Sommelier Statistics

Introduction:

Wine is a drink that is consumed worldwide by many distinct cultures. Some can even say the drink has lasted the test of time and has even tasted better by doing so. There are many ways to produce wine and every critic has an opinion on what tastes best to them. Factors that have impact on the fundamental taste include region, year, climate, price, type, year, etc. Many people have different criteria when they decide to purchase a bottle of wine. We want to identify some key insights by exploring these factors and their relationships.

With Kaggle's data set, we will look at the global wine industry through 'WineEnthusiast'.

Research Questions:

1. We looked descriptive factors that consumers consider when deciding to purchase wine:
   - What region has the most expensive wine?
   - What region has the best rated wine?
   - What type of wine is rated the best?
   - What types of wine are most prevalent based on region?

2. We also sought to identify correlations, especially to external weather and climate data:
   - Does the price of wine correlate with better ratings?
   - Does the vineyard's Köppen climate classification correlate to wine rating??
   - Do ratings correlate to major climate deviations in the wine's country of origin for its vintage year?

Our Data:

Primary Dataset: 'WineEnthusiast' - https://www.kaggle.com/zynicide/wine-reviews
The dataset was scraped as of November 22nd, 2017 with the following critical parameters:

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)

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)

Description: A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.

Country: The country that the wine is from

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 (i.e. Rutherford inside the Napa Valley), but this value can sometimes be blank

Winery: The winery that made the wine

Designation: The vineyard within the winery where the grapes that made the wine are from

Price: The cost for a bottle of the wine

Taster Name: Name of the person who tasted and reviewed the wine
Taster Twitter Handle: Twitter handle for the person who tasted and reviewed the wine

Supplemental Datasets:

http://hanschen.org/koppen/#data
Köppen climate classification by decade in a 0.5 degree lat/long resolution. The dataset includes latitude, longitude, and decadal classification columns. We will use the data from 2001-2010 as it is most representative of the vintages of our wine data. Furthermore, we feel it is unnecessary to retrieve the Köppen climate classification for precisely the vintage year, as the climate classification is very gradual in its changes.

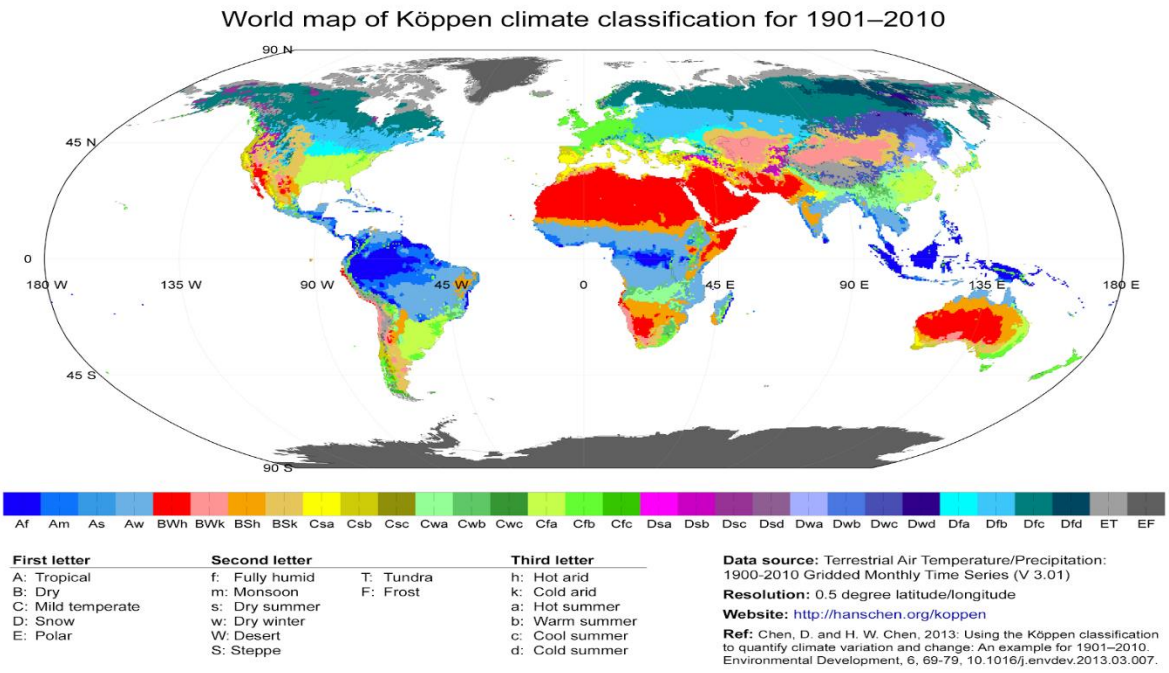

Figure 1: Visual Map of Köppen Climate Dataset

https://www.ncdc.noaa.gov/cdo-web/datasets
https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/gsom-gsoy_documentation.pdf
Global Historical Climate Data (Global Summary of the Year) aggregated by the National Oceanic and Atmospheric Association (NOAA). The dataset involves 75,965 individual .csv files, each containing data from a unique weather station. Collectively, the stations have more than 2 million rows of data, with each row containing summary statistics of a year's weather data. We investigated fields containing annual rainfall (mm), days/year when max temperature exceeded 90F, and days/year when max temperature did not exceed 32F. The dataset proved to be larger than we expected, but by focusing on various subsections we were able to still work with it.

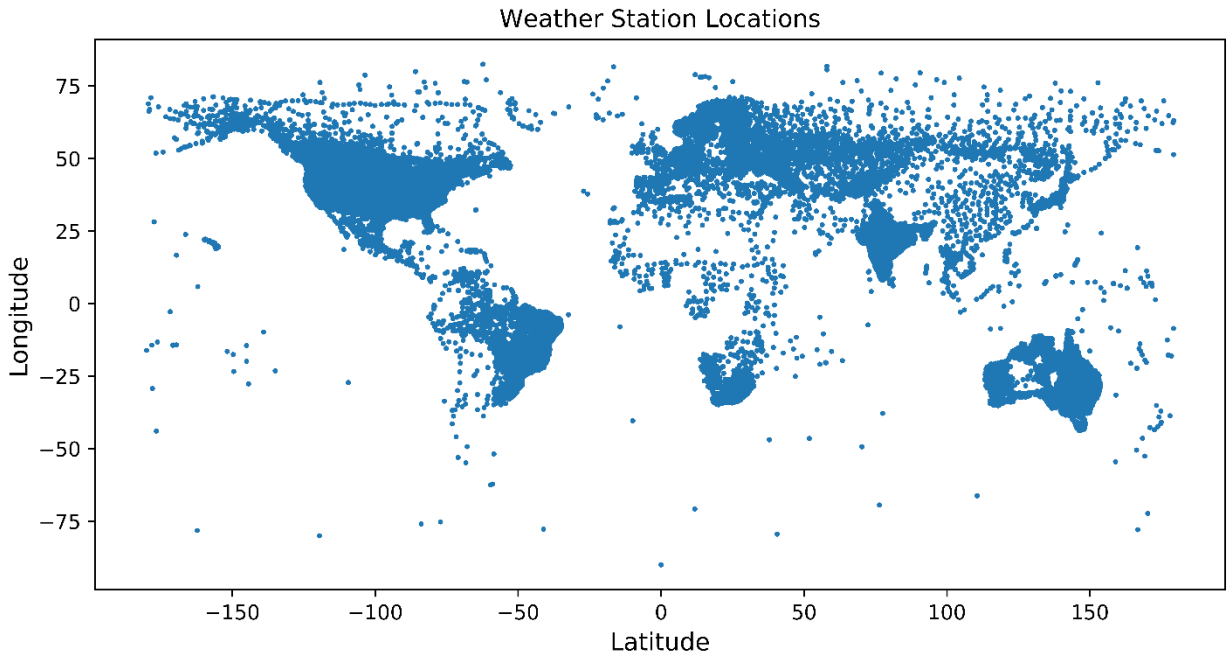

Figure 2: Visual Map of NOAA Weather Station Locations

Initial Data Exploration, Validation, and Cleaning:

Cleaning Primary Dataset:

The original dataset was read into a pandas dataframe from its .csv source file. From here we removed an unnamed column that was created during the import and set each data series to an appropriate data type.

