Understanding the Quality of Wine Revealed In Chemistry

Iris Cannary

2024-08-19

Abstract

asdf

Context

The quality of wine can be mystifying despite being of central importance to the practice of making it. However, that quality may be no enigma whatsoever for vintners and sommeliers, whose professions require diverse experience with and deep understanding of wine; and the well-to-do, whose means allow them to have similarly diverse experiences with the potation and the social impetus to seek wines of the highest perceived quality.

Vintners seek to acquire a map to lead their grapes to becoming the most sought-after wines on the market. Wine consumers, like society as a whole, highly value wine as a social drink and a quenching garnish to their meals, and because of that value system pursue ever more pleasing and compatible wines. The most wealthy and high-status of these consumers add to that system the use of the most valuable wines as markers of their wealth and status—consider the banker in the film *The Big Short* asking a colleague "What's with the Dom?", referring to a celebration of a lucrative deal with the famously expensive champagne Dom Pérignon. One may also consider the scene in the limited series *Anatomy of A Scandal* during which a collegiate social club of aristocratic young men launch into a boisterous chant about wasting "Bolly", referring to the similarly expensive Bollinger. Sommeliers, then, have a reputation and business liasonship to maintain as middleperson to the enterprising vintners pursuing firm and lasting footing in the market and the diverse and eager consumers pursuing the best wines they can acquire.

In this study, we aim to leverage the statistical sciences to uncover just *how* a wine, on account of its constituent substances and their combination, achieves a particular level of quality. As society continues its long march from a worldview grounded in religion, myth, and superstition to one built on a fundament of science, the work achieved by this study offers a valuable transition of the basis of wine quality from a subjective realm to an objective one. Foremost among the practical benefits to this are the improved cost efficiency for vintners to adjust their grape growing and winemaking techniques and the bolstered confidence of the consumer when choosing a wine.

Analysis: Data Preparation and EDA

The training data file includes fifteen total variables for 5,463 records in Comma-Separated Values (CSV) format. The testing data set contains 1,034 records of all of the same except for quality values. The target, of course, is quality, which makes the test set peculiar; eleven of the other variables are numeric measures of chemical attributes. The remaining three comprise each record's ID, and then the wine's type and location.

```
# check NAs
sum(is.na(trainFull))
```

```
## [1] 0
```

```
sum(is.na(testFull))
```

```
## [1] 0
```

```r
# find all levels of cat vars
unique(trainFull[["type"]])
```

```
## [1] "white" "red"
```

```r
unique(trainFull[["location"]])
```

```
## [1] "Texas"      "California" "Californnia"
```

```r
unique(trainFull[["quality"]])
```

```
## [1] 5 4 6 7 8 3 9
```

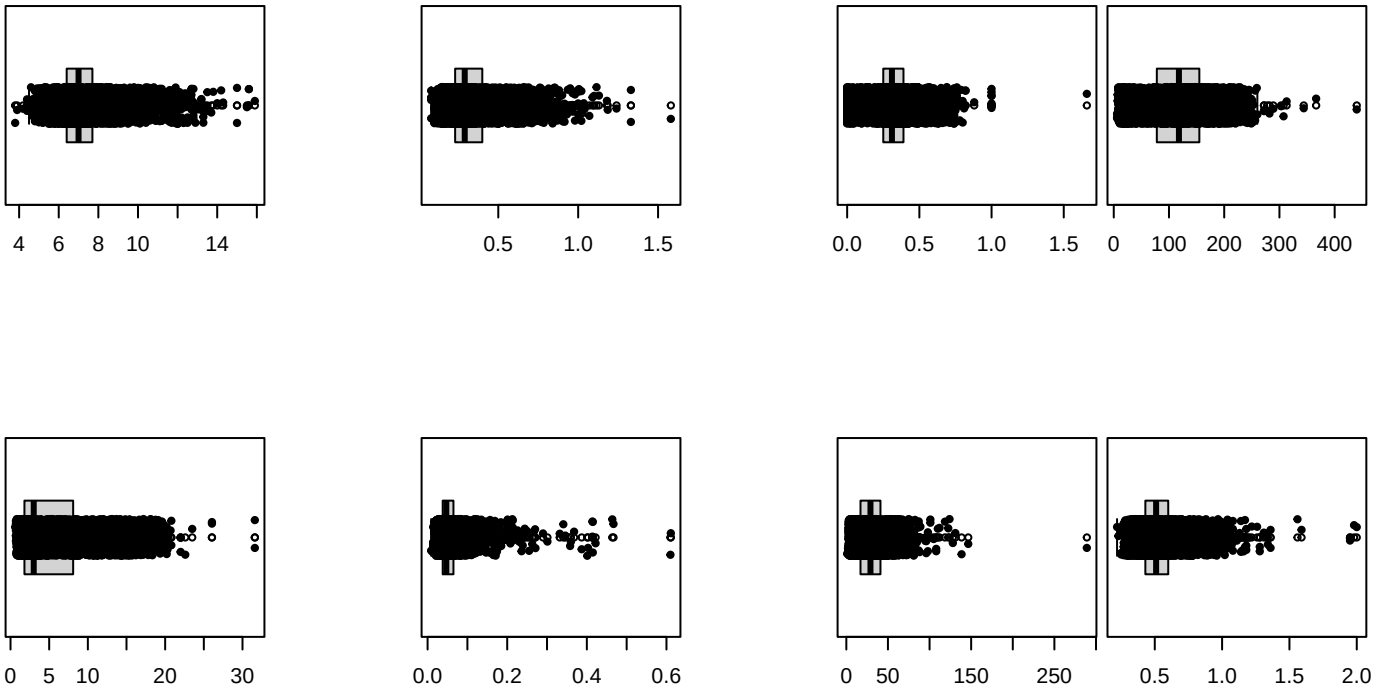
```r
unique(testFull[["type"]])
```

```
## [1] "red"   "white"
```

```r
unique(testFull[["location"]])
```

```
## [1] "California" "Texas"      "Californnia"
```

```r
# find outliers
par(mfrow = c(2, 3))
invisible(lapply(2:6, function(i) {
  boxplot(trainFull[, i], horizontal = TRUE)
  stripchart(trainFull[, i], method = "jitter", pch = 19, add = TRUE)}))
invisible(lapply(7:12, function(i) {
  boxplot(trainFull[, i], horizontal = TRUE)
  stripchart(trainFull[, i], method = "jitter", pch = 19, add = TRUE)}))
```



The data is conveniently quite clean, having no null values at the outset. However, we can see that in the location variable for both data sets, some number of records contain the misspelled state name "Californnia". We can also see by the boxplot-stripchart combination that several of the numerical variables show outliers. One's first instinct may be to only chase down the outliers in the attributes where they are obvious on the plot, but for consistency, every attribute will have any outliers removed subject to the conventional $1.5 \cdot \text{IQR}$

rule, eliminating any data more than 150% of the **i**nter**q**uartile **r**ange under the first quartile or over the third.

```
# fix cat var errors
```

Analysis: Identifying & Predicting Relationships

Revelations