# Analysis of the Wine Dataset

## 5 actionable insights

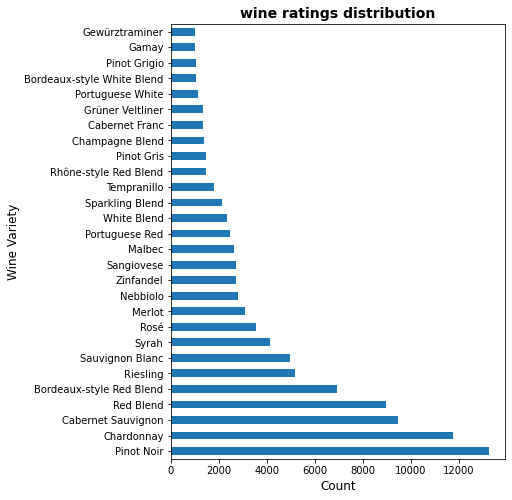
1. Data is highly imbalanced! Graph [[1]](#_[1]_Wine_Ratings) shows the distribution of the wine variety with respect to the amount of reviews that we have. We can see that Chardonnay and Pinot Noir have the highest amount of reviews, followed by Cabernet Sauvignon and Red Blend. Training a model on such data can lead to overfitting to the majority of the reviews.
2. Word cloud for the wine rating description [[2]](#_[2]_Word_Cloud) shows that the words use similar context. This means that there is no point in identifying the word-sense disambiguation and synonyms in the reviews. Word cloud for highly rated wines [[3]](#_[3]_Word_Cloud) (points > 90) tend to use words like rich, palate, long, racy, flavors, finish and then less frequent words like juicy, gold, texture, defined etc. This indicates that the Excellent wine reviews tend to have positive sentiment. While word cloud for average rated wines [[4]](#_[4]_Word_Cloud) (points <= 90) tend to use words like acidity, dry, dominate, along with words that shows frequent shift from positive to negative sentiment like still, although, underscoring. These reviews show positivity (because all reviews are rated >= 80) but overall explain negative sentiment.
3. The description used in the reviews of various countries [[5]](#_[5]_Bigrams_of) shows that the users tend to use similar group of words in their reviews. Bigrams for countries where there are more no of reviews tend to use words like fruit flavors, black cherry. Biggest difference between most frequently used bigrams is observed in the countries like France and Portugal where Ready drink is used most frequently. The bigrams of the review title [[6]](#_[6]_Bigrams_of) shows similar behavior. Hence, I decided to go with tf-idf vectorizer and use bigrams for preprocessing to encode the review description text.
4. If we observe the description length and the points given [[7]](#_[7]_Description_Length), We can see that the users tend to describe more when they give a very high rating to a wine variety and users tend to describe less when they think that the wine is average.
5. The points given to wine variety [[8]](#_[8]_Wine_Points) are mostly clouded in between the range 85 and 92. Median rating for any variety is at most 90 and at least 86. There are some varieties of wines that are given ratings that are beyond their typical range. i.e. outliers. 9 types of wines were given the highest rating in at least one review. E.g. Bordeaux-style Red Blend. And 24 out of 28 wines were given the lowest rating in at least one review. Most common are Chardonnay and Cabernet Sauvignon.

### *Bonus Insights*

1. User @vossroger has given more than 20 thousand reviews [[9]](#_[9]_Reviews_per) making him the person who has purchased and rated most wines. He should take it is easy as it is not good for his health. 😛 . @paulgwine is the person who has tried the most number of wine varieties [[10]](#_[10]_Most_unique) immediately followed by @JoeCz, @mattkettmann and our drinking hero @vossroger. If we check the ratings given by these users [[11]](#_[11]_Users_and) @vossroger tend to give below average rating to most of the wines while @mattkettmann gives slightly better reviews. @winechristna has the longest error bar because they have given less no of reviews and hence the uncertainty about the points [[12]](#_[12]_Central_Tendency) is higher
2. 69% of the wine reviews are from the Central Coast region [[13]](#_[13]_Reviews_per). If we check this region in a bit detail then we can observe that Pinot Noir, Chardonnay and Red Blend are the most reviewed wines. This shows why we have more no of reviews. Amongst these three Pinot Noir has the higher median (90) as compared to other (88) [[14]](#_[14]_Reviews_in). Most prices are between 0 to 100 (say Dollars). While there are some wines that have higher price [[15].](#_[15]_Boxplot_of)

