JAYPEE INSTITUE OF INFORMATION TECHNOLOGY

SECTOR-62, NOIDA

OPEN-SOURCE SOFTWARE LAB

Global Air Pollution using using Linear Regression

PROJECT SYNOPSIS

Submitted to: Dr. Sonal CST

Submitted by:

Guneet Gandhi- 21103245

Ananya Sharma- 21103216

Vinayak Kanojia- 21103206

Naman Garg-21103244

ABSTRACT:

Air Pollution is contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere. Household combustion devices, motor vehicles, industrial facilities and forest fires are common sources of air pollution. Pollutants of major public health concern include particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulphur dioxide. Outdoor and indoor air pollution cause respiratory and other diseases and are important sources of morbidity and mortality.

LINEAR REGRESSION MODEL:

Linear regression is a statistical technique that aims to model the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to the observed data. In the context of our project, we are employing linear regression to analyse and predict global air pollution trends by examining how various factors contribute to changes in air quality.

Key Concepts and Components:

Dependent Variable (Y): In the context of our project, the dependent variable is air pollution, which is measured using parameters such as PM2.5 (Particulate Matter 2.5), PM10, nitrogen dioxide (NO2), sulphur dioxide (SO2), or other relevant air quality indicators. This variable represents what we are trying to predict or model.

Independent Variables (X): Independent variables are the factors that we believe influence air pollution levels. These can include industrial emissions, traffic density, meteorological conditions (e.g., temperature, humidity, wind speed), population density, urbanization rates, and more. Each of these variables serves as a potential predictor of air pollution.

Linear Regression Equation: The fundamental equation of linear regression is:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \varepsilon$$

Y: The predicted air pollution level.

β0: The intercept (constant) term.

β1, β2, ... βn: The coefficients of the independent variables, indicating the strength and direction of their influence.

X1, X2, ... Xn: The values of the independent variables.

ε: The error term, representing the difference between the observed and predicted values.

Simple Linear Regression: In simple linear regression, there is only one independent variable. It can help us understand the linear relationship between air pollution and a single predictor, for instance, examining how temperature affects air quality.

Multiple Linear Regression: Multiple linear regression extends the analysis to include multiple independent variables, allowing us to assess the combined impact of various factors on air pollution.

Dataset Features:

- **Country**: Name of the country
- **City**: Name of the city
- **AQI Value**: Overall AQI value of the city
- **AQI Category**: Overall AQI category of the city
- **CO AQI Value**: AQI value of Carbon Monoxide of the city
- **CO AQI Category**: AQI category of Carbon Monoxide of the city
- **Ozone AQI Value**: AQI value of Ozone of the city
- **Ozone AQI Category**: AQI category of Ozone of the city
- **NO2 AQI Value**: AQI value of Nitrogen Dioxide of the city
- **NO2 AQI Category**: AQI category of Nitrogen Dioxide of the city

- **PM2.5 AQI Value**: AQI value of Particulate Matter with a diameter of 2.5 micrometres or less of the city
- **PM2.5 AQI Category**: AQI category of Particulate Matter with a diameter of 2.5 micrometres or less of the city

This **dataset** provides **geolocated information** about the following **pollutants**:

1. **Nitrogen Dioxide [NO2]** : Nitrogen Dioxide is one of the several nitrogen oxides. It is introduced into the air by natural phenomena like entry from stratosphere or lighting. At the surface level, however, NO2 forms from cars, trucks and buses emissions, power plants and off-road equipment. Exposure over short periods can aggravate respiratory diseases, like asthma. Longer exposures may contribute to develoment of asthma and respiratory infections. People with asthma, children and the elderly are at greater risk for the health effects of NO2.
2. **Ozone [O3]** : The Ozone molecule is harmful for outdoor air quality (if outside of the ozone layer). At surface level, ozone is created by chemical reactions between oxides of nitrogen and volatile organic compounds (VOC). Differently from the good ozone located in the upper atmosphere, ground level ozone can provoke several health problems like chest pain, coughing, throat irritation and airway inflammation. Furthermore it can reduce lung function and worsen bronchitis, emphysema, and asthma. Ozone affects also vegetation and ecosystems. In particular, it damages sensitive vegetation during the growing season.
3. **Carbon Monoxide [CO]** : Carbon Monoxide is a colorless and odorless gas. Outdoor, it is emitted in the air above all by cars, trucks and other vehicles or machineries that burn fossil fuels. Such items like kerosene

and gas space heaters, gas stoves also release CO affecting indoor air quality.

Breathing air with a high concentration of CO reduces the amount of oxygen that can be transported in the blood stream to critical organs like the heart and brain. At very high levels, which are not likely to occur outdoor but which are possible in enclosed environments. CO can cause dizziness, confusion, unconsciousness and death.

4. **Particulate Matter [PM2.5]** : Atmospheric Particulate Matter, also known as atmospheric aerosol particles, are complex mixtures of small solid and liquid matter that get into the air. If inhaled they can cause serious heart and lungs problem. They have been classified as group 1 carcinogen by the International Agengy for Research on Cancer (IARC). PM10 refers to those particules with a diameter of 10 micrometers or less. PM2.5 refers to those particles with a diameter of 2.5 micrometers or less.

IMPLEMENTATION:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
df = pd.read_csv('global air pollution dataset.csv')
df.head()
```

Out[4]:

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
0	Russian Federation	Praskoveya	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
1	Brazil	Presidente Dutra	41	Good	1	Good	5	Good	1	Good	41	Good
2	Italy	Priolo Gargallo	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
3	Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
4	France	Punaauia	22	Good	0	Good	22	Good	0	Good	6	Good

ef = df.drop(['NO2 AQI Value','NO2 AQI Category','CO AQI Value','CO AQI Category'],axis=1, inplace= False)

ef.head()

Out[6]:

	Country	City	AQI Value	AQI Category	Ozone AQI Value	Ozone AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
0	Russian Federation	Praskoveya	51	Moderate	36	Good	51	Moderate
1	Brazil	Presidente Dutra	41	Good	5	Good	41	Good
2	Italy	Priolo Gargallo	66	Moderate	39	Good	66	Moderate
3	Poland	Przasnysz	34	Good	34	Good	20	Good
4	France	Punaauia	22	Good	22	Good	6	Good

ef.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23463 entries, 0 to 23462
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Country             23036 non-null  object
 1   City                23462 non-null  object
 2   AQI Value           23463 non-null  int64 
 3   AQI Category        23463 non-null  object
 4   Ozone AQI Value     23463 non-null  int64 
 5   Ozone AQI Category  23463 non-null  object
 6   PM2.5 AQI Value     23463 non-null  int64 
 7   PM2.5 AQI Category  23463 non-null  object
dtypes: int64(3), object(5)
memory usage: 1.4+ MB
```

ef.describe()

Out[8]:

	AQI Value	Ozone AQI Value	PM2.5 AQI Value
count	23463.000000	23463.000000	23463.000000
mean	72.010868	35.193709	68.519755
std	56.055220	28.098723	54.796443
min	6.000000	0.000000	0.000000
25%	39.000000	21.000000	35.000000
50%	55.000000	31.000000	54.000000
75%	79.000000	40.000000	79.000000
max	500.000000	235.000000	500.000000

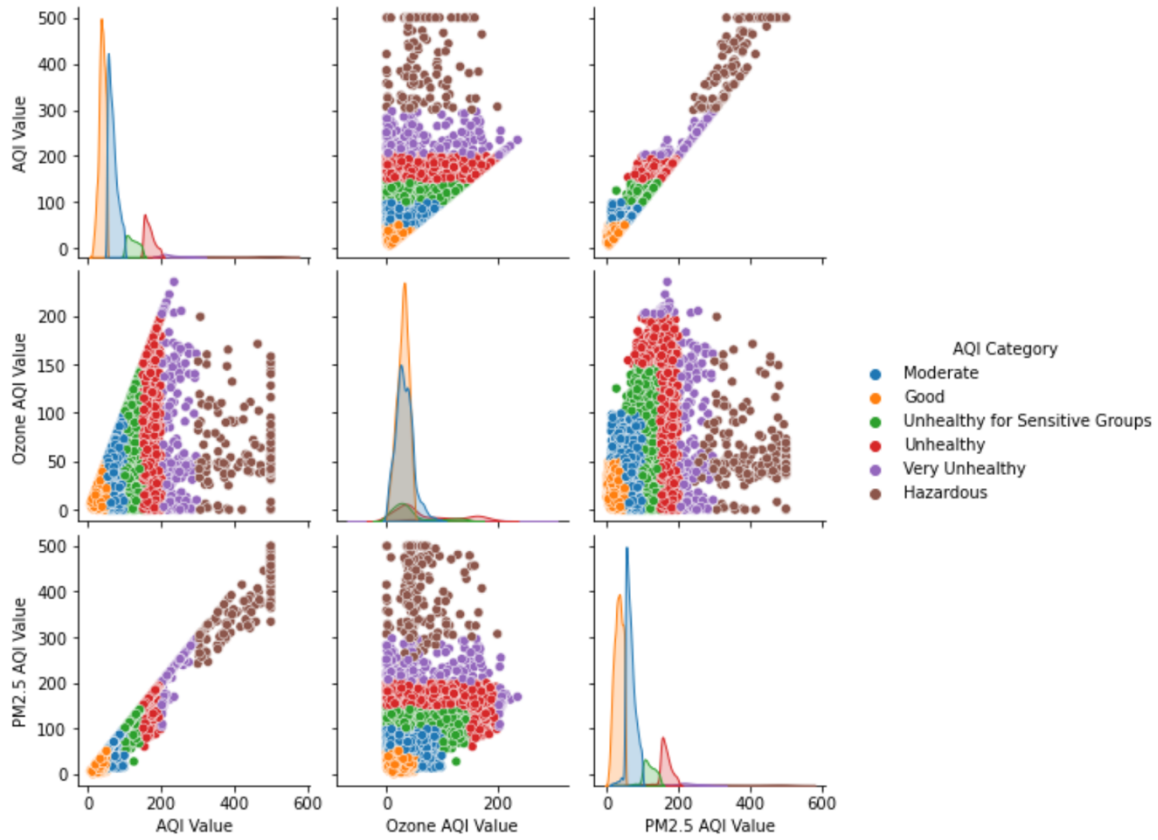
list(ef.columns)

```
Out[9]: ['Country',
         'City',
         'AQI Value',
         'AQI Category',
         'Ozone AQI Value',
         'Ozone AQI Category',
         'PM2.5 AQI Value',
         'PM2.5 AQI Category']
```

plt.figure(figsize=(12,12))

