conclusions_groupby

October 18, 2017

1 Drawing Conclusions Using Groupby

```
In [1]: # Load `winequality_edited.csv`
        import pandas as pd

        df = pd.read_csv('winequality_edited.csv')

        df.head()
```

```
Out[1]:    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
        0      9.4        5   RED
        1      9.8        5   RED
        2      9.8        5   RED
        3      9.8        6   RED
        4      9.4        5   RED
```

1.0.1 Is a certain type of wine associated with higher quality?

```
In [2]: # Find the mean quality of each wine type (red and white) with groupby
        df.groupby('color')['quality'].mean()
```

```
Out[2]: color
        RED      5.636023
```

```
WHITE    5.877909
Name: quality, dtype: float64
```

1.0.2 What level of acidity receives the highest average rating?

In [7]:
```
# View the min, 25%, 50%, 75%, max pH values with Pandas describe
df.describe()['pH']
```

Out[7]:
```
count    6497.000000
mean        3.218501
std         0.160787
min         2.720000
25%         3.110000
50%         3.210000
75%         3.320000
max         4.010000
Name: pH, dtype: float64
```

In [8]:
```
# Bin edges that will be used to "cut" the data into groups
bin_edges = [ 2.72, 3.11, 3.21, 3.32, 4.0] # Fill in this list with five values you just
```

In [19]:
```
# Labels for the four acidity level groups
bin_names = [ 'high', 'med-high', 'med-low', 'low'] # Name each acidity level category
```

In [20]:
```
# Creates acidity_levels column
df['acidity_levels'] = pd.cut(df['pH'], bin_edges, labels=bin_names)

# Checks for successful creation of this column
df.head()
```

Out[20]:
```
   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
```

```
In [21]: # Find the mean quality of each acidity level with groupby
         df.groupby('acidity_levels')['quality'].mean()

Out[21]: acidity_levels
         high        5.783343
         med-high    5.784540
         med-low     5.850832
         low         5.859415
         Name: quality, dtype: float64

In [22]: # Save changes for the next section
         df.to_csv('winequality_edited.csv', index=False)

In [ ]:
```