DATA SCIENCE DAY2

In [38]:

```
#exercise 1
#1
import pandas as pd
a=pd.read_excel('coalpublic2013.xlsx')
print(a.dtypes)
```

```
Year                          int64
MSHA ID                       int64
Mine Name                    object
Mine State                   object
Mine County                  object
Mine Status                  object
Mine Type                    object
Company Type                 object
Operation Type               object
Operating Company            object
Operating Company Address    object
Union Code                   object
Coal Supply Region           object
Production (short tons)       int64
Average Employees             int64
Labor Hours                   int64
dtype: object
--------------------------------------------------
   MSHA ID                         Mine Status
0   103381  Active, men working, not producing
1   103404               Permanently abandoned
2   100759  Active, men working, not producing
3   103246                              Active
4   103451                              Active
```

In [62]:

```
a=pd.read_excel('coalpublic2013.xlsx',usecols=[1,5])
print(a.head(5))
```

```
   MSHA ID                         Mine Status
0   103381  Active, men working, not producing
1   103404               Permanently abandoned
2   100759  Active, men working, not producing
3   103246                              Active
4   103451                              Active
```

In [57]:

```
#2
a=pd.read_excel('coalpublic2013.xlsx')
print('sum:',a['Production (short tons)'].sum())
print('mean:',a['Production (short tons)'].mean())
print('max:',a['Production (short tons)'].max())
print('min:',a['Production (short tons)'].min())
```

```
sum: 984841779
mean: 679201.2268965517
max: 111005549
min: 0
```

In [75]:

```
#3
import numpy as np
a=pd.read_excel('coalpublic2013.xlsx')
a.insert(6,"new_column",np.nan)
print(a.head(5))
```

```
   Year  MSHA ID                       Mine Name  ... Production (short ton
s) Average Employees Labor Hours
0  2013   103381            Tacoa Highwall Miner  ...                   5600
4                 10       22392
1  2013   103404                Reid School Mine  ...                   2880
7                 18       28447
2  2013   100759  North River #1 Underground Min  ...                 144011
5                183      474784
3  2013   103246                      Bear Creek  ...                   8758
7                 13       29193
4  2013   103451                     Knight Mine  ...                  14749
9                 27       46393

[5 rows x 17 columns]
```

In [76]:

```
#exercise 2
#1
a=pd.read_excel('coalpublic2013.xlsx',skiprows=20)
print(a)
```

```
      2013   102976 Piney Woods Preparation Plant          Alabama  ...  App
alachia Southern        0    9    23193
0      2013   103380                        Calera          Alabama  ...  App
alachia Southern        0    6    12621
1      2013   103380                        Calera          Alabama  ...  App
alachia Southern        0    1     1402
2      2013   103422               Clark No 1 Mine          Alabama  ...  App
alachia Southern   122727   61   140250
3      2013   103467           Helena Surface Mine          Alabama  ...  App
alachia Southern    59664   16    30539
4      2013   101247                     No 4 Mine          Alabama  ...  App
alachia Southern  2622528  643  1551141
...    ...      ...                              ...                ...  ...
...       ...  ...      ...
1425  2013  1103254             Fidelity Mine  Refuse Recovery  ...
Illinois Basin    18532    4     8249
1426  2013  1102636                       Wfi  Refuse Recovery  ...
Illinois Basin     5070    4     1449
1427  2013  4407233                  Gobco #8  Refuse Recovery  ...   Ap
palachia Central   377607   16    43684
1428  2013  1518524            Turkey Pen Refuse  Refuse Recovery  ...   Ap
palachia Central     7744    2      622
1429  2013  1519685        Fedscreek Refuse Pile  Refuse Recovery  ...   Ap
palachia Central    17357    3     1020

[1430 rows x 16 columns]
```

In [82]:

```
#2
a=pd.read_excel('coalpublic2013.xlsx')
b=a[['Production (short tons)','Labor Hours']].sum()
c=pd.DataFrame(data=b).T
d=c.reindex(columns=a.columns)
print(d)
```

```
   Year  MSHA ID  Mine Name  Mine State  ...  Coal Supply Region  Production
(short tons)  Average Employees  Labor Hours
0    NaN      NaN        NaN         NaN  ...                 NaN
984841779                NaN    177910757

[1 rows x 16 columns]
```

In [116]:

```
#3
a=pd.read_excel('coalpublic2013.xlsx')
print(a.tail(10))
```

```
      Year  MSHA ID              Mine Name  ... Production (short tons) Aver
age Employees Labor Hours
1440  2013  3609405                Phoenix  ...                    4473
5        5670
1441  2013   100515      Mary Lee # 1 Mine  ...                    8400
4        6240
1442  2013  3609337     Marco Gfcc Project  ...                    6809
4        5175
1443  2013  1518401                  No. 1  ...                   94748
4        6337
1444  2013  1519713             # 1 Refuse  ...                    1879
2         200
1445  2013  1103254          Fidelity Mine  ...                   18532
4        8249
1446  2013  1102636                    Wfi  ...                    5070
4        1449
1447  2013  4407233               Gobco #8  ...                  377607
16       43684
1448  2013  1518524      Turkey Pen Refuse  ...                    7744
2         622
1449  2013  1519685  Fedscreek Refuse Pile  ...                   17357
3        1020

[10 rows x 16 columns]
```

In [86]:

```
#4
a=pd.read_excel('coalpublic2013.xlsx')
a[['MSHA ID','Labor Hours']].groupby('MSHA ID').sum()
```

Out[86]:

```
         Labor Hours
MSHA ID
  100329      144002
  100347      215295
  100515        6240
  100759      474784
  100851     1001809
     ...         ...
 4801353     2811138
 4801429      161270
 4801645       35687
 4801646      661265
 5000030      286079

1321 rows × 1 columns
```

In [95]:

```
#exercise 3
#1,2
a=pd.read_excel('coalpublic2013.xlsx')
a[a["Labor Hours"] > 20000]
```

Out[95]:

	Year	MSHA ID	Mine Name	Mine State	Mine County	Mine Status	Mine Type	Company Type
0	2013	103381	Tacoa Highwall Miner	Alabama	Bibb	Active, men working, not producing	Surface	Indepedent Producer Operator
1	2013	103404	Reid School Mine	Alabama	Blount	Permanently abandoned	Surface	Indepedent Producer Operator
2	2013	100759	North River #1 Underground Min	Alabama	Fayette	Active, men working, not producing	Underground	Indepedent Producer Operator
3	2013	103246	Bear Creek	Alabama	Franklin	Active	Surface	Indepedent Producer Operator
4	2013	103451	Knight Mine	Alabama	Franklin	Active	Surface	Indepedent Producer Operator
...	...	...	...	...	...	...	...	...
1418	2013	4800677	Jim Bridger Mine	Wyoming	Sweetwater	Active	Surface	Operating Subsidiary
1419	2013	4801180	Black Butte And Leucite Hills	Wyoming	Sweetwater	Active	Surface	Operating Subsidiary
1420	2013	4801646	Bridger Underground Coal Mine	Wyoming	Sweetwater	Active	Underground	Operating Subsidiary
1428	2013	3603561	Mcclure Strip	Refuse Recovery	Jefferson	Active	Refuse	Indepedent Producer Operator
1447	2013	4407233	Gobco #8	Refuse Recovery	Russell	Active	Refuse	Indepedent Producer Operator

893 rows × 16 columns

In [111]:

```
#3
a=pd.read_excel('coalpublic2013.xlsx')
a[a["Mine State"].map(lambda a: a.startswith('P'))]
```

911	2013	3608517	Adc, Inc Pit 008	Pennsylvania (Bituminous)	Westmoreland	Active	Surface	Indepedent Producer Operator	Mine only
912	2013	3609057	Shearer Mine	Pennsylvania (Bituminous)	Westmoreland	Active, men working, not producing	Surface	Indepedent Producer Operator	Mine only
913	2013	3609175	Kellar #1	Pennsylvania (Bituminous)	Westmoreland	Temporarily closed	Surface	Indepedent Producer Operator	Mine only
914	2013	3609275	Bertovich Surface Mine	Pennsylvania (Bituminous)	Westmoreland	Active	Surface	Indepedent Producer Operator	Mine only
915	2013	3610009	Kingston	Pennsylvania	Westmoreland	Active	Underground	Indepedent Producer	Mine only

