Schema/Summary Statistics

Schema and stats for Air Pollution Data

```
root
 |-- coord: struct (nullable = true)
 |    |-- lat: double (nullable = true)
 |    |-- lon: double (nullable = true)
 |-- list: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- components: struct (nullable = true)
 |    |    |    |-- co: double (nullable = true)
 |    |    |    |-- nh3: double (nullable = true)
 |    |    |    |-- no: double (nullable = true)
 |    |    |    |-- no2: double (nullable = true)
 |    |    |    |-- o3: double (nullable = true)
 |    |    |    |-- pm10: double (nullable = true)
 |    |    |    |-- pm2_5: double (nullable = true)
 |    |    |    |-- so2: double (nullable = true)
 |    |    |-- dt: long (nullable = true)
 |    |    |-- main: struct (nullable = true)
 |    |    |    |-- aqi: long (nullable = true)
```

	latitude	longitude	date	AQI	CO	NH3	NO	NO2	O3	PM10	PM2.5	SO2
0	42.3601	-71.0589	2024-02-05 22:00:00	2	280.38	0.77	0.18	11.65	84.40	1.89	1.30	1.61
1	42.3601	-71.0589	2024-02-05 23:00:00	2	293.73	0.90	0.00	13.54	80.11	2.23	1.56	1.67
2	42.3601	-71.0589	2024-02-06 00:00:00	2	303.75	1.06	0.00	14.74	77.25	2.83	2.03	1.79
3	42.3601	-71.0589	2024-02-06 01:00:00	2	300.41	1.08	0.00	13.54	76.53	2.90	2.08	1.64
4	42.3601	-71.0589	2024-02-06 02:00:00	2	290.39	1.00	0.00	11.31	77.25	2.46	1.77	1.45
5	42.3601	-71.0589	2024-02-06 03:00:00	2	283.72	0.92	0.00	9.68	77.96	2.08	1.49	1.31
6	42.3601	-71.0589	2024-02-06 04:00:00	2	273.70	0.75	0.00	7.63	79.39	1.51	1.05	1.15
7	42.3601	-71.0589	2024-02-06 05:00:00	2	267.03	0.66	0.00	6.34	79.39	1.12	0.76	1.15
8	42.3601	-71.0589	2024-02-06 06:00:00	2	263.69	0.64	0.00	5.65	79.39	0.98	0.66	1.15
9	42.3601	-71.0589	2024-02-06 07:00:00	2	263.69	0.66	0.00	5.57	78.68	0.94	0.63	1.19

Schema and stats for Historical Weather Data

```
root
 |-- calctime: double (nullable = true)
 |-- city_id: long (nullable = true)
 |-- cnt: long (nullable = true)
 |-- cod: string (nullable = true)
 |-- list: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- clouds: struct (nullable = true)
 |    |    |    |-- all: long (nullable = true)
 |    |    |-- dt: long (nullable = true)
 |    |    |-- main: struct (nullable = true)
 |    |    |    |-- feels_like: double (nullable = true)
 |    |    |    |-- humidity: long (nullable = true)
 |    |    |    |-- pressure: long (nullable = true)
 |    |    |    |-- temp: double (nullable = true)
 |    |    |    |-- temp_max: double (nullable = true)
 |    |    |    |-- temp_min: double (nullable = true)
 |    |    |-- rain: struct (nullable = true)
 |    |    |    |-- 1h: double (nullable = true)
 |    |    |    |-- 3h: double (nullable = true)
 |    |    |-- snow: struct (nullable = true)
 |    |    |    |-- 1h: double (nullable = true)
 |    |    |    |-- 3h: double (nullable = true)
 |    |    |-- weather: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |-- icon: string (nullable = true)
 |    |    |    |    |-- id: long (nullable = true)
```

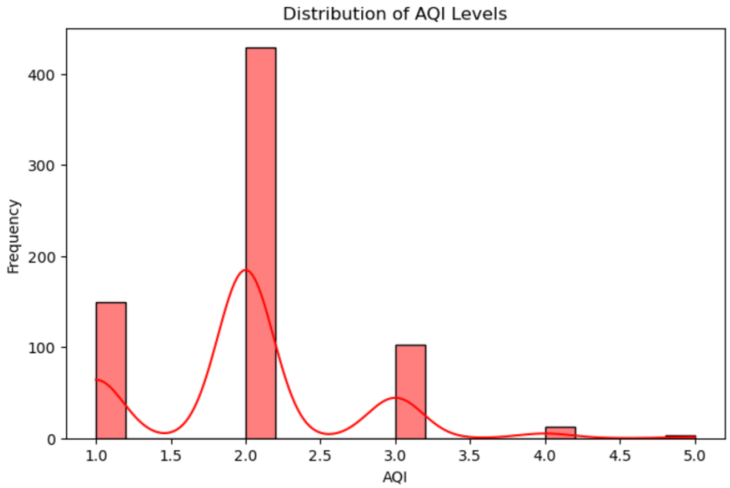
| | | | |-- main: string (nullable = true) | | |-- wind: struct (nullable = true) | | | |-- deg: long (nullable = true) | | | |-- gust: double (nullable = true) | | | |-- speed: double (nullable = true) |-- message: string (nullable = true)

