How is Wine Quality Affected by Physiochemical Compounds?

Bho Bhat Chhetri

2022-12-01

Contents

1 Introduction

Wine has been historically consumed by humans as one of the important alcoholic beverages highly popular throughout the world. About 7.9 billion of the world's population drink wine as an alcoholic beverage. In the USA, about 118 million people drink wine, among them females 45% and males 15%. While many tout its health benefits, researchers and medical experts continue to debate whether the drink can offer a wholesome advantage to human health.

1.1 Background

This study considers data from **Vinho Verde**, a unique wine production from the Northwest region of Portugal. A total of **6497** wines were collected from May 2004 to February 2007 using only protected designation of origin samples that were tested at the official certification entity **Comissao de Viticultura da Regiao dos Vinhos Verdes**(CVRVV) and recorded by a computerized system (iLab). The most common physiological test and the sensory test are conducted to separate data sets of white and red wine. A sensory test was done via blind test, which graded wine on a scale from zero(very bad) to 10(excellent)link. The result of this test is useful for improving the production process and improve the quality of the wine.

1.2 Description

In this dataset, the goal is to make prediction of wine quality based on some chemical facts. Despite the initial data, the wine is scored from (0-10) some of which may guess that data is much more suited to be predicted using regression, we need to remember that the score is discrete, which makes classification model more suitable for this analysis.

The purpose of this analysis is to construct the prediction of wine quality based on some chemical facts and interpretation of the model by deriving five most important factors for wine quality. Will it be percentage of alcohol, acidity or other variables?this analysis will try to answer these questions after analyzing this dataset.

The essential **owner** of this dataset were Paulo Cortez, University of Minho, Guimares, Portugal, link A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal and the data was **collected by** A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the **Vinho Verde** Region(**CVRVV**), Porto, Portugal. The data representative was included all samples stored in CVRVV.

The strength of the dataset represents only one product of wine produced in **Minho** region of Portugal,first dataset of this kind, highly popular, with 2 million UCI download. The data set is large, but the proportion of red and white wine data is significantly different. The white wine representation is three times higher than the red wine data. The number of data set is a huge strength of this study as the probability of inclusion is higher. The separate data set for red and white wine gives clear demarcation on data characters. The inclusion of sensory tests along with physio-chemical tests gives more clarity on the result. One of the best things about this data set is they did not have null values in the observations section of all the variables, which helps in easy analysis of the data by preventing the statistical power, biased estimates and invalid inclusions. The dataset is publicly available, so everyone can use it to analyze in different programming languages.

The weakness was lack of other relevant features, such as selling price, production cost, etc. The relationship between human experts and the physio-chemical lab test data set is not clear yet. This data set does not represent any other brand of wine available in the market, other than this unique product from Portugal, and the data set contains only numeric variables without a categorical variable, which creates complications in ggplot.Similarly the variable names are in the column names without row names that creates the trouble in proper analysis and plot.Similarly the main missingness of this dataset is no row names and categorical variables. When it comes to understanding whether wine can be useful to human health and to what extent it can be healthy for the human body, it is very important to understand the various chemical components that make up the wine. It is also important to understand the distinction between **white and red wine**, the two general wine types that are produced and marketed today, not only because they are produced and inventoried in different ways. Also, their chemical composition might not be the same. This can evidently offer whether one is better than the other.

Image:1.1 Wine testing process by experts

This image shows data collected over the last several months by VineSleuth. The VineSleuth's data shows that expert wine elevators "are able to repeat their observations on individual wine samples about 90% of the time" when tasting wines blind.source.This image describes the testing process of wine.

Image: 1.2 Wine samples in the dataset

This image shows the sample of the red and white wine based on the color, which represents the types of wine in the data set.The red wine looks dark red grapes juice and white wine looks like green grape juice due to their sources are red and white grapes. source

2 Figures

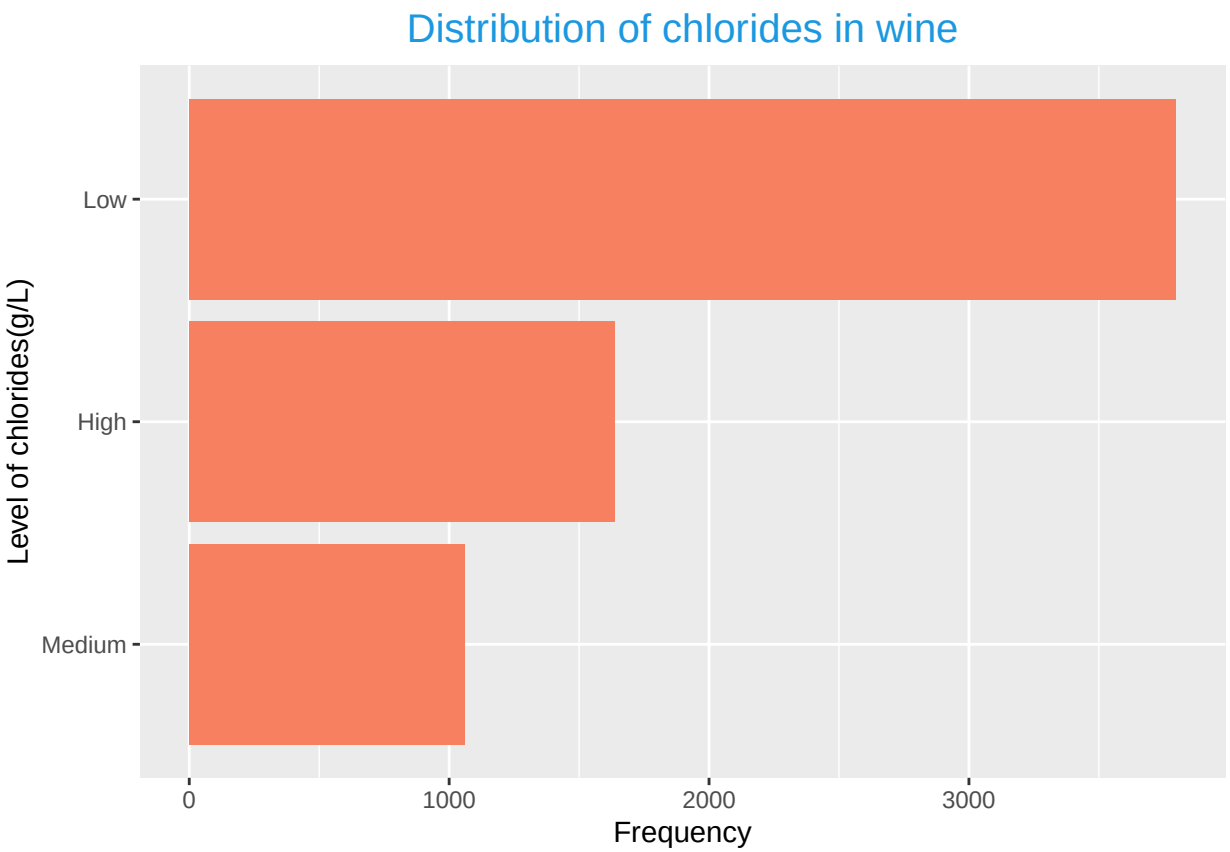

Figure:2.1 Bar graph for chlorides in wine dataset

This bar plot represents the distribution of chlorides based on its amount present in the wine.As we can observe less amount of chlorides is good for the wine. Most of the wine in the data has less amount of chloride categorized as low(<0.050g/L), maximum amount as high(>0.065g/L) and moderate amount categorized as medium(>0.051g/L).However, low amount of chlorides was reported in maximum wine samples.

