Wine Ratings – Which Wine is the Best?

# **Andrea Fox Summer 2021 https://github.com/anfox86/Wine-Ratings-Which-Wine-is-the-Best**

# Which Domain?

My data is coming from Kaggle and will consist of four different datasets all on the same topic. The following resources are the ones I plan to use as they cover machine learning ideas, various analysis approaches, and provide some background on wine reviews. Several of the links go through author’s approach to wine reviews, but they are all using a different dataset than I am. I am hoping the machine learning sources will give me an idea on direction when it comes to modeling and digging deeper into my data.

1. <https://nycdatascience.com/blog/student-works/visualizing-wine-reviews/>
2. <https://towardsdatascience.com/wine-ratings-prediction-using-machine-learning-ce259832b321>
3. <https://datascience.wine/2017/07/19/is-robert-parker-high-comparing-wine-raters/>
4. <https://www.eatingwell.com/article/152611/5-most-popular-wines/>
5. <https://medium.com/@henrik.warfvinge/picking-your-next-bottle-of-wine-using-machine-learning-70f5f5cc10eb>
6. <https://dataconomy.com/2016/03/the-perfect-pairing-machine-learning-and-wine/>
7. <https://www.forbes.com/sites/bernardmarr/2021/05/07/vivino-choose-your-next-great-wine-with-big-data-and-artificial-intelligence/?sh=653ae0d46b94>
8. <https://towardsdatascience.com/predicting-wine-quality-using-text-reviews-8bddaeb5285d>
9. <https://medium.com/@larserikbolstad/predicting-wine-ratings-using-machine-learning-aa64167d25b9>
10. <https://medium.com/@zszhuangsheng/wine-reviews-data-analysis-e414a4d39646>

# Which Data?

The data that I am planning to use for this project came from Kaggle, Wine Rating & Price <https://www.kaggle.com/budnyak/wine-rating-and-price>. There are five csv files in this project space, but I only plan to use four of them. One of the csv files is a single row of different varieties of wine. The other four csv files contain specific data for each wine type broken out by red, rose, sparkling, and white. All four of the csv files have the same columns, so it will be easy enough to combine them into one dataset for further analysis.

# Research Questions? Benefits? Why analyze these data?

My approach for this project is I am the client. I have never been big into wine, but I have been told it is because I have never found the right one. For me to determine this I need to look at several of the variables like country, year, and ratings. Making the assumption that rating is the measure of quality I want to look at the following:

1. I have seen on tv and read in books that the older the wine the better, so my first question is fairly EDA driven and I want to compare year of make to rating to see if that is true.
2. Is there a relationship between the wine and the price? I always look at the cheap bottles at the store and wonder if those are just as good as going to a winery and paying several hundred.
3. Can I use machine learning to predict the rating of a wine based on location, price, or rating?
4. Are the different countries/regions known for a certain type of wine? If so, does this have an effect on the ratings?
5. This one is a reach, but depending on ratings I want to see if there is a consistent wine or winery that typically does well, but has an off year? My thought process here is California is experiencing a major drought right now, and Napa Valley one of the biggest wine producers in the United States is in California. I wonder if anything bottled in 2021 would be considered bad or not. This will take some research along with data analysis to answer.

# What Method?

I have been starting with visual analysis in all of my projects because then I get a chance to see the data and gain some ideas on potential relationships, as well as see some of my theories either prove correct or crash and burn. From there I want to do something like I did in project 1 and look at a few different machine learning techniques to see which scores the best with this data. One of my resources tried Ordinal Classification, K-Nearest Neighbors, and Random Forest for their evaluation with Random Forest scoring the best (within less than a percent). The nice thing is here I have my independent variable being the wine rating, so I should be able to look at various methods easily unlike Project 2 where all of my variables were independent. I want to do a predictor based on one of my questions, but I also want to see the relationships between rating and whether price or location or year has any effect.

# Potential Issues?

Because I am making this project more personal, my biggest worry is I start falling down a rabbit hole and mess up my schedule. I also always worry about getting into the various machine learning types, and whether I am doing them correctly and if so am I understanding the output correctly.

# Concluding Remarks

As I mentioned above this is really a research topic for myself. I want to see if the ratings can guide me to a type of wine I might like, but I also want to see what variables can affect the ratings. If it is within possibility, I may go buy and try the type of wine that comes out on top just to see if my project data is accurate.