Data with Python - Examples

May 5, 2018

```
In [57]: #ipython
```

```
In [58]: import pandas as pd
```

1 pandas: load data

```
In [59]: DATA_PATH = '/usr/lib/python3/dist-packages/pandas/tests/data/tips.csv'
```

```
In [60]: #!cat /usr/lib/python3/dist-packages/pandas/tests/data/tips.csv
```

```
In [61]: data = pd.read_csv(DATA_PATH)
```

```
In [62]: data
```

Out[62]:

	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
6	8.77	2.00	Male	No	Sun	Dinner	2
7	26.88	3.12	Male	No	Sun	Dinner	4
8	15.04	1.96	Male	No	Sun	Dinner	2
9	14.78	3.23	Male	No	Sun	Dinner	2
10	10.27	1.71	Male	No	Sun	Dinner	2
11	35.26	5.00	Female	No	Sun	Dinner	4
12	15.42	1.57	Male	No	Sun	Dinner	2
13	18.43	3.00	Male	No	Sun	Dinner	4
14	14.83	3.02	Female	No	Sun	Dinner	2
15	21.58	3.92	Male	No	Sun	Dinner	2
16	10.33	1.67	Female	No	Sun	Dinner	3
17	16.29	3.71	Male	No	Sun	Dinner	3
18	16.97	3.50	Female	No	Sun	Dinner	3
19	20.65	3.35	Male	No	Sat	Dinner	3
20	17.92	4.08	Male	No	Sat	Dinner	2
21	20.29	2.75	Female	No	Sat	Dinner	2

```
22         15.77  2.23  Female     No   Sat  Dinner     2
23         39.42  7.58    Male     No   Sat  Dinner     4
24         19.82  3.18    Male     No   Sat  Dinner     2
25         17.81  2.34    Male     No   Sat  Dinner     4
26         13.37  2.00    Male     No   Sat  Dinner     2
27         12.69  2.00    Male     No   Sat  Dinner     2
28         21.70  4.30    Male     No   Sat  Dinner     2
29         19.65  3.00  Female     No   Sat  Dinner     2
..           ...   ...     ...    ...   ...     ...   ...
214        28.17  6.50  Female    Yes   Sat  Dinner     3
215        12.90  1.10  Female    Yes   Sat  Dinner     2
216        28.15  3.00    Male    Yes   Sat  Dinner     5
217        11.59  1.50    Male    Yes   Sat  Dinner     2
218         7.74  1.44    Male    Yes   Sat  Dinner     2
219        30.14  3.09  Female    Yes   Sat  Dinner     4
220        12.16  2.20    Male    Yes   Fri   Lunch     2
221        13.42  3.48  Female    Yes   Fri   Lunch     2
222         8.58  1.92    Male    Yes   Fri   Lunch     1
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
227        20.45  3.00    Male     No   Sat  Dinner     4
228        13.28  2.72    Male     No   Sat  Dinner     2
229        22.12  2.88  Female    Yes   Sat  Dinner     2
230        24.01  2.00    Male    Yes   Sat  Dinner     4
231        15.69  3.00    Male    Yes   Sat  Dinner     3
232        11.61  3.39    Male     No   Sat  Dinner     2
233        10.77  1.47    Male     No   Sat  Dinner     2
234        15.53  3.00    Male    Yes   Sat  Dinner     2
235        10.07  1.25    Male     No   Sat  Dinner     2
236        12.60  1.00    Male    Yes   Sat  Dinner     2
237        32.83  1.17    Male    Yes   Sat  Dinner     2
238        35.83  4.67  Female     No   Sat  Dinner     3
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 x 7 columns]
```

```
In [63]: data.sort_values('tip')
```

```
Out[63]:      total_bill   tip     sex smoker   day    time  size
67          3.07  1.00  Female    Yes   Sat  Dinner     1
236        12.60  1.00    Male    Yes   Sat  Dinner     2
92          5.75  1.00  Female    Yes   Fri  Dinner     2
```

111	7.25	1.00	Female	No	Sat	Dinner	1
0	16.99	1.01	Female	No	Sun	Dinner	2
215	12.90	1.10	Female	Yes	Sat	Dinner	2
237	32.83	1.17	Male	Yes	Sat	Dinner	2
235	10.07	1.25	Male	No	Sat	Dinner	2
75	10.51	1.25	Male	No	Sat	Dinner	2
135	8.51	1.25	Female	No	Thur	Lunch	2
43	9.68	1.32	Male	No	Sun	Dinner	2
146	18.64	1.36	Female	No	Thur	Lunch	3
218	7.74	1.44	Male	Yes	Sat	Dinner	2
195	7.56	1.44	Male	No	Thur	Lunch	2
30	9.55	1.45	Male	No	Sat	Dinner	2
233	10.77	1.47	Male	No	Sat	Dinner	2
126	8.52	1.48	Male	No	Thur	Lunch	2
190	15.69	1.50	Male	Yes	Sun	Dinner	2
57	26.41	1.50	Female	No	Sat	Dinner	2
97	12.03	1.50	Male	Yes	Fri	Dinner	2
99	12.46	1.50	Male	No	Fri	Dinner	2
117	10.65	1.50	Female	No	Thur	Lunch	2
132	11.17	1.50	Female	No	Thur	Lunch	2
217	11.59	1.50	Male	Yes	Sat	Dinner	2
145	8.35	1.50	Female	No	Thur	Lunch	2
130	19.08	1.50	Male	No	Thur	Lunch	2
53	9.94	1.56	Male	No	Sun	Dinner	2
12	15.42	1.57	Male	No	Sun	Dinner	2
224	13.42	1.58	Male	Yes	Fri	Lunch	2
168	10.59	1.61	Female	Yes	Sat	Dinner	2
..	...	...	...	...	...	...	...
5	25.29	4.71	Male	No	Sun	Dinner	4
95	40.17	4.73	Male	Yes	Fri	Dinner	4
46	22.23	5.00	Male	No	Sun	Dinner	2
39	31.27	5.00	Male	No	Sat	Dinner	3
142	41.19	5.00	Male	No	Thur	Lunch	5
143	27.05	5.00	Female	No	Thur	Lunch	6
156	48.17	5.00	Male	No	Sun	Dinner	6
185	20.69	5.00	Male	No	Sun	Dinner	5
11	35.26	5.00	Female	No	Sun	Dinner	4
83	32.68	5.00	Male	Yes	Thur	Lunch	2
197	43.11	5.00	Female	Yes	Thur	Lunch	4
73	25.28	5.00	Female	Yes	Sat	Dinner	2
116	29.93	5.07	Male	No	Sun	Dinner	4
155	29.85	5.14	Female	No	Sun	Dinner	5
172	7.25	5.15	Male	Yes	Sun	Dinner	2
211	25.89	5.16	Male	Yes	Sat	Dinner	4
85	34.83	5.17	Female	No	Thur	Lunch	4
52	34.81	5.20	Female	No	Sun	Dinner	4
44	30.40	5.60	Male	No	Sun	Dinner	4
181	23.33	5.65	Male	Yes	Sun	Dinner	2

```
88         24.71   5.85    Male     No  Thur   Lunch     2
239        29.03   5.92    Male     No   Sat  Dinner     3
47         32.40   6.00    Male     No   Sun  Dinner     4
183        23.17   6.50    Male    Yes   Sun  Dinner     4
214        28.17   6.50  Female    Yes   Sat  Dinner     3
141        34.30   6.70    Male     No  Thur   Lunch     6
59         48.27   6.73    Male     No   Sat  Dinner     4
23         39.42   7.58    Male     No   Sat  Dinner     4
212        48.33   9.00    Male     No   Sat  Dinner     4
170        50.81  10.00    Male    Yes   Sat  Dinner     3

[244 rows x 7 columns]
```

```
In [64]: data.head()
```

```
Out[64]:    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 [65]: data['tip']
```

```
Out[65]: 0      1.01
1      1.66
2      3.50
3      3.31
4      3.61
5      4.71
6      2.00
7      3.12
8      1.96
9      3.23
10     1.71
11     5.00
12     1.57
13     3.00
14     3.02
15     3.92
16     1.67
17     3.71
18     3.50
19     3.35
20     4.08
21     2.75
22     2.23
23     7.58
24     3.18
```

```
25     2.34
26     2.00
27     2.00
28     4.30
29     3.00
       ...
214    6.50
215    1.10
216    3.00
217    1.50
218    1.44
219    3.09
220    2.20
221    3.48
222    1.92
223    3.00
224    1.58
225    2.50
226    2.00
227    3.00
228    2.72
229    2.88
230    2.00
231    3.00
232    3.39
233    1.47
234    3.00
235    1.25
236    1.00
237    1.17
238    4.67
239    5.92
240    2.00
241    2.00
242    1.75
243    3.00
Name: tip, Length: 244, dtype: float64
```

```
In [66]: data['tip'] / data['total_bill']
```

```
Out[66]: 0      0.059447
         1      0.160542
         2      0.166587
         3      0.139780
         4      0.146808
         5      0.186240
         6      0.228050
         7      0.116071
```

```
8      0.130319
9      0.218539
10     0.166504
11     0.141804
12     0.101816
13     0.162778
14     0.203641
15     0.181650
16     0.161665
17     0.227747
18     0.206246
19     0.162228
20     0.227679
21     0.135535
22     0.141408
23     0.192288
24     0.160444
25     0.131387
26     0.149589
27     0.157604
28     0.198157
29     0.152672
         ...
214    0.230742
215    0.085271
216    0.106572
217    0.129422
218    0.186047
219    0.102522
220    0.180921
221    0.259314
222    0.223776
223    0.187735
224    0.117735
225    0.153657
226    0.198216
227    0.146699
228    0.204819
229    0.130199
230    0.083299
231    0.191205
232    0.291990
233    0.136490
234    0.193175
235    0.124131
236    0.079365
237    0.035638
238    0.130338
```

```
239    0.203927
240    0.073584
241    0.088222
242    0.098204
243    0.159744
Length: 244, dtype: float64
```

```
In [67]: data['perc_tip'] = data['tip'] / data['total_bill']
```

```
In [68]: data.head()
```

Out[68]:

	total_bill	tip	sex	smoker	day	time	size	perc_tip
0	16.99	1.01	Female	No	Sun	Dinner	2	0.059447
1	10.34	1.66	Male	No	Sun	Dinner	3	0.160542
2	21.01	3.50	Male	No	Sun	Dinner	3	0.166587
3	23.68	3.31	Male	No	Sun	Dinner	2	0.139780
4	24.59	3.61	Female	No	Sun	Dinner	4	0.146808

```
In [69]: data[data.sex == 'Female']
```

Out[69]:

	total_bill	tip	sex	smoker	day	time	size	perc_tip
0	16.99	1.01	Female	No	Sun	Dinner	2	0.059447
4	24.59	3.61	Female	No	Sun	Dinner	4	0.146808
11	35.26	5.00	Female	No	Sun	Dinner	4	0.141804
14	14.83	3.02	Female	No	Sun	Dinner	2	0.203641
16	10.33	1.67	Female	No	Sun	Dinner	3	0.161665
18	16.97	3.50	Female	No	Sun	Dinner	3	0.206246
21	20.29	2.75	Female	No	Sat	Dinner	2	0.135535
22	15.77	2.23	Female	No	Sat	Dinner	2	0.141408
29	19.65	3.00	Female	No	Sat	Dinner	2	0.152672
32	15.06	3.00	Female	No	Sat	Dinner	2	0.199203
33	20.69	2.45	Female	No	Sat	Dinner	4	0.118415
37	16.93	3.07	Female	No	Sat	Dinner	3	0.181335
51	10.29	2.60	Female	No	Sun	Dinner	2	0.252672
52	34.81	5.20	Female	No	Sun	Dinner	4	0.149382
57	26.41	1.50	Female	No	Sat	Dinner	2	0.056797
66	16.45	2.47	Female	No	Sat	Dinner	2	0.150152
67	3.07	1.00	Female	Yes	Sat	Dinner	1	0.325733
71	17.07	3.00	Female	No	Sat	Dinner	3	0.175747
72	26.86	3.14	Female	Yes	Sat	Dinner	2	0.116902
73	25.28	5.00	Female	Yes	Sat	Dinner	2	0.197785
74	14.73	2.20	Female	No	Sat	Dinner	2	0.149355
82	10.07	1.83	Female	No	Thur	Lunch	1	0.181728
85	34.83	5.17	Female	No	Thur	Lunch	4	0.148435
92	5.75	1.00	Female	Yes	Fri	Dinner	2	0.173913
93	16.32	4.30	Female	Yes	Fri	Dinner	2	0.263480
94	22.75	3.25	Female	No	Fri	Dinner	2	0.142857
100	11.35	2.50	Female	Yes	Fri	Dinner	2	0.220264
101	15.38	3.00	Female	Yes	Fri	Dinner	2	0.195059

```
102       44.30  2.50  Female    Yes   Sat  Dinner     3  0.056433
103       22.42  3.48  Female    Yes   Sat  Dinner     2  0.155219
..          ...   ...     ...    ...   ...     ...   ...       ...
155       29.85  5.14  Female     No   Sun  Dinner     5  0.172194
157       25.00  3.75  Female     No   Sun  Dinner     4  0.150000
158       13.39  2.61  Female     No   Sun  Dinner     2  0.194922
162       16.21  2.00  Female     No   Sun  Dinner     3  0.123381
164       17.51  3.00  Female    Yes   Sun  Dinner     2  0.171331
168       10.59  1.61  Female    Yes   Sat  Dinner     2  0.152030
169       10.63  2.00  Female    Yes   Sat  Dinner     2  0.188147
178        9.60  4.00  Female    Yes   Sun  Dinner     2  0.416667
186       20.90  3.50  Female    Yes   Sun  Dinner     3  0.167464
188       18.15  3.50  Female    Yes   Sun  Dinner     3  0.192837
191       19.81  4.19  Female    Yes  Thur   Lunch     2  0.211509
197       43.11  5.00  Female    Yes  Thur   Lunch     4  0.115982
198       13.00  2.00  Female    Yes  Thur   Lunch     2  0.153846
201       12.74  2.01  Female    Yes  Thur   Lunch     2  0.157771
202       13.00  2.00  Female    Yes  Thur   Lunch     2  0.153846
203       16.40  2.50  Female    Yes  Thur   Lunch     2  0.152439
205       16.47  3.23  Female    Yes  Thur   Lunch     3  0.196114
209       12.76  2.23  Female    Yes   Sat  Dinner     2  0.174765
213       13.27  2.50  Female    Yes   Sat  Dinner     2  0.188395
214       28.17  6.50  Female    Yes   Sat  Dinner     3  0.230742
215       12.90  1.10  Female    Yes   Sat  Dinner     2  0.085271
219       30.14  3.09  Female    Yes   Sat  Dinner     4  0.102522
221       13.42  3.48  Female    Yes   Fri   Lunch     2  0.259314
223       15.98  3.00  Female     No   Fri   Lunch     3  0.187735
225       16.27  2.50  Female    Yes   Fri   Lunch     2  0.153657
226       10.09  2.00  Female    Yes   Fri   Lunch     2  0.198216
229       22.12  2.88  Female    Yes   Sat  Dinner     2  0.130199
238       35.83  4.67  Female     No   Sat  Dinner     3  0.130338
240       27.18  2.00  Female    Yes   Sat  Dinner     2  0.073584
243       18.78  3.00  Female     No  Thur  Dinner     2  0.159744

[87 rows x 8 columns]
```

```
In [70]: data[data.sex == 'Female'].to_csv('waitresses.csv')

In [71]: !cat waitresses.csv

,total_bill,tip,sex,smoker,day,time,size,perc_tip
0,16.99,1.01,Female,No,Sun,Dinner,2,0.05944673337257211
4,24.59,3.61,Female,No,Sun,Dinner,4,0.14680764538430255
11,35.26,5.0,Female,No,Sun,Dinner,4,0.14180374361883155
14,14.83,3.02,Female,No,Sun,Dinner,2,0.20364126770060686
16,10.33,1.67,Female,No,Sun,Dinner,3,0.1616650532429816
18,16.97,3.5,Female,No,Sun,Dinner,3,0.20624631703005306
21,20.29,2.75,Female,No,Sat,Dinner,2,0.13553474618038444
```

```
22,15.77,2.23,Female,No,Sat,Dinner,2,0.14140773620798985
29,19.65,3.0,Female,No,Sat,Dinner,2,0.15267175572519084
32,15.06,3.0,Female,No,Sat,Dinner,2,0.199203187250996
33,20.69,2.45,Female,No,Sat,Dinner,4,0.11841469308844853
37,16.93,3.07,Female,No,Sat,Dinner,3,0.1813349084465446
51,10.29,2.6,Female,No,Sun,Dinner,2,0.2526724975704568
52,34.81,5.2,Female,No,Sun,Dinner,4,0.14938236139040506
57,26.41,1.5,Female,No,Sat,Dinner,2,0.05679666792881484
66,16.45,2.47,Female,No,Sat,Dinner,2,0.1501519756838906
67,3.07,1.0,Female,Yes,Sat,Dinner,1,0.32573289902280134
71,17.07,3.0,Female,No,Sat,Dinner,3,0.1757469244288225
72,26.86,3.14,Female,Yes,Sat,Dinner,2,0.11690245718540582
73,25.28,5.0,Female,Yes,Sat,Dinner,2,0.19778481012658228
74,14.73,2.2,Female,No,Sat,Dinner,2,0.14935505770536323
82,10.07,1.83,Female,No,Thur,Lunch,1,0.1817279046673287
85,34.83,5.17,Female,No,Thur,Lunch,4,0.14843525696238874
92,5.75,1.0,Female,Yes,Fri,Dinner,2,0.17391304347826086
93,16.32,4.3,Female,Yes,Fri,Dinner,2,0.26348039215686275
94,22.75,3.25,Female,No,Fri,Dinner,2,0.14285714285714285
100,11.35,2.5,Female,Yes,Fri,Dinner,2,0.22026431718061676
101,15.38,3.0,Female,Yes,Fri,Dinner,2,0.19505851755526657
102,44.3,2.5,Female,Yes,Sat,Dinner,3,0.05643340857787811
103,22.42,3.48,Female,Yes,Sat,Dinner,2,0.15521855486173058
104,20.92,4.08,Female,No,Sat,Dinner,2,0.1950286806883365
109,14.31,4.0,Female,Yes,Sat,Dinner,2,0.2795248078266946
111,7.25,1.0,Female,No,Sat,Dinner,1,0.13793103448275862
114,25.71,4.0,Female,No,Sun,Dinner,3,0.15558148580318942
115,17.31,3.5,Female,No,Sun,Dinner,2,0.20219526285384173
117,10.65,1.5,Female,No,Thur,Lunch,2,0.14084507042253522
118,12.43,1.8,Female,No,Thur,Lunch,2,0.14481094127111827
119,24.08,2.92,Female,No,Thur,Lunch,4,0.1212624584717608
121,13.42,1.68,Female,No,Thur,Lunch,2,0.12518628912071533
124,12.48,2.52,Female,No,Thur,Lunch,2,0.20192307692307693
125,29.8,4.2,Female,No,Thur,Lunch,6,0.14093959731543623
127,14.52,2.0,Female,No,Thur,Lunch,2,0.13774104683195593
128,11.38,2.0,Female,No,Thur,Lunch,2,0.17574692442882248
131,20.27,2.83,Female,No,Thur,Lunch,2,0.13961519486926494
132,11.17,1.5,Female,No,Thur,Lunch,2,0.13428827215756492
133,12.26,2.0,Female,No,Thur,Lunch,2,0.1631321370309951
134,18.26,3.25,Female,No,Thur,Lunch,2,0.17798466593647316
135,8.51,1.25,Female,No,Thur,Lunch,2,0.14688601645123384
136,10.33,2.0,Female,No,Thur,Lunch,2,0.1936108422071636
137,14.15,2.0,Female,No,Thur,Lunch,2,0.1413427561837456
139,13.16,2.75,Female,No,Thur,Lunch,2,0.20896656534954408
140,17.47,3.5,Female,No,Thur,Lunch,2,0.20034344590726963
143,27.05,5.0,Female,No,Thur,Lunch,6,0.18484288354898337
144,16.43,2.3,Female,No,Thur,Lunch,2,0.1399878271454656
145,8.35,1.5,Female,No,Thur,Lunch,2,0.17964071856287425
```

```
146,18.64,1.36,Female,No,Thur,Lunch,3,0.07296137339055794
147,11.87,1.63,Female,No,Thur,Lunch,2,0.13732097725358045
155,29.85,5.14,Female,No,Sun,Dinner,5,0.1721943048576214
157,25.0,3.75,Female,No,Sun,Dinner,4,0.15
158,13.39,2.61,Female,No,Sun,Dinner,2,0.19492158327109782
162,16.21,2.0,Female,No,Sun,Dinner,3,0.12338062924120913
164,17.51,3.0,Female,Yes,Sun,Dinner,2,0.17133066818960593
168,10.59,1.61,Female,Yes,Sat,Dinner,2,0.15203021718602455
169,10.63,2.0,Female,Yes,Sat,Dinner,2,0.18814675446848542
178,9.6,4.0,Female,Yes,Sun,Dinner,2,0.4166666666666667
186,20.9,3.5,Female,Yes,Sun,Dinner,3,0.1674641148325359
188,18.15,3.5,Female,Yes,Sun,Dinner,3,0.1928374655647383
191,19.81,4.19,Female,Yes,Thur,Lunch,2,0.21150933871781932
197,43.11,5.0,Female,Yes,Thur,Lunch,4,0.1159823706796567
198,13.0,2.0,Female,Yes,Thur,Lunch,2,0.15384615384615385
201,12.74,2.01,Female,Yes,Thur,Lunch,2,0.15777080062794346
202,13.0,2.0,Female,Yes,Thur,Lunch,2,0.15384615384615385
203,16.4,2.5,Female,Yes,Thur,Lunch,2,0.15243902439024393
205,16.47,3.23,Female,Yes,Thur,Lunch,3,0.19611414693381907
209,12.76,2.23,Female,Yes,Sat,Dinner,2,0.17476489028213166
213,13.27,2.5,Female,Yes,Sat,Dinner,2,0.18839487565938207
214,28.17,6.5,Female,Yes,Sat,Dinner,3,0.23074192403265883
215,12.9,1.1,Female,Yes,Sat,Dinner,2,0.08527131782945736
219,30.14,3.09,Female,Yes,Sat,Dinner,4,0.10252156602521566
221,13.42,3.48,Female,Yes,Fri,Lunch,2,0.2593144560357675
223,15.98,3.0,Female,No,Fri,Lunch,3,0.18773466833541927
225,16.27,2.5,Female,Yes,Fri,Lunch,2,0.15365703749231716
226,10.09,2.0,Female,Yes,Fri,Lunch,2,0.19821605550049554
229,22.12,2.88,Female,Yes,Sat,Dinner,2,0.13019891500904157
238,35.83,4.67,Female,No,Sat,Dinner,3,0.13033770583310075
240,27.18,2.0,Female,Yes,Sat,Dinner,2,0.07358351729212656
243,18.78,3.0,Female,No,Thur,Dinner,2,0.1597444089456869
```

```
In [72]: data['perc_tip'].mean()

Out[72]: 0.16080258172250478

In [73]: # Slow version:
         #the_sum = 0
         #for row in data:
         #    the_sum += row['perc_tip']
         #
         #the_mean = the_sum / len(data)

In [74]: data.groupby('size')['perc_tip'].mean()

Out[74]: size
         1    0.217292
```

```
2    0.165719
3    0.152157
4    0.145949
5    0.141495
6    0.156229
Name: perc_tip, dtype: float64
```

```
In [75]: data.groupby(['size', 'sex'])['perc_tip'].mean()
```

```
Out[75]: size  sex
1     Female    0.215131
      Male      0.223776
2     Female    0.170830
      Male      0.162694
3     Female    0.159899
      Male      0.147641
4     Female    0.132734
      Male      0.150197
5     Female    0.172194
      Male      0.133821
6     Female    0.162891
      Male      0.149567
Name: perc_tip, dtype: float64
```

```
In [76]: means = data.groupby(['size', 'sex'])['perc_tip'].mean()
```

```
In [77]: means.unstack('sex')
```

```
Out[77]: sex     Female      Male
size
1     0.215131  0.223776
2     0.170830  0.162694
3     0.159899  0.147641
4     0.132734  0.150197
5     0.172194  0.133821
6     0.162891  0.149567
```

```
In [78]: data.groupby(['size', 'sex'])['perc_tip'].mean().unstack().to_latex()
```

```
Out[78]: '\\begin{tabular}{lrr}\n\\toprule\nsex &    Female &      Male \\\\\nsize
```

