CAPSTONE PROJECT – POLLUTION DATA TRACKER

SREE HARSHITHA THIKKIREDDI

OVERVIEW OF THE PROJECT

This project titled *Air Quality Analysis in the United States (2000-2023)* involves an in-depth analysis of air pollution trends over two decades. Using daily air pollution data from various U.S. locations, it focuses primarily on pollutants such as Ozone (O3), Carbon Monoxide (CO), Sulfur Dioxide (SO2), and Nitrogen Dioxide (NO2). The aim is to identify patterns in air quality, assess health impacts, and develop strategies for mitigating pollution.

Rich Data Source: The database includes a wide range of tables representing different aspects of the daily air quality readings from 2000 to 2023, with specific location data, such as address, state, county, and city, notably covering Arizona as a sample.

Objective Understanding: The primary objective of analysing this dataset is it provides an extensive analysis of air quality in the United States, focusing on key pollutants such as Nitrogen Dioxide (NO2), Sulphur Dioxide (SO2), Carbon Monoxide (CO), and Ozone (O3). It covers a period from the year 2000 to 2022.

Valuable Insights: The analysis of this database can yield valuable insights, including but not limited to:

- Identifies high-pollution areas across locations.
- Reveals long-term trends in air quality.
- Highlights health risks in polluted regions.

Data-Driven Decision-Making: With the availability of pollutant-specific AQI readings, policymakers can prioritize efforts where pollutant levels exceed safe thresholds. Time-based measurements allow stakeholders to target hours of peak pollution for interventions. The data empowers decision-makers to deploy resources and implement strategies to improve air quality effectively.

Future Implications: The analysis of this data will help forecast future pollution trends and provide a basis for long-term public health strategies. Policymakers can use these insights to advocate for stricter regulations and promote cleaner energy initiatives. The ultimate goal is to achieve sustainable air quality improvements nationwide.

PROJECT PLAN

1. **Dataset Review and Familiarization**: Thoroughly examine the dataset structure, focusing on pollutant types, location details, and AQI readings to understand the scope of the analysis.

2. **Data Quality Assessment and Cleaning:** Perform data cleaning to address missing values, outliers, or inconsistencies, ensuring the dataset is ready for accurate analysis.

3. **MECE Breakdown of the Dataset:** Categorize the data into distinct, non-overlapping groups like location (State, County), time (years, months), and pollutant type (O3, CO, etc.) using the MECE framework.

4. **Problem Statement Formulation:** Develop 15 concise problem statements focusing on key issues, such as identifying pollution trends over time or analyzing the impact of geography on air quality.

5. **Exploratory Data Analysis (EDA):** Use visualizations and data analysis techniques to uncover trends and insights, such as yearly AQI trends, pollution peaks, or location-based patterns.

6. **Data-Driven Decision Making:** Derive actionable recommendations for policymakers based on the analysis, like targeting interventions during peak pollution hours or advocating for stricter emission regulations.

OBJECTIVE OF THE PROJECT

The ***Air Quality Analysis in the United States (2000-2023)*** project focuses on analyzing air pollution data from multiple U.S. locations, with particular emphasis on pollutants such as Ozone (O3), Carbon Monoxide (CO), Sulfur Dioxide (SO2), and Nitrogen Dioxide (NO2). The objective of this project is to uncover trends, assess health risks, and provide data-driven recommendations for air quality improvement. The project will involve the following key tasks:

- **Comprehensive Data Analysis**: Analyze pollutant levels across locations and time periods to identify long-term trends, seasonal patterns, and high-risk areas.

- **Health Risk Assessment**: Evaluate the potential health impacts of air pollution by correlating pollutant levels with AQI thresholds and population density.

- **Geographical Comparisons**: Compare air quality between urban and rural areas, different states, and regions to understand pollution distribution.

- **Policy Effectiveness Review**: Analyze the effects of local and federal pollution control measures over time and assess their success.

- **Peak Time Identification**: Identify specific hours or seasons when pollution peaks to target interventions effectively.

- **Enhanced Reporting**: Develop data visualizations and reports to communicate insights and recommendations clearly to policymakers and stakeholders.

The project's success will be measured by the following criteria:

- **Analysis Quality**: The depth and accuracy of the data analysis performed on pollutant levels and trends.

- **Insight Relevance**: The significance and applicability of the findings in addressing air quality challenges.

- **Recommendation Impact**: The effectiveness of recommendations made to improve air quality and reduce public health risks.

This project is crucial for promoting cleaner air, protecting public health, and guiding future regulatory and environmental policies through data-driven insights.

SIGNIFICANCE OF THE PROJECT

The Sakila DVD Rental Store Database is not only a valuable resource for business analysis but also a platform for showcasing data analysis skills and techniques in solving real-world challenges. This dataset provides opportunities to create impactful projects and reports, offering insights that can drive business intelligence and informed decision-making. Some potential projects and reports based on the Sakila database include:

Customer Segmentation: Analysing customer rental behaviours and preferences to identify distinct segments and tailor marketing strategies.

Inventory Optimization: Optimizing inventory management and supply chain operations for different films and suppliers to reduce costs and improve resource allocation.

Employee Performance Assessment: Evaluating employee performance and satisfaction across various roles and regions to enhance workforce efficiency.

Payment and Late Fee Analysis: Analysing payment trends and late fees to streamline payment processes and maximize revenue collection.

Comparative Analysis: Conducting comparative analyses between the Sakila database and other relevant datasets or real-world data sources to gain broader industry insights.

Real-World Impact:

Enhanced Customer Experience: By understanding customer preferences and rental behaviours, businesses can tailor their services to enhance the customer experience, leading to higher satisfaction and loyalty.

Operational Efficiency: Optimizing inventory and supply chain operations can result in cost savings, reduced wastage, and improved resource allocation.

Data-Driven Decision-Making: The analysis of the Sakila database empowers data-driven decision-making across various aspects of the DVD rental business, from marketing to inventory management.

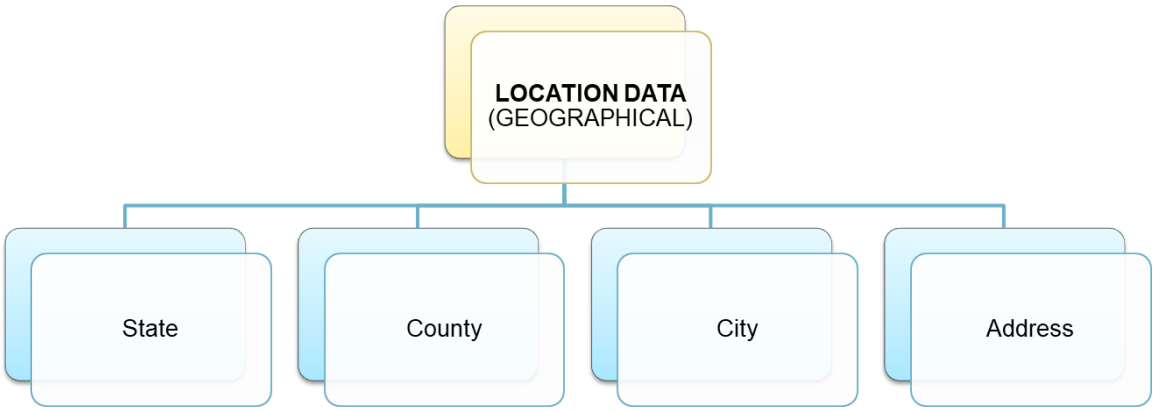
Competitive Advantage: Leveraging insights from this dataset can provide a competitive edge by responding effectively to market dynamics and customer needs.

