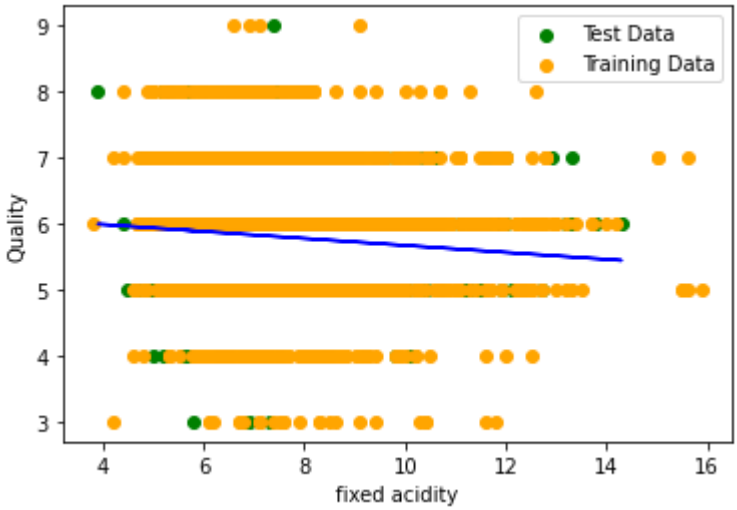
Regression analysis

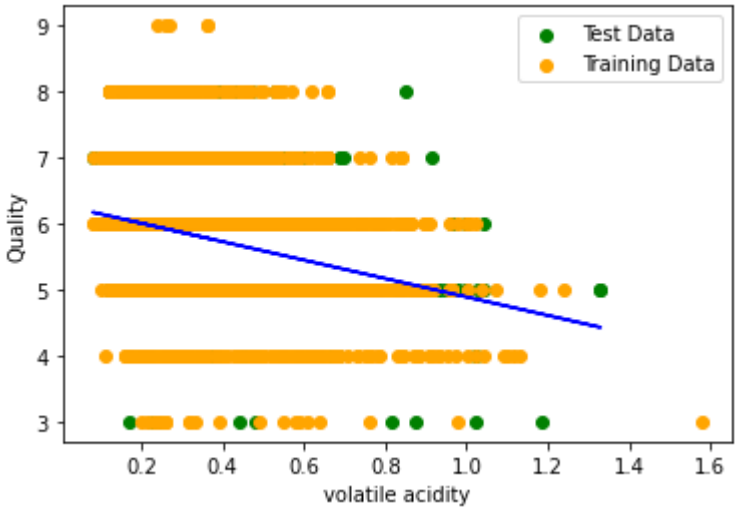
A check for missing/null values was performed in the classification notebook and note duplicated here.

Generate regression model for each feature independently. First for red and white together, then red only, then white only

```
Regression models for red and white wine
both data for feature: fixed acidity
Coefficients:
 [-0.05290104]
Mean Squared Error:
5.819022226981535
Variance score:
0.0051585347776136325
```

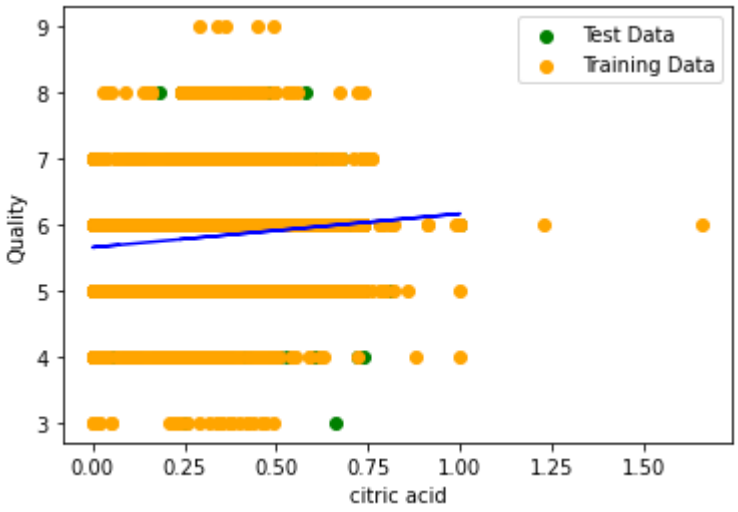



```
both data for feature: volatile acidity
Coefficients:
 [-1.39581448]
Mean Squared Error:
5.822450927502022
Variance score:
0.08642909501442242
```

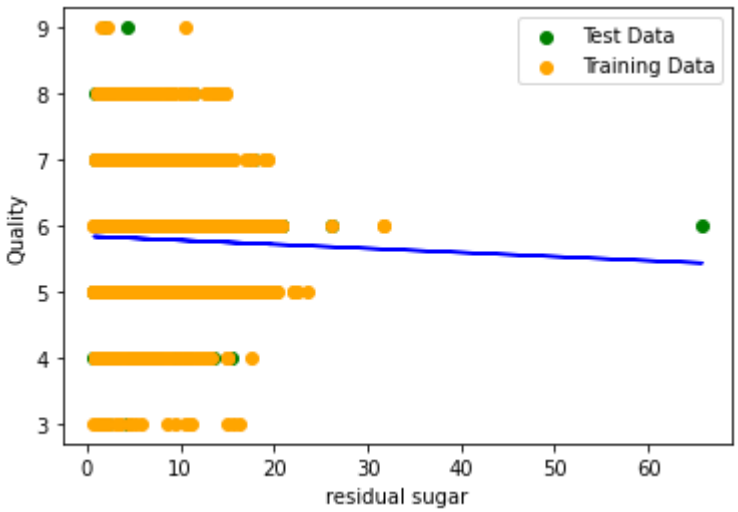



```
both data for feature: citric acid
Coefficients:
 [0.50895235]
Mean Squared Error:
5.818589093792448
```

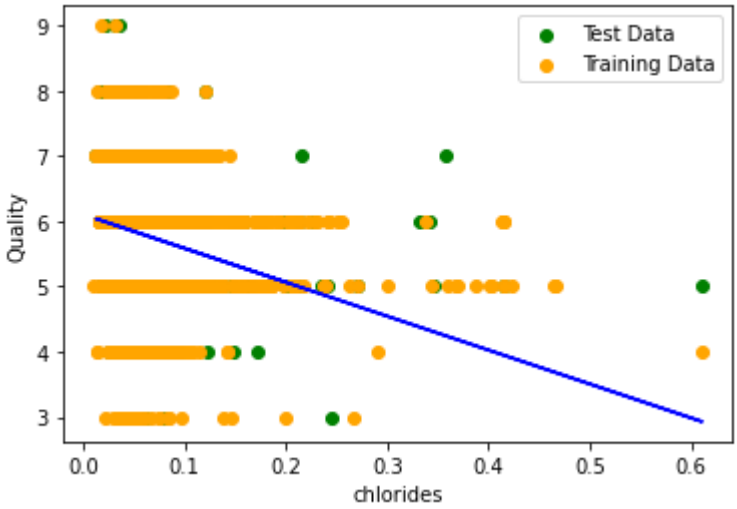
```
Variance score:
0.008388305047866584
```



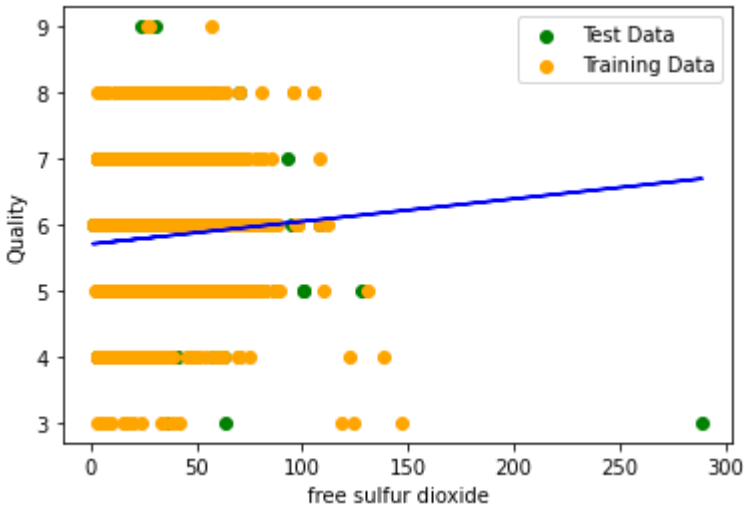

```
both data for feature: residual sugar
Coefficients:
 [-0.00617331]
Mean Squared Error:
5.8356463269298375
Variance score:
-0.002125409500149944
```



