

Project: Wine Quality

*Jovan Burgos
Adrienne Eddy
Chad Michael Eyer
Valorie Myer
Tyler Watson*

08 December 2019

Introduction

Wine testing is important to both sellers and consumers. By testing wine samples, vineyards are able to ensure that their wines are safe for consumers, and that the taste of each type of wine stays consistent. It is even used to help set pricing scales. This project will be comparing different physicochemical properties and sensory elements of wine in order to see what elements impact quality.

The chemical composition of wine can impact both the quality of the wine and our bodies. This study will be looking at 11 different physicochemical properties: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. We will specifically look to see if there is an interaction between volatile acidity and citric acidity. Volatile acidity is how wine spoilage is measured (Neeley, 2004). Typically, wine makers attempt to keep volatile acidity to a minimum to improve the quality of wine (MoreFlavor, n.d.). Citric acid keeps wine tasting fresh, while imparting a slight citric taste to the wine (University of California Davis, n.d.).

Hypotheses

This project answers the following research questions:

- Are variances of quality of red wines equal to variances of quality of white wines?
- Are mean quality of red wines equal to mean quality of white wines?
- Do physicochemical attributes significantly affect quality?
- If so which physicochemical attributes significantly affect quality?
- Is there an interaction between citric acid and volatile acidity for red wines?
- What physicochemical attribute differences might lead to better quality?

Methods

All tests are run at the 95% confidence level unless otherwise noted.

Below, you will find the tests for each hypothesis and a brief explanation and justification for each test.

Variances of quality

We used the *F* test because this test allows us to compare the variances of each data set to see if they are equal. We used this to guide us in determining which tests to use along the way.

Mean quality

Provided that the data met the assumptions, we used *t*-tests because the two sets of data are independent and this test allows us to compare means.

Do any physicochemical attributes affect quality?

We used multivariable regression and a multivariable ANOVA because this allowed us to create a model and analyze this model that will show if there is any relationship between any of the attributes and the quality of the wine. We created pairwise scatterplots which give us a visual idea of what the data looks like.

Which physicochemical attributes affect quality?

We used the results from the multivariable ANOVA from the previous hypothesis to determine which attributes were significant predictors of quality.

Is there an interaction between citric acid and volatile acidity?

We created a regression model to determine if there is an interaction between citric acid and volatile acidity.

physicochemical attributes potentially improving quality

We ran t-tests after testing assumptions to compare physicochemical attributes between red and white wine to determine the kind of difference.

Results

This project compared different physicochemical and sensory elements of wine. The wine in this data set are Vinho Verde wines, which comes from Minho region of Portugal. The data was collected between May 2004 and February 2007 and includes 1599 red wine samples and 4898 white wine samples. This study looked at 11 different physicochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. This data also contains a sensory score variable called “quality.” Each wine sample was tasted by at least 3 judges, each of whom scored the wine from 0 (very bad) to 10 (excellent). The quality variable was the median of the scores (Cortez, Cerdeira, Almeida, Matos, & Reis, 2009).

The first research question asked if the variance of quality was the same for both red and white wines. There is sufficient evidence to suggest that the variability of the quality scores of red and white wines is significantly different, with $p < .001$ at the $\alpha = .05$ level. Because there is a difference in variability, it impacted which tests were run in the following hypothesis tests.

The next research question was to determine if the mean quality of the two types of wines were different. There is sufficient evidence to suggest that the mean quality is different, with $p < .001$ at the $\alpha = .05$ level.

The natural follow-up question is, “Which wine has a greater mean quality?” From several t-tests, it was determined that there is sufficient evidence to suggest that the mean quality of red wine is less than the mean quality of white wine at the $\alpha = .05$ level ($p < .001$).