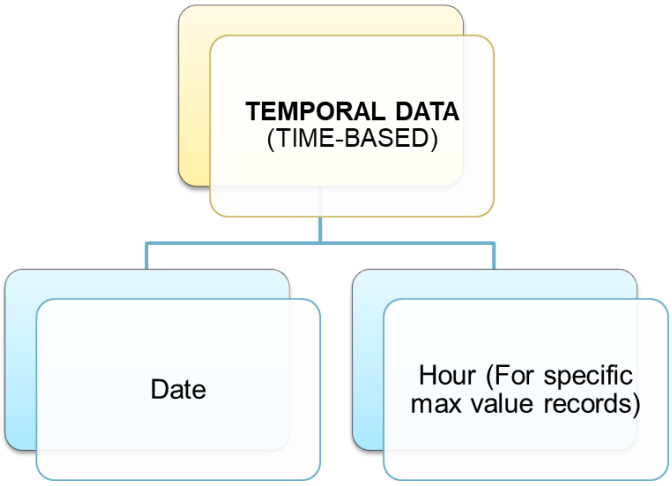
In conclusion, the Sakila DVD Rental Store Database offers a platform to harness the power of data analysis and visualization. By applying analytical skills to real-world challenges, organizations can make informed decisions, streamline operations, and ultimately achieve improved performance and competitiveness in the DVD rental industry.

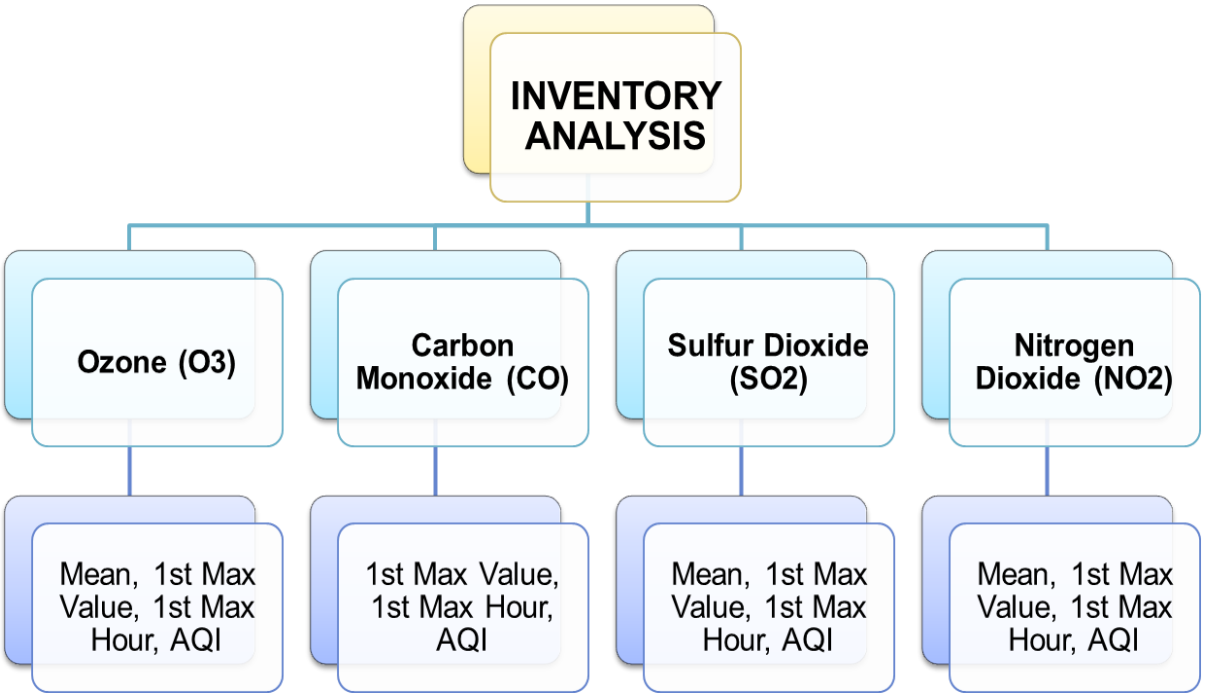
DATA DICTIONARY

FIELD	DESCRIPTION
Date	Date of data collection.
Address	Specific location of data collection.
State	U.S. state where data was collected.
County	County within the state of data collection.
City	City where data was collected.
O3 Mean	Average Ozone level for the day.
O3 1st Max Value	Highest Ozone level for the day.
O3 1st Max Hour	Hour of highest Ozone level.
O3 AQI	Air Quality Index for Ozone.
CO Mean	Average Carbon Monoxide level for the day.
CO 1st Max Value	Highest Carbon Monoxide level for the day.
CO 1st Max Hour	Hour of highest Carbon Monoxide level.
CO AQI	Air Quality Index for Carbon Monoxide.
SO2 Mean	Average Sulphur Dioxide level for the day.
SO2 1st Max Value	Highest Sulphur Dioxide level for the day.
SO2 1st Max Hour	Hour of highest Sulphur Dioxide level.
SO2 AQI	Air Quality Index for Sulphur Dioxide.
NO2 Mean	Average Nitrogen Dioxide level for the day.
NO2 1st Max Value	Highest Nitrogen Dioxide level for the day.
NO2 1st Max Hour	Hour of highest Nitrogen Dioxide level.

