Principal Component Analysis (PCA) and Statistical Tests Using Factoshiny and R Commander.

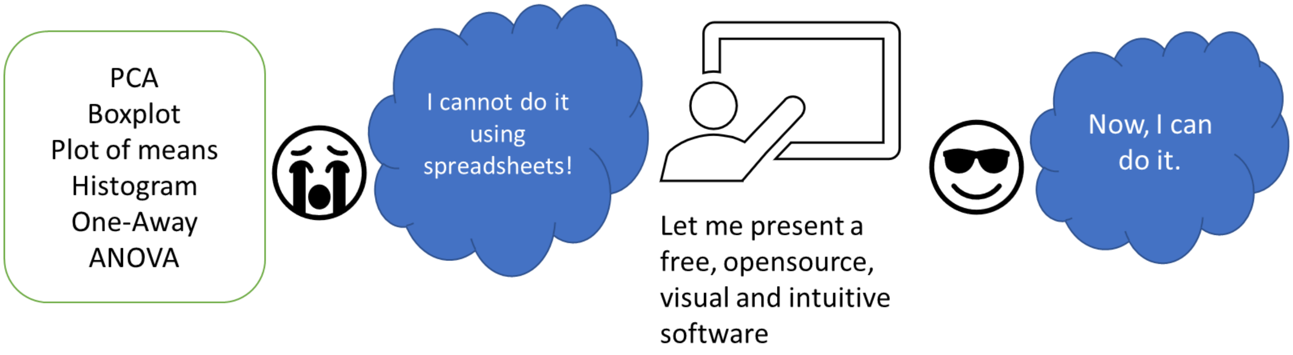
Matheus Fernandes Filgueiras[‡] and [‡]Endler Marcel Borges[‡]

[‡]*Departamento de Química, Fundação Universidade Regional de Blumenau, FURB, Campus 1, Rua Antônio da Veiga, 140, Victor Konder, 89012-900 Blumenau-SC, Brasil.*

ABSTRACT

Spreadsheets are commonly used for data handling. However, in huge data sets, spreadsheets cannot do statistical tests, such as one-way ANOVA, boxplot, plot of means, principal component analysis (PCA). Most of the students had never worked with programming software such as MATLAB, Phyton, Octave and R project. Hence, in this lab experiment, students analyzed large data sets using R Commander and Factoshiny plugins. Commander and Factoshiny are packages which gives graphical user interface (GUI). GUI plugins allows students with no programming knowledge to run statical tests quickly and easily without having to type a single command line. The class was divided into three parts. First, students analyzed a red wine data set (1599 samples, 11 physicochemical variables, and one qualitative variable) to find correlations between wine quality (qualitative variable) and its physicochemical variables (quantitative variable). Second, they analyzed a white wine data set (4898 samples, 11 physicochemical variables, and one qualitative variable) to find correlations between white wine quality and its physicochemical variables. Third, they analyzed a red wine and white wine data set and found correlations between wine's physicochemical variables and their quality and type. Statistical tests and PCA were carried out using R Commander and Factoshiny, respectively. Due to the graphical interface and simplicity of these two plugins, the class can be concluded in 200 min.

GRAPHICAL ABSTRACT



PCA
Boxplot
Plot of means
Histogram
One-Away
ANOVA

I cannot do it using spreadsheets!

Let me present a free, opensource, visual and intuitive software

Now, I can do it.

KEYWORDS:

Graduate Education / Research < Audience

Upper-Division Undergraduate < Audience

Analytical Chemistry < Domain

Chemoinformatics < Domain

Interdisciplinary / Multidisciplinary < Domain

Chemometrics < Topics

Computational Chemistry < Topics

Introduction

As R projetc is freeware, it does not offer a ‘pretty-face’ and students must type a series of line command lines to run statistical tests. However, there are many graphical user interface (GUI) plugins which run into R-project, these plugins have push-button menus like contemporary software. R Commander is the most popular GUI for R project.[1]

We did not have any free software with an easy-to-use graphical interface for data handing, so we published our first paper describing how to do principal component analysis (PCA) with Factoshiny.[2] After that, we had to do statistical tests and plots which were difficult to do with spreadsheets, such as one-way ANOVA, boxplot, plot of means and histograms. Therefore, we have released this manuscript that demonstrates how to do statical tests and graphs using R Commander and Factoshiny.

Statistical tests were carried out using the R Commander plugin[3] and exploratory analysis was carried out using the Factoshiny plugin.[4,5] Both are plugins that run inside the R project.

Most of our students do not have experience with programming, and they can use these apps without typing a single command. In addition, the R project is free, open-source, widely available, and widely used, making it accessible to all students.

Students may do several statistical tests and PCA using these plugins without having to type a single command. Both plugins have a graphical interface and a friendly interface.

Statistical tests were carried out using the R Commander[3] and exploratory analysis was carried out using the Factoshiny.[4] Both are plugins that run inside the R project and have a visual interface. R Commander is an R package dedicated to statistical analysis of large data sets.[3]

PCA is a mathematical approach that converts a data matrix into a new space of principal components (PCs). It is a multivariate analysis technique that transforms the data and reduces its dimensionality into a form that highlights the variance of the data. It reveals the underlying patterns and groups within the data. This is achieved by constructing a new set of variables obtained from linear combinations of the original ones. The new variables are projection coefficients of the old variables in new PCs.[6]

There are several examples of PCA as a learning model in this *Journal* such as analysis of stainless-steel standard using laser-induced breakdown spectroscopy,[7] classification of vegetable oils using FTIR,[8] analysis of air quality,[9] identification of edible oils using ^{1}H NMR,[10–12] classification of river sediments using its heavy metals concentration[13] Identification of the arabica and robusta varieties of green coffee beans using ^{1}H NMR,[14] recognizing the origin of sand specimens by diffuse reflectance infrared fourier transform spectroscopy,[15] Classification of representative elements[16] and lanthanides[17] using its chemical properties, discover chocolates origin using HPLC,[18] metabolomics of horse blood using HPLC-MS,[19]

When we were working with a large data set and looking to extract relevant information from it, we could use boxplots and plots of means, because these plots provide better visualization of the data set than spreadsheets. [20,21] However, spreadsheets cannot do boxplots and graphics of means, where these plots can be done using R Commander. Here, it was shown that correlations between quantitative and qualitative variables obtained using PCA could be explored using these plots.

Experimental Overview

Here, we choose examples of wine analysis. First, students analyze a data set containing 11 quantitative variables (physicochemical properties) and one qualitative variable (quality) for 1599 red wines (Table S1). They found correlations between red wine quality and its quality using PCA. Then, these correlations were analyzed using one-away ANOVA, pie charts, histograms, plot of means and boxplots.

In the second example, students found correlations between physicochemical properties and white wine quality as done in the first example, but the data set contains information about 4898 white wines (Table S2).

In the third example, students found correlations between wine type (red and white) and its quantitative variables (Table S3).

The data sets were imported from form Microsoft Excel into R Commander and the data set was used directly in Factoshiny. R Commander recognizes numbers as quantitative variables and words (such as red and white) as qualitative variables. Thus, when data sets were imported "quality" (1 to 10) must be assigned as a qualitative variable. R Commander can import data set in different formats such as SPSS, SAS, Minitab, STATA, and Excel.

We have not described the mathematical background of PCA and boxplots, but a description of PCA has been shown in reviews,[22–26] and boxplots were explained in papers previously published in this *Journal*.[20,27]

Students' Learning Goals

Students run PCA in large wine data sets using Factoshiny. As a result of performing PCA in those data sets, it is expected that students:

- Identify the importance of each variable in the loading plot
- Identify correlations between variables in loading plot
- Identify similarities and dissimilarities among samples in the score plot
- Understand the influence of each variable in the position of samples in the score plot.
- Found correlations between physicochemical properties and wine quality
- Found correlations between physicochemical properties and wine type.

After running the PCA, students use R Commander to perform a statistical analysis on the data set, guiding the selection of the most important variables to differentiate wine samples based on their quality. As a result of performing PCA in those data sets, it is expected that students:

- Recognize differences between means using one-away ANOVA
- Build pie charts, boxplots, plots of means, and histograms.
- Understand how each physicochemical property was correlated with wine quality and type using boxplots and graphic of means.

Materials and Methods

Data set

In Table S1 to S3, the data set explored is a collection of white and red wine from the "Vinho Verde" wine region in Portugal.[28–30] The data set for red wine consists of 1599 samples (Table S1) and the one for white wine consists of 4898 samples (Table S2). Each sample has 11 physicochemical variables: Fixed acidity (g(tartaric acid)/dm^3), Volatile acidity (g(acetic acid)/dm^3), Citric acid (g/dm^3), Residual sugar (g/dm^3), Chlorides (g(sodium chloride)/dm^3), Free sulfur dioxide (mg/dm^3), Total sulfur dioxide (mg/dm^3), Density (g/cm^3), pH, Sulphates (g(potassium sulphate)/dm^3), Alcohol (vol.%) and one qualitative variable: quality. Quality was determined by sensory assessors, they (a minimum of three) used blind tastes to grade the wine on a scale of quality that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations. This data set is free and available on the web. Table S3 contains red and white samples, it has 11 quantitative variables and two qualitative variables: quality and type, where the type is red or white.[28]

HAZARDS

In this dry lab, there are no hazards

Results and Discussion

Red wines PCA

Analyzing the data for red wines (Table S1), we have 11 physicochemical variables (quantitative variables) and one qualitative variable (quality). There was any relation between 11 physicochemical variables and red wine quality? How could I visualize these 11 variables at the same time?

