Lec 45 - Splitting, Applying and Combining

April 28, 2015

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

```
In [2]: # Let's grab the wine data again
        dframe_wine = pd.read_csv('winequality_red.csv',sep=';')

        #Preview
        dframe_wine.head()
```

```
Out[2]:    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
```

What if we wanted to know the highest alcohol content for each quality range?
We can use groupby mechanics to split-apply-combine

```
In [4]: # Create a function that assigns a rank to each wine based on alcohol content, with 1 being the
        def ranker(df):
            df['alc_content_rank'] = np.arange(len(df)) + 1
            return df
```

```
In [8]: # Now sort the dframe by alcohol in ascending order
        dframe_wine.sort('alcohol',ascending=False,inplace=True)

        # Now we'll group by quality and apply our ranking function
        dframe_wine = dframe_wine.groupby('quality').apply(ranker)
```

```
In [9]: #Preview
        dframe_wine.head()

Out[9]:       fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
        652            15.9              0.36         0.65             7.5      0.096
        588             5.0              0.42         0.24             2.0      0.060
        142             5.2              0.34         0.00             1.8      0.050
        144             5.2              0.34         0.00             1.8      0.050
        1270            5.0              0.38         0.01             1.6      0.048

              free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
        652                    22                    71  0.99760  2.98       0.84
        588                    19                    50  0.99170  3.72       0.74
        142                    27                    63  0.99160  3.68       0.79
        144                    27                    63  0.99160  3.68       0.79
        1270                   26                    60  0.99084  3.70       0.75

              alcohol  quality  alc_content_rank
        652      14.9        5                 1
        588      14.0        8                 1
        142      14.0        6                 1
        144      14.0        6                 2
        1270     14.0        6                 3
```

```
In [13]: # Now finally we can just call for the dframe where the alc_content_rank == 1

         # Get the numebr of quality counts
         num_of_qual = dframe_wine['quality'].value_counts()

         #Show
         num_of_qual

Out[13]: 5    681
         6    638
         7    199
         4     53
         8     18
         3     10
         dtype: int64
```

```
In [15]: # Now we'll show the combined info for teh wines that had the highest alcohol content for thei
         dframe_wine[dframe_wine.alc_content_rank == 1].head(len(num_of_qual))

Out[15]:      fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
         652           15.9              0.36         0.65             7.5      0.096
         588            5.0              0.42         0.24             2.0      0.060
         142            5.2              0.34         0.00             1.8      0.050
         821            4.9              0.42         0.00             2.1      0.048
         45             4.6              0.52         0.15             2.1      0.054
         899            8.3              1.02         0.02             3.4      0.084

              free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
         652                   22                    71  0.99760  2.98       0.84
         588                   19                    50  0.99170  3.72       0.74
         142                   27                    63  0.99160  3.68       0.79
```

```
821                  16                   42  0.99154  3.71       0.74
45                    8                   65  0.99340  3.90       0.56
899                   6                   11  0.99892  3.48       0.49

     alcohol  quality  alc_content_rank
652     14.9        5                 1
588     14.0        8                 1
142     14.0        6                 1
821     14.0        7                 1
45      13.1        4                 1
899     11.0        3                 1
```

```
In [ ]: # Awesome! Ask yourself if there are any trends you would like to find in this data?
        # Is there a relationship between wine ranking and alcohol content?
```