Final Project Report

Favorable Reception of Wine

STA5703

By

Carlos Cocuy, Karol Martinez, and Joseph Chakko

Research Question

What features are most important in determining the favorable reception of a wine? We are leveraging the wine dataset provided by Kaggle.

Overview of the Dataset

Our response variable is *Points*. It gives a numeric way for someone to convey how much they enjoyed the bottle of wine.

Variable	**Type**	**Description**	**Distinct Values**
Country	String	Country of Origin	43
Description	String	Text printed on bottle label by the manufacturer	119955
Designation	String	Vineyard within the winery	37979
Points (Response)	Number	The number of points WineEnthusiast rated the wine on a scale of 80-100	21
Price	Number	Retail price	390
Province	String	The province or state of origin	425
Region_1	String	The wine growing area in a province or state	1229
Region_2	String	The subregion within region_1 (if relevant)	17
Taster_Name	String	The name of the rater	19
Taster_Twitter_Handle	String	The twitter handle of the rater	15
Title	String	The name of the wine	118840
Variety	String	Type of grape	707
Winery	String	Manufacturer	16757

Most of the data is categorical. Encoding these fields yields thousands of predictors.

Data Cleaning

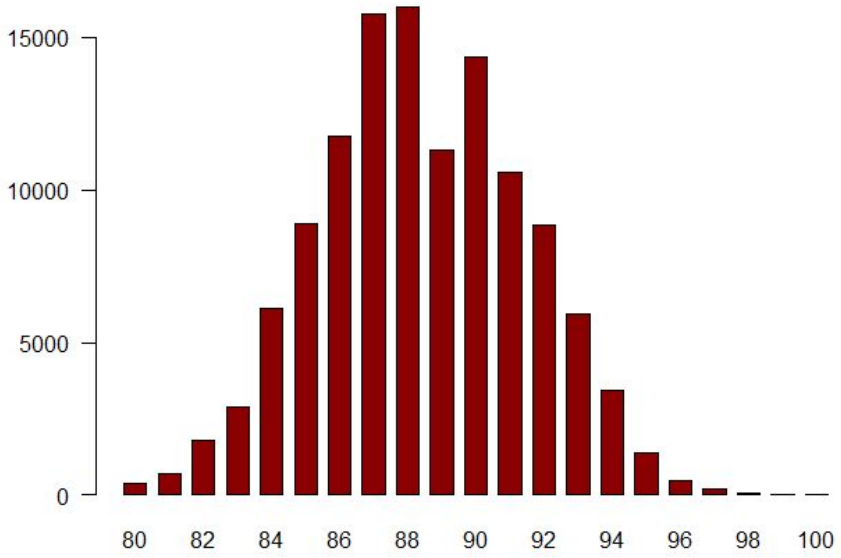
The Wine Dataset is in good condition but had some issues that needed to be addressed. First was extraneous columns. There are many columns of data that we could not find a way to utilize such as twitter handle and reviewer name. Other columns such as Region_2 did not provide much more information than already existing columns. We chose to omit these columns when training our models.

The dataset had preexisting “NA” values. This was troublesome in the column ‘price’ which was one of our most important predictors. There were around 6,000 entries where the

price was not available. We considered filling these values with an average number but ultimately decided to omit these entries.

