

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

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)
```

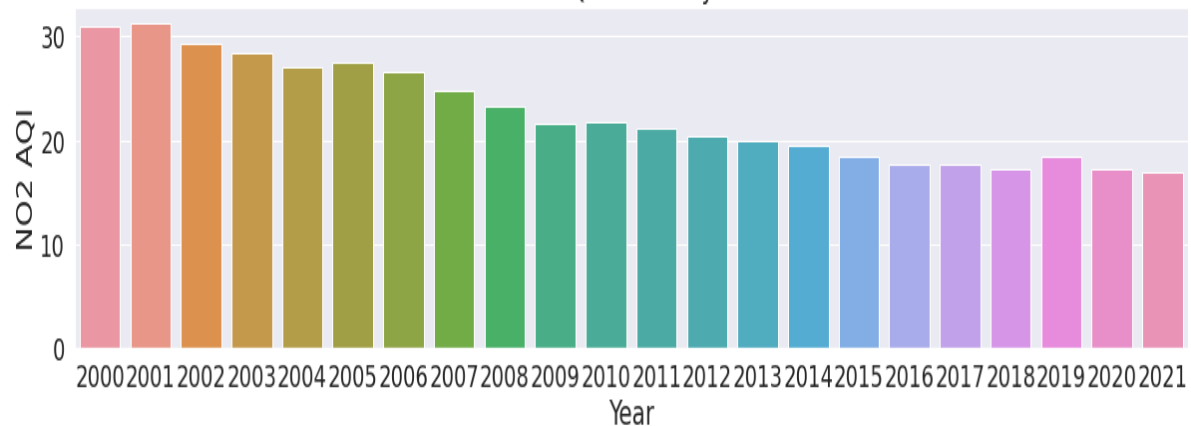
Displaying top row data

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

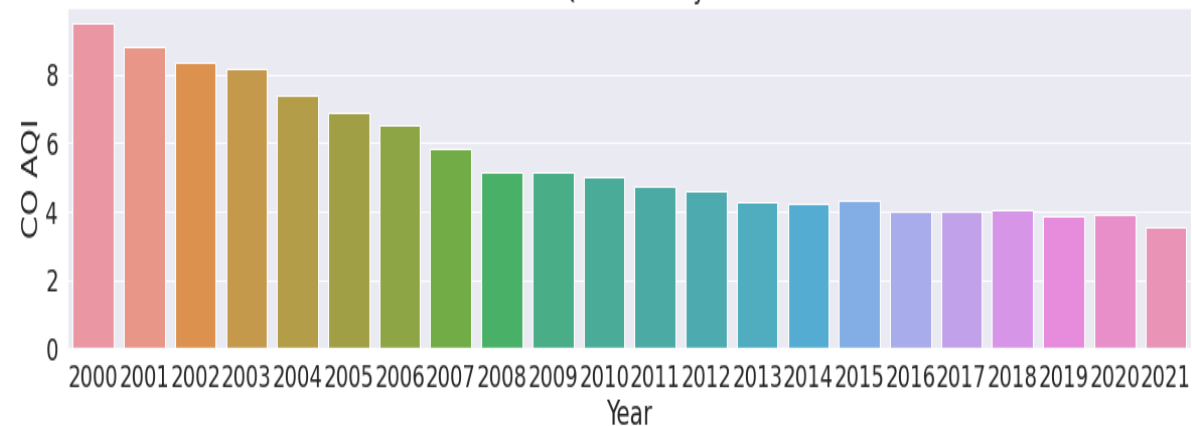
```
-RECORD 0-----  
Date          | 2000-01-01  
Year          | 2000  
Month         | 1  
Day           | 1  
Address       | 1645 E ROOSEVELT ST-CENTRAL PHOENIX STN  
State         | Arizona  
County        | Maricopa  
City          | Phoenix  
O3 Mean       | 0.019765  
O3 1st Max Value | 0.04  
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