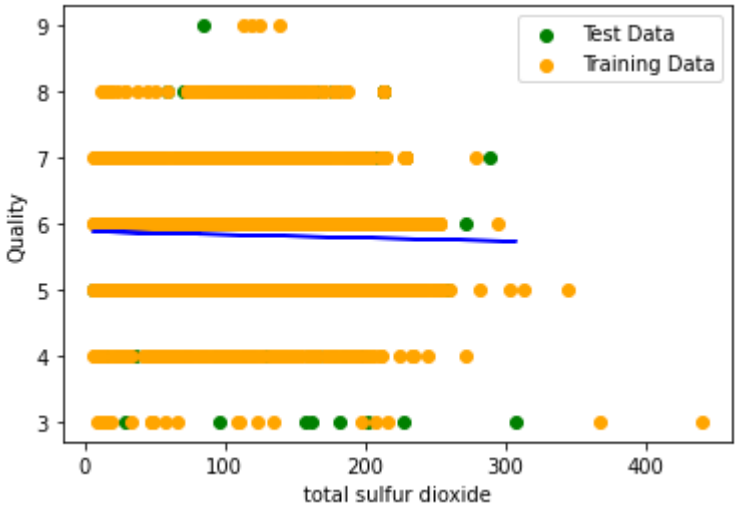

```
both data for feature: chlorides
Coefficients:
 [-5.1963059]
Mean Squared Error:
5.828010119237531
Variance score:
0.02946677125556163
```



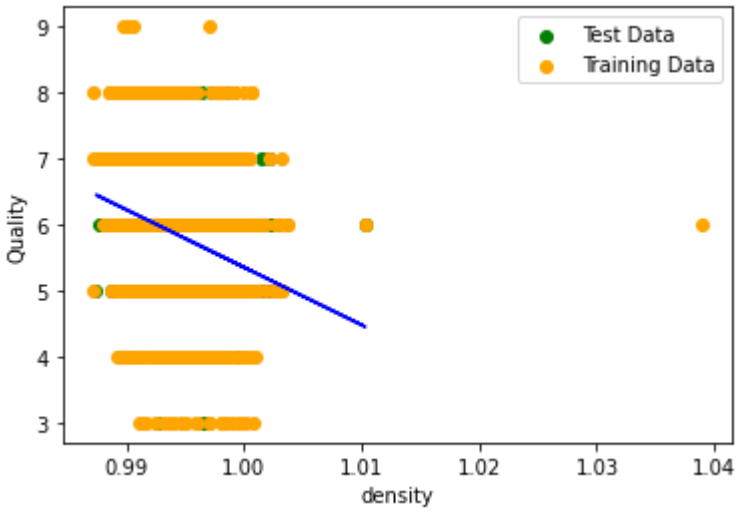

```
both data for feature: free sulfur dioxide
Coefficients:
 [0.00342908]
Mean Squared Error:
5.835329275865919
Variance score:
-0.007013367698172512
```



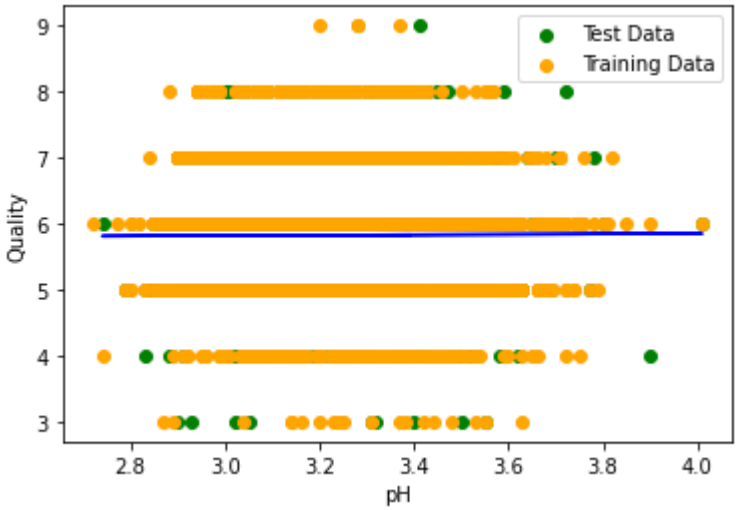

```
both data for feature: total sulfur dioxide
Coefficients:
 [-0.00048703]
Mean Squared Error:
5.8064554928690715
Variance score:
0.002416325030431987
```



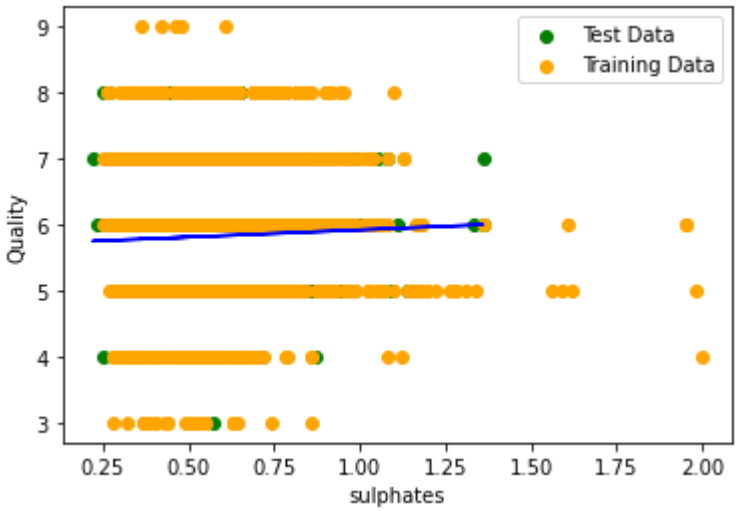

```
both data for feature: density
Coefficients:
 [-87.20937353]
Mean Squared Error:
5.829169489689966
Variance score:
0.10355023791278939
```



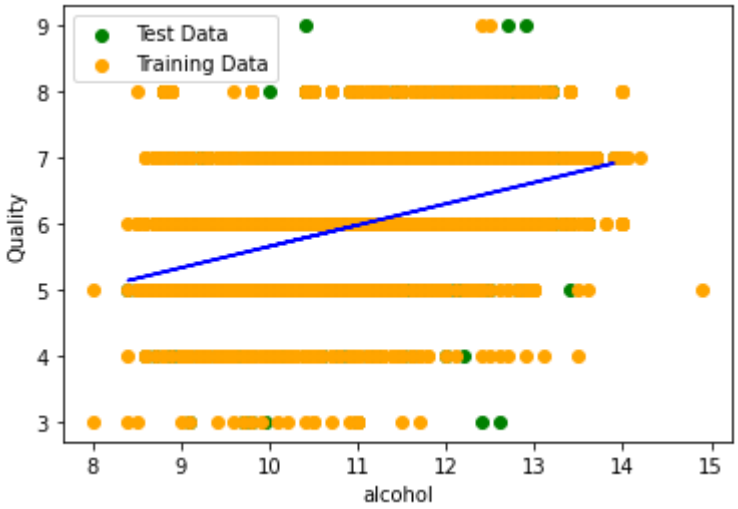

```
both data for feature: pH
Coefficients:
 [0.03134559]
Mean Squared Error:
5.808589453661387
Variance score:
-0.0005452896356985537
```



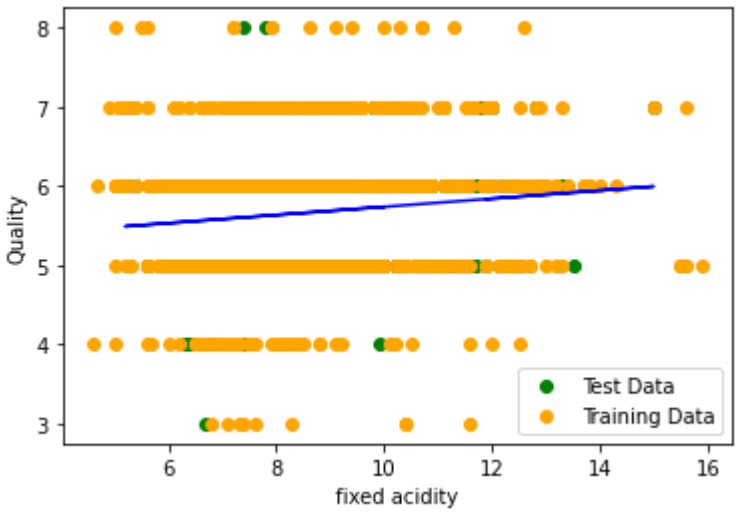

```
both data for feature: sulphates
Coefficients:
 [0.21891502]
Mean Squared Error:
5.819334487948806
Variance score:
0.0019284208174265016
```