# Graphs

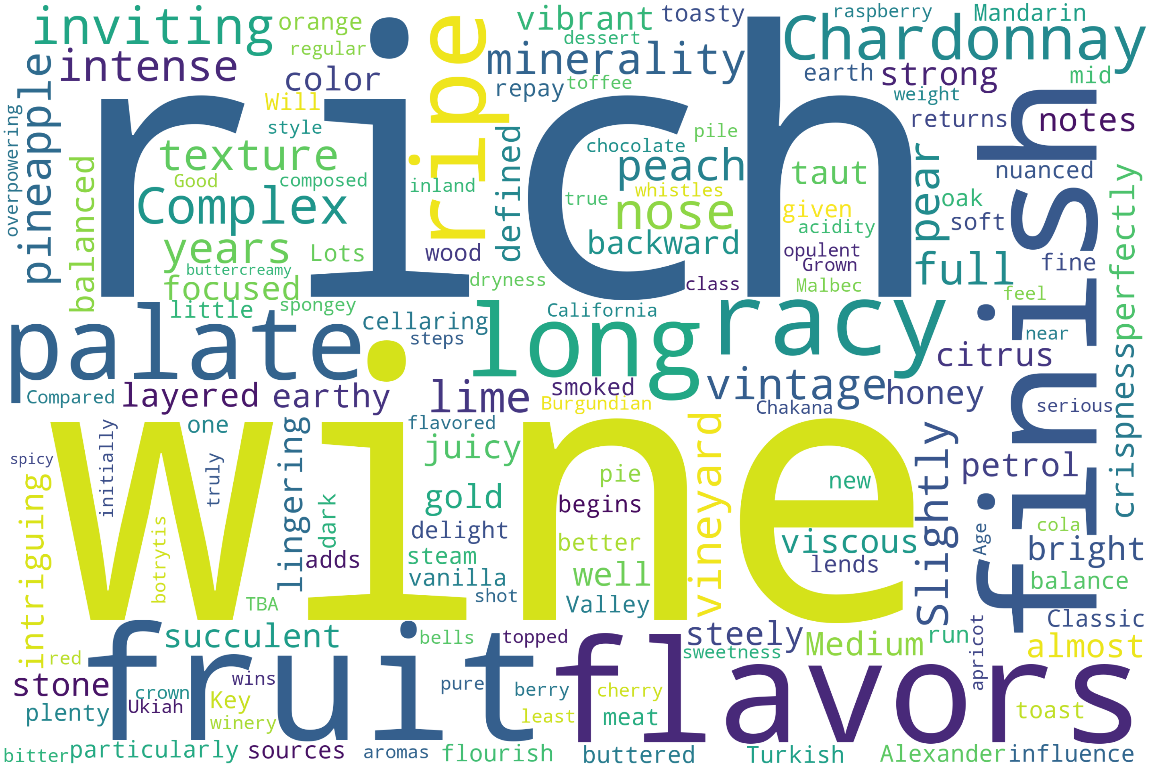
## [1] Wine Ratings Distribution



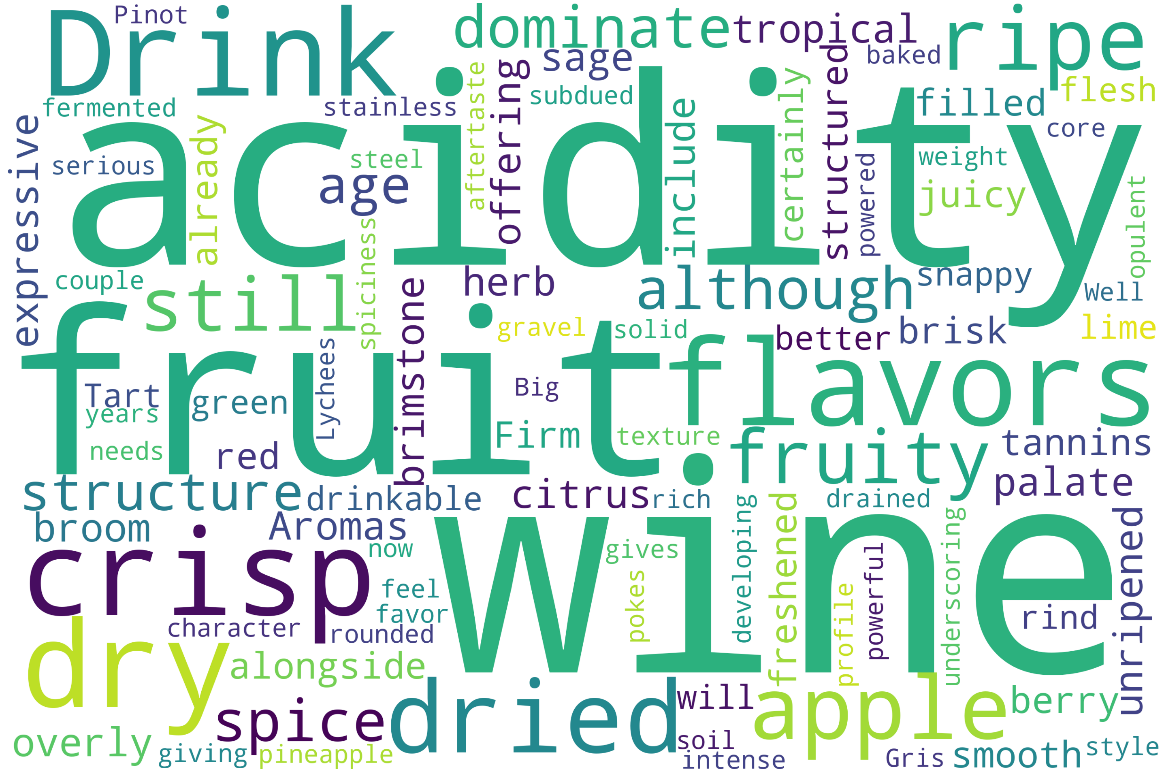
## [2] Word Cloud for wine rating description



