Air Pollution

DATA SCIENCE PROJECT

Merey Bolat
Umargaliyeva Leila
Zhumabek Zhaina

Why Air pollution?

Air pollution is a significant contributor to global mortality, with studies linking it to respiratory and cardiovascular diseases. By analyzing air quality data and predicting mortality rates, this project aimes to highlight the urgent need for measures to improve air quality and protect public health.

Research questions

1. What percentage of children die as a result of polluted air?
2. How is air pollution connected to sustainable development?
3. Is there an effect of poverty and average income on air quality degradation?
4. What are the main sources and types of air pollutants in a specific region or city, and how have they changed over time?

Main dataset

air_pollution_data.csv

- 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 micrometers or less of the city
- PM2.5 AQI Category : AQI category of Particulate Matter with a diameter of 2.5 micrometers or less of the city

This dataset was taken from kaggle.com
(https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset)

Dataset.csv for model training and predicting

Id - a unique id

region - an identifier of a region

O3 mean - ozone, daily average computed for a particular region

PM10 mean - particulate matter 10 micrometers or less in diameter, daily average

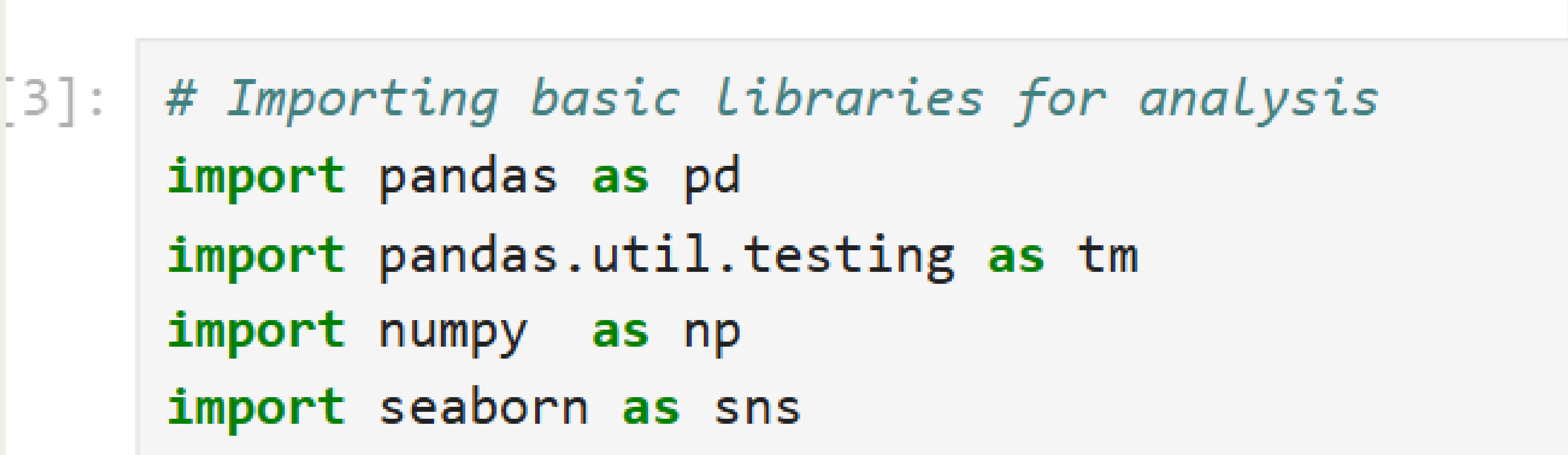
PM25 mean - particulate matter 2.5 micrometers or less in diameter, daily average

NO2 mean - nitrogen dioxide, daily average

Temperature mean - Temperature at 2 m, daily average

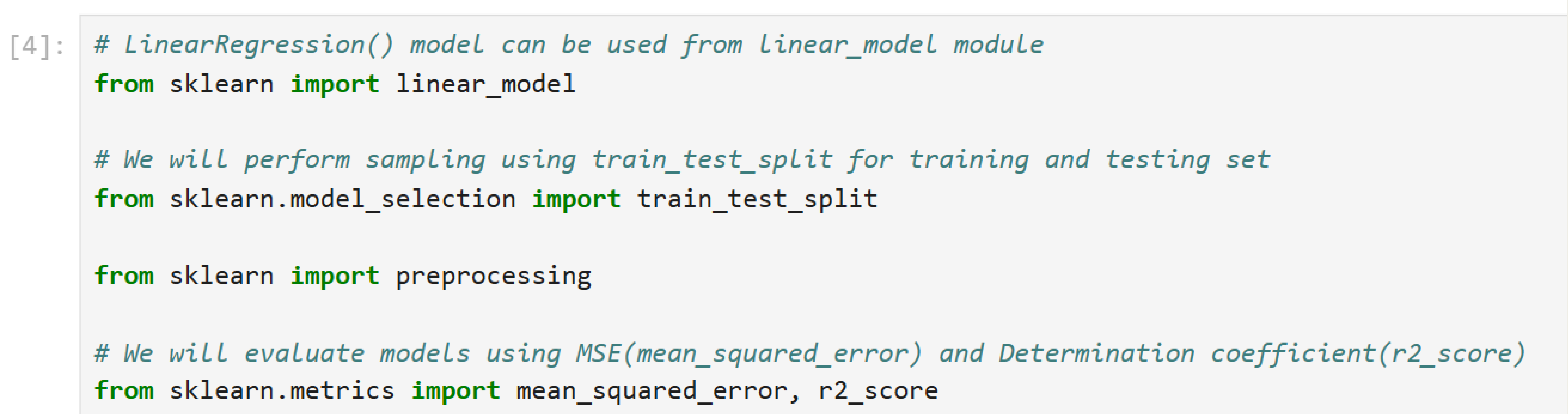
mortality rate - number of deaths per 100000 people.

Python Libraries



[3]:

```
# Importing basic libraries for analysis
import pandas as pd
import pandas.util.testing as tm
import numpy  as np
import seaborn as sns
```

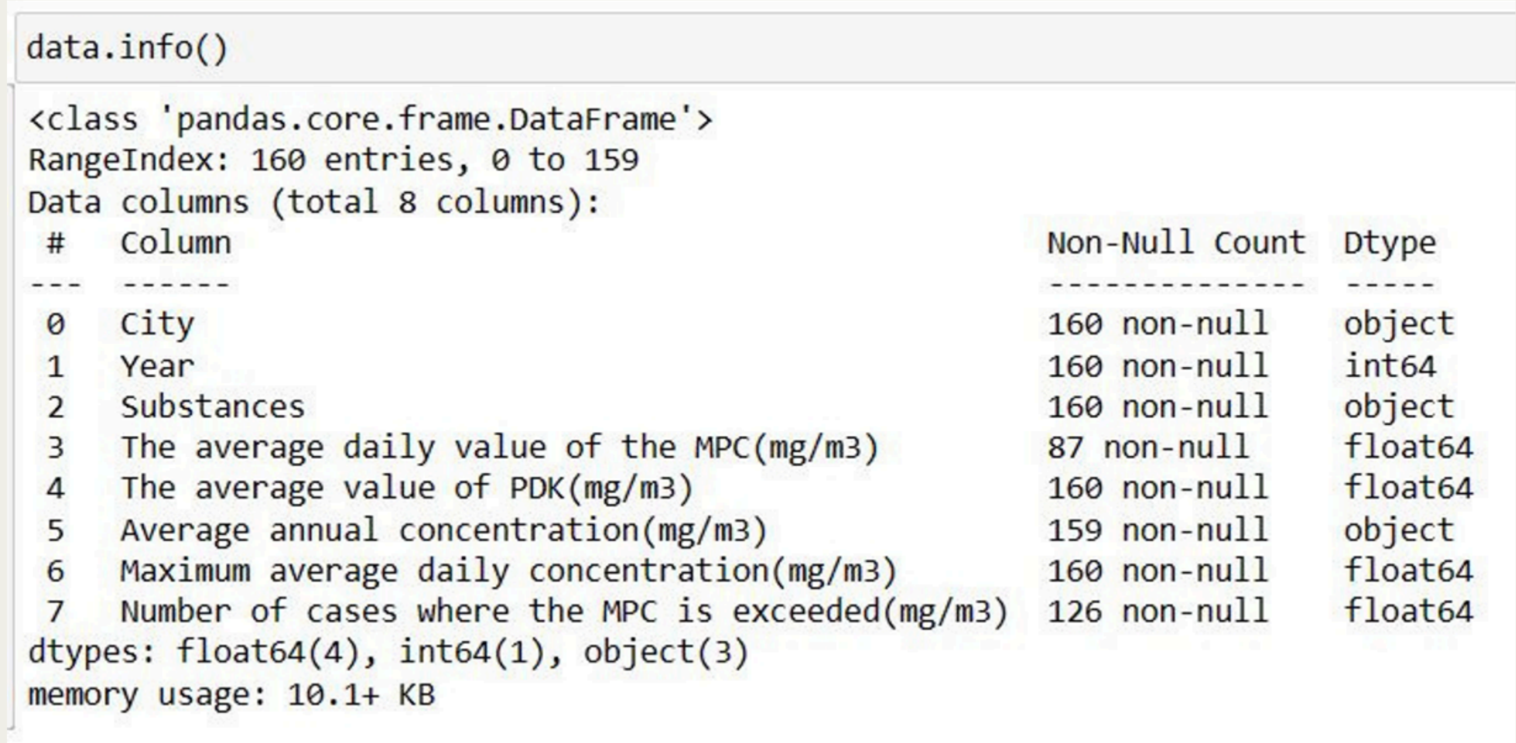


[4]:

```
# LinearRegression() model can be used from linear_model module
from sklearn import linear_model

# We will perform sampling using train_test_split for training and testing set
from sklearn.model_selection import train_test_split

from sklearn import preprocessing

# We will evaluate models using MSE(mean_squared_error) and Determination coefficient(r2_score)
from sklearn.metrics import mean_squared_error, r2_score
```

Exploratory Data Analysis

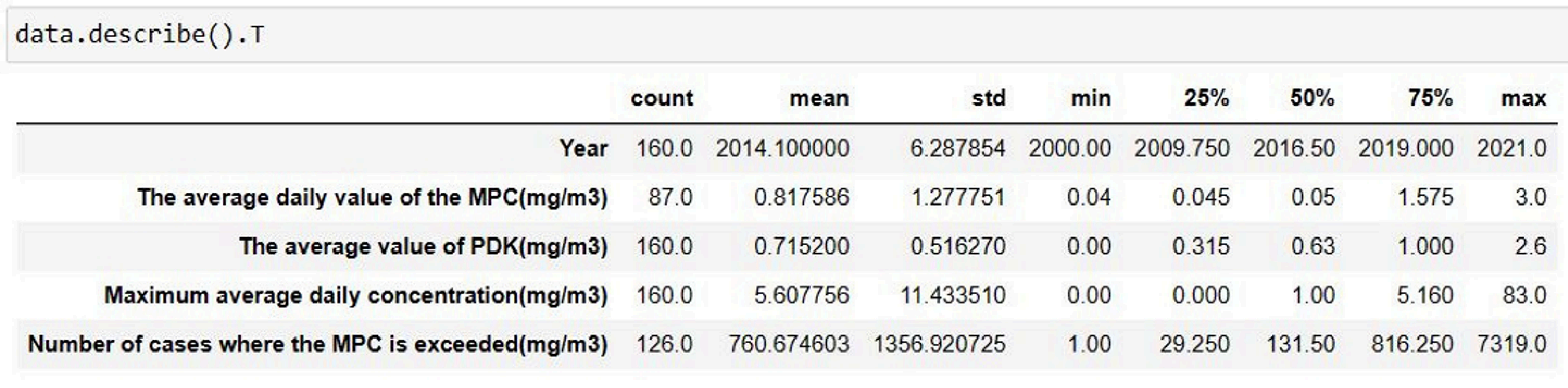
1.Data Profiling



```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 160 entries, 0 to 159
Data columns (total 8 columns):
 #   Column                                       Non-Null Count  Dtype  
---  ------                                       --------------  -----  
 0   City                                         160 non-null    object 
 1   Year                                         160 non-null    int64  
 2   Substances                                   160 non-null    object 
 3   The average daily value of the MPC(mg/m3)    87 non-null     float64
 4   The average value of PDK(mg/m3)              160 non-null    float64
 5   Average annual concentration(mg/m3)          159 non-null    object 
 6   Maximum average daily concentration(mg/m3)   160 non-null    float64
 7   Number of cases where the MPC is exceeded(mg/m3)  126 non-null    float64
dtypes: float64(4), int64(1), object(3)
memory usage: 10.1+ KB
```

2. General Statistics



```
data.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Year	160.0	2014.100000	6.287854	2000.00	2009.750	2016.50	2019.000	2021.0
The average daily value of the MPC(mg/m3)	87.0	0.817586	1.277751	0.04	0.045	0.05	1.575	3.0
The average value of PDK(mg/m3)	160.0	0.715200	0.516270	0.00	0.315	0.63	1.000	2.6
Maximum average daily concentration(mg/m3)	160.0	5.607756	11.433510	0.00	0.000	1.00	5.160	83.0
Number of cases where the MPC is exceeded(mg/m3)	126.0	760.674603	1356.920725	1.00	29.250	131.50	816.250	7319.0

3. Data cleaning

Possible problems:

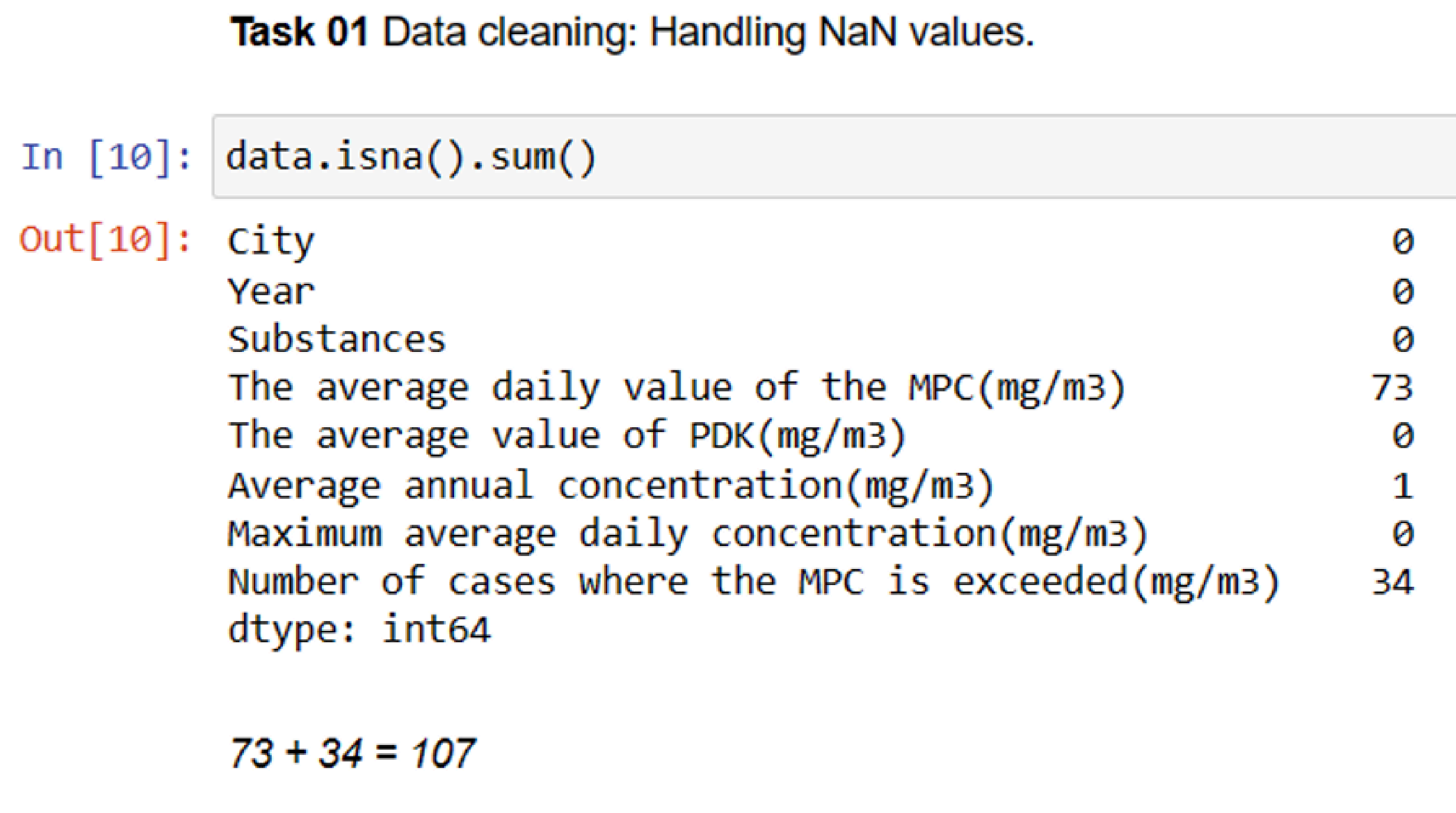
- Missing values
- Outliers

Missing values

- Can be removed if they are in range 0-5%
- Can be restored if they are in range 5-20% (by replacing with median, mean and etc.)

3. Data Cleaning

Handling empty values

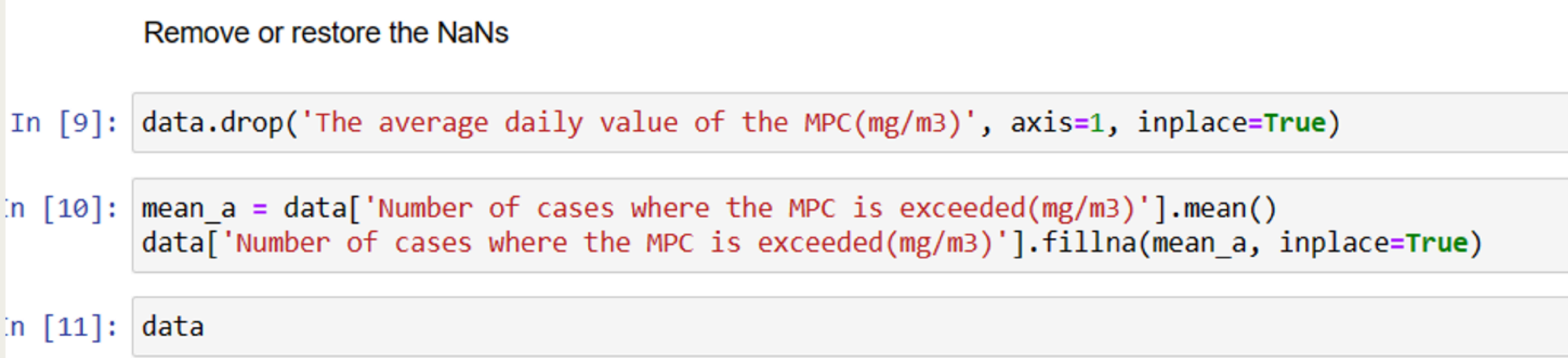


Task 01 Data cleaning: Handling NaN values.

```
In [10]: data.isna().sum()

Out[10]: City                                                0
         Year                                                0
         Substances                                          0
         The average daily value of the MPC(mg/m3)          73
         The average value of PDK(mg/m3)                     0
         Average annual concentration(mg/m3)                 1
         Maximum average daily concentration(mg/m3)          0
         Number of cases where the MPC is exceeded(mg/m3)   34
         dtype: int64
```

73 + 34 = 107

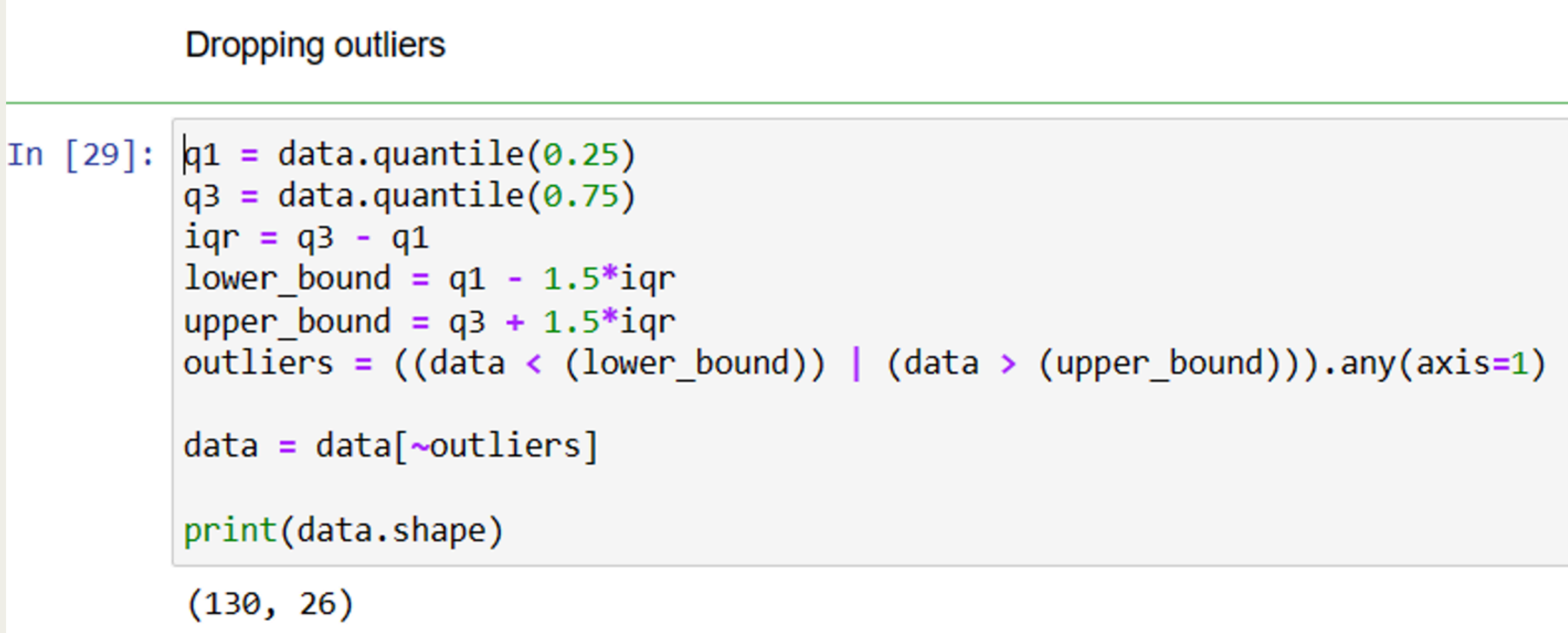
We will restore the NaNs by using means, because it is commonly used, also it can be a reasonable way when the amount of missing data is small.



Remove or restore the NaNs

```
In [9]: data.drop('The average daily value of the MPC(mg/m3)', axis=1, inplace=True)

n [10]: mean_a = data['Number of cases where the MPC is exceeded(mg/m3)'].mean()
        data['Number of cases where the MPC is exceeded(mg/m3)'].fillna(mean_a, inplace=True)

n [11]: data
```

Handling outliers

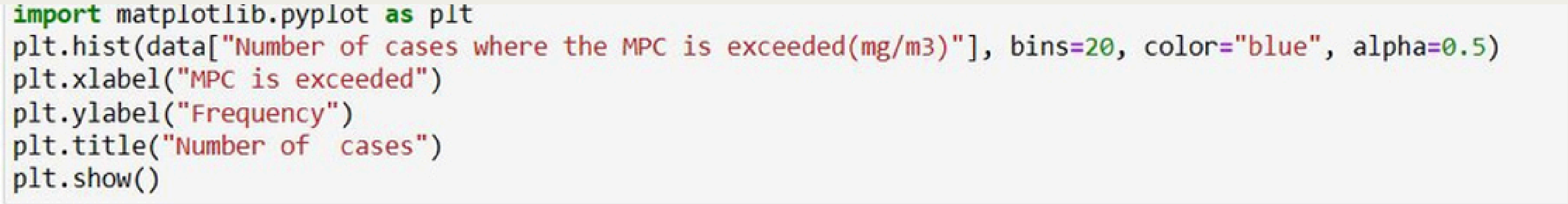


Dropping outliers

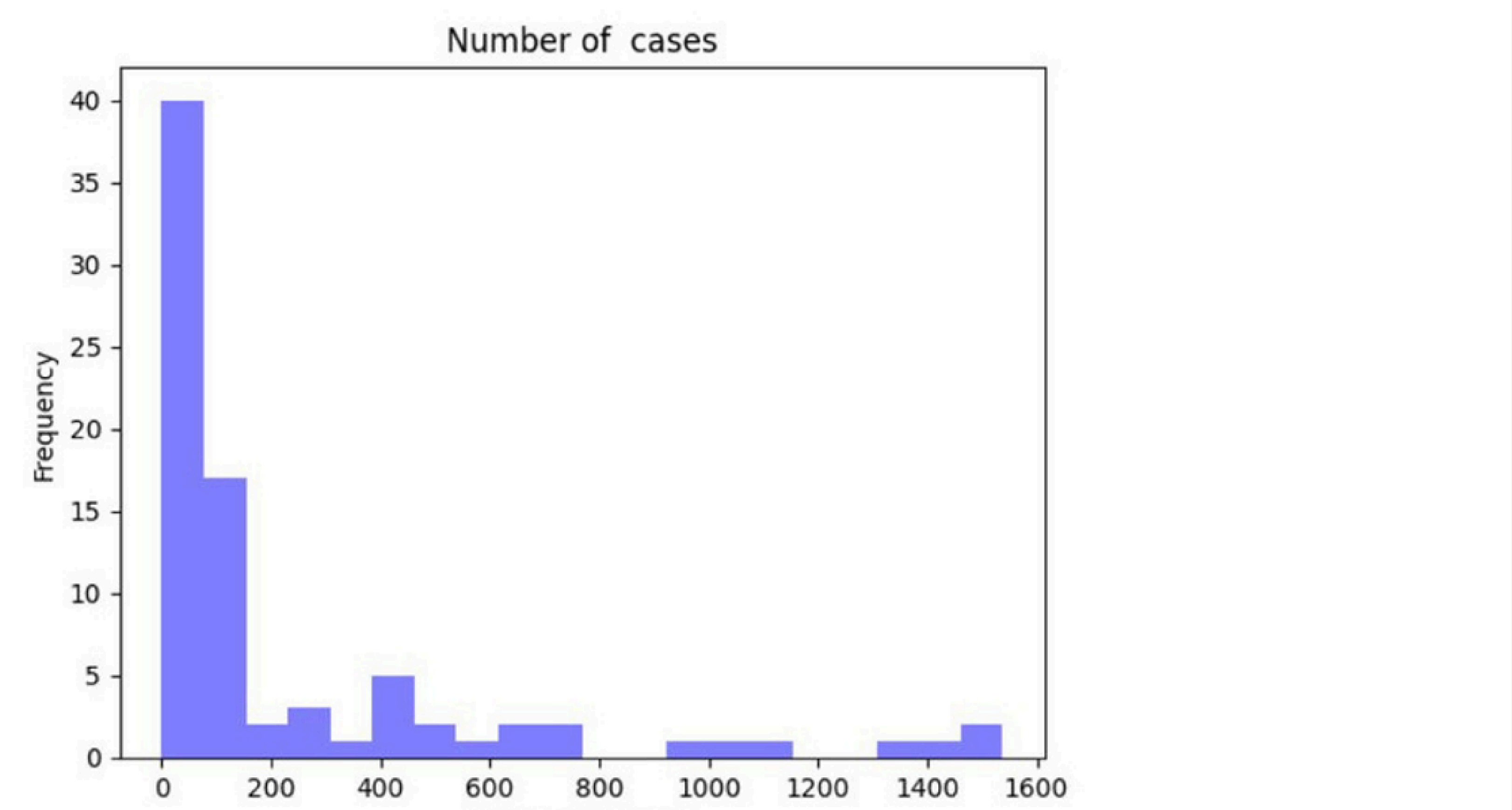
```
In [29]: q1 = data.quantile(0.25)
         q3 = data.quantile(0.75)
         iqr = q3 - q1
         lower_bound = q1 - 1.5*iqr
         upper_bound = q3 + 1.5*iqr
         outliers = ((data < (lower_bound)) | (data > (upper_bound))).any(axis=1)

         data = data[~outliers]

         print(data.shape)

         (130, 26)
```

Firstly, we should remove the column which is called 'The average daily value of the MPC(mg/m3)'
Secondly, we should restore the rows of 'Number of cases where the MPC is exceeded(mg/m3)'

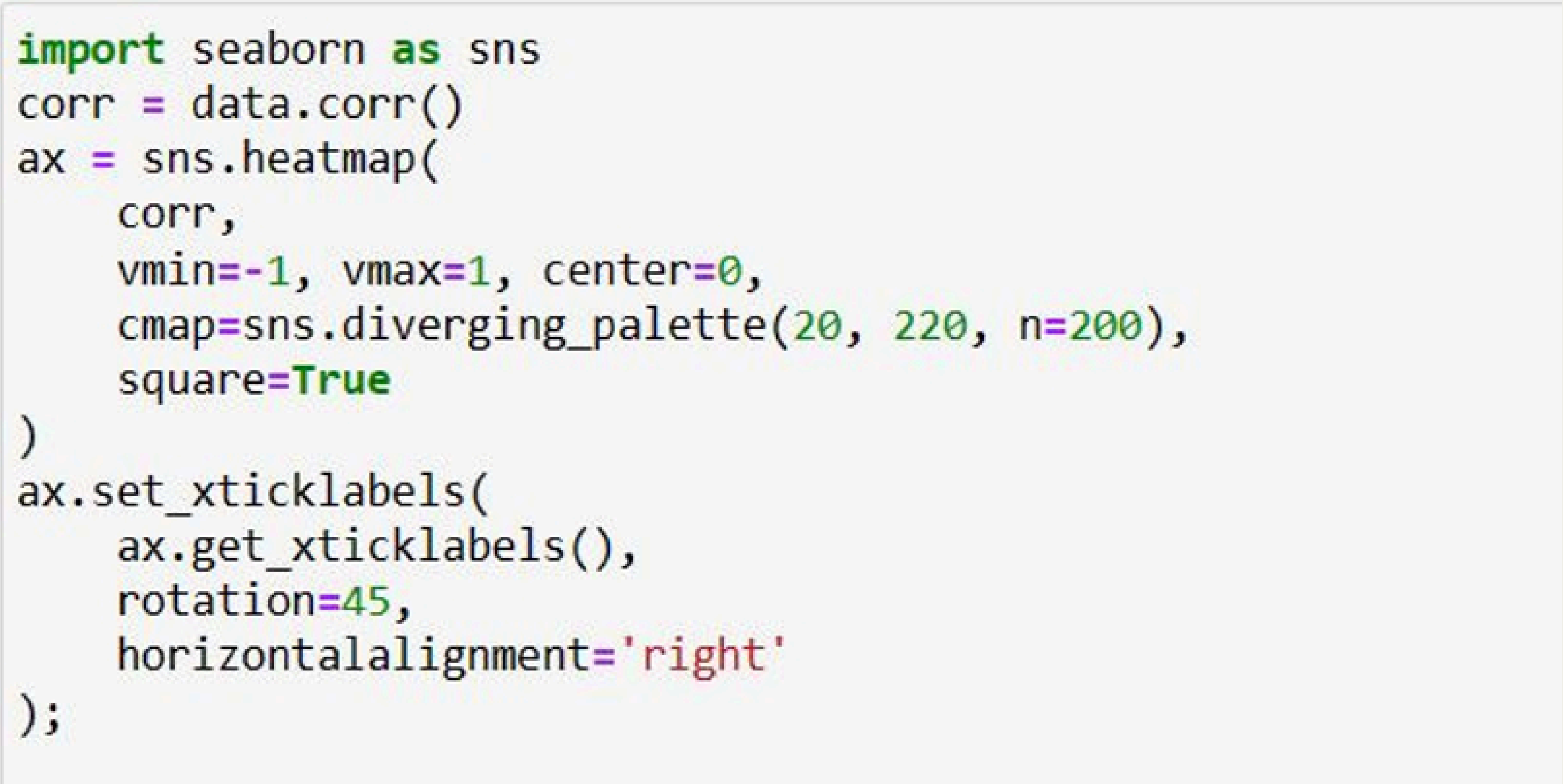
4.Relationship Visualization



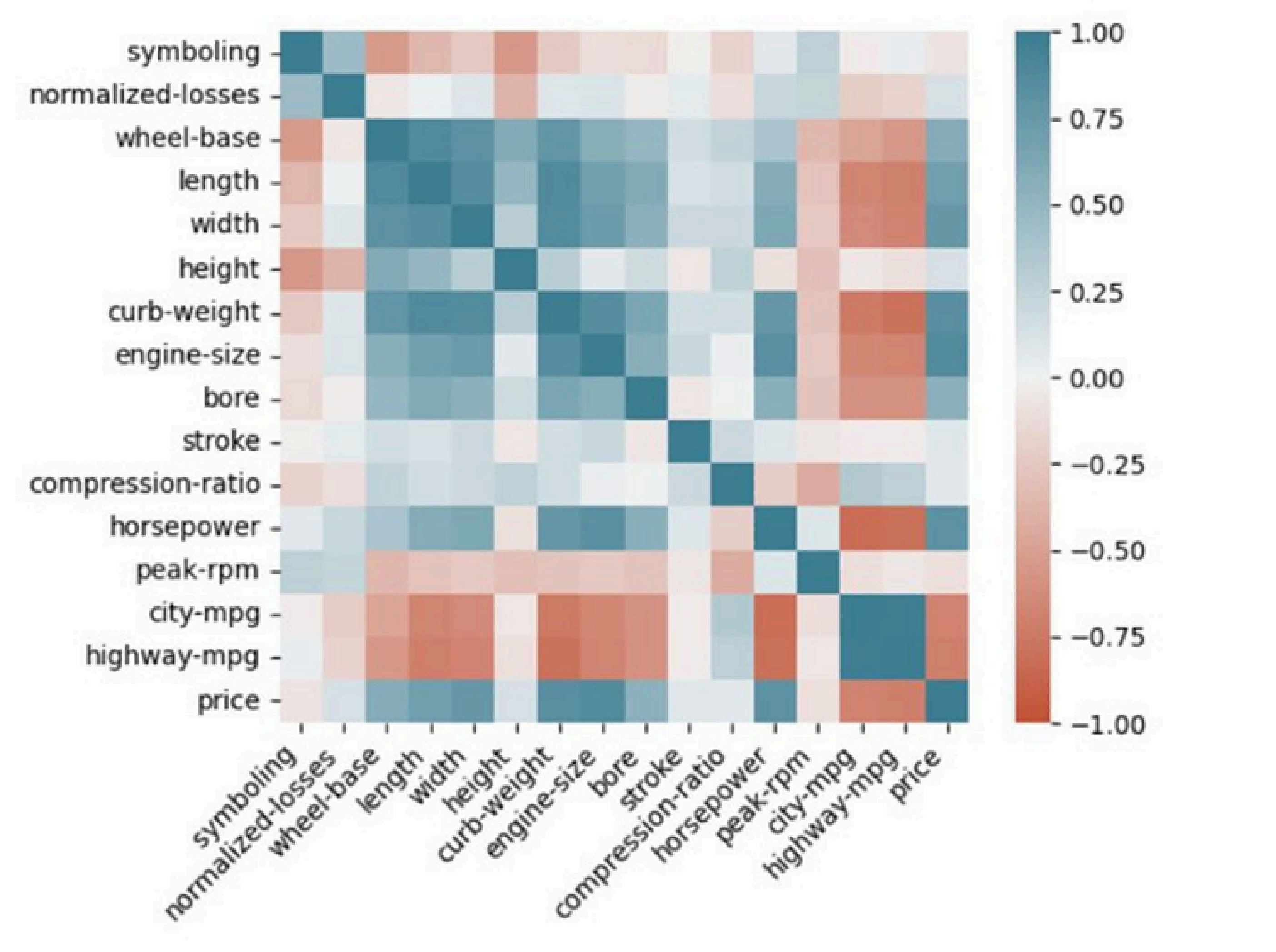
```
import matplotlib.pyplot as plt
plt.hist(data["Number of cases where the MPC is exceeded(mg/m3)"], bins=20, color="blue", alpha=0.5)
plt.xlabel("MPC is exceeded")
plt.ylabel("Frequency")
plt.title("Number of  cases")
plt.show()
```



Number of cases

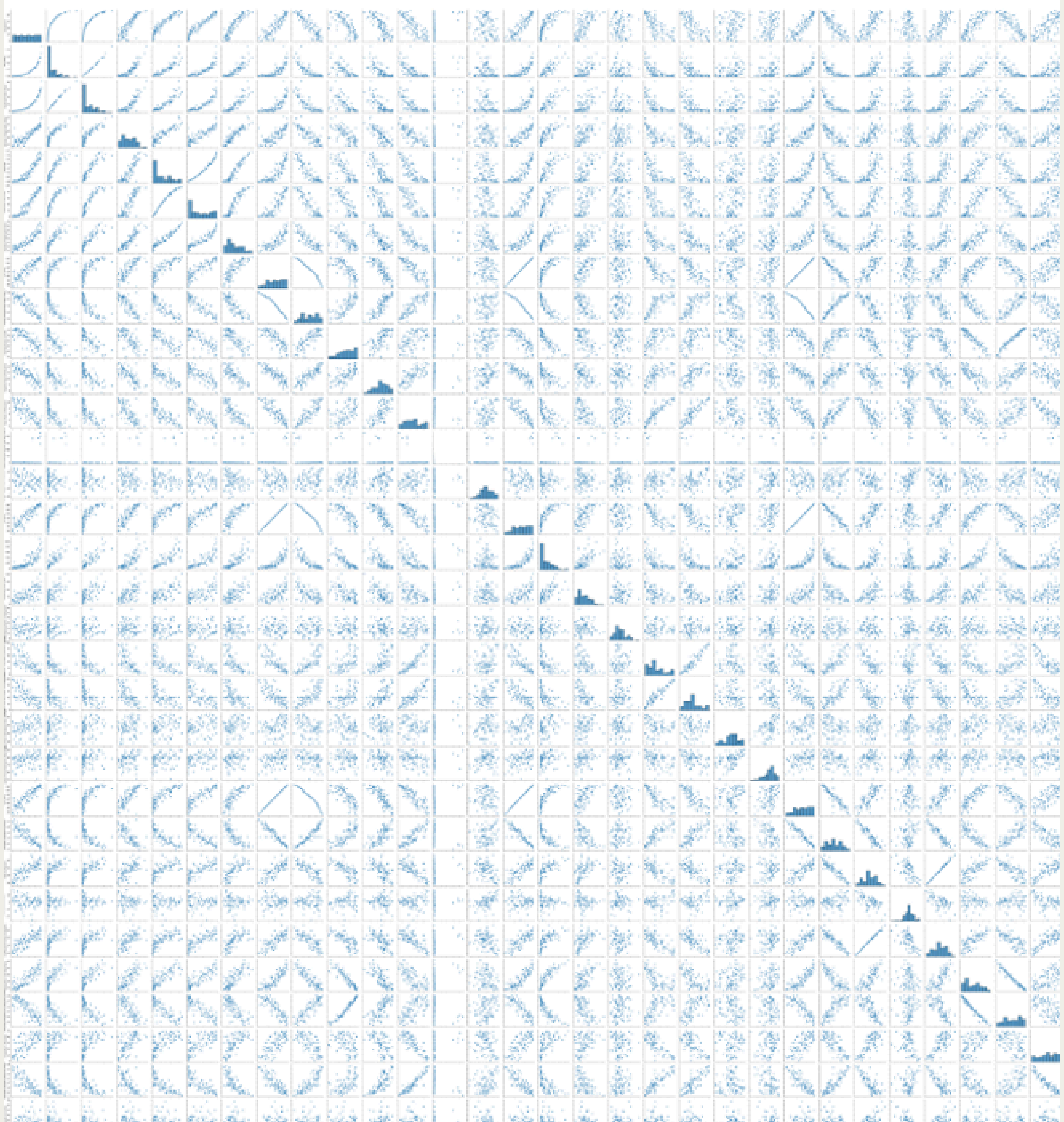


```
import seaborn as sns
corr = data.corr()
ax = sns.heatmap(
    corr,
    vmin=-1, vmax=1, center=0,
    cmap=sns.diverging_palette(20, 220, n=200),
    square=True
)
ax.set_xticklabels(
    ax.get_xticklabels(),
    rotation=45,
    horizontalalignment='right'
);
```



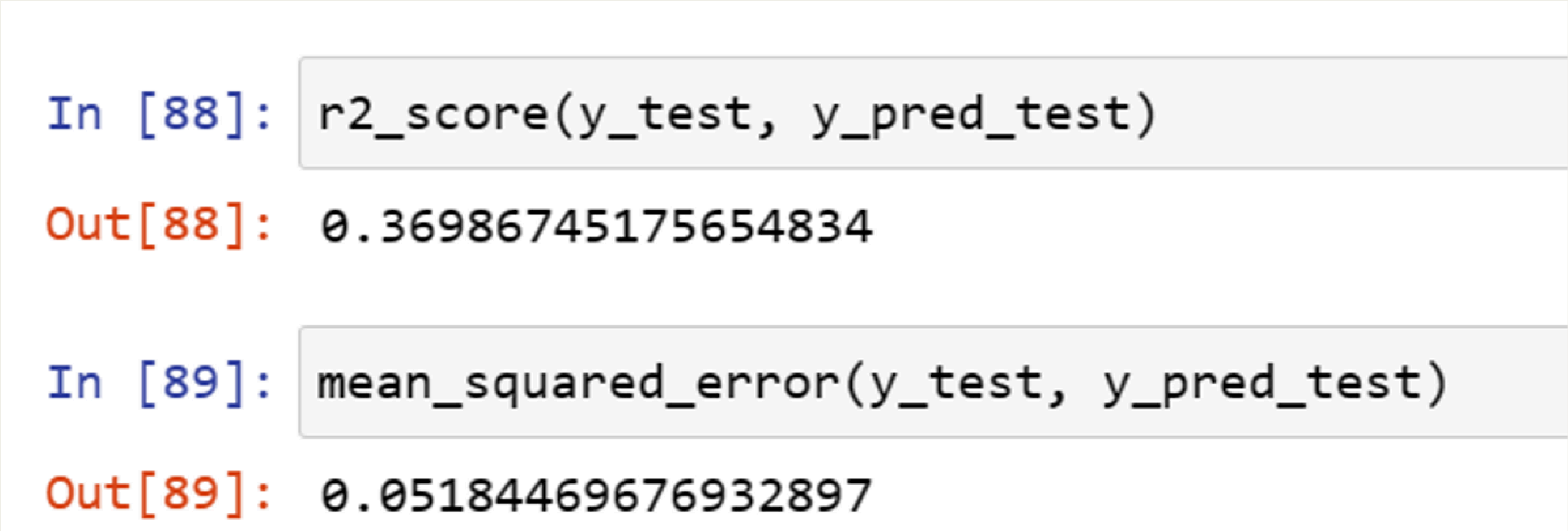
Model Selection

As you can see from this pairplot there is a lot of linear relationship in our dataset, so we decided to choose Linear Regression Model

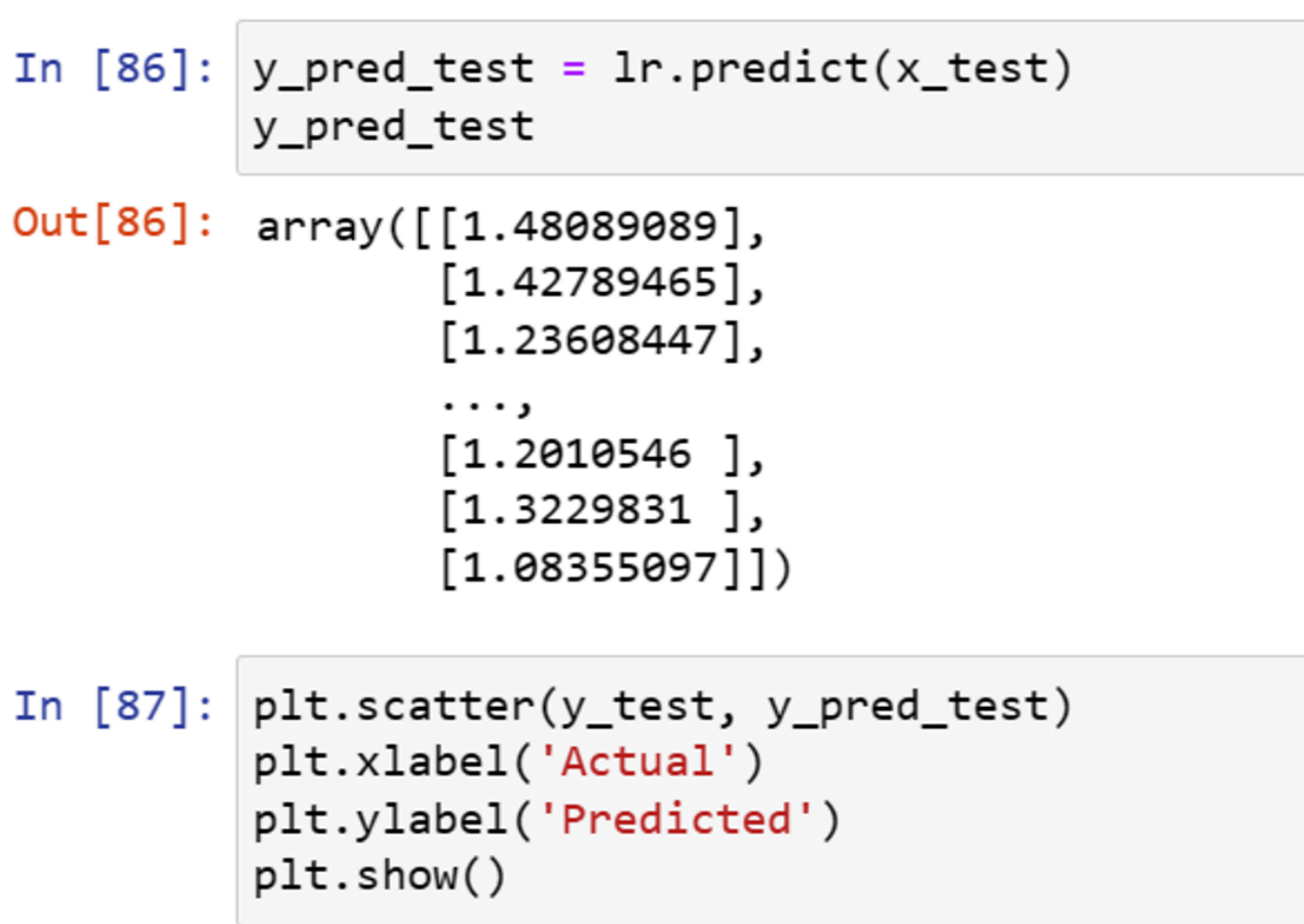


Model evaluation and improvement

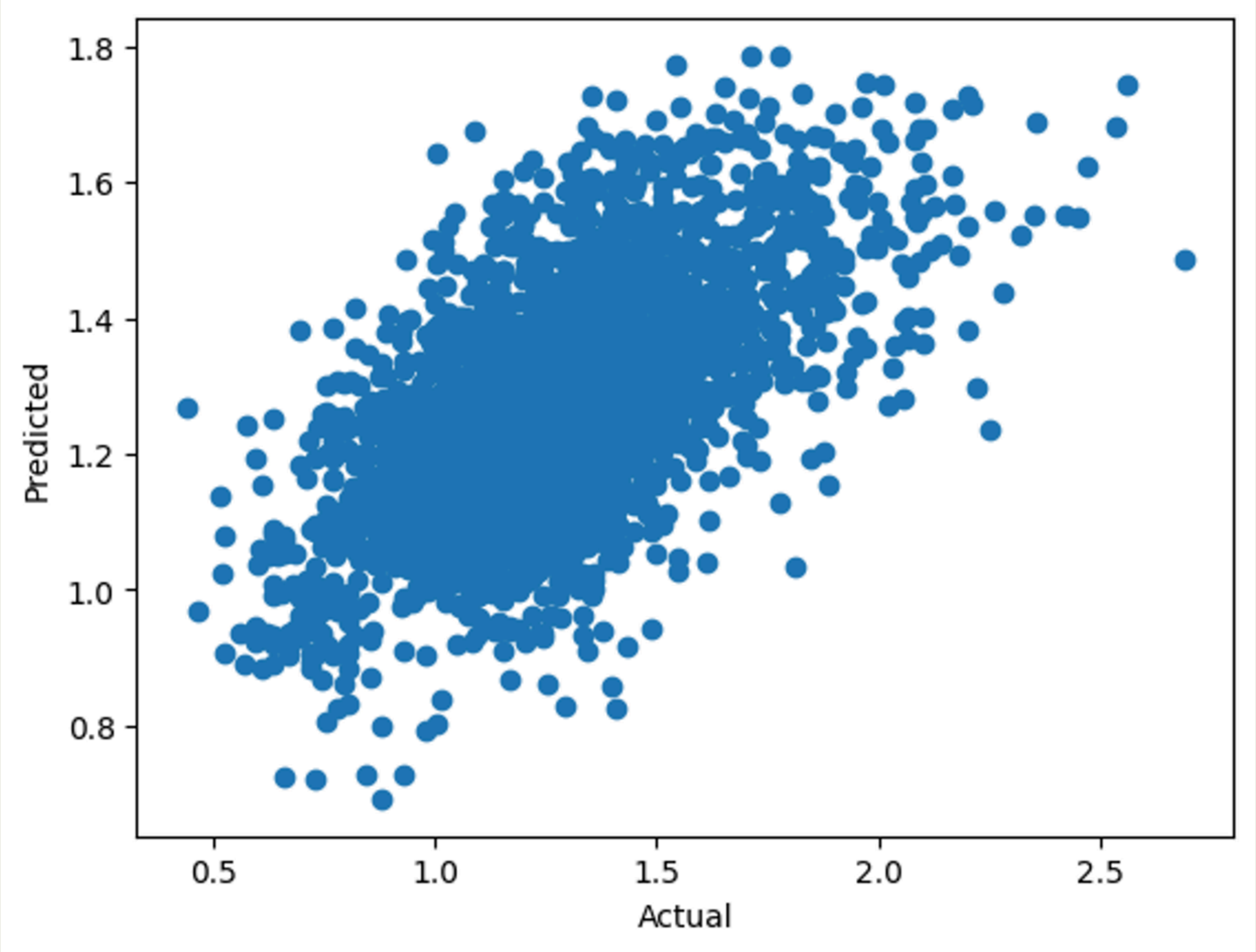
The goal is to predict mortality rates (number of deaths per 100,000 people) for each region using O3, nitrogen dioxide (NO2), PM10 PM2.5. We have used Linear Regression model



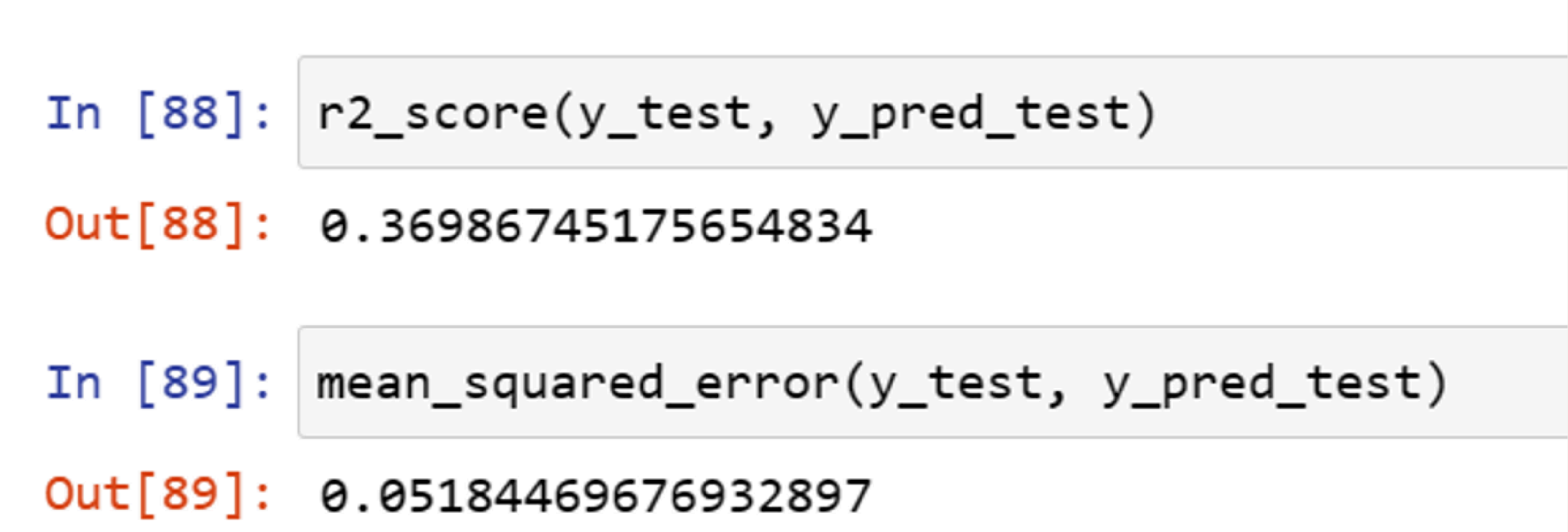
```
In [88]: r2_score(y_test, y_pred_test)

Out[88]: 0.36986745175654834

In [89]: mean_squared_error(y_test, y_pred_test)

Out[89]: 0.05184469676932897
```



```
In [86]: y_pred_test = lr.predict(x_test)
         y_pred_test

Out[86]: array([[1.48089089],
                [1.42789465],
                [1.23608447],
                ...,
                [1.2010546 ],
                [1.3229831 ],
                [1.08355097]])

In [87]: plt.scatter(y_test, y_pred_test)
         plt.xlabel('Actual')
         plt.ylabel('Predicted')
         plt.show()
```



Model evaluation and improvement



```
In [88]: r2_score(y_test, y_pred_test)

Out[88]: 0.36986745175654834

In [89]: mean_squared_error(y_test, y_pred_test)

Out[89]: 0.05184469676932897
```

Conclusion

- The current linear regression model explains 36.99% of the variability in mortality rates, which is relatively low. The MSE of 0.0518 suggests moderate prediction accuracy, but there is room for improvement.
- To improve the model, it would be better to incorporate additional predictors, exploring nonlinear relationships, or using more advanced modeling techniques.

Thank you