Air Quality Analysis and Prediction in Tamil Nadu - Guidelines

Documentation

Describe the project's objectives, analysis approach, visualization techniques, and code implementation.
Include example outputs of data analysis and visualizations.
Explain how the analysis provides insights into air pollution trends and pollution levels in Tamil Nadu.

Submission

Share the GitHub repository link containing the project's code and dataset.
Provide instructions on how to replicate the analysis, load the dataset, perform calculations, and create visualizations using Python.

Problem Definition:

The project aims to analyze and visualize air quality data from monitoring stations in Tamil Nadu. The objective is to gain insights into air pollution trends, identify areas with high pollution levels, and develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels.

Design Thinking:

Project Objectives:

Define objectives such as analyzing air quality trends, identifying pollution hotspots, and building a predictive model for RSPM/PM10 levels.

Design Thinking:

Project Objectives: Define objectives such as analyzing air quality trends, identifying pollution hotspots, and building a predictive model for RSPM/PM10 levels.

Analysis Approach: Plan the steps to load, preprocess, analyze, and visualize the air quality data.

Visualization Selection: Determine visualization techniques (e.g., line charts, heatmaps) to effectively represent air quality trends and pollution levels.

Import and load the data
Import necessary packages: Here I have imported packages needed for preprocessing and modelling

```python
import pandas as pd
import os
import zipfile
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.graph_objs as go
import numpy as np
from plotly.offline import init_notebook_r
import missingno as msno
from sklearn.impute import KNNImputer
from sklearn import preprocessing
import pylab
import scipy.stats as stats
from scipy.special import boxcox1p
import pylab
import scipy.stats as stats
%matplotlib inline
# Transformation and modelling packages
from sklearn.model_selection import train_
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import Standar
from sklearn.ensemble import RandomForestC
from sklearn.linear_model import LogisticR
from sklearn.naive_bayes import GaussianNB
```

Histogram using given dataset:

A histogram is a graph that shows the frequency of numerical data using rectangles. The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (the amount, or how often that variable appears).

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv("C:\\Users\\MUJAHID\\Downloads\\cpcb_dly_aq_tamil_nadu-2014.csv")
plt.figure(figsize=(10,6))
sns.histplot(data=df)
plt.xlabel("Avg air quality")
plt.ylabel("RSPM/PM 10 level")
plt.show()
```

THANK YOU