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.

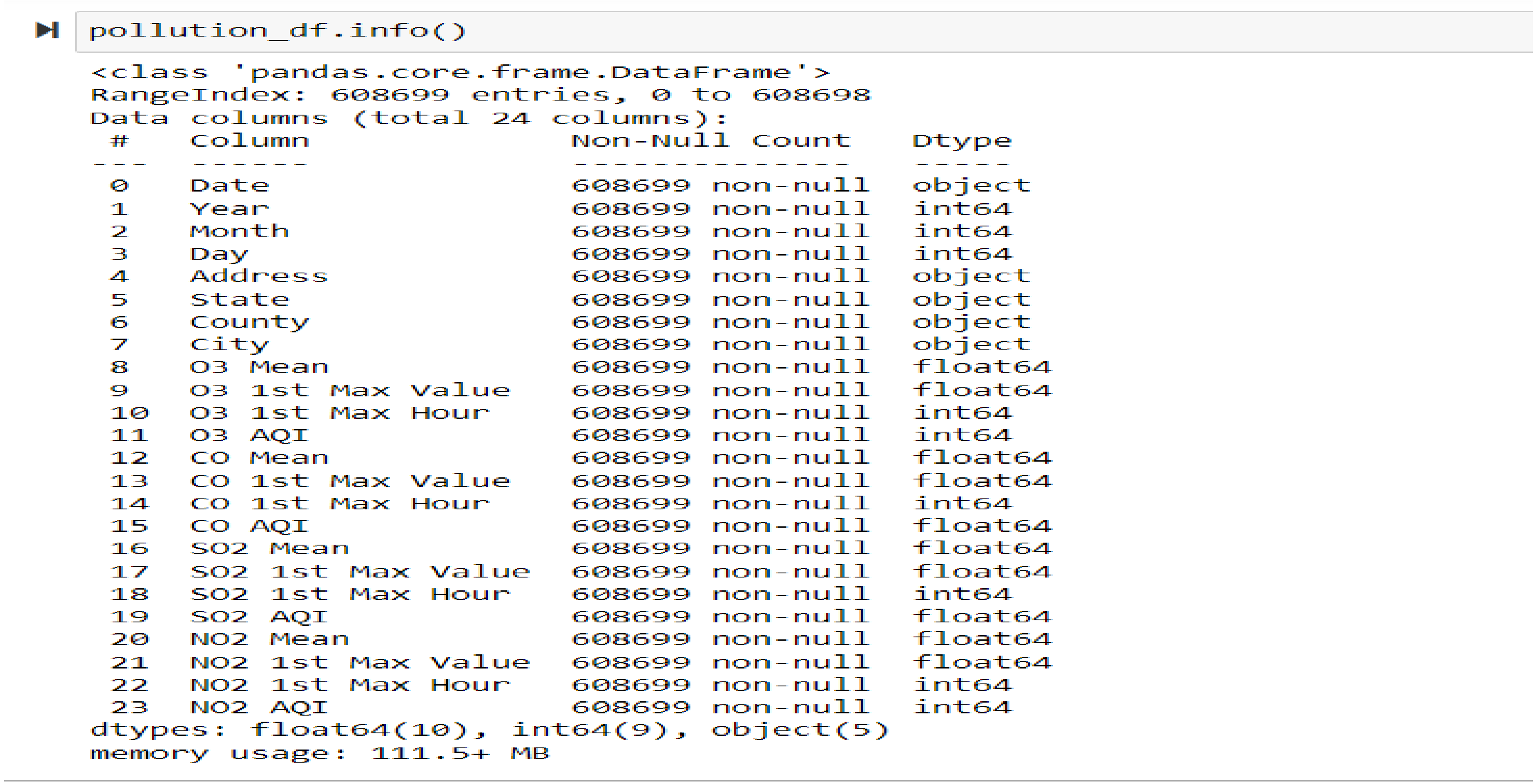
DATA SOURCE

- 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
- Date, Year, Month, Day, Address, State, County, City, 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
- Size of the dataset: 97.76 MB

