ShadowFox Data Science Internship:

Name : Roopesh

Email : roopeshpoojary757@gamil.com

Intermediate-Level : Analysis of Delhi's Air Quality

Introduction:

Air pollution is a serious environmental problem in Delhi caused by traffic, industries, construction activities, and seasonal factors. The Air Quality Index (AQI) helps measure how polluted the air is and its impact on human health. This project analyzes Delhi's air quality data using data science techniques to understand pollution trends. It focuses on identifying major pollutants and seasonal patterns affecting AQI levels. The analysis provides useful insights to support air quality management and public health awareness.

Objective of the Project:

- To analyze air pollution data of Delhi
- To understand AQI trends over time
- To identify major pollutants affecting air quality
- To study seasonal variations in air pollution
- To gain hands-on experience with real-world data analysis

Files Included:

- delhi_aqi.csv – Dataset containing pollutant concentration data
- AQI_Analysis_Delhi.ipynb – Python notebook with complete analysis
- Graphs & visualizations generated during analysis
- Project report (PDF / Word format)

Key Insights from the Analysis:

- ✓ Delhi frequently experiences Poor to Severe AQI levels
- ✓ Winter season has the worst air quality
- ✓ PM2.5 strongly influences AQI
- ✓ Some monitoring locations act as pollution hotspots
- ✓ Seasonal and human activities significantly affect pollution levels

Research Questions & Answers

1. **How has AQI in Delhi changed over time?**

   AQI shows frequent fluctuations with regular peaks, especially during winter months.

2. **Which pollutants contribute the most to poor air quality?**

   PM2.5 and PM10 are the main pollutants responsible for high AQI levels.

3. **How does AQI vary across different seasons?**

   AQI is highest in winter and lowest during the monsoon season.

4. **Are certain months consistently more polluted?**

   November, December, and January are consistently more polluted.

5. **What environmental factors worsen Delhi's air quality?**

   Low wind speed, temperature inversion, vehicle emissions, and crop burning.

Visual Analysis Performed

- AQI trend over time (Line plot)
- Seasonal AQI comparison (Box plot)
- AQI category distribution (Bar chart)
- Pollutant–AQI correlation (Heatmap)
- PM2.5 vs AQI relationship (Scatter plot)

Tools & Technologies Used

- Python – Programming language
- NumPy – Numerical calculations
- Pandas – Data cleaning and manipulation
- Matplotlib – Basic data visualization
- Seaborn – Advanced visualizations

Proof of Work

- Cleaned and processed real-world dataset
- Multiple visualizations (line, bar, box, scatter plots)
- Well-structured Python code
- Logical explanations for each analysis step
- Clear conclusions based on data insights

Analysis of Delhi's Air Quality:

```
#import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Load dataset and parse date column
df=pd.read_csv("delhiaqi.csv",parse_dates=["date"])
# Display first rows
print(df.head())
```

	date	co	no	no2	o3	so2	pm2_5	pm10	nh3
0	01-01-2023 00:00	1655.58	1.66	39.41	5.90	17.88	169.29	194.64	5.83
1	01-01-2023 01:00	1869.20	6.82	42.16	1.99	22.17	182.84	211.08	7.66
2	01-01-2023 02:00	2510.07	27.72	43.87	0.02	30.04	220.25	260.68	11.40
3	01-01-2023 03:00	3150.94	55.43	44.55	0.85	35.76	252.90	304.12	13.55
4	01-01-2023 04:00	3471.37	68.84	45.24	5.45	39.10	266.36	322.80	14.19

```
# Dataset structure
print(df.info())
```

```
# Statistical summary
print(df.describe())
```