	city_id	date	temperature	feels_like	temp_min	temp_max	humidity	pressure	wind_speed	wind_direction	cloud_coverage	weather_condition
0	1	2025-02-12 23:00:00	272.94	266.88	271.83	273.79	77	1029	7.20	60	100	Clouds
1	1	2025-02-13 00:00:00	273.05	267.02	271.85	273.79	78	1029	7.20	70	100	Clouds
2	1	2025-02-13 01:00:00	273.32	266.96	272.18	273.96	76	1029	8.23	90	100	Clouds
3	1	2025-02-13 02:00:00	273.40	266.87	272.28	274.26	77	1029	8.75	90	100	Snow
4	1	2025-02-13 03:00:00	273.52	267.22	272.28	274.31	77	1028	8.23	100	100	Clouds
5	1	2025-02-13 04:00:00	273.40	266.69	272.23	274.26	80	1027	9.26	100	100	Snow
6	1	2025-02-13 05:00:00	273.09	266.09	272.20	273.96	82	1027	10.29	100	100	Snow
7	1	2025-02-13 06:00:00	273.12	266.32	271.85	273.96	84	1026	9.26	100	100	Snow
8	1	2025-02-13 07:00:00	273.34	266.44	272.25	274.26	86	1024	9.77	100	100	Snow
9	1	2025-02-13 08:00:00	273.58	267.11	272.36	274.38	88	1024	8.75	90	100	Snow

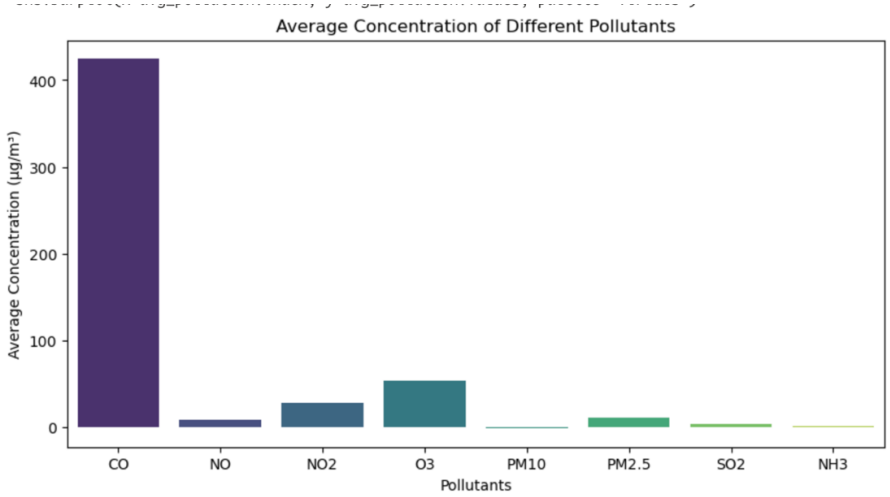
EDA Report for Air Pollution and Historical Weather Data

Air Pollution Data Analysis

1) The below graph gives a distribution of AQI levels vs frequency (how many times a particular level appears on the chart). Here we can see that most of the observations fall under level 2.5 which means the AQI on most days is considered good to excellent according the weatherAPI AQI levels. However, few observations at level 3 or more shows that for particular instances the AQI was from bad to worst.

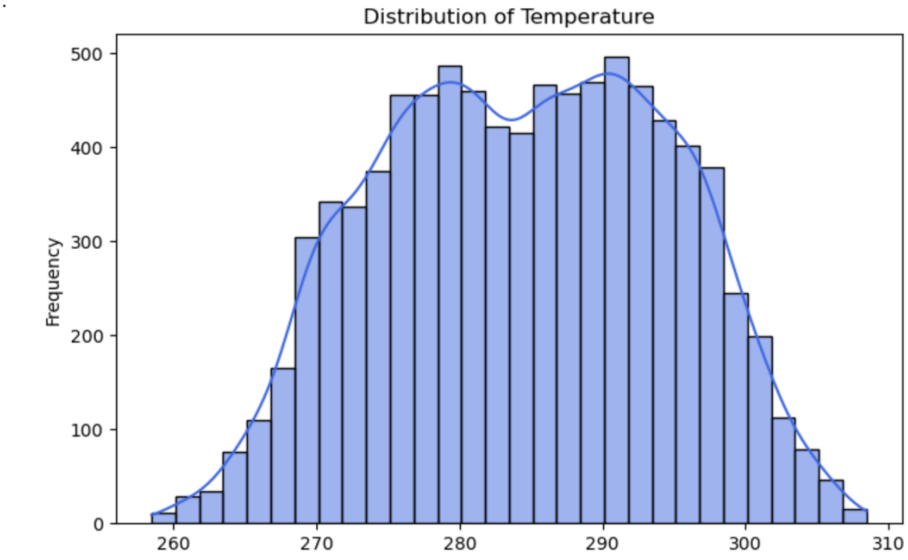


2) This graph gives out a distribution of Average concentration of different pollutants found in the air. As we can clearly see the difference between CO and other pollutants, we can say that CO is the main contributor for air pollution. All other pollutants are in very negligible concentrations compared to CO.

