Red Wine

SUMMARY STATISTICS AND DISTRIBUTIONS

Summary Statistics Table

	mean	trim_mean	sd	median	min	max	n
fixed acidity	8.31963727	8.15253708	1.741096318	7.90000	4.60000	15.90000	1599
volatile acidity	0.52782051	0.51806792	0.179059704	0.52000	0.12000	1.58000	1599
citric acid	0.27097561	0.26128806	0.194801137	0.26000	0.00000	1.00000	1599
residual sugar	2.53880550	2.25835285	1.409928060	2.20000	0.90000	15.50000	1599
chlorides	0.08746654	0.08023497	0.047065302	0.07900	0.01200	0.61100	1599
free sulfur dioxide	15.87492183	14.57728337	10.460156970	14.00000	1.00000	72.00000	1599
total sulfur dioxide	46.46779237	41.84309133	32.895324478	38.00000	6.00000	289.00000	1599
density	0.99674668	0.99673621	0.001887334	0.99675	0.99007	1.00369	1599
pH	3.31111320	3.30909446	0.154386465	3.31000	2.74000	4.01000	1599
sulphates	0.65814884	0.63744731	0.169506980	0.62000	0.33000	2.00000	1599
alcohol	10.42298311	10.31003123	1.065667582	10.20000	8.40000	14.90000	1599
quality	5.63602251	5.58860265	0.807569440	6.00000	3.00000	8.00000	1599

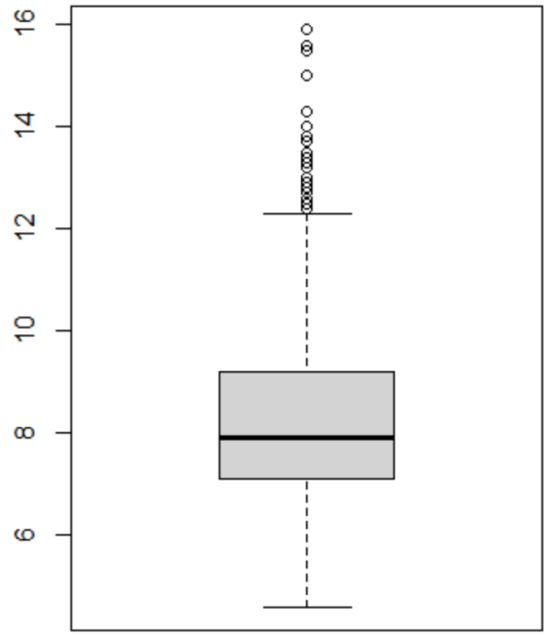
Initial Observations:

The sample size of this data is 1599, because there are this many observations. There are no missing observations for any column of data.

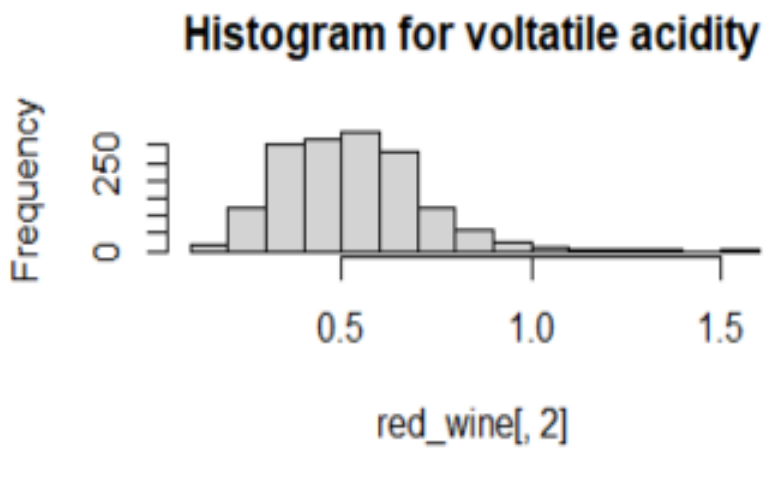
There are some outliers in this data. This is clear by taking the summary statistics of the data set. The max values in some of the variables are far larger than the other values in the same set. There are some concerns with the data quality, because outliers can cause the data to be misrepresented by summary statistics.

Fixed Acidity

	mean	trim_mean	sd	median	min	max	n
fixed acidity	8.319637	8.152537	1.741096	7.9	4.6	15.9	1599

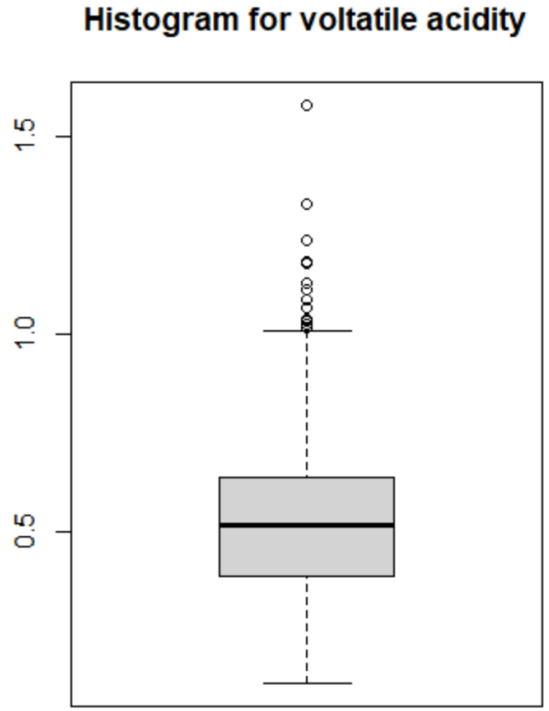


The fixed acidity follows a close to normal distribution, though, there are many values significantly higher than the mean and a very large max value. Because the data is shaped this way, the trimmed mean will provide a more accurate way to describe the average of this data by removing some of the outliers on both ends.

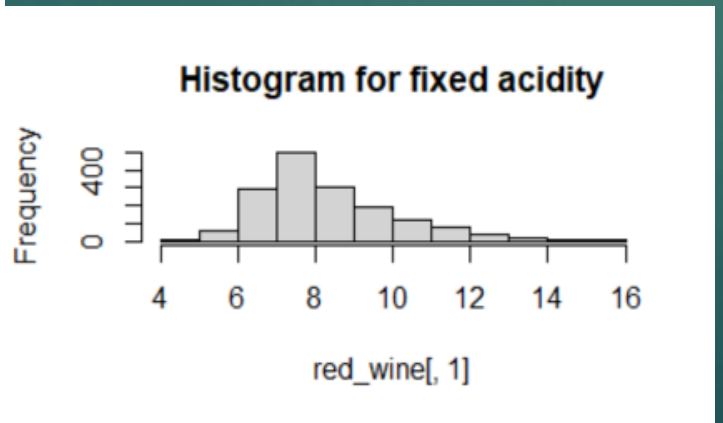


Volatile Acidity

	mean	trim_mean	sd	median	min	max	n
volatile acidity	0.5278205	0.5180679	0.1790597	0.52	0.12	1.58	1599

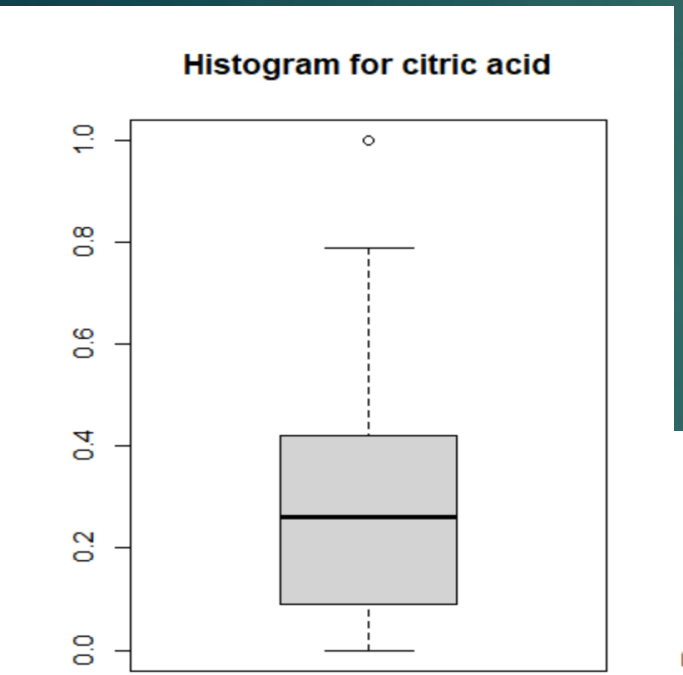


Volatile acidity seems to follow a slightly skewed distribution. The histogram and extended upper whisker on the box plot gives this intuition as well as the mean being slightly higher than the median. There do not appear to be any significant outliers as the skew justifies a high max value.

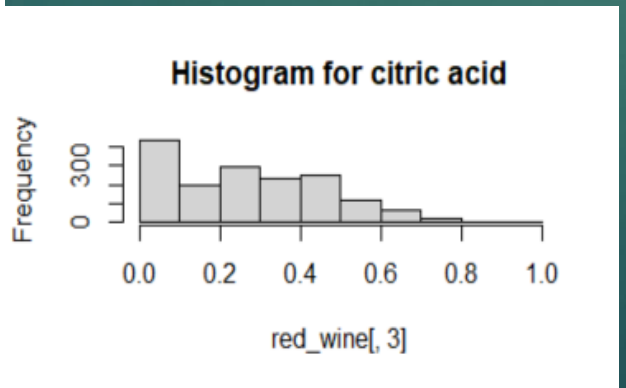


Citric Acid

	mean	trim_mean	sd	median	min	max	n
citric acid	0.2709756	0.2612881	0.1948011	0.26	0	1	1599



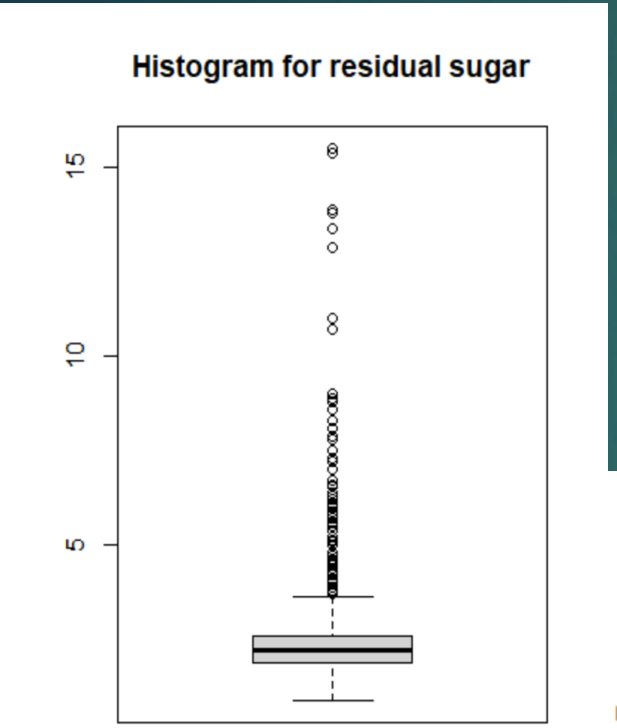
The curve for citric acid appears to be quite skewed given by the mean being greater than the median and the two plots, though the values seem to be quite uniform from 0.1 – 0.5. This could be due to a selection of wines having the characteristic of very low citric acid while most others tending to hover within the previously specified range of values.