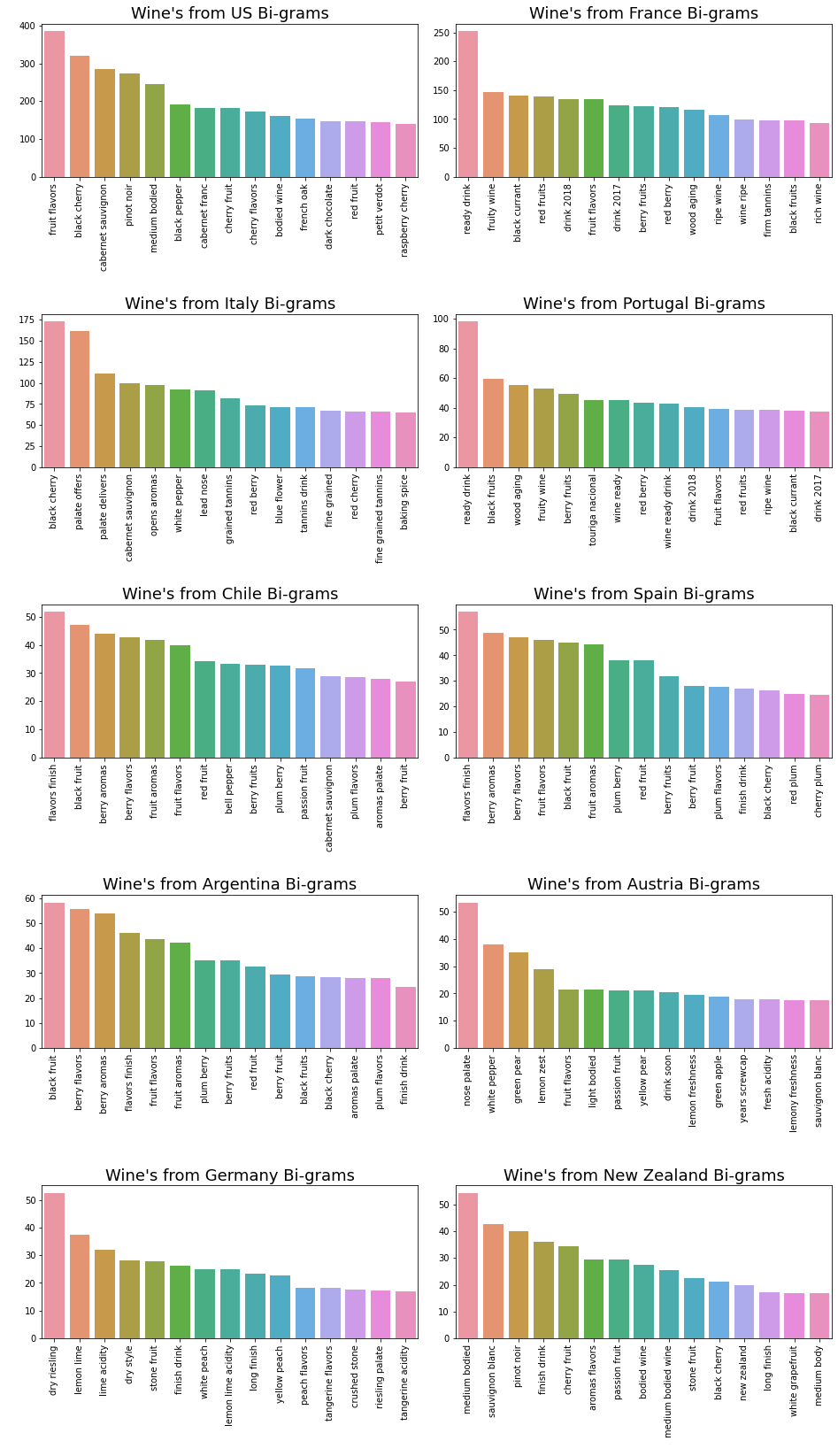
## [3] Word Cloud for Excellent Wines



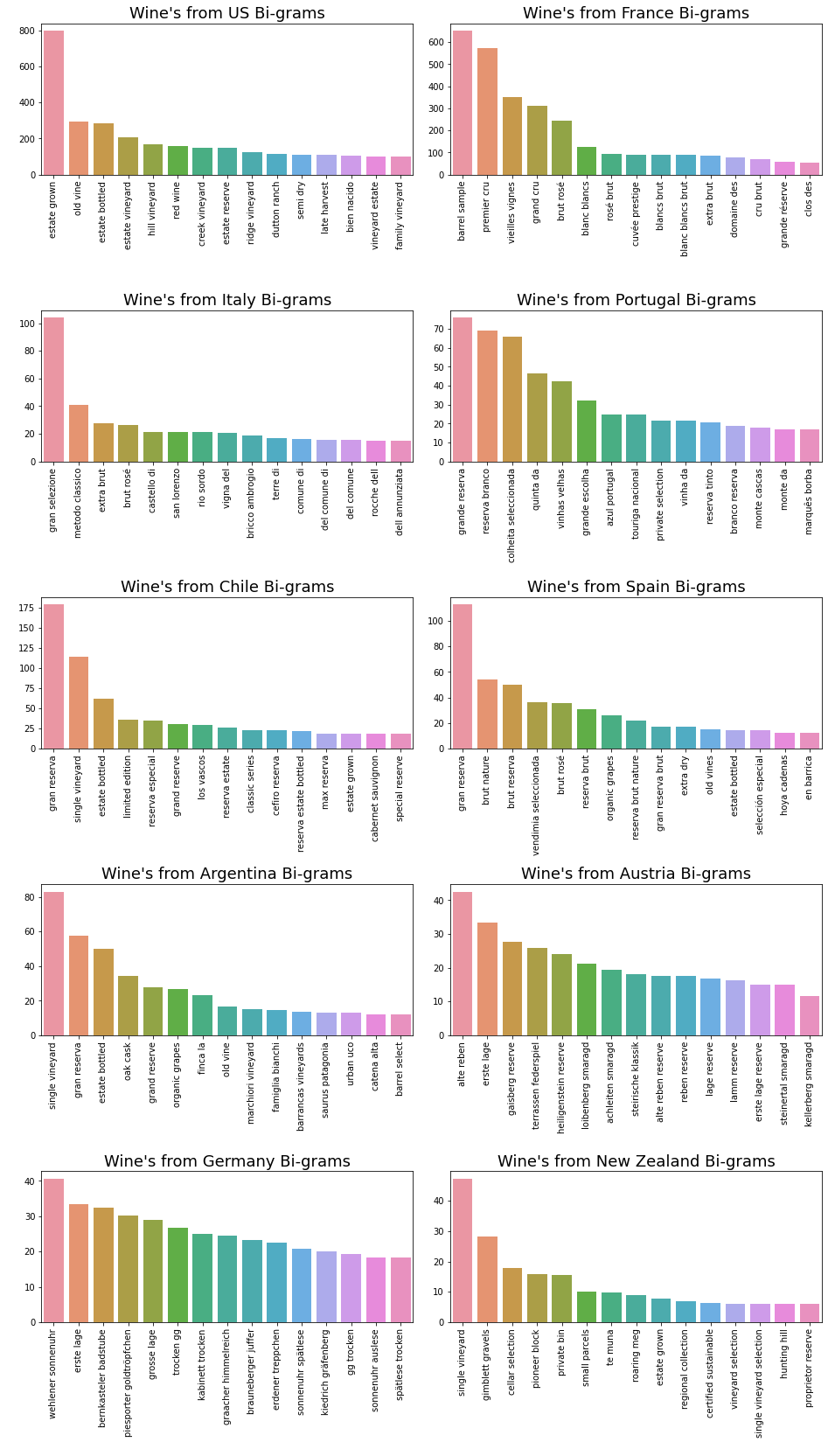
