Red Wine Quality

Made by Maja
Updated 20/7/2021
Made in R/R Studio

Hello, all!

In this R notebook, I will show you all the details about statistical analysis I performed based on a dataset called Red Wine Quality.

Citation info: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Note – some boxplots and summaries are a bit different than it sounds in written. When you knit out the code, it only runs once (cleaning the outliers), whereas I've cleaned outliers several times, and hence the different numbers. The ones that are in written are the right ones (only 2 boxplots are like that, the other ones are fine).

The variables are as in follows:

1. **Fixed acidity** – most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2. **Volatile acidity** – the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.
3. **Citric acid** – found in small quantities, citric acid can add „freshness" and flavor to wines
4. **Residual sugar** – the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 g/L and wines with greater than 45g/L are considered sweet
5. **Chlorides** – the amount of salt in the wine
6. **Free sulfur** dioxide – the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. **Total sulfur dioxide** - amount of free and bound forms of So2; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. **Density** - the density of water is close to that of water depending on the percent alcohol and sugar content
9. **ph** - describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. **Sulphates** - a wine additive which can contribute to sulfur dioxide gas (So2) levels, which acts as an antimicrobial and antioxidant
11. **Alcohol** - the percent alcohol content of the wine
12. **Quality** - output variable (based on sensory data, score between 0 and 10)

Summaries of the variables

In this chapter, I am going to create summaries of the variables in the sample. The sample itself has 1599 observations or wines, and 12 variables which have been presented and explained above.

```
boxplot (dfrm$fixed.acidity, plot = FALSE)$out # this is the formula with which you can see the actual outliers i
n the variable fixed acidity. I plotted false because there is no need to see the boxplot again.
```

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8
## [16] 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9
## [31] 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9
## [46] 13.3 12.9 12.6 12.6
```

```
outliers = boxplot (dfrm$fixed.acidity, plot = FALSE)$out # we are going to attach the formula above to a varia
ble named outliers.
dfrm = dfrm [-which(dfrm$fixed.acidity %in% outliers),] # with this formula, I removed the outliers from the da
ta
summary (dfrm$fixed.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.600   7.100   7.900   8.163   9.100  12.300
```

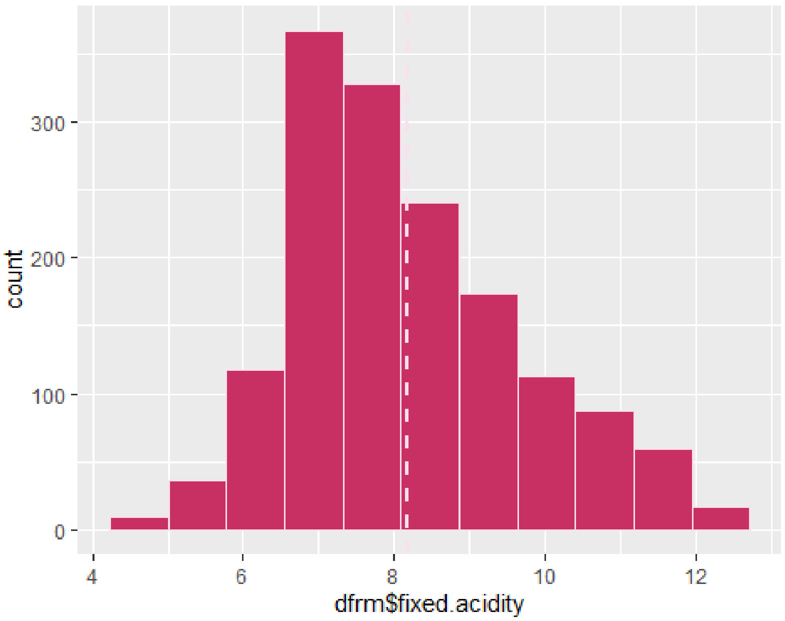
```
sd (dfrm$fixed.acidity)
```

```
## [1] 1.513582
```

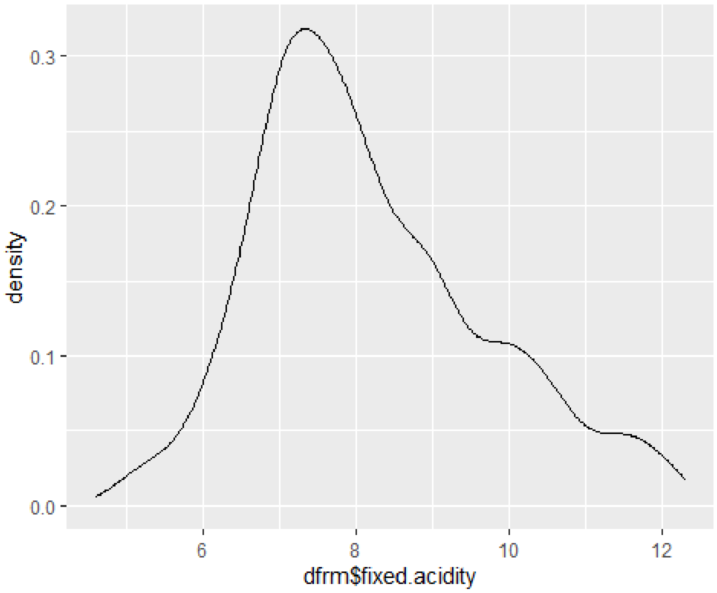
```
IQR (dfrm$fixed.acidity)
```

```
## [1] 2
```

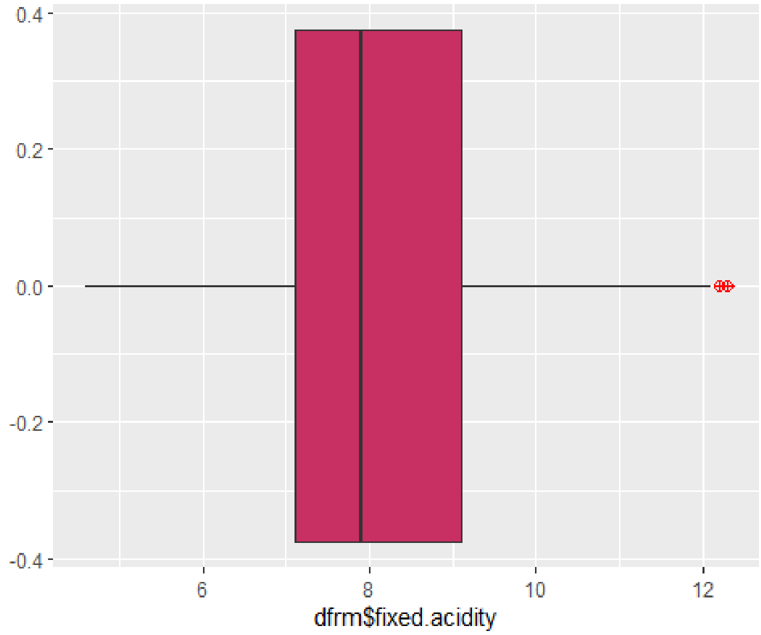
```
binsize = diff (range(dfrm$fixed.acidity))/10 # code for bin number on histogram
ggplot (dfrm, aes(x = dfrm$fixed.acidity)) + geom_histogram (binwidth = binsize, fill = "#c82f63",colour = "#f
9e0ec")+ geom_vline (aes(xintercept = mean(dfrm$fixed.acidity)), colour = "#f9e0ec", linetype = "dashed", s
ize = 1)
```



```
ggplot (dfrm, aes (x = dfrm$fixed.acidity)) + geom_density()
```



```
ggplot (dfrm, aes(x = dfrm$fixed.acidity)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape = 10, outlier.size = 2)
```



Typical fixed acidity level of red wines in the dataset is 8.163, with range of 4.6 to 12.3, which is a pretty wide range. The total variability is 1.5, which is low. You can see on the histogram and density plot that the distribution is positively skewed, and after removing outliers - it looks like a pretty normal distribution (still skewed, though). Boxplot doesn't show any outliers.

