Predicting Wine Quality

Capstone Project by Alberto Chaves

Introduction

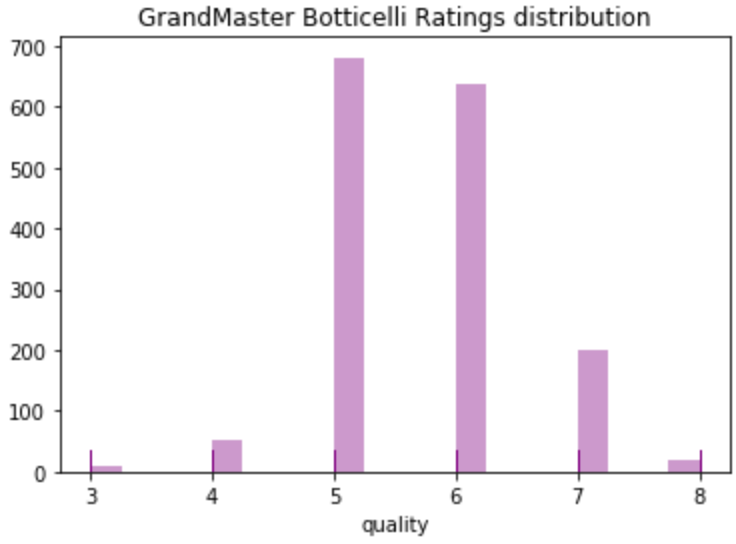
Chaviticus Wineries, produces at least 30 wines with a variety of flavours. His owner, Alberto, has set himself the goal of getting into the top 10 wines for the year, given by GrandMaster Botticelli. For this contest, wineries can only send 3 wines, so he has to choose among his whole stock, which three to send.

Given that wine ranking is so subjective to the taster, Alberto wants to use the chemical compositions and the ratings given by GrandMaster Botticelli in the past, and try to predict the ratings of his own stock, and then send the ones the GrandMaster would probably like best.

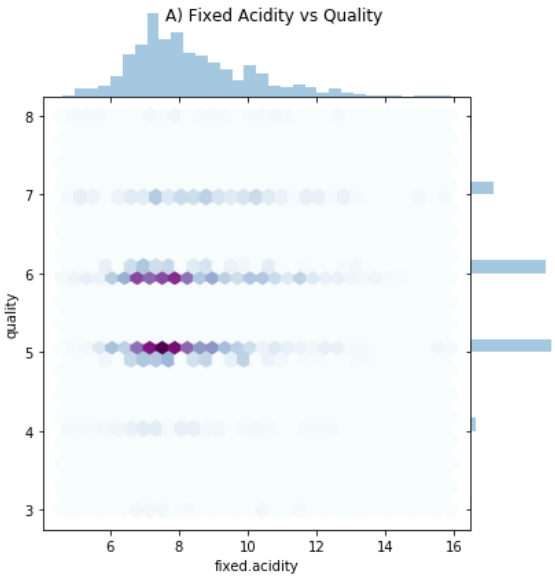
Data

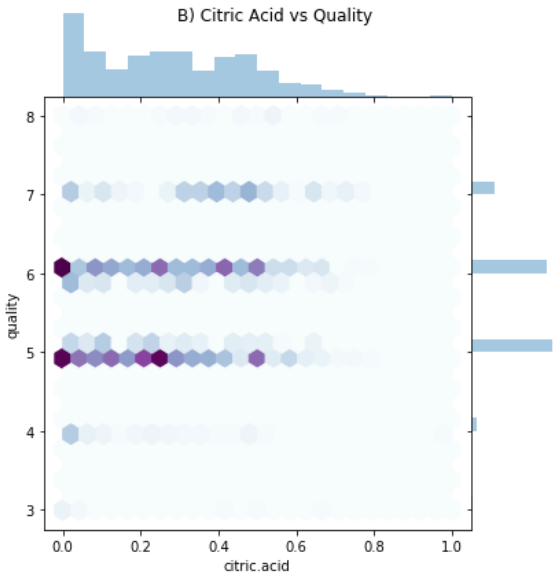
Quality

As we can observe, Grandmaster's ratings are mostly 5 or 6, and seldom gives a rating above 7. This is important to consider when we try to predict, since the predictions should deliver a similar distribution.