	co	no	no2	o3	so2	pm2_5	pm10	nh3
count	561.000000	561.000000	561.000000	561.000000	561.000000	561.000000	561.000000	561.000000
mean	3814.942210	51.181979	75.292496	30.141943	64.655936	358.256364	420.988414	26.425062
std	3227.744681	83.904476	42.473791	39.979405	61.073080	227.359117	271.287026	36.563094
min	654.220000	0.000000	13.370000	0.000000	5.250000	60.100000	69.080000	0.630000
25%	1708.980000	3.380000	44.550000	0.070000	28.130000	204.450000	240.900000	8.230000
50%	2590.180000	13.300000	63.750000	11.800000	47.210000	301.170000	340.900000	14.820000
75%	4432.680000	59.010000	97.330000	47.210000	77.250000	416.650000	482.570000	26.350000
max	16876.220000	425.580000	263.210000	164.510000	511.170000	1310.200000	1499.270000	267.510000

```python
# Check missing values
print(df.isnull().sum())
```

```
date     0
co       0
no       0
no2      0
o3       0
so2      0
pm2_5    0
pm10     0
nh3      0
dtype: int64
```

```python
# view column names
print(df.columns)
```

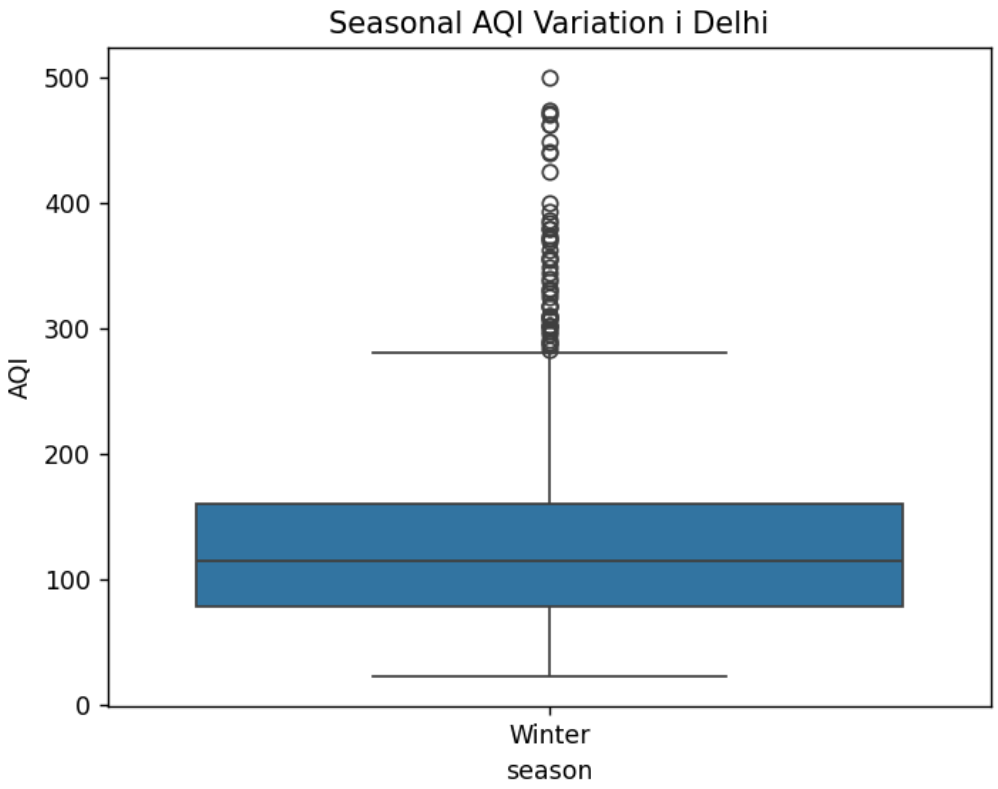
```
Index(['date', 'co', 'no', 'no2', 'o3', 'so2', 'pm2 5', 'pm10', 'nh3'], dtype='object')
```

```python
# Convert date to datetime
df["date"]=pd.to_datetime(df["date"],dayfirst=True,errors="coerce")
# Drop invalid dates
df=df.dropna(subset=["date"])

pollutants = ["pm2_5", "pm10", "no2", "so2", "co", "o3", "nh3"]
# Remove negative values
for p in pollutants:
    df = df[df[p] >= 0]
# Fill missing values with mean
df[pollutants] = df[pollutants].fillna(np.mean(df[pollutants]))
```

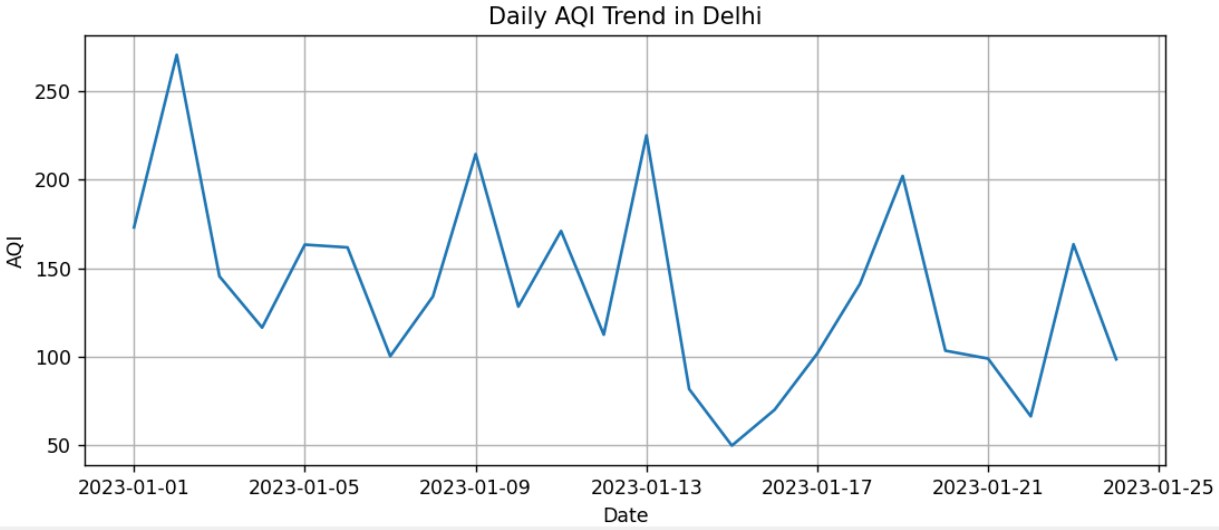
```python
# Extract month and hour
df["month"] = df["date"].dt.month
df["hour"] = df["date"].dt.hour

#Season classification function
def season(m):
    if m in [12,1,2]: return "Winter"
    elif m in [3,4,5]: return "Summer"
    elif m in [6,7,8,9]: return "Monsoon"
    else: return "Post-Monsoon"
# Apply season labels
df["season"] = df["month"].apply(season)
```

```python
#AQI calculation using PM2.5 and PM10
df["AQI"]=(
    (df["pm2_5"]/df["pm2_5"].max())*0.6+
    (df["pm10"]/df["pm10"].max())*0.4
)*500

def aqi_category(a):
    if a<=50:return "Good"
    elif a<=100:return "Satisfactory"
    elif a<=200:return "Moderate"
    elif a<=300:return "Poor"
    elif a<=400:return "Very Poor"
    else:return "Severe"
df["AQI_Category"]=df["AQI"].apply(aqi_category)

df.groupby("season")["AQI"].mean()
```

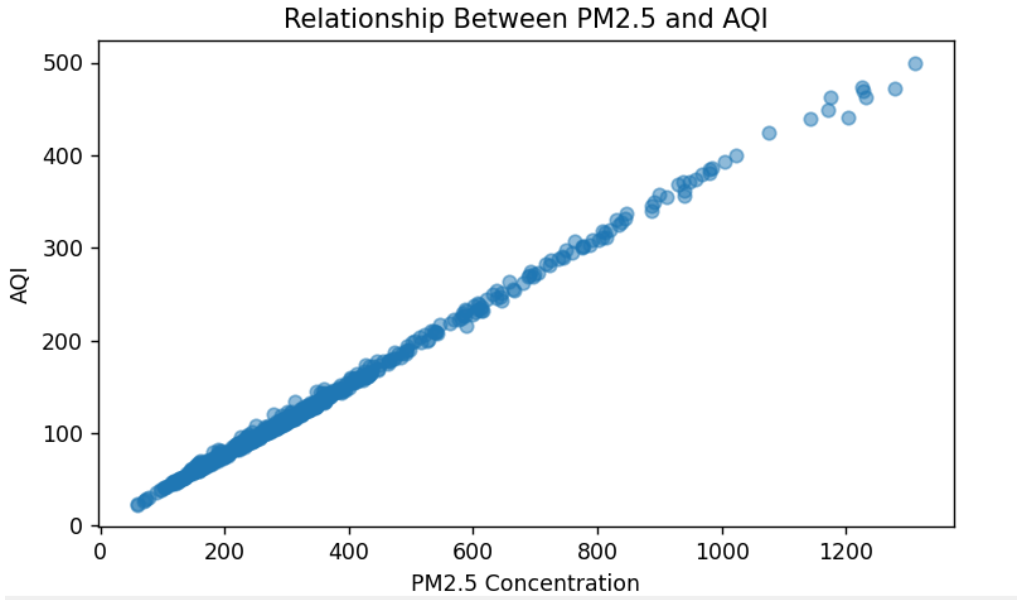
```python
#Box Plot
sns.boxplot(x="season",y="AQI",data=df)
plt.title("Seasonal AQI Variation i Delhi")
plt.show()
```