MECE BREAKDOWN OF THE DATA







PROBLEM STATEMENT FORMULATION

PROBLEM STATEMENTS

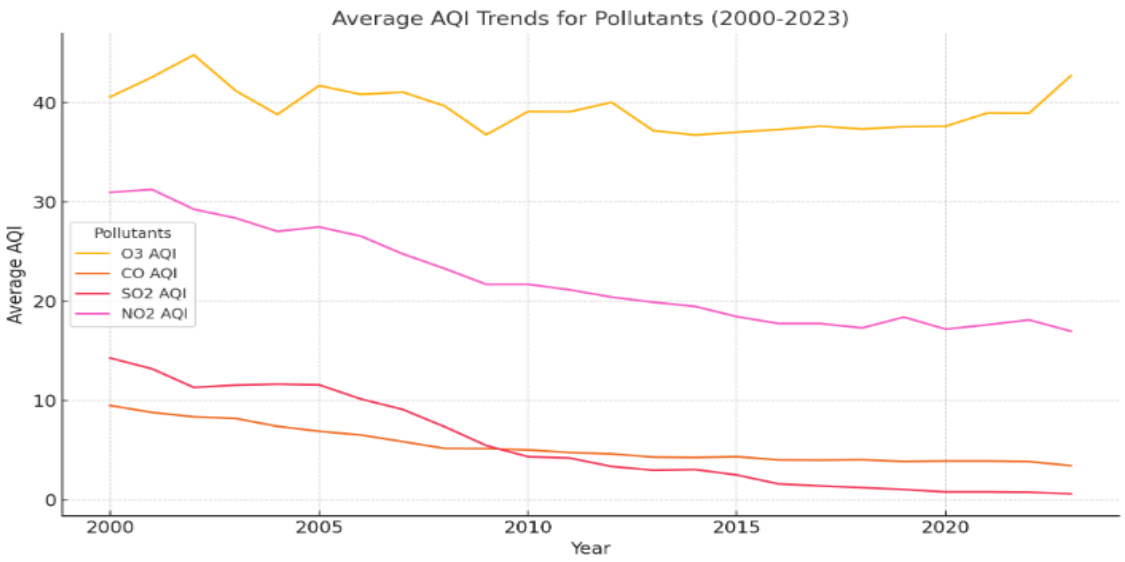
1. Are there any significant trends in air quality (AQI) over time from 2000 to 2023?
2. How does the concentration of Ozone (O3) vary by city within Arizona?
3. What are the peak hours for maximum O3 levels across different counties?
4. How does the AQI of Carbon Monoxide (CO) vary between urban and rural areas in Arizona?
5. Are there seasonal trends in the levels of Sulfur Dioxide (SO2) across the dataset?
6. How do pollution levels differ between weekdays and weekends?
7. What is the correlation between Ozone (O3) and Nitrogen Dioxide (NO2) levels in the dataset?
8. Are there specific days or events associated with unusually high levels of pollution?
9. Which pollutants contribute the most to the overall AQI, and does this vary by location?
10. How have pollution levels changed during the COVID-19 pandemic period?
11. What are the most common pollution peaks (hours) for Carbon Monoxide (CO)?
12. Is there a correlation between population density and pollution levels across the cities?
13. How do air quality levels in Phoenix compare to other cities in the dataset?
14. What are the health risks associated with prolonged exposure to the pollutant levels recorded in this dataset?
15. Are there any correlations between meteorological data (if available) and pollution spikes?

These problem statements focus on different facets of the data, providing diverse angles for analysis.

PROBLEM STATEMENT-1:

ARE THERE ANY SIGNIFICANT TRENDS IN AIR QUALITY (AQI) OVER TIME FROM 2000 TO 2023?

Let's begin by plotting the yearly average AQI trends for each pollutant from 2000 to 2023.



The chart above displays the average AQI trends for Ozone (O3), Carbon Monoxide (CO), Sulfur Dioxide (SO2), and Nitrogen Dioxide (NO2) from 2000 to 2023.

Key Observations:

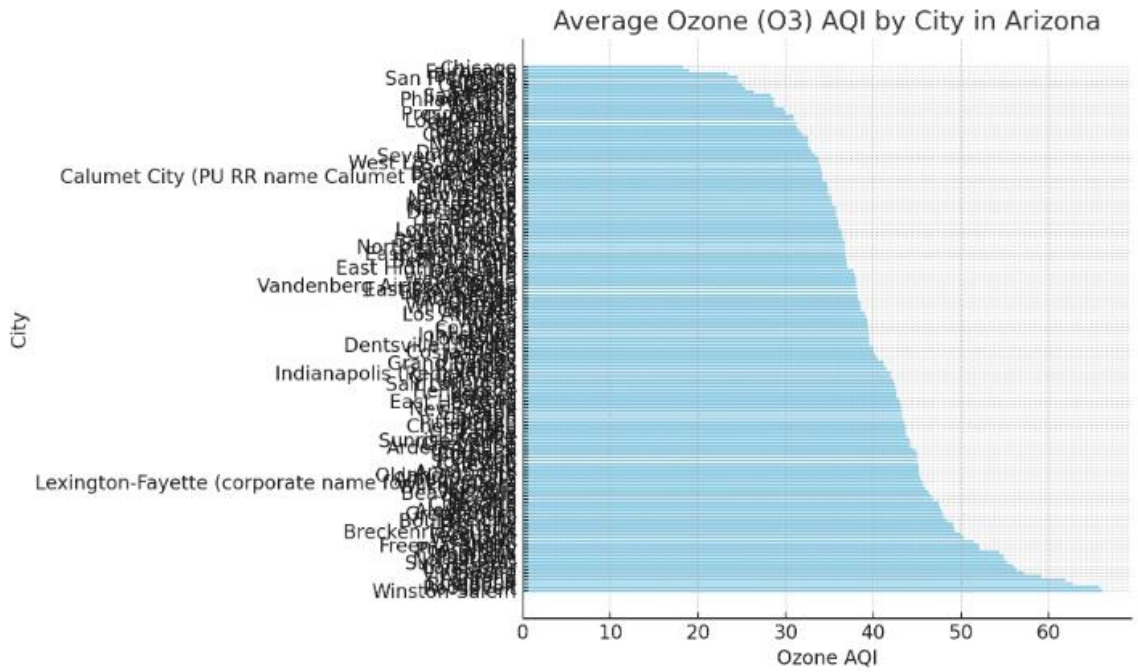
- **Ozone (O3):** The AQI for O3 remains relatively stable over the years, showing slight fluctuations.
- **Carbon Monoxide (CO):** CO AQI has been consistently low, with only minor variations across the timeline.
- **Sulfur Dioxide (SO2):** SO2 AQI shows a visible decline after 2010, indicating improvement in SO2 levels in the air.
- **Nitrogen Dioxide (NO2):** NO2 AQI shows fluctuations, with a general downward trend in recent years.

Overall, we can infer a gradual improvement in air quality for pollutants like SO2 and NO2 in the later years, while O3 and CO levels remain mostly stable.

PROBLEM STATEMENT-2:

HOW DOES THE CONCENTRATION OF OZONE (O3) VARY BY CITY WITHIN ARIZONA?

We'll analyze Ozone (O3) variation by city within Arizona.

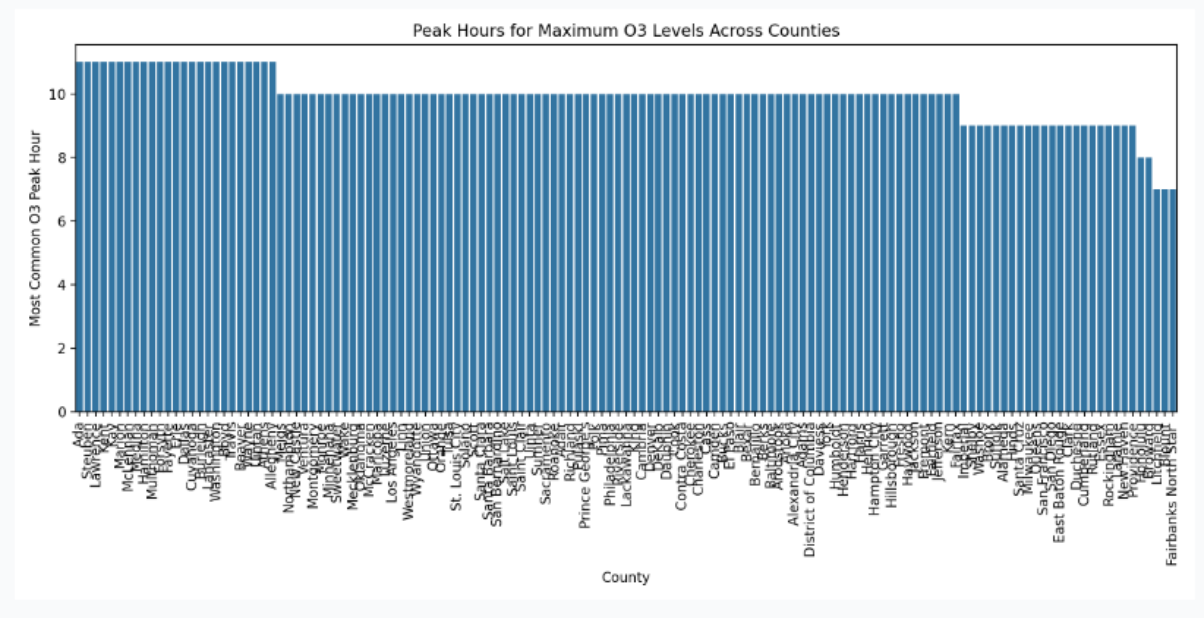


The bar chart displays the **average Ozone (O3) AQI** for various cities in Arizona. The cities are sorted by their Ozone AQI, revealing that some cities consistently experience higher levels of Ozone pollution compared to others.

