Investigate a dataset on wine quality using Python

November 12, 2019

1 Data Analysis on Wine Quality Data Set

Investigate the dataset on physicochemical properties and quality ratings of red and white wine samples.

1.0.1 Gathering Data

[103]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
red_df = pd.read_csv("winequality-red.csv",sep=';')
white_df = pd.read_csv('winequality-white.csv',sep=';')
```

### Assessing Data > 1.Number of samples in each data set.

2.Number of columns in each data set.

[8]:
```
print(red_df.shape)
red_df.head()
```

```
(1599, 12)
```

[8]:
```
   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
0      9.4        5
1      9.8        5
2      9.8        5
3      9.8        6
4      9.4        5
```

[9]:
```
print(white_df.shape)
white_df.head()
```

```
(4898, 12)
```

[9]:
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.0              0.27         0.36            20.7      0.045
1            6.3              0.30         0.34             1.6      0.049
2            8.1              0.28         0.40             6.9      0.050
3            7.2              0.23         0.32             8.5      0.058
4            7.2              0.23         0.32             8.5      0.058

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 45.0                 170.0   1.0010  3.00       0.45
1                 14.0                 132.0   0.9940  3.30       0.49
2                 30.0                  97.0   0.9951  3.26       0.44
3                 47.0                 186.0   0.9956  3.19       0.40
4                 47.0                 186.0   0.9956  3.19       0.40

   alcohol  quality
0      8.8        6
1      9.5        6
2     10.1        6
3      9.9        6
4      9.9        6
```

Checking for features with missing values.

[10]:
```
red_df.isnull().sum()
```

[10]:
```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
```

```
dtype: int64
```

[11]:
```
white_df.isnull().sum()
```

[11]:
```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

Are there any duplicate rows in these datasets significant/need to be dropped?

[14]:
```
white_df.duplicated().sum()
```

[14]:
```
937
```

[15]:
```
red_df.duplicated().sum()
```

[15]:
```
240
```

Finding the number of unique values for quality in eeach dataset?

[16]:
```
red_df.quality.nunique()
```

[16]:
```
6
```

[17]:
```
white_df.quality.nunique()
```

[17]:
```
7
```

What is the mean density in the red wine dataset?

[19]:
```
red_df.density.mean()
```

[19]:
```
0.996746679174484
```

1.0.2 Appending Data

merging the two datasets, red and white wine data, into a single data.

Create Color Columns Create two arrays as long as the number of rows in the red and white dataframes that repeat the value "red" or "white."

[24]:
```
# create color array for red dataframe
color_red = np. repeat('red',red_df.shape[0])
# create color array for white dataframe
color_white = np.repeat ('white',white_df.shape[0])
```

Adding arrays to the white and red dataframes

[25]:
```
red_df['color']=color_red
red_df.head()
```

[25]:
```
   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
```

[27]:
```
white_df['color']=color_white
white_df.head()
```

[27]:
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.0              0.27         0.36            20.7      0.045
1            6.3              0.30         0.34             1.6      0.049
2            8.1              0.28         0.40             6.9      0.050
3            7.2              0.23         0.32             8.5      0.058
4            7.2              0.23         0.32             8.5      0.058

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 45.0                 170.0   1.0010  3.00       0.45
1                 14.0                 132.0   0.9940  3.30       0.49
2                 30.0                  97.0   0.9951  3.26       0.44
3                 47.0                 186.0   0.9956  3.19       0.40
4                 47.0                 186.0   0.9956  3.19       0.40

   alcohol  quality  color
0      8.8        6  white
1      9.5        6  white
2     10.1        6  white
3      9.9        6  white
4      9.9        6  white
```

Combine DataFrames with Append

[34]:
```
# append dataframes
wine_df = red_df.append(white_df)
# view dataframe to check for success
wine_df.head()
wine_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6497 entries, 0 to 4897
Data columns (total 13 columns):
fixed acidity           6497 non-null float64
volatile acidity        6497 non-null float64
citric acid             6497 non-null float64
residual sugar          6497 non-null float64
chlorides               6497 non-null float64
free sulfur dioxide     6497 non-null float64
total sulfur dioxide    6497 non-null float64
density                 6497 non-null float64
pH                      6497 non-null float64
sulphates               6497 non-null float64
alcohol                 6497 non-null float64
quality                 6497 non-null int64
color                   6497 non-null object
dtypes: float64(11), int64(1), object(1)
memory usage: 710.6+ KB
```

Save Combined Dataset
Save newly combined dataframe as winequality_edited.csv.

