In [1]:
```
import pandas as pd
path = './Datasets/'
USE_CSV = False
if not USE_CSV:
    air_quality = pd.read_pickle(path+'air_quality.pkl')
else:
    air_quality = pd.read_csv(path+'air_quality.csv',dtype = {'PM2.5_cat'
```

Q1: show the average PM2.5 over months

In [2]:
```
air_quality.columns
```

Out[2]:
```
Index(['date_time', 'PM2.5', 'PM10', 'SO2', 'NO2', 'CO', 'O3', 'TEMP',
'PRES',
       'DEWP', 'RAIN', 'wd', 'WSPM', 'station', 'year', 'month', 'day',
'hour',
       'quarter', 'day_of_week_num', 'day_of_week_name', 'time_until_202
2',
       'time_until_2022_days', 'time_until_2022_weeks', 'prior_2016_in
d',
       'PM2.5_cat', 'TEMP_category'],
      dtype='object')
```

In [3]:
```
air_quality[['PM2.5','month']].groupby(by='month').mean()
```

Out[3]:

	PM2.5
month	
1	98.547996
2	83.785755
3	98.300096
4	74.878336
5	65.205977
6	71.619663
7	75.996913
8	56.235265
9	64.049654
10	95.848617
11	101.436812
12	115.889403

Q2

`air_quality[['TEMP']].groupby(by='quarter').mean()`

What's wrong here?

Answer: because we didn't put the quarter feature in the selected DataFrame[air_quality[['TEMP']]]

Q3

Use lambda function to check the span of PM2.5 and PM10 with groupby.agg()

In [4]:

```
air_quality[['quarter', 'PM2.5', 'PM10']].groupby(by='quarter').agg(lambd
```

Out[4]:

	PM2.5	PM10
quarter		
1	818.0	992.0
2	531.0	984.0
3	372.0	860.0
4	738.0	791.0

In [5]:

```
def max_min(col):
    return col.max() - col.min()
```

Q4

In [6]:

```
air_quality[["PM2.5", "RAIN", "TEMP", "quarter"]].groupby(by="quarter").a
    {"PM2.5": "describe", "RAIN": "sum", "TEMP": max_min}
)
air_quality[['month','PM2.5','RAIN','TEMP']].groupby(by='month').agg(
    {'PM2.5': ['max','mean'],
     'RAIN': ['min', 'sum'],
     'TEMP': max_min }
)
```

Out[6]:

	PM2.5			RAIN	TEMP
	max	mean	min	sum	max_min
month					
1	767.0	98.547996	0.0	2.0	29.4
2	821.0	83.785755	0.0	82.3	27.0
3	520.0	98.300096	0.0	51.4	33.5
4	533.0	74.878336	0.0	202.8	32.0
5	408.0	65.205977	0.0	342.0	36.0
6	525.0	71.619663	0.0	925.7	23.1
7	375.0	75.996913	0.0	2135.4	22.3
8	283.0	56.235265	0.0	729.9	22.0
9	311.0	64.049654	0.0	1152.7	27.5
10	465.0	95.848617	0.0	392.0	29.7
11	685.0	101.436812	0.0	146.9	31.2
12	741.0	115.889403	0.0	8.2	26.4

Q5

In [7]:

```
student = pd.read_csv(path + "student.csv")
pd.pivot_table(student,
               index = ['sex', 'internet'],
               values = 'score'
                )
```

Out[7]:

		score
sex	internet	
F	no	9.184211
	yes	10.141176
M	no	9.714286
	yes	11.125786

Q6

In [8]:

```
pd.pivot_table(
    student,
    index=['sex',"internet"],
    values=["age", "score"],
    aggfunc={"age": ["max", "min",'mean'], "score": max_min},

)
```

Out[8]:

		age			score
		max	mean	min	max_min
sex	internet				
F	no	19	17.078947	15	18
	yes	20	16.652941	15	19
M	no	21	16.928571	15	18
	yes	22	16.610063	15	20

In [9]:

```
pd.pivot_table(student,
               index=[ "studytime"],
               values=['age',"score"],
               aggfunc = {'age' : ['max', 'min'],
                          'score': 'median'},
               columns = 'sex'
               )
```

Out[9]:

	age				score	
	max		min		median	
sex	F	M	F	M	F	M
studytime						
1. <2 hours	18	22	15	15	11.0	10.0
2. 2 - 5 hours	19	20	15	15	10.0	11.0
3. 5 - 10 hours	20	18	15	15	11.0	14.0
4. >10 hours	19	18	15	15	11.0	12.5

Q7

In [10]:

```
pd.pivot_table(air_quality,
               index=[ "station",'quarter'],
               values=['PM2.5',"wd"],
               aggfunc = {'PM2.5' : 'mean',
                        #    'wd' : pd.Series.mode
                            'wd' : lambda x : x.mode()},
           #    columns = 'sex'
               )
```

Out[10]:

		PM2.5	wd
station	quarter		
Dongsi	1	93.588818	NE
	2	71.323896	SW
	3	67.782695	ENE
	4	106.970248	ENE
Gucheng	1	96.199428	NW
	2	71.678643	SSW
	3	63.241872	N
	4	104.526163	N
Tiantan	1	92.363923	NE
	2	68.861329	SW
	3	64.771197	ENE
	4	100.806901	ENE

your findings here:

Answer

- there is no difference between station
-

Q8

In [14]:

```
pd.pivot_table(
    air_quality,
    index=["station", "wd"],
    columns="quarter",
    values=["PM2.5", "RAIN"],
    aggfunc="mean",
)
```

Out[14]:

		PM2.5					
	quarter	1	2	3	4	1	
station	wd						
Dongsi	E	140.388974	84.696234	75.788126	154.685030	0.008576	0.12
	ENE	124.255266	73.881776	68.913477	144.769452	0.002602	0.09
	ESE	141.971983	89.249390	84.234797	140.986689	0.015733	0.06
	N	50.373967	54.446602	50.389831	95.494949	0.008678	0.10
	NE	109.084091	65.551102	59.304825	123.646941	0.000909	0.04
	NNE	83.196562	57.521994	53.169557	98.993182	0.003152	0.10
	NNW	29.011673	45.190083	47.786047	40.968883	0.004280	0.13
	NW	33.411765	37.955340	36.985663	30.053254	0.007250	0.09
	S	108.116592	89.624390	80.923077	117.140625	0.005381	0.03
	SE	124.059459	86.913223	90.964419	123.921127	0.008108	0.03
	SSE	120.509434	88.874016	88.294751	126.095455	0.009906	0.08
	SSW	104.113095	81.439739	76.962213	93.407258	0.011310	0.02
	SW	91.983425	69.395702	68.681913	97.578036	0.000921	0.01
	W	96.135135	65.299595	53.459893	81.763485	0.000541	0.04
	WNW	48.059441	44.350181	42.791667	45.749347	0.000000	0.02
	WSW	93.577437	69.370000	54.445946	84.320395	0.000000	0.02
Gucheng	E	117.195122	77.055164	56.969855	107.279365	0.008537	0.04
	ENE	106.619355	76.626398	57.628947	111.227166	0.001290	0.07
	ESE	108.786550	79.212500	67.730272	129.788779	0.011404	0.07
	N	85.341351	72.702914	68.372682	121.751313	0.006351	0.08
	NE	95.757909	72.699703	58.485677	104.191744	0.006190	0.10
	NNE	91.473520	62.073913	51.498542	82.774697	0.007788	0.06
	NNW	66.119816	53.664804	50.510204	82.292308	0.000230	0.02
	NW	70.620447	55.271739	50.186161	96.270156	0.002233	0.05
	S	120.243762	82.053799	75.040678	126.536993	0.008637	0.04
	SE	115.097059	82.078125	74.402439	126.185771	0.011176	0.06
	SSE	119.716129	87.134703	79.131298	126.271242	0.004839	0.02
	SSW	101.126984	73.674825	68.627286	119.757871	0.004762	0.01
	SW	96.810624	77.075251	73.234000	101.450000	0.003926	0.03
	W	110.838922	64.577869	58.884752	85.491468	0.000798	0.07
	WNW	74.438017	53.360000	53.328358	74.430912	0.001983	0.05
	WSW	111.535385	70.111872	63.936404	114.357895	0.003385	0.03

		PM2.5				
	quarter	**1**	**2**	**3**	**4**	**1**
station	**wd**					
Tiantan	**E**	136.331579	83.585695	72.509052	143.099202	0.013421	0.11
	ENE	122.841943	74.089286	67.515805	134.898804	0.003863	0.09
	ESE	138.237817	87.682216	81.561254	134.608838	0.018324	0.06
	N	51.577491	59.202671	50.857798	84.624724	0.007749	0.09
	NE	106.080330	65.923077	56.680116	116.870715	0.000824	0.04
	NNE	84.017722	55.515670	52.550442	96.185714	0.002785	0.12
	NNW	30.027972	40.862360	43.696498	41.892944	0.003846	0.13
	NW	34.615285	34.389362	35.434084	33.624161	0.006865	0.10
	S	105.344569	82.751773	74.491903	116.431298	0.004494	0.03
	SE	120.551807	81.417154	84.535714	119.956403	0.011566	0.03
	SSE	119.698745	82.895522	83.067873	127.488106	0.012971	0.09
	SSW	99.629333	77.212598	71.773462	88.926923	0.023733	0.02
	SW	88.196837	63.830380	63.584699	91.475524	0.000879	0.01
	W	93.973822	64.088983	50.877934	76.175182	0.000524	0.05
	WNW	48.903427	40.664179	41.390135	45.821826	0.000000	0.03
	WSW	90.612466	64.133470	53.929224	83.929577	0.000000	0.02

Insight here

1. Significant Seasonal Impact on PM2.5

   - Winter pollution is the worst: All stations recorded the highest PM2.5 values in Q1 and Q4 (winter).
   - For example, at Dongsi station, wind direction E: Q1=140.39, Q4=154.69 vs Q2=84.70, Q3=75.79
   - Summer air quality is the best: PM2.5 values were lowest in Q2 and Q3.