PROJECT

AIR QUALITY FORECAST IN USA

Problem Statement

- Air pollution is one of the most serious problems in the world. It refers to the contamination of the atmosphere by harmful chemicals or biological materials.
- Air pollution can cause long-term and short-term health effects. It's found that the elderly and young children are more affected by air pollution. Short-term health effects include eye, nose, and throat irritation, headaches, allergic reactions, and upper respiratory infections. Some long-term health effects are lung cancer, brain damage, liver damage, kidney damage, heart disease, and respiratory disease.
- This project is about the Exploratory Data Analysis of the Air Quality across states in USA using Pyspark. From the year 2000 through 2021, this dataset contains daily statistics on four important gas pollutants: carbon monoxide, nitrogen dioxide, ground-level ozone, and sulfur dioxide. This project predicts the most hazardous gas O3 AQI value using pyspark ML.

Keywords in the Data set

- **Ozone molecule** (O3) is harmful to air quality outside of the ozone layer.
- **Carbon Monoxide** (CO) is a colorless, odorless gas that can be harmful when inhaled in large amounts.
- **Sulfur dioxide** (SO2) is a colorless, reactive air pollutant with a strong odor. This gas can be a threat to human health, animal health, and plant life.
- **Nitrogen dioxide** (NO2) is a gaseous air pollutant composed of nitrogen and oxygen and is one of a group of related gases called nitrogen oxides.
- **Air Quality Index** (AQI)

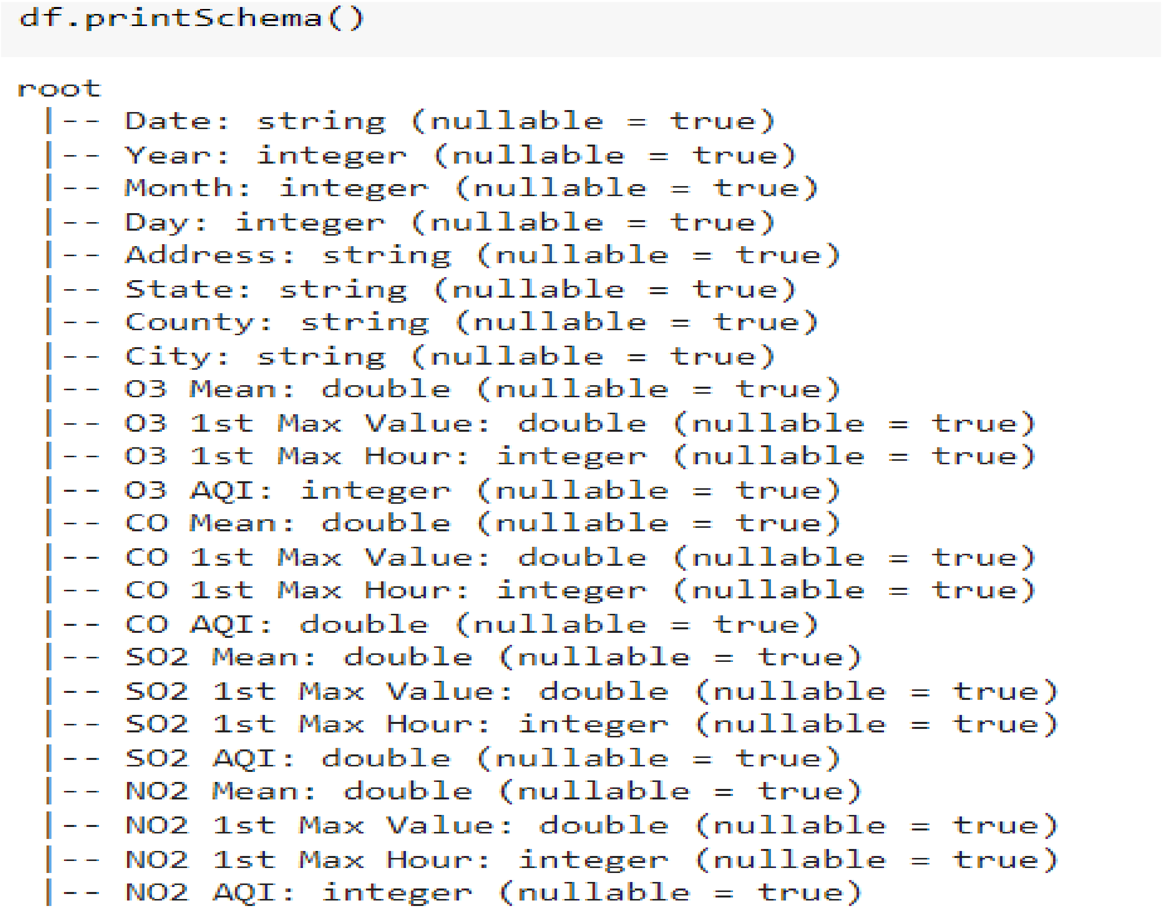
DATA SOURCE

- https://www.kaggle.com/alpacanonymous/us-pollution-20002021/download
- https://aqs.epa.gov/aqsweb/airdata/download_files.html

DATASET DETAILS:

- Number of rows: 608700
- Number of columns: 24
- Size of the dataset: 97.76 MB

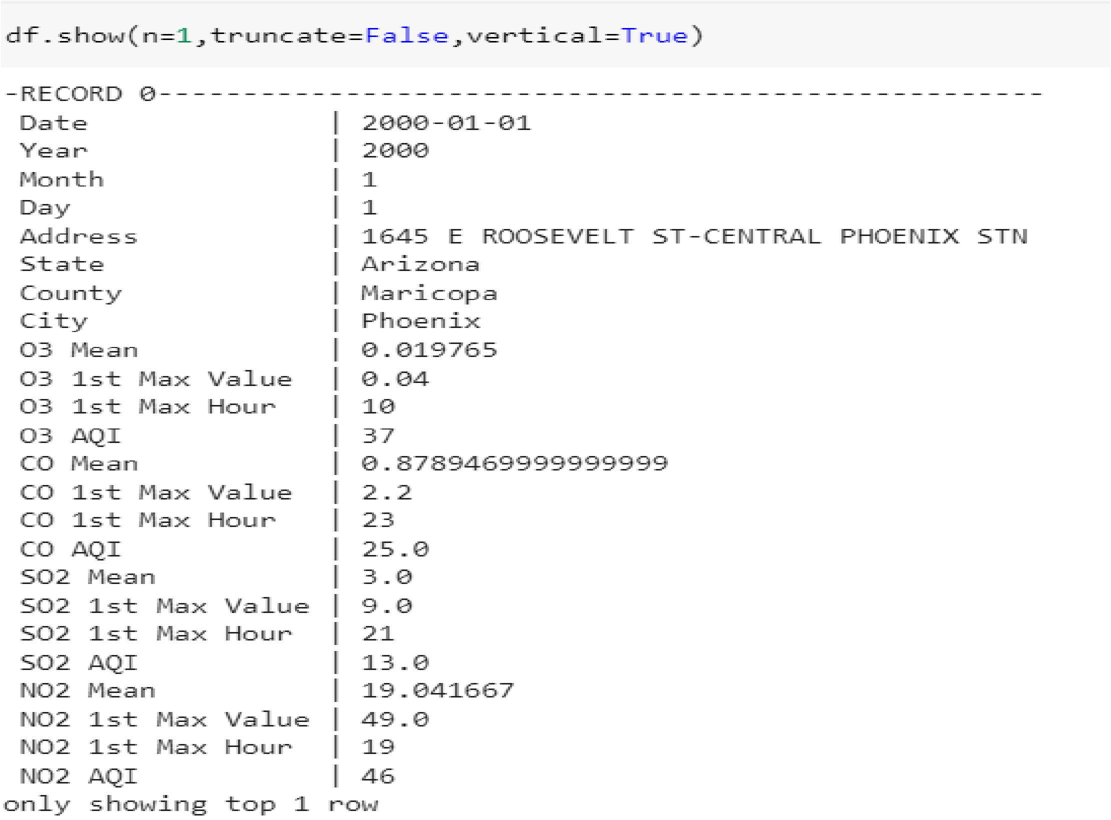
Column Name and its Data type



```
df.printSchema()
```

```
root
 |-- Date: string (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Month: integer (nullable = true)
 |-- Day: integer (nullable = true)
 |-- Address: string (nullable = true)
 |-- State: string (nullable = true)
 |-- County: string (nullable = true)
 |-- City: string (nullable = true)
 |-- O3 Mean: double (nullable = true)
 |-- O3 1st Max Value: double (nullable = true)
 |-- O3 1st Max Hour: integer (nullable = true)
 |-- O3 AQI: integer (nullable = true)
 |-- CO Mean: double (nullable = true)
 |-- CO 1st Max Value: double (nullable = true)
 |-- CO 1st Max Hour: integer (nullable = true)
 |-- CO AQI: double (nullable = true)
 |-- SO2 Mean: double (nullable = true)
 |-- SO2 1st Max Value: double (nullable = true)
 |-- SO2 1st Max Hour: integer (nullable = true)
 |-- SO2 AQI: double (nullable = true)
 |-- NO2 Mean: double (nullable = true)
 |-- NO2 1st Max Value: double (nullable = true)
 |-- NO2 1st Max Hour: integer (nullable = true)
 |-- NO2 AQI: integer (nullable = true)
```

Displaying top row data

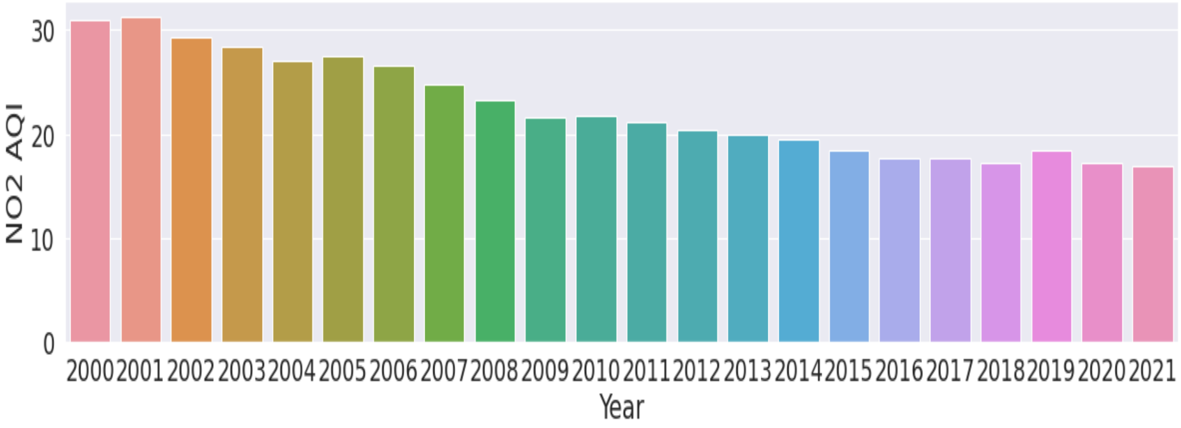


```
df.show(n=1,truncate=False,vertical=True)
```

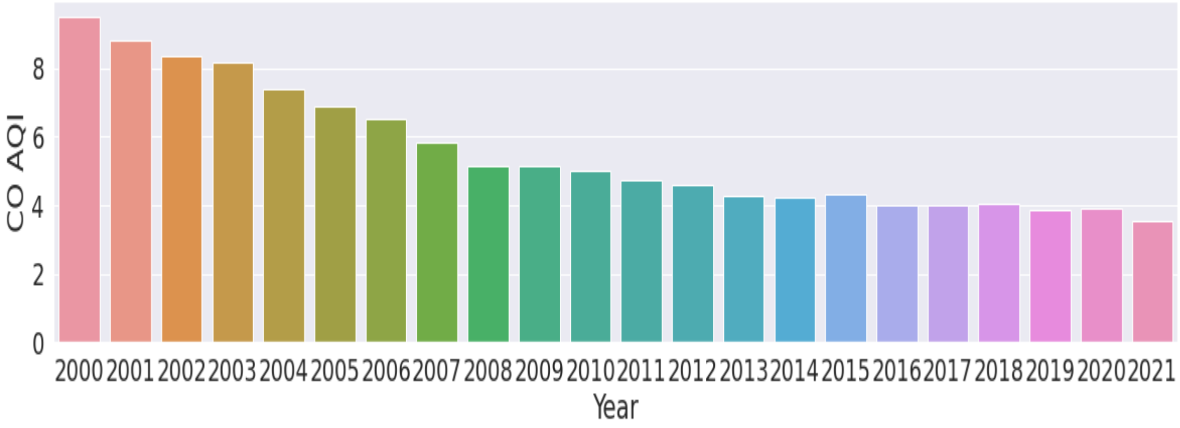
```
-RECORD 0---------------------------------------------------
 Date              | 2000-01-01
 Year              | 2000
 Month             | 1
 Day               | 1
 Address           | 1645 E ROOSEVELT ST-CENTRAL PHOENIX STN
 State             | Arizona
 County            | Maricopa
 City              | Phoenix
 O3 Mean           | 0.019765
 O3 1st Max Value  | 0.04
 O3 1st Max Hour   | 10
 O3 AQI            | 37
 CO Mean           | 0.8789469999999999
 CO 1st Max Value  | 2.2
 CO 1st Max Hour   | 23
 CO AQI            | 25.0
 SO2 Mean          | 3.0
 SO2 1st Max Value | 9.0
 SO2 1st Max Hour  | 21
 SO2 AQI           | 13.0
 NO2 Mean          | 19.041667
 NO2 1st Max Value | 49.0
 NO2 1st Max Hour  | 19
 NO2 AQI           | 46
only showing top 1 row
```

Year wise AQI(Air Quality Index) for NO_2, CO, O_3, SO_2

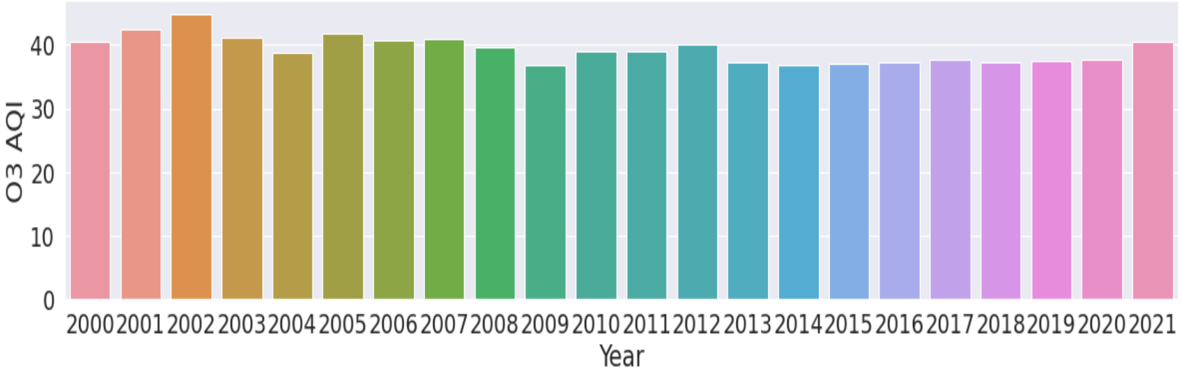
NO2 AQI Over the years



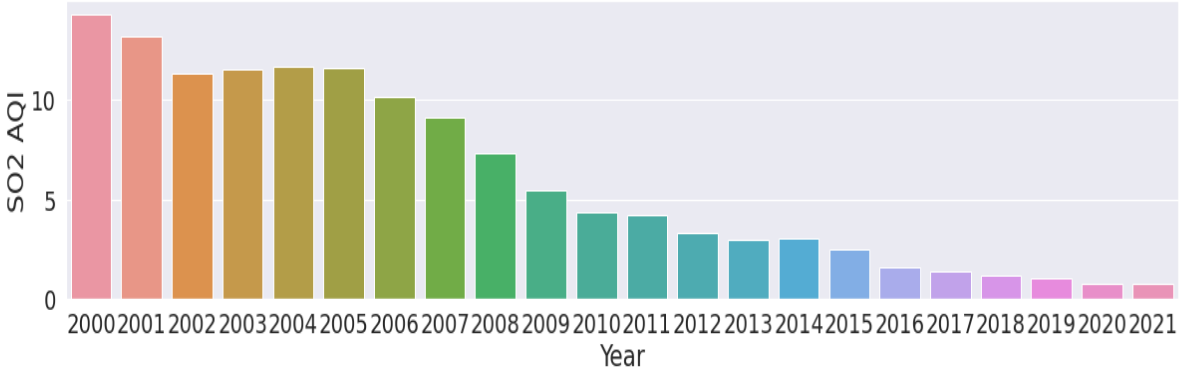
CO AQI Over the years



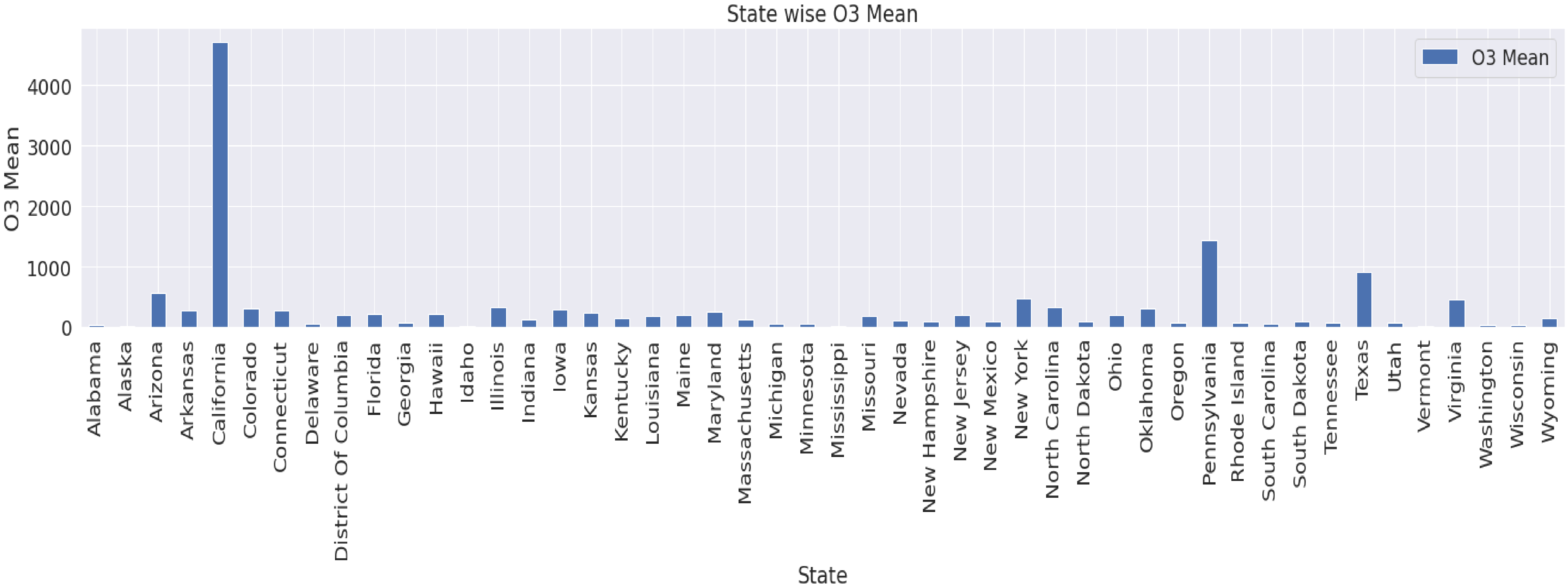
O3 AQI Over the years



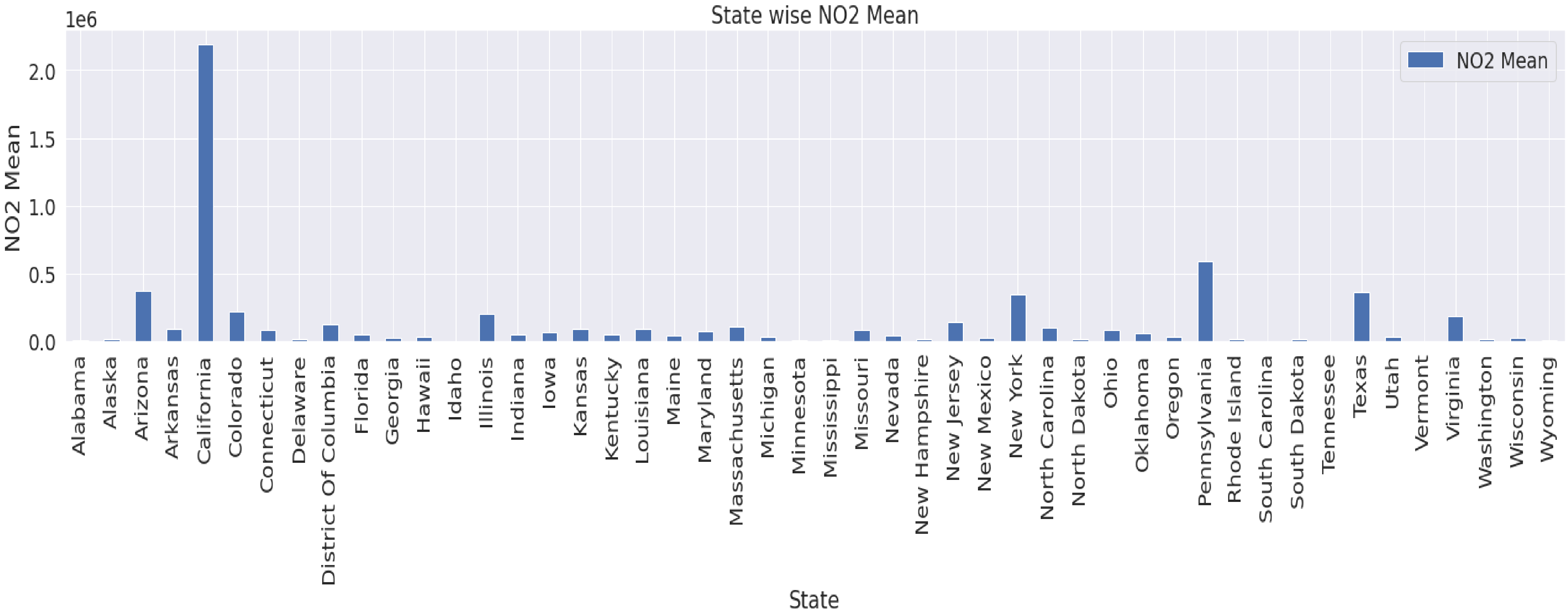
SO2 AQI Over the years



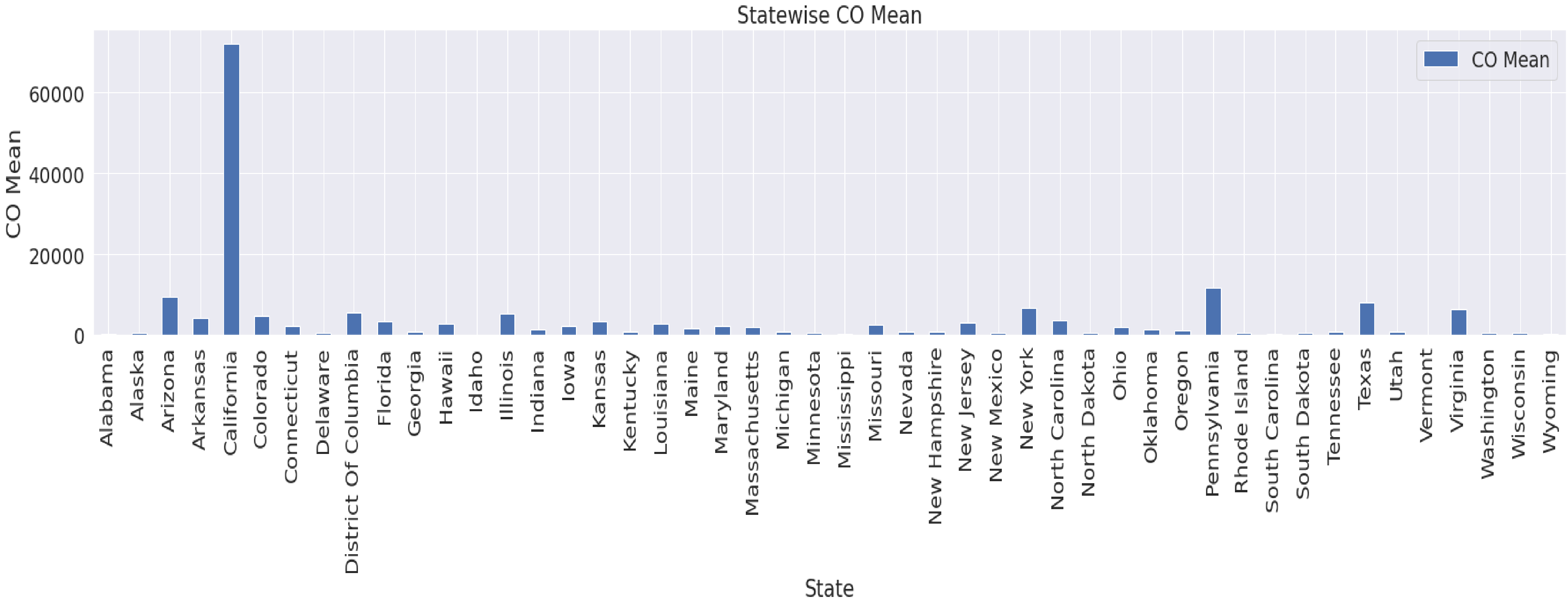
State wise Ozone molecule mean



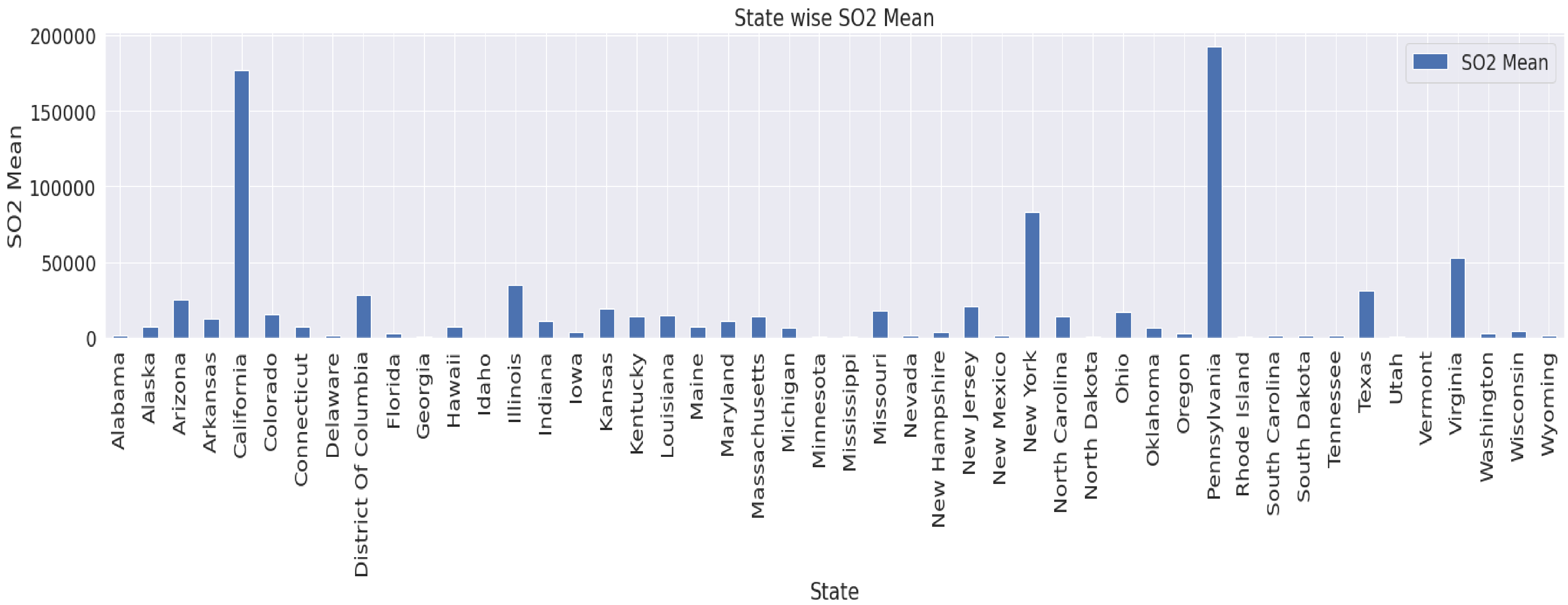
State wise Nitrogen dioxide molecule mean

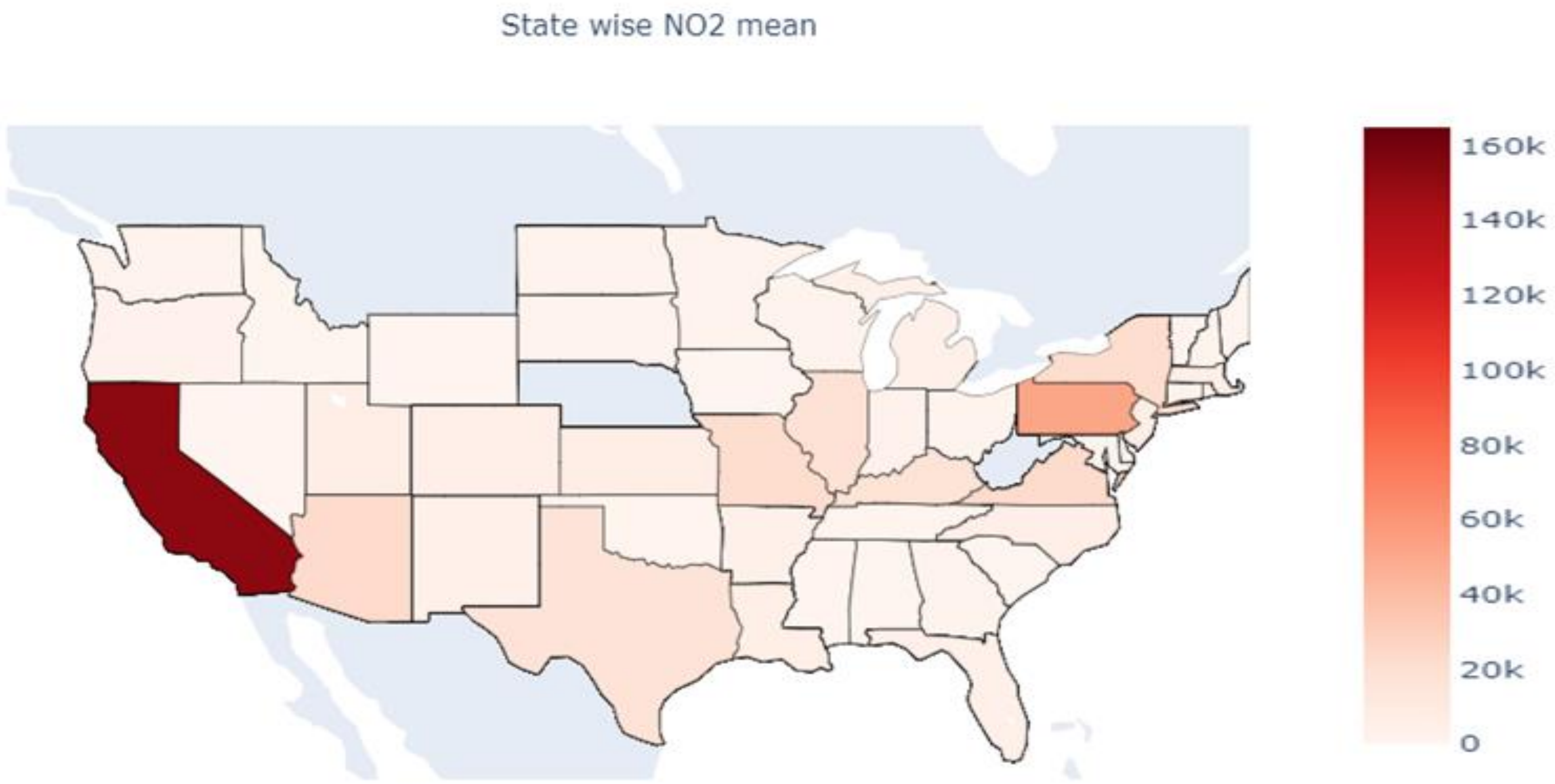


State wise Carbon Monoxide molecule mean

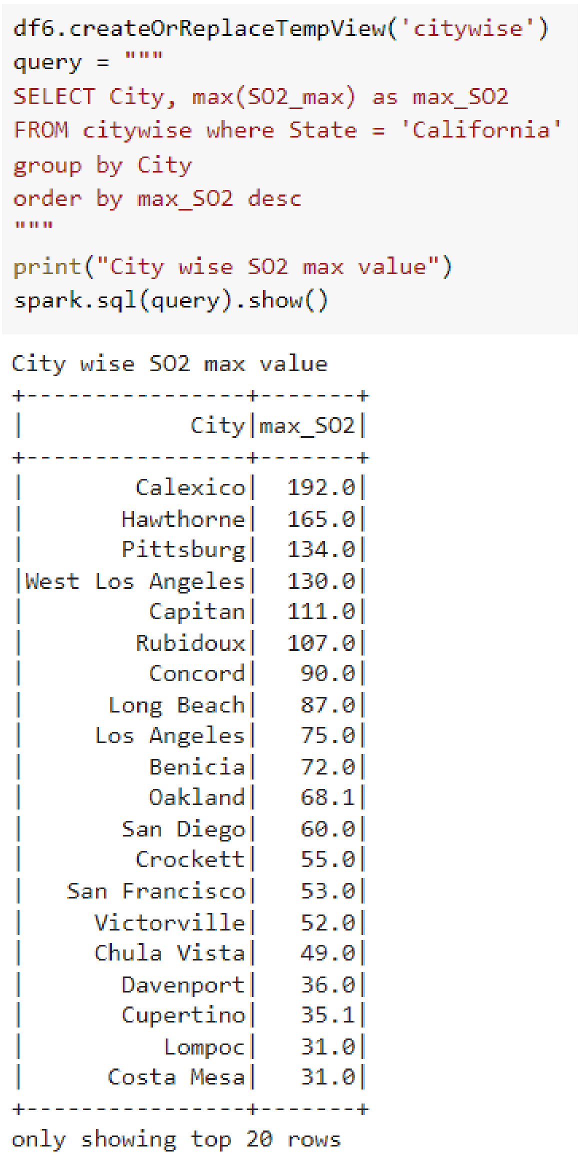


State wise sulfur dioxide molecule mean



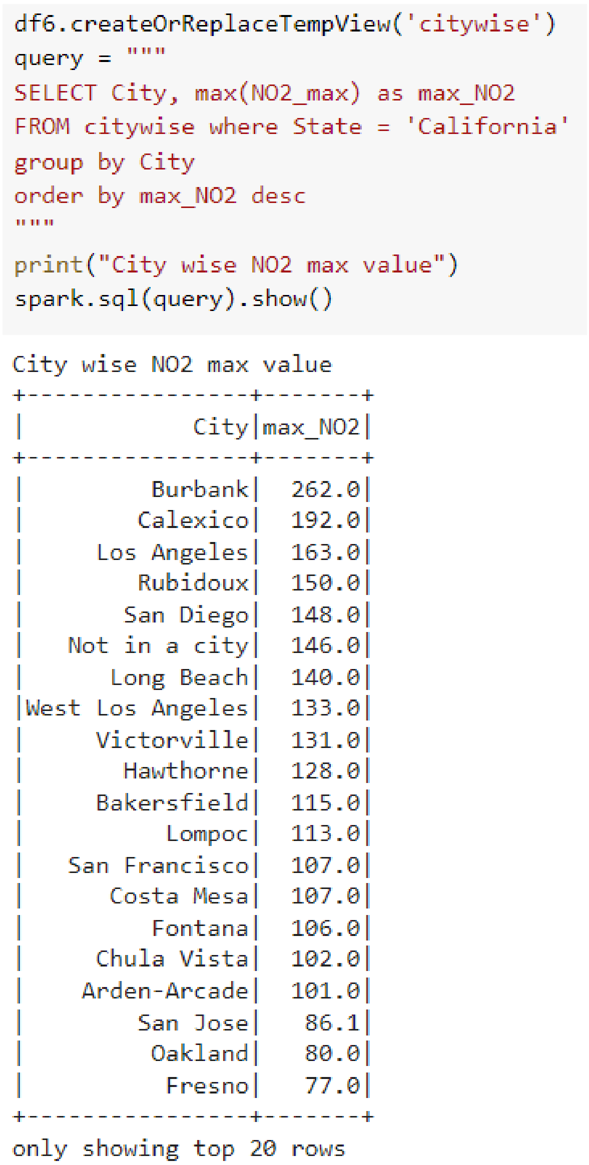


According to the choropleth map above, California has the greatest levels of air pollution.



```
df6.createOrReplaceTempView('citywise')
query = """
SELECT City, max(SO2_max) as max_SO2
FROM citywise where State = 'California'
group by City
order by max_SO2 desc
"""
print("City wise SO2 max value")
spark.sql(query).show()
```

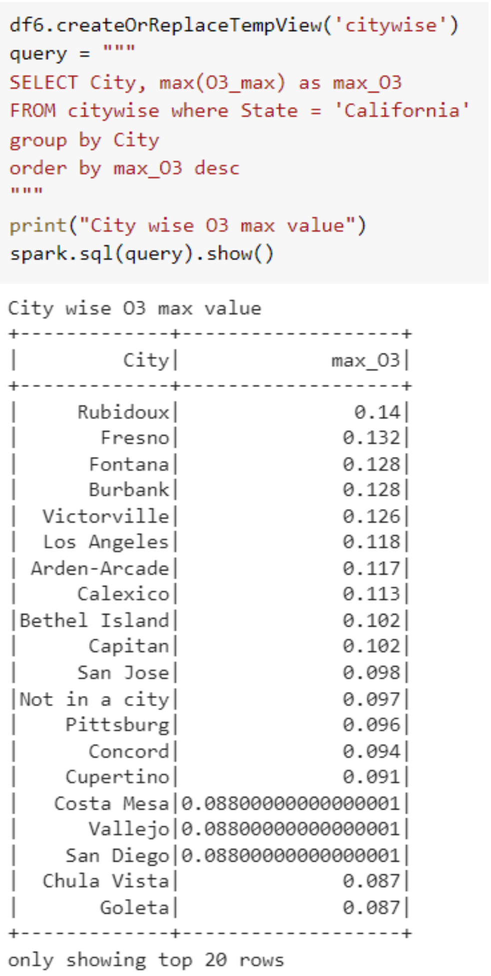
```
City wise SO2 max value
+----------------+-------+
|            City|max_SO2|
+----------------+-------+
|        Calexico|  192.0|
|       Hawthorne|  165.0|
|       Pittsburg|  134.0|
|West Los Angeles|  130.0|
|         Capitan|  111.0|
|        Rubidoux|  107.0|
|         Concord|   90.0|
|      Long Beach|   87.0|
|     Los Angeles|   75.0|
|         Benicia|   72.0|
|         Oakland|   68.1|
|       San Diego|   60.0|
|        Crockett|   55.0|
|   San Francisco|   53.0|
|     Victorville|   52.0|
|     Chula Vista|   49.0|
|       Davenport|   36.0|
|       Cupertino|   35.1|
|          Lompoc|   31.0|
|      Costa Mesa|   31.0|
+----------------+-------+
only showing top 20 rows
```



```
df6.createOrReplaceTempView('citywise')
query = """
SELECT City, max(NO2_max) as max_NO2
FROM citywise where State = 'California'
group by City
order by max_NO2 desc
"""
print("City wise NO2 max value")
spark.sql(query).show()
```

