**IBM NAANMUDHALVAN PHASE-3**

**Air Quality Analysis & prediction**

**Using Data Science**

Air Quality Analysis:

Air quality is a measure of how clean or polluted the air is. Monitoring air quality is important because polluted air can be bad for our health—and the health of the environment. Air quality is measured with the Air Quality Index, or AQI. The AQI works sort of like a thermometer that runs from 0 to 500 degrees.

Air Quality Analysis & prediction:

Air quality analysis and prediction using data science techniques involves leveraging data to understand historical patterns, identify trends, and build predictive models for forecasting air quality levels.

By 2050, outdoor air pollution (particulate matter and ground-level ozone) is projected to become the top cause of environmentally related deaths worldwide

Air pollution contributes to a wide variety of adverse health effects. The six most common air pollutants are called “criteria” air pollutants and include carbon monoxide, lead, ground-level ozone, particulate matter, nitrogen dioxide, and sulfur dioxide.

These steps and utilizing appropriate data science techniques, you can analyze historical air quality data, build accurate predictive models, and contribute to ongoing efforts to monitor and improve air quality.

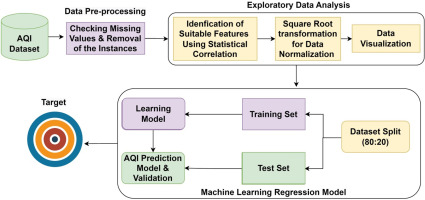
Tools We are Using for this Project

Of course, we're using Python to build our project

Air quality analysis and prediction using data science involves various tools and technologies to collect, process, analyze, and visualize data.

1. Jupyter environment (Jupyter Lab or Jupyter notebook) – for experimenting with our project.
2. Pandas – for loading data as a dataframe and wrangling the data.
3. NumPy and SciPy - For performing some data manipulation and analysis.
4. Scikit-Learn, TensorFlow, and PyTorch - The are used for building machine learning models to predict air quality based on historical data.
5. Seaborn, Matplotlib and Plotly Express – It used for creating static, interactive, and animated visualizations of air quality data.
6. Git- Git is used for version control, allowing collaborative work on code and data analysis projects.

The features of the dataset are:



**Exploratory Data Analysis (EDA):**

As you might know, EDA is the key to performing well as a data analyst or data scientist. It gives you first-hand information about the whole dataset, and it helps you understand all the relationships between the features in our dataset.

We are performing the three phases of EDA here which are:

1. Univariate Analysis.

2. Bivariate Analysis.

3. Multivariate Analysis

Firstly we are imporrting all the necessary libraries that we are using in this project. We also need to load the dataset into a dataframe so we can see all the features that are present in it.

import pandas as pd

import seaborn as sns

import matplotlib. pyplot as plt

import plotly.express as px

import numpy as np

from scipy.stats import iqr

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

df = pd.read\_csv('../input/air-quality-data-set/AirQuality.csv',sep=';')

df.head()

First, since the air quality is based on the total amount prediction have spent, we'll add the amount spent on the product:

df.info()

df.drop('NMHC(GT)', axis=1, inplace=True)

df[ar.columns[2:13]][(lower\_outliers | upper\_outliers)].info()

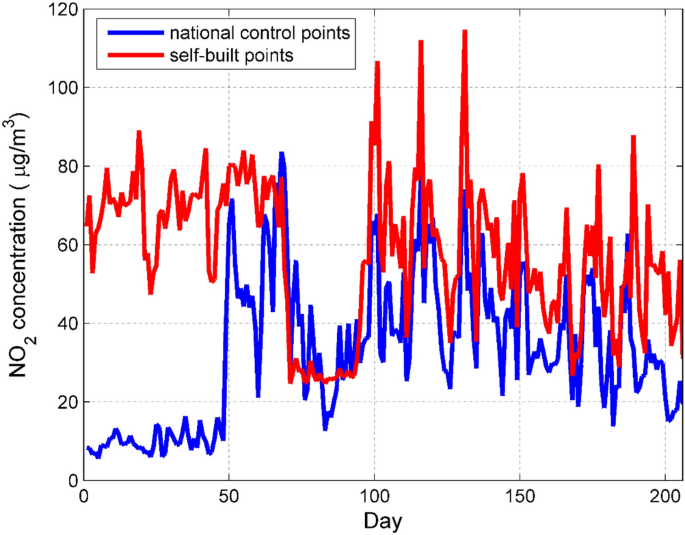
df\_filt.info()

**Univariate analysis:**

Univariate analysis entails evaluating a single feature in order to get insights about it. So, the initial step in performing EDA is to undertake univariate analysis, which includes evaluating descriptive or summary statistics about the feature.

sns.histplot(data=df, x="Day", bins = list(range(10, 150, 10)))

plt.title("National control points")



**Bivariate Analysis:**

The next step is to perform bivariate analysis. This involves comparing two attributes at the same time.

Bivariate analysis entails determining the correlation between two features, for example.

