# DATA ANALYSIS FOR US AIR POLLUTION

XIANG LIU

XUE XU

# AGENDA

- Project Overview
- Inspirations
- Data Pre-processing
- Data Analysis
- Conclusion
- Reference

# PROJECT OVERVIEW

This dataset includes four major pollutants (Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 - 2016 and place them neatly in a CSV  file

| Pollutants | Cause |
|---|---|
| NO2 | are expelled from high temperature combustion, and are also produced during thunderstorms by electric discharge |
| SO2 | produced by volcanoes and in various industrial processes |
| O3 | brought about by human activities (largely the combustion of fossil fuel) |
| CO | vehicular exhaust contributes to the majority of carbon monoxide let into our atmosphere |

# TOXITY

| Pollutants | Toxity |
| --- | --- |
| NO2 | Breathing difficulties, Throat spasms, Headache, Fatigue, Nausea |
| SO2 | Temporary respiratory problems, chronic bronchitis, emphysema, decreased fertility, coughing |
| O3 | Irritant effect on the respiratory tract; stimulation of atherosclerosis |
| CO | Confusion, vision and balance problems, Loss of consciousness, Nausea and vomiting, Headaches |

# DATASET

There is a total of 28 fields. The four pollutants (NO2, O3, SO2 and CO) each has 5 specific columns. Observations totaled to over 1.4 million

| Column Name | Description |
|---|---|
| NO2 Units | The units measured for NO2 |
| NO2 Mean | The arithmetic mean of concentration of NO2 within a given day |
| NO2 AQI | The calculated air quality index of NO2 within a given day |
| NO2 1st Max Value | The maximum value obtained for NO2 concentration in a given day |
| NO2 1st Max Hour | The hour when the maximum NO2 concentration was recorded in a given day |

# AQI --- AIR QUALITY INDEX

| Air Quality Index Levels of Health Concern | Numerical Value | Meaning |
|---|---|---|
| Good | 0 to 50 | Air quality is considered satisfactory, and air pollution poses little or no risk. |
| Moderate | 51 to 100 | Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution. |
| Unhealthy for Sensitive Groups | 101 to 150 | Members of sensitive groups may experience health effects. The general public is not likely to be affected. |
| Unhealthy | 151 to 200 | Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects. |
| Very Unhealthy | 201 to 300 | Health warnings of emergency conditions. The entire population is more likely to be affected. |
| Hazardous | 301 to 500 | Health alert: everyone may experience more serious health effects. |

# DATASET

| Column Name | Description |
|---|---|
| State Code | The code allocated by US EPA to each state |
| County Code | The code of counties in a specific state allocated by US EPA |
| Site Num | The site number in a specific county allocated by US EPA |
| Address | Address of the monitoring site |
| State | New Jersey… |
| County | County of monitoring site |
| City | City of the monitoring site |
| Date | LocalDate of monitoring |

# INSPIRATIONS

1. How is the concentration of each pollutant in US states?
2. Which time in a day has the maximum density of air pollutants?
3. From 2000-2016, the concentration of each pollutant increase or decrease?
4. Can we predict the trend of each pollutant in the future?

# DATA PREPROCESSING

## Step 1 check if the data has missing value

```
In [3]: le=data['CO AQI']
        le[le.isnull()]
```

```
Out[3]: 0          NaN
        2          NaN
        4          NaN
        6          NaN
        8          NaN
        10         NaN
        12         NaN
```

```
data = pd.read_csv('/Users/xiangliu/Desktop/CSC560 Data/pollution_us_2000_2016.csv')
data.shape
```

```
(1746661, 29)
```

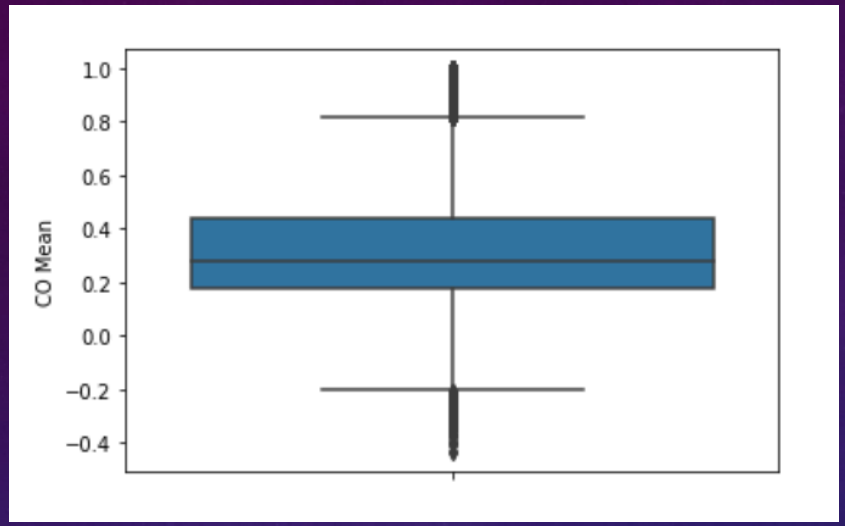
## Step 2 Use dropna to drop missing value

```
##Look Into each value
data=data.dropna()
data.to_csv("/Users/xiangliu/Desktop/CSC560 Data/pollution_AQI.csv",index=True,sep=',')
```

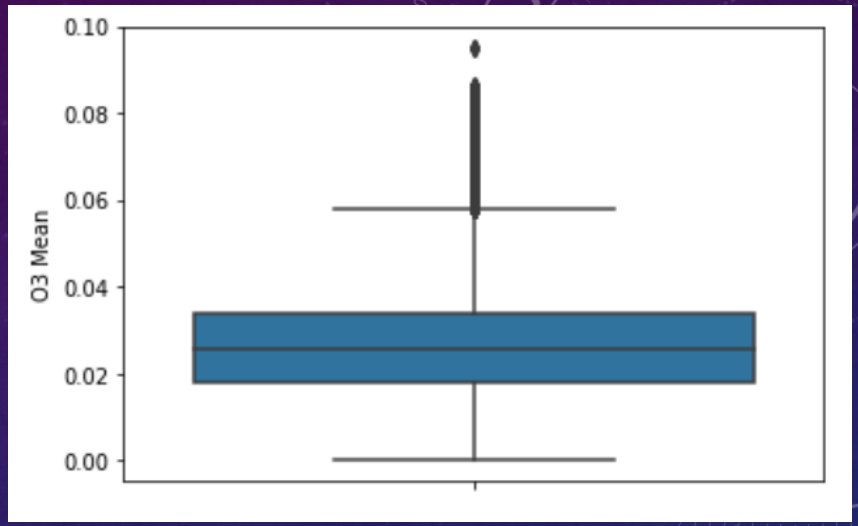
```
data = pd.read_csv('/Users/xiangliu/Desktop/CSC560 Data/pollution_AQI.csv')
data.shape
```

```
(436876, 30)
```

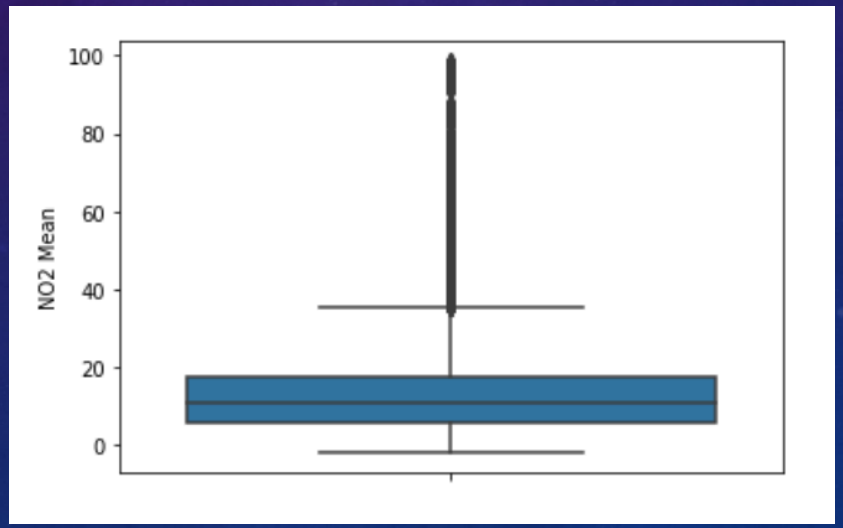
## Step 3 Check unusual value



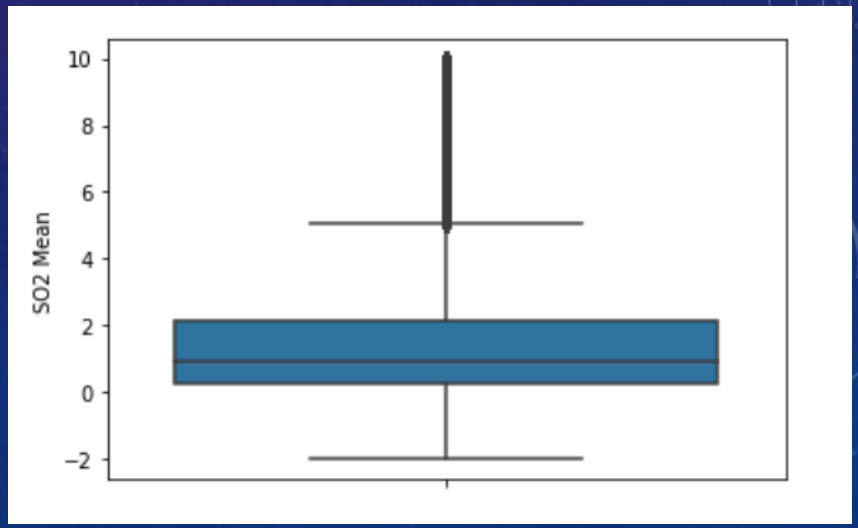

CO Mean




O3 Mean




NO2 Mean




SO2 Mean

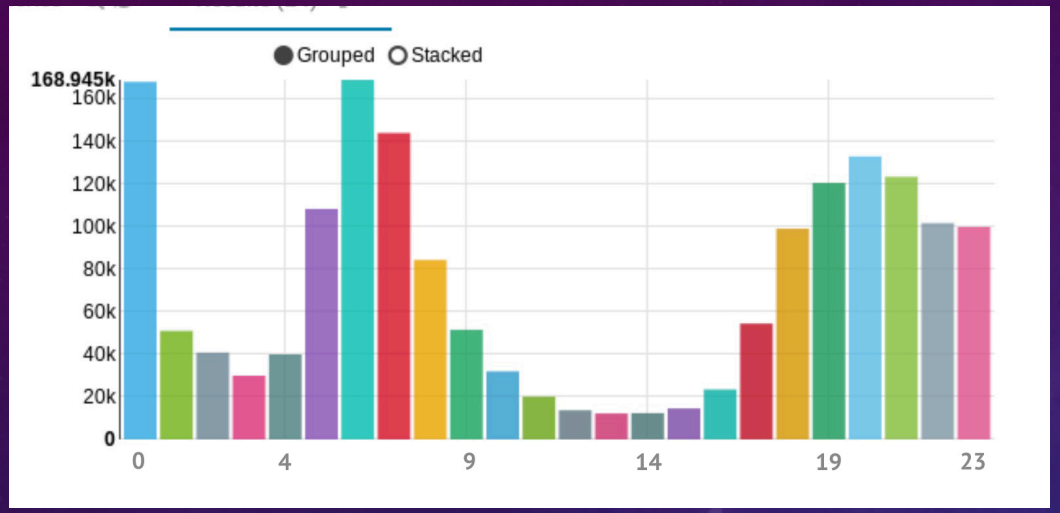
# HADOOP M/R + HIVE

```python
#!/usr/bin/env python
import sys

# input comes from STDIN (standard input)

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split('|')
    # write the results to STDOUT (standard output);
    # pick the volumn index 13, 18, 19 for NO2 1st Max Hour,03 1st Max Hour,
    #NO2 1st Max Hour, CO 1st Max Hour
    print('%d\t%d\t%d\t%d' % (words[13],words[18],words[23], words[28]))
```

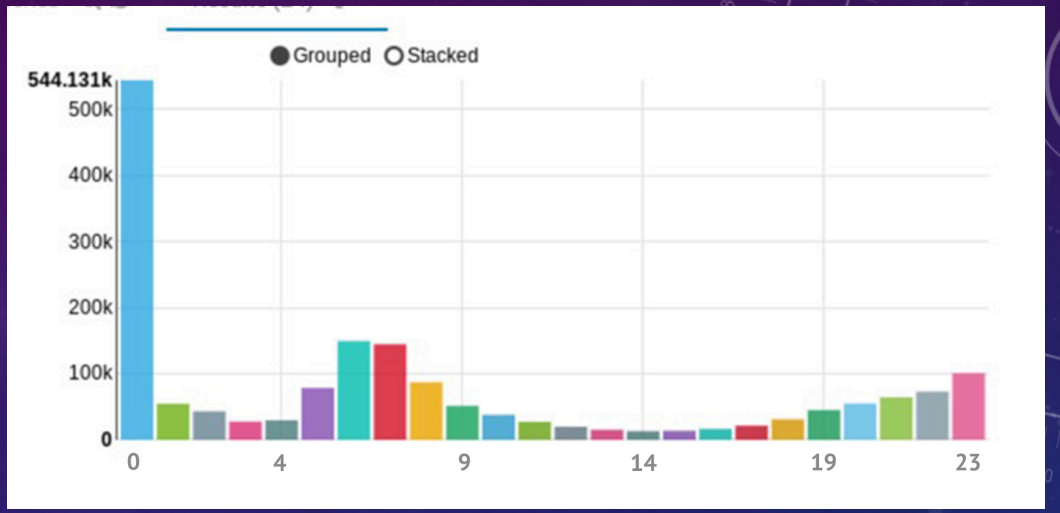
hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-2.6.0-mr1-cdh5.13.0.jar
-input /tmp/data/pollu1.csv
-output /tmp/data/out_put01
-mapper /home/cloudera/myPractice/mapper.py
-reducer /home/cloudera/myPractice/reducer.py

```python
#!/usr/bin/env python
from operator import itemgetter
import sys

NO2=[]
O3=[]
SO2=[]
CO=[]
dic_NO2={}
dic_O3={}
dic_SO2={}
dic_CO={}
current_count = 0
word = None
# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # parse the input we got from mapper.py
    hour = line.split('\t')

    # convert hour (currently a string) to int
    try:
        hour = int(hour)
    except ValueError:
        # hour was not a number, so silently ignore/discard this line
        continue
    # put the four number into four different lists
    NO2.append(hour[0])
    O3.append(hour[1])
    SO2.append(hour[2])
    CO.append(hour[3])

# read and count the frequency of hours
for key in NO2:
    dic_NO2[key] = dic_NO2.get(key, 0) + 1
print(dic_NO2[key])

for key in O3:
    dic_O3[key] = dic_O3.get(key, 0) + 1
print(dic_O3[key])

for key in SO2:
    dic_SO2[key] = dic_SO2.get(key, 0) + 1
print(dic_SO2[key])

for key in CO:
    dic_CO[key] = dic_CO.get(key, 0) + 1
print(dic_CO[key])
```

