Homework1

February 3, 2024

1 Homework 1

Dataset contains information on New York City air quality surveillance data.

Air pollution is one of the most important environmental threats to urban populations and while all people are exposed, pollutant emissions, levels of exposure, and population vulnerability vary across neighborhoods. Exposures to common air pollutants have been linked to respiratory and cardiovascular diseases, cancers, and premature deaths. These indicators provide a perspective across time and NYC geographies to better characterize air quality and health in NYC. Data can also be explored online at the Environment and Health Data Portal: http://nyc.gov/health/environmentdata.

https://catalog.data.gov/dataset/?res_format=CSV —-> Air Quality—>https://catalog.data.gov/dataset/air-quality ——-> Download CSV

[71]:
```
import pandas as pd
```

[73]:
```
AirQuality = pd.read_csv ('Air_Quality.csv')
```

[75]:
```
AirQuality.shape
```

[75]: (16218, 12)

[77]:
```
print(AirQuality.head(n = 21))
```

```
    Unique ID  Indicator ID                   Name Measure Measure Info  \
0      172653           375  Nitrogen dioxide (NO2)    Mean          ppb
1      172585           375  Nitrogen dioxide (NO2)    Mean          ppb
2      336637           375  Nitrogen dioxide (NO2)    Mean          ppb
3      336622           375  Nitrogen dioxide (NO2)    Mean          ppb
4      172582           375  Nitrogen dioxide (NO2)    Mean          ppb
5      667327           375  Nitrogen dioxide (NO2)    Mean          ppb
6      172607           375  Nitrogen dioxide (NO2)    Mean          ppb
7      172675           375  Nitrogen dioxide (NO2)    Mean          ppb
8      175345           375  Nitrogen dioxide (NO2)    Mean          ppb
9      176689           375  Nitrogen dioxide (NO2)    Mean          ppb
10     176682           375  Nitrogen dioxide (NO2)    Mean          ppb
11     336507           375  Nitrogen dioxide (NO2)    Mean          ppb
12     740910           375  Nitrogen dioxide (NO2)    Mean          ppb
```

13	175348	375	Nitrogen dioxide (NO2)	Mean	ppb
14	175894	375	Nitrogen dioxide (NO2)	Mean	ppb
15	175895	375	Nitrogen dioxide (NO2)	Mean	ppb
16	175349	375	Nitrogen dioxide (NO2)	Mean	ppb
17	176693	375	Nitrogen dioxide (NO2)	Mean	ppb
18	741006	375	Nitrogen dioxide (NO2)	Mean	ppb
19	550028	375	Nitrogen dioxide (NO2)	Mean	ppb
20	336723	375	Nitrogen dioxide (NO2)	Mean	ppb

	Geo Type Name	Geo Join ID	Geo Place Name	\
0	UHF34	203	Bedford Stuyvesant - Crown Heights	
1	UHF34	203	Bedford Stuyvesant - Crown Heights	
2	UHF34	204	East New York	
3	UHF34	103	Fordham - Bronx Pk	
4	UHF34	104	Pelham - Throgs Neck	
5	UHF34	104	Pelham - Throgs Neck	
6	UHF34	306308	Chelsea-Village	
7	UHF34	306308	Chelsea-Village	
8	UHF42	206	Borough Park	
9	UHF42	206	Borough Park	
10	UHF42	106	High Bridge - Morrisania	
11	UHF42	106	High Bridge - Morrisania	
12	UHF42	106	High Bridge - Morrisania	
13	UHF42	209	Bensonhurst - Bay Ridge	
14	UHF42	209	Bensonhurst - Bay Ridge	
15	UHF42	210	Coney Island - Sheepshead Bay	
16	UHF42	210	Coney Island - Sheepshead Bay	
17	UHF42	210	Coney Island - Sheepshead Bay	
18	UHF42	410	Rockaways	
19	CD	201	Mott Haven and Melrose (CD1)	
20	CD	101	Financial District (CD1)	

	Time Period	Start_Date	Data Value	Message
0	Annual Average 2011	12/01/2010	25.30	NaN
1	Annual Average 2009	12/01/2008	26.93	NaN
2	Annual Average 2015	01/01/2015	19.09	NaN
3	Annual Average 2015	01/01/2015	19.76	NaN
4	Annual Average 2009	12/01/2008	22.83	NaN
5	Annual Average 2020	01/01/2020	16.19	NaN
6	Annual Average 2009	12/01/2008	38.16	NaN
7	Annual Average 2011	12/01/2010	34.96	NaN
8	Winter 2010-11	12/01/2010	30.10	NaN
9	Annual Average 2013	12/01/2012	20.23	NaN
10	Annual Average 2013	12/01/2012	23.73	NaN
11	Winter 2014-15	12/01/2014	26.00	NaN
12	Annual Average 2021	01/01/2021	18.04	NaN
13	Winter 2010-11	12/01/2010	28.44	NaN
14	Summer 2009	06/01/2009	18.95	NaN

```
15          Summer 2009  06/01/2009       15.22      NaN
16       Winter 2010-11  12/01/2010       25.70      NaN
17  Annual Average 2013  12/01/2012       16.36      NaN
18  Annual Average 2021  01/01/2021       11.41      NaN
19  Annual Average 2017  01/01/2017       21.25      NaN
20       Winter 2014-15  12/01/2014       30.40      NaN
```

[79]:
```
print(AirQuality.tail())
```

```
       Unique ID  Indicator ID                                   Name  \
16213     130750           647           Outdoor Air Toxics - Formaldehyde
16214     130780           647           Outdoor Air Toxics - Formaldehyde
16215     131020           652  Cardiac and respiratory deaths due to Ozone
16216     131026           652  Cardiac and respiratory deaths due to Ozone
16217     325247           643                Annual vehicle miles traveled

                          Measure Measure Info Geo Type Name  Geo Join ID  \
16213  Annual average concentration        µg/m3         UHF42          211
16214  Annual average concentration        µg/m3       Borough            5
16215         Estimated annual rate  per 100,000         UHF42          504
16216         Estimated annual rate  per 100,000       Borough            5
16217                million miles      per km2            CD          107

                 Geo Place Name Time Period  Start_Date  Data Value  Message
16213    Williamsburg - Bushwick        2005  01/01/2005         3.1      NaN
16214              Staten Island        2005  01/01/2005         2.3      NaN
16215  South Beach - Tottenville   2005-2007  01/01/2005         7.5      NaN
16216              Staten Island   2005-2007  01/01/2005         7.8      NaN
16217      Upper West Side (CD7)        2016  01/01/2016        50.0      NaN
```

[81]:
```
print(AirQuality.describe())
```

```
          Unique ID  Indicator ID   Geo Join ID    Data Value  Message
count   16218.000000  16218.000000  1.621800e+04  16218.000000      0.0
mean   372730.417746    427.803613  6.097103e+05     19.975917      NaN
std    215507.613560    110.921411  7.893388e+06     21.322349      NaN
min    121644.000000    365.000000  1.000000e+00      0.000000      NaN
25%    173211.250000    365.000000  2.020000e+02      9.050000      NaN
50%    325262.500000    375.000000  3.030000e+02     15.300000      NaN
75%    605270.750000    386.000000  4.040000e+02     26.037500      NaN
max    799868.000000    661.000000  1.051061e+08    424.700000      NaN
```

[83]:
```
print(AirQuality.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16218 entries, 0 to 16217
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
```

```
---  ------          --------------  -----
 0   Unique ID       16218 non-null  int64
 1   Indicator ID    16218 non-null  int64
 2   Name            16218 non-null  object
 3   Measure         16218 non-null  object
 4   Measure Info    16218 non-null  object
 5   Geo Type Name   16218 non-null  object
 6   Geo Join ID     16218 non-null  int64
 7   Geo Place Name  16218 non-null  object
 8   Time Period     16218 non-null  object
 9   Start_Date      16218 non-null  object
 10  Data Value      16218 non-null  float64
 11  Message         0 non-null      float64
dtypes: float64(2), int64(3), object(7)
memory usage: 1.5+ MB
None
```