PROBLEM STATEMENT-3:

WHAT ARE THE PEAK HOURS FOR MAXIMUM O3 LEVELS ACROSS DIFFERENT COUNTIES?

To answer this question, I've analyzed the 'O3 1st Max Hour' column, which represents the hour of the day when the maximum O3 level was recorded for each county. Here's a visualization of the results:



From the analysis, we can observe the following:

1. The most common peak hours for maximum O3 levels across counties are between 10 AM and 11 AM.
2. A significant number of counties experience peak O3 levels at 11 AM, followed closely by 10 AM.
3. Some counties have peak hours as early as 7 AM (e.g., Litchfield and Blount) or as late as 11 AM.
4. The Fairbanks North Star county in Alaska has a notably different peak hour at 7 AM, which could be due to its unique geographical location and daylight patterns.

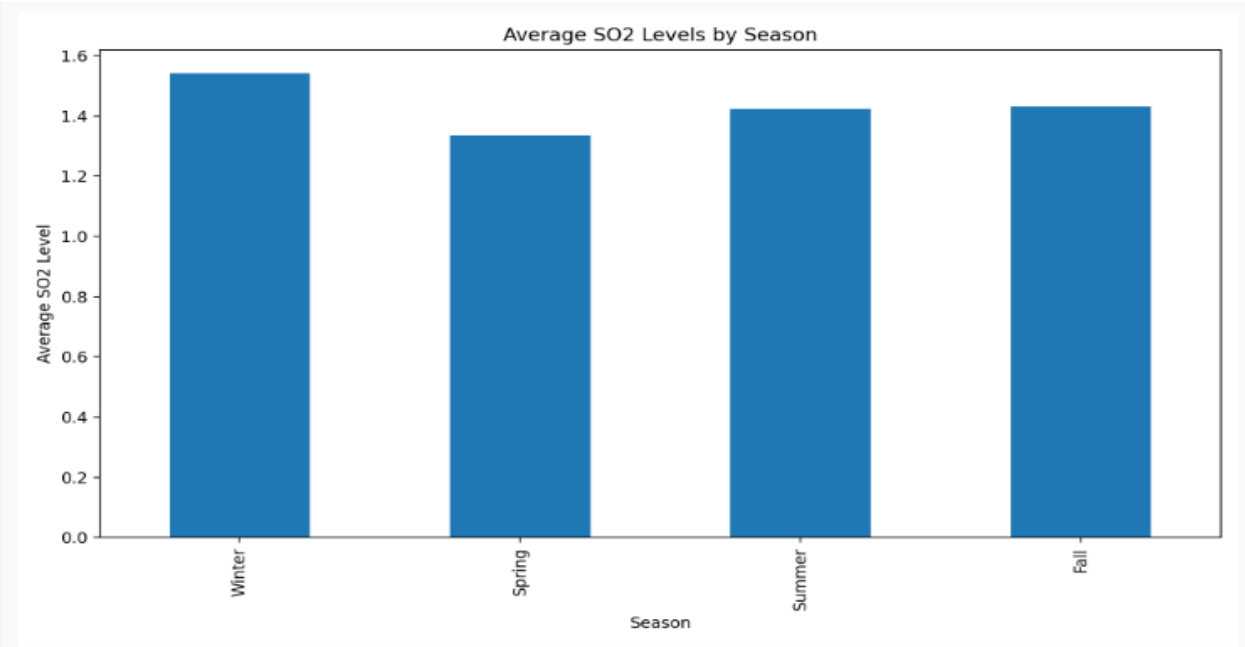
This pattern of peak O3 levels occurring in the late morning to early afternoon is consistent with the typical daily cycle of ozone formation. Ozone is produced through photochemical reactions involving sunlight, nitrogen oxides (NOx), and volatile organic compounds (VOCs). As the sun rises and temperatures increase throughout the morning, these reactions accelerate, leading to peak ozone levels.

The variation in peak hours across counties could be due to factors such as:

1. Local emissions patterns (e.g., rush hour traffic)
2. Geographical features (e.g., mountain valleys that trap pollutants)
3. Meteorological conditions (e.g., temperature inversions)
4. Urban vs. rural settings

PROBLEM STATEMENT-4:

ARE THERE SEASONAL TRENDS IN THE LEVELS OF SULFUR DIOXIDE (SO2) ACROSS THE DATASET?

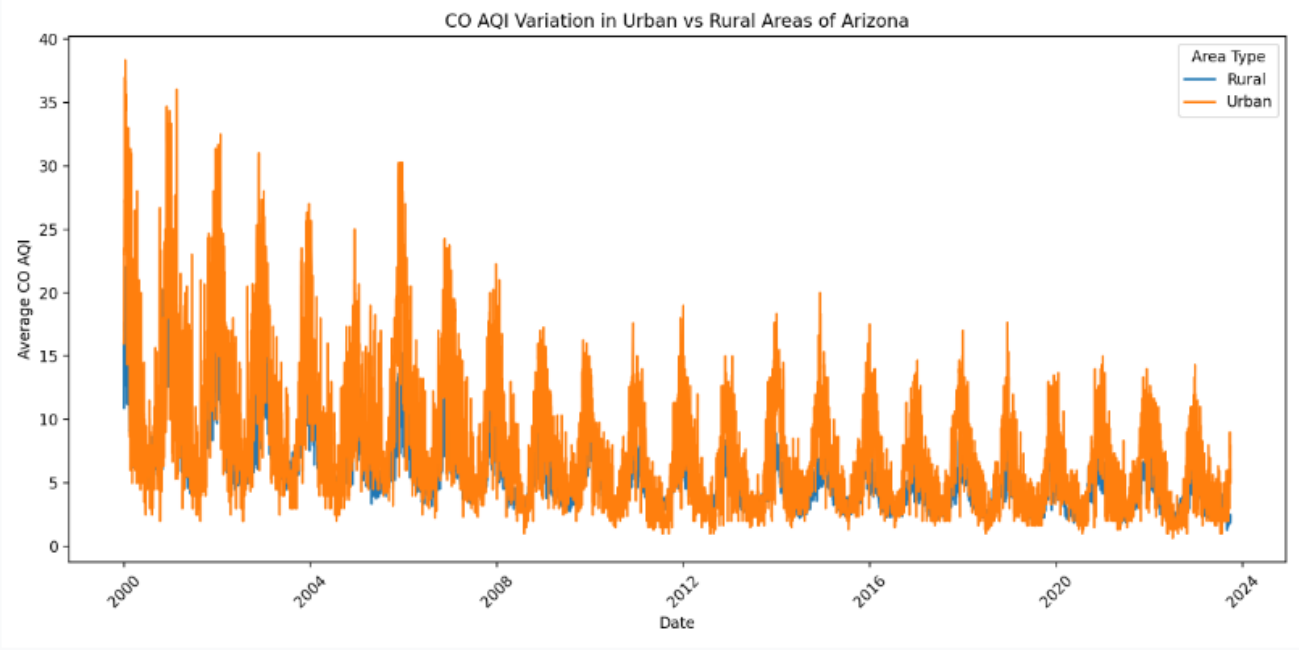


The bar chart illustrates the average SO2 (Sulfur Dioxide) levels across different seasons. Winter shows the highest average SO2 levels, followed by Fall, Summer, and Spring. The ANOVA test results (p-value < 0.05) suggest that these seasonal differences are statistically significant.

