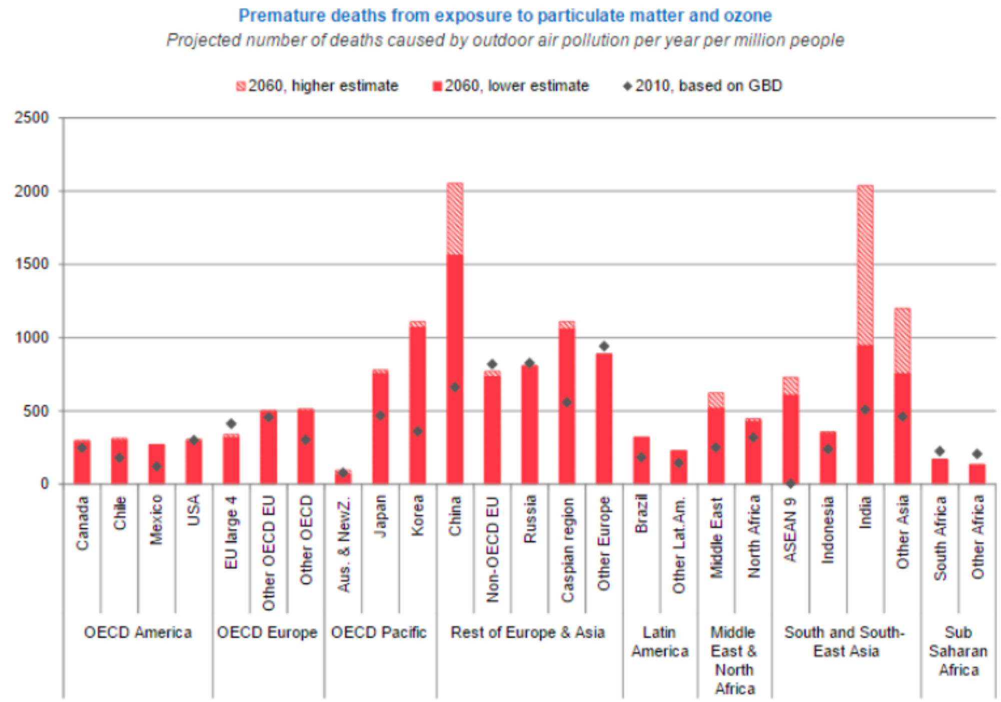
Final_Project (*Clean_air*)

DATA-603: Platforms for Big Data Processing
Instructor: Prof. Andy Enkeboll
By: Chiranjib Dutta

The Problem Statement

A new OECD report, *The Economic Consequences of Outdoor Air Pollution*, estimates that outdoor air pollution will cause 6-9 million premature deaths annually by 2060, compared to three million in 2010. That is equivalent to a person dying every 4-5 seconds. Cumulatively, more than 200 million people will die prematurely in the next 45 years as a result of air pollution.
Cleaner technologies are available, with the potential to improve air quality considerably. But policymakers tend to focus myopically on the costs of action, rather than the costs of inaction. With economic growth and rising energy demand set to fuel a steady rise in emissions of air pollutants and rapidly rising concentrations of particulate matter (PM) and ozone in the coming decades, this approach is untenable.(World Economic Forum)



(WEF)

Proposals

The motto of this project is:

- To determine the Air Quality Index(AQI) at different regions of the United States.
- Grade AQI as per quality
- Devise a machine learning model to predict the air quality in advance so that people become aware.

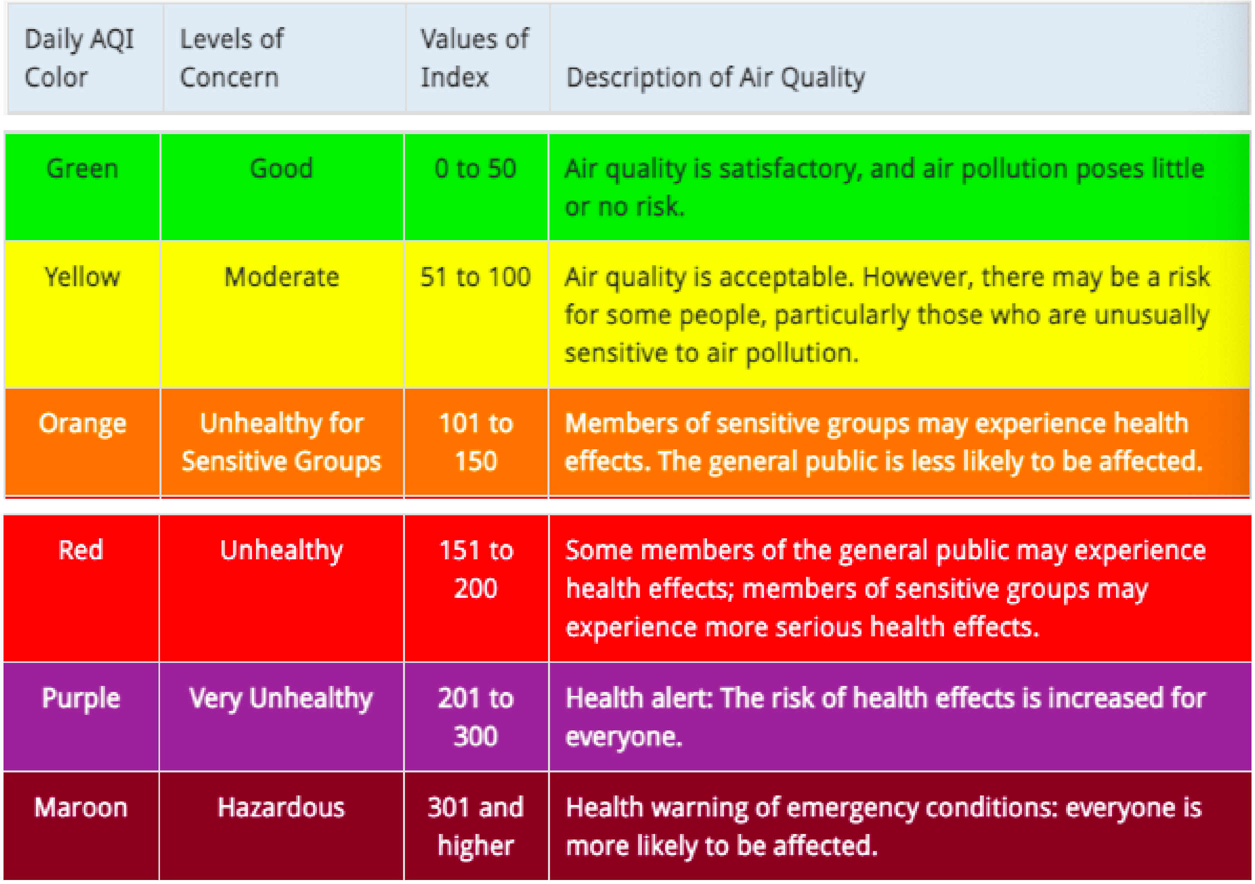
Data Source

For this project we have used US pollution data from kaggle:
https://www.kaggle.com/sogun3/uspollution
This dataset deals with pollution in the U.S. Pollution in the U.S. has been well documented by the U.S. EPA but as it is a pain to download all the data and arrange them in a format that interests data scientists. The data provider gathered four major pollutants (Nitrogen Dioxide, Sulfur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 - 2016 and placed them neatly in a CSV file. The **file contains 1746661 (~1.75 million)rows and 28 columns.**
The original data is scraped from the database of the U.S. EPA :
https://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

The basics

How does the AQI work?

Think of the AQI as a yardstick that runs from 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern. For example, an AQI value of 50 or below represents good air quality, while an AQI value over 300 represents hazardous air quality.
For each pollutant an AQI value of 100 generally corresponds to an ambient air concentration that equals the level of the short-term national ambient air quality standard for protection of public health. AQI values at or below 100 are generally thought of as satisfactory. When AQI values are above 100, air quality is unhealthy: at first for certain sensitive groups of people, then for everyone as AQI values get higher.
The AQI is divided into six categories. Each category corresponds to a different level of health concern. Each category also has a specific color. The color makes it easy for people to quickly determine whether air quality is reaching unhealthy levels in their communities



Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

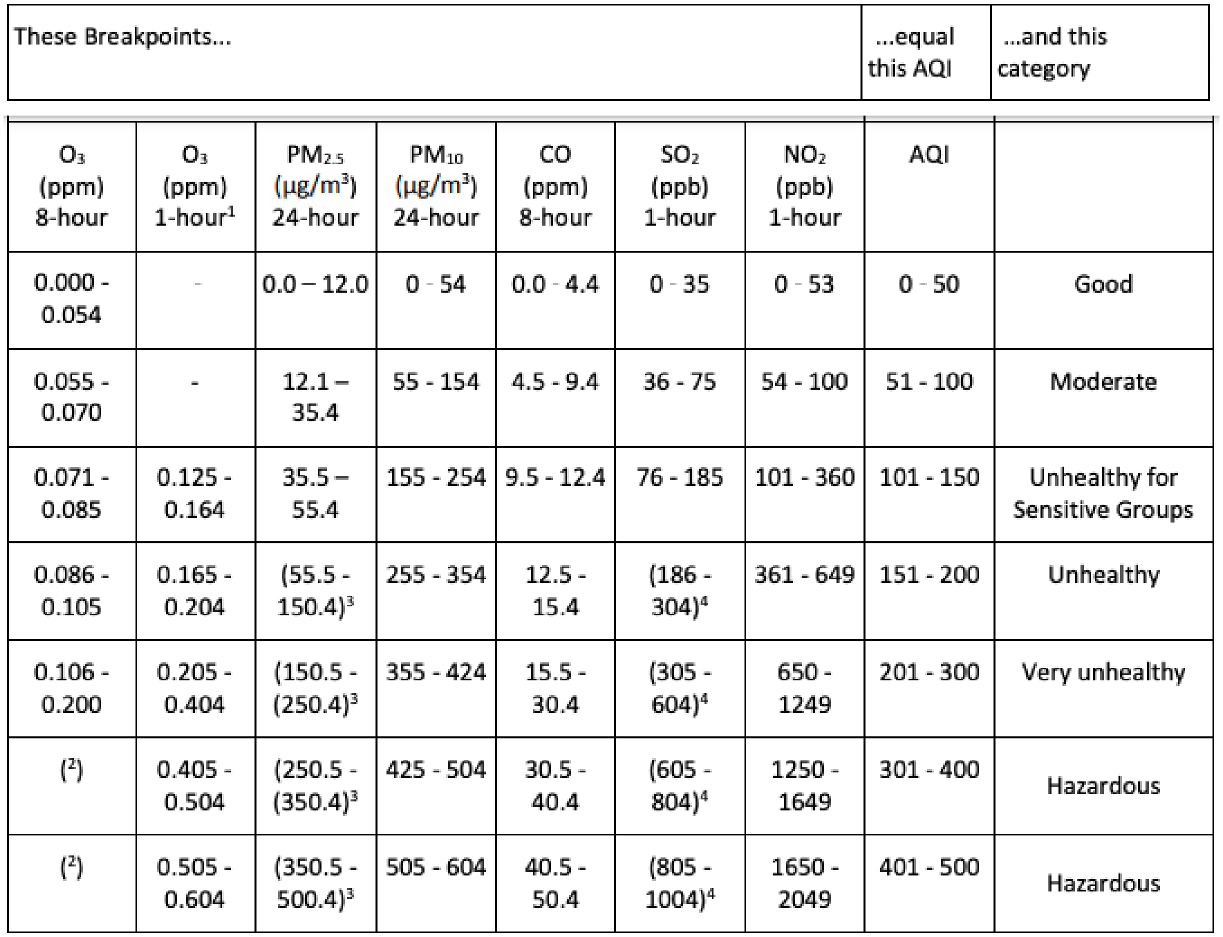
airnow.gov

Five major pollutants

EPA establishes an AQI for five major air pollutants regulated by the Clean Air Act. Each of these pollutants has a national air quality standard set by EPA to protect public health:

- ground-level ozone
- particle pollution (also known as particulate matter, including PM2.5 and PM10)
- carbon monoxide
- sulfur dioxide
- nitrogen dioxide