[85]:
```
AirQuality['index'] = pd.Series(range(0,16217))
```

[87]:
```
print(AirQuality.head())
```

```
   Unique ID  Indicator ID                     Name Measure Measure Info  \
0     172653           375  Nitrogen dioxide (NO2)    Mean          ppb
1     172585           375  Nitrogen dioxide (NO2)    Mean          ppb
2     336637           375  Nitrogen dioxide (NO2)    Mean          ppb
3     336622           375  Nitrogen dioxide (NO2)    Mean          ppb
4     172582           375  Nitrogen dioxide (NO2)    Mean          ppb

  Geo Type Name  Geo Join ID                     Geo Place Name  \
0         UHF34          203  Bedford Stuyvesant - Crown Heights
1         UHF34          203  Bedford Stuyvesant - Crown Heights
2         UHF34          204                       East New York
3         UHF34          103                  Fordham - Bronx Pk
4         UHF34          104                Pelham - Throgs Neck

           Time Period  Start_Date  Data Value  Message  index
0  Annual Average 2011  12/01/2010       25.30      NaN    0.0
1  Annual Average 2009  12/01/2008       26.93      NaN    1.0
2  Annual Average 2015  01/01/2015       19.09      NaN    2.0
3  Annual Average 2015  01/01/2015       19.76      NaN    3.0
4  Annual Average 2009  12/01/2008       22.83      NaN    4.0
```

[89]:
```
AirQuality.shape
```

[89]:
```
(16218, 13)
```

[91]:
```
print(AirQuality.head())
```

```
   Unique ID  Indicator ID                   Name Measure Measure Info  \
0     172653           375  Nitrogen dioxide (NO2)    Mean          ppb
1     172585           375  Nitrogen dioxide (NO2)    Mean          ppb
2     336637           375  Nitrogen dioxide (NO2)    Mean          ppb
3     336622           375  Nitrogen dioxide (NO2)    Mean          ppb
4     172582           375  Nitrogen dioxide (NO2)    Mean          ppb

  Geo Type Name  Geo Join ID                    Geo Place Name  \
0         UHF34          203  Bedford Stuyvesant - Crown Heights
1         UHF34          203  Bedford Stuyvesant - Crown Heights
2         UHF34          204                       East New York
3         UHF34          103                  Fordham - Bronx Pk
4         UHF34          104                Pelham - Throgs Neck

           Time Period  Start_Date  Data Value  Message  index
0  Annual Average 2011  12/01/2010       25.30      NaN    0.0
1  Annual Average 2009  12/01/2008       26.93      NaN    1.0
2  Annual Average 2015  01/01/2015       19.09      NaN    2.0
3  Annual Average 2015  01/01/2015       19.76      NaN    3.0
4  Annual Average 2009  12/01/2008       22.83      NaN    4.0
```

[93]:
```
print(AirQuality["Data Value"].unique())
```

```
[25.3  26.93 19.09 … 45.43 70.7  29.6 ]
```

