Lec 44 - Aggregation

April 28, 2015

In [1]:
```
import numpy as np
import pandas as pd
from pandas import Series,DataFrame
```

In [6]:
```
# Data Agrregation consists of operations that result in a scalar (e.g. mean(),sum(),count(), e

#Let's get a csv data set to play with
url = 'http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/'


# Save thewinquality.csv file in the same folder as your ipython notebooks, note the delimiter
dframe_wine = pd.read_csv('winequality_red.csv',sep=';')
```

In [7]:
```
# Let's get a preview
dframe_wine.head()
```

Out[7]:
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.4              0.70         0.00             1.9      0.076
1            7.8              0.88         0.00             2.6      0.098
2            7.8              0.76         0.04             2.3      0.092
3           11.2              0.28         0.56             1.9      0.075
4            7.4              0.70         0.00             1.9      0.076

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                   11                    34   0.9978  3.51       0.56
1                   25                    67   0.9968  3.20       0.68
2                   15                    54   0.9970  3.26       0.65
3                   17                    60   0.9980  3.16       0.58
4                   11                    34   0.9978  3.51       0.56

   alcohol  quality
0      9.4        5
1      9.8        5
2      9.8        5
3      9.8        6
4      9.4        5
```

In [8]:
```
# How about we find out the average alcohol content for the wine
dframe_wine['alcohol'].mean()
```

Out[8]: 10.422983114446529

In [25]:
```
# That was an example of an aggregate, how about we make our own?
def max_to_min(arr):
    return arr.max() - arr.min()
```

```python
# Let's group the wines by "quality"
wino = dframe_wine.groupby('quality')

