Analysis and Modeling of the Spanish Wine Dataset

Bogdan Tesileanu

Johns Hopkins University

Table of Contents

Introduction 3

Data 4

Methodology 5

Analysis 8

Results 13

Conclusion and Recommendations 17

Introduction

The following is an analysis of Spanish wine to assess quality among different categories of wine. It represents an overview of the processes taken to analyze a dataset and build a predictive model, and how to deal with hurdles along the way. A production model would require far more tuning than is demonstrated here, but the process and steps are the same.

Data

The “Spanish Wine Quality Dataset” was obtained from Kaggle, published by user “fedesoriano”. It consists of 7500 entries and 11 features. The abridged data dictionary is below:

- winery: Winery name
- wine: Name of the wine
- year: Year in which the grapes were harvested
- rating: Average rating given to the wine by the users [from 1-5]
- num_reviews: Number of users that reviewed the wine
- country: Country of origin [Spain]
- region: Region of the wine
- price: Price in euros [€]
- type: Wine variety
- body: Body score, defined as the richness and weight of the wine in your mouth [from 1-5]
- acidity: Acidity score, defined as wine's “pucker” or tartness; it's what makes a wine refreshing and your tongue salivate and want another sip [from 1-5]

The data was web-scraped by the user and published in April 2022. It contains null values in the year (290 times), type (545 times), body (1169 times), and acidity (1169 times) columns. The rating, price, acidity, and body are all classified as float datatypes; num_reviews is classified as an integer datatype, while the rest of the features are objects.

Methodology

The data requires some preprocessing before it can be used. To start, the null values will be analyzed. At first, the year feature only came up as having 2 null values, but further inspection revealed that 288 of the non-null values were encoded as “N.V.”, which were subsequently changed in nulls. None of the other columns had any other inappropriate labels.

The decision was made to impute the missing values instead of deleting whole rows or columns as doing so would have reduced the size of the dataset by over 15%. However, since some of the missing values were categorical, methods such as finding the mean or median would not have yielded results. Instead, a clustering algorithm was used to group similar entries together, and where null values existed, impute the attributes of the local cluster. This was done by first label-encoding the categorical values, and then performing a k-mean algorithm with Expectation Maximization until convergence (k-POD algorithm by Chi, Chi, and Baraniuk).

The next step was to focus on the perceived quality of wine. While the dataset includes a rating and number of reviews for each type, this was deemed insufficient to assess quality. As a result, the Bayesian Average was computed for each entry, which took into account both the ratings and the number of reviews and combine them into a new score, given the name “quality”. The score is continuous from 4.2 to 4.7, with 4.7 being the best. The formula for the Bayesian Average is as follows:

$$\text{bayesAvg} = \frac{\text{productRatingsAvg} \times \text{productRatingsCount} + C \times m}{\text{productRatingsCount} + C}$$

where

C = 25th percentile of the number of ratings, 389 for this dataset

m = Overall average rating for all rows, 4.254933 in this case

Additionally, since the original rating feature was not continuous, another column was added named “qualcat” (for quality, but categorical), which was produced by binning the quality feature into 5 categories:

1: for quality < 4.3

2: for 4.3 ≤ quality < 4.4

3: for 4.4 ⩽ quality < 4.5

4: for 4.5 ⩽ quality < 4.6

5: for 4.6 ⩽ quality

The new features and clean data make it possible to analyze the data and create the model. After analyzing the data, both classification and regression models were used to identify the best one.

For the classification round, a label encoder was used to quantify the categorical variables so that they can be understood by the classifiers. Next, all the values were standardized to have a mean of 0 and a standard deviation of 1. This was done to prevent higher values from having a greater impact on the algorithm. Lastly, because of the large differences between the number of samples for each “qualcat”, undersampling and oversampling were performed to ameliorate some of the biases created toward minority classes. Many machine learning algorithms are influenced by quantity of samples, and a small sample size may cause the algorithm to completely ignore that class. As the last step of preparation, the data was split into 80% training set, and 20% testing set.

For the classification 9 different algorithms were ran through hyperparameter tuning and cross-fold validation to achieve the best response. The algorithms that were tested were:

- Random Forest Classifier

- Support Vector Classifier
- Logistic Regression
- Decision Tree Classifier
- K-Neighbors Classifier
- Gradient Boosting Classifier
- Extra Trees Classifier
- Bagging Classifier
- XGBoost Classifier

Once the best parameters were achieved, feature importances were extracted and insights were made. Next, a similar process was performed but using regression models. In this case, the target variable “quality” was used.

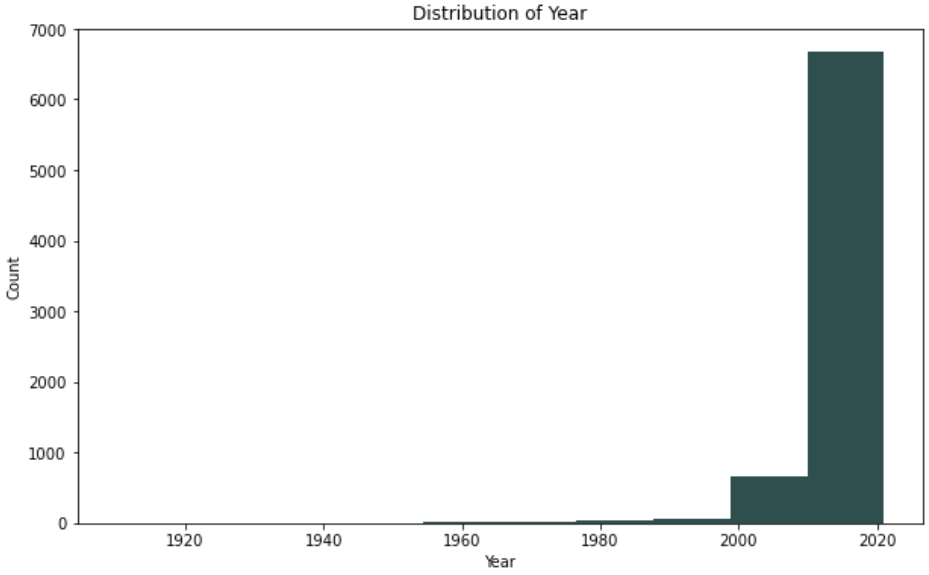
First, instead of label encoding the variables, dummy variables were used to avoid any kind of ranking influence perceived by the algorithm. This caused an increase to 1425 variables, which some algorithms may struggle with. Nevertheless, the algorithms chosen were:

- Linear Regression
- Lasso Regression
- Ridge Regression
- Bayesian Ridge Regression
- Decision Tree Regression
- Random Forest Regression

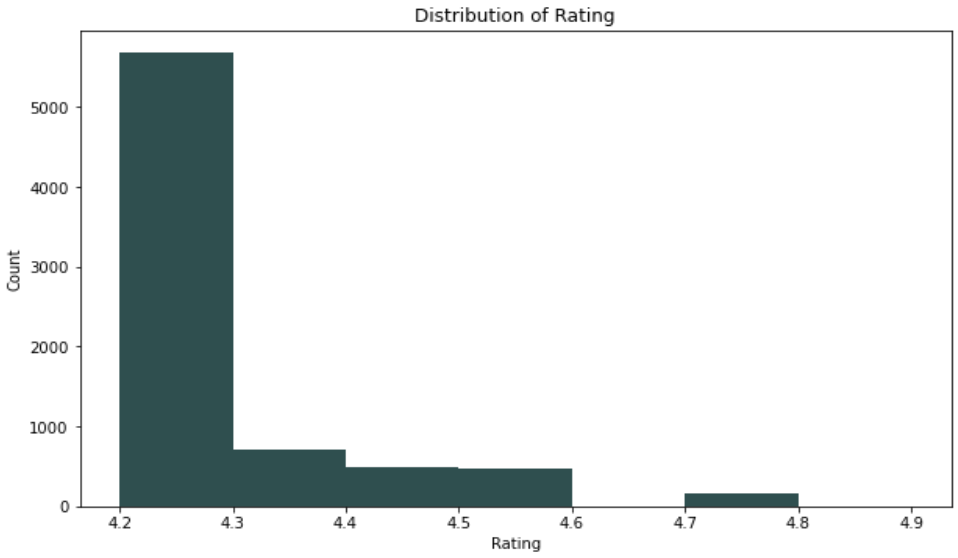
Similarly to the classifiers, the feature importances were gathered and conclusions were made.

Analysis

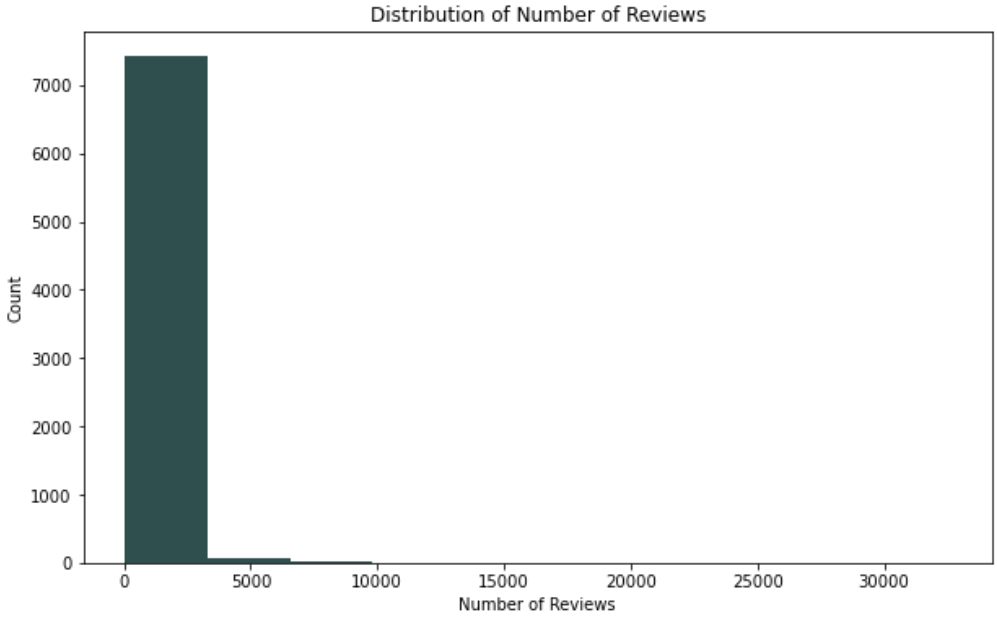
Initial familiarization with the variables yielded some interesting findings. First, more than 75% of the wines in the dataset are from year 2011 or later, with the average being 2013. Nevertheless, the oldest wine included in the dataset is from 1910.



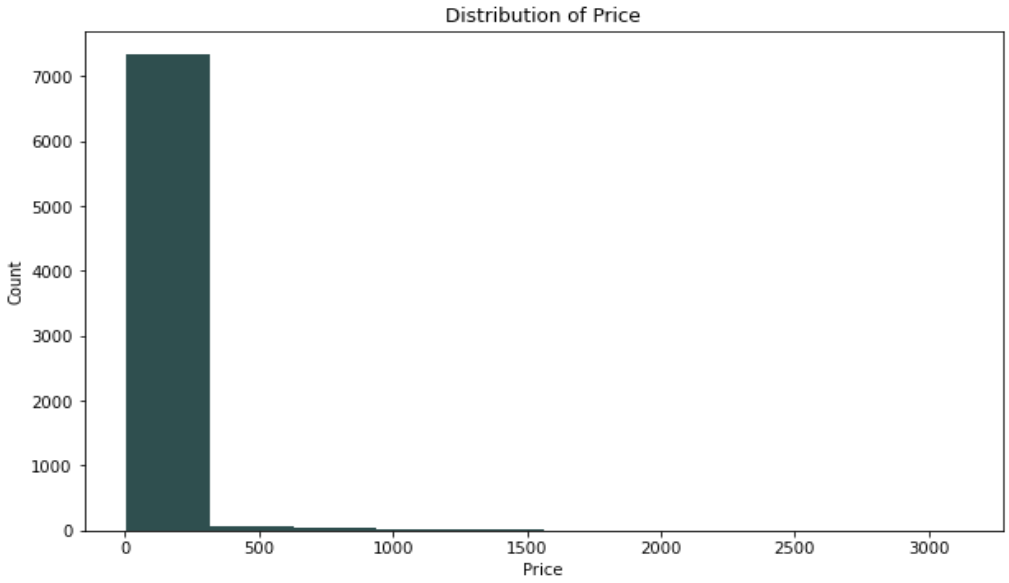
Next, the average ratings for all the wines is about 4.25 on a scale from 4.2 to 4.9. The standard deviation for rating is about 0.12. Over 75% of the wines have a rating of 4.2. Because of this, the newly created "quality" feature may be more useful in finding insights.



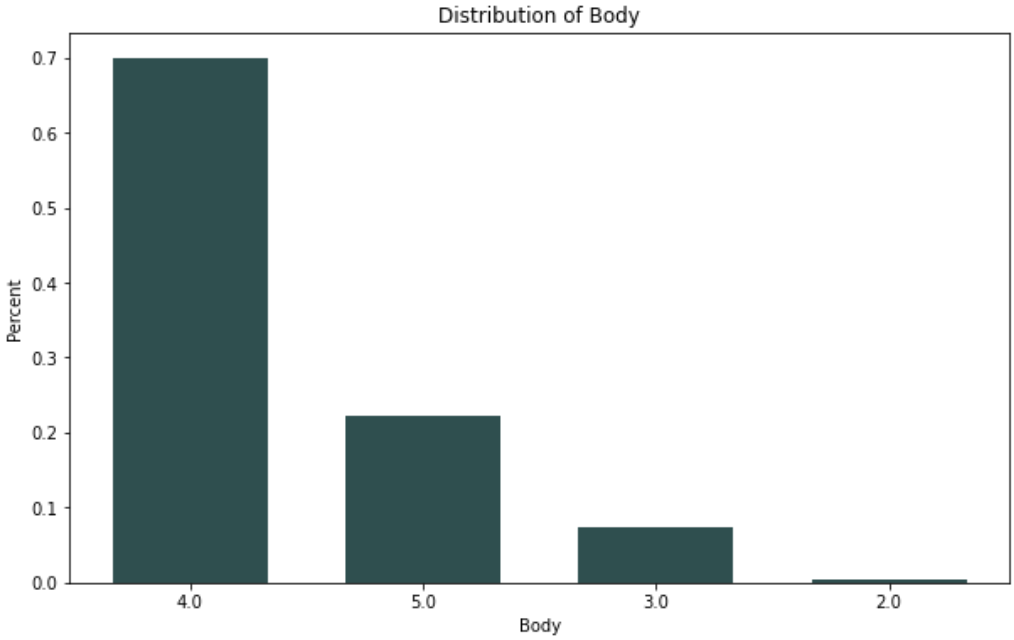
Likewise, most of the wines received under 500 reviews, but the largest amount was 32624. The smallest amount was 25, but there may still be signal in the data even with such a low number. The average number of reviews was 451.



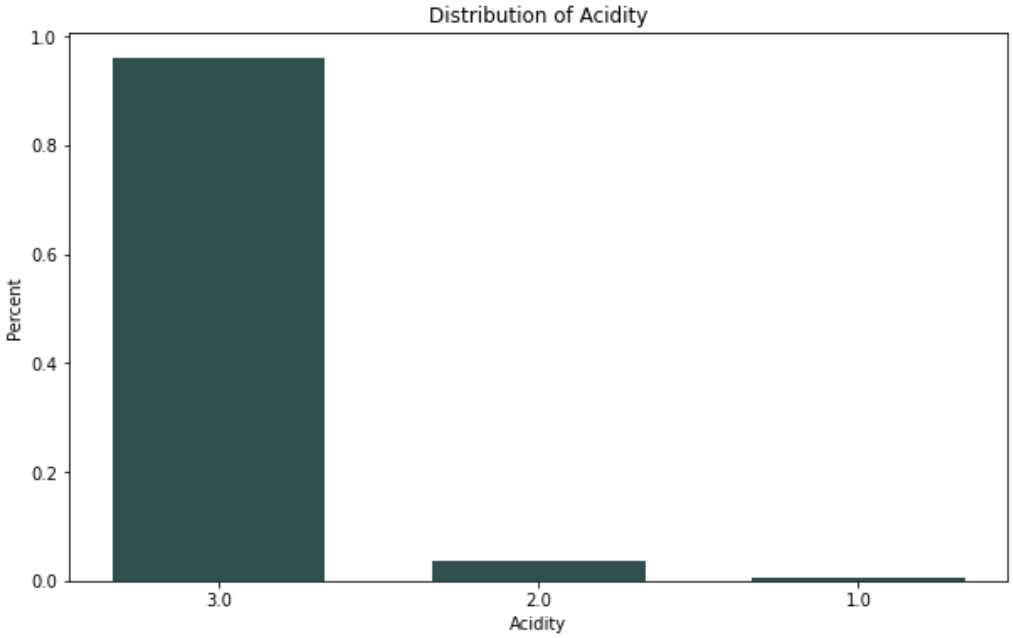
Next, the price also had a similar distribution to the reviews, with 75% of the wine costing less than $51.35, but the most expensive being $3119.00. The average price was about $60.00, thus we can assume that the data was largely right-skewed.



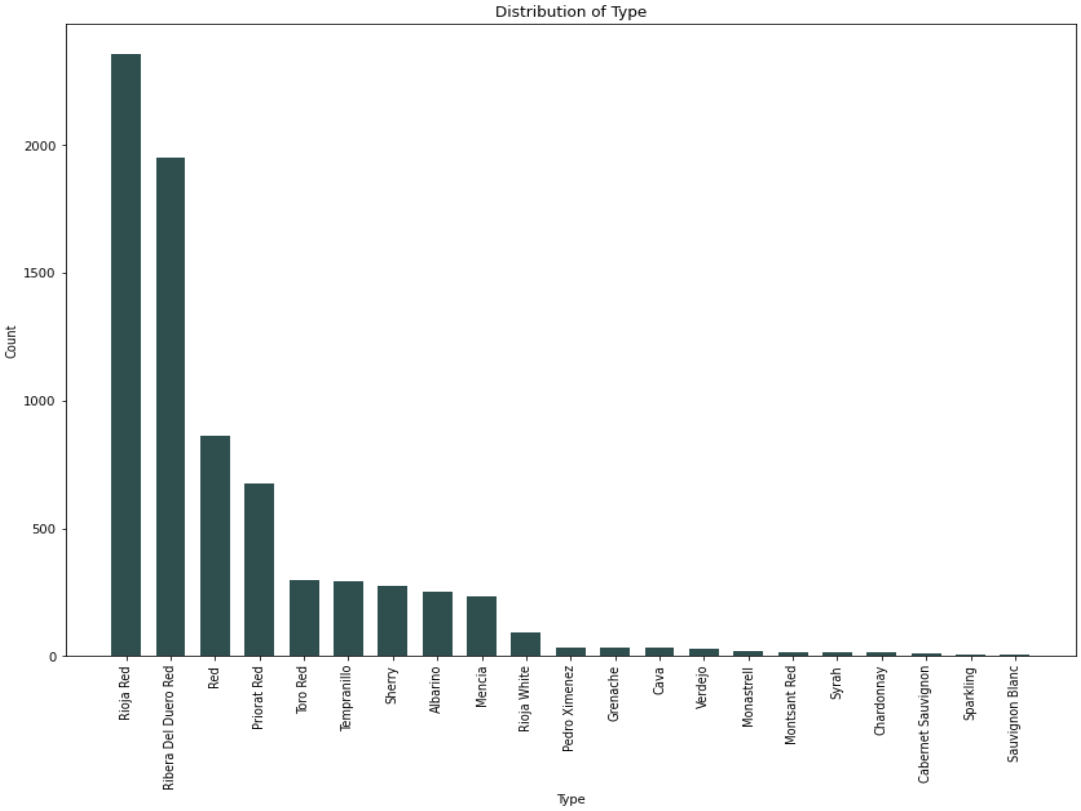
The body of the wine refers to the richness and weight of the wine. In this dataset, the body score was between 2 and 5, with the average wine being about a 4. The two graphs below are normalized to show percent of total distribution.



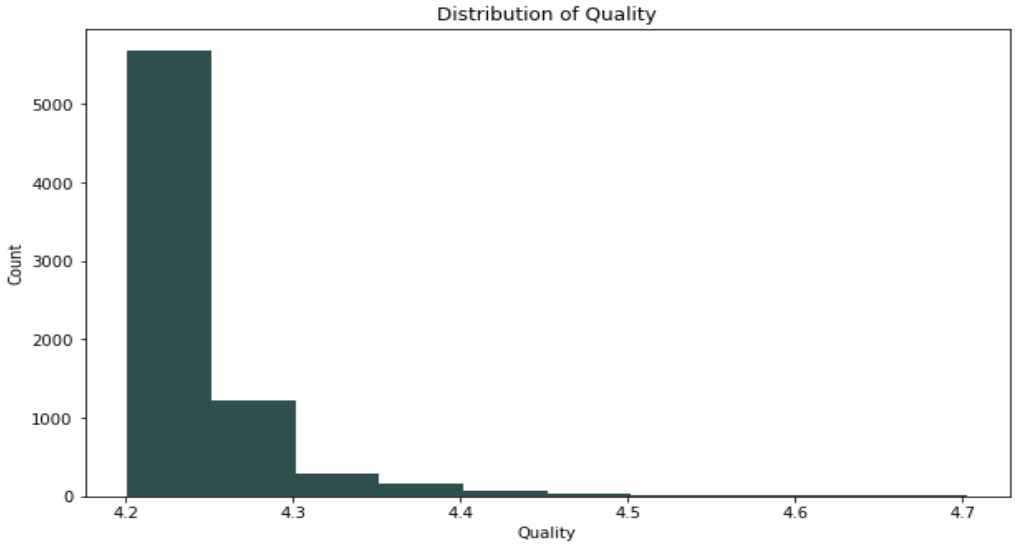
Similarly, acidity was measured on a scale from 1 to 3, with the average wine having an acidity of 3.



There are 21 different types of wines on the list, with red wines making up the top 5, or more than 85% of the wines included.



Finally, our target variable “quality” had an average of 4.25, not unlike the ratings. However, since the values range only from 4.2 to 4.7, the standard deviation is smaller at 0.05.



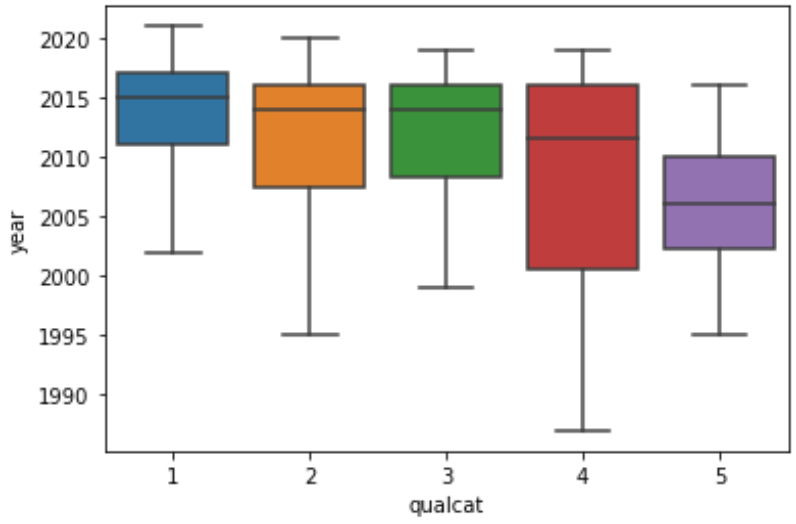
Next, we will perform some pairwise analysis with our variable of interest “quality”. To get an overall idea, a heatmap of the correlations between the variables was created.

The high correlation between quality and qualcat can be ignored, but we do notice a relatively high correlation between price and qualcat of 0.39. Others that stand out are year and body, with both being about 0.2. The above findings, if accurate, should be confirmed by the models.

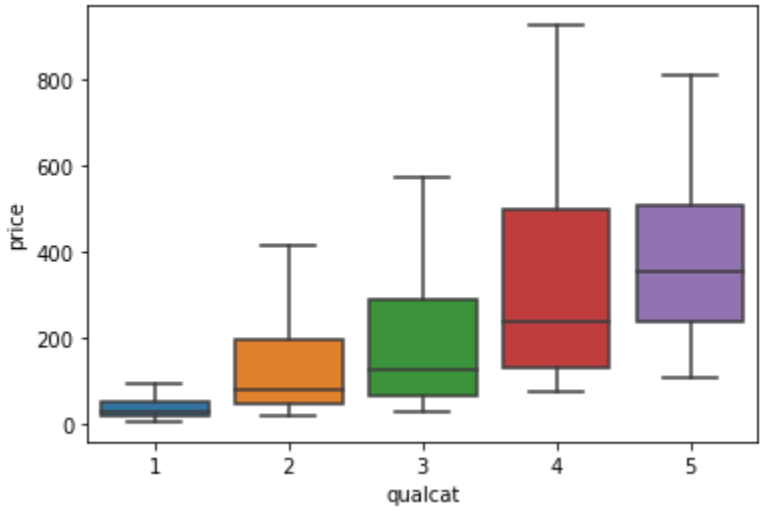
Results

For the classification models, the best performance was delivered by the Random Forest Classifier with n_estimators = 100 and max_depth = 20. After performing 5-fold cross validation, the accuracy score was 0.85 on average. Upon extracting some of the data from the model, it was discovered that the most important features that contributed to quality were year, price, winery, and wine.

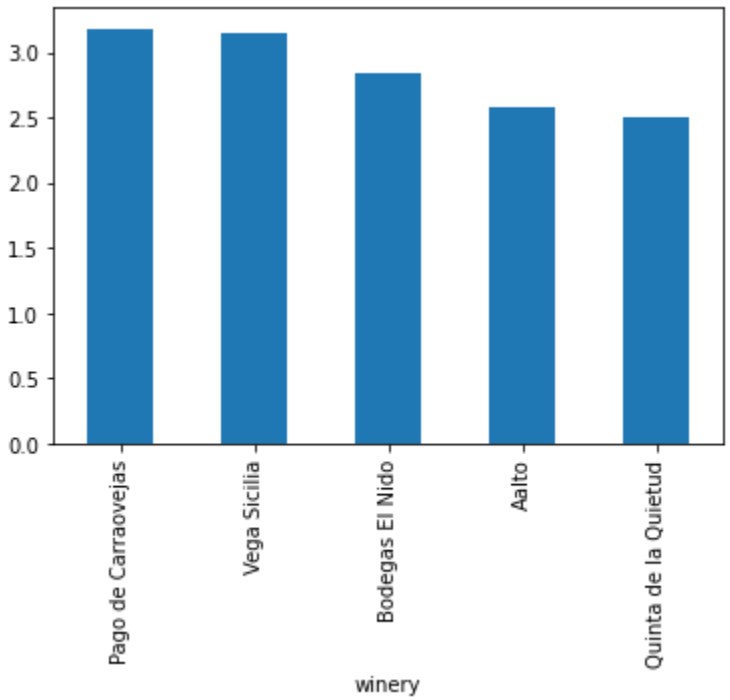
Digging a little deeper revealed that for qualcat of 4 or 5, the average year of the wine was 2005.



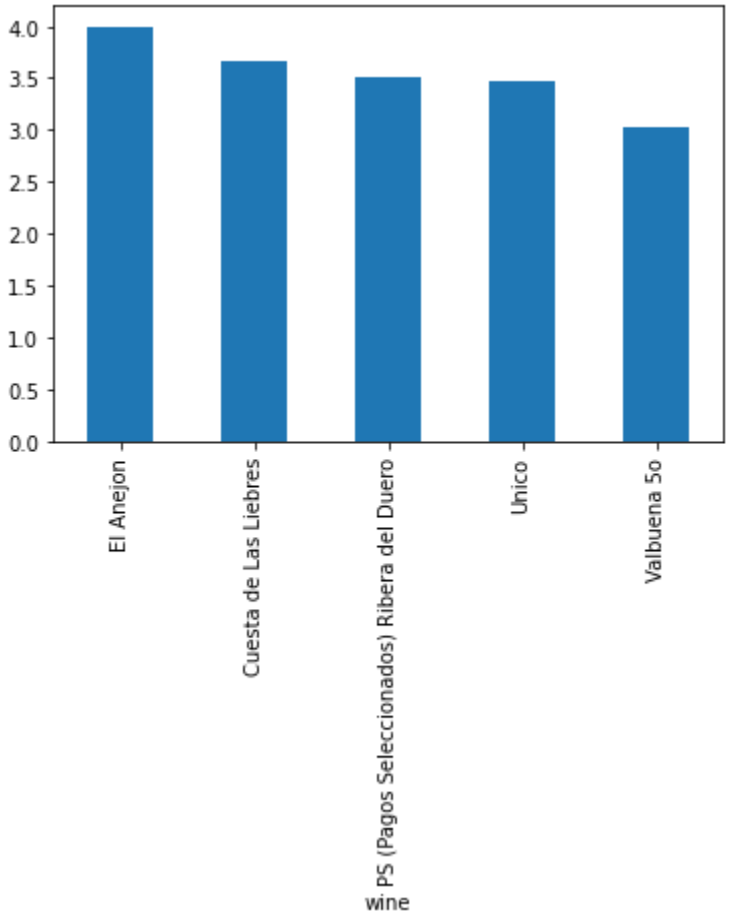
Similarly, the average price of qualcat 3, 4, and 5 were in the $300s, while for 2 and 1 were much lower.



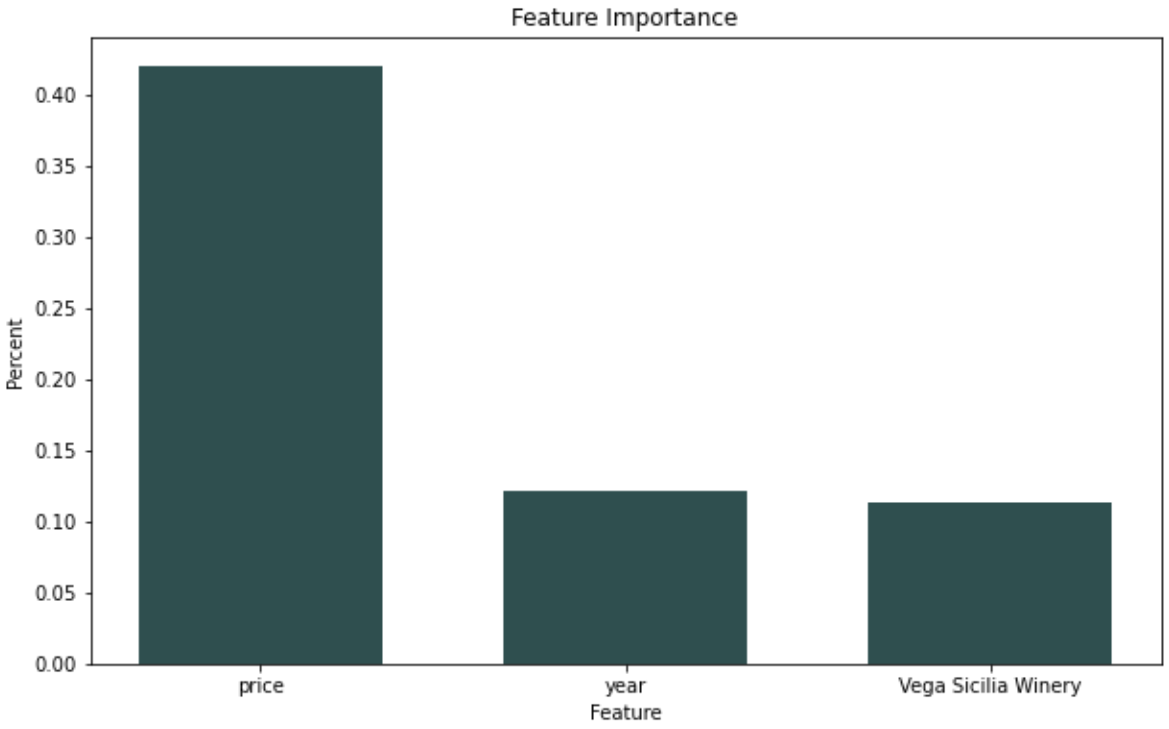
As for the winery, it seems that the top 5 wineries with the highest quality of wine were Pago de Carraovejas, Vega Siciliana, Bodegas El Nido, Aalto, Quinta de la Quietud. Below is each of these wineries with their average wine score.



Lastly, the type of wine also made a different on the quality. The 5 highest quality wines were El Anejon, Cuesta de Las Liebres, PS Ribera del Duero, Unico, and Valbuena 5o.



Not surprisingly, the results of the regression model were not much different. However, an interesting insight was discovered by creating dummy variables instead of label encoding. The best regression model turned out to be the Random Forrest Regressor with an R^2 score of 0.82. However, what was most impressive about the model was not its accuracy, but the top 3 features it revealed. As expected, price accounted for 42% of the data, while year accounted for 12% of the data. Surprisingly, the winery Vega Siciliana accounted for 11% of the quality of wine. This winery also appeared in the classification model with an average qualcat of wine of 3.1.


Feature Importance
Percent
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
price
year
Vega Sicilia Winery
Feature

Conclusion and Recommendations

This preliminary analysis and modeling of the Spanish Wine dataset revealed several interesting findings. The accuracy of the models and agreement between the two approaches inspire confidence that the model would be successful in predicting quality if taken to production, but not after more tuning and analysis. With more time, it would be interesting to analyze the price point of each wine and a sell-through rate. To achieve this, more data would be required such as the selling price, distributor price, price of manufacturing, as well as consumer data for each winery. Combining the important features of the model with some optimization of the price and quantity of bottles produced could aid a store selling wine increase revenues by buying and selling the right quantity of each wine.

To maximize quality, the CEO should focus wines in the $300 per bottle price range from around 2005. It is also advisable to create relationships with the top 3 wineries, but especially Vega Siciliana since it seems to produce the highest quality wine. Additionally, a focus on the 3 highest quality wines El Anejon, Cuesta de Las Liebres, PS Ribera del Duero would be advisable.