```
boxplot (dfrm$volatile.acidity, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035
## [13] 1.025 1.115 1.010 1.020 1.020 1.580 1.180 1.040
```

```
outliers = boxplot (dfrm$volatile.acidity, plot = FALSE)$out # we are going to attach the formula above to a variable named outliers.
dfrm = dfrm [-which (dfrm$volatile.acidity %in% outliers),] # with this formula, I removed the outliers from the data
summary (dfrm$volatile.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1200  0.3900  0.5200  0.5225  0.6350  1.0050
```

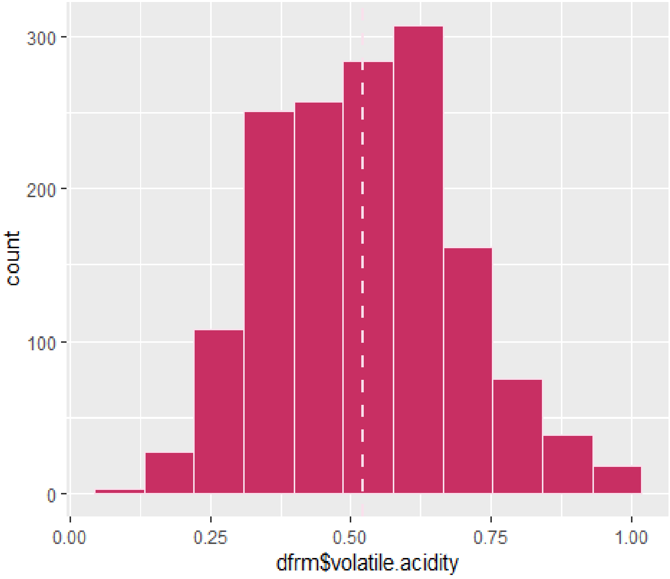
```
sd(dfrm$volatile.acidity)
```

```
## [1] 0.1665945
```

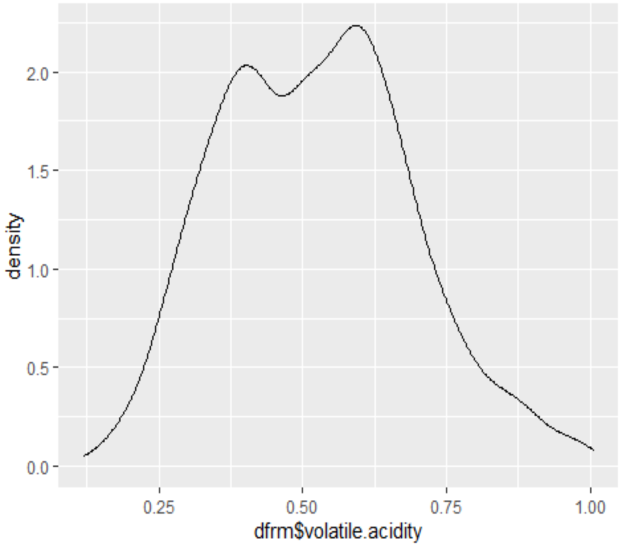
```
IQR(dfrm$volatile.acidity)
```

```
## [1] 0.245
```

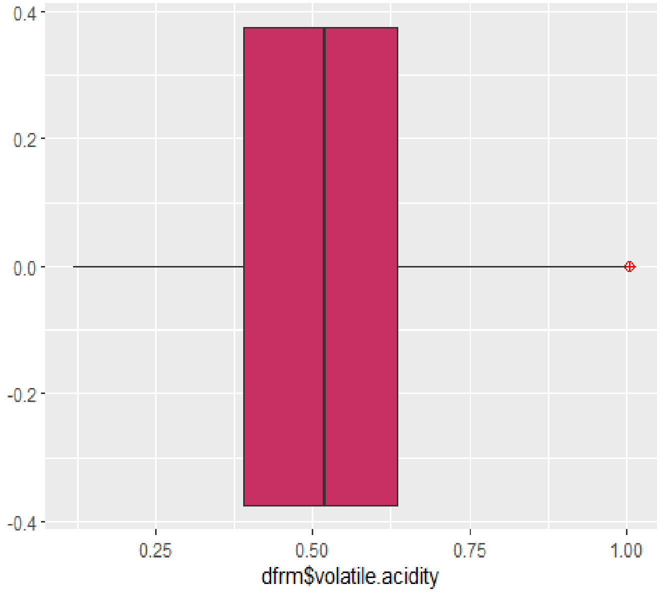
```
binsize = diff (range(dfrm$volatile.acidity))/10 # code for bin number on histogram
ggplot (dfrm, aes(x = dfrm$volatile.acidity)) + geom_histogram (binwidth = binsize, fill = "#c82f63", colour = "#f9e0ec")+ geom_vline(aes (xintercept = mean (dfrm$volatile.acidity)), colour = "#f9e0ec", linetype = "dashed", size = 1)
```



```
ggplot (dfrm, aes(x = dfrm$volatile.acidity)) + geom_density()
```



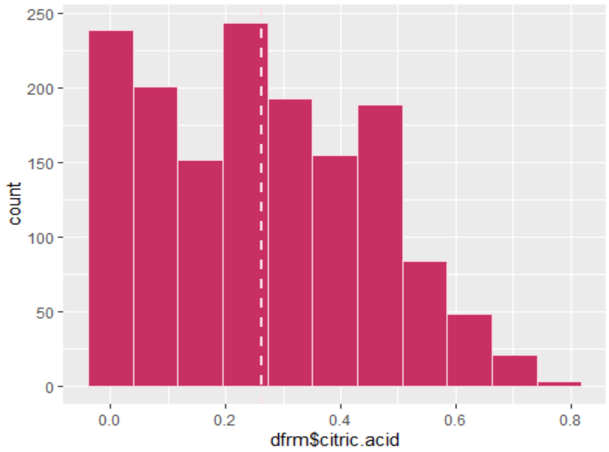
```
ggplot (dfrm, aes(x = dfrm$volatile.acidity)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape = 10, outlier.size = 2)
```



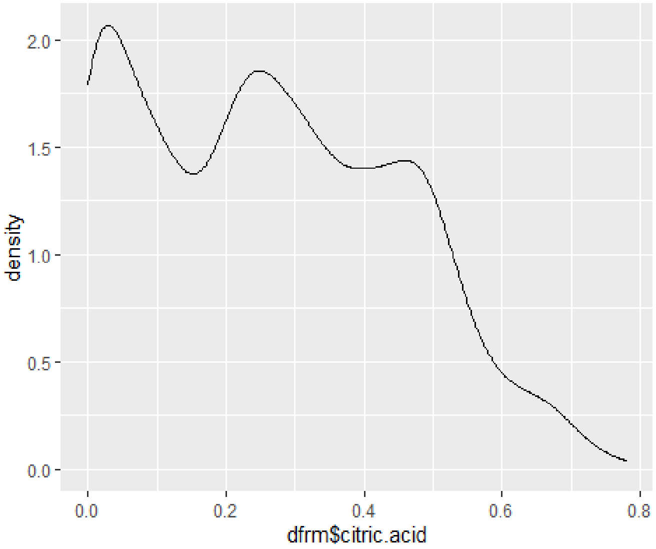
Volatile acidity is the amount of acetic acid in wine which, at too high of levels, can lead to an unpleasant and vinegar taste. Typical value for this dataset is 0.5225, with median being very close to it. But, since first and third quartile are very different, I suppose that there is a skewness in the data. Histogram proves it, and density plot shows it is a bimodal distribution, with positive skewness. After deleting outliers, It remains only one, but I shall leave it as it is for now. Regardless of the wide range, standard deviation is 0.16, with middle variability being higher than the overall.

```
boxplot (dfrm$citric.acid, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.