[33]:
```
wine_df.to_csv('winequality_edited.csv', index=False)
```

1.0.3 Exploring with visuals

Based on histograms of columns in this dataset, which of the following feature variables appear skewed to the right?

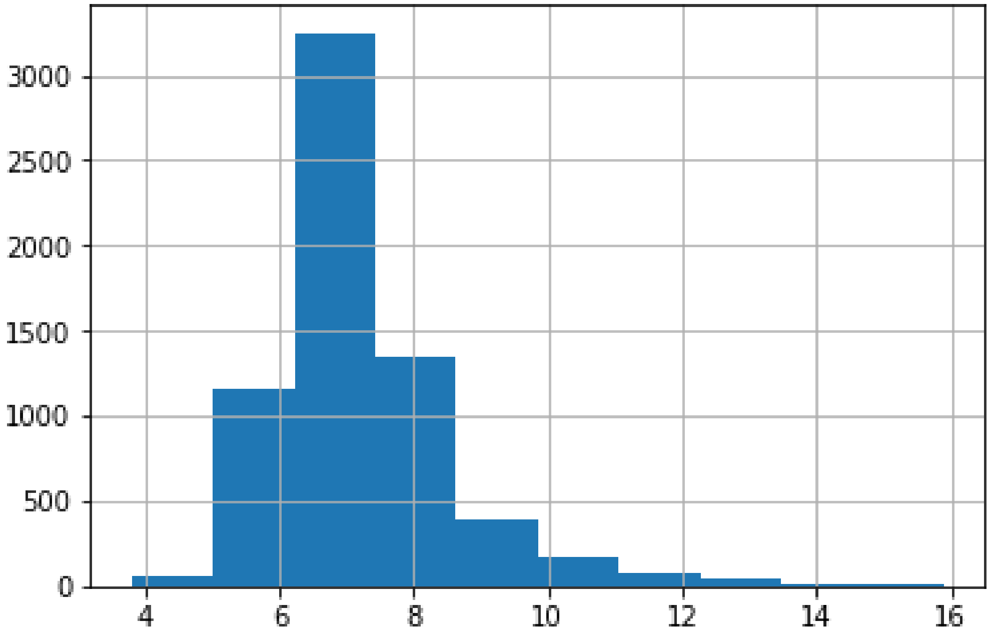
[41]:
```
# Load dataset
df = pd.read_csv('winequality_edited.csv')
df.head()
```

[41]:
```
   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
```

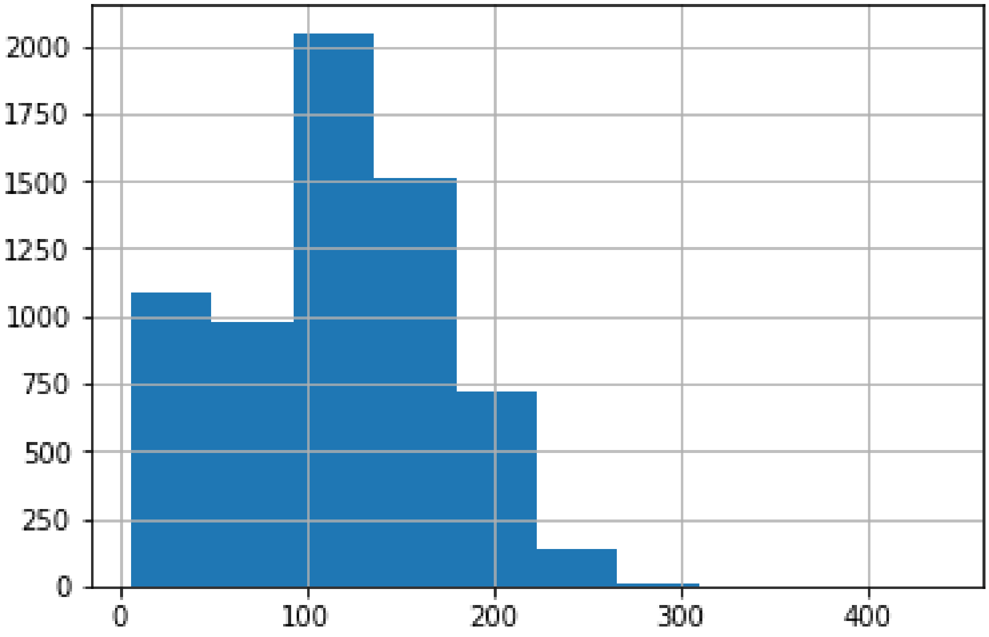
```
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56

   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
```

Histograms for Various Features

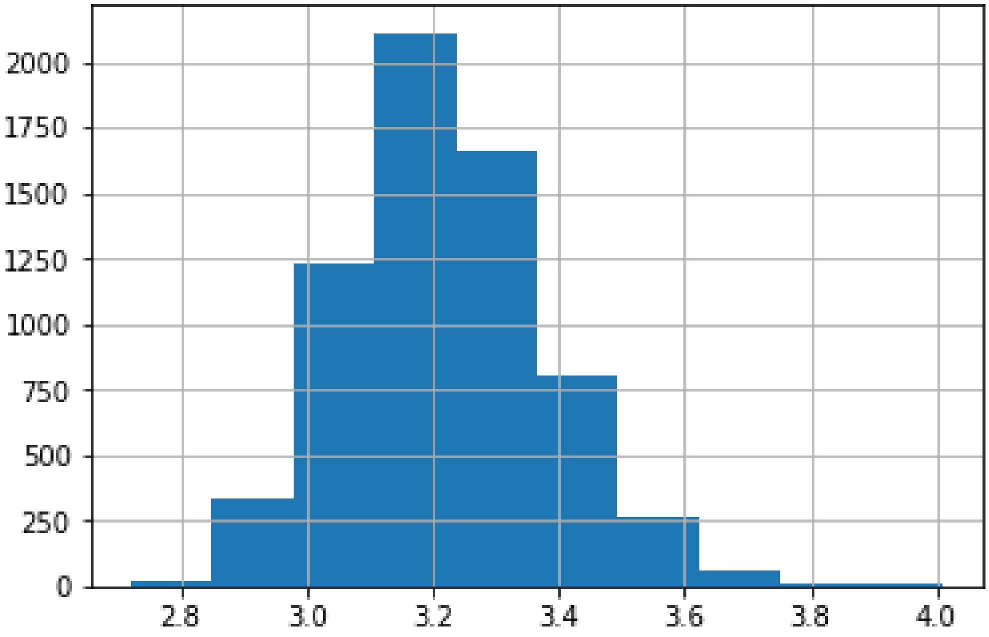
[43]:
```
df['fixed acidity'].hist();
```



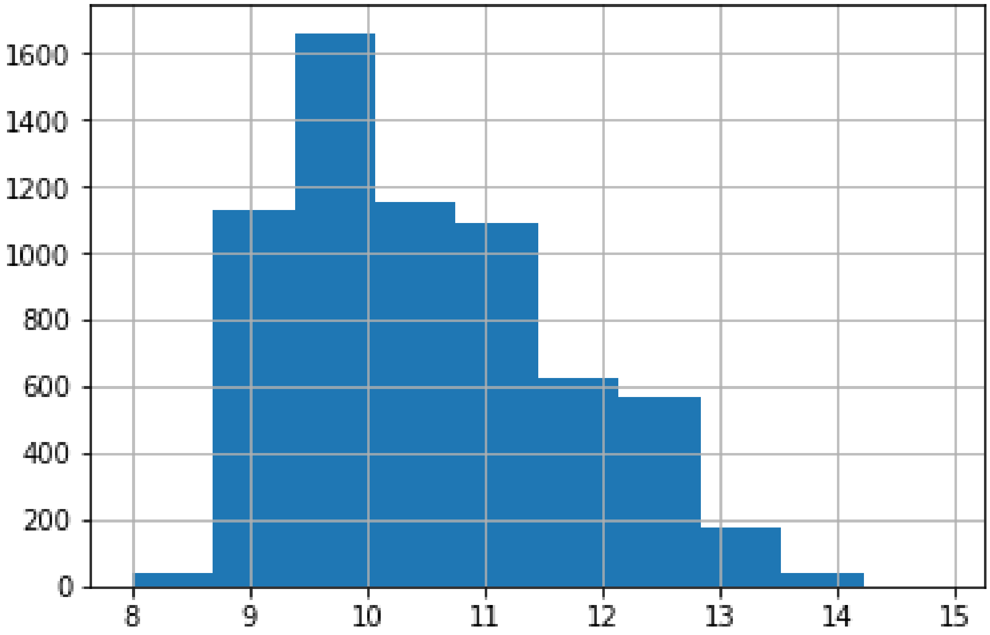
[44]:
```
df['total sulfur dioxide'].hist();
```



