What Makes the Best Rated Wine?

Congyi Fan, Sep 2020

> "Wine is one of the most civilized things in the world and one of the most natural things of the world that has been brought to the greatest perfection, and it offers a greater range for enjoyment and appreciation than, possibly, any other purely sensory thing. "
>
> -Ernest Hemingway

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?

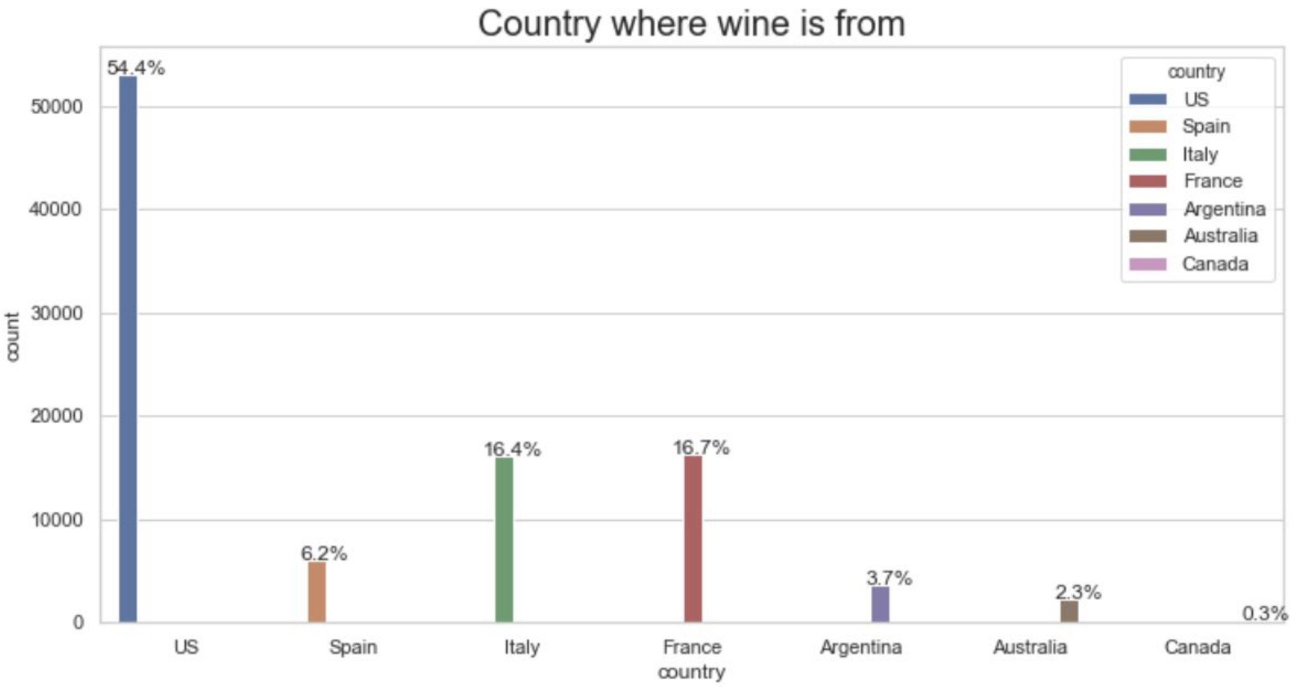
Predicting wine scores - Assumptions

- higher priced bottles get higher points
- older bottles get higher points
- French and Italian wines have highest points