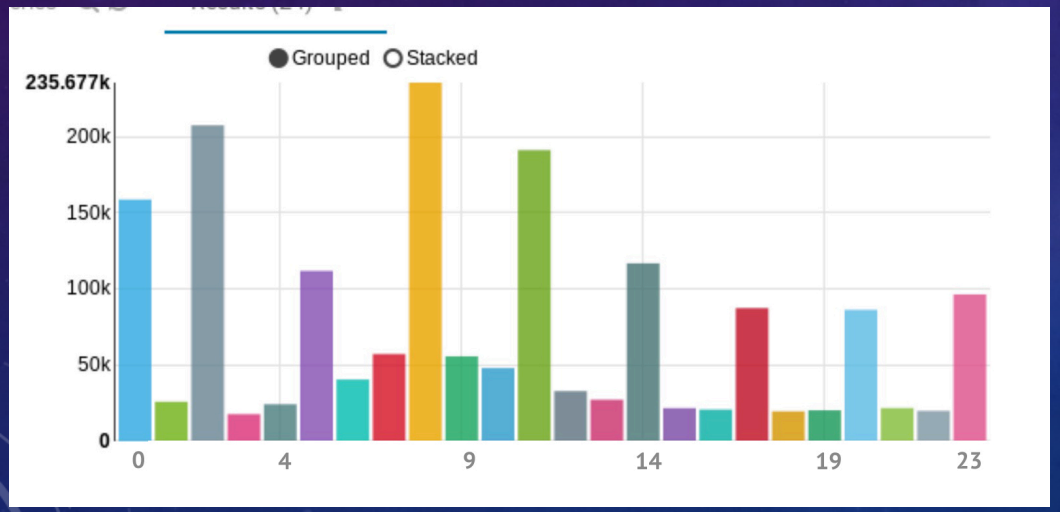
# DATA ANALYSIS IN HIVE



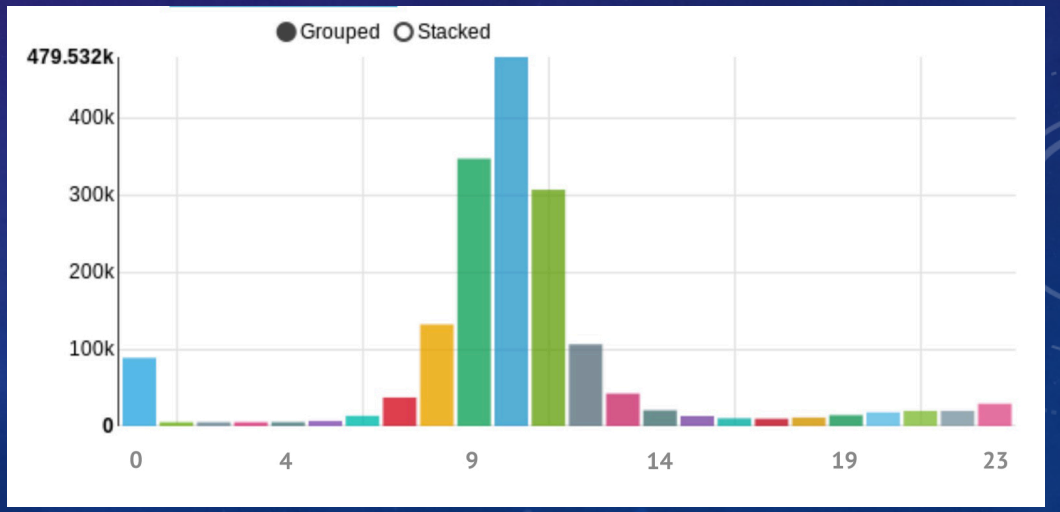

NO2_1st_max_hour




CO_1st_max_hour




SO2_1st_max_hour

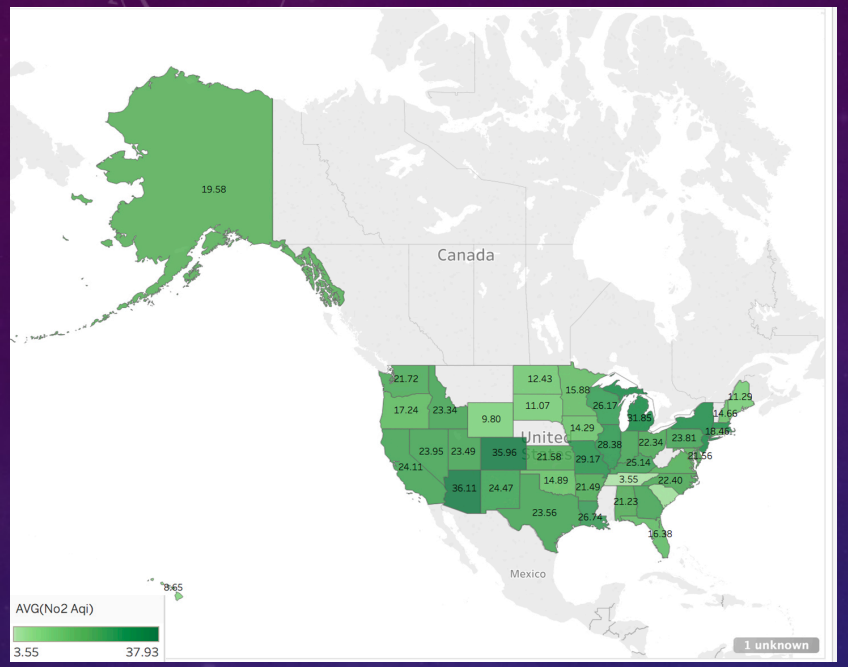



O3_1st_max_hour

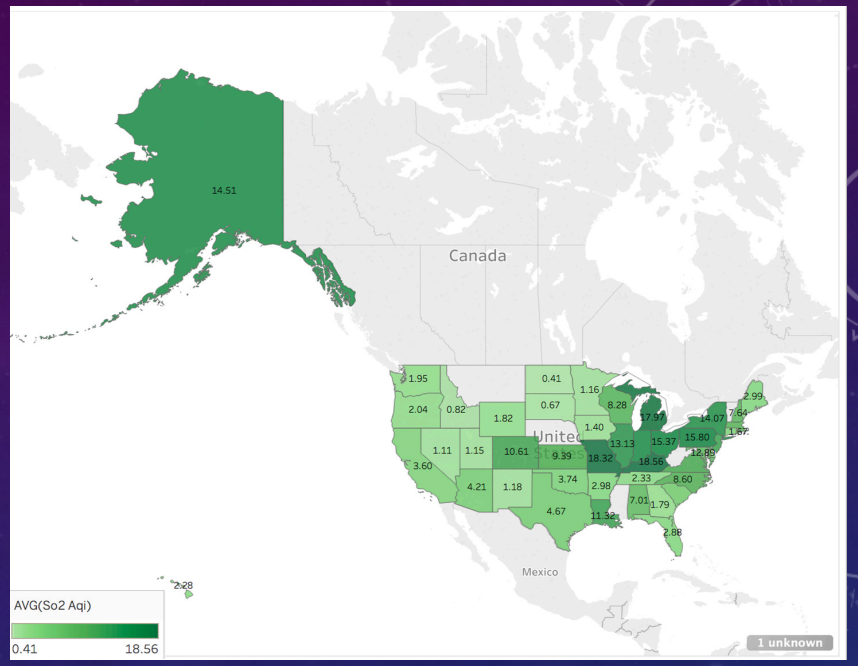
# AQI AVG VALUE
## (STATE)

Except Montana, Nebraska, Mississippi, West Vigina, Vermont these 5 states

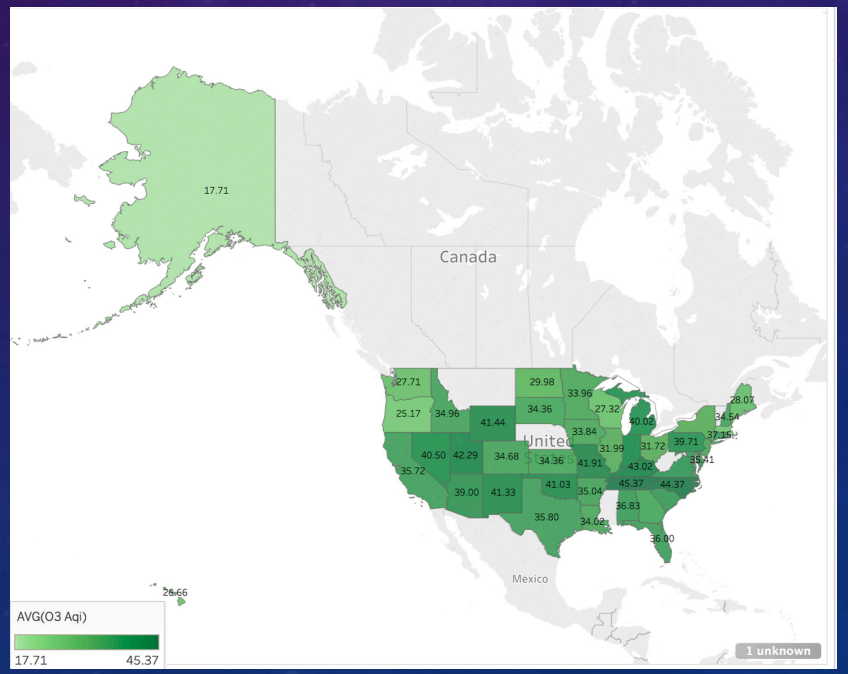
NO2 AQI



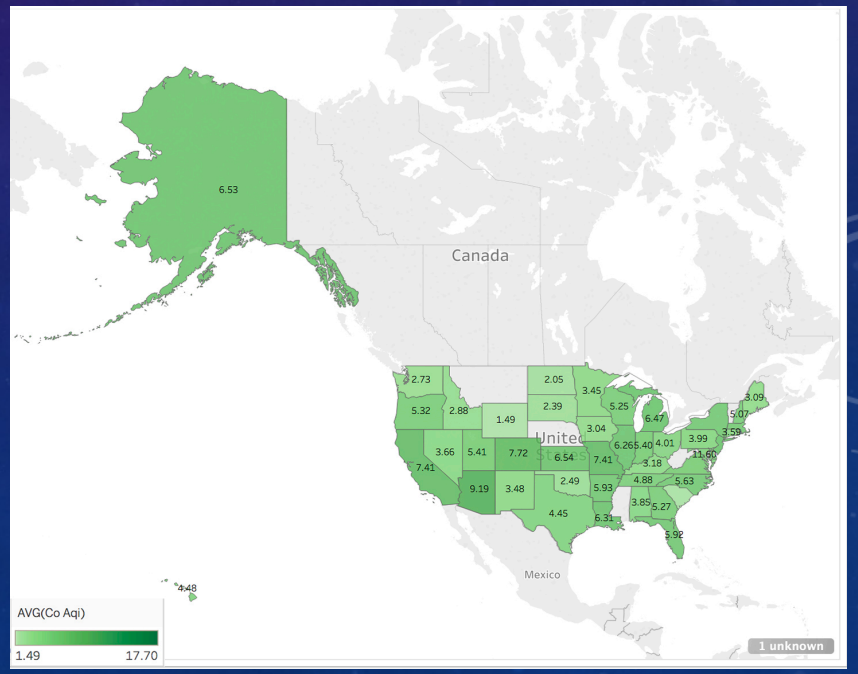

SO2 AQI




