What Makes the Best Rated Wine?

Congyi Fan

September 2020

1. Introduction

For fellow wine lovers out there, when browsing in a wine shop and seeing the sign that has a 98 point next to it - aren't you more inclined to buy it compared to the bottle next to it that says 88 point?

In this project, I want to explore what makes the highest rated wine and one can use the model to predict the point of a bottle given its region, variety, price and year.

This project is intended for anyone who's interested in wine tasting, and how 'experts' in the industry tend to rate wines (any potential bias such as higher priced wine getting higher ratings), and can use the model to predict an unknown bottle.

2. Data

2.1 Data source

Here I will be using the winemag-data-130k-v2.csv dataset from Kaggle. Columns in the dataset:

country: the country that the wine is from

description: a long description/commentary of the wine

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 (mostly >=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 (ie Rutherford inside the Napa Valley), but this value can sometimes be blank

taster_name: name of the taster

taster_twitter_handle: twitter handle of taster

title: name of the wine

variety: The type of grapes used to make the wine (i.e. Pinot Noir)

winery: winery the wine is from

2.2 Data cleaning

There are missing values in some column, so I took out the rows with missing values because it does not make sense to fill with say, most frequent value for categorical value in the data set; the data set itself does not have year value, so I extracted it from the title of the wine and got rid of the rare/inaccurate values; lastly, most of the data come from US, French, and Italian wines, so I focused on more granular level feature region_1 to perform the analysis. I did not choose region_2 because there are more missing values in this column.

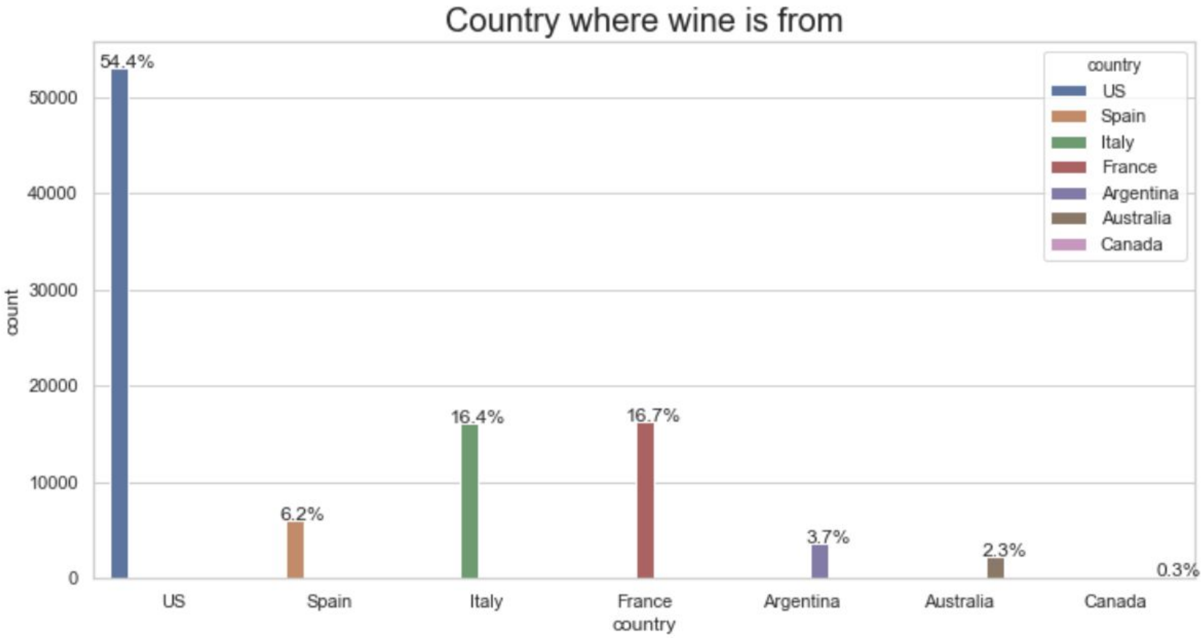
2.3 Feature selection

I picked price, region_1 and year for the list of features above, because description is a long description of what the wine is and cannot be easily utilized in a regression model; designation & region_2 have a lot of missing values ; taster info is not super relevant and also has missing values; title & winery are also too specific.