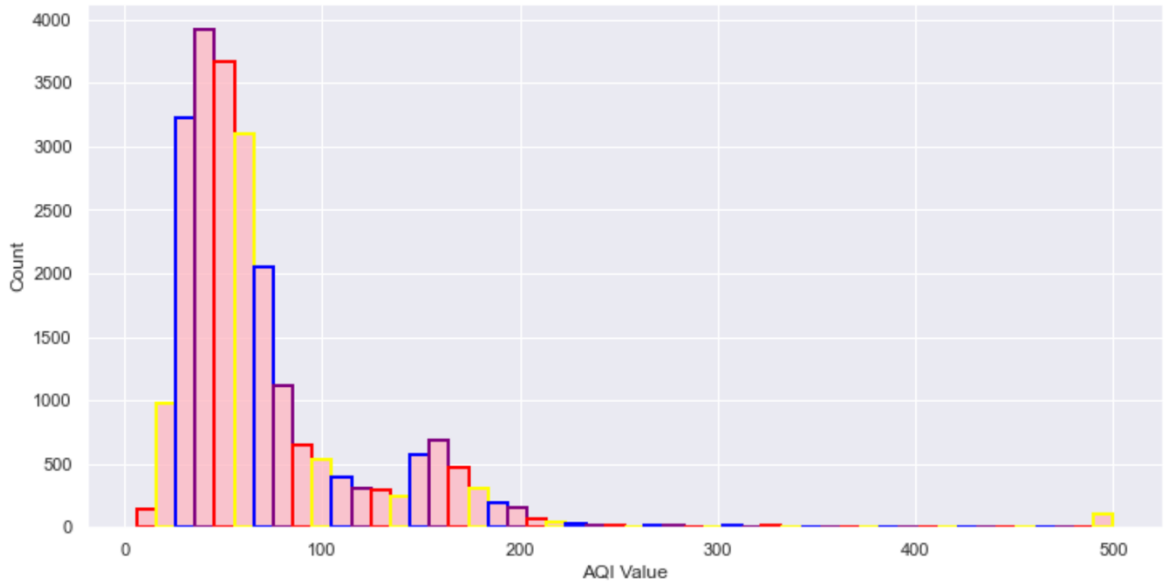
sns.pairplot(ef, hue='AQI Category')

```
<seaborn.axisgrid.PairGrid at 0x7f7d3daa67f0>
<Figure size 864x864 with 0 Axes>
```




```
plt.figure(figsize=(12,6))
sns.set_style(style='darkgrid')
edge_colors = ["red", "yellow", "blue", "purple"]
sns.histplot(ef['AQI Value'],bins = 50,color='lightpink',edgecolor=edge_colors,linewidth=2)
```

Out[22]: <AxesSubplot:xlabel='AQI Value', ylabel='Count'>



ef.corr()

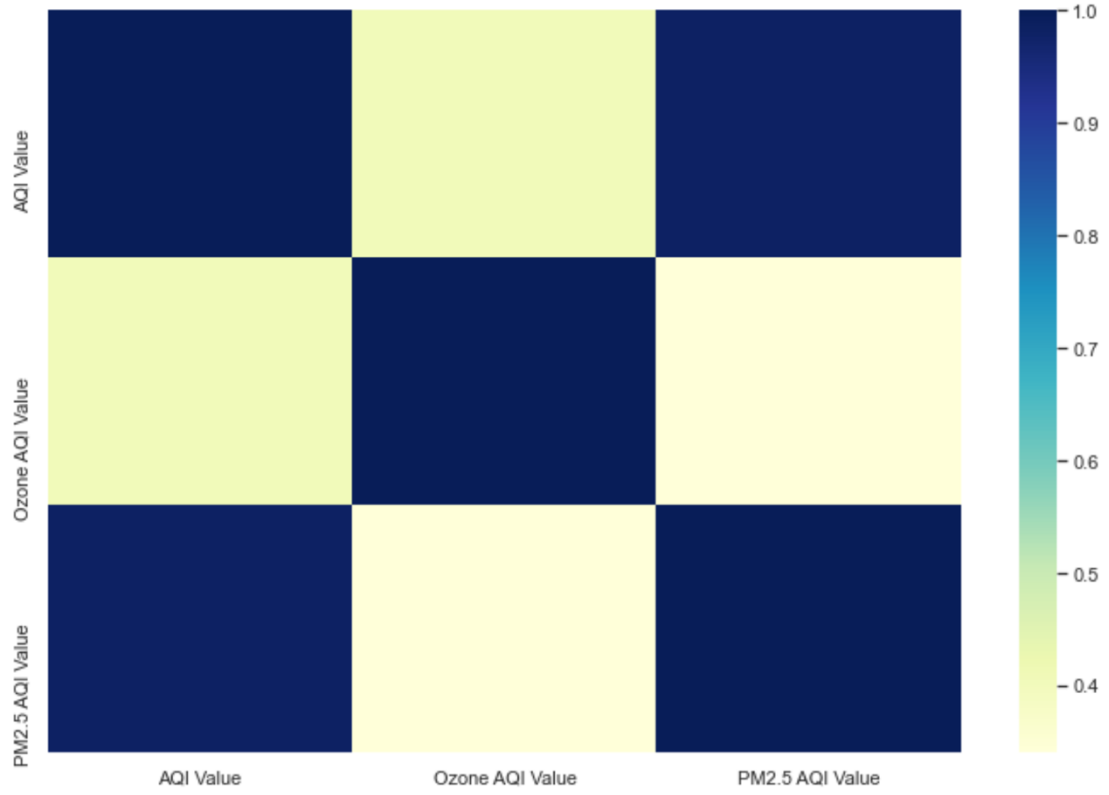
Out[23]:

	AQI Value	Ozone AQI Value	PM2.5 AQI Value
AQI Value	1.000000	0.405310	0.984327
Ozone AQI Value	0.405310	1.000000	0.339887
PM2.5 AQI Value	0.984327	0.339887	1.000000

plt.figure(figsize=(12,8))

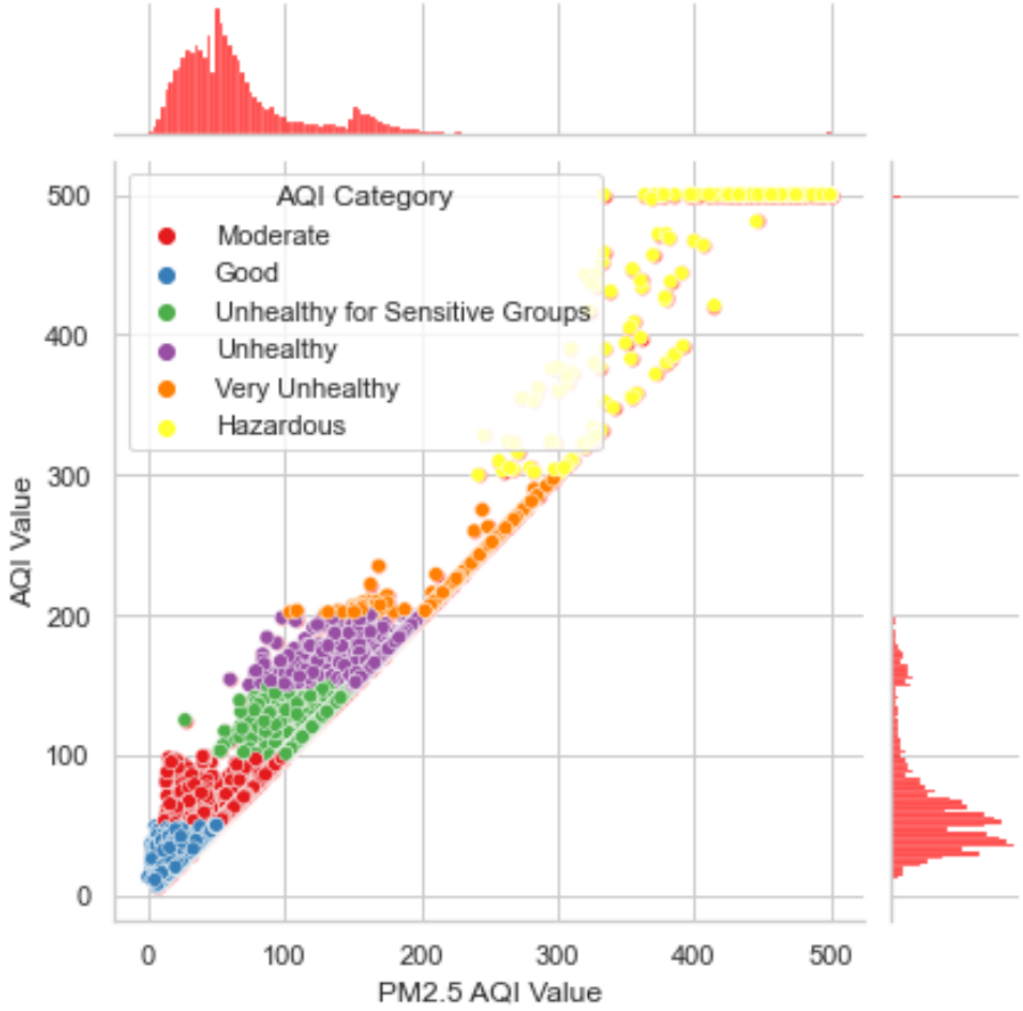
sns.heatmap(ef.corr(), cmap="YlGnBu")

```
<AxesSubplot:>
```

