**CSE 158 Assignment 2 report**

**Abstract**

Wine Enthusiast Companies, self described as “the world’s number one source for wine accessories, storage, information, education, events and travel, and is a driving force in the marketplace”. With a dataset consist of over ten thousands reviews written by wine connoisseurs, it should be possible to use a well known data-mining technique to predict price of wine

**I. DataSet**

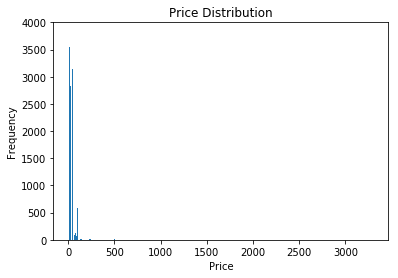
The dataset is in Json format with about 130k reviews with 13 different fields. These 13 fields with their description of each field is stated below:

|  |  |
| --- | --- |
| Points | In scale of 1 - 100 |
| Title | Name of the wine |
| Variety | Grapes used |
| Description | Describing taste, smell, etc. |
| Country | Country the wine is from |
| Province | Province or state the wine is from |
| Region 1 | Wine growing area |
| Region 2 | More specific region |
| Winery | Winery that made that wine |
| Designation | Vineyard within winery |
| Price | Cost of a bottle |
| Taster Name | Name of the reviewer |
| Taster Twitter Handler | Twitter handle of reviewer |

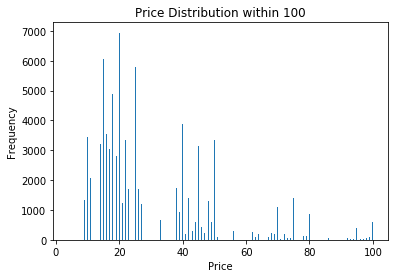
The total size of the dataset is 129971. After scanning the dataset, we found that some of the 13 fields had no entries. Therefore, we discarded those reveiw and obtained a dataset size of 120975 reviews, which is sufficient enough for the purpose of this project.

**IA. Statistics**

First, lets look at the price distribution of the dataset.



The most expansive wine costs 3,300; however, most of the wines cost within 100 dollars. Therefore, the histogram does not look well representable. Then, we retrieve the data that the price is within 100 dollars. The size is 117609, which means around 3000 wines, just a small portion of data, costs more than 100 dollars. We plot the price distribution histogram again.



This histogram shows more interesting result for our predictive task. Therefore, we going to resize our data to only contain price range from 1 - 100. The size of data with price range from 1-100 is 117609.

The number of country in the data is 43. The country that have most number of wines is US, which has 53221, and the next two countries are Italy and France, have 16238 and 16774. Below is a histogram of the top 100 countries with the most frequency in the data set.

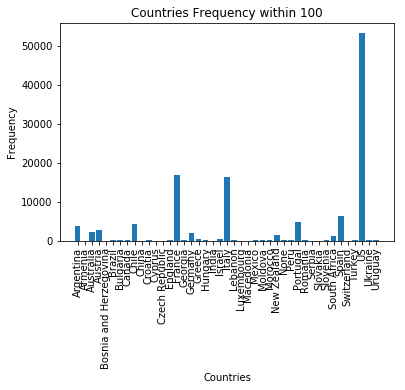
****

Figure a, wines’ frequency in each country

Figure a shows that only the top 3 countries contain the most reviews in the data. From the histogram, we can determine that countries field would not be a good field for predictive task.

**IB. Finding Relationship**

**a. Price vs Year**

From figure b, we see that most wines are from 1990 to 2017. The price range in each year are basically the same (0-100). Therefore, we think that there does not exist any obvious relationship between years and prices.

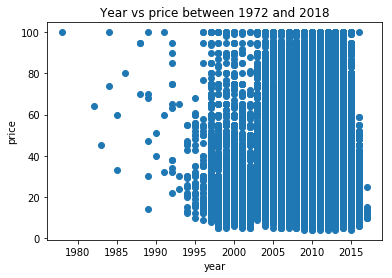
****

Figure b, years vs prices between 1972 and 2018

**b. Winery vs Average Price**

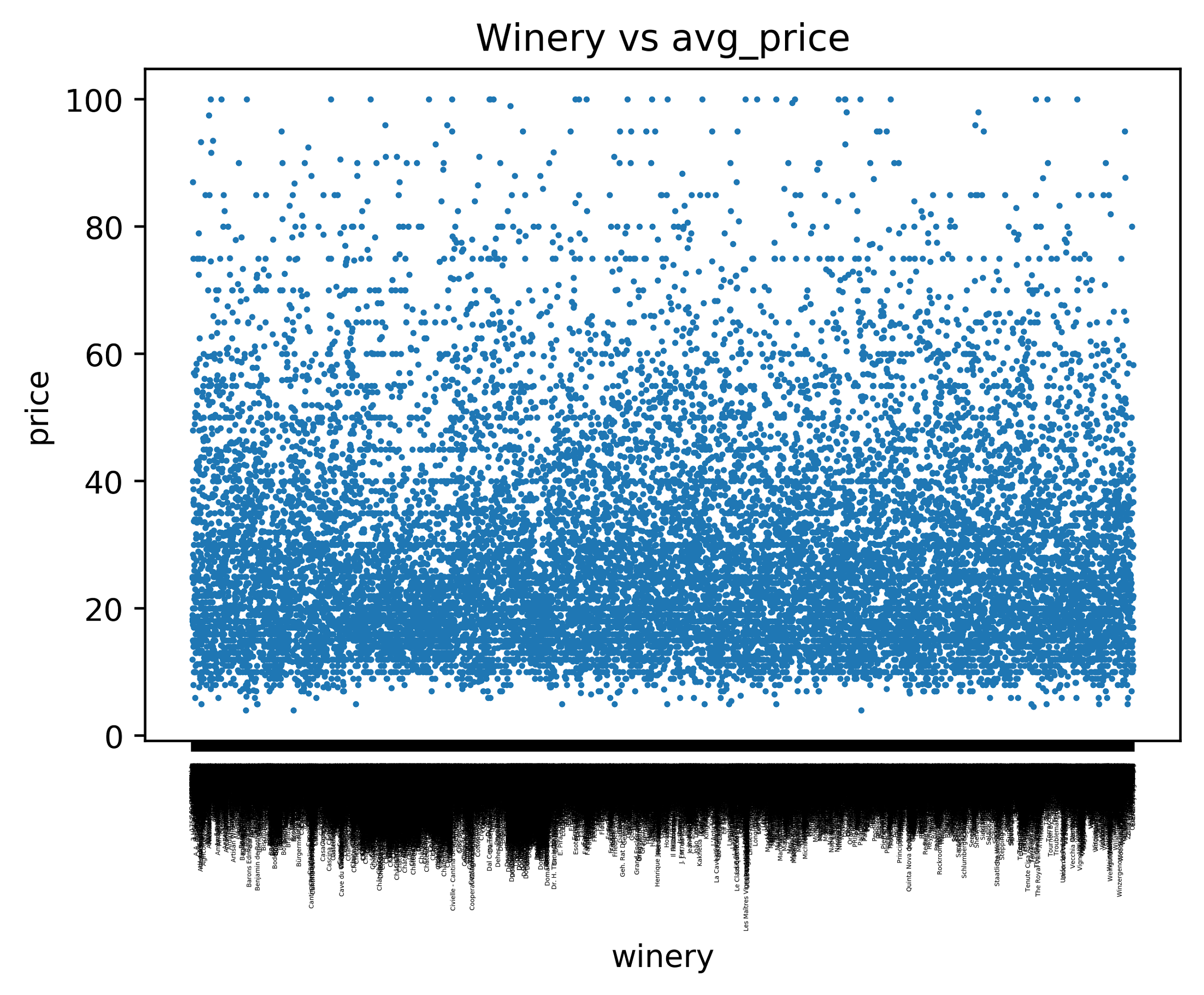
****

Figure c, Winery vs average price in each winery

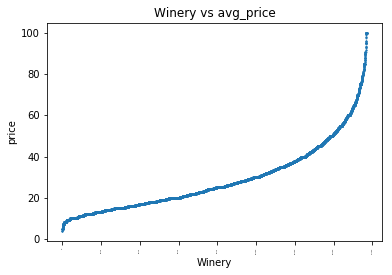
****

Figure d, Winery vs average price in each winery sorted by average price

By plotting the graphs above, winery vs average price, we can see that there is a relationship between winery and price. From figure c, it is harder to see the relationship between these two field, since there are many different winery which makes the graph to be harder to read. After organizing the data with increasing price of wine from different winery, we obtained figure d. From this graph, we can determine that winery has some relationship with price, thus we can use winery as a feature to predict price.

**c. Variety vs Price**

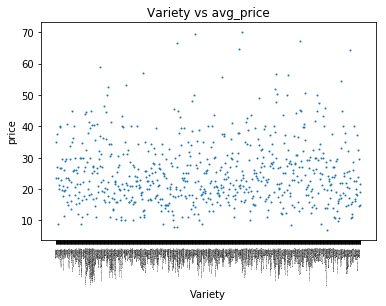
****

Figure e, Variety vs average price in each winery

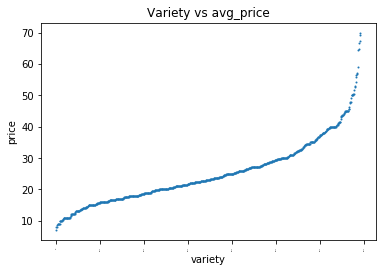
****

Figure f, Variety vs average price in each variety sorted by average price

After finding that there exist a relationship between winery and price, we further discover the relationship between variety or what type of grape was used in the wine, and price. Similar to the initial finding of winery and price, figure e also show a similar result from figure c. However, figure e shows that each data point is much more scatter from each other which indicate that there are less variety in the dataset. We also tried the same method on the figure f, which plot variety vs average price in increasing order. As we can see from figure f, we also obtained a similar relationship as figure d in the variety vs average price graph.

**II. Predictive Task**

In the project, we will try to predict the price of the wine given other field of a review.

Since the variety and winery vs price are on a line after sorting, clustering may not be a good idea, so we will use linear regression as our model.



Where X is the matrix of features we select and is the corresponding coefficient for each feature. y would be predicted label(price) of a given data. The baseline model that we use to be a reference of our accuracy is by using a linear regression model which used “points” as the only feature in the Matrix.



The primary objective for using linear regression model is to determine some good features that would be relevant for our predictive task.

Since we are going to predict the price, it is hard to use accuracy to evaluate the performance. Therefore, to evaluate our model and the validity of our prediction, we would try to minimize the Mean Squared Error (MSE) with respect to the baseline that we defined above. Moreover, we would defined MSE by using the following formula:



Other than regular linear regression model that we learned in class. We also try to use random forest regression to be one of our model. This model is based on the decision tree method but it will use multiple trees as part of the vote to determine the predictive result. First, the algorithm will first pick m features randomly from n features of the training matrix. Then, use the m features to build rules for the tree until there are p leafs. Repeat the previous steps until q trees are built. The algorithm will use those trees to predict the result from given testing data. Since each tree may get different result, the final prediction of each data point will be based on voting scheme. Although the predicting result is continuous value, the algorithm will determine the final result by taking the mean of all trees.

Features that we use for these different linear regression model are mainly variety and winery average price. However, we would have to optimal the predictive task by incorporating different feature to the model such as year, variety price stand deviation, point average price and country average price. The methodology used to obtained the value as follow:

1. Split the data into half, 1st half is the training set and 2nd half is the testing set.
2. Find the average price of each winery, variety, point and country
3. Find the stand deviation of price for each variety
4. Extract the year from the title, since the title contain the name and year of the wine.
5. From each data in the training set, transform each data as a feature factor by adding the average price of the winery and variety of each wine.
6. Use the training data to train the linear regression model and predict the result on the testing data
7. Find the MSE of the result and compare the performance of each linear model



**III. Model**

Using point as a feature to be our baseline, while using winery average price and variety average price as a feature for our model. In order to optimize the predictor, we will be trying to use variety and winery to predict the price of a wine. We used variety and winery to be our features because from the data section of our report, we showed from figure b to e, that there are some kind of correlations between price and variety and winery. Therefore, our model would try to utilize that information we found and elaborate it on our predictive task. Other than using average price of the winery and variety to be a feature vector, we have also tried to use binary to represent winery and variety in the feature vector. However, by doing so we would run into the problem of scalability, since there are 693 variety and more than ten thousand winery. The resulting vector would be extremely large

Other than these linear regression models that we build, we have also consider using SVM by using the most 500 common words as a feature vector for comparison. However, not everything went smoothly as it should. Before we analyze the data that we gathered, we think that the wine that produced earlier should cost more than those produced later. Thus, we use the tile field, which contain the name of the wine and the year it produce, to get the year of the wine produced and use it as a feature for our linear regression model. However, the resulting mse is extremely high and from figure b, we notice that there may not be an obvious correlation between year and price.

To optimize our predictive task, we tried to incorporate more fields that seem to be useful and correlated with price. We have created two different linear regression model that use RandomForestClassifier and Ridge along with 4 different datafield to predict the price of wine. The strengths of these linear regression predator is that is easy to train and less time consuming compare to other models. However linear regression model would only be efficient for linear relationship and the data must be independent.

**IV. Literature**

Wine Enthusiast Companies data set was used to by varies people to study the prediction of the quality of the wine, the wine type and other different predictive task. Moreover, it was studied to see correlations between different field of the wine and the quality of the wine. The data set that we are using for this predictive task is an existing data from Wine Enthusiast Companies and we downloaded it from kaggle in the link: <https://www.kaggle.com/zynicide/wine-reviews>. This data set was used for a recommendation system which introduce wine to a user who may want to try based on the review of the wine. These data was also used for as guide that help beginner to search for different wine and check what kind of food goes well with that wine.

There are many dataset that are similar to the one that we are using for our predictive task. One data set that we found and have been studied in the past is the wine review from CellarTracker. It was studied in various way, and many of the report written for this dataset are using linear regression to predict rating based on other information obtained from the review.

In the Economic Journal, Orley Ashenfelter the author of *Predicting the Quality and Prices of Bordeaux Wine*, have also used linear regression model to predict the quality and the price of a wine. He created this predictive task with many different aspect and data field to perform the predictive task. Instead of only using variety, he also used the weather statistic of different season to predict the quality and price. He used a much more complete model with linear regression because he can obtain data that is more relevant. The conclusion from existing work are somewhat similar to our finding. Many of the existing work have suggested that vintage of a wine would affect the quality and price. In our finding, we have also show that variety, which is our field for vintage, is correlated with the price of a wine. Therefore, our finding does have some interaction with other existing work on similar data set.

**V. Results and Conclusion**

The result of our models are shown below.

Model Features MSE

|  |  |  |
| --- | --- | --- |
| Linear regression | points | 259 |
| Linear regression | Winery avg price | 154 |
| Linear regression | variety avg price, winery avg price | 146.73 |
| Random Forest Regressor | Year, variety average price, variety price sd, point average price, winery average price,  Country average price | 108.28 |
| Linear regression | Year, variety average price, variety price sd, point average price, winery average price,  Country average price | 126.99 |

From this results, we can confirm our finding in the beginning of our report. We found that variety and winery do affect the price of the wine. By comparing the MSE, we can also conclude that the taste or the point of the wine does not have a clear correlation with price of the wine, since the MSE only improved slightly when incorporated with 4 different fields to the feature for the linear model.

Compare to the baseline model that we build, our linear regression models have a much better performance since the MSE reduced significantly. From the MSE that we get from the model and the result that we obtained, we can conclude that winery average price and variety average price have the most correlation with price. In the feature vector, variety and winery worked the best on prediction. Moreover, the 4 different field have less significances as a feature since the MSE for those model only decrease slightly compared to the model that only used variety and winery as a feature. For the parameters of our model,

N\_estimators = 200, which is the number of trees to build, max\_depth = 30, the maximum depth of each tree, min\_samples\_leaf = 5, min number of samples required at each leaf node and N\_jobs = 100, the number of jobs to run parallel for both fit and predict.

Many of our proposed model succeed because the feature that we used for linear regression does contain a linear relationship with the price that we are trying to predict. Since many of the data fields have a linear relationship with price, our models would be able to predict the price more accurately based on the concept that we learned in class. However, some of the features failed because they do not have a linear relationship to price. Without a linear correlation, it would be impossible for linear regression model to obtain a good result on the data set.

**References**

[1] Polamuri [Saimadhu](http://dataaspirant.com/author/saimadhu/). HOW THE RANDOM FOREST ALGORITHM WORKS IN MACHINE LEARNING [<http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/>].

[2] Scikit learn. RandomForestRegressor [<http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>].

[3] Ashenfelter Orley. Predicting the Quality and Prices of Bordeaux Wine [https://www.jstor.org/stable/20108831?seq=8#page\_scan\_tab\_contents].

In order to prevent overfitting, we would also introduce a regularization parameter , to penalize an complex models. After inserting this new parameter to our equation, out new predictor would be:

