**Introduction**

While seeing relatives out in Oregon, one of our initial group members, Michael Leon, went around visiting vineyards and wineries in the Oregon countryside with his uncle. His uncle taught him the ins and outs of wine tasting, and what makes certain wines different from others in taste, quality, and other aspects. As we were looking around at public datasets to do our project, we came across the wine dataset, and after we heard his story, we were curious if wine quality could be determined based on the chemical components of the wine, as opposed to just a connoisseur's taste. All of our group members showed an interest in the topic, and we were all in to do our project on wine.

Previous work using machine learning on the wine dataset is extensive in the literature [1-3]. Varied techniques have been used: from common classification techniques like k-nearest neighbors, random forests, and SVMs [2], to uncommon techniques like fuzzy ones [1,3]. In our project, we benchmark these techniques against more modern ones.

First, the team uses statistical methods to analyze the data and perform principal component analysis (PCA). Then, the team uses the best features from PCA in various classification and regression algorithms. Algorithm 2 is an implementation of ridge regression on the wine quality values. Method two implements the k-means algorithm, where the assigned cluster mean is used to predict a wine's quality. Technique 3 is a decision tree implementation that classifies the wine qualities based on groups of “bad” and “good” quality.

Together, these techniques demonstrate a comprehensive analysis of modern machine learning techniques on the wine dataset. Overall, we believe the methods developed during this project enable classifying wines qualitatively rather than subjectively. These models can be used to make business decisions on the variable inputs that companies need to consider when bringing a new wine to market. They can reduce the amount of research and development costs necessary to develop new varieties of wine. Additionally, they could be used to aid wine sellers in pricing their products. The value of the wine industry makes this project very useful to those who may want to consider a scientific approach to man

## Methods:

**Correlation Matrix**

The correlation matrix was made in order to help visualize the mean and standard deviation of each feature with respect to the quality variable. Using it, one can determine the correlations between various feature’s and a blend of wine’s quality. Qualitatively, correlation coefficient with magnitude greater than 0.2 show features that have strong predictive power. For the red wine data set alone, the features that have a perceived relationship of some type with the quality rating are volatile acidity, alcohol content, citric acid, and sulphates. For the wine data set alone, the features that have a perceived relationship with the quality rating are chlorides, density, and alcohol content. Already we can see that the two types of wine may need to be analyzed separately in order to create an accurate model, but we will still consider the full data set of both types. The full data set shows strong correlation of volatile acidity, chlorides, density, and alcohol content with the quality rating of the wine.

**PCA and Ridge Regression**

Say how we used the correlation matrix to decide qw needed PCA

Next, we decided to use dimension reduction to clean up the data and reduce the number of random variables in consideration. We performed principal component analysis(PCA) on the features on both types of wine and obtained the retained variance normalized to total variance in order to select only the variables that contribute the most to the total variance. For red wine, PCA was able to select 4 of the 11 components in order to retain 99% of the total variance of that data set. For white wine, PCA narrowed down the components to 6 from 11 while maintaining 99% of the variance in this data set. With this information, we know we can later on build a simpler, more efficient model(linear regression, decision tree, etc.) with less data.
Ridge regression was then incorporated into our analysis (using a lambda equal to 0 based on the results of our cross validation algorithm) with and without the prior PCA method. The red wine data had an rmse of .657 without PCA and .823 with PCA. We further broke the data into the individual rating groups, and each of these had a lower rmse. The white wine data had an rmse of .715 without PCA and .839 with PCA. Breaking this data down into the rating groups gave a lower rmse for each individual group. Looking into this, we can see that the rmse is higher for rating groups with a low number of data points, which makes sense because the model would be less accurate with fewer test points. Another take away that we have from this model is that white wine tends to have a higher rmse because the training error is increased with more data points(white wine has 3x as many data points as red wine) because of overfitting. We conclude that PCA is less necessary with more data

We also used the retained variance calculations on each feature to represent how correlated two features are. For example, we performed this on the graph below, which details two features as the variables (Free Sulfur Dioxide vs. Total Sulfur Dioxide). The red points represent the red wine data set, and the blue points represent the white wine data set. We can note a couple key points from this graph. First, the data varies little in the diagonal directional which implies strong correlation between the two features. Second, the two wine types occupy separate clusters in the graph, which further argues that the different wine types should be analyzed separately.

