Capstone Project

winequalityx--- R

Project Title: **Predict what makes a good white wine**

Table Of Contents

Capstone Project – winequalityx-R

Predict what makes a good white wine

Introduction

Our analyses focus in a Portuguese white wine database consisting of 4, 898 observations, The data set contains eleven
explanatory variables that measure wine attributes and one response variable:" wine quality". In more detail,
Fixed acidity: a measurement of the total concentration of titratable acids and free hydrogen ions present in the wine. Theoretically, having a low acidity will result in a flat and boring wine while having too much acid can lead to tartness or even a sour wine. These acids either occur naturally in the grapes or are created through the fermentation process.

• Volatile acidity: a measure of steam distillable acids present in a wine. In theory, our palates are quite sensitive to the presence of volatile acids and for that reason a good wine should keep their concentrations as low as possible.

• Citric acid: one of the many acids that are measured to obtained fixed acidity.

• Residual sugar: measurement of any natural grape sugars that are left over after fermentation ceases. In theory residual sugar can help wines age well.

• Chlorides: the amount of salt in the wine. • Free sulfuric dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.

• Total sulfuric dioxide: amount of free and bound forms of SO2; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.

• Density: measure of density of wine.

• pH: value for pH.

• Sulfates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant.

• Alcohol: the percentage of alcohol present in the wine. • Quality: subjective measurement ranging from 1 to 10 (although the observed data ranges from 3 to These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.
Good wine generally has a higher acidity level. Alcohol level and sugar are also important features in the analysis.

Number of observations in the given dataset: 4898

Summary of the data/ Review of Literature

```
summary(white)
```

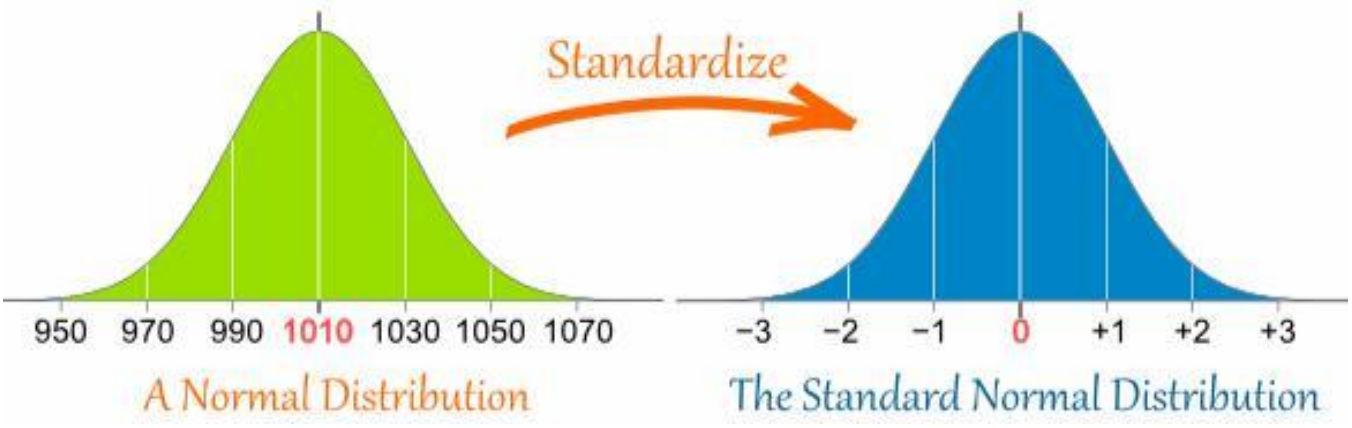
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

1) There is a big range for sulphur.dioxide (both Free and Total) across the samples.
2) The sample consists of 4898 White wine.
3) The alcohol content varies from 8.00 to 14.20 for the samples in dataset.
4) The quality of the samples range from 3 to 9 with 6 being the median.
5) The range for fixed acidity is quite high with minimum being 3.8 and maximum being 14.2
6) pH value varies from 2.720 to 3.820 with a median being 3.180

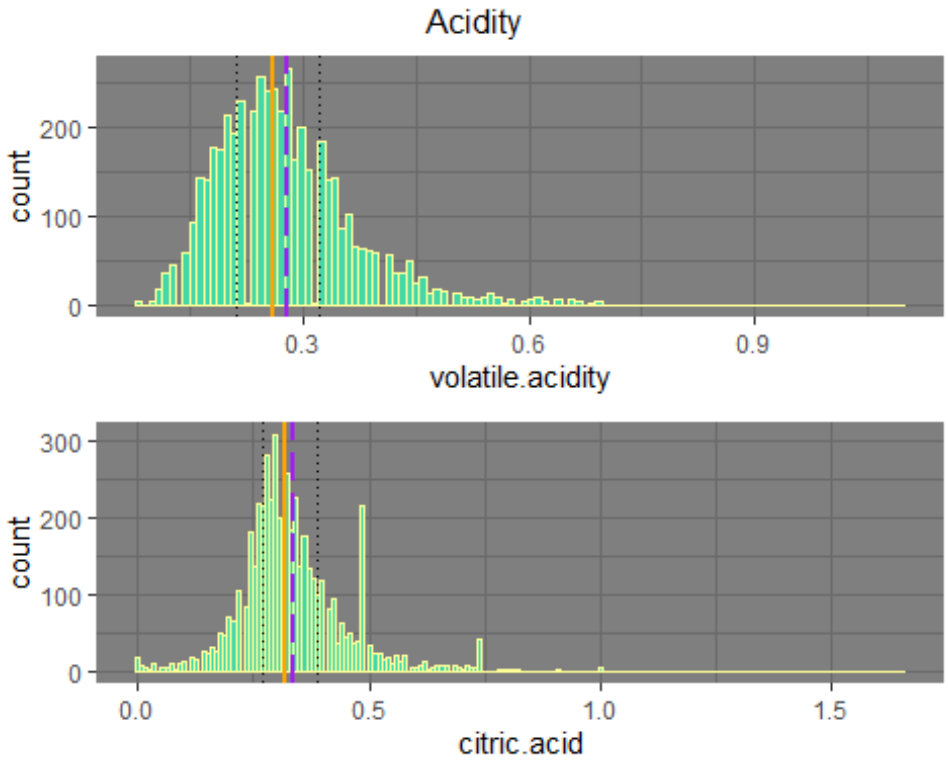
Transforming Data

Data Standardization

Data standardization is a process in which data attributes within a data model are organized to increase the cohesion of entity types. In other words, the goal of data standardization is to reduce and even eliminate data redundancy, an important consideration for application developers because it is incredibly difficult to store objects in a database that maintains the same information in several places.