[95]:
```
import numpy as np
```

[97]:
```
np.NaN
```

[97]:
```
nan
```

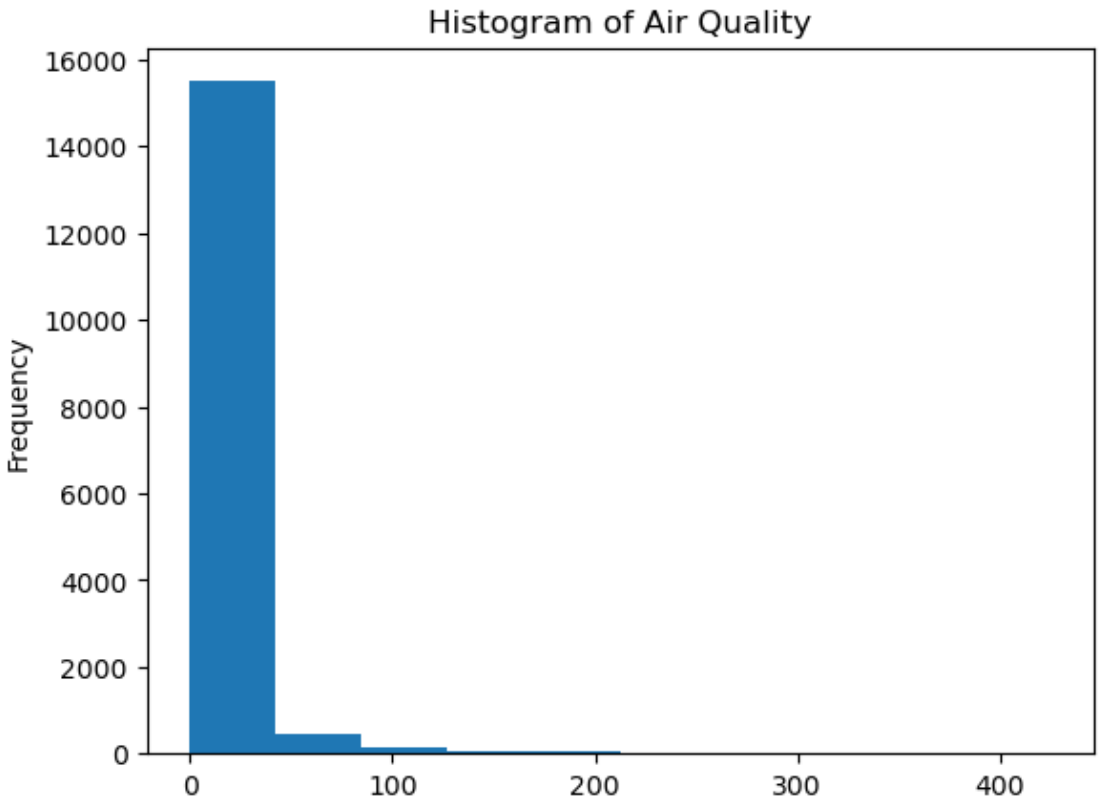
[205]:
```
AirQuality['Data Value'] = AirQuality['Data Value'].replace({70.7: np.NaN})
```

[206]:
```
print(AirQuality['Data Value'].unique())
```

```
[25.3  26.93 19.09 … 45.43   nan 29.6 ]
```

[207]:
```
AirQuality['Data Value'].plot(kind = 'hist', title = 'Histogram of Air Quality')
```

[207]:
```
<Axes: title={'center': 'Histogram of Air Quality'}, ylabel='Frequency'>
```



[208]:
```
AirQuality["Data Value"] = AirQuality['Data Value']
```

[209]:
```
print(AirQuality.tail())
```

```
       Unique ID  Indicator ID                                        Name  \
16213     130750           647           Outdoor Air Toxics - Formaldehyde
16214     130780           647           Outdoor Air Toxics - Formaldehyde
16215     131020           652  Cardiac and respiratory deaths due to Ozone
16216     131026           652  Cardiac and respiratory deaths due to Ozone
16217     325247           643                Annual vehicle miles traveled

                          Measure Measure Info Geo Type Name  Geo Join ID  \
16213  Annual average concentration        µg/m3         UHF42          211
16214  Annual average concentration        µg/m3       Borough            5
16215         Estimated annual rate  per 100,000         UHF42          504
16216         Estimated annual rate  per 100,000       Borough            5
16217                 million miles      per km2            CD          107

                Geo Place Name Time Period  Start_Date  Data Value  Message  \
16213    Williamsburg - Bushwick        2005  01/01/2005         3.1      NaN
16214             Staten Island        2005  01/01/2005         2.3      NaN
```

```
16215  South Beach - Tottenville   2005-2007  01/01/2005         7.5      NaN
16216              Staten Island   2005-2007  01/01/2005         7.8      NaN
16217      Upper West Side (CD7)         2016  01/01/2016        50.0      NaN

         index  Quality_z
16213  16213.0        NaN
16214  16214.0        NaN
16215  16215.0        NaN
16216  16216.0        NaN
16217      NaN        NaN
```