## [1] 1

outliers = boxplot (dfrm$citric.acid, plot = FALSE)$out # we are going to attach the formula above to a variable named outliers.
dfrm = dfrm [-which(dfrm$citric.acid %in% outliers),] # with this formula, I removed the outliers from the data
summary (dfrm$citric.acid)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2500  0.2622  0.4100  0.7800

sd (dfrm$citric.acid)

## [1] 0.1877939

IQR (dfrm$citric.acid)

## [1] 0.32
```

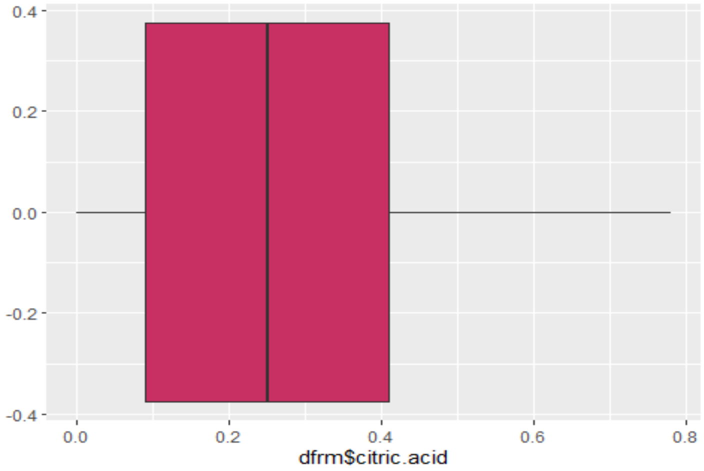
```
binsize = diff(range(dfrm$citric.acid))/10 # code for bin number on histogram
ggplot (dfrm, aes(x = dfrm$citric.acid)) + geom_histogram (binwidth = binsize, fill = "#c82f63", colour = "#f9e0ec") + geom_vline (aes(xintercept = mean(dfrm$citric.acid)), colour = "#f9e0ec", linetype = "dashed", size = 1)
```



```
ggplot (dfrm, aes(x = dfrm$citric.acid)) + geom_density()
```

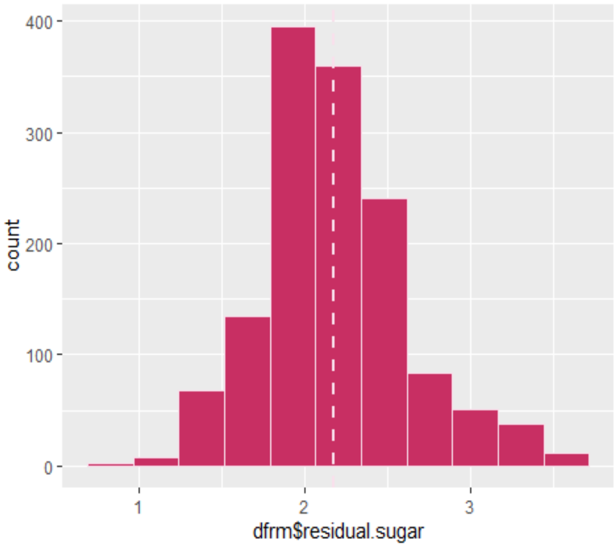


```
ggplot (dfrm, aes(x = dfrm$citric.acid)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape = 10, outlier.size = 2)
```

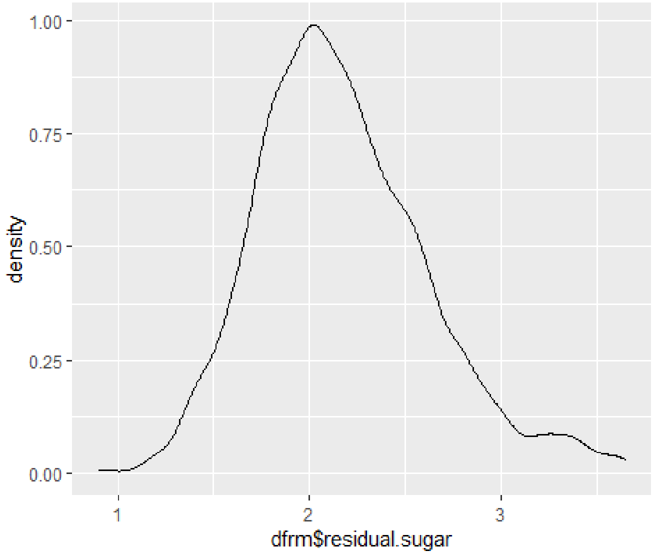


Citric acid is found in small quantities in wines, adding "freshness" and flavor to wines. After deleting extreme outliers, the typical value is 0.26, with median not being too far from it. Again, the quartiles are different, and the range is wide. The middle, 50%, variability of data is larger than overall variability. Even after deleting outliers, the distribution is positively skewed, without any visible outliers. Still, because the density plot shows that the distribution clearly isn't normally distributed, I won't use this variable in the further analyses.

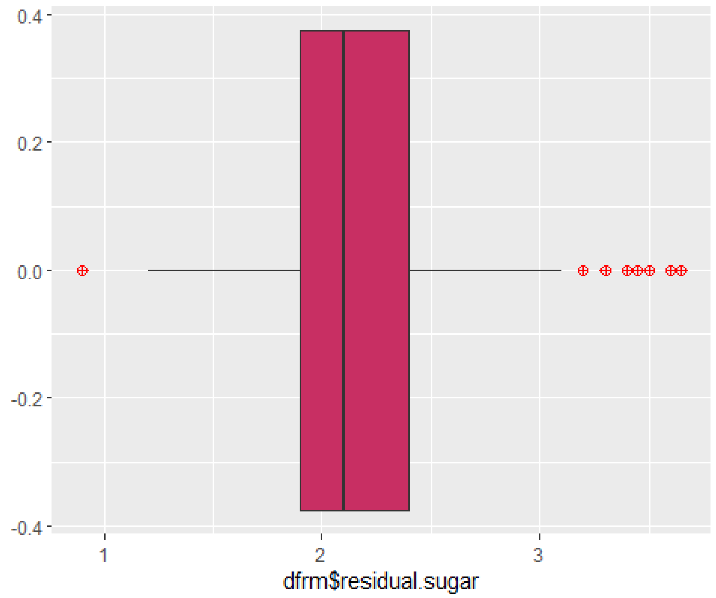
```
boxplot (dfrm$residual.sugar, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.

outliers = boxplot (dfrm$residual.sugar, plot = FALSE)$out # we are going to attach the formula above to a variable named outliers.
dfrm = dfrm [-which (dfrm$residual.sugar %in% outliers),] # with this formula, I removed the outliers from the data
summary (dfrm$residual.sugar)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.100   2.171   2.400   3.650

sd (dfrm$residual.sugar)

## [1] 0.4459984

IQR (dfrm$residual.sugar)

## [1] 0.5

binsize = diff(range(dfrm$residual.sugar))/10 # code for bin number on histogram
ggplot (dfrm, aes(x = dfrm$residual.sugar)) + geom_histogram (binwidth = binsize, fill = "#c82f63", colour = "#f9e0ec") + geom_vline (aes(xintercept = mean(dfrm$residual.sugar)), colour = "#f9e0ec", linetype = "dashed", size = 1)
```



```
ggplot (dfrm, aes(x = dfrm$residual.sugar)) + geom_density()
```

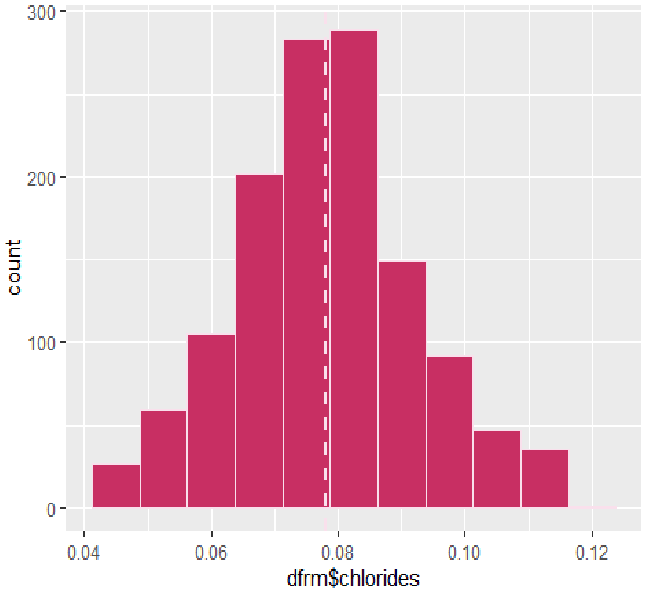