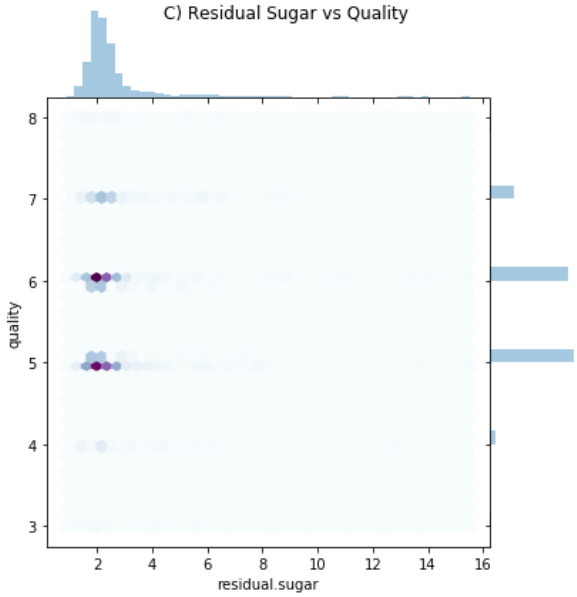
Chemical components distribution vs Quality

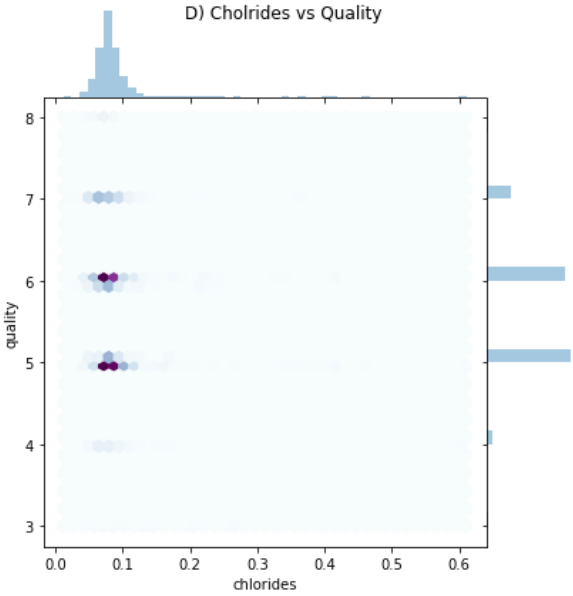




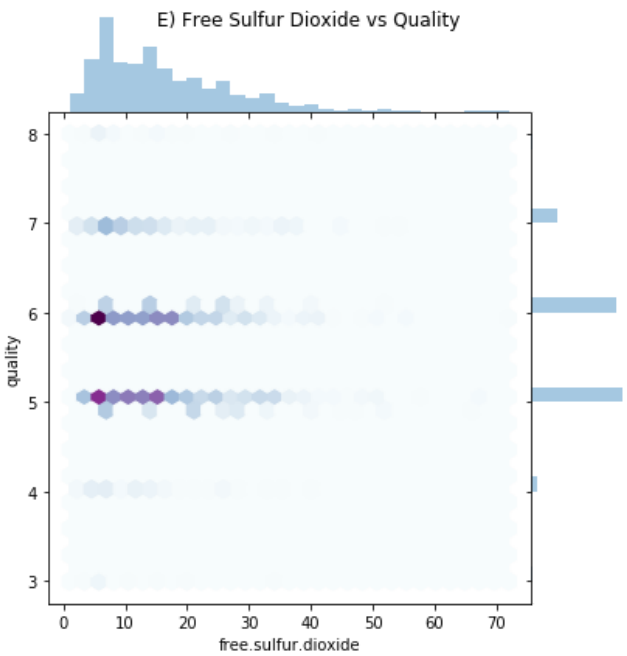
A)- **fixed acidity** - most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

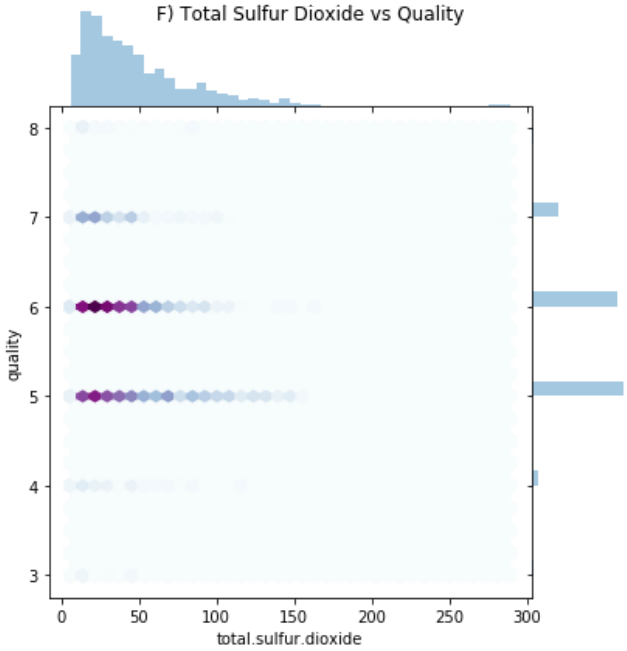
B)- **citric acid** - found in small quantities, citric acid can add 'freshness' and flavor to wines





