Pandas Analysis Part-2

In [3]:
```
import pandas as pd
import os
os.getcwd()
```
Out[3]:
```
'C:\\Users\\Amit'
```

In [4]:
```
os.chdir(r'C:\Users\Amit\Desktop\Data_sets')
```

In [5]:
```
tips = pd.read_csv('tips.csv')
```

In [6]:
```
tips
```
Out[6]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [7]:
```
type(tips)
```
Out[7]:
```
pandas.core.frame.DataFrame
```

In [8]:
```
len(tips)
```
Out[8]:
```
244
```

In [9]:
```
len(tips.columns)
```
Out[9]:
```
7
```

In [10]:
```
tips.head(6)
```
Out[10]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
5	25.29	4.71	Male	No	Sun	Dinner	4

In [9]:
```
tips.tail()
```
Out[9]:

	total_bill	tip	sex	smoker	day	time	size
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

In [10]:
```
tips.info()
```
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   total_bill  244 non-null    float64
 1   tip         244 non-null    float64
 2   sex         244 non-null    object
 3   smoker      244 non-null    object
 4   day         244 non-null    object
 5   time        244 non-null    object
 6   size        244 non-null    int64
dtypes: float64(2), int64(1), object(4)
memory usage: 13.5+ KB
```

In [11]:
```
tips.describe()
```
Out[11]:

	total_bill	tip	size
count	244.000000	244.000000	244.000000
mean	19.785943	2.998279	2.569672
std	8.902412	1.383638	0.951100
min	3.070000	1.000000	1.000000
25%	13.347500	2.000000	2.000000
50%	17.795000	2.900000	2.000000
75%	24.127500	3.562500	3.000000
max	50.810000	10.000000	6.000000

In [12]:
```
tips
```
Out[12]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [13]:
```
"""change the name of columns it is just executing.
to change permanently at the end ('inplace=True')"""