PCA is a mathematical tool that reduces the data dimensionality, making its visualization possible, while retaining as much information as possible that was already present in the original data.[9] The loading plot (Figure 1) and the score plot (Figure 2) were used to answer these questions

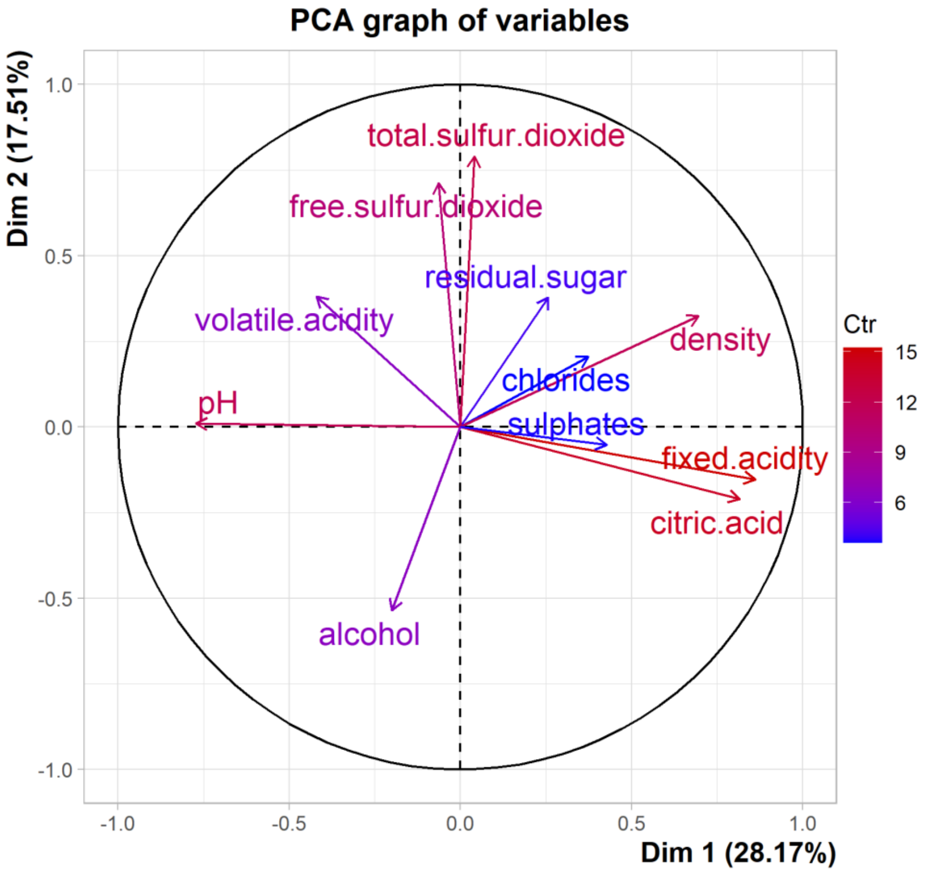


Figure 1: Loading plot built using data set in Table S1. PC1 vs PC2

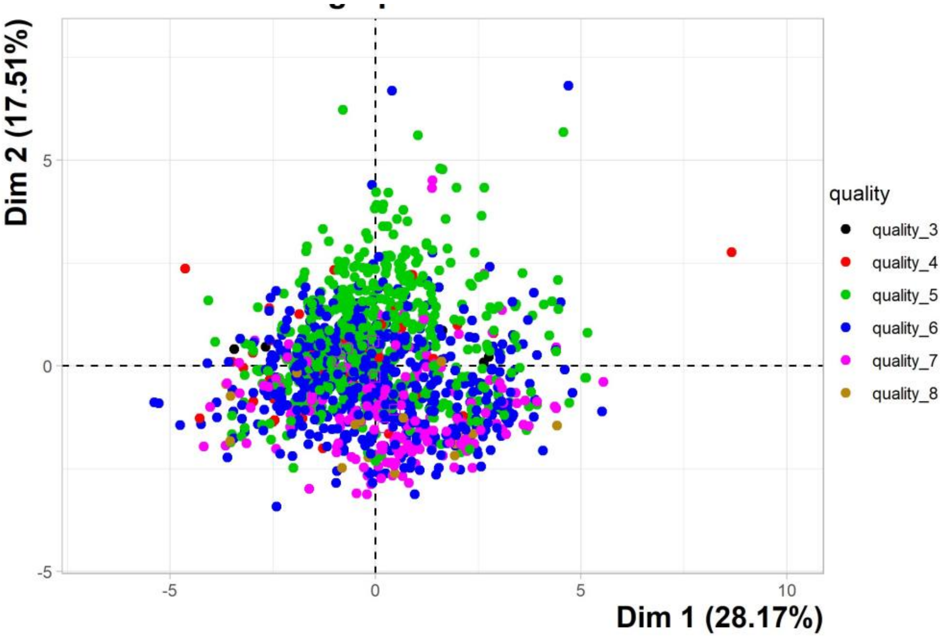


Figure 2: Score plot built using data set in Table S1, PC1 vs PC2

The PCA employs a mathematical procedure that transforms a set of possibly correlated variables into a new set of uncorrelated variables, called principal components, PCs.[2,31] The amount of each of the original variables included in the PC is described by the loading (Figure 1). By plotting the loadings for the two PCs, it was possible to assess the relative importance of each of the variables in the PCA model, variables with higher impact have larger vectors than variables with lower impact. In Figure 1, total sulfur dioxide, free sulfur dioxide, density, fixed acidity, citric acid, and pH (red) were variables with a larger impact in the PCA model. Volatile acidity and alcohol were variables (purple) with intermediate impact. Sulphates, chlorides, and residual sugar were variables with a smaller impact in the PCA model.

In the loading plot (Figure 1), correlations among variables were observed in the loading plot, positively correlated variables were located close together, and inversely correlated variables were at 180° to one another,[2,32] for example, fixed acidity and citric acid are directly correlated, both variables were inversely correlated to pH. Total sulfur dioxide and free sulfur dioxide were directly correlated.

The score plot (Figure 2) showed the projection of samples (red wine) along with the principal components (PCs). The location of wine samples in the score plot was directly related to the loading plot (Figure 1). The loading plot showed that samples on the right-hand side have larger density, fixed acidity, citric acid sulphates, and chlorides than samples placed on the left-hand side, while samples on the left-hand side have larger pH and volatile acids than samples placed on the right-hand side.

Samples on the top side have larger total sulfur dioxide, free sulfur dioxide, and residual sugar than samples placed on the bottom, while samples placed on the bottom have larger alcohol than samples placed on the top (Figure 2).

In Figure 2, most of the high-quality red wines (quality 6, 7, and 8) were placed on the bottom, and low-quality red wines (quality 3, 4, and 5) were placed on top. Thus, we can conclude that high-quality red wines have larger alcohol and smaller total sulfur dioxide and free sulfur dioxide than low-quality red wines.

Red wines data set statistical evaluation

A series of statical tests and plots can be carried out using R Commander. A pie chart (Figure 3) showed the number of red wine samples on each quality (Table S1). Most of the red wine samples were quality 5, 6 and 7, quality 3 and 8 had few samples.

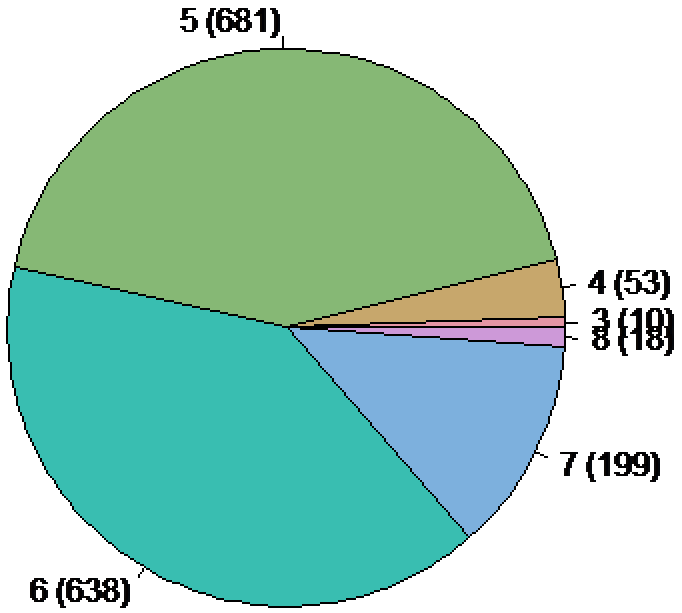


Figure 3: Pie chart of red wines quality (Table S1)

One-way ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. Spreadsheets do one-way ANOVA between columns and lines, for example, we could insert alcohol of quality 4 in one line, quality 5 in another one, and so on. Then, we could do it using a spreadsheet, but in the format that the data was distributed in Table S1, it was a hard task. The R Commander does one-way ANOVA for each qualitative variable in the data set, and differences between means could be easily checked without any manipulation of the data set.

The one-way ANOVA of the 11 physicochemical variables for each red wine quality (**Table 1**) shown that just residual sugar had equivalent means for each red wine quality, while alcohol was the physicochemical variable that had the larger difference between means.

Table 1: One-away ANOVA of 11 physicochemical variables for each red wine quality in Table S1

Physicochemical variable	F	p-value
alcohol	115.9	2E-16
chlorides	6.036	1.53E-05
citric acid	19.69	2.00E-16
density	13.4	8.12E-13
fixed acidity	6.283	8.79E-06
Free sulfur dioxide	4.754	2.57 E-04
pH	4.342	6.28 E-04
residual sugar	1.053	0.385
sulphates	22.27	>2E-16
total sulfur dioxide	25.48	>2E-16
volatile acidity	60.91	>2E-16

The boxplot of residual sugars (Figure 4) showed that means for each red wine quality were equivalent, and qualities 5, 6, and 7 have several outlier samples. The plot of means (Figure 5) showed that red wines quality cannot be differentiated using residual sugar due to its close means and larger standards deviations. The plot of means (Figure 4) was a useful tool for visualizing the means and standard deviations of residual sugars according to their quality, but outliers and interquartile range were not shown in these plots (Figure 5).

