Wine-O-Mania

Our group’s project was to discover how to select a bottle of wine. We started by developing a variety of hypothesis and/or impact statements to research. Our dataset was provided by Kaggle. A dataset with 150K observations, wrangled to 137, from Wine Reviews was used. Once the data set was downloaded and cleansed, we began developing the dictionaries, lists, and data frames to understand that the data was telling us. We geo-coded the country, province and region (sub-area within the province) provided in the dataset via Google Maps Cloud API and Google Maps Text-to-Search to identify the latitude and longitudes of wine-growing locations across the world. We looked at the number of varieties of wine grown at each of these locations.

The first question we posed was does the price of the wine impact a consumer’s decision? We looked at the pricing and rating. Based on the information we found, majority of consumers purchase wines within the $30-50 range. Over 50% of the wine purchase where in this range; $30 was 37.3% and $50 was 26.3%. Next, we looked at the location where the wine was purchased.

The next question we posed was the relationship on price and locations; to display this we used different scatter and box of the points and prices were run, showing a loose linear/power correlation between price and points. The boxplot showed the spread of the prices increased significantly as the points ratings got close to 100, while observations wine’s for with points from 80 and 94 showed little price spread. A hypothesis test on the correlation between price and point showed there is no significant correlation between the price of a wine and its ranking. In addition, different regressions showed points can’t predict wine price. Conclusion based on the hypothesis test and observation of the box plot, wine enthusiast can get Very Good (85-89) to Outstanding (90-94) wines for the prices similar to those charged for Good wines. I.e., don’t pay more just because the wine has a higher point value.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pearson’s r coefficient and regression results for each wine sub group.** | | | | |
| **All wines** | **Good Wines** | **Very Good Wines** | **Outstanding Wines** | **Classic Wines** |
| **Pearson correlation coefficient between points and price** | | | | |
| **0.459867** | **0.313647** | **0.453946** | **0.254986** | **0.03738** |
| **R-square for linear regression** | | | | |
| **0.470** | **0.689** | **0.691** | **0.614** | **0.477** |

The next question was how the soil conditions impacts the wine price and rating. We thought of identifying the climate conditions (annual average temperature, precipitation) and soil (pH and moisture content) characteristics of these wine-growing locations and if at all they contribute to Wine ratings or not. It was challenging to gather the data for those locations all over the world. Therefore, we decided to focus only on US locations, as the data for these annual conditions were available for free (USDA, API calls to Agro-monitoring {partner website of OpenWeatherMap}). It is important to report annual data instead of reporting current conditions which is not an accurate representation of climate and soil-characteristics of wine-growing areas. We found the ideal conditions for the soil were:

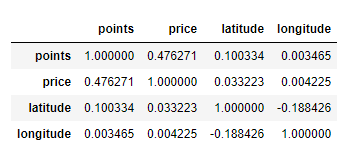
Temperature – 6-12°C

Rainfall – 200-800 mm

Soil pH - 5.5-8.5

Soil moisture – 0.15-0.37

The final question was what the relationship between the location and rating is. We developed the following hypothesis statements; Ho: The location of the winery does affect the wine’s rating. Ha: The location of the winery does not impact the wine’s rating. We conducted a correlation analysis using Pearson’s method and learned that the winery’s location has a weak positive impact on how the wine is rated.



The conclusion of our project:

- pricing influences wine purchases the most

- pricing doesn’t guarantee a high rating wine

- there are ideal soil conditions ensure the quality of wine

- the location of the winery does not guarantee a high rating