Breakpoints for the AQI

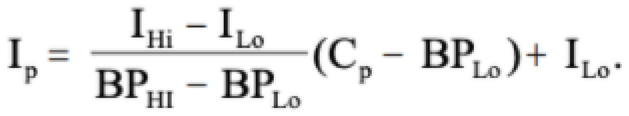


These Breakpoints...	...equal this AQI	...and this category

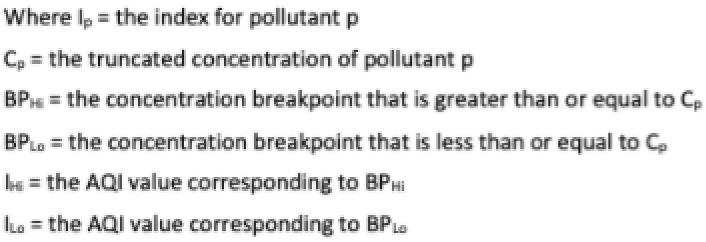
O_3 (ppm) 8-hour	O_3 (ppm) 1-hour[1]	$PM_{2.5}$ ($\mu g/m^3$) 24-hour	PM_{10} ($\mu g/m^3$) 24-hour	CO (ppm) 8-hour	SO_2 (ppb) 1-hour	NO_2 (ppb) 1-hour	AQI	
0.000 - 0.054	-	0.0 – 12.0	0 - 54	0.0 - 4.4	0 - 35	0 - 53	0 - 50	Good
0.055 - 0.070	-	12.1 – 35.4	55 - 154	4.5 - 9.4	36 - 75	54 - 100	51 - 100	Moderate
0.071 - 0.085	0.125 - 0.164	35.5 – 55.4	155 - 254	9.5 - 12.4	76 - 185	101 - 360	101 - 150	Unhealthy for Sensitive Groups
0.086 - 0.105	0.165 - 0.204	(55.5 - 150.4)[3]	255 - 354	12.5 - 15.4	(186 - 304)[4]	361 - 649	151 - 200	Unhealthy
0.106 - 0.200	0.205 - 0.404	(150.5 - (250.4)[3]	355 - 424	15.5 - 30.4	(305 - 604)[4]	650 - 1249	201 - 300	Very unhealthy
([2])	0.405 - 0.504	(250.5 - (350.4)[3]	425 - 504	30.5 - 40.4	(605 - 804)[4]	1250 - 1649	301 - 400	Hazardous
([2])	0.505 - 0.604	(350.5 - 500.4)[3]	505 - 604	40.5 - 50.4	(805 - 1004)[4]	1650 - 2049	401 - 500	Hazardous

Airnow.gov

The formula to determine the air quality index of a place using the table above



$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{HI} - BP_{Lo}}(C_p - BP_{Lo}) + I_{Lo}.$$



Where I_p = the index for pollutant p

C_p = the truncated concentration of pollutant p

BP_{Hi} = the concentration breakpoint that is greater than or equal to C_p

BP_{Lo} = the concentration breakpoint that is less than or equal to C_p

I_{Hi} = the AQI value corresponding to BP_{Hi}

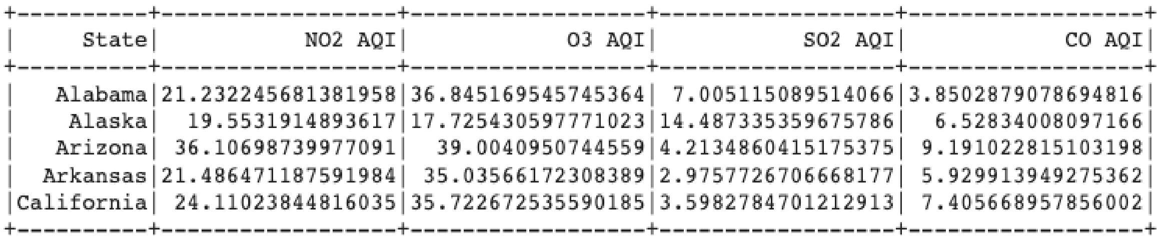
I_{Lo} = the AQI value corresponding to BP_{Lo}

Determining State wise and month wise air quality index (AQI)

AQI for SO2, NO2, O3 and CO is segregated out for all the states of the United states. The segregation is further splitted month wise for each state.

State wise AQI

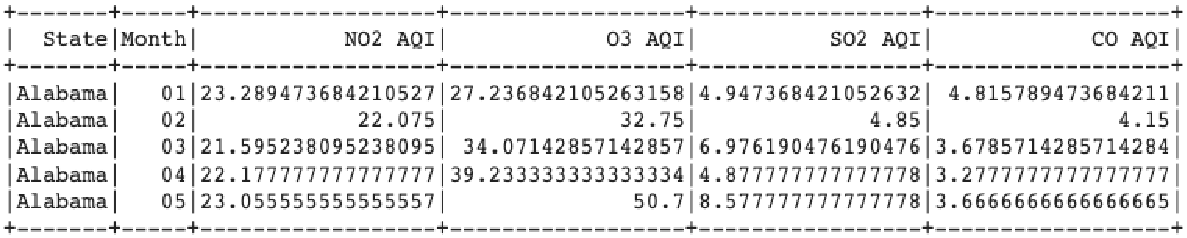
```
from pyspark.sql.functions import countDistinct, avg
from pyspark.sql.functions import desc
dfGroupBy = lines.groupBy("State").agg(avg("NO2 AQI").alias("NO2 AQI"), \
                                      avg("O3 AQI").alias("O3 AQI"), \
                                      avg("SO2 AQI").alias("SO2 AQI"), \
                                      avg("CO AQI").alias("CO AQI")) \
.sort("State") \
.show(5)
```



```
+----------+------------------+------------------+------------------+------------------+
|     State|           NO2 AQI|            O3 AQI|           SO2 AQI|            CO AQI|
+----------+------------------+------------------+------------------+------------------+
|   Alabama|21.232245681381958|36.845169545745364| 7.005115089514066|3.8502879078694816|
|    Alaska|  19.5531914893617|17.725430597771023|14.487335359675786|   6.52834008097166|
|   Arizona| 36.10698739977091|  39.0040950744559|4.2134860415175375| 9.191022815103198|
|  Arkansas|21.486471187591984| 35.03566172308389|2.9757726706668177| 5.929913949275362|
|California| 24.11023844816035|35.722672535590185|3.5982784701212913| 7.405668957856002|
+----------+------------------+------------------+------------------+------------------+
```