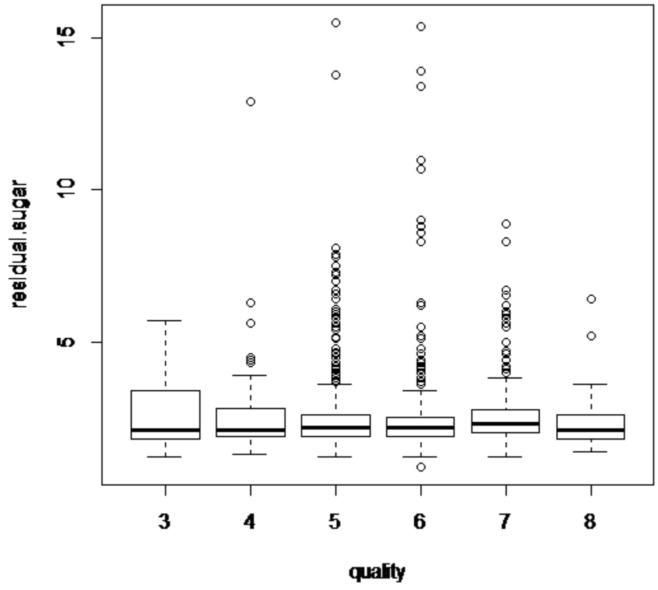


Figure 4: Boxplot of the residual sugar in red wines

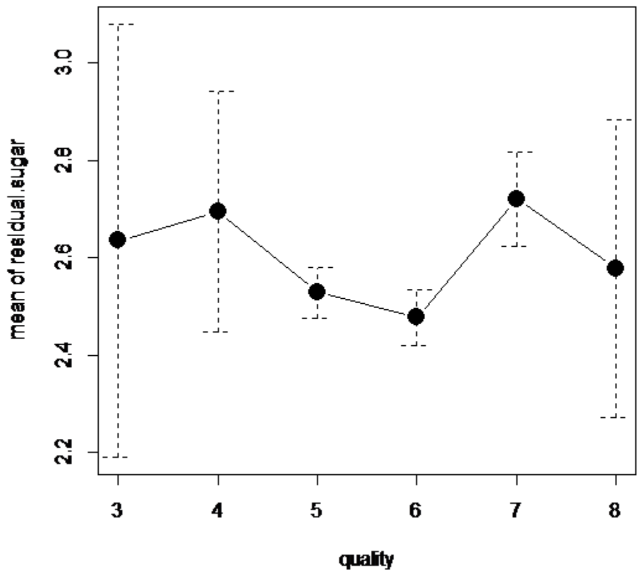


Figure 5: Plot of means of residual sugar in red wines

The boxplot of alcohol (Figure 6) showed that qualities 3, 4, and 5 had equivalent alcohol content. In qualities 6, 7, and 8, the mean alcohol increases as the wine quality increases. It was difficult to distinguish quality five from the other qualities because it has many outliers.

In quality 6, 7, and 8, the graphic of means (Figure 7) showed a positive correlation between red wine quality and alcohol. It also showed that qualities 3, 4, and 5 had equivalent alcohol means.

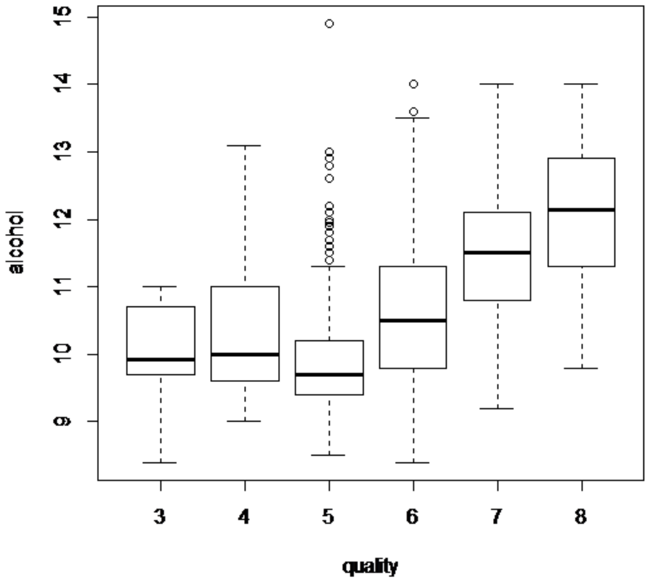
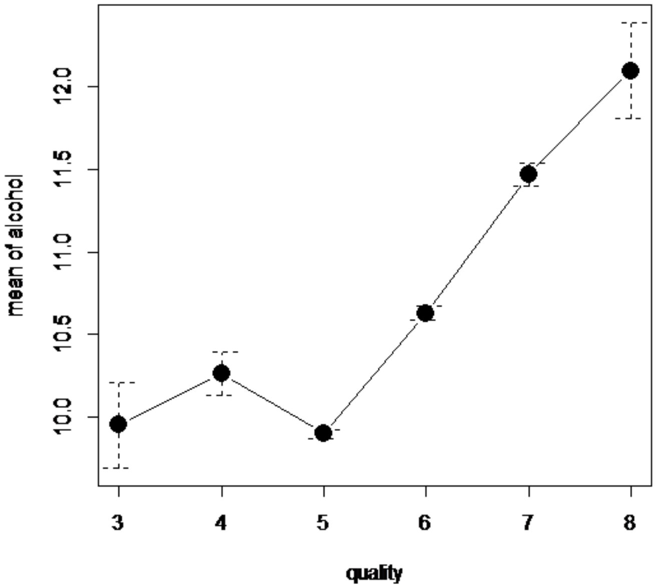
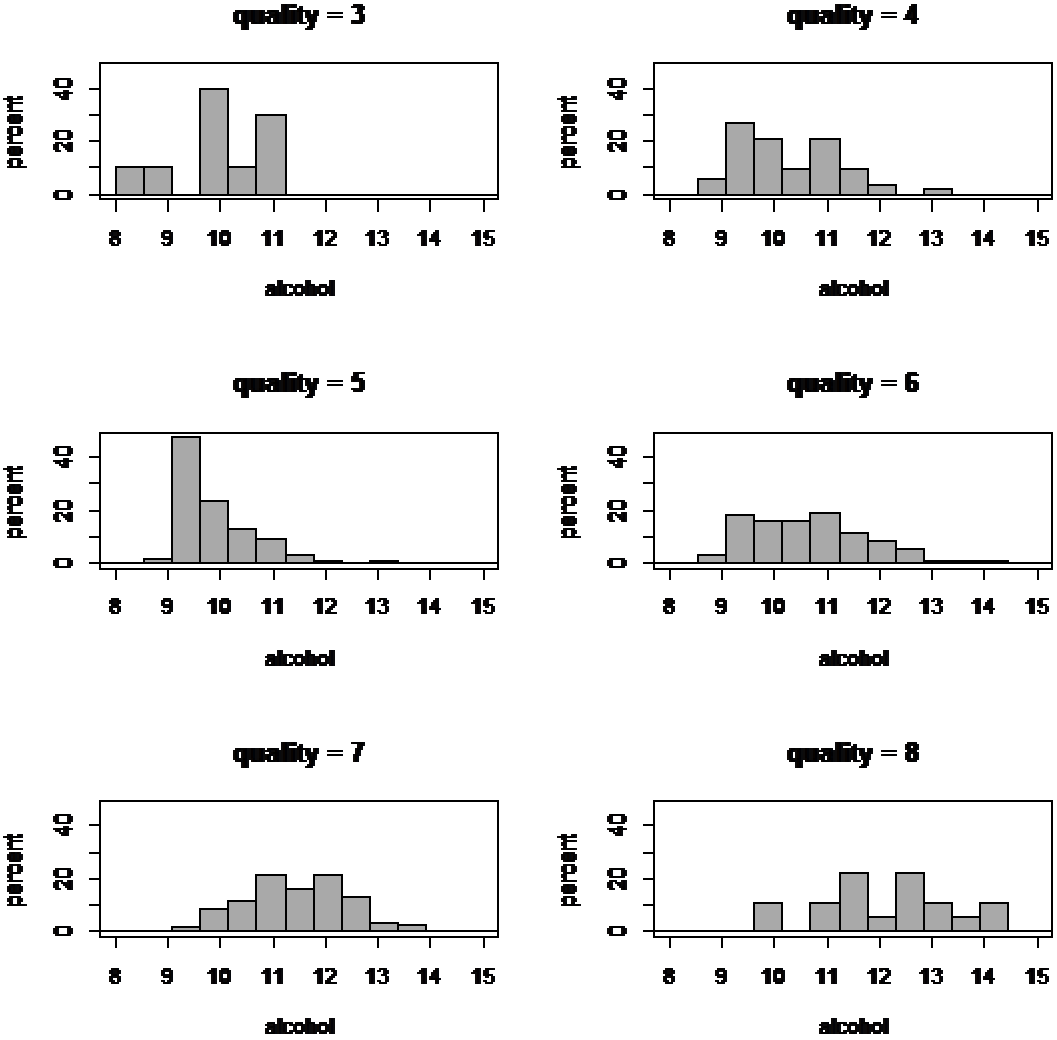


Figure 6: Boxplot of alcohol in red wines (Table S1)



190 Figure 7: plot of means for alcohol in red wines (Table S1)

The R Commander also does histograms of each variable according to its qualitative variables. The histogram of alcohol (Figure 8) showed that quality 5 alcohol follows a logarithm distribution. In qualities 4, 6, and 7 alcohol follows a Gaussian distribution. In qualities 3 and 8, we cannot observe the kind of distribution due to its small number of



samples.

Figure 8: Alcohol histogram of red wine quality (Table S1)

The plot of means showed a series of trends of each physicochemical variable according to wine quality. High-quality red wines have small chlorides. In quality 5 to 8, there was an inverse correlation between quality and chlorides. Quality 4 cannot be differentiated from other qualities because it has a larger standard deviation (Figure 9).

pH affects wine's taste and lifetime. Quality 3 and 4 have a larger pH than other qualities, but quality eight cannot be differentiated from qualities 5, 6, and 7 due to its higher standard deviation (Figure 10).

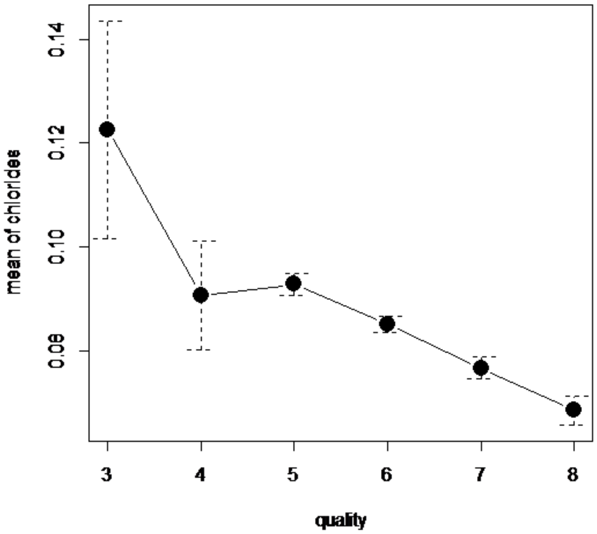


Figure 9: Chloride plot of means of red wines (Table S1)

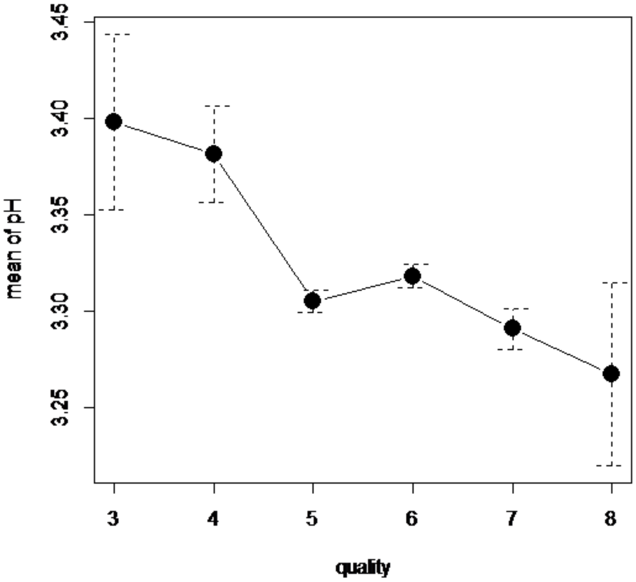


Figure 10: pH plot of means of red wines (Table S1)

Withe wines PCA

The loading plot (Figure *11*). showed that white wine samples placed on the right-hand side have larger free sulfur dioxide, total sulfur dioxide, residual sugar, density, and chlorides than samples placed on the left-hand side, while samples placed on the left-hand side have larger alcohol than samples placed on the right-hand side. Samples placed on the top side have higher pH than samples placed on the bottom, while samples on the bottom have larger fixed acidity and citric acid than samples placed on the top (Figure *11*).

pH was inversely correlated with citric acid and fixed acidity. It was expected that wines that had higher citric acid and fixed acidity would have a lower pH than samples that had smaller citric acid and fixed acidity (Figure *11*). As expected, density and residual sugar were correlated, while alcohol and density were inversely correlated (Figure *11*).

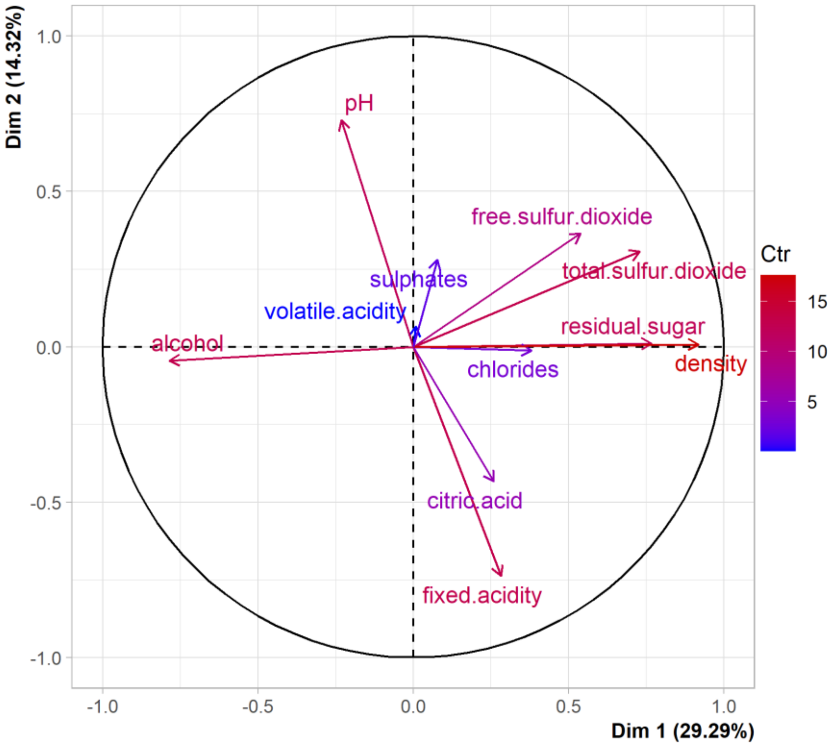


Figure 11: Loading plot built using data set in Table S2, PC1 vs PC2

The score plot (Figure *12*) showed that high-quality white wines were placed on the left-hand side, which means that high-quality white wines have larger alcohol than low-quality white wines, while low-quality white wines have larger residual sugars, density, chlorides, free sulfur dioxide and total sulfur dioxide than samples placed than high-quality white wines.

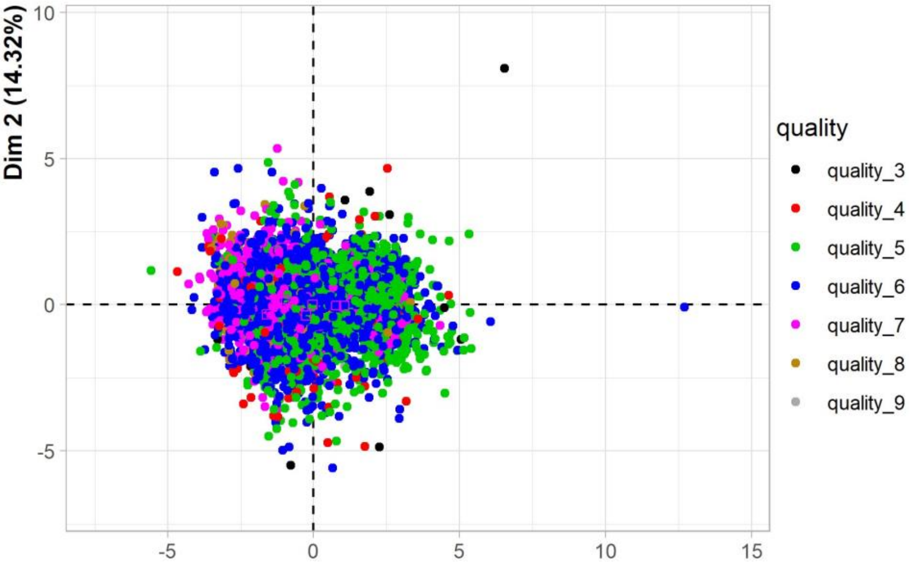
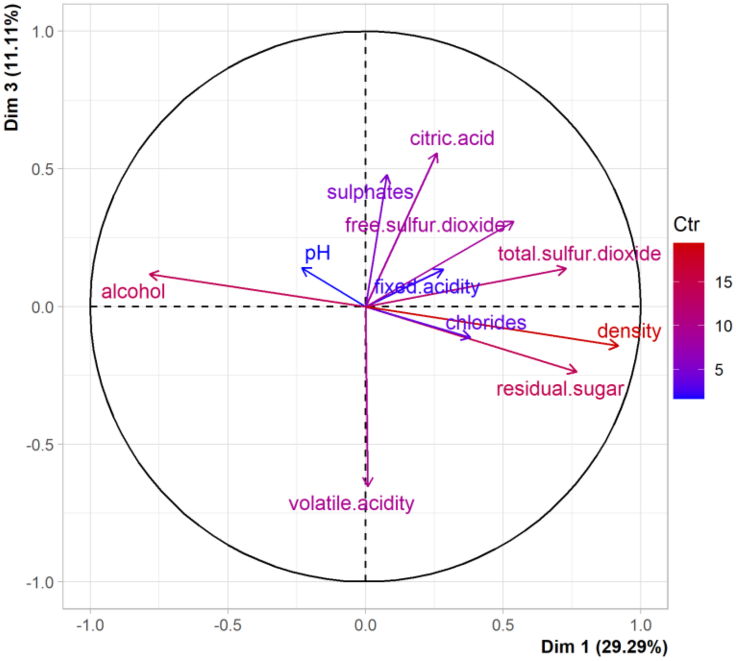


Figure 12: Score plot built using data set in Table S2, PC1 vs PC2

PC1 and PC2 explain just 43% of the total variance, and additional information can be obtained by using one more PC. The PC3 in the loading plot (Figure *13*) showed that white wine samples placed on
230 the bottom had higher volatile acidity than white wines placed on top, while white wines placed on the top had larger sulphates and citric acid than samples placed on the bottom. The score plot (Figure *14*) showed that most of the quality 4 white wines were placed on the bottom. Thus, quality 4 white wines can be differentiated from other qualities due to their larger volatile acidity.



235 Figure 13: Loading plot built using data set in Table S2, PC1 vs PC3

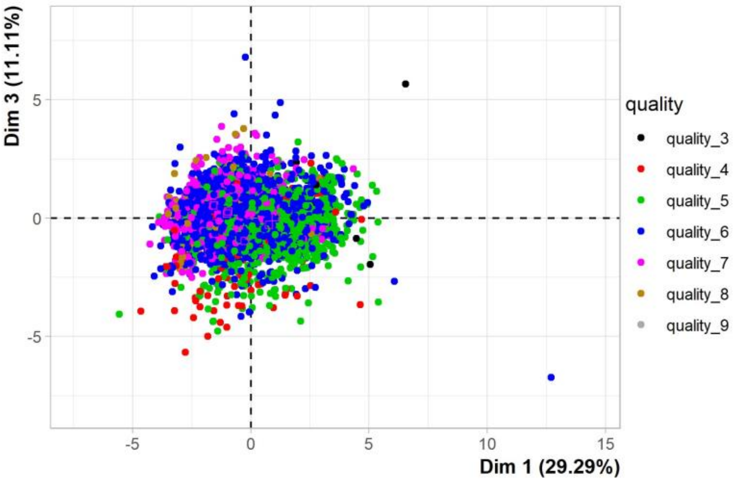


Figure 14: Score plot built using data set in Table S2, PC1 vs PC3

White wines data set statistical evaluation

Most of the white wine samples were quality 5 (29.7%), 6 (44.9%), and 7 (18%). There were few samples of quality 3 (0.4%), 4 (3.3%), 8 (3.6%) and 9 (0.1%). The distribution of white wines according to their quality was observed using a bar plot (Figure *15*).

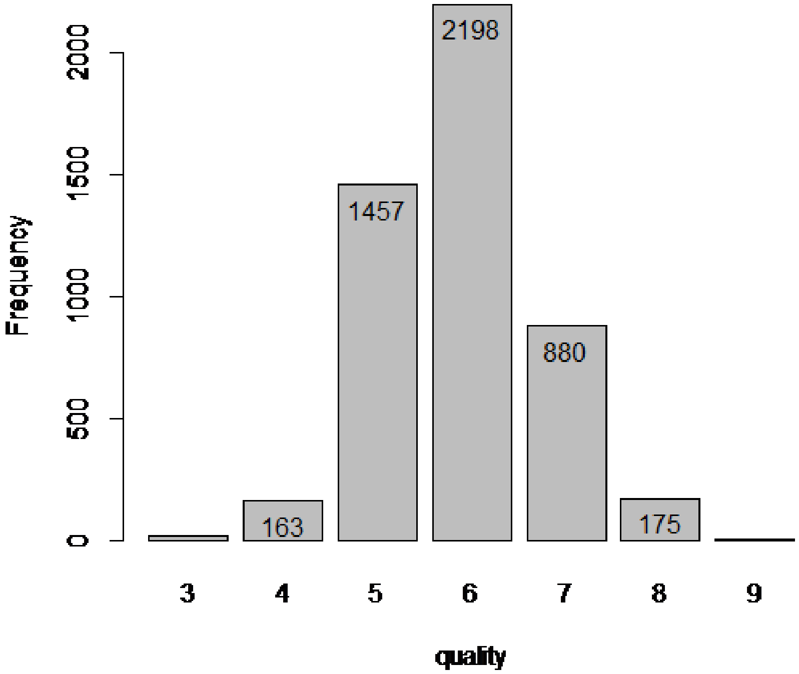


Figure 15: Bars graphic of white wines quality

In white wines (Table S2), all physicochemical variables had different means for each wine quality (**Table *2***). One-away ANOVA showed that it was a large difference between alcohol means in white wines.

Table 2: One-away ANOVA of 11 physicochemical variables for each white wine quality in Table S2

Physicochemical variable	F	p-value
alcohol	229.7	<2e-16
chlorides	42.47	<2e-16
citric acid	3.246	3.48 e-16
density	105.9	<2e-16
fixed acidity	12.89	1.64e-14
Free sulfur dioxide	19.72	<2e-16
pH	10.1	3.96e-11
residual sugar	21.27	<2e-16
sulphates	3.642	1.31e-3
total sulfur dioxide	45.2	<2e-16
volatile acidity	61.92	<2e-16

The plot of means (Figure *16*) showed that qualities 3 to 5 had equivalent alcohol. In qualities 5 to 9, alcohol increase as quality increase. The boxplot (Figure 17), confirmed results obtained with the plot of means. In addition, it showed that, in quality 5, most white wines had alcohol above the mean and many outliers, representing a log distribution.

The alcohol histogram confirmed the log distribution of alcohol in quality 5. In quality 8, we cannot observe its distribution pattern because it has few samples.

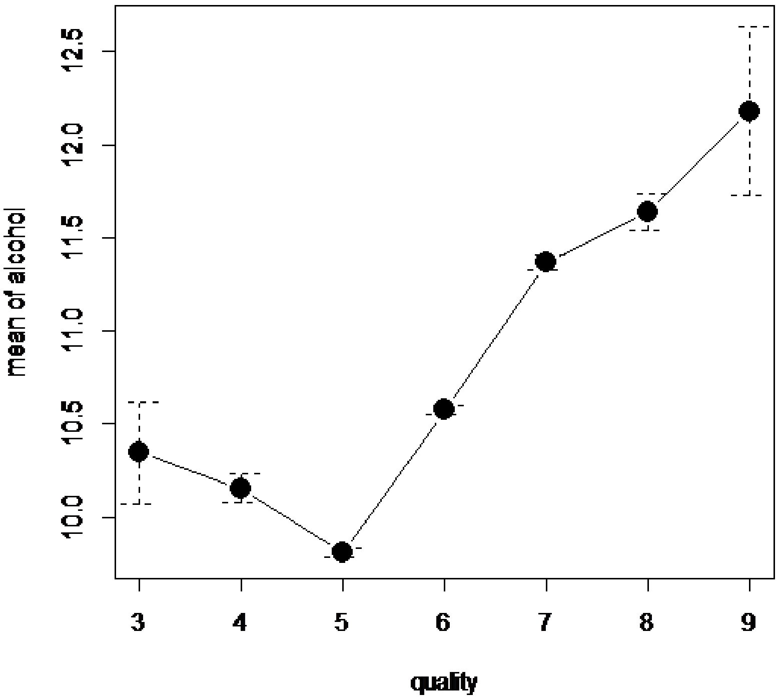


Figure 16: The alcohol plot of means of white wines qualities.

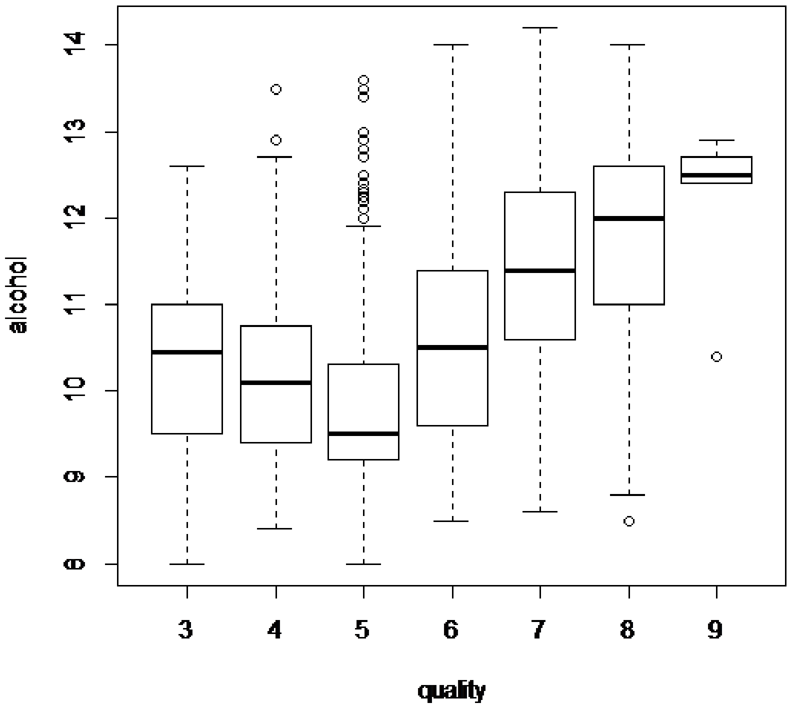


Figure 17: The alcohol boxplot of white wines qualities (Table S2)

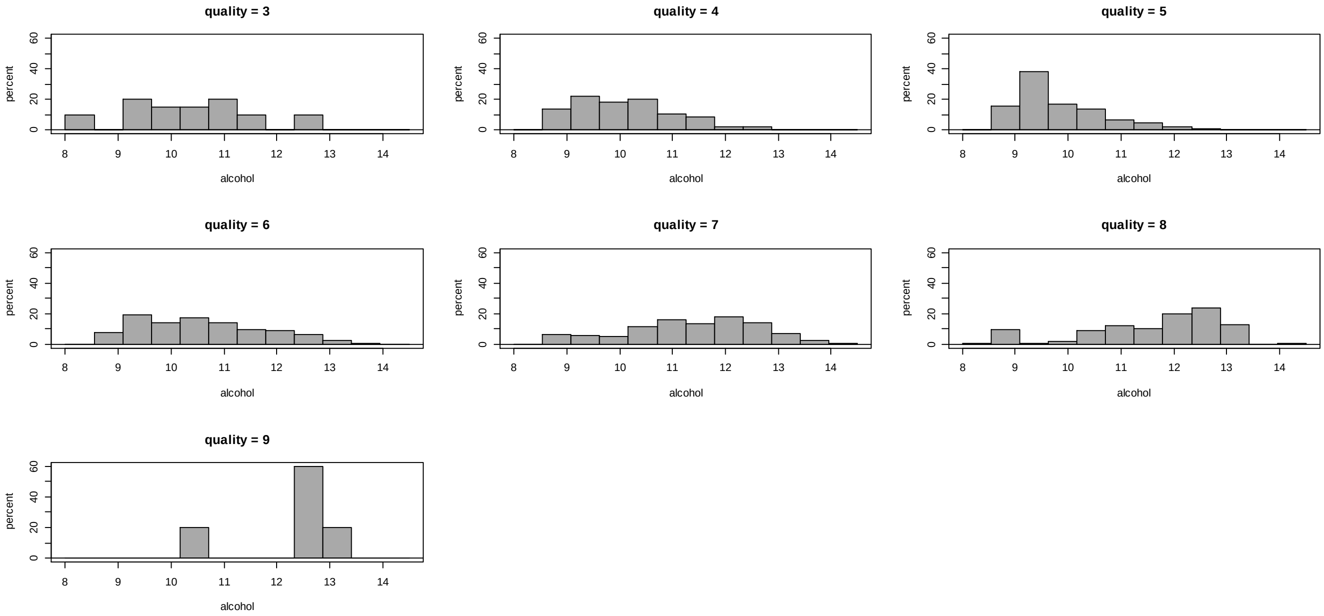


Figure 18: Alcohol histogram of white wines qualities.

Figure *13* showed that quality 4 was differentiated from other qualities because it has larger volatile acidity than other qualities. The volatile acidity boxplot (*Figure 19*) showed that quality 4 white wines have larger mean volatile acidity than other qualities. However, volatile acidity alone cannot be used to classify a white wine as quality 4, because quality 4 mean was in the upper quartile range of other qualities, and quality 5 and 6 have many outliers.

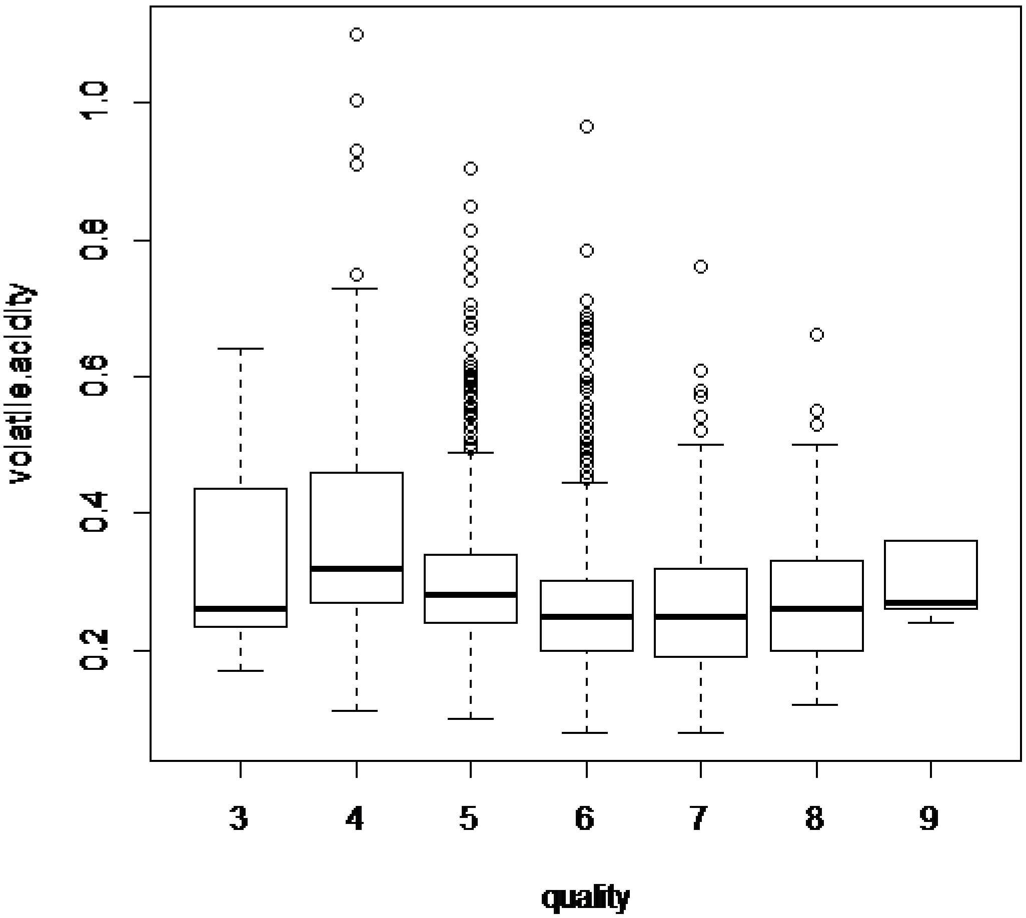


Figure 19: The volatile acid box plot of white wines qualities.

Red and white wines PCA

Red and white wines have different tastes[29] but can we represent their difference in taste in terms of the 11 physicochemical variables? PCA can respond to this question. The loading plot (Figure *20*) showed that samples placed on the left-hand side have larger total sulfur dioxide, free sulfur dioxide, and residual sugars than samples placed on the right-hand side, while samples placed on the right-hand side have larger volatile acidity, sulfates, chlorides, and fixed acidity than samples placed on the left-hand side.

The loading plot Figure *20* also showed that total sulfur and sulfur dioxide were correlated, density and alcohol were inversely correlated, fixed acidity and residual sugar were directly correlated.

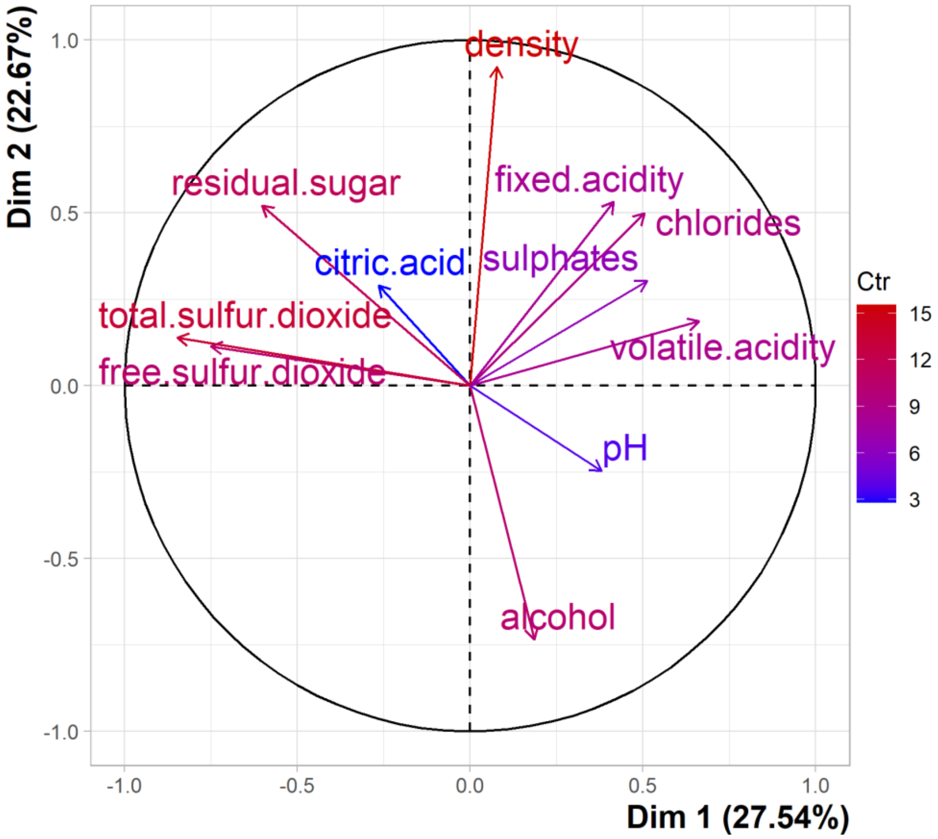


Figure 20: Loading plot built using data set in Table S3. PC1 vs PC2

The score plot (Figure 21) showed that white wines were placed on the left-hand side and red wines were placed on the right-hand side. Thus, the taste difference between white and red wines was explained using 11 physicochemical properties. Red wines have larger chlorides, fixed acidity, sulfates, and volatile acidity than white wines, and white wines have larger residual sugar, citric acid, total sulfur dioxide, and free sulfur than white wines.