## [4] Word Cloud for Average Wines



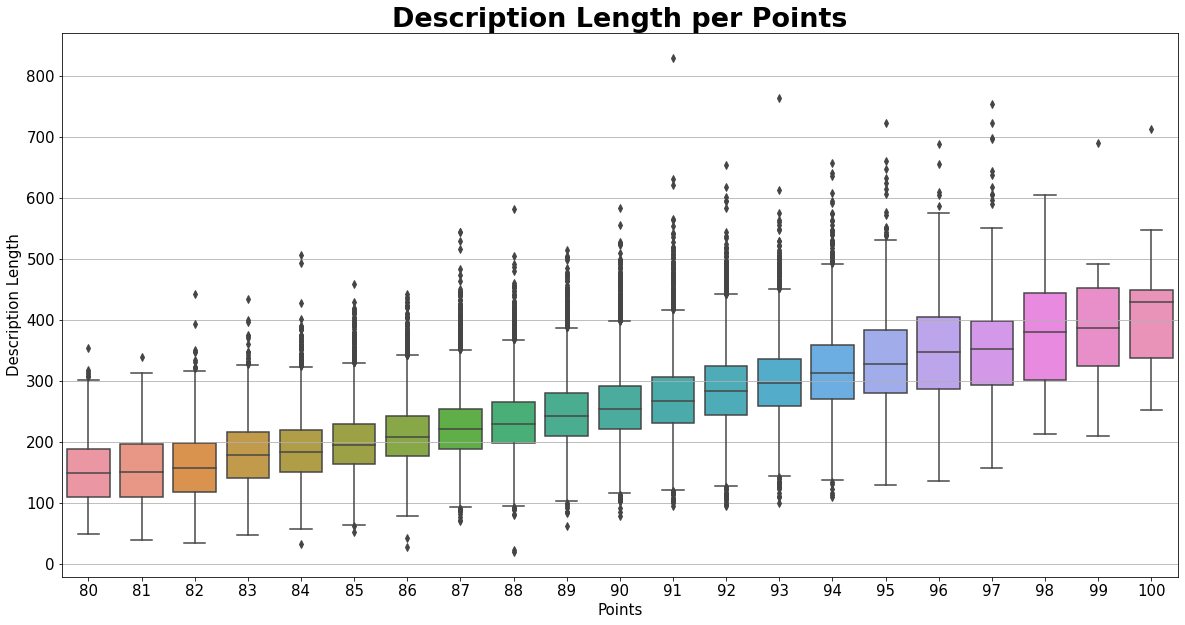
## [5] Bigrams of reviews