Determining Ratings from Wine Reviewers

By Isiah Fimbrez, Kumann Liu, and Joshua Vilela

I. Abstract

In our project, we used a wine dataset to see if we can determine the ratings of the different wines tasted by analyzing other factors of the dataset. The features that we used for the project included the country, description, the people that tasted the wine, type of variety, the winery it came from, and the price. We implemented a way to convert the features to usable data by implementing transformers from the ML Features package to preprocess the dataset and are able to be used in our models. The model we used was Linear Regression and other methods of looking at some of the features such as finding the highest rated wineries or the ones with the average highest prices of the wine. Our champion model would be the one that included price, variety,winery and country that it came from but since the results we got were not great we could not fully predict the ratings of the wine based on the information shown in the dataset.

II. Data and Methods

Background of Data

The original dataset was scraped from WineEnthusiast on June 15th, 2017. The data we obtained from Kaggle contains 129,975 rows and 13 columns describing wine details such as country origins, provinces, and the regions they stem from as well as their reviews from critics. Each review contains a numerical rating, on a scale from 1 to 100, and a description of the wine. The dataset was read as type ‘dataframe’, and most analysis is done with the ML package, with some use of RDDs for exploratory data analysis.

Data Cleaning

In order to have relevant data for our project, we removed certain columns that we decided would not help with determining the quality of the wine and what type of rating they would achieve. We removed columns that would not affect ratings such as the provinces, the regions, tasters’ twitter handles as well as the name of the wine which would help trim the data to only the useful parts of datasets.

We kept price and points as the first columns to keep as we wanted to evaluate if prices showed the quality of the wine and its rating. We also kept other columns in the dataset which we think could be good indicators of the wine having a good rating:

- **country:** original country of where the wine was created
- **description:** shows the characteristics of the wine

- **taster_name:** Name of the taster for that particular wine review
- **variety:** type of wine tasted
- **winery:** the winery in which the wine was created
- **price:** the price of a particular wine

We also saw that each of the columns had null values so we filtered each column to only include rows with non-empty strings so instead of 129,975 rows we instead have 129,879 rows.

Exploratory Data Analysis

After reading in the data, we examined the data to find out which wineries were the highest quality using price and rating points. We also analyzed the descriptions of the reviews to create a word cloud and find the most commonly used words in the dataset.

Highest Quality Wineries

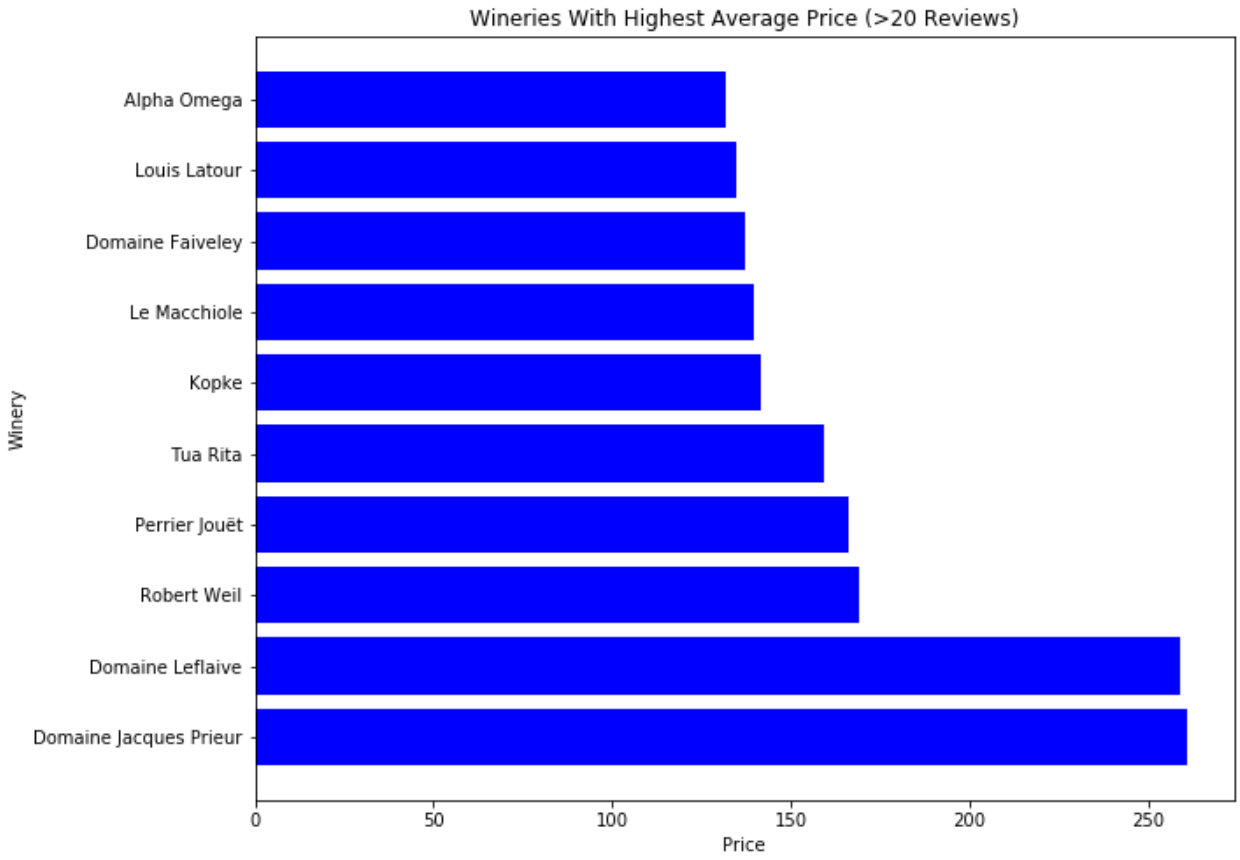
We were curious to see which wineries had the highest quality wines. We filtered the wineries to have a minimum of twenty reviews.

The first metric we used was the average rating point amount. The following table displays the ten wineries with the highest average rating points.

Winery	Average Points
Domaine Leflaive	94.35
Franz Hirtzberger	94.10
Paul Hobbs	94.04
Cayuse	93.89
Château Margaux	93.64
Château Léoville Las Cases	93.57
Shea	93.29
Louis Roederer	93.27
Massolino	93.26

Ramey	93.13

The second metric we used was the average price of the wines. Once again, we only included wineries with at least twenty reviews. The following graph displays the wineries with the ten highest average prices.

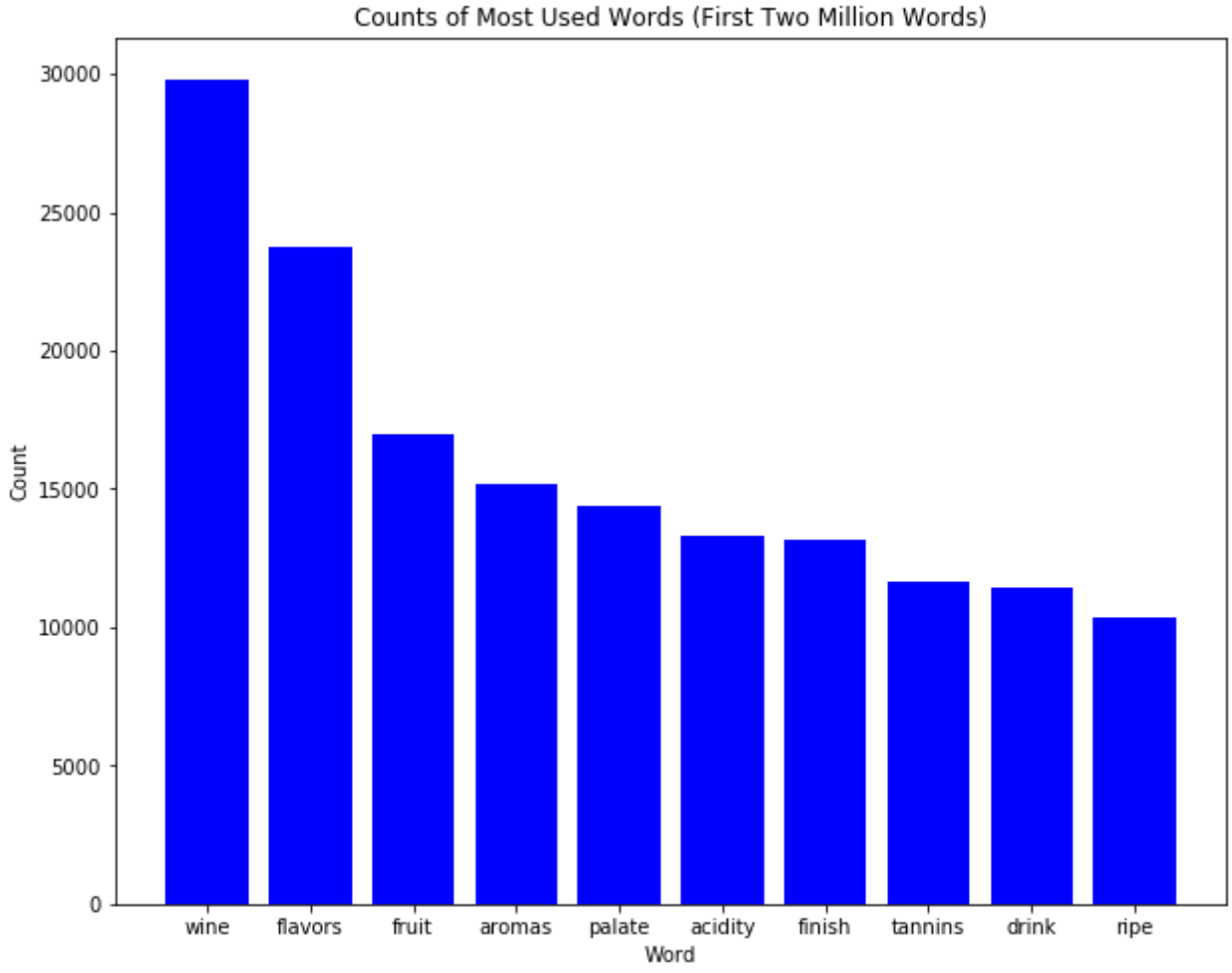


Domaine Leflaive appears on both the highest average points and highest average price lists. This suggests that this winery has wine quality that matches its high average price at over two hundred and fifty dollars.

Description Analysis

We examined the descriptions of the wine reviews to determine the most commonly used words in the reviews. We first used the wordcloud package to create a word cloud of the most common words.

We also conducted our own analysis of the descriptions to find the most common words. Due to the size of the RDD after splitting the descriptions, we only used the first two million words of the descriptions. Ignoring common words such as ‘and’ and ‘the’, the most common words can be seen below.



Data Preprocessing

In order to properly utilize our data, we first had to convert the string-valued variables into a form we could use to perform linear regression in predicting the score of each wine. The only numerical column which we would use as a predictor variable was the price column. The other predictor variables were categorical-based variables, leaving us to find a work around through encoding the **country**, **variety**, and **winery** variables. This was accomplished through the use of the StringIndexer() function. Keeping in mind the enormous amount of data points we have present, we included a short list of 20 rows of translated encoded values below for a reference.

variety	winery	country	points	country_index	winery_index	variety_index
White Blend	Nicosia	Italy	87	2.0	3227.0	15.0
Portuguese Red	Quinta dos Avidagos	Portugal	87	4.0	4476.0	14.0
Pinot Gris	Rainstorm	US	87	0.0	2454.0	19.0
Riesling	St. Julian	US	87	0.0	3655.0	5.0
Pinot Noir	Sweet Cheeks	US	87	0.0	1813.0	0.0
Tempranillo-Merlot	Tandem	Spain	87	3.0	2112.0	335.0
Frappato	Terre di Giurfo	Italy	87	2.0	3669.0	142.0
Gewürztraminer	Trimbach	France	87	1.0	975.0	27.0
Gewürztraminer	Heinz Eifel	Germany	87	9.0	1634.0	27.0
Pinot Gris	Jean-Baptiste Adam	France	87	1.0	156.0	19.0
Cabernet Sauvignon	Kirkland Signature	US	87	0.0	55.0	2.0
Gewürztraminer	Leon Beyer	France	87	1.0	1277.0	27.0
Cabernet Sauvignon	Louis M. Martini	US	87	0.0	657.0	2.0
Nerello Mascalese	Masseria Setteporte	Italy	87	2.0	9278.0	84.0
Chardonnay	Mirassou	US	87	0.0	661.0	1.0
Riesling	Richard Böcking	Germany	87	9.0	3631.0	5.0
Malbec	Felix Lavaque	Argentina	87	6.0	7405.0	13.0
Malbec	Gaucho Andino	Argentina	87	6.0	8976.0	13.0
Tempranillo Blend	Pradorey	Spain	87	3.0	1293.0	38.0
Meritage	Quiévremont	US	87	0.0	9487.0	52.0

As you can see, White Blend variety has an index of 15.0, the Nicosia winery has an encoded index of 3227.0 and Italy has an index of 2.0.

Following the encoding of string-valued columns, we had to remove the original columns as they no longer served a purpose in our analysis. We then began to remove every column with a null value and replace it with the mean value of that particular column in order to maintain a sizable number of data points which would still provide meaningful contributions to our analysis.

Afterwards, we had to typecast the **price** variable as a *double* in order to maintain consistency with the format of the encoded variables.

We then applied the VectorAssembler() and transform() functions to create an assembler which would create a **features** column, which encompasses every index value of a particular row inside an array, in order to prepare the predictor variables for scaling and its use in linear regression.

Following, we utilized the StandardScaler(), fit(), and transform() functions to create a scaler which would be enacted on the new **features** column, to then compute the mean and standard deviation for each array (in each row), then append the new column labeled **features_scaled**, including these new metrics in another smaller array.

III. Results

Models and Results

We decided to split our data into 70% for training data and 30% for test data. Our problem is focused on predicting the resulting rating for a wine dependent on factors contributing to its conception (the use of **country**, **variety**, **price**, **winery** as predictors). Our models consisted of using multiple regression fits with varied combinations of these variables as predictors.

Benchmark Model

Our benchmark model consisted of using **Ratings** as the response variable and all of the following as predictor variables: **country_index**, **variety_index**, **price**, **winery_index**. We utilized the LinearRegression() function with maxIterations = 10, regularParameters = 0.3, and elasticNetParameters = 0.8, along with the fit() and transform() functions to create a linear regression model involving the mentioned variables. Regression Evaluator was then used to assess the Mean-Squared Error and R-Squared value of the model. The resulting values are 7.570 and 2.751 respectively.

Comparing Models

We created three models using similar methods as described above, with the main differences being which predictor variables were used. For our first model, we used only the **price** of the wine as a predictor variable. The second model consisted of the

string-valued variables **country**, **variety**, and **winery** while the third models had all of the variables as predictors.

To assess accuracy, we utilized RegressionEvaluator() to find the mean-squared error (MSE) and R-squared values of each model. MSE is a measurement which finds the square distance between the estimated values and actual values based on a linear regression model. A lower value typically indicates a model closer to predicting the actual, underlying values. R-squared values, in regression analysis, indicate how close data points are to the fitted regression line. In general, the higher the R-Squared value, the better the model. Hence, the model with the best tradeoff between having a high R-squared value and a low MSE would serve as the best model, which was the third model (**Country**, **Variety**, **Winery**, **Prices** as predictor variables).

The second model has the highest R-squared value but also has a significantly higher MSE in comparison to the other models. The first and third model have relatively similar values for the MSE and R-squared, but the tradeoff between MSE and R-Squared for the third model is better than that of the first model, hence why we chose the third as our benchmark model.

For a further breakdown of the results from the models, view the table below:

Model (list of predictors)	**R-Squared Value**	**MSE**
1. Prices	0.176	7.708
2. Country, Variety, Winery	0.021	8.996
3. Country, Variety, Winery, Prices	0.189	7.570

IV. Conclusion

The model that included country, variety, winery, and prices was overall our champion model as it provided the lowest MSE as well as the highest R-Squared Value. We could see that price as the sole predictor is only slightly more than when adding other features like country,variety, and winery which implies that adding these features to evaluate the data does not help that much compared to just price alone when trying to determine

ratings of the wine. This could be because the other features did not influence the ratings of the wine as much as the price indicates for it. Since the MSE and R-Squared value are still high on all models, we can conclude that we cannot predict the ratings of the wines from the data provided in the dataset. This could be due to the tasters' bias on wines which makes it harder to determine ratings as they are a variable that cannot be controlled for. For future tests, we could also change the response variable to one of the other columns like variety or the type of winery they could be in. We could have also done logistic regression and check for any misclassifications on the confusion matrix and determine if that would have been a better model.