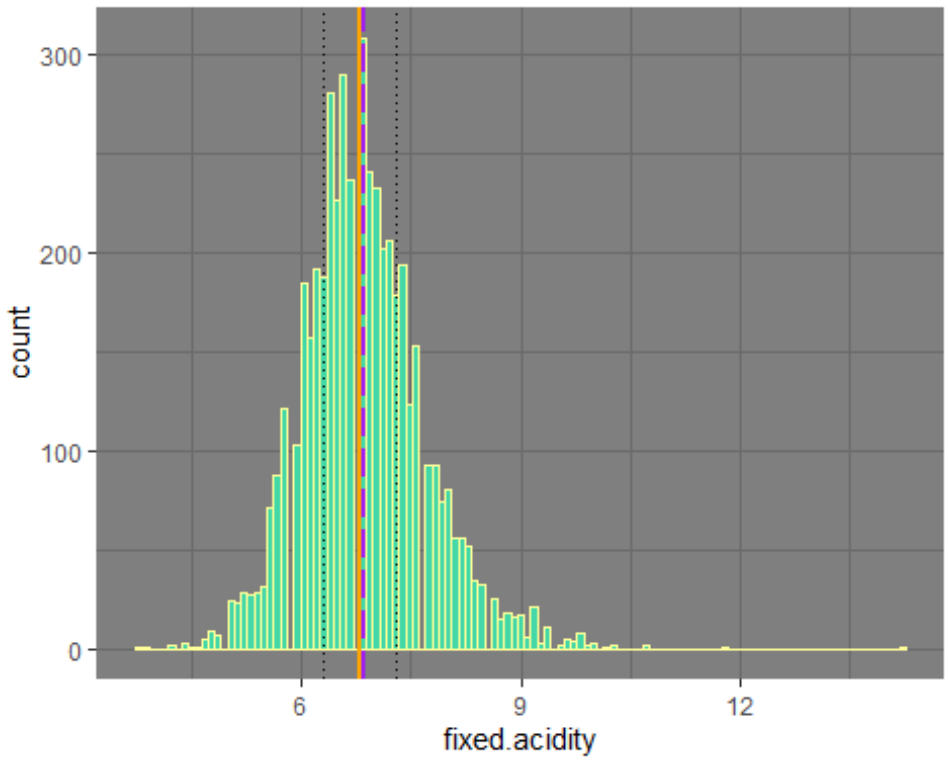
Data visualization



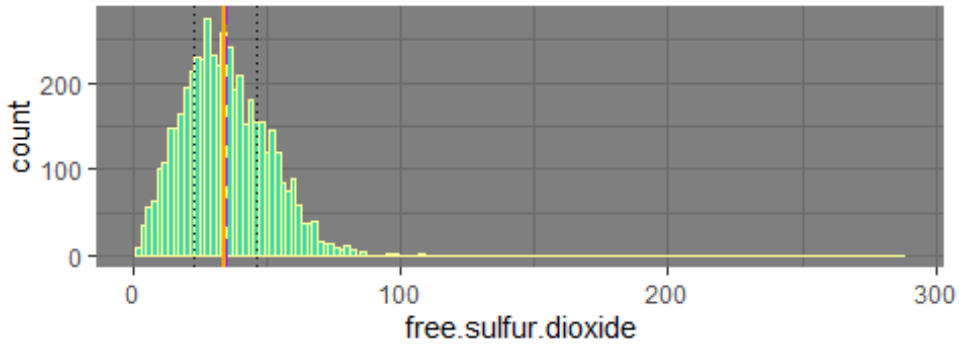
*Purple-Dashed line indicates the mean, orange indicates the median, while the dotted lines are for quartiles. *

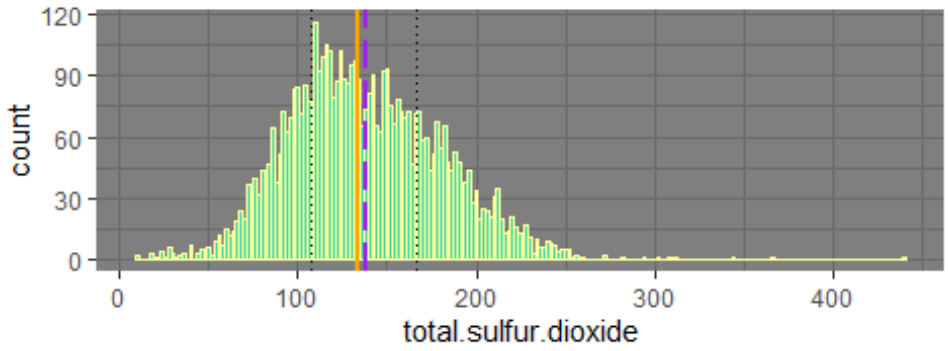
1 dm^3 = 1 Litre

Both Citric and Acetic acid (indicated by volatile acid) are measured in gm/dm^3
and have an almost normal distribution. Citric acid has an unusual peak at 0.49 gm/dm^3, this may be because of presence of many wines from one particular
winemaker or because of any regulations.

