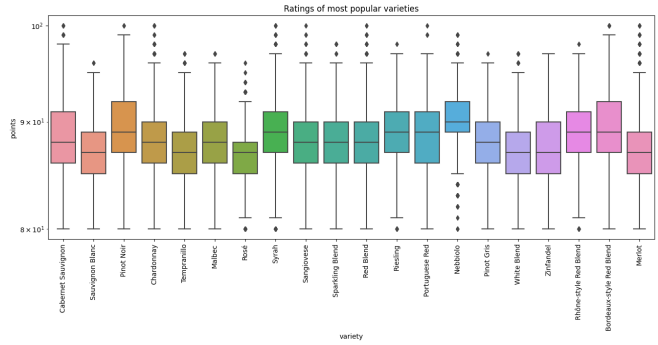
Wine Recommendations

Dataset - https://www.kaggle.com/datasets/zynicide/wine-reviews

DATASET:

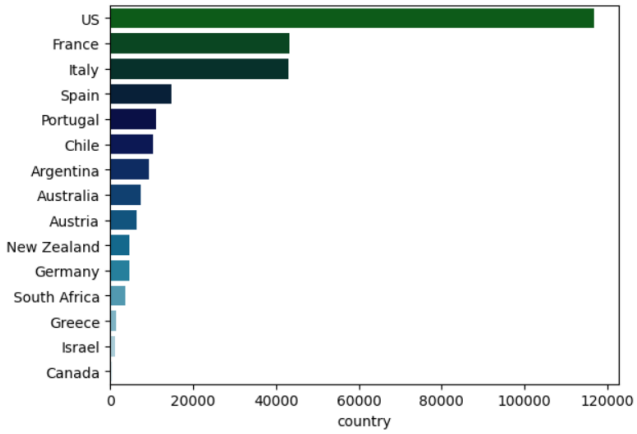
For this project, we decided to create a recommender system based on wines around the world. Given a multitude of features, we wanted to be able to output a recommendation for a wine based on the given features. We also wanted to be able to predict the points, or how good the wine was, based on features that would normally be on the wine label. This would include features such as winery, designation, variety, and region.

We created a graph displaying the ratings of the most popular varieties of wines. The box plot displays the wines that have more points, which is how the popularity is represented. Some of the most popular wines include Pinot Noir, Nebbiolo, and Bordeaux-style Red Blend.



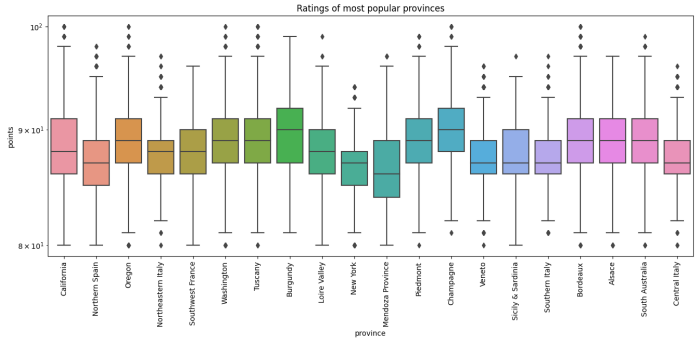
For the majority of popular wines, the median lies between 850 to 900. The lower quartile is not less than 825 for any of the wines in the graph above, and the upper quartile is no more than 925, so there is not much variance across some of the more popular wines. The range also does not have a stark difference, leading to there being very little variation in price over the top rated wines. This suggests that price is relatively constant over the majority of most popular wines.

The following bar graph breaks down the number of reviews by the countries that the wines they reviewed came from. Considering that the dataset has reviews of wines from 50 countries, the top 15 alone are depicted in the graph below.