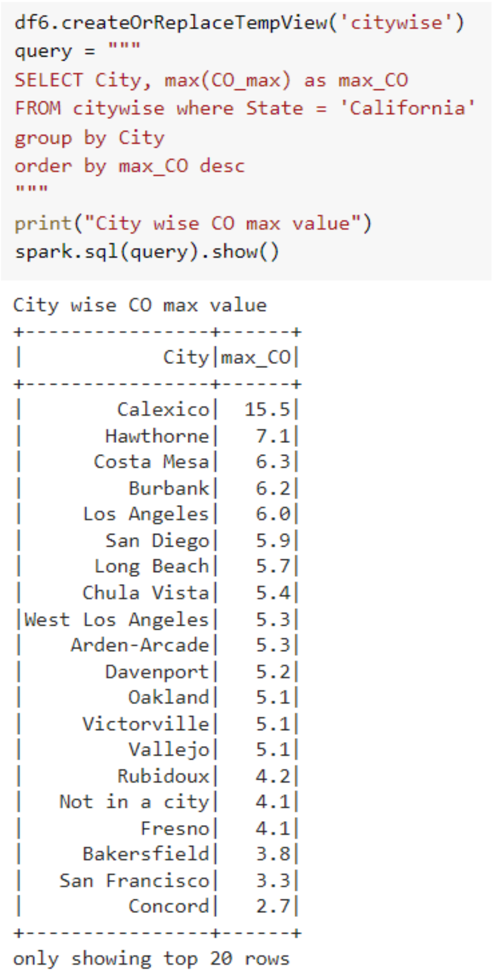
```
City wise NO2 max value
+----------------+-------+
|            City|max_NO2|
+----------------+-------+
|         Burbank|  262.0|
|        Calexico|  192.0|
|     Los Angeles|  163.0|
|        Rubidoux|  150.0|
|       San Diego|  148.0|
|     Not in a city|  146.0|
|      Long Beach|  140.0|
|West Los Angeles|  133.0|
|     Victorville|  131.0|
|       Hawthorne|  128.0|
|     Bakersfield|  115.0|
|          Lompoc|  113.0|
|   San Francisco|  107.0|
|      Costa Mesa|  107.0|
|         Fontana|  106.0|
|     Chula Vista|  102.0|
|    Arden-Arcade|  101.0|
|        San Jose|   86.1|
|         Oakland|   80.0|
|          Fresno|   77.0|
+----------------+-------+
only showing top 20 rows
```

It can be inferred that Burbank City has Max NO2 value and Calexico City has Max SO2 value in California State



```
df6.createOrReplaceTempView('citywise')
query = """
SELECT City, max(O3_max) as max_O3
FROM citywise where State = 'California'
group by City
order by max_O3 desc
"""
print("City wise O3 max value")
spark.sql(query).show()
```

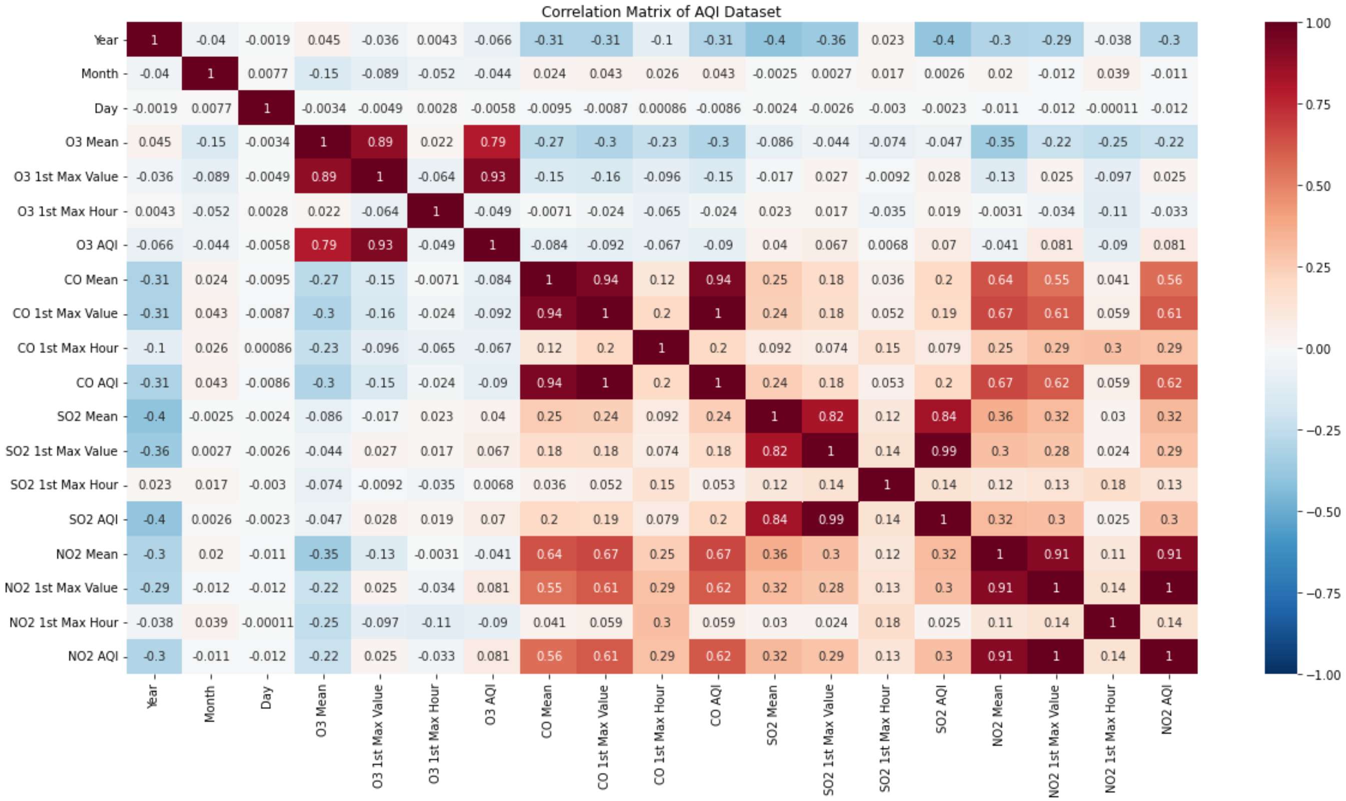
```
City wise O3 max value
+-------------+------------------+
|         City|            max_O3|
+-------------+------------------+
|     Rubidoux|              0.14|
|       Fresno|             0.132|
|      Fontana|             0.128|
|      Burbank|             0.128|
|  Victorville|             0.126|
|  Los Angeles|             0.118|
| Arden-Arcade|             0.117|
|     Calexico|             0.113|
|Bethel Island|             0.102|
|      Capitan|             0.102|
|     San Jose|             0.098|
|Not in a city|             0.097|
|    Pittsburg|             0.096|
|      Concord|             0.094|
|    Cupertino|             0.091|
|   Costa Mesa|0.0880000000000001|
|      Vallejo|0.0880000000000001|
|    San Diego|0.0880000000000001|
|  Chula Vista|             0.087|
|       Goleta|             0.087|
+-------------+------------------+
only showing top 20 rows
```



```
df6.createOrReplaceTempView('citywise')
query = """
SELECT City, max(CO_max) as max_CO
FROM citywise where State = 'California'
group by City
order by max_CO desc
"""
print("City wise CO max value")
spark.sql(query).show()
```

```
City wise CO max value
+---------------+------+
|           City|max_CO|
+---------------+------+
|       Calexico|  15.5|
|      Hawthorne|   7.1|
|     Costa Mesa|   6.3|
|        Burbank|   6.2|
|    Los Angeles|   6.0|
|      San Diego|   5.9|
|     Long Beach|   5.7|
|    Chula Vista|   5.4|
|West Los Angeles|   5.3|
|   Arden-Arcade|   5.3|
|      Davenport|   5.2|
|        Oakland|   5.1|
|    Victorville|   5.1|
|        Vallejo|   5.1|
|       Rubidoux|   4.2|
|  Not in a city|   4.1|
|         Fresno|   4.1|
|    Bakersfield|   3.8|
|  San Francisco|   3.3|
|        Concord|   2.7|
+---------------+------+
only showing top 20 rows
```

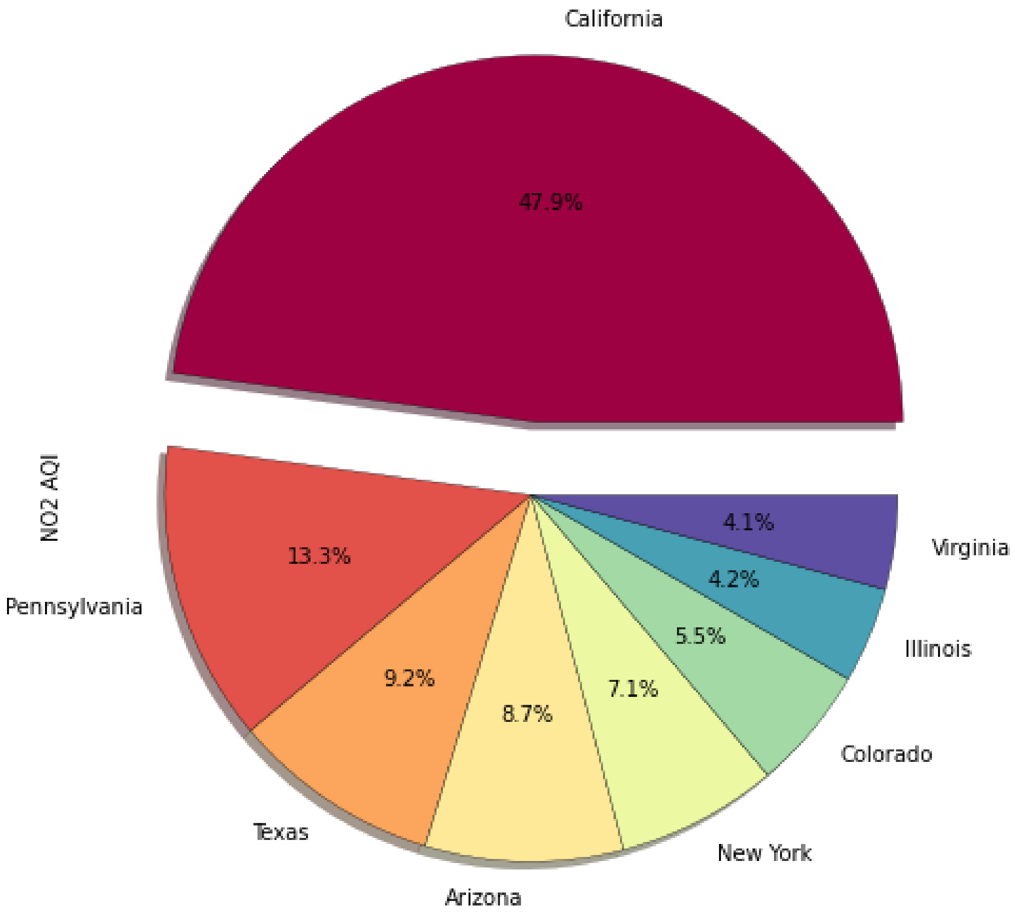
It can be inferred that Rubidoux City has Max O3 value and Calexico City has Max CO value in California State



Correlation Matrix of AQI Dataset

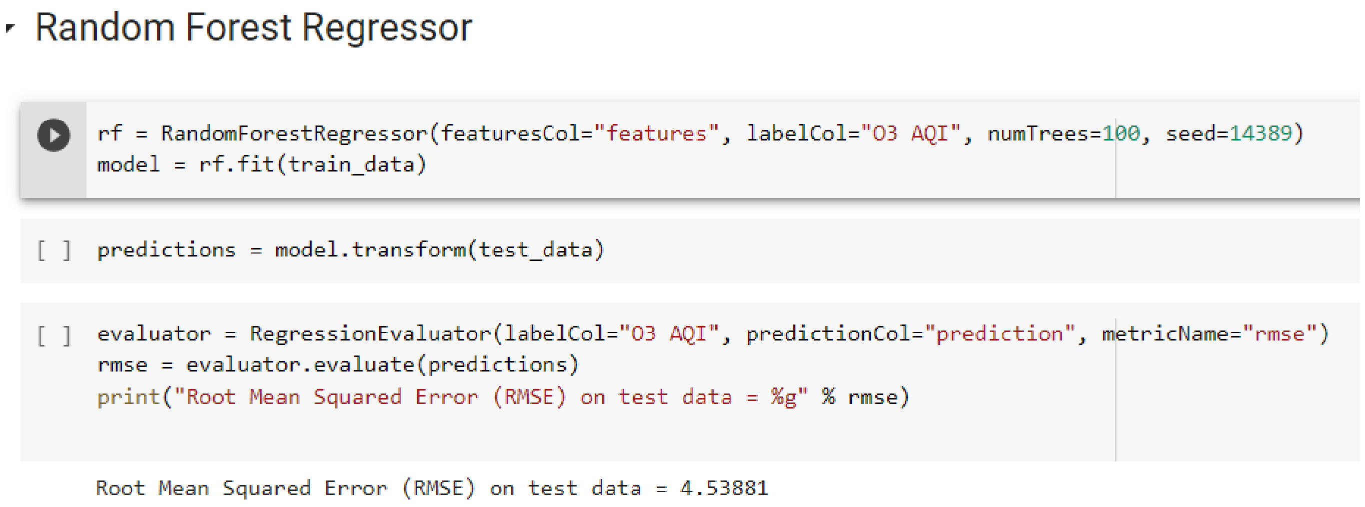
	Year	Month	Day	O3 Mean	O3 1st Max Value	O3 1st Max Hour	O3 AQI	CO Mean	CO 1st Max Value	CO 1st Max Hour	CO AQI	SO2 Mean	SO2 1st Max Value	SO2 1st Max Hour	SO2 AQI	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI
Year	1	-0.04	-0.0019	0.045	-0.036	0.0043	-0.066	-0.31	-0.31	-0.1	-0.31	-0.4	-0.36	0.023	-0.4	-0.3	-0.29	-0.038	-0.3
Month	-0.04	1	0.0077	-0.15	-0.089	-0.052	-0.044	0.024	0.043	0.026	0.043	-0.0025	0.0027	0.017	0.0026	0.02	-0.012	0.039	-0.011
Day	-0.0019	0.0077	1	-0.0034	-0.0049	0.0028	-0.0058	-0.0095	-0.0087	0.00086	-0.0086	-0.0024	-0.0026	-0.003	-0.0023	-0.011	-0.012	-0.00011	-0.012
O3 Mean	0.045	-0.15	-0.0034	1	0.89	0.022	0.79	-0.27	-0.3	-0.23	-0.3	-0.086	-0.044	-0.074	-0.047	-0.35	-0.22	-0.25	-0.22
O3 1st Max Value	-0.036	-0.089	-0.0049	0.89	1	-0.064	0.93	-0.15	-0.16	-0.096	-0.15	-0.017	0.027	-0.0092	0.028	-0.13	0.025	-0.097	0.025
O3 1st Max Hour	0.0043	-0.052	0.0028	0.022	-0.064	1	-0.049	-0.0071	-0.024	-0.065	-0.024	0.023	0.017	-0.035	0.019	-0.0031	-0.034	-0.11	-0.033
O3 AQI	-0.066	-0.044	-0.0058	0.79	0.93	-0.049	1	-0.084	-0.092	-0.067	-0.09	0.04	0.067	0.0068	0.07	-0.041	0.081	-0.09	0.081
CO Mean	-0.31	0.024	-0.0095	-0.27	-0.15	-0.0071	-0.084	1	0.94	0.12	0.94	0.25	0.18	0.036	0.2	0.64	0.55	0.041	0.56
CO 1st Max Value	-0.31	0.043	-0.0087	-0.3	-0.16	-0.024	-0.092	0.94	1	0.2	1	0.24	0.18	0.052	0.19	0.67	0.61	0.059	0.61
CO 1st Max Hour	-0.1	0.026	0.00086	-0.23	-0.096	-0.065	-0.067	0.12	0.2	1	0.2	0.092	0.074	0.15	0.079	0.25	0.29	0.3	0.29
CO AQI	-0.31	0.043	-0.0086	-0.3	-0.15	-0.024	-0.09	0.94	1	0.2	1	0.24	0.18	0.053	0.2	0.67	0.62	0.059	0.62
SO2 Mean	-0.4	-0.0025	-0.0024	-0.086	-0.017	0.023	0.04	0.25	0.24	0.092	0.24	1	0.82	0.12	0.84	0.36	0.32	0.03	0.32
SO2 1st Max Value	-0.36	0.0027	-0.0026	-0.044	0.027	0.017	0.067	0.18	0.18	0.074	0.18	0.82	1	0.14	0.99	0.3	0.28	0.024	0.29
SO2 1st Max Hour	0.023	0.017	-0.003	-0.074	-0.0092	-0.035	0.0068	0.036	0.052	0.15	0.053	0.12	0.14	1	0.14	0.12	0.13	0.18	0.13
SO2 AQI	-0.4	0.0026	-0.0023	-0.047	0.028	0.019	0.07	0.2	0.19	0.079	0.2	0.84	0.99	0.14	1	0.32	0.3	0.025	0.3
NO2 Mean	-0.3	0.02	-0.011	-0.35	-0.13	-0.0031	-0.041	0.64	0.67	0.25	0.67	0.36	0.3	0.12	0.32	1	0.91	0.11	0.91
NO2 1st Max Value	-0.29	-0.012	-0.012	-0.22	0.025	-0.034	0.081	0.55	0.61	0.29	0.62	0.32	0.28	0.13	0.3	0.91	1	0.14	1
NO2 1st Max Hour	-0.038	0.039	-0.00011	-0.25	-0.097	-0.11	-0.09	0.041	0.059	0.3	0.059	0.03	0.024	0.18	0.025	0.11	0.14	1	0.14
NO2 AQI	-0.3	-0.011	-0.012	-0.22	0.025	-0.033	0.081	0.56	0.61	0.29	0.62	0.32	0.29	0.13	0.3	0.91	1	0.14	1

Correlation matrix of the AQI dataset



Which state has the highest O3 AQI?