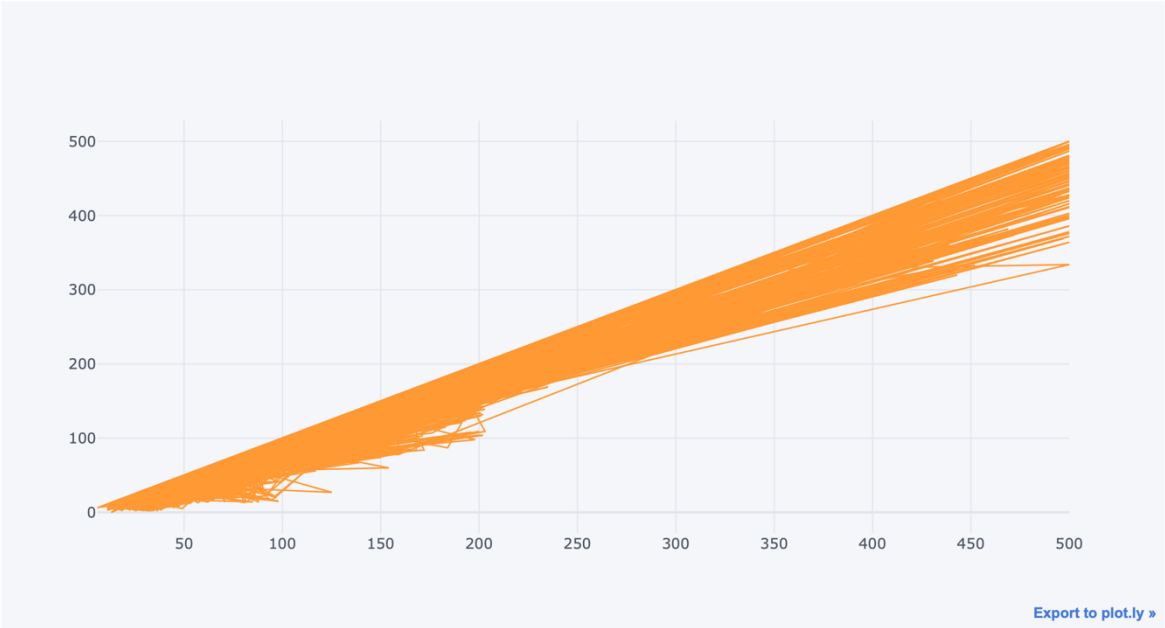



```
plt.figure(figsize=(12,8))

sns.set(style="whitegrid")

g = sns.jointplot(x=ef['PM2.5 AQI Value'], y=ef['AQI Value'], ratio=5, color='red')


sns.scatterplot(data=ef, x='PM2.5 AQI Value', y='AQI Value', hue='AQI Category', palette='Set1', ax=g.ax_joint)


plt.show()
```

```
<Figure size 864x576 with 0 Axes>
```

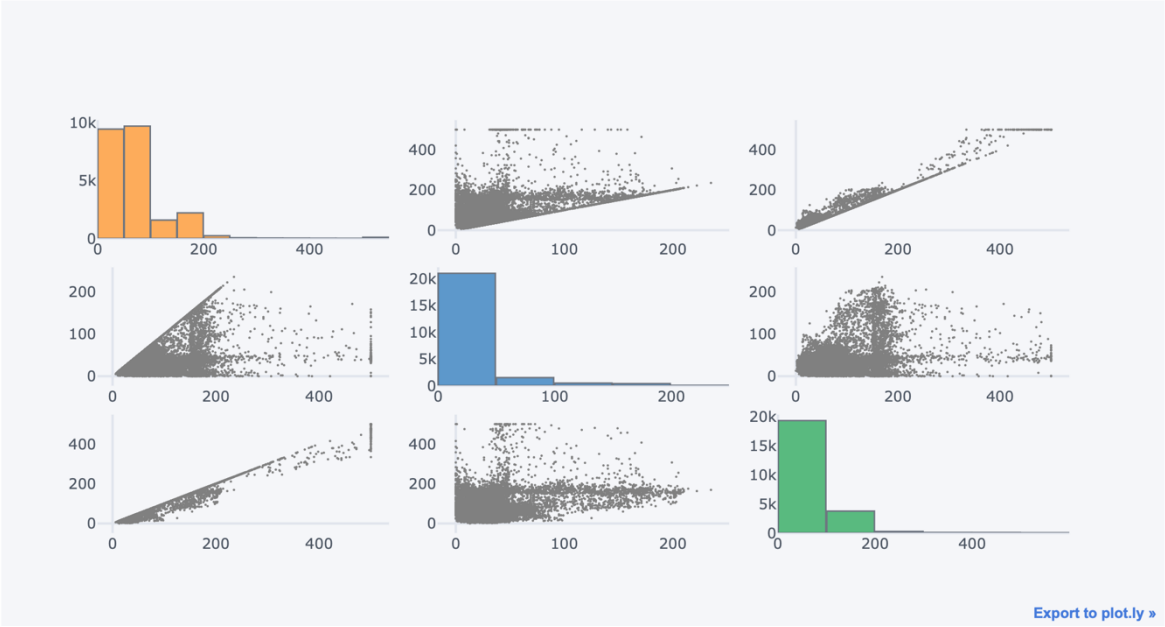



```
import cufflinks as cf
cf.go_offline()
ef.iplot(kind='scatter', x ='AQI Value', y = 'PM2.5 AQI Value')
```