O3 AQI

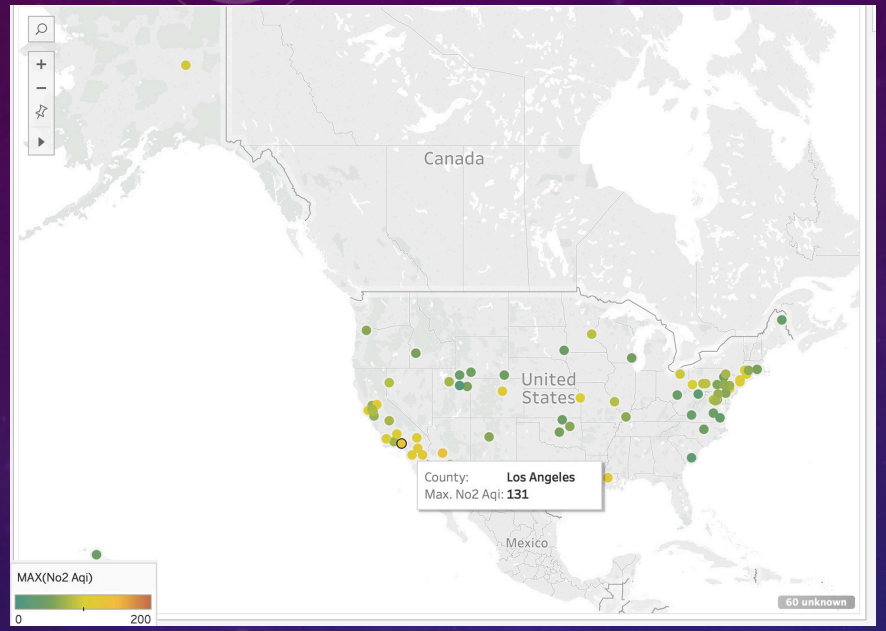



CO AQI

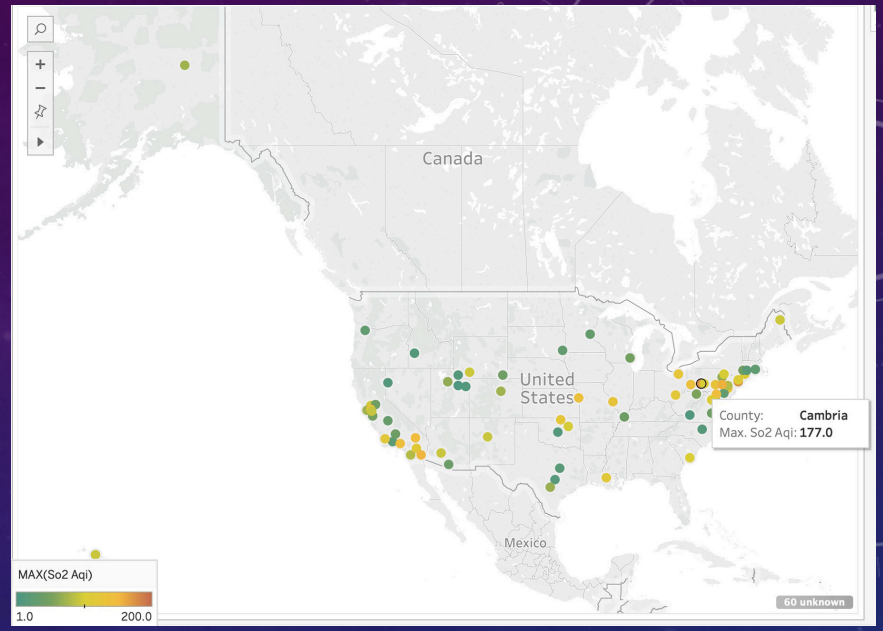


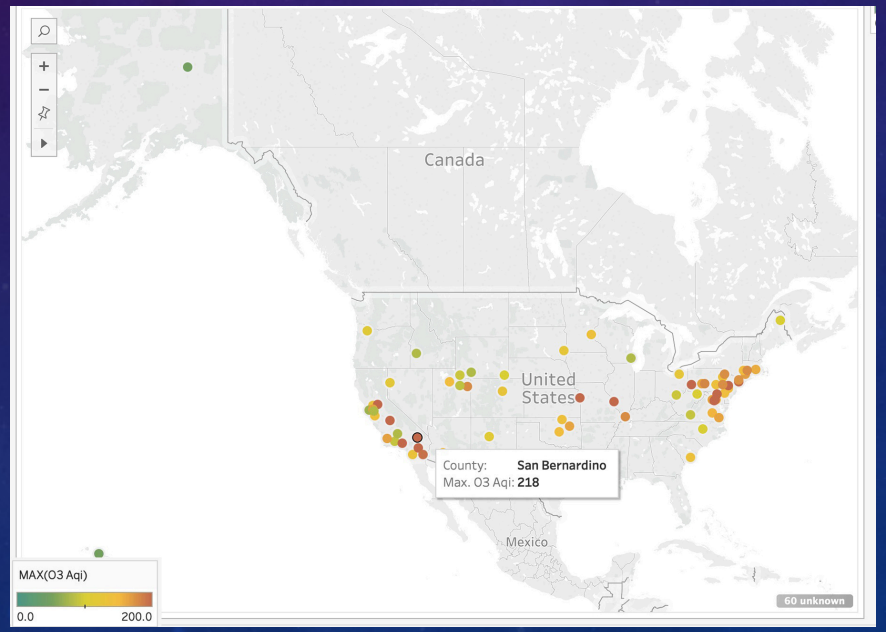
# AQI MAX VALUE
## (COUNTY)

NO2 AQI

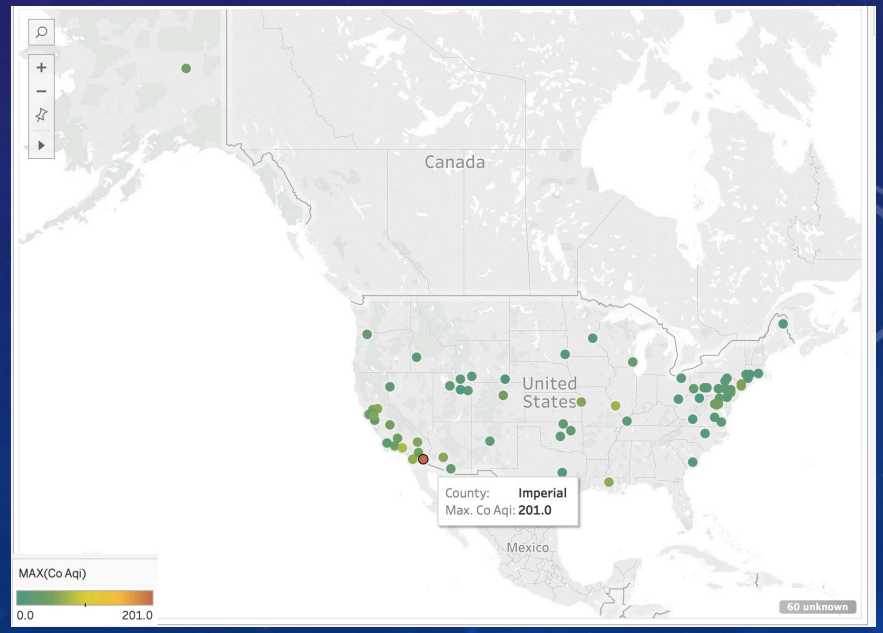



SO2 AQI



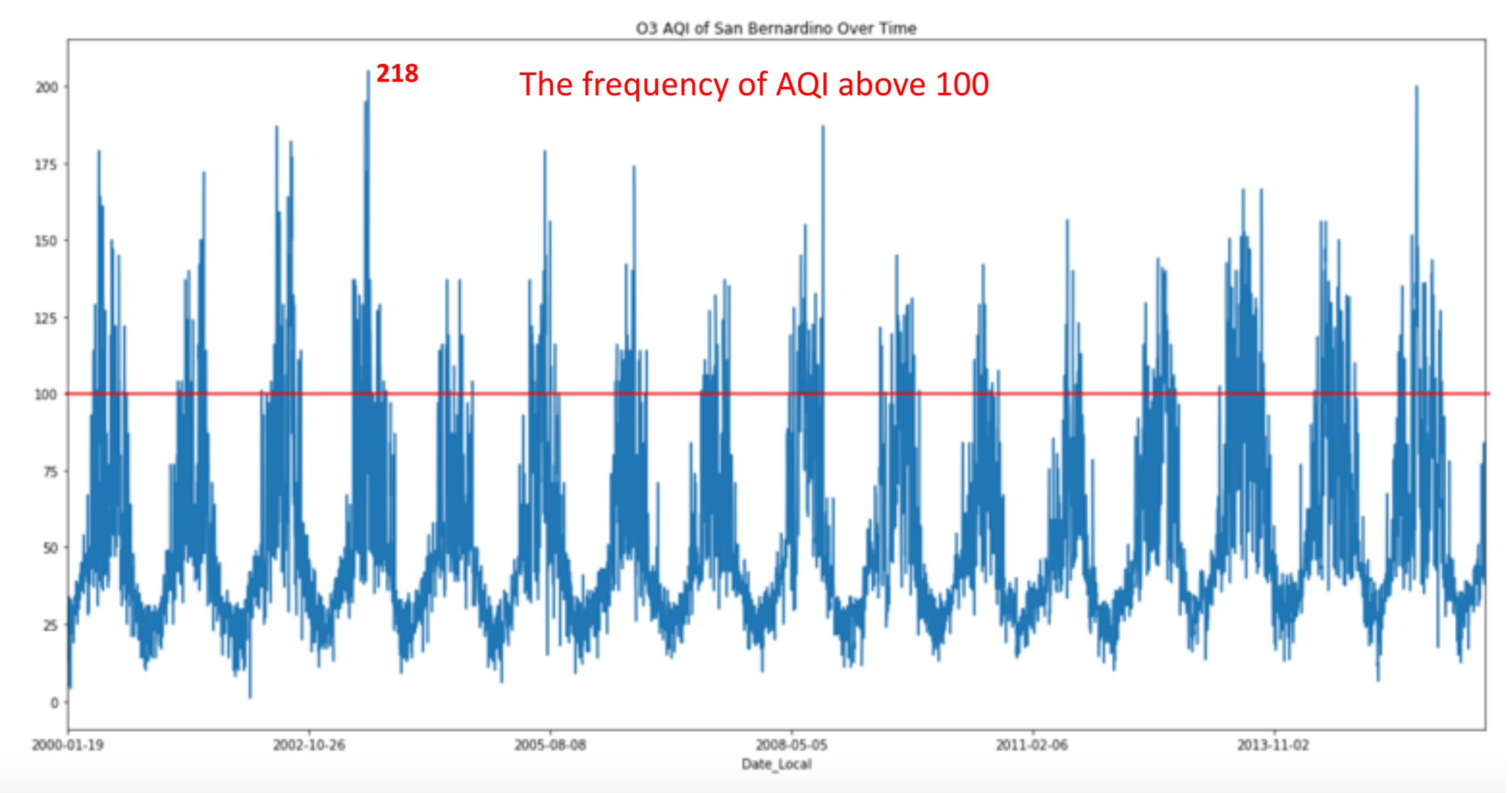




O3 AQI

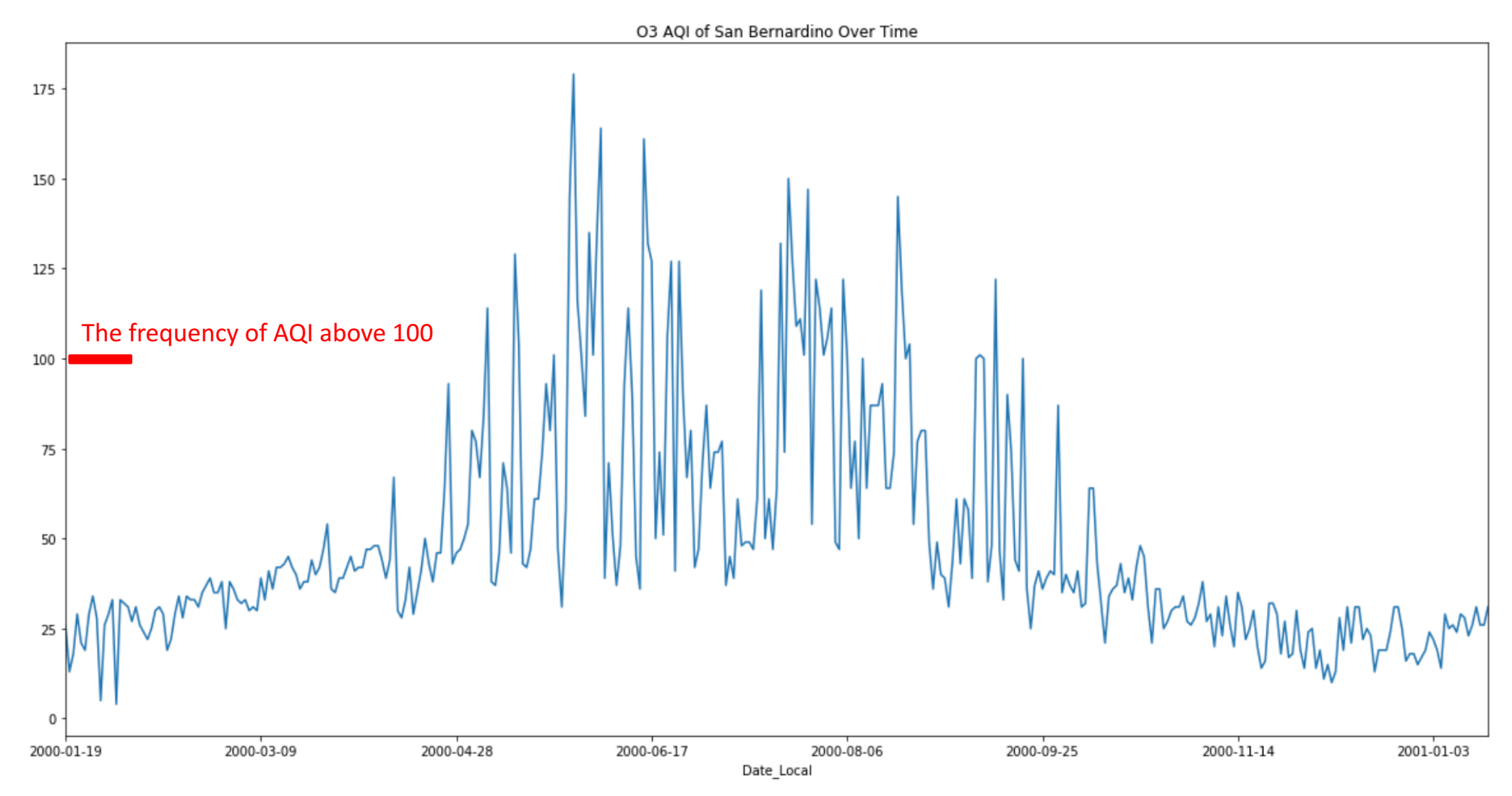



CO AQI

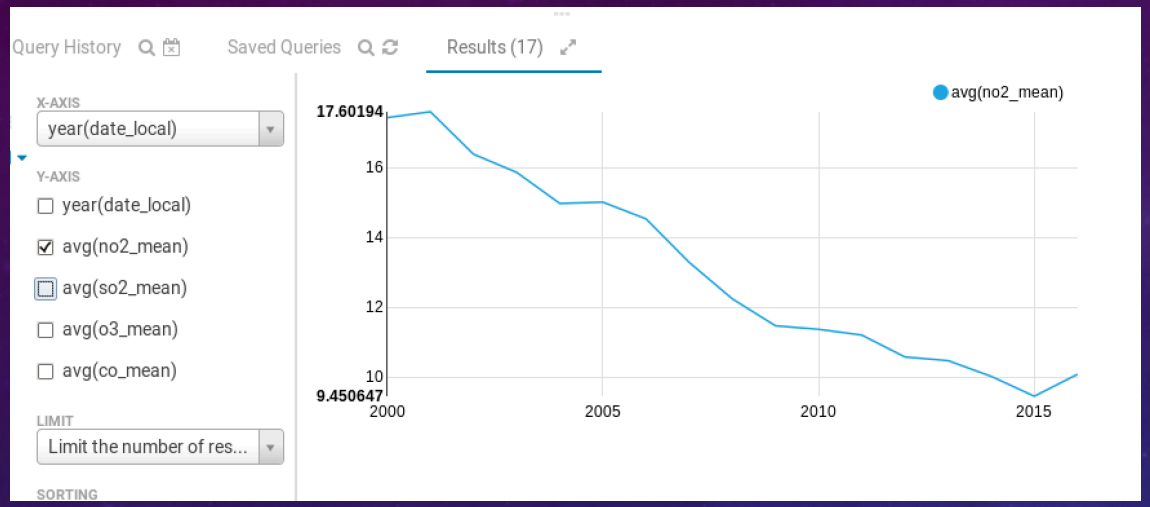
# San Bernardino O3 AQI Analysis (2000-2016)



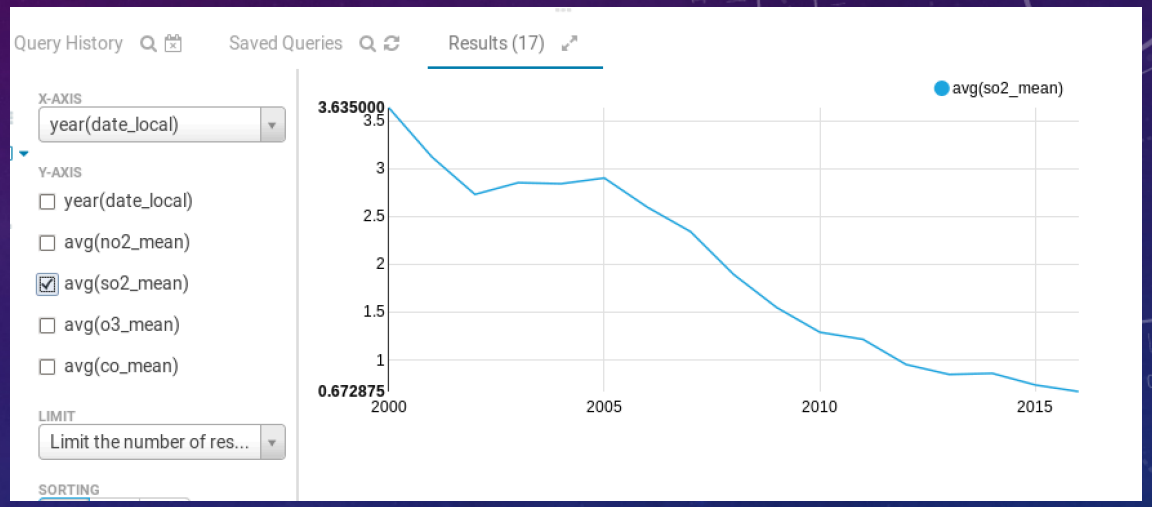
# San Bernardino O3 AQI Analysis (2000)