Random Forest Regressor to Predict the O3 AQI value.



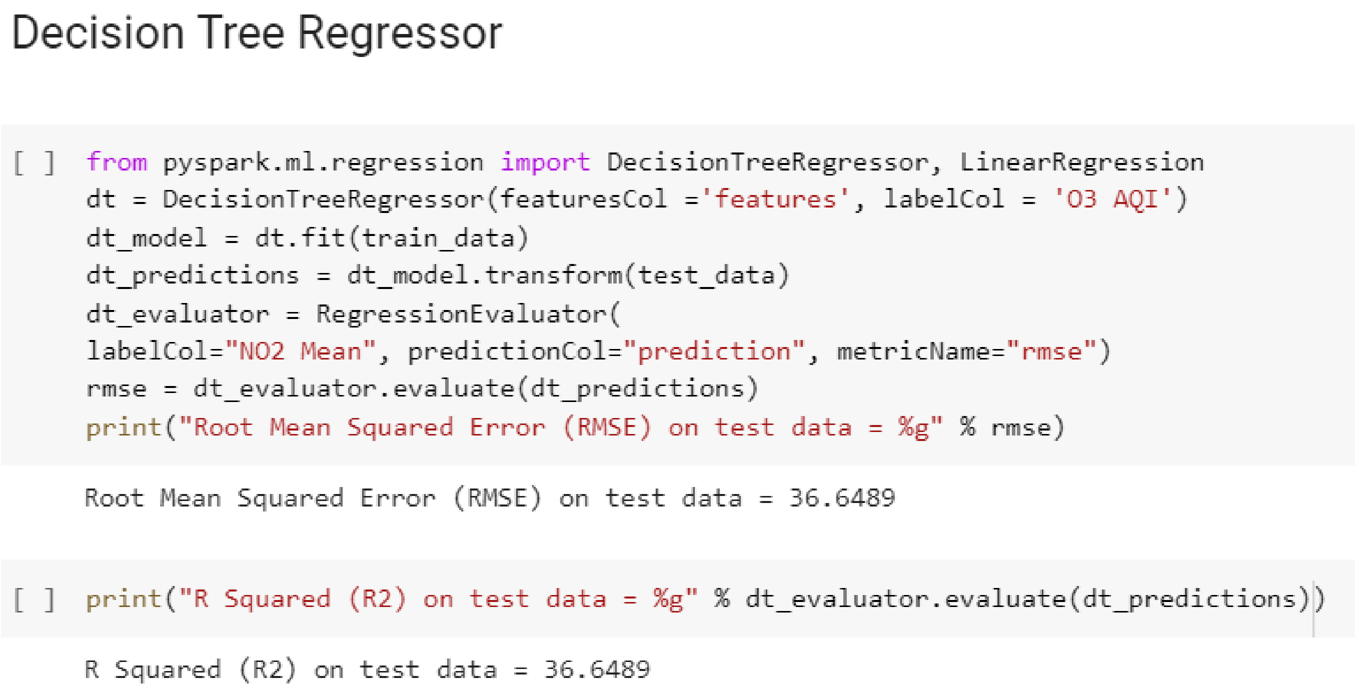
```
rf = RandomForestRegressor(featuresCol="features", labelCol="O3 AQI", numTrees=100, seed=14389)
model = rf.fit(train_data)
```

```
predictions = model.transform(test_data)
```

```
evaluator = RegressionEvaluator(labelCol="O3 AQI", predictionCol="prediction", metricName="rmse")
rmse = evaluator.evaluate(predictions)
print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)
```

```
Root Mean Squared Error (RMSE) on test data = 4.53881
```

Decision Tree Regressor



Decision Tree Regressor

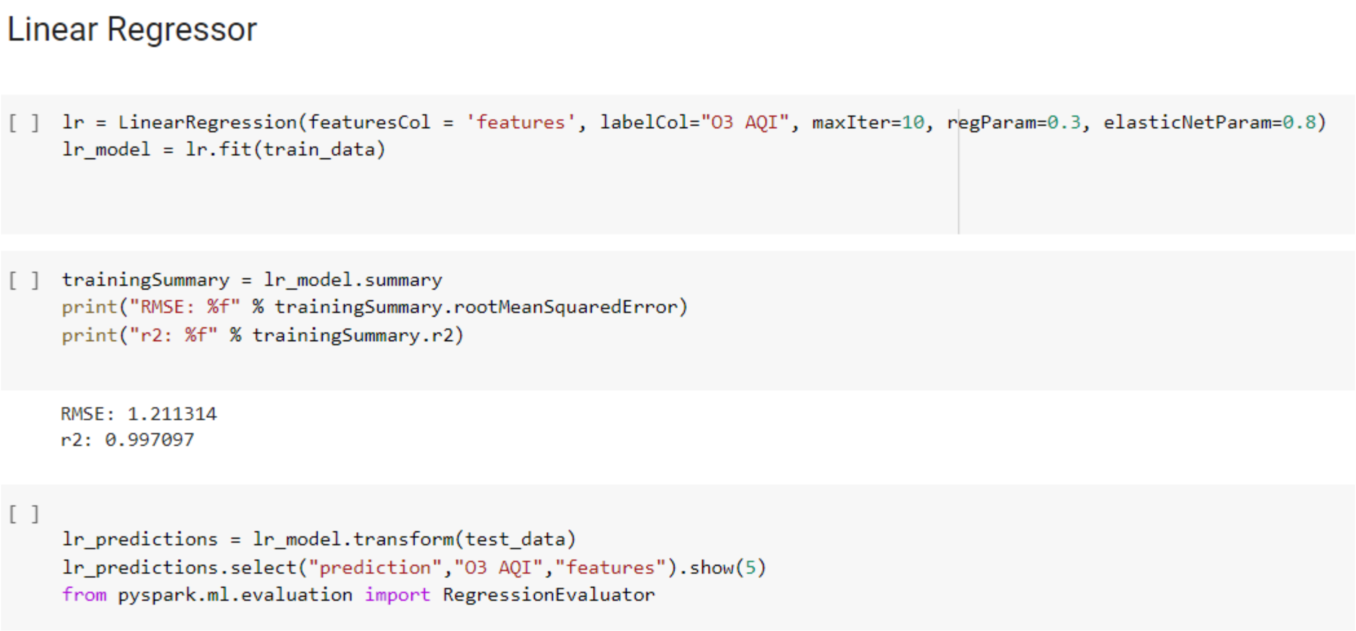
```
[ ]  from pyspark.ml.regression import DecisionTreeRegressor, LinearRegression
     dt = DecisionTreeRegressor(featuresCol ='features', labelCol = 'O3 AQI')
     dt_model = dt.fit(train_data)
     dt_predictions = dt_model.transform(test_data)
     dt_evaluator = RegressionEvaluator(
     labelCol="NO2 Mean", predictionCol="prediction", metricName="rmse")
     rmse = dt_evaluator.evaluate(dt_predictions)
     print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)

     Root Mean Squared Error (RMSE) on test data = 36.6489
```

```
[ ]  print("R Squared (R2) on test data = %g" % dt_evaluator.evaluate(dt_predictions))

     R Squared (R2) on test data = 36.6489
```

Linear Regressor



Linear Regressor

```
[ ]  lr = LinearRegression(featuresCol = 'features', labelCol="O3 AQI", maxIter=10, regParam=0.3, elasticNetParam=0.8)
     lr_model = lr.fit(train_data)
```

```
[ ]  trainingSummary = lr_model.summary
     print("RMSE: %f" % trainingSummary.rootMeanSquaredError)
     print("r2: %f" % trainingSummary.r2)
```

```
     RMSE: 1.211314
     r2: 0.997097
```

```
[ ]
     lr_predictions = lr_model.transform(test_data)
     lr_predictions.select("prediction","O3 AQI","features").show(5)
     from pyspark.ml.evaluation import RegressionEvaluator
```

THANK YOU