[45]:
```
df['pH'].hist();
```



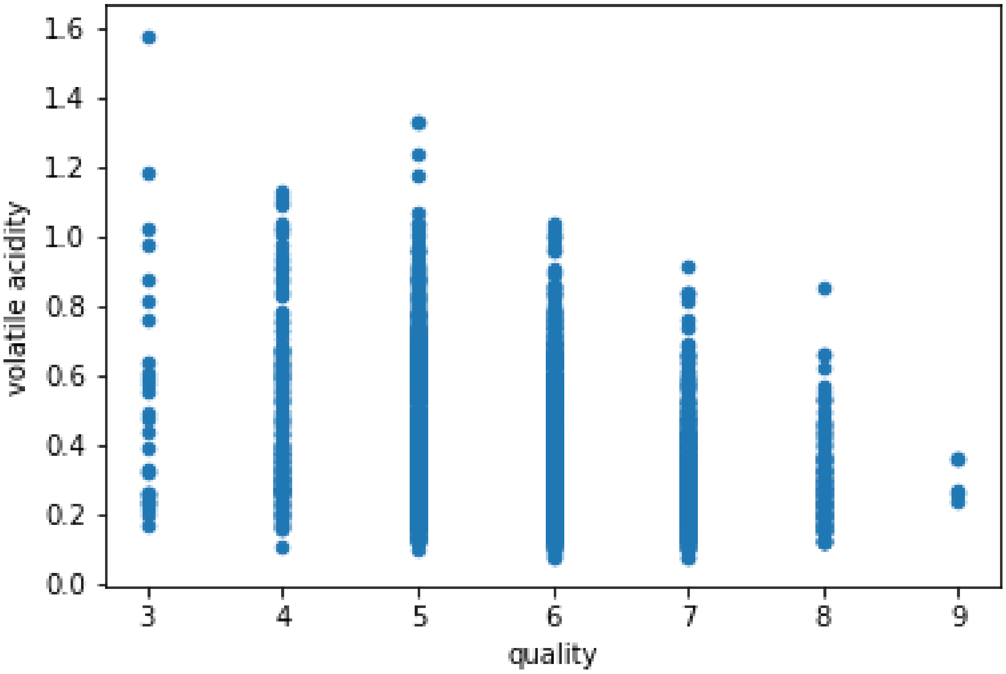
[46]:
```
df['alcohol'].hist();
```



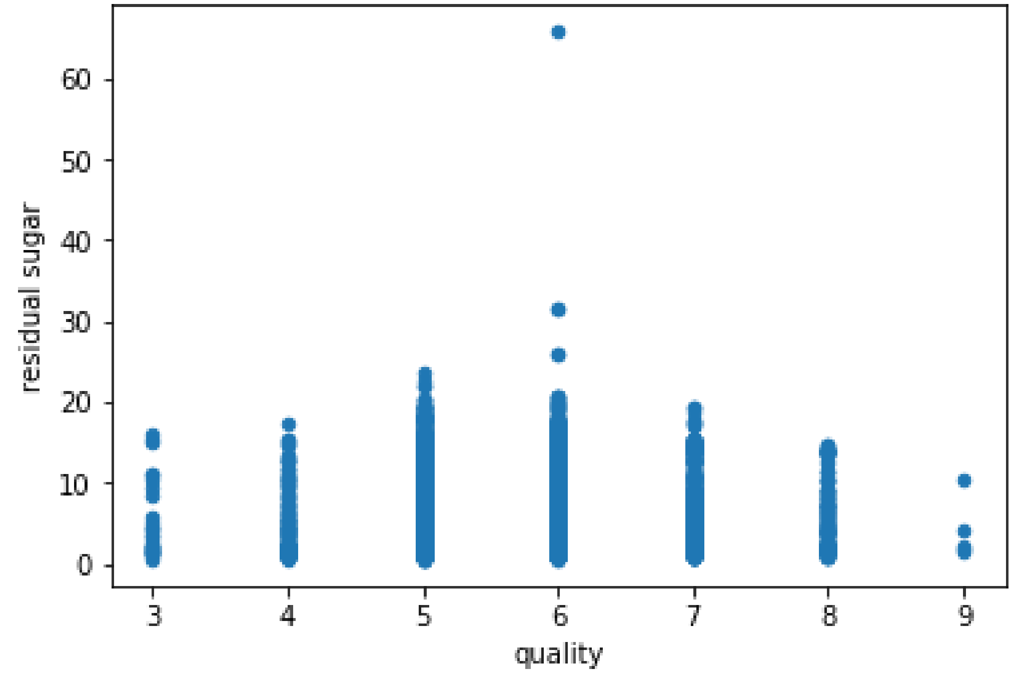
Based on the above plots Fixed Acidity appears skewed to right.