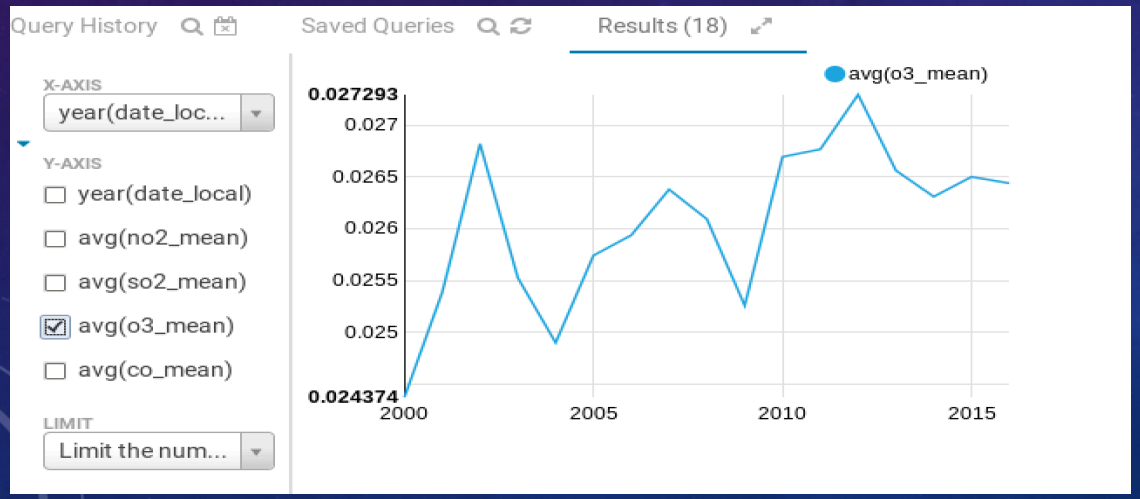
# DATA ANALYSIS IN HIVE



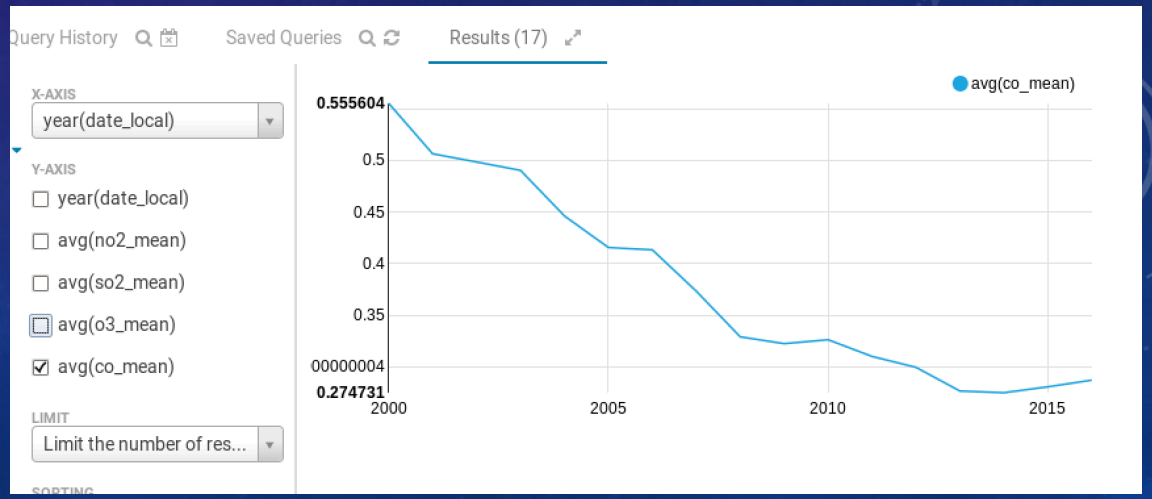

NO2_mean




SO2_mean



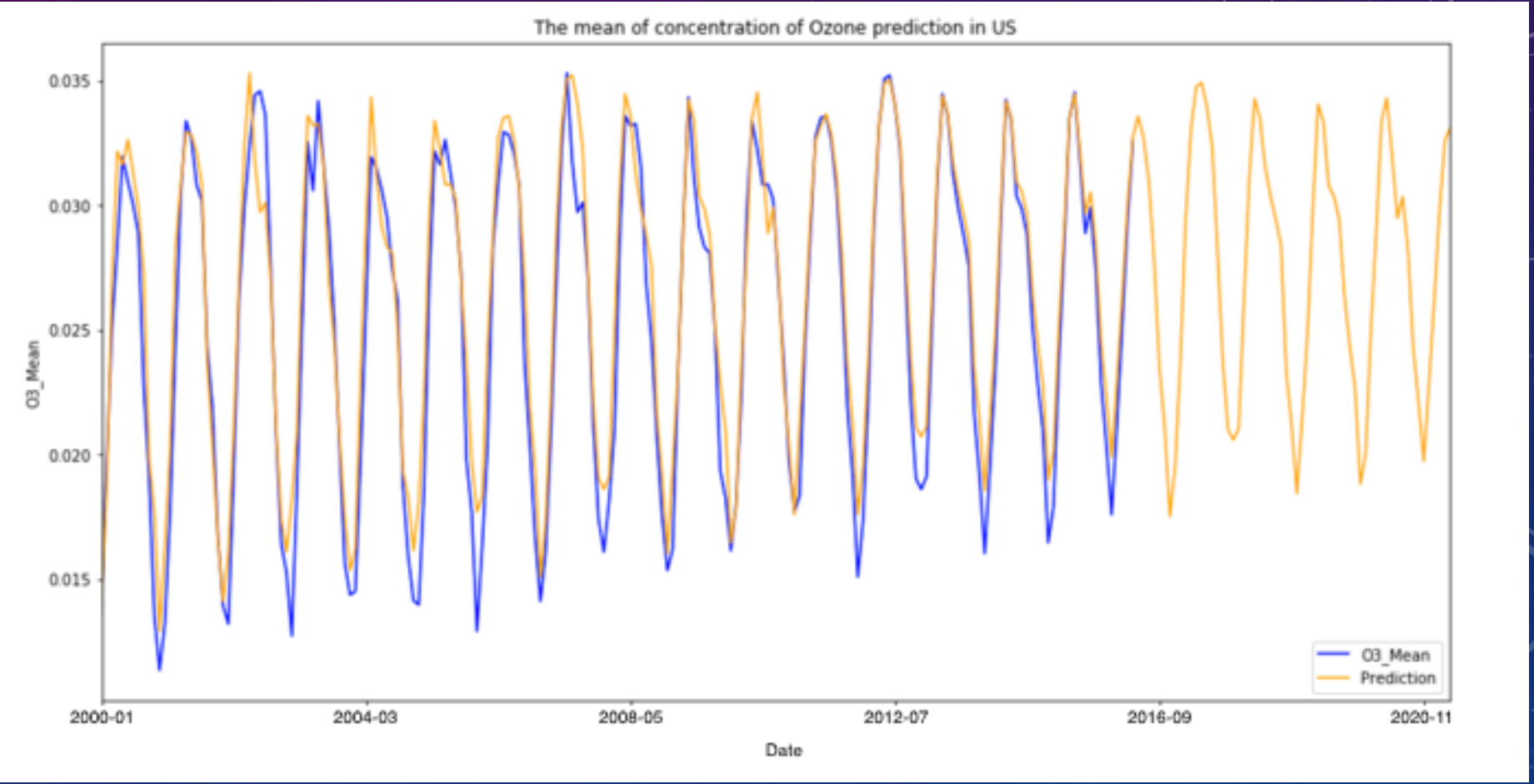

O3_mean




CO_mean

# LINEAR REGRESSION MODEL

- Sklearn for the regression algorithm
- Split our testing and training data sets, test_size equal to 20% of the data
- Model accuracy is 0.95




US O3_Mean prediction

# SUMMARY

1.Max Value Hour's study
-For NO2, and CO, most of the places have their max value between 5-9 am and 5- 12 pm
-For O3, the max value generally appears at 8 am- 1pm
-For SO2, the peak appears every three hours

2. Mean Value by year study
-NO2, SO2 and CO are declining every year gradually
-O3 is in a stable level with fluctuations

3. Pollution distribution
-Overall, the average AQI for each state is under the safe line (<50)
-Counties in East and West coast have slight pollution on NO2, SO2 ; CO is under the safe level; O3 values in some of the counties have exceed the safe level, might cause health issue

4. Study of O3 in a specific area (San Bernardino)
-O3 AQI presents periodical changes, peak appears between May - October

# REFERENCE


- https://en.wikipedia.org/wiki/Air_pollution
- https://enlight.nyc/projects/stock-market-prediction
- https://www.kaggle.com/

# THANK YOU FOR YOUR ATTENTION