CS 199

HW1 - Data representation

February 10, 2014

Sam Laane <laane2@illinois.edu>
José Vicente Ruiz <ruizcep2@illinois.edu>

1 Dataset

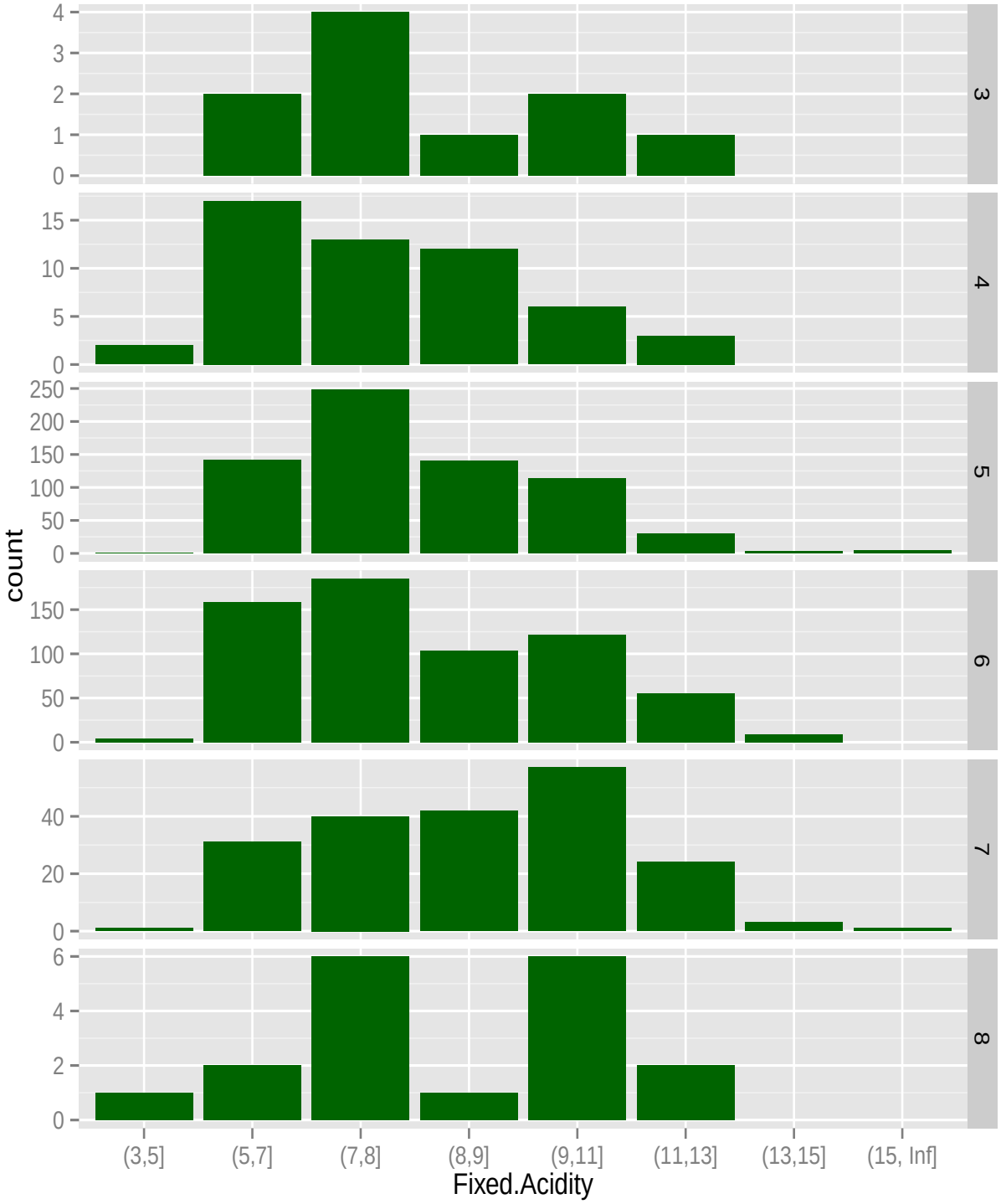
The dataset that we have chosen is about the red wine variant of the Portuguese “Vinho Verde". The number of instances is 1599 and the available features of each of them are the following:

- **Input variables** (based on physicochemical tests):
  1. fixed acidity
  2. volatile acidity
  3. citric acid
  4. residual sugar
  5. chlorides
  6. free sulfur dioxide
  7. total sulfur dioxide
  8. density
  9. pH
  10. sulphates
  11. alcohol
- **Output variable** (based on sensory data):
  12. quality (score between 0 and 10)

So our goal was to finger out what input variables determine the quality of a “good” red wine. Since we don’t have details about the scoring process, we can just trust that is reliable.

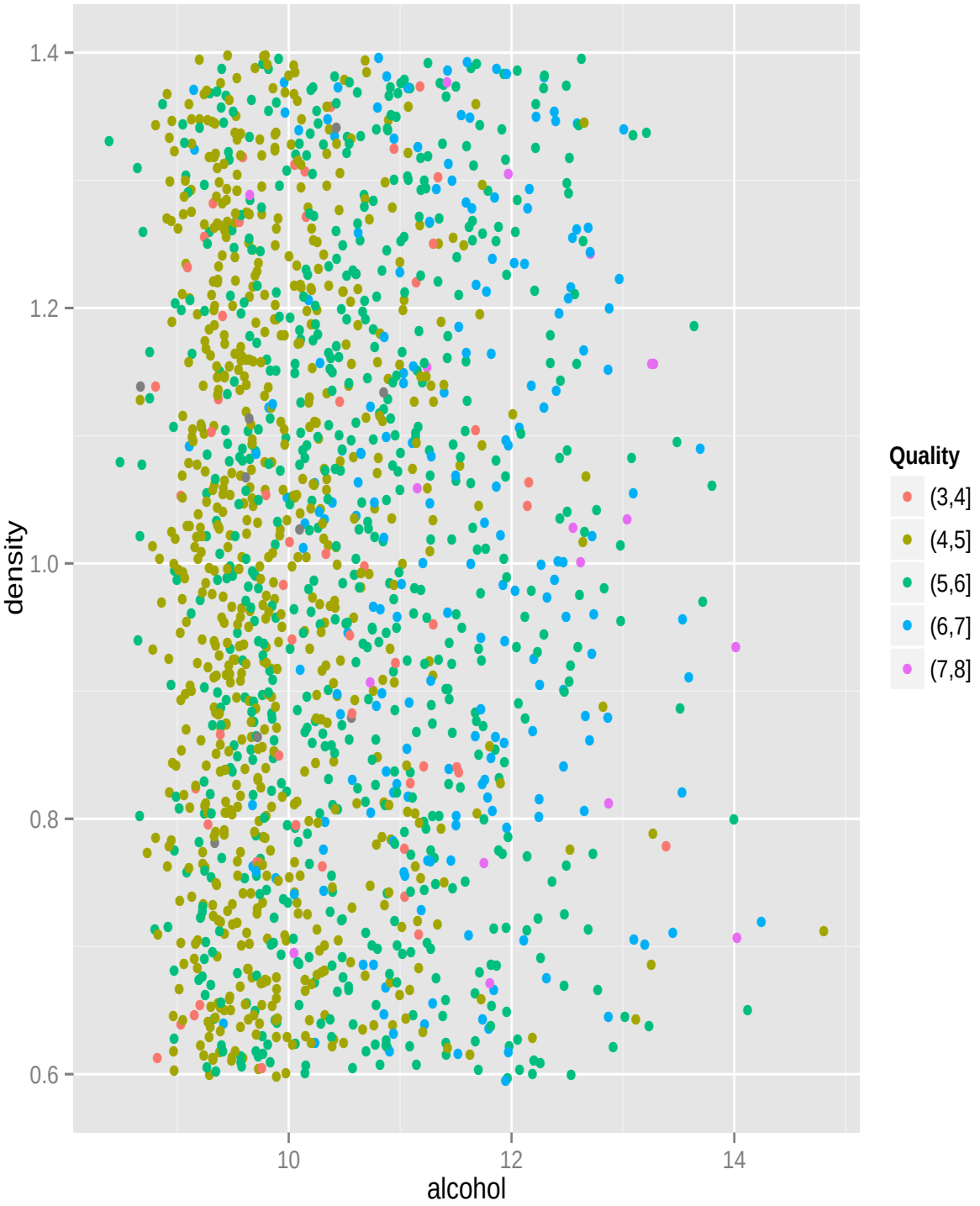
In the beginning, we compared the remaining sugar with respect to the alcohol, however it proved useless. We have omitted the figure since it lacks of interest.

2 Histogram - Fixed acidity and Quality



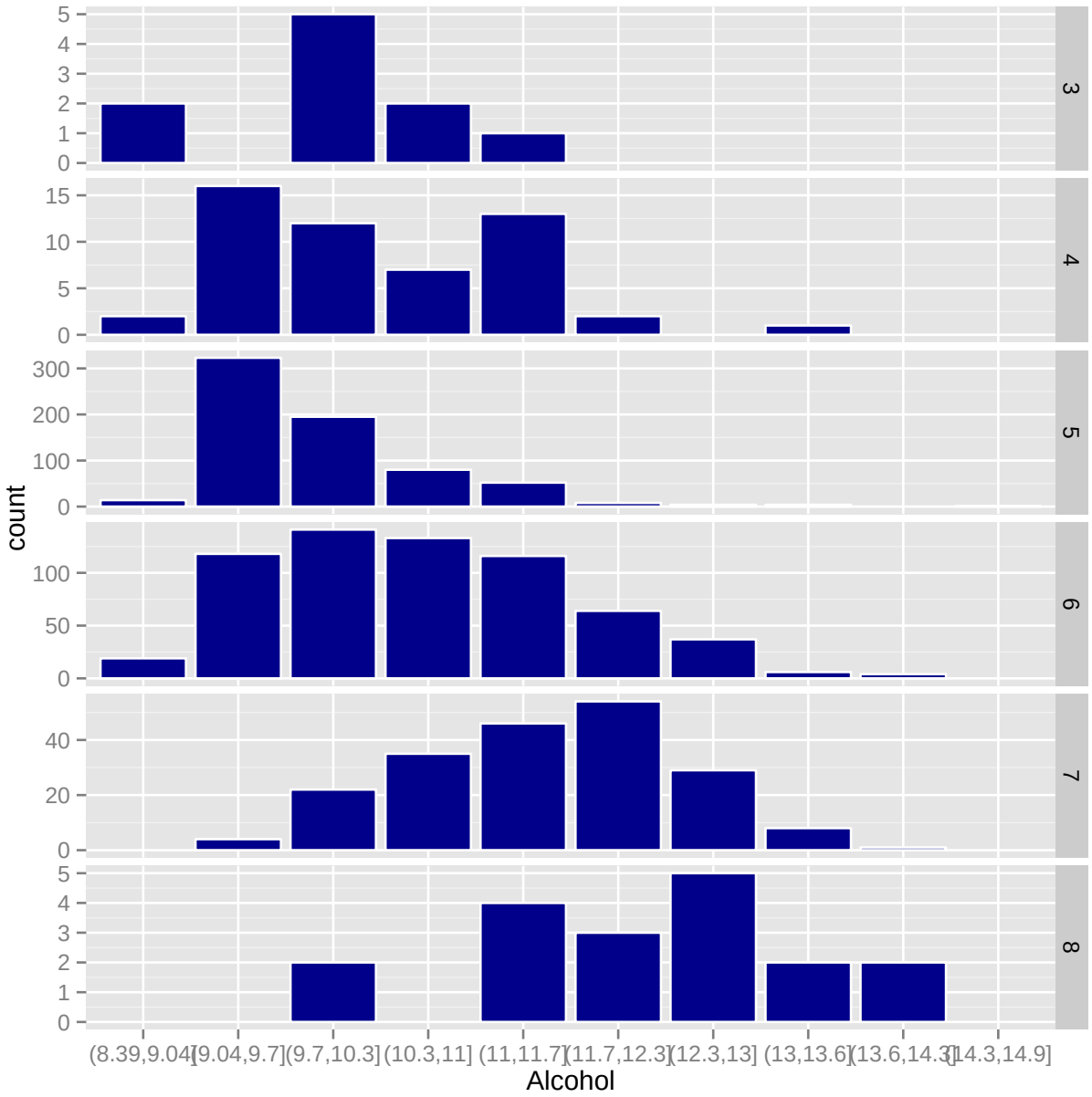
Another thing we have looked into was how the acidity affects the quality. While the over all trend seems to be that acidity does not matter that much we do observe that the highest quality wine has a slightly higher or lower acidity score. It looks like people want a little excitement and find the the 8-9 range boring but don't want there red wines too strong.

3 Scatterplot - Alcohol, Density and Quality



It looks like density does not matter much in terms of alcohol content or quality. Wines have a large but even distribution of densities that have no correlation. Alcohol, however, seems quite interesting. Most wines seem to fall on the lower end of the alcohol scale. However, the few that have highest amounts of alcohol seem to have a higher quality.

4 Histogram - Alcohol and Quality



The problem with our scatter plot was that, though we could see the thin distribution of high alcohol wines were seen as good, it was hard to tell whether or not the spread of good wines was also present under the large amounts of low quality, low alcohol wines.

This bar chart clearly shows that this is not the case. The best wines all seem to have larger amounts of alcohol. I would like to know why alcohol is the sole variable with such an impact. Could it be that the more alcohol the drinkers consumed the more they over looked the wines flaws? Or was it that quality was measured in an economic sense such that the more expensive the vino were said to be higher quality? Maybe people just like the test of alcohol. Whatever the reason it is clean that if one wanted to make a quality red they should focus on alcohol content.