Classifying wine varieties by class
Project

Aleksandr Trohhatsov
IVSB32

ICY0006
Margarita Spitsakova

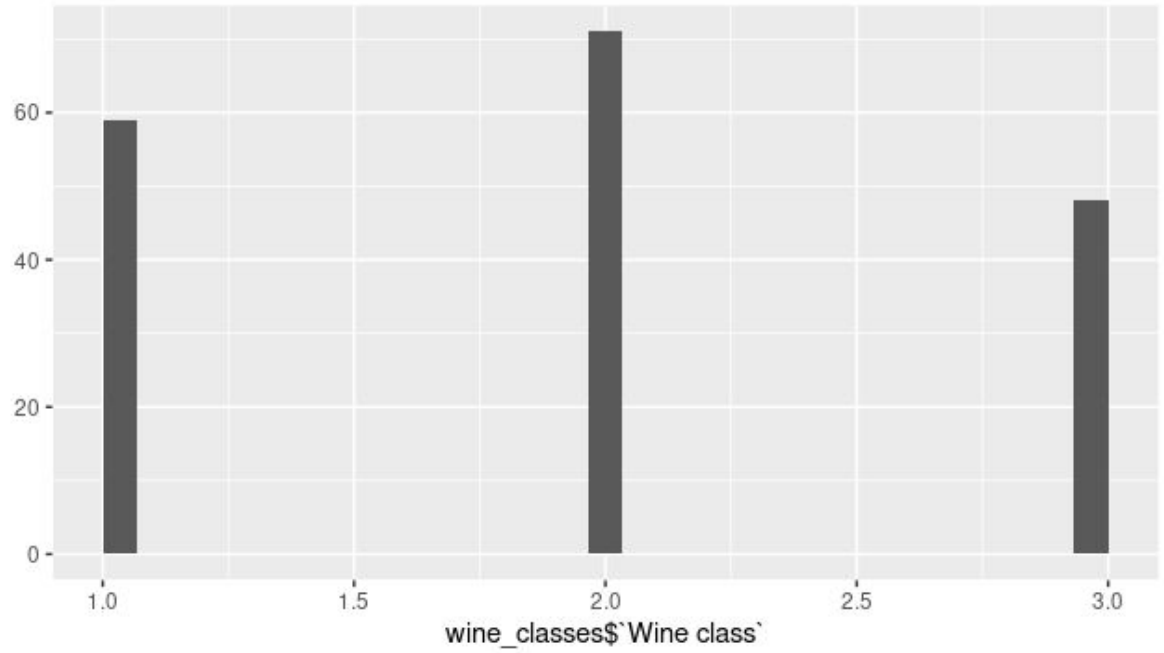
2020

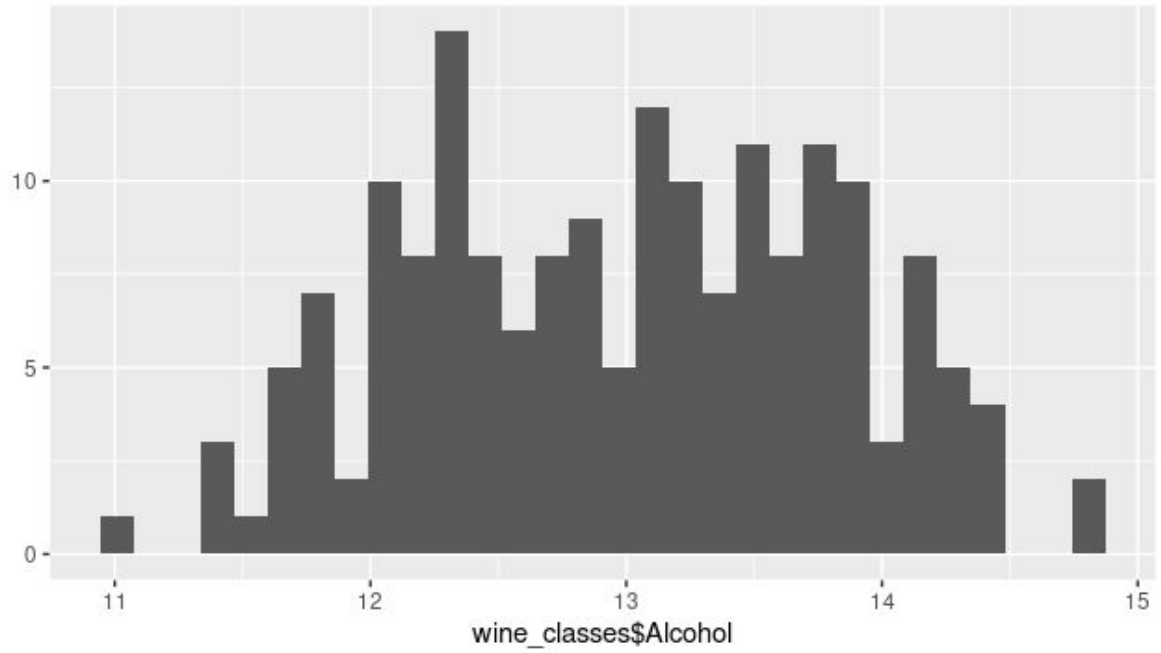
----------STAGE 1----------

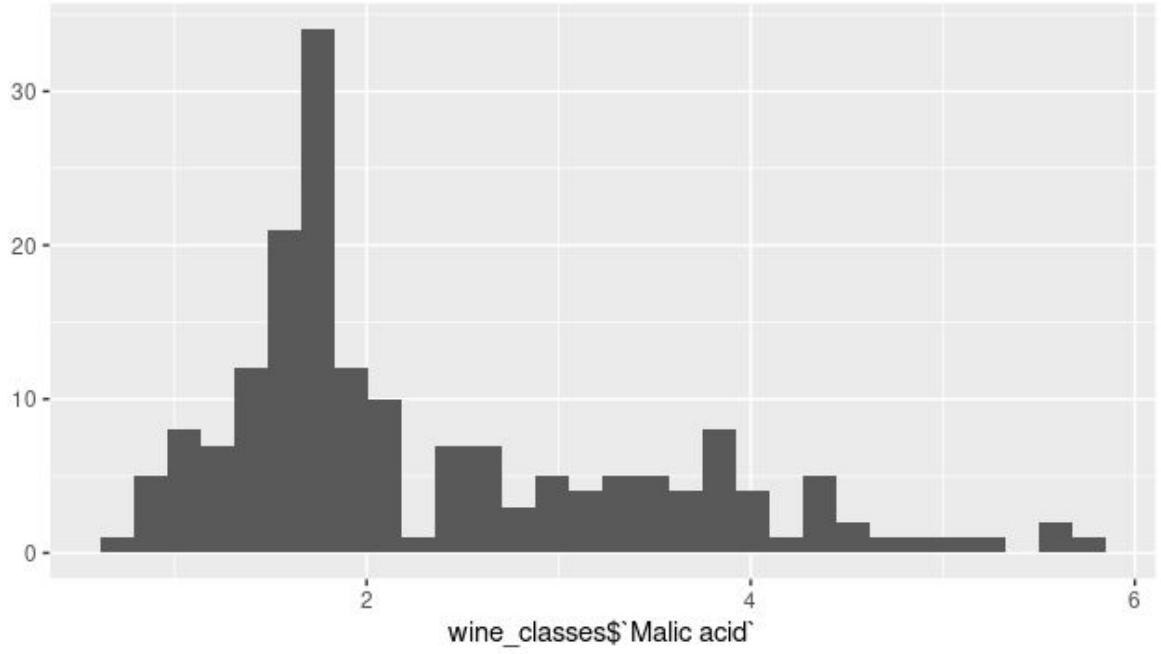
Source:
https://www.kaggle.com/brynja/wineuci
https://archive.ics.uci.edu/ml/datasets/wine

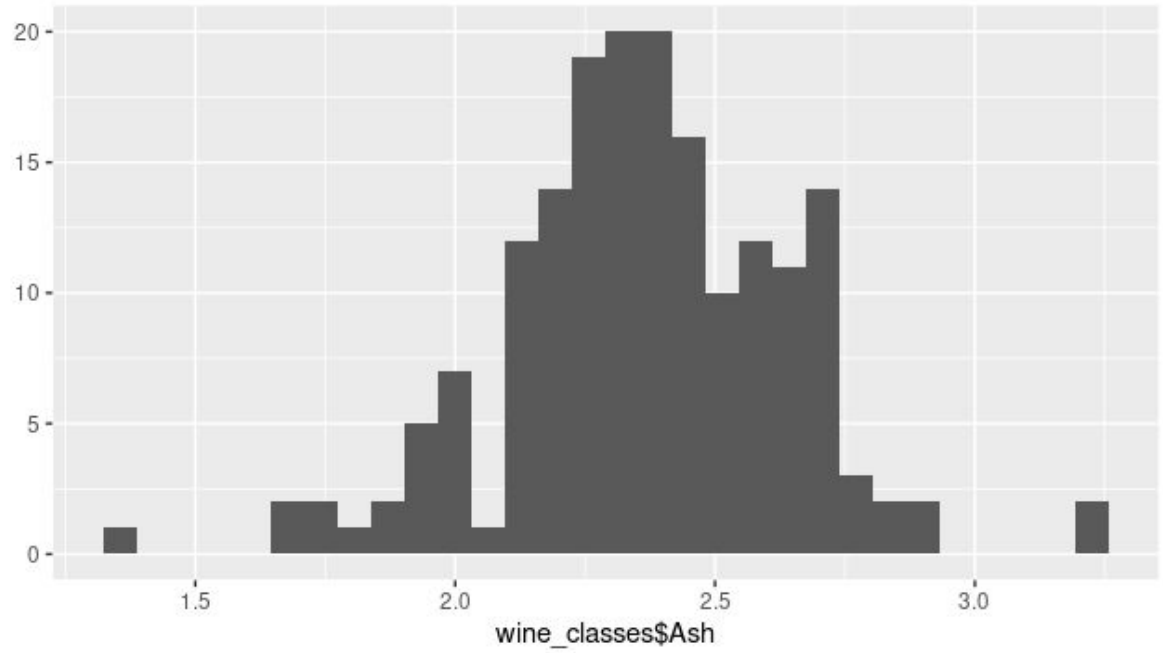
Variables:
1) Wine class
2) Alcohol
3) Malic acid
4) Ash
5) Alcalinity of ash
6) Magnesium
7) Total phenols
8) Flavanoids
9) Nonflavanoid phenols
10) Proanthocyanins
11)Color intensity
12)Hue
13)OD280/OD315 of diluted wines
14)Proline

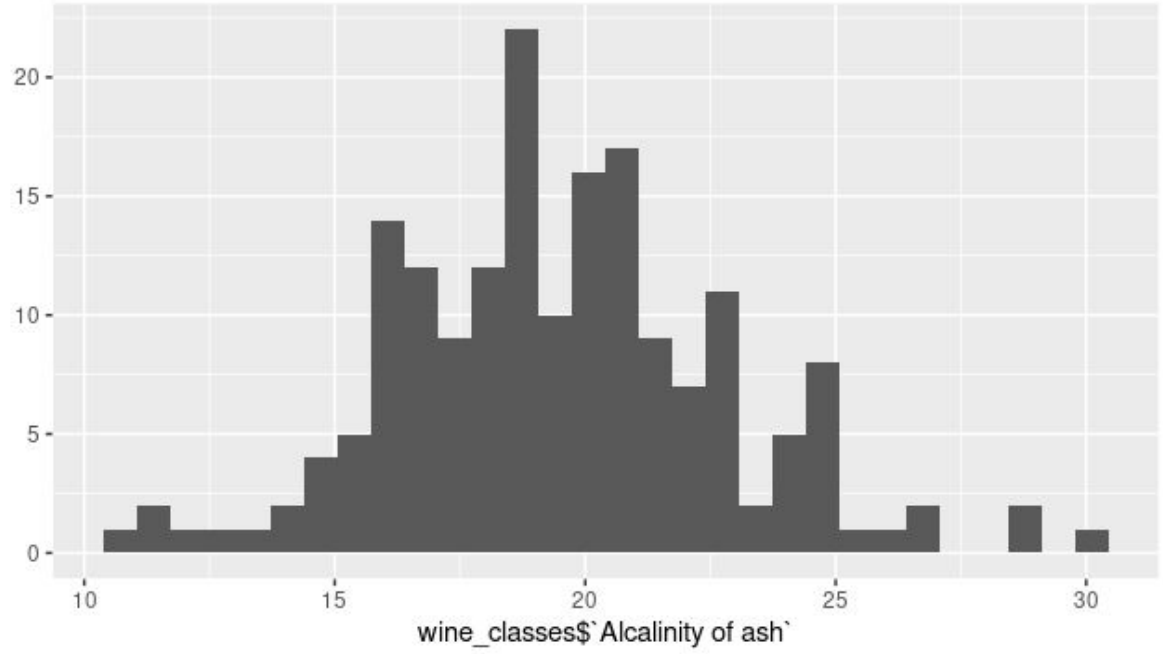
This dataset concerns wine qualities by class, which give the drink its specific properties. The initial CSV document did not have names for all the variables, they had to be acquired on the second link specified in sources and inserted into the CSV file.

