2.1 - DataFrame

April 11, 2017

```
In [1]: import pandas as pd

In [2]: # Modo canonico - voi potete non farlo!
        df = pd.DataFrame([[1,2], [3,4]],
                          index=['a', 'b'],
                          columns=['A', 'B'])

In [3]: df
```

```
   A  B
a  1  2
b  3  4
```

```
In [4]: df['A']
```

```
a    1
b    3
Name: A, dtype: int64
```

```
In [5]: df.loc['a', 'A']

Out[5]: 1

In [6]: # /usr/lib/python3/dist-packages/pandas/tests/data/tips.csv
        df = pd.read_csv("tips.csv")

In [7]: # Un'altra cosa fondamentale che distingue pandas da numpy
        df.dtypes
```

```
total_bill    float64
tip           float64
sex            object
smoker         object
day            object
```

```
time           object
size            int64
dtype: object
```

```
In [8]: # Attenzione: che cos'è? Una Series! E che tipo ha? object!
        df.loc[1]
```

```
total_bill     10.34
tip             1.66
sex             Male
smoker            No
day              Sun
time          Dinner
size               3
Name: 1, dtype: object
```

```
In [9]: df.iloc[1]
```

```
total_bill     10.34
tip             1.66
sex             Male
smoker            No
day              Sun
time          Dinner
size               3
Name: 1, dtype: object
```

```
In [10]: df['sex']
```

```
0      Female
1        Male
2        Male
3        Male
4      Female
5        Male
6        Male
7        Male
8        Male
9        Male
10       Male
11     Female
12       Male
13       Male
```

```
14     Female
15       Male
16     Female
17       Male
18     Female
19       Male
20       Male
21     Female
22     Female
23       Male
24       Male
25       Male
26       Male
27       Male
28       Male
29     Female
        ...
214    Female
215    Female
216      Male
217      Male
218      Male
219    Female
220      Male
221    Female
222      Male
223    Female
224      Male
225    Female
226    Female
227      Male
228      Male
229    Female
230      Male
231      Male
232      Male
233      Male
234      Male
235      Male
236      Male
237      Male
238    Female
239      Male
240    Female
241      Male
242      Male
243    Female
Name: sex, dtype: object
```

```
In [11]: df['sex'].dtype

Out[11]: dtype('O')

In [12]: df.sex is df['sex']

Out[12]: True

In [13]: df.loc[df.sex == 'Female']
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
11	35.26	5.00	Female	No	Sun	Dinner	4
14	14.83	3.02	Female	No	Sun	Dinner	2
16	10.33	1.67	Female	No	Sun	Dinner	3
18	16.97	3.50	Female	No	Sun	Dinner	3
21	20.29	2.75	Female	No	Sat	Dinner	2
22	15.77	2.23	Female	No	Sat	Dinner	2
29	19.65	3.00	Female	No	Sat	Dinner	2
32	15.06	3.00	Female	No	Sat	Dinner	2
33	20.69	2.45	Female	No	Sat	Dinner	4
37	16.93	3.07	Female	No	Sat	Dinner	3
51	10.29	2.60	Female	No	Sun	Dinner	2
52	34.81	5.20	Female	No	Sun	Dinner	4
57	26.41	1.50	Female	No	Sat	Dinner	2
66	16.45	2.47	Female	No	Sat	Dinner	2
67	3.07	1.00	Female	Yes	Sat	Dinner	1
71	17.07	3.00	Female	No	Sat	Dinner	3
72	26.86	3.14	Female	Yes	Sat	Dinner	2
73	25.28	5.00	Female	Yes	Sat	Dinner	2
74	14.73	2.20	Female	No	Sat	Dinner	2
82	10.07	1.83	Female	No	Thur	Lunch	1
85	34.83	5.17	Female	No	Thur	Lunch	4
92	5.75	1.00	Female	Yes	Fri	Dinner	2
93	16.32	4.30	Female	Yes	Fri	Dinner	2
94	22.75	3.25	Female	No	Fri	Dinner	2
100	11.35	2.50	Female	Yes	Fri	Dinner	2
101	15.38	3.00	Female	Yes	Fri	Dinner	2
102	44.30	2.50	Female	Yes	Sat	Dinner	3
103	22.42	3.48	Female	Yes	Sat	Dinner	2
..	...	...	...	...	...	...	...
155	29.85	5.14	Female	No	Sun	Dinner	5
157	25.00	3.75	Female	No	Sun	Dinner	4
158	13.39	2.61	Female	No	Sun	Dinner	2
162	16.21	2.00	Female	No	Sun	Dinner	3