# Show
wino.describe()
```

Out[25]:

quality		alcohol	chlorides	citric acid	density	fixed acidity \
3	count	10.000000	10.000000	10.000000	10.000000	10.000000
	mean	9.955000	0.122500	0.171000	0.997464	8.360000
	std	0.818009	0.066241	0.250664	0.002002	1.770875
	min	8.400000	0.061000	0.000000	0.994710	6.700000
	25%	9.725000	0.079000	0.005000	0.996150	7.150000
	50%	9.925000	0.090500	0.035000	0.997565	7.500000
	75%	10.575000	0.143000	0.327500	0.998770	9.875000
	max	11.000000	0.267000	0.660000	1.000800	11.600000
4	count	53.000000	53.000000	53.000000	53.000000	53.000000
	mean	10.265094	0.090679	0.174151	0.996542	7.779245
	std	0.934776	0.076192	0.201030	0.001575	1.626624
	min	9.000000	0.045000	0.000000	0.993400	4.600000
	25%	9.600000	0.067000	0.030000	0.995650	6.800000
	50%	10.000000	0.080000	0.090000	0.996500	7.500000
	75%	11.000000	0.089000	0.270000	0.997450	8.400000
	max	13.100000	0.610000	1.000000	1.001000	12.500000
5	count	681.000000	681.000000	681.000000	681.000000	681.000000
	mean	9.899706	0.092736	0.243686	0.997104	8.167254
	std	0.736521	0.053707	0.180003	0.001589	1.563988
	min	8.500000	0.039000	0.000000	0.992560	5.000000
	25%	9.400000	0.074000	0.090000	0.996200	7.100000
	50%	9.700000	0.081000	0.230000	0.997000	7.800000
	75%	10.200000	0.094000	0.360000	0.997900	8.900000
	max	14.900000	0.611000	0.790000	1.003150	15.900000
6	count	638.000000	638.000000	638.000000	638.000000	638.000000
	mean	10.629519	0.084956	0.273824	0.996615	8.347179
	std	1.049639	0.039563	0.195108	0.002000	1.797849
	min	8.400000	0.034000	0.000000	0.990070	4.700000
	25%	9.800000	0.068250	0.090000	0.995402	7.000000
	50%	10.500000	0.078000	0.260000	0.996560	7.900000
	75%	11.300000	0.088000	0.430000	0.997893	9.400000
	max	14.000000	0.415000	0.780000	1.003690	14.300000
7	count	199.000000	199.000000	199.000000	199.000000	199.000000
	mean	11.465913	0.076588	0.375176	0.996104	8.872362
	std	0.961933	0.029456	0.194432	0.002176	1.992483
	min	9.200000	0.012000	0.000000	0.990640	4.900000
	25%	10.800000	0.062000	0.305000	0.994765	7.400000
	50%	11.500000	0.073000	0.400000	0.995770	8.800000
	75%	12.100000	0.087000	0.490000	0.997360	10.100000
	max	14.000000	0.358000	0.760000	1.003200	15.600000
8	count	18.000000	18.000000	18.000000	18.000000	18.000000
	mean	12.094444	0.068444	0.391111	0.995212	8.566667
	std	1.224011	0.011678	0.199526	0.002378	2.119656
	min	9.800000	0.044000	0.030000	0.990800	5.000000
	25%	11.325000	0.062000	0.302500	0.994175	7.250000

50%	12.150000	0.070500	0.420000	0.994940	8.250000
75%	12.875000	0.075500	0.530000	0.997200	10.225000
max	14.000000	0.086000	0.720000	0.998800	12.600000

quality		free sulfur dioxide	pH	residual sugar	sulphates \
3	count	10.000000	10.000000	10.000000	10.000000
	mean	11.000000	3.398000	2.635000	0.570000
	std	9.763879	0.144052	1.401596	0.122020
	min	3.000000	3.160000	1.200000	0.400000
	25%	5.000000	3.312500	1.875000	0.512500
	50%	6.000000	3.390000	2.100000	0.545000
	75%	14.500000	3.495000	3.100000	0.615000
	max	34.000000	3.630000	5.700000	0.860000
4	count	53.000000	53.000000	53.000000	53.000000
	mean	12.264151	3.381509	2.694340	0.596415
	std	9.025926	0.181441	1.789436	0.239391
	min	3.000000	2.740000	1.300000	0.330000
	25%	6.000000	3.300000	1.900000	0.490000
	50%	11.000000	3.370000	2.100000	0.560000
	75%	15.000000	3.500000	2.800000	0.600000
	max	41.000000	3.900000	12.900000	2.000000
5	count	681.000000	681.000000	681.000000	681.000000
	mean	16.983847	3.304949	2.528855	0.620969
	std	10.955446	0.150618	1.359753	0.171062
	min	3.000000	2.880000	1.200000	0.370000
	25%	9.000000	3.200000	1.900000	0.530000
	50%	15.000000	3.300000	2.200000	0.580000
	75%	23.000000	3.400000	2.600000	0.660000
	max	68.000000	3.740000	15.500000	1.980000
6	count	638.000000	638.000000	638.000000	638.000000
	mean	15.711599	3.318072	2.477194	0.675329
	std	9.940911	0.153995	1.441576	0.158650
	min	1.000000	2.860000	0.900000	0.400000
	25%	8.000000	3.220000	1.900000	0.580000
	50%	14.000000	3.320000	2.200000	0.640000
	75%	21.000000	3.410000	2.500000	0.750000
	max	72.000000	4.010000	15.400000	1.950000
7	count	199.000000	199.000000	199.000000	199.000000
	mean	14.045226	3.290754	2.720603	0.741256
	std	10.175255	0.150101	1.371509	0.135639
	min	3.000000	2.920000	1.200000	0.390000
	25%	6.000000	3.200000	2.000000	0.650000
	50%	11.000000	3.280000	2.300000	0.740000
	75%	18.000000	3.380000	2.750000	0.830000
	max	54.000000	3.780000	8.900000	1.360000
8	count	18.000000	18.000000	18.000000	18.000000
	mean	13.277778	3.267222	2.577778	0.767778
	std	11.155613	0.200640	1.295038	0.115379
	min	3.000000	2.880000	1.400000	0.630000
	25%	6.000000	3.162500	1.800000	0.690000
	50%	7.500000	3.230000	2.100000	0.740000
	75%	16.500000	3.350000	2.600000	0.820000
	max	42.000000	3.720000	6.400000	1.100000

```
                total sulfur dioxide  volatile acidity
quality
3       count             10.000000         10.000000
        mean              24.900000          0.884500
        std               16.828877          0.331256
        min                9.000000          0.440000
        25%               12.500000          0.647500
        50%               15.000000          0.845000
        75%               42.500000          1.010000
        max               49.000000          1.580000
4       count             53.000000         53.000000
        mean              36.245283          0.693962
        std               27.583374          0.220110
        min                7.000000          0.230000
        25%               14.000000          0.530000
        50%               26.000000          0.670000
        75%               49.000000          0.870000
        max              119.000000          1.130000
5       count            681.000000        681.000000
        mean              56.513950          0.577041
        std               36.993116          0.164801
        min                6.000000          0.180000
        25%               26.000000          0.460000
        50%               47.000000          0.580000
        75%               84.000000          0.670000
        max              155.000000          1.330000
6       count            638.000000        638.000000
        mean              40.869906          0.497484
        std               25.038250          0.160962
        min                6.000000          0.160000
        25%               23.000000          0.380000
        50%               35.000000          0.490000
        75%               54.000000          0.600000
        max              165.000000          1.040000
7       count            199.000000        199.000000
        mean              35.020101          0.403920
        std               33.191206          0.145224
        min                7.000000          0.120000
        25%               17.500000          0.300000
        50%               27.000000          0.370000
        75%               43.000000          0.485000
        max              289.000000          0.915000
8       count             18.000000         18.000000
        mean              33.444444          0.423333
        std               25.433240          0.144914
        min               12.000000          0.260000
        25%               16.000000          0.335000
        50%               21.500000          0.370000
        75%               43.000000          0.472500
        max               88.000000          0.850000
```

In [22]:
```
# We can now apply our own aggregate function, this function takes the max value of the col an
wino.agg(max_to_min)
```

Out[22]:
```
         fixed acidity  volatile acidity  citric acid  residual sugar  \
quality
3                  4.9             1.140         0.66             4.5
4                  7.9             0.900         1.00            11.6
5                 10.9             1.150         0.79            14.3
6                  9.6             0.880         0.78            14.5
7                 10.7             0.795         0.76             7.7
8                  7.6             0.590         0.69             5.0

         chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
quality
3            0.206                   31                    40  0.00609  0.47
4            0.565                   38                   112  0.00760  1.16
5            0.572                   65                   149  0.01059  0.86
6            0.381                   71                   159  0.01362  1.15
7            0.346                   51                   282  0.01256  0.86
8            0.042                   39                    76  0.00800  0.84

         sulphates  alcohol
quality
3             0.46      2.6
4             1.67      4.1
5             1.61      6.4
6             1.55      5.6
7             0.97      4.8
8             0.47      4.2
```

In [26]:
```
# We can also pass string methods through aggregate
wino.agg('mean')
```

Out[26]:
```
         fixed acidity  volatile acidity  citric acid  residual sugar  \
quality
3             8.360000          0.884500     0.171000        2.635000
4             7.779245          0.693962     0.174151        2.694340
5             8.167254          0.577041     0.243686        2.528855
6             8.347179          0.497484     0.273824        2.477194
7             8.872362          0.403920     0.375176        2.720603
8             8.566667          0.423333     0.391111        2.577778

         chlorides  free sulfur dioxide  total sulfur dioxide   density  \
quality
3         0.122500            11.000000             24.900000  0.997464
4         0.090679            12.264151             36.245283  0.996542
5         0.092736            16.983847             56.513950  0.997104
6         0.084956            15.711599             40.869906  0.996615
7         0.076588            14.045226             35.020101  0.996104
8         0.068444            13.277778             33.444444  0.995212

               pH  sulphates    alcohol
quality
3        3.398000   0.570000   9.955000
4        3.381509   0.596415  10.265094
5        3.304949   0.620969   9.899706
6        3.318072   0.675329  10.629519
7        3.290754   0.741256  11.465913
```

```
8        3.267222   0.767778  12.094444
```

In [27]:
```python
# Let's go back to the original dframe
dframe_wine.head()
```

Out[27]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	7.4	0.70	0.00	1.9	0.076	
1	7.8	0.88	0.00	2.6	0.098	
2	7.8	0.76	0.04	2.3	0.092	
3	11.2	0.28	0.56	1.9	0.075	
4	7.4	0.70	0.00	1.9	0.076	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	11	34	0.9978	3.51	0.56	
1	25	67	0.9968	3.20	0.68	
2	15	54	0.9970	3.26	0.65	
3	17	60	0.9980	3.16	0.58	
4	11	34	0.9978	3.51	0.56	

	alcohol	quality
0	9.4	5
1	9.8	5
2	9.8	5
3	9.8	6
4	9.4	5

In [28]:
```python
# Let's adda  quality to alcohol content ratio
dframe_wine['qual/alc ratio'] = dframe_wine['quality']/dframe_wine['alcohol']
```

In [29]:
```python
# Show
dframe_wine.head()
```

Out[29]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	7.4	0.70	0.00	1.9	0.076	
1	7.8	0.88	0.00	2.6	0.098	
2	7.8	0.76	0.04	2.3	0.092	
3	11.2	0.28	0.56	1.9	0.075	
4	7.4	0.70	0.00	1.9	0.076	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	11	34	0.9978	3.51	0.56	
1	25	67	0.9968	3.20	0.68	
2	15	54	0.9970	3.26	0.65	
3	17	60	0.9980	3.16	0.58	
4	11	34	0.9978	3.51	0.56	

	alcohol	quality	qual/alc ratio
0	9.4	5	0.531915
1	9.8	5	0.510204
2	9.8	5	0.510204
3	9.8	6	0.612245
4	9.4	5	0.531915

In [32]:
```python
# WE can also use pivot tables instead of groupby

