Project #3 – Interactive Visualization Using Tableau

Amy Hoffman
March 8, 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 and which tasters they can trust to provide accurate ratings.

Data

The wine review data set was scraped from WineEnthusiast website. The secondary data set contains thirteen 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	Type	Description[1]	Min	Max	Mean	St. Dev.	Number of Category Values	Percent Missing
country	categorical	The country that the wine is from	-	-	-	-	49	0.0%
description	categorical	A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.	-	-	-	-	97,821	0.0%
designation	categorical	The vineyard within the winery where the grapes that made the wine are from	-	-	-	-	30,622	0.0%
points	quantitative	The number of points WineEnthusiast rates the wine on a scale of 1-100	80	100	87	3.2	-	0.0%
price	quantitative	The cost for a bottle of the wine	$4.00	$2,300.00	$33.13	$36.32	-	9.1%

province	categorical	The province or state that the wine is from	-	-	-	-	456	0.0%
region_1	categorical	The wine growing area in a province or state	-	-	-	-	1237	16.6%
region_2	categorical	Sometimes there are more specific regions specified within a wine growing area	-	-	-	-	19	59.6%
taster_name	categorical	The first and last name of the taster who rated the wine	-	-	-	-	20	20.2%
taster_ twitter_ handle	categorical	The Twitter hand of the taster	-	-	-	-	16	24.2%
title	categorical	The title/name of the wine	-	-	-	-	118,840	0.0%
variety	categorical	The type of grapes used to make the wine	-	-	-	-	632	0.0%
winery	categorical	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, title, 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

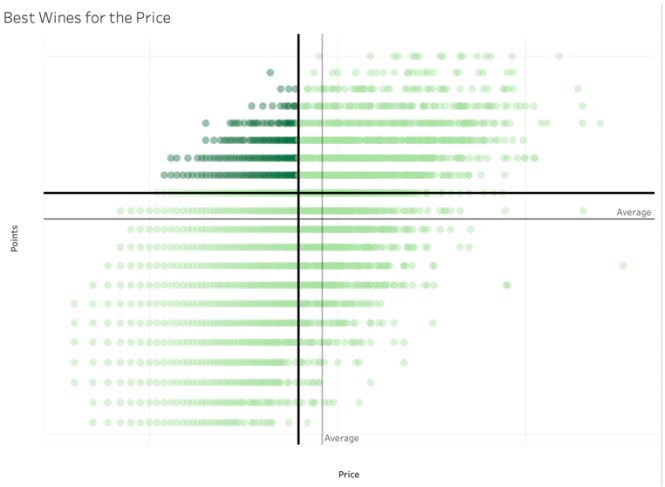
While this data set contains lots of text and categorical variables, there are still several important questions this data set can answer without entirely adventuring into the realm of text analysis. For example, this data set can provide insight into the following five questions:

1. What wines provide the most bang for the consumer's buck? Meaning what wines have higher ratings but lower prices?
2. Which tasters are most strict and least strict?
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.

Design



As discussed in the prior assignment, an easy way to identify cheap yet highly rated wines is to compare price against rating. In this visualization the log of price is plotted against the log of points for reasons discussed in the data section above. From this view with the average price and points reference lines the user can quickly identify quadrant III where all the higher-than-average rated wines exist at a lower-than-average cost. However, the user may be willing to spend more or want a wine with a higher rating; therefore, the visualization allows the user to move the crosshairs upon click. This updates the visualization such that the selected points are dark green and the non-selected are a lighter color. This makes it obvious which wines the user should be looking at.