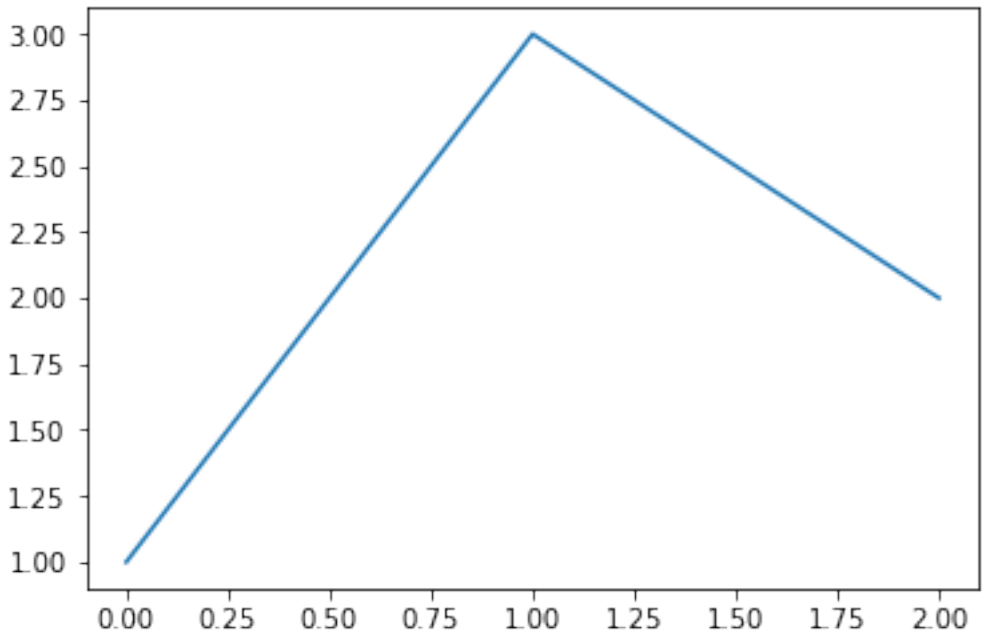
```
In [79]: # "readable" version:
         (data
          .groupby(['size', 'sex'])
          ['perc_tip']
          .mean()
          .unstack()
          .to_latex())
```

```
Out[79]: '\\begin{tabular}{lrr}\n\\toprule\nsex &    Female &      Male \\\\\nsize
```