In [112]:

```
#4
a=pd.read_excel('coalpublic2013.xlsx')
a[a['MSHA ID'].isin([3609833,3608517])]
```

Out[112]:

	Year	MSHA ID	Mine Name	Mine State	Mine County	Mine Status	Mine Type	Company Type	Ope
600	2013	3609833	Christner Project	Pennsylvania (Bituminous)	Allegheny	Temporarily closed	Surface	Indepedent Producer Operator	Min
911	2013	3608517	Adc, Inc Pit 008	Pennsylvania (Bituminous)	Westmoreland	Active	Surface	Indepedent Producer Operator	Min

In [114]:

```
#5
a=pd.read_excel('coalpublic2013.xlsx')
a[a['Mine Name'].isin(['Cherep #1','Bertovich Surface Mine'])]
```

Out[114]:

	Year	MSHA ID	Mine Name	Mine State	Mine County	Mine Status	Mine Type	Company Type	Opera
599	2013	3607443	Cherep #1	Pennsylvania (Bituminous)	Allegheny	Active, men working, not producing	Surface	Indepedent Producer Operator	Mine
914	2013	3609275	Bertovich Surface Mine	Pennsylvania (Bituminous)	Westmoreland	Active	Surface	Indepedent Producer Operator	Mine

In [115]:

```
#6
a=pd.read_excel('coalpublic2013.xlsx')
b=pd.read_excel('coalpublic2013.xlsx')
c=pd.read_excel('coalpublic2013.xlsx')
pd.concat([a,b,c])
```

Out[115]:

	Year	MSHA ID	Mine Name	Mine State	Mine County	Mine Status	Mine Type	Company Type	O
0	2013	103381	Tacoa Highwall Miner	Alabama	Bibb	Active, men working, not producing	Surface	Indepedent Producer Operator	M
1	2013	103404	Reid School Mine	Alabama	Blount	Permanently abandoned	Surface	Indepedent Producer Operator	M
2	2013	100759	North River #1 Underground Min	Alabama	Fayette	Active, men working, not producing	Underground	Indepedent Producer Operator	M Pre
3	2013	103246	Bear Creek	Alabama	Franklin	Active	Surface	Indepedent Producer Operator	M
4	2013	103451	Knight Mine	Alabama	Franklin	Active	Surface	Indepedent Producer Operator	M
...	...	...	...	...	...	...	...	...	
1445	2013	1103254	Fidelity Mine	Refuse Recovery	Perry	Active, men working, not producing	Refuse	Operating Subsidiary	M Pre
1446	2013	1102636	Wfi	Refuse Recovery	Saline	Active, men working, not producing	Refuse	Indepedent Producer Operator	M
1447	2013	4407233	Gobco #8	Refuse Recovery	Russell	Active	Refuse	Indepedent Producer Operator	M
1448	2013	1518524	Turkey Pen Refuse	Refuse Recovery	Pike	Active	Refuse	Indepedent Producer Operator	M
1449	2013	1519685	Fedscreek Refuse Pile	Refuse Recovery	Pike	Active, men working, not producing	Refuse	Indepedent Producer Operator	M

4350 rows × 16 columns

In [121]:

```
#exercise 4
#1,2
a=pd.read_excel('employee.xlsx')
a[a["hire_date"] > '01-01-07']
```

Out[121]:

	emp_id	first_name	last_name	hire_date
4	104	Bruce	Ernst	2007-05-21
7	107	Diana	Lorentz	2007-02-07
13	113	Luis	Popp	2007-12-07
19	119	Karen	Colmenares	2007-08-10

In [122]:

```
#3
a=pd.read_excel('employee.xlsx')
a.sort_values('hire_date')
```

Out[122]:

	emp_id	first_name	last_name	hire_date
2	102	Lex	De Haan	2001-01-13
9	109	Daniel	Faviet	2002-08-16
8	108	Nancy	Greenberg	2002-08-17
14	114	Den	Raphaely	2002-12-07
15	115	Alexander	Khoo	2003-05-18
0	100	Steven	King	2003-06-17
5	105	David	Austin	2005-06-25
17	117	Sigal	Tobias	2005-07-24
1	101	Neena	Kochhar	2005-09-21
10	110	John	Chen	2005-09-28
11	111	Ismael	Sciarra	2005-09-30
16	116	Shelli	Baida	2005-12-24
3	103	Alexander	Hunold	2006-01-03
6	106	Valli	Pataballa	2006-02-05
12	112	Jose Manuel	Urman	2006-03-07
18	118	Guy	Himuro	2006-11-15
7	107	Diana	Lorentz	2007-02-07
4	104	Bruce	Ernst	2007-05-21
19	119	Karen	Colmenares	2007-08-10
13	113	Luis	Popp	2007-12-07

In [123]:

```
#4
a=pd.read_excel('employee.xlsx')
a[(a['hire_date'] >='Mar-2002') & (a['hire_date'] <= 'Dec-2005')]
```

Out[123]:

	emp_id	first_name	last_name	hire_date
0	100	Steven	King	2003-06-17
1	101	Neena	Kochhar	2005-09-21
5	105	David	Austin	2005-06-25
8	108	Nancy	Greenberg	2002-08-17
9	109	Daniel	Faviet	2002-08-16
10	110	John	Chen	2005-09-28
11	111	Ismael	Sciarra	2005-09-30
14	114	Den	Raphaely	2002-12-07
15	115	Alexander	Khoo	2003-05-18
17	117	Sigal	Tobias	2005-07-24

In [125]:

```
#5
a=pd.read_excel('employee.xlsx')
b=a.set_index(['hire_date'])
b["2005"]
```

Out[125]:

	emp_id	first_name	last_name
hire_date			
2005-09-21	101	Neena	Kochhar
2005-06-25	105	David	Austin
2005-09-28	110	John	Chen
2005-09-30	111	Ismael	Sciarra
2005-12-24	116	Shelli	Baida
2005-07-24	117	Sigal	Tobias

In [126]:

```
#6
a=pd.read_excel('employee.xlsx')
a.set_index(['hire_date'])
```

Out[126]:

	emp_id	first_name	last_name
hire_date			
2003-06-17	100	Steven	King
2005-09-21	101	Neena	Kochhar
2001-01-13	102	Lex	De Haan
2006-01-03	103	Alexander	Hunold
2007-05-21	104	Bruce	Ernst
2005-06-25	105	David	Austin
2006-02-05	106	Valli	Pataballa
2007-02-07	107	Diana	Lorentz
2002-08-17	108	Nancy	Greenberg
2002-08-16	109	Daniel	Faviet
2005-09-28	110	John	Chen
2005-09-30	111	Ismael	Sciarra
2006-03-07	112	Jose Manuel	Urman
2007-12-07	113	Luis	Popp
2002-12-07	114	Den	Raphaely
2003-05-18	115	Alexander	Khoo
2005-12-24	116	Shelli	Baida
2005-07-24	117	Sigal	Tobias
2006-11-15	118	Guy	Himuro
2007-08-10	119	Karen	Colmenares

In [127]:

```python
#7
a=pd.read_excel('employee.xlsx')
a.sort_values(['first_name','last_name'])
```

Out[127]:

	emp_id	first_name	last_name	hire_date
3	103	Alexander	Hunold	2006-01-03
15	115	Alexander	Khoo	2003-05-18
4	104	Bruce	Ernst	2007-05-21
9	109	Daniel	Faviet	2002-08-16
5	105	David	Austin	2005-06-25
14	114	Den	Raphaely	2002-12-07
7	107	Diana	Lorentz	2007-02-07
18	118	Guy	Himuro	2006-11-15
11	111	Ismael	Sciarra	2005-09-30
10	110	John	Chen	2005-09-28
12	112	Jose Manuel	Urman	2006-03-07
19	119	Karen	Colmenares	2007-08-10
2	102	Lex	De Haan	2001-01-13
13	113	Luis	Popp	2007-12-07
8	108	Nancy	Greenberg	2002-08-17
1	101	Neena	Kochhar	2005-09-21
16	116	Shelli	Baida	2005-12-24
17	117	Sigal	Tobias	2005-07-24
0	100	Steven	King	2003-06-17
6	106	Valli	Pataballa	2006-02-05

In [129]:

```
#8
pd.read_excel('employee.xlsx',sheet_name=1)
```

Out[129]:

	emp_id	first_name	last_name	hire_date
0	120	Matthew	Weiss	2004-07-18
1	121	Adam	Fripp	2005-04-10
2	122	Payam	Kaufling	2003-05-01
3	123	Shanta	Vollman	2005-10-10
4	124	Kevin	Mourgos	2007-11-16
5	125	Julia	Nayer	2005-07-16
6	126	Irene	Mikkilineni	2006-09-28
7	127	James	Landry	2007-01-14
8	128	Steven	Markle	2008-03-08
9	129	Laura	Bissot	2005-08-20
10	130	Mozhe	Atkinson	2005-10-30
11	131	James	Marlow	2005-02-16
12	132	TJ	Olson	2007-04-10
13	133	Jason	Mallin	2004-06-14
14	134	Michael	Rogers	2006-08-26
15	135	Ki	Gee	2007-12-12
16	136	Hazel	Philtanker	2008-02-06
17	137	Renske	Ladwig	2003-07-14
18	138	Stephen	Stiles	2005-10-26

In [130]:

```
#9
a=pd.read_excel('employee.xlsx',sheet_name=0)
b=pd.read_excel('employee.xlsx',sheet_name=1)
c=pd.read_excel('employee.xlsx',sheet_name=2)
pd.concat([a,b,c])
```

Out[130]:

	emp_id	first_name	last_name	hire_date
0	100	Steven	King	2003-06-17
1	101	Neena	Kochhar	2005-09-21
2	102	Lex	De Haan	2001-01-13
3	103	Alexander	Hunold	2006-01-03
4	104	Bruce	Ernst	2007-05-21
5	105	David	Austin	2005-06-25
6	106	Valli	Pataballa	2006-02-05
7	107	Diana	Lorentz	2007-02-07
8	108	Nancy	Greenberg	2002-08-17
9	109	Daniel	Faviet	2002-08-16
10	110	John	Chen	2005-09-28
11	111	Ismael	Sciarra	2005-09-30
12	112	Jose Manuel	Urman	2006-03-07
13	113	Luis	Popp	2007-12-07
14	114	Den	Raphaely	2002-12-07
15	115	Alexander	Khoo	2003-05-18
16	116	Shelli	Baida	2005-12-24
17	117	Sigal	Tobias	2005-07-24
18	118	Guy	Himuro	2006-11-15
19	119	Karen	Colmenares	2007-08-10
0	120	Matthew	Weiss	2004-07-18
1	121	Adam	Fripp	2005-04-10
2	122	Payam	Kaufling	2003-05-01
3	123	Shanta	Vollman	2005-10-10
4	124	Kevin	Mourgos	2007-11-16
5	125	Julia	Nayer	2005-07-16
6	126	Irene	Mikkilineni	2006-09-28
7	127	James	Landry	2007-01-14
8	128	Steven	Markle	2008-03-08
9	129	Laura	Bissot	2005-08-20
10	130	Mozhe	Atkinson	2005-10-30
11	131	James	Marlow	2005-02-16

	emp_id	first_name	last_name	hire_date
12	132	TJ	Olson	2007-04-10
13	133	Jason	Mallin	2004-06-14
14	134	Michael	Rogers	2006-08-26
15	135	Ki	Gee	2007-12-12
16	136	Hazel	Philtanker	2008-02-06
17	137	Renske	Ladwig	2003-07-14
18	138	Stephen	Stiles	2005-10-26
0	141	Trenna	Rajs	2003-10-17
1	142	Curtis	Davies	2005-01-29
2	143	Randall	Matos	2006-03-15
3	144	Peter	Vargas	2006-07-09
4	145	John	Russell	2004-10-01
5	146	Karen	Partners	2005-01-05
6	147	Alberto	Errazuriz	2005-03-10
7	148	Gerald	Cambrault	2007-10-15
8	149	Eleni	Zlotkey	2008-01-29