Data

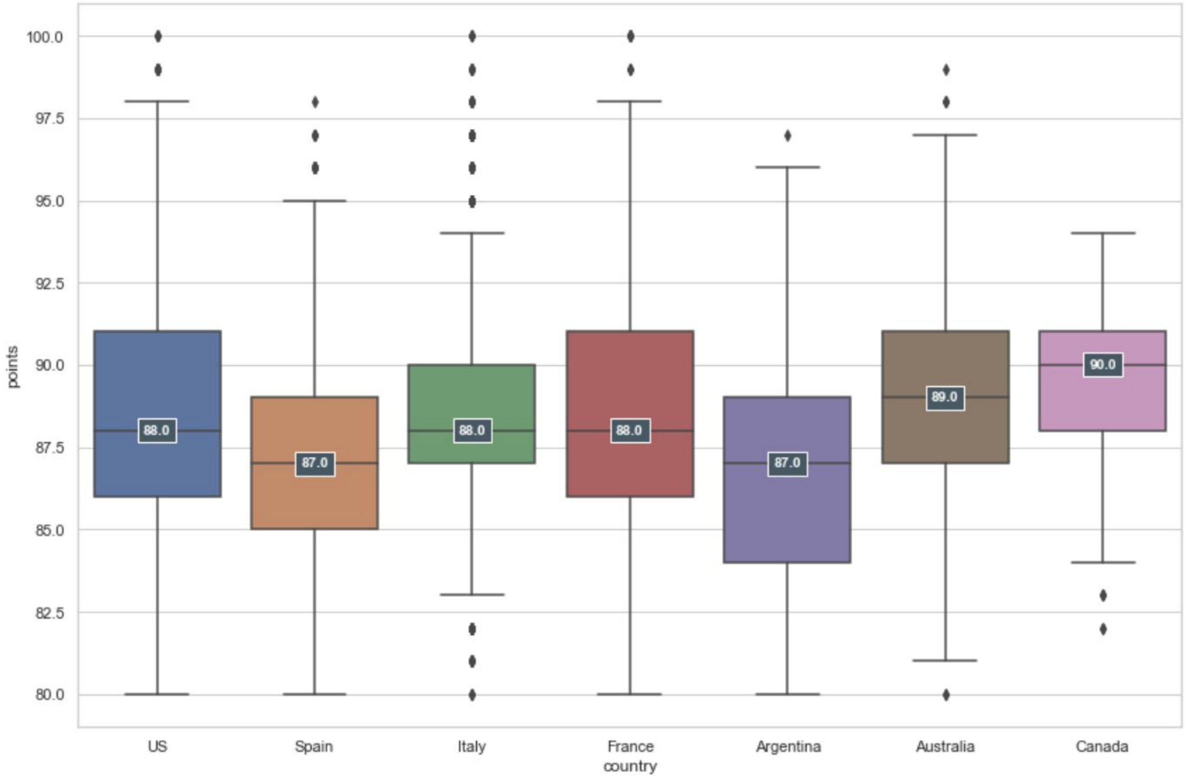
- winemag-data-130k-v2.csv dataset from Kaggle
- 129,971 rows and 15 features in the raw dataset
- Any row with null in relevant features were dropped
- Added year feature extracted from wine title
- Cleaned data for prediction contains 4 independent variables - price, variety, region and year