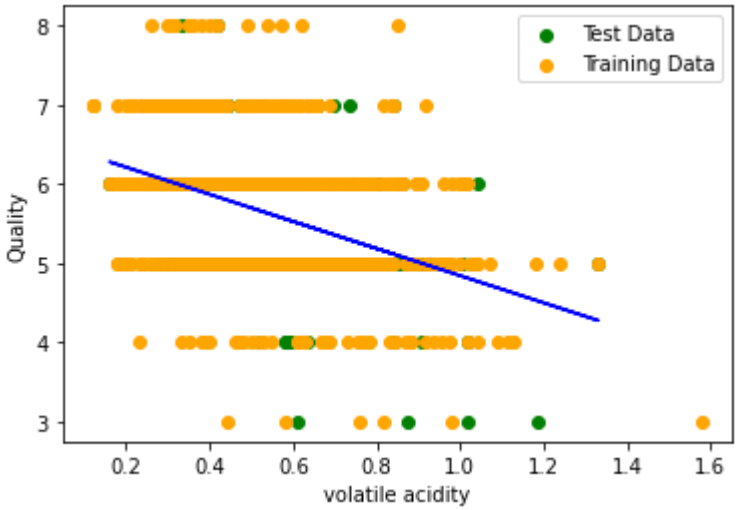

```
both data for feature: alcohol
Coefficients:
 [0.32316386]
Mean Squared Error:
5.828397041513352
Variance score:
0.19554155113265093
```



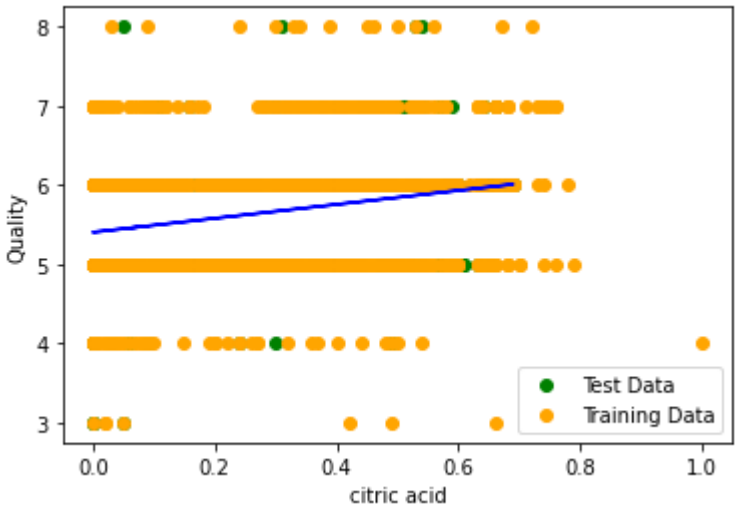

```
Regression models for red wine
red data for feature: fixed acidity
Coefficients:
 [0.05153547]
Mean Squared Error:
5.609065693089349
Variance score:
0.019239942973329982
```



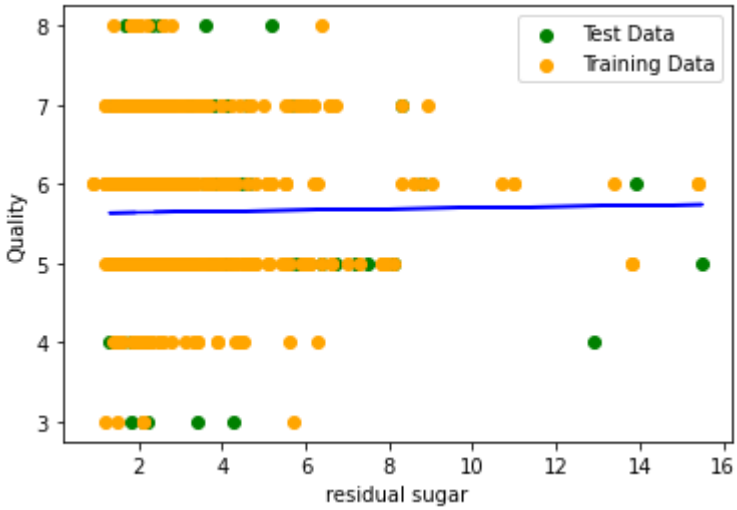

```
red data for feature: volatile acidity
Coefficients:
 [-1.71119853]
Mean Squared Error:
5.6246088667470255
Variance score:
0.16196747006304724
```



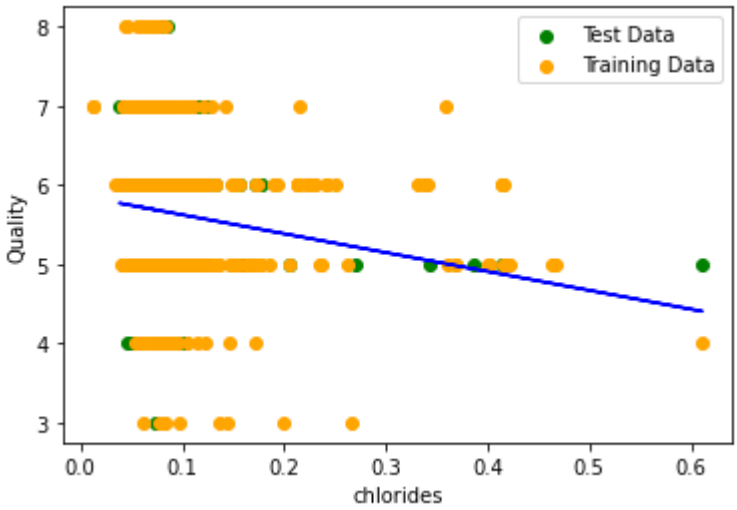

```
red data for feature: citric acid
Coefficients:
 [0.87805888]
Mean Squared Error:
5.623541456469117
Variance score:
0.07127717788572474
```



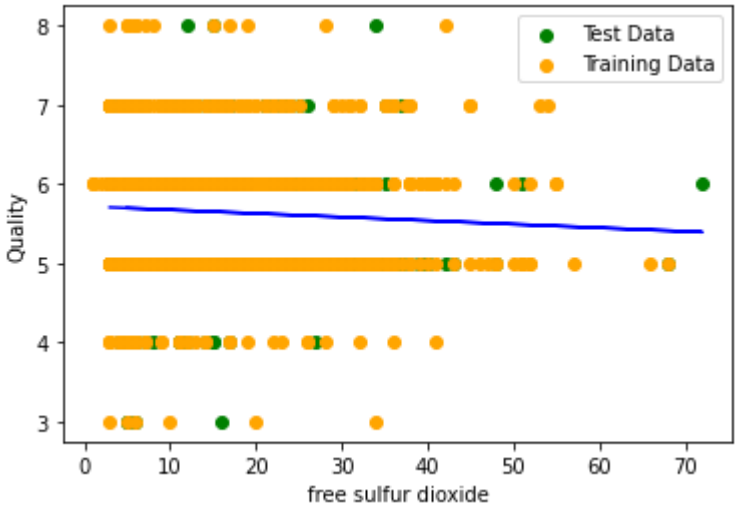

```
red data for feature: residual sugar
Coefficients:
 [0.00741804]
Mean Squared Error:
5.622449228557955
Variance score:
-0.0025780773092685116
```



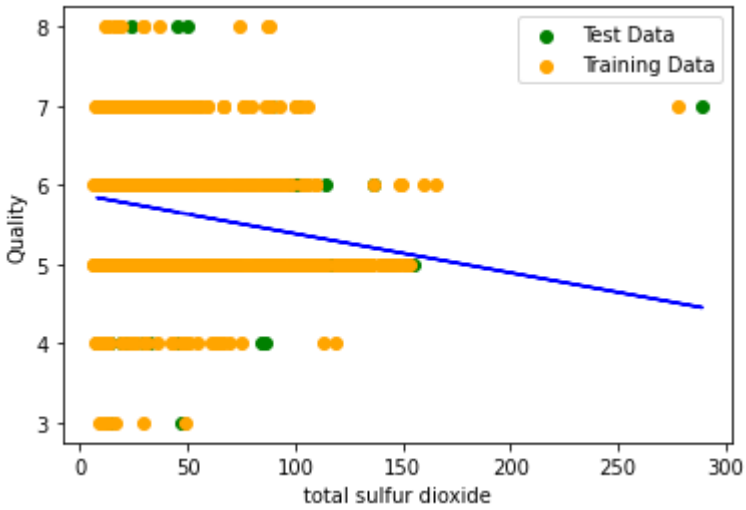

```
red data for feature: chlorides
Coefficients:
 [-2.38415527]
Mean Squared Error:
5.613234362929125
Variance score:
-0.005762013457862869
```