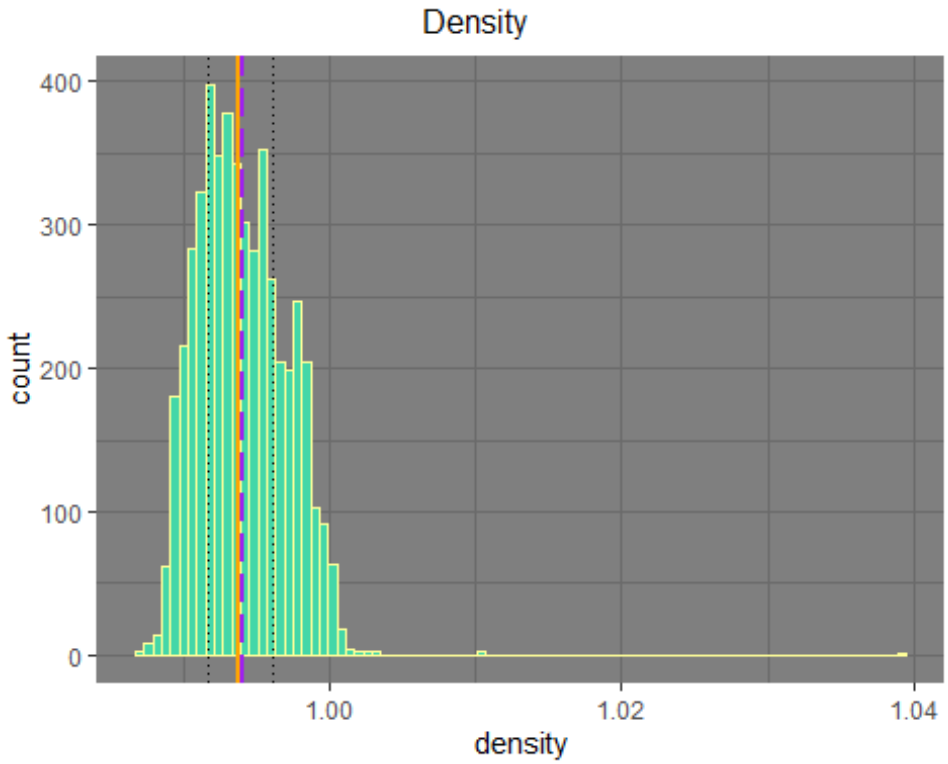


Sulphur Dioxide

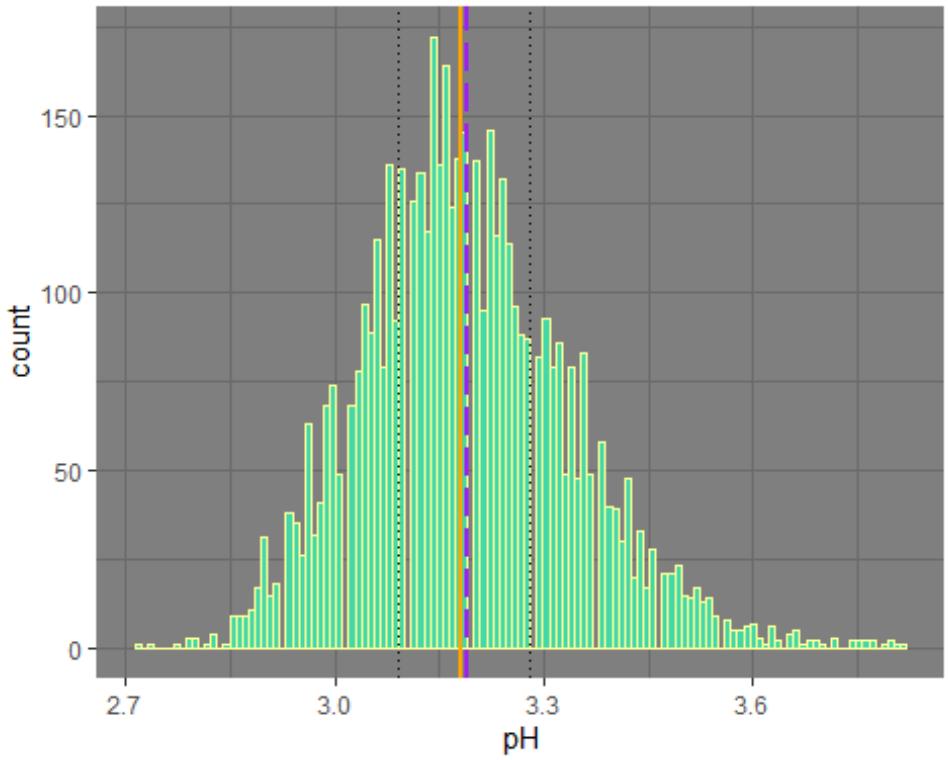