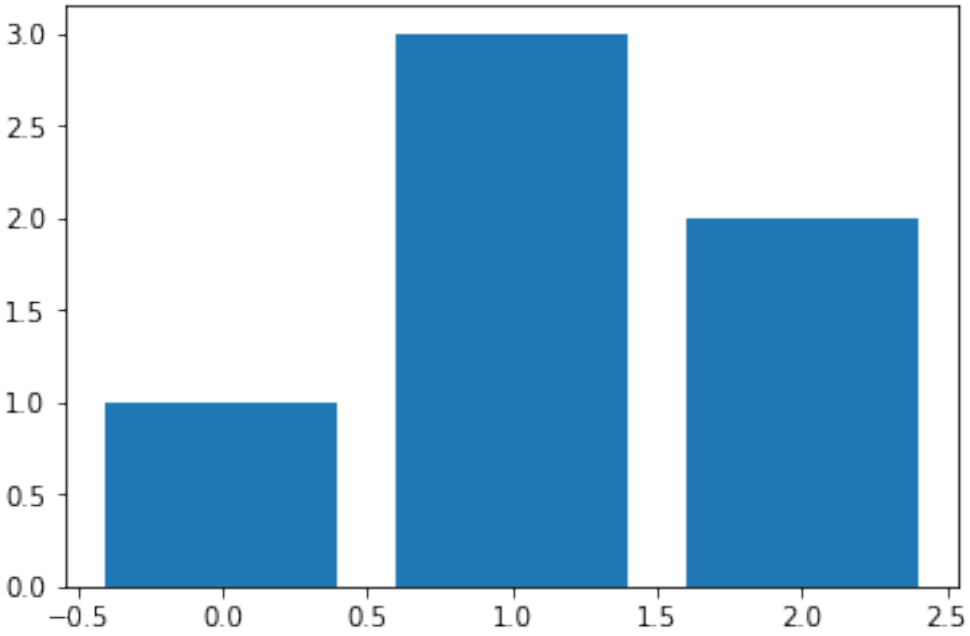
2 → Slide

3 matplotlib: visualize data

```
In [80]: from matplotlib import pyplot as plt
         %matplotlib inline

In [81]: plt.plot([1,3,2])

Out[81]: [<matplotlib.lines.Line2D at 0x7fa9fcc22160>]
```