```
ggplot (dfrm, aes (x = dfrm$residual.sugar)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape = 10, outlier.size = 2)
```

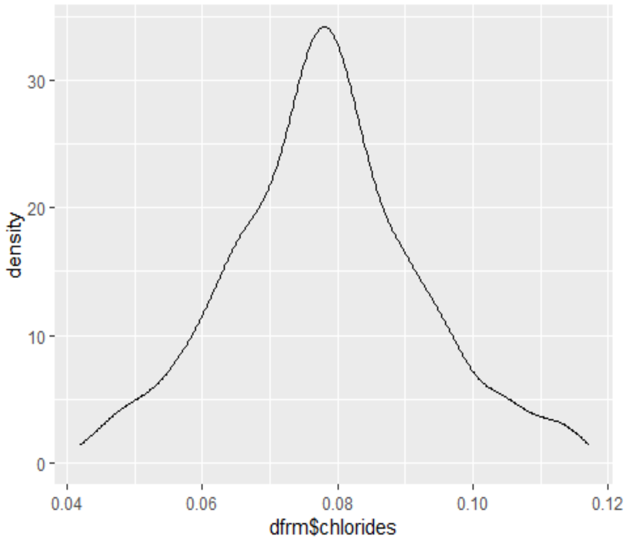


Residual sugar is the amount of sugar remaining after fermentation stops, it is very rare to find wines with less than 1g/L and wines with greater than 45g/L (those are considered sweet). After I deleted extreme outliers, there was no extremes like that, and it all stayed between 1.2 and 3.1 grams/L of red wine. Mean is just slightly higher than the median, which gives us positive skewness. Total variability shown by standard deviation is smaller than middle variability. The distribution shown by density plot, histogram and boxplot looks almost perfect.

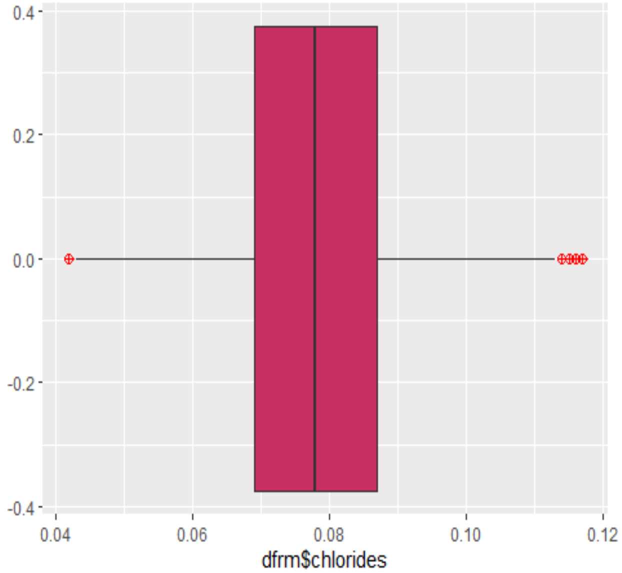
```
boxplot (dfrm$chlorides, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.

outliers = boxplot (dfrm$chlorides, plot = FALSE)$out # we are going to attach the formula above to a variable named outliers.
dfrm = dfrm [-which (dfrm$chlorides %in% outliers),] # with this formula, I removed the outliers from the data
summary (dfrm$chlorides)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04200 0.06900 0.07800 0.07809 0.08700 0.11700

sd (dfrm$chlorides)

## [1] 0.01421088

IQR (dfrm$chlorides)

## [1] 0.018

binsize = diff (range(dfrm$chlorides))/10 # code for bin number on histogram
ggplot (dfrm, aes(x = dfrm$chlorides)) + geom_histogram (binwidth = binsize, fill = "#c82f63", colour = "#f9e
0ec") + geom_vline(aes(xintercept = mean (dfrm$chlorides)), colour = "#f9e0ec", linetype = "dashed", size =
1)
```



```
ggplot (dfrm, aes(x = dfrm$chlorides)) + geom_density()
```

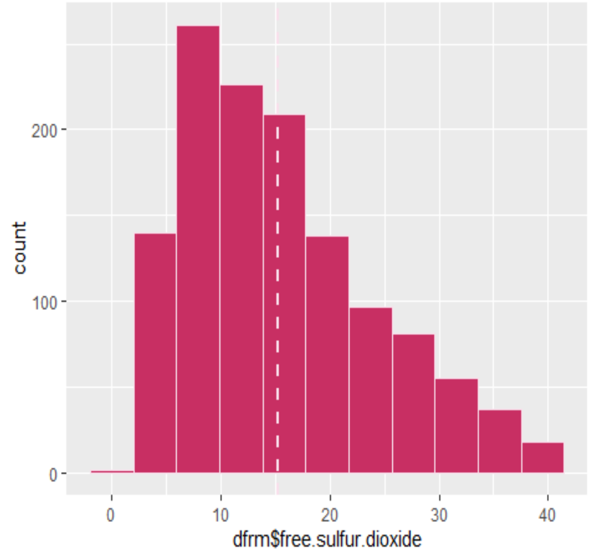


```
ggplot (dfrm, aes(x = dfrm$chlorides)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape
= 10, outlier.size = 2)
```

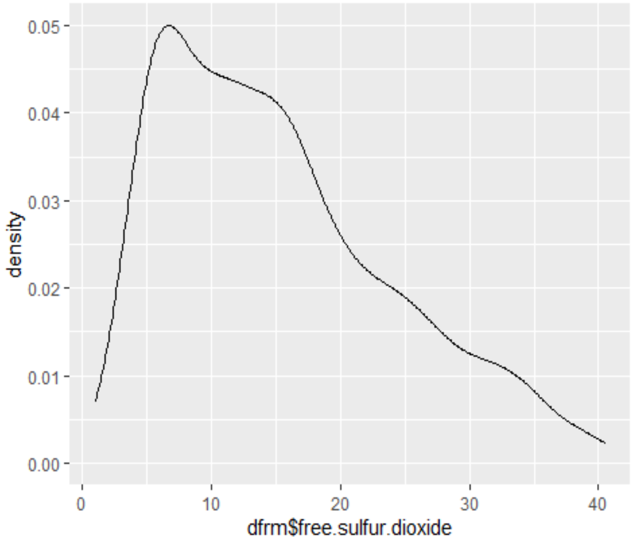


Chlorides represent the amount of salt in the wine. The average amount of chlorides in these red wines is 0.07, with minimum values of salt in wines overall. Median is just slightly higher than the mean, which gives us negative skewness that is almost not visible on the plots. It looks, once again, like perfect normal distribution, after I've deleted outliers.

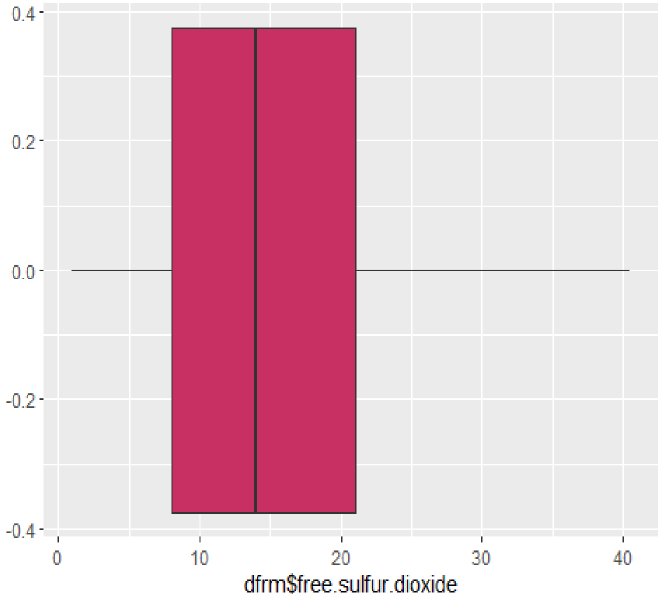
```
boxplot (dfrm$free.sulfur.dioxide, plot = FALSE)$out # this is the formula with which you can see the actual ou
tliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.

## [1] 42 43 47 41 54 46 45 41 41 41 53 52 51 41 45 42 57 50 45 48 41 43 48 43 42

outliers = boxplot(dfrm$free.sulfur.dioxide, plot = FALSE)$out # we are going to attach the formula above to a
variable named outliers.
dfrm = dfrm [-which (dfrm$free.sulfur.dioxide %in% outliers),] # with this formula, I removed the outliers from
the data
summary (dfrm$free.sulfur.dioxide)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     8.0    14.0    15.2    21.0    40.5

sd (dfrm$free.sulfur.dioxide)

## [1] 8.838114

IQR(dfrm$free.sulfur.dioxide)

## [1] 13

binsize = diff(range(dfrm$free.sulfur.dioxide))/10 # code for bin number on histogram
ggplot (dfrm, aes(x = dfrm$free.sulfur.dioxide)) + geom_histogram (binwidth = binsize, fill = "#c82f63", colo
ur = "#f9e0ec") + geom_vline(aes(xintercept = mean(dfrm$free.sulfur.dioxide)), colour = "#f9e0ec", linetype
= "dashed", size = 1)
```



```
ggplot (dfrm, aes(x = dfrm$free.sulfur.dioxide)) + geom_density()
```



```
ggplot (dfrm, aes(x = dfrm$free.sulfur.dioxide)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outli
er.shape = 10, outlier.size = 2)
```

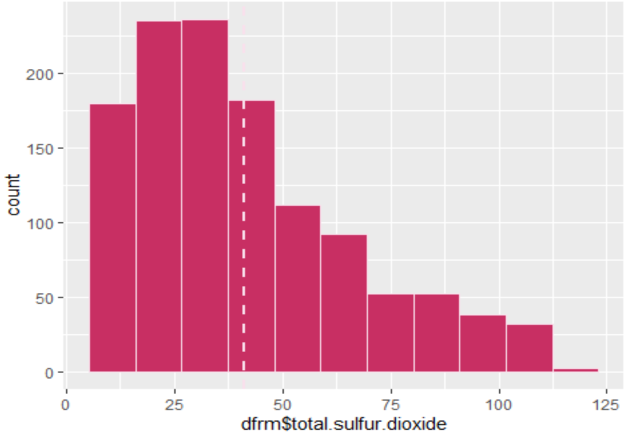


Free sulfur dioxide is the free form of SO2 which exists in the equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion, it also prevents microbial grown and the oxidation of the wine. Typical level of free sulfur dioxide in these wines is 15.2 (positive skewness), but the minimum and maximum value here is differentiating a lot. The standard deviation and IQR are high. It is visible on the plots, that this distribution is positively skewed, and not a normal distribution. Because of that reason, I'll exclude it from further analyses.

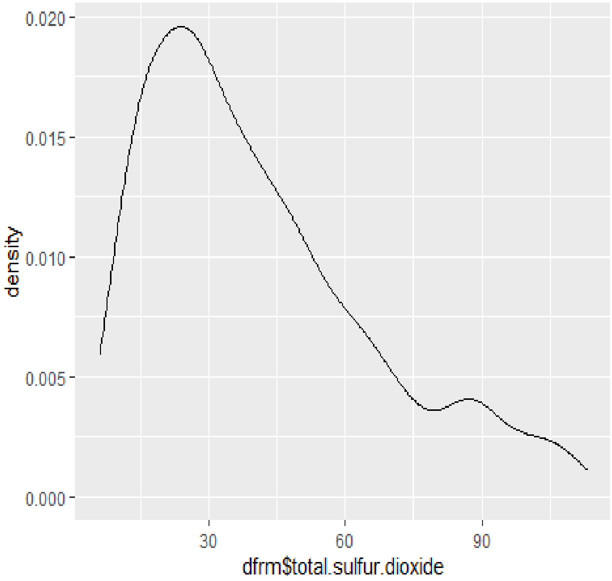
```
boxplot (dfrm$total.sulfur.dioxide, plot = FALSE)$out # this is the formula with which you can see the actual o
utliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.

##  [1] 114 119 119 136 125 136 133 141 143 144 127 120 145 120 144 119 135 114 165
## [20] 122 124 129 142 116 121 149 147 145 155 152 122 125 127 139 143 144 119 130
## [39] 122 115 119 119 119 141 141 133 147 147 131 131 131

outliers = boxplot (dfrm$total.sulfur.dioxide, plot = FALSE)$out # we are going to attach the formula above to
a variable named outliers.
dfrm = dfrm [-which (dfrm$total.sulfur.dioxide %in% outliers),] # with this formula, I removed the outliers fro
m the data
summary (dfrm$total.sulfur.dioxide)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   35.00   41.11   54.00  113.00

sd (dfrm$total.sulfur.dioxide)

## [1] 24.92582

IQR (dfrm$total.sulfur.dioxide)
```

```
## [1] 32
```

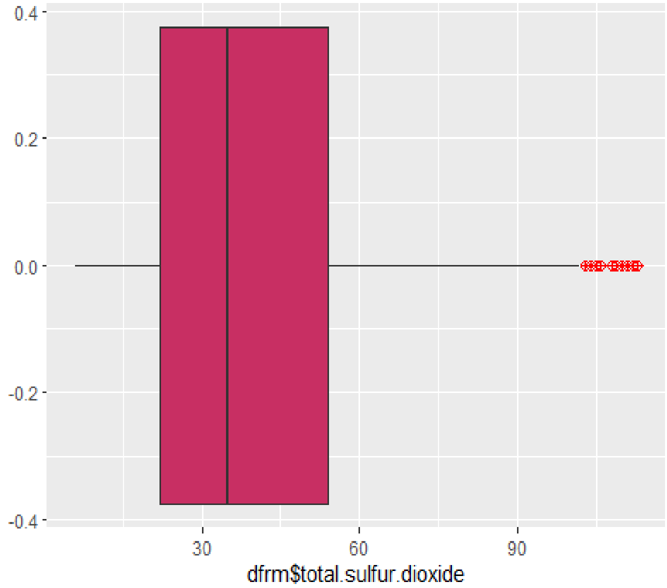
```
binsize = diff(range(dfrm$total.sulfur.dioxide))/10 # code for bin number on histogram
ggplot (dfrm, aes(x = dfrm$total.sulfur.dioxide)) + geom_histogram (binwidth = binsize, fill = "#c82f63", colour = "#f9e0ec") + geom_vline(aes(xintercept = mean(dfrm$total.sulfur.dioxide)), colour = "#f9e0ec", linetype = "dashed", size = 1)
```



```
ggplot (dfrm, aes(x = dfrm$total.sulfur.dioxide)) + geom_density()
```



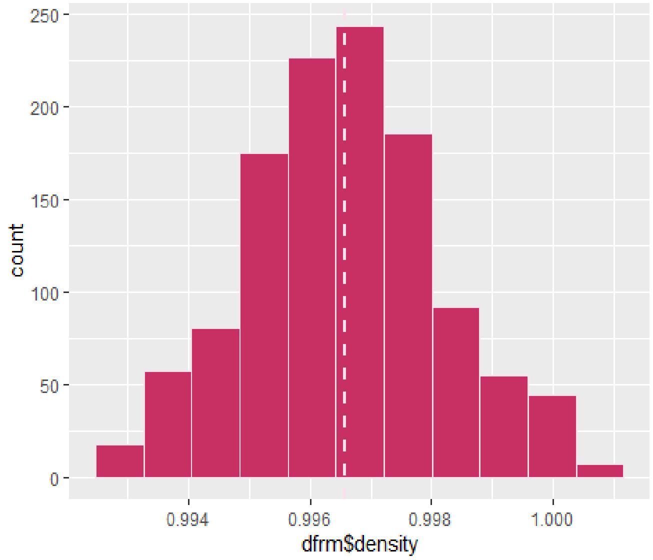
```
ggplot (dfrm, aes(x = dfrm$total.sulfur.dioxide)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape = 10, outlier.size = 2)
```



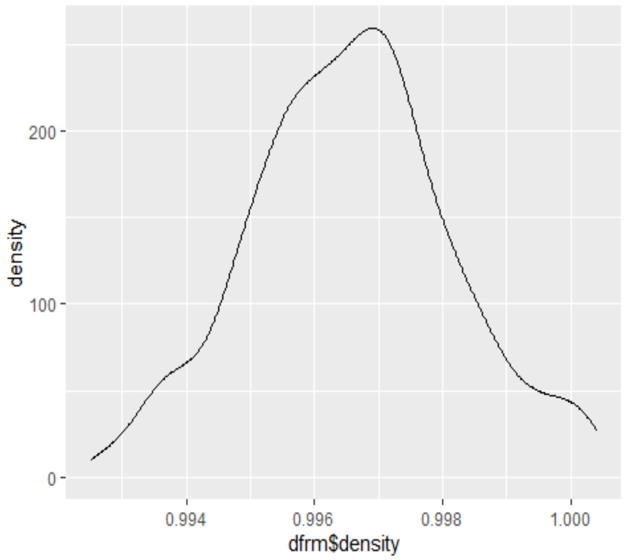
Total sulfur dioxide is the amount of free and bound forms of SO2. In low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine. Typical value in these wines here is 38.37 (positive skewness), but the maximum value exceeds the concentration above 50 ppm. Both standard deviation and IQR are high, and this variable is going to be excluded from further analyses.

```
boxplot (dfrm$density, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.

##  [1] 0.99160 0.99160 0.99240 1.00100 1.00060 0.99240 1.00080 1.00060 0.99170
## [10] 1.00100 0.99154 0.99162 0.99007 0.99007 0.99235 0.99220 0.99150 0.99240
## [19] 0.99157 0.99242 0.99242 0.99080 0.99084 0.99191 0.99236 0.99182 0.99182

outliers = boxplot (dfrm$density, plot = FALSE)$out # we are going to attach the formula above to a variable named outliers.
dfrm = dfrm [-which (dfrm$density %in% outliers),] # with this formula, I removed the outliers from the data
summary (dfrm$density)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9925  0.9955  0.9966  0.9966  0.9976  1.0004

sd (dfrm$density)

## [1] 0.001575209

IQR (dfrm$density)

## [1] 0.002075
```

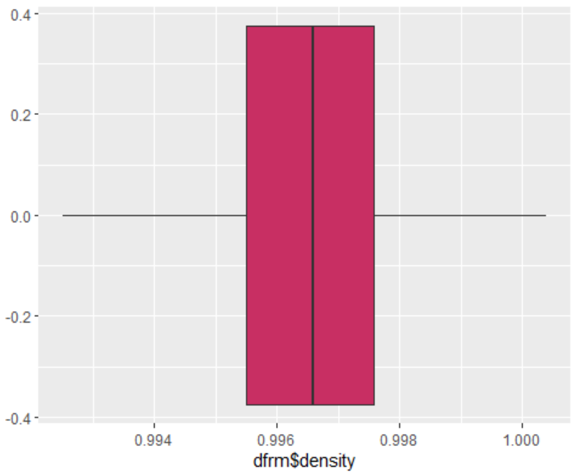
```r
binsize = diff(range(dfrm$density))/10 # code for bin number on histogram
ggplot (dfrm, aes(x = dfrm$density)) + geom_histogram(binwidth = binsize, fill = "#c82f63",colour = "#f9e0ec") + geom_vline(aes(xintercept = mean(dfrm$density)), colour = "#f9e0ec", linetype = "dashed", size = 1)
```



```r
ggplot (dfrm, aes(x = dfrm$density)) + geom_density()
```

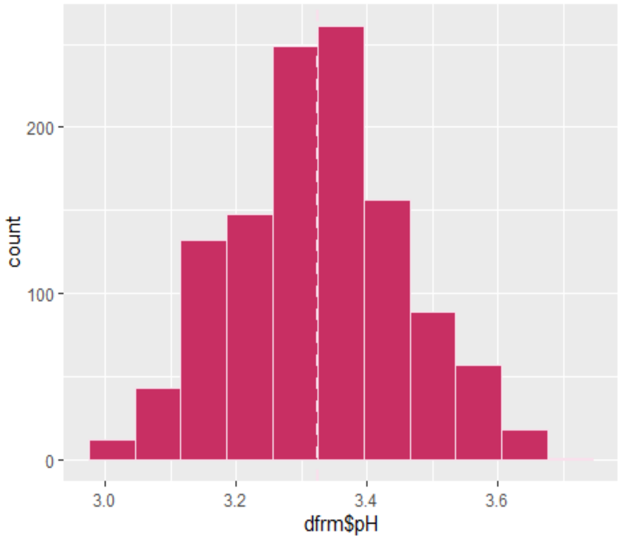


```r
ggplot(dfrm, aes(x = dfrm$density)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape = 10, outlier.size = 2)
```

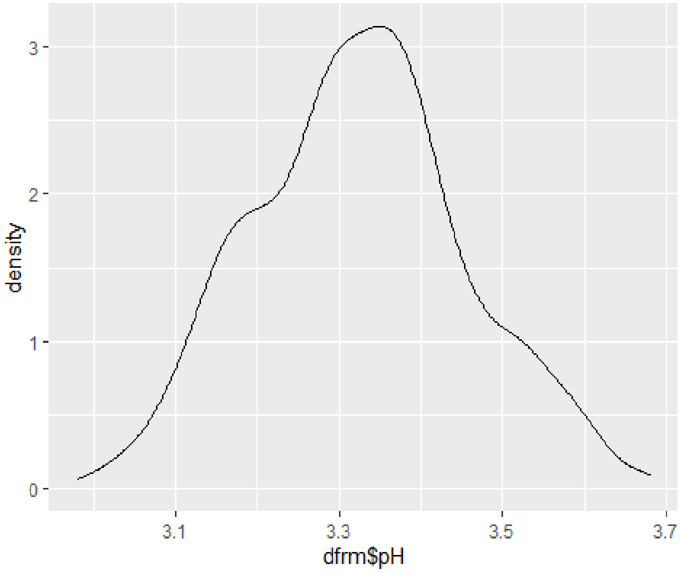


Density represents density of water, which is close to that of the water, depending on the percent alcohol and sugar content. All values are very close together, with median being just slightly over mean (negative skewness). But, this skewness isn't visible on the plots = normal distribution.

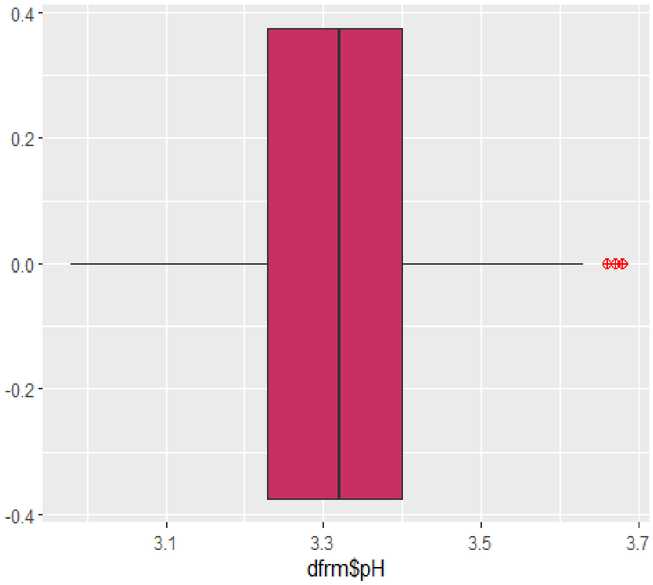
```
boxplot (dfrm$pH, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the vari
able fixed acidity. I plotted false because there is no need to see the boxplot again.

##  [1] 3.90 3.85 3.69 3.69 2.89 2.89 2.92 2.94 2.94 3.69 3.69 3.71 3.71 3.78 2.94
## [16] 3.78 3.71 2.88 3.72 3.72

outliers = boxplot (dfrm$pH, plot = FALSE)$out # we are going to attach the formula above to a variable name
d outliers.
dfrm = dfrm [-which (dfrm$pH %in% outliers),] # with this formula, I removed the outliers from the data
summary (dfrm$pH)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.980   3.230   3.320   3.324   3.400   3.680

sd (dfrm$pH)

## [1] 0.1301453

IQR (dfrm$pH)

## [1] 0.17

binsize = diff (range(dfrm$pH))/10 # code for bin number on histogram
ggplot(dfrm, aes(x = dfrm$pH)) + geom_histogram (binwidth = binsize, fill = "#c82f63", colour = "#f9e0ec")
+ geom_vline(aes(xintercept = mean(dfrm$pH)), colour = "#f9e0ec", linetype = "dashed", size = 1)
```



```
ggplot (dfrm, aes(x = dfrm$pH)) + geom_density()
```



```
ggplot (dfrm, aes(x = dfrm$pH)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape = 10, o
utlier.size = 2)
```



pH describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic), and most wines are between 3-4 on the pH scale. Here, the typical value is 3.327, which is within the range of 3-4. Again, on the plots we can see that this distribution is normal.