The next research question was to see if physicochemical properties affect the quality of wine. The resulting multivariable regression line is:

$$\hat{y} = 55.763 + 0.068 \text{ fixed acidity} - 1.328 \text{ volatile acidity} - 0.11 \text{ citric acid} + 0.044 \text{ residual sugar} - 0.484 \text{ chlorides} + 0.006 \text{ free sulfur dioxide} - 0.002 \text{ total sulfur dioxide} - 54.967 \text{ density} + 0.439 \text{ pH} + 0.768 \text{ sulphates} + 0.267 \text{ alcohol}$$

The multivariable ANOVA test (results shown below, along with pairwise comparisons of the different physicochemical attributes) shows that there is sufficient evidence to suggest that wine quality scores are significantly affected by fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. Quality scores are not dependent on citric acid or chlorides.

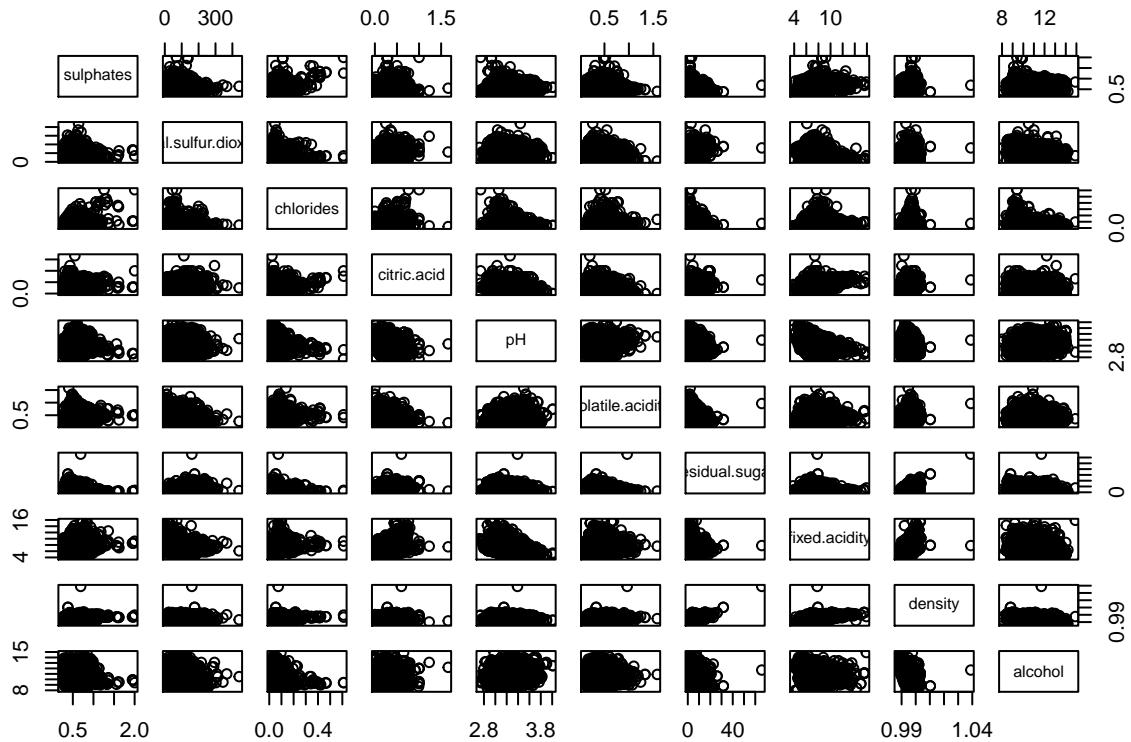
```
kable(h3_anova, digits = 3, format = "pandoc", caption = "ANOVA table",
      align=c('l', rep('r', 5))) # Output ANOVA table
```

Table 1: ANOVA table

	SS	df	MS	F	p
Fixed acidity	10.214	1	10.214	18.890	< 0.0001
Volatile acidity	159.265	1	159.265	294.545	< 0.0001
Citric acid	1.026	1	1.026	1.897	0.1685
Residual sugar	38.595	1	38.595	71.378	< 0.0001
Chlorides	1.143	1	1.143	2.114	0.1460

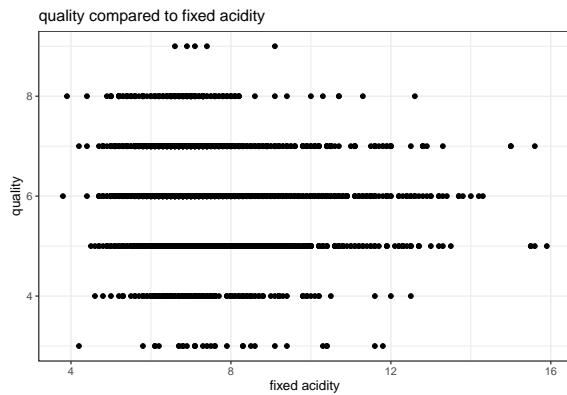
	SS	df	MS	F	p
Free sulfur dioxide	34.155	1	34.155	63.166	< 0.0001
Total sulfur dioxide	43.495	1	43.495	80.441	< 0.0001
Density	11.090	1	11.090	20.509	< 0.0001
pH	12.777	1	12.777	23.630	< 0.0001
Sulphates	55.073	1	55.073	101.853	< 0.0001
Alcohol	137.789	1	137.789	254.829	< 0.0001
Error	3506.531	6485	0.541		
Total	4011.154	6496	505.163		

```
pairs(~ sulphates + total.sulfur.dioxide +
      chlorides + citric.acid + pH +
      volatile.acidity + residual.sugar +
      fixed.acidity + density +
      alcohol, data=wine)
```



Quality compared to fixed acidity

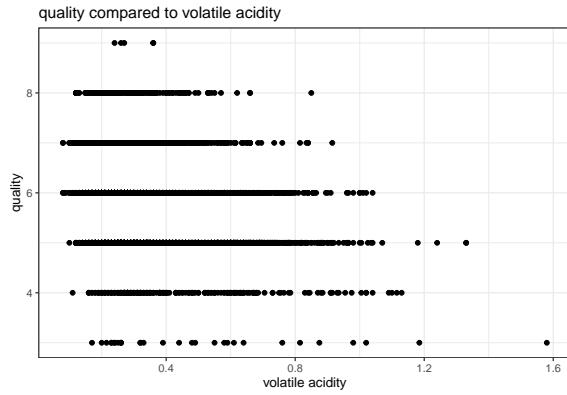
```
wine %>%
  ggplot(aes(x=wine$fixed.acidity, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to fixed acidity")+
  xlab("fixed acidity")+
  ylab("quality")+
  theme_bw()
```



When we look at fixed acidity, we see a slope of 0.068. This means for every 1 unit change of fixed acidity, we can expect to see a 0.068 change in quality. Or, for every 14.775 units change in fixed acidity, we can expect a 1 unit change in quality.

Quality compared to volatile acidity

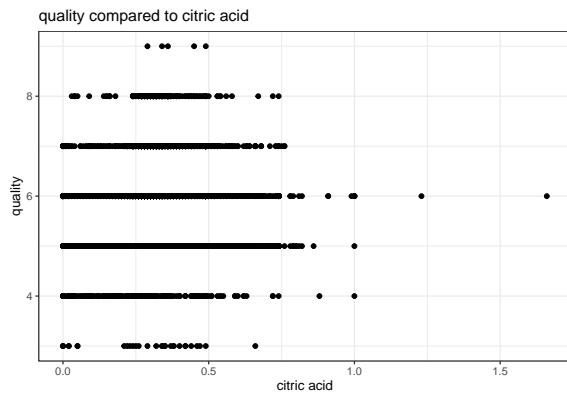
```
wine %>%
  ggplot(aes(x=wine$volatile.acidity, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to volatile acidity")+
  xlab("volatile acidity")+
  ylab("quality")+
  theme_bw()
```



When we look at volatile acidity, we see a slope of -1.328 . This means for every 1 unit change of volatile acidity, we can expect to see a -1.328 change in quality. Or, for every -0.753 units change in volatile acidity we can expect a 1 unit change in quality.

Quality compared to citric acid

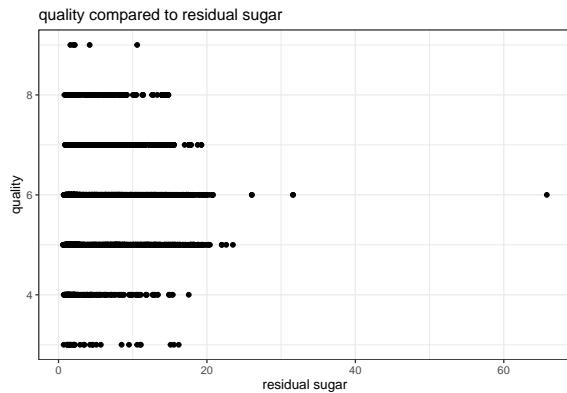
```
wine %>%
  ggplot(aes(x=wine$citric.acid, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to citric acid")+
  xlab("citric acid")+
  ylab("quality")+
  theme_bw()
```



When we look at citric acid, we see a slope of -0.11 . This means for every 1 unit change of citric acid, we can expect to see a -0.11 change in quality. Or, for every -9.119 units change in citric acid, we can expect a 1 unit change in quality.

Quality compared to residual sugar

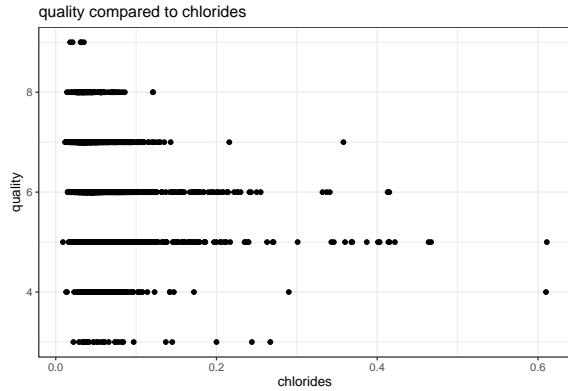
```
wine %>%
  ggplot(aes(x=wine$residual.sugar, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to residual sugar")+
  xlab("residual sugar")+
  ylab("quality")+
  theme_bw()
```



When we look at residual sugar, we see a slope of 0.044 . This means for every 1 unit change of residual sugar, we can expect to see a 0.044 change in quality. Or, for every 22.957 units change in residual sugar, we can expect a 1 unit change in quality.

Quality compared to chlorides

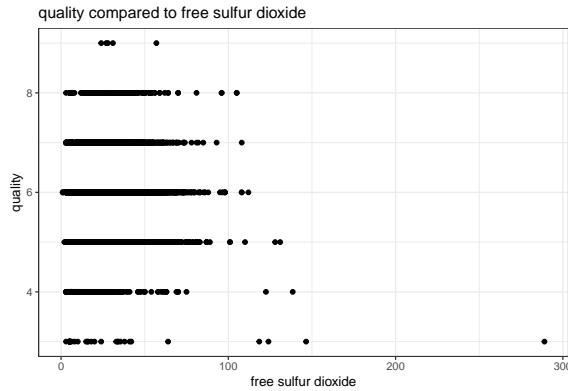
```
wine %>%
  ggplot(aes(x=wine$chlorides, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to chlorides")+
  xlab("chlorides")+
  ylab("quality")+
  theme_bw()
```



When we look at chlorides, we see a slope of -0.484 . This means for every 1 unit change of chlorides, we can expect to see a -0.484 change in quality. Or, for every -2.067 units change in chlorides, we can expect a 1 unit change in quality.

Quality compared to free sulfur dioxide

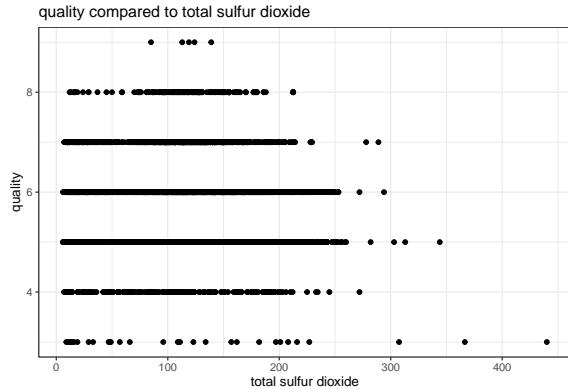
```
wine %>%
  ggplot(aes(x=wine$free.sulfur.dioxide, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to free sulfur dioxide")+
  xlab("free sulfur dioxide")+
  ylab("quality")+
  theme_bw()
```



When we look at free sulfur dioxide, we see a slope of 0.006 . This means for every 1 unit change of free sulfur dioxide, we can expect to see a 0.006 change in quality. Or, for every 167.507 units change in free sulfur dioxide, we can expect a 1 unit change in quality.

Quality compared to total sulfur dioxide

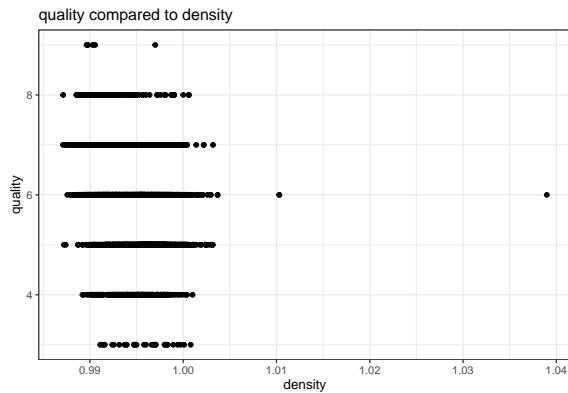
```
wine %>%
  ggplot(aes(x=total.sulfur.dioxide, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to total sulfur dioxide")+
  xlab("total sulfur dioxide")+
  ylab("quality")+
  theme_bw()
```



When we look at total sulfur dioxide, we see a slope of -0.0025 . This means for every 1 unit change of total sulfur dioxide, we can expect to see a -0.0025 change in quality. Or, for every -403.015 units change in total sulfur dioxide, we can expect a 1 unit change in quality.

Quality compared to density

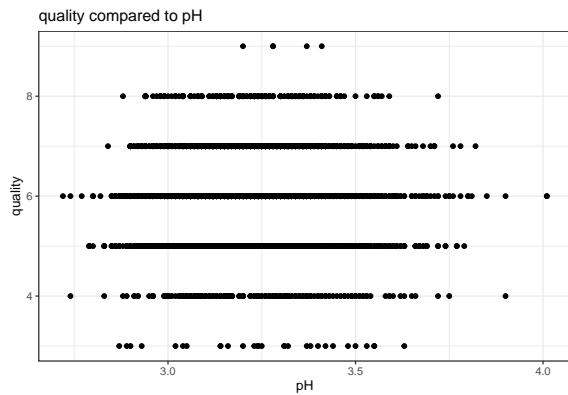
```
wine %>%
  ggplot(aes(x=wine$density, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to density")+
  xlab("density")+
  ylab("quality")+
  theme_bw()
```



When we look at density, we see a slope of -54.967 . This means for every 1 unit change of density, we can expect to see a -54.967 change in quality. Or, for every -0.018 units change in density, we can expect a 1 unit change in quality.

Quality compared to pH

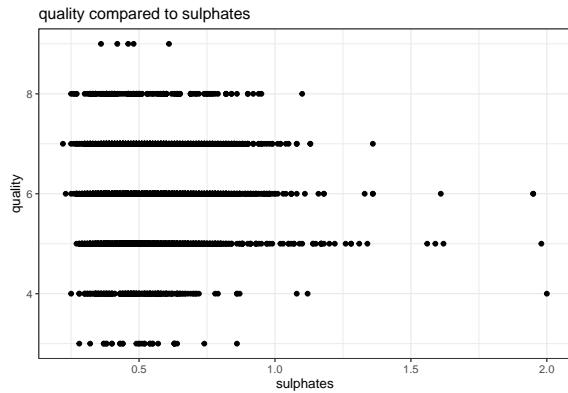
```
wine %>%
  ggplot(aes(x=wine$pH, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to pH")+
  xlab("pH")+
  ylab("quality")+
  theme_bw()
```



When we look at pH, we see a slope of 0.439. This means for every 1 unit change of pH, we can expect to see a 0.439 change in quality. Or, for every 2.276 units change in pH, we can expect a 1 unit change in quality.

Quality compared to sulphates

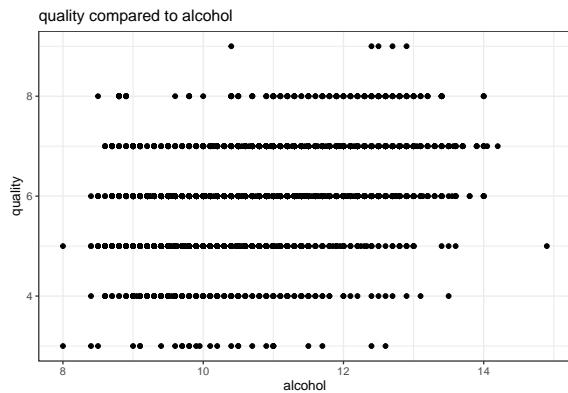
```
wine %>%
  ggplot(aes(x=wine$sulphates, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to sulphates")+
  xlab("sulphates")+
  ylab("quality")+
  theme_bw()
```



When we look at sulphates, we see a slope of 0.768. This means for every 1 unit change of sulphates, we can expect to see a 0.768 change in quality. Or, for every 1.302 units change in sulphates, we can expect a 1 unit change in quality.

Quality compared to alcohol

```
wine %>%
  ggplot(aes(x=wine$alcohol, y=wine$quality
             )) +
  geom_point() +
  labs(title="quality compared to alcohol")+
  xlab("alcohol")+
  ylab("quality")+
  theme_bw()
```



When we look at alcohol, we see a slope of 0.267. This means for every 1 unit change of alcohol, we can expect to see a 0.267 change in quality. Or, for every 3.745 units change in alcohol, we can expect a 1 unit change in quality.

Winemakers use citric acid to sterilize equipment between batches, and it is found in fruitier wines. However, citric acid can contribute to volatile acidity which leads to wine spoilage. Even though citric acid is not a significant predictor of quality, we wanted to see if there is an interaction between citric acid and volatile acidity. There is sufficient evidence to suggest that there is an interaction between citric acid and volatile acidity because $p = 0.0495$.

Knowing that all the physicochemical properties except for citric acid or chlorides are significant predictors of quality and that white wine has the higher quality, the last research question was to see how the physicochemical composition differs between white and red wines. There is sufficient evidence to suggest that white wine has less fixed acidity ($p < .001$), less volatile acidity (spoilage) ($p < .001$), more residual sugar ($p < .001$), greater free sulfur dioxide ($p < .001$), greater total sulfur dioxide ($p < .001$), less dense ($p < .001$), lower pH ($p < .001$), fewer sulphates ($p < .001$), and more alcohol ($p = 0.0021$).