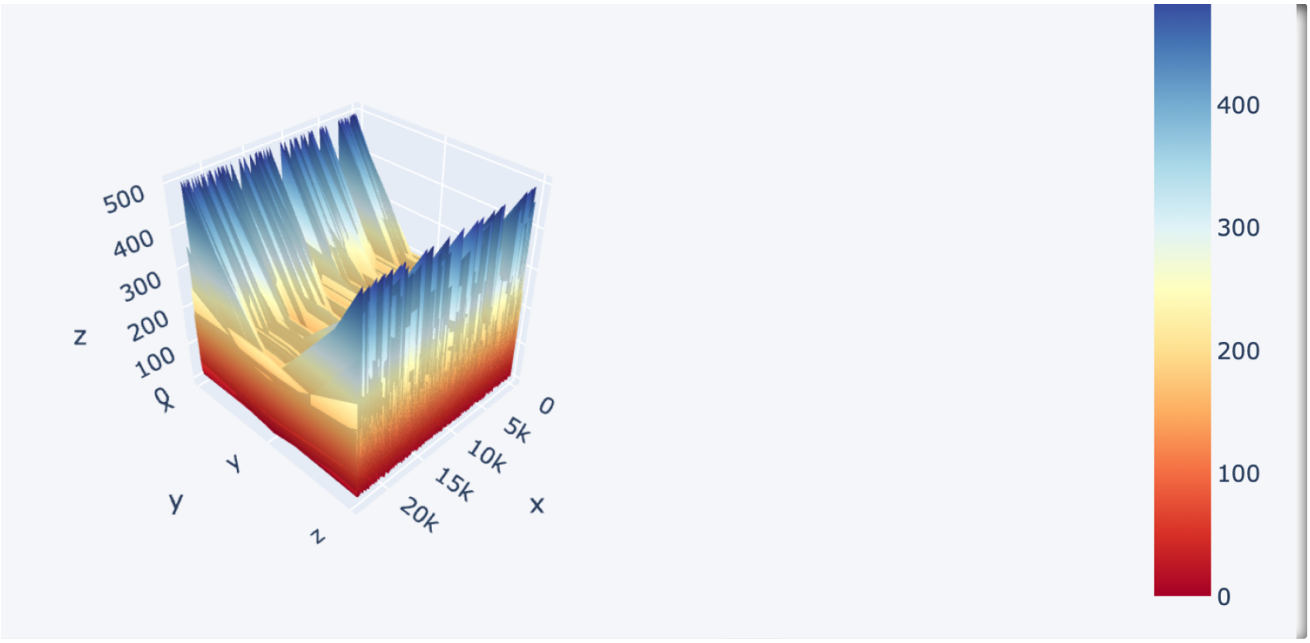


df3 = pd.DataFrame({'x':ef['AQI Value'],'y':ef['Ozone AQI Value'],'z':ef['PM2.5 AQI Value']})

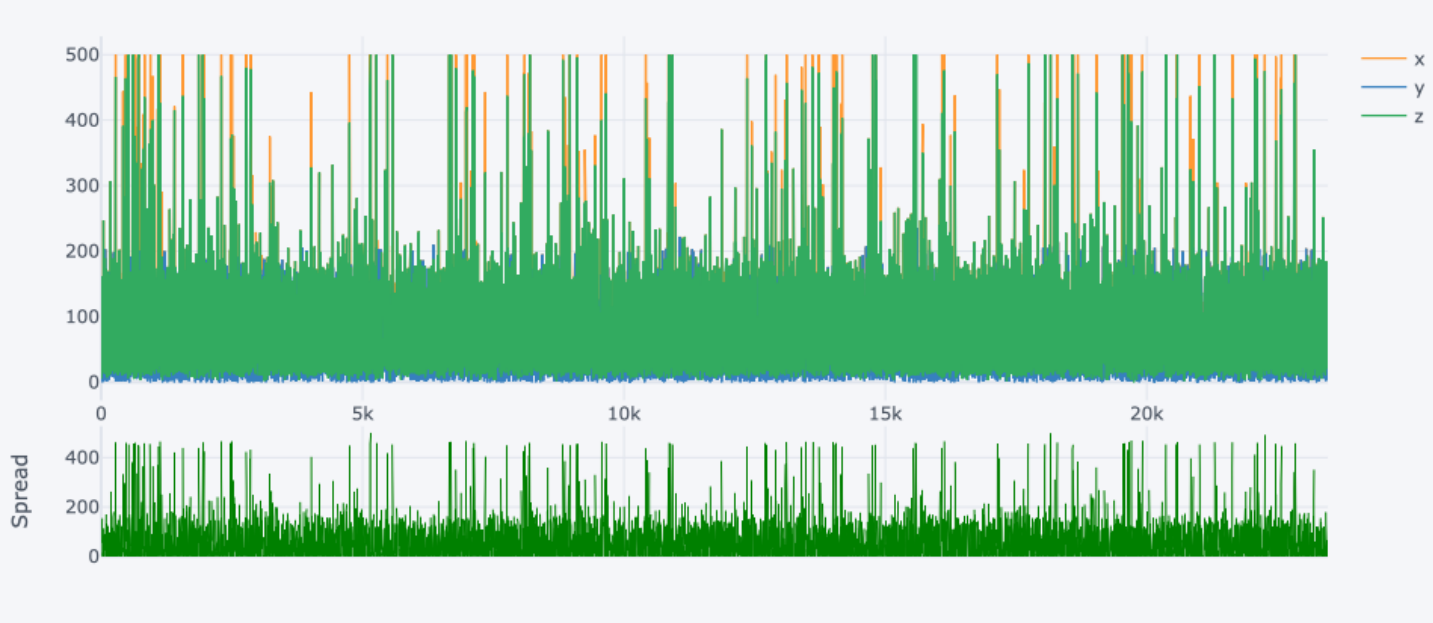
df3.scatter_matrix()



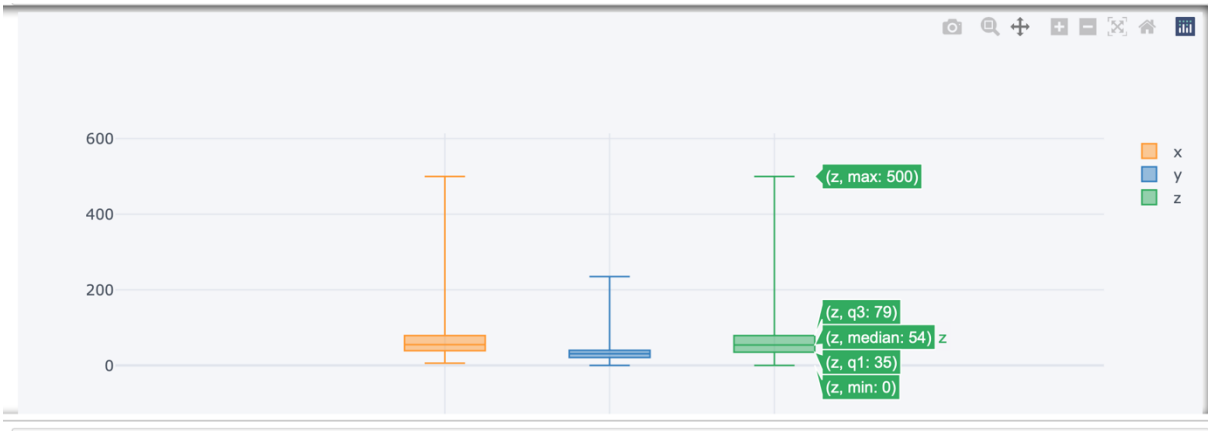
df3.iplot(kind='surface',colorscale='rdylbu')



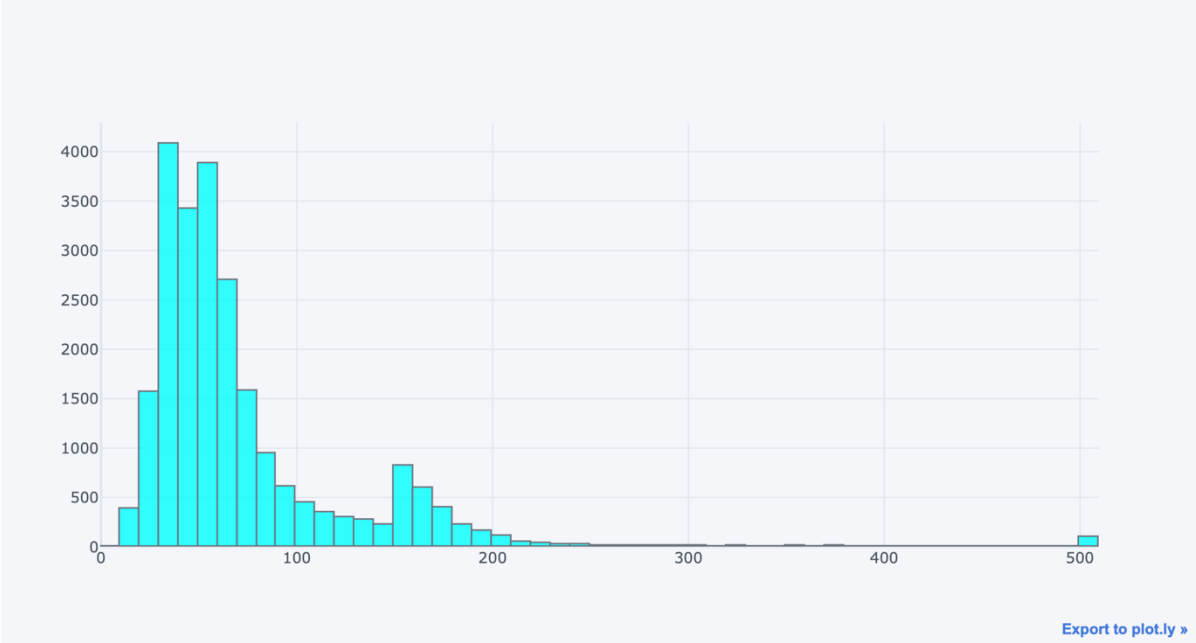
df3.iplot(kind ='spread')



df3.iplot(kind='box')



df3['x'].iplot(kind='hist',bins=75, colors='Aqua')



df3.head()

