WineEnthusiast Reviews:
An Analysis of Economic Value, Geographic Variations, and Descriptors

Economist - Andrew Dively
GIS Data Scientist - Ashley Moss
NLP Data Scientist - Patti Degner

DataSet

The main dataset we would like to analyze is the “Wine Reviews” dataset from Kaggle.

Our data is 129971 observations of 14 variables describing:

- Index
- Country: the country the wine is from
- Description
- Designation: The vineyard within the winery where the grapes that made the wine are from
- Points: The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
- Price: The cost for a bottle of the wine
- Province: The province or state that the wine is from
- Region_1: The wine growing area in a province or state (ie Napa)
- Region_2: Sometimes there are more specific regions specified within a wine growing area
- Taster_name
- Taster_twitter_handle
- Title: The title of the wine review, which often contains the vintage if you're interested in extracting that feature
- Variety: The type of grapes used to make the wine (ie Pinot Noir)
- Winery: The winery that made the wine

Notes:

- There is missing data in nearly every column
- Possible data points to analyze:
  - Price
  - Points
  - Country
  - Variety

Wine notes were scraped from Wikipedia: https://en.wikipedia.org/wiki/Wine_tasting_descriptors

Research Interests

We explored the quality and value of wine from several angles. We analyzed which countries produced the highest scoring wines. We dove into economic value measures and determined which regions and brands had the best value from an economic standpoint. Lastly, we looked at the way particular wine descriptors such as “fruity” and “herbaceous” correlated to wine quality.

Value

- Which countries/regions had the best value in terms of points/dollar?
- How does rating distribution change along price?
- How are values (points per dollar) distributed across the data set?

Variations by Region

- Which countries/regions produced particular varieties of wines?
- Which countries/regions produced the highest scoring wines?
- Which countries/regions had the best value in terms of points/dollar?

Tasting Notes

- Are there descriptors that affect price?
- Which wine notes lead to higher value?

Data Cleaning & Metadata

Overall, the dataset didn’t require much cleaning, as it was acquired from Kaggle. We did end up creating a lot of Metadata variables (Wine Age, Sex, Wine Year, Rating Estimate, and Value), which significantly increased our ability to analyze the dataset.

Taster Name: We had to remove a lot of special characters that were probably the result of some encoding error.

Description: We had to use text analysis to remove key wine descriptor words from the reviews. A list of descriptors was scraped from Wikipedia. Then the descriptors were extracted from each description and analyzed.

Wine Year & Wine Age: We noticed that the wine titles contained the year, for example, “Failla 2010 Estate Vineyard Chardonnay (Sonoma Coast)”, so we were able to separate all of the numbers from the text, but in many cases, there were other numbers inside of the text, not related to the year, which required extensive work, but in the end, 97% of our data was able to have an accurate wine year. From the wine year, we performed a simple calculation 2020 - Wine Year = Wine Age, to create the Wine Age variable.

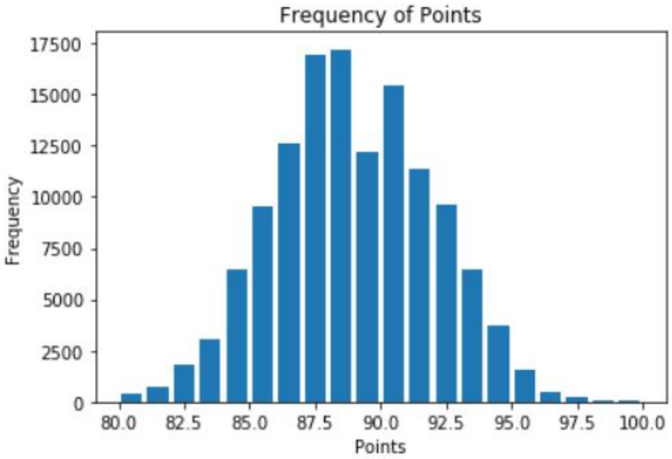
Expected Points & Value: As you will see below, we created a model for points vs log(price), which gave us a regression line that fit the data well. We ended up with an equation for the regression line: expected points = 78.9 + 6.55 * log(price), by plugging in price we were able to create the expected points variable. The value is simply how much the actual points for a particular wine over or underperformed the expected points: value = actual points - expected points.

Country Polygons & Country Codes: Map data showing the detailed polygon outline of a country was obtained from this open source resource and transformed to a Python GeoJSON format. In order to join to the geographic data, the original wine data was augmented with a new column for 3-letter country code.

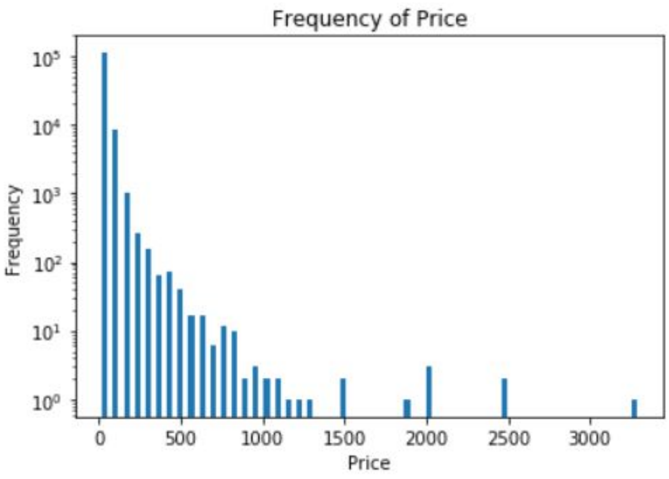
Data cleansing: N/A values were checked for minimal counts and then removed or adjusted as appropriate. 0-denominator values were scrubbed as well when calculating ratios.

Exploratory Analysis

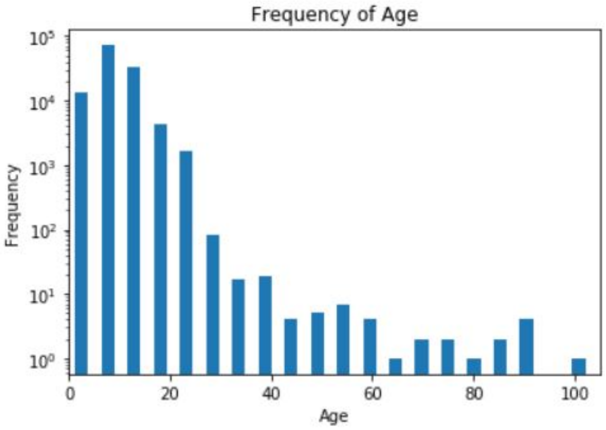
Points Analysis: In general the wine rating scale starts at 50, but WineEnthusiast only posts reviews for wines scoring 80+ points. Below is a frequency plot of wine ratings. They are normally distributed around 88-89 points.



Price Analysis: Price follows a logarithmic distribution under $3500, so we decided to use log adjustments on price throughout the paper. In this diagram, the y-axis is on a log scale to show key outliers in the data.

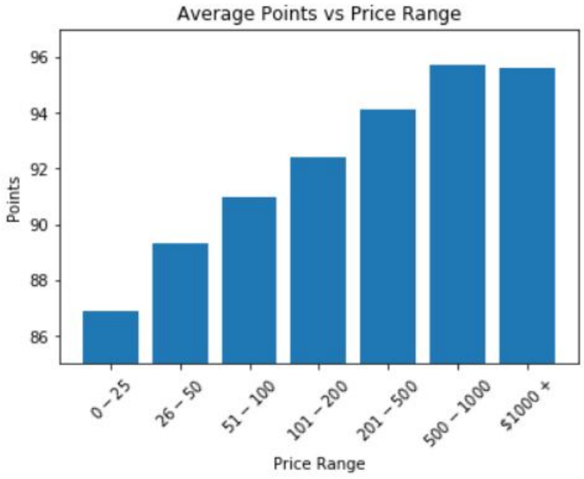


Age Analysis: Reviewers overwhelmingly sampled wines 20 years or younger, we put the y-axis on a log scale here because of the large amount of data.

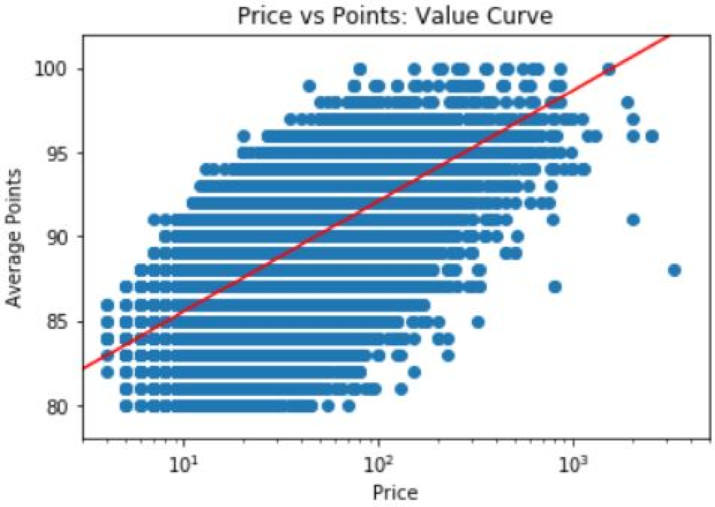


Wine Value

Looking at the data, we observe that there is a wide variation between the points for certain price ranges.

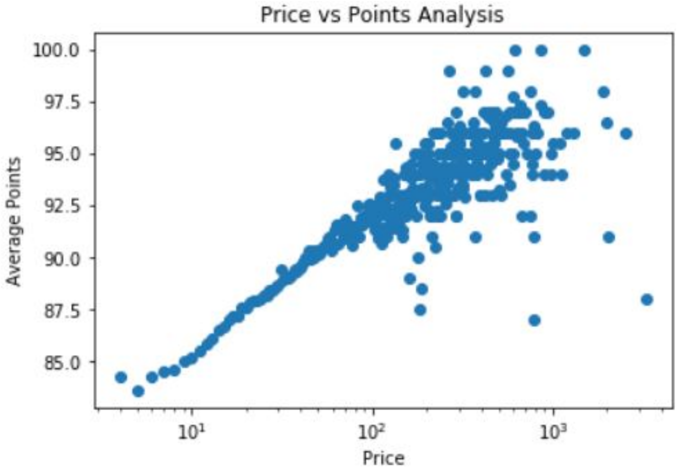


To better understand this relationship, we used linear regression to develop a model for the relationship between points and log(price). We used log in this case because it produced a model of much better fit, which can be seen below:

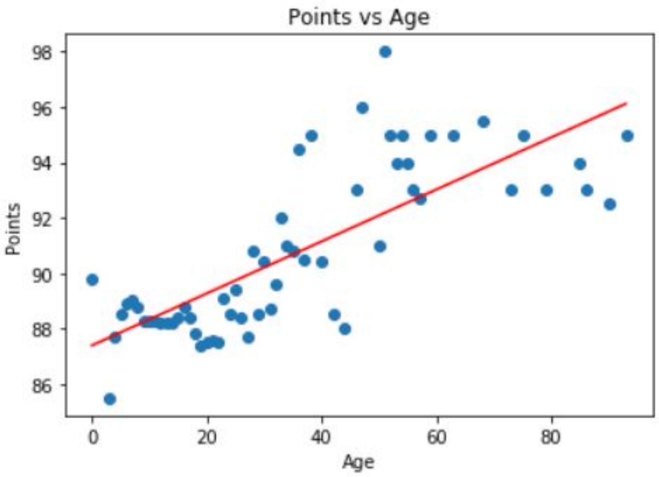


From the consumer's perspective, anything above the **Value Curve** means that a particular wine's rating exceeds that which would be expected for a particular price. Anything below the value curve indicates the wine's rating is below what would be expected for a particular price.

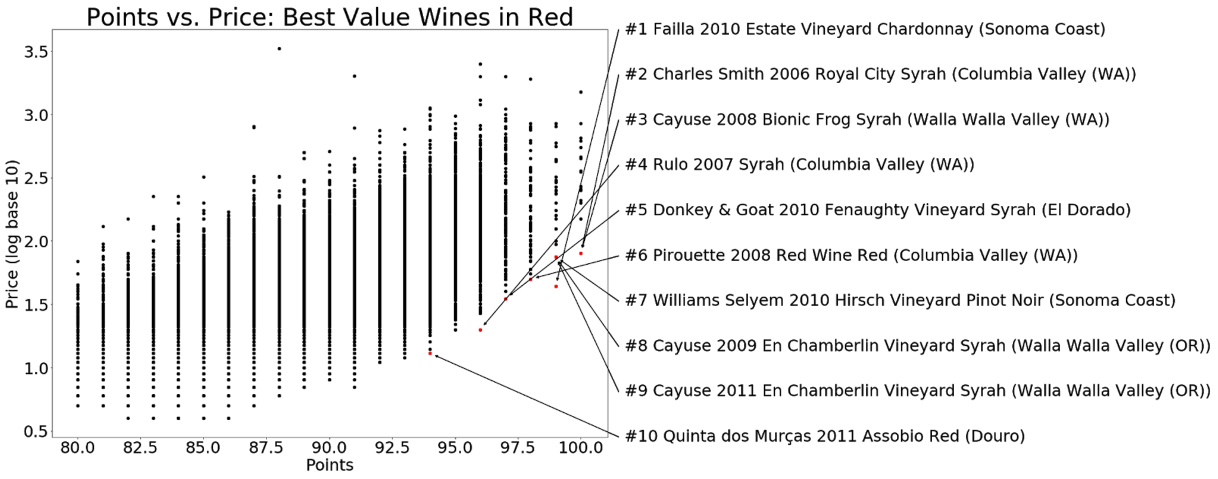
Price vs Average Points: When taking the average of points across each price value, the relationship becomes even clearer, this gives each price equal weight, since our distribution has a heavy concentration of lower priced wines.



Points vs Age: We also noticed a clear relationship between points and age. Points increased as the age increased. Again, we used average age to compensate for the small proportion of younger wines sampled.



Best Values: By graphing the points against the log of price, we can visualize the best value wines. The highlighted top 10 wines below outperform expectations from an economic standpoint and are graphed with a red dot to make them stand out in the visualization. Several wines from the top 10 list were red wine varieties from the Pacific Northwest of the United States.



Wine Geography

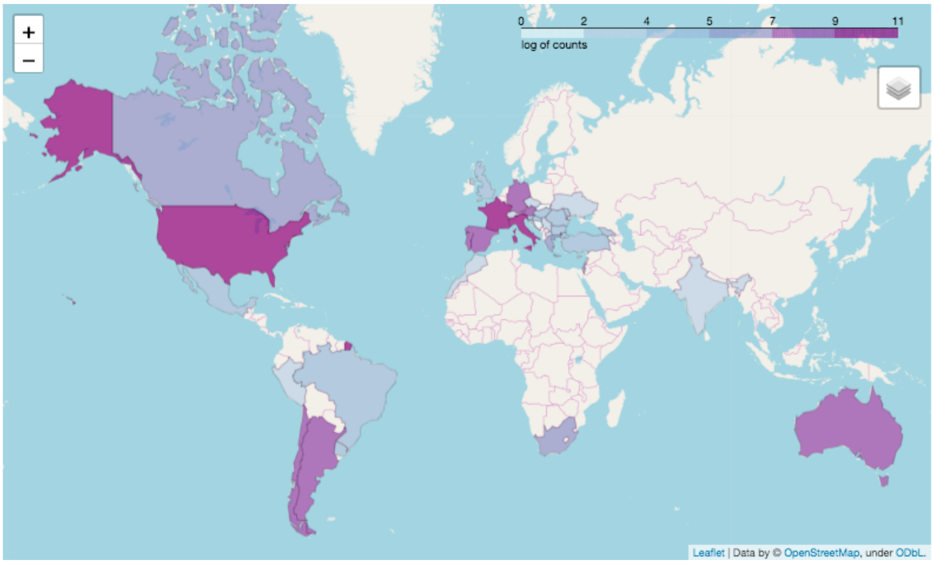
Exploratory Steps: Countries in the Dataset

To analyze the data by location, mostly we created choropleth maps where areas are shaded in proportion to the value being measured. Map data showing the outline of a country was obtained from an open source resource and joined to an augmented version of the wine data. Finally, we built interactive graphics using the package Folium.

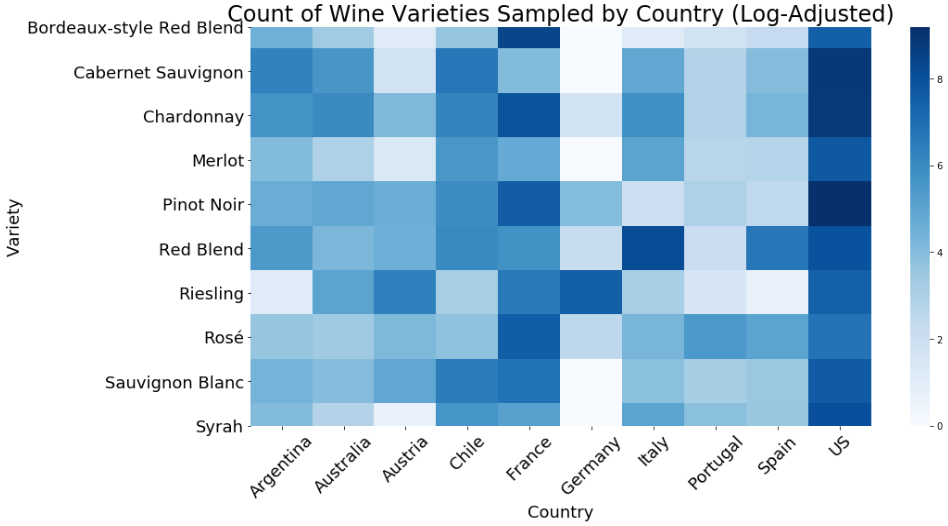
When going through the choropleth process it was helpful to pre-aggregate the data to be visualized in order to cut down on computational requirements in the visualization. Another best practice is to first test the result with a very small subset of the data, e.g. filtering to one country.

One question is whether all countries are represented in the data set and whether they have an equal number of reviews. There was uneven coverage with only 44 countries represented and the US heavily favored. Russia certainly has at least one winery, but none were reviewed. In summary this data represents a Westernized view of the wine world. Below find a log-adjusted graph of count of reviews by country.

Count of Reviews by Country, on a Logarithmic Scale



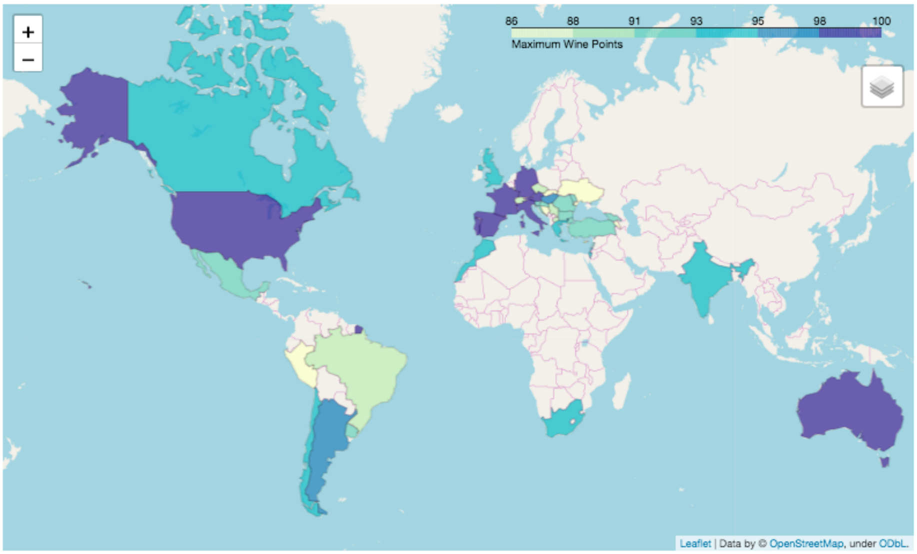
Which wine varietals correspond to each country? Below is a heatmap showing the log-adjusted count of wines in top categories. The log transform tones down the dominance of America in the counts. Climatic factors influence which types of grapes can grow. Germany is noted for Riesling; France, Rosé and Bordeaux; and America, Pinot Noir and Chardonnay.



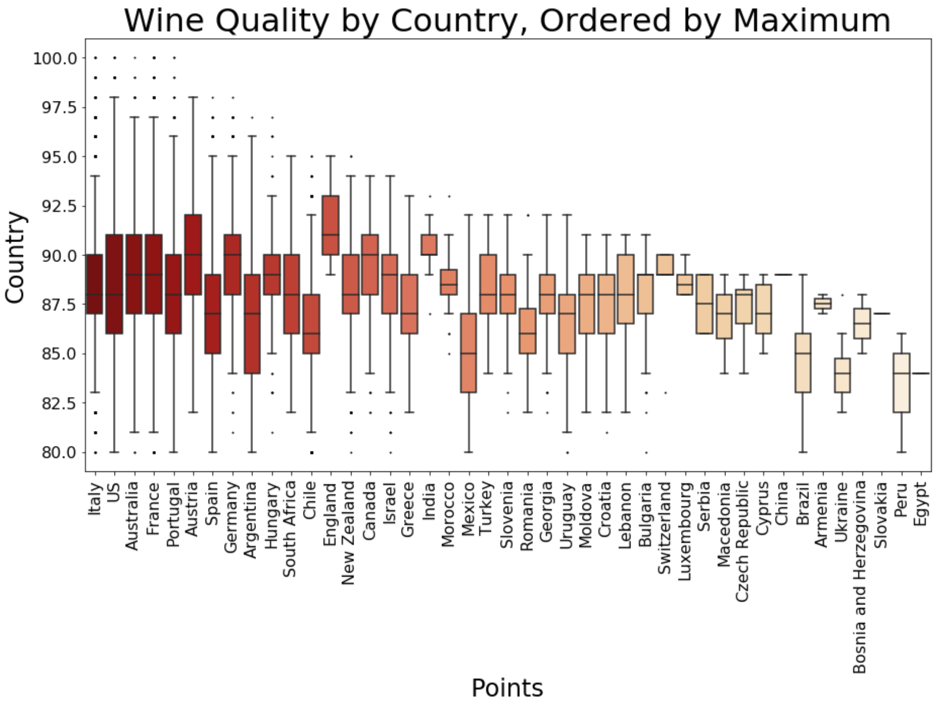
Analysis: Maximum Quality by Geography

The maximum wine point ratings fall roughly line with global wine reputations, showing marked contrast between best and worst countries. The U.S., Australia, and several Mediterranean nations ranked high on the list. Eastern Europe and South American countries ranked lower.

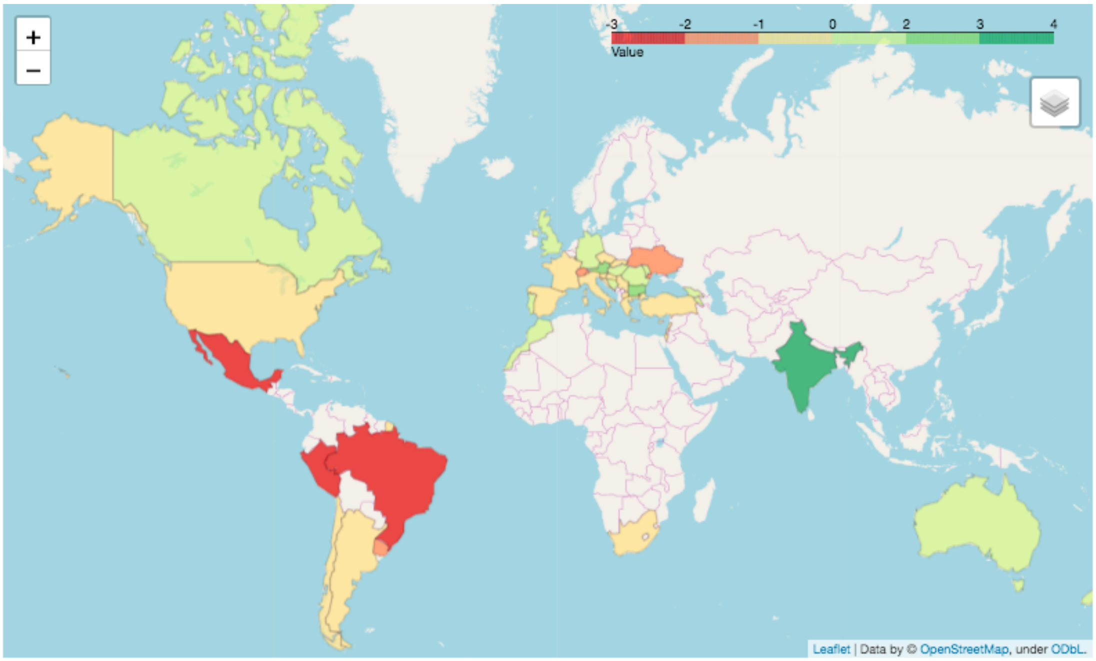
Maximum Wine Rating by Country



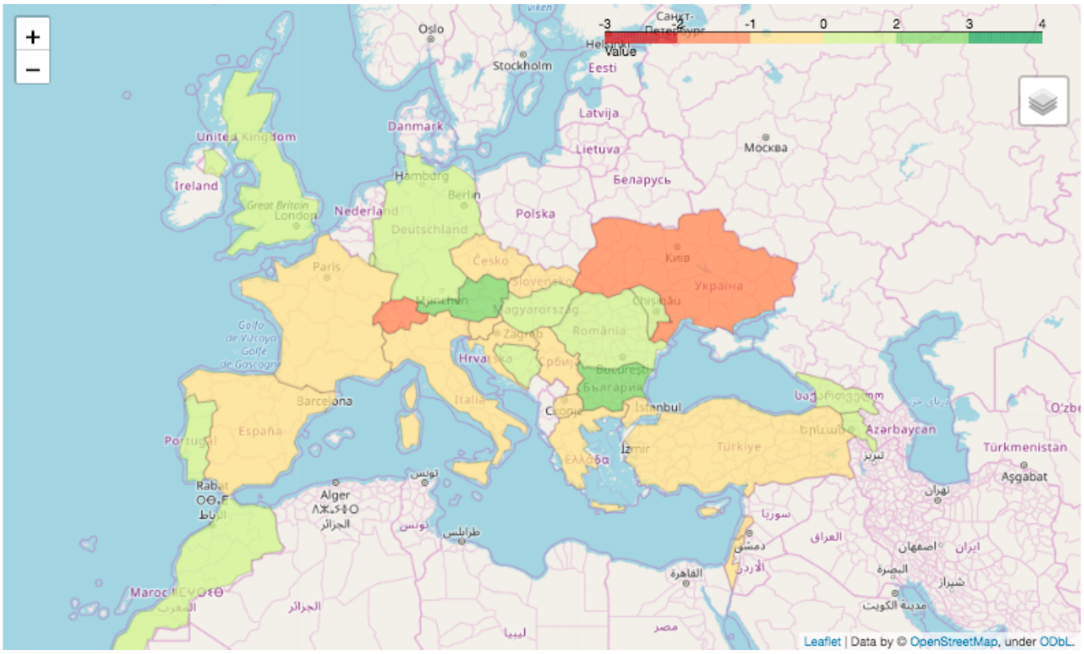
A connoisseur may be curious about detailed rating distributions by country. Here we can see a boxplot of wine ratings by country ordered by maximum. Five countries (Italy, US, Australia, France, and Portugal) had 100-point wines in the sample.



Lastly we integrated the economic value analysis with the mapping method. As mentioned in the value analysis section, the derived wine score gives a measure of the economic bargain associated with each wine, which we then aggregated by region. Of countries with moderate sample sizes, Austria, Germany, and Portugal showed the best value scores. Our hypothesis is that these countries produce well-made wines without attracting the kind of global attention that drives up prices elsewhere.



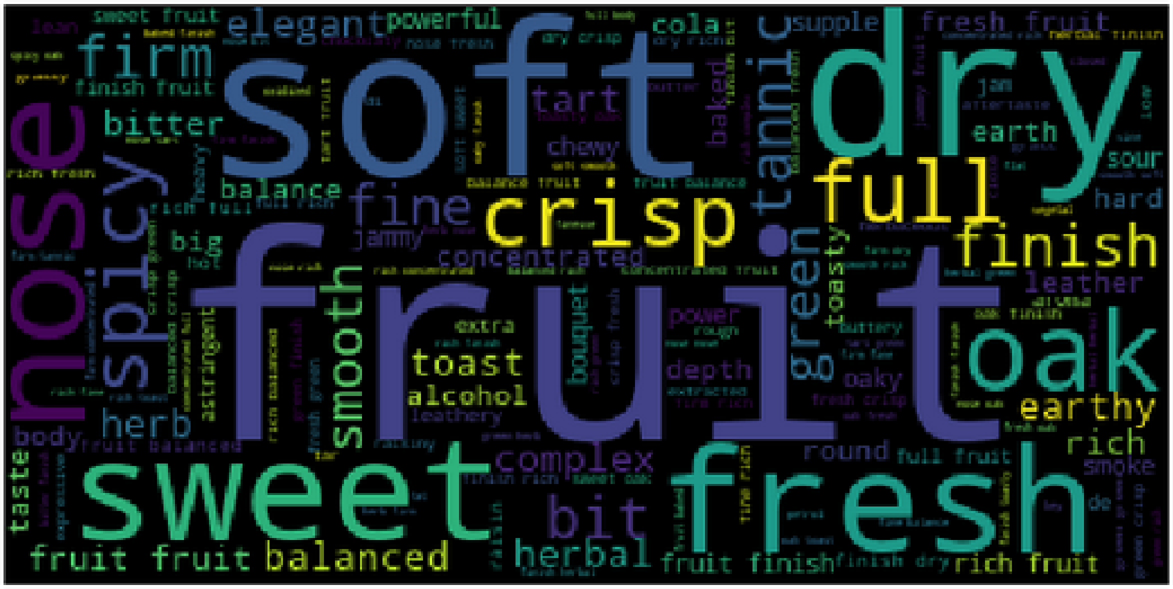
Here is a zoomed-in view of Europe showing Austria and Germany scoring high.



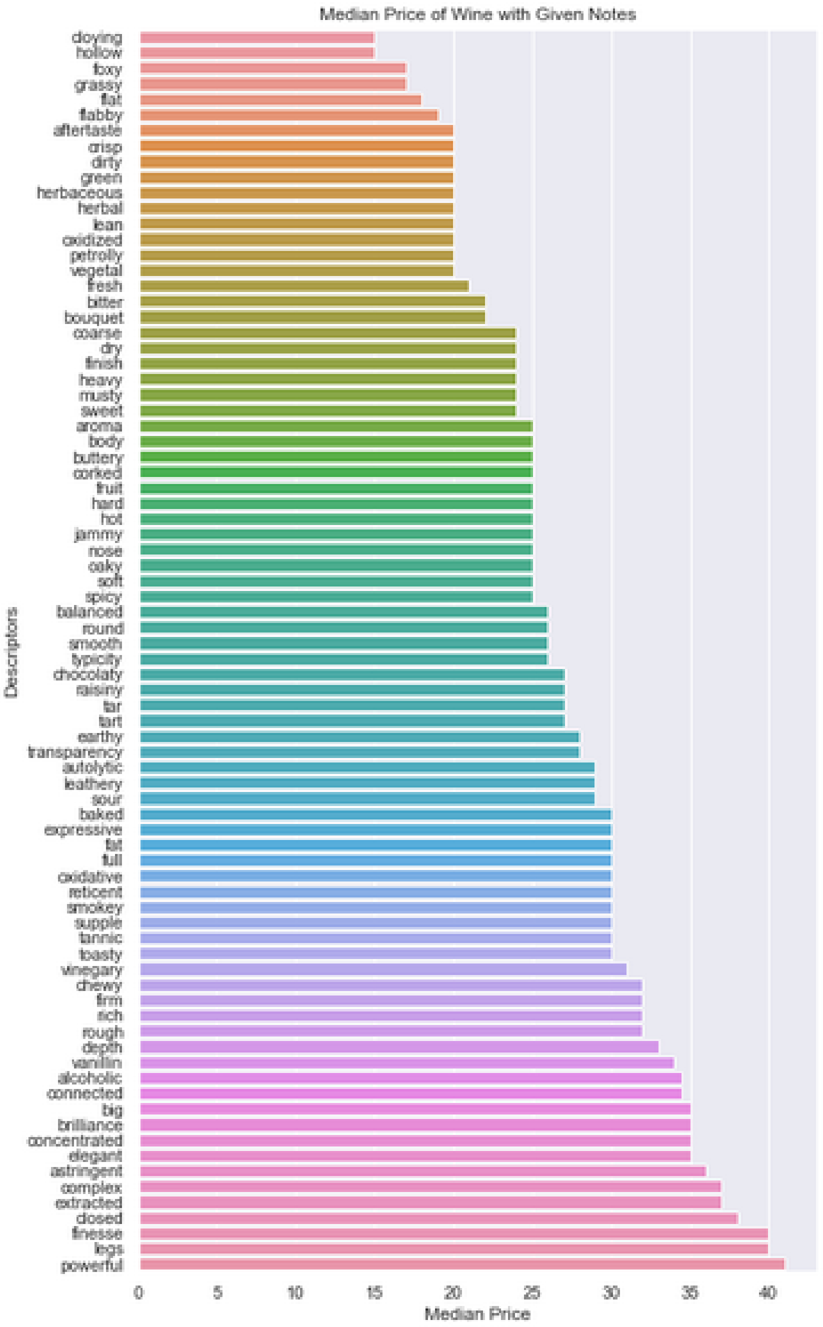
Wine Descriptors

We wanted to see how the description of wines changed the price and value. To begin, we scraped a list of wine descriptors from this Wikipedia page: https://en.wikipedia.org/wiki/Wine_tasting_descriptors

To begin this analysis, we created a wordcloud. This wordcloud only contains words that are considered “descriptors”, and we pulled the text from the “description” column of the dataset. This exploratory data analysis gave us an idea of how frequently various descriptors come up in this dataset.

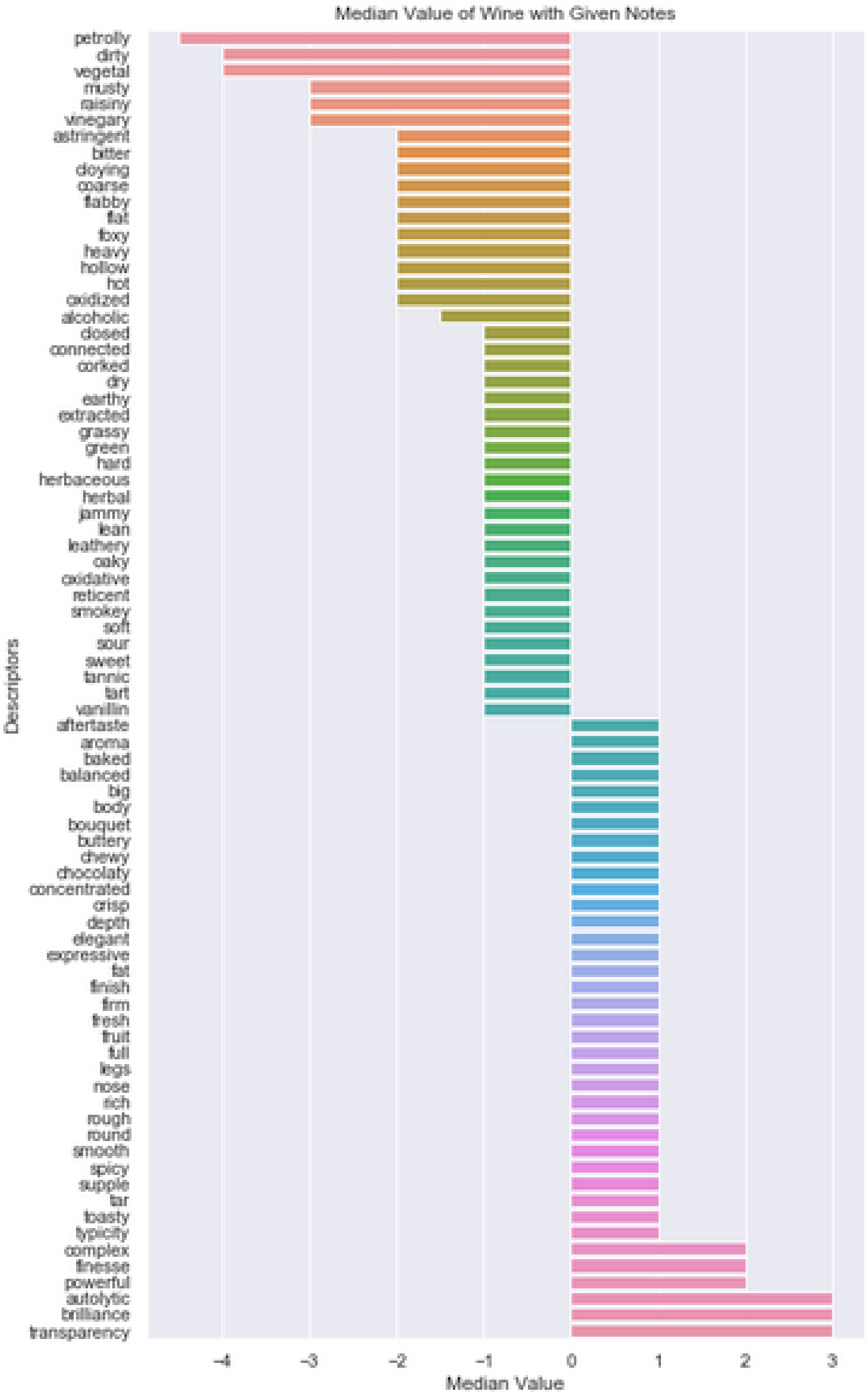


Next, we created a graph of median price of wine with a given note. For example, wines with the note “dirty” had a median price of $20.



Here, we see some interesting results. Wines described as “doying”, “hollow”, and “foxy” tended to have lower prices, where wines that are “powerful”, had “legs”, and “finesse” tended to be more expensive.

We made a similar chart based on wine value by note (where wine value is calculated as stated above).



Wines described with “autolytic”, “brilliance”, and “transparency” tended to have higher value, whereas wines described as “petrolly”, “dirty”, or “vegetal” tended to have lower value. Powerful wines tend to cost more, but still provide a good overall value. Wines described as “foxy” are cheap, but you don’t get good overall value. Wines described as “smooth” are mid-range and provide good value for your money.

We cannot assess how strongly these descriptors affect the price and value without doing regressions. However, this initial analysis indicates that there may be significance to the descriptors of wine and their effect on price and value.