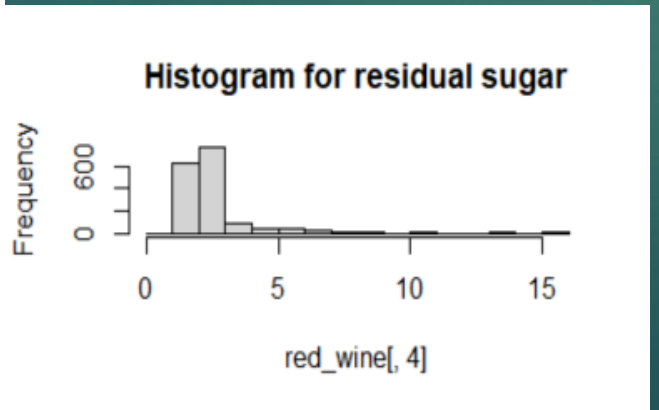


Residual Sugar

	mean	trim_mean	sd	median	min	max	n
residual sugar	2.538806	2.258353	1.409928	2.2	0.9	15.5	1599

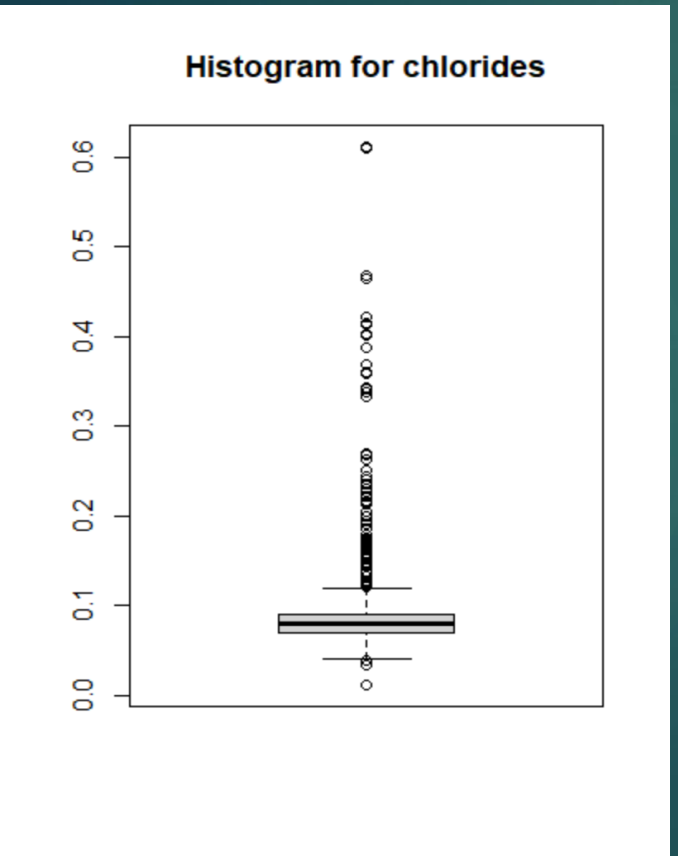
The residual sugar distribution is sparse and wide. The values tend to hover between 1 and 3.5 while there is an extensive tale on the high end of the histogram and many values outside the interquartile range in the boxplot. These extended values are likely due to niche wine selections, vineyard locations, or processes. Some could be true outliers, but an additional explanations seems justified. A trimmed mean would be a good option to represent the majority of wines, because it seems the skew of this data is made up by only a small fraction of the observations.



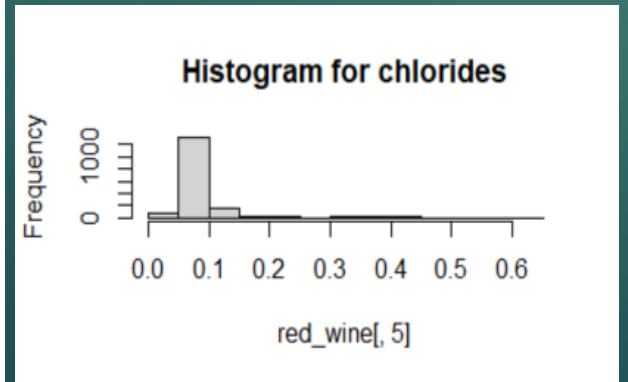


Chlorides

	mean	trim_mean	sd	median	min	max	n
chlorides	0.08746654	0.08023497	0.0470653	0.079	0.012	0.611	1599



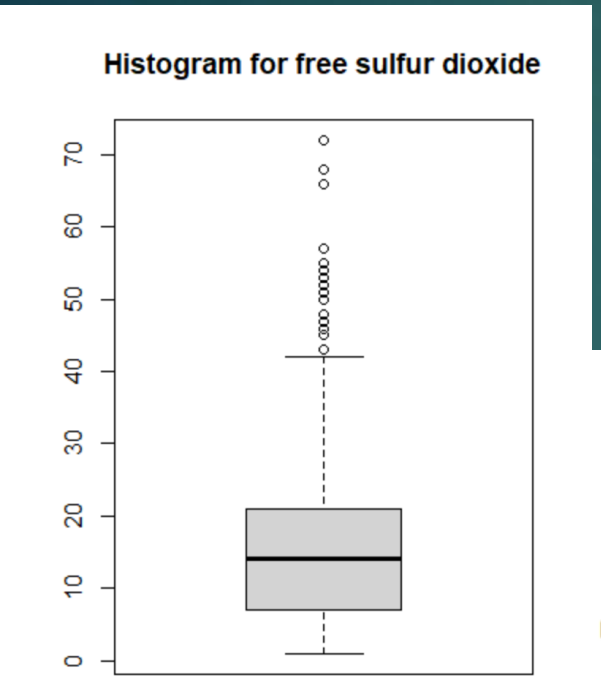
This is another category which has a very sparse distribution. For the most part, the distribution looks normal when trimming off the outliers. Some of these outliers, again, seem like they could be explained by a niche cause. As with residual sugar, the trimmed mean would be a good option for summarizing the average.

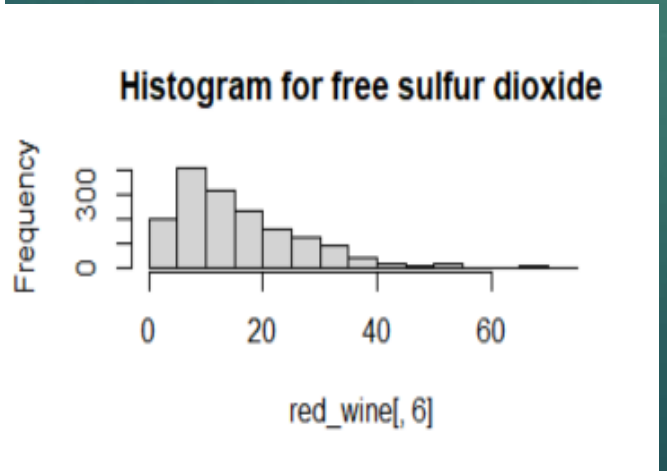


Free Sulfur Dioxide

	mean	trim_mean	sd	median	min	max	n
free sulfur dioxide	15.87492	14.57728	10.46016	14	1	72	1599

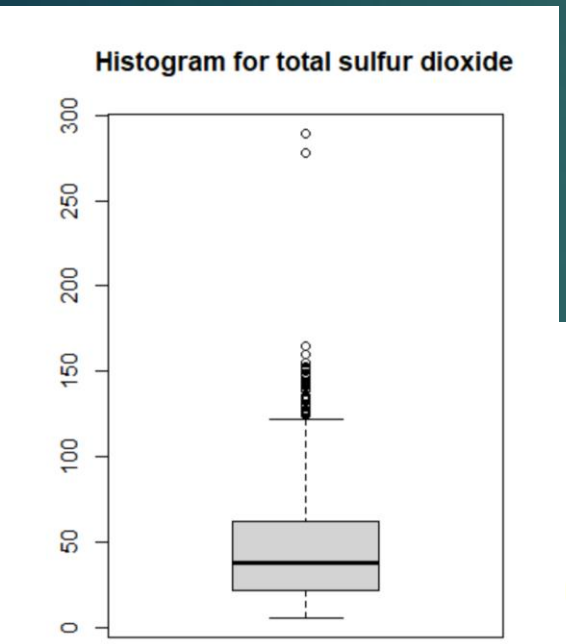
Sulfur dioxide follows a positively skewed distribution as indicated by the plots and the high mean value relative to the median. Even though the distribution is skewed, some of these high values do not appear justified by its skewedness. At least the maximum value appears as though it should be considered an outlier. As revealed by the boxplot, there are a few other points which hover significantly over the high values. Because this distribution is skewed, the best option to summarize the average of this data would be the median.



