Wine Scores

A report by Jorge Londono

Problem Statement

- With different wineries competing against each other to provide their best ingredients and mixtures, it can often be overwhelming for consumers to get a grasp of what to look for in a bottle of wine
- Luckily, there are expert reviewers and sommeliers who can assign point values based on the quality of a wine
- However, while we can look at specific reviews to figure out which wines perform well for these reviewers, this is a surface level into analyzing what makes wine great.
- What are the best predictors we can find to foresee the quality of a bottle of wine?

Benefactors of this Project

- Wine Producers
- Wine Consumers and enthusiasts
- Wine Industry Media

Data and Data Cleaning

Origin of Data

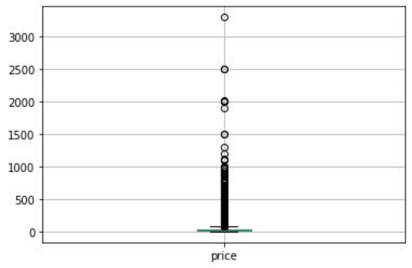
This data was provided by a kaggle user who scraped it from the Wine Enthusiast Magazine in 2017. It contains over 129,000 wine reviews

Missing Values

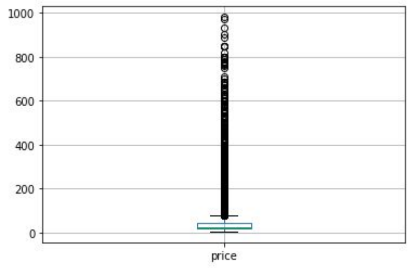
Using the ‘info’ method showed that there were around 9,000 observations where the price variable was missing.

In order to have an accurate predictive model, the observations with these null values were removed using the ‘dropna’ method on the data frame.

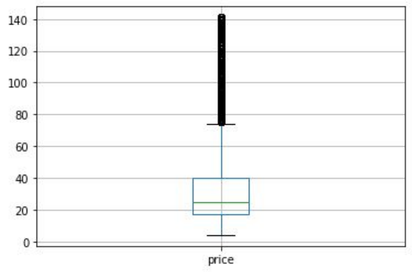
Dealing with Outliers



Initial data

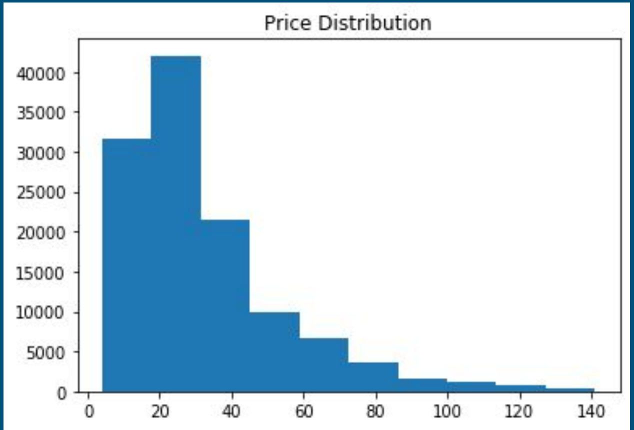


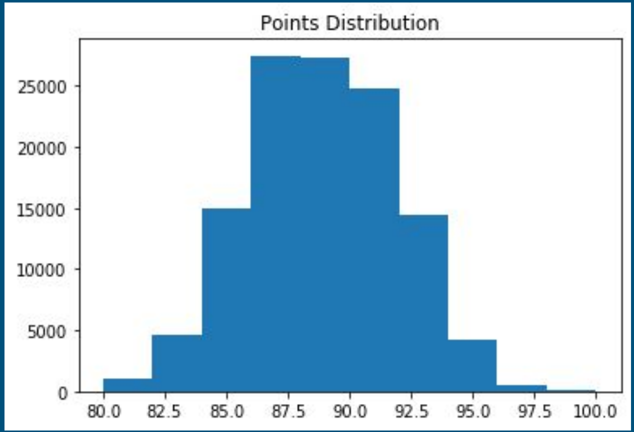
Data with wine priced over $1,000 removed



Data with wine priced with a Z-score over 3

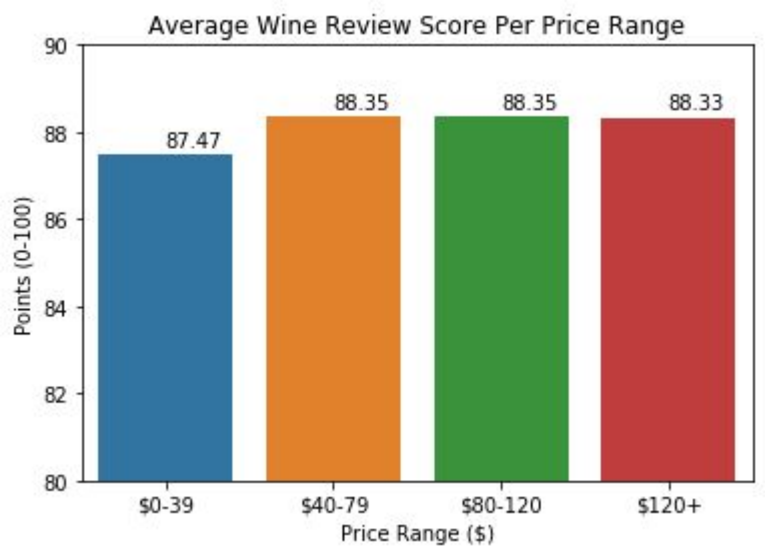
Exploratory Data Analysis





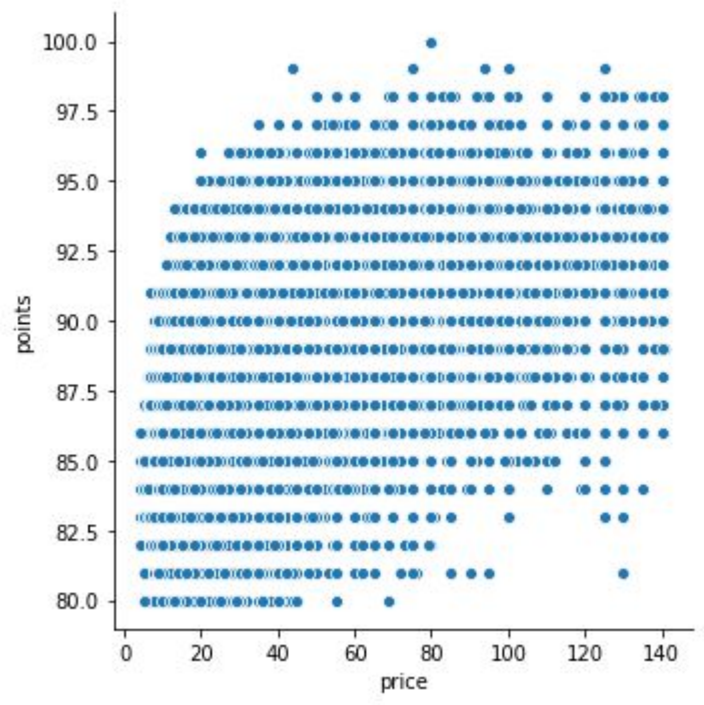
From the graphs above, we can see that the points column of our data frame appears to be normally distributed. However, the price column of our data set seems to be right skewed and might need to be transformed via a logarithmic function.

Comparing Price Groups



Separating the wines by prices shows us that there is a difference between the most expensive and least expensive wines.

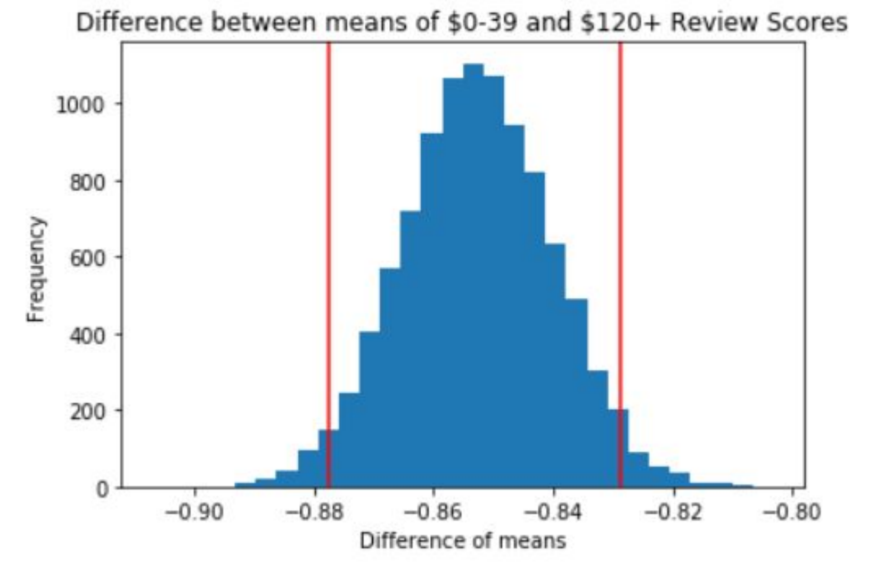
Correlation analysis



- Pearson correlation coefficient: 0.55 between price and review points.

- Moderately positive correlation

Bootstrapping



- Null Hypothesis: wine under priced $40 has the same mean score as wine priced $120 and over
- Alternate Hypothesis: he mean score of wine priced under $40 is not the same as the mean score of wine priced at $120 and over.
- Because this interval does not include 0, at 95% we can reject the null hypothesis that the wine priced under $40 has a mean review score equal to that of wine priced at $120 and over.

Algorithm Selection

- This is a regression problem since we are trying to predict a continuous value with our features.

- Lasso and Ridge regression are two regression algorithms that will be optimal seeing as how our model might suffer from high dimensionality.

Parameter Tuning

Using GridSearchCV() we found the best alphas for both Ridge and Lasso models.

Ridge alpha:1

Lasso alpha: 0.00001

Results

Model	R^2
Algorithm: Ridge Features:Price Parameters: Default	0.304
Algorithm: Ridge Features: Price(Log) Parameters: Default	0.366
Algorithm: Ridge Features: Price(Log), Country Parameters: Default	0.393
Algorithm: Ridge Features: Price(Log), Province Parameters: Default	0.415
Algorithm: Ridge Features: Price(Log), Province, Variety Parameters: Default	0.430
Algorithm: Lasso Features: Price, Province, Variety Parameters:normalize =True, alpha =0.00001	0.430

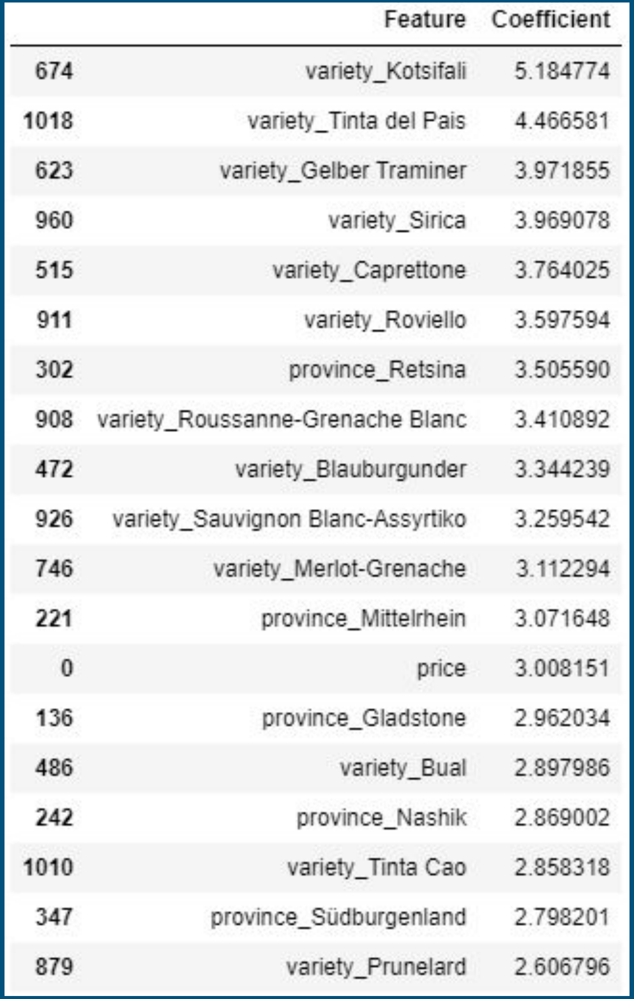
The last two models are pretty much the same.

RMSE Ridge: 2.270

RMSE Lasso: 2.265

Conclusion

- The best model is a Lasso regression with the price, the province where it came from, and the variety as its features, as well as the alpha set to 0.00001 and normalized.
- The province of Kotsifali has the greater coefficient in our model (therefore the the largest predictor)



	Feature	Coefficient
674	variety_Kotsifali	5.184774
1018	variety_Tinta del Pais	4.466581
623	variety_Gelber Traminer	3.971855
960	variety_Sirica	3.969078
515	variety_Caprettone	3.764025
911	variety_Roviello	3.597594
302	province_Retsina	3.505590
908	variety_Roussanne-Grenache Blanc	3.410892
472	variety_Blauburgunder	3.344239
926	variety_Sauvignon Blanc-Assyrtiko	3.259542
746	variety_Merlot-Grenache	3.112294
221	province_Mittelrhein	3.071648
0	price	3.008151
136	province_Gladstone	2.962034
486	variety_Bual	2.897986
242	province_Nashik	2.869002
1010	variety_Tinta Cao	2.858318
347	province_Südburgenland	2.798201
879	variety_Prunelard	2.606796

Recommendations

- Producers should focus on releasing higher price wine if they want to score better for a review.
- Wine producers should note that the top 3 wine varieties based on their coefficient values are: Kotsifali, Tinta del Pais, and Gelber Traminer.
- Wine from the provinces of Retsina, Mittelrhein, and Gladstone will score better .

Areas of Further Study

- Ingredients in wine that can predict score.
- The effect of reviews on wine sales.
- Further research into broad agricultural patterns