Wine score vs. country



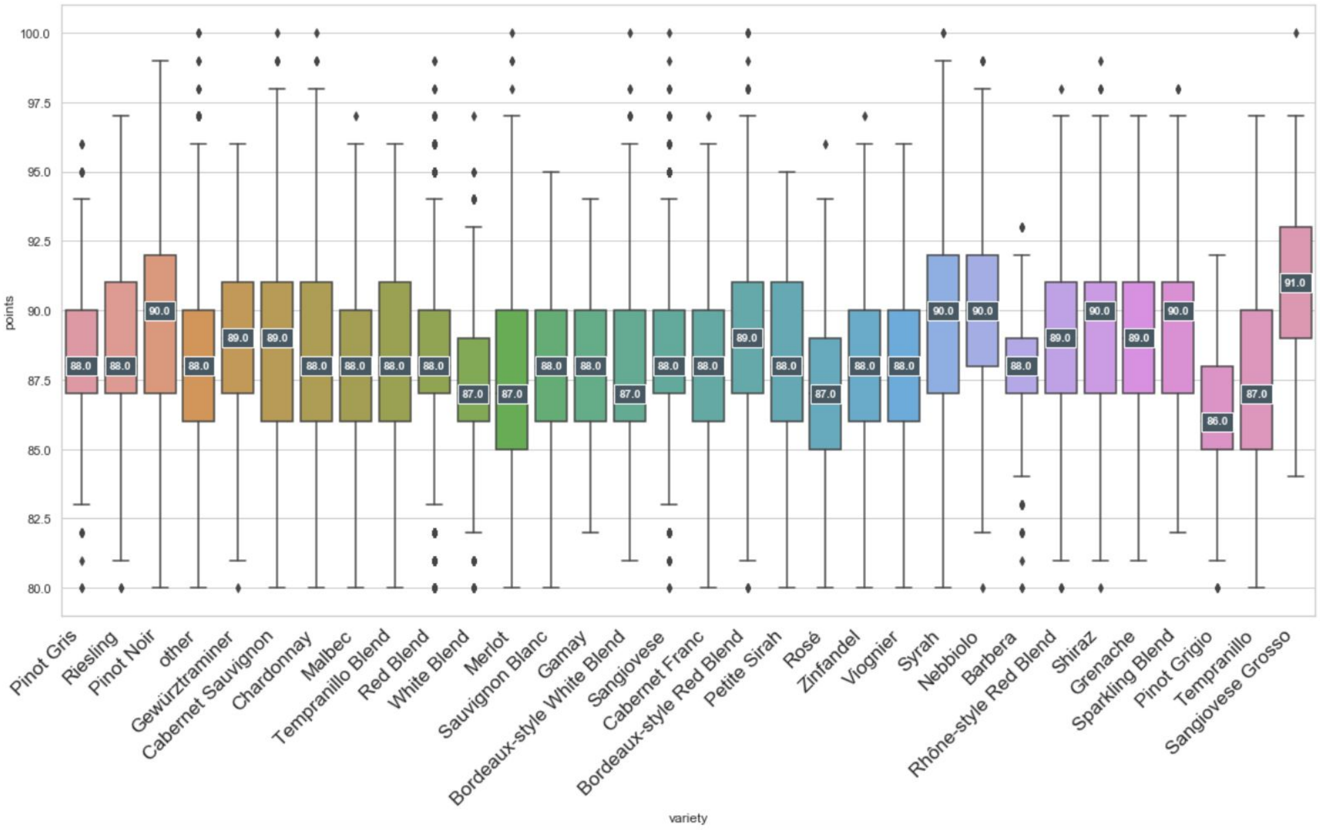
Looks like most of the data come from US, French, and Italian wines; we should keep that in mind when we think about bias in the data

Wine score vs. country



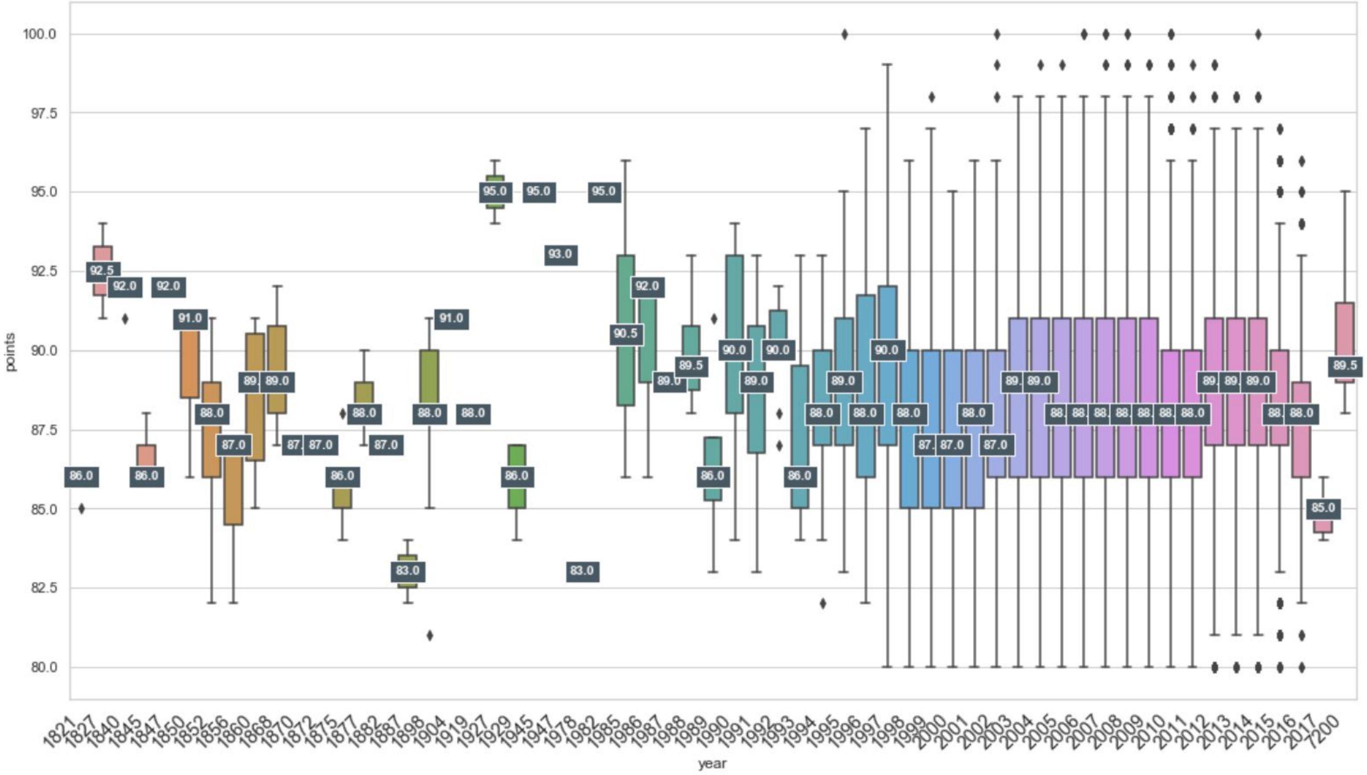
1. US, French and Italian wine have the same median points, and have outliers that have 100-point perfect scores
2. Italian wine has the smallest IQR while both US and French wine have the biggest
3. Australian & Canadian wine do not have large amount of data but have better median points than other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

Wine score vs. variety



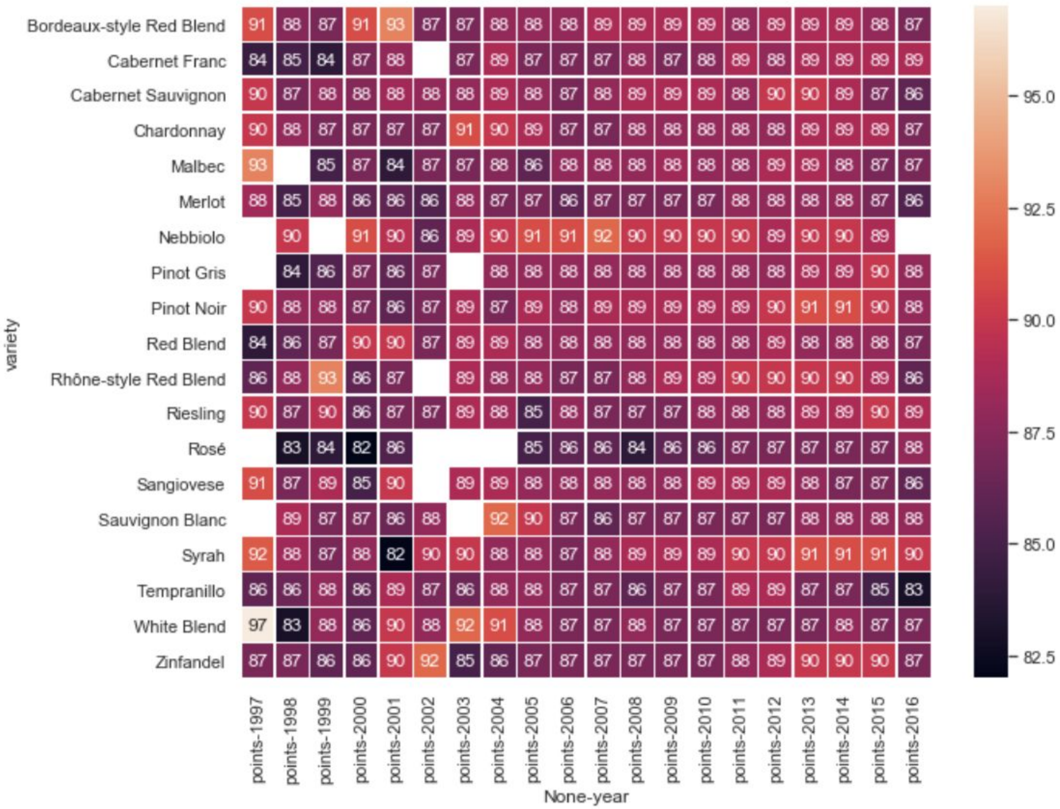
Here we can see Pinor Noir, Syrah, Nebbiolo, Shiraz & Sangiovese Grosso have the best median points.

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.

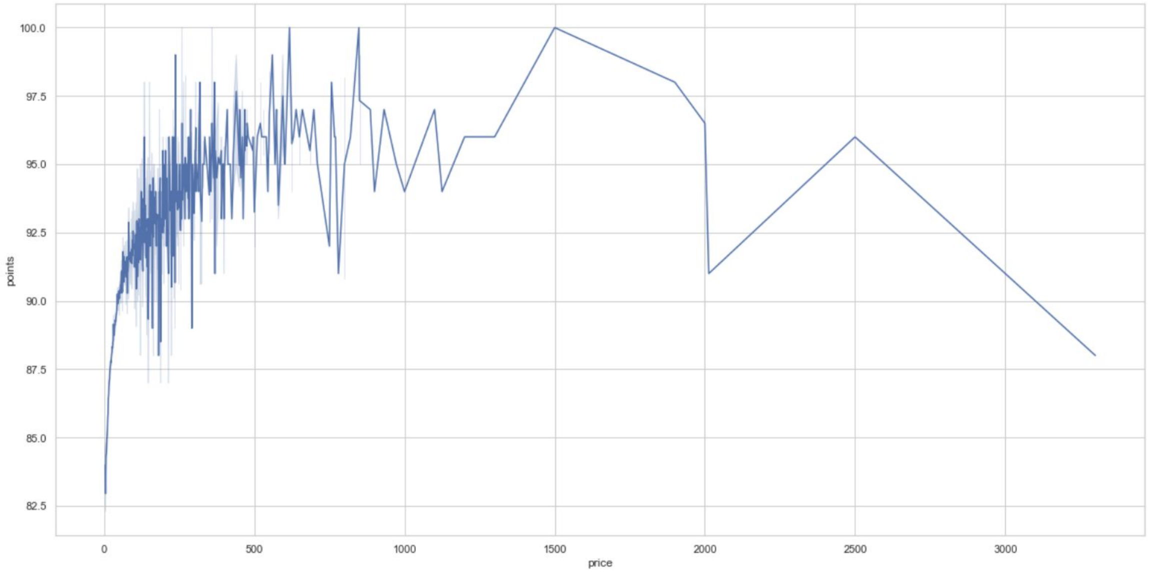
Wine score vs. year & variety



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

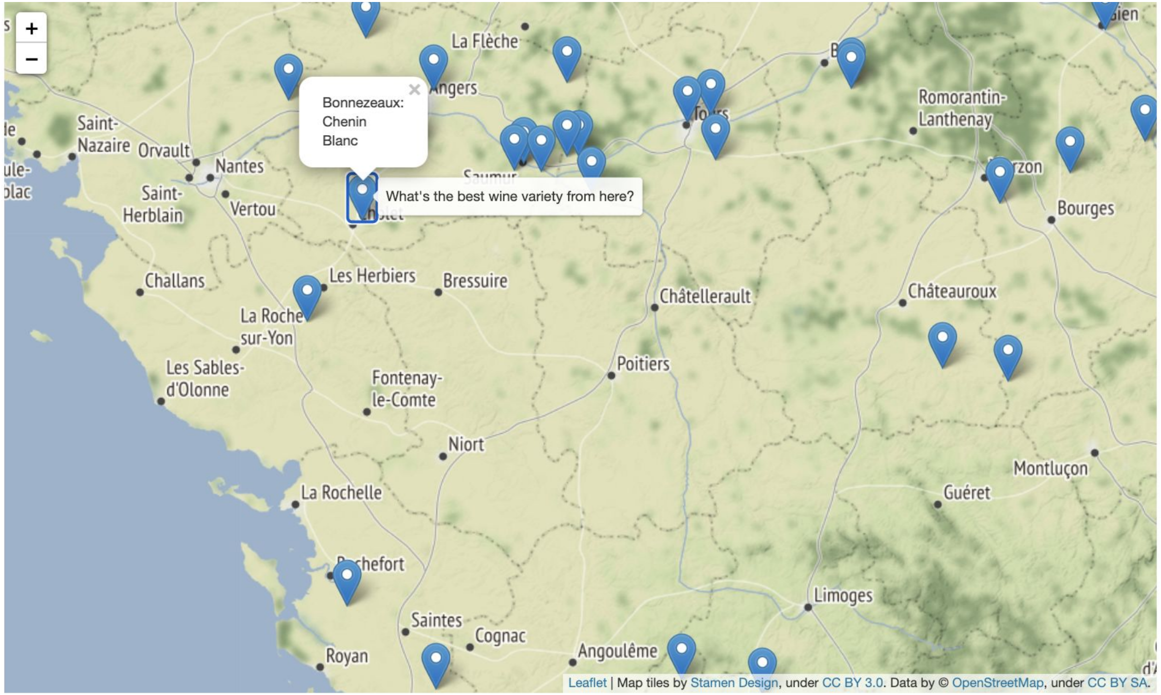
Seems like year combined with variety still is a bit arbitrary, but to our surprise, 1998-2005 has more wine varieties with low score vs 2006-2016.

Wine score vs. price



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.

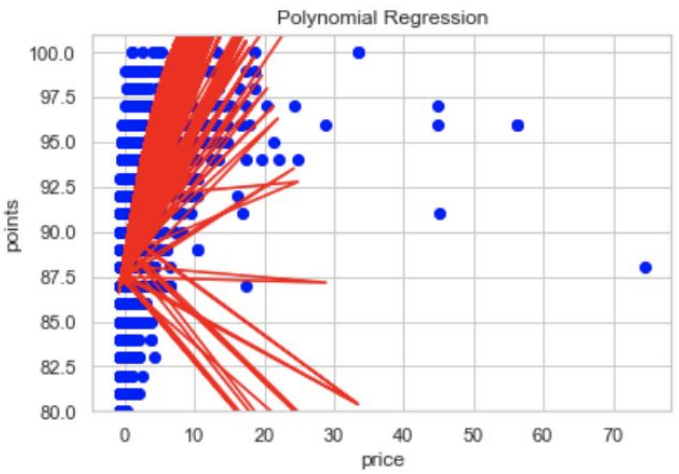
Best wine variety from each region



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.

Linear Regression MOdels

- Single regression (price vs. points)
  - Polynomial Degree of 3
  - R^2 of around .28 for both training and test data.
- Multiple regression (price, variety, year vs. points)
  - Polynomial Degree of 3
  - R^2 of .42 with training set and .34 in testing set.



MULTI-LAYER PERCEPTRON REGRESSOR

R^2 of 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.

Conclusion

- There is a lot of subjectivities that goes into rating wine
- Can build better model with more robust features related to wine quality, such as wine flavor
- Built decent model to predict wine score based on price, region and variety for use this as starting point