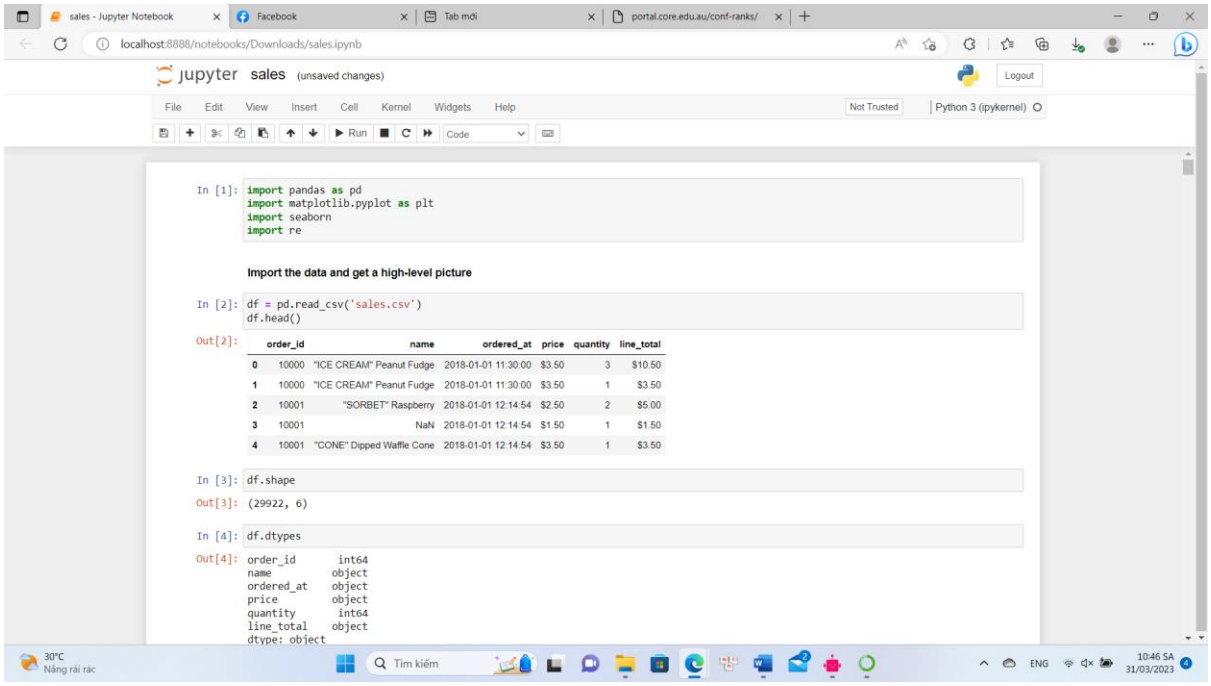
Nguyen Hoang Long_19521787

Github: https://github.com/glong25/Lab2_datamining.git



In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn
import re
```

Import the data and get a high-level picture

In [2]:
```
df = pd.read_csv('sales.csv')
df.head()
```

Out[2]:

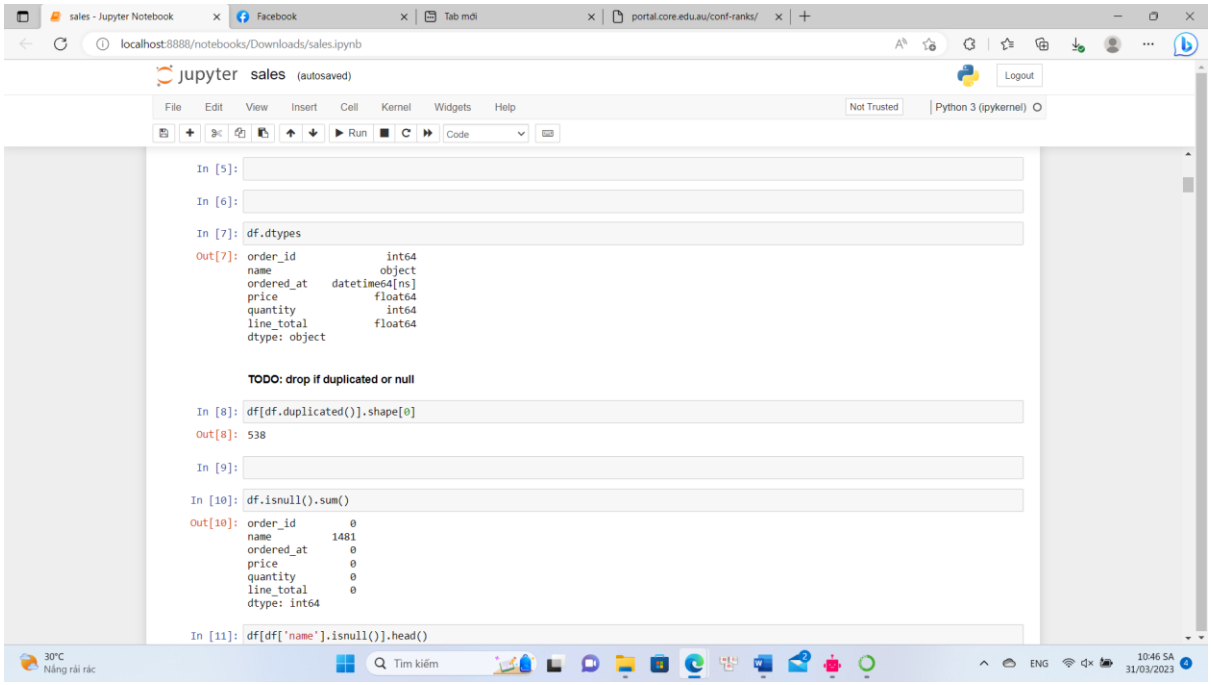
	order_id	name	ordered_at	price	quantity	line_total
0	10000	"ICE CREAM" Peanut Fudge	2018-01-01 11:30:00	$3.50	3	$10.50
1	10000	"ICE CREAM" Peanut Fudge	2018-01-01 11:30:00	$3.50	1	$3.50
2	10001	"SORBET" Raspberry	2018-01-01 12:14:54	$2.50	2	$5.00
3	10001	NaN	2018-01-01 12:14:54	$1.50	1	$1.50
4	10001	"CONE" Dipped Waffle Cone	2018-01-01 12:14:54	$3.50	1	$3.50

In [3]:
```
df.shape
```

Out[3]:
```
(29922, 6)
```

In [4]:
```
df.dtypes
```

Out[4]:
```
order_id       int64
name          object
ordered_at    object
price         object
quantity       int64
line_total    object
dtype: object
```



In [5]:

In [6]:

In [7]:
```
df.dtypes
```

Out[7]:
```
order_id              int64
name                 object
ordered_at    datetime64[ns]
price                float64
quantity               int64
line_total           float64
dtype: object
```

TODO: drop if duplicated or null

In [8]:
```
df[df.duplicated()].shape[0]
```

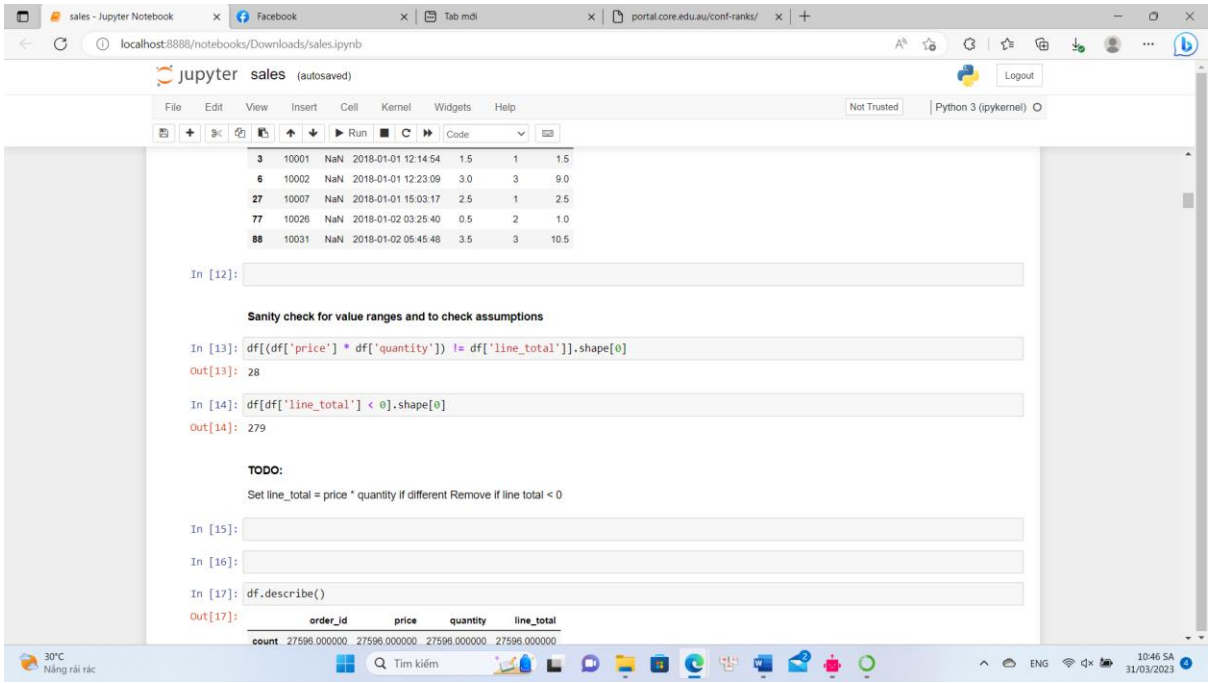
Out[8]:
```
538
```

In [9]:

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
order_id        0
name         1481
ordered_at      0
price           0
quantity        0
line_total      0
dtype: int64
```

In [11]:
```
df[df['name'].isnull()].head()
```



3	10001	NaN	2018-01-01 12:14:54	1.5	1	1.5
6	10002	NaN	2018-01-01 12:23:09	3.0	3	9.0
27	10007	NaN	2018-01-01 15:03:17	2.5	1	2.5
77	10026	NaN	2018-01-02 03:25:40	0.5	2	1.0
88	10031	NaN	2018-01-02 05:45:48	3.5	3	10.5

In [12]:

Sanity check for value ranges and to check assumptions

In [13]:
```
df[(df['price'] * df['quantity']) != df['line_total']].shape[0]
```
Out[13]: 28

In [14]:
```
df[df['line_total'] < 0].shape[0]
```
Out[14]: 279

TODO:

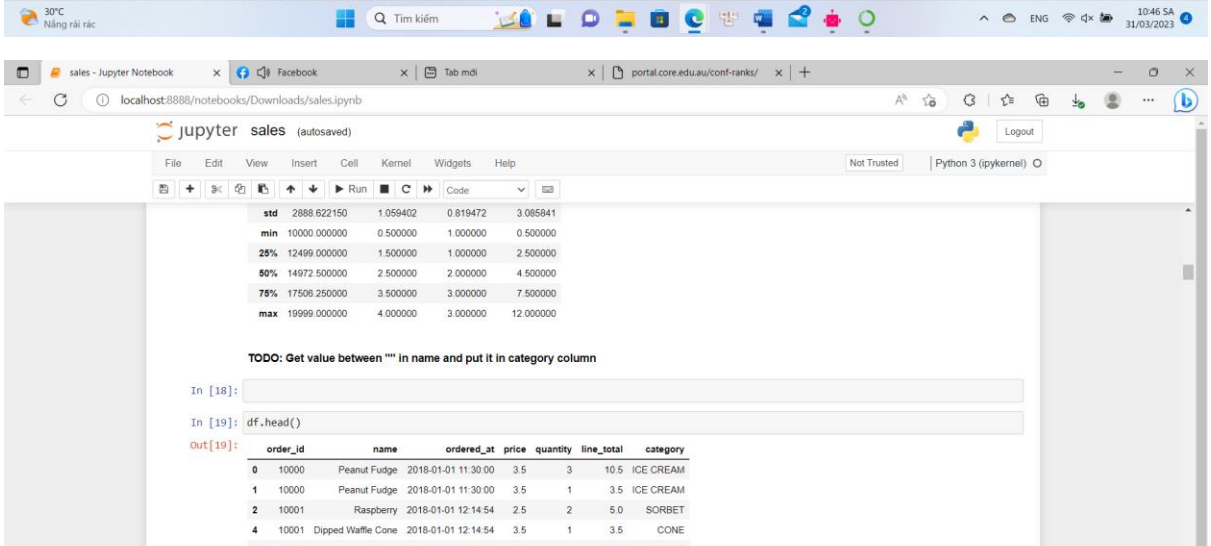
Set line_total = price * quantity if different Remove if line total < 0

In [15]:

In [16]:

In [17]:
```
df.describe()
```
Out[17]:

	order_id	price	quantity	line_total
count	27596.000000	27596.000000	27596.000000	27596.000000



std	2888.622150	1.059402	0.819472	3.085841
min	10000.000000	0.500000	1.000000	0.500000
25%	12499.000000	1.500000	1.000000	2.500000
50%	14972.500000	2.500000	2.000000	4.500000
75%	17506.250000	3.500000	3.000000	7.500000
max	19999.000000	4.000000	3.000000	12.000000

TODO: Get value between "" in name and put it in category column

In [18]:

In [19]:
```
df.head()
```
Out[19]:

	order_id	name	ordered_at	price	quantity	line_total	category
0	10000	Peanut Fudge	2018-01-01 11:30:00	3.5	3	10.5	ICE CREAM
1	10000	Peanut Fudge	2018-01-01 11:30:00	3.5	1	3.5	ICE CREAM
2	10001	Raspberry	2018-01-01 12:14:54	2.5	2	5.0	SORBET
4	10001	Dipped Waffle Cone	2018-01-01 12:14:54	3.5	1	3.5	CONE
5	10002	Lychee	2018-01-01 12:23:09	3.0	1	3.0	SORBET

Analysis, finally!

In [20]:
```
f, ax = plt.subplots(figsize=(10, 6))
df.groupby('name')['line_total'].sum().sort_values(ascending=False).head(10).plot(kind='bar')
f.autofmt_xdate()
plt.show()
```

sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved)

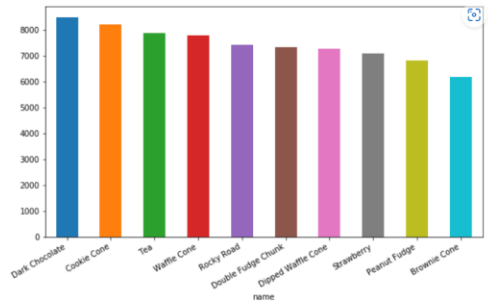
Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

```
plt.show()
```



In [2]:
```
%matplotlib inline
import numpy as np
import pandas as pd
df = pd.read_csv("PastHires.csv")
df.head()
```

Out[2]:

	Years Experience	Employed?	Previous employers	Level of Education	Top-tier school	Interned	Hired
0	10	Y	4	BS	N	N	Y
1	0	N	0	BS	Y	Y	Y

30°C Nắng rải rác | Tìm kiếm | ENG | 10:46 SA 31/03/2023

sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

	Years Experience	Employed?	Previous employers	Level of Education	Top-tier school	Interned	Hired
0	10	Y	4	BS	N	N	Y
1	0	N	0	BS	Y	Y	Y
2	7	N	6	BS	N	N	N
3	2	Y	1	MS	Y	N	Y
4	20	N	2	PhD	Y	N	N

In [3]:
```
df.head(10)
```

Out[3]:

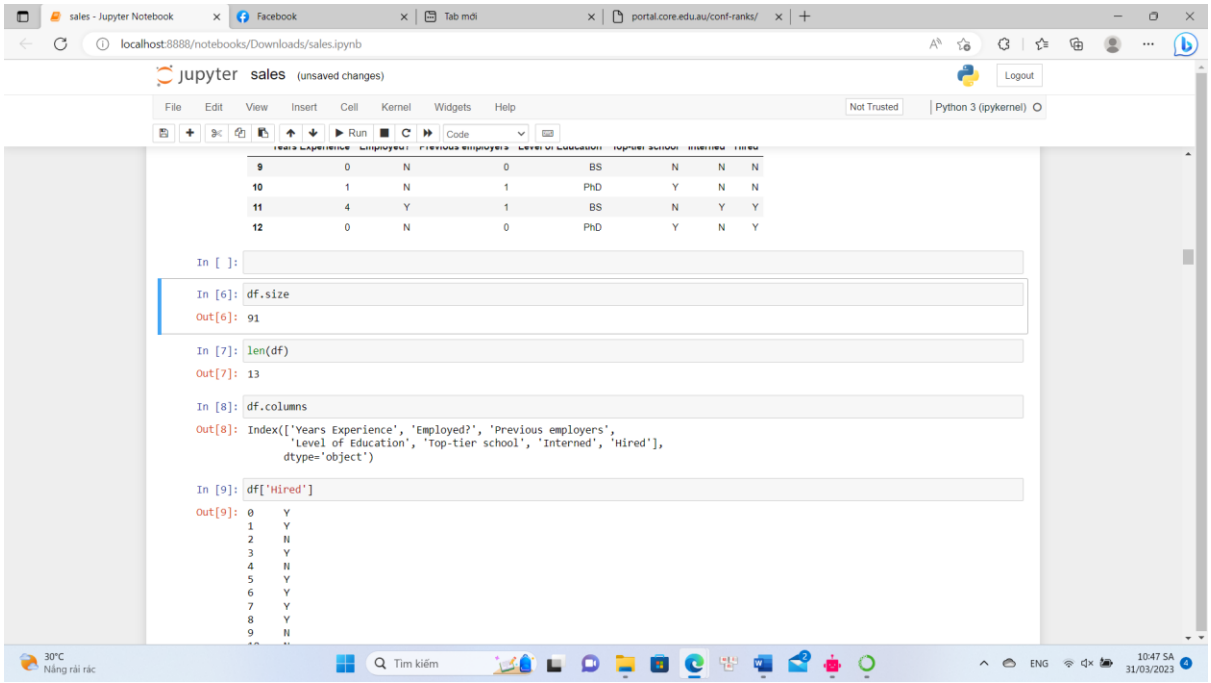
	Years Experience	Employed?	Previous employers	Level of Education	Top-tier school	Interned	Hired
0	10	Y	4	BS	N	N	Y
1	0	N	0	BS	Y	Y	Y
2	7	N	6	BS	N	N	N
3	2	Y	1	MS	Y	N	Y
4	20	N	2	PhD	Y	N	N
5	0	N	0	PhD	Y	Y	Y
6	5	Y	2	MS	N	Y	Y
7	3	N	1	BS	N	Y	Y
8	15	Y	5	BS	N	N	Y
9	0	N	0	BS	N	N	N

In [4]:
```
df.tail(4)
```

Out[4]:

	Years Experience	Employed?	Previous employers	Level of Education	Top-tier school	Interned	Hired
9	0	N	0	BS	N	N	N
10	1	N	1	PhD	Y	N	N

30°C Nắng rải rác | Tìm kiếm | ENG | 10:46 SA 31/03/2023



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

9	0	N	0	BS	N	N	N
10	1	N	1	PhD	Y	N	N
11	4	Y	1	BS	N	Y	Y
12	0	N	0	PhD	Y	N	Y

In [ ]:

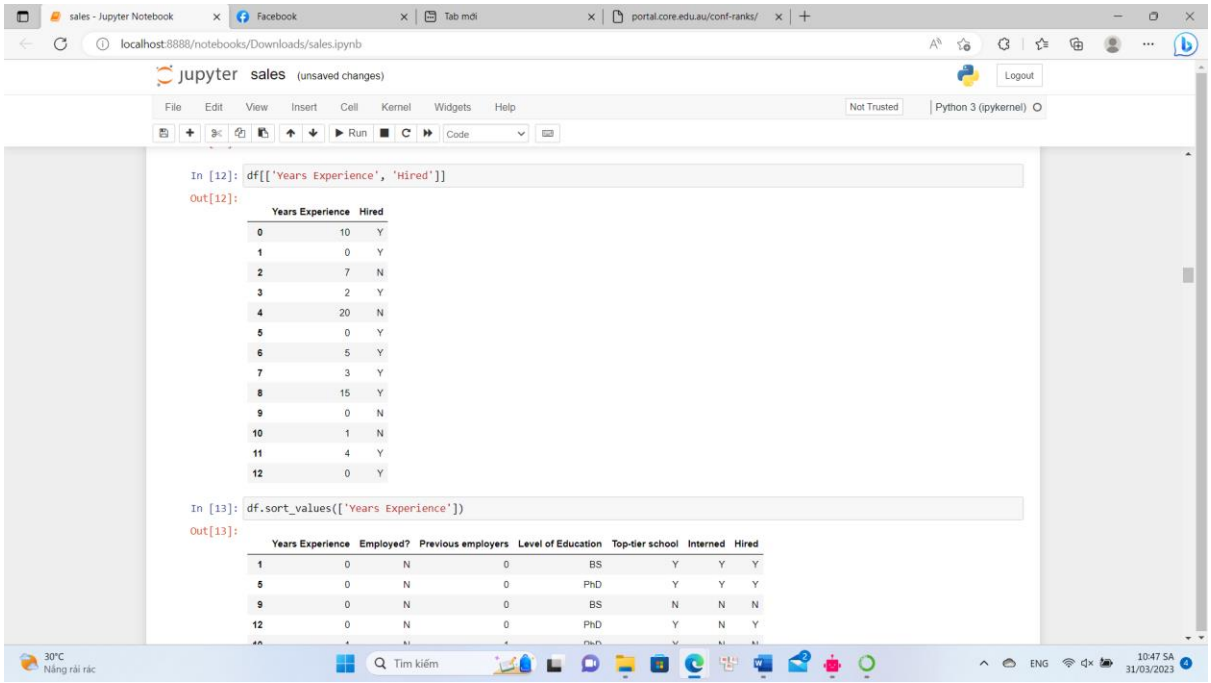
In [6]:
```
df.size
```
Out[6]: 91

In [7]:
```
len(df)
```
Out[7]: 13

In [8]:
```
df.columns
```
Out[8]:
```
Index(['Years Experience', 'Employed?', 'Previous employers',
       'Level of Education', 'Top-tier school', 'Interned', 'Hired'],
      dtype='object')
```

In [9]:
```
df['Hired']
```
Out[9]:
```
0     Y
1     Y
2     N
3     Y
4     N
5     Y
6     Y
7     Y
8     Y
9     N
```

30°C Nắng rải rác | Tìm kiếm | ENG | 10:47 SA 31/03/2023



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [12]:
```
df[['Years Experience', 'Hired']]
```
Out[12]:

	Years Experience	Hired
0	10	Y
1	0	Y
2	7	N
3	2	Y
4	20	N
5	0	Y
6	5	Y
7	3	Y
8	15	Y
9	0	N
10	1	N
11	4	Y
12	0	Y

In [13]:
```
df.sort_values(['Years Experience'])
```
Out[13]:

	Years Experience	Employed?	Previous employers	Level of Education	Top-tier school	Interned	Hired
1	0	N	0	BS	Y	Y	Y
5	0	N	0	PhD	Y	Y	Y
9	0	N	0	BS	N	N	N
12	0	N	0	PhD	Y	N	Y

30°C Nắng rải rác | Tìm kiếm | ENG | 10:47 SA 31/03/2023

sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter **sales** (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help | Not Trusted | Python 3 (ipykernel)

Run | Code

Out[13]:

	Years Experience	Employed?	Previous employers	Level of Education	Top-tier school	Interned	Hired
1	0	N	0	BS	Y	Y	Y
5	0	N	0	PhD	Y	Y	Y
9	0	N	0	BS	N	N	N
12	0	N	0	PhD	Y	N	Y
10	1	N	1	PhD	Y	N	N
3	2	Y	1	MS	Y	N	Y
7	3	N	1	BS	N	Y	Y
11	4	Y	1	BS	N	Y	Y
6	5	Y	2	MS	N	Y	Y
2	7	N	6	BS	N	N	N
0	10	Y	4	BS	N	N	Y
8	15	Y	5	BS	N	N	Y
4	20	N	2	PhD	Y	N	N

In [17]:
```
Degree_counts= df['Level of Education'].value_counts()
Degree_counts
```

Out[17]:
```
BS     7
PhD    4
MS     2
Name: Level of Education, dtype: int64
```

In [18]:
```
Degree_counts.plot(kind='bar')
```

Out[18]: <AxesSubplot:>

30°C Nắng rải rác | Tìm kiếm | ENG | 10:47 SA 31/03/2023

sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

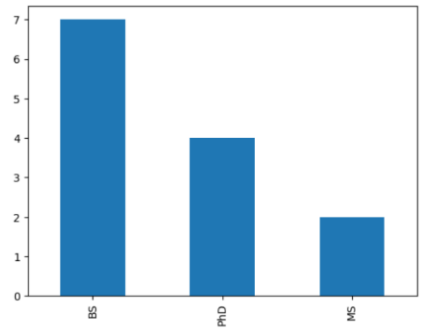
jupyter **sales** (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help | Not Trusted | Python 3 (ipykernel)

Run | Code

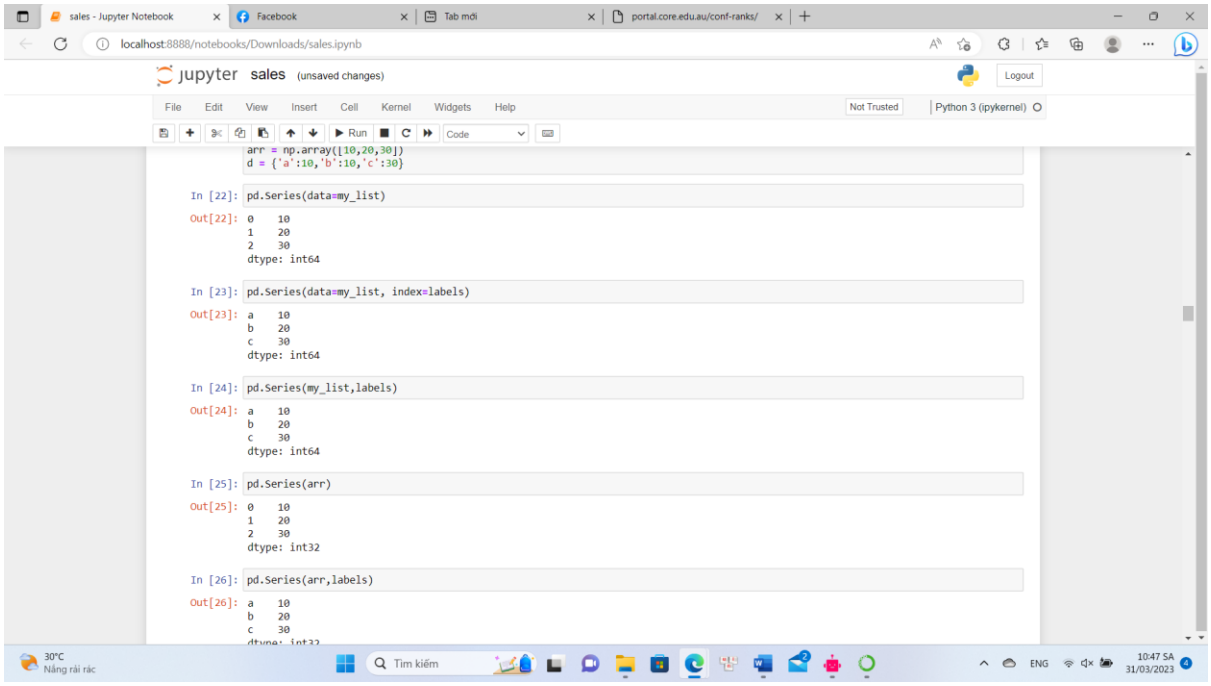
Out[18]: <AxesSubplot:>

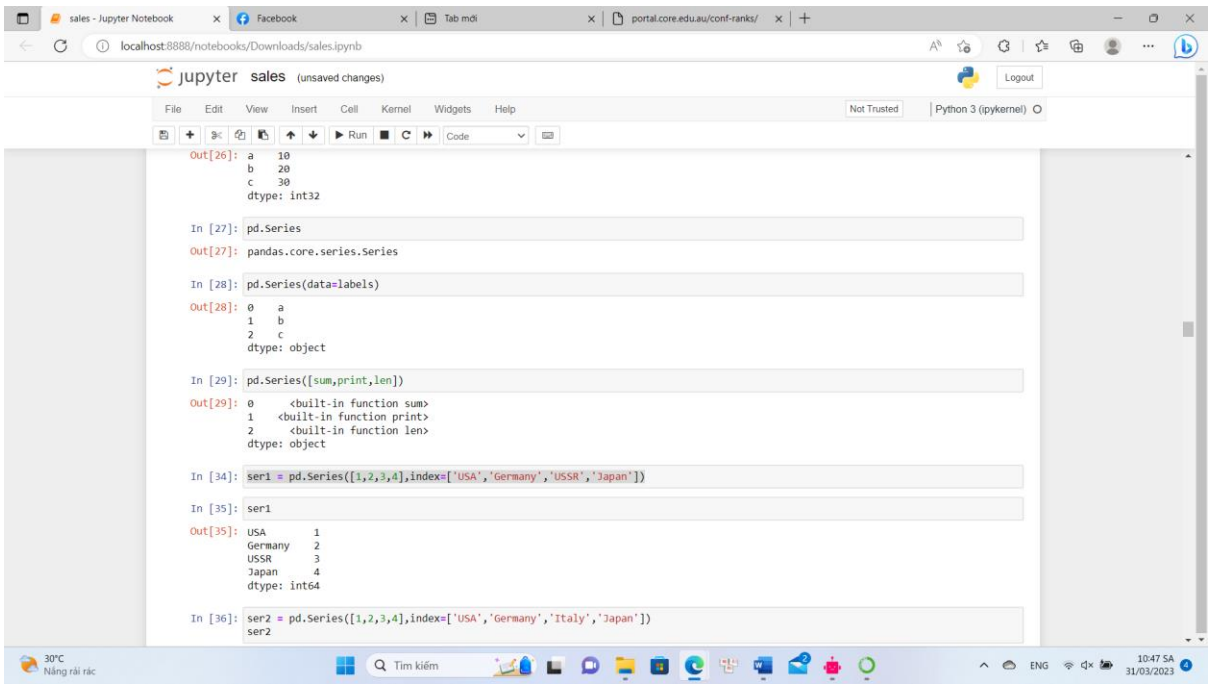


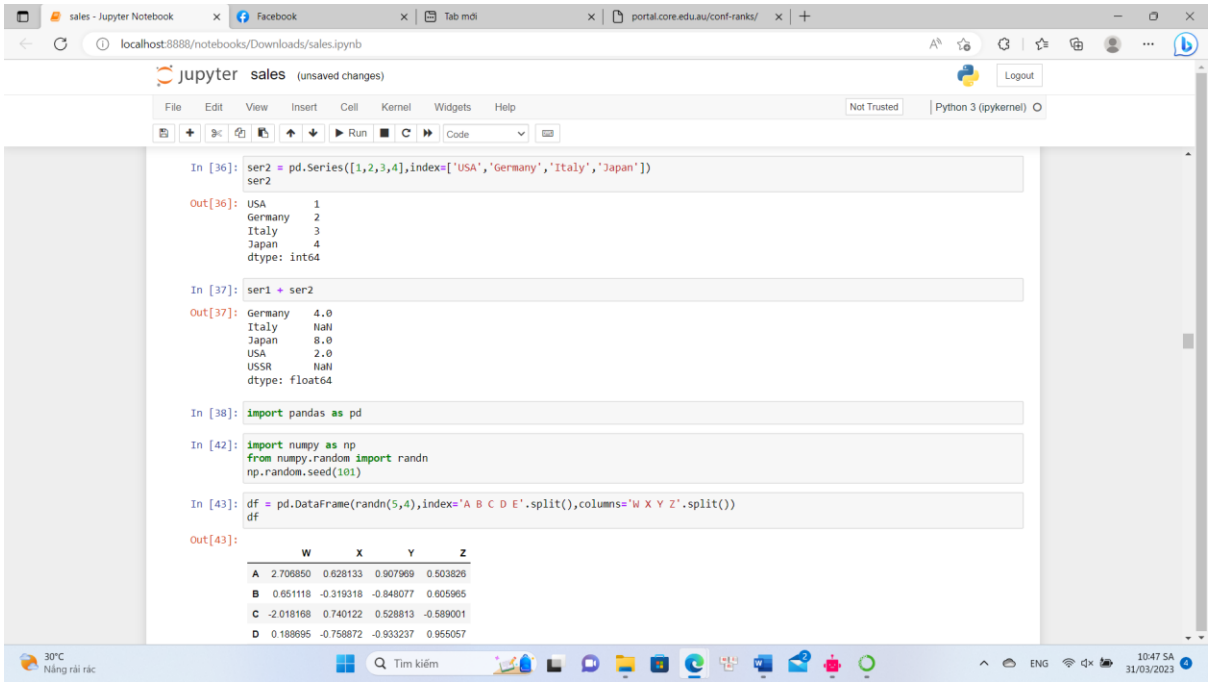
In [19]:
```
import numpy as np
import pandas as pd
```

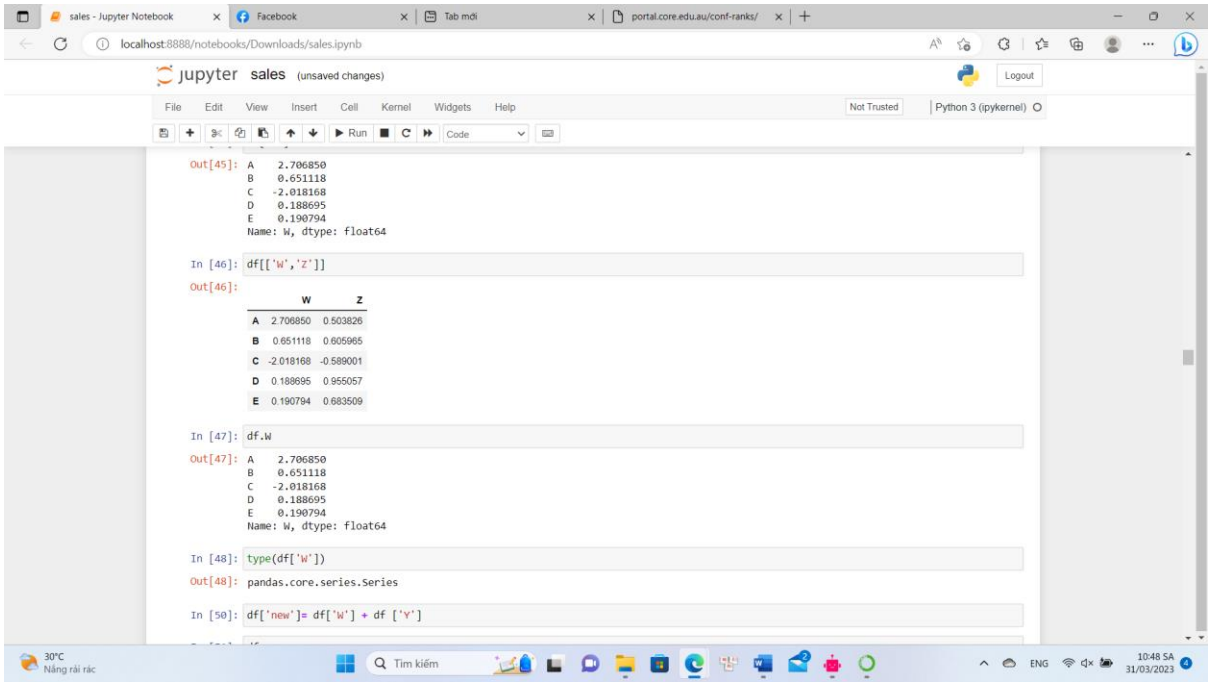
In [21]:
```
labels = ['a','b','c']
my_list = [10,20,30]
arr = np.array([10,20,30])
d = {'a':10,'b':10,'c':30}
```

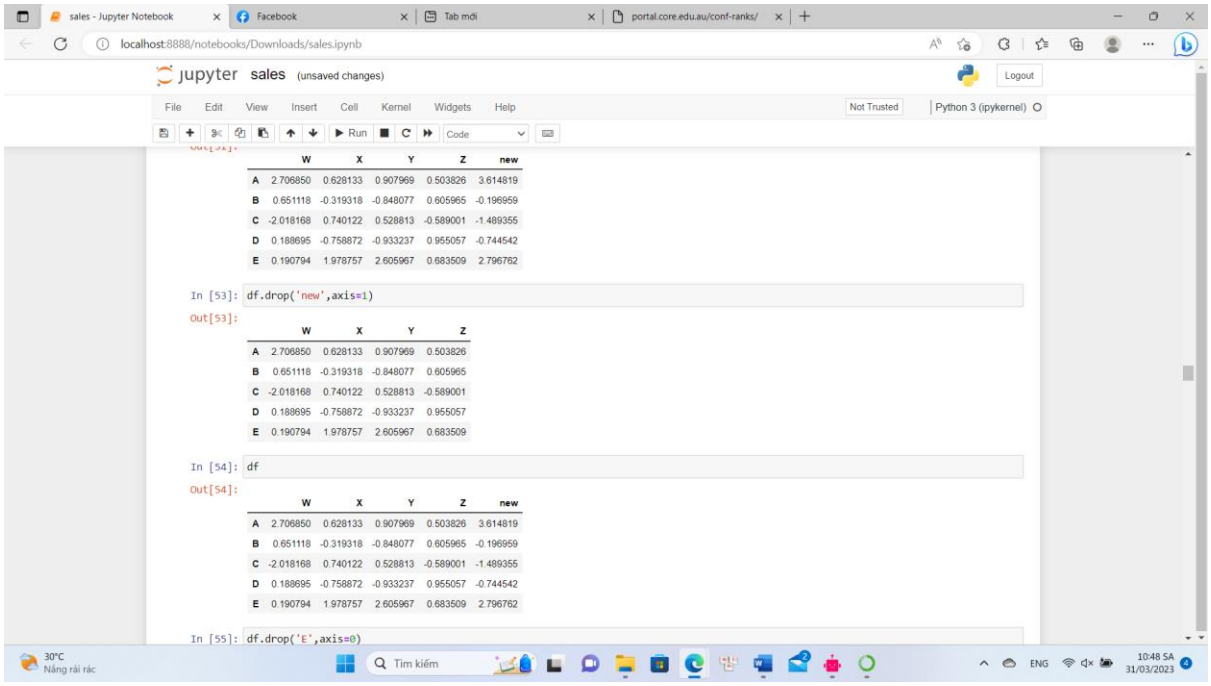
30°C Nắng rải rác | Tìm kiếm | ENG | 10:47 SA 31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
arr = np.array([10,20,30])
d = {'a':10,'b':10,'c':30}
In [22]: pd.Series(data=my_list)
Out[22]: 0 10
1 20
2 30
dtype: int64
In [23]: pd.Series(data=my_list, index=labels)
Out[23]: a 10
b 20
c 30
dtype: int64
In [24]: pd.Series(my_list,labels)
Out[24]: a 10
b 20
c 30
dtype: int64
In [25]: pd.Series(arr)
Out[25]: 0 10
1 20
2 30
dtype: int32
In [26]: pd.Series(arr,labels)
Out[26]: a 10
b 20
c 30
30°C
Nắng rải rác
Tìm kiếm
ENG
10:47 SA
31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Out[26]: a 10
b 20
c 30
dtype: int32
In [27]: pd.Series
Out[27]: pandas.core.series.Series
In [28]: pd.Series(data=labels)
Out[28]: 0 a
1 b
2 c
dtype: object
In [29]: pd.Series([sum,print,len])
Out[29]: 0 <built-in function sum>
1 <built-in function print>
2 <built-in function len>
dtype: object
In [34]: ser1 = pd.Series([1,2,3,4],index=['USA','Germany','USSR','Japan'])
In [35]: ser1
Out[35]: USA 1
Germany 2
USSR 3
Japan 4
dtype: int64
In [36]: ser2 = pd.Series([1,2,3,4],index=['USA','Germany','Italy','Japan'])
ser2
30°C
Nắng rải rác
Tìm kiếm
ENG
10:47 SA
31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [36]:
ser2 = pd.Series([1,2,3,4],index=['USA','Germany','Italy','Japan'])
ser2
Out[36]:
USA 1
Germany 2
Italy 3
Japan 4
dtype: int64
In [37]:
ser1 + ser2
Out[37]:
Germany 4.0
Italy NaN
Japan 8.0
USA 2.0
USSR NaN
dtype: float64
In [38]:
import pandas as pd
In [42]:
import numpy as np
from numpy.random import randn
np.random.seed(101)
In [43]:
df = pd.DataFrame(randn(5,4),index='A B C D E'.split(),columns='W X Y Z'.split())
df
Out[43]:
W X Y Z
A 2.706850 0.628133 0.907969 0.503826
B 0.651118 -0.319318 -0.848077 0.605965
C -2.018168 0.740122 0.528813 -0.589001
D 0.188695 -0.758872 -0.933237 0.955057
30°C
Nắng rải rác
Tìm kiếm
ENG
10:47 SA
31/03/2023

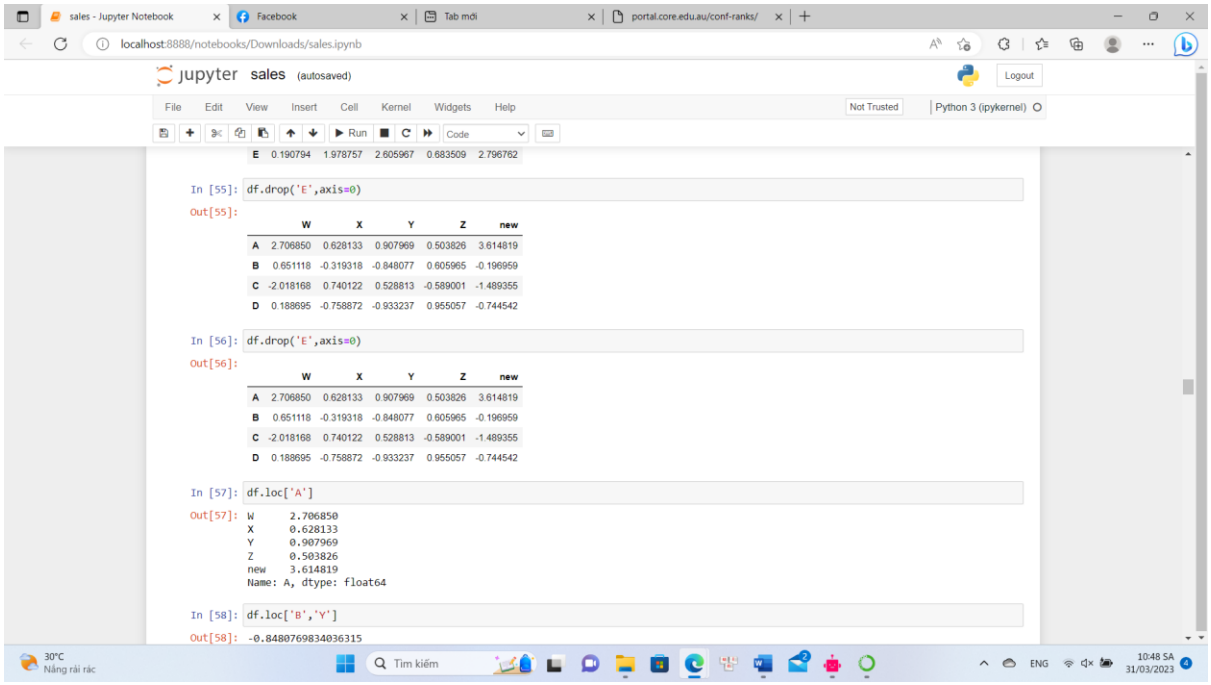

sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Out[45]:
A 2.706850
B 0.651118
C -2.018168
D 0.188695
E 0.190794
Name: W, dtype: float64
In [46]:
df[['W','Z']]
Out[46]:
W Z
A 2.706850 0.503826
B 0.651118 0.605965
C -2.018168 -0.589001
D 0.188695 0.955057
E 0.190794 0.683509
In [47]:
df.W
Out[47]:
A 2.706850
B 0.651118
C -2.018168
D 0.188695
E 0.190794
Name: W, dtype: float64
In [48]:
type(df['W'])
Out[48]:
pandas.core.series.Series
In [50]:
df['new']= df['W'] + df ['Y']
30°C
Nắng rải rác
Tìm kiếm
ENG
10:48 SA
31/03/2023



	W	X	Y	Z	new
A	2.706850	0.628133	0.907969	0.503826	3.614819
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959
C	-2.018168	0.740122	0.528813	-0.589001	-1.489355
D	0.188695	-0.758872	-0.933237	0.955057	-0.744542
E	0.190794	1.978757	2.605967	0.683509	2.796762

	W	X	Y	Z
A	2.706850	0.628133	0.907969	0.503826
B	0.651118	-0.319318	-0.848077	0.605965
C	-2.018168	0.740122	0.528813	-0.589001
D	0.188695	-0.758872	-0.933237	0.955057
E	0.190794	1.978757	2.605967	0.683509

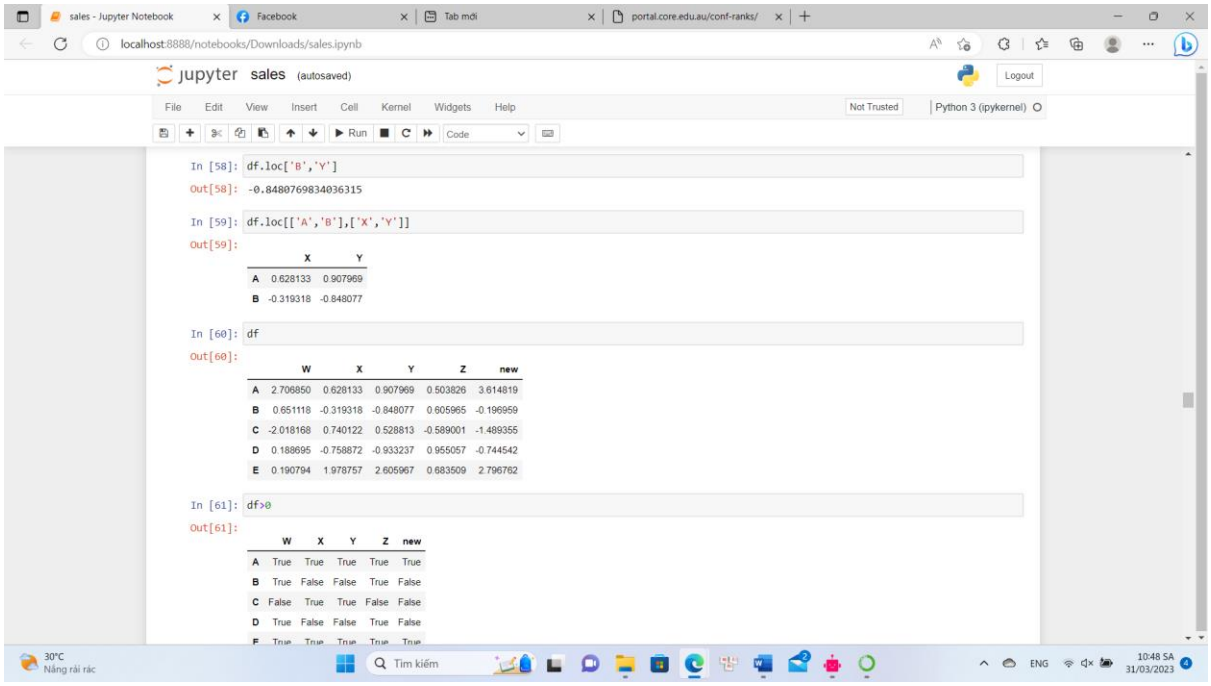
	W	X	Y	Z	new
A	2.706850	0.628133	0.907969	0.503826	3.614819
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959
C	-2.018168	0.740122	0.528813	-0.589001	-1.489355
D	0.188695	-0.758872	-0.933237	0.955057	-0.744542
E	0.190794	1.978757	2.605967	0.683509	2.796762



	W	X	Y	Z	new
E	0.190794	1.978757	2.605967	0.683509	2.796762

	W	X	Y	Z	new
A	2.706850	0.628133	0.907969	0.503826	3.614819
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959
C	-2.018168	0.740122	0.528813	-0.589001	-1.489355
D	0.188695	-0.758872	-0.933237	0.955057	-0.744542

	W	X	Y	Z	new
A	2.706850	0.628133	0.907969	0.503826	3.614819
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959
C	-2.018168	0.740122	0.528813	-0.589001	-1.489355
D	0.188695	-0.758872	-0.933237	0.955057	-0.744542



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [58]:
```
df.loc['B','Y']
```

Out[58]:
```
-0.8480769834036315
```

In [59]:
```
df.loc[['A','B'],['X','Y']]
```

Out[59]:

	X	Y
A	0.628133	0.907969
B	-0.319318	-0.848077

In [60]:
```
df
```

Out[60]:

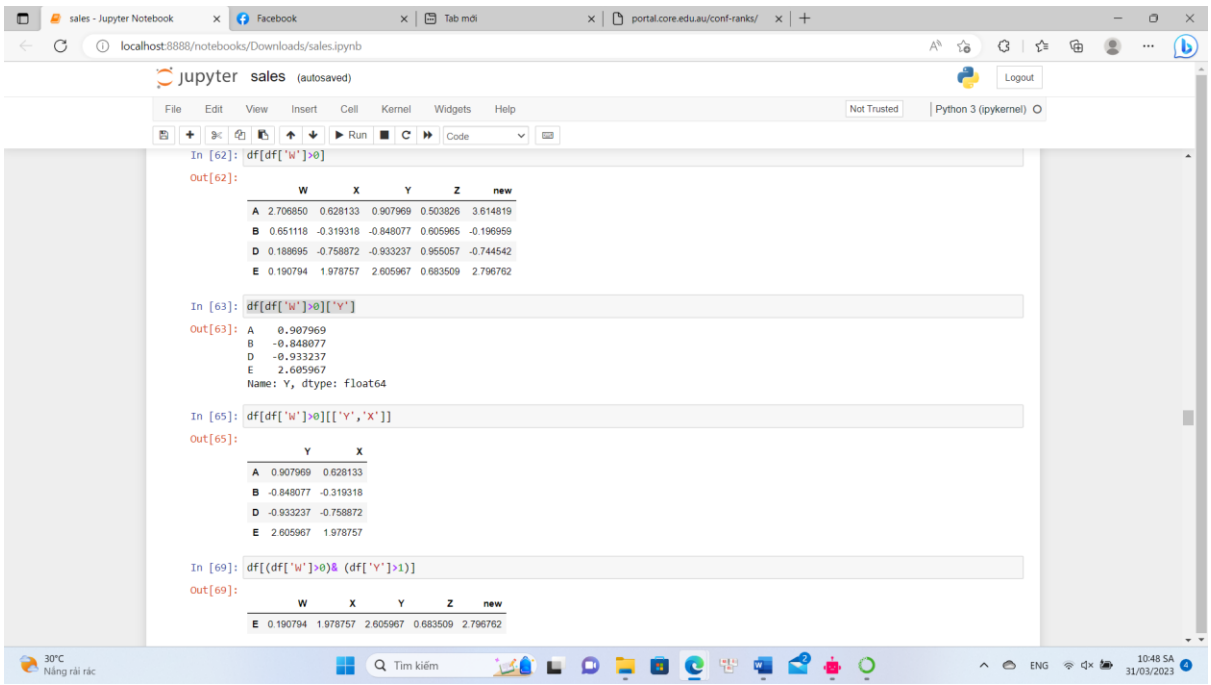
	W	X	Y	Z	new
A	2.706850	0.628133	0.907969	0.503826	3.614819
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959
C	-2.018168	0.740122	0.528813	-0.589001	-1.489355
D	0.188695	-0.758872	-0.933237	0.955057	-0.744542
E	0.190794	1.978757	2.605967	0.683509	2.796762

In [61]:
```
df>0
```

Out[61]:

	W	X	Y	Z	new
A	True	True	True	True	True
B	True	False	False	True	False
C	False	True	True	False	False
D	True	False	False	True	False
E	True	True	True	True	True

30°C Nắng rải rác | Tìm kiếm | ENG | 10:48 SA 31/03/2023



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [62]:
```
df[df['W']>0]
```

Out[62]:

	W	X	Y	Z	new
A	2.706850	0.628133	0.907969	0.503826	3.614819
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959
D	0.188695	-0.758872	-0.933237	0.955057	-0.744542
E	0.190794	1.978757	2.605967	0.683509	2.796762

In [63]:
```
df[df['W']>0]['Y']
```

Out[63]:
```
A    0.907969
B   -0.848077
D   -0.933237
E    2.605967
Name: Y, dtype: float64
```

In [65]:
```
df[df['W']>0][['Y','X']]
```

Out[65]:

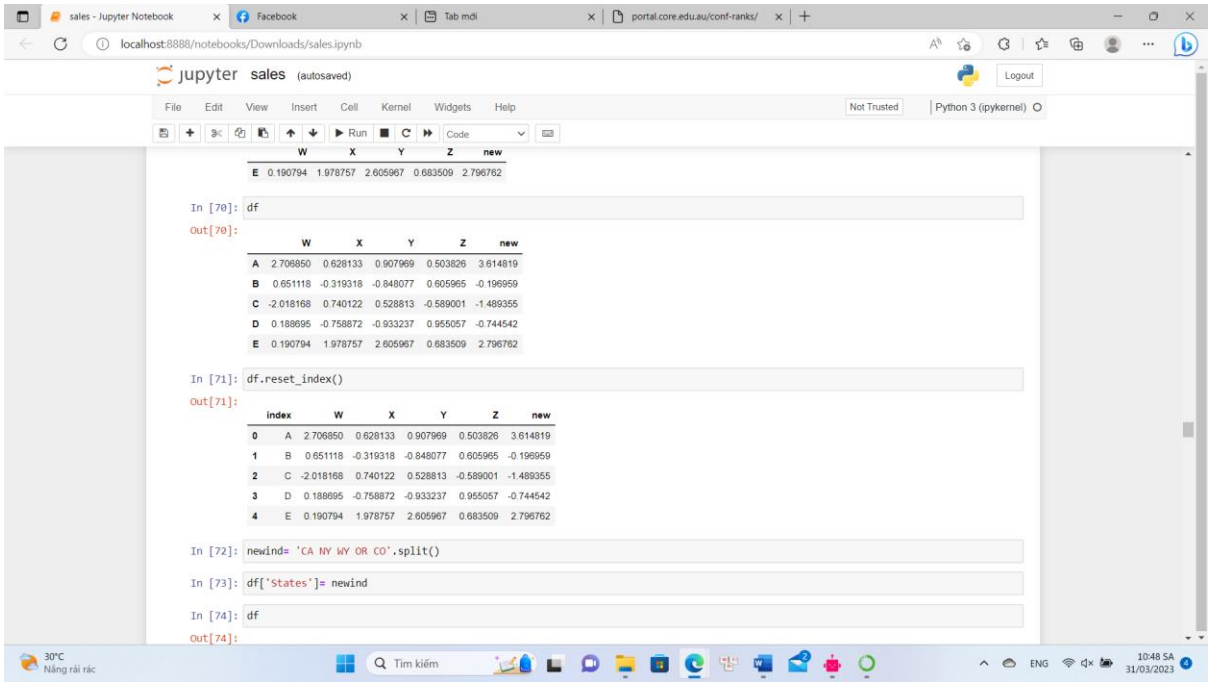
	Y	X
A	0.907969	0.628133
B	-0.848077	-0.319318
D	-0.933237	-0.758872
E	2.605967	1.978757

In [69]:
```
df[(df['W']>0)& (df['Y']>1)]
```

Out[69]:

	W	X	Y	Z	new
E	0.190794	1.978757	2.605967	0.683509	2.796762

30°C Nắng rải rác | Tìm kiếm | ENG | 10:48 SA 31/03/2023



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

	W	X	Y	Z	new
E	0.190794	1.978757	2.605967	0.683509	2.796762

In [70]:
```
df
```

Out[70]:

	W	X	Y	Z	new
A	2.706850	0.628133	0.907969	0.503826	3.614819
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959
C	-2.018168	0.740122	0.528813	-0.589001	-1.489355
D	0.188695	-0.758872	-0.933237	0.955057	-0.744542
E	0.190794	1.978757	2.605967	0.683509	2.796762

In [71]:
```
df.reset_index()
```

Out[71]:

	index	W	X	Y	Z	new
0	A	2.706850	0.628133	0.907969	0.503826	3.614819
1	B	0.651118	-0.319318	-0.848077	0.605965	-0.196959
2	C	-2.018168	0.740122	0.528813	-0.589001	-1.489355
3	D	0.188695	-0.758872	-0.933237	0.955057	-0.744542
4	E	0.190794	1.978757	2.605967	0.683509	2.796762

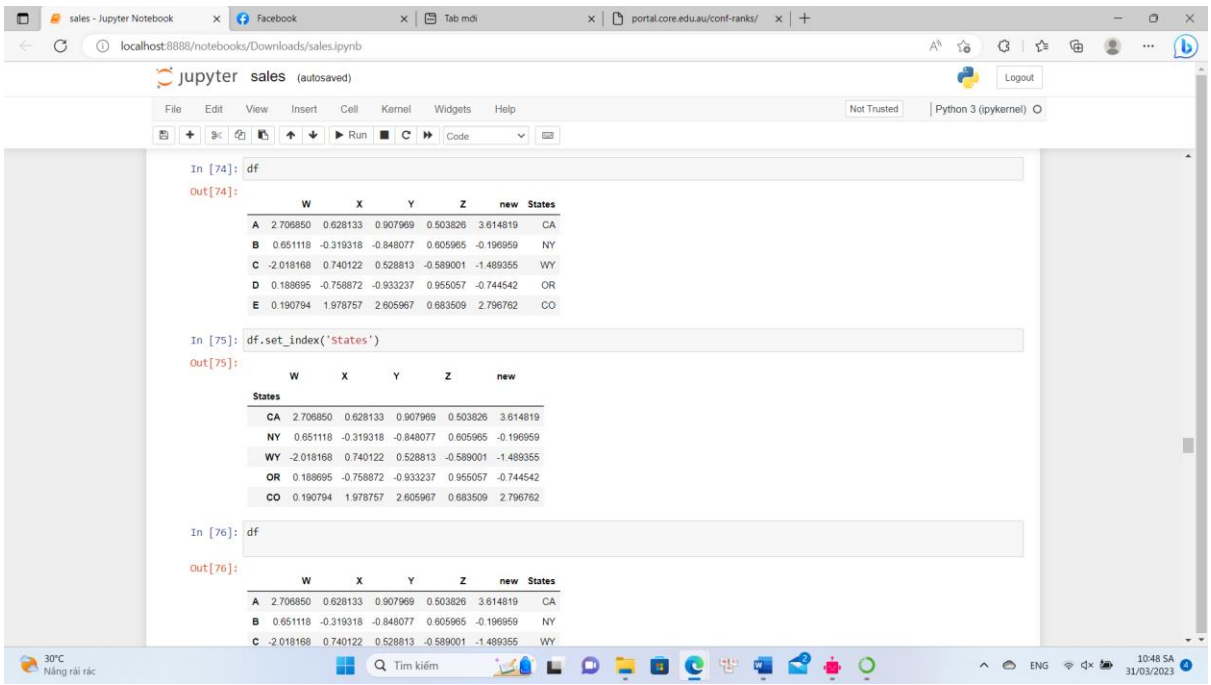
In [72]:
```
newind= 'CA NY WY OR CO'.split()
```

In [73]:
```
df['States']= newind
```

In [74]:
```
df
```

Out[74]:

30°C Nắng rải rác | Tìm kiếm | ENG | 10:48 SA 31/03/2023



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [74]:
```
df
```

Out[74]:

	W	X	Y	Z	new	States
A	2.706850	0.628133	0.907969	0.503826	3.614819	CA
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959	NY
C	-2.018168	0.740122	0.528813	-0.589001	-1.489355	WY
D	0.188695	-0.758872	-0.933237	0.955057	-0.744542	OR
E	0.190794	1.978757	2.605967	0.683509	2.796762	CO

In [75]:
```
df.set_index('States')
```

Out[75]:

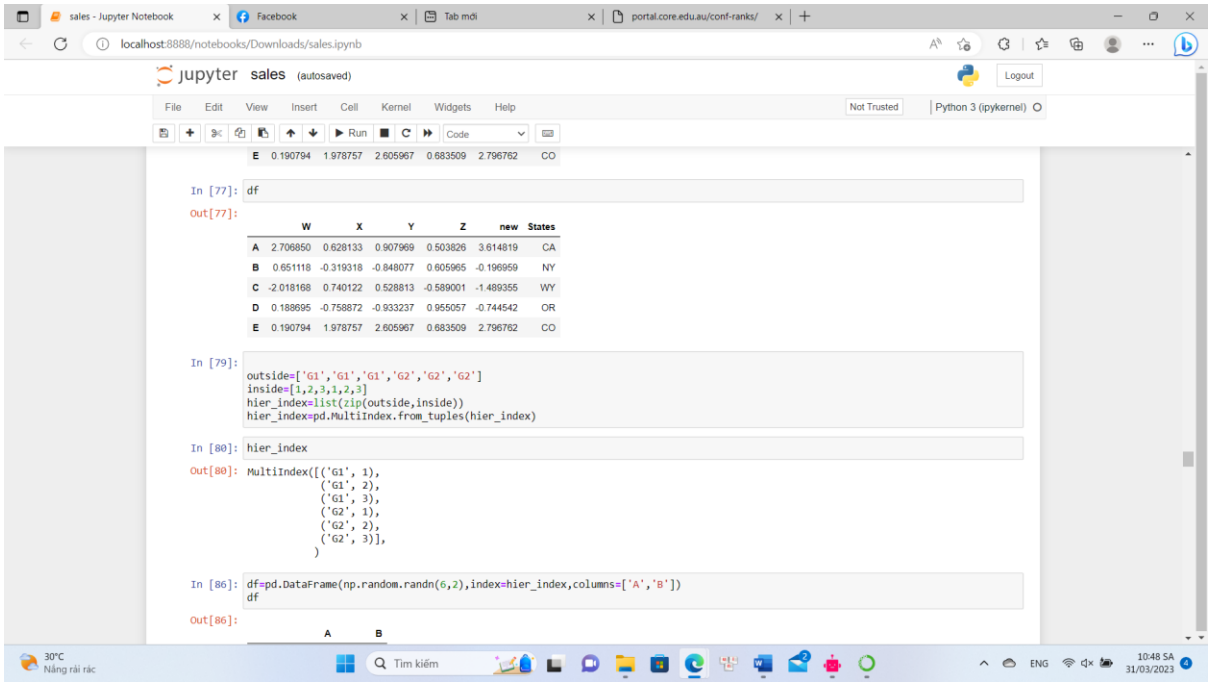
States	W	X	Y	Z	new
CA	2.706850	0.628133	0.907969	0.503826	3.614819
NY	0.651118	-0.319318	-0.848077	0.605965	-0.196959
WY	-2.018168	0.740122	0.528813	-0.589001	-1.489355
OR	0.188695	-0.758872	-0.933237	0.955057	-0.744542
CO	0.190794	1.978757	2.605967	0.683509	2.796762

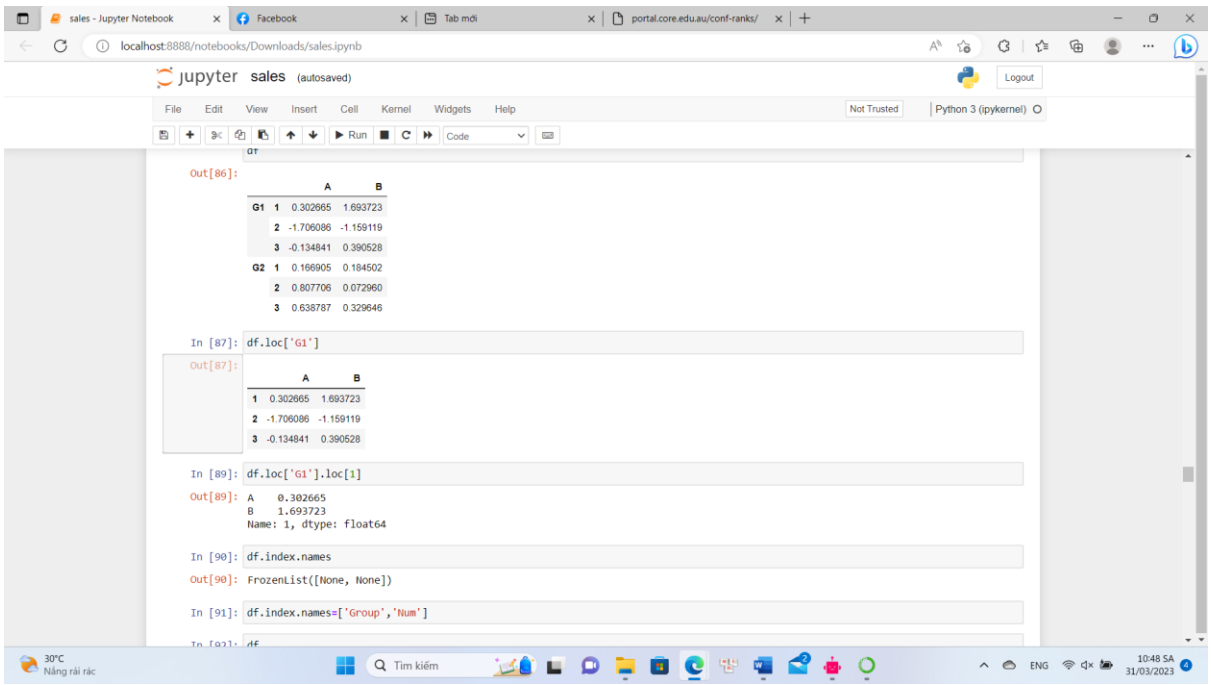
In [76]:
```
df
```

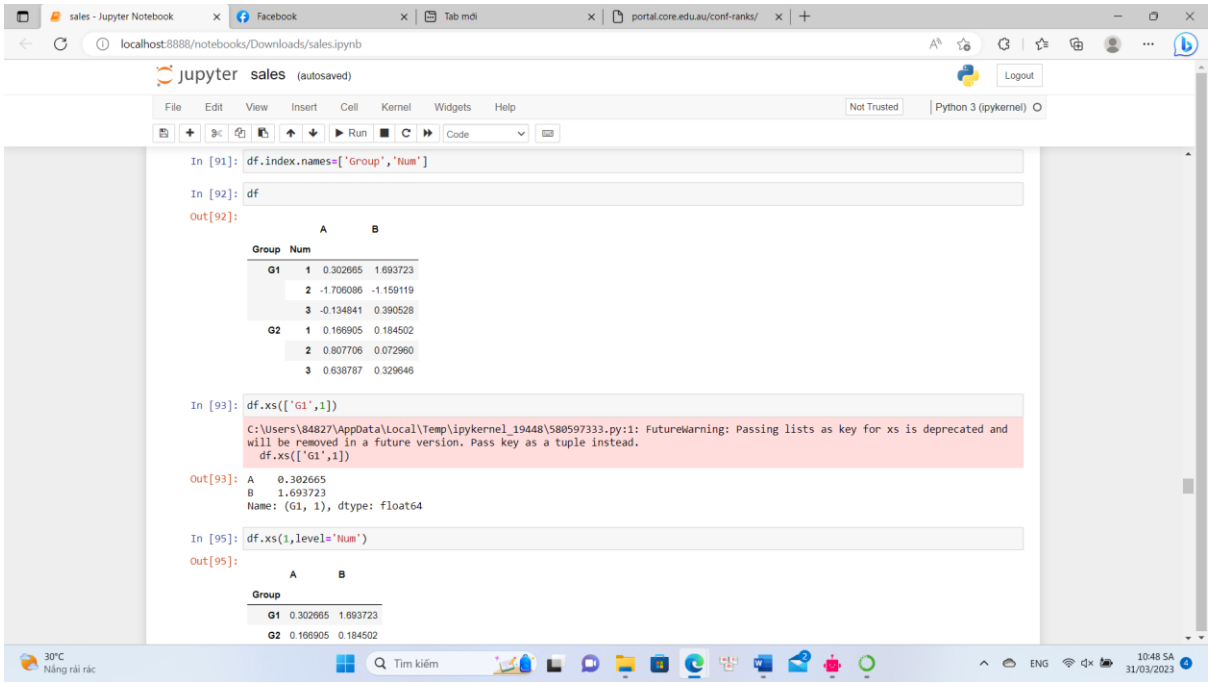
Out[76]:

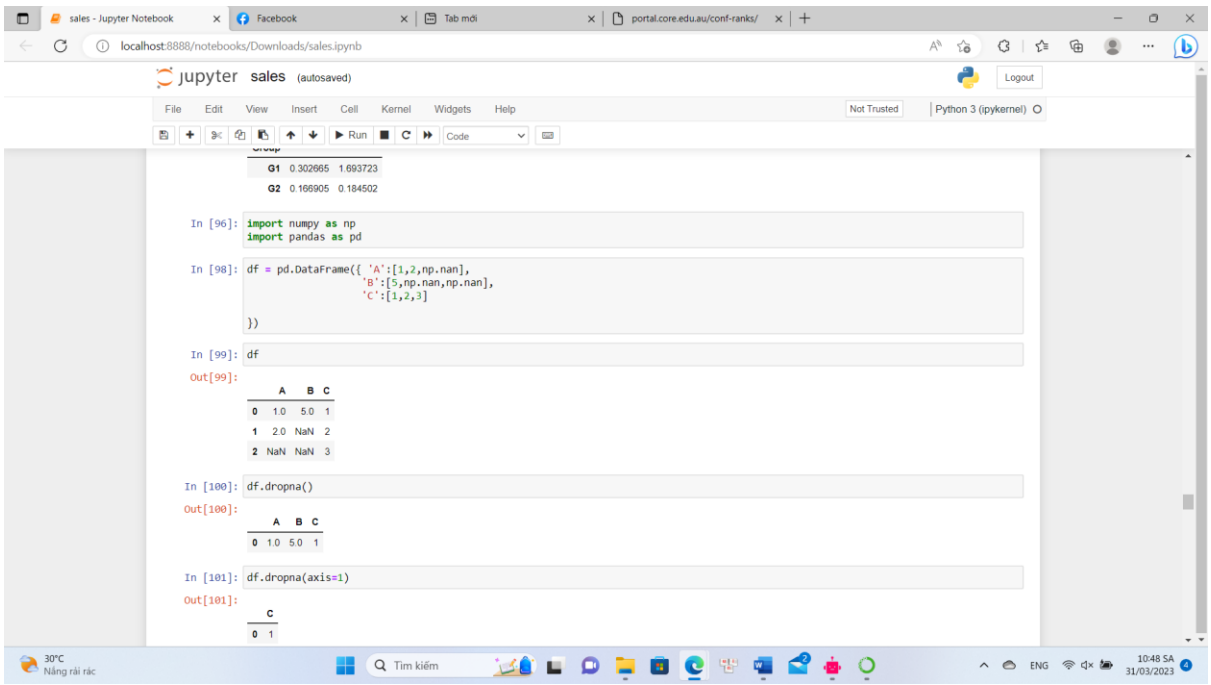
	W	X	Y	Z	new	States
A	2.706850	0.628133	0.907969	0.503826	3.614819	CA
B	0.651118	-0.319318	-0.848077	0.605965	-0.196959	NY
C	-2.018168	0.740122	0.528813	-0.589001	-1.489355	WY

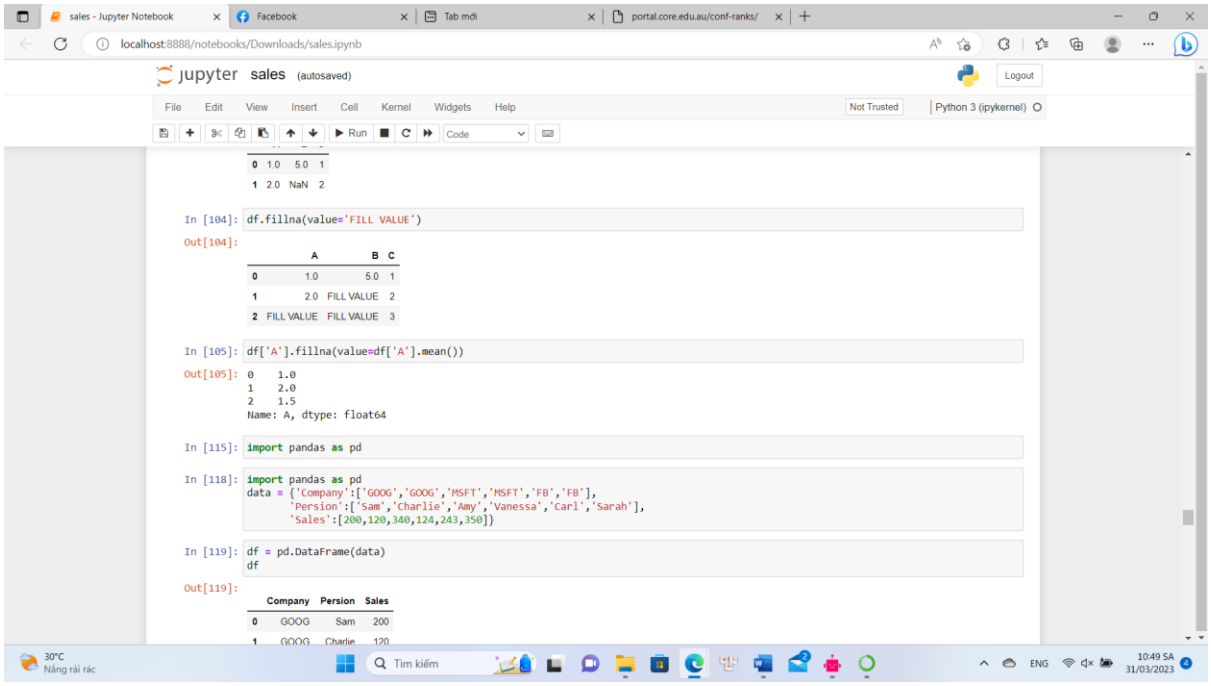
30°C Nắng rải rác | Tìm kiếm | ENG | 10:48 SA 31/03/2023

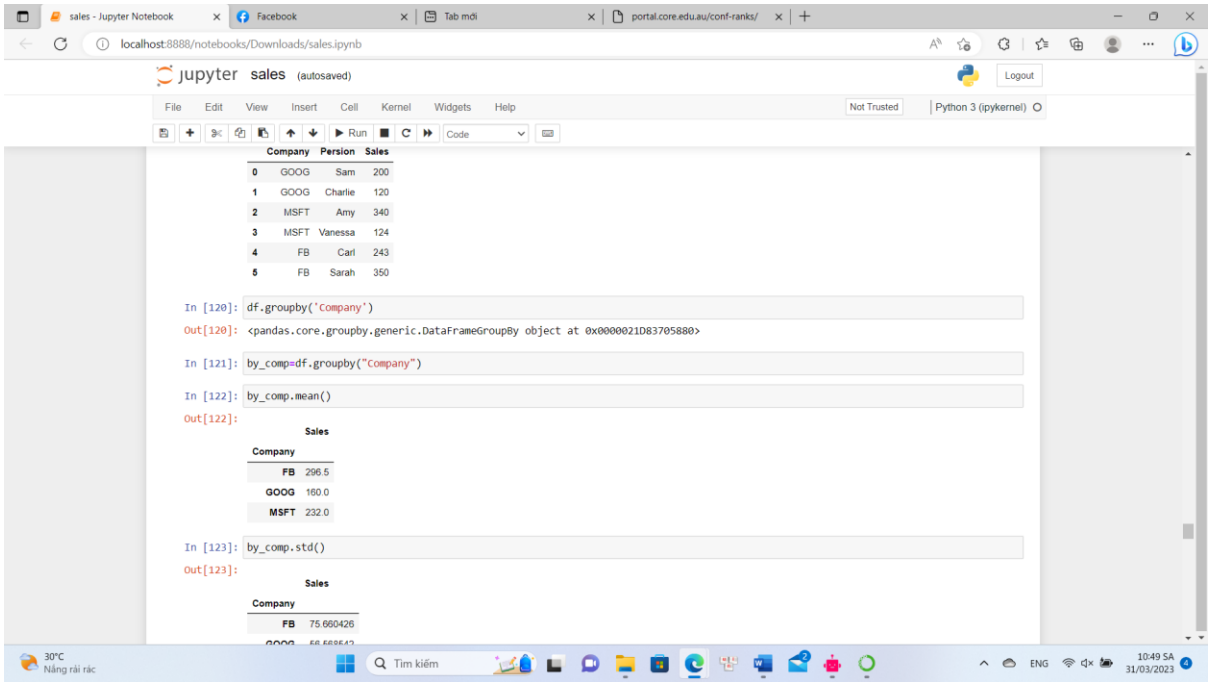

sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
E 0.190794 1.978757 2.605967 0.683509 2.796762 CO
In [77]: df
Out[77]:
W X Y Z new States
A 2.706850 0.628133 0.907969 0.503826 3.614819 CA
B 0.651118 -0.319318 -0.848077 0.605965 -0.196959 NY
C -2.018168 0.740122 0.528813 -0.589001 -1.489355 WY
D 0.188695 -0.758872 -0.933237 0.955057 -0.744542 OR
E 0.190794 1.978757 2.605967 0.683509 2.796762 CO
In [79]:
outside=['G1','G1','G1','G2','G2','G2']
inside=[1,2,3,1,2,3]
hier_index=list(zip(outside,inside))
hier_index=pd.MultiIndex.from_tuples(hier_index)
In [80]: hier_index
Out[80]:
MultiIndex([('G1', 1),
('G1', 2),
('G1', 3),
('G2', 1),
('G2', 2),
('G2', 3)],
)
In [86]:
df=pd.DataFrame(np.random.randn(6,2),index=hier_index,columns=['A','B'])
df
Out[86]:
A B
30°C
Nắng rải rác
Tìm kiếm
ENG
10:48 SA
31/03/2023

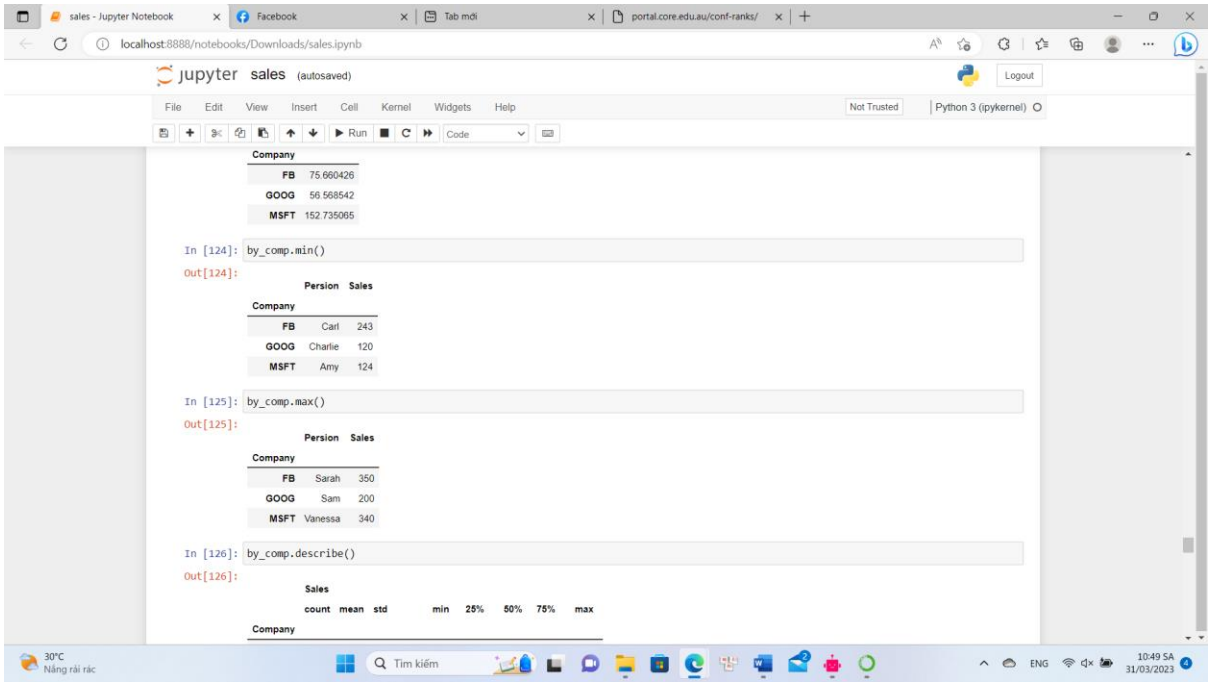

sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
df
Out[86]:
A B
G1 1 0.302665 1.693723
2 -1.706086 -1.159119
3 -0.134841 0.390528
G2 1 0.166905 0.184502
2 0.807706 0.072960
3 0.638787 0.329646
In [87]: df.loc['G1']
Out[87]:
A B
1 0.302665 1.693723
2 -1.706086 -1.159119
3 -0.134841 0.390528
In [89]: df.loc['G1'].loc[1]
Out[89]:
A 0.302665
B 1.693723
Name: 1, dtype: float64
In [90]: df.index.names
Out[90]: FrozenList([None, None])
In [91]: df.index.names=['Group','Num']
30°C
Nắng rải rác
Tìm kiếm
ENG
10:48 SA
31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [91]: df.index.names=['Group','Num']
In [92]: df
Out[92]:
A B
Group Num
G1 1 0.302665 1.693723
2 -1.706086 -1.159119
3 -0.134841 0.390528
G2 1 0.166905 0.184502
2 0.807706 0.072960
3 0.638787 0.329646
In [93]: df.xs(['G1',1])
C:\Users\84827\AppData\Local\Temp\ipykernel_19448\580597333.py:1: FutureWarning: Passing lists as key for xs is deprecated and will be removed in a future version. Pass key as a tuple instead.
df.xs(['G1',1])
Out[93]: A 0.302665
B 1.693723
Name: (G1, 1), dtype: float64
In [95]: df.xs(1,level='Num')
Out[95]:
A B
Group
G1 0.302665 1.693723
G2 0.166905 0.184502
30°C
Nắng rải rác
Tìm kiếm
ENG
10:48 SA
31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
G1 0.302665 1.693723
G2 0.166905 0.184502
In [96]: import numpy as np
import pandas as pd
In [98]: df = pd.DataFrame({ 'A':[1,2,np.nan],
'B':[5,np.nan,np.nan],
'C':[1,2,3]
})
In [99]: df
Out[99]:
A B C
0 1.0 5.0 1
1 2.0 NaN 2
2 NaN NaN 3
In [100]: df.dropna()
Out[100]:
A B C
0 1.0 5.0 1
In [101]: df.dropna(axis=1)
Out[101]:
C
0 1
30°C
Nắng rải rác
Tìm kiếm
ENG
10:48 SA
31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
0 1.0 5.0 1
1 2.0 NaN 2
In [104]: df.fillna(value='FILL VALUE')
Out[104]:
A B C
0 1.0 5.0 1
1 2.0 FILL VALUE 2
2 FILL VALUE FILL VALUE 3
In [105]: df['A'].fillna(value=df['A'].mean())
Out[105]: 0 1.0
1 2.0
2 1.5
Name: A, dtype: float64
In [115]: import pandas as pd
In [118]: import pandas as pd
data = {'Company':['GOOG','GOOG','MSFT','MSFT','FB','FB'],
'Persion':['Sam','Charlie','Amy','Vanessa','Carl','Sarah'],
'Sales':[200,120,340,124,243,350]}
In [119]: df = pd.DataFrame(data)
df
Out[119]:
Company Persion Sales
0 GOOG Sam 200
1 GOOG Charlie 120
30°C
Nắng rải rác
Tìm kiếm
ENG
10:49 SA
31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Company Persion Sales
0 GOOG Sam 200
1 GOOG Charlie 120
2 MSFT Amy 340
3 MSFT Vanessa 124
4 FB Carl 243
5 FB Sarah 350
In [120]: df.groupby('Company')
Out[120]: <pandas.core.groupby.generic.DataFrameGroupBy object at 0x0000021D83705880>
In [121]: by_comp=df.groupby("Company")
In [122]: by_comp.mean()
Out[122]:
Sales
Company
FB 296.5
GOOG 160.0
MSFT 232.0
In [123]: by_comp.std()
Out[123]:
Sales
Company
FB 75.660426
30°C
Nắng rải rác
Tìm kiếm
ENG
10:49 SA
31/03/2023



jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

Company	
FB	75.660426
GOOG	56.568542
MSFT	152.735065

In [124]:
```
by_comp.min()
```

Out[124]:

Company	Persion	Sales
FB	Carl	243
GOOG	Charlie	120
MSFT	Amy	124

In [125]:
```
by_comp.max()
```

Out[125]:

Company	Persion	Sales
FB	Sarah	350
GOOG	Sam	200
MSFT	Vanessa	340

In [126]:
```
by_comp.describe()
```

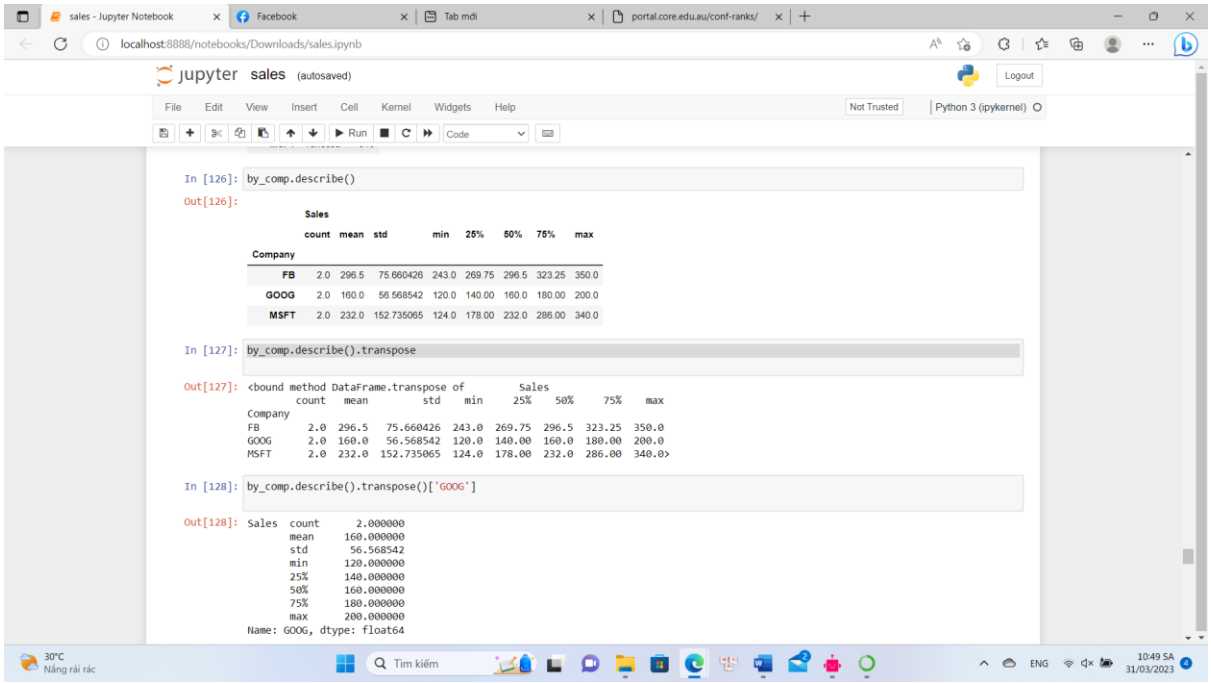
Out[126]:

	Sales							
	count	mean	std	min	25%	50%	75%	max
Company								

30°C Nắng rải rác

Tìm kiếm

ENG 10:49 SA 31/03/2023



jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [126]:
```
by_comp.describe()
```

Out[126]:

	Sales							
	count	mean	std	min	25%	50%	75%	max
Company								
FB	2.0	296.5	75.660426	243.0	269.75	296.5	323.25	350.0
GOOG	2.0	160.0	56.568542	120.0	140.00	160.0	180.00	200.0
MSFT	2.0	232.0	152.735065	124.0	178.00	232.0	286.00	340.0

In [127]:
```
by_comp.describe().transpose
```

Out[127]:
```
<bound method DataFrame.transpose of         Sales
          count   mean         std    min     25%    50%     75%    max
Company
FB          2.0  296.5   75.660426  243.0  269.75  296.5  323.25  350.0
GOOG        2.0  160.0   56.568542  120.0  140.00  160.0  180.00  200.0
MSFT        2.0  232.0  152.735065  124.0  178.00  232.0  286.00  340.0>
```

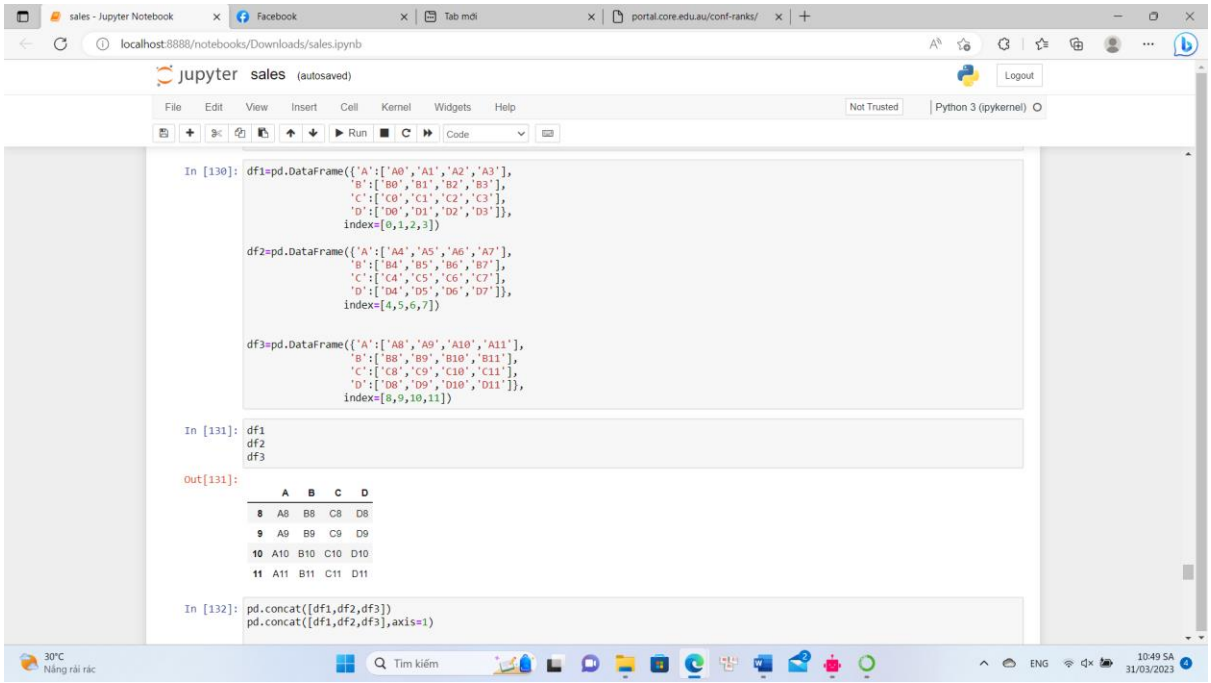
In [128]:
```
by_comp.describe().transpose()['GOOG']
```

Out[128]:
```
Sales  count      2.000000
       mean     160.000000
       std       56.568542
       min      120.000000
       25%      140.000000
       50%      160.000000
       75%      180.000000
       max      200.000000
Name: GOOG, dtype: float64
```

30°C Nắng rải rác

Tìm kiếm

ENG 10:49 SA 31/03/2023



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [130]:
```
df1=pd.DataFrame({'A':['A0','A1','A2','A3'],
                  'B':['B0','B1','B2','B3'],
                  'C':['C0','C1','C2','C3'],
                  'D':['D0','D1','D2','D3']},
                 index=[0,1,2,3])

df2=pd.DataFrame({'A':['A4','A5','A6','A7'],
                  'B':['B4','B5','B6','B7'],
                  'C':['C4','C5','C6','C7'],
                  'D':['D4','D5','D6','D7']},
                 index=[4,5,6,7])


df3=pd.DataFrame({'A':['A8','A9','A10','A11'],
                  'B':['B8','B9','B10','B11'],
                  'C':['C8','C9','C10','C11'],
                  'D':['D8','D9','D10','D11']},
                 index=[8,9,10,11])
```

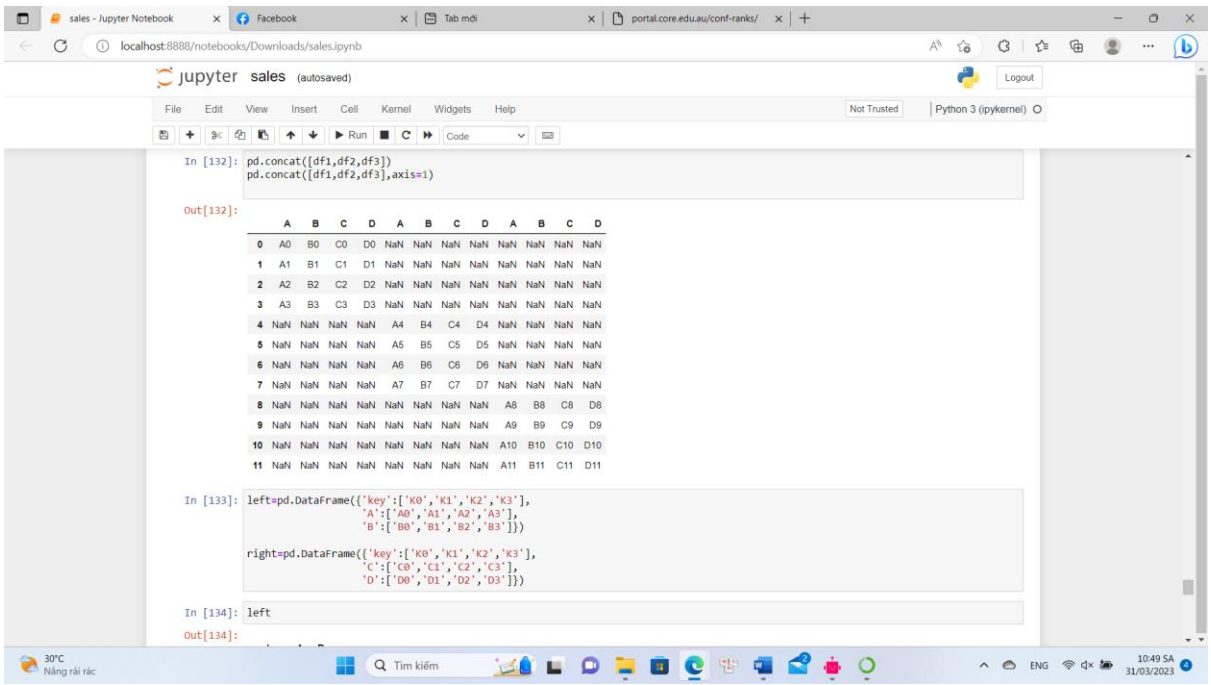
In [131]:
```
df1
df2
df3
```

Out[131]:

	A	B	C	D
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

In [132]:
```
pd.concat([df1,df2,df3])
pd.concat([df1,df2,df3],axis=1)
```

30°C Nắng rải rác | Tìm kiếm | ENG | 10:49 SA 31/03/2023



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [132]:
```
pd.concat([df1,df2,df3])
pd.concat([df1,df2,df3],axis=1)
```

Out[132]:

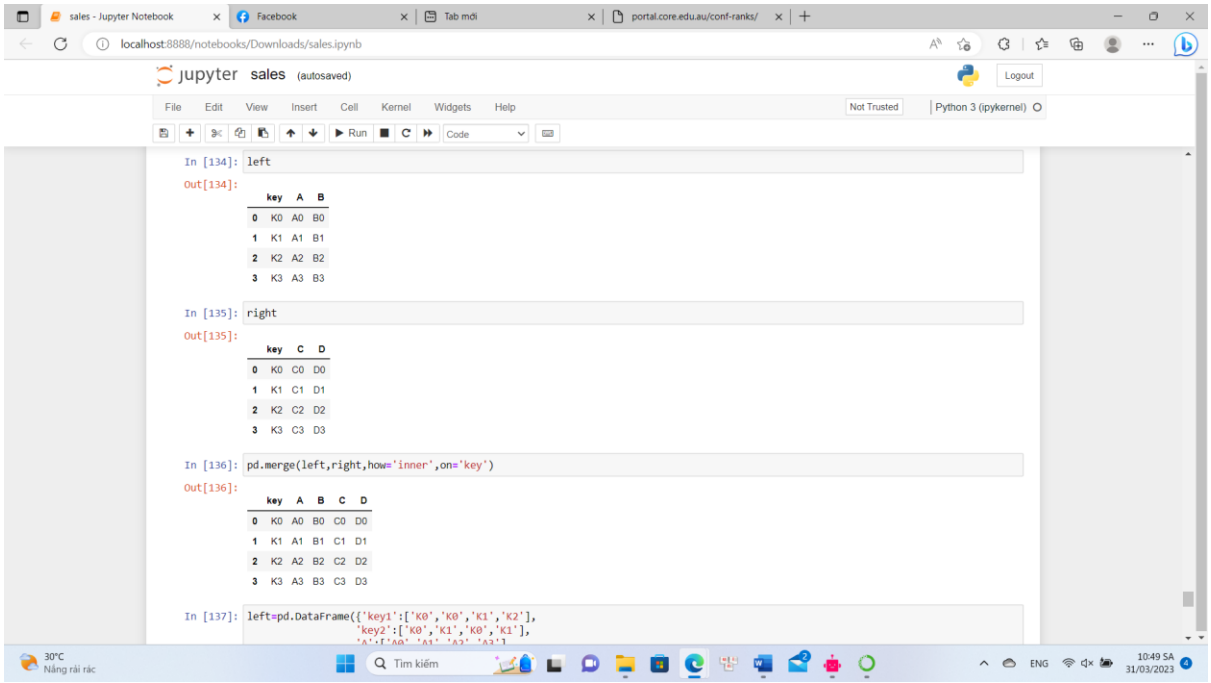
	A	B	C	D	A	B	C	D	A	B	C	D
0	A0	B0	C0	D0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	A1	B1	C1	D1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	A2	B2	C2	D2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	A3	B3	C3	D3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	A4	B4	C4	D4	NaN	NaN	NaN	NaN
5	NaN	NaN	NaN	NaN	A5	B5	C5	D5	NaN	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	A6	B6	C6	D6	NaN	NaN	NaN	NaN
7	NaN	NaN	NaN	NaN	A7	B7	C7	D7	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	A8	B8	C8	D8
9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	A9	B9	C9	D9
10	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	A10	B10	C10	D10
11	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	A11	B11	C11	D11