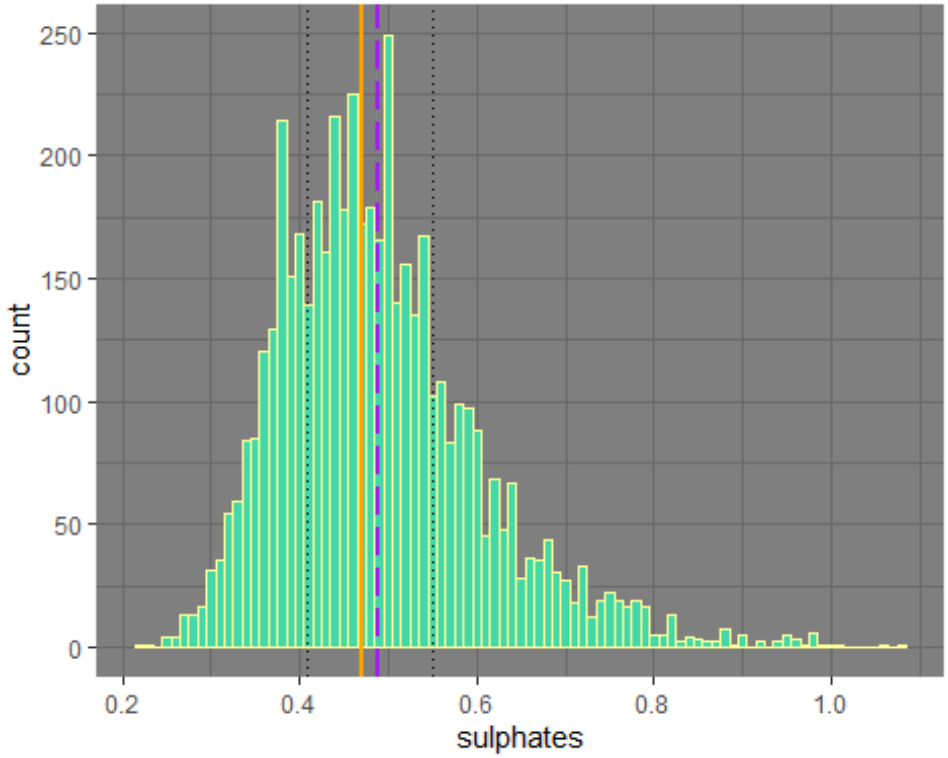
Both Free and Total Sulphur dioxide have symmetric data and both are measured in
mg/dm^3

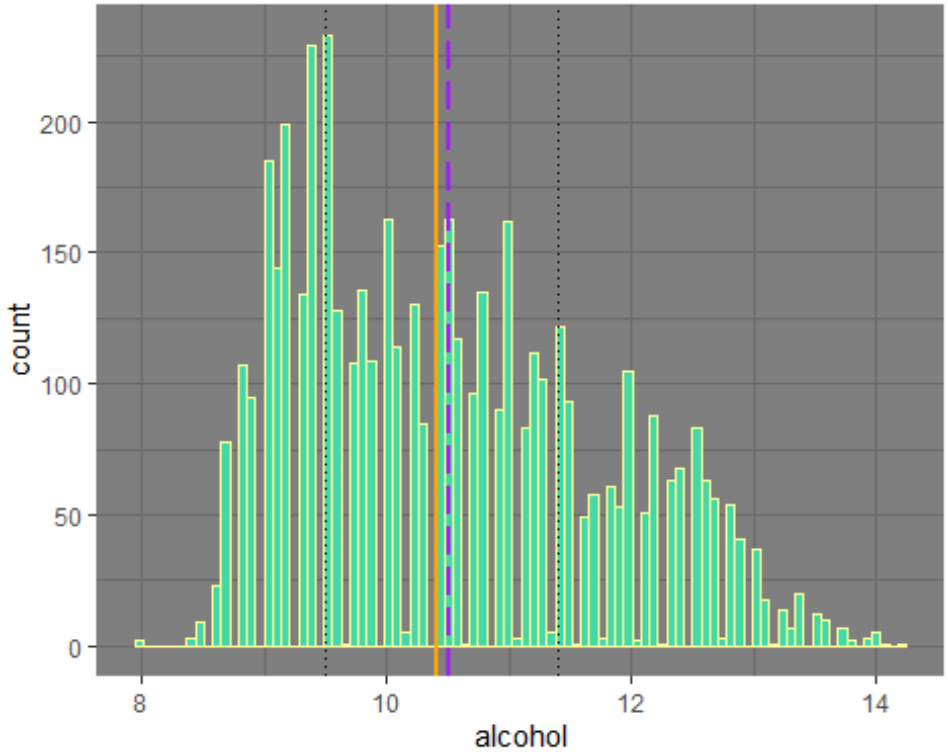


The density data is present in extremely narrow range. As indicated from summary
and the visualisation above min value is .9871 and 3 Qt is at .9961. This is an very interesting observation because with almost 4900 observations this small range
can help making a generalised statement like wines have density of .99 gm/cm^3.

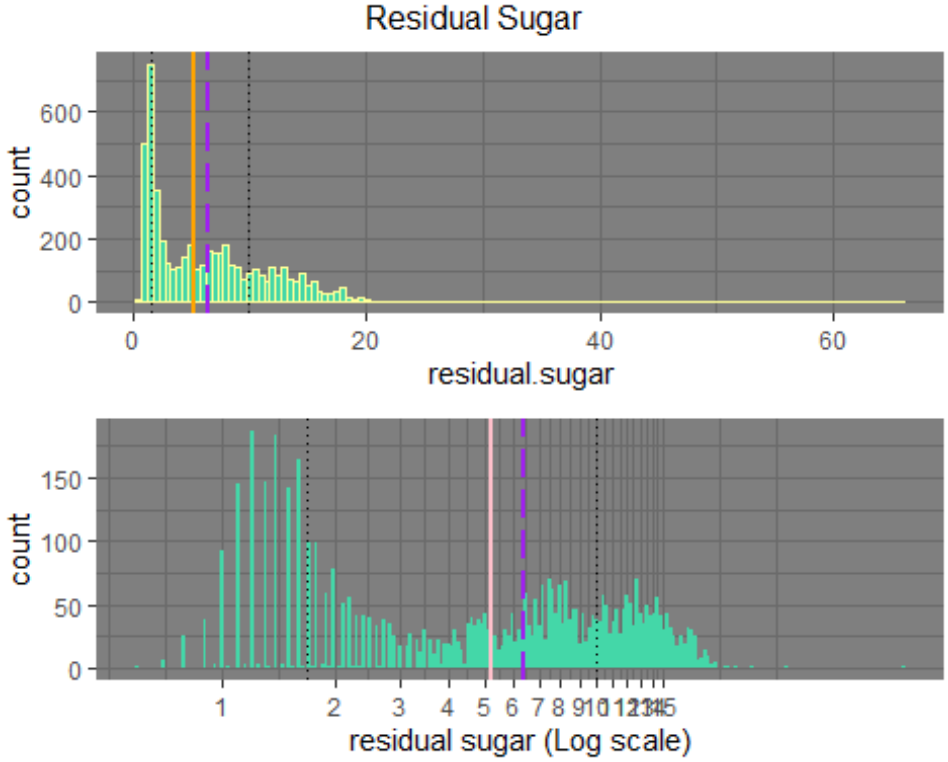


pH range of wine in our dataset 2.7 and 3.8, with average of 3.18.
pH is basically a measure of acidity with 0 being most acidic , 7, neutral and 14 being highly basic.





Alcohol measured in % volume doesn't show much symmetry and have values between
8 and 14 with 75% lying below 11.4. As visible we have most no. of wines with alcohol percent between 9.5 and 9.6.



Purple-Dashed line indicates the mean, orange indicates the median, while the dotted lines are for quartiles.

Residual Sugar, or RS for short, refers to any natural grape sugars that is left over after fermentation ceases. When plotted on log scale showed much better visualisation. Half of the wines have RS between 0.5 and 5 gm/litre while the other half lies between 5 and 20 gm/litre. The data seems to
have a normal distribution between 0 and 5 gm/litre on log scale.