```
164        17.51  3.00  Female    Yes   Sun  Dinner     2
168        10.59  1.61  Female    Yes   Sat  Dinner     2
169        10.63  2.00  Female    Yes   Sat  Dinner     2
178         9.60  4.00  Female    Yes   Sun  Dinner     2
186        20.90  3.50  Female    Yes   Sun  Dinner     3
188        18.15  3.50  Female    Yes   Sun  Dinner     3
191        19.81  4.19  Female    Yes  Thur   Lunch     2
197        43.11  5.00  Female    Yes  Thur   Lunch     4
198        13.00  2.00  Female    Yes  Thur   Lunch     2
201        12.74  2.01  Female    Yes  Thur   Lunch     2
202        13.00  2.00  Female    Yes  Thur   Lunch     2
203        16.40  2.50  Female    Yes  Thur   Lunch     2
205        16.47  3.23  Female    Yes  Thur   Lunch     3
209        12.76  2.23  Female    Yes   Sat  Dinner     2
213        13.27  2.50  Female    Yes   Sat  Dinner     2
214        28.17  6.50  Female    Yes   Sat  Dinner     3
215        12.90  1.10  Female    Yes   Sat  Dinner     2
219        30.14  3.09  Female    Yes   Sat  Dinner     4
221        13.42  3.48  Female    Yes   Fri   Lunch     2
223        15.98  3.00  Female     No   Fri   Lunch     3
225        16.27  2.50  Female    Yes   Fri   Lunch     2
226        10.09  2.00  Female    Yes   Fri   Lunch     2
229        22.12  2.88  Female    Yes   Sat  Dinner     2
238        35.83  4.67  Female     No   Sat  Dinner     3
240        27.18  2.00  Female    Yes   Sat  Dinner     2
243        18.78  3.00  Female     No  Thur  Dinner     2

[87 rows x 7 columns]
```

```
In [14]: df.tail()
```

```
     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 [15]: df.loc[243, 'smoker'] = 'Yes'

In [16]: df.tail()
```

```
     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    Yes  Thur  Dinner     2
```

```
In [17]: df.loc[242, ['sex', 'size']] = ['Female', 4]
```

```
In [18]: df.tail()
```

```
     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  Female     No   Sat  Dinner     4
243        18.78  3.00  Female    Yes  Thur  Dinner     2
```

```
In [19]: df.loc[244] = df.loc[243]
```

```
In [20]: df.tail()
```

```
     total_bill   tip     sex smoker   day    time  size
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  Female     No   Sat  Dinner     4
243        18.78  3.00  Female    Yes  Thur  Dinner     2
244        18.78  3.00  Female    Yes  Thur  Dinner     2
```

```
In [21]: df2 = df.copy()
```

```
In [22]: df2.loc[245, 'tip'] = 1.2
```

```
In [23]: df2.tail()
```

```
     total_bill   tip     sex smoker   day    time  size
241        22.67  2.00    Male    Yes   Sat  Dinner   2.0
242        17.82  1.75  Female     No   Sat  Dinner   4.0
243        18.78  3.00  Female    Yes  Thur  Dinner   2.0
244        18.78  3.00  Female    Yes  Thur  Dinner   2.0
245          NaN  1.20     NaN    NaN   NaN     NaN   NaN
```

```
In [24]: df2.dtypes
```

```
total_bill    float64
tip           float64
sex            object
smoker         object
day            object
time           object
size          float64
dtype: object
```

```
In [25]: df2.loc[245, 'size'] = 3

In [26]: df2['size'] = df2['size'].astype(int)

In [27]: df2.tail()
```

	total_bill	tip	sex	smoker	day	time	size
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Female	No	Sat	Dinner	4
243	18.78	3.00	Female	Yes	Thur	Dinner	2
244	18.78	3.00	Female	Yes	Thur	Dinner	2
245	NaN	1.20	NaN	NaN	NaN	NaN	3

```
In [28]: df2.loc[244, 'pioggia'] = True

In [29]: df2.loc[244, 'anni'] = 31

In [30]: df2 = df.copy()

In [31]: df2['anni'] = 'x'