1.0.4 Scatterplots of Quality Against Various Features

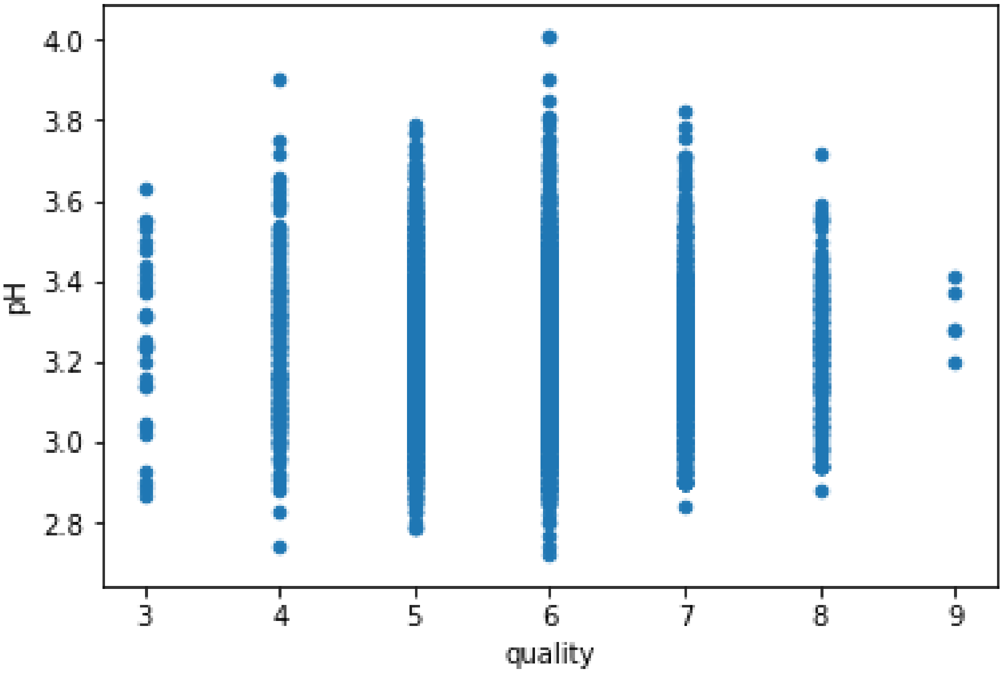
[50]:
```
df.plot(x='quality',y='volatile acidity',kind ='scatter');
```