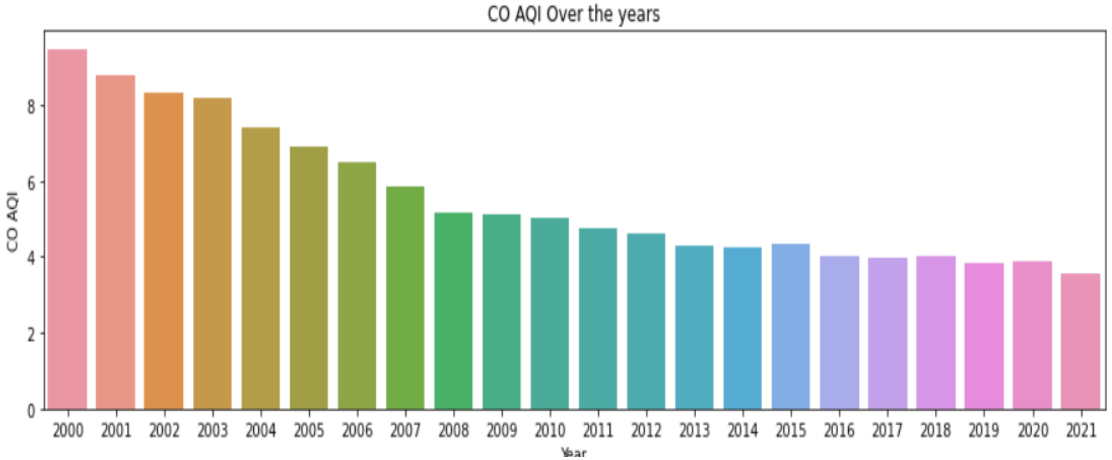
Exploratory Data Analysis – Columns and datatypes