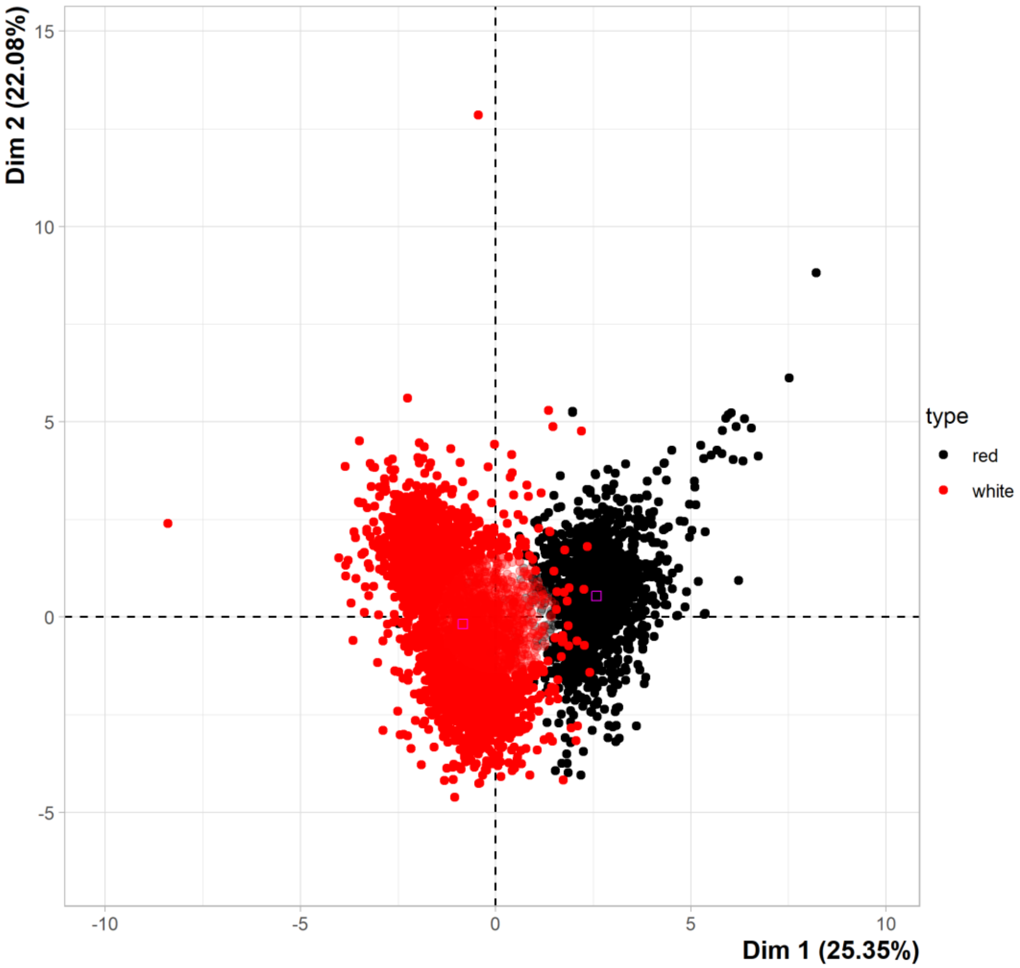


Figure 21: Score plot built using data set in Table S3, PC1 vs PC2,

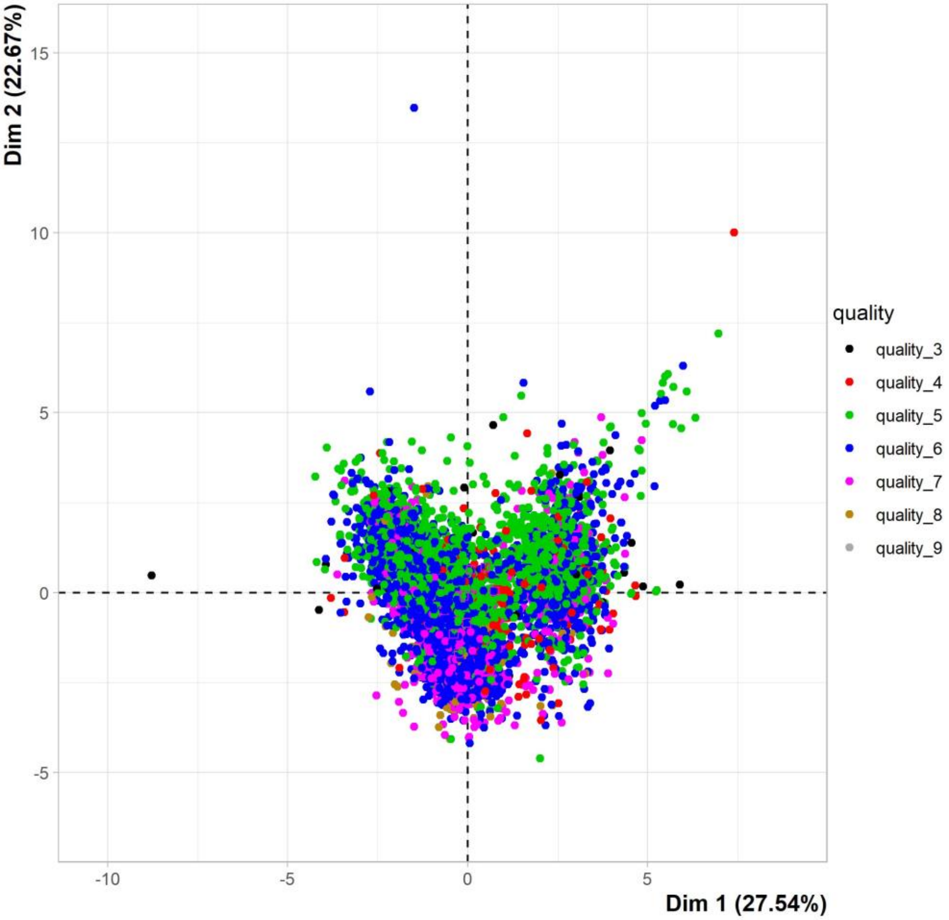
285 

Figure 22: Score plot built using data set in Table S3, PC1 vs PC2

The score plot (Figure 22) also showed that most high-quality wines (Quality 6 and 7) were placed on the bottom, where larger alcohol and smaller density were correlated with high-quality.

Red and white wines statistical evaluation

290 In the previous section, when we applied PCA to red and white wines data set (Table S3), it was observed that red wines have larger total sulfur dioxide and smaller volatile acidity than white wines, while white wines. These previous findings were confirmed using boxplots (Figure 23 and Figure 24).

White wines have larger sulfur dioxide than red wines and just some outlier red wines have sulfur dioxide in the mean range of white wines (Figure *23*)

295 Red wines have larger chlorides than white wines, but white wines have several outlier samples with chlorides above the red wine’s upper quartile (Figure *24*).

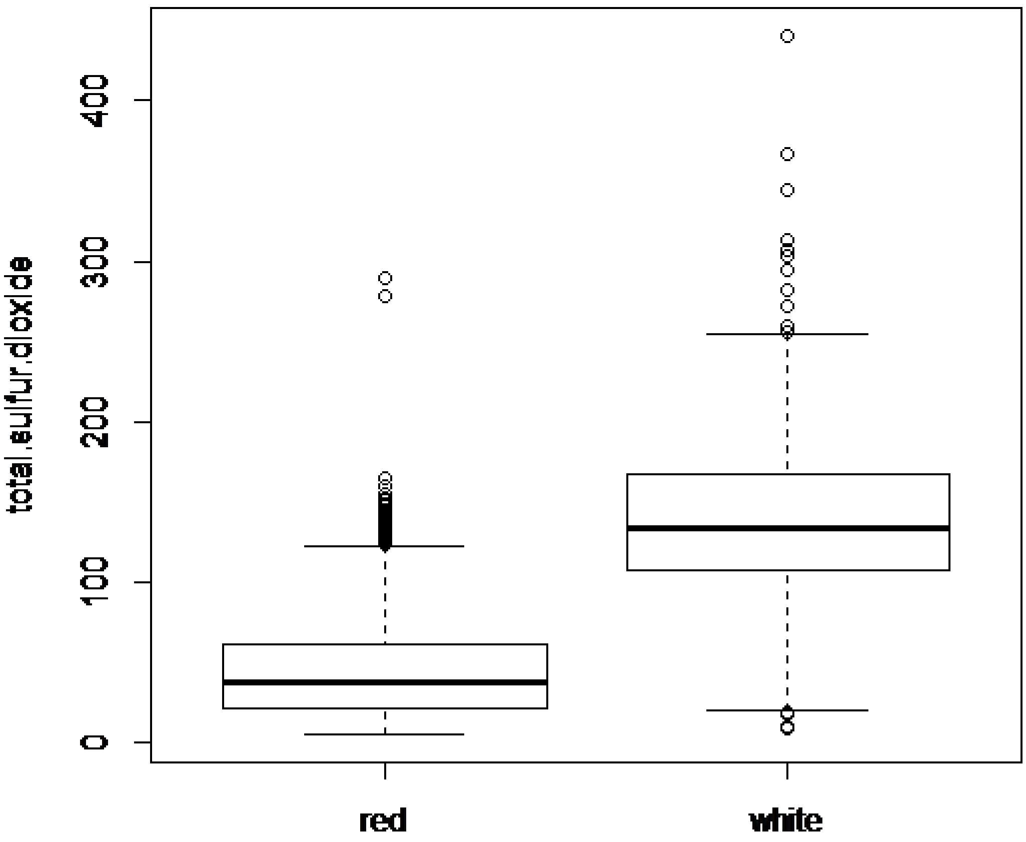


Figure 23:Total sulfur dioxide boxplot red and white wines

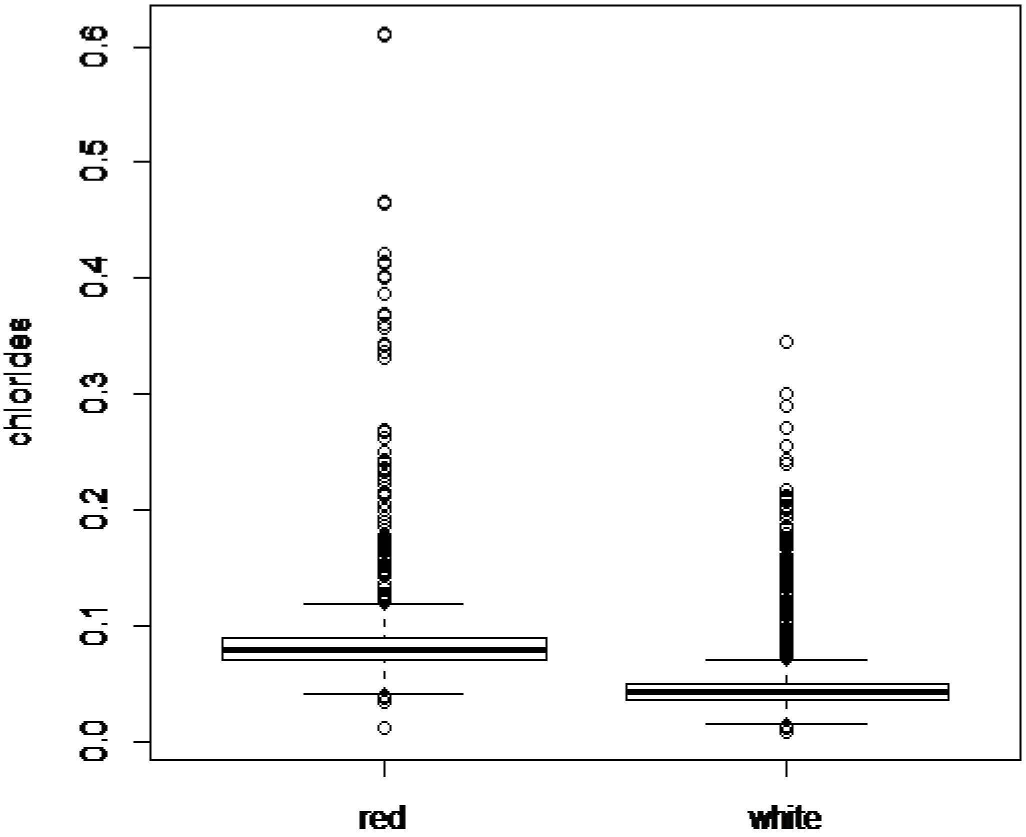


Figure 24: Chlorides Boxplot of red and white wines

Conclusion

This laboratory experiment can be realized in any school since it uses freeware (R project). Students without experience in programming can do it easily and without typing a single command line. Statistical tests such as boxplot, plot of means, and PCA cannot be done using spreadsheets. However, these tests were easily done using R Commander and Factoshiny plugins. Due to their graphical, easy, intuitive, and simple interface, students did all statical tests presented throughout the manuscript without any inconvenience.