as per country



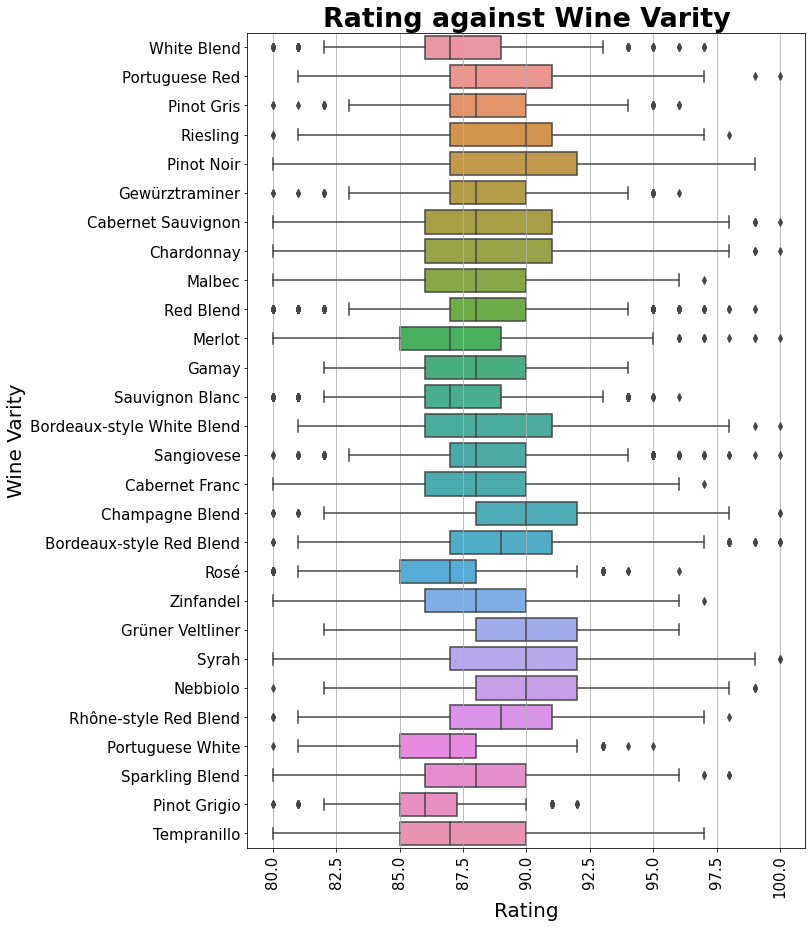
## [6] Bigrams of review titles as per country