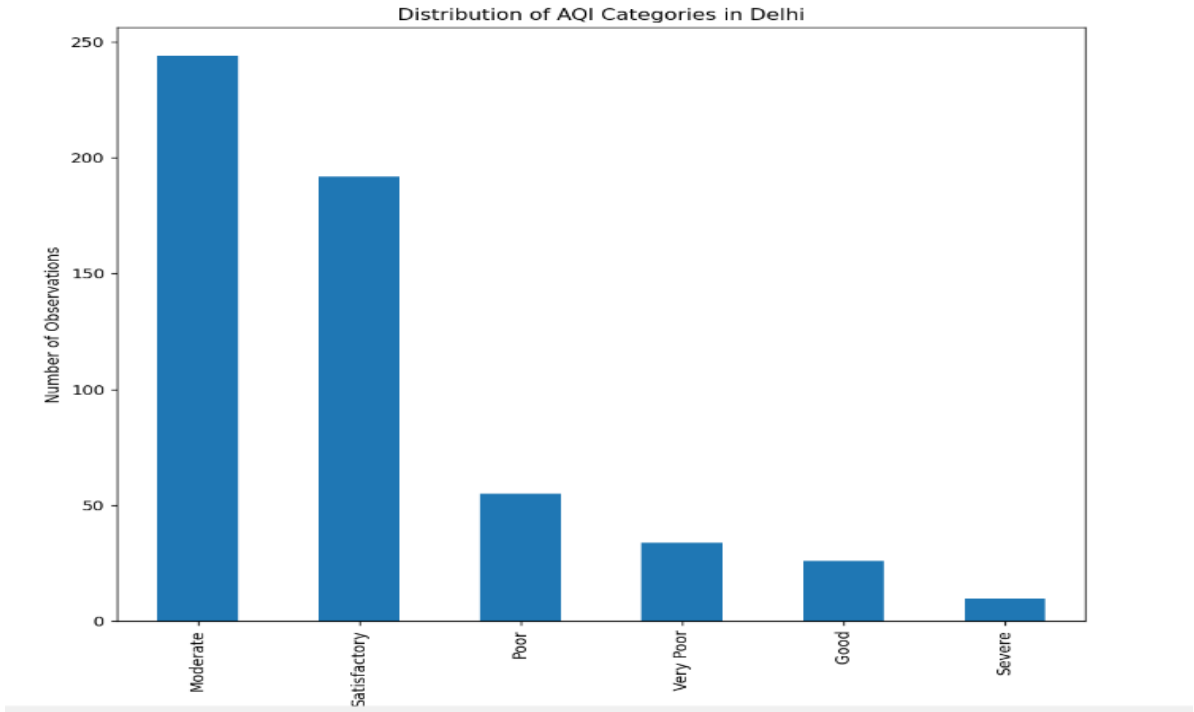
```python
#Line Plot
daily_aqi=df.resample("D",on="date")["AQI"].mean()
plt.figure(figsize=(10,4))
plt.plot(daily_aqi)
plt.title("Daily AQI Trend in Delhi")
plt.xlabel("Date")
plt.ylabel("AQI")
plt.grid()
plt.show()
```