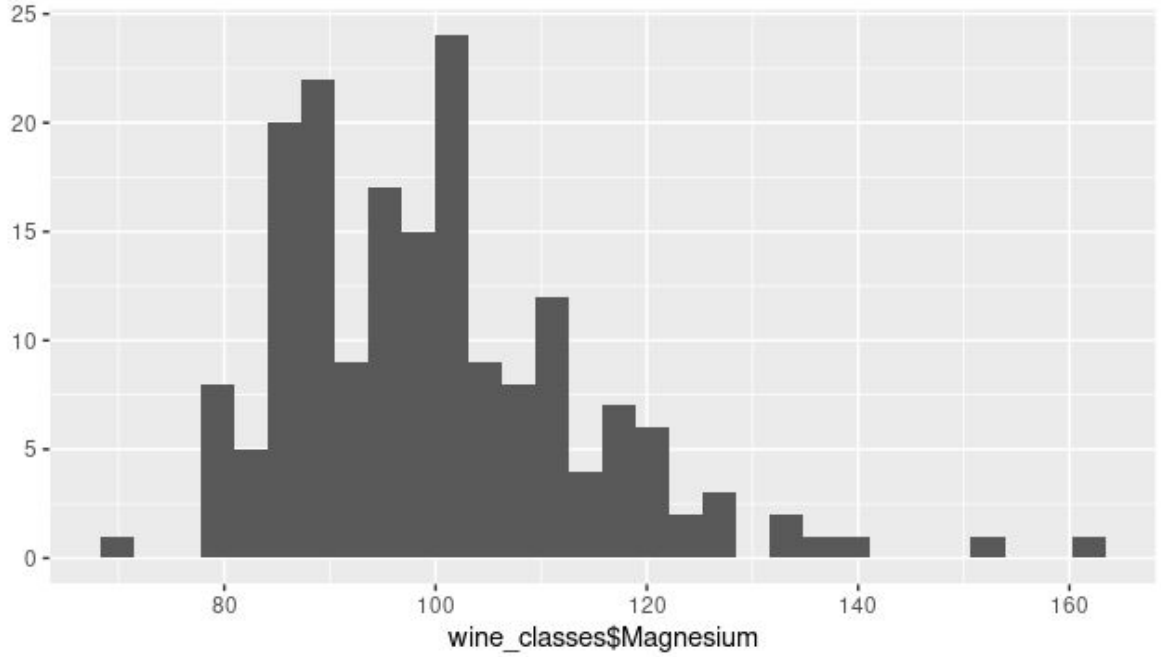

60
40
20
0
1.0
1.5
2.0
2.5
3.0
wine_classes$`Wine class`

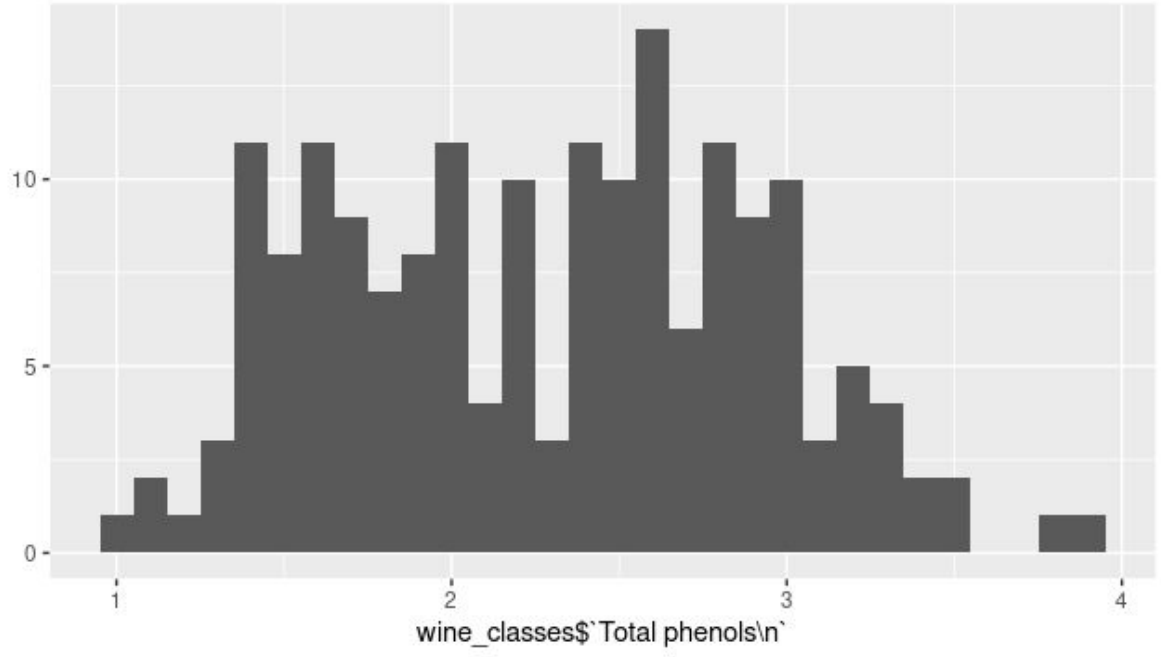

10
5
0
11
12
13
14
15
wine_classes$Alcohol

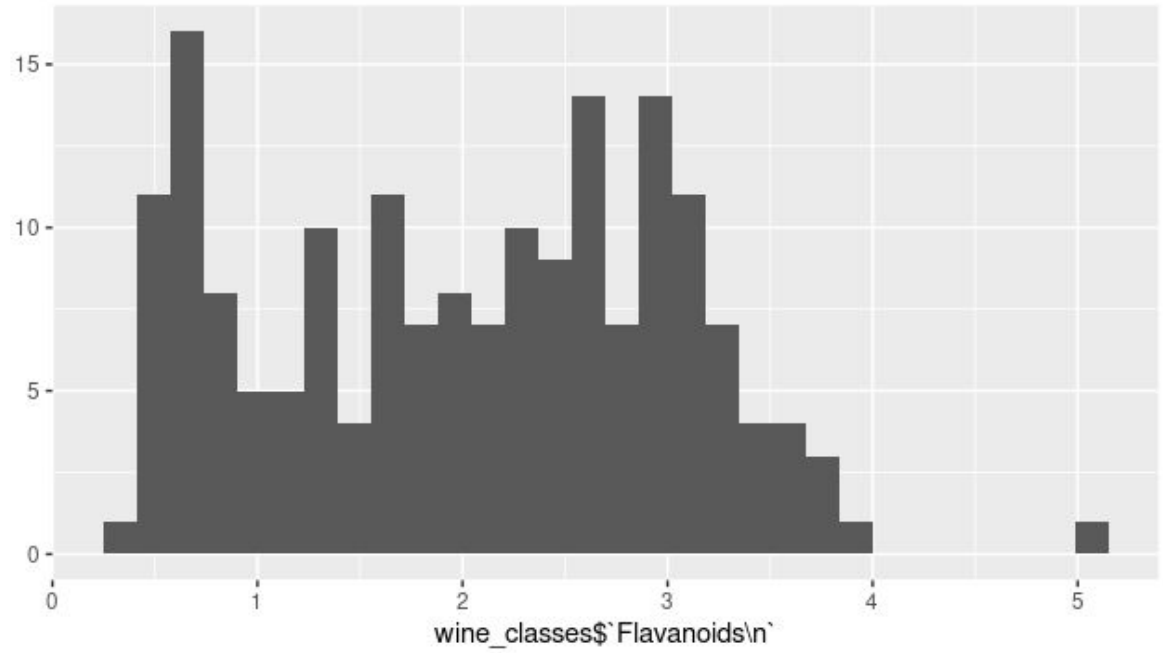

30
20
10
0
2
4
6
wine_classes$`Malic acid`

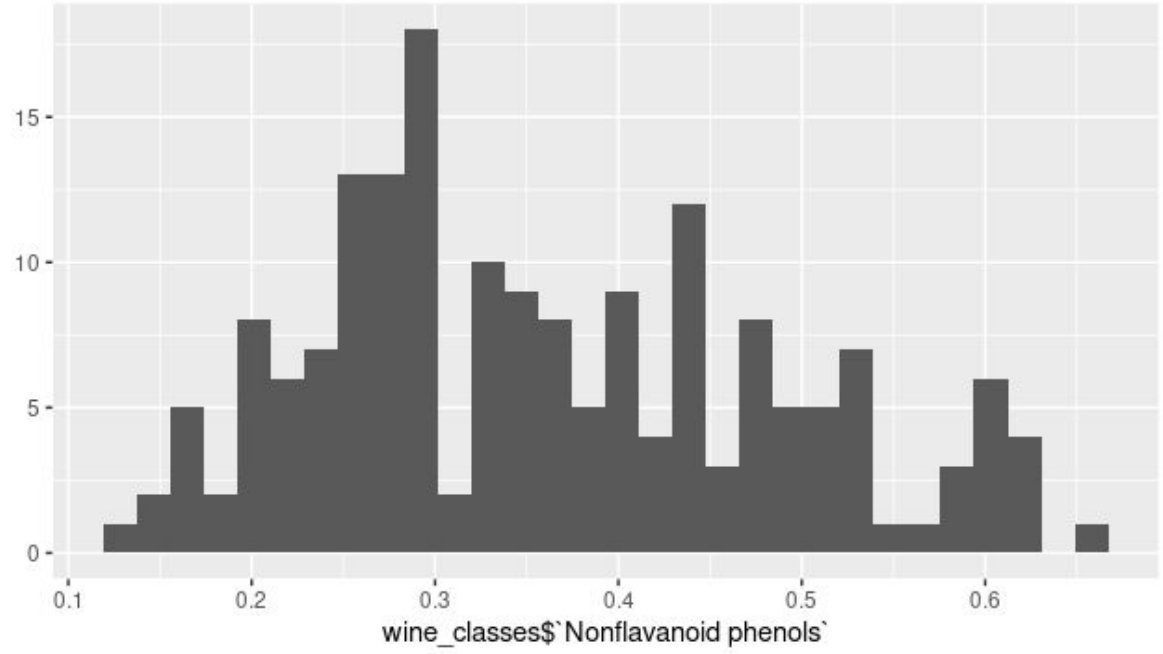

20
15
10
5
0
1.5
2.0
2.5
3.0
wine_classes$Ash

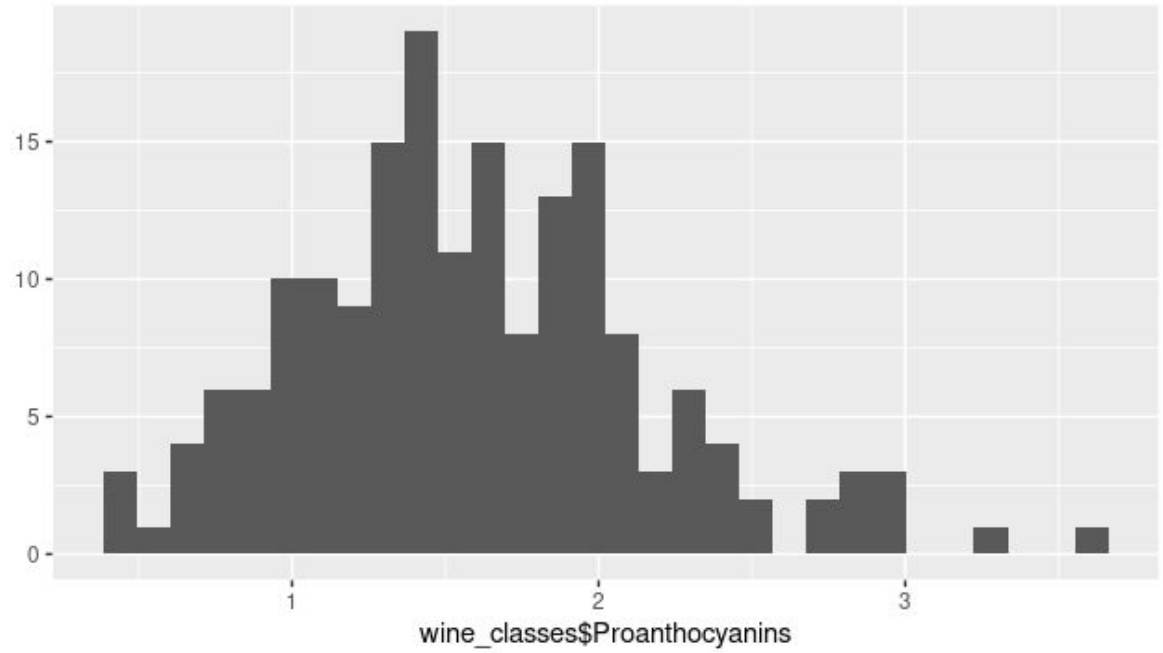

20
15
10
5
0
10
15
20
25
30
wine_classes$`Alcalinity of ash`

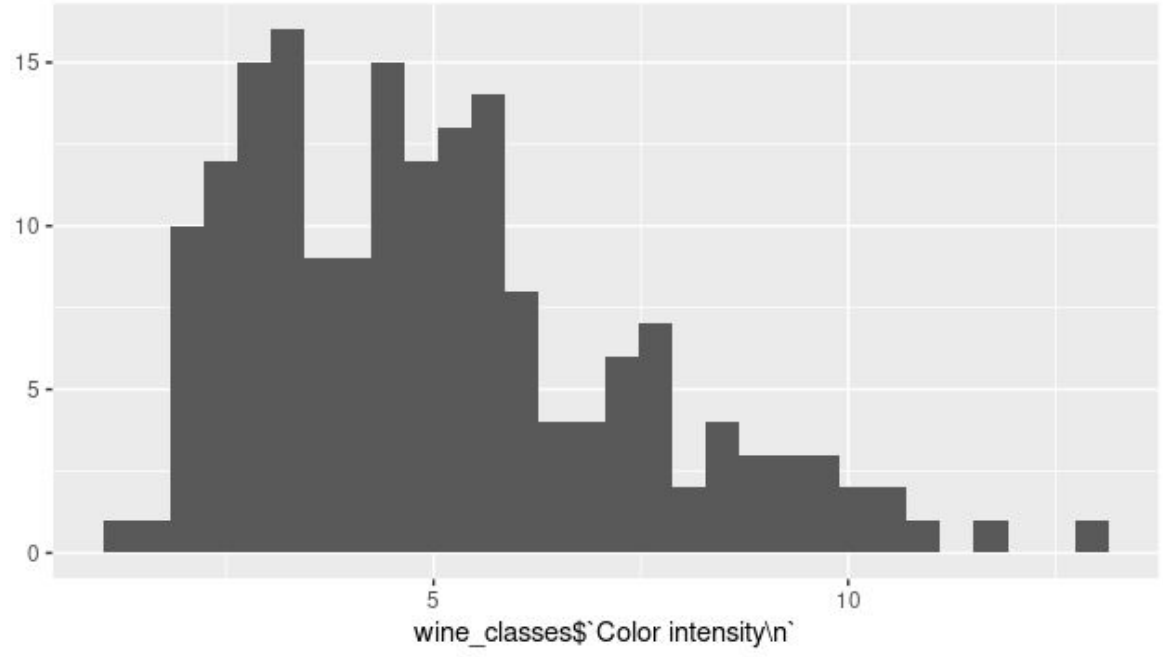

25
20
15
10
5
0
80
100
120
140
160
wine_classes$Magnesium

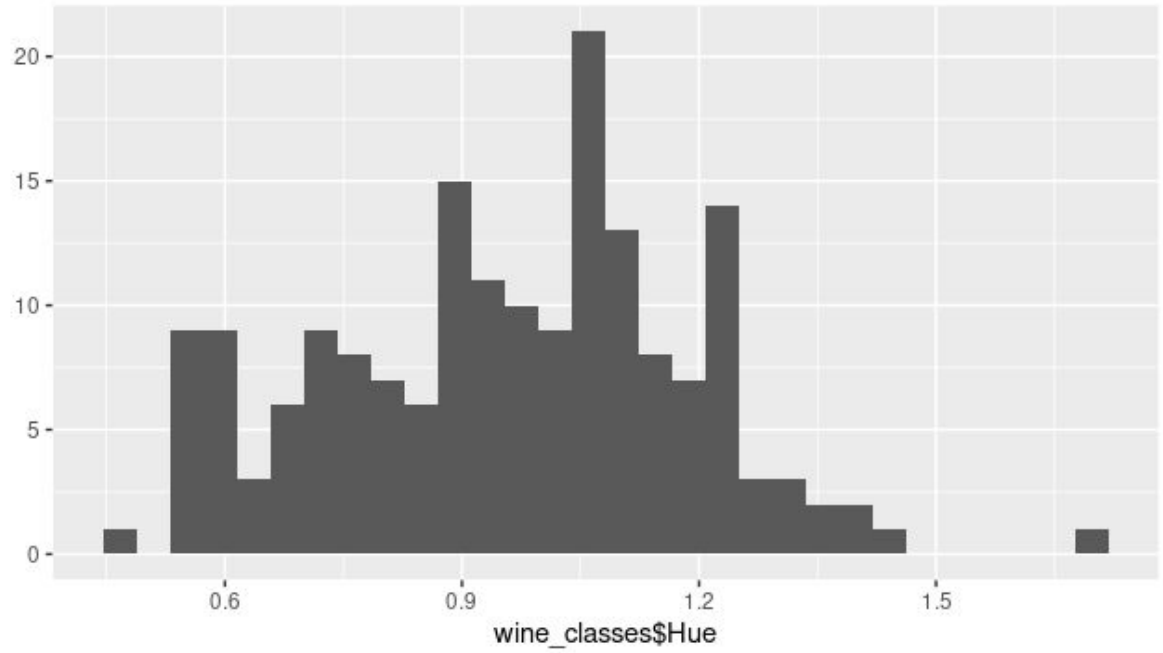

10
5
0
1
2
3
4
wine_classes$`Total phenols\n`

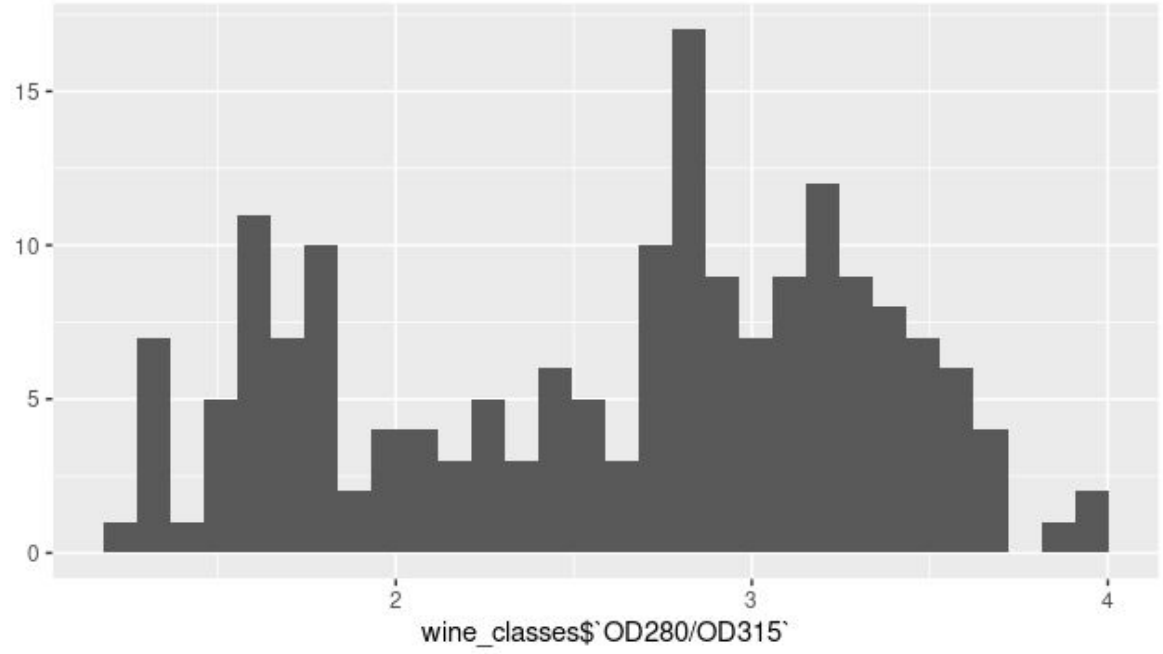

15
10
5
0
0
1
2
3
4
5
wine_classes$`Flavanoids\n`

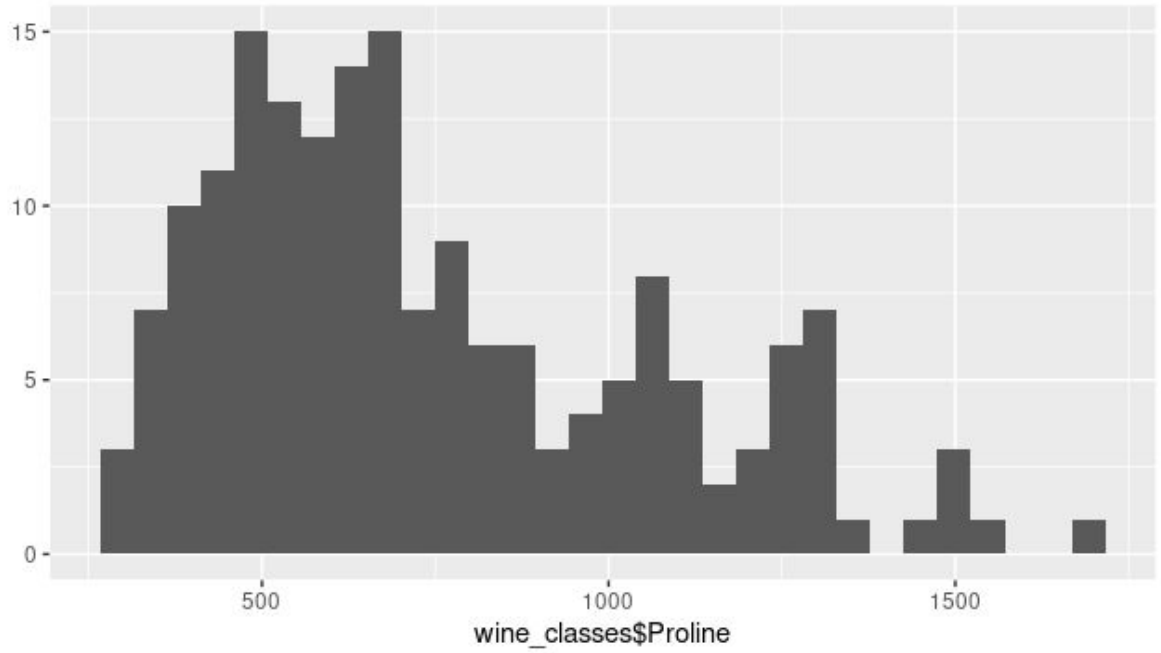

15
10
5
0
0.1
0.2
0.3
0.4
0.5
0.6
wine_classes$`Nonflavanoid phenols`


15
10
5
0
1
2
3
wine_classes$Proanthocyanins


15
10
5
0
5
10
wine_classes$`Color intensity\n`


20
15
10
5
0
0.6
0.9
1.2
1.5
wine_classes$Hue


15
10
5
0
2
3
4
wine_classes$`OD280/OD315`



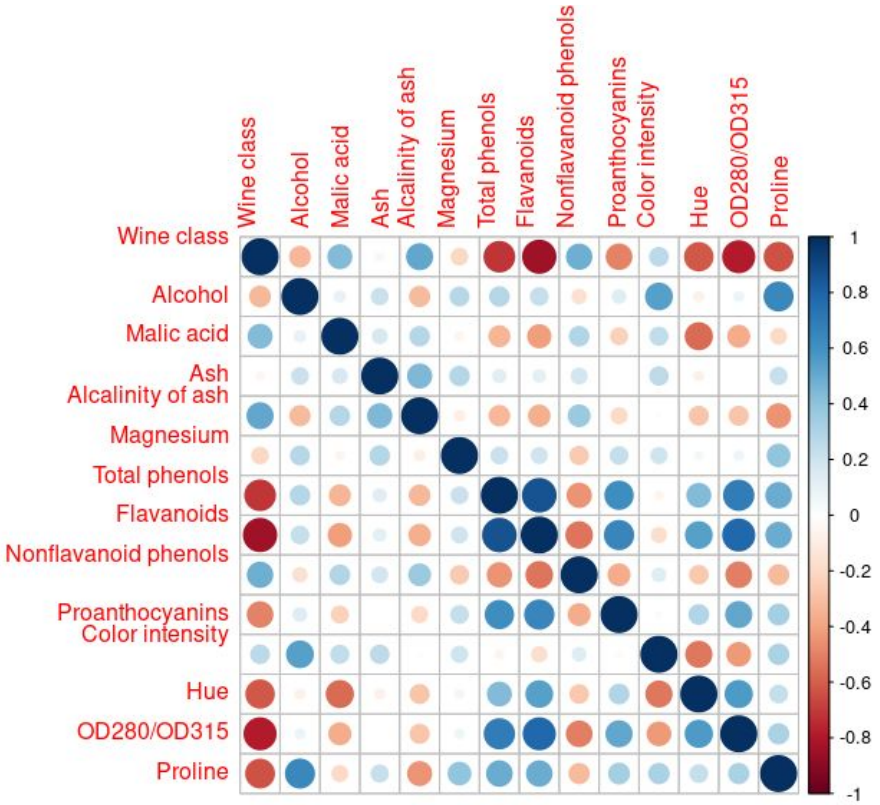
----------STAGE 2----------

Plot descriptions will be numbered based on column number in the previous stage:

1) Mode shows that the most popular wine class out of the ones we have here is class ‘2’. Geometric and harmonic mean both round out to 2, which corresponds to the mode. Frequency tables shows that class 2 is the most common, with class 1 in second place and class 3 finishing last.
2) Alcohol percentages are equal to roughly 13 in both Median and Mean. The maximum percentage is 14.83. Mode prints out two values: 12.37 and 13.05.
3) Malic acid percentage seems to vary widely between each wine, as the minimum and maximum value are 0.74 and 5.8 respectively. The frequency table shows widely varying values. Median and mean are equal to 1.865 and 2.336 respectively.
4) Ash percentage median and mean both roughly equal 2.4. Mode prints out two values: 2.28 and 2.3. Frequency again shows a large amount of different values.
5) Ash alkalinity seems to be concentrated at a value of either 20 or 21, the latter of which is shown in the median and mean values, as well as in the frequency values of 15 and 11 respectively.
6) Magnesium contents in wine tend to hover around 90-100, with the bear minimum being 70, the former being reaffirmed by median and mean values. The mode of the current dataset is 88.
7) Total phenol content is quite varied, with mode and mean equaling 2.2 and 2.1 respectively. The maximum value is 3.88.
8) Flavanoids all have differing values, with the mode being 2.65 with 4 occurrences. Variance is very low with a value of 0.59.
9) Non-Flavanoids have a lot of differing values, the majority of them being 0.26, 0.29 and 0.43. Median and mean are 0.34 and 0.36 respectively.

10) Proanthocyanins have a lot of differing values, majority of them being 1.87, 1.35 and 1.46. Median and mean are 1.555 and 1.591 respectively.
11) Color intensity is almost all different, as the mode returns 3 values, as the most common entries appear only 4 times. Those values are 2.6, 3.8, 4.6. Mean function returns a value of 5.058.
12) Hue varies in wine with every type of wine, with the mode being 1.04. Median and mean values are extremely similar, being 0.9650 and 0.9574.
13) OD280/OD315 have widely differing values, majority of them being seen only once or twice in the whole dataset. Mode function returns 2.87. Median and mean are 2.780 and 2.612 respectively.
14) Proline have widely differing values, majority of them being seen only once or twice in the whole dataset. Mode function returns 520 and 680. Median and mean are 673.5 and 746.9 respectively.

