

▶ `df.info()`

```
>>> <class 'pandas.core.DataFrame'>
Int64Index: 2009 entries, 7834 to 9842
Data columns (total 16 columns):
#   Column              Non-Null Count  Dtype
---  -
0   City                2009 non-null   object
1   Date                2009 non-null   object
2   PM2.5               1892 non-null   float64
3   PM10                302 non-null    float64
4   NO                  1973 non-null   float64
5   NO2                 1973 non-null   float64
6   NOx                 1973 non-null   float64
7   NH3                 1547 non-null   float64
8   CO                  1984 non-null   float64
9   SO2                 1973 non-null   float64
10  O3                   1960 non-null   float64
11  Benzene              1724 non-null   float64
12  Toluene              1756 non-null   float64
13  Xylene               0 non-null      float64
14  AQI                  1884 non-null   float64
15  AQI_Bucket           1884 non-null   object
dtypes: float64(13), object(3)
memory usage: 266.8+ KB
```

```
[ ] df.isnull().sum()
```

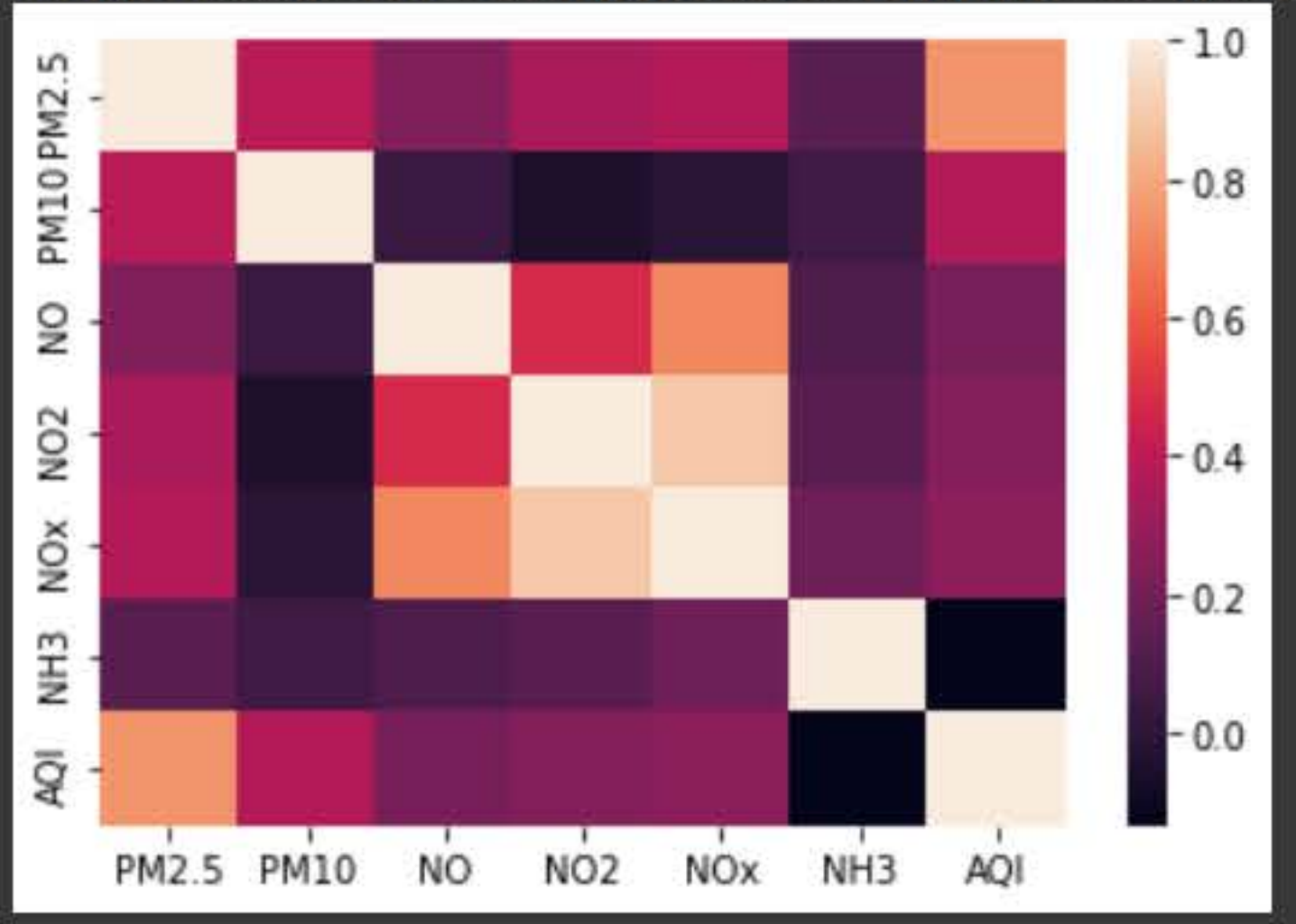
```
City      0
Date      0
PM2.5     117
PM10      1707
NO        36
NO2       36
NOx       36
NH3       462
CO        25
SO2       36
O3        49
Benzene   285
Toluene   253
Xylene    2009
AQI       125
AQI_Bucket 125
dtype: int64
```



```
[ ] #df1['City'].unique()
df = df.dropna()
```

```
#df2 = df1[df1['City']=='Chennai']
sns.heatmap(df.corr())
```

<matplotlib.axes._subplots.AxesSubplot at 0x7efeb1591250>



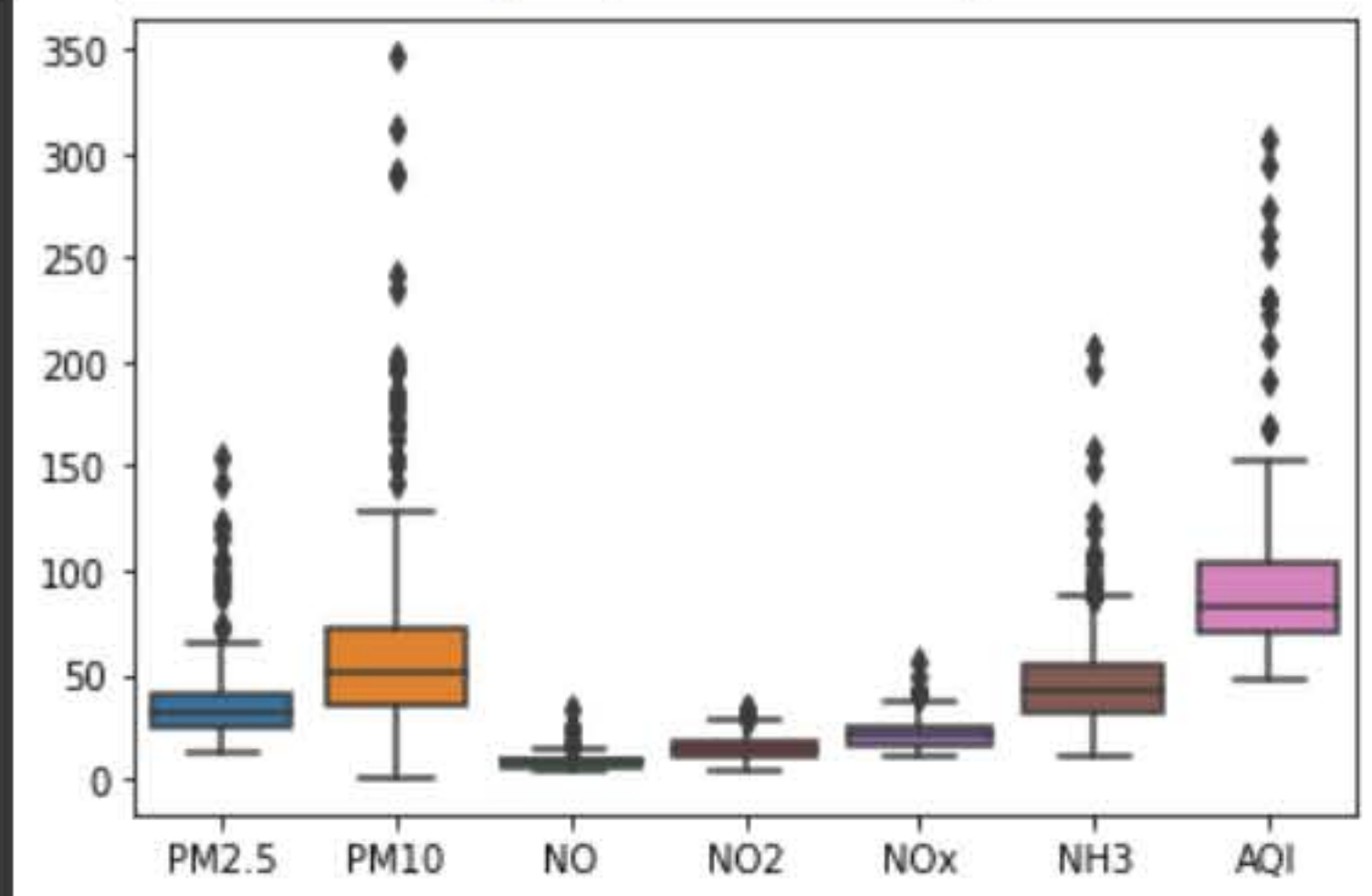
```
[ ] df.describe()
```

	PM2.5	PM10	NO	NO2	NOx	NH3	AQI
count	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000
mean	36.351391	62.834404	8.427781	14.721424	21.549106	47.081325	91.569536
std	20.209951	49.785339	3.346104	4.875727	6.757622	26.721170	38.166972
min	11.900000	0.210000	3.650000	4.380000	9.790000	10.290000	47.000000
25%	24.160000	34.557500	6.150000	11.085000	16.535000	31.667500	70.000000
50%	31.665000	50.245000	7.955000	14.065000	20.580000	42.085000	83.000000
75%	40.752500	71.975000	9.697500	17.747500	24.985000	53.897500	104.000000
max	153.730000	346.730000	33.320000	33.680000	56.010000	207.140000	306.000000

```
[ ] sns.boxplot(data=df)
```



```
[ ] <matplotlib.axes._subplots.AxesSubplot at 0x7efeb15ccc10>
```



```
[ ] drop_outlier = df[(df['AQI']>500) | (df['PM2.5']>180) | (df['NO']>65) | (df['NH3']>50) | (df['NO2']>90) | (df['NOx']>100) | (df['PM10']>450)].index
```

```
[ ] df = df.drop(drop_outlier)
```

```
[ ] df.info()
```

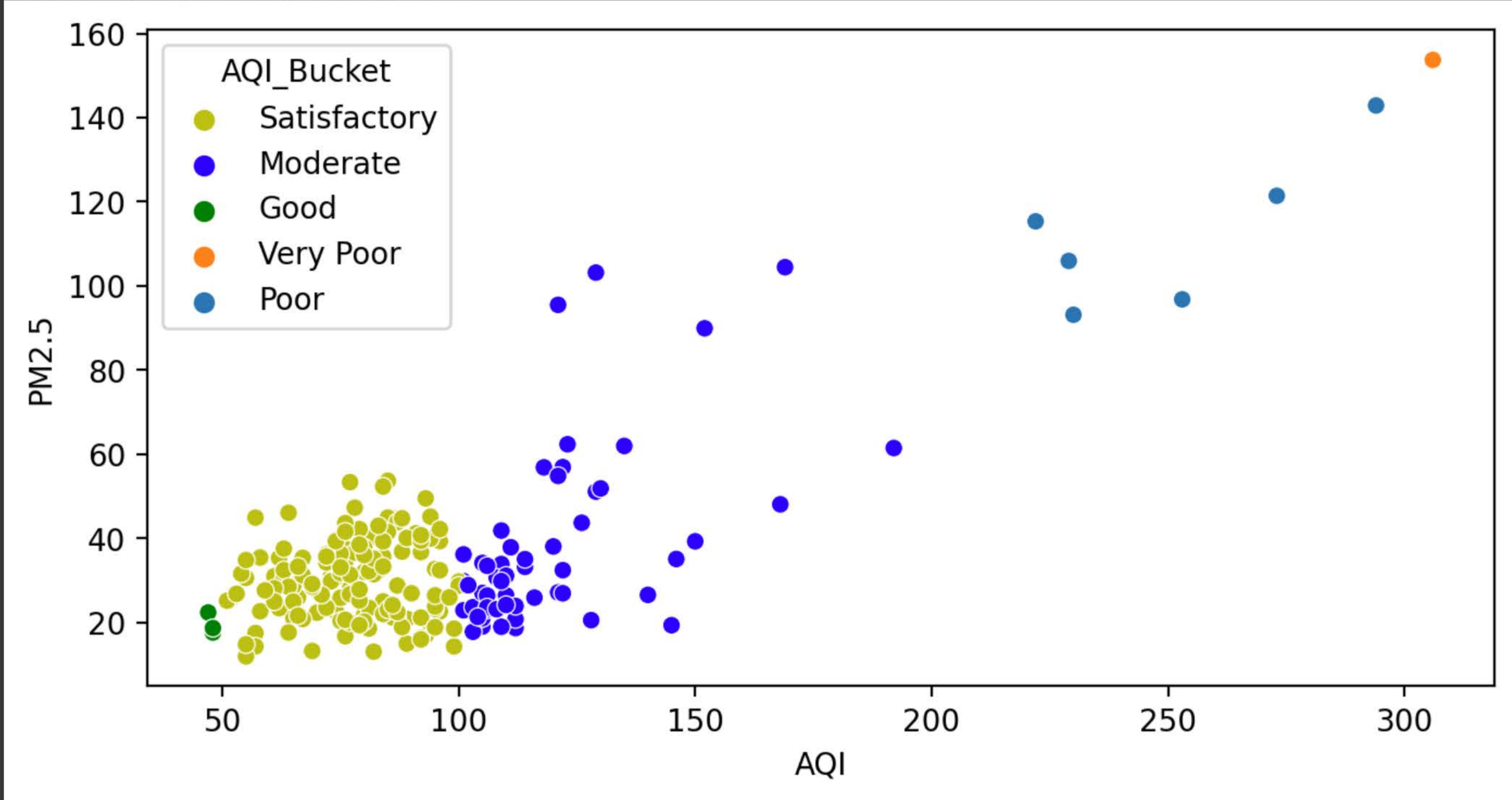
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 213 entries, 9464 to 9842
Data columns (total 10 columns):
#   Column      Non-Null Count  Dtype
---  ---
0    City        213 non-null    object
1    Date        213 non-null    object
2    PM2.5       213 non-null    float64
3    PM10        213 non-null    float64
4    NO          213 non-null    float64
5    NO2         213 non-null    float64
6    NOx         213 non-null    float64
7    NH3         213 non-null    float64
8    AQI         213 non-null    float64
9    AQI_Bucket  213 non-null    object
dtypes: float64(7), object(3)
memory usage: 18.3+ KB
```

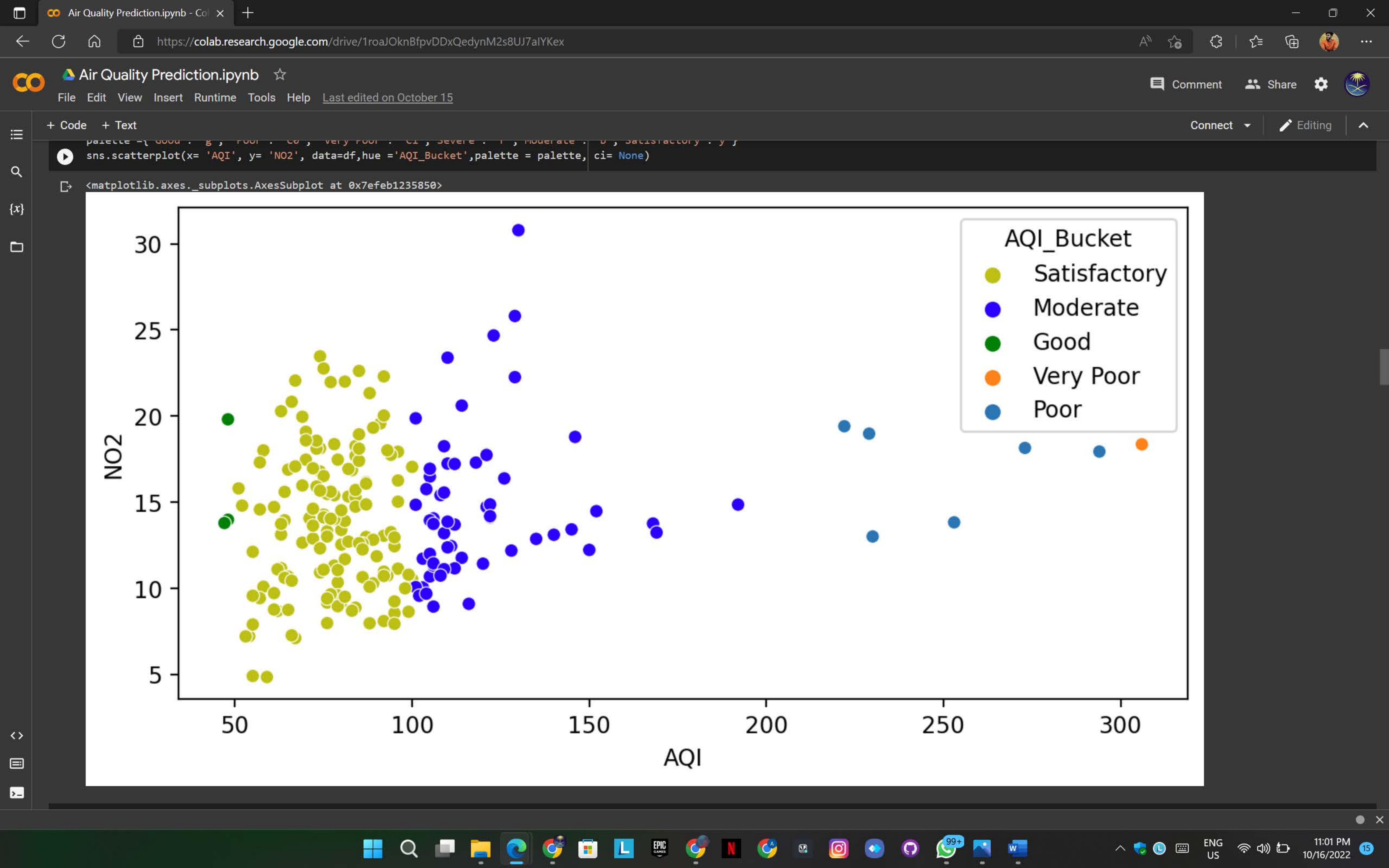
```
[ ] sns.boxplot(data= df)
```

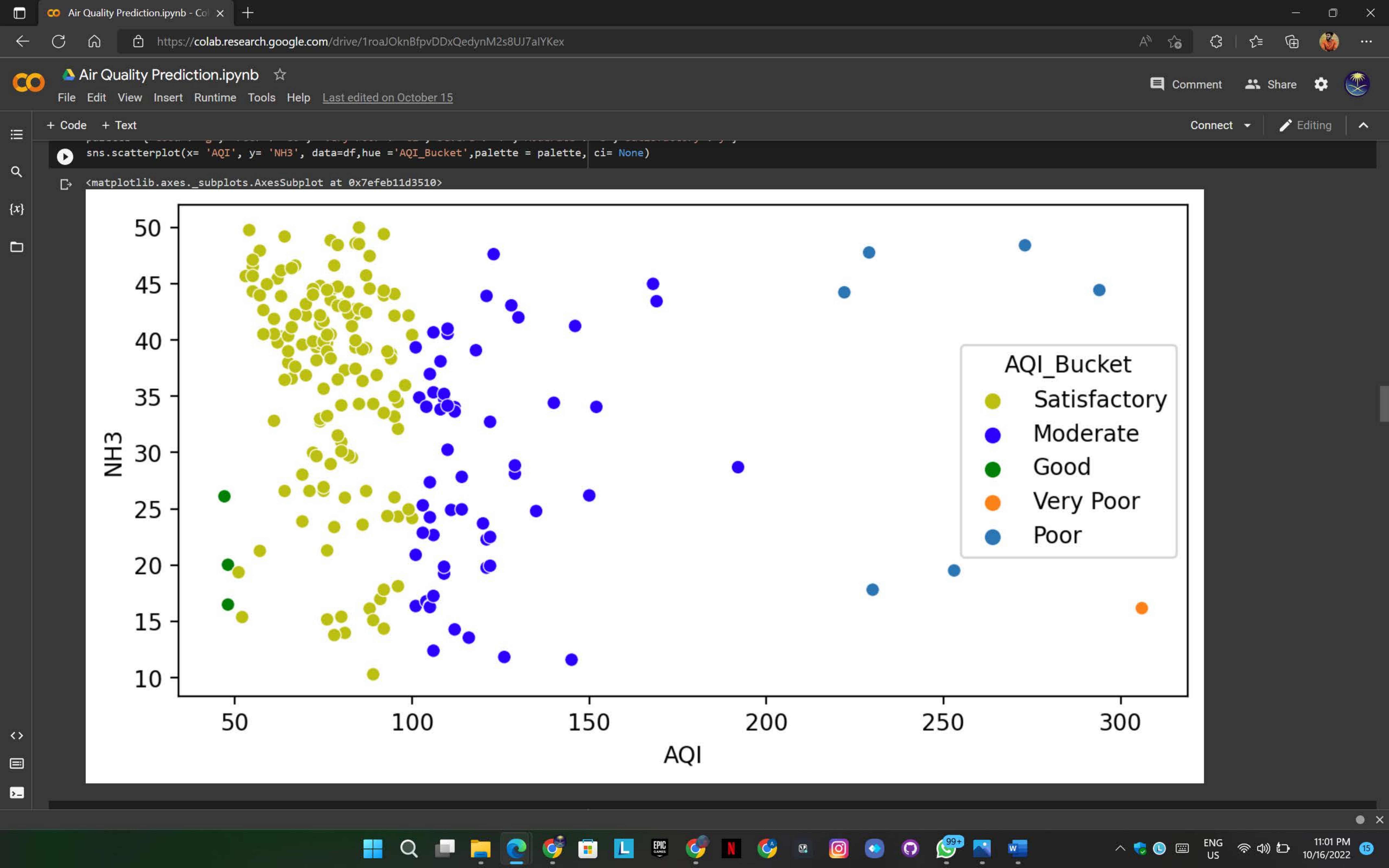


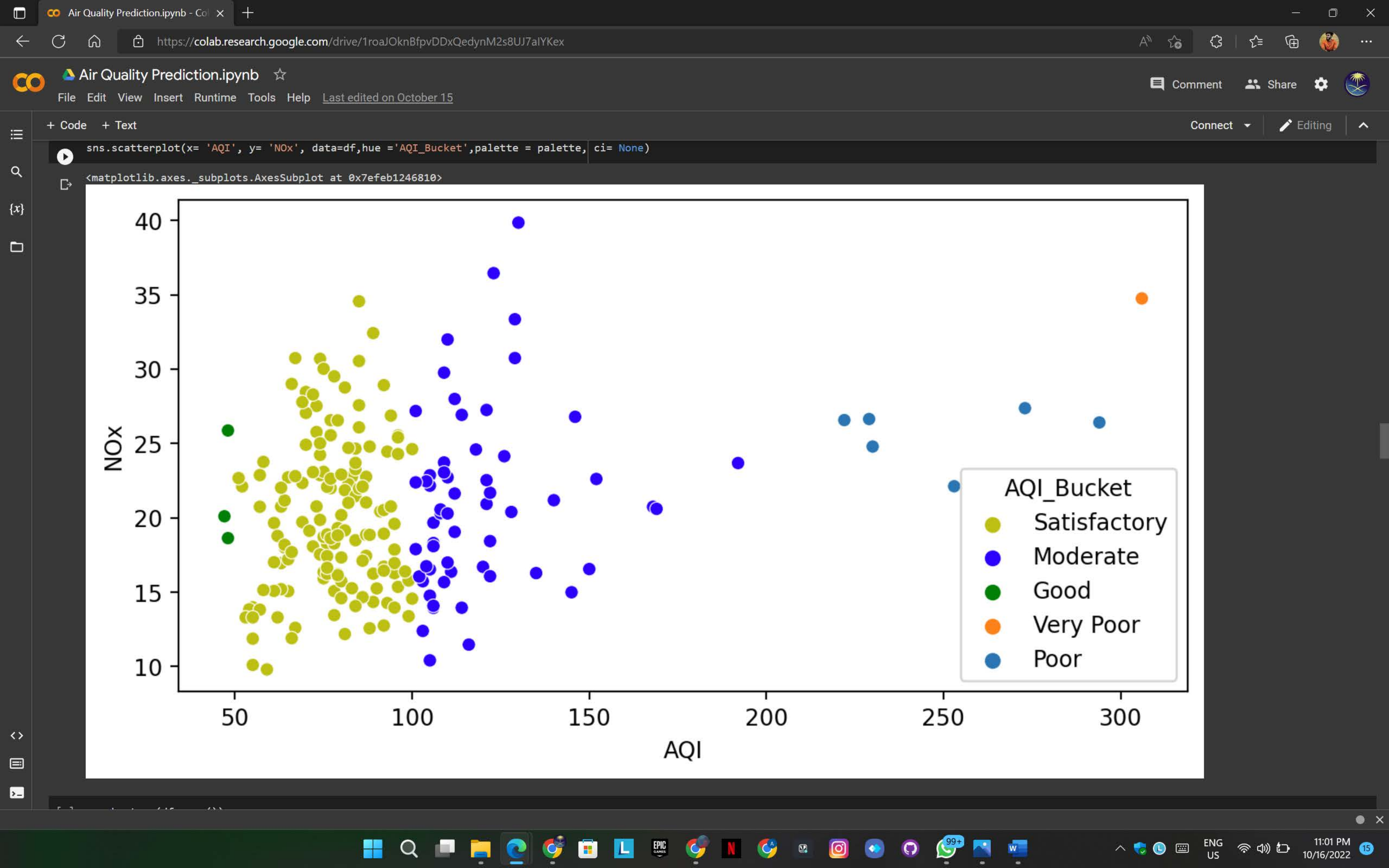
```
palette={'Good': 'g', 'Poor': 'C0', 'Very Poor': 'C1', 'Severe': 'r', 'Moderate': 'b', 'Satisfactory': 'y'}  
sns.scatterplot(x= 'AQI', y= 'PM2.5', data=df, hue = 'AQI_Bucket', palette = palette, ci= None)
```

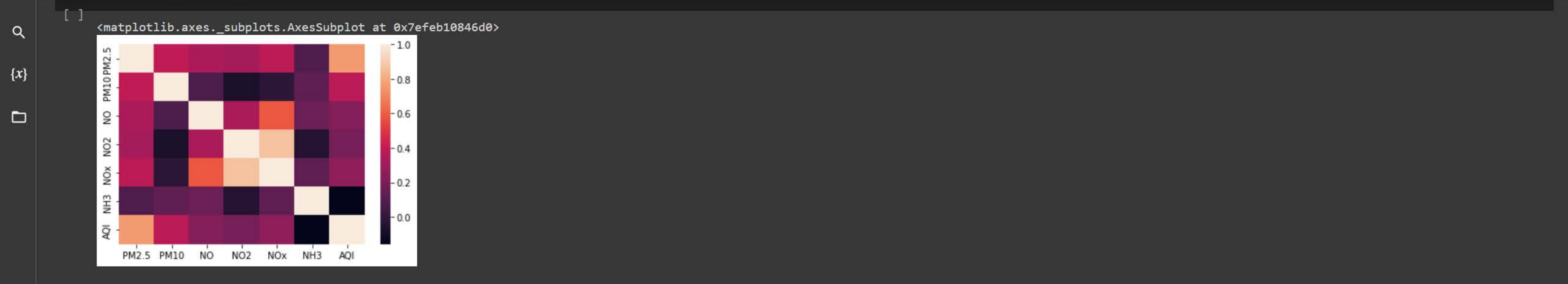
<matplotlib.axes._subplots.AxesSubplot at 0x7efeb12add50>











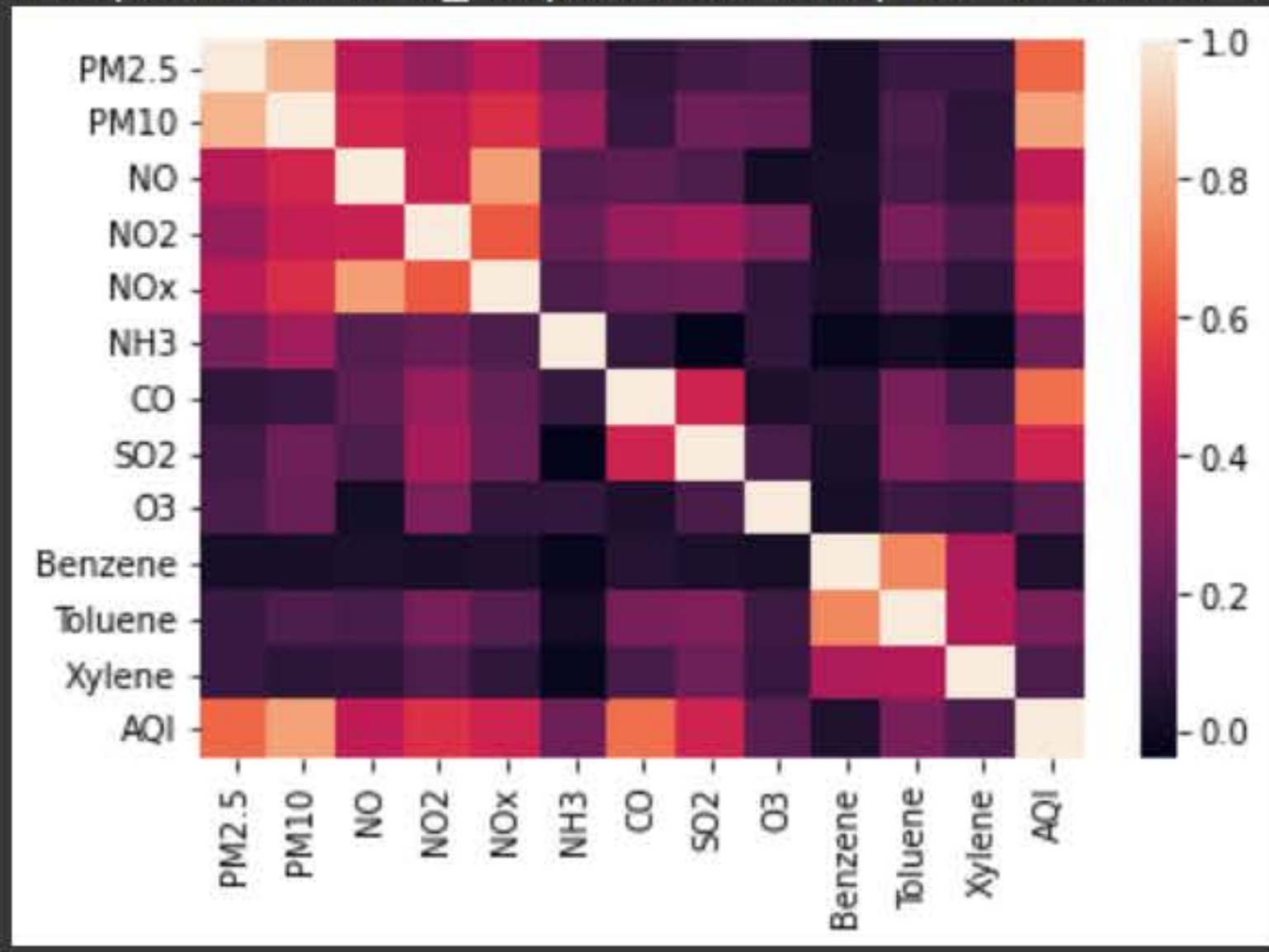
```
df1
```

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	S02	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN
...
29526	Visakhapatnam	2020-06-27	15.02	50.94	7.68	25.06	19.54	12.47	0.47	8.55	23.30	2.24	12.07	0.73	41.0	Good
29527	Visakhapatnam	2020-06-28	24.38	74.09	3.42	26.06	16.53	11.99	0.52	12.72	30.14	0.74	2.21	0.38	70.0	Satisfactory
29528	Visakhapatnam	2020-06-29	22.91	65.73	3.45	29.53	18.33	10.71	0.48	8.42	30.96	0.01	0.01	0.00	68.0	Satisfactory
29529	Visakhapatnam	2020-06-30	16.64	49.97	4.05	29.26	18.80	10.03	0.52	9.84	28.30	0.00	0.00	0.00	54.0	Satisfactory
29530	Visakhapatnam	2020-07-01	15.00	66.00	0.40	26.85	14.05	5.20	0.59	2.10	17.05	NaN	NaN	NaN	50.0	Good

29531 rows × 16 columns

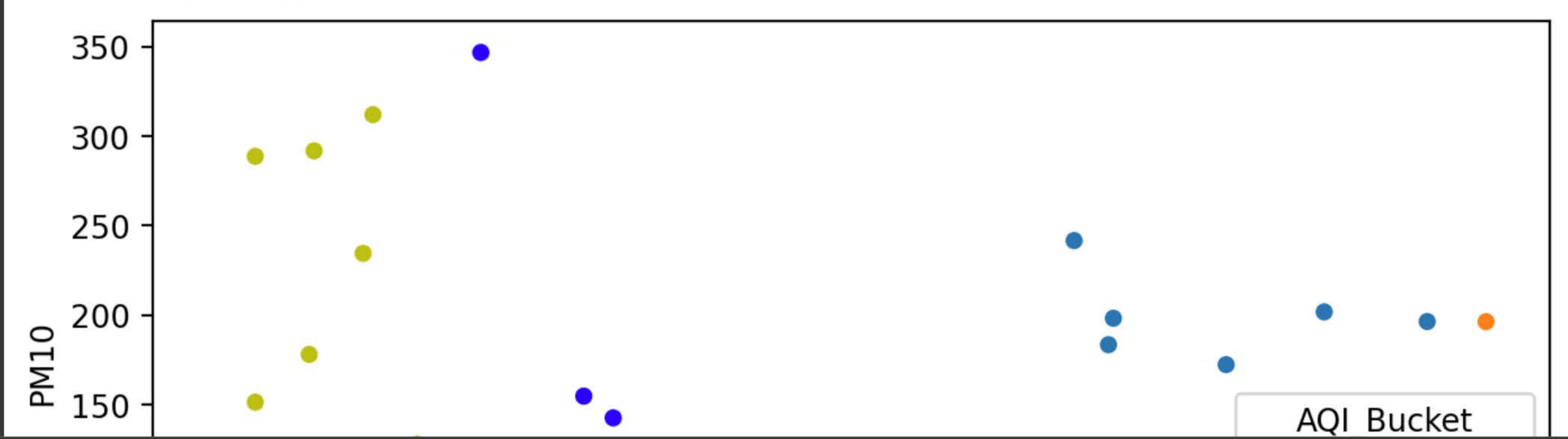
```
[ ] sns.heatmap(df1.corr())
```

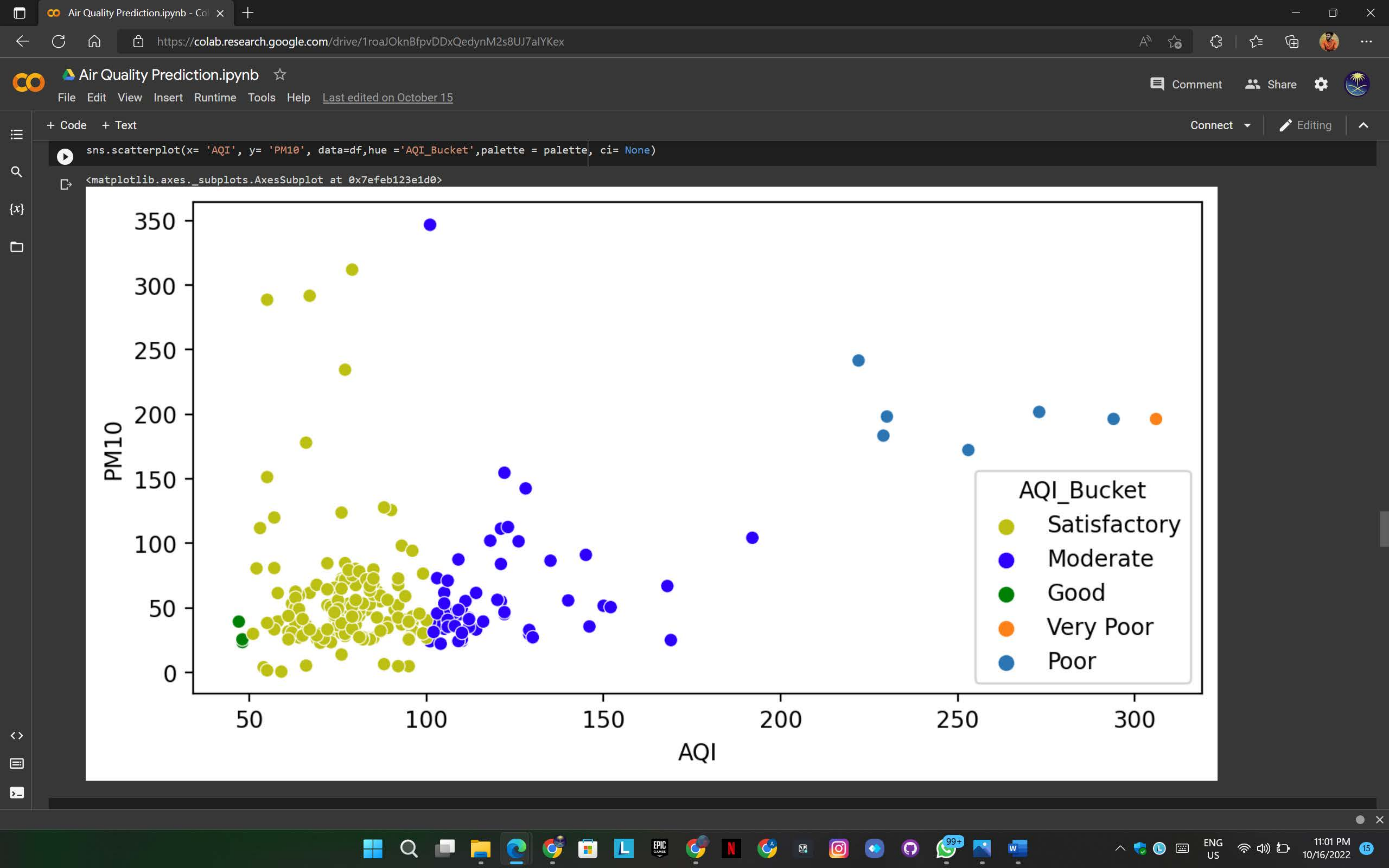

[] <matplotlib.axes._subplots.AxesSubplot at 0x7efeb0faab10>

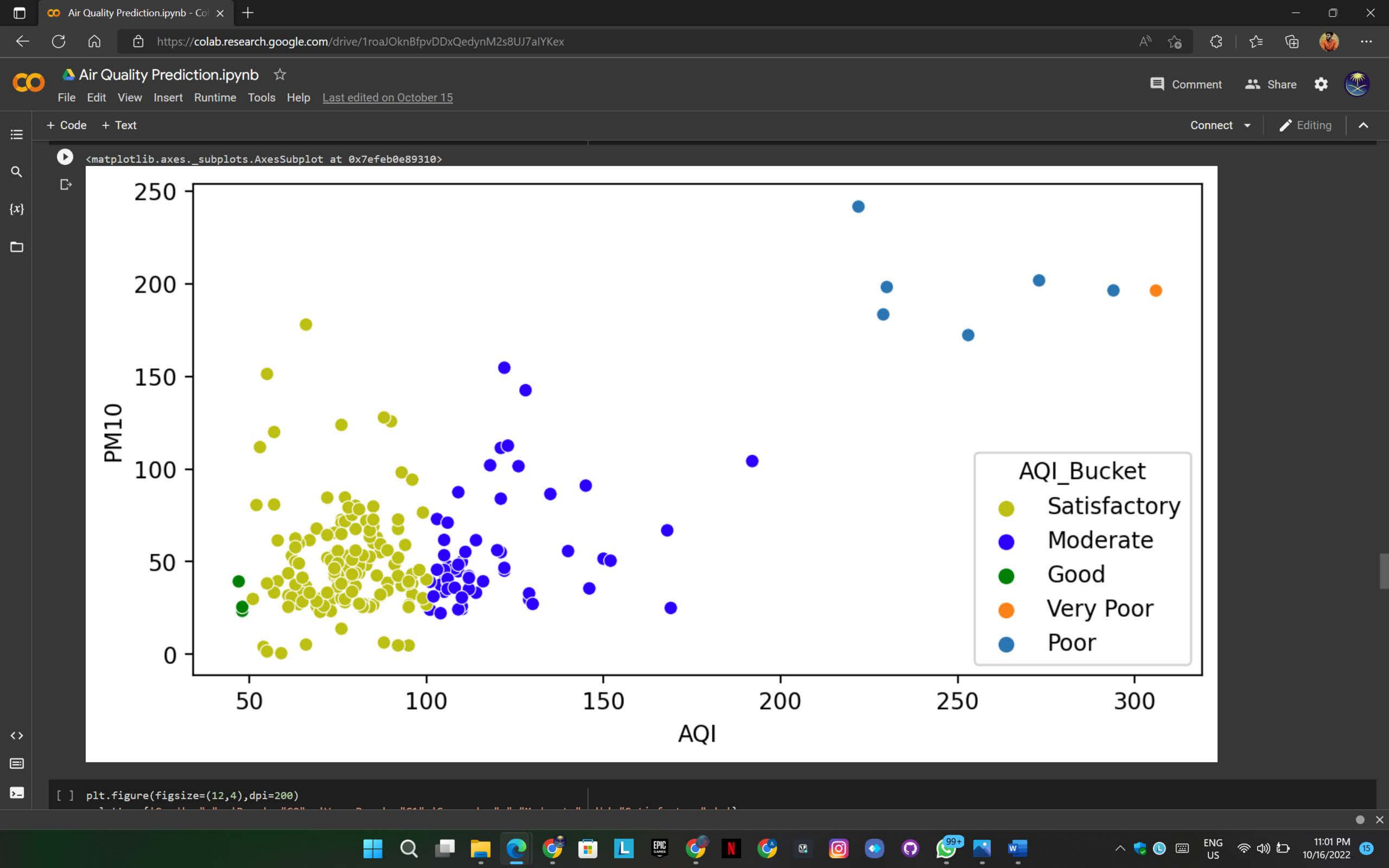


```
plt.figure(figsize=(8,4),dpi=200)
palette ={'Good': "g", 'Poor': "C0", 'Very Poor': "C1",'Severe': "r","Moderate": 'b',"Satisfactory":'y'}
sns.scatterplot(x= 'AQI', y= 'PM10', data=df,hue='AQI_Bucket',palette = palette, ci= None)
```

[] <matplotlib.axes._subplots.AxesSubplot at 0x7efeb123e1d0>

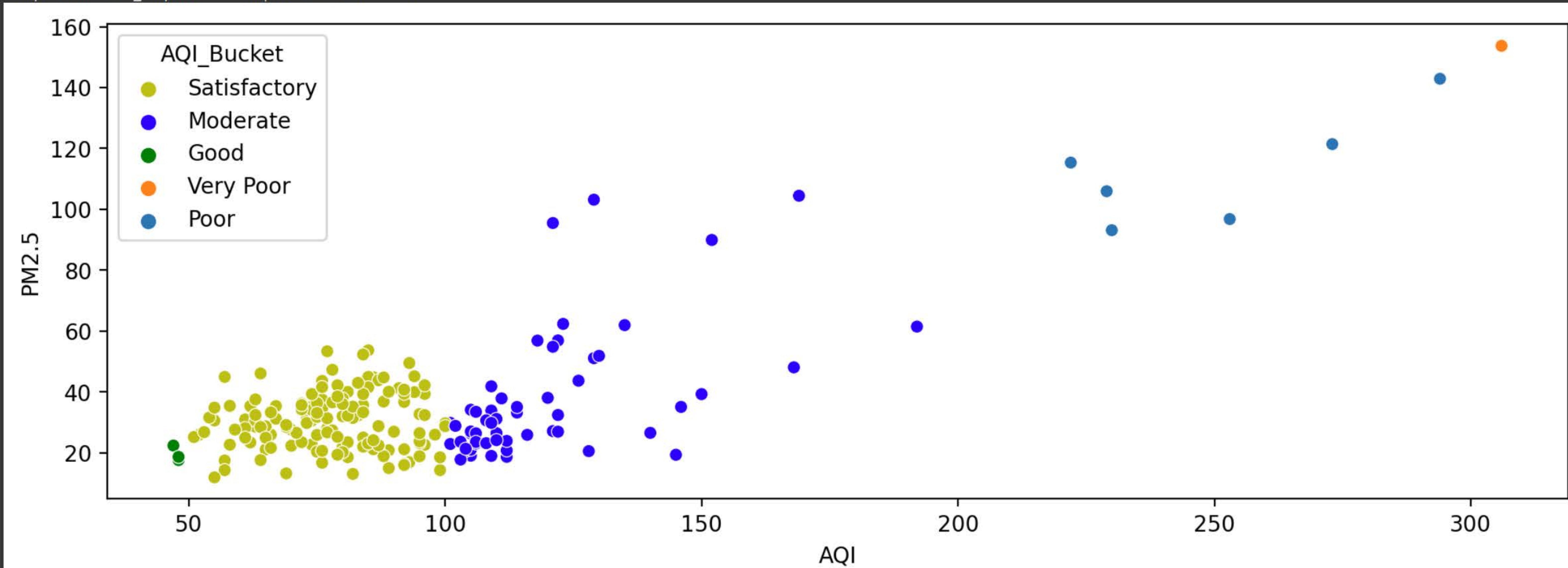







```
plt.figure(figsize=(12,4),dpi=200)
palette={'Good': "g", 'Poor': "C0", 'Very Poor': "C1", 'Severe': "r", "Moderate": 'b', "Satisfactory": 'y'}
sns.scatterplot(x= 'AQI', y= 'PM2.5', data=df, hue = 'AQI_Bucket', palette = palette, ci= None)
```

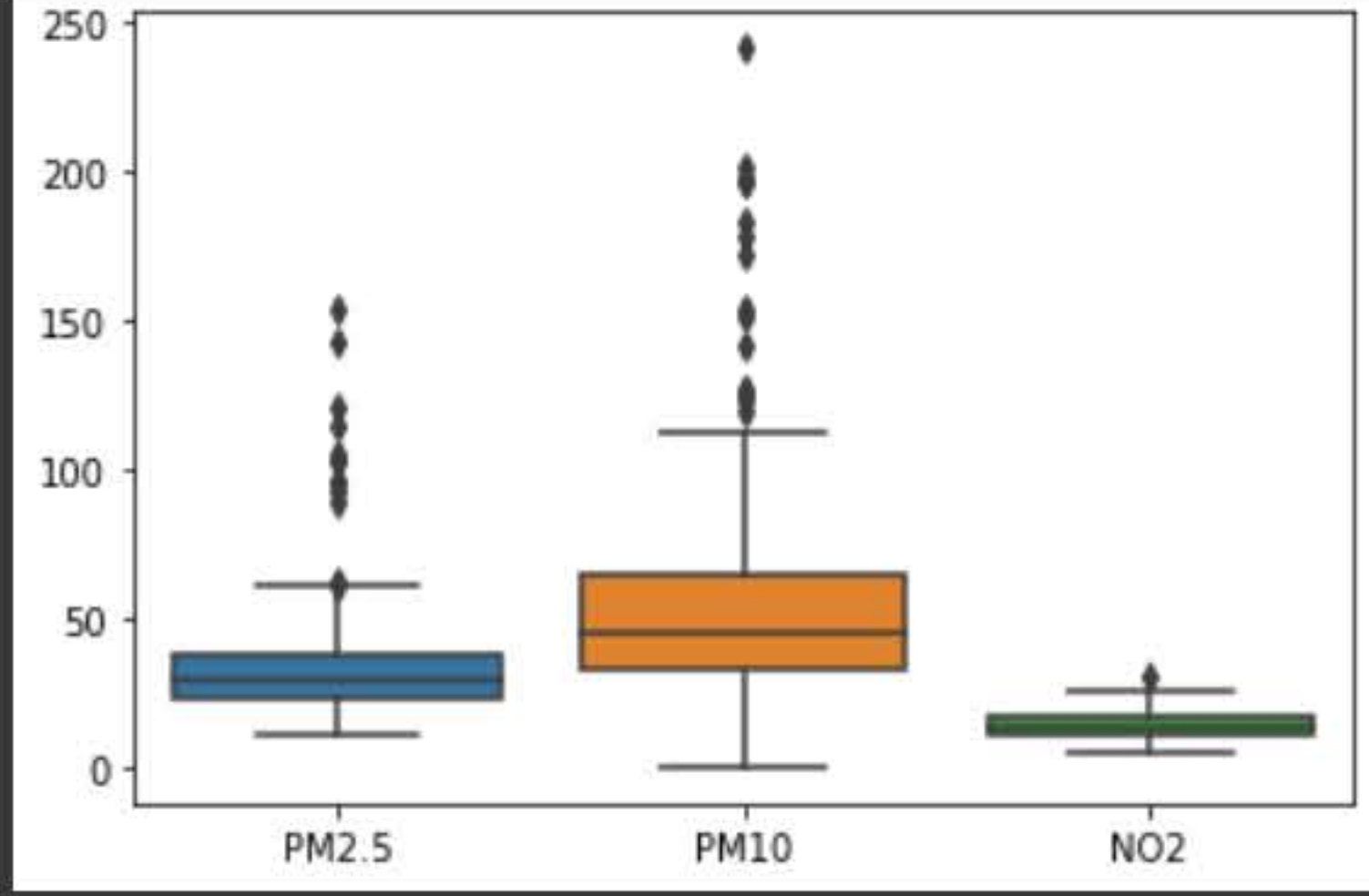
```
[ ] <matplotlib.axes._subplots.AxesSubplot at 0x7efeb0e08d50>
```



```
[ ] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 208 entries, 9464 to 9842
Data columns (total 10 columns):
```


[] <matplotlib.axes._subplots.AxesSubplot at 0x7efeb1515190>



▶ sns.pairplot(data=X)

[>] <seaborn.axisgrid.PairGrid at 0x7efeb0d515d0>