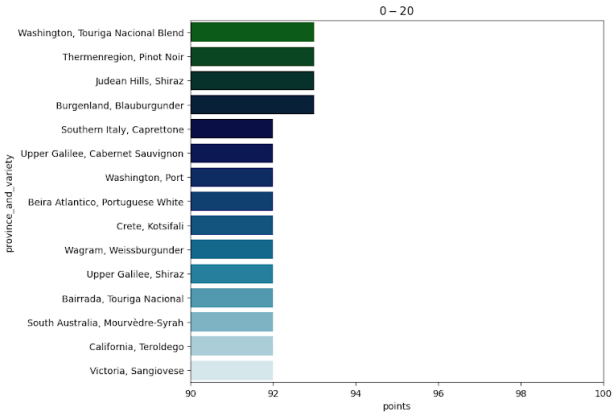


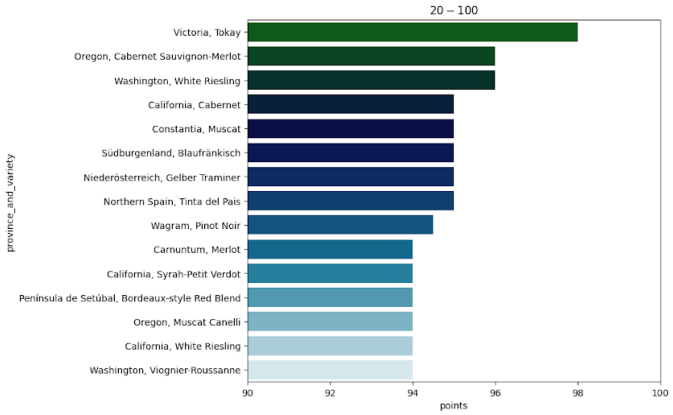
Of all the wines reviewed in our datasets, the most wines have come from the United States by far. The second most, France, has less than half of the wines reviewed from the United States. The second and third countries (France and Italy) have a very similar number of reviews that were concerned with wines from their countries. From that point onwards, there is another sharp drop off in popularity, with around 3000 fewer reviews for the fourth most popular county in Spain. From Spain onwards, there is a steady trend of decreasing reviews, indicating that a majority of all reviews examine wines from three main countries: the United States, France, and Italy.

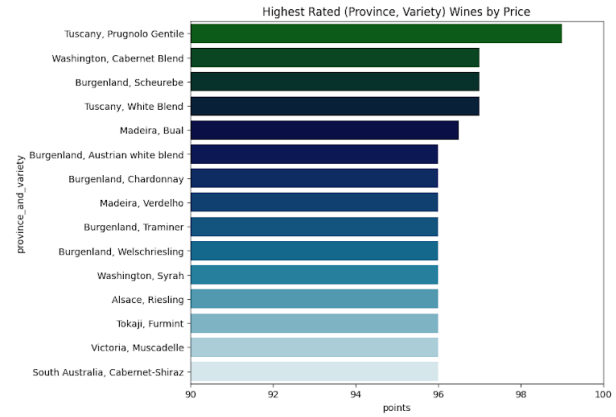
Looking beyond country popularity, another point of relevance with wine ratings is the specific province where they come from. Because each province produces multiple wine varieties, the bar graph below breaks down the average ratings of the 15 highest rated wine varieties and the provinces where they were made.



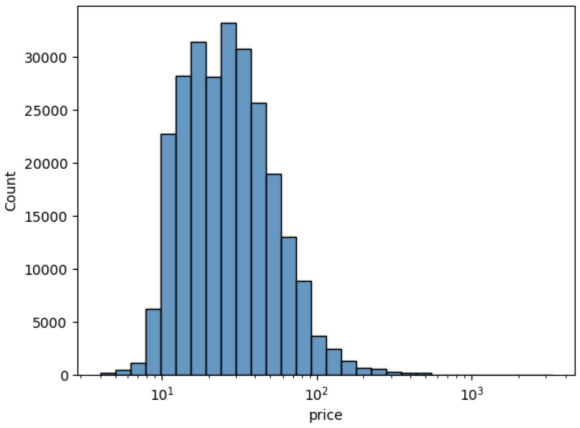
In addition to the comprehensive data analysis, we created another box plot illustrating the preeminent provinces, where Burgundy and Champagne notably emerged as the foremost regions based on accumulated points, thereby providing a clear and compelling visualization of the provinces' popularity within the context of our analysis.

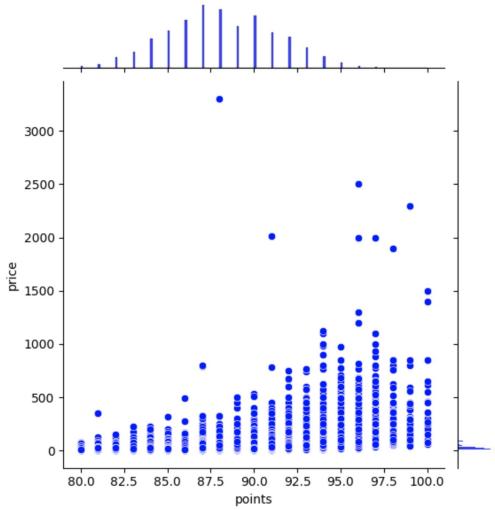


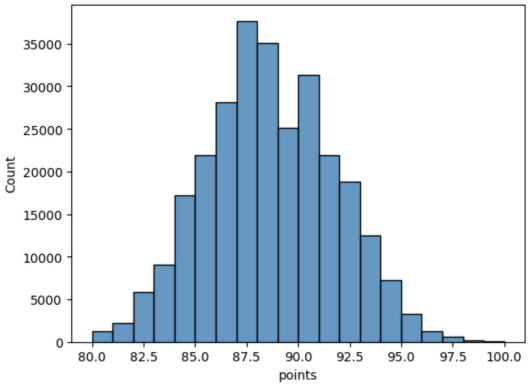




These (province, variety) tuples were split up into three separate price ranges in order to investigate the highest rated wines one could buy within these budget ranges. The wines with the highest rated medians were the Touriga Nacional Blend from Washington, Tokay variety from the Victoria region, and Prugiono Gentle variety from Tuscany for the $0 to $20, $20 to $100, and $100 to $500 ranges respectively range, the wine with the highest rated median is the Touriga Nacional Blend from Washington. There were only clear winners in two of the three price ranges. Overall, the graphs seem to have a trend where, for the most part, there are groups of bars that represent wines having similar average ratings, with slight drops between each group. As a matter of fact, the difference between the highest and the 15th highest rated wine is only around 4 or 5 points, even though the provinces that are in question here come from all around the world, with provinces of Canada, the United States, South Africa, Austria, and many more all making an appearance. This may indicate that, looking at the data from a larger geographic scale, the country as a whole does not play the largest role in determining how high the average rating of a variety of wine is.



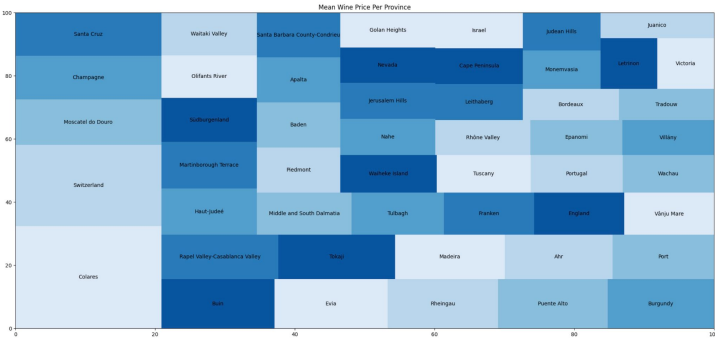




From these graphs, we can observe that both price and points are close to a normal distribution. The price distribution is slightly skewed right, with some wines costing significantly more than others ($100-$3000). The vast majority of wines are in the range of ($10-$100). The

points scored for wines are normally distributed with a median of 88.0. This is a promising indication since it will be easier for us to predict the scores of wines with a normal distribution. Combining these two distributions together, we see a trend that higher priced wines score higher, as expected.

In the word cloud of words in wine reviews, we see that the most popular words are those used to describe the characteristics of wine. For example, acidity, complex, structure, aroma, spice, and vintage are all terms used by reviewers to provide the reader with a sense of how the wine tastes relative to others.



Lastly, this square plot illustrates the mean price per province for the 50 most expensive provinces. Surprisingly, Colares and Switzerland are the most expensive provinces for wine, even though they are not as renowned as provinces like Napa Valley and Champagne. A possible reason for this is that these provinces have a smaller selection of wines, which all happen to be expensive due to the distinctive location.

PREDICTIVE TASK:

In order to process this multitude of data, we had to first parse through two csv files and concatenate them in order to obtain the final dataset. Once we acquired that, we then wanted to analyze which features were more important when attempting to recommend a wine based on similar characteristics. While examining the data, we found that there were some variables such as the taster's name, their twitter handle, the title, and the region2 variables to be less than necessary, so we dropped their columns. We also inspected which variables contained many null values, and dropped those to analyze our data better.

Our team identified a few different predictive tasks that would be beneficial to wine enthusiasts. We approached this problem from the perspective of an app that could identify enjoyable wines. Of course, there are many apps where people can rate their preferences for a wine, but the goal of this app would be to use the data that came from these ratings and preferences to build a recommender system that would provide users with suggestions about what wines they might enjoy. A recommender would be useful here since it's possible that the wine in question may have no reviews or very few reviews, and in either case, the available data would not be able to make a prediction. Thus, these models assist in using the other available data to make recommendations.To accomplish this ultimate goal, this project aims to break the process down into a series of intermediary tasks. From the exploratory data analysis that came earlier, it is clear that the ratings vary across a variety of variables including price, winery, variety, province, etc., and using the available data, this project will aim to build models that can:

1) Predict the price of wine using linear regression
2) Predict the rating of wine using linear regression
3) Predict whether a wine should be recommended to a user purely based off of its features using logistic regression
4) Predict whether a wine should be recommended to a user based on the interactions between users and items.

MODEL:

The goal of this project is to build a model that would provide recommendations to the user regarding which wine to purchase. To start simple, the first model made wine recommendations based on one factor: is the wine rated highly or not? This first model performs binary classification, placing each wine into one of two classes: would recommend purchasing and would not recommend purchasing. A logistic regression based model seemed appropriate considering this is a classification task. Before using the LogisticRegression class from sklearn's linear_model class, the data first has to be prepped. For simplicity, only two predictor variables were used: the variety and province of the wine. Since they are both characteristic features that help differentiate each wine,

they seemed like a good choice. They are both categorical variables so they needed to be one-hot encoded. There are 756 unique values in the variety variable and 490 unique values in the province variable. A dictionary was created to assign an index value to each of these unique variables, and these dictionaries were used in a feature function to create feature vectors for each observation. The truth vector was created by using list comprehension to assign 1 to every observation that had a rating higher than the 75th percentile of all ratings (which was 90.0) and 0 to all other observations. However, due to the size of the feature vectors, there were performance issues, so only a random 25% of the 280901 total observations were used for training, and a random 10% were used for validation. The size of the training and validation sets was constant across all linear and logistic regression models. The model was then fit on this training data, and when tested on a validation set, it resulted in an accuracy of around 0.769. With an accuracy greater than 0.75, this model provided a good start, but there is still room for improvement.

The next step of the process is to see if there is any way to bring the accuracy up using the same logistic regression model. The most logical next step is to add another predictor variable. Looking back at the exploratory data analysis, there seemed to be some relation between the price and the rating. Since the classification is based on rating, price was added as a feature to the new model. A new feature function was written to include price as part of the feature vector along with variety and province, which were both one hot encoded in the same way as the previous model. The feature and truth vectors were transformed in the same way as the previous model, with the only difference being the use of the new feature function. After fitting the new model on the training data, the accuracy on the validation set was increased to around 0.813.

With price being such a crucial element of the improvement in accuracy, it begged the question of whether price could be predicted, as that would help make recommendations of wines where neither price nor rating are not readily available. Since price is a continuous variable, this becomes a regression problem, and again, variety and province are used as the predictor variables. The same feature function as the first logistic regression model is used, with variety and province being one hot encoded using the dictionaries that map their unique values to indexes. This is again used on the training data to get the feature vector, while the truth vector this time was simply the price values of each observation. The LinearRegression class from sklearn's linear_model was used, and it was fitted on the training data before being used on the validation set, where it performed poorly with a root mean square error (RMSE) of around 31.572, indicating that it is difficult to predict price using this model.

While it is possible that continuing to optimize the price predicting model by doing things like adding more predictor variables could help the ultimate goal, it would be too much of a sidebar from the ultimate goal. However, the idea of using linear regression did lead to the next step of improving the logistic regression model used earlier. Optimizing a linear regression model that predicts rating seemed like it would help the logistic regression since the same features could be used for the logistic regression, which is also highly dependent on rating. After experimenting, a combination of the variables of winery and designation in addition to the previously used variety, province, and price predictors was revealed to be the most valuable. A new feature function was needed, and efficiency once again became a problem because winery and designation also needed to be one-hot encoded. Because there are 47239 unique designation values and 19186 unique winery values, it is not feasible to one-hot encode each of these. Hence, the dictionary that mapped unique values to index was only created for the 350 most popular designations and the 75 most popular wineries. The new function one-hot encoded these variables in the same way, but if a winery or designation was not found in the corresponding dictionary, the part of the feature vector describing that variable was filled with 0s. After transforming the training data to obtain feature vectors, the truth vector was created to consist of the raw rating values. The model was then fitted on the training data, and it performed well on the validation set, with an RMSE of around 3.625.

Considering the performance of the linear regression model, the same feature function was used to create features for the logistic regression model. Using the same processing pipeline as the logistic regression models from earlier, this new logistic regression model performed even better with a validation accuracy of around 0.862, which is an impressive jump of almost 0.05.

Lastly, a Jaccard similarity prediction model was implemented. Its purpose was to predict whether a taster would enjoy a certain wine depending on its Jaccard similarity to other wines tasted by the taster. First, an augmented validation set was created with entries of "not tasted" wines for every "tasted" entry. Generating these

negative entries is essential to training a stronger and more accurate model. When testing this predictor, the resulting accuracy was fairly low (0.55). Upon deeper study, it was discovered that there were only a select few tasters who had tasted massive amounts of wines. For instance, one taster had tasted north of 25,000 wines. This explains the similarity model's performance because a majority of the wines had identical taster sets that were very small.