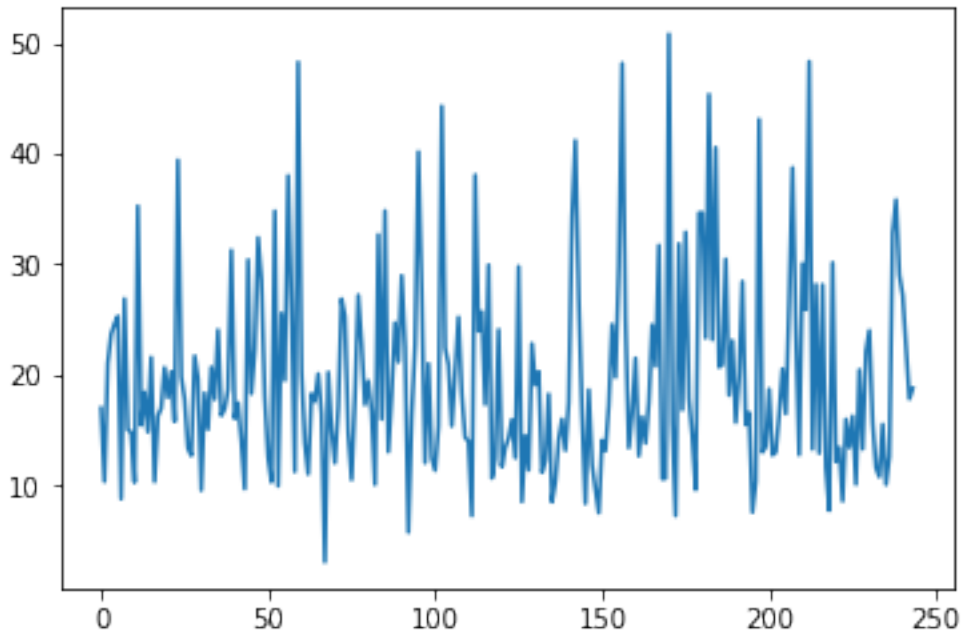



```
In [82]: plt.bar([0,1,2], [1,3,2])

Out[82]: <Container object of 3 artists>
```

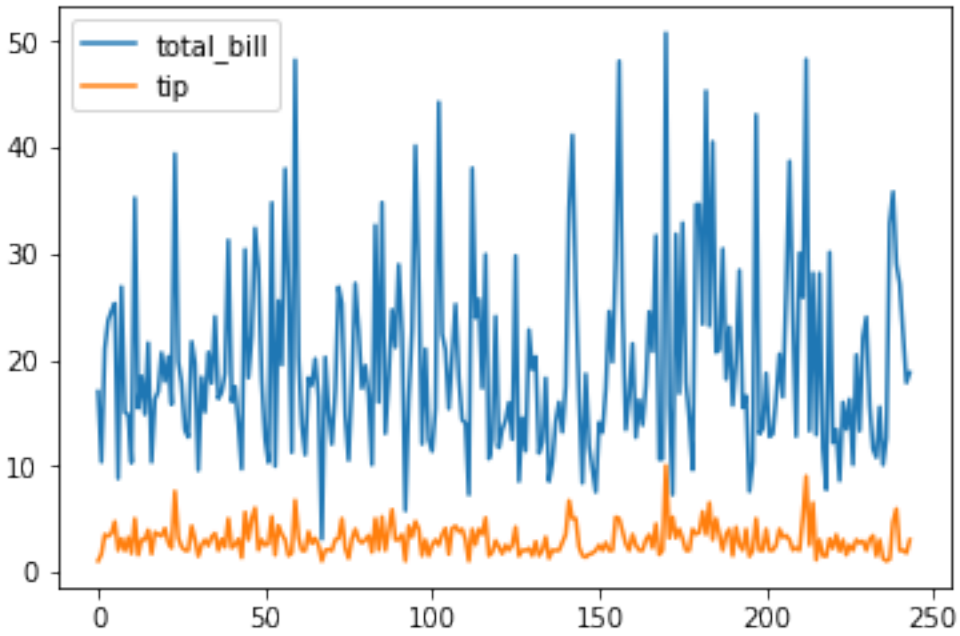



```
In [83]: data['total_bill'].plot()
```

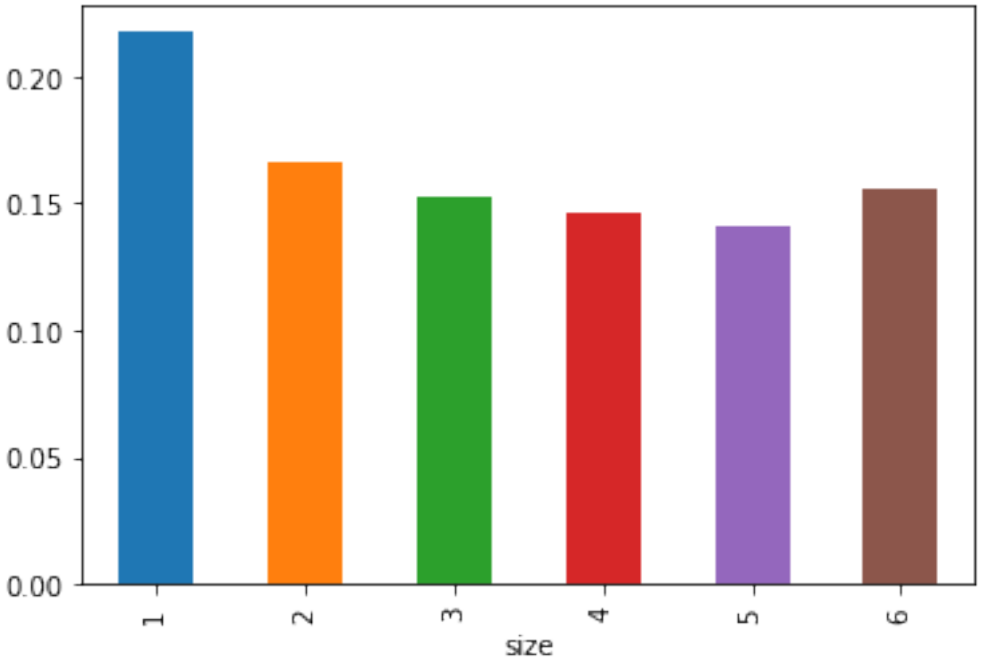
```
Out[83]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa9fcb29320>
```



```
In [84]: data[['total_bill', 'tip']].plot()
Out[84]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa9fcb26dd8>
```

