**Title: Analysis of Air Quality Data in Tamil Nadu**

**1. Introduction:**

🡪Briefly introduce the project and its objectives.

**2. Objectives:**

**State the project objectives:**

🡪 Analyze air quality data from monitoring stations in Tamil Nadu.

🡪 Gain insights into air pollution trends.

🡪 Identify areas with high pollution levels.

🡪 Develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels.

**3. Data Source:**

🡪 Mention the source of the air quality data (e.g., dataset location or organization).

**4. Analysis Approach:**

**Describe the analysis approach:**

🡪 Data loading and preprocessing.

🡪 Exploratory data analysis (EDA).

🡪 Visualization of key air quality indicators.

🡪 Building predictive models.

**5. Data Preprocessing:**

**Explain the data preprocessing steps:**

🡪Loading the dataset.

🡪 Handling missing values.

🡪 Splitting the data into features and target variables.

**6. Data Analysis and Visualization:**

**Present the code implementation and visualizations:**

🡪 Code for loading and preprocessing the data.

🡪 Visualizations to understand the data distribution:

🡪 Histogram of RSPM/PM10 levels.

🡪 Scatter plots for SO2 vs. RSPM/PM10 and NO2 vs. RSPM/PM10.

**7. Predictive Modeling:**

**If applicable, explain the implementation of predictive models:**

🡪 Code for building a Linear Regression model.

🡪 Mention the model evaluation metrics (e.g., R-squared and RMSE).

🡪 Interpret the model's performance and discuss whether it's suitable for the dataset.

**8. Interpretation and Insights:**

**Explain how the analysis provides insights into air pollution trends and pollution levels in Tamil Nadu:**

🡪 Discuss the patterns observed in visualizations.

🡪 Highlight any correlations between variables.

🡪 Conclude on what the data analysis reveals about air quality in Tamil Nadu.

**9. Conclusion:**

🡪 Summarize the findings and project outcomes.

🡪 Reflect on the project's objectives and whether they were achieved.

**10. References:**

**Include example outputs of data analysis and visualization:**

- Display the summary statistics of the dataset.

- Show sample visualizations and plots of air quality indicators

**Program:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score

from sklearn.model\_selection import train\_test\_split

# Load the air quality data

data = pd.read\_csv("C:/Users/balur/Downloads/pollution.csv")

# Data Preprocessing

data = data.dropna(subset=['SO2', 'NO2', 'RSPM/PM10'])

# Split the data into features (SO2, NO2) and target variable (RSPM/PM10)

x = data[['SO2', 'NO2']]

y = data['RSPM/PM10']

# Split the data into training and testing sets

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=42)

# Perform an initial analysis of the dataset

# For example, calculate basic statistics

summary\_statistics = data.describe()

print("Summary Statistics:")

print(summary\_statistics)

# Visualize the distribution of RSPM/PM10 levels

plt.figure(figsize=(8, 6))

sns.histplot(data=data, x='RSPM/PM10', bins=20, kde=True)

plt.title('Distribution of RSPM/PM10 Levels')

plt.xlabel('RSPM/PM10 Levels')

plt.ylabel('Frequency')

plt.show()

# Create a scatter plot of SO2 vs. RSPM/PM10

plt.figure(figsize=(8, 6))

sns.scatterplot(data=data, x='SO2', y='RSPM/PM10')

plt.title('SO2 vs. RSPM/PM10')

plt.xlabel('SO2 Levels')

plt.ylabel('RSPM/PM10 Levels')

plt.show()

# Create a scatter plot of NO2 vs. RSPM/PM10

plt.figure(figsize=(8, 6))

sns.scatterplot(data=data, x='NO2', y='RSPM/PM10')

plt.title('NO2 vs. RSPM/PM10')

plt.xlabel('NO2 Levels')

plt.ylabel('RSPM/PM10 Levels')

plt.show()

# Build and evaluate a Linear Regression model for prediction

regr = LinearRegression()

regr.fit(x\_train, y\_train)

y\_pred = regr.predict(x\_test)

# Evaluate the model

r2 = r2\_score(y\_test, y\_pred)

print("Linear Regression - R-squared:", r2)

# Optional: You can further analyze the relationships between variables and explore more advanced models.

# Finally, summarize your findings and conclusions based on the analysis and visualizations.

**Output:**

**Summary Statistics:**

**Stn Code SO2 NO2 RSPM/PM10 PM 2.5**

**count 2862.000000 2862.000000 2862.000000 2862.000000 0.0**

**mean 475.484277 11.506988 22.135220 62.437456 NaN**

**std 277.741701 5.050855 7.133291 31.277419 NaN**

**min 38.000000 2.000000 5.000000 12.000000 NaN**

**25% 238.000000 8.000000 17.000000 41.000000 NaN**

**50% 366.000000 12.000000 21.500000 55.000000 NaN**

**75% 764.000000 15.000000 25.000000 78.000000 NaN**

**max 773.000000 49.000000 71.000000 269.000000 NaN**

**Linear Regression - R-squared: 0.20658507746336507**

**Visualization:**







