Project #2 – Data Exploration and Design

Amy Hoffman
February 21, 2022

Introduction

All wine shoppers want to buy a good bottle of wine at a good price. But this is a more difficult challenge than one might anticipate due to the enormous number of options, the hundreds of grape varietals, and the geographic diversity of wines. To help consumers select a good bottle of wine, they need to know what is considered a bargain, where the wine comes from, and what countries are known for producing their grape varietal.

Data

The wine review data set was scraped from WineEnthusiast website and contains ten fields[1]. Table 1 provides an overview of the data by providing column descriptions, distribution information for quantitative variables, number of unique categories for categorical variables, and the percentage of values missing for all variables. Price and points are quantitative variables, and the remaining variables are all categorical.

Variable	Description[1]	Min	Max	Mean	St. Dev.	Number of Category Values	Percent Missing
country	The country that the wine is from	-	-	-	-	49	0.0%
description	A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.	-	-	-	-	97,821	0.0%
designation	The vineyard within the winery where the grapes that made the wine are from	-	-	-	-	30,622	0.0%
points	The number of points WineEnthusiast rates the wine on a scale of 1-100	80	100	87	3.2	-	0.0%
price	The cost for a bottle of the wine	$4.00	$2,300.00	$33.13	$36.32	-	9.1%
province	The province or state that the wine is from	-	-	-	-	456	0.0%
region_1	The wine growing area in a province or state	-	-	-	-	1237	16.6%
region_2	Sometimes there are more specific regions	-	-	-	-	19	59.6%

	specified within a wine growing area						
variety	The type of grapes used to make the wine	-	-	-	-	632	0.0%
winery	The winery that created the wine	-	-	-	-	14,810	0.0%

***Table 1**: Overview of the data contained in the winemag-data_first150k.csv file along with the column descriptions provided by zacktoutt, the creator of the data set.*

Importantly, the country, description, variety, and winery are all complete and not missing values. Understandably, region_1 and region_2 have the highest rate of missing variables because by law, winemakers cannot include a region_1 or region_2 on the wine label unless the wine meets specific regional standards.

Notice the distribution of price and points. Based on these descriptive statistics it is hypothesized that prince and points both have right skewed distributions where the maximums are outliers.

In the real world, wines are grouped into white and red wines. While this data set does not explicitly have red or white wine indicators, this can easily be deduced from the grape variety.

Visualizations

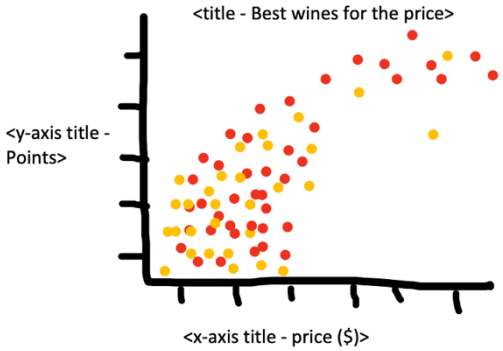
While this data set contains lots of text and categorical variables, there are still several important questions this data set can answer without entirely adventuring into the realm of text analysis. For example, this data set can provide insight into the following five questions:

1. What wines provide the most bang for the consumer's buck? Meaning what wines have higher ratings but lower prices?
2. What countries are the largest red, white, and overall wine producers?
3. What grape varieties are most popular by country?
4. How do points ratings change by grape variety?
5. How do prices change by grape variety and/or country?

Now, let's look at how to visualize the data to answer some of these questions.

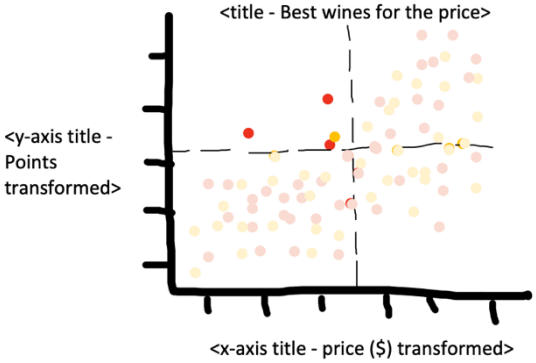
What wines provide the most bang for the consumer's buck?

The best place to start would be a simple scatter plot with price along one axis and rating points along the other, both with major tick marks and labels. Based on the hypothesis of the price and points distributions, it is expected most wine will exist in Quadrant III as shown in this example chart. The use of red and gold colors visually indicates if the point represents a white or red wine. This basic scatter plot will likely display trends that will inform changes to this visualization such as those shown in the next visualization.



If the hypothesis that price and points are right skewed, a log transformation would showcase a more linear trend as shown in this visualization. This then enables the creation of linear regression model to more clearly define the trend. This can be done for red, white, and all wines. Wines above the line are bargain wines for the price, and those below are not. Like the first visualization, this visualization leverages color to visually distinguish between red and white wines where black describes all wines. This visualization also introduces the use of saturation where points below the regression lines are lighter and those above are darker, which draws attention to the higher quality wines at each price point. Ideally, this visualization would work best with user interaction. The ability to select to whether to view either just red, just white, red and white, or all wines would reduce clutter and confusion.

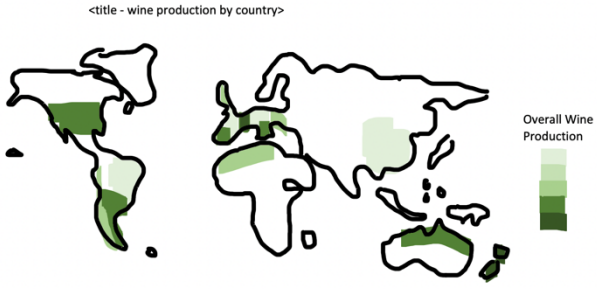
Lastly, if the hypothesis of the price and points distributions does not hold true, dividing the scatter plot into quadrants would provide a simple classification. This approach would also work on the log transformed data. The center of the quadrants would by default be the average price and average rating points. Like the second visualization, user interaction would enable the points to filter to just red, just white, or all wines and update the quadrants accordingly. The use of saturation plays an important role in drawing the eye to Quadrant II where the highest quality wines for the lowest prices exist.



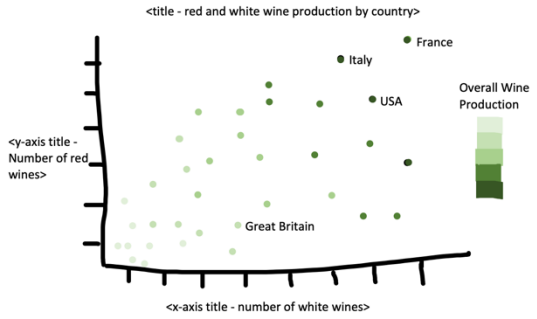
In each of these visualizations, the ability to hover or click on a point and see details about the wine is important for the user to learn which wines provide the biggest bang for their buck. The best visualization approach for this question relies in part on the analysis technique. The second visualization enables the user to identify the highest rated wine at all price points, whereas the third visualization only recommends the above average rated and below average priced wines.

What countries are the largest red, white, and overall wine producers?

One way to visualize this is by using a map, saturation, and a legend to define the number of wines produced. User interaction would be key in order to switch between seeing overall, just red, or just white production. Hover or click capabilities are necessary to view details about how many wines were produced. However, this visualization falls short for large producing smaller countries just as Portugal. Yet, this visualization excels in providing the

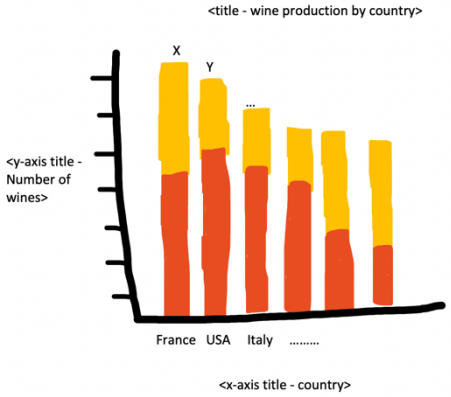


opportunity to view geographical patterns. For example, most wine is produced between 30 to 50 degrees longitude.



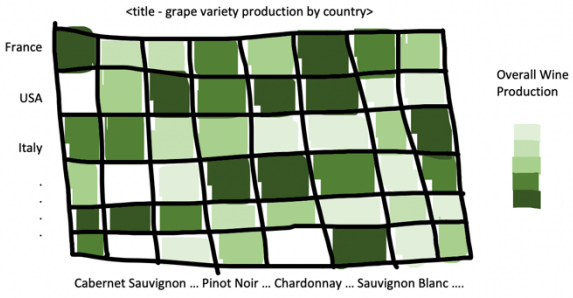
An alternative way to view this would be a scatter plot where the number of red wines is on one axis and the number of white wines is on the other, where the axes have major tick marks and labels. Saturation and a legend further help define overall wine production. One could use size rather than saturation but using both could create too much redundancy and cause clutter. Labeling some of the top, bottom, and well-known producing countries allows the user to quickly identify largest producers without having to rely on comparing hover text using short term memory. Unlike the map, the scatter plot allows the user to easily compare a country's white and red wine production.

Lastly, a bar chart sorted from greatest to least producing countries could answer this question. Unlike the prior to visualizations, the bar chart allows the user to easily compare overall wine production and quickly identify any number of the top or bottom overall wine producing countries. In the instance of a stacked bar chart, comparing red and white wine production across countries is more difficult and requires more short-term memory use. The addition of user interaction could remove the need for a stacked bar chart by simply allowing the user to select if they wish to view red, white, or overall wine production and updating the chart accordingly. In either case, the text at the top of the bars removes the need to compare the height of each bar against the y-axis labels.

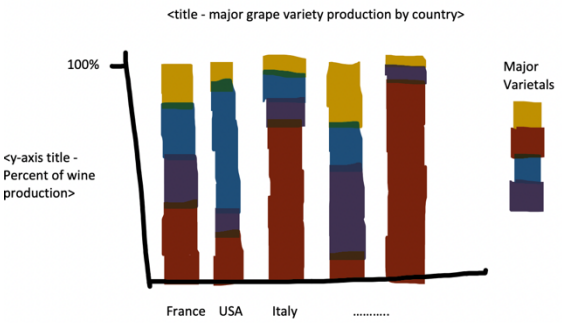


What grape varieties are most popular by country?

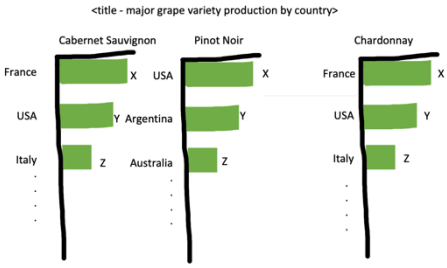
Since answering this question requires the use two categorical and one quantitative variable, a heatmap would work well. This heat map shows primary varietals along one side and countries along the other. The countries would be grouped by region or in order of greatest producers. The varietals would be grouped by black or white grapes or by order of popularity. The saturation and key indicate the number of wines made from each varietal. This heat map would also scale well to include all 632 varietals and 49 countries. The heat map works well to distinguish between popular and not popular grape varieties but relies on short term memory and hover text to identify the single most popular varieties.



A scaled stacked bar chart simplifies the task of identifying the single most popular variety by country but does not scale as easily to view all varietals grown in a country. Therefore, this view would be best to show for the 6-9 primary varietals. The bars are all the same height where the length of each section is equal to the percentage of wines from each country made with each varietal. The color scale would need to be discrete since each color describes a different varietal. However, the varietals could be grouped within the bar such that the black and white grapes are together. Like the heat map, the countries could be grouped by region or ordered by wine production. Adding hover text or click interaction would provide the additional details for the user to make comparisons. Like the heat map. However, this requires the user to exercise short term memory to make comparisons across countries or even varietals grown in the same country.



One way to counteract this is to create individual bar charts, creating a row or matrix of bar charts. Each bar chart is individually sorted from largest to smallest production and contains labels for easier comparison. This format would work well as a drill down for the two prior to visualizations, were a chart displays a single varietal. This scales decently for the number of countries but would not scale well to display all varietals. Therefore, this visualization would need to serve as part of a drill down interaction or only display the primary varietals.



Conclusion

Hopefully consumers can explore and understand global wine production through these simple visualizations. These visualizations provide the consumer with aggregated and detailed information about prices and wine production. The more the consumer understands, the more confident they can be in selecting new wines they enjoy.

Sources

Zackthoutt. Nov 27, 2107. Wine Reviews. Version 4. Web. https://www.kaggle.com/zynicide/wine-reviews