Adam Foley

Topics in Management: Data Analytics Boot Camp

Final Project

January 19, 2018

I. Dataset

The data set I chose to analyze was a collection of over 150,000 wine reviews. The wines came from a variety of countries including: United States, Spain, Italy, France, New Zealand, etc. The data consisted of 8 variables from each wine review, these included country, province, region, variety, winery, points for the wine review (ranging 1-100), price of the wine ($USD). My motivation for selecting this dataset can be derived from my lack of knowledge in the wine market. The goal of this project was to understand whether variety, country, and or price can be a good determinant of a wines rating.

II. Questions

-Is there any indication that price is the best determinant of the review a certain wine will receive?

-Does country and variety have a correlation with the points/price multiple on each wine?

III. Data Preparation

Fortunately, the data set I analyzed did not require much preparation in order to perform my analysis. However, I did manipulate some aspects which I have outlined below:

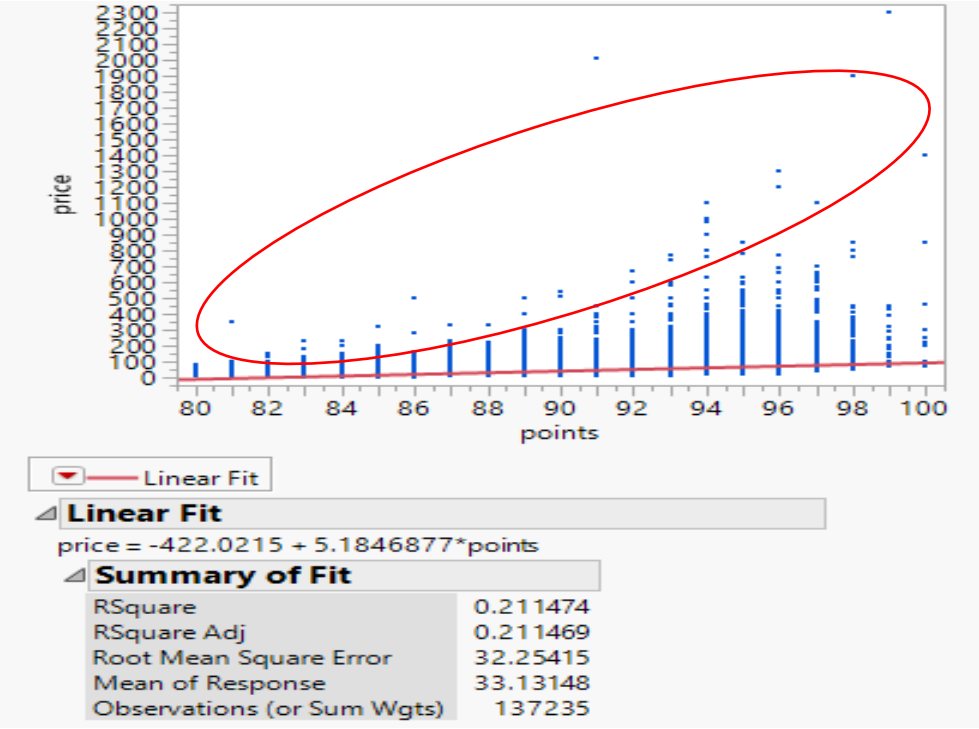
1) **Deletion of description of the wine** – I chose to remove this variable as my analysis would be derived from the country, variety, price, and points.
2) **Created Points/Price Multiple –** This variable was the center of my analysis; I used this to assess whether a premium was given to wines from specific countries or varieties.
3) **Generated three random samples –** Due to the nature of my data set, including over 150,000 review, I chose to generate three random samples in order to perform analysis.

IV. Analysis

Is there any indication that price is a way to determine the review a certain wine will receive?

The short answer is no. Using basic linear regression, I looked at the relationship between price and points of each wine. The expectation of this regression was that as price rose, the amount of points that it would receive would increase, as price in many cases is a good way to determine the quality of a product. My expectation of the data can be seen by the red oval. My reasoning being I would not expect low price wines ranging from $100-200 to receive scores in the 90-point range, as other wines that have prices in the range of $500-600. (See Figure 1) However, this was not the case when I ran the regression. Shown in the summary of fit, the Adjusted Rsqaure value was only .21 which does not represent a strong correlation.