'Country' field:

The country field was originally null in 63 records. By grouping these records by their winery, we were able to manually search for 27 unique wineries and populate their country. After filling in null values, we examined the unique country values to determine that none were slightly misspelled repeats of each other.

'Description' field:

The description field was null in zero records. While many fields were similar, most were unique to the wine being sampled. The questions we pose shouldn't require further cleaning of this series, but could be useful for creating a wine classification algorithm based on the frequency of words in this field.

'Designation' field:

The designation field was null in 37,465 records, with many unique designations repeated in slightly different spelling or format. For example, the "Reserve" designation was accompanied by "Reserva," and "Riserva" in the five most frequent designations, with many more variations in the remainder of the dataset. While the designation field purportedly describes individual vineyards within a winery, we find that it often fails to do so (i.e. "Reserve"

and “Extra Dry”). If we intended to use this field to in our analysis we should parse the values that include “vineyard” or “ranch” and store them in another field. At this time, our questions posed should not require it.

‘Points’ field:

The points field was null in zero records. As described by the dataset documentation, the point values are limited to the range 80-100. While this field doesn’t require any additional cleaning or validation, we will need to keep in mind this limitation. For example, examining the relationship between price and points will be limited in scope to the dataset’s high-quality wines.

‘Price’ field:

The price field had 8,996 missing values, which were populated by NaN values to ensure they didn’t affect the statistical methods we planned on using. The range of values present were logical, ranging from $4.00 to $3,300, with a mean of $35.36.

‘Province’ field:

The province field was null in most records. Of the records that contained data in this field, most were U.S. states. Using this field could be useful for state comparisons, but it mostly useless as a refinement to the wine’s geographic origin. Instead of using the province field, using either the country field or our calculated latitude and longitude fields will provide a more complete indicator of wine origin at different resolutions.

‘Region_1’ field:

The province field was null in most records. Of the records that contained data in this field, most were relevant to subregions within U.S. states. See the above description of the province field.

‘Region_2’ field:

The province field was null in most records. Of the records that contained data in this field, most were again relevant to subregions within U.S. states. See above description of the province field.

‘Taster_Name’ field:

The taster name field was null in most records, but where present, it was precise. There were only 19 unique tasters listed in the dataset, making it relatively easy to compare on this dimension. However, any discovered insights - such as the sommelier bias question posed above - will be limited to the individuals and will not hold to statistical inference to sommeliers in general.

‘Taster_Twitter_Handle’ field:

The taster twitter handle field is a subset of the taster name field. Not all taster names are associated with a twitter handle (some tasters probably don’t have twitter), but all twitter handles also have a name associated to the record. For this reason, this field adds no information to our analysis unless we intend to contact the sommeliers.

‘Title’ field:

The title field was null in zero records and was unique for about 90% of the records. The title is usually a combination of data from the variety, winery and designation fields, but also includes wine vintages.

'Variety' field:

The variety field was null in one record, with 707 unique varieties. Over 100k wine records fell into the top 25 varieties, with varieties becoming ever more descriptive as their relative frequency decreased. Analysis on the most common varieties should not require addition data cleaning.

'Winery' field:

The winery field was null in zero records and held 16,757 unique values. The field is fairly diverse: the 8,500 most common wineries account for 90% of the records.

Calculated and Merged Data Columns:

Vintage field:

The vintage field needed to be parsed from the title field. We first used a regular expression to search for the first four-digit number in the title field, setting this as the corresponding record's vintage. After reviewing the vintages resulting from this method, we found we had vintages as large as 7200 and as small as 1000. Exploring these anomalies, we found many titles included four digit numbers that were not vintages, such as "Foxen 7200" (a type of wine), "1000 Stories" (the name of a winery), or "1503" (the year the winery's manor was built). Next, we attempted to search for the four-digit vintage starting from the end of the title field. This caused similar errors, especially where there was no vintage listed and the title still included a four-digit number. After conducting research of extremely old wines, their palatability, and their cost, we assumed that we would not have data on vintages older than 1900. Limiting our regular expression search to the range of 1900-2020, we then individually examined the oldest vintages, finding that 1934 was the oldest vintage in the dataset. Finally, we constrained our regular expression to the year range of 1934-2020. 2020 was chosen for simplicity in writing the regular expression, as no wines resulted in a vintage newer than 2017. There is a potential that some of these parsed values in the range 1934-2017 do not, in fact, represent a vintage, but rather a different aspect of the wine's or winery's name. We will remain cognizant of this fact throughout our analysis.

Köppen Climate Classification (koppen):

In order to utilize our external dataset to determine the climate classification for each record, we first determined the latitude and longitude for each winery in our dataset. To do this, we used the Google Maps Geocoding API to attempt to retrieve the latitude and longitude for each winery. This latitude and longitude data was then transformed to match the gridded format of the Köppen dataset, and then was used to lookup the corresponding classification and assign it to a "koppen" series for all wine records that shared that originating winery. After removing this new data for entries that did not place the winery in the correct country or U.S. state, we were left with 99,164 successful additions of latitude, longitude, and the Köppen climate classification to our dataset. Given our time constraints, we chose to limit our analysis that required these fields to this subset of records.

Global Historical Climate Data:

To work with the NOAA data, we needed to read in all the files and stitch together the data into a dataframe containing the station's latitude, longitude, and various annual weather measurements. We then validated that the station coordinates looked reasonable by plotting them on a map with dots on a lat/long plane. What emerged was a mostly complete outline of the Earth's landmasses – highlighting just how extensive these stations are. We then looked for

anomalies in various weather measurements we were interested in, including annual precipitation, days/year of max temps exceeding 90F, and days/year of max temps below 32F. For instance, we determined one station in New Caledonia had bogus precipitation readings based on external climate descriptions and visual inspection of the vegetation with Google Earth. Finally, we worked to merge the weather data to our wine dataset by choosing various subsets of wine records, calculating the nearest weather station to the winery, and appending the station's weather data for the wine's vintage. In this way we were able to know the approximate rainfall, heat, and freezes when specific wines were produced. We then spent considerable time trying to find correlations between weather and the price or rating of wines.

Results:

1. Does the price of wine correlate with better ratings?

As non-sommeliers, we were particularly interested in the value of different wines. We visualized this relationship in Figure 3, where we find that wine rating and price distribution is correlated, but not deterministic. Note that price outliers are not displayed for the sake of clarity. The correlation is clearly positive, with the distribution of prices moving upward as wine rating is increased. However, the increasing variation in the data indicates that you don't need to splurge to buy the best rated wines. Further examination of this relationship would be interesting. For example, what variables underlie both a sommelier's rating and a wine's price? Is this correlation largely an effect of wineries charging more for well rated wines?

2. Does the vineyard's Köppen climate classification correlate to wine rating?

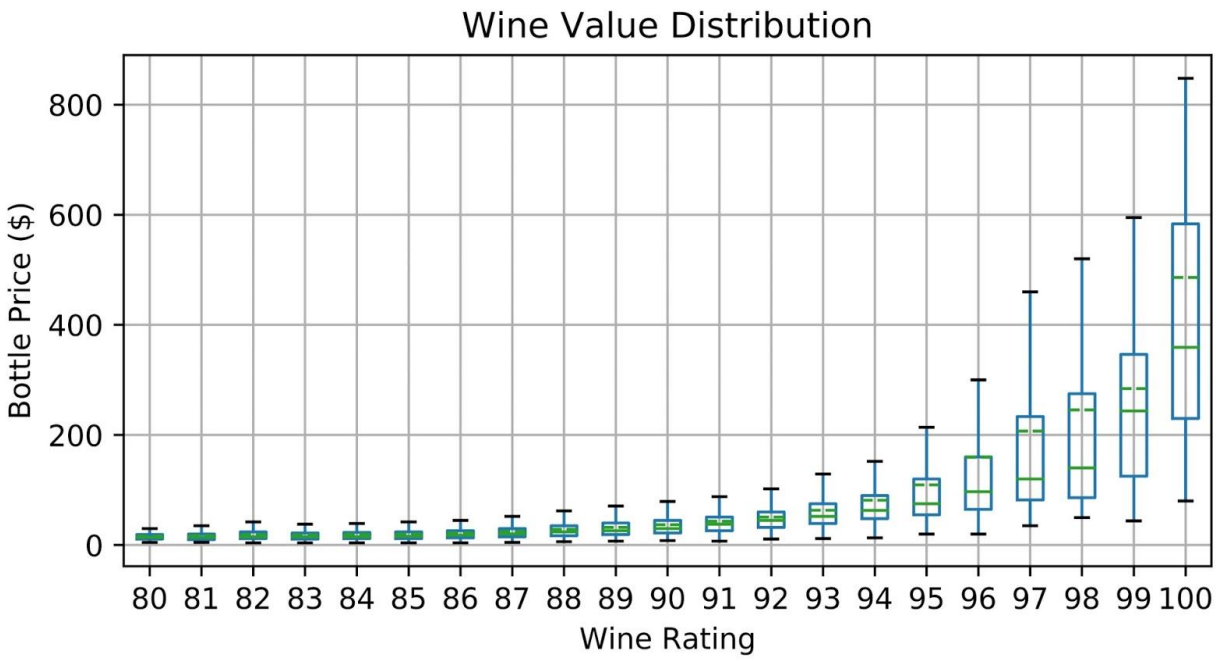

Figure 3

To begin exploring the relationship between the climate classification and wine rating, we first performed the data calculation and dataset merging discussed above. Once these attributes were combined into our main wine dataset, we quickly discovered a high correlation with quality wine and climate classification. It was readily apparent that while there are many distinct climates that support grape growth and wine production, a few climates accounted for most of the production, with Csb (mild temperate, dry summer, warm summer), Cfb (mild temperate, fully humid, warm summer), and Csa (mild temperate, dry summer, hot summer) accounting for over 75% of the wines analyzed. To determine if these climates were the best for wine production, and not just most accessible to wine producers, we standardized the proportion of wine record climates against the total land area that each climate represented in the world. The climates that are overrepresented in wine production (ratio > 1) compared to the world distribution of climates are displayed in Figure 4.

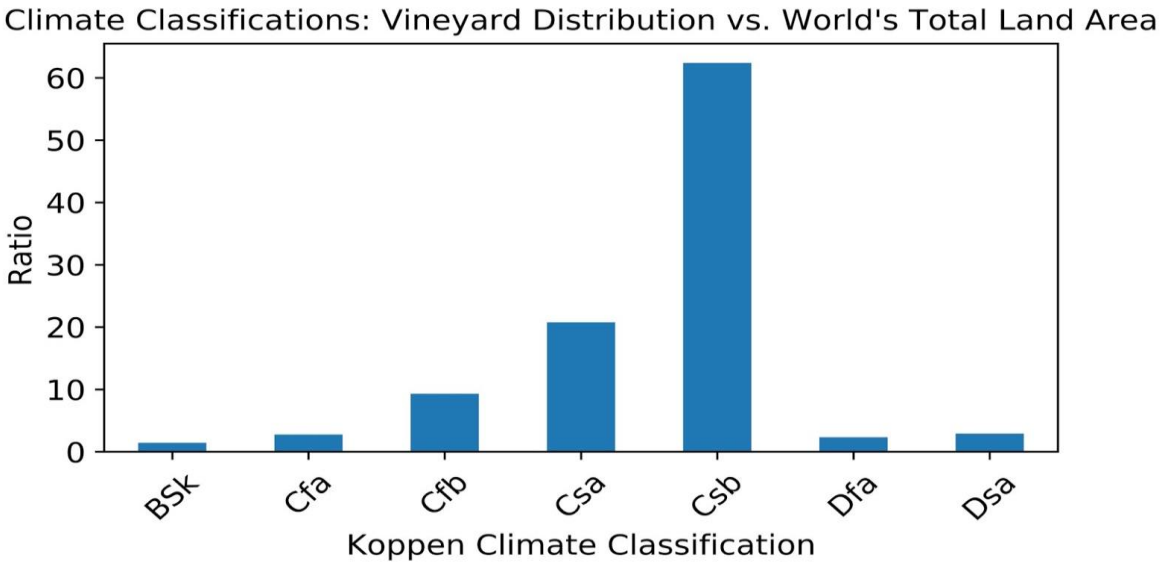

Figure 4

Here we see, for example, that Csb climates produce about sixty times the amount of wine than would be expected from a random distribution, with Csa and Cfb climates also showing significant overrepresentation. While all wines in our dataset are of high quality (>80 points), we wanted to see if the popular wine producing climates produced better wine.

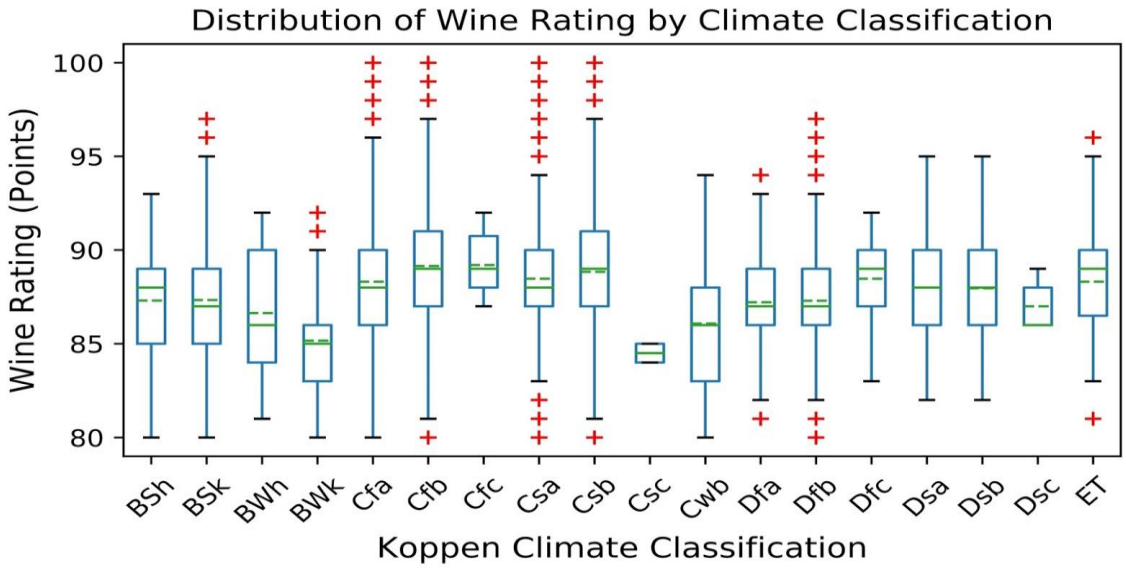

Figure 5

To further explore this phenomenon, we examined the distribution of wine ratings across each climate category, as seen in Figure 5. Notably, the four most overrepresented climates also produced wine with the highest maximum ratings, and some of the highest means and medians.

Our findings led us to explore where these wine regions are, and specifically, if there are any potential wine regions that are not being used for production. Figure 6 displays more or less what we'd expect from our life experience with wine, with climates Csb, Csa, and Cfb plotted in red, and analyzed wineries plotted in black.

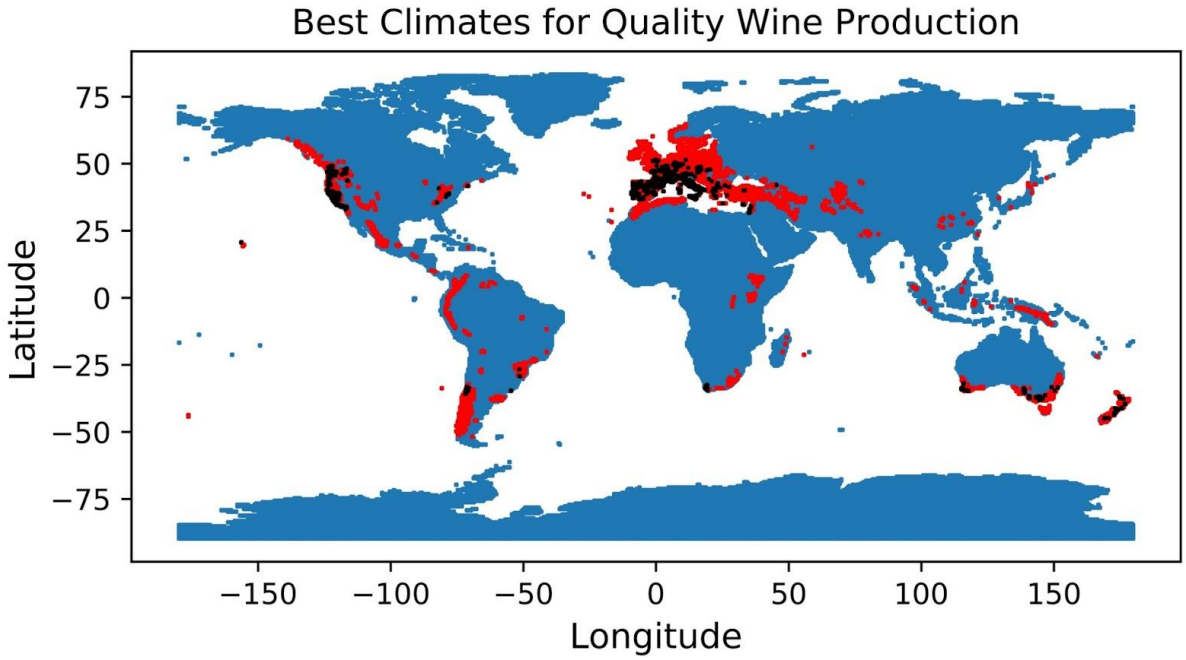

Figure 6

A few suitable wine regions are suspiciously winery sparse, including areas in and around Iran, Chile, northern Europe, and Central Africa. Some of this absence is easily attributable to cultural norms or historical events, such as the Iranian Revolution, when once famous Iranian wine stopped production. The low density of wine production in Chile is likely an artifact of our data manipulation: when locations were pulled from the Google Maps Geocoding API, they were pulled from the top result which was sometimes, in error, a wine distributor or winery headquarters. When validating this data, any inconsistencies with country or U.S. state were programmatically removed from the sample. Any bias in Google's algorithm causing a winery, rather than a vineyard, to be returned would then cause our data to underrepresent vineyards that export their grapes or use a western distributor.

Closing out this course of discovery, we wanted to illustrate where wine grapes could be grown, but where vineyards are proportionally underrepresented Figure 7.

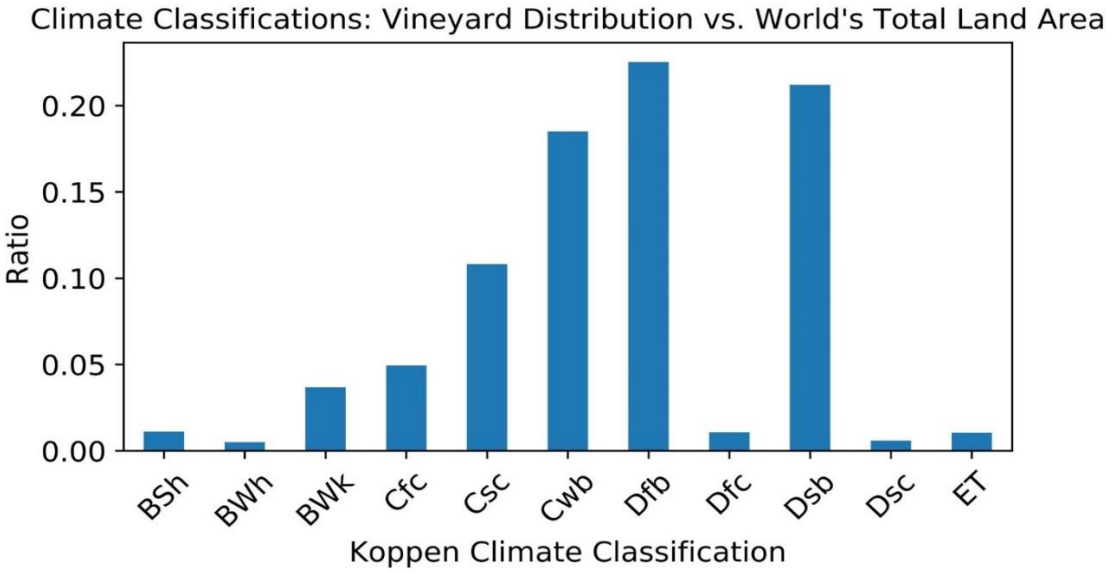

Figure 7

All but the least frequent climates are similarly displayed in map format in Figure 8.

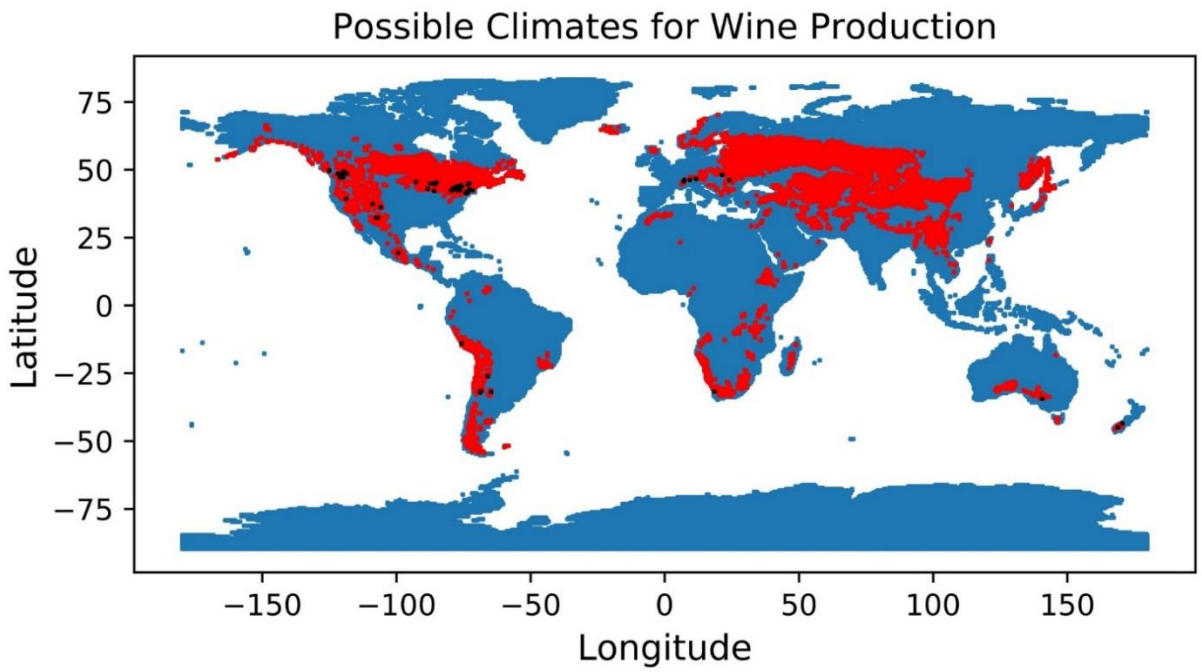

Figure 8

These climates are generally more inland than the best wine producing climates. Furthermore, all vineyards falling into these climates skirt the outsides of the regions, presumably near the climate border with the more popular climate designations. Finally, as we'll see in our further analysis, there are many variations in weather at the resolution a vintner would be interested in, that simply cannot be captured by the relatively broad Köppen classifications.

When we began to investigate the NOAA data for areas around Napa Valley in California, we were surprised to see the variations in annual rainfall, excessive heat, and freezes from locations that were not terribly distant – despite knowing that California has many microclimates. We decided to see just how much variation you could expect to see within the various Köppen classifications (Figure 9).

As expected, climates like Csb (includes the Mediterranean) have less variation than more temperate zone like Cfb (includes Germany). This finding helped validate both NOAA and the Köppen datasets. But even across the Csb region in the year 2000, there were wineries getting almost 0 rain while others received more than 2,000mm.

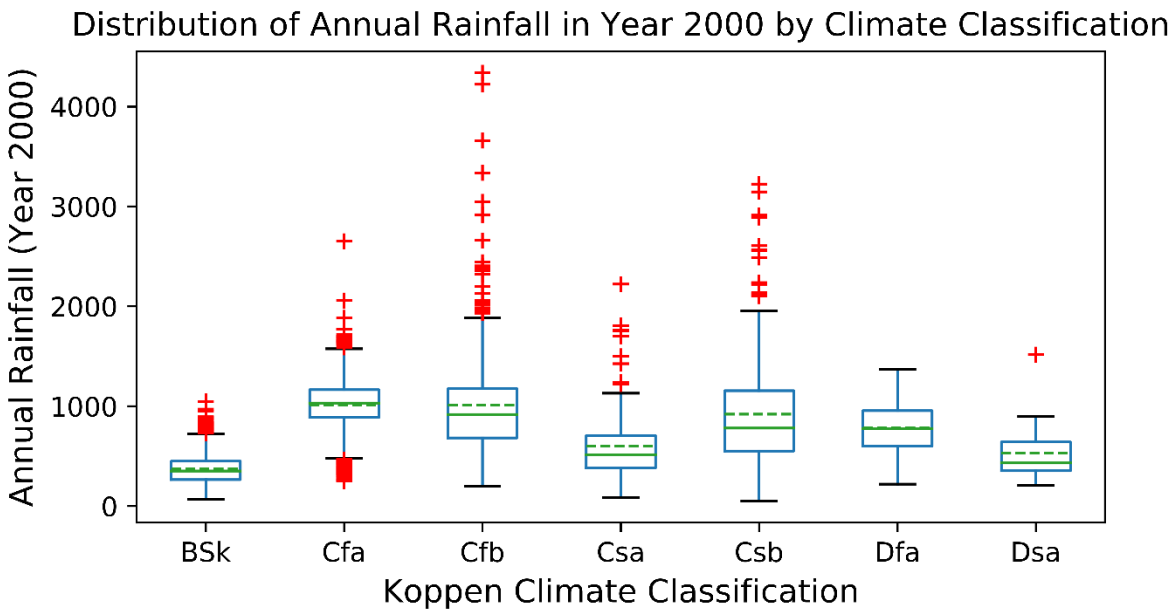

Figure 9

3. Does the weather during a given year impact a wine's price or ratings?

Given the weather variations within Köppen classifications, we were curious if we could locate correlations between these weather variations and prices or ratings of wine. To make the dataset manageable, we focused on the most common variety of wine, pinot noir. We then looked to see if some 11,000+ ratings for pinot noir showed any pattern based on rainfall near the winery during the year of production. There turned out to be less than 0.1 correlation. That held true across a number of other variables, including different wine types, specific regions, number of days of extreme heat, and number of days of freezes. The charts all would take a form similar to Figure 10.

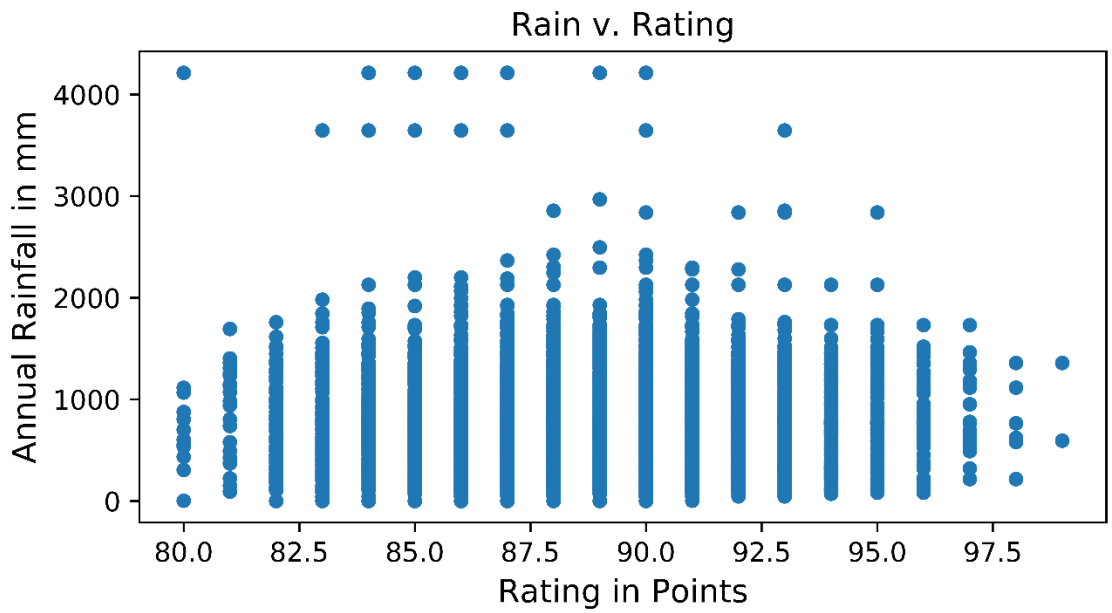

Figure 10

The Wine Enthusiast dataset is not helpful for determining this kind of information. Crucially, the magazine only publishes about wines that receive a rating of 80 or better, so we the dataset is highly biased toward wines with vintages that were successful. There are no failures in these wines to show strong impacts from weather events.

To try to get around this limitation, we focused on counting the numbers of rated wines that a region produced in the hopes of capturing data that would show the impact of a bad weather year. We focused on German Riesling wines since the wine is widely produced across Germany, and the country falls entirely within one Köppen classification. However, we discovered that the number of rated Rieslings from Germany grew rapidly starting in 2010 – not necessarily reflecting an increase in quality but rather an increase in popularity for the once-derided grape. We had one more trick up our sleeve – rather than count the number of rated wines, we could count the percent of rated Rieslings that Germany produced each year. That would eliminate the effects of Riesling mania and give us insight into the quality of the country's annual output.

By doing this, we did find a small 0.32 correlation between number of freezes and ratings. It appears having more days of freezing weather during the year is associated slightly with better rated wine – to a point. See Figure 11.

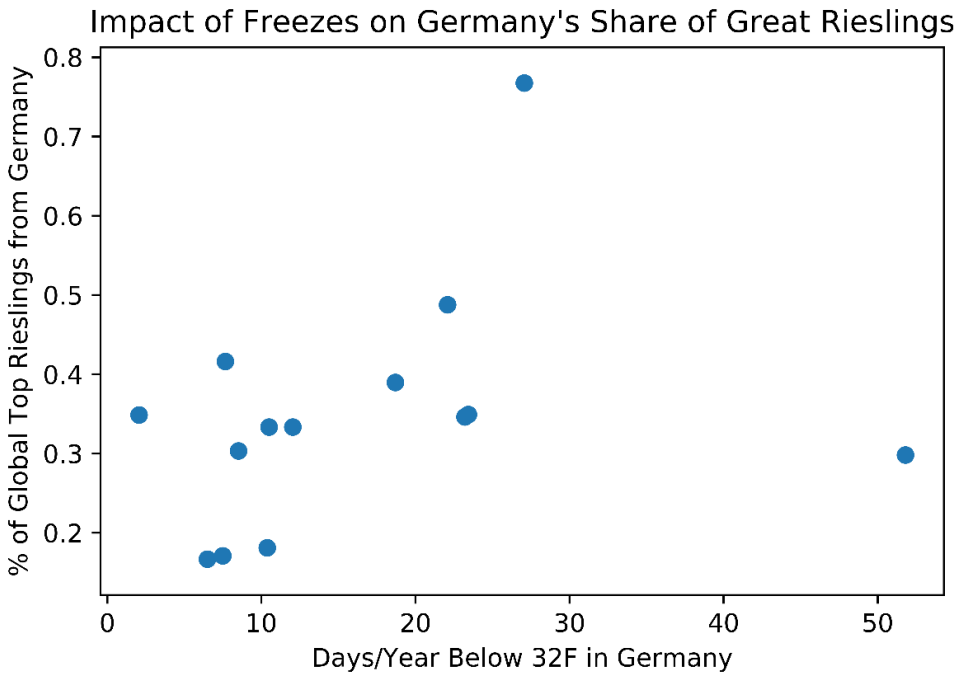

Figure 11

4. What region has the most expensive wine?

We started by looking at some summary level statistics of our data based on regions. What did the prices of wine look like from a global view? Given our large dataset, we were able to take a look at the price of every single bottle in the dataset and compare on a global scale as seen in Figure 12.

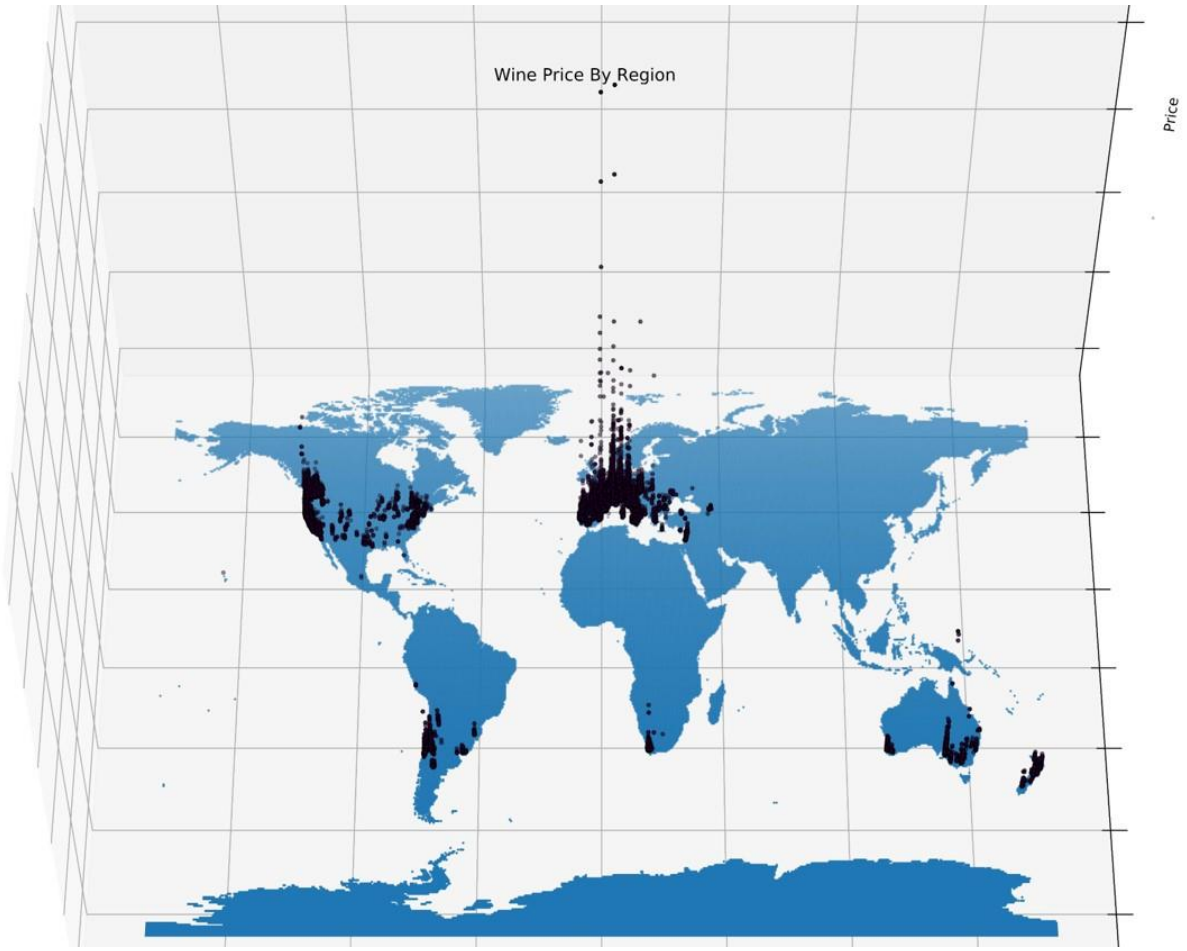

Figure 12

The black dots symbolize the region from where the bottles of wine are from. The prices line up vertically to show the price differential. Some countries that were represented in the dataset include US, Canada, Argentina, Chile, South Africa, Austria, Germany, Italy, France, and Australia. The prices of wine in Europe, on average, were higher than anywhere else in the world. The country with the highest average priced wine is France, followed by the US, Italy, Austria, Germany, and Portugal.

5. What region has the best rated wine?

We wanted to know which countries were producing the best rated wine. In our dataset, we looked at the 'points' field for each bottle of wine. Scores were represented with a points distribution that was greater than or equal to 95. We also took a look at the countries with the exclusive perfect rating of 100 from 'Wine Enthusiast'. As seen in Figure 13, the US had the most bottles with a rating of 95 or better, followed by France.

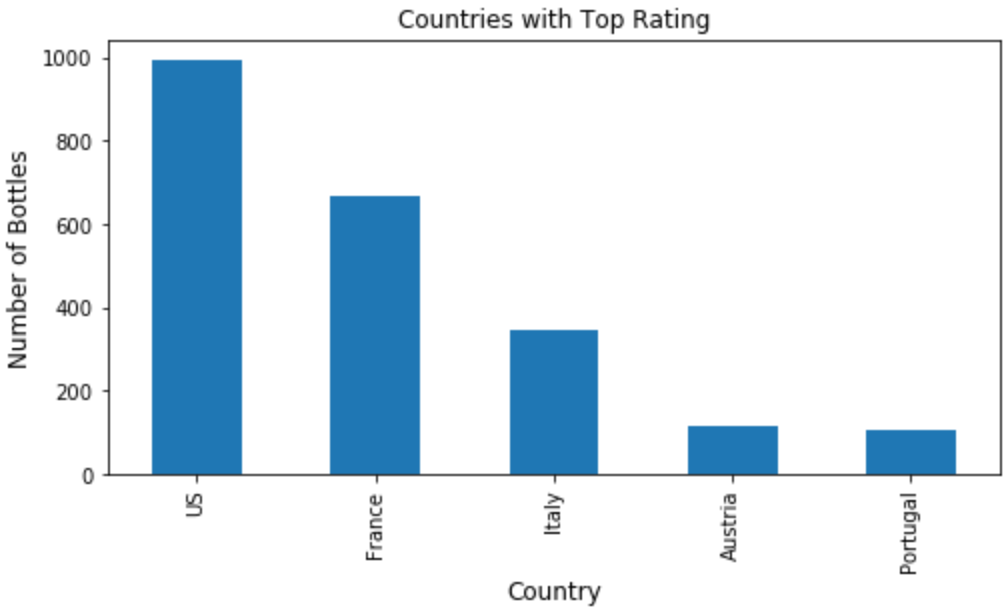

Figure 13

However, as depicted in Figure 14, France led all countries when it came to a perfect rating of 100.

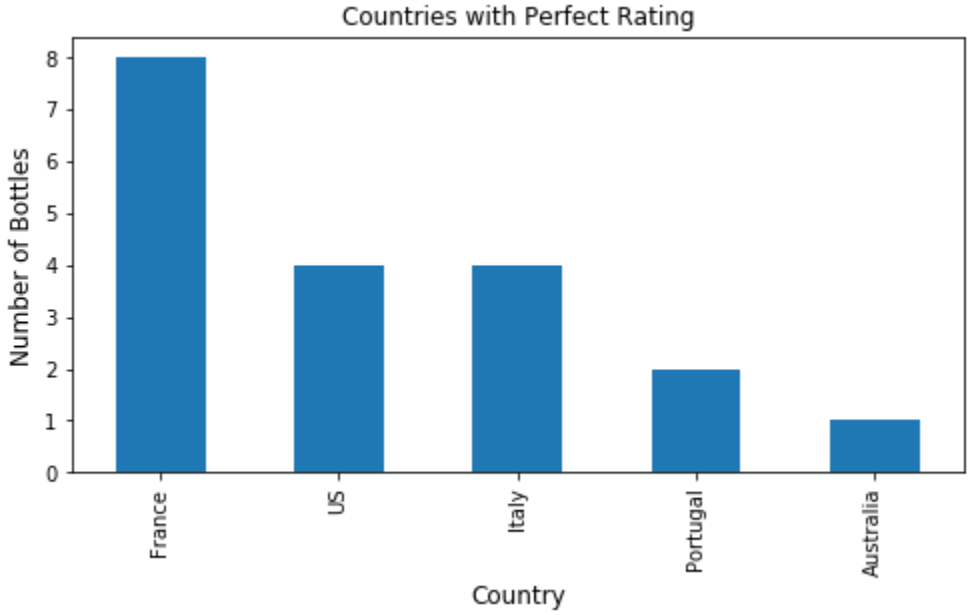

Figure 14

The United States had more wines within the top five possible ratings compared to other countries by a large margin. This is likely attributable to the fact that our dataset had more US wines compared to other nations. However, when we took a look at the perfect rating of 100, France had more bottles of wine with that perfect rating. This observation begs another look at ratings by country of origin.

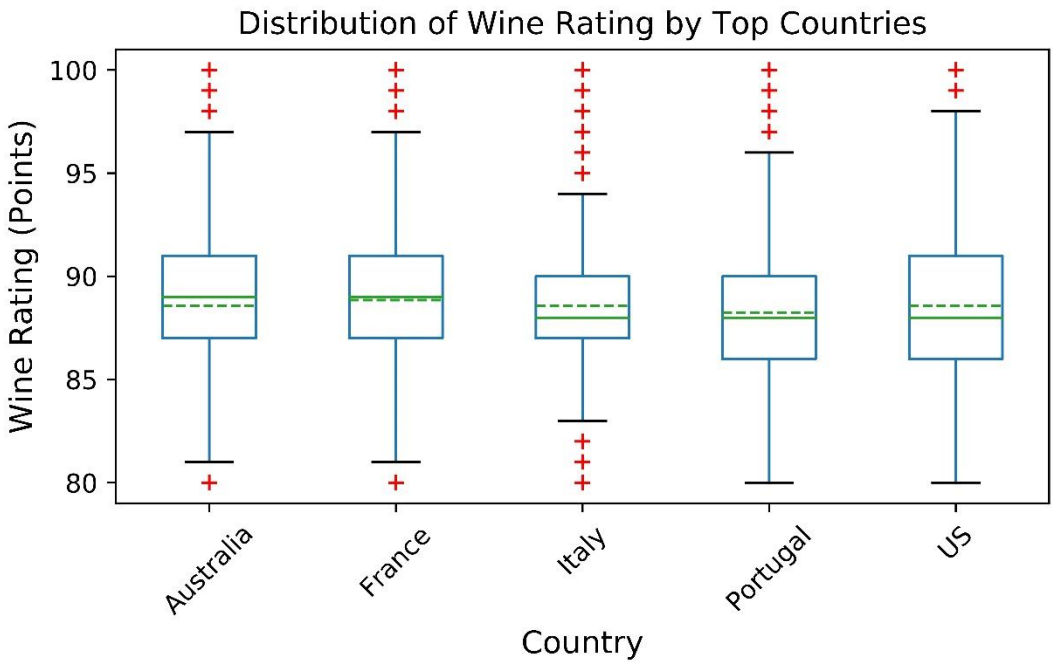

Figure 15

With a final look at the distribution of ratings by country in Figure 15, we find that the differences encountered in Figure 13 and Figure 14 were probably due to overrepresentation in our dataset, as no country's rating distribution particularly stands out from the crowd.

6. What types of wine are most prevalent by country?

Different countries generally have different wines that cater better to their region due to various attributes. Naturally, we wanted to explore the types of wines that were most prevalent in each country. We decided to plot the same five countries and the three most prevalent wine types that each country offers, as shown in Figure 16.

Within our dataset, US mainly produced Pinot Noir, Cabernet Sauvignon, and Chardonnay. France mainly produced Bordeaux-style Red Blend, Chardonnay, and Pinot Noir. Italy favored Red Blend, Nebbiolo, and Sangiovese. Portugal mainly produced Portuguese Red, Portuguese White, and Port. Finally, Austria came in with Gruner Veltliner, Riesling, and Blaufränkisch.

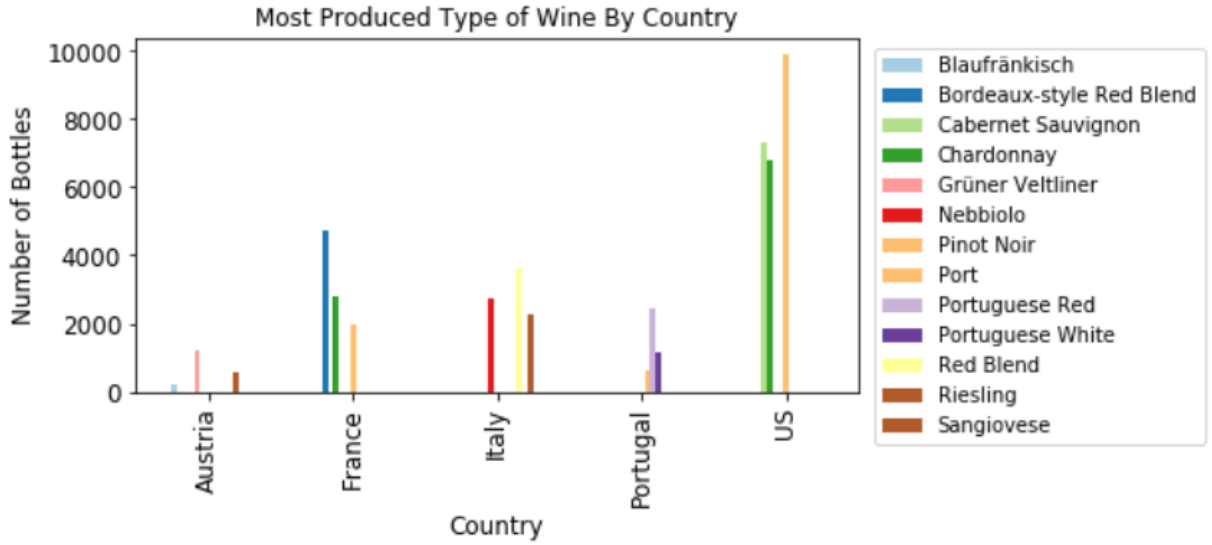

Figure 16

This graphic shows the most prevalent types of wines in each country. Generally, distinct types of wine require a different set of requirements during harvesting. Different tastes in wine types can also be attributable to the region as well. We were able to pull the 'description' of each wine to gain a better understanding of what embodies that specific type of wine. This is an important observation as we look to uncover why these regions specifically produce those types of wines. We will take a closer look at weather patterns of regions to see if there is a correlation based on our initial observations.

7. What variety of wine is rated best?

We will conclude with the question that might help you most at the grocery store: What variety of wine receives the best ratings? We took the mean of the ratings for each variety of wine and looked to see which were the top five and the bottom five. But most of the names would not be familiar to casual consumers. Instead, in Figure 17, we plotted box plots of the ratings earned by popular varieties of wine. You will see that the movie "Sideways" wasn't wrong – the sommeliers generally give Merlot lower reviews. Meanwhile, if you don't know anything about the wines on the rack, you could do worse than picking up a bottle of Syrah.

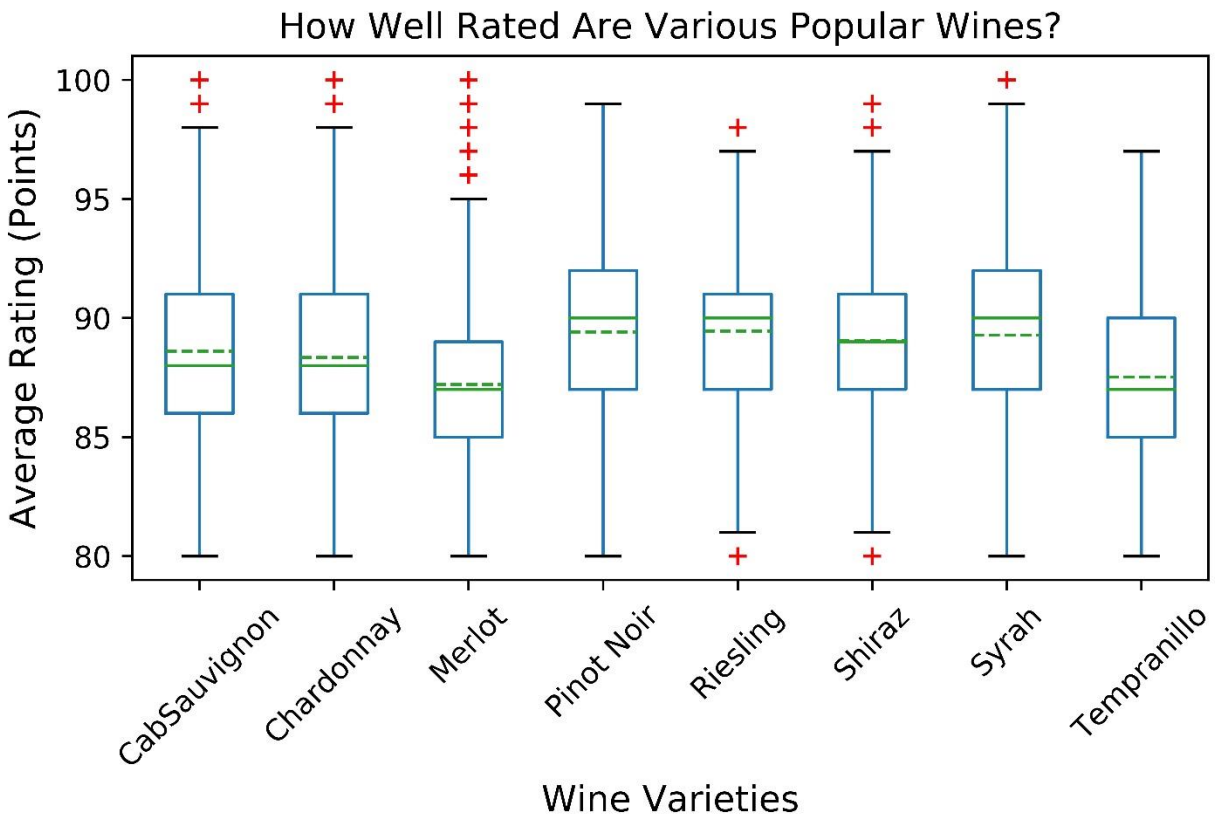

Figure 17

Happy imbibing!