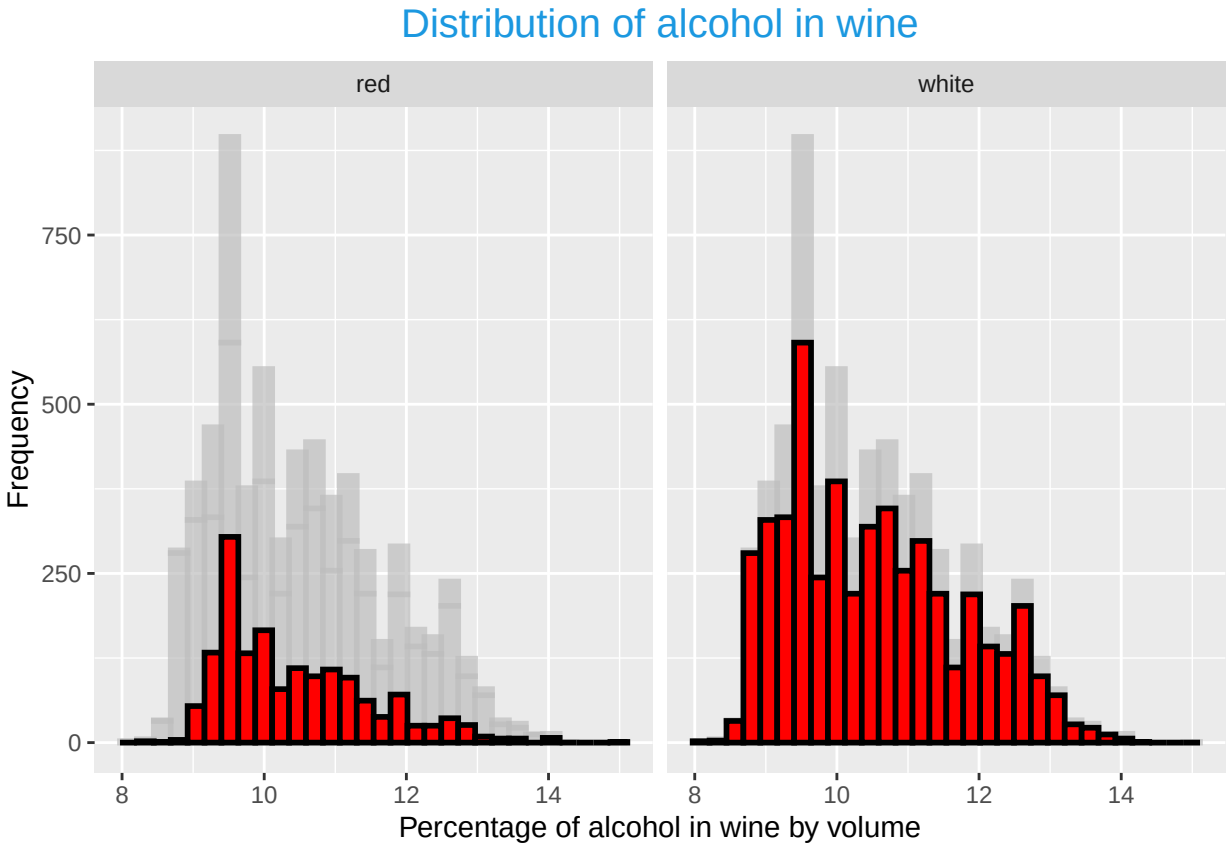

Figure:2.2 The histogram of alcohol in wine style

This histogram represents the distribution of red and white wine based on percentage of alcohol by volume. As seen from the plot most of the wine has alcohol amount less then 10%.The gghighlight fuction hilighted the maximum distribution based on amount of alcohol more than 10%.There is very less number of red wine identified at alcohol level (14%)compared to white wine.

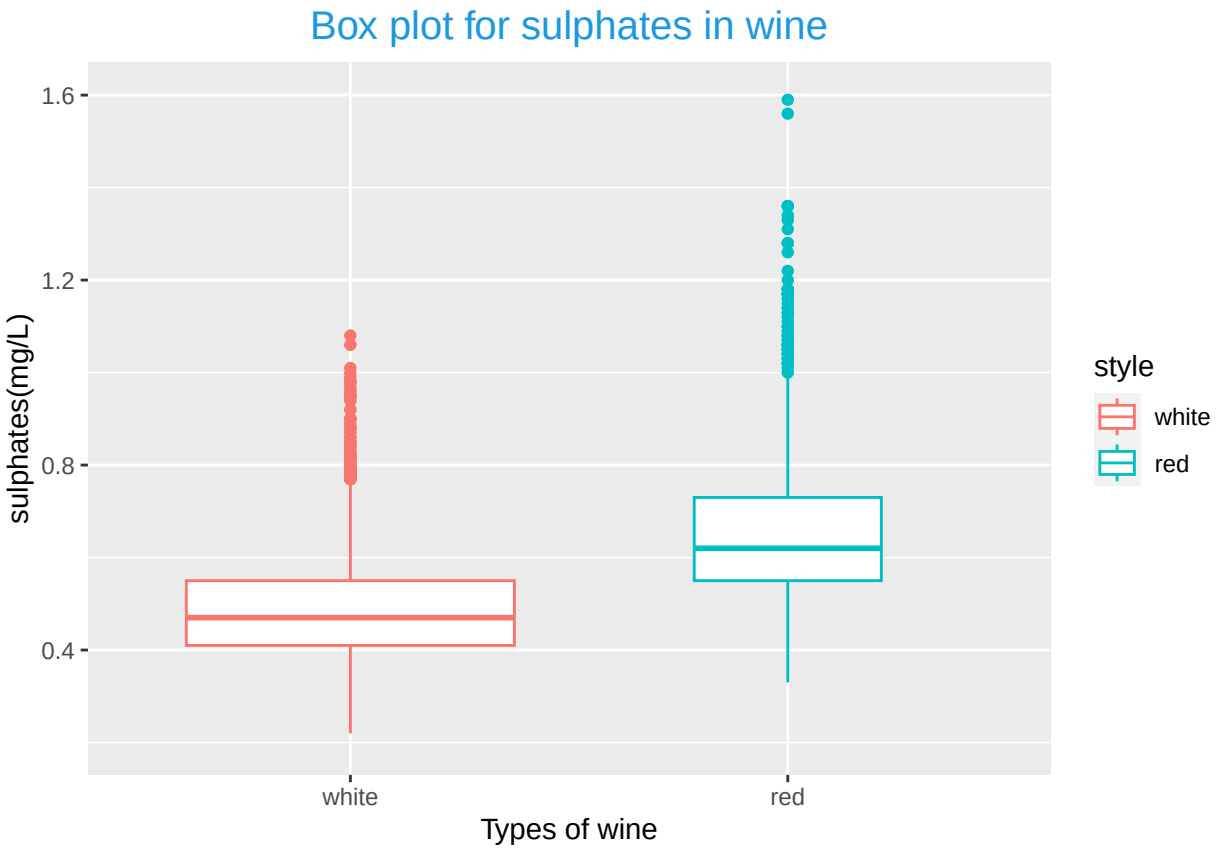

Figure:2.3 Box plot of sulphates in wine style

The box plot of sulphates across the wine type indicates that the amount of sulphates present in the wine dataset is not uniform. The median value of sulphates for the white wine is slightly above the first quartile where as the red wine is above the lower median quartile,this indicates there is significant difference of sulphates in wine type.In white wine median value lies slightly about the lower quartile(25%), where as the red wine lies lower half quartile(50%) which help to easily identify the sulphates amount in wine dataset.

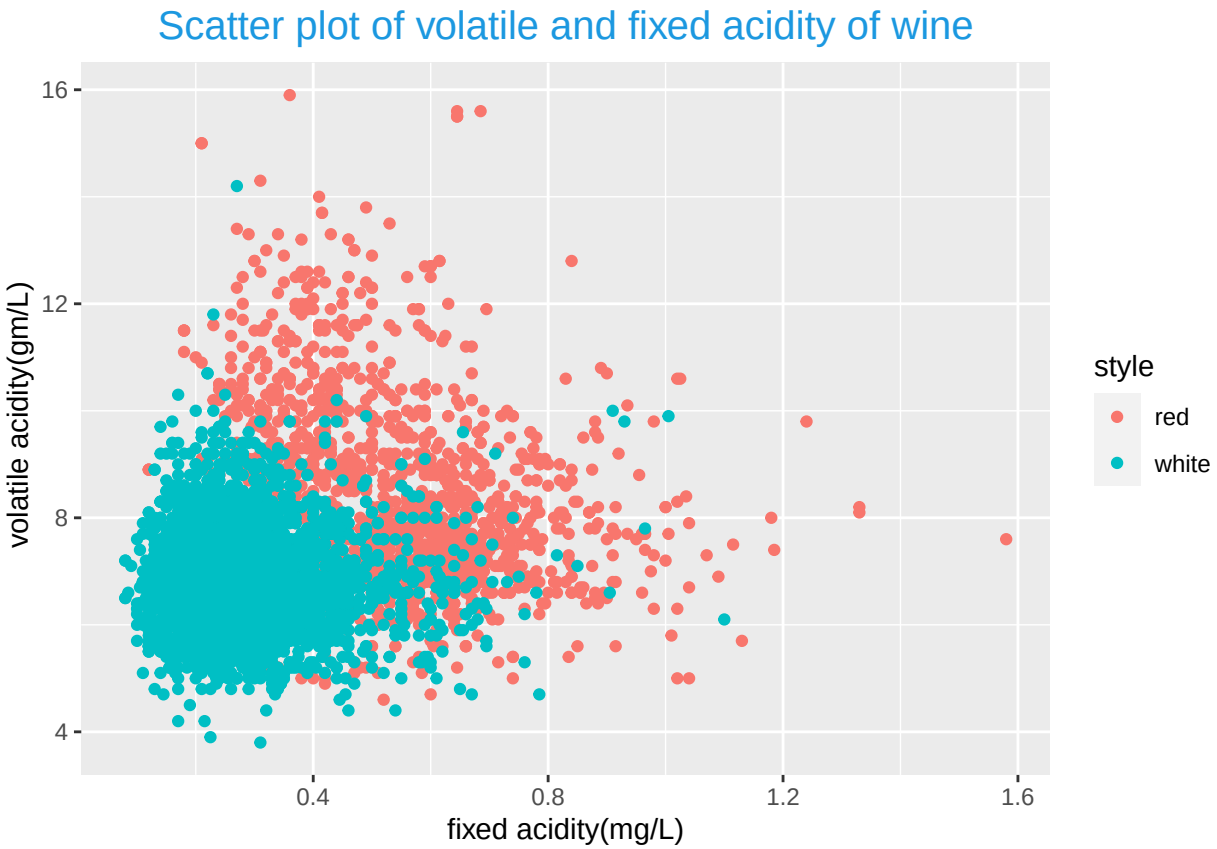

Figure: 2.4 Scatter plot for fixed and volatile acidity

The scatter plot shows the relationship between the volatile and fixed acidity in both wines. Acidities are highly scattered in red wine where as white wine is more concentrated at level of 8gm/lit on volatile and 0.4 mg/lit on fixed site.Graph shows that red wine acidity (volatile and fixed) are significantly higher compared to white wine.

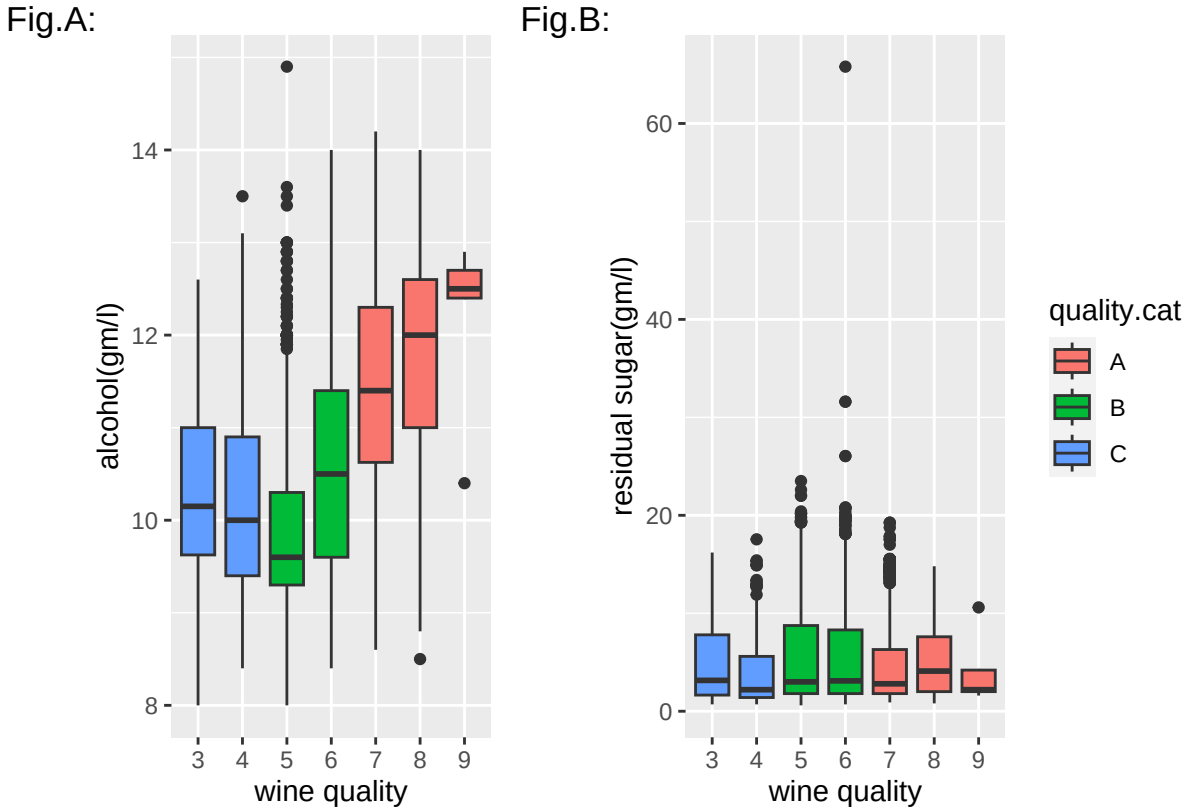

Figure:2.5 Box plot for quality Vs alcohol and residual sugar of wine(A=rating>6.5,B>4.9 & C<4.8)

The box graph 'A' shows the minimum, maximum,first quartile, second quartile(median) and third quartile of the alcohol present in the wine dataset. As observed in quality rating 'C' the wine quality is evenly distributed and no high variation is observed. While in category 'B', amount of alcohol has wide variations. The amount of alcohol above the upper quartile is very high. Besides in 'A', the distribution is similar as in 'C'.Clearly, the wine with quality 'B' has large amount of alcohol.The quality is determined based on rating of the wine. The quality 'A' has rating>6.5, B>4.9 and C<4.8.

Similarly, the box graph figure 'B' shows the minimum, maximum,first quartile, second quartile(median)and third quartile of the residual sugar present in the wine dataset.As we observed in quality rating 'B' the wine quality is evenly distributed without any variations and upper quartile is high.While in category 'A' amount of residual sugar has wide variation,the amount of sugar observed below the median quartile. Besides in'B' the distribution is similar in 'C'.The wine quality'B' has large amount of residual sugar.The quality rating is same in graph figure 'A'.