# Pivot table of quality
dframe_wine.pivot_table(index=['quality'])
```

Out[32]:

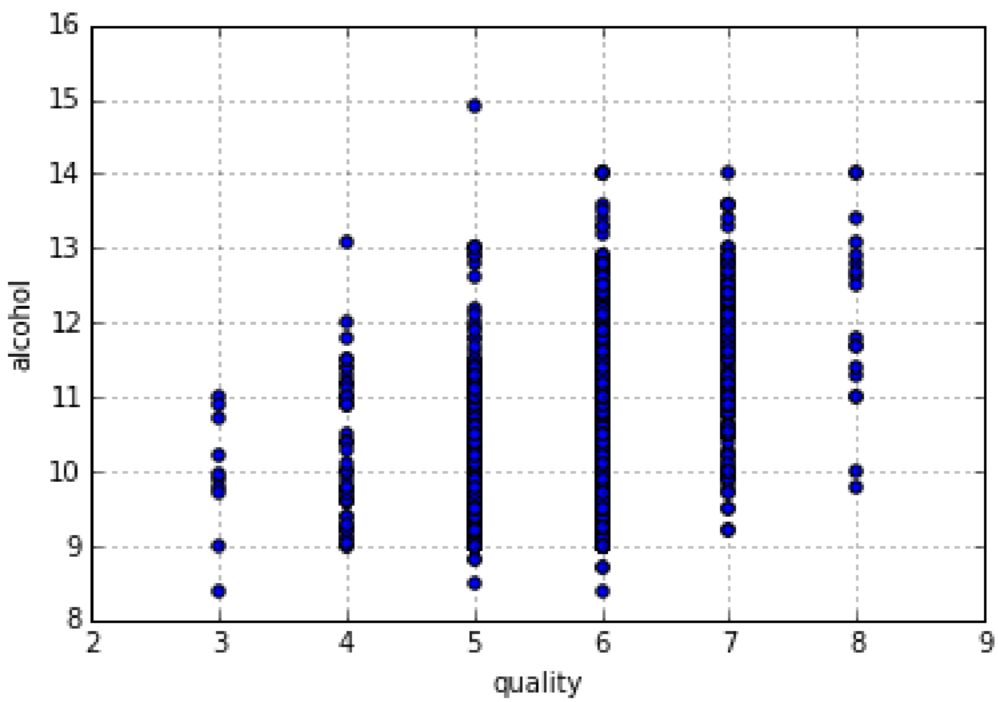
quality	alcohol	chlorides	citric acid	density	fixed acidity \
3	9.955000	0.122500	0.171000	0.997464	8.360000
4	10.265094	0.090679	0.174151	0.996542	7.779245
5	9.899706	0.092736	0.243686	0.997104	8.167254
6	10.629519	0.084956	0.273824	0.996615	8.347179
7	11.465913	0.076588	0.375176	0.996104	8.872362
8	12.094444	0.068444	0.391111	0.995212	8.566667

quality	free sulfur dioxide	pH	qual/alc ratio	residual sugar \
3	11.000000	3.398000	0.303286	2.635000
4	12.264151	3.381509	0.392724	2.694340
5	16.983847	3.304949	0.507573	2.528855
6	15.711599	3.318072	0.569801	2.477194
7	14.045226	3.290754	0.614855	2.720603
8	13.277778	3.267222	0.668146	2.577778

quality	sulphates	total sulfur dioxide	volatile acidity
3	0.570000	24.900000	0.884500
4	0.596415	36.245283	0.693962
5	0.620969	56.513950	0.577041
6	0.675329	40.869906	0.497484
7	0.741256	35.020101	0.403920
8	0.767778	33.444444	0.423333

In [38]:
```
%matplotlib inline
dframe_wine.plot(kind='scatter',x='quality',y='alcohol')
```

Out[38]: <matplotlib.axes._subplots.AxesSubplot at 0xecb6470>



We can see that the data is probably better fit for a box plot for a more concise view of the data See if you can figure how to get a boxplot using the pandas documentation and what you have learned so far

Don't worry if you can't quite figure it out just yet, the next section will cover all sorts of data visualizations!

```
In [ ]:
```