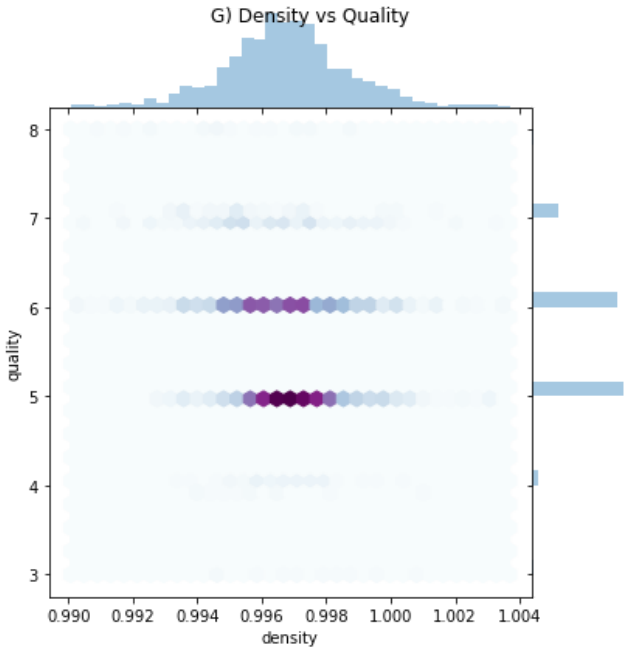
C) - **residual sugar** - the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered.
D) - **chlorides** - the amount of salt in the wine

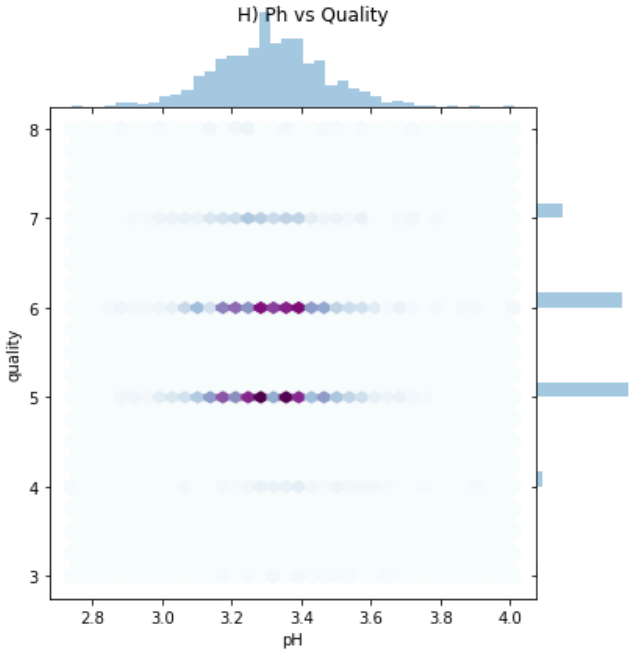




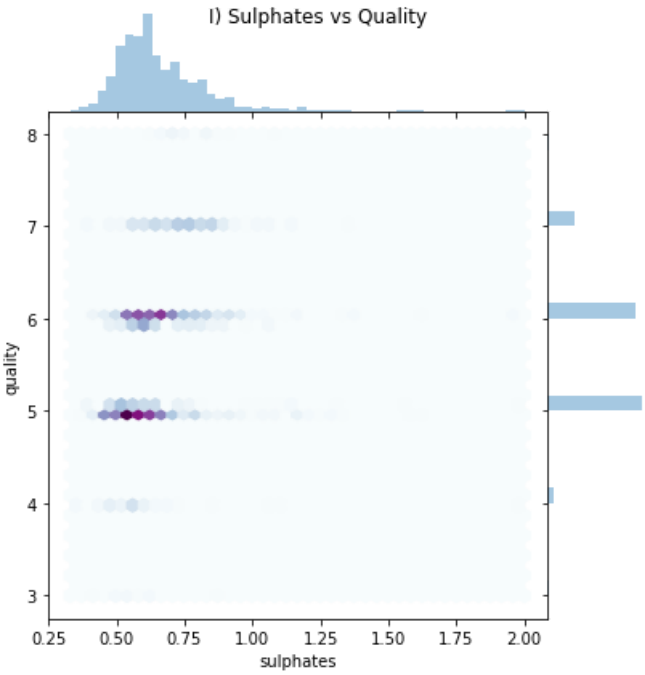
E) - **free sulfur dioxide** - the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.

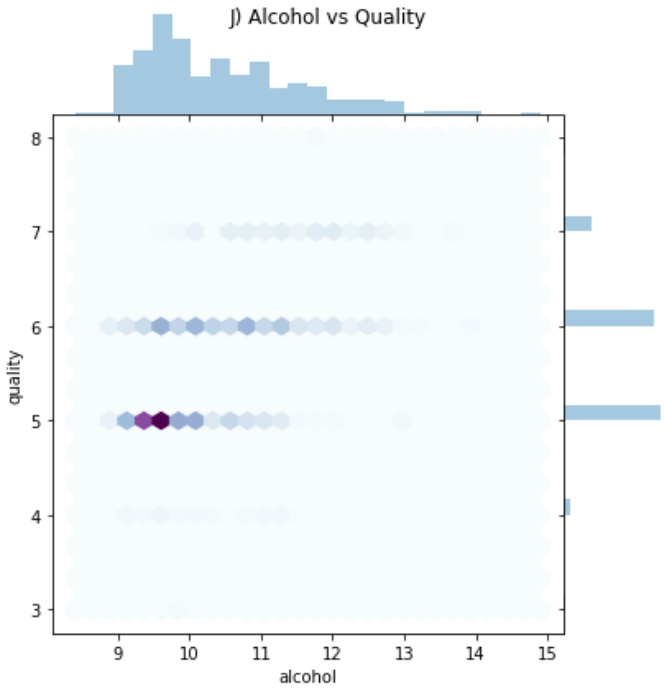
F) - **total sulfur dioxide** - amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.





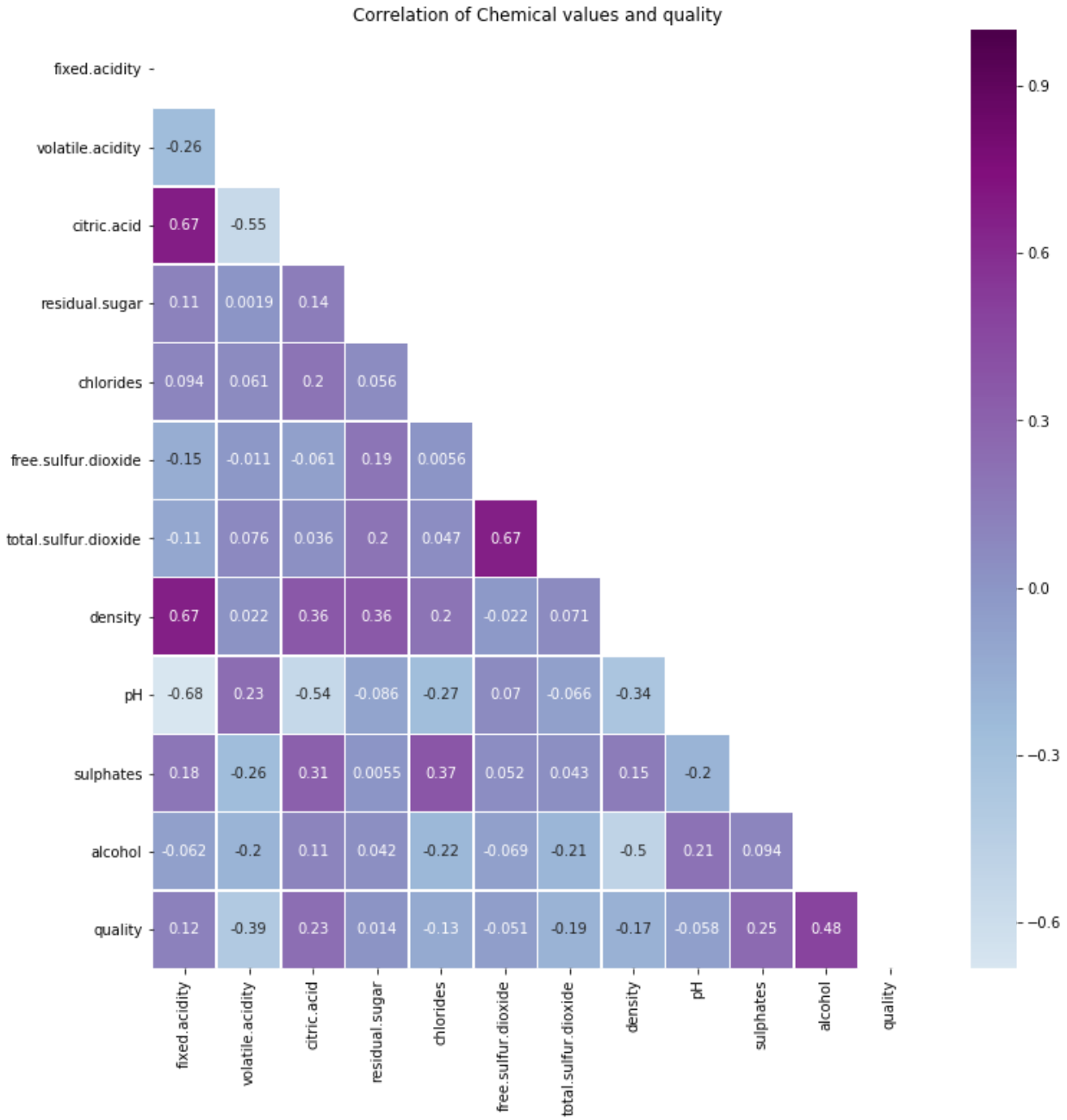
G) - **density** - the density of water is close to that of water depending on the percent alcohol and sugar content.
H) - **pH** - describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale





I) - sulphates - a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
J) - alcohol - the percent alcohol content of the wine

Preprocessing and feature engineering



Correlation information shows alcohol with the highest positive correlation to quality, but it is not a very strong one. The lowest negative correlation is volatile acidity, which at high levels can lead to unpleasant vinegar taste, so it makes sense that the correlation is negative. No features need to be dropped since there seems to be no risk of overfitting.
Data was normalized for those models who perform better with normalized data (linear regression, support vector machines)

Modeling

Four models were used for this study:

- Linear regression
- Logistic regression
- Support Vector Machines
- Random Forest

They were tuned using the GridsearchCV exhaustive search over specified parameter values for an estimator.
The following parameters were used for each model:

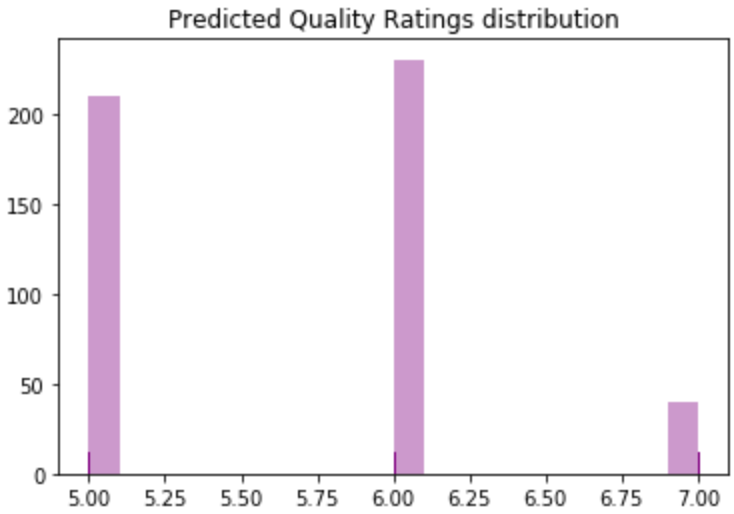
Logistic regression - 'C': 0.505060404040404, 'penalty': 'l2', 'solver': 'liblinear'
Support Vector Machines - C=100, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma=0.001, kernel='rbf'
Random Forest - 'max_depth': 8, 'max_features': 'auto', 'n_estimators': 200
Linear regression

Model	MSE	RMSE
Linear Regression	0.46	0.67
Logistic Regression	0.47	0.68
Support Vector Machines	0.37	0.61
Random Forest	0.36	0.60

Random Forest was chosen for the prediction, because of the lower RMSE value. It is important to remember that RMSE is on the same scale as the Quality variable, which means the value predicted differs from the real value in up to 0.6.

Results

Predicted Quality ratings were rounded and plotted to see the distribution shown in the next figure. This distribution is similar to the one seen on the quality ratings in the data analysis, but no outliers.



Conclusions

The error is not optimal, and could be diminished by other models available. Quality in this case is a subjective rating, given by the same GrandMaster through different years. It makes sense that it is a little hard even for the models to have an exact result. Nonetheless, the predictions can give Alberto a clearer picture of which ones to filter out from the initial list of 30 options, and some insights into the GrandMasters' taste for his wines.

Dataset credits:
P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.
In Decision Support Systems, Elsevier, 47(4):547-553, 2009.