[210]:
```
print(AirQuality['Data Value'].unique())
```

```
[25.3  26.93 19.09 … 45.43   nan 29.6 ]
```

[211]:
```
Dictionary = {"Data Value"}
```

[ ]:
```

```

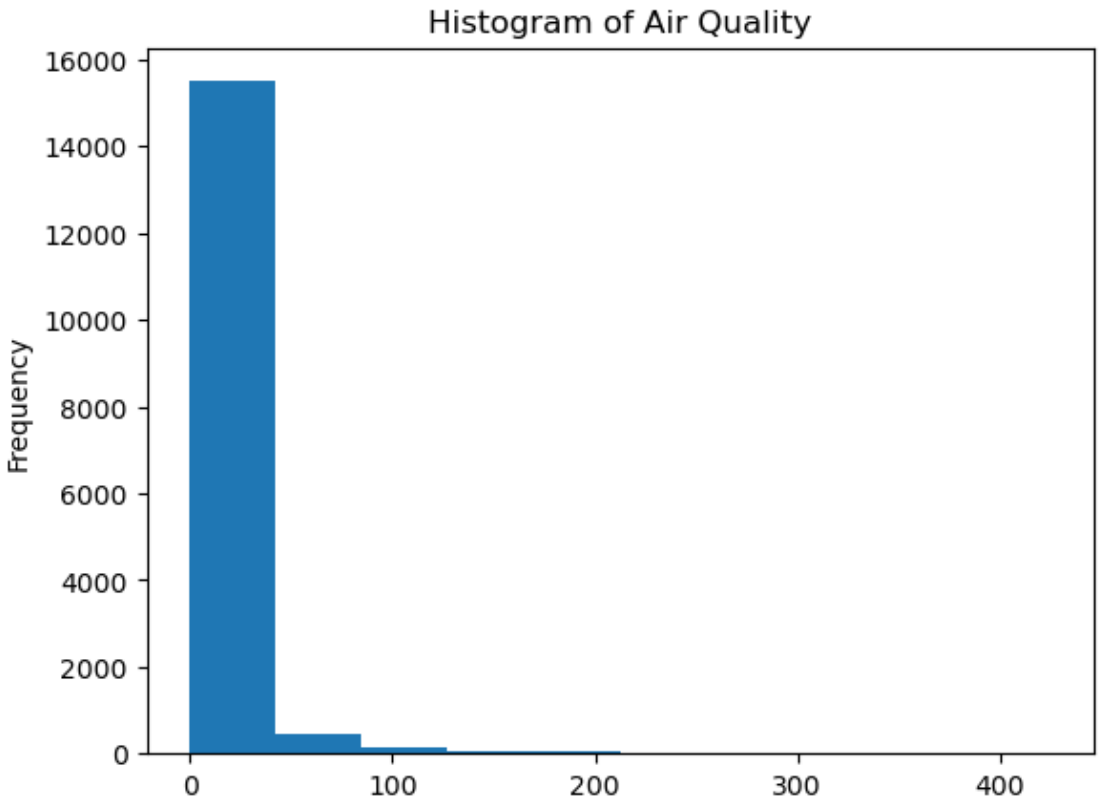
[213]:
```
print(AirQuality.head())
```

```
   Unique ID  Indicator ID                    Name Measure Measure Info  \
0     172653           375  Nitrogen dioxide (NO2)    Mean          ppb
1     172585           375  Nitrogen dioxide (NO2)    Mean          ppb
2     336637           375  Nitrogen dioxide (NO2)    Mean          ppb
3     336622           375  Nitrogen dioxide (NO2)    Mean          ppb
4     172582           375  Nitrogen dioxide (NO2)    Mean          ppb

  Geo Type Name  Geo Join ID                      Geo Place Name  \
0         UHF34          203  Bedford Stuyvesant - Crown Heights
1         UHF34          203  Bedford Stuyvesant - Crown Heights
2         UHF34          204                        East New York
3         UHF34          103                  Fordham - Bronx Pk
4         UHF34          104                Pelham - Throgs Neck

            Time Period  Start_Date  Data Value  Message  index  Quality_z
0  Annual Average 2011  12/01/2010       25.30      NaN    0.0        NaN
1  Annual Average 2009  12/01/2008       26.93      NaN    1.0        NaN
2  Annual Average 2015  01/01/2015       19.09      NaN    2.0        NaN
3  Annual Average 2015  01/01/2015       19.76      NaN    3.0        NaN
4  Annual Average 2009  12/01/2008       22.83      NaN    4.0        NaN
```

[214]:
```
from scipy import stats
```

[215]:
```
AirQuality['Data Value'].plot(kind = 'hist', title = 'Histogram of Air Quality')
```

[215]:
```
<Axes: title={'center': 'Histogram of Air Quality'}, ylabel='Frequency'>
```



[216]:
```
AirQuality["Quality_z"] = stats.zscore(AirQuality['Data Value'])
```

[217]:
```
print(AirQuality['Quality_z'].head(n=50))
```

```
0    NaN
1    NaN
2    NaN
3    NaN
4    NaN
5    NaN
6    NaN
7    NaN
8    NaN
9    NaN
10   NaN
11   NaN
12   NaN
13   NaN
14   NaN
15   NaN
16   NaN
```

```
17   NaN
18   NaN
19   NaN
20   NaN
21   NaN
22   NaN
23   NaN
24   NaN
25   NaN
26   NaN
27   NaN
28   NaN
29   NaN
30   NaN
31   NaN
32   NaN
33   NaN
34   NaN
35   NaN
36   NaN
37   NaN
38   NaN
39   NaN
40   NaN
41   NaN
42   NaN
43   NaN
44   NaN
45   NaN
46   NaN
47   NaN
48   NaN
49   NaN
Name: Quality_z, dtype: float64
```

[218]:
```
AirQuality.query('Quality_z > 3 | Quality_z < -3')
```

[218]:
```
Empty DataFrame
Columns: [Unique ID, Indicator ID, Name, Measure, Measure Info, Geo Type Name,
Geo Join ID, Geo Place Name, Time Period, Start_Date, Data Value, Message,
index, Quality_z]
Index: []
```

[219]:
```
AirQuality_outliers = AirQuality.query('Quality_z > 3 | Quality_z < -3')
```

[220]:
```
print(AirQuality.head())
```

```
   Unique ID  Indicator ID                   Name Measure Measure Info  \
```

```
0     172653           375  Nitrogen dioxide (NO2)    Mean          ppb
1     172585           375  Nitrogen dioxide (NO2)    Mean          ppb
2     336637           375  Nitrogen dioxide (NO2)    Mean          ppb
3     336622           375  Nitrogen dioxide (NO2)    Mean          ppb
4     172582           375  Nitrogen dioxide (NO2)    Mean          ppb

  Geo Type Name  Geo Join ID                     Geo Place Name  \
0         UHF34          203  Bedford Stuyvesant - Crown Heights
1         UHF34          203  Bedford Stuyvesant - Crown Heights
2         UHF34          204                       East New York
3         UHF34          103                 Fordham - Bronx Pk
4         UHF34          104               Pelham - Throgs Neck

            Time Period  Start_Date  Data Value  Message  index  Quality_z
0  Annual Average 2011  12/01/2010       25.30      NaN    0.0        NaN
1  Annual Average 2009  12/01/2008       26.93      NaN    1.0        NaN
2  Annual Average 2015  01/01/2015       19.09      NaN    2.0        NaN
3  Annual Average 2015  01/01/2015       19.76      NaN    3.0        NaN
4  Annual Average 2009  12/01/2008       22.83      NaN    4.0        NaN
```

[221]:
```
Quality_sort = AirQuality.sort_values(['Quality_z'], ascending = False)
```

[222]:
```
print(Quality_sort[['Data Value',]].head(n=15))
```

```
    Data Value
0        25.30
1        26.93
2        19.09
3        19.76
4        22.83
5        16.19
6        38.16
7        34.96
8        30.10
9        20.23
10       23.73
11       26.00
12       18.04
13       28.44
14       18.95
```

[ ]:
```

```

[ ]:
```

```