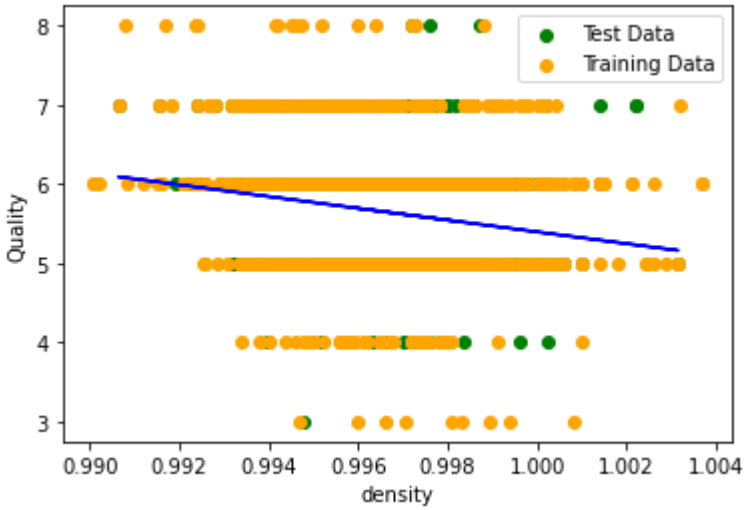

```
red data for feature: free sulfur dioxide
Coefficients:
 [-0.00452804]
Mean Squared Error:
5.6183260034522196
Variance score:
-0.005553519467990542
```



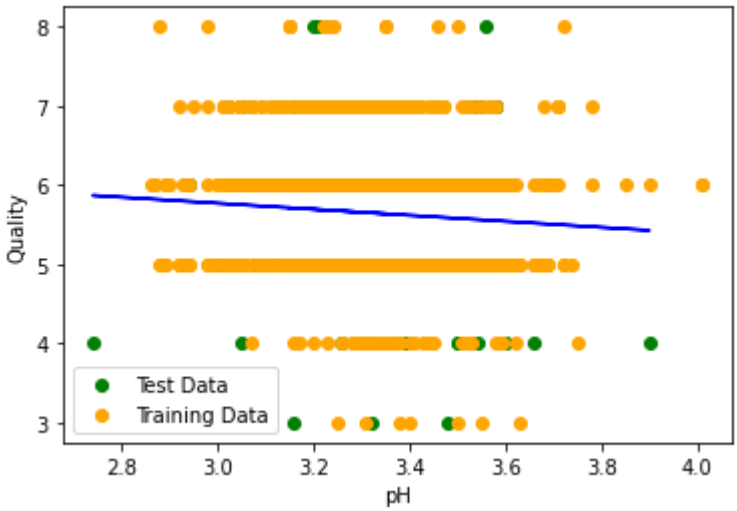

```
red data for feature: total sulfur dioxide
Coefficients:
 [-0.00493608]
Mean Squared Error:
5.607216863565033
Variance score:
0.006576398971764541
```



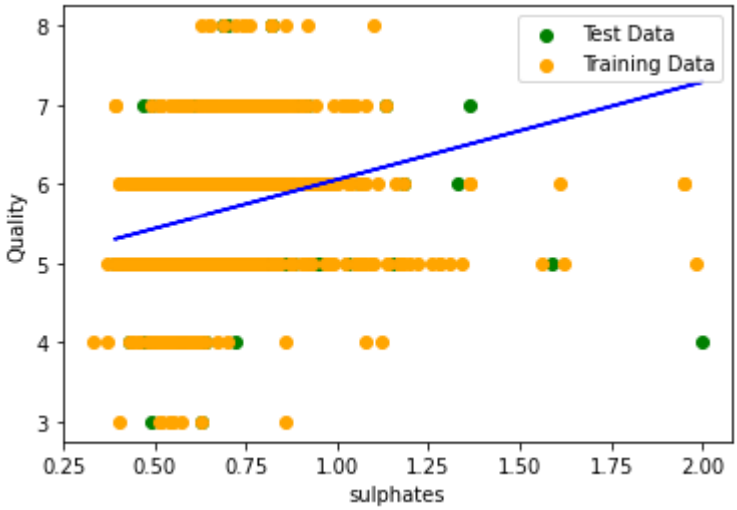

```
red data for feature: density
Coefficients:
 [-74.43724293]
Mean Squared Error:
5.623506374739074
Variance score:
0.03087804409563133
```



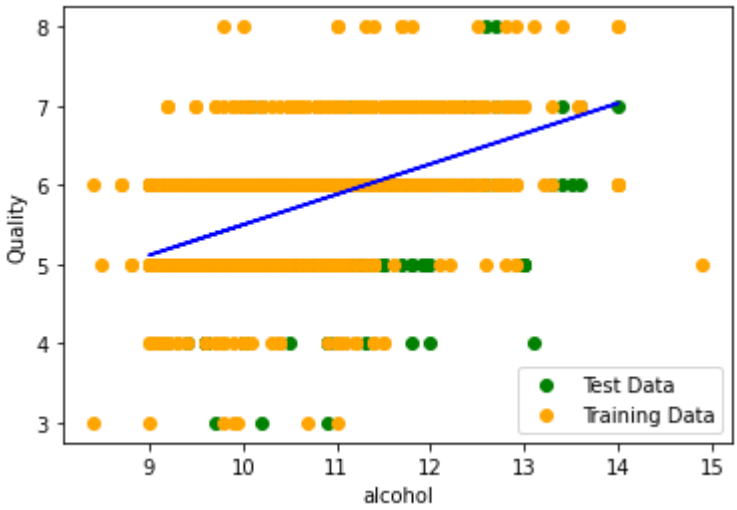

```
red data for feature: pH
Coefficients:
 [-0.38046389]
Mean Squared Error:
5.614271906611424
Variance score:
-0.014935084228442141
```



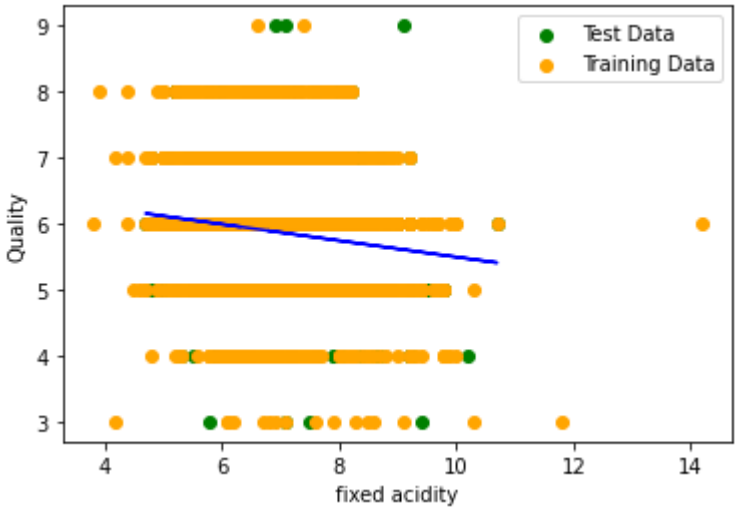

```
red data for feature: sulphates
Coefficients:
 [1.23244556]
Mean Squared Error:
5.628268402388561
Variance score:
0.04692424097326808
```



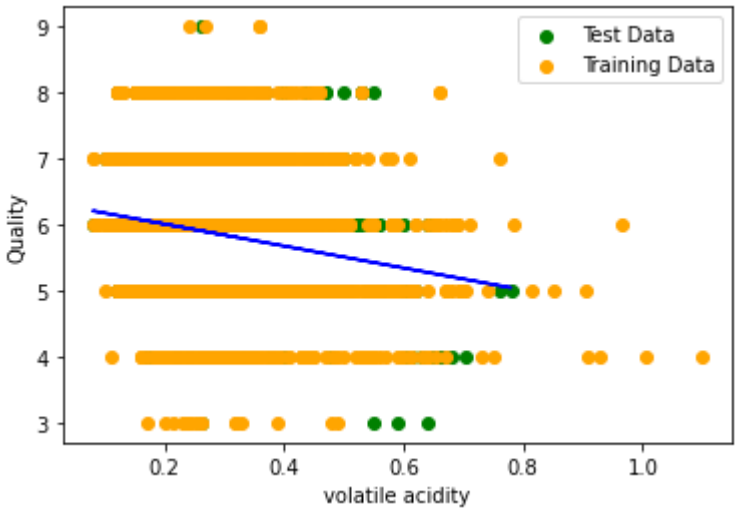

```
red data for feature: alcohol
Coefficients:
 [0.38368334]
Mean Squared Error:
5.6434134022849545
Variance score:
0.10954480469123573
```



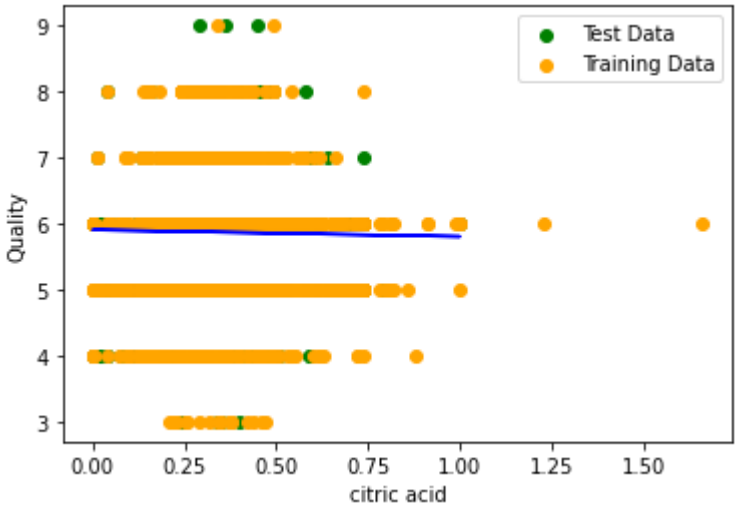

```
Regression models for white wine
white data for feature: fixed acidity
Coefficients:
 [-0.12430697]
Mean Squared Error:
5.86784682252961
Variance score:
0.005770361874784968
```



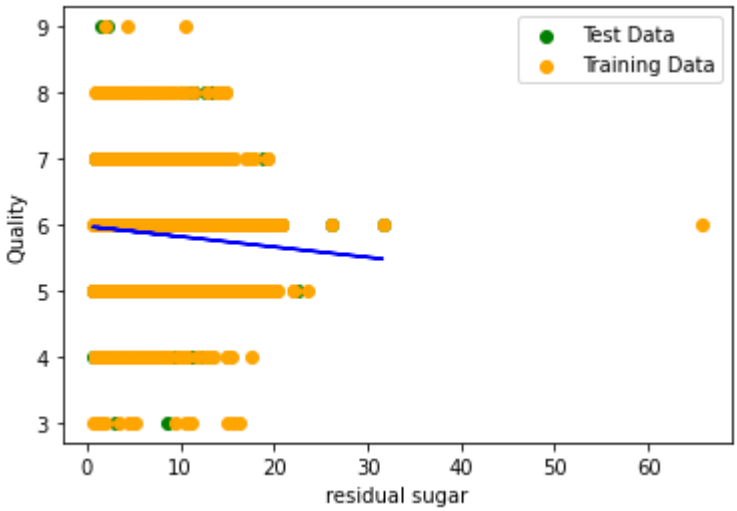

```
white data for feature: volatile acidity
Coefficients:
 [-1.66447563]
Mean Squared Error:
5.868652533026381
Variance score:
0.04509161291981989
```



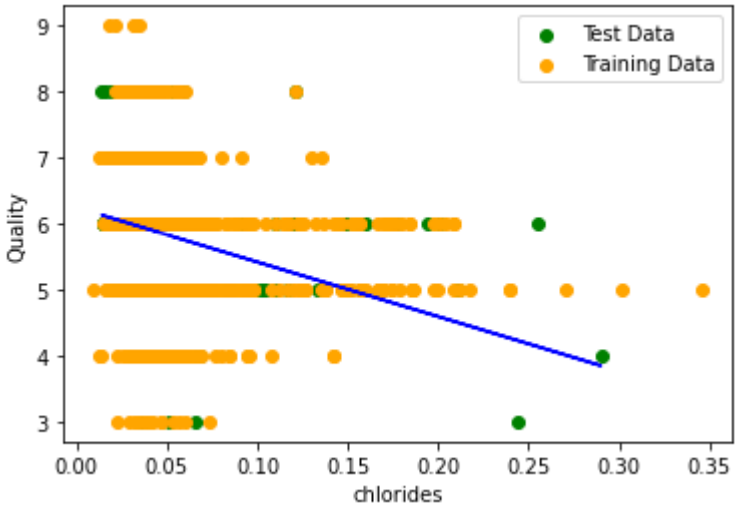

```
white data for feature: citric acid
Coefficients:
 [-0.10200337]
Mean Squared Error:
5.886624467506095
Variance score:
-0.0014758388329447758
```



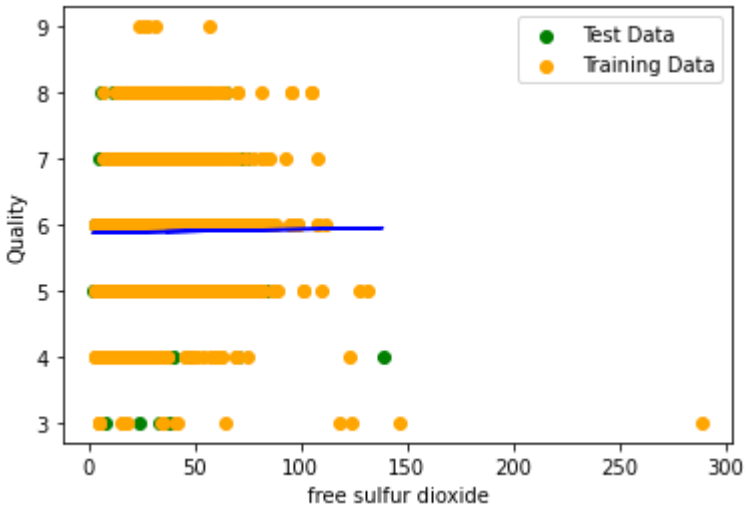

```
white data for feature: residual sugar
Coefficients:
 [-0.01559132]
Mean Squared Error:
5.882678401049191
Variance score:
0.015681949396942785
```



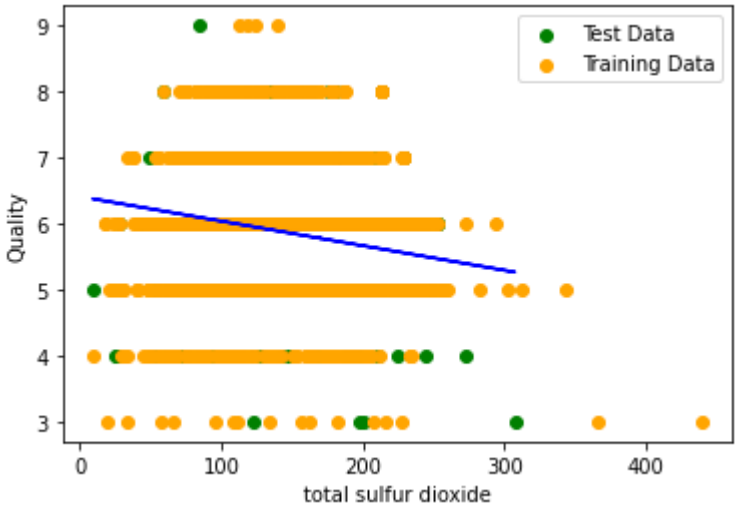

```
white data for feature: chlorides
Coefficients:
 [-8.26163935]
Mean Squared Error:
5.898281322662372
Variance score:
0.04602389484661995
```



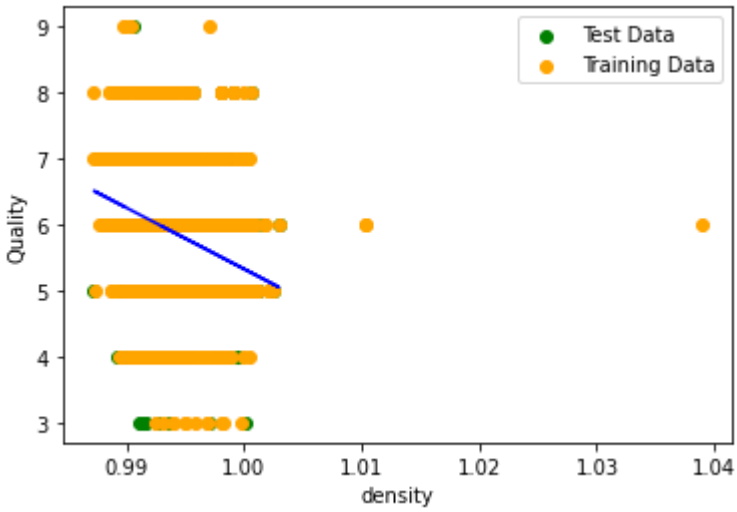

```
white data for feature: free sulfur dioxide
Coefficients:
 [0.00055579]
Mean Squared Error:
5.855643239890341
Variance score:
-0.0070741766242818915
```



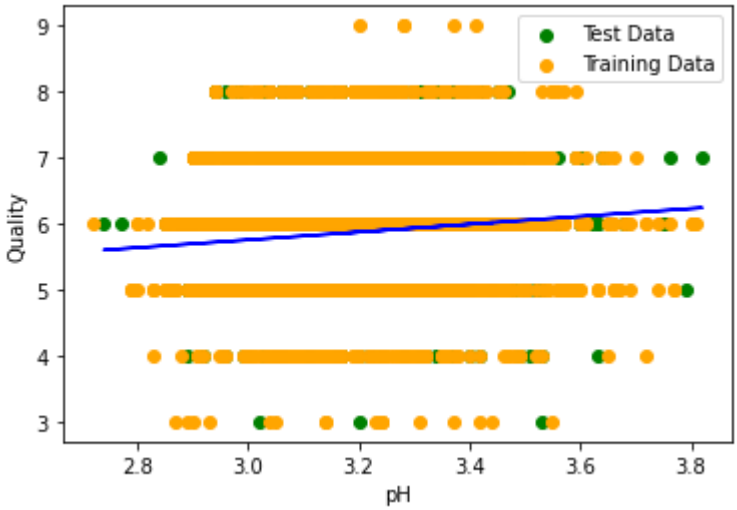

```
white data for feature: total sulfur dioxide
Coefficients:
 [-0.00371973]
Mean Squared Error:
5.854213334522278
Variance score:
0.017623638598817215
```



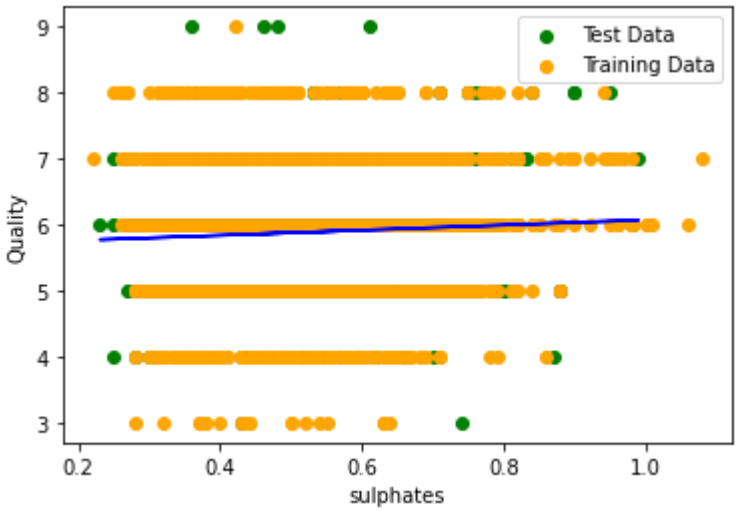

```
white data for feature: density
Coefficients:
 [-92.61802882]
Mean Squared Error:
5.8728488114914486
Variance score:
0.07559740044715335
```



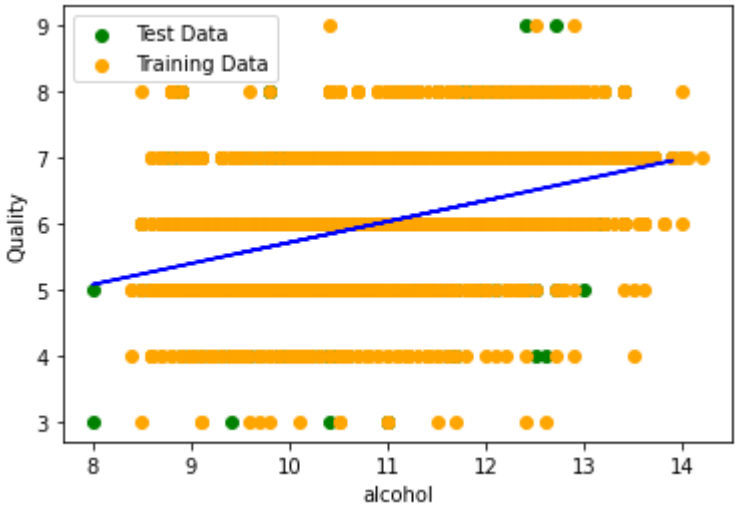

```
white data for feature: pH
Coefficients:
 [0.59281523]
Mean Squared Error:
5.8986332584467
Variance score:
0.0031658934299807484
```




```
white data for feature: sulphates
Coefficients:
 [0.38838912]
Mean Squared Error:
5.884710856114476
Variance score:
0.003320831163004634
```

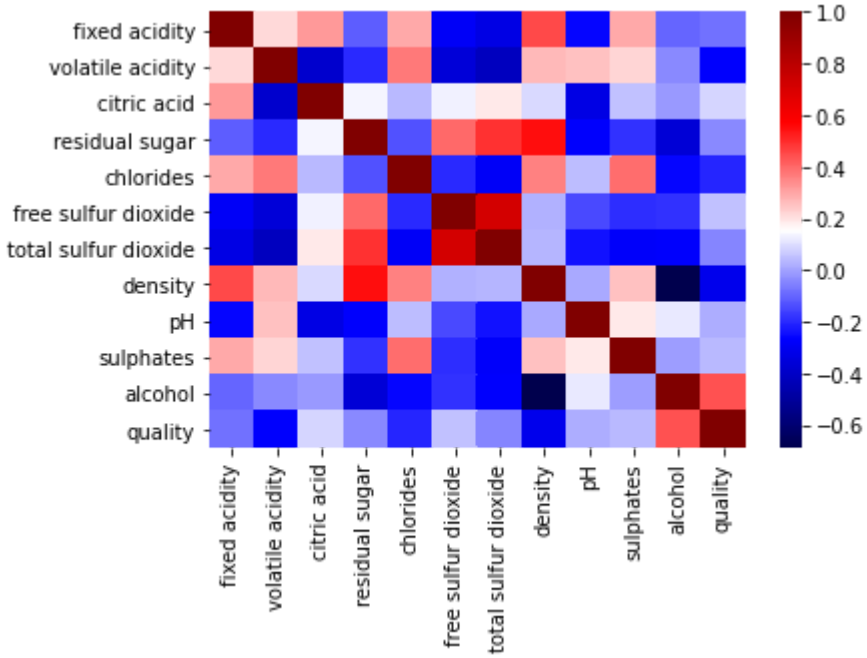



```
white data for feature: alcohol
Coefficients:
 [0.31830327]
Mean Squared Error:
5.864094147588735
Variance score:
0.16814724106462986
```



Check our regression models against correlation values

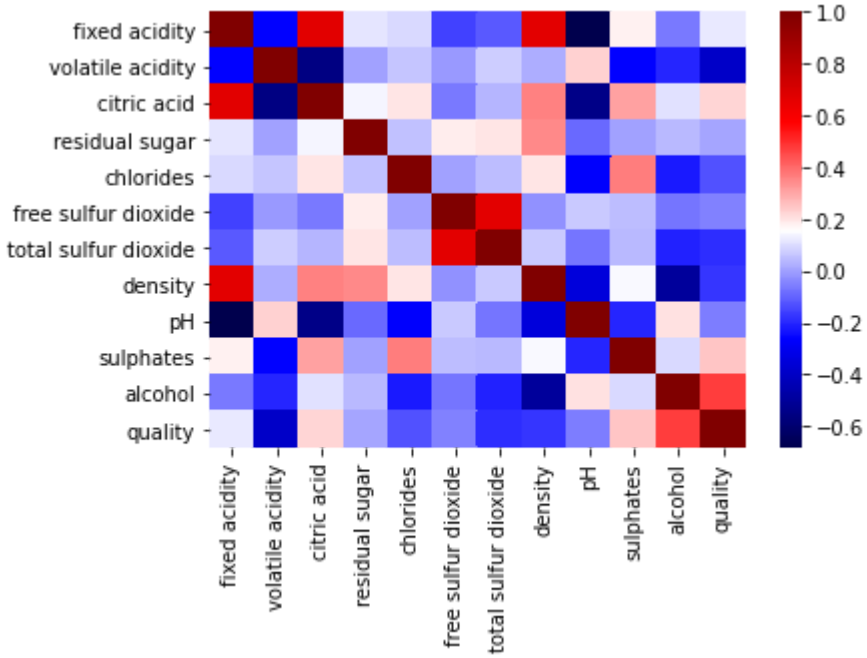
```
Correlation values for red and white wine data
alcohol                0.444319
citric acid            0.085532
free sulfur dioxide    0.055463
sulphates              0.038485
pH                     0.019506
residual sugar        -0.036980
total sulfur dioxide  -0.041385
fixed acidity         -0.076743
chlorides             -0.200666
volatile acidity      -0.265699
density               -0.305858
Name: quality, dtype: float64
<AxesSubplot:>
```



The 3 strongest correlations (alcohol, density, and volatile acidity) match the 3 regression models with the highest variance, which is what we expect.

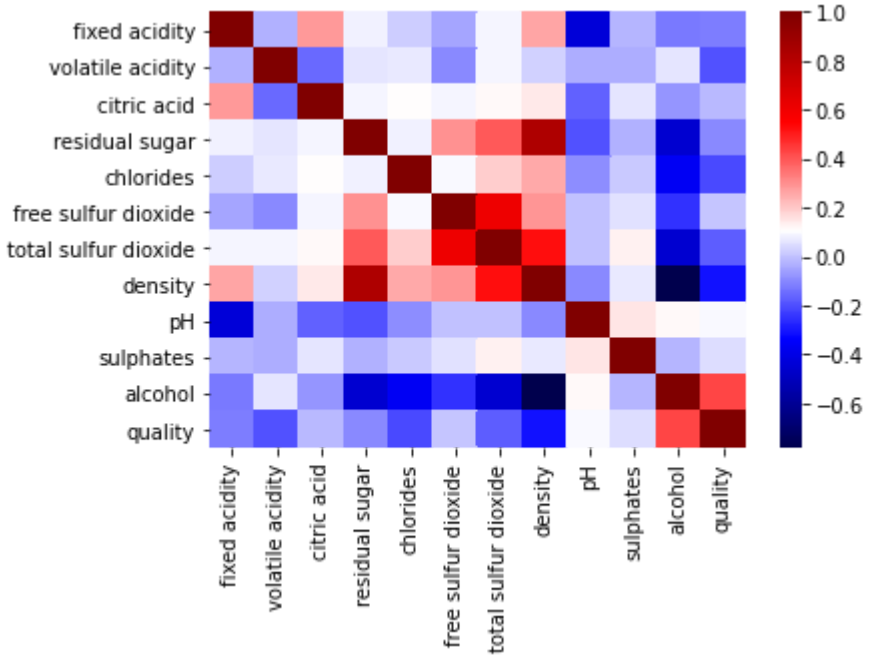
```
Correlation values for red wine data
alcohol                0.476166
sulphates              0.251397
```

```
citric acid             0.226373
fixed acidity           0.124052
residual sugar          0.013732
free sulfur dioxide    -0.050656
pH                     -0.057731
chlorides              -0.128907
density                -0.174919
total sulfur dioxide   -0.185100
volatile acidity       -0.390558
Name: quality, dtype: float64
<AxesSubplot:>
```



The 3 strongest correlations (alcohol, volatile acidity, and sulphate) match the 3 regression models with the highest variance, which is what we expect.

```
Correlation values for white wine data
alcohol                 0.435575
pH                      0.099427
sulphates               0.053678
free sulfur dioxide     0.008158
citric acid            -0.009209
residual sugar         -0.097577
fixed acidity          -0.113663
total sulfur dioxide   -0.174737
volatile acidity       -0.194723
chlorides              -0.209934
density                -0.307123
Name: quality, dtype: float64
<AxesSubplot:>
```



The 3 strongest correlations (alcohol, density, and chlorides) match the 3 regression models with the highest variance, which is what we expect.

Next, we generate regression models using all features. First for Red and white wine, then red only, then white only.

```
Regression model using all features for red and white wine
Coefficients:
 [ 7.40485544e-02 -1.35299503e+00 -1.39055977e-01  4.70848541e-02
 -2.54216254e-01  5.30344593e-03 -2.29965177e-03 -5.86556615e+01
  4.59673266e-01  7.47131282e-01  2.69596447e-01]
Mean Squared Error:
5.821900999353147
Variance score:
0.2971152744642864
```

The single feature regression model using alcohol had the highest variance, at ~.20. In this case, using all features produces a higher variance than any one feature independently

```
Regression model using all features for wine
Coefficients:
 [ 0.0123794  -1.12258419 -0.31268554  0.01244403 -2.19051723  0.00540102
 -0.00332445  2.71279013 -0.60982308  0.86327179  0.30736636]
Mean Squared Error:
5.614095820732772
Variance score:
0.3412702325087531
```

The single feature regression model using alcohol had the highest variance, at ~.25. In this case, using all features produces a higher variance than any one feature independently

```
Regression model using all features for white wine
Coefficients:
 [ 6.23610520e-02 -1.93810886e+00  2.06387546e-02  7.77341511e-02
 -4.82175601e-01  3.59879703e-03 -4.56089231e-04 -1.36688384e+02
  6.09359093e-01  5.62017332e-01  2.11238594e-01]
Mean Squared Error:
5.847296462516752
Variance score:
0.2722397371250632
```

The single feature regression model using alcohol had the highest variance, at ~.22. In this case, using all features produces a higher variance than any one feature independently

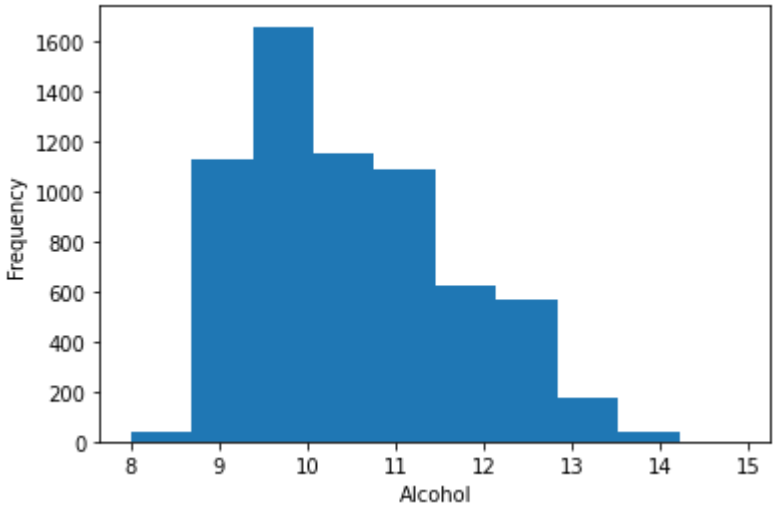
Next, we will attempt to use a combination of features to create a regression models whose variances are higher than the regression models which use all features. First for red and white wine, then red only, then white only. This could be computationally expensive. To decrease the amount of work, we will only consider features where abs(correlation) >= .1.

```
Regression models using high correlation features for red and white wine
Coefficients:
 [ 6.05361523e-02 -1.36296363e+00 -1.60053517e-01  4.32955144e-02
 -3.09493968e-01  5.17152834e-03 -2.43895259e-03 -5.19066281e+01
  3.84340744e-01  7.14072208e-01  2.66870371e-01]
Mean Squared Error:
5.827227816983976
Variance score:
0.3033883131435452
Features used: ()

Regression models using high correlation features for red wine
Coefficients:
 [-1.40994964e-02 -1.08753500e+00 -2.43514267e-03  2.52543198e-03
 -1.88918899e+00  3.97603436e-03 -3.12152461e-03  1.16410468e+01
 -4.72188318e-01  8.13372709e-01  2.83876598e-01]
Mean Squared Error:
5.68927170491286
Variance score:
0.4386011843260166
Features used: ()

Regression models using high correlation features for white wine
Coefficients:
 [ 9.88198343e-02 -1.76213148e+00  5.82724324e-02  9.31361866e-02
  2.20703705e-01  3.55885291e-03 -2.55447809e-04 -1.87132066e+02
  7.89301008e-01  6.44514927e-01  1.52484929e-01]
Mean Squared Error:
5.885992922052552
Variance score:
0.2775941766226955
Features used: ()
```

Discussion

It would seem from our analysis that alcohol content is the most important predictor for wine quality. Let's look into that a bit more.

```
Text(0.5, 0, 'Alcohol')
```

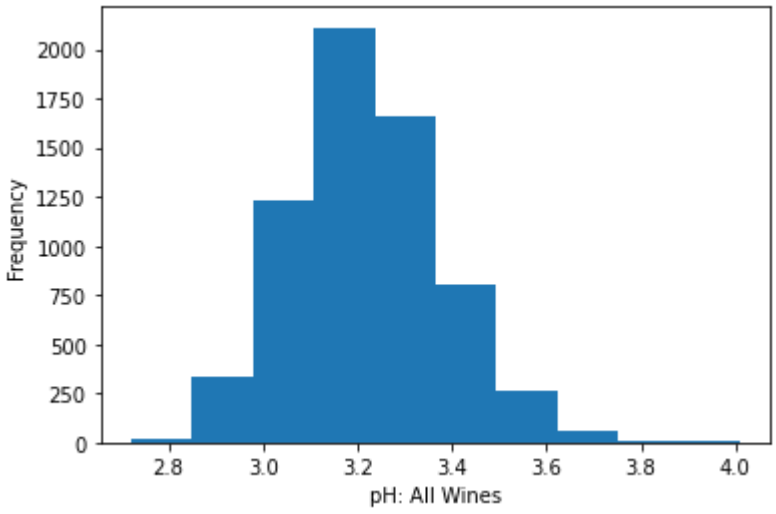



```
Mean quality score for wines with alcohol content greater than 13%
6.688
Mean quality score for all wines
5.818377712790519
Data for wines with alcohol content higher than 14%
```

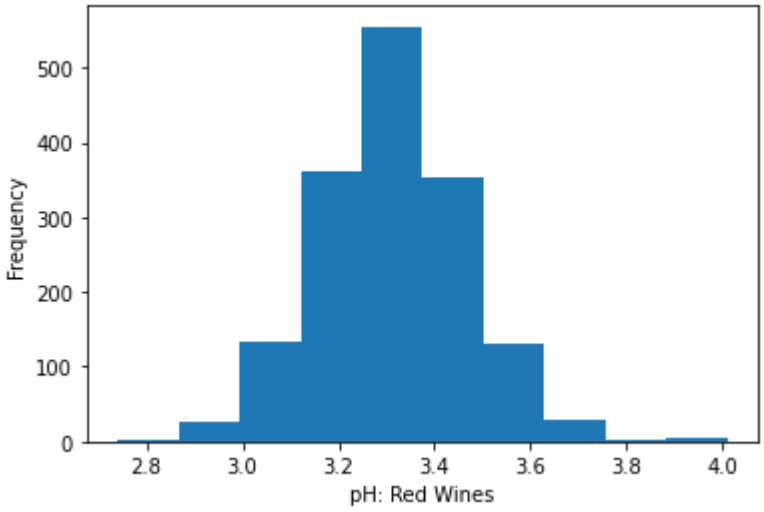
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
3918	6.4	0.35	0.28	1.6	0.037	31.0	113.0	0.98779	3.12	0.40	14.20	7
4503	5.8	0.61	0.01	8.4	0.041	31.0	104.0	0.99090	3.26	0.72	14.05	7
652	15.9	0.36	0.65	7.5	0.096	22.0	71.0	0.99760	2.98	0.84	14.90	5

It is interesting that pH did not come up as relevant in any of our models. While this analysis is not meant to be an exercise in chemistry, that does seem odd. It is also interesting that there's not much discernable difference between the pH of red wine and white wine.

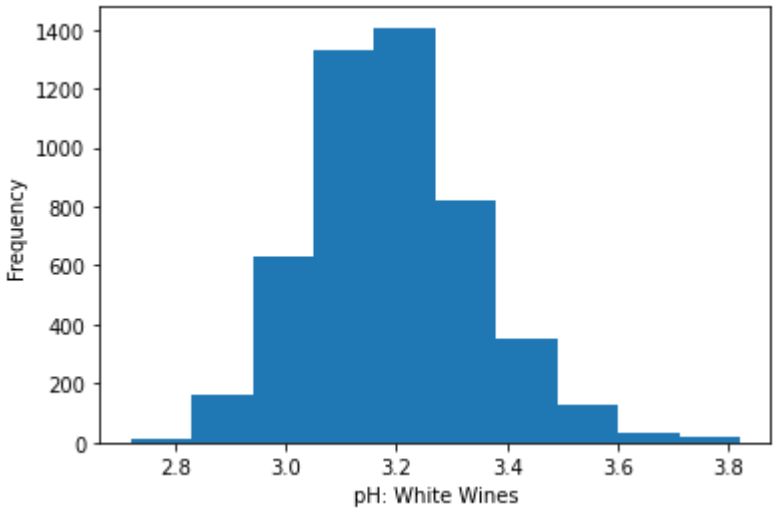
```
Average pH of all wine: 3.2185008465445586
```




```
Average pH of red wine: 3.3111131957473416
```

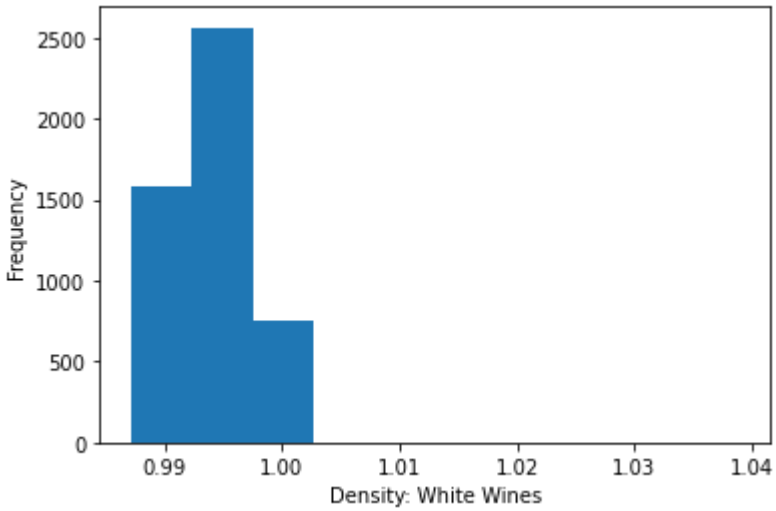



```
Average pH of white wine: 3.1882666394446715
```

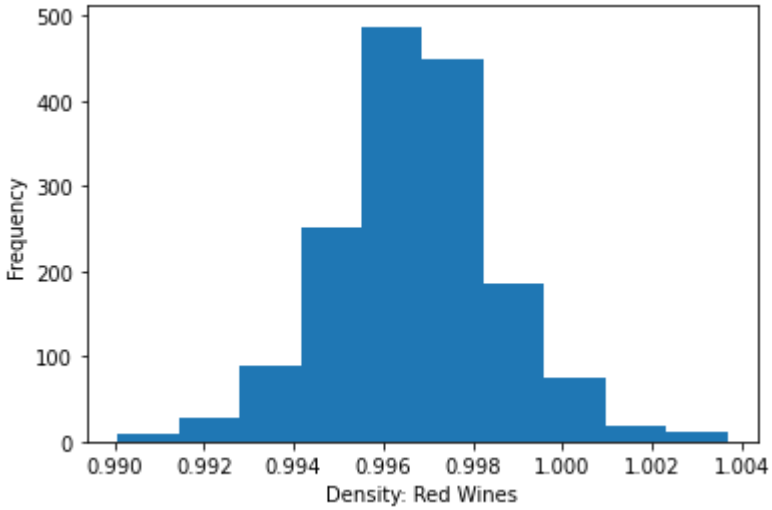


Density is negatively correlated with quality in red and white wines, but it seems more important in white than red. See coefficients in the regression models using all features

```
Text(0.5, 0, 'Density: White Wines')
```




```
Text(0.5, 0, 'Density: Red Wines')
```




```
Average quality of white wines with density >1
5.552631578947368

Average quality of white wines with density <.995
6.042954767328344
```