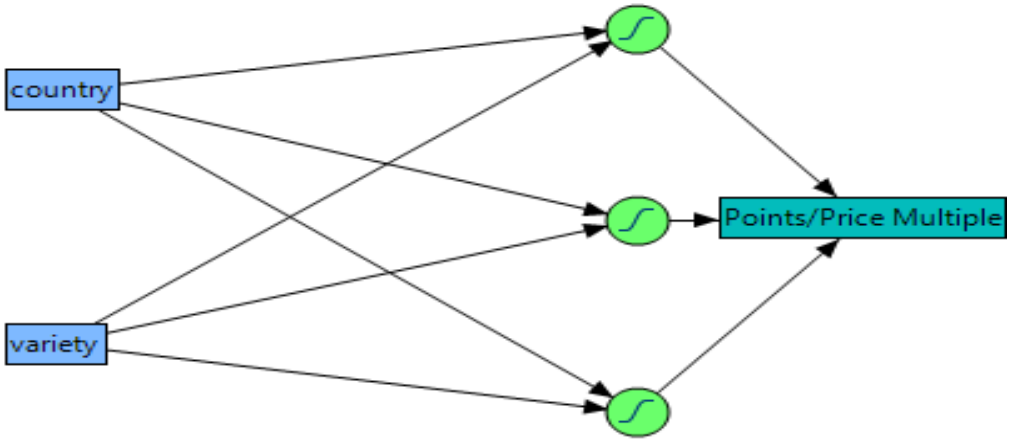
Figure 1. Linear Regression of Price of wine to the review in received.



Does country and variety have a correlation with the points/price multiple on each wine?

Yes, there is a connection between the country and variety which leads to a price premium. I completed this analysis by using neural networks. The factors I considered where country and variety and I the outcome I used was the price/points multiple I computed during my data preparation. Using three random samples from the data, I found a .31 correlation in all three neural networks I ran. Although this is not the strongest correlation, using neural networks illustrated that country and variety and a better determinant of points a wine will received compared to simply basing off price.

Figure 2. Neural Network of Country and variety in determining the Points/Price Multiple



V. Data Visualization

Figure 3. Top 10 Price-Point Multiple By Country

This figure illustrates the premium given to the top 10 countries based on the price to points multiple, calculated using the dated from 150,000 wine reviews.

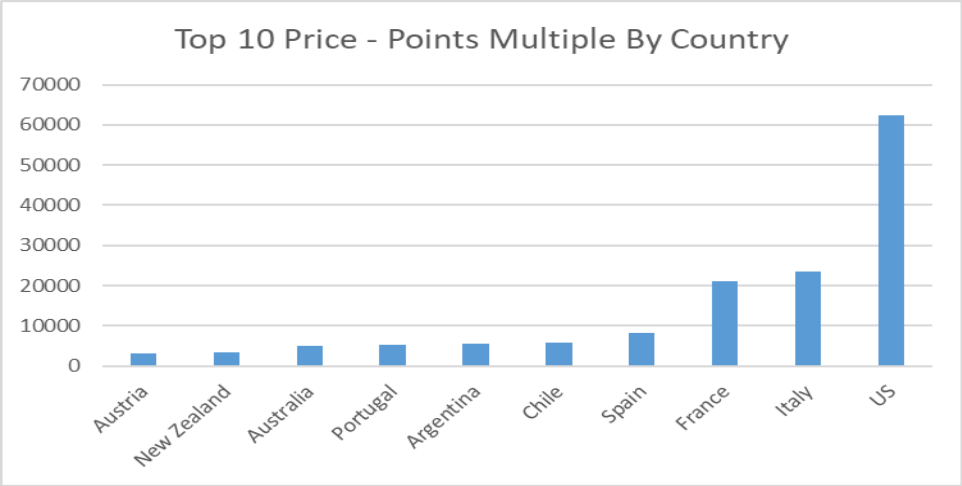


Figure 4. Percentage of Wine Reviews By Country

This figure illustrates the countries which were represented in the 150,000 reviews included in the data set.