[51]:
```
df.plot(x='quality',y='residual sugar',kind ='scatter');
```

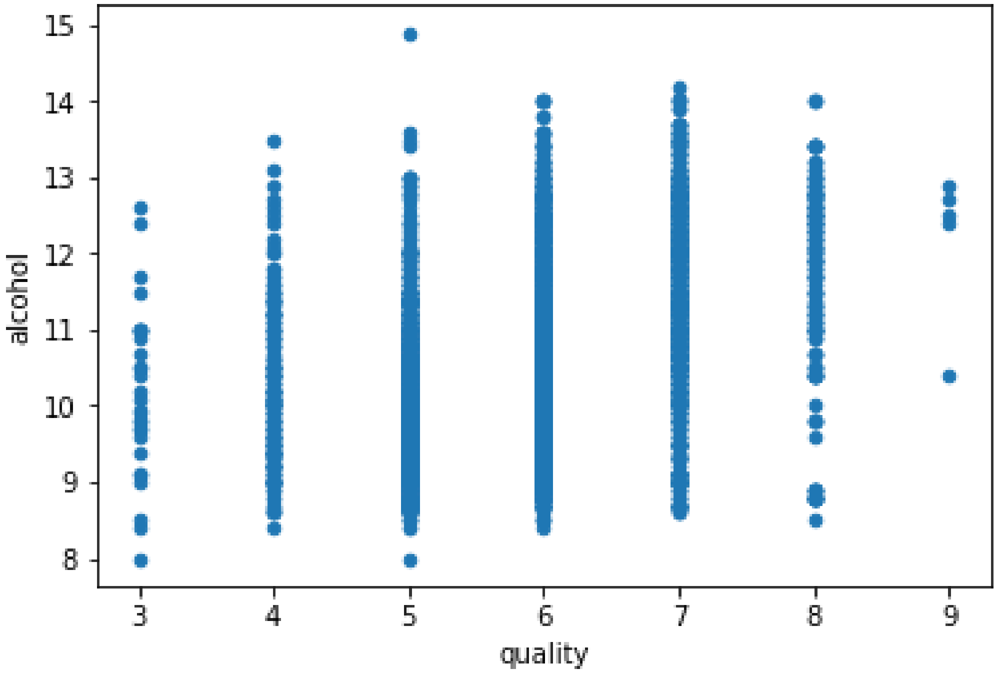


[52]:
```
df.plot(x='quality',y='pH',kind ='scatter');
```



[53]:

```
df.plot(x='quality',y='alcohol',kind ='scatter');
```



Based on scatterplots of quality against different feature variables, Alcohol is most likely to have a positive impact on quality.

1.0.5 Conclusions using Groupby

Q1: Is a certain type of wine (red or white) associated with higher quality?

[54]:

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

[54]:

```
color
red      5.636023
white    5.877909
Name: quality, dtype: float64
```

the mean quality of red wine is less than that of white wine.

Q2: What level of acidity (pH value) receives the highest average rating?

[55]:

```
# View the min, 25%, 50%, 75%, max pH values with Pandas describe
df.describe().pH
```

[55]:

```
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
```

[56]:
```
# Bin edges that will be used to "cut" the data into groups
bin_edges = [2.72, 3.11, 3.21, 3.32, 4.01] # Fill in this list with five values␣
 ↪you just found
```

[57]:
```
# Labels for the four acidity level groups
bin_names = ['high', 'mod_high', 'medium', 'low'] # Name each acidity level␣
 ↪category
```

[58]:
```
# 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()
```

[58]:
```
   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       mod_high
2      9.8        5   red         medium
3      9.8        6   red       mod_high
4      9.4        5   red            low
```

[61]:
```
What level of acidity receives the highest mean quality rating?
```

```
Object `rating` not found.
```

[ ]:
```
What level of acidity receives the highest mean quality rating
```

[59]:
```
# Find the mean quality of each acidity level with groupby
df.groupby('acidity_levels').mean().quality
```

[59]:
```
acidity_levels
high        5.783343
mod_high    5.784540
medium      5.850832
low         5.859593
Name: quality, dtype: float64
```

Low level of acidity recieves the highest mean quality rating.

[60]:
```
# Save changes for the next section
df.to_csv('winequality_edited.csv', index=False)
```

1.0.6 Conclusions Using Query

Q1: Do wines with higher alcoholic content receive better ratings?

[63]:
```
df = pd.read_csv('winequality_edited.csv')
```

[66]:
```
# get the median amount of alcohol content
df.alcohol.median()
```

[66]:
```
10.3
```

[71]:
```
# select samples with alcohol content less than the median
## low_alcohol = df[df.alcohol < 10.3]
low_alcohol = df.query('alcohol < 10.3')
```

[72]:
```
# select samples with alcohol content greater than or equal to the median
## high_alcohol = df[df.alcohol >= 10.3]
high_alcohol= df.query('alcohol >= 10.3')
```

[79]:
```
# ensure these queries included each sample exactly once
num_samples = df.shape[0]
num_samples == low_alcohol['quality'].count() + high_alcohol['quality'].count()␣
 ↪# should be True
```

[79]:
```
True
```

[74]:
```
# get mean quality rating for the low alcohol and high alcohol groups
low_alcohol.quality.mean(), high_alcohol.quality.mean()
```

[74]:
```
(5.475920679886686, 6.146084337349397)
```

wines with higher alcoholic content generally receive better ratings

1.0.7 Plotting with Matplotlib

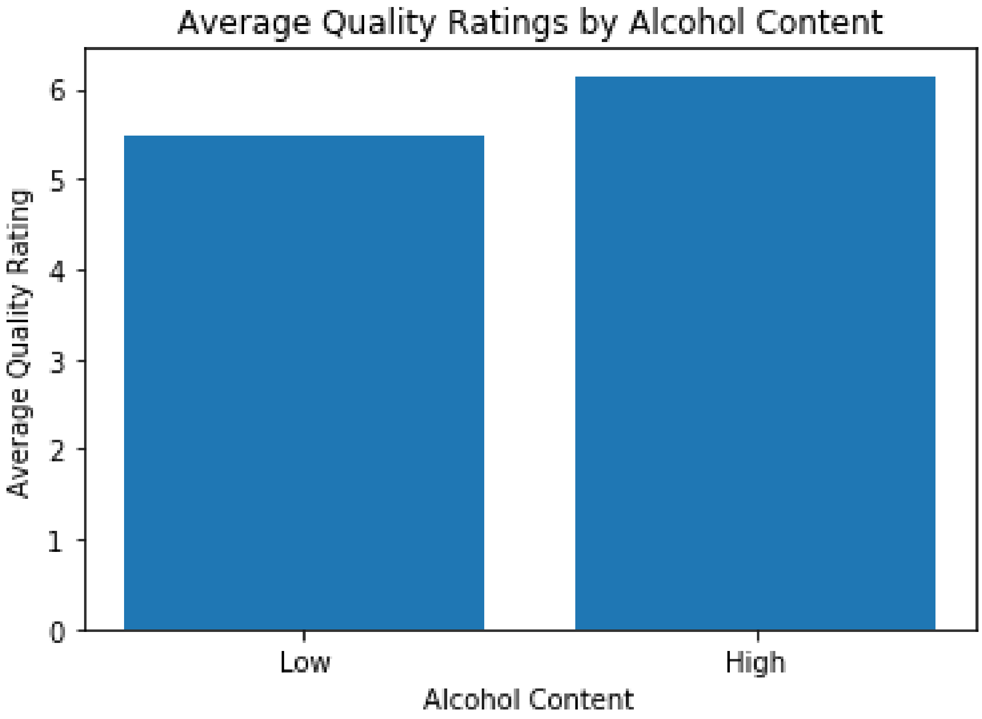
Use Matplotlib to create bar charts that visualize the conclusions made with groupby and query.

[94]:
```
# Use query to select each group and get its mean quality
median = df['alcohol'].median()
low = df.query('alcohol < {}'.format(median))
high = df.query('alcohol >= {}'.format(median))
```

```
mean_quality_low = low['quality'].mean()
mean_quality_high = high['quality'].mean()
```

[95]:

```
# Create a bar chart with proper labels
locations = [1, 2]
heights = [mean_quality_low, mean_quality_high]
labels = ['Low', 'High']
plt.bar(locations, heights, tick_label=labels)
plt.title('Average Quality Ratings by Alcohol Content')
plt.xlabel('Alcohol Content')
plt.ylabel('Average Quality Rating');
```



What level of acidity receives the highest average rating? > Create a bar chart with a bar for each of the four acidity levels.

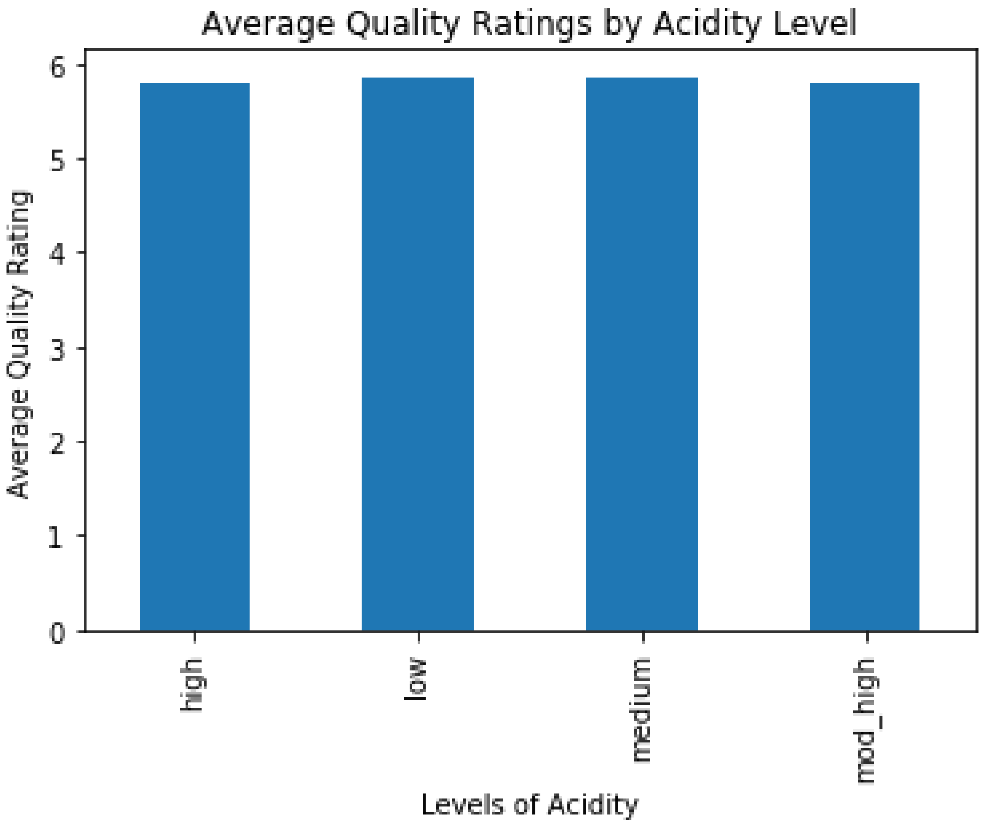
[99]:

```
# Use groupby to get the mean quality for each acidity level
mean = df.groupby('acidity_levels').mean().quality
```

[100]:

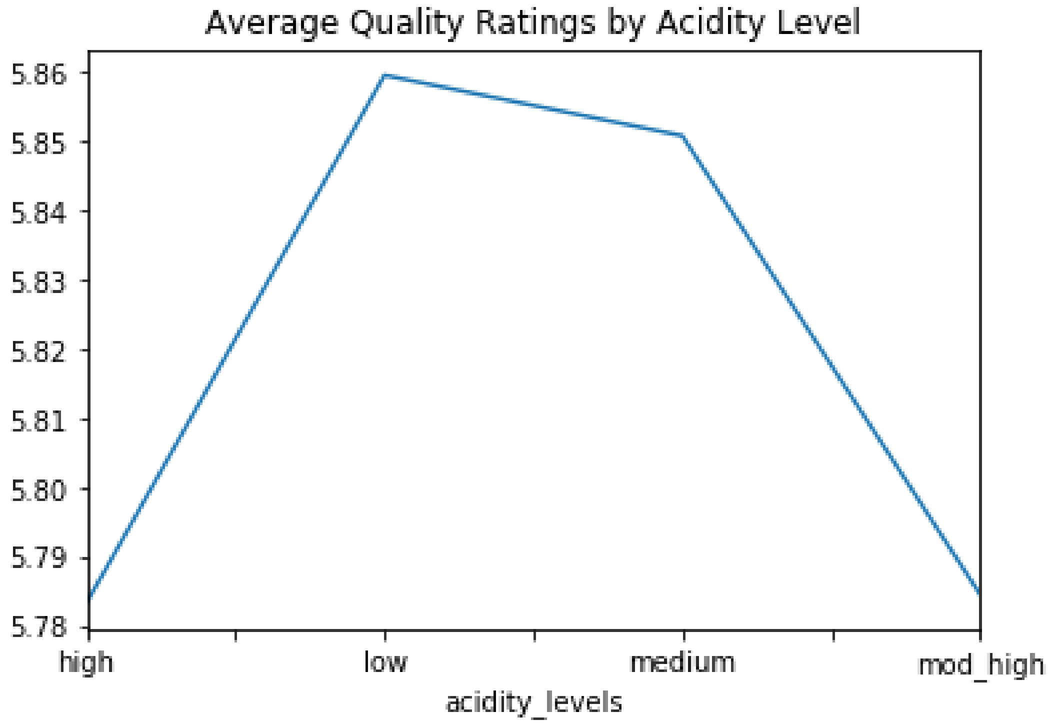
```
# Create a bar chart with proper labels
df.groupby('acidity_levels')['quality'].mean().plot(kind='bar',title='Average␣
 ↪Quality Ratings by Acidity Level')

#locations = [1, 2,3,4]
#heights = [High,Low,Medium,Moderately_High]
labels = ["High","Low","Medium","Moderately_High"]
#plt.bar(locations, heights, tick_label=labels)
```

```
#plt.title('Average Quality Ratings by Residual Sugar')
plt.xlabel('Levels of Acidity')
plt.ylabel('Average Quality Rating');
```



Create a line plot for each of the four acidity levels.

[102]:
```
df.groupby('acidity_levels')['quality'].mean().plot(kind='line',title='Average␣
 ↪Quality Ratings by Acidity Level');
```


Average Quality Ratings by Acidity Level
5.86
5.85
5.84
5.83
5.82
5.81
5.80
5.79
5.78
high
low
medium
mod_high
acidity_levels