3. Exploratory Data Analysis

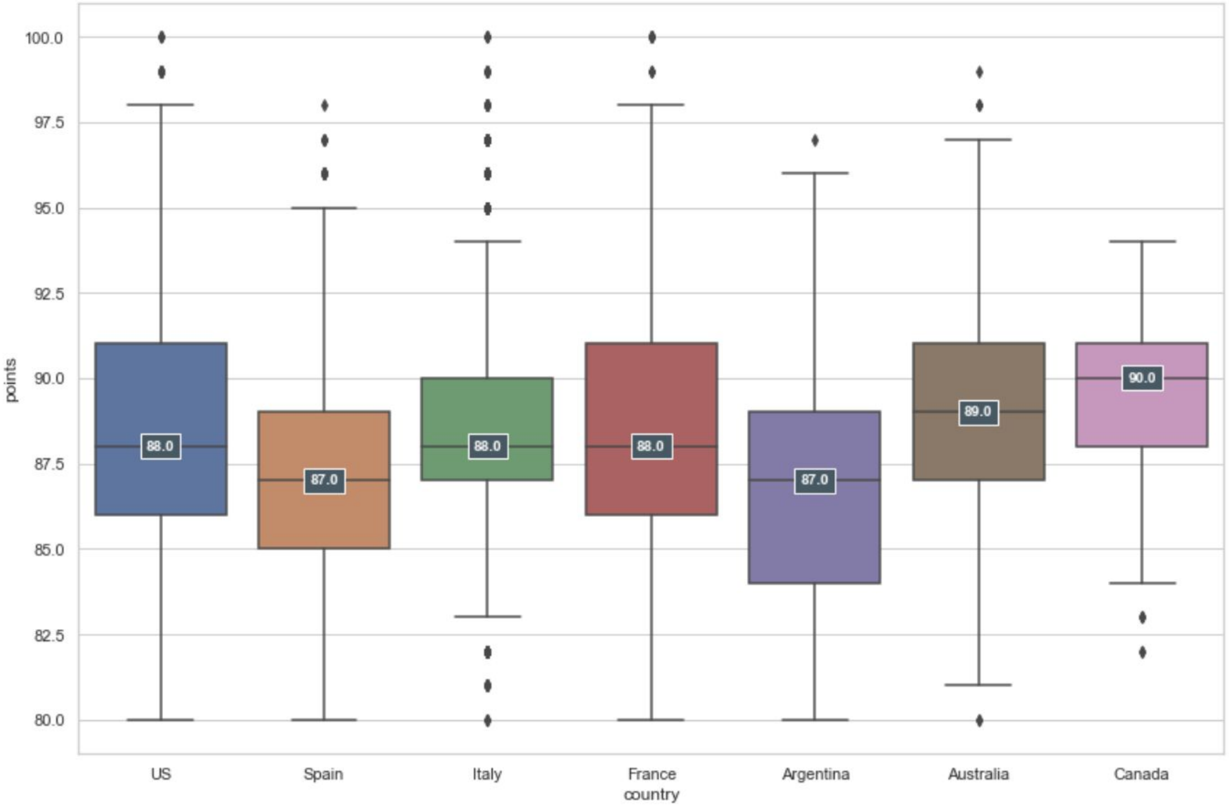
3.1 Wine score vs. country

Looks like most of the data come from US, French, and Italian wines - we should keep that in mind when we think about bias in the data.

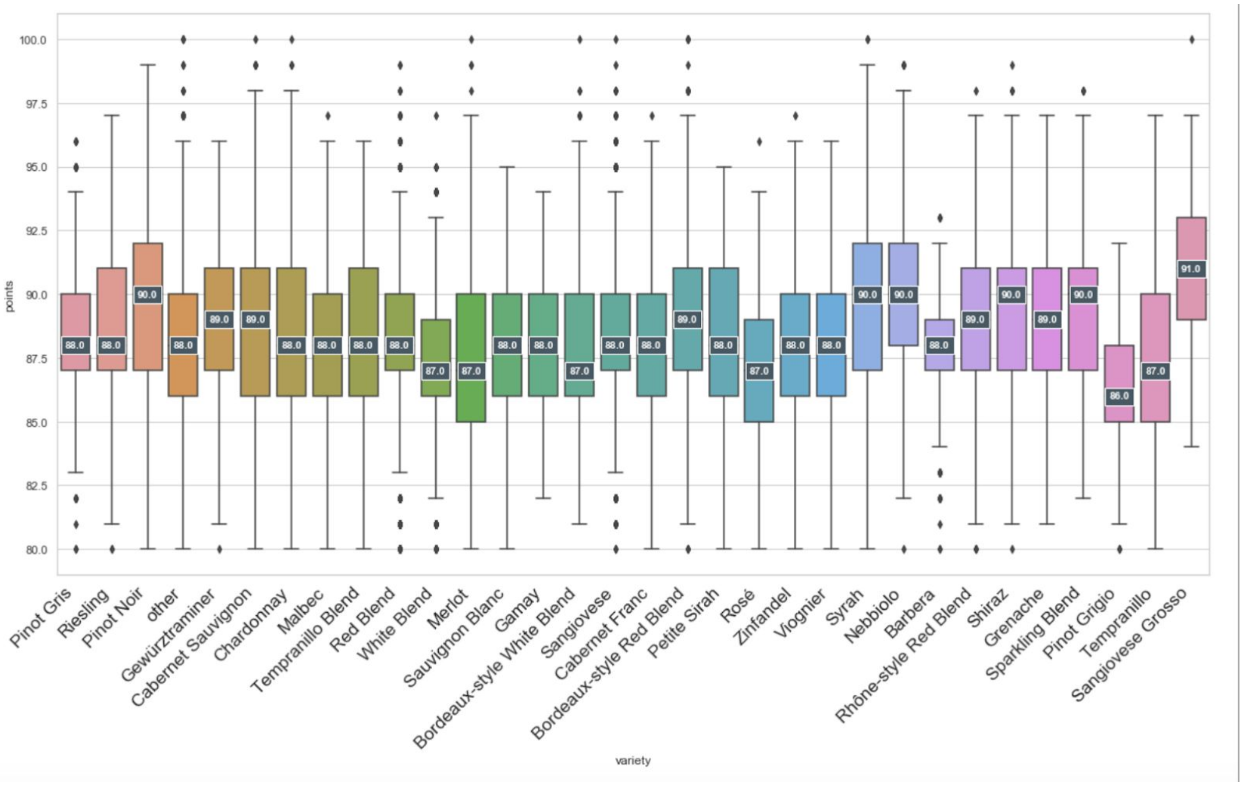


Plotting the points of wines based on country

1. US, French and Italian wine have the same median points, and have outliers that have 100-point perfect scores
2. Italian wine has the smallest IQR while both US and French wine have the biggest
3. Australian & Canadian wine do not have large amount of data but have better median points than other major countries

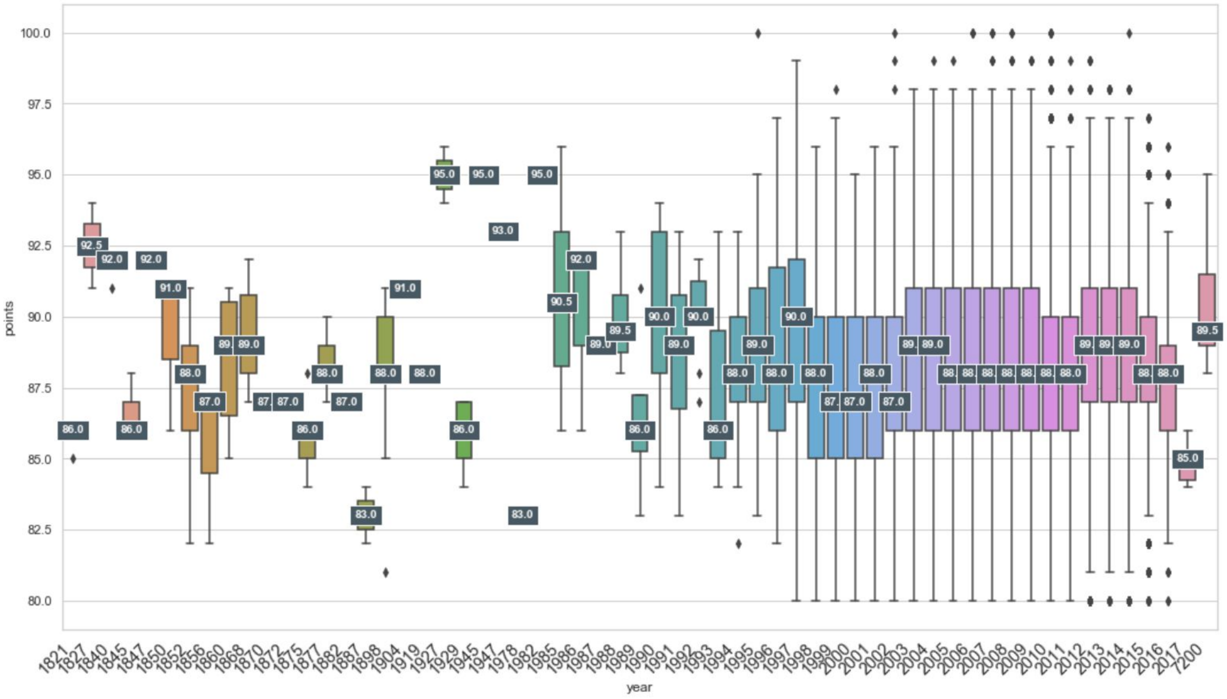


3.2 Wine score vs. variety



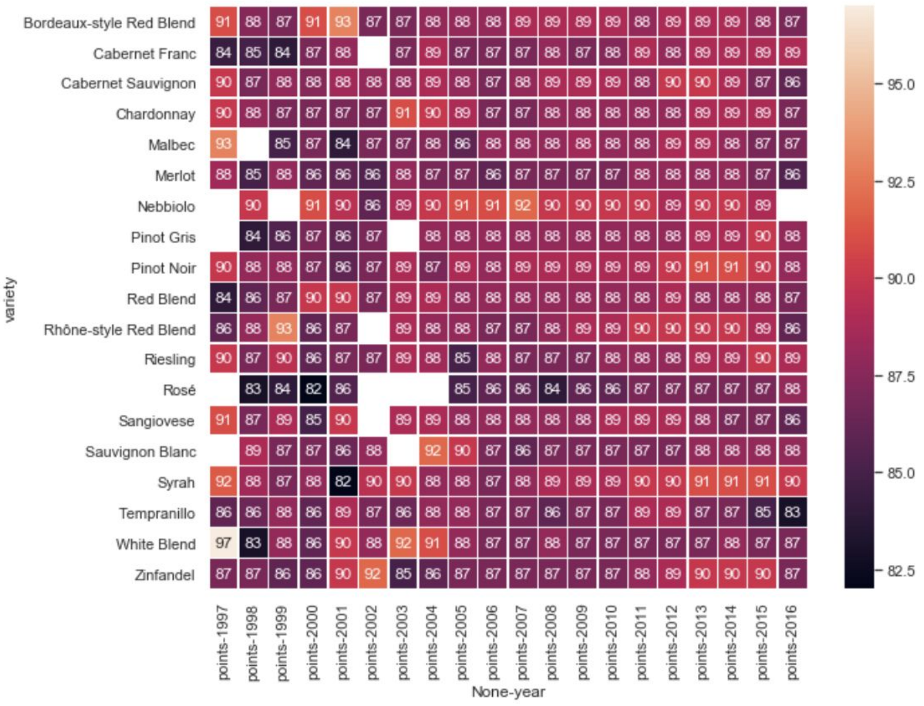
Here we can see Pinot Noir, Syrah, Nebbiolo, Shiraz & Sangiovese Grosso have the best median points.

3.3 Wine score vs. year



Here we can see the year feature doesn't seem to have a huge impact for recent wines (1994-2017), but points of vintage wines may vary a lot.

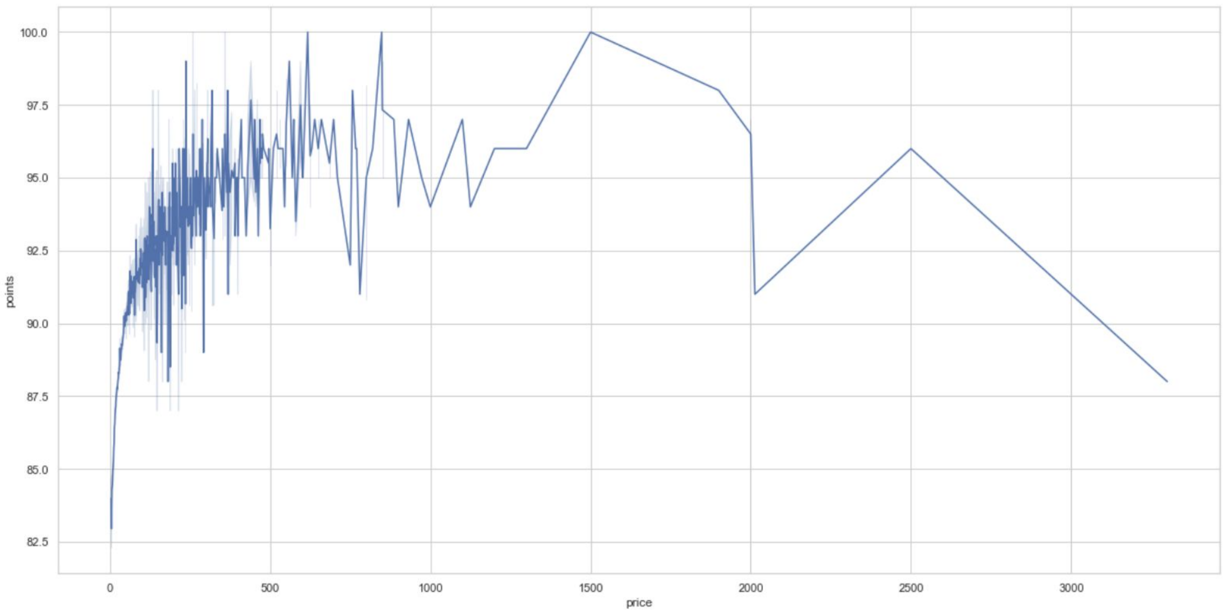
3.4 Wine score vs. year & variety



variety	points-1997	points-1998	points-1999	points-2000	points-2001	points-2002	points-2003	points-2004	points-2005	points-2006	points-2007	points-2008	points-2009	points-2010	points-2011	points-2012	points-2013	points-2014	points-2015	points-2016
Bordeaux-style Red Blend	91	88	87	91	93	87	87	88	88	88	89	89	89	89	88	89	89	89	88	87
Cabernet Franc	84	85	84	87	88		87	89	87	87	87	88	87	88	89	88	89	89	89	89
Cabernet Sauvignon	90	87	88	88	88	88	88	89	88	87	88	89	89	89	88	90	90	89	87	86
Chardonnay	90	88	87	87	87	87	91	90	89	87	87	88	88	88	88	88	89	89	89	87
Malbec	93		85	87	84	87	87	88	86	88	88	88	88	88	88	89	89	88	87	87
Merlot	88	85	88	86	86	86	88	87	87	86	87	87	87	87	88	88	88	88	87	86
Nebbiolo		90		91	90	86	89	90	91	91	92	90	90	90	90	89	90	90	89	
Pinot Gris		84	86	87	86	87		88	88	88	88	88	88	88	88	88	89	89	90	88
Pinot Noir	90	88	88	87	86	87	89	87	89	88	89	89	89	89	89	90	91	91	90	88
Red Blend	84	86	87	90	90	87	89	89	88	88	88	88	88	88	88	89	88	88	88	87
Rhône-style Red Blend	86	88	93	86	87		89	88	88	87	87	88	89	89	90	90	90	90	89	86
Riesling	90	87	90	86	87	87	89	88	85	88	87	87	87	88	88	88	89	89	90	89
Rosé		83	84	82	86				85	86	86	84	86	86	87	87	87	87	87	88
Sangiovese	91	87	89	85	90		89	89	88	88	88	88	88	89	89	89	88	87	87	86
Sauvignon Blanc		89	87	87	86	88		92	90	87	86	87	87	87	87	87	88	88	88	88
Syrah	92	88	87	88	82	90	90	88	88	87	88	89	89	89	90	90	91	91	91	90
Tempranillo	86	86	88	86	89	87	86	88	88	87	87	86	87	87	89	89	87	87	85	83
White Blend	97	83	88	86	90	88	92	91	88	87	87	88	87	87	87	87	87	88	87	87
Zinfandel	87	87	86	86	90	92	85	86	87	87	87	87	87	87	88	89	90	90	90	87

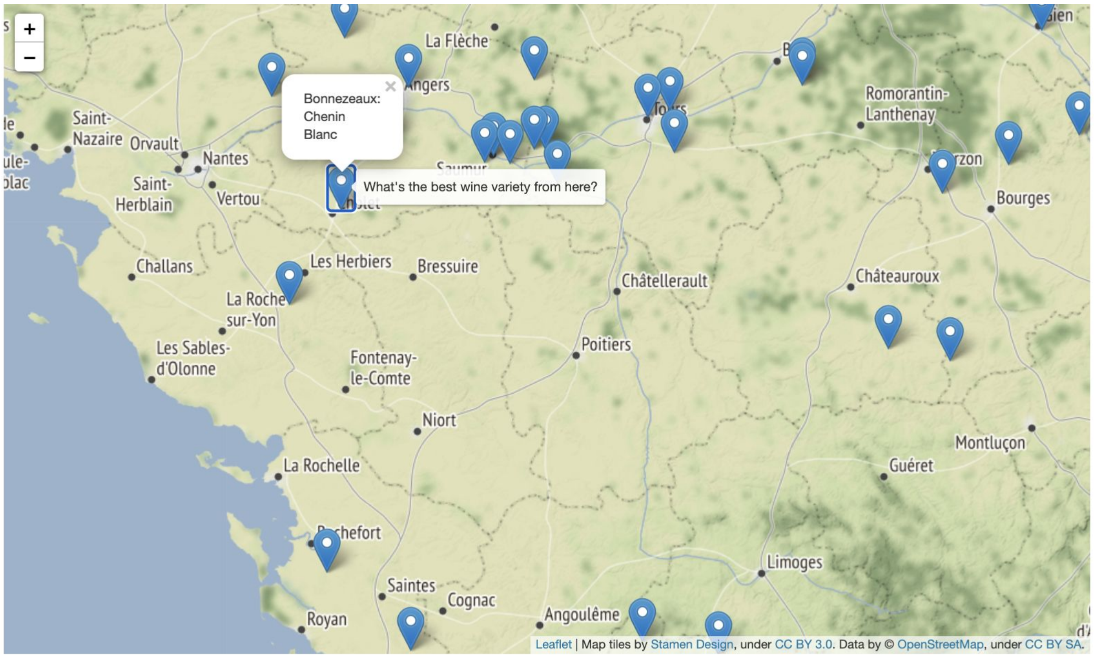
Seems like year combined with variety still is a bit arbitrary, but to our surprise, 1998-2005 has more wine varieties with low score vs 2006-2016.

3.5 Wine score vs. price



Again - quite interesting that outliers included we see what's close to a downward parabola; there's a general tendency of increasing of points when price increases up til around $500, then the deviation also is very drastic onwards, this might suggest that the price feature when combined with others may be a better indicator of points.

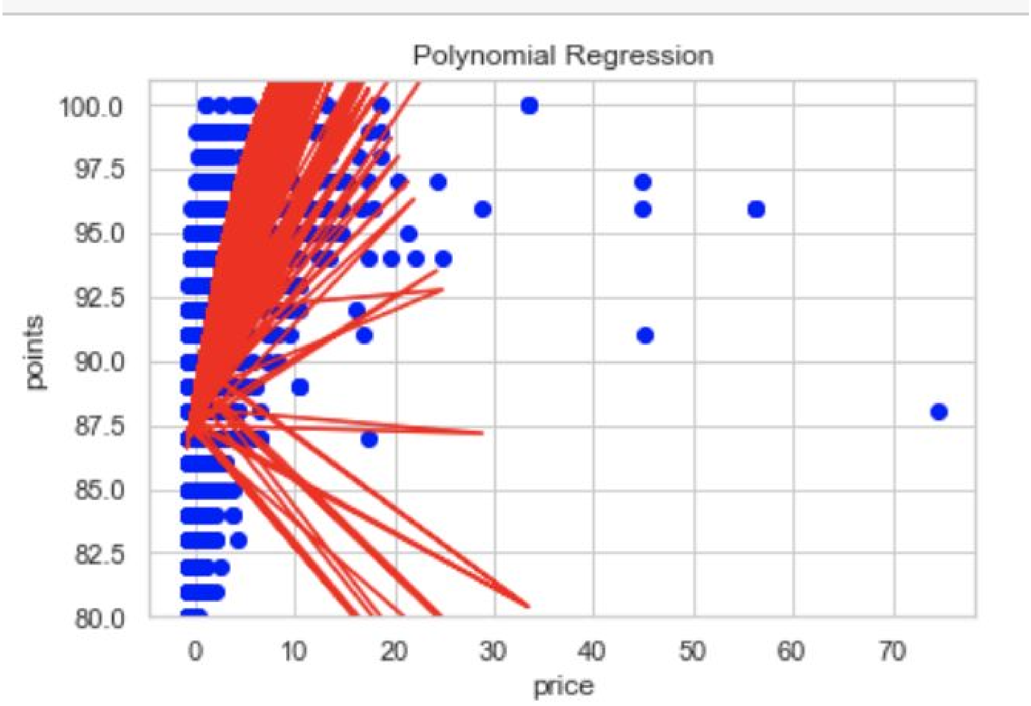
I also made an interactive map using geopy and folium that you can view the best variety (based on median score) of wine coming from a specific region.



4. Predictive Modeling

4.1 Single Regression

First I wanted to try to explore a single regression model w./ price vs. points; from the visualization above I knew I'd probably need a polynomial regression, so I tried that with a degree of 3. The result is a model with R^2 of around .28 for both training and test data.



4.2 Multiple Regression

Then I used the same degree for multiple features - price, variety and region. The result is a model with R^2 of .42 with training set and .34 in testing set.

4.3 MLP Regressor

Finally I set up a multi-layer perceptron regressor to predict points on variety, price & region and the result is pretty decent for training data of .44 but has a lower score for testing data of .42, which means approximately 42% of the variation of points is explained by this model.

5. Conclusion

This project is an attempt to uncover how wines are getting scored based on its price, region and year. I learned that price nor year by itself is not necessarily a definite indicator of good wine. There are certainly a lot more to what makes a good wine than just these features. And wine score does not necessarily dictate wine being good/bad either. It may also depend on when you open a particular bottle and how you taste it. Overall, there are many subjectivities in how good a wine bottle is, but this model can be used as a starting point for beginners to wine to learn about wine is historically scored based on region, price, year.