**Classification through K-means Clustering Analysis**
In order for us to classify the wine into different groups, we had to create these different groups while still considering the quality rating. We decided to classify the wines into 3 groups: Bad Wine (0-4), Good Wine (5-6), and Great Wine (7+). We then applied the k-means clustering algorithm.

With the picture, we can see there is still a lot of overlap between the three clusters. To deal with this, we then performed a distortion score analysis that plots the sum of squared distances from each point to its assigned center in order to decide the optimal number of clusters. The best number of clusters occurs at the “elbow” of the graph, which can be seen at.....(2 or 3)? While we have correctly applied the K-means clustering algorithm on the quality groups of the data and clusters are dense in some locations, we can still build a better model that more accurately predicts the quality based on the given features.

**Decision Tree**

Next, we decided to implement the supervised learning analysis of decision trees. For our decision trees, the dependent variable(classification) is going to be the quality of the wine broken into two groups: Bad Wine (0-5) and Good Wine (6+). We used 80% of the data to train the tree, and the other 20% to test it. First, we considered the red wine alone. We split the data into a train group for building the tree, and a test group for testing the accuracy of the decision tree. The red wine tree had an accuracy of 90.63% for predicting if the wine was good and the first split attribute was the alcohol content. We performed the same analysis using the white wine data set, but we achieved a much lower accuracy of 71.02% with a primary split attribute of Sulphates. Creating and testing a decision tree using the full data set gave us an accuracy of 80.31%. It is surprising that the first split on the total tree was alcohol as the white wine data set has almost triple the data points as the red wine set. However, the difference in the two split attributes among the different trees implies it makes more sense to analyze them separately as more information is gained by different attributes for each respective tree.

The one concern with using a decision tree was the correlation between the variables. This would make each split less independent of the others which would decrease the information gain at each level. This is evidenced by multiple splits on the same attribute occuring in sequence. For example, in the red wine data set, the tree splits on alcohol content 6 seperate times, showing the dominance of this variable over the others. It follows that in the prior PCA analyses, alcohol content was a retained feature in each iteration of the algorithm we ran. Overall, this was the best way we found to classify the data; however, it only breaks it down into two classes.

Correlation makes model less accurate

- **Data multicollinearity**: This type of multicollinearity is present in the data itself rather than being an artifact of our model. Observational experiments are more likely to exhibit this kind of multicollinearity.
- The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.
- Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model. You might not be able to trust the p-values to identify independent variables that are statistically significant.

**Conclusion:**
Overall, the decision tree showed the most promising results. The model was able to achieve a strong average classification rate of 80.31%, which indicates that wines could be effectively classified into "good" and "bad" quality baskets. However, the limitation of the decision tree method was that the data was binary classified. The model could be even more useful with further work into classifying with more nuanced labels.

Still, the ridge regression and k-means algorithms were effective in making predictions as to the wine quality. The average RMSE of 0.777 for the red wines and 1.19 for the white wines demonstrates the predictive strength of the model. Similarly, the k-means algorithm was able to effectively divide the dataset into 3 groups that all had distinct average quality among the groups.

In general, the models for the red wine datasets had lower errors and better accuracy than the models for the white wines. By performing PCA, it is possible to effectively reduce the dataset features from 11 to 4 for the red wines and 6 for the white wines while retaining 99% of the data variance. For the red wine data set alone, the features of predictive importance are volatile acidity, alcohol content, citric acid, and sulphates are elements of the principal components. For the wine data set alone, the features of predictive components are chlorides, density, and alcohol content. Additionally, this analysis indicates that the types of wine are different enough that they need to be separated out in the machine learning process. Finally, the greater effectiveness of the models on the red wine dataset seem to indicate that red wines are, in general, are better suited to machine learning techniques.

In totality, the models pushed the limits of the wine dataset. Validation on the ridge regression algorithm showed that lambda values at, or close to, zero are optimal, indicating tendencies to overfit the data. Similarly, the models performed significantly better on data that had been dimension-reduced by PCA. These patterns show the promise of future work in further preventing model overfitting.

[1] Escobet, Antoni, et al. "Modeling Wine Preferences from Physicochemical Properties Using Fuzzy Techniques ." Scitepress, pp. 1–7.

[2] Er, Yeşim & Atasoy, Ayten. (2016). The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. International Journal of Intelligent Systems and Applications in Engineering. 23-23. 10.18201/ijisae.265954.

[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemica