chengdu
Hangzhou
Nanjing
Shanghai
Shenzhen
93.61
110.39
107.51
108.09
100.03
99.41


2019 Q1 Air Quality in US
median
city
Chicago
Los Angeles
Manhattan
San Francisco
Seattle
Washington D.C.
111.35
102.0
110.3
118.94
108.27
103.4


2020 Q1 Air Quality in US
median
city
Chicago
Los Angeles
Manhattan
San Francisco
Seattle
Washington D.C.
97.89
100.9
107.49
108.03
100.87
96.88

Discussion of relations between covid19 confirmed numbers and air quality in three different countries: United states, India, China.

First, viewing the PM10 changing V.S. covid19 confirmed numbers in each country, assumed the humidity and wind speed can be neglectable, we can see plotting in time series dates only India (Figure 1) has the obvious curve going downward when confirmed numbers of covid19 goes up; in United states (Figure 2) and China (Figure 3), the interaction plots show there are no strong relations between PM10 and confirmed numbers.

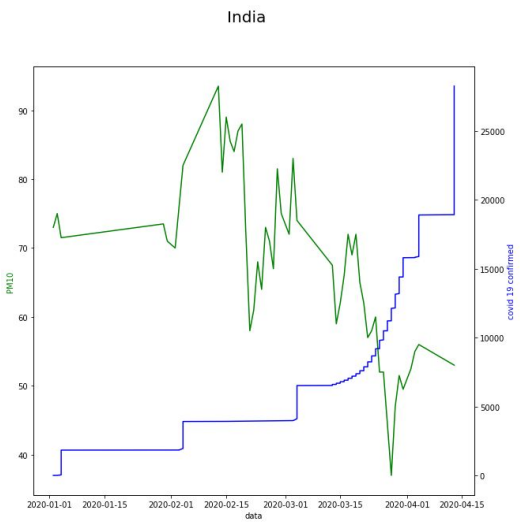


Figure1: PM10 vs covid19 in India

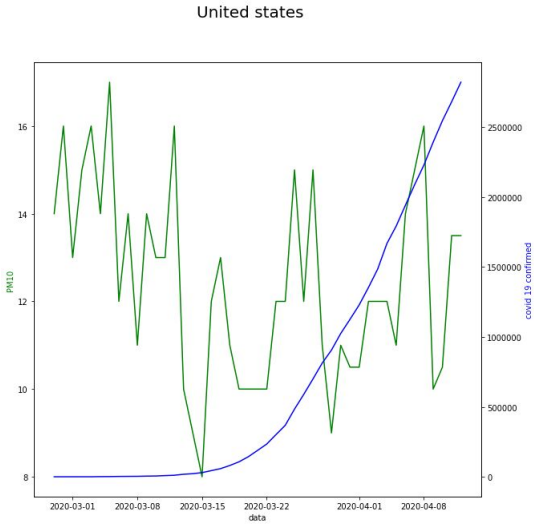


Figure2: PM10 vs covid19 in USA

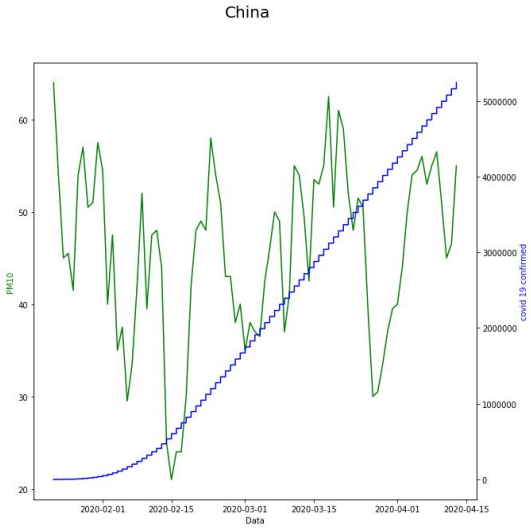


Figure3: PM10 vs covid19 in China

Second, viewing the PM2.5 changing V.S. covid19 confirmed numbers in each country, we can see the result is similar as PM2.5 has downward trend when confirmed numbers goes up (Figure 4); surprisingly results in China has downward curve reflecting the confirmed numbers. However the trend shows not strong relations in USA.