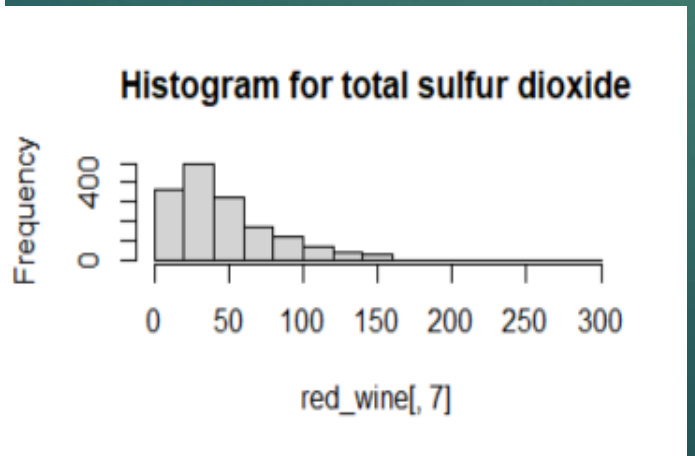


Total Sulfur Dioxide

	mean	trim_mean	sd	median	min	max	n
total sulfur dioxide	46.46779	41.84309	32.89532	38	6	289	1599

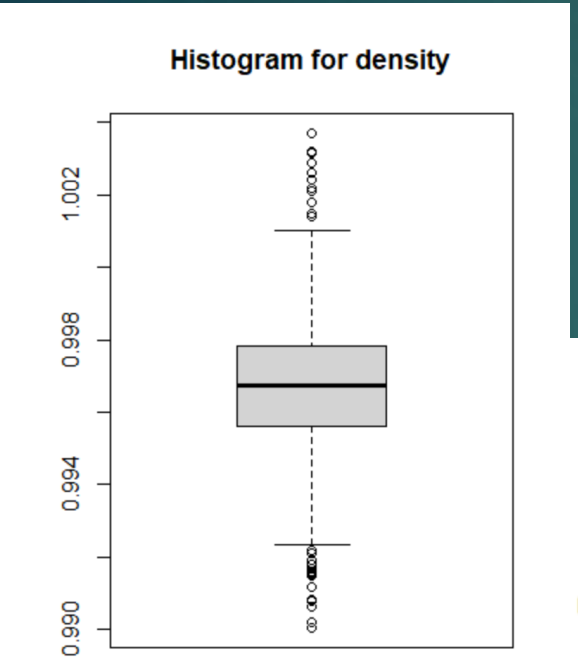


The distribution for total sulfur dioxide is reminiscent of free sulfur dioxide. The distribution is neatly skewed with just a few outliers on the high end that is revealed by the boxplot.

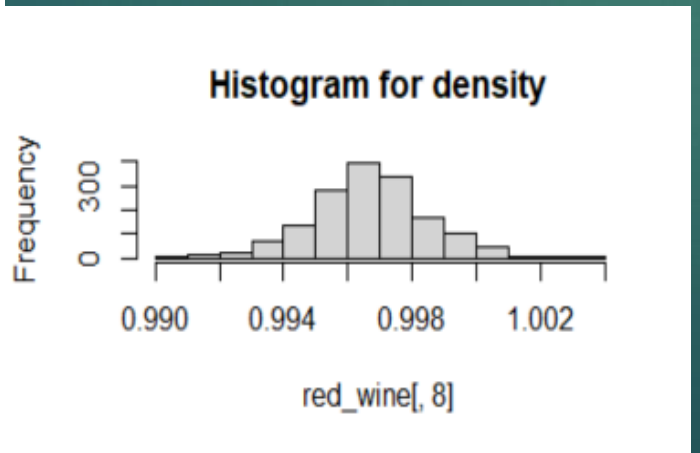


Density

	mean	trim_mean	sd	median	min	max	n
density	0.9967467	0.9967362	0.001887334	0.99675	0.99007	1.00369	1599

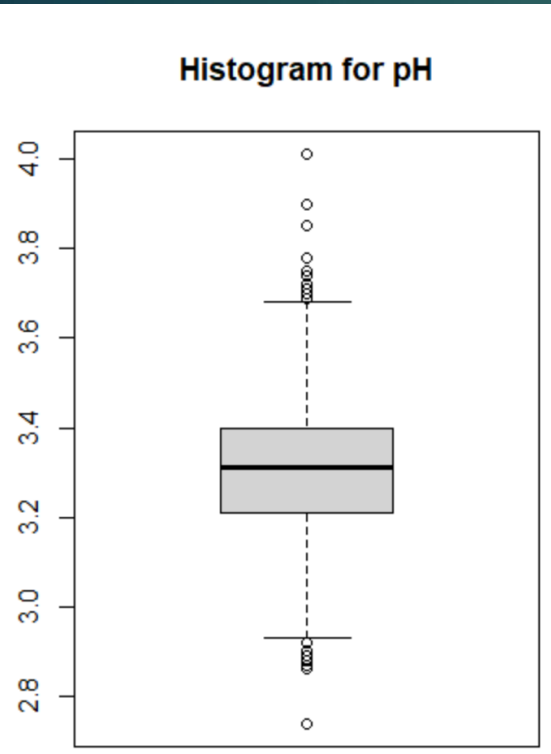


The distribution for the density of the wine follows a normal distribution. The intuition for this is gained by looking at the equal lengths on the whiskers of the box plot, the symmetric curve of the histogram, and very close mean-median values relative to the standard deviation. There do not appear to be any significant outliers so the mean would summarize the average adequately.

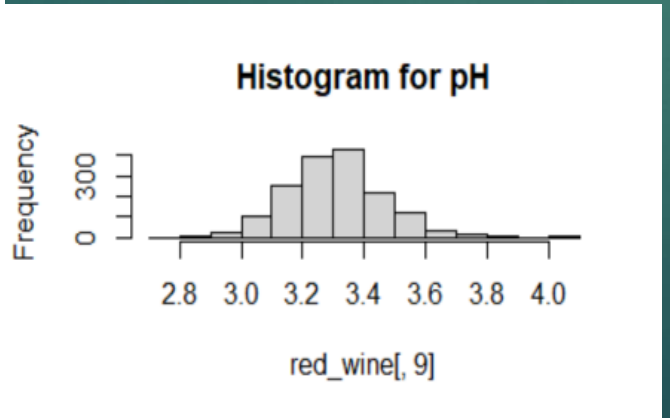


pH

	mean	trim_mean	sd	median	min	max	n
pH	3.311113	3.309094	0.1543865	3.31	2.74	4.01	1599

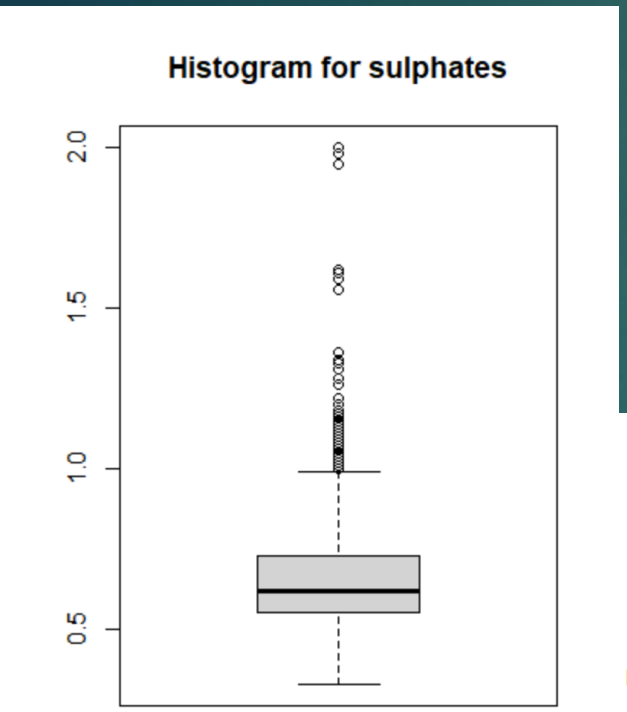


The pH distribution curve follows a normal distribution. In this case, it does appear there are some outliers on the high and low end as indicated by the whiskers and standard deviation. The maximum value is about 4.5 standard deviations away from the mean and the minimum value is 3.7 standard deviation. The trimmed mean would be a great way to summarize the average here.

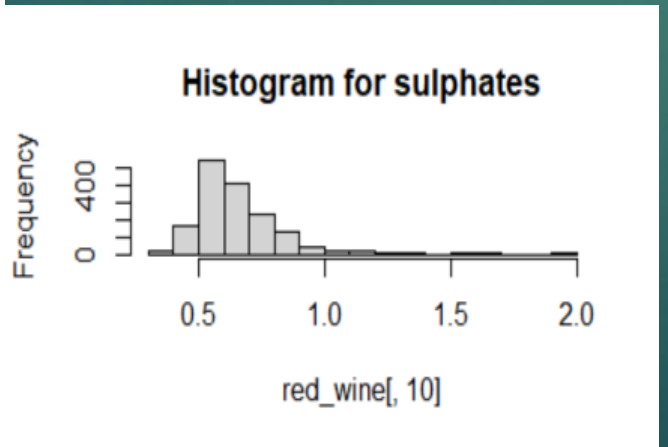


Sulphates

	mean	trim_mean	sd	median	min	max	n
sulphates	0.6581488	0.6374473	0.169507	0.62	0.33	2	1599

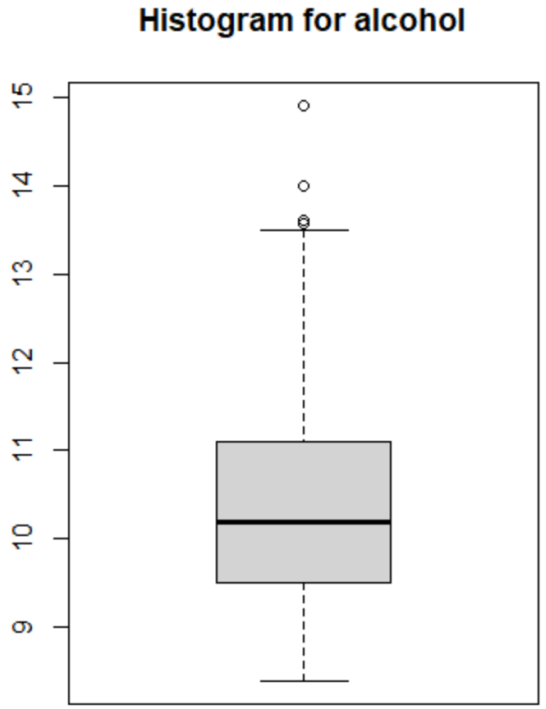


The distribution of sulphates follows a skewed distribution, and it contains multiple outliers on the high end. The outliers are clear when you look at the box plot, because they sit far above the values on the high end. Using the median would be the best option for summarizing the average of this data.



Alcohol Content

	mean	trim_mean	sd	median	min	max	n
alcohol	10.42298	10.31003	1.065668	10.2	8.4	14.9	1599



The distribution of the alcohol content follows a neatly skewed distribution. There does not appear to be any blatant outliers. The median would be the best option to summarize the average of this data.