tips.rename(columns={'total_bill':'totalbill',
                     'sex':'gender',
                     'size':'count'}, inplace=True)
```

In [14]:
```
tips
```
Out[14]:

	totalbill	tip	gender	smoker	day	time	count
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [15]:
```
tips
```
Out[15]:

	totalbill	tip	gender	smoker	day	time	count
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [16]:
```
#change to origional name
tips.rename(columns={'totalbill':'total_bill',
                     'gender':'sex',
                     'count':'size'}, inplace=True)
```

In [17]:
```
#selecting required reccords form file
tips[1:4][['total_bill','size']]
```
Out[17]:

	total_bill	size
1	10.34	3
2	21.01	3
3	23.68	2

In [18]:
```
tips[['total_bill','tip','smoker','time']].head()
```
Out[18]:

	total_bill	tip	smoker	time
0	16.99	1.01	No	Dinner
1	10.34	1.66	No	Dinner
2	21.01	3.50	No	Dinner
3	23.68	3.31	No	Dinner
4	24.59	3.61	No	Dinner

In [19]:
```
#we want all the 'Dinner' from 'time' (filter)
tips[tips['time']=='Dinner']
```
Out[19]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

176 rows × 7 columns

In [20]:
```
len(tips[tips['time']=='Dinner'])
```
Out[20]:
```
176
```

In [21]:
```
#if we want tip>5 with time == 'Dinner'
tips[(tips['time']=='Dinner') & (tips['tip']>5)]
```
Out[21]:

	total_bill	tip	sex	smoker	day	time	size
23	39.42	7.58	Male	No	Sat	Dinner	4
44	30.40	5.60	Male	No	Sun	Dinner	4
47	32.40	6.00	Male	No	Sun	Dinner	4
52	34.81	5.20	Female	No	Sun	Dinner	4
59	48.27	6.73	Male	No	Sat	Dinner	4
116	29.93	5.07	Male	No	Sun	Dinner	4
155	29.85	5.14	Female	No	Sun	Dinner	5
170	50.81	10.00	Male	Yes	Sat	Dinner	3
172	7.25	5.15	Male	Yes	Sun	Dinner	2
181	23.33	5.65	Male	Yes	Sun	Dinner	2
183	23.17	6.50	Male	Yes	Sun	Dinner	4
211	25.89	5.16	Male	Yes	Sat	Dinner	4
212	48.33	9.00	Male	No	Sat	Dinner	4
214	28.17	6.50	Female	Yes	Sat	Dinner	3
239	29.03	5.92	Male	No	Sat	Dinner	3

In [22]:
```
len(tips[(tips['time']=='Dinner') & (tips['tip']>5)])
```
Out[22]:
```
15
```

In [23]:
```
# in this '|' represents 'or'
tips[(tips['size']>5) | (tips['total_bill']>45)]
```
Out[23]:

	total_bill	tip	sex	smoker	day	time	size
59	48.27	6.73	Male	No	Sat	Dinner	4
125	29.80	4.20	Female	No	Thur	Lunch	6
141	34.30	6.70	Male	No	Thur	Lunch	6
143	27.05	5.00	Female	No	Thur	Lunch	6
156	48.17	5.00	Male	No	Sun	Dinner	6
170	50.81	10.00	Male	Yes	Sat	Dinner	3
182	45.35	3.50	Male	Yes	Sun	Dinner	3
212	48.33	9.00	Male	No	Sat	Dinner	4

In [24]:
```
#this gives unique values from 'day'
tips.day.unique()
```
Out[24]:
```
array(['Sun', 'Sat', 'Thur', 'Fri'], dtype=object)
```

In [25]:
```
tips[(tips['day']=='Thur') | (tips['day']=='Fri')]
```
Out[25]:

	total_bill	tip	sex	smoker	day	time	size
77	27.20	4.00	Male	No	Thur	Lunch	4
78	22.76	3.00	Male	No	Thur	Lunch	2
79	17.29	2.71	Male	No	Thur	Lunch	2
80	19.44	3.00	Male	Yes	Thur	Lunch	2
81	16.66	3.40	Male	No	Thur	Lunch	2
...	...	...	...	...	...	...	...
223	15.98	3.00	Female	No	Fri	Lunch	3
224	13.42	1.58	Male	Yes	Fri	Lunch	2
225	16.27	2.50	Female	Yes	Fri	Lunch	2
226	10.09	2.00	Female	Yes	Fri	Lunch	2
243	18.78	3.00	Female	No	Thur	Dinner	2

81 rows × 7 columns

In [26]:
```
#it is used to enter multiple record to filter
tips[tips['day'].isin(['Thur','Fri'])]
```
Out[26]:

	total_bill	tip	sex	smoker	day	time	size
77	27.20	4.00	Male	No	Thur	Lunch	4
78	22.76	3.00	Male	No	Thur	Lunch	2
79	17.29	2.71	Male	No	Thur	Lunch	2
80	19.44	3.00	Male	Yes	Thur	Lunch	2
81	16.66	3.40	Male	No	Thur	Lunch	2
...	...	...	...	...	...	...	...
223	15.98	3.00	Female	No	Fri	Lunch	3
224	13.42	1.58	Male	Yes	Fri	Lunch	2
225	16.27	2.50	Female	Yes	Fri	Lunch	2
226	10.09	2.00	Female	Yes	Fri	Lunch	2
243	18.78	3.00	Female	No	Thur	Dinner	2

81 rows × 7 columns

In [27]:
```
tips[tips['day'].isin(['Thur','Fri','sun'])]
```
Out[27]:

	total_bill	tip	sex	smoker	day	time	size
77	27.20	4.00	Male	No	Thur	Lunch	4
78	22.76	3.00	Male	No	Thur	Lunch	2
79	17.29	2.71	Male	No	Thur	Lunch	2
80	19.44	3.00	Male	Yes	Thur	Lunch	2
81	16.66	3.40	Male	No	Thur	Lunch	2
...	...	...	...	...	...	...	...
223	15.98	3.00	Female	No	Fri	Lunch	3
224	13.42	1.58	Male	Yes	Fri	Lunch	2
225	16.27	2.50	Female	Yes	Fri	Lunch	2
226	10.09	2.00	Female	Yes	Fri	Lunch	2
243	18.78	3.00	Female	No	Thur	Dinner	2

81 rows × 7 columns

In [28]:
```
#to find null value
tips.isnull().sum()
```
Out[28]:
```
total_bill    0
tip           0
sex           0
smoker        0
day           0
time          0
size          0
dtype: int64
```

In [29]:
```
#to find how many Male and Female are there
tips.groupby('sex').size()
```
Out[29]:
```
sex
Female     87
Male      157
dtype: int64
```

In [30]:
```
tips.groupby(['sex', 'smoker']).size()
```
Out[30]:
```
sex     smoker
Female  No        54
        Yes       33
Male    No        97
        Yes       60
dtype: int64
```

In [31]:
```
import numpy as np
tips.groupby('sex').agg({'tip':np.mean})
```
Out[31]:

	tip
sex	
Female	2.833448
Male	3.089618

In [32]:
```
"""how many male with or without smoker giving tip
   and also for female --> this is average tip"""

tips.groupby(['sex','smoker']).agg({'tip':np.mean})
```
Out[32]:

sex	smoker	tip
Female	No	2.773519
	Yes	2.931515
Male	No	3.113402
	Yes	3.051167

In [33]:
```
#overall average tip
np.mean(tips[['tip']])
```
Out[33]:
```
tip    2.998279
dtype: float64
```

In [34]:
```
sum(tips['tip'])/len(tips)
```
Out[34]:
```
2.9982786885245902
```

In [55]:
```
#iloc means 'i' for index 'loc' for location --> index location
tips.iloc[0:5]
```
Out[55]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

In [36]:
```
tips.iloc[0:4, 0:3]
```
Out[36]:

	total_bill	tip	sex
0	16.99	1.01	Female
1	10.34	1.66	Male
2	21.01	3.50	Male
3	23.68	3.31	Male

In [37]:
```
#if we want 0,10,20,30 row only and 0,3,5 gives columns
tips.iloc[[0,10,20,30],[0,3,5]]
```
Out[37]:

	total_bill	smoker	time
0	16.99	No	Dinner
10	10.27	No	Dinner
20	17.92	No	Dinner
30	9.55	No	Dinner

In [11]:
```
#converts dataframe to array
tips.iloc[[0,10,20,30],[0,3,5]].values
```
Out[11]:
```
array([[16.99, 'No', 'Dinner'],
       [10.27, 'No', 'Dinner'],
       [17.92, 'No', 'Dinner'],
       [9.55, 'No', 'Dinner']], dtype=object)
```

In [ ]:
```

```