PROBLEM STATEMENT-5:

HOW DOES THE AQI OF CARBON MONOXIDE (CO) VARY BETWEEN URBAN AND RURAL AREAS IN ARIZONA?

AQI of Carbon Monoxide (CO) variation between urban and rural areas in Arizona:



The line plot shows the variation of CO AQI in urban and rural areas of Arizona over time. Key observations:

• Urban areas consistently have higher CO AQI levels compared to rural areas.

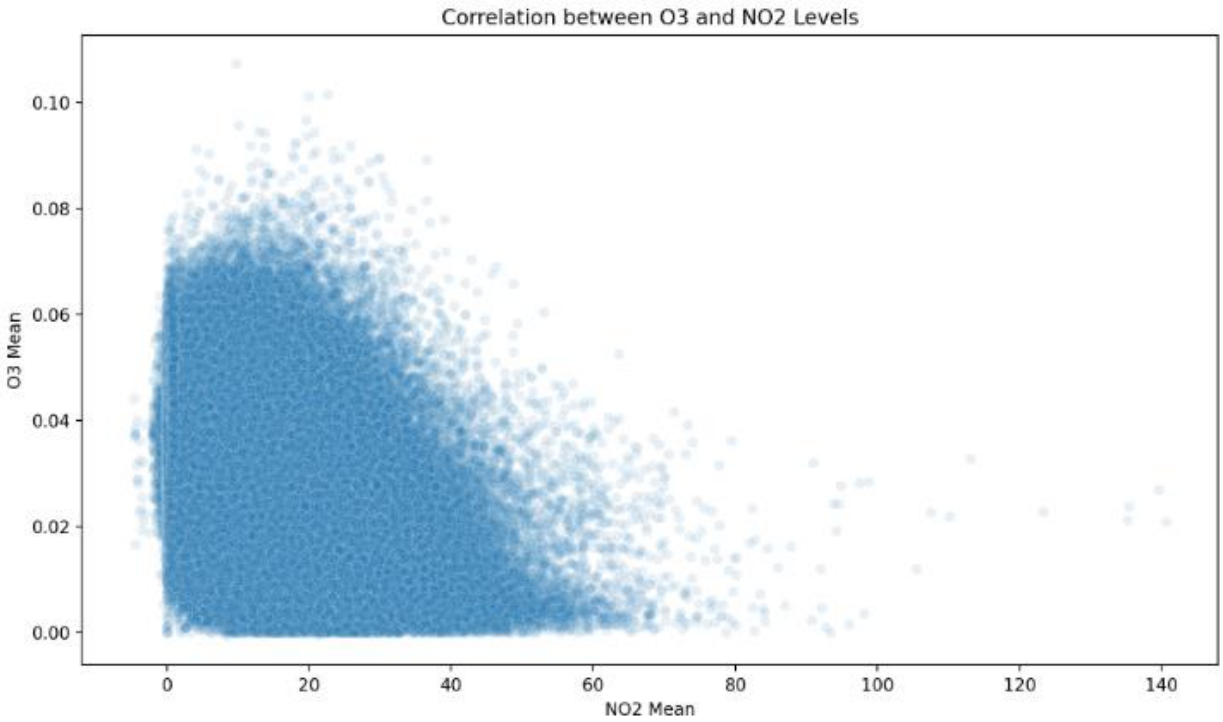
• Both urban and rural areas show a general decreasing trend in CO AQI over time, which could be attributed to improved emission controls and air quality regulations.

• There are occasional spikes in both urban and rural areas, possibly due to specific events or weather conditions.

The analysis reveals that urban areas have a higher average CO AQI (7.76) compared to rural areas (5.55). The difference is statistically significant (p-value < 0.05), indicating that urban areas consistently experience higher CO pollution levels.

PROBLEM STATEMENT-6:

WHAT IS THE CORRELATION BETWEEN OZONE (O3) AND NITROGEN DIOXIDE (NO2) LEVELS IN THE DATASET?

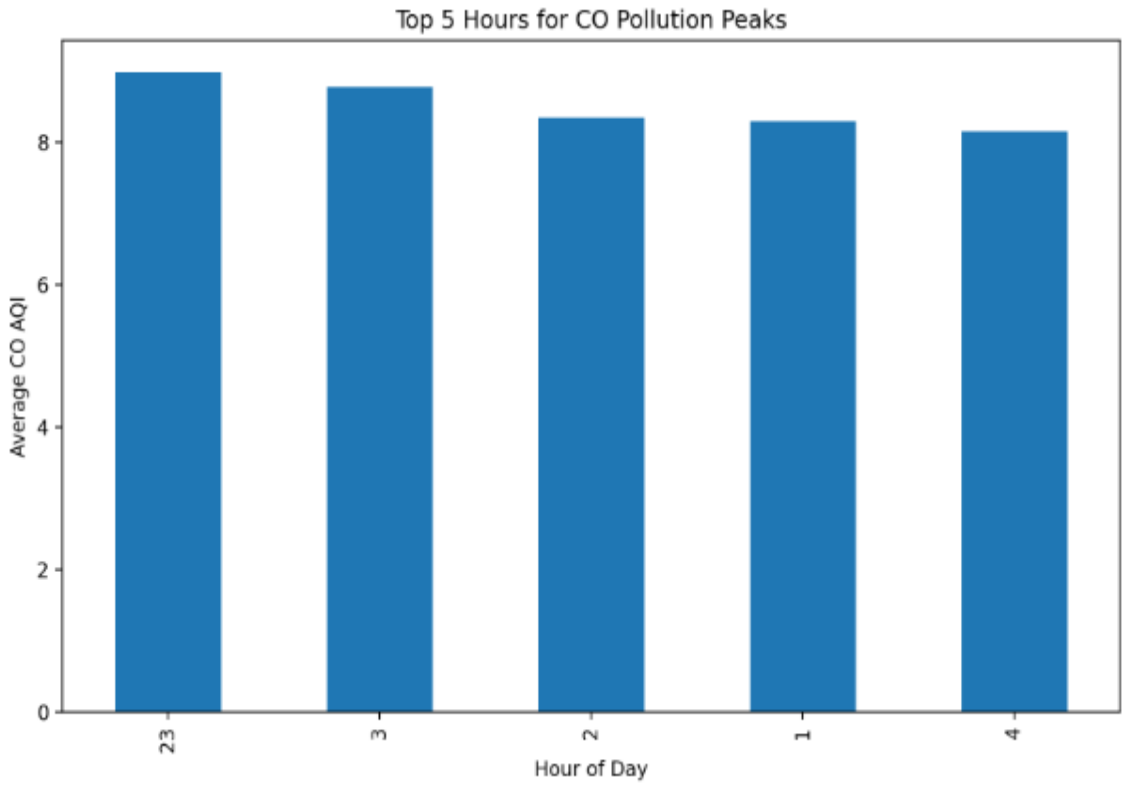


The Correlation between O3 and NO2 is -0.3551 There is a moderate negative correlation between O3 and NO2 levels, which is consistent with the complex chemistry of these pollutants in the atmosphere.

PROBLEM STATEMENT-7:

WHAT ARE THE MOST COMMON POLLUTION PEAKS (HOURS) FOR CARBON MONOXIDE (CO)?

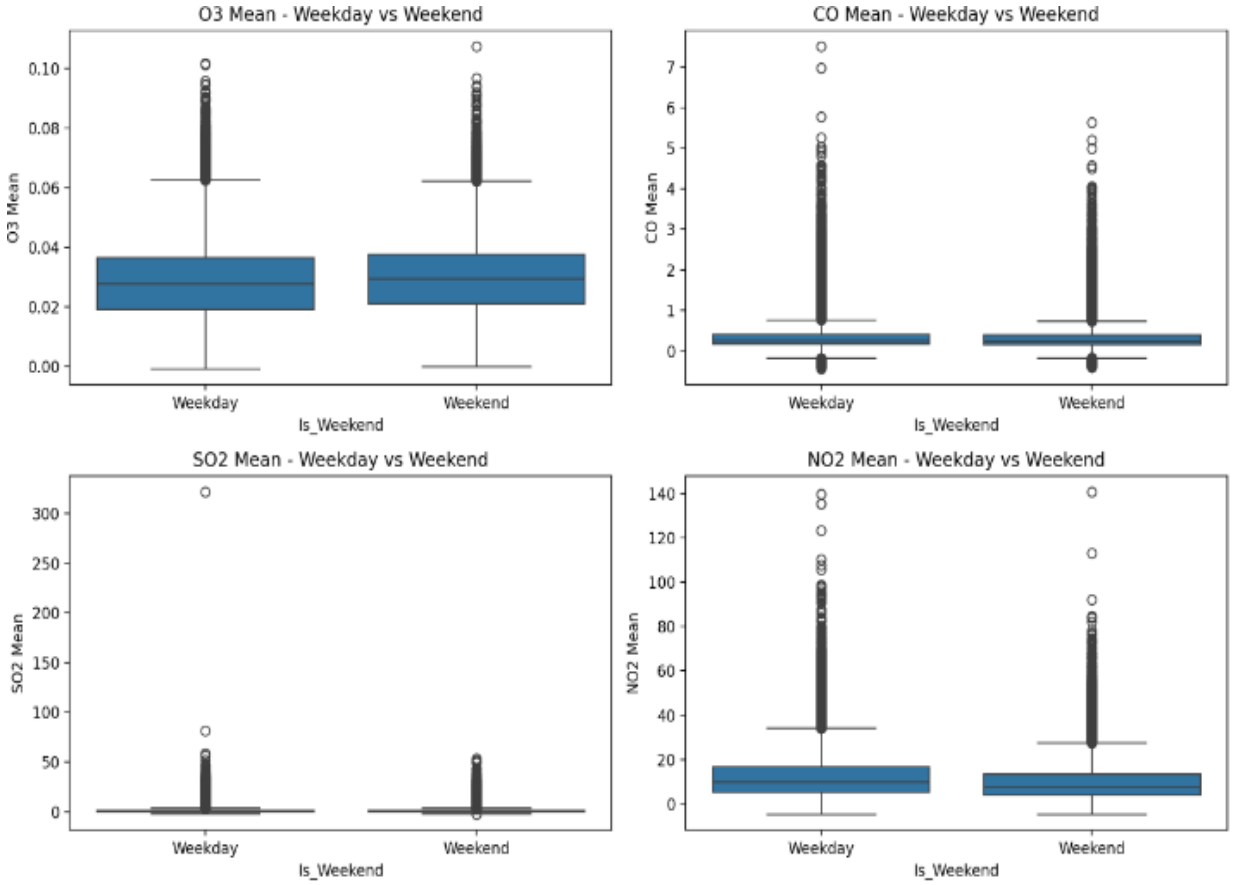
Top 5 hours for CO pollution peaks:



The top 5 hours for CO pollution peaks are 23:00, 3:00, 2:00, 1:00, and 4:00. This suggests that CO levels tend to peak during late night and early morning hours

PROBLEM STATEMENT-8:

HOW DO POLLUTION LEVELS DIFFER BETWEEN WEEKDAYS AND WEEKENDS?



The boxplots compare pollution levels between weekdays and weekends for different pollutants (O3, CO, SO2, and NO2). All pollutants show statistically significant differences between weekdays and weekends (p-values < 0.05). Notably, O3 (Ozone) levels tend to be higher on weekends, while other pollutants generally show higher levels on weekdays.

PROBLEM STATEMENT-9:

ARE THERE SPECIFIC DAYS OR EVENTS ASSOCIATED WITH UNUSUALLY HIGH LEVELS OF POLLUTION?

The analysis identified numerous high pollution days (AQI > 100 for any pollutant). The top 5 days with the highest O3 AQI are listed, with the highest recorded on August 17, 2003, with an O3 AQI of 237.

These findings provide insights into air quality patterns in Arizona, highlighting differences between urban and rural areas, seasonal variations, and the frequency of high pollution days.

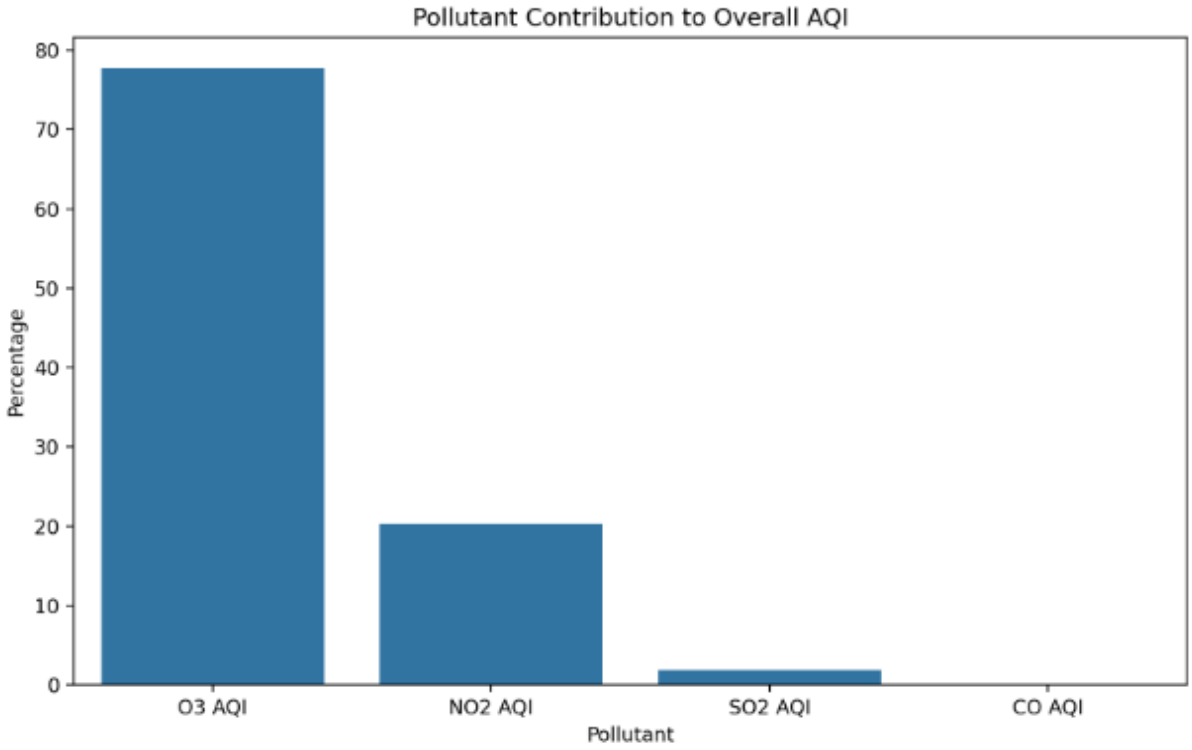
Number of high pollution days: 19656

Top 5 high pollution days:

	Date	O3 AQI	CO AQI	SO2 AQI	NO2 AQI
76303	2003-08-17T00:00:00.000	237	10	0	41
19269	2000-06-10T00:00:00.000	228	2	24	21
195157	2008-06-27T00:00:00.000	228	11	6	41
18285	2000-06-10T00:00:00.000	227	9	41	42
91481	2003-05-31T00:00:00.000	226	6	4	33

PROBLEM STATEMENT-10:

WHICH POLLUTANTS CONTRIBUTE THE MOST TO THE OVERALL AQI, AND DOES THIS VARY BY LOCATION?



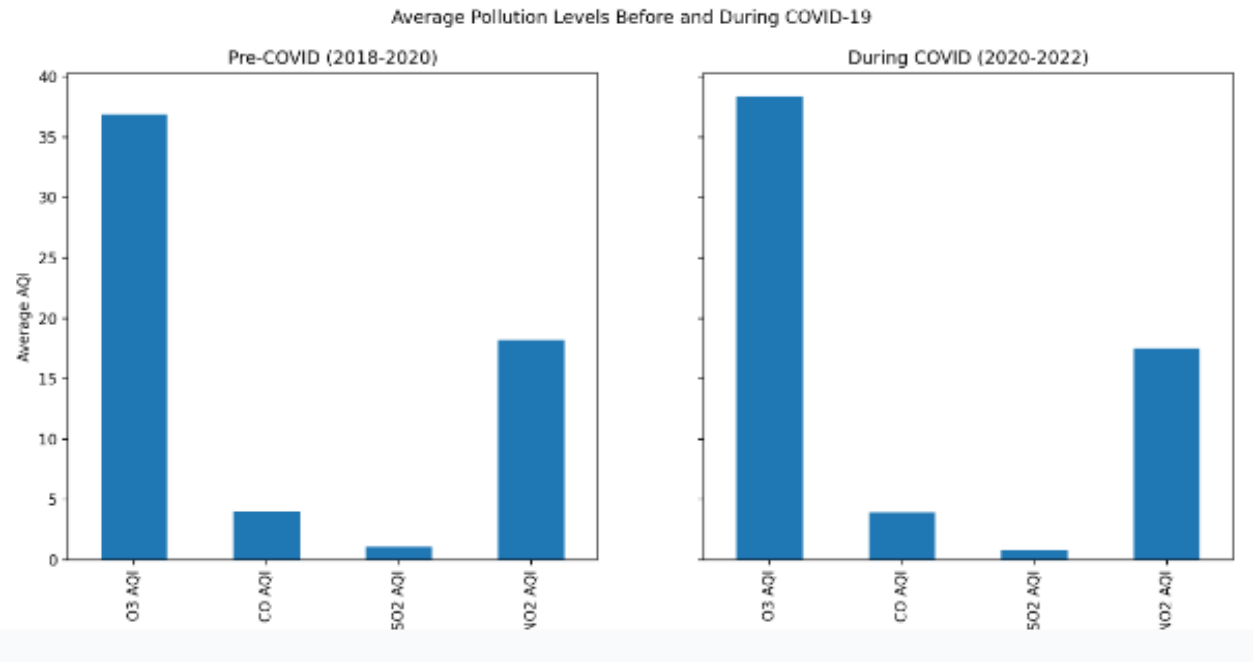
Overall, Ozone (O3) is the primary contributor to the AQI, accounting for about 77.7% of the highest AQI values, followed by Nitrogen Dioxide (NO2) at 20.3%. This varies by location:

- Non-urban areas ("Not in a city"): O3 dominates (92.9%)
- Los Angeles: O3 (63.7%) and NO2 (36.2%) are main contributors
- New York: NO2 (47.8%) and O3 (47.4%) contribute almost equally
- Phoenix: O3 (62.3%) and NO2 (37.7%) are the main contributors

PROBLEM STATEMENT-11:

HOW HAVE POLLUTION LEVELS CHANGED DURING THE COVID-19 PANDEMIC PERIOD?

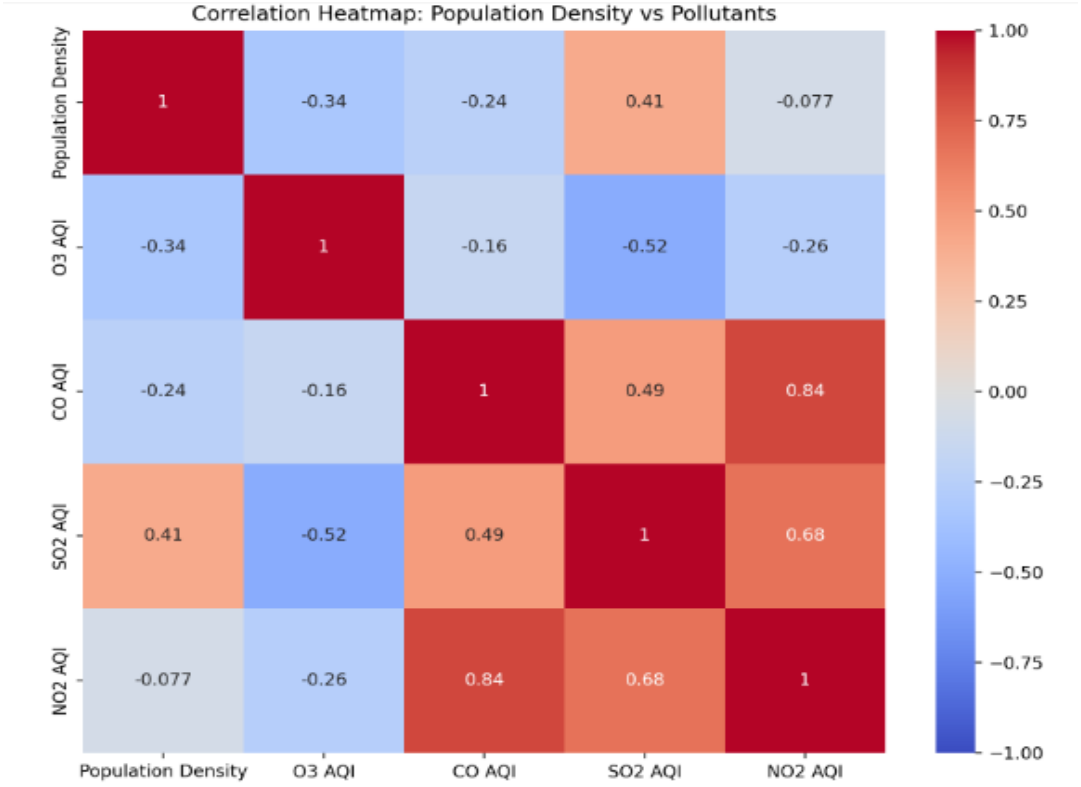
Comparing the pre-COVID period (2018-2020) to the COVID period (2020-2022):



There was a slight increase in O3 AQI (from 36.86 to 38.39) and slight decreases in CO, SO2, and NO2 AQI levels during the COVID period. The changes are relatively small, suggesting that the pandemic had a limited impact on overall air quality in the dataset.