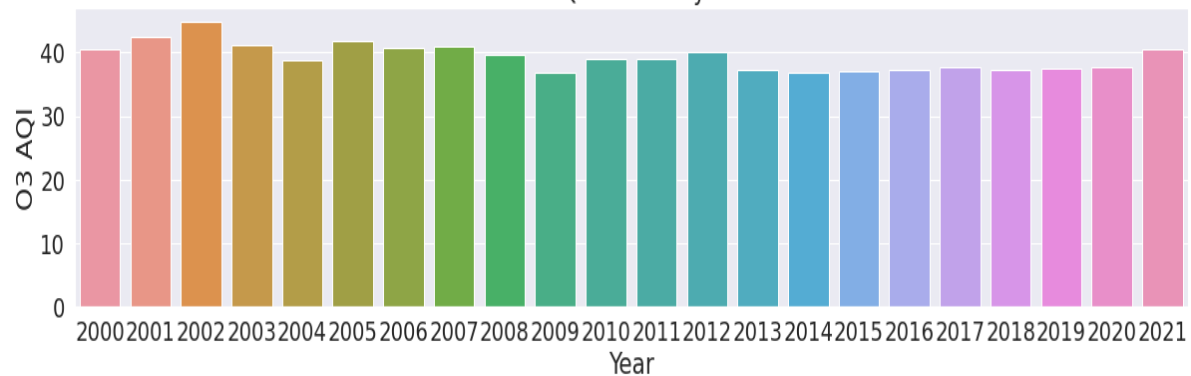
NO_2 AQI Over the years



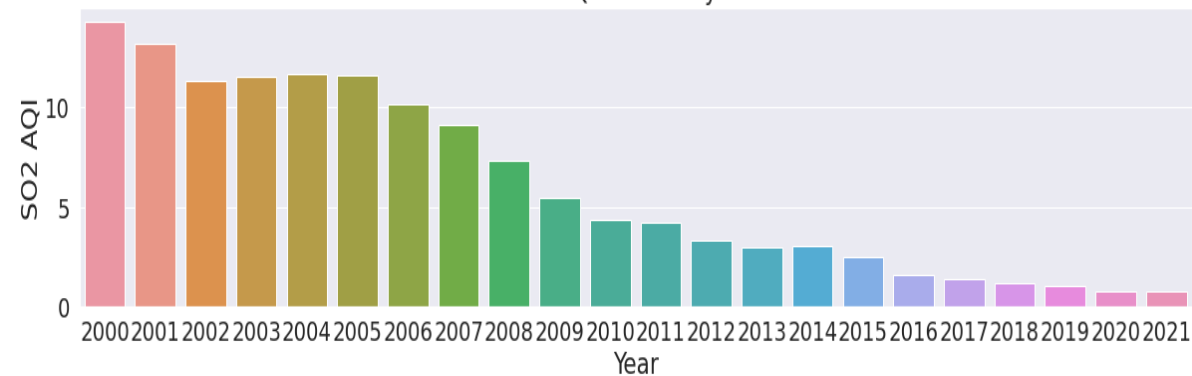
CO AQI Over the years



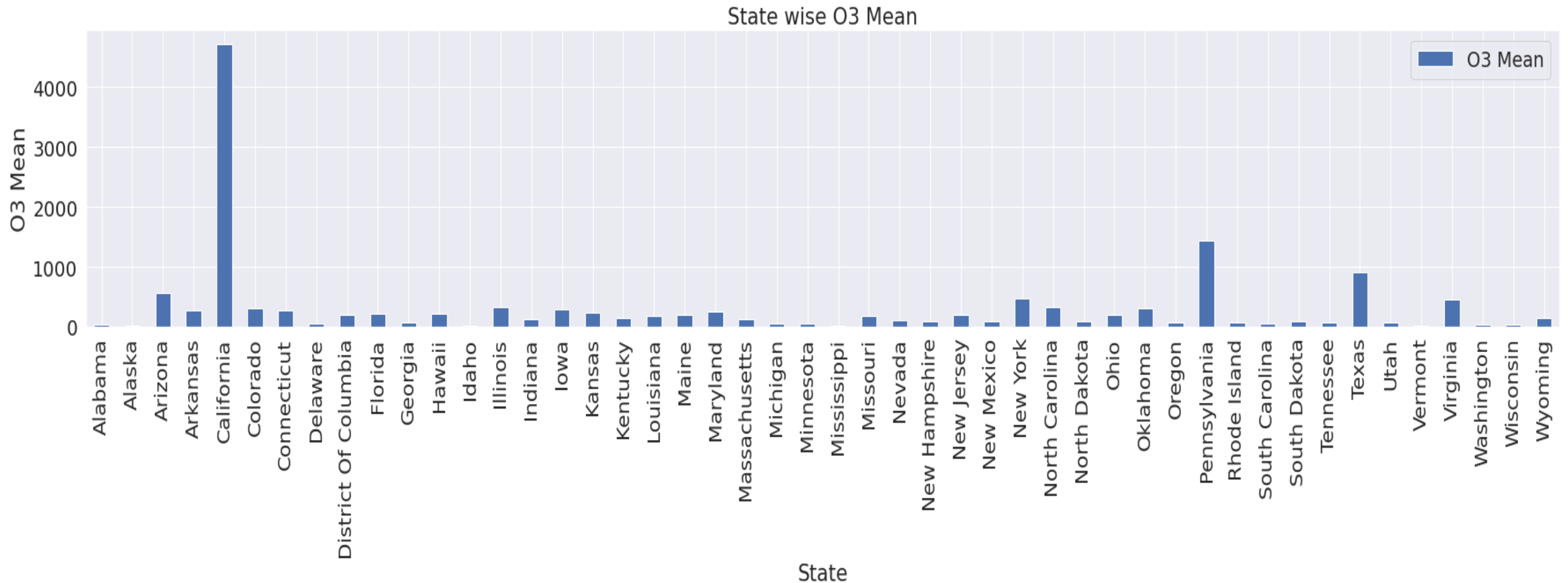
O_3 AQI Over the years



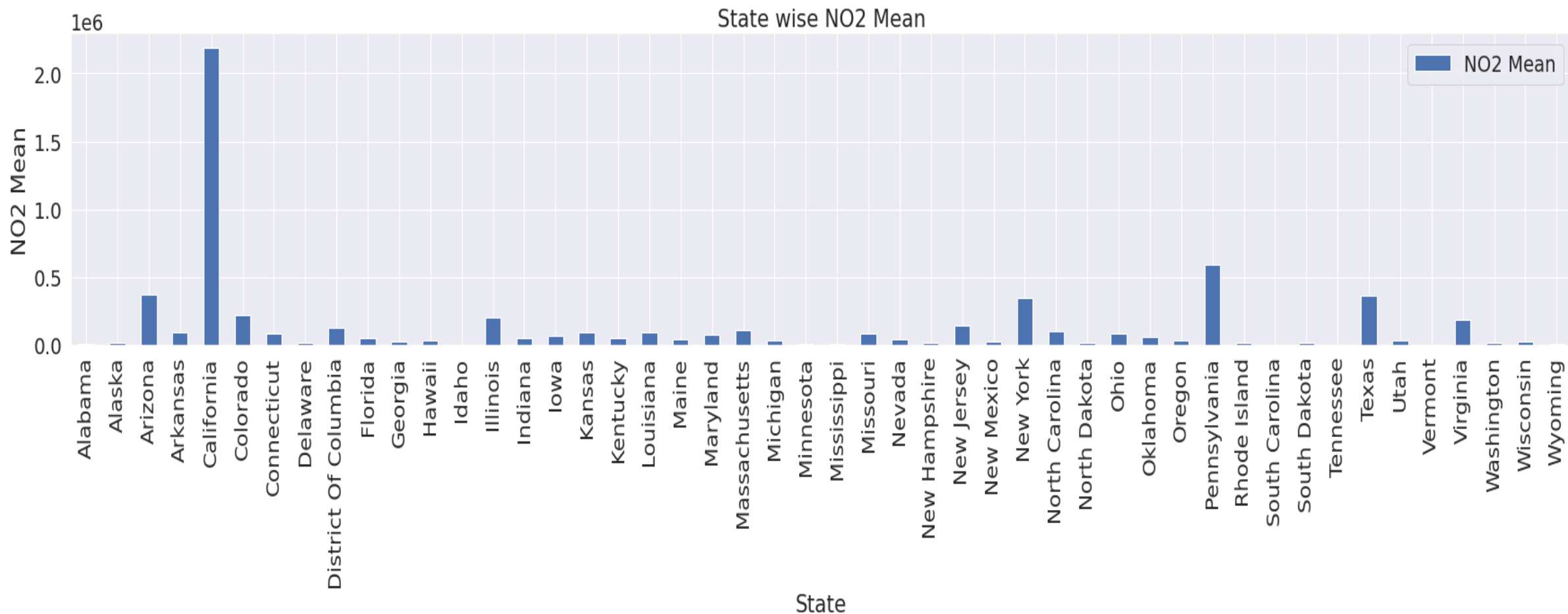
SO_2 AQI Over the years



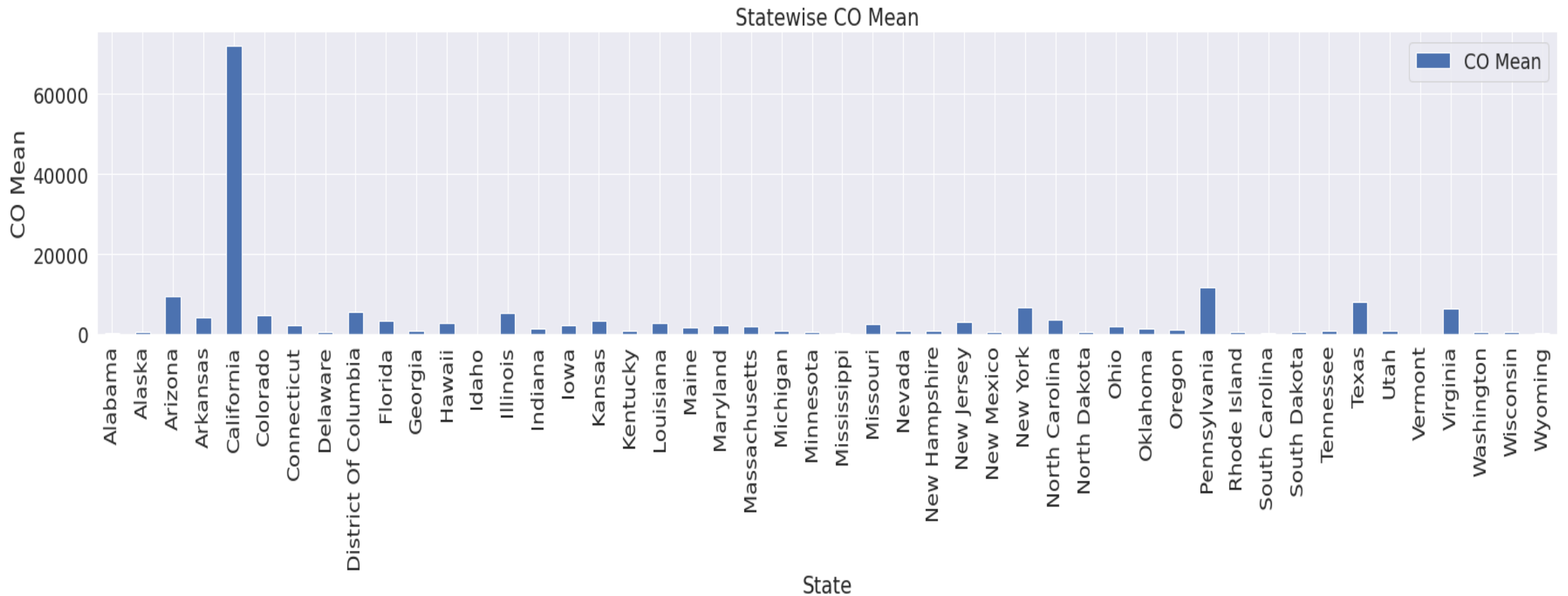
State wise Ozone molecule mean



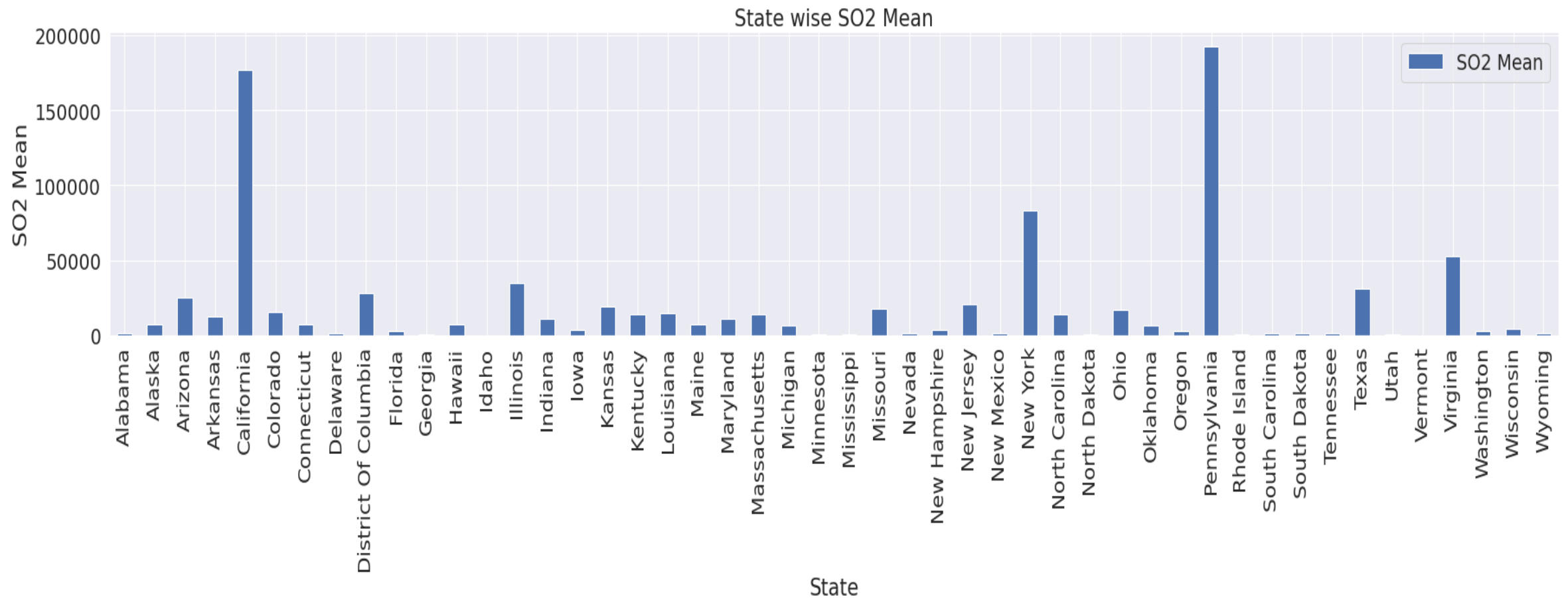
State wise Nitrogen dioxide molecule mean