3 Demographic table

	red	white	Overall
	(N=1599)	(N=4898)	(N=6497)
density			
Mean (SD)	1.0 (0.0019)	0.99 (0.0030)	0.99 (0.0030)
Median [Min, Max]	1.0 [0.99, 1.0]	0.99 [0.99, 1.0]	1.0 [0.99, 1.0]
chlorides			
Mean (SD)	0.087 (0.047)	0.046 (0.022)	0.056 (0.035)
Median [Min, Max]	0.079 [0.012, 0.61]	0.043 [0.0090, 0.35]	0.047 [0.0090, 0.61]
pH			
Mean (SD)	3.3 (0.15)	3.2 (0.15)	3.2 (0.16)
Median [Min, Max]	3.3 [2.7, 4.0]	3.2 [2.7, 3.8]	3.2 [2.7, 4.0]
sulphates			
Mean (SD)	0.66 (0.17)	0.49 (0.11)	0.53 (0.15)
Median [Min, Max]	0.62 [0.33, 2.0]	0.47 [0.22, 1.1]	0.51 [0.22, 2.0]
alcohol			
Mean (SD)	10 (1.1)	11 (1.2)	10 (1.2)
Median [Min, Max]	10 [8.4, 15]	10 [8.0, 14]	10 [8.0, 15]
quality			
Mean (SD)	5.6 (0.81)	5.9 (0.89)	5.8 (0.87)
Median [Min, Max]	6.0 [3.0, 8.0]	6.0 [3.0, 9.0]	6.0 [3.0, 9.0]

Table:3.1 Demographic table for numeric variables of wine dataset

The demographic table shows that there is large differences in white and red wine in terms of some variables like chlorides,sulphates, and quality, where as less significance differences in alcohol,density, and pH. No significant differences seen among mean value density value of red and white wine where as mean value of chlorides in red wine is almost two times higher than white. Similarly, no differences seen between mean PH values of white and red wine. Sulfate is noted higher in red wine. Percentage of alcohol in red and white wine is not much difference in its mean value. Mean quality value of white wine is higher than red.

	red	white	Overall
	(N=1599)	(N=4898)	(N=6497)
chlorides.cat			
High	1357 (84.9%)	281 (5.7%)	1638 (25.2%)
Low	62 (3.9%)	3735 (76.3%)	3797 (58.4%)
Medium	180 (11.3%)	882 (18.0%)	1062 (16.3%)
alcohol.cat			
Best	162 (10.1%)	813 (16.6%)	975 (15.0%)
Better	1434 (89.7%)	4071 (83.1%)	5505 (84.7%)
good	3 (0.2%)	14 (0.3%)	17 (0.3%)
quality.cat			
A	217 (13.6%)	1060 (21.6%)	1277 (19.7%)
B	1319 (82.5%)	3655 (74.6%)	4974 (76.6%)
C	63 (3.9%)	183 (3.7%)	246 (3.8%)

Figure:3.2 Demographic table for categorical variables of wine dataset

The descriptive statistic on demographic table of categorical variables of the wine dataset shows the frequencies, percentage and fraction of the given variables. In chlorides categorical variable most of the red wine samples

lies on higher category with(84.9%) that means amount of chlorides in those wine more than (0.065 gm/L), where as in white wine it is about(5.7%)on this category.Similarly low chloride category is higher(76.3%)in white wine with chlorides amount (between 0.051-0.065 gm/L) where as in red wine with(3.9%).In alcohol category, white wine with best category (16.6%) having alcohol amount more than (12%) by volume compared to red wine(10.1%). In quality category white wine shows higher 'A' with(21.6%)based on the rating more than 6.5 for each sample, where as in red wine with(13.6%). There is insignificance differences in 'B' and 'C' category in both wines.

4 Conclusion

This data set was collected from kaggle.com, one of the publicly available site.This dataset only contains 11 physio chemical and one sensory variables and all of them have numeric values.Analysis shows that the distribution of chlorides is higher in low category group which represent the good quality of wine. Study shows that high quality of wine has higher percentage of alcohol and on the other hand it shows that sulphate percentage in wine does not affect the wine quality. Analysis concludes that white wines are of high quality than red wine.