----------STAGE 3----------



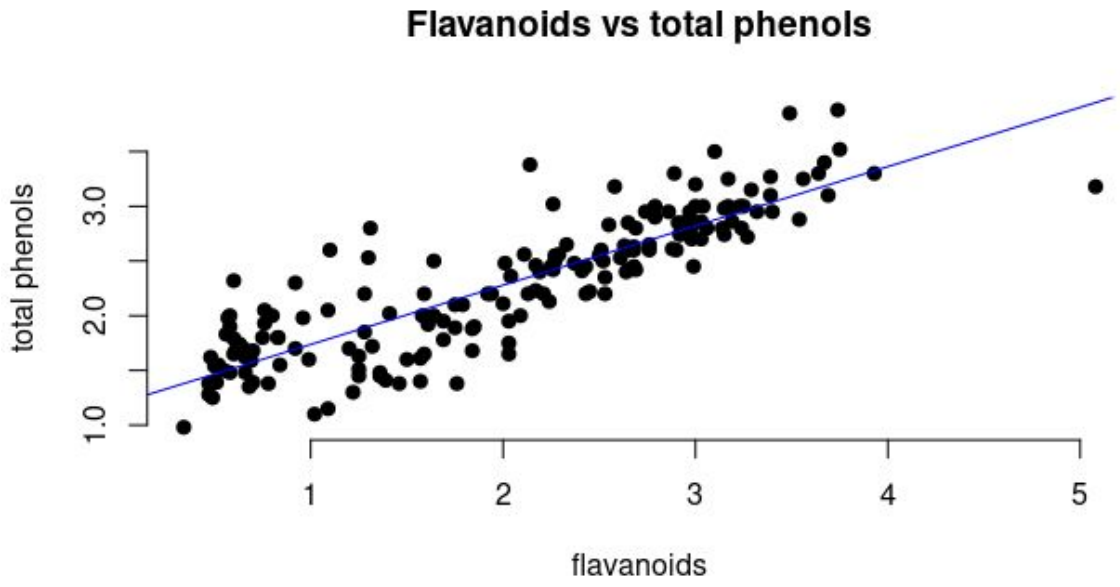
High correlation pairs:

- Flavanoids vs total phenols
- Flavanoids vs wine class
- Flavanoids vs hue
- wine class vs hue
- wine class vs OD280/OD315

Now I will describe what I see in each scatterplot:

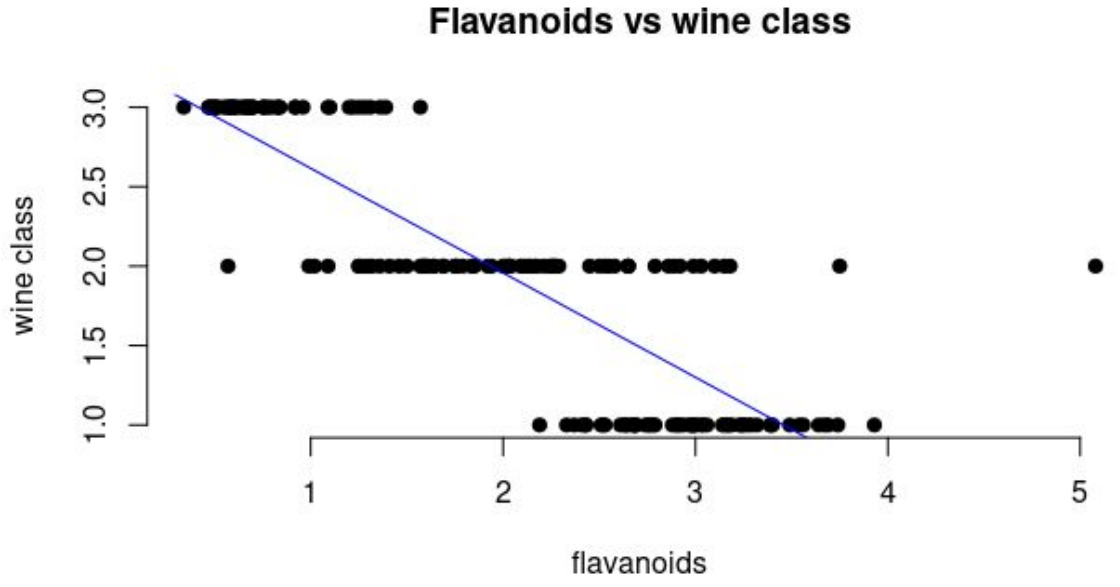
Flavanoids vs total phenols

The variables seem to closely connected, as they are spread quite equally throughout the length of the graph. Also, as flavanoids increase, total phenols increase as well



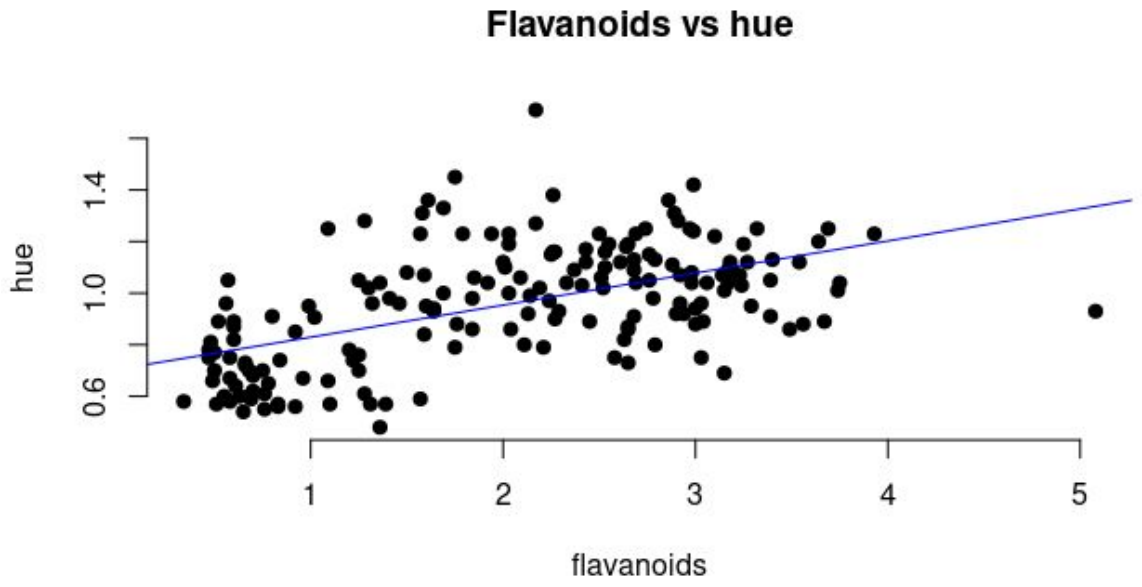
Flavanoids vs wine class

From the graph, we can see that class 1 wines have roughly 2 times the flavanoids that class 3 wines have. For class 2 wines, they flavanoid content varies widely, mostly the same levels as level 1 or 3 wines.



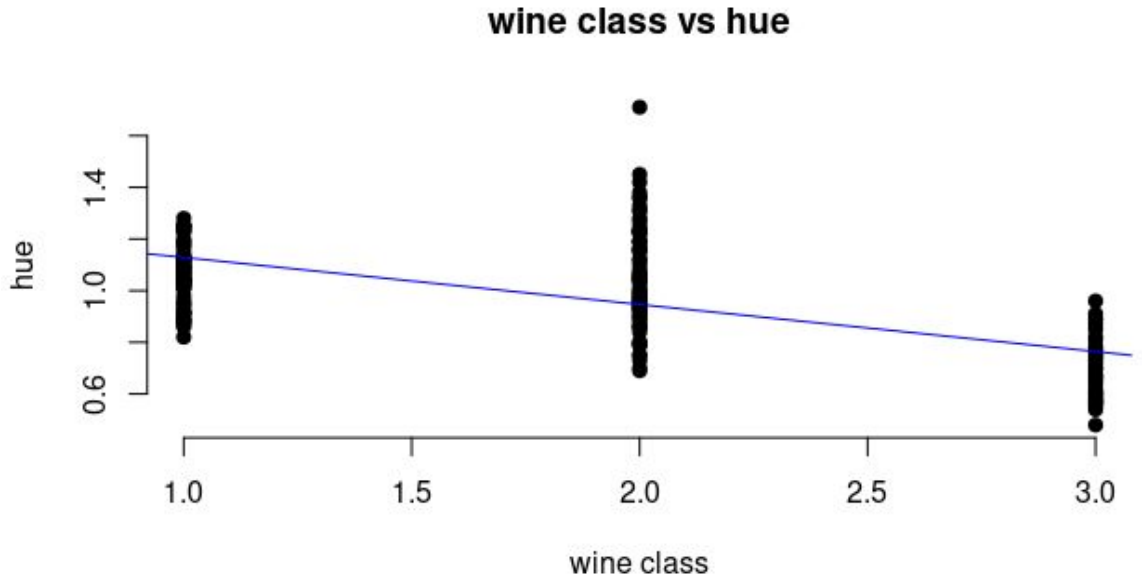
Flavanoids vs hue

Analyzing the plot, we can see that as the amount of flavanoids increases, the hue slightly also increases in general, although there are also exceptions.



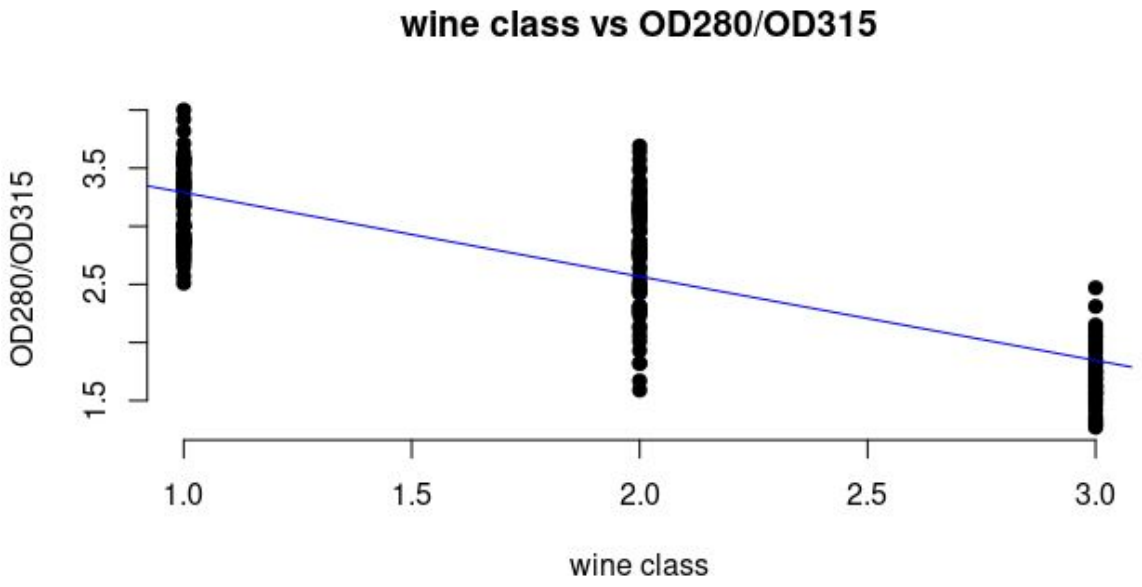
Wine class vs hue

And now, to contrast the previous plot, we can see that as you go from class 1 to 2 and class 2 to 3, the hue of the wine decreases. The hue of class 2 wine seems to be the highest in some cases.



Wine class vs OD280/OD315

Finally, in this graph, we can see that OD280/OD315 contents decrease as you go from class 1 to 2 and class 2 to 3. This time, The highest values can be attributed to class 1 wines, whereas class 3 wines almost lack these substances.



----------STAGE 4----------

I chose VIKINGLOTTO for my lottery. The lottery uses the Powerball method, where the person playing the game chooses 6 balls from a pool of 48 and then chooses 1 from a different pool of 6. The balls are then drawn each Wednesday. If the player gets at least 1 of the 6 main balls correct he wins a prize. In addition, if the player gets the bonus number, aka powerball, right, he gets a better prize. To win the jackpot, you have to get all the balls correct, including the powerball.

The computation is quite simple once you understand it. We have to pick out 6 out of 48 main balls, while at the same time getting some of them right and while having a result for the powerball. And we need to find all of this out considering we have 48 total balls. So we need to multiply them so the formula is thus:

*(chance of getting/not getting the powerball right * amount of balls gotten right * amount of balls picked out)/all of the balls*

Chances for all the wins (in order of regular-powerball(1-0, 1-1, 2-0, 2-1, 3-0 etc)):
0.3466028, 0.06932055, 0.1140141, 0.02280281, 0.01559167, 0.003118333, 0.0008770313, 0.0001754063, 1.711281e-05.
After computing all of the values, it is clear that after the win where you have 2 of 6 balls and no powerball, where the chance is already only 11.4%, the chance of you winning something becomes so low that it is almost impossible to win.

----------STAGE 5----------

The training of the train set and its deployment with the test set has been a success. Once you use the summary() function you can see that the values in the column change. Also when running a plot function, we can see that the progression goes in the same direction, albeit at a bit of a different angle due to a value being a bit off scale. The split was done with the 80/20 approach in mind. Only linear regression has been tested.

Coefficient of determination is roughly equal to 0.62, which is above average. RMSE returns 0.9191024, which is really good for a linear model. MAE returns 0.7624571, which is again above average accuracy. Minimum-maximum accuracy computation produces a similar result: 0.7424423, which is also a good indicator. MAPE equals roughly 0.385, which is an above average result as well. All together, they paint a picture of above average performance.