Month wise AQI

```
from pyspark.sql.functions import countDistinct, avg, col
from pyspark.sql.functions import desc
dfGroupBy = lines.withColumn("Month", col("Date Local").substr(6, 2)) \
.groupBy("State", "Month").agg(avg("NO2 AQI").alias("NO2 AQI"), \
                                      avg("O3 AQI").alias("O3 AQI"), \
                                      avg("SO2 AQI").alias("SO2 AQI"), \
                                      avg("CO AQI").alias("CO AQI")) \
.sort("State", "Month") \
.show(5)
```



```
+-------+-----+------------------+------------------+------------------+------------------+
|  State|Month|           NO2 AQI|            O3 AQI|           SO2 AQI|            CO AQI|
+-------+-----+------------------+------------------+------------------+------------------+
|Alabama|   01|23.289473684210527|27.236842105263158|4.947368421052632| 4.815789473684211|
|Alabama|   02|            22.075|             32.75|             4.85|              4.15|
|Alabama|   03|21.595238095238095| 34.07142857142857|6.976190476190476|3.6785714285714284|
|Alabama|   04|22.177777777777777|39.233333333333334|4.877777777777778|3.2777777777777777|
|Alabama|   05|23.055555555555557|              50.7|8.577777777777778|3.6666666666666665|
+-------+-----+------------------+------------------+------------------+------------------+
```

Determining AQI date wise and color grading as per quality

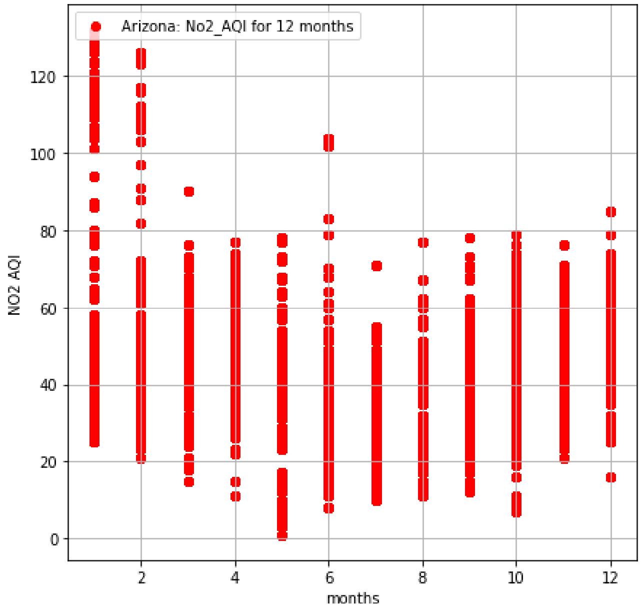
After determining date wise air quality index (AQI) for each ingredient. The highest AQI among the contributors is chosen as the AQI for the place for that date as per norms. It is color coded as per norms for awareness.

```
# Determining AQI datewise and colour grading as per quality

from pyspark.sql.functions import *

dfCoalesce = lines.withColumn("AQI", greatest(coalesce(col("NO2 AQI"),
lit(0)), coalesce(col("O3 AQI"), lit(0)), coalesce(col("SO2 AQI"),
lit(0)), coalesce(col("CO AQI"), lit(0)))) \
.withColumn("AQI Color", when((col("AQI") >= 0) & (col("AQI") <= 50),
"GREEN")
           .when((col("AQI") >= 51) & (col("AQI") <= 100), "YELLOW")
           .when((col("AQI") >= 101) & (col("AQI") <= 150), "ORANGE")
           .when((col("AQI") >= 151) & (col("AQI") <= 200), "RED")
           .when((col("AQI") >= 201) & (col("AQI") <= 300), "PURPLE")
           .otherwise(lit("MAROON")))\
.select("State", "Date Local", col("NO2 AQI"), col("O3 AQI"), col("SO2
AQI"),  col("CO AQI"), "AQI", "AQI Color")\
.where(col("State") == "Arizona") \
.show(50)
```