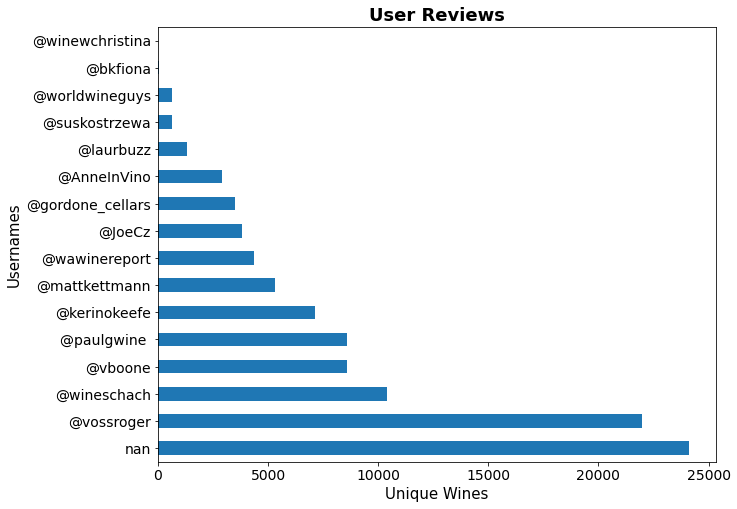
## [7] Description Length per rating



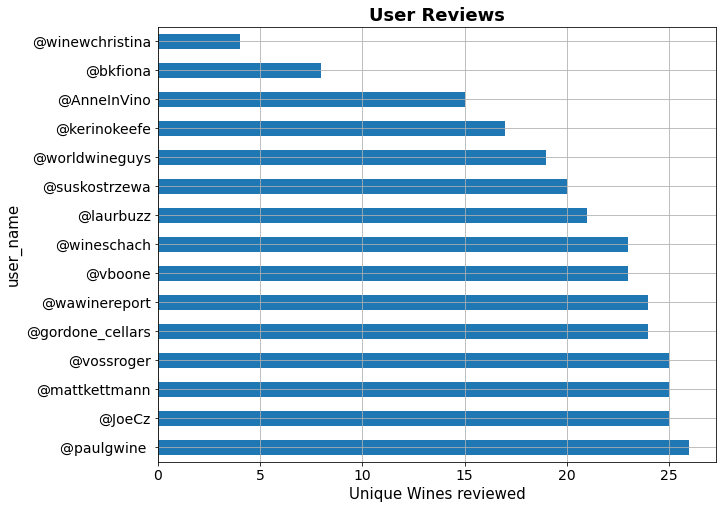
## [8] Wine Points Distribution



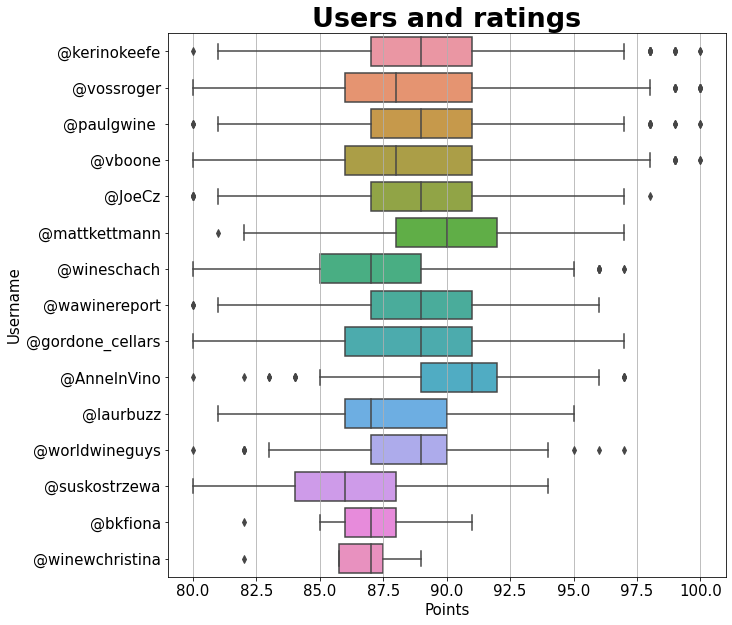
## [9] Reviews per user



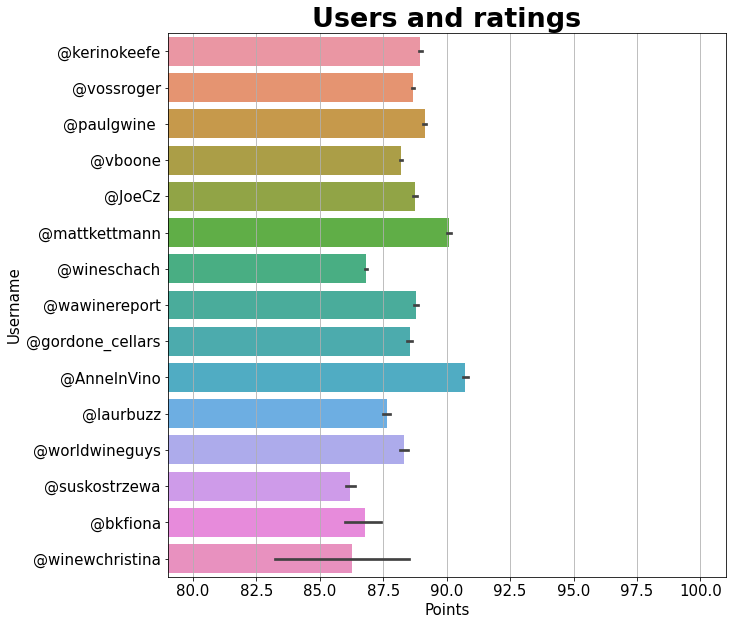
## [10] Most unique wine tried per user



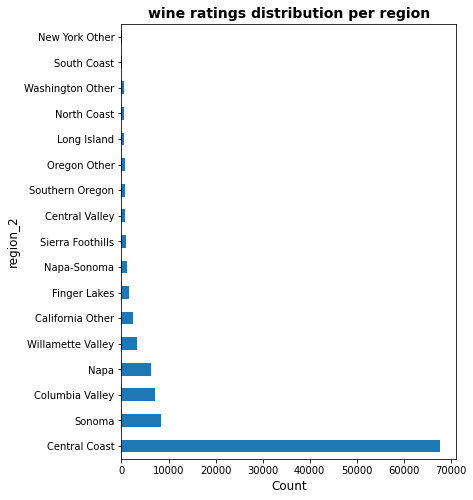
## [11] Users and their ratings



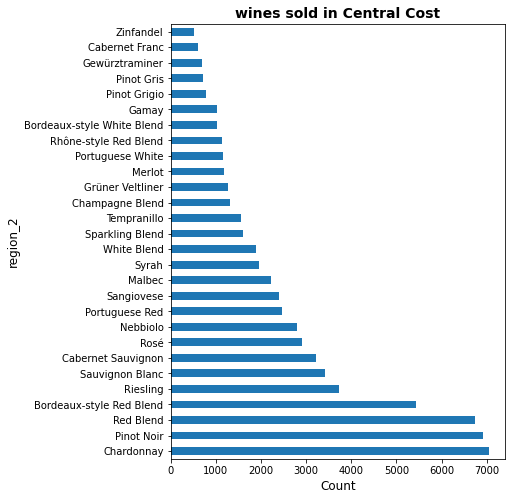
## [12] Central Tendency and error



## [13] Reviews per region



## [14] Reviews in the Central Coast



## [15] Boxplot of Central Coast Region