Steps performed

1.Installing packages: GGally, gridExtra, ggplot2, dplyr,tree,class,randomForest,rOCR,caret
2.Fetching the required packages from the library
3. **Developing various plots for exploratory analysis.**
4.Develop a standardization function that will help to standardize certain non----standardized variables.
5. It's important to note that all of these numeric predictor variables (fixed .acidity, volatile.acidity, citric.acid, chlorides, free.sulfur.dioxide, pH, sulphates, alcohol) are not all scaled the same. As such, it's appropriate to scale them before running any analyses.
6. **Create a cleaner dataset for analysis**

7.Creating a Training and Test dataset (Note: Both Training and Test data should wholly represent the original dataset.

7.Building models

a. **Model 1 and 2** – using *tree* function. Decision tree model.
b. **Model 3 and 4**– using *knn* function.
c. **Model 5 and 6** – using *randomForest* function

8.Creating the actual confusion Matrix based on predicted values.

9.Interpreting results.

Results

```
##                     Accuracy Rate Error Rate       AUC
## tree                        0.749      0.251 0.7885253
## pruned.tree                 0.749      0.251 0.7604509
## k=10 kNN                    0.756      0.244 0.6893241
## k=9 kNN                     0.769      0.231 0.6796150
## full.randomForest           0.829      0.171 0.8899367
## small.randomForest          0.806      0.194 0.8517639
```

The random Forest model is highly accurate.

Conclusion

So judging from all of our findings, we have seen that in this case, randomForest is the best algorithm (out of the three we've compared) for classifying this wine dataset. So we have answered the question of what among these three classification algorithms is truly the best.

The decision tree algorithm is useful but ultimately, randomForest is superior version of it since it aggregates many decision trees to create an optimized model that is not susceptible to overfitting. When it comes to interpretability however, a decision tree is preferred. When using a decision tree however it is important to use cross-validation to prune the tree in order to narrow it down to the most important variables.

Compared to decision trees, the k-nearest neighbor algorithm has a slightly greater accuracy rate but a worse AUC. The decision tree method did however help to narrow down the three most relevant attributes: `alcohol`, `volatile.acidity`, and `free.sulfur.dioxide`. This finding was consistent with when we took a look at the most important variables in the randomForest model.

We were able to apply this subset of attributes to the randomForest algorithm and come out with a strong model that only utilizes a few independent variables in order to classify at a high success rate. This lends strength to the argument that these three variables are the most relevant when it comes to determining the content of a good wine.

As far as what these variables' importance is in reality, is that sulfur dioxide is crucial for killing bacteria in wine when creating it. On the other hand, volatile acidity is an undesired trait in wine that affects flavor, that can be caused by such bacteria. So it makes sense that wine that is high in sulfur dioxide, and low in volatile acidity, is considered good.

The pending questions that remain are, did we overfit or underfit to the training data when testing these different classification methods? It is also worth determining exactly the threshold for the amounts of these variables such as `alcohol`, for example finding the optimal amount of alcohol content to create a good wine.

We would also like to delve more into how best to select some kk for kNN that maintains a high level of accuracy while also having a balance between bias and variance without either over or underfitting. We would also posit a similar question for the number of nodes in a decision tree. Finally, is dropping variables in randomForest really necessary, if the randomization inherent in it already accounts for overfitting?

If we can only compare models that utilize the same set of predictors, then we should look at the pruned classification tree against the randomForest model utilizing the same attributes. We see even there that the randomForest model is superior.

In conclusion we have found that randomForest is best for binary classification and that alcohol, volatile acidity, and free sulfur dioxide are the most important predictors when attempting to classify a good wine.

References

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction To Data Mining. 1st ed. Addison Wesley: Pearson, 2005. Print.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Sing T, Sander O, Beerenwinkel N and Lengauer T (2005). "ROCR: visualizing classifier performance in R." *Bioinformatics*, *21*(20), pp. 7881. <URL: http://rocr.bioinf.mpi-sb.mpg.de>.

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

Brian Ripley (2016). tree: Classification and Regression Trees. R package version 1.0-37. https://CRAN.R-project.org/package=tree

Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.