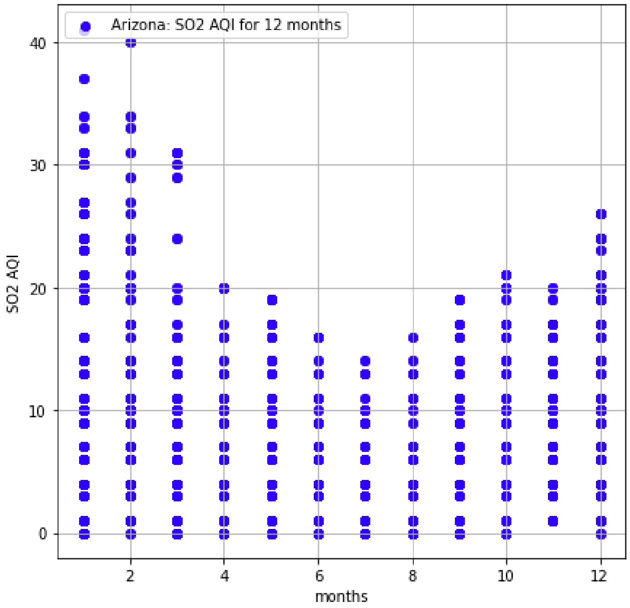
```
+-------+----------+-------+------+-------+------+-----+---------+
|  State|Date Local|NO2 AQI|O3 AQI|SO2 AQI|CO AQI|  AQI|AQI Color|
+-------+----------+-------+------+-------+------+-----+---------+
|Arizona|2000-01-01|     46|    34|   13.0|  null| 46.0|    GREEN|
|Arizona|2000-01-01|     46|    34|   13.0|  25.0| 46.0|    GREEN|
|Arizona|2000-01-01|     46|    34|   null|  null| 46.0|    GREEN|
|Arizona|2000-01-01|     46|    34|   null|  25.0| 46.0|    GREEN|
|Arizona|2000-01-02|     34|    27|    4.0|  null| 34.0|    GREEN|
|Arizona|2000-01-02|     34|    27|    4.0|  26.0| 34.0|    GREEN|
|Arizona|2000-01-02|     34|    27|   null|  null| 34.0|    GREEN|
|Arizona|2000-01-02|     34|    27|   null|  26.0| 34.0|    GREEN|
|Arizona|2000-01-03|     48|    14|   16.0|  null| 48.0|    GREEN|
|Arizona|2000-01-03|     48|    14|   16.0|  28.0| 48.0|    GREEN|
|Arizona|2000-01-03|     48|    14|   null|  null| 48.0|    GREEN|
|Arizona|2000-01-03|     48|    14|   null|  28.0| 48.0|    GREEN|
|Arizona|2000-01-04|     72|    28|   23.0|  null| 72.0|   YELLOW|
|Arizona|2000-01-04|     72|    28|   23.0|  34.0| 72.0|   YELLOW|
|Arizona|2000-01-04|     72|    28|   null|  null| 72.0|   YELLOW|
|Arizona|2000-01-04|     72|    28|   null|  34.0| 72.0|   YELLOW|
|Arizona|2000-01-05|     58|    10|   21.0|  null| 58.0|   YELLOW|
|Arizona|2000-01-05|     58|    10|   21.0|  42.0| 58.0|   YELLOW|
|Arizona|2000-01-05|     58|    10|   null|  null| 58.0|   YELLOW|
|Arizona|2000-01-05|     58|    10|   null|  42.0| 58.0|   YELLOW|
|Arizona|2000-01-06|     71|    21|   24.0|  null| 71.0|   YELLOW|
|Arizona|2000-01-06|     71|    21|   24.0|  41.0| 71.0|   YELLOW|
|Arizona|2000-01-06|     71|    21|   null|  null| 71.0|   YELLOW|
|Arizona|2000-01-06|     71|    21|   null|  41.0| 71.0|   YELLOW|
|Arizona|2000-01-07|     41|    20|   30.0|  null| 41.0|    GREEN|
|Arizona|2000-01-07|     41|    20|   30.0|  40.0| 41.0|    GREEN|
|Arizona|2000-01-07|     41|    20|   null|  null| 41.0|    GREEN|
|Arizona|2000-01-07|     41|    20|   null|  40.0| 41.0|    GREEN|
|Arizona|2000-01-08|     39|    17|   26.0|  null| 39.0|    GREEN|
|Arizona|2000-01-08|     39|    17|   26.0|  57.0| 57.0|   YELLOW|
|Arizona|2000-01-08|     39|    17|   null|  null| 39.0|    GREEN|
|Arizona|2000-01-08|     39|    17|   null|  57.0| 57.0|   YELLOW|
|Arizona|2000-01-09|     35|    19|   19.0|  null| 35.0|    GREEN|
|Arizona|2000-01-09|     35|    19|   19.0|  32.0| 35.0|    GREEN|
|Arizona|2000-01-09|     35|    19|   null|  null| 35.0|    GREEN|
|Arizona|2000-01-09|     35|    19|   null|  32.0| 35.0|    GREEN|
|Arizona|2000-01-10|     68|    13|   30.0|  null| 68.0|   YELLOW|
|Arizona|2000-01-10|     68|    13|   30.0|  42.0| 68.0|   YELLOW|
|Arizona|2000-01-10|     68|    13|   null|  null| 68.0|   YELLOW|
|Arizona|2000-01-10|     68|    13|   null|  42.0| 68.0|   YELLOW|
|Arizona|2000-01-11|     80|    14|   34.0|  null| 80.0|   YELLOW|
|Arizona|2000-01-11|     80|    14|   34.0|  51.0| 80.0|   YELLOW|
|Arizona|2000-01-11|     80|    14|   null|  null| 80.0|   YELLOW|
|Arizona|2000-01-11|     80|    14|   null|  51.0| 80.0|   YELLOW|
|Arizona|2000-01-12|     80|    12|   37.0|  null| 80.0|   YELLOW|
|Arizona|2000-01-12|     80|    12|   37.0|  48.0| 80.0|   YELLOW|
|Arizona|2000-01-12|     80|    12|   null|  null| 80.0|   YELLOW|
|Arizona|2000-01-12|     80|    12|   null|  48.0| 80.0|   YELLOW|
|Arizona|2000-01-13|    104|     8|   30.0|  null|104.0|   ORANGE|
|Arizona|2000-01-13|    104|     8|   30.0|  52.0|104.0|   ORANGE|
+-------+----------+-------+------+-------+------+-----+---------+
only showing top 50 rows
```

Plotting month wise AQI

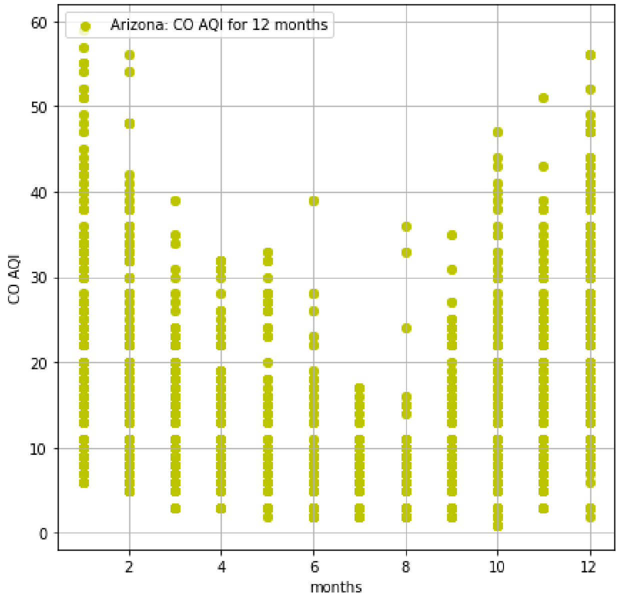
For these plots, data for the state of Arizona is taken. We found **an interesting aspect** from the plots. Since the data is from the year 2000 to 2016, 16 years data is included in the plot. The interesting aspect is that from the month of March to mid-september the pollution level due to NO2, SO2 and CO generally remained lower as compared to other months of the year, the highest is observed for the month of December and January, whereas pollution due to O3 behaves exactly in an opposite cycle.



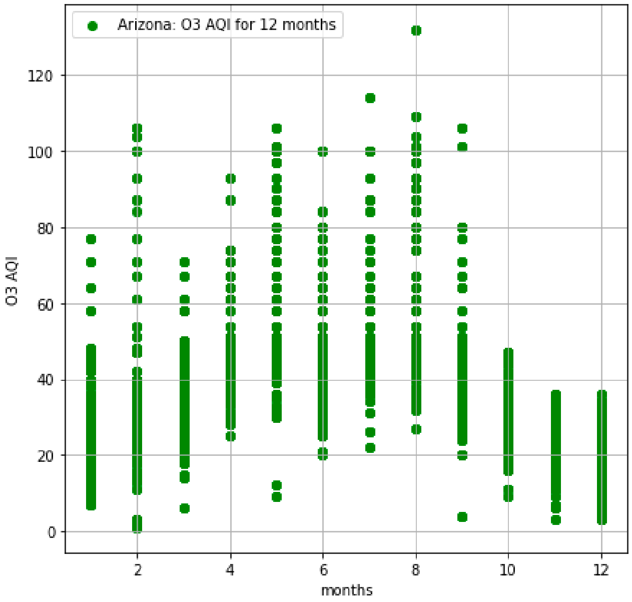
NO2 AQI



SO2 AQI



CO AQI



O3 AQI

Following through some articles from the internet some reason for the interesting observation were explained

Less pollution in summer

In summer, the air in the planetary boundary layer (the lowest part of the atmosphere) is warmer and lighter and rises upwards more easily. This carries pollutants away from the ground and mixes them with cleaner air in the upper layers of the atmosphere in a process called '**vertical mixing**'.(weather.com)

More pollution in winter

During winters the planetary boundary layer is thinner as the cooler air near the earth's surface is dense. The cooler air is trapped under the warm air above that forms a kind of atmospheric 'lid'. This phenomenon is called **winter inversion**. Since the vertical mixing of air happens only within this layer, the pollutants released lack enough space to disperse in the atmosphere.(weather.com)

More Ozone pollution during summer

There's always some ground level ozone floating around. With that said, the levels of this dangerous gas rise significantly as the weather warms. The reason for this is that ground level ozone is produced by chemical reactions powered by sunlight. Compounds found in vehicle and industrial air pollution, when exposed to sunlight and hot temperatures, can react to form ozone. The combination of more direct sunlight and longer daylight hours creates the summer ozone season. Since more people travel in the summer, there is also more vehicle exhaust in the air. As a result, the compounds that react to form ozone become more available.(usairpurifiers)

The Plot code:

```
import matplotlib.pyplot as plt
import matplotlib

matplotlib.rcParams['figure.figsize'] = [7, 7] # for square canvas
x_month=df1_Arizona['month'].to_list()
y_NO2=df1_Arizona['NO2 AQI'].to_list()

plt.scatter(x_month,y_NO2, color='r',label='NO2 AQI for 12 months')
plt.legend(loc=2)
plt.xlabel("months")
plt.ylabel("NO2 AQI")
plt.grid()
```

A Machine learning model for predicting AQI

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

dfP = df_New
# Preparing the data for training the model
X = pd.DataFrame(np.c_[dfP['AQI']], columns = ['AQI'])
Y = dfP['month']

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2,
random_state=1)
print(X_train.shape)
print(X_test.shape)
print(Y_train.shape)
print(Y_test.shape)
```

```
(800000, 1)
(200000, 1)
(800000,)
(200000,)
```

```
lin_model = LinearRegression()
lin_model.fit(X_train, Y_train)
```

```
LinearRegression()
```

```
# Model evaluation
# model evaluation for training set
y_train_predict = lin_model.predict(X_train)
max_relative_error = 100*np.amax(np.absolute((Y_train-y_train_predict)/Y_t
r2 = r2_score(Y_train, y_train_predict)

print("The model performance for training set")
print("--------------------------------------")
print('max_relative_error is {}'.format(max_relative_error))
print('R2 score is {}'.format(r2))
print("\n")
```

```
# model evaluation for testing set
y_test_predict = lin_model.predict(X_test)
max_relative_error = 100*np.amax(np.absolute((Y_train-y_train_predict)/Y_t
r2 = r2_score(Y_test, y_test_predict)

print("The model performance for testing set")
print("--------------------------------------")
print('max_relative_error is {}'.format(max_relative_error))
print('R2 score is {}'.format(r2))
```

```
The model performance for training set
--------------------------------------
max_relative_error is 577.3099801829809
R2 score is 0.0008093409521113815


The model performance for testing set
--------------------------------------
max_relative_error is 577.3099801829809
R2 score is 0.0007799770359860903
```

SUMMARY

To avoid health hazards due to pollution we should use air quality index(AQI) predictions before going outdoors or making any holiday trip planning.

Regarding ozone exposure since most ground level ozone is found outdoors, there are limits to how much you can do to protect yourself. One very useful step you can take, though, is to limit your time outdoors on especially clear days. If there’s a bit of cloud cover, the ozone levels won’t be quite as high due to limited sunlight. Ozone also tends to be a bit lower in the evenings, though this can change if the night is especially hot or very still.

REFERENCE:

- https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf
- https://www.airnow.gov/aqi/aqi-basics/
- https://www.weforum.org/agenda/2016/08/air-pollution-deaths-oecd/
- https://www.oecd.org/environment/the-economic-consequences-of-outdoor-air-pollution-9789264257474-en.htm
- https://weather.com/en-IN/india/science/news/2018-10-30-why-do-pollution-levels-skyrocket-during-winter
- https://www.usairpurifiers.com/blog/summer-ozone-season-why-is-ozone-worse-in-hot-weather/