ASSOCIATED CONTENT

Supporting Information

Tables used in the examples given in the paper (ZIP)

https://drive.google.com/drive/folders/1q95vAnXO0oAgpzARvA_H62A1D2NYDAuj?usp=sharing

AUTHOR INFORMATION

Corresponding Author

*E-mail: marcelborgesb@gmail.com and embsouza@furb.br

ACKNOWLEDGMENTS

The authors acknowledge financial support and fellowships from the Brazilian agencies FAPESC (Fundação de Amparo a Pesquisa do Estado de Santa Catarina). CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) Grant Number 399402226/2016-0. and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

REFERENCES

(1) Wilson, J. Statistical Computing with R: Selecting the Right Tool for the Job—R Commander or Something Else? *Wiley Interdisciplinary Reviews: Computational Statistics* **2012**, *4* (6), 518–526. https://doi.org/10.1002/WICS.1228.

(2) Sidou, L. F.; Borges, E. M. Teaching Principal Component Analysis Using a Free and Open Source Software Program and Exercises Applying PCA to Real-World Examples. *Journal of Chemical Education* **2020**, *97* (6), 1666–1676. https://doi.org/10.1021/acs.jchemed.9b00924.

(3) FactoMineR: Exploratory Multivariate Data Analysis with R http://factominer.free.fr/ (accessed 2021 -09 -20).

(4) Factoshiny http://factominer.free.fr/graphs/factoshiny.html (accessed 2021 -09 -20).

(5) Antonelli, T. M.; Olivieri, A. C. Developing and Implementing an R Shiny Application to Introduce Multivariate Calibration to Advanced Undergraduate Students. *Journal of Chemical Education* **2020**, *97* (4), 1176–1180. https://doi.org/10.1021/ACS.JCHEMED.9B00850.

(6) Maher, C.; Schazmann, B.; Gornushkin, I. B.; Rurack, K.; Gojani, A. B. Exploring an Application of Principal Component Analysis to Laser-Induced Breakdown Spectroscopy of Stainless-Steel Standard Samples as a Research Project. *Journal of Chemical Education* **2021**. https://doi.org/10.1021/ACS.JCHEMED.1C00563.

(7) Maher, C.; Schazmann, B.; B. Gornushkin, I.; Rurack, K.; B. Gojani, A. Exploring an Application of Principal Component Analysis to Laser-Induced Breakdown Spectroscopy of Stainless-Steel Standard Samples as a Research Project. *Journal of Chemical Education* **2021**, *0* (0). https://doi.org/10.1021/acs.jchemed.1c00563.

(8) Rusak, D. A.; Brown, L. M.; Martin, S. D. Classification of Vegetable Oils by Principal Component Analysis of FTIR Spectra. *Journal of Chemical Education* **2003**, *80* (5), 541–543. https://doi.org/10.1021/ED080P541.

(9) Pérez-Arribas, L. V.; León-González, M. E.; Rosales-Conrado, N. Learning Principal Component Analysis by Using Data from Air Quality Networks. *Journal of Chemical Education* **2017**, *94* (4), 458–464. https://doi.org/10.1021/ACS.JCHEMED.6B00550.

(10) Anderson, S. L.; Rovnyak, D.; Strein, T. G. Identification of Edible Oils by Principal Component Analysis of ^{1} H NMR Spectra. *Journal of Chemical Education* **2017**, *94* (9), 1377–1382. https://doi.org/10.1021/acs.jchemed.7b00012.

(11) Yeh, T.-S. Comment on “Identification of Edible Oils by Principal Component Analysis of ^{1} H NMR Spectra.” *Journal of Chemical Education* **2019**, *96* (8), 1790–1792. https://doi.org/10.1021/acs.jchemed.9b00133.

(12) Rovnyak, D.; Strein, T. G. Reply to “Comment on ‘Identification of Edible Oils by Principal Component Analysis of ^{1} H NMR Spectra.’” *Journal of Chemical Education* **2019**, *96* (8), 1793–1795. https://doi.org/10.1021/acs.jchemed.9b00557.

(13) Cazar, R. A. An Exercise on Chemometrics for a Quantitative Analysis Course. *Journal of Chemical Education* **2003**, *80* (9), 1026. https://doi.org/10.1021/ed080p1026.

(14) Sandusky, P. O. Introducing Undergraduate Students to Metabolomics Using a NMR-Based Analysis of Coffee Beans. *Journal of Chemical Education* **2017**, *94* (9), 1324–1328. https://doi.org/10.1021/ACS.JCHEMED.6B00559.

(15) de Lorenzi Pezzolo, A. To See the World in a Grain of Sand: Recognizing the Origin of Sand Specimens by Diffuse Reflectance Infrared Fourier Transform Spectroscopy and Multivariate Exploratory Data Analysis. *Journal of Chemical Education* **2011**, *88* (9), 1304–1308. https://doi.org/10.1021/ed9000409.

(16) Besalú, E. From Periodic Properties to a Periodic Table Arrangement. *Journal of Chemical Education* **2013**, *90* (8), 1009–1013. https://doi.org/10.1021/ed3004534.

(17) Horovitz, O.; Sârbu, C. Characterization and Classification of Lanthanides by Multivariate-Analysis Methods. *Journal of Chemical Education* **2005**, *82* (3), 473. https://doi.org/10.1021/ed082p473.

(18) Stitzel, S. E.; Sours, R. E. High-Performance Liquid Chromatography Analysis of Single-Origin Chocolates for Methylxanthine Composition and Provenance Determination. *Journal of Chemical Education* **2013**, *90* (9), 1227–1230. https://doi.org/10.1021/ED3003918.

(19) Boyce, M. C.; Lawler, N. G.; Tu, Y.; Reinke, S. N. Introducing Undergraduate Students to Metabolomics Using Liquid Chromatography–High Resolution Mass Spectrometry Analysis of Horse Blood. *Journal of Chemical Education* **2019**, *96* (4), 745–750. https://doi.org/10.1021/ACS.JCHEMED.8B00625.

(20) Ferreira, J. E. v.; Miranda, R. M.; Figueiredo, A. F.; Barbosa, J. P.; Brasil, E. M. Box-and-Whisker Plots Applied to Food Chemistry. *Journal of Chemical Education* **2016**, *93* (12), 2026–2032. https://doi.org/10.1021/ACS.JCHEMED.6B00300.

(21) Canaes, L. S.; Brancalion, M. L.; Rossi, A. v.; Rath, S. Using Candy Samples To Learn about Sampling Techniques and Statistical Data Evaluation. *Journal of Chemical Education* **2008**, *85* (8), 1083–1088. https://doi.org/10.1021/ED085P1083.

(22) Bro, R.; Smilde, A. K. Principal Component Analysis. *Anal. Methods* **2014**, *6* (9), 2812–2831. https://doi.org/10.1039/C3AY41907J.

(23) Granato, D.; Santos, J. S.; Escher, G. B.; Ferreira, B. L.; Maggio, R. M. Use of Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) for Multivariate Association between Bioactive Compounds and Functional Properties in Foods: A Critical Perspective. *Trends in Food Science & Technology* **2018**, *72*, 83–90. https://doi.org/10.1016/J.TIFS.2017.12.006.

(24) Nunes, C. A.; Alvarenga, V. O.; de Souza Sant’Ana, A.; Santos, J. S.; Granato, D. The Use of Statistical Software in Food Science and Technology: Advantages, Limitations and Misuses.

Food Research International **2015**, *75*, 270–280. https://doi.org/10.1016/J.FOODRES.2015.06.011.

(25) Brereton, R. G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J. M.; Walczak, B.; Tauler, R. Chemometrics in Analytical Chemistry—Part II: Modeling, Validation, and Applications. *Analytical and Bioanalytical Chemistry* **2018**, *410* (26), 6691–6704. https://doi.org/10.1007/s00216-018-1283-4.

(26) Brereton, R. G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J. M.; Walczak, B.; Tauler, R. Chemometrics in Analytical Chemistry—Part I: History, Experimental Design and Data Analysis Tools. *Analytical and Bioanalytical Chemistry* **2017**, *409* (25), 5891–5899. https://doi.org/10.1007/s00216-017-0517-1.

(27) Larsen, R. D. Box-and-Whisker Plots. *Journal of Chemical Education* **1985**, *62* (4), 302–305. https://doi.org/10.1021/ED062P302.

(28) Wine Quality Datasets http://www3.dsi.uminho.pt/pcortez/wine/ (accessed 2021 -09 -20).

(29) Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; Reis, J. Modeling Wine Preferences by Data Mining from Physicochemical Properties. *Decision Support Systems* **2009**, *47* (4), 547–553. https://doi.org/10.1016/J.DSS.2009.05.016.

(30) Lafuente, D.; Cohen, B.; Fiorini, G.; Alejo García, A.; Bringas, M.; Morzan, E.; Onna, D. A Gentle Introduction to Machine Learning for Chemists: An Undergraduate Workshop Using Python Notebooks for Visualization, Data Processing, Analysis, and Modeling. *Journal of Chemical Education* **2021**, *98* (9), 2892–2898. https://doi.org/10.1021/acs.jchemed.1c00142.

(31) Cozzolino, D.; Power, A.; Chapman, J. Interpreting and Reporting Principal Component Analysis in Food Science Analysis and Beyond. *Food Analytical Methods* **2019**, 1–5. https://doi.org/10.1007/s12161-019-01605-5.

(32) Borges, E. M. How to Select Equivalent and Complimentary Reversed Phase Liquid Chromatography Columns from Column Characterization Databases. *Analytica Chimica Acta* **2014**, *807*, 143–152. https://doi.org/10.1016/J.ACA.2013.11.010.