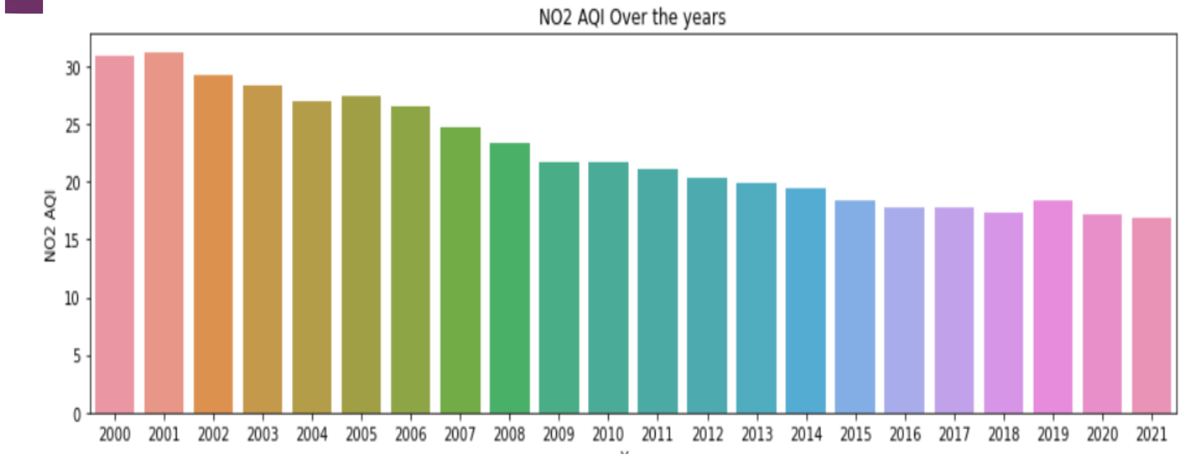


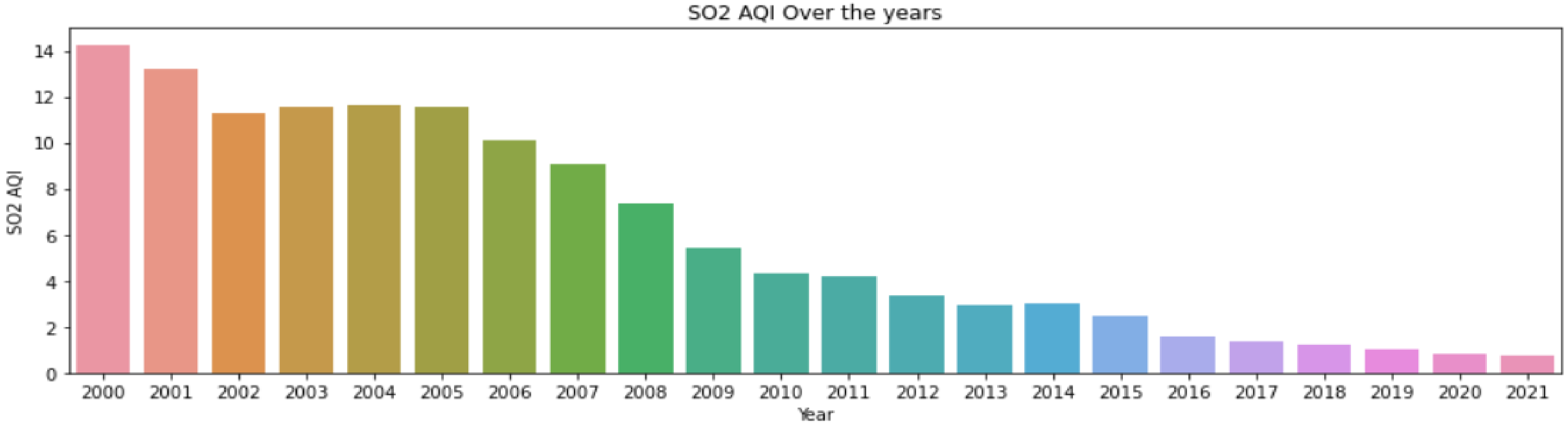
```
pollution_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 608699 entries, 0 to 608698
Data columns (total 24 columns):
 #   Column              Non-Null Count   Dtype  
---  ------              --------------   -----  
 0   Date                608699 non-null  object 
 1   Year                608699 non-null  int64  
 2   Month               608699 non-null  int64  
 3   Day                 608699 non-null  int64  
 4   Address             608699 non-null  object 
 5   State               608699 non-null  object 
 6   County              608699 non-null  object 
 7   City                608699 non-null  object 
 8   O3 Mean             608699 non-null  float64
 9   O3 1st Max Value    608699 non-null  float64
 10  O3 1st Max Hour     608699 non-null  int64  
 11  O3 AQI              608699 non-null  int64  
 12  CO Mean             608699 non-null  float64
 13  CO 1st Max Value    608699 non-null  float64
 14  CO 1st Max Hour     608699 non-null  int64  
 15  CO AQI              608699 non-null  float64
 16  SO2 Mean            608699 non-null  float64
 17  SO2 1st Max Value   608699 non-null  float64
 18  SO2 1st Max Hour    608699 non-null  int64  
 19  SO2 AQI             608699 non-null  float64
 20  NO2 Mean            608699 non-null  float64
 21  NO2 1st Max Value   608699 non-null  float64
 22  NO2 1st Max Hour    608699 non-null  int64  
 23  NO2 AQI             608699 non-null  int64  
dtypes: float64(10), int64(9), object(5)
memory usage: 111.5+ MB
```

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







Proposed Solution

- Predict which city will have the highest AQI value in the coming years.
- Identifying which pollutant gas is affecting the environment more in each city and state.
- Check which air pollutant has increased over a period and which air pollutant has decreased across the states in USA.

THANK YOU