5 References

1. 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.(https://www.sciencedirect.com/science/article/pii/S0167923609001377)
2. Stackoverflow.com(https://stackoverflow.com/questions/48799074/how-to-reorder-a-factor-in-a-dataframe-with-fct-reorder)
3. Sthsa.com(http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization)
4. Tidyverse.org (https://ggplot2.tidyverse.org/reference/geom_histogram.html)
5. Forcats tidyverse(https://forcats.tidyverse.org/articles/forcats.html)
6. Histogram and frequency polygon(https://ggplot2.tidyverse.org/reference/geom_histogram.html)
7. Create the HTML of descriptive statistics table(https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html)

6 Package Citations

Analyses were conducted using the R Statistical language (version 4.2.2; R Core Team, 2022) on Windows 10 x64 (build 19044), using the packages report (version 0.5.5; Makowski D et al., 2021), patchwork (version 1.1.2; Pedersen T, 2022), psych (version 2.2.9; Revelle W, 2022), table1 (version 1.4.2; Rich B, 2021), ggplot2 (version 3.4.0; Wickham H, 2016), forcats (version 0.5.2; Wickham H, 2022), dplyr (version 1.0.10; Wickham H et al., 2022), readr (version 2.1.3; Wickham H et al., 2022), knitr (version 1.40; Xie Y, 2022) and gghighlight (version 0.4.0; Yutani H, 2022).

6.1 References

- Makowski D, Ben-Shachar M, Patil I, Lüdecke D (2021). "AutomatedResults Reporting as a Practical Tool to Improve Reproducibility andMethodological Best Practices Adoption." *CRAN*.https://github.com/easystats/report.
- Pedersen T (2022). *patchwork: The Composer of Plots*. R packageversion 1.1.2, https://CRAN.R-project.org/package=patchwork.

- R Core Team (2022). *R: A Language and Environment for StatisticalComputing*. R Foundation for Statistical Computing, Vienna, Austria.https://www.R-project.org/.
- Revelle W (2022). *psych: Procedures for Psychological, Psychometric,and Personality Research*. Northwestern University, Evanston,Illinois. R package version 2.2.9,https://CRAN.R-project.org/package=psych.
- Rich B (2021). *table1: Tables of Descriptive Statistics in HTML*. Rpackage version 1.4.2, https://CRAN.R-project.org/package=table1.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*.Springer-Verlag New York. ISBN 978-3-319-24277-4,https://ggplot2.tidyverse.org.
- Wickham H (2022). *forcats: Tools for Working with CategoricalVariables (Factors)*. R package version 0.5.2,https://CRAN.R-project.org/package=forcats.
- Wickham H, François R, Henry L, Müller K (2022). *dplyr: A Grammar ofData Manipulation*. R package version 1.0.10,https://CRAN.R-project.org/package=dplyr.
- Wickham H, Hester J, Bryan J (2022). *readr: Read Rectangular TextData*. R package version 2.1.3,https://CRAN.R-project.org/package=readr.
- Xie Y (2022). *knitr: A General-Purpose Package for Dynamic ReportGeneration in R*. R package version 1.40, https://yihui.org/knitr/.Xie Y (2015). *Dynamic Documents with R and knitr*, 2nd edition.Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1498716963,https://yihui.org/knitr/.Xie Y (2014). "knitr: A Comprehensive Tool for Reproducible Research inR." In Stodden V, Leisch F, Peng RD (eds.), *Implementing ReproducibleComputational Research*. Chapman and Hall/CRC. ISBN 978-1466561595,http://www.crcpress.com/product/isbn/9781466561595.
- Yutani H (2022). *gghighlight: Highlight Lines and Points in'ggplot2'*. R package version 0.4.0,https://CRAN.R-project.org/package=gghighlight.