LITERATURE:

The dataset employed in our analysis was sourced from Kaggle, and its utilization mirrored a comparable methodology employed in the construction of a recommender system. Similar datasets have historically been leveraged to ascertain recommendations for wines, albeit with different performance evaluation functions. Because of the enduring popularity and historical depth of wine, the extensive array of available varieties facilitates the creation of numerous datasets, each contributing to a nuanced understanding of the diverse facets within this vast and longstanding domain. There are various methods in analyzing such data, including but not limited to basic descriptive statistics and data visualization through scatter plots and histograms similar to ours. Additionally, machine learning models such as classification and regression models to predict wine quality scores based on attributes have been created with clustering algorithms like K-means or hierarchical clustering. Even recommender systems can be created with this type of dataset, through collaborative filtering or content-based filtering, which involves building profiles for wine drinkers and recommending wines based on the similar characteristics to their likings. Performance evaluation is also popular, with mean absolute error (MAE) and mean squared error(MSE) being popular amongst the literature relating to wine that we have noticed. There were also numerous models used on the dataset, such as the probabilistic matrix factorization (PMF), non-negative matrix factorization (NMF), maximum margin matrix factorization (MMMF), Bayesian personalized ranking (BPR), indexable Bayesian personalized ranking (IBPR), item k-nearest-neighbors (ItemKNN), multi-layer perceptron (MLP), neural matrix factorization (NeuMF/NCF), hidden factors and hidden topics (HFT), collaborative topic regression (CTR), and others included in multiple studies. While these methods are prevalent, we also saw that there was deep learning with neural networks involved as well as natural language processing. Within these various findings, there were some conclusions that had details slightly similar to those that we encountered. For instance, one study utilized collaborative information filtering consisting of non-parametric supervised learning using K-Dimensional binary tree specific algorithms to recommend wines in Portugal and Brazil (https://www.mdpi.com/2504-2289/7/1/20) . Wines in Portugal that were highly recommended fell under the province of Madeira, which was highly rated in terms of points in our findings. There were various wines that overlapped between the two results, showcasing that some conclusions were similar to existing findings. That study also had measurements made using MAE and RMSE, and we also developed an RMSE function to help our model deal with outliers in its prediction. Another feature was modifying hyperparameters as well to improve results in our accuracy.

RESULTS:

Our final model achieved a high level of performance for the tasks of predicting whether or not a wine is considered to be quality (logistic regression) and to predict the average points of a wine (linear regression). Based on the features incorporated in these models, we find there is a strong correlation between the locality of the wine, the variety of grape, and the price of the wine to predict the rating of the wine.

This makes intuitive sense because we know certain regions have particularly ideal conditions for certain varieties of wine, but may not be ideal for other varieties of wine. There are also some regions that are in general well regarded for the quality of their wines such as Napa Valley or Champagne. Based on both the location and the specificity of location information, it can be a strong indicator for the quality of the wine. If a wine label specifies that it's from California, but does not provide the specific region of the province the grapes were harvested from, it's typically a sign of inferior quality. Thus, in our model, we incorporated as much geographical information as possible, along with the price of the wine. Of course, some wines may be overpriced and other wines may deliver quality well above their price. Nevertheless, the price of the wine is still a strong indicator for wine quality since it's a representation of the production cost involved with producing the wine.

Given this context of how we can predict wine quality from its label, our results were aligned with our expectations for model performance. Our best model's features were price, province, variety, designation, and winery. For a threshold of predicting whether a wine is or

is not above the 75th percentile of scores, we achieved an accuracy of 0.862 using logistic regression. The same features in a linear regression model resulted in a point prediction RMSE of 3.625. We believe this would provide a reasonably high confidence prediction to indicate whether we believe a wine will be of quality.

One possible way to improve the accuracy of our linear regression model that we did not explore would be to change our one-hot encoding to a target encoding implementation. Since some of the features had 1000s of categorical values, we had to restrict the size of our one-hot encoding by only providing a non-zero encoding for the most popular instances of that feature's value. If we were able to include more of our data as non-zero representations in the model, we may have been able to improve our model. However, it's important to note that it's unlikely to achieve a near perfect prediction accuracy. Wines are inherently subjective so two tasters may not agree on the rating for a particular wine. This level of variability presents a challenge to making predictions. Additionally, there is such a large variety of wines and the quality may simply come down to price and/or supplier. Two wineries producing the same variety from similar regions could vary in score, even if they sell for the same price. Thus, if we haven't seen a wine before, it is quite hard to determine whether or not the wine will be of quality.