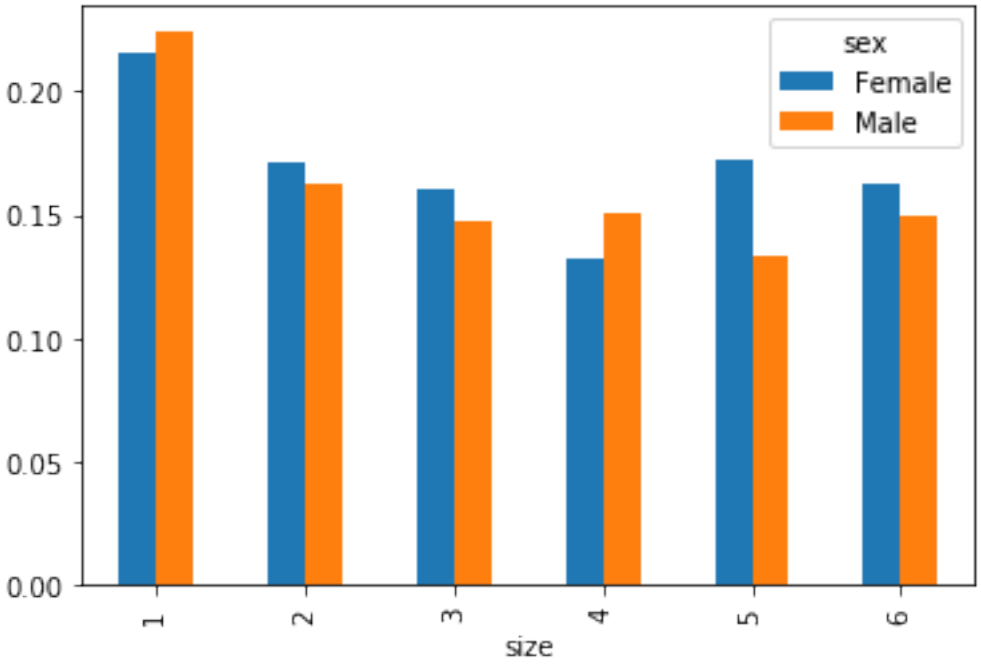



```
In [85]: data.groupby('size')['perc_tip'].mean().plot(kind='bar')
Out[85]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa9fcb1e128>
```



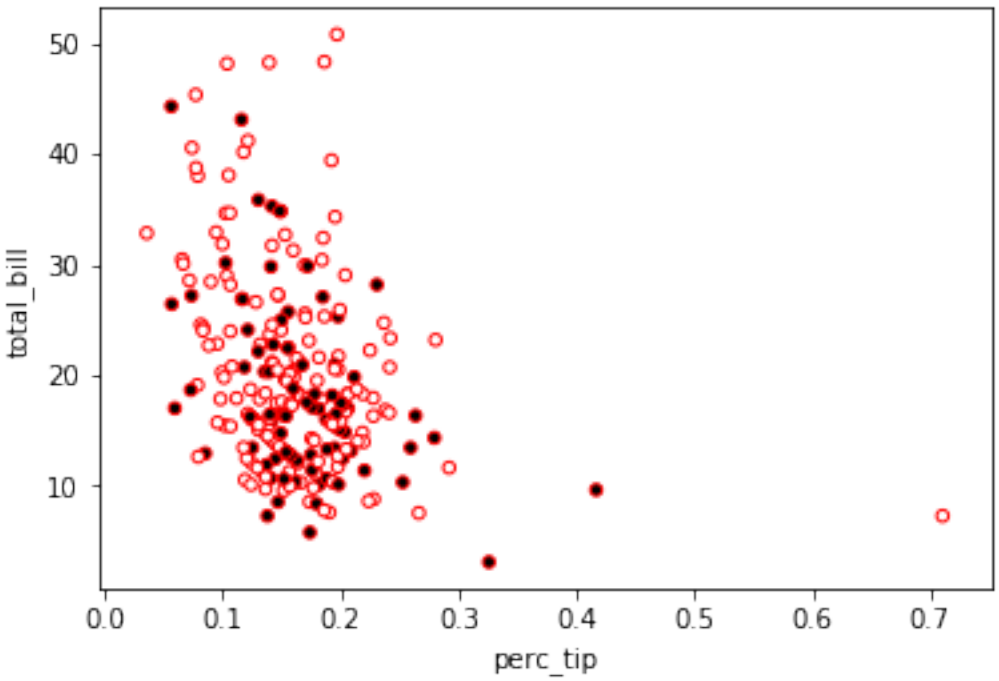
```
In [86]: data.groupby(['size', 'sex'])['perc_tip'].mean().unstack().plot(kind='bar'
```

```
Out[86]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa9fca84e48>
```




```
In [87]: female = (data['sex'] == 'Female')
         data.plot(kind='scatter', x='perc_tip', y='total_bill',
                   c=female, edgecolor='r'
                   )
```

```
Out[87]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa9fcb264e0>
```



4 → Slide

5 Statsmodels

```
In [88]: import statsmodels.api as sm

In [89]: res = sm.OLS.from_formula('tip ~ total_bill + sex + day + size', data=data

In [90]: res.summary()

Out[90]: <class 'statsmodels.iolib.summary.Summary'>
         """
                                     OLS Regression Results
         ==============================================================================
         Dep. Variable:                    tip   R-squared:                       0
         Model:                            OLS   Adj. R-squared:                  0
         Method:                 Least Squares   F-statistic:                     3
         Date:                Sat, 05 May 2018   Prob (F-statistic):           4.04
         Time:                        10:36:20   Log-Likelihood:                -34
         No. Observations:                 244   AIC:                             7
         Df Residuals:                     237   BIC:                             7
         Df Model:                           6
         Covariance Type:            nonrobust
         ==============================================================================
```

```
                   coef    std err          t      P>|t|      [0.025      0
------------------------------------------------------------------------------
Intercept       0.7601      0.289      2.633      0.009       0.191
sex[T.Male]    -0.0326      0.141     -0.231      0.818      -0.311
day[T.Sat]     -0.1202      0.261     -0.461      0.645      -0.634
day[T.Sun]     -0.0064      0.269     -0.024      0.981      -0.536
day[T.Thur]    -0.0789      0.269     -0.293      0.770      -0.609
total_bill      0.0932      0.009     10.019      0.000       0.075
size            0.1867      0.087      2.137      0.034       0.015
==============================================================================
Omnibus:                       26.163   Durbin-Watson:                   2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49
Skew:                           0.571   Prob(JB):                     1.87
Kurtosis:                       4.886   Cond. No.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
"""
```

In [91]:
```
data['day'].unique()
```

Out[91]:
```
array(['Sun', 'Sat', 'Thur', 'Fri'], dtype=object)
```

6 → Slide

7 scikit-learn

In [92]:
```
from sklearn.neural_network import MLPClassifier
```

In [93]:
```
clf = MLPClassifier()
```

In [94]:
```
data.head()
```

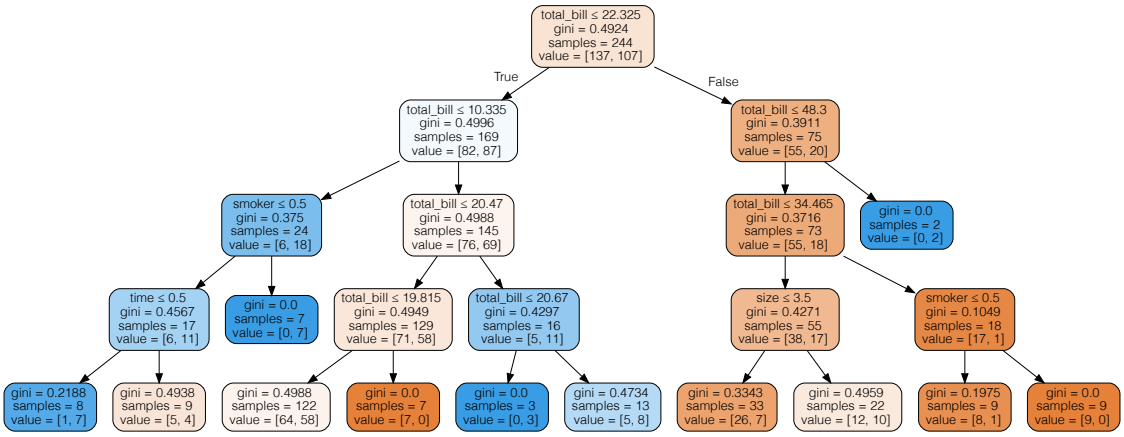
Out[94]:

	total_bill	tip	sex	smoker	day	time	size	perc_tip
0	16.99	1.01	Female	No	Sun	Dinner	2	0.059447
1	10.34	1.66	Male	No	Sun	Dinner	3	0.160542
2	21.01	3.50	Male	No	Sun	Dinner	3	0.166587
3	23.68	3.31	Male	No	Sun	Dinner	2	0.139780
4	24.59	3.61	Female	No	Sun	Dinner	4	0.146808

In [95]:
```
data['sex'] = data['sex'] == 'Female'
data['smoker'] = data['smoker'] == 'Yes'
data['time'] = data['time'] == 'Dinner'
```

In [96]:
```
data.head()
```

```
Out[96]:    total_bill   tip    sex  smoker  day  time  size  perc_tip
          0       16.99  1.01   True   False  Sun  True     2  0.059447
          1       10.34  1.66  False   False  Sun  True     3  0.160542
          2       21.01  3.50  False   False  Sun  True     3  0.166587
          3       23.68  3.31  False   False  Sun  True     2  0.139780
          4       24.59  3.61   True   False  Sun  True     4  0.146808

In [97]: data['good_tip'] = data['perc_tip'] > data['perc_tip'].mean()

In [98]: x = data.drop(['good_tip', 'day', 'perc_tip', 'tip'], axis=1)
         y = data['good_tip']

In [99]: data.head()

Out[99]:    total_bill   tip    sex  smoker  day  time  size  perc_tip  good_tip
          0       16.99  1.01   True   False  Sun  True     2  0.059447     False
          1       10.34  1.66  False   False  Sun  True     3  0.160542     False
          2       21.01  3.50  False   False  Sun  True     3  0.166587      True
          3       23.68  3.31  False   False  Sun  True     2  0.139780     False
          4       24.59  3.61   True   False  Sun  True     4  0.146808     False

In [100]: res = clf.fit(x, y)

In [101]: # I'M CHEATING! I'M CHEATING!
          res.score(x, y)

Out[101]: 0.5860655737704918

In [102]: from sklearn.tree import DecisionTreeClassifier, export_graphviz

In [103]: tree = DecisionTreeClassifier(max_depth=4)

In [104]: res = tree.fit(x, y)

In [105]: res.score(x, y)

Out[105]: 0.6475409836065574

In [106]: dot_data = export_graphviz(tree, out_file=None,
                                     feature_names=x.columns,
                                     filled=True, rounded=True,
                                     special_characters=True)

In [107]: import graphviz
          graph = graphviz.Source(dot_data)

In [108]: graph

Out[108]:
```




```
In [109]: from sklearn.ensemble import RandomForestClassifier

In [110]: forest = RandomForestClassifier(max_depth=4)

In [111]: res = forest.fit(x, y)

In [112]: res.score(x, y)

Out[112]: 0.6680327868852459
```