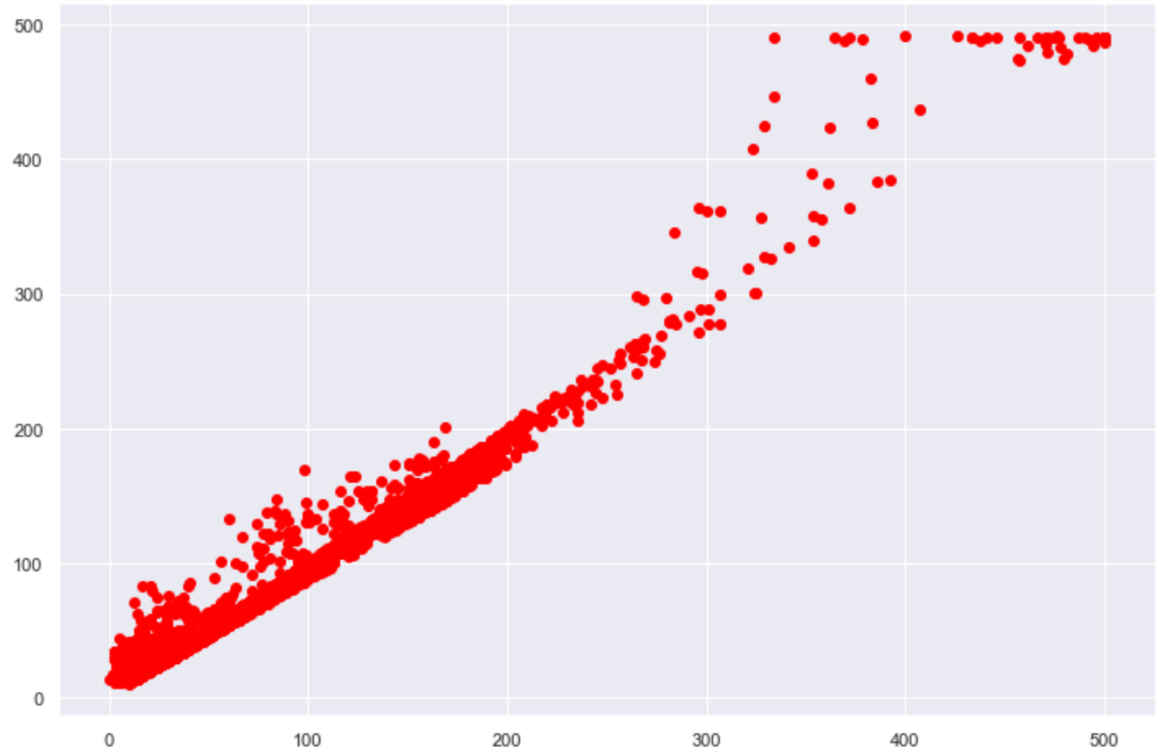
	x	y	z
0	51	36	51
1	41	5	41
2	66	39	66
3	34	34	20
4	22	22	6

```
ef.dropna(inplace = True)

features = ef[['AQI Value', 'Ozone AQI Value']]

target = ef['PM2.5 AQI Value']

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(features, target,
test_size=0.4, random_state=101)

from sklearn.linear_model import LinearRegression

lm = LinearRegression()


lm.fit(X_train,y_train)

print(lm.intercept_)

list(features.columns)

coeff_df =
pd.DataFrame(lm.coef_,features.columns,columns=['Coefficient'])

coeff_df
```

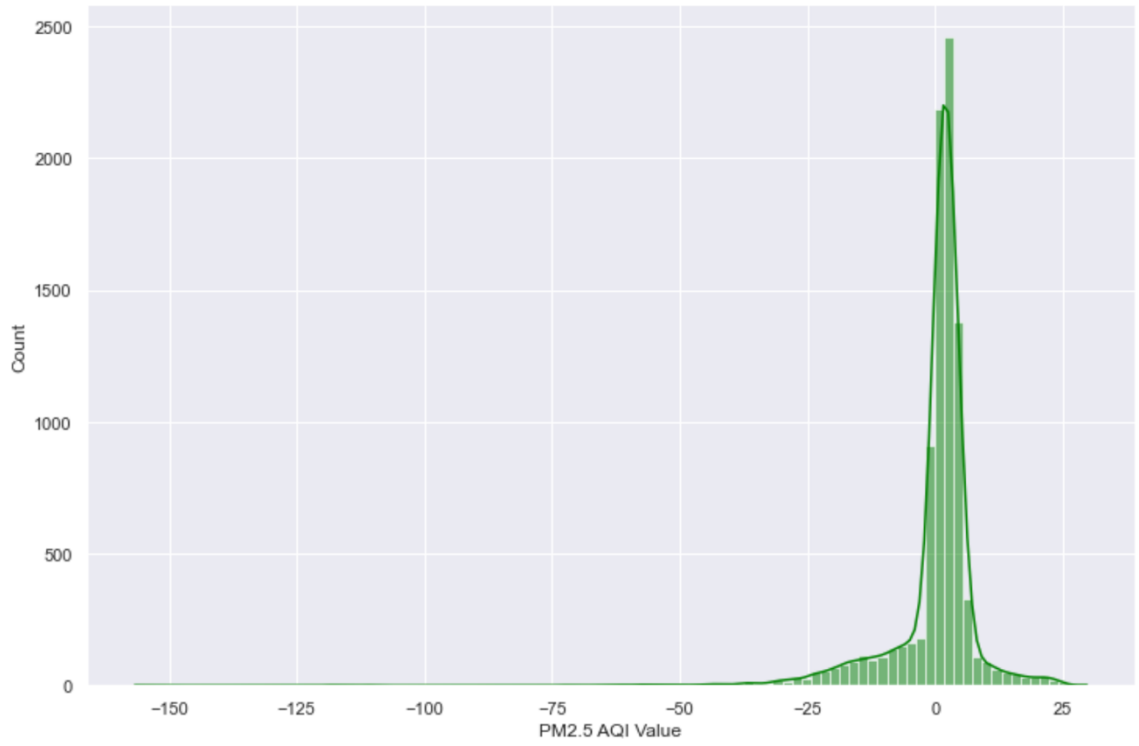
	Coefficient
AQI Value	0.987163
Ozone AQI Value	-0.140707

```
predictions = lm.predict(X_test)

plt.figure(figsize=(12,8))

sns.set_style(style='darkgrid')
```

```
plt.scatter(y_test,predictions, color='red')
```




```
plt.figure(figsize=(12,8))
sns.histplot((y_test-predictions),bins=100,kde=True, color ='green');
```




```
from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(y_test, predictions))

print('MSE:', metrics.mean_squared_error(y_test, predictions))

print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test,
predictions)))

new_sample = np.array([25, 35]).reshape(1, -1)

predicted_pm25_aqi = lm.predict(new_sample)

print(f'Predicted PM2.5 AQI Value: {predicted_pm25_aqi[0]}')
```