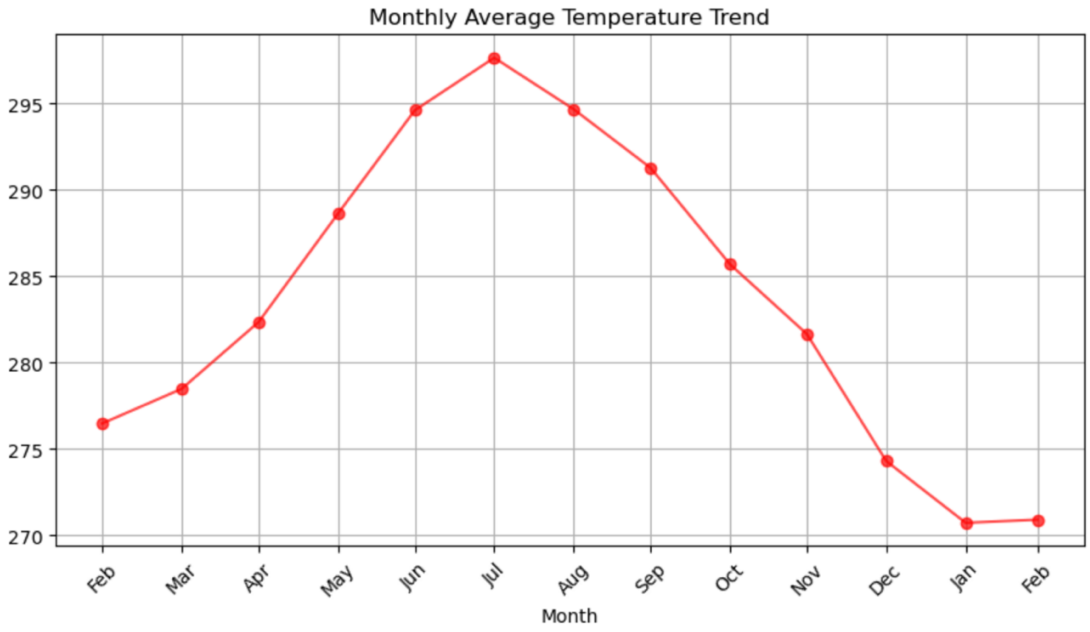


Historical Weather Data Analysis

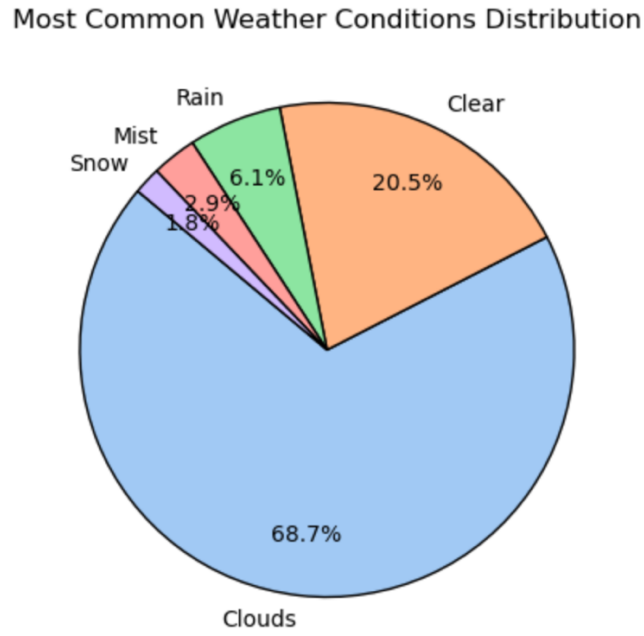
1) The graph below shows the spread of temperature in Kelvin. As we see that for most days, the temperature is between 270-300 K interval. For Boston city, the chunk of observations is in between 275-295 K which makes sense because it is on the colder side for most days and in summer the average temperature goes around 290-300K. This distribution follows a normal distribution of temperature ranges.



2) This graph below shows the monthly average temperatures in Kelvin. The average temperatures are around 272 to 280 K for the months of November till April. And for the rest of the months, the temperatures are around the 285 to 300 K range during summers.



3) The below graph shows us how the most common weather conditions are distributed year-round. The topmost being cloudy conditions having 68.7%, followed by clear skies with 20.5%. The other categories like rain, mist, and snow cover about 21% of the distribution.



4) From the below correlation matrix, we can derive that 1) higher humidity has more cloud covers on an average 2) Whenever there is low air pressure, the wind speed increases resulting in more windier days.