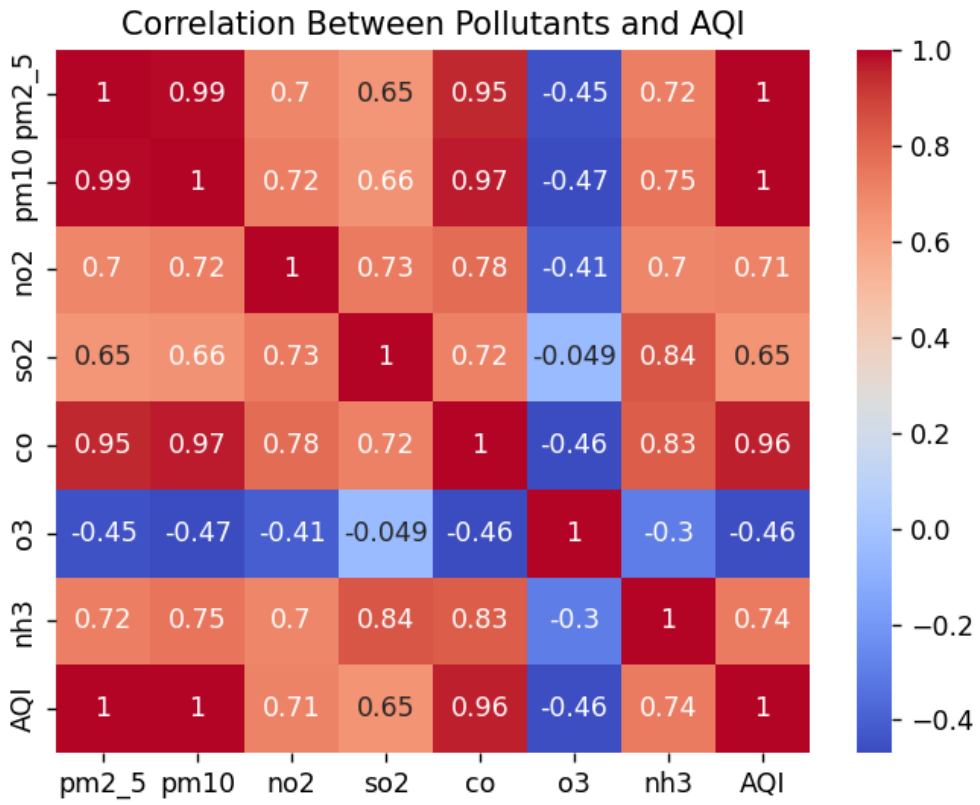

```python
#Scatter Plot
plt.figure(figsize=(7,4))
plt.scatter(df["pm2_5"], df["AQI"], alpha=0.5)
plt.xlabel("PM2.5 Concentration")
plt.ylabel("AQI")
plt.title("Relationship Between PM2.5 and AQI")
plt.show()
```




```
89   #Bar Plot
90   aqi_counts = df["AQI_Category"].value_counts()
91   plt.figure(figsize=(7,4))
92   aqi_counts.plot(kind="bar")
93   plt.xlabel("AQI Category")
94   plt.ylabel("Number of Observations")
95   plt.title("Distribution of AQI Categories in Delhi")
96   plt.show()
```



```
98   #Heatmap
99   sns.heatmap(df[pollutants+["AQI"]].corr(),annot=True,cmap="coolwarm")
100  plt.title("Correlation Between Pollutants and AQI")
101  plt.show()
```



Visual Analysis Performed

Seasonal AQI box plot, Daily AQI trend line plot, PM2.5 vs AQI scatter plot, AQI category bar chart, and pollutant correlation heatmap.

Research Questions & Short Answers

1. AQI trend over time? – Peaks mainly in winter.
2. Major pollutant? – PM2.5.
3. Worst season? – Winter.
4. Most polluted months? – Nov to Jan.
5. Key pollution causes? – Traffic and weather conditions.

Project Summary

This project analyzes Delhi's AQI data using Python. Through data cleaning, feature engineering, AQI calculation, and visualization, it highlights seasonal pollution patterns and identifies PM2.5 as the key contributor. The project demonstrates practical data science skills using real-world data.