List of Wines

Contains only wines selected in chart above

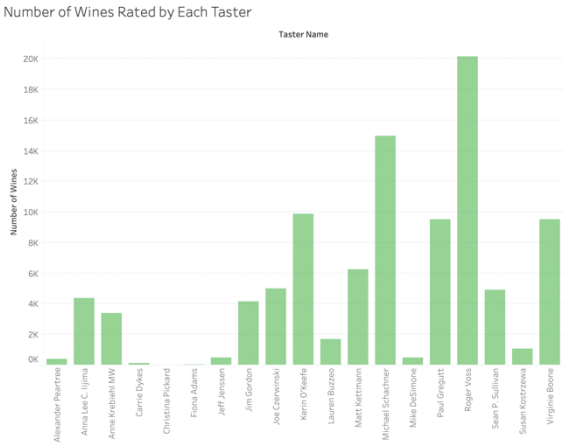
Title	Variet..	Countr..	Points	
Anglim 2013 Hasting..	Mourvèdre	US	93	$38.00
Anglim 2013 St. Pete..	Syrah	US	93	$42.00
Anna Maria Abbona ..	Nebbiolo	Italy	93	$48.00
Anthill 2005 Tina Ma..	Pinot Noir	US	93	$43.00
Antica 2007 Antinori ..	Bordeaux-s..	US	94	$55.00
Antichi Vigneti di Ca..	Nebbiolo	Italy	93	$55.00
Antico Colle 2010 Vi..	Sangiovese	Italy	93	$22.00
Antigal 2010 Aduent..	Red Blend	Argent..	93	$28.00
Antiquum Farm 2014..	Pinot Noir	US	93	$55.00
AntoLin Cellars 2010 ..	Viognier	US	93	$19.00
Anton Bauer 2011 Re..	Red Blend	Austria	93	$30.00
Anton Bauer 2012 Re..	Pinot Noir	Austria	96	$50.00
Anton Bauer 2013 Re..	Pinot Noir	Austria	94	$50.00

0 50 100

Avg. Price

However, multiple wines can exist at each point in the scatter plot above. Thus, the table on the dashboard automatically filters to only the wines selected in the scatter plot. The table provides all the basic information such as the title of the wine, country, variety type, average points, and average price. While points do not drastically change, the price does. To account for this, the table shows the price as a horizontal bar chart so the user can quickly distinguish between a more expensive and a cheaper bottle of wine. Further, the user can sort the table by any of the five columns to aid in their search.

Moreover, an integral part of identifying a good wine for a good price is trusting the ratings of the wine. Like everyone else, tasters' preferences influence their ratings. It would help most to know what those preferences are; however, that data is not contained in the data set. Therefore, we must rely on a taster's familiarity with a type of wine and their history of ratings. The bar chart provides a clean visualization of the number of wines the taster rated. At first glance it is easy to determine which tasters have tasted the most wines and which ones have tasted the fewest.



Below the bar chart is a box and whisker plot grouped by taster. The box and whisker plot visually shows the range of ratings given by a taster. Most tasters rate wines from the lower bound of 80 up to the upper bound of 100. The tasters who have not tasted many wines tend to have smaller ranges. Despite this, there is still a small variation in the average ratings by taster. Filtering this visualization results in even larger disparities between the ratings by each taster.

All four of these charts are filterable by country, grape variety, taster, price, and points. While some users are specifically interested in minimizing price or maximizing points, others are more interested in finding a good wine of a specific grape variety or from a specific location. The combination of these five filters enables the users to identify wines regardless of their intent.

Discussion

The visualizations in the dashboard answer the first two technical questions posed above:

1. What wines provide the most bang for the consumer's buck? Meaning what wines have higher ratings but lower prices?
2. Which tasters are most strict and least strict?

The scatter plot paired with the table enables the user to quickly identify the wines that are above a selected rating but below the selected price. A difference in saturation indicates the difference between selected and non-selected points on the scatter plot. The table provides detailed information about the selected points. The filtering capabilities allows the user to further narrow their search. For example, the user can say they only want to see wines that have a rating of 90 or above. From there, they can identify which wines have a lower-than-average price for wines with a rating more than 90 points in the scatter plot and view detailed information about these wines in the table.

The bar chart and box and whisker plots collectively help answer the second question. A taster with a large range of ratings and a large volume may have a more sensitive palette and rate wines accordingly. Whereas a taster with a small range and large volume likely has a similar opinion of all the wines. More analysis would be needed to determine if this is true or if the range of ratings is simply from tasting more wines that legitimately vary in price. Moreover, if a taster rates primarily wines of the same variety, the user can likely have more confidence in the taster's ratings. For example, if the user enjoys Malbecs, they can filter to see which tasters commonly rate Malbecs and see which tasters are most strict on their ratings. This enables the user to then select a higher rated Malbec wine by a taster with more experience in tasting Malbecs.

Conclusion

Hopefully consumers can explore and understand the relationship between price and points through these simple visualizations in order to select an enjoyable wine. Further, the hopefully the user can find confidence in selected a wine rated by an experienced taster. 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