plotting_type_quality

October 24, 2017

1 Plotting Wine Type and Quality with Matplotlib

```
In [5]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        % matplotlib inline
        import seaborn as sns
        sns.set_style('darkgrid')

        wine_df = pd.read_csv('winequality_edited.csv')
        wine_df.head()
```

```
Out[5]:    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                 11.0                  34.0   0.9978  3.51       0.56
        1                 25.0                  67.0   0.9968  3.20       0.68
        2                 15.0                  54.0   0.9970  3.26       0.65
        3                 17.0                  60.0   0.9980  3.16       0.58
        4                 11.0                  34.0   0.9978  3.51       0.56

           alcohol  quality color acidity_levels
        0      9.4        5   RED            low
        1      9.8        5   RED       med-high
        2      9.8        5   RED        med-low
        3      9.8        6   RED       med-high
        4      9.4        5   RED            low
```

1.0.1 Create arrays for red bar heights white bar heights

Remember, there's a bar for each combination of color and quality rating. Each bar's height is based on the proportion of samples of that color with that quality rating. 1. Red bar proportions =

counts for each quality rating / total # of red samples 2. White bar proportions = counts for each quality rating / total # of white samples

```
In [3]: # get counts for each rating and color
        color_counts = wine_df.groupby(['color', 'quality']).count()['pH']
        color_counts
```

```
Out[3]: color  quality
        RED    3            10
               4            53
               5           681
               6           638
               7           199
               8            18
        WHITE  3            20
               4           163
               5          1457
               6          2198
               7           880
               8           175
               9             5
        Name: pH, dtype: int64
```

```
In [4]: # get total counts for each color
        color_totals = wine_df.groupby('color').count()['pH']
        color_totals
```

```
Out[4]: color
        RED      1599
        WHITE    4898
        Name: pH, dtype: int64
```

```
In [7]: # get proportions by dividing red rating counts by total # of red samples
        red_proportions = color_counts['RED'] / color_totals['RED']
        red_proportions
```

```
Out[7]: quality
        3    0.006254
        4    0.033146
        5    0.425891
        6    0.398999
        7    0.124453
        8    0.011257
        Name: pH, dtype: float64
```

```
In [8]: # get proportions by dividing white rating counts by total # of white samples
        white_proportions = color_counts['WHITE'] / color_totals['WHITE']
        white_proportions
```

```
Out[8]: quality
        3    0.004083
        4    0.033279
        5    0.297468
        6    0.448755
        7    0.179665
        8    0.035729
        9    0.001021
        Name: pH, dtype: float64
```

1.0.2 Plot proportions on a bar chart

Set the x coordinate location for each rating group and and width of each bar.

```
In [12]: ind = np.arange(len(red_proportions))  # the x locations for the groups
         width = 0.35       # the width of the bars
```

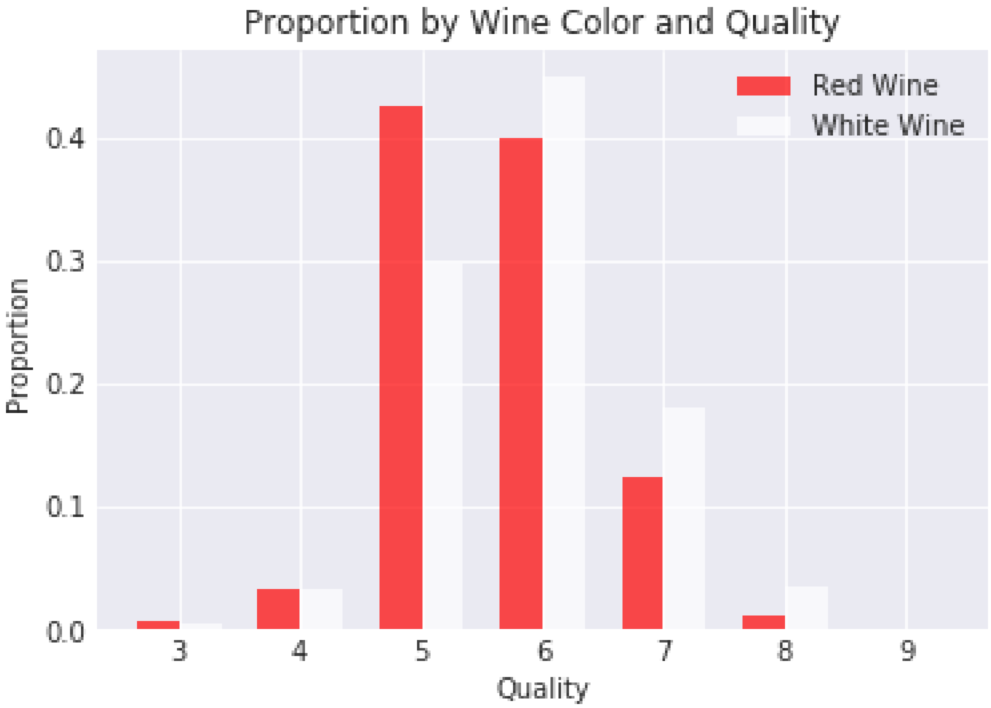
Now let's create the plot.

```
In [13]: # plot bars
         red_bars = plt.bar(ind, red_proportions, width, color='r', alpha=.7, label='Red Wine')
         white_bars = plt.bar(ind + width, white_proportions, width, color='w', alpha=.7, label=

         # title and labels
         plt.ylabel('Proportion')
         plt.xlabel('Quality')
         plt.title('Proportion by Wine Color and Quality')
         locations = ind + width / 2  # xtick locations
         labels = ['3', '4', '5', '6', '7', '8', '9']  # xtick labels
         plt.xticks(locations, labels)

         # legend
         plt.legend()

Out[13]: <matplotlib.legend.Legend at 0x7fdf69c0eda0>
```



Oh, that didn't work because we're missing a red wine value for a the 9 rating. Even though this number is a 0, we need it for our plot. Run the last two cells after running the cell below.

```
In [11]: red_proportions['9'] = 0
         red_proportions

Out[11]: quality
         3    0.006254
         4    0.033146
         5    0.425891
         6    0.398999
         7    0.124453
         8    0.011257
         9    0.000000
         Name: pH, dtype: float64

In [ ]:
```