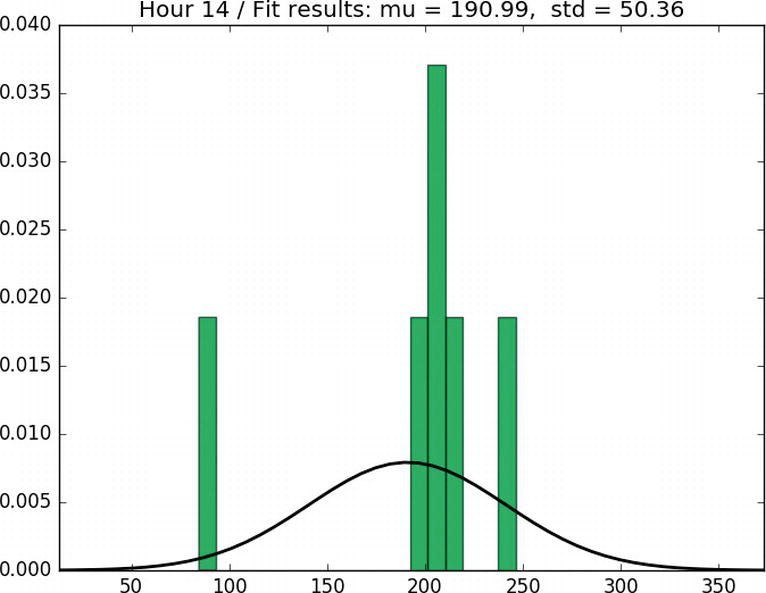
In our case, some of the bivariate analysis we'll perform in the project include observing the average total spent across different client age groups, determining a correlation between customer income and total amount spent, and so on, as shown below.

for i **in** df.columns[0:0.40]:

sns.boxplot(x=df[i])

plt.title('Boxplot of the sensors data')

plt.show()



**Multivariate Analysis:**

After you've completed univariate (analysis of single feature) and bivariate (analysis of two features) analysis, the last phase of EDA is to perform Multivariate Analysis. Multivariate Analysis consists of understanding the relationship between two or more variables.

fig = px.scatter(

data\_frame=df\_cut,

x = "Dataset Type",

y= "Number of publications",

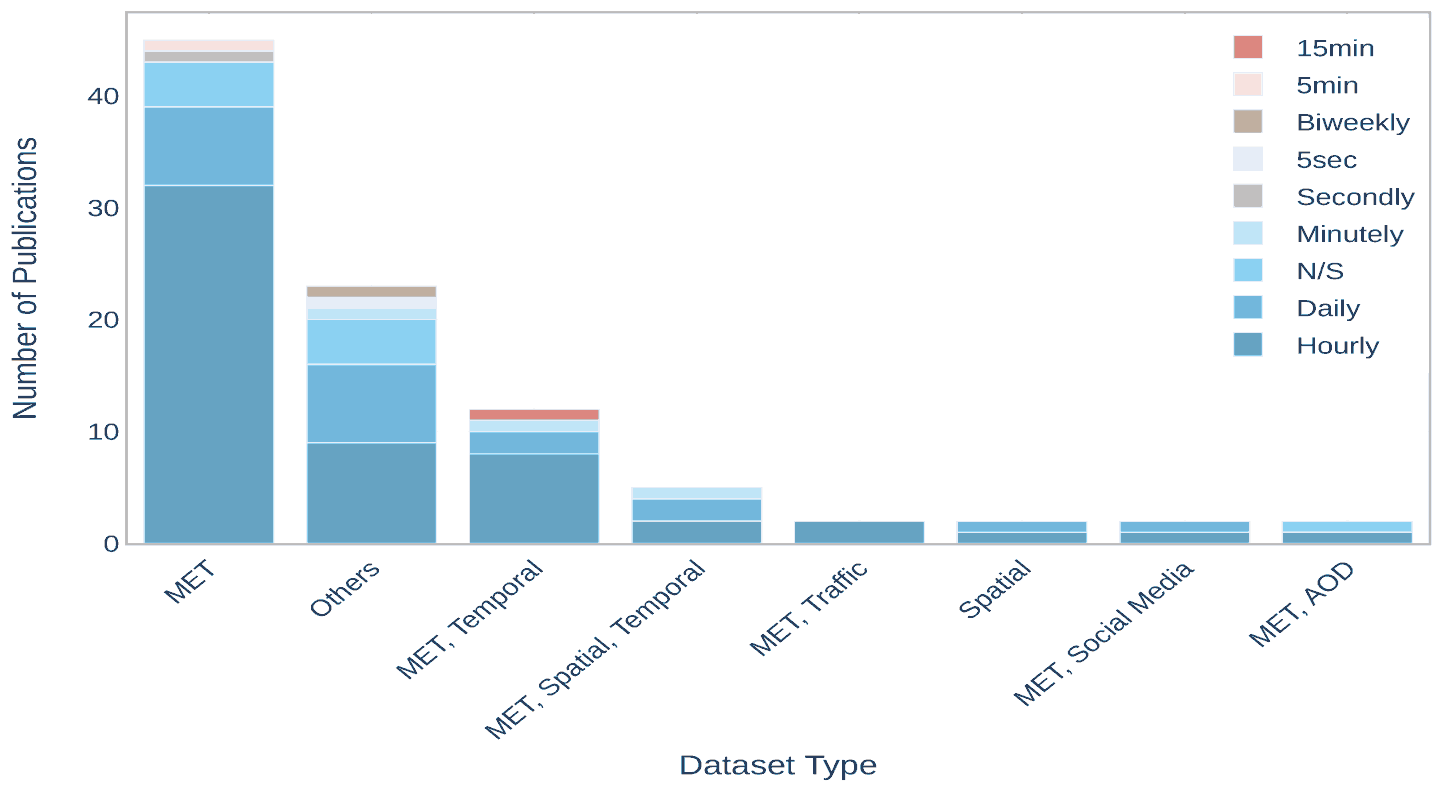
title = "Relationship between Dataset TypeVS Number of publications",

color = "publication",

height=40

)

fig.show()



Analysis of relationship between Dataset Type, Number of publication.We have done three phases of EDA for this air quality analysis and prediction using python.