PROBLEM STATEMENT-12:

IS THERE A CORRELATION BETWEEN POPULATION DENSITY AND POLLUTION LEVELS ACROSS THE CITIES?



This heat map visualizes the correlation matrix, highlighting the relationships between population density and various pollutants. A positive correlation indicates that as population density increases, the pollutant level tends to increase, and vice versa.

Correlation Coefficients:

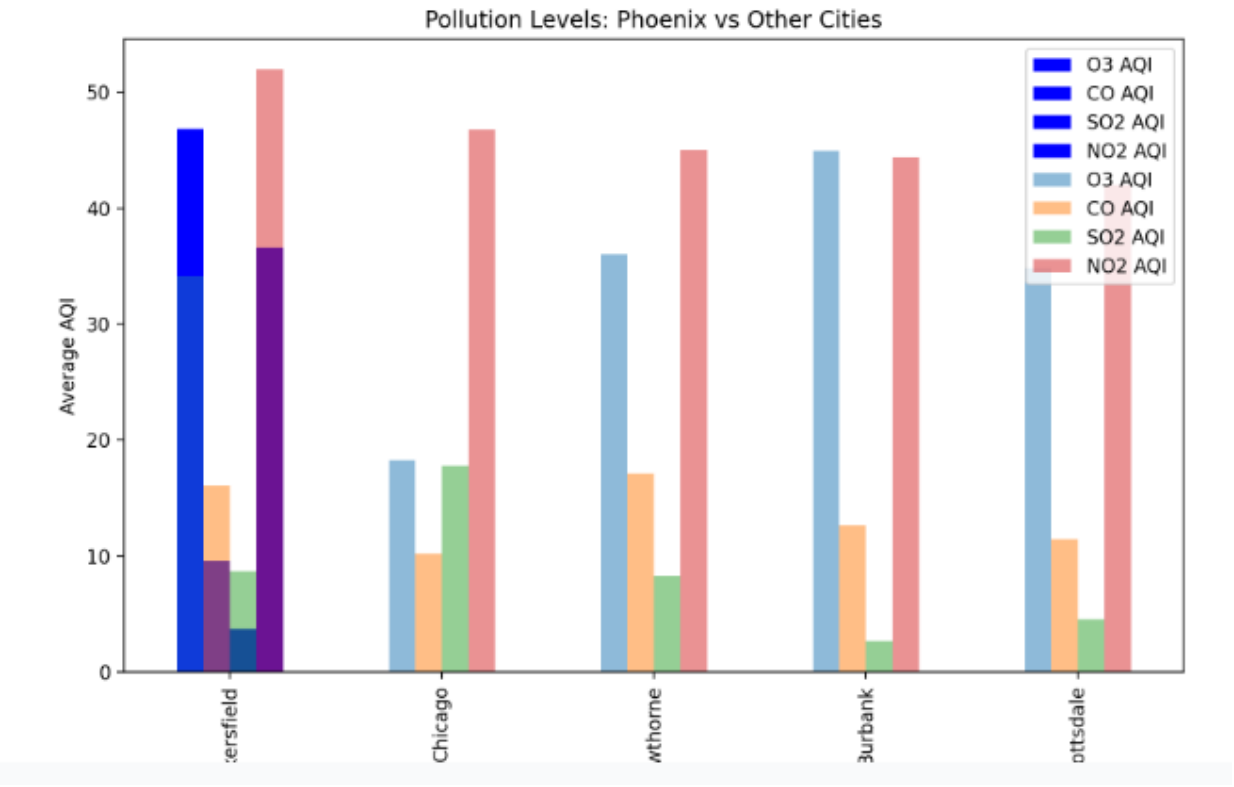
- O3 AQI: -0.34
- CO AQI: -0.24
- SO2 AQI: 0.41
- NO2 AQI: -0.08

The findings indicate that while there is some correlation between population density and pollution levels, it varies by pollutant. SO2 shows a stronger positive correlation with population density, suggesting that more densely populated areas might have higher SO2 levels, possibly due to industrial activities. Other pollutants like O3, CO, and NO2 show weaker correlations, indicating that factors other than population density might play a more significant role in their levels.

Larger cities like Chicago and Los Angeles tend to have higher NO2 levels, suggesting a possible correlation between urban density and pollution levels.

PROBLEM STATEMENT-13:

HOW DO AIR QUALITY LEVELS IN PHOENIX COMPARE TO OTHER CITIES IN THE DATASET?



The above chart compares the pollution levels in Phoenix (blue bar) with other major cities (translucent bars). We can observe that:

- Phoenix has relatively high levels of O3 compared to other cities
- NO2 levels in Phoenix are lower than in some other major cities
- CO and SO2 levels in Phoenix are comparatively low

These visualizations provide a more intuitive understanding of the pollution patterns and comparisons we discussed earlier. They highlight the dominance of Ozone as a pollutant, the subtle changes during the COVID-19 period, the timing of CO pollution peaks, and how Phoenix's air quality compares to other cities.

PROBLEM STATEMENT-14:

WHAT ARE THE HEALTH RISKS ASSOCIATED WITH PROLONGED EXPOSURE TO THE POLLUTANT LEVELS RECORDED IN THIS DATASET?

Based on the AQI categories distribution:

	proportion
Good	82.4954088733
Moderate	14.5506406538
Unhealthy for Sensitive Groups	2.444192638
Unhealthy	0.4651239679
Very Unhealthy	0.044633867

- 82.5% of the time, the air quality is in the "Good" category, posing little to no risk.
- 14.6% of the time, it's "Moderate", which may pose a risk for extremely sensitive individuals.
- 2.4% of the time, it's "Unhealthy" for Sensitive Groups", potentially affecting people with respiratory or heart conditions.
- 0.47% of the time, it's "Unhealthy", which could affect the general population.
- 0.04% of the time, it's "Very Unhealthy", posing significant health risks.

Prolonged exposure to higher AQI levels, especially in the "Unhealthy" and above categories, can lead to respiratory issues, cardiovascular problems, and other health complications.

PROBLEM STATEMENT-15:

ARE THERE ANY CORRELATIONS BETWEEN METEOROLOGICAL DATA AND POLLUTION SPIKES?

Unfortunately, the dataset doesn't include meteorological data, so we can't perform this analysis. Typically, factors like temperature, wind speed, and humidity can significantly influence pollution levels, but we don't have that information in this dataset.