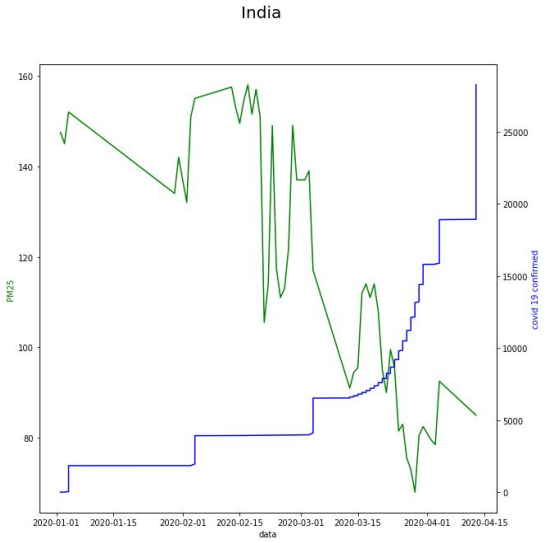


Figure4: PM2.5 vs covid19 in India

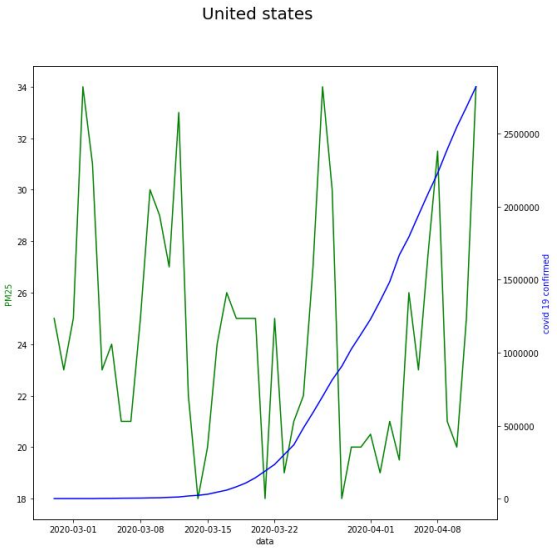


Figure5: PM2.5 vs covid19 in United states

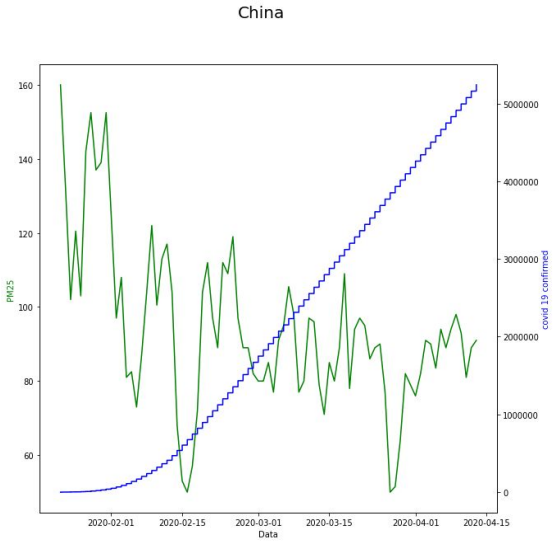


Figure6: PM2.5 vs covid19 in China

Possible factors to explain:

From the result above, assumed confirmed number increases cause human activity decrease in China and India, the PM 2.5 and PM 10 changes are caused by human activity, however we can explain in USA where has less air pollution industries, thus the PM2.5 and PM10 are not related to human activity.

Machine learning

We plan to implement an ARIMA model to predict future improvements in air quality as social distancing policies likely will continue to stand. We will build the model on a subset of the data and use AWS in order to run it on the entire dataset.