```
boxplot (dfrm$sulphates, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.
```

```
##  [1] 1.56 1.20 0.95 1.08 0.97 0.96 0.96 1.18 0.98 1.13 1.04 1.11 1.13 1.06 1.04
## [16] 1.05 1.02 1.02 0.96 1.36 1.16 1.18 1.10 1.01 0.97 0.97
```

```
outliers = boxplot (dfrm$sulphates, plot = FALSE)$out # we are going to attach the formula above to a variable named outliers.
dfrm = dfrm [-which (dfrm$sulphates %in% outliers),] # with this formula, I removed the outliers from the data
summary (dfrm$sulphates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6100  0.6291  0.7000  0.9400
```

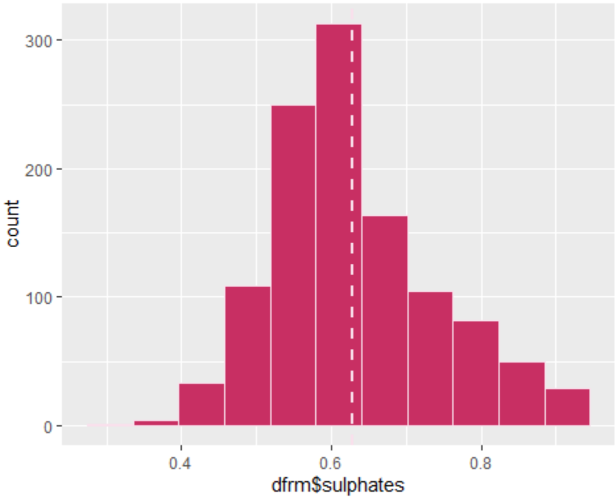
```
sd (dfrm$sulphates)
```

```
## [1] 0.1125968
```

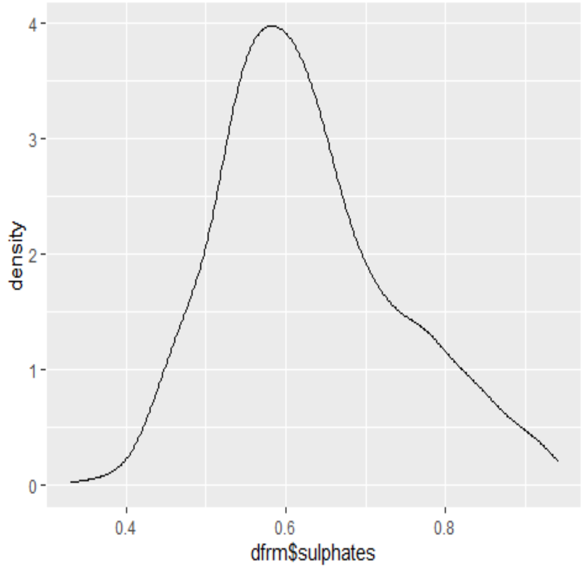
```
IQR (dfrm$sulphates)
```

```
## [1] 0.15
```

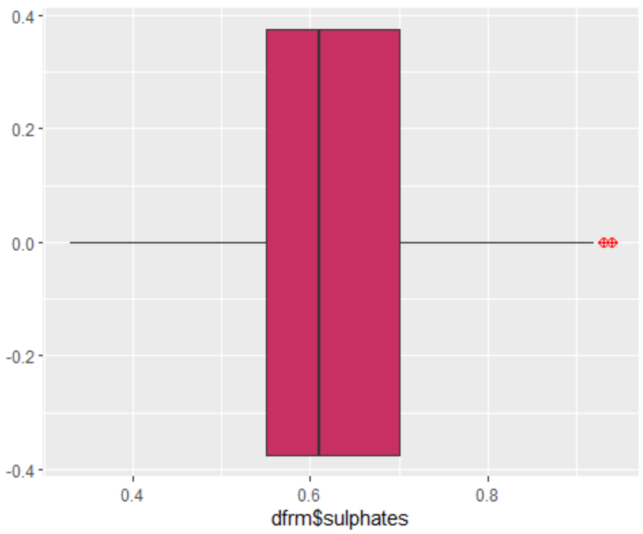
```
binsize = diff(range (dfrm$sulphates))/10 # code for bin number on histogram
ggplot(dfrm, aes(x = dfrm$sulphates)) + geom_histogram(binwidth = binsize, fill = "#c82f63", colour = "#f9eoec") + geom_vline(aes(xintercept = mean(dfrm$sulphates)), colour = "#f9eoec", linetype = "dashed", size = 1)
```



```
ggplot (dfrm, aes(x = dfrm$sulphates)) + geom_density()
```

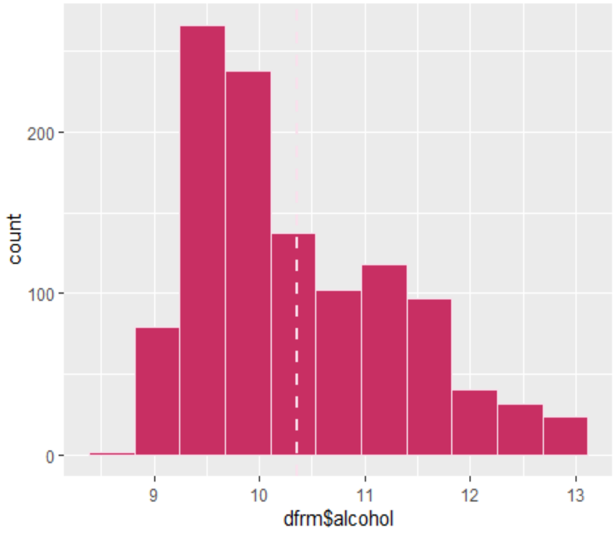


```
ggplot (dfrm, aes(x = dfrm$sulphates)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape = 10, outlier.size = 2)
```

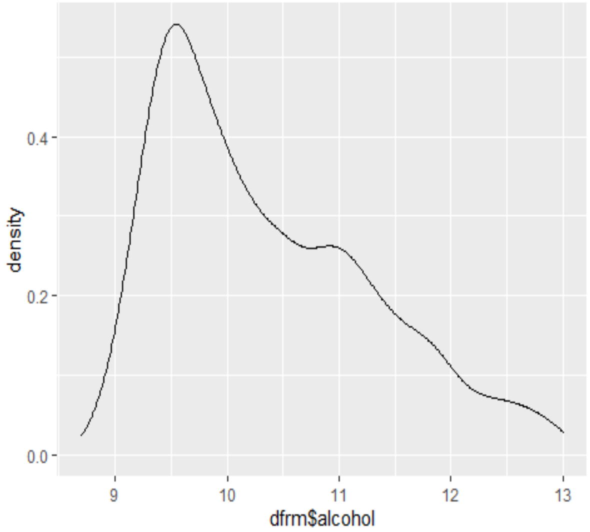


Sulphates are a wine additive which can contribute to sulfur dioxide gas (SO2) levels, which acts as an antimicrobial and antioxidant. Typical value here is 0.63, with median slightly lower than it. The minimum and maximum value differentiate high, but inner (middle) variability is higher. Plots show positive skewness, but density plot appears to show pretty normal distribution.

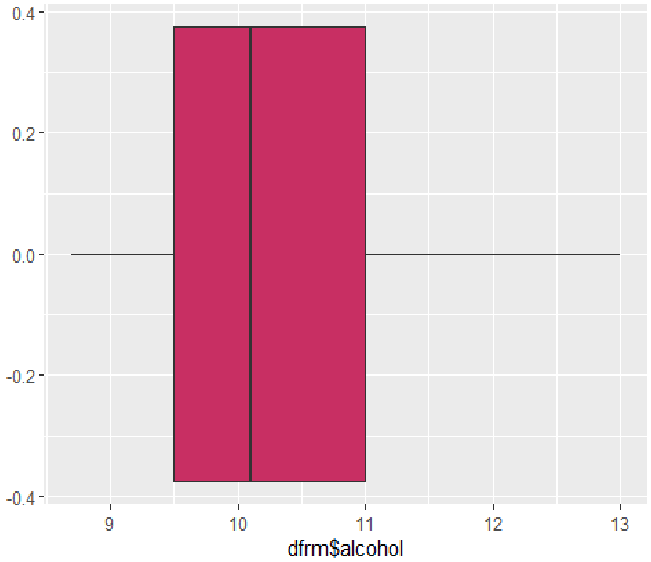
```
boxplot (dfrm$alcohol, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the variable fixed acidity. I plotted false because there is no need to see the boxplot again.

## [1] 14.0 13.3 13.4 13.3 13.6

outliers = boxplot(dfrm$alcohol, plot = FALSE)$out # we are going to attach the formula above to a variable named outliers.
dfrm = dfrm [-which (dfrm$alcohol %in% outliers),] # with this formula, I removed the outliers from the data
summary (dfrm$alcohol)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.70    9.50   10.10   10.35   11.00   13.00

sd (dfrm$alcohol)

## [1] 0.9536173

IQR (dfrm$alcohol)

## [1] 1.5

binsize = diff(range(dfrm$alcohol))/10 # code for bin number on histogram
ggplot(dfrm, aes (x = dfrm$alcohol)) + geom_histogram (binwidth = binsize, fill = "#c82f63", colour = "#f9e0ec") + geom_vline(aes(xintercept = mean(dfrm$alcohol)), colour = "#f9e0ec", linetype = "dashed", size = 1)
```



```
ggplot (dfrm, aes(x = dfrm$alcohol)) + geom_density()
```

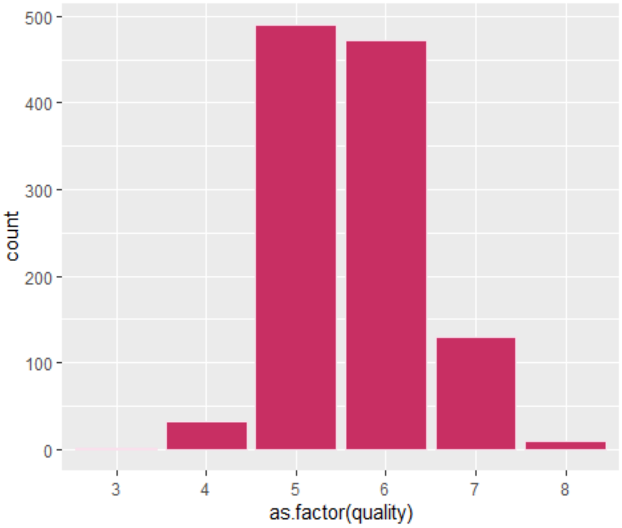


```
ggplot (dfrm, aes(x = dfrm$alcohol)) + geom_boxplot (fill = "#c82f63", outlier.colour = "red", outlier.shape =
10, outlier.size = 2)
```



Alcohol variable represents the percent alcohol of the wine. Typical percentage of alcohol in these wines is 10.36%, with range from 8 to 13. The density plot doesn't show the right shape of normal distribution, so I'll keep this variable out of further analyses, although this variable plays a big role in wines.

```
summary (dfrm$quality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.637   6.000   8.000

sd (dfrm$quality)

## [1] 0.7585177

IQR (dfrm$quality)

## [1] 1

binsize = diff(range(dfrm$quality))/10 # code for bin number on histogram
ggplot(dfrm, aes(x = as.factor(quality))) + geom_bar(fill = "#c82f63", colour = "#f9e0ec")
```



The quality variable is an output variable, based on sensory data, and it scores between 0 and 10. The highest quality score here is 8, and the lowest 3. This is the output variable we are trying to predict (y-variable), so we'll keep it.

With this, I have finished summaries of all variables in the sample. Since we want to predict the quality variable, this is going to be our y-variable or dependent variable, whereas all other variables will be independent variables (x-variables). More info on that in the next chapter!

Regression analysis and correlation

In the prior chapter, I have shown summaries and basic graphics of each variable. The main goal of this dataset analysis is that we determine which variables make a wine good, or which variables give higher quality of red wines in the sample, but also in the population as well.

For that, I am going to calculate correlations, to see what the relationship between variables is. After that, I am going to calculate linear regression models where the y-variable (dependent variable) is quality variable, and all other variables will be independent variables or x-variables.

So, let us start with correlation first. Since there are 66 combination of correlation, I am making a correlation matrix, and then proceed from there to extract only significant correlations (above 0,5) and make a regression model.

```
cormat <- cor(dfrm)
print (cormat, digits=2)
```

```
##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity               1.000          -0.2691       0.659          0.227
## volatile.acidity           -0.269           1.0000      -0.634          0.028
## citric.acid                 0.659          -0.6337       1.000          0.151
## residual.sugar              0.227           0.0283       0.151          1.000
## chlorides                   0.169           0.1209       0.063          0.235
## free.sulfur.dioxide        -0.169          -0.0021      -0.095          0.081
## total.sulfur.dioxide       -0.091           0.0914      -0.014          0.176
## density                     0.597           0.0597       0.292          0.388
## pH                         -0.693           0.2244      -0.476         -0.043
## sulphates                   0.160          -0.3375       0.265          0.049
## alcohol                    -0.012          -0.2474       0.169          0.127
## quality                     0.131          -0.3654       0.239          0.042
##                      chlorides free.sulfur.dioxide total.sulfur.dioxide density
## fixed.acidity            0.169             -0.1686               -0.091   0.597
## volatile.acidity         0.121             -0.0021                0.091   0.060
## citric.acid              0.063             -0.0950               -0.014   0.292
## residual.sugar           0.235              0.0813                0.176   0.388
## chlorides                1.000              0.0145                0.193   0.395
## free.sulfur.dioxide      0.015              1.0000                0.635  -0.027
## total.sulfur.dioxide     0.193              0.6352                1.000   0.161
## density                  0.395             -0.0273                0.161   1.000
## pH                      -0.151              0.1607                0.032  -0.217
## sulphates               -0.096              0.1165               -0.057   0.054
## alcohol                 -0.286             -0.0117               -0.248  -0.528
## quality                 -0.179             -0.0031               -0.190  -0.206
##                           pH sulphates alcohol quality
## fixed.acidity        -0.6930    0.1603  -0.012  0.1306
## volatile.acidity      0.2244   -0.3375  -0.247 -0.3654
## citric.acid          -0.4756    0.2649   0.169  0.2394
## residual.sugar       -0.0431    0.0490   0.127  0.0417
## chlorides            -0.1513   -0.0959  -0.286 -0.1788
## free.sulfur.dioxide   0.1607    0.1165  -0.012 -0.0031
## total.sulfur.dioxide  0.0323   -0.0573  -0.248 -0.1899
## density              -0.2175    0.0543  -0.528 -0.2062
## pH                    1.0000   -0.0069   0.098 -0.0890
## sulphates            -0.0069    1.0000   0.281  0.4349
## alcohol               0.0978    0.2812   1.000  0.4979
## quality              -0.0890    0.4349   0.498  1.0000
```

From the correlation matrix, I several correlation coefficients higher than 0.4, but the number of combination is too high. Because of that, I want to check how much variability is accounted by all of the variables together (x), versus Quality (y) variable. In order to do that, I'm going to create a multiple regression model.

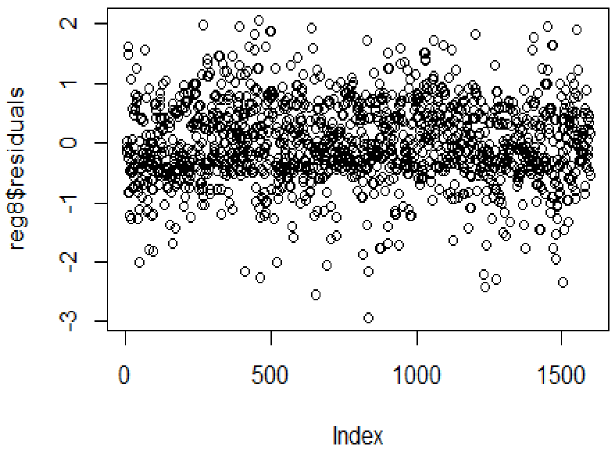
```
reg8 <- lm (quality~fixed.acidity + volatile.acidity + residual.sugar + chlorides + density + pH + sulphates)
print (summary(reg8))
```

```
## 
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar + 
##     chlorides + density + pH + sulphates)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.93826 -0.41178 -0.06903  0.45897  2.06735 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      205.31883   13.97723  14.690  < 2e-16 ***
## fixed.acidity      0.20828    0.01883  11.062  < 2e-16 ***
## volatile.acidity  -1.04221    0.10617  -9.816  < 2e-16 ***
## residual.sugar     0.08765    0.01317   6.653 3.93e-11 ***
## chlorides         -2.04416    0.41989  -4.868 1.24e-06 ***
## density         -205.35225   14.30815 -14.352  < 2e-16 ***
## pH                 0.89646    0.16367   5.477 5.02e-08 ***
## sulphates          1.22568    0.11354  10.795  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6761 on 1591 degrees of freedom
## Multiple R-squared:  0.3021, Adjusted R-squared:  0.2991 
## F-statistic: 98.41 on 7 and 1591 DF,  p-value: < 2.2e-16
```

As you can see, at first, R squared isn't particularly high - only 0.2961, which means that there are other factors and variables in real life that are having an effect on quality of red wines. But, all p-values of x variables are lower than 0.05, which means that they are significant in the model, and the model itself has a low p-value.

Residuals should be scattered around, forming no specific form. As we can see below, there is no specific form taken.

```
plot(reg8$residuals)
```



The linear model is –>

Quality = 205,31 + 0,20828 * fixed acidity - 1,04221 * volatile acidity + 0,08765 * residual sugar - 2,04416 * chlorides - 205,35225 * density + 0,89646 * pH + 1,22568 * sulphates