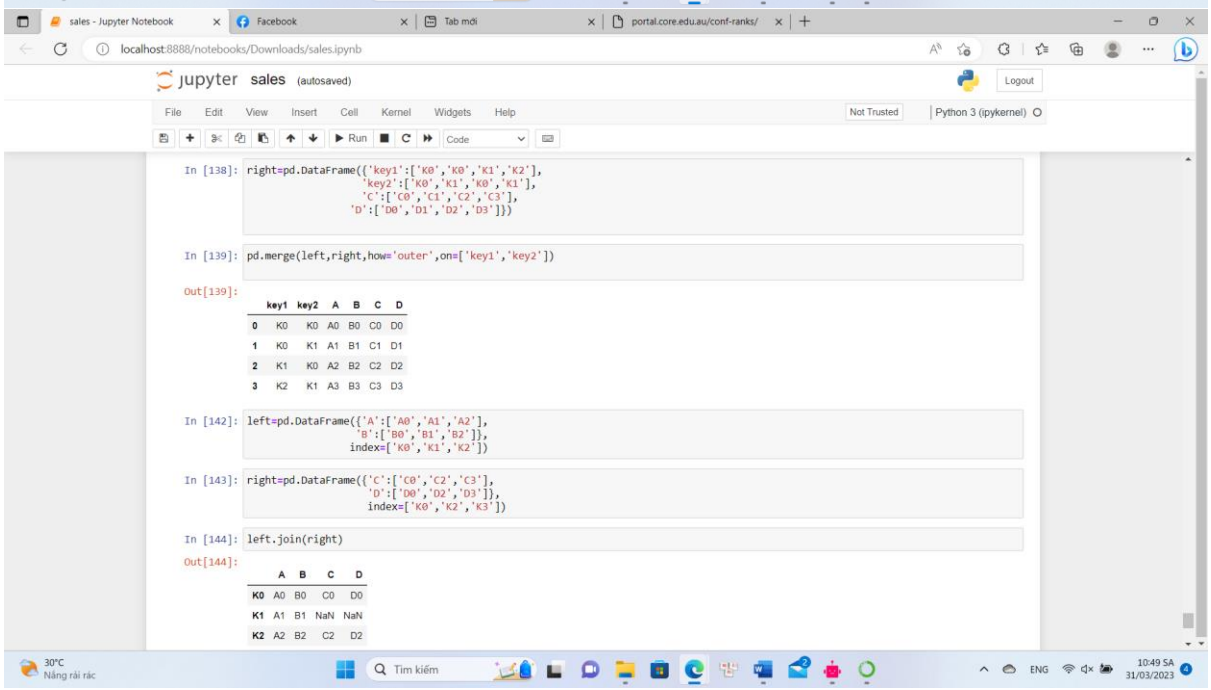
In [133]:
```
left=pd.DataFrame({'key':['K0','K1','K2','K3'],
                   'A':['A0','A1','A2','A3'],
                   'B':['B0','B1','B2','B3']})

right=pd.DataFrame({'key':['K0','K1','K2','K3'],
                    'C':['C0','C1','C2','C3'],
                    'D':['D0','D1','D2','D3']})
```

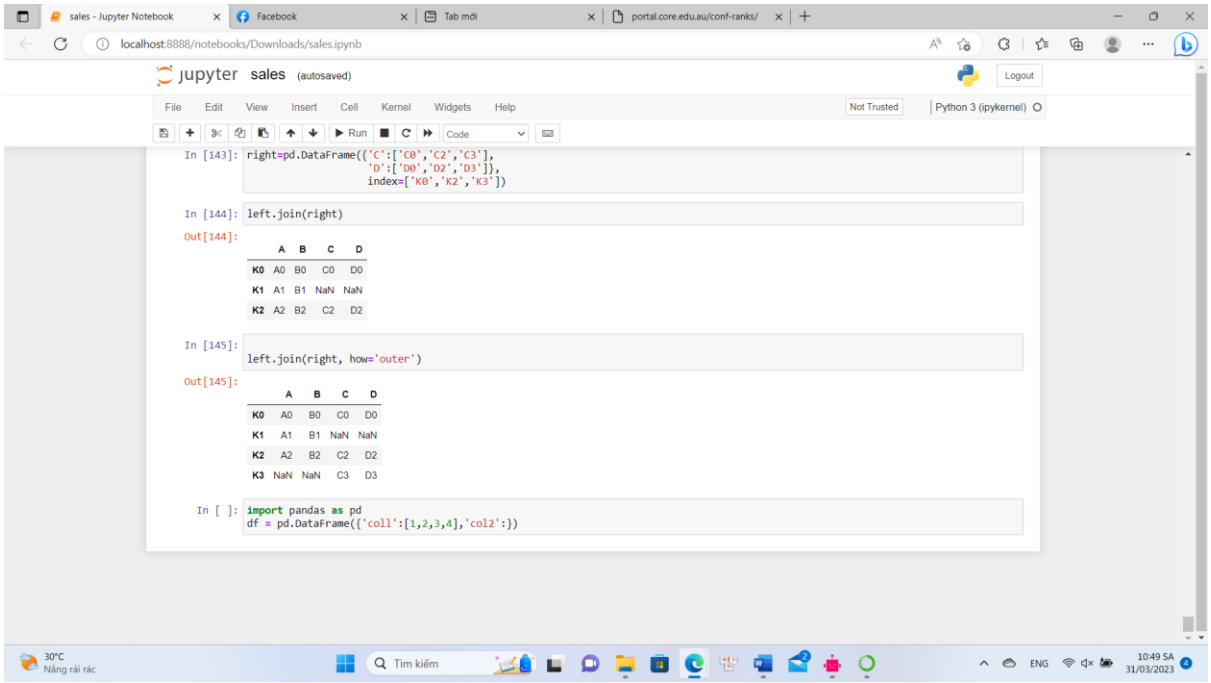
In [134]:
```
left
```

Out[134]:

30°C Nắng rải rác | Tìm kiếm | ENG | 10:49 SA 31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [134]: left
Out[134]:
key A B
0 K0 A0 B0
1 K1 A1 B1
2 K2 A2 B2
3 K3 A3 B3
In [135]: right
Out[135]:
key C D
0 K0 C0 D0
1 K1 C1 D1
2 K2 C2 D2
3 K3 C3 D3
In [136]: pd.merge(left,right,how='inner',on='key')
Out[136]:
key A B C D
0 K0 A0 B0 C0 D0
1 K1 A1 B1 C1 D1
2 K2 A2 B2 C2 D2
3 K3 A3 B3 C3 D3
In [137]: left=pd.DataFrame({'key1':['K0','K0','K1','K2'],
'key2':['K0','K1','K0','K1'],
30°C
Nắng rải rác
Tìm kiếm
ENG
10:49 SA
31/03/2023


sales - Jupyter Notebook
Facebook
Tab mới
portal.core.edu.au/conf-ranks/
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [138]: right=pd.DataFrame({'key1':['K0','K0','K1','K2'],
'key2':['K0','K1','K0','K1'],
'C':['C0','C1','C2','C3'],
'D':['D0','D1','D2','D3']})
In [139]: pd.merge(left,right,how='outer',on=['key1','key2'])
Out[139]:
key1 key2 A B C D
0 K0 K0 A0 B0 C0 D0
1 K0 K1 A1 B1 C1 D1
2 K1 K0 A2 B2 C2 D2
3 K2 K1 A3 B3 C3 D3
In [142]: left=pd.DataFrame({'A':['A0','A1','A2'],
'B':['B0','B1','B2']},
index=['K0','K1','K2'])
In [143]: right=pd.DataFrame({'C':['C0','C2','C3'],
'D':['D0','D2','D3']},
index=['K0','K2','K3'])
In [144]: left.join(right)
Out[144]:
A B C D
K0 A0 B0 C0 D0
K1 A1 B1 NaN NaN
K2 A2 B2 C2 D2
30°C
Nắng rải rác
Tìm kiếm
ENG
10:49 SA
31/03/2023



sales - Jupyter Notebook | Facebook | Tab mới | portal.core.edu.au/conf-ranks/

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

In [143]:
```
right=pd.DataFrame({'C':['C0','C2','C3'],
                    'D':['D0','D2','D3']},
                   index=['K0','K2','K3'])
```

In [144]:
```
left.join(right)
```

Out[144]:

	A	B	C	D
K0	A0	B0	C0	D0
K1	A1	B1	NaN	NaN
K2	A2	B2	C2	D2

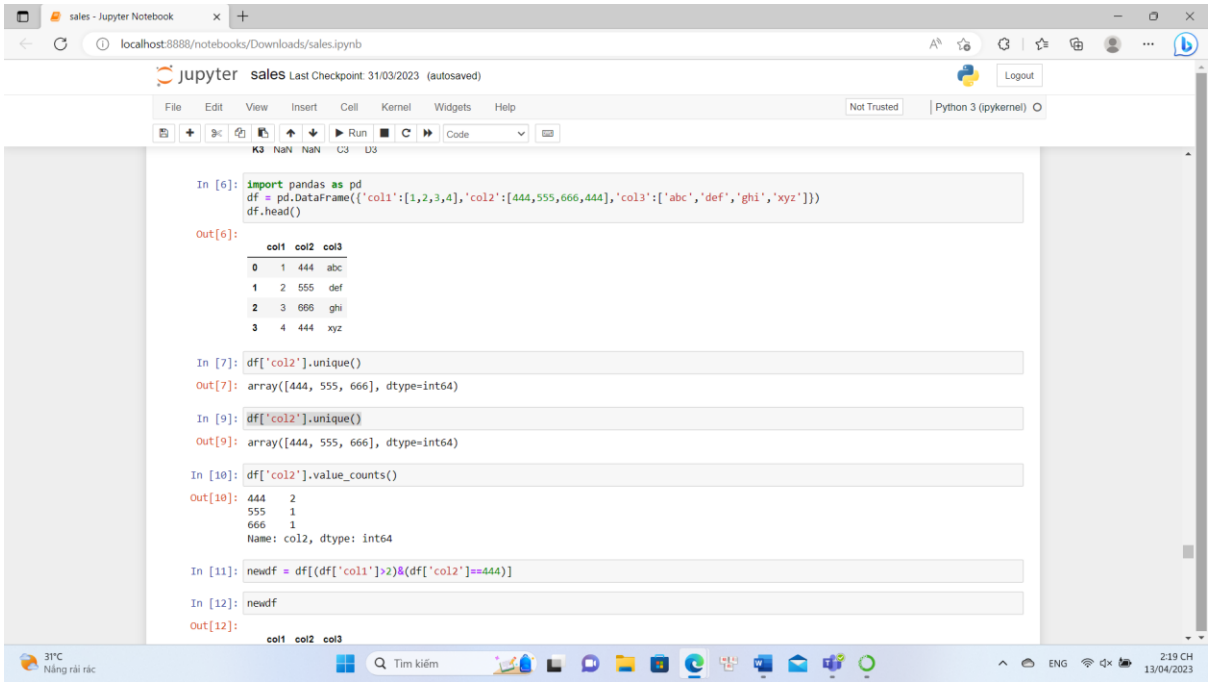
In [145]:
```
left.join(right, how='outer')
```

Out[145]:

	A	B	C	D
K0	A0	B0	C0	D0
K1	A1	B1	NaN	NaN
K2	A2	B2	C2	D2
K3	NaN	NaN	C3	D3

In [ ]:
```
import pandas as pd
df = pd.DataFrame({'col1':[1,2,3,4],'col2':})
```

30°C Nắng rải rác | Tìm kiếm | ENG | 10:49 SA 31/03/2023



sales - Jupyter Notebook

localhost:8888/notebooks/Downloads/sales.ipynb

jupyter sales Last Checkpoint: 31/03/2023 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

K3	NaN	NaN	C3	D3

In [6]:
```
import pandas as pd
df = pd.DataFrame({'col1':[1,2,3,4],'col2':[444,555,666,444],'col3':['abc','def','ghi','xyz']})
df.head()
```

Out[6]:

	col1	col2	col3
0	1	444	abc
1	2	555	def
2	3	666	ghi
3	4	444	xyz

In [7]:
```
df['col2'].unique()
```

Out[7]:
```
array([444, 555, 666], dtype=int64)
```

In [9]:
```
df['col2'].unique()
```

Out[9]:
```
array([444, 555, 666], dtype=int64)
```

In [10]:
```
df['col2'].value_counts()
```

Out[10]:
```
444    2
555    1
666    1
Name: col2, dtype: int64
```

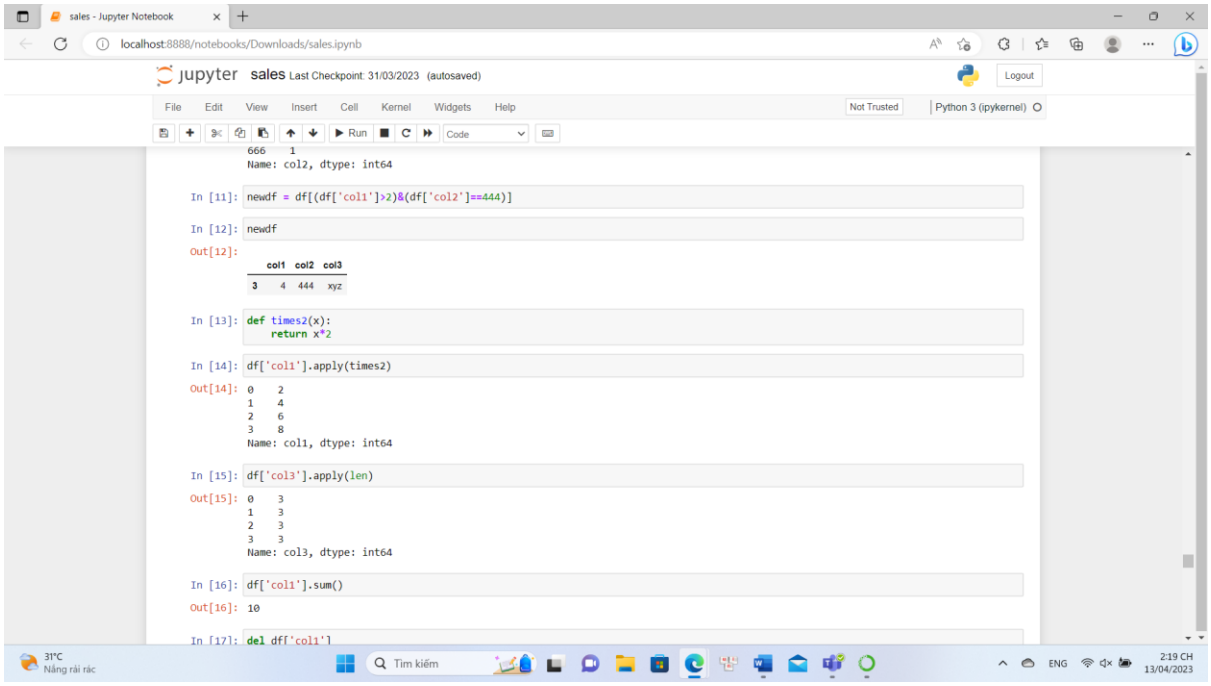
In [11]:
```
newdf = df[(df['col1']>2)&(df['col2']==444)]
```

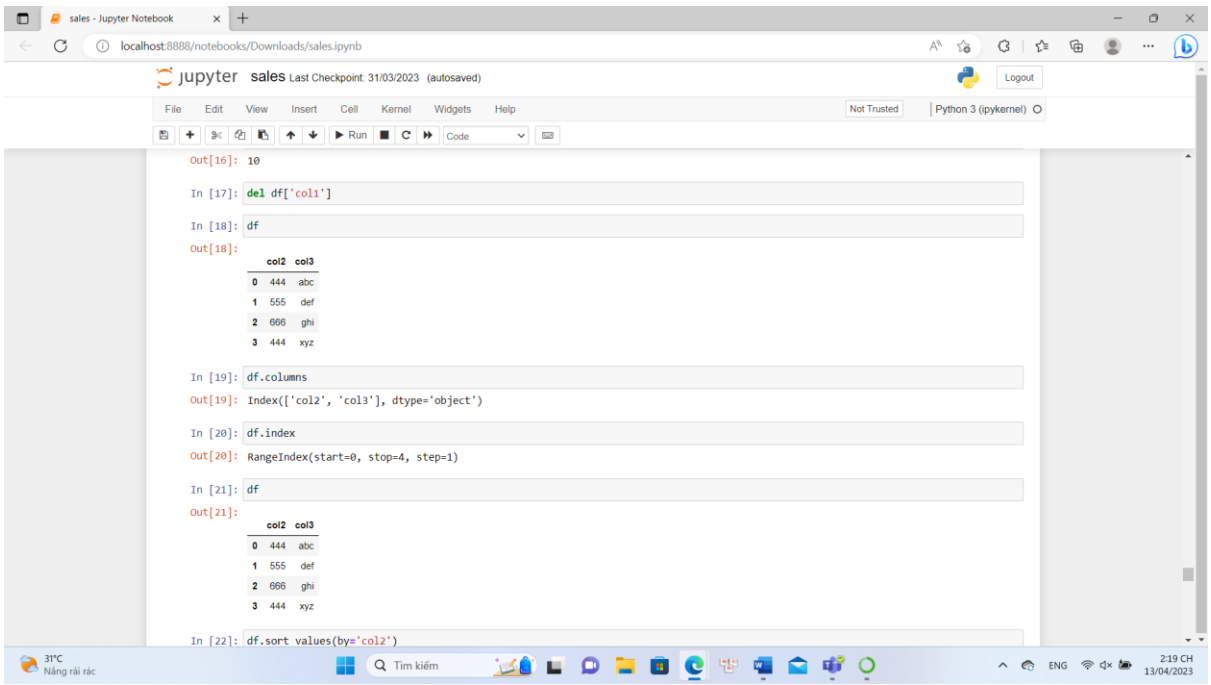
In [12]:
```
newdf
```

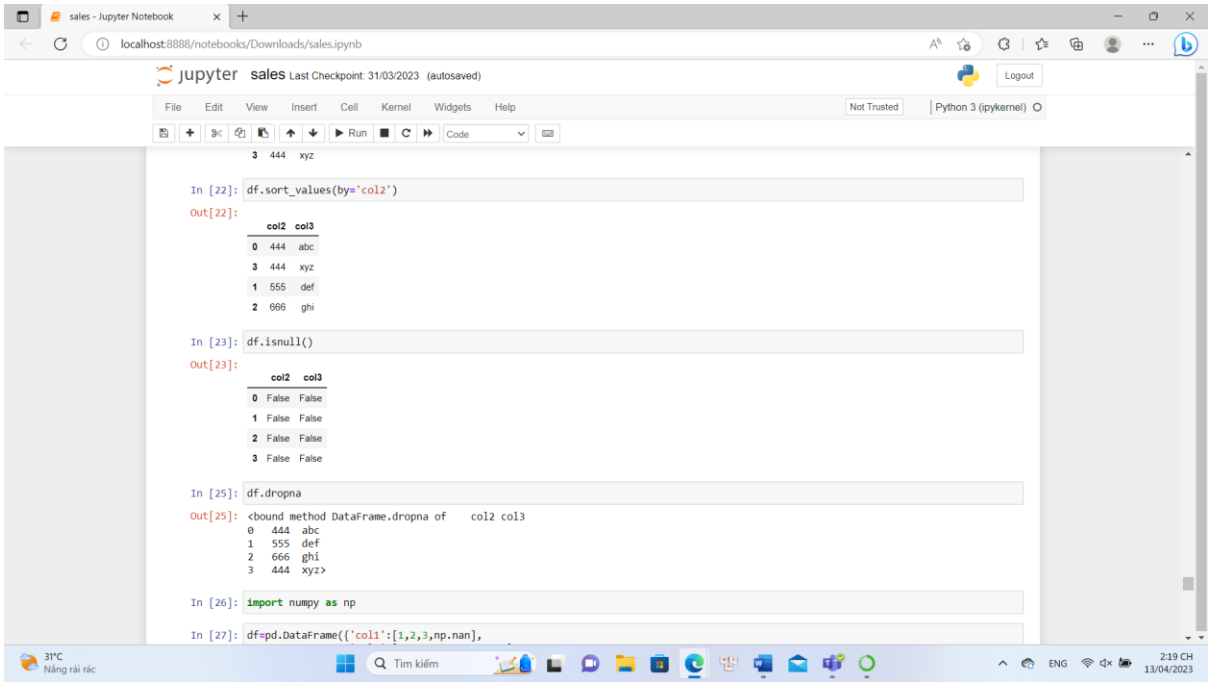
Out[12]:

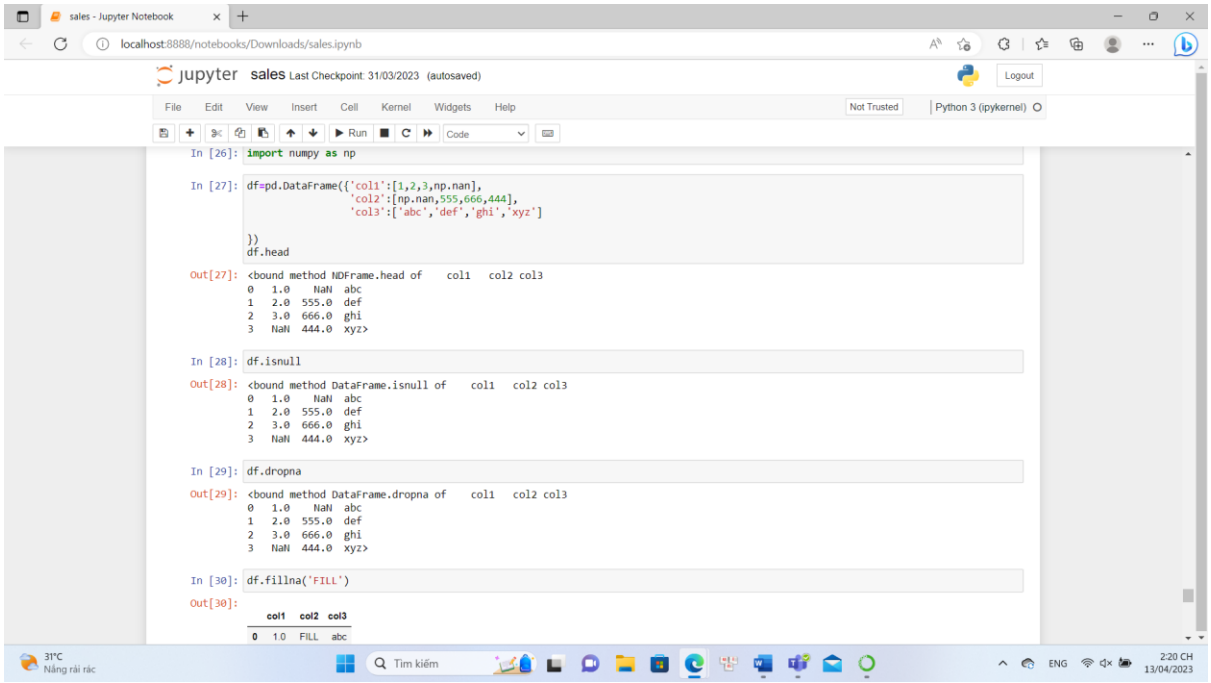
	col1	col2	col3

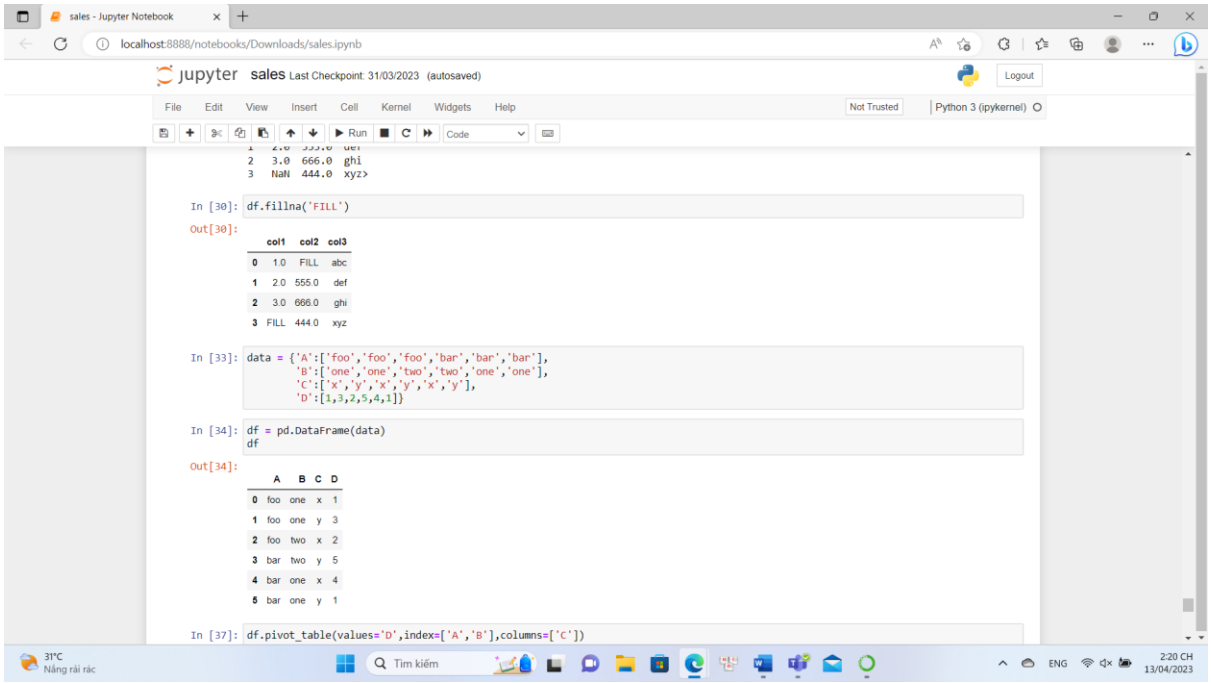
31°C Nắng rải rác | Tìm kiếm | ENG | 2:19 CH 13/04/2023

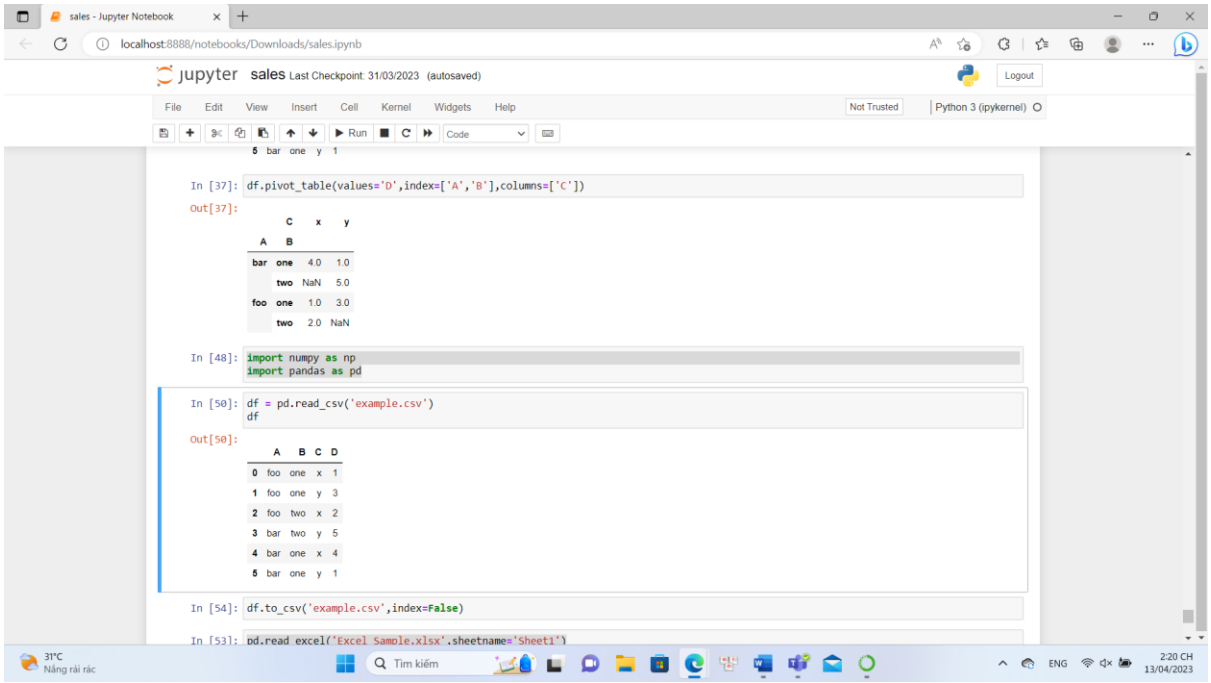

sales - Jupyter Notebook
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales Last Checkpoint: 31/03/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
666 1
Name: col2, dtype: int64
In [11]: newdf = df[(df['col1']>2)&(df['col2']==444)]
In [12]: newdf
Out[12]:
col1 col2 col3
3 4 444 xyz
In [13]: def times2(x):
return x*2
In [14]: df['col1'].apply(times2)
Out[14]: 0 2
1 4
2 6
3 8
Name: col1, dtype: int64
In [15]: df['col3'].apply(len)
Out[15]: 0 3
1 3
2 3
3 3
Name: col3, dtype: int64
In [16]: df['col1'].sum()
Out[16]: 10
In [17]: del df['col1']
31°C
Nắng rải rác
Tìm kiếm
ENG
2:19 CH
13/04/2023


sales - Jupyter Notebook
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales Last Checkpoint: 31/03/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Out[16]: 10
In [17]: del df['col1']
In [18]: df
Out[18]:
col2 col3
0 444 abc
1 555 def
2 666 ghi
3 444 xyz
In [19]: df.columns
Out[19]: Index(['col2', 'col3'], dtype='object')
In [20]: df.index
Out[20]: RangeIndex(start=0, stop=4, step=1)
In [21]: df
Out[21]:
col2 col3
0 444 abc
1 555 def
2 666 ghi
3 444 xyz
In [22]: df.sort_values(by='col2')
31°C
Nắng rải rác
Tìm kiếm
ENG
2:19 CH
13/04/2023


sales - Jupyter Notebook
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales Last Checkpoint: 31/03/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
3 444 xyz
In [22]: df.sort_values(by='col2')
Out[22]:
col2 col3
0 444 abc
3 444 xyz
1 555 def
2 666 ghi
In [23]: df.isnull()
Out[23]:
col2 col3
0 False False
1 False False
2 False False
3 False False
In [25]: df.dropna
Out[25]: <bound method DataFrame.dropna of col2 col3
0 444 abc
1 555 def
2 666 ghi
3 444 xyz>
In [26]: import numpy as np
In [27]: df=pd.DataFrame({'col1':[1,2,3,np.nan],
31°C
Nắng rải rác
Tìm kiếm
ENG
2:19 CH
13/04/2023


sales - Jupyter Notebook
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales Last Checkpoint: 31/03/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [26]: import numpy as np
In [27]: df=pd.DataFrame({'col1':[1,2,3,np.nan],
'col2':[np.nan,555,666,444],
'col3':['abc','def','ghi','xyz']
})
df.head
Out[27]: <bound method NDFrame.head of col1 col2 col3
0 1.0 NaN abc
1 2.0 555.0 def
2 3.0 666.0 ghi
3 NaN 444.0 xyz>
In [28]: df.isnull
Out[28]: <bound method DataFrame.isnull of col1 col2 col3
0 1.0 NaN abc
1 2.0 555.0 def
2 3.0 666.0 ghi
3 NaN 444.0 xyz>
In [29]: df.dropna
Out[29]: <bound method DataFrame.dropna of col1 col2 col3
0 1.0 NaN abc
1 2.0 555.0 def
2 3.0 666.0 ghi
3 NaN 444.0 xyz>
In [30]: df.fillna('FILL')
Out[30]:
col1 col2 col3
0 1.0 FILL abc
31°C
Nắng rải rác
Tìm kiếm
ENG
2:20 CH
13/04/2023


sales - Jupyter Notebook
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales Last Checkpoint: 31/03/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
2 3.0 666.0 ghi
3 NaN 444.0 xyz
In [30]: df.fillna('FILL')
Out[30]:
col1 col2 col3
0 1.0 FILL abc
1 2.0 555.0 def
2 3.0 666.0 ghi
3 FILL 444.0 xyz
In [33]: data = {'A':['foo','foo','foo','bar','bar','bar'],
'B':['one','one','two','two','one','one'],
'C':['x','y','x','y','x','y'],
'D':[1,3,2,5,4,1]}
In [34]: df = pd.DataFrame(data)
df
Out[34]:
A B C D
0 foo one x 1
1 foo one y 3
2 foo two x 2
3 bar two y 5
4 bar one x 4
5 bar one y 1
In [37]: df.pivot_table(values='D',index=['A','B'],columns=['C'])
31°C
Nắng rải rác
Tìm kiếm
ENG
2:20 CH
13/04/2023


sales - Jupyter Notebook
localhost:8888/notebooks/Downloads/sales.ipynb
jupyter sales Last Checkpoint: 31/03/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
5 bar one y 1
In [37]: df.pivot_table(values='D',index=['A','B'],columns=['C'])
Out[37]:
C x y
A B
bar one 4.0 1.0
two NaN 5.0
foo one 1.0 3.0
two 2.0 NaN
In [48]: import numpy as np
import pandas as pd
In [50]: df = pd.read_csv('example.csv')
df
Out[50]:
A B C D
0 foo one x 1
1 foo one y 3
2 foo two x 2
3 bar two y 5
4 bar one x 4
5 bar one y 1
In [54]: df.to_csv('example.csv',index=False)
In [53]: pd.read_excel('Excel Samole.xlsx',sheetname='Sheet1')
31°C
Nắng rải rác
Tìm kiếm
ENG
2:20 CH
13/04/2023