Data Analysis

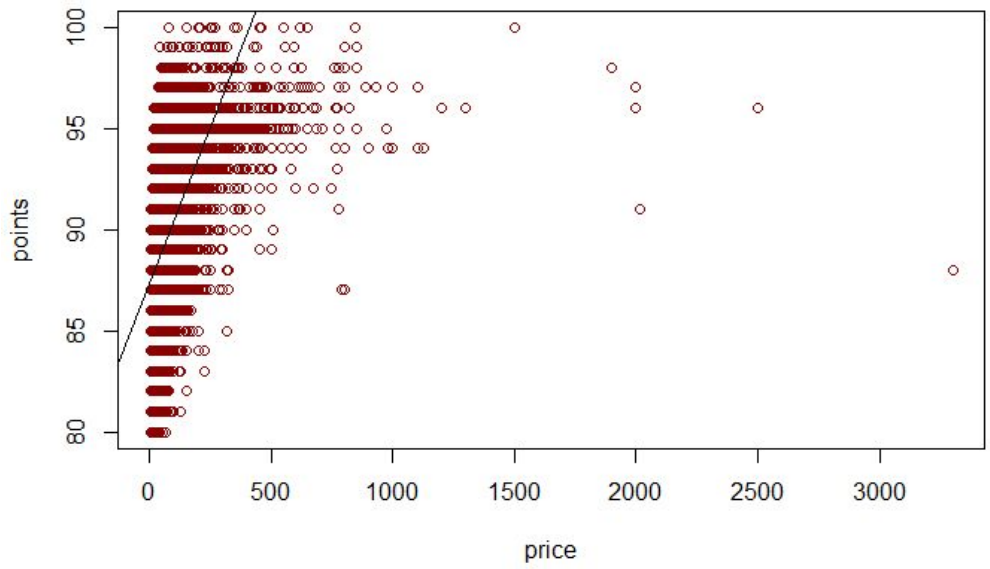
The first step in our analysis was to visualize the data by creating plots in R.



The distribution of *points* is about normal centered on 88 as seen above. With most of the values being in the 84 to 92 range with very few ratings at the top end.
The summary of *points* is:

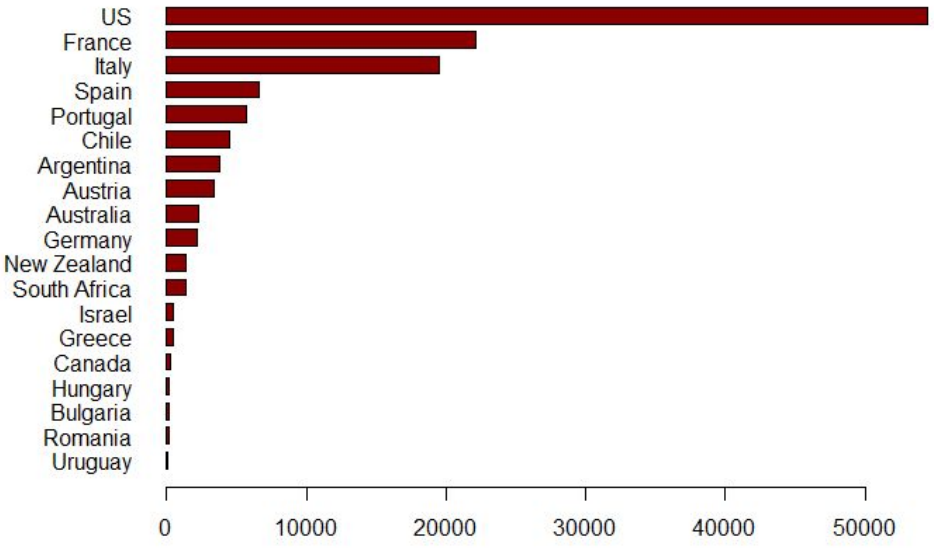
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
80.00	86.00	88.00	88.42	91.00	100.00

The first variable we look at, price, is the only quantitative variable in our possible list of predictors. We created a scatterplot using the response variable *points* and independent variable price. We then placed a trendline over the data to visualize if there is a correlation between them.



As the figure above shows there is a positive correlation between *price* and *points*, yet there seems to be many outliers such as the wine that cost over $3,000.

The remainder of our predictors are categorical. We created a bar chart that shows the frequency distribution of *Country*.



This revealed an issue with the dataset: the distribution of predictors is quite skewed. For example, the vast majority of wine in the dataset is from the United States. This causes the predictor to be not significant even though wines from the United States typically score higher

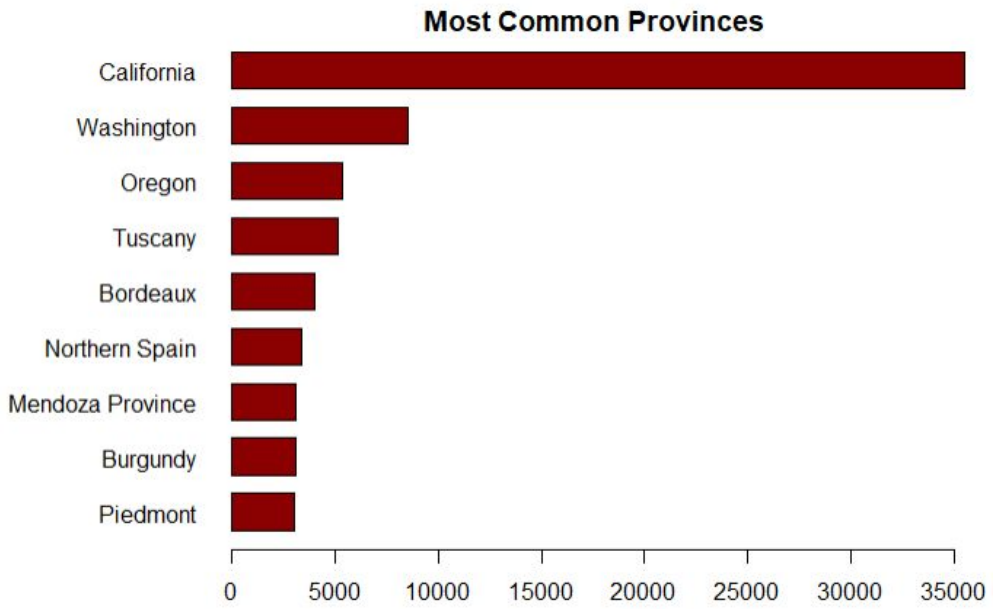
than wines from other countries. Luckily the data includes more specific province/region names which can be leveraged for countries that are highly represented in the dataset. Unfortunately, it does not have a way to group the countries that are underrepresented. We imported a new csv that includes Continent and Region for each *Country* allowing us to categorize the Countries into larger groups.

The second step in analysis is determining which predictors are statistically significant. This was mainly conducted through trial and error. This process proved to be tedious because of the large number of categorical variables. For example, there are 707 distinct values for the variety of wine. Categorical predictors are considered using dummy encoding. In some cases this results in thousands of predictors.

We had to create variables for each category value that seemed significant. These variables are binary containing a value of 1 if it belongs to that category and 0 otherwise. To not lose hours on creating a variable for each possibility, we fit the model with the variable that utilizes every possible category. Then only created dummy variables for those with p-values below 0.05. For other predictors we grouped the encoding together. For example, countries that do not have many entries in the dataset may be grouped by Continent or Region.

Similarly, the *Country* predictor needed to be adjusted to become useful. The majority of the wines are from the United States. Thus, it is not particularly useful information for the model. However the dataset also includes information about the province where the wine was bottled. There were many provinces that were statistically relevant and worth testing in our model.In this case, it is definitely important to examine the individual provinces.

California is one of the biggest producers of wine in the entire country. When dealing with wines that are from California we felt like it was worthwhile to utilize the region predictors.



Sonoma County, Napa Valley, and Russian River Valley were all statistically significant and used as their own predictor.

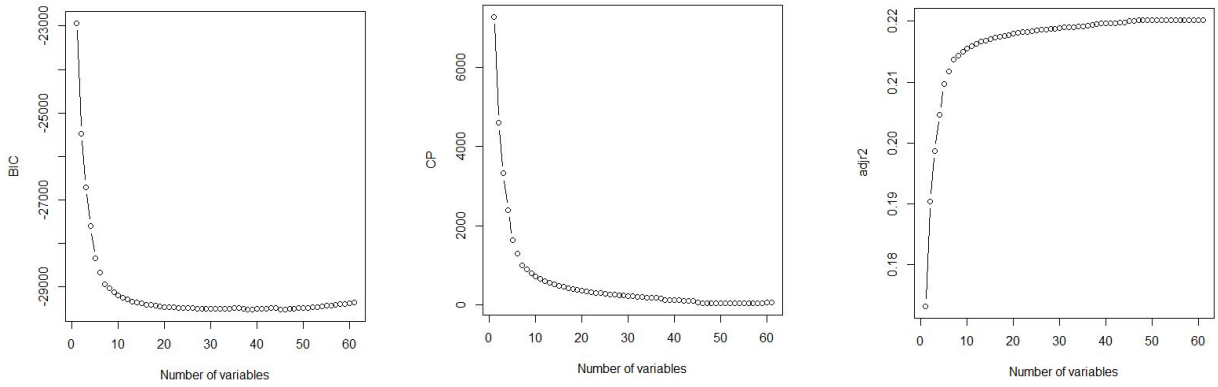
An important consideration when deciding which variables are important is collinearity between predictors. The categorical nature of our dataset made determining correlation somewhat cumbersome. We decided to use Point Biserial Correlation using price and points as the continuous variables.

Predictor	Price Correlation	Points Correlation
Latin America	0.09510046	-0.1664711
South Europe	-0.01264894	-0.03337917
Brazil	-0.005578498	-0.02437385
Mexico	-0.005037015	-0.02502222
Peru	-0.00485377	-0.01836684
Brachetto	-0.004770895	-0.01691418
Nevada	0.0003334726	-0.007883478
Iowa	-0.002786033	-0.01024618
Sonoma	0.02010974	0.01515963

After calculating the correlation between many predictors, we determined that there is no significant correlation between binary predictors and price or points.

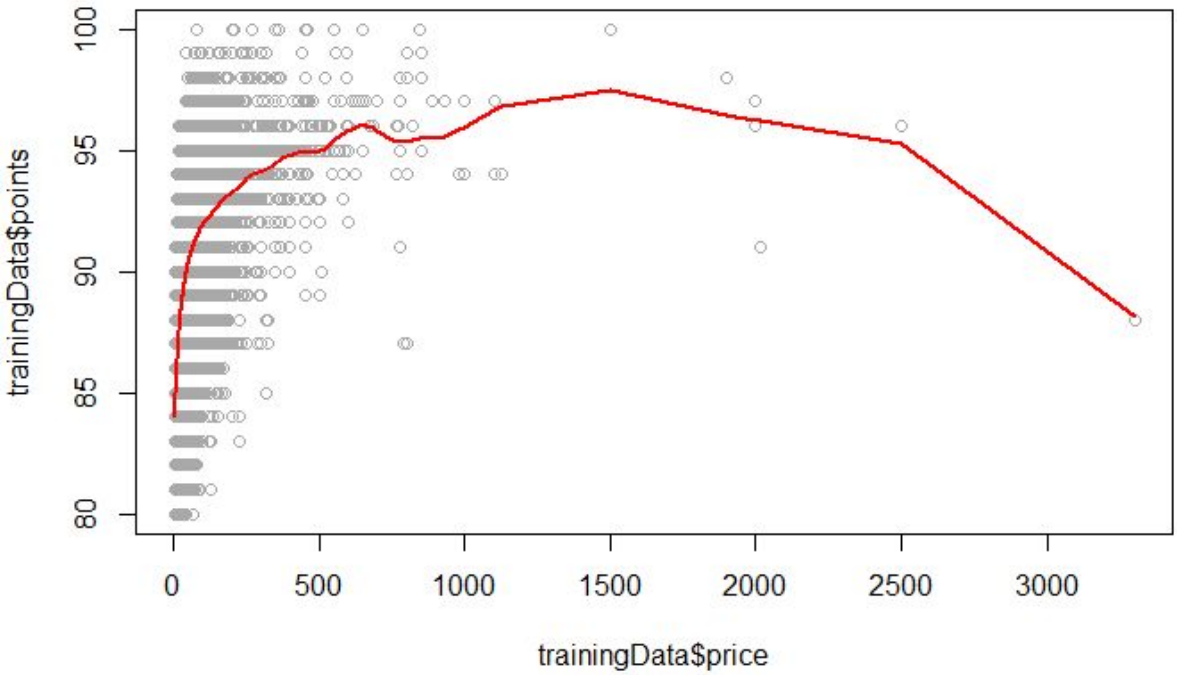
Parameter Selection

We used forward stepwise selection for the models that require us to manually provide a subset of parameters. The parameters were chosen using the R function regsubsets. The number of parameters were chosen by calculating the Bayesian Information Criteria, Adjusted R-Squared and Mallow's Cp.



Another issue we faced when selecting parameters was controlling the strength of the price variable. There is a clear relationship between price and the response variable points. Unfortunately, the relation proved to be too strong. As the price increased, the model would predict values well above the 100 point maximum. Our first reaction was to remove outlier prices. However, removing wines with prices above $2,000 did not improve the test MSE. There were not enough of these cases to warrant removing them.

Instead we decided to apply a stepwise function to price to divide it into 6 levels of cost. This greatly reduced the effect that high priced wines had on the value the model predicted. It stopped the model from making predictions over 100 points.



Model Fitting

The first model we implemented was Multiple Linear Regression. This is a simple model that allows us to utilize multiple predictors and output a quantitative value. The model was a good starting point to begin our testing and conduct further analysis. However the test MSE and R-Squared value were not competitive and we decided to implement other models.

Intuitively, the next step was to employ polynomial linear regression. This provided a better R-Squared score and accounted for more of the variation in the data. The test MSE also decreased.

One of the difficulties with fitting these models is selecting the parameters. Manually choosing subsets was tedious and difficult to determine when an optimal selection of predictors was selected. Using regsubsets to automate forward selection was much easier but it does not guarantee that the best subset of predictors was chosen. This challenge made us excited to implement Ridge Regression and the Lasso since they do not require us to specify a subset of predictors. These models use all predictors but modify their coefficients using the shrinkage parameter λ. The shrinkage parameter was chosen using the glmnet library and a grid.

The most complex models implemented were Smoothing Spline and GAM. Smoothing splines provide 18 degrees of freedom and more flexibility that we hoped would better fit our data. Finally we implemented a Generative Additive model that utilizes the smoothing spline. These models would perform poorly if they overfit the training data.

Results

Model	Adjusted R-Squared	Test MSE
Always Guess Mean	0	9.17
Multiple Linear Regression	0.2111	7.221275
Polynomial Regression	0.3171	6.350079
Smoothing Spline	0.3823715	5.71027
GAM	0.3894903	5.644454
Ridge Regression	0.2378359	7.091365
The Lasso	-7.358984e-07	9.218316

Discussion

The model that performed the best is the Generative Additive Model. It utilized the already successful smoothing splines and made predictions using additional predictors. Not only does it have the lowest Test MSE, it also has the highest adjusted R-Squared value. This result did not surprise us. The relationship between points and price is polynomial and price is one of the most prominent predictors in the dataset. The GAM provides more flexibility and better fits the data than other models. While it is difficult to interpret, it has a strong grasp of the data without overfitting.

We were expecting Ridge Regression to outperform all models since it can handle many predictors, even if they are not all related to the response. Additionally it would not require us to do any parameter selection. The results were decent but lead us to believe that many of the predictors are related. The Lasso's poor results imply the predictors are not sparse.

Despite their simplicity, both linear and polynomial regression performed well. Of course polynomial had a stronger performance since the relationship between price and points is non-linear. We believe that polynomial regression performed well enough that it can be used to help in the interpretation of the dataset.

Conclusion

After fitting multiple models on our dataset, we felt ready to answer our research question. The predictors that are most helpful in determining the reception of a wine is price and location. Both of these predictors had to be tuned to achieve good results. Price was contained through the use of a stepwise function. Location was changed to be more generic when a country was underrepresented and more specific when a country was overrepresented.

The best model was a Generative Additive Model that utilized a smoothing spline. The complexity gave it a large amount of flexibility and the ability to make accurate predictions about the dataset. Simpler models like polynomial regression also performed well and can be used to help interpret the dataset.

In the future, this research can be advanced by a dataset with more entries. Currently, each observation is distinct, meaning that each wine is only reviewed one time. It would be useful to have multiple reviewers assign scores to the same bottle of wine.