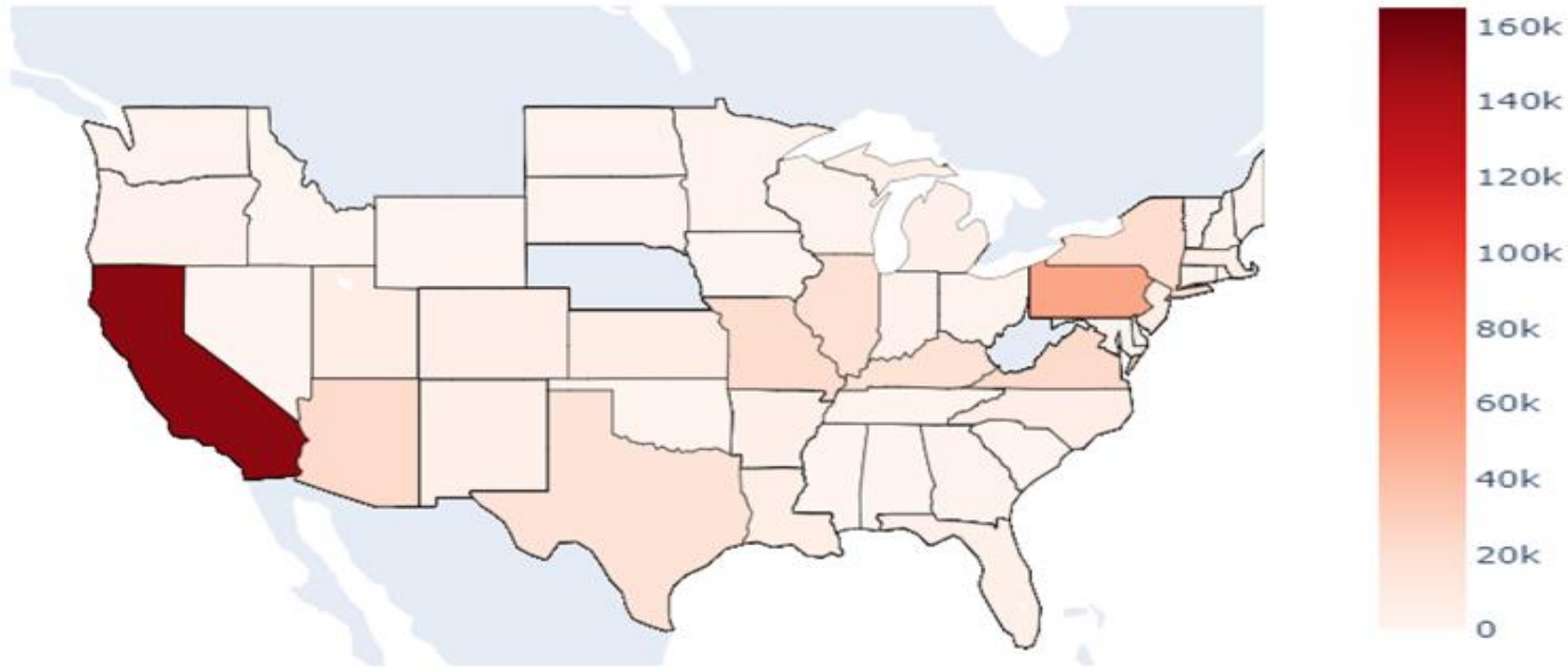
State wise Carbon Monoxide molecule mean



State wise sulfur dioxide molecule mean



State wise NO2 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	0.08800000000000001
Vallejo	0.08800000000000001
San Diego	0.08800000000000001
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



THANK YOU