In [32]: df2.loc[243, 'anni'] = 31

In [33]: df2.tail()
```

	total_bill	tip	sex	smoker	day	time	size	anni
240	27.18	2.00	Female	Yes	Sat	Dinner	2	x
241	22.67	2.00	Male	Yes	Sat	Dinner	2	x
242	17.82	1.75	Female	No	Sat	Dinner	4	x
243	18.78	3.00	Female	Yes	Thur	Dinner	2	31
244	18.78	3.00	Female	Yes	Thur	Dinner	2	x

```
In [34]: # ?
         df2.dtypes
```

```
total_bill    float64
tip           float64
sex            object
smoker         object
day            object
time           object
size            int64
anni           object
dtype: object
```

```
In [35]: df['perc'] = df['tip'] / df['total_bill']
```

```
In [36]: df
```

	total_bill	tip	sex	smoker	day	time	size	perc
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
5	25.29	4.71	Male	No	Sun	Dinner	4	0.186240
6	8.77	2.00	Male	No	Sun	Dinner	2	0.228050
7	26.88	3.12	Male	No	Sun	Dinner	4	0.116071
8	15.04	1.96	Male	No	Sun	Dinner	2	0.130319
9	14.78	3.23	Male	No	Sun	Dinner	2	0.218539
10	10.27	1.71	Male	No	Sun	Dinner	2	0.166504
11	35.26	5.00	Female	No	Sun	Dinner	4	0.141804
12	15.42	1.57	Male	No	Sun	Dinner	2	0.101816
13	18.43	3.00	Male	No	Sun	Dinner	4	0.162778
14	14.83	3.02	Female	No	Sun	Dinner	2	0.203641
15	21.58	3.92	Male	No	Sun	Dinner	2	0.181650
16	10.33	1.67	Female	No	Sun	Dinner	3	0.161665
17	16.29	3.71	Male	No	Sun	Dinner	3	0.227747
18	16.97	3.50	Female	No	Sun	Dinner	3	0.206246
19	20.65	3.35	Male	No	Sat	Dinner	3	0.162228
20	17.92	4.08	Male	No	Sat	Dinner	2	0.227679
21	20.29	2.75	Female	No	Sat	Dinner	2	0.135535
22	15.77	2.23	Female	No	Sat	Dinner	2	0.141408
23	39.42	7.58	Male	No	Sat	Dinner	4	0.192288
24	19.82	3.18	Male	No	Sat	Dinner	2	0.160444
25	17.81	2.34	Male	No	Sat	Dinner	4	0.131387
26	13.37	2.00	Male	No	Sat	Dinner	2	0.149589
27	12.69	2.00	Male	No	Sat	Dinner	2	0.157604
28	21.70	4.30	Male	No	Sat	Dinner	2	0.198157
29	19.65	3.00	Female	No	Sat	Dinner	2	0.152672

```
..          ...   ...     ...    ...   ...     ...   ...       ...
215       12.90  1.10  Female    Yes   Sat  Dinner     2  0.085271
216       28.15  3.00    Male    Yes   Sat  Dinner     5  0.106572
217       11.59  1.50    Male    Yes   Sat  Dinner     2  0.129422
218        7.74  1.44    Male    Yes   Sat  Dinner     2  0.186047
219       30.14  3.09  Female    Yes   Sat  Dinner     4  0.102522
220       12.16  2.20    Male    Yes   Fri   Lunch     2  0.180921
221       13.42  3.48  Female    Yes   Fri   Lunch     2  0.259314
222        8.58  1.92    Male    Yes   Fri   Lunch     1  0.223776
223       15.98  3.00  Female     No   Fri   Lunch     3  0.187735
224       13.42  1.58    Male    Yes   Fri   Lunch     2  0.117735
225       16.27  2.50  Female    Yes   Fri   Lunch     2  0.153657
226       10.09  2.00  Female    Yes   Fri   Lunch     2  0.198216
227       20.45  3.00    Male     No   Sat  Dinner     4  0.146699
228       13.28  2.72    Male     No   Sat  Dinner     2  0.204819
229       22.12  2.88  Female    Yes   Sat  Dinner     2  0.130199
230       24.01  2.00    Male    Yes   Sat  Dinner     4  0.083299
231       15.69  3.00    Male    Yes   Sat  Dinner     3  0.191205
232       11.61  3.39    Male     No   Sat  Dinner     2  0.291990
233       10.77  1.47    Male     No   Sat  Dinner     2  0.136490
234       15.53  3.00    Male    Yes   Sat  Dinner     2  0.193175
235       10.07  1.25    Male     No   Sat  Dinner     2  0.124131
236       12.60  1.00    Male    Yes   Sat  Dinner     2  0.079365
237       32.83  1.17    Male    Yes   Sat  Dinner     2  0.035638
238       35.83  4.67  Female     No   Sat  Dinner     3  0.130338
239       29.03  5.92    Male     No   Sat  Dinner     3  0.203927
240       27.18  2.00  Female    Yes   Sat  Dinner     2  0.073584
241       22.67  2.00    Male    Yes   Sat  Dinner     2  0.088222
242       17.82  1.75  Female     No   Sat  Dinner     4  0.098204
243       18.78  3.00  Female    Yes  Thur  Dinner     2  0.159744
244       18.78  3.00  Female    Yes  Thur  Dinner     2  0.159744

[245 rows x 8 columns]
```

```
In [37]: # .columns e .index sono esattamente lo stesso tipo di oggetto!
         df.columns
```

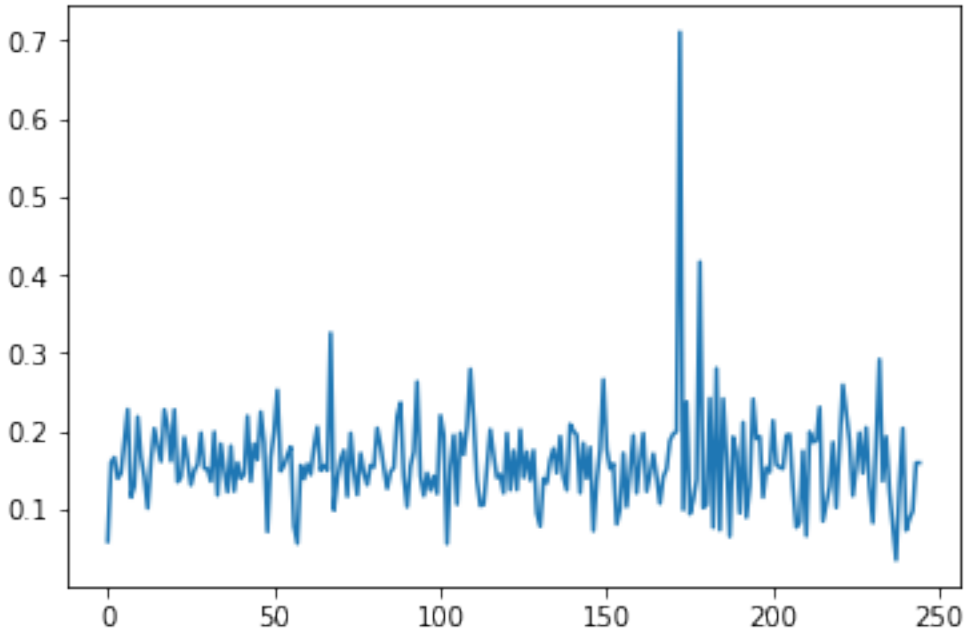
```
Out[37]: Index(['total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'size', 'perc'
```

0.1 Plot

```
In [38]: import matplotlib.pyplot as plt
         %matplotlib inline
```

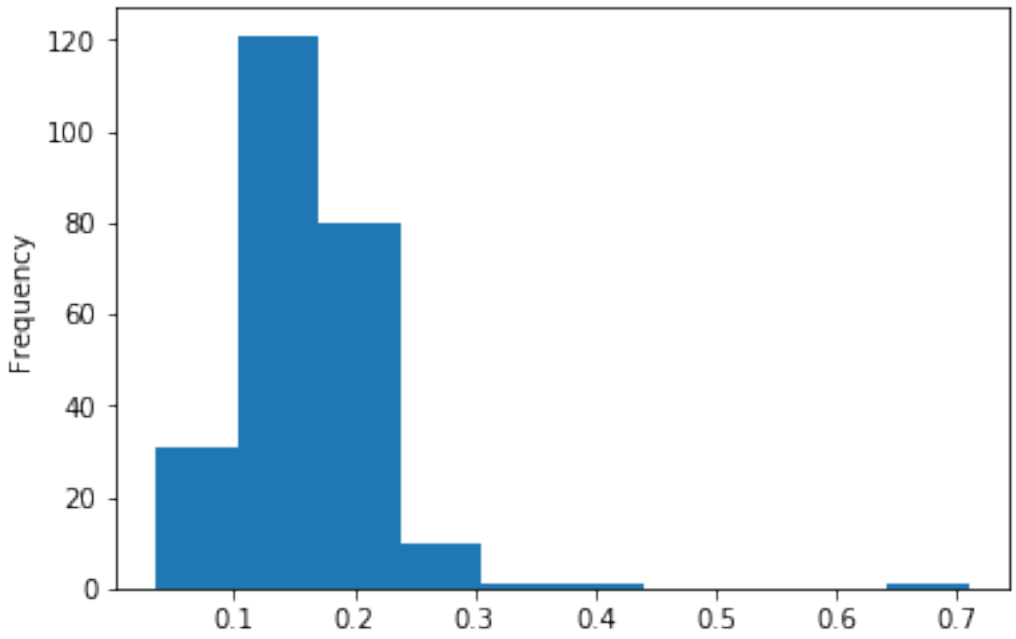
```
In [39]: # Schifo - non è quello che cerchiamo
         plt.plot(df['perc'])
```

```
Out[39]: [<matplotlib.lines.Line2D at 0x7fac5e3632b0>]
```



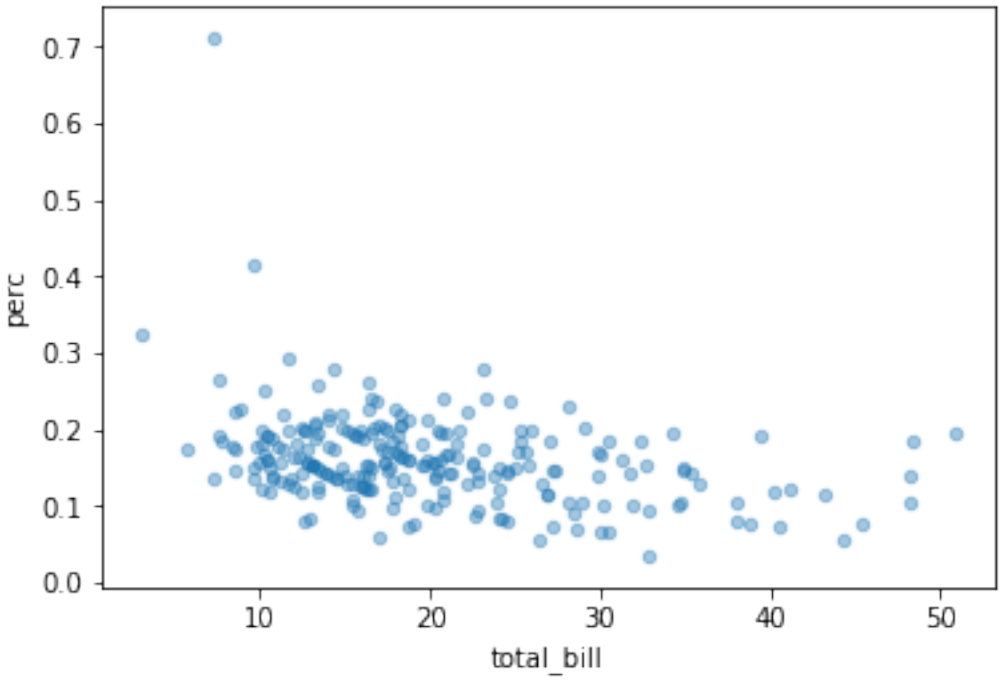

```
In [40]: df['perc'].plot(kind='hist')
```

```
Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x7fac5e363588>
```



```
In [41]: df.plot(x='total_bill', y='perc', kind='scatter', alpha=.4)
```

```
Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x7fac5aa95780>
```



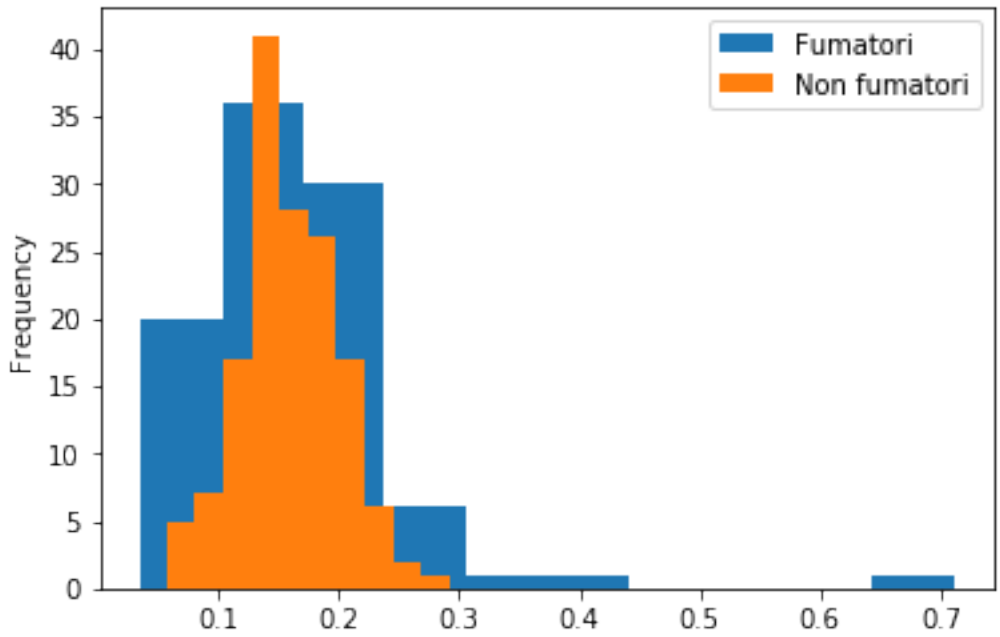

```
In [42]: df.loc[df['smoker'] == 'No', 'smoker'] = False
         df.loc[df['smoker'] == 'Yes', 'smoker'] = True
         df['smoker'] = df['smoker'].astype(bool)
```

```
In [43]: df.dtypes
```

```
total_bill    float64
tip           float64
sex            object
smoker           bool
day            object
time           object
size            int64
perc          float64
dtype: object
```

```
In [44]: df.loc[df.smoker, 'perc'].plot(kind='hist', label='Fumatori')
         df.loc[~df.smoker, 'perc'].plot(kind='hist', label='Non fumatori')
         plt.legend()
```

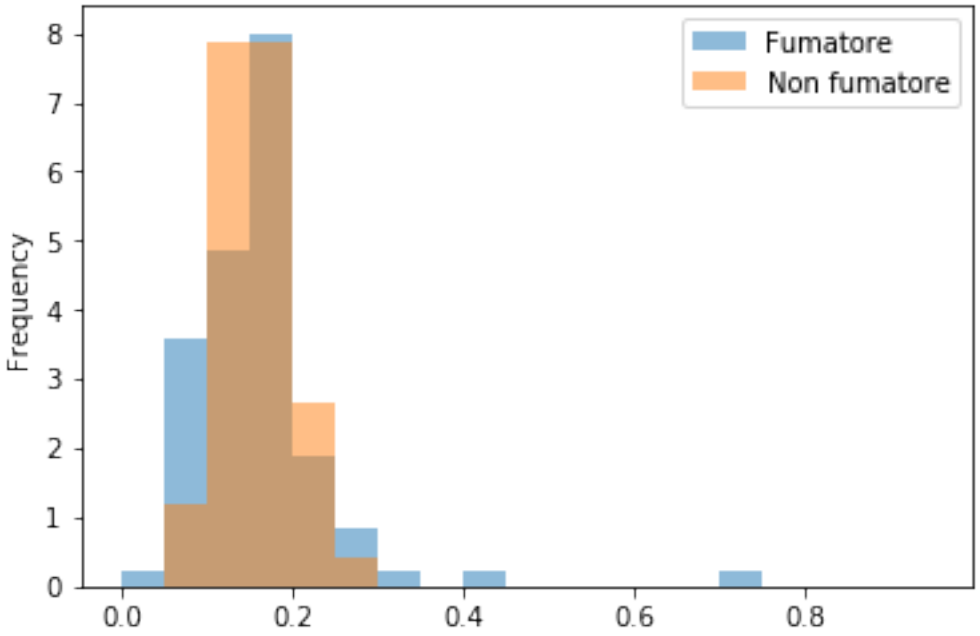
Out[44]: <matplotlib.legend.Legend at 0x7fac5a9be550>



In [45]:
```
# Versione ripulita
import numpy as np

bins = np.arange(0, 1, .05)
df.loc[df.smoker, 'perc'].plot(kind='hist', label='Fumatore', bins=bins, a
                                    normed=True)
df.loc[~df.smoker, 'perc'].plot(kind='hist', label='Non fumatore', bins=bi
                                     normed=True)
plt.legend()
```

Out[45]: <matplotlib.legend.Legend at 0x7fac5a9be240>



In [46]: *# Menzionare seaborn*