Data Understanding

Cody Frisby

9/4/2017

I chose a data set from the the UC Irvine machine learning repository website, the “Wine Quality Data Set” (red).

A data set such as this might be useful for a vineyard owner/wine business. They may want to understand what factors result in the “best” tasting wines according to a panel of wine-tasting experts. Wine experts might be consulted to determine which variables can and should be included in such a study. Can we measure the color of the wine? (I know these are all red wines but the color can vary). What factors from a chemical analysis of the wines are the most important? Which factors, that we can capture, are most predictive of wine “quality” as rated by the panel of wine experts? Further, if there are certain characteristcs of the “best” wines, can I control for them at my vineyard? Or is it just luck where my vineyard is?

These are some of the questions I might want to ask as a vineyard owner. If the data can answer any of them, then the investment in understanding the data is well worth it.

The “Wine Quality” data set was created using red and white wine samples of the Portuguese “Vinho Verde” wine. The predictor variables are objective measures of the chemical charateristics of the wine (e.g pH, citric acid, etc) and the dependent variable, **quality**, is a rating on a scale from 0 to 10 by a panel of wine-tasting experts.

Note: Due to privacy and logistic concerns there is no data about wine brands, selling price, and grape types.

After posing some questions and/or hypothesis we can begin to understand our data. It’s important to understand how the data was collected. Was it input manuall? Was it mined from much larger, raw data? As it turns out, the 11 variables (not including the output variable **quality**) in this dataset, come from a physiochemical analysis of the wine. These physiochemical properties include **fixed acidity**, **volatile acidity**, **citric acid**, **residual sugar**, **chlorides**, **free sulfer dioxide**, **total sulfur dioxide**, **density**, **pH**, **sulfates**, and **alcohol**.

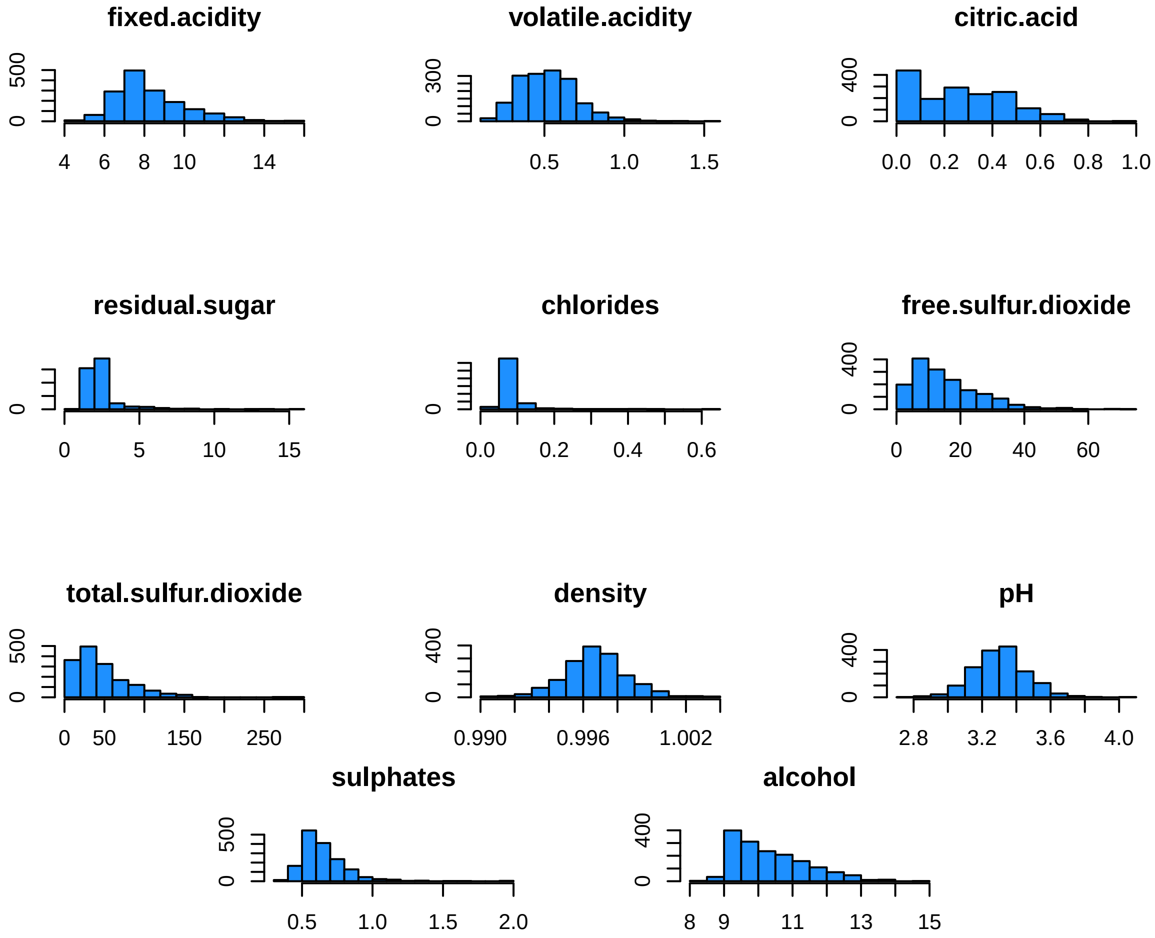
Among the 11 variables that resulted from the physiochemical tests, are there any outliers among them? **Note**: An outlier is defined as being either above or below $\bar{x} \pm 2\hat{\sigma}$ where $\bar{x}$ is the average and $\hat{\sigma}$ is the standard deviation of the variable of interest.

The below table displays the percentage of observations that are considered outliers according to the above definition. As we can see, none of the variables contain greater than 5.1% observations beyond the outlier criteria. The table also contains each variables mean, standard deviation, and the difference betwen the max and the min (range_delta).

characteristic	average	std_dev	range_delta	percent_out
fixed.acidity	8.3196373	1.7410963	11.30000	5.003127
volatile.acidity	0.5278205	0.1790597	1.46000	3.502189
citric.acid	0.2709756	0.1948011	1.00000	2.188868
residual.sugar	2.5388055	1.4099281	14.60000	4.690431
chlorides	0.0874665	0.0470653	0.59900	2.814259
free.sulfur.dioxide	15.8749218	10.4601570	71.00000	4.127580
total.sulfur.dioxide	46.4677924	32.8953245	283.00000	5.003127
density	0.9967467	0.0018873	0.01362	5.065666
pH	3.3111132	0.1543865	1.27000	4.690431
sulphates	0.6581488	0.1695070	1.67000	3.689806
alcohol	10.4229831	1.0656676	6.50000	4.377736

Each of these variables should have associated units, unless they are explicitly indicated by the name of the variable (e.g. pH), but the data information file does not contain the units. While a chemist might readily know the standard units for these measures, most people will not.

I like to look at the shape of the histograms for each variable. This is usually a quick way to identify stange behavior and/or outliers. There appears to be a mix of bell shapes AND distributions with some skew to the right. Before moving on, we should understand why some of these show skew behavior and keep it in mind before running any analysis.


fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol