Wine Score Milestone Report

Jorge Londono

Problem Statement

Wine production is a global multi-billion dollar industry that has a history spanning over 8,000 years. 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?

The benefactors of the results of this project can be several types of people. The first one is the wine producer. By seeing what type of variables result in better outcomes, the producer is able to decide how they want to run their business. Should they focus on making a specific variety of wine? Should they build a new winery in a specific province?

Next there is the consumer. This project can help bring insight into the variables that can answer, what makes wine great? This way they can make more informed decisions on what to buy, how much to value a specific wine, and if said wine is worth the price.

Lastly there is the wine industry media. If certain variables are found to show better scores with reviewers, this can inform the media of what types of new wine projects they need to spotlight and pay attention to. Should reviewers have certains standards when trying a bottle of wine, if they can estimate where that wine should rank based on the features of this project? Also, can this predictive tool highlight biases in the wine industry media?.

Data

The data that will be used for this project can be found in: https://www.kaggle.com/zynicide/wine-reviews . 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.

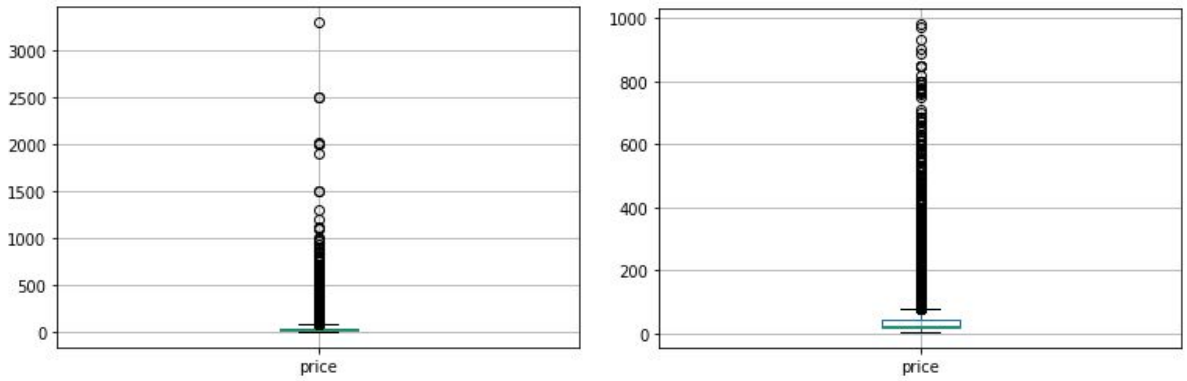
Cleaning and Wrangling

SInce the data frame obtained for this project came from a Kaggle submission, there was no substantial data wrangling or cleaning required. However the data was analyzed for certain conditions such as outliers and missing values for the numerical variables.

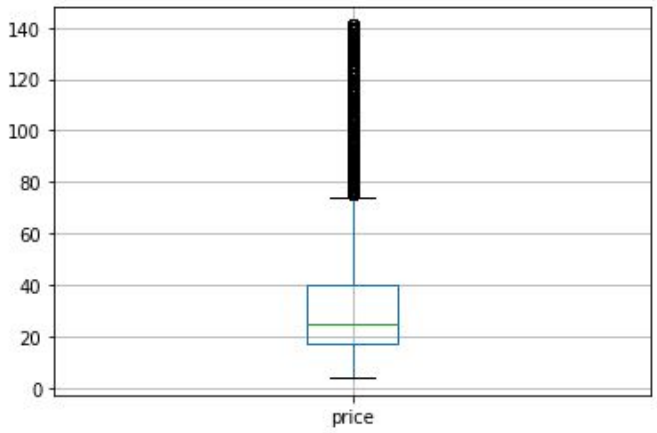
In order to check for outliers and missing values I first took a glimpse at the data using the ‘info’ and ‘describe’ methods. This showed that there were around 9,000 observations where the price variable was null. Additionally it showed that while the median price of the bottle of wine was only $25, the max value in that column was $3,300, a possible outlier.

As stated previously, around 9,000 observations showed the price as ‘NaN’. Therefore in order to have an accurate predictive model, the observations with these null values were removed using the ‘dropna’ method on the data frame. To confirm that all of the null values were removed, the ‘tail’ method was used with the price column sorted using the ‘sort_values’ method. This showed that all of the null values for price were not in the dataframe anymore.

Outliers proved to be slightly more difficult to remove accurately. Creating a boxplot showed that there were a few observations that had a price higher than $1,000. Therefore, using the ‘drop’ method for the data frame, all of the observations with wine over $1,000 were removed. The total observations removed was 16.



However, this was not enough to correctly remove the skewness of the boxplot. Therefore, I performed a Z-test using the scipy method ‘zscore’ to find and remove all of the observations that were above a Z-score of 3. This left me with the boxplot below, which looks like a better set of data to use for this project.



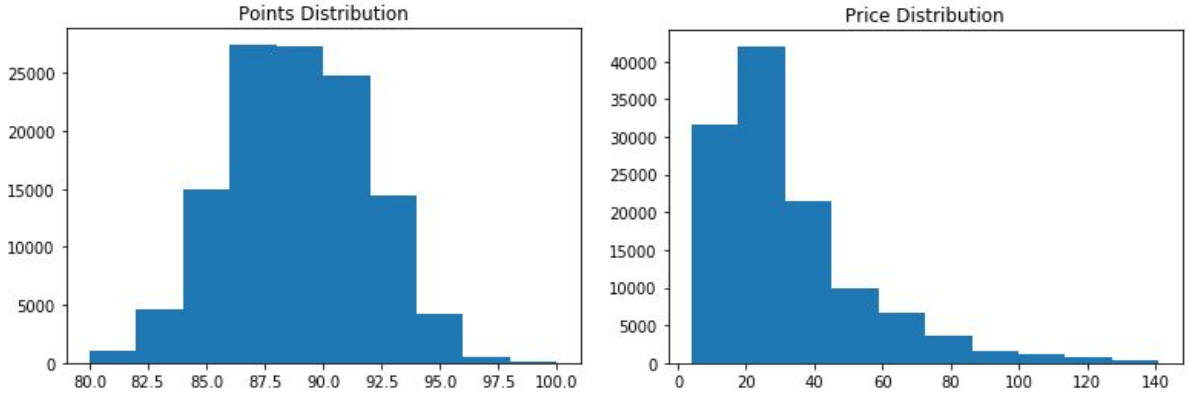
Exploratory Data Analysis

Descriptive statistics for this data set can be found below.

	Unnamed: 0	points	price
count	119390.000000	119390.000000	119390.000000
mean	65047.159151	88.354259	32.417958
std	37505.237588	2.991389	21.957717
min	1.000000	80.000000	4.000000
25%	32584.250000	86.000000	17.000000
50%	65138.500000	88.000000	25.000000
75%	97496.750000	90.000000	40.000000
max	129970.000000	100.000000	141.000000

While the mean and median values of the 'points' column are very similar, at 88.35 and 88.0 points respectively, the mean and median values of the 'price' column seem to show a larger difference, at 32.42 and 25.0 respectively, which could mean that our right skew could still be significant.Additionally, the minimum and max values for the price column show a greater spread of values than that of the points column.

With this in mind, it would be a good idea to visualize what this data looks like to further understand how this project should proceed.



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.

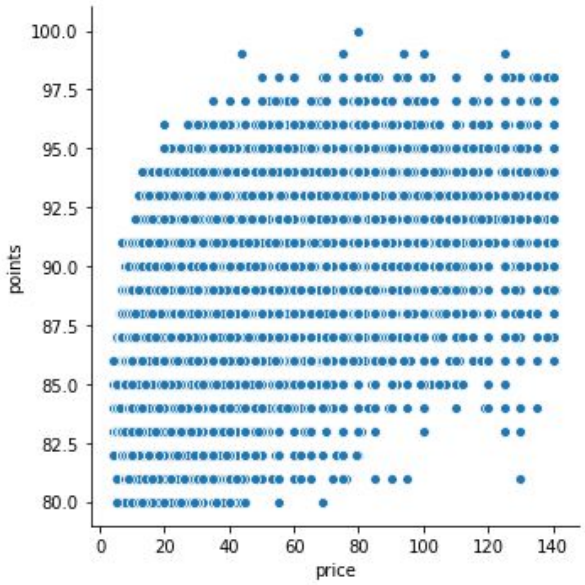
Statistical Analysis

While this project will include variables such as: Country of origin, province of origin, wine variety, and winery, in its model, the main feature appears to be the price of the wine. This statistical analysis will aim to uncover if there is a statistical significance in the difference between the least expensive wines and the most expensive wines when it comes to review score.

The first step is to create subgroups based on price in order to compare their mean score. Creating a bar plot will help us visualize the difference in average score between these subgroups. There seems to be a difference in the mean score from the lowest priced group to the highest priced group.

There seems to be a difference in the mean score from the lowest priced group to the highest priced group.

Next we can try to find a correlation between price and review score.

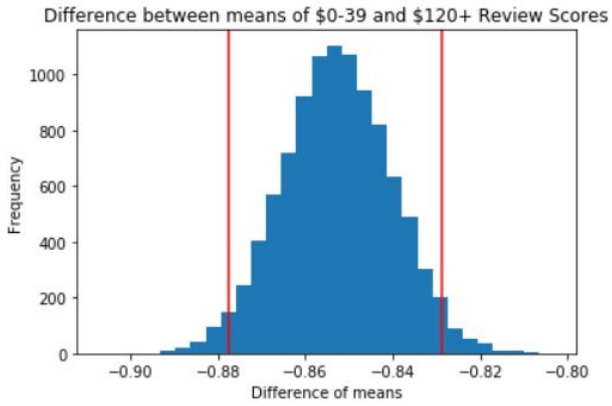


Creating a plot chart helps us visualise the existence of a relationship between 2 different variables. In this case, while we can see that there appears to be a positive correlation, it is not clear how strong it is.

Using the numpy method ‘corrcoef’ we can find out what the correlation between these variables is. At 0.55 there appears to be a moderate positive correlation between the score received and the price of the wine.

While a good hypothesis would be that the price of the wine affects how highly it is rated, a bootstrapping test will show us if the data is not giving us these numbers by chance.

At this point we have identified 2 different things. Our data shows that there is a positive correlation between how well a wine scores and how much the wine is priced at. Additionally, that the mean score of wine under $40 is lower than the mean score of wine $120 and over. A bootstrapping hypothesis test will help us replicate the data 10000 times to see if the difference between group 1 and group 4 exists not purely by chance. By using a Bootstrapping method we will test the following null hypothesis: with 95% confidence we can state that wine under priced $40 has the same mean score as wine priced $120 and over. The alternative hypothesis would be that the mean score of wine priced under $40 is not the same as the mean score of wine priced at $120 and over.



A two tailed test with a 95% confidence interval shows that the difference between the means of the two groups is between -0.88 and -0.82.

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.