Conclusion

This study of Portuguese wines shows that winemaking is both an art and a science. As vineyards strive to develop better tasting wines, they need to consider the physicochemical make-up of their product. The study found that the physicochemical properties of wine play an important part in the quality. Out of the eleven physicochemical properties studied, all but two, citric acid and chlorides, significantly predicted the quality of the wine. The data from this study showed that the white wine quality was better than red wine quality. When we looked at how the red and white wines differed in physicochemical properties, we found that white wine has less fixed acidity, less volatile acidity (spoilage), more residual sugar, greater free sulfur dioxide, greater total sulfur dioxide, less dense, lower pH, fewer sulphates, and more alcohol. Another notable find is that these physicochemical properties don't just impact the quality of the wines—they also impact each other, as seen by the interaction between citric acid and volatile acidity. These are important concepts to keep in mind for large vineyards all the way to those who want to try making their own wine at home.

One of the limitations of this study was that all the wine was from the same region of Portugal. Another limitation was that the quality variable was based on a score of small groups of judges—only 3 judges in some cases. Something important to keep in mind when considering the quality score is that the score is based on the median, rather than the average, of the judges' scores.

Another limitation was our regression model. We ran a lack of fit test and found that the model, while showing several attributes being significant, was not an appropriate model ($p < .001$).

Future studies could be done to include wines from different parts of the globe and to include a larger panel of judges.

Session info

```
sessionInfo()

## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin19.0.0 (64-bit)
## Running under: macOS Catalina 10.15.2
##
## Matrix products: default
## BLAS/LAPACK: /usr/local/Cellar/openblas/0.3.7/lib/libopenblas-r0.3.7.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats      graphics   grDevices utils      datasets   methods    base
##
## other attached packages:
## [1] knitr_1.26     alr3_2.0.8     car_3.0-3      carData_3.0-2
## [5] pander_0.6.3   BSDA_1.2.0     lattice_0.20-38 forcats_0.4.0
## [9] stringr_1.4.0   dplyr_0.8.3    purrr_0.3.2    readr_1.3.1
## [13] tidyverse_1.2.1 tidyr_0.8.3    tibble_2.1.3   ggplot2_3.2.1
## [17] tidyselect_0.2.5 xfun_0.11      haven_2.1.1    colorspace_1.4-1
## [21] generics_0.0.2   vctrs_0.2.0    htmltools_0.4.0 yaml_2.2.0
## [25] rlang_0.4.0     e1071_1.7-2    pillar_1.4.2    foreign_0.8-71
## [29] glue_1.3.1      withr_2.1.2    modelr_0.1.5    readxl_1.3.1
## [33] munsell_0.5.0   gtable_0.3.0    cellranger_1.1.0 zip_2.0.4
## [37] rvest_0.3.4     evaluate_0.14   labeling_0.3    rio_0.5.16
## [41] curl_4.0        class_7.3-15   highr_0.8      broom_0.5.2
## [45] Rcpp_1.0.2      scales_1.0.0    backports_1.1.4 jsonlite_1.6
## [49] abind_1.4-5     hms_0.5.1      digest_0.6.20   openxlsx_4.1.0.1
## [53] stringi_1.4.3   grid_3.6.1     cli_1.1.0      tools_3.6.1
## [57] magrittr_1.5     lazyeval_0.2.2   crayon_1.3.4    pkgconfig_2.0.2
## [61] zeallot_0.1.0   data.table_1.12.6 xml2_1.2.1    lubridate_1.7.4
## [65] assertthat_0.2.1 rmarkdown_1.14   httr_1.4.1     rstudioapi_0.10
## [69] R6_2.4.0        nlme_3.1-140   compiler_3.6.1
```

References

- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547–553. doi:10.1016/j.dss.2009.05.016
- MoreFlavor. (n.d.). Citric acid. Retrieved from <https://morewinemaking.com/products/citric-acid.html>
- Neeley, E. (2004). Wine spoilage is legally defined by volatile acidity, largely composed of acetic acid. Retrieved from <https://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity>
- University of California Davis. (n.d.). Citric acid. Retrieved from <https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid>