Jacob Gottesfeld (jbg272), Vaughn Campos (vac62), Fisher Bricker (gfb53)

ORIE 4740 Project Proposal

For our final project, our group has harvested data from the API of the popular online wine marketplace Vivino. Vivino has information on over 70,000 wines from around the world and users can rate and purchase them directly on the website. By iterating through Vivino, we created a dataset of 8500+ unique wines and 14 variables. These variables include: the price of the bottle, user rating, number of user ratings, type (red, white, rose, etc.), style (i.e. Californian Red Blend), year, body (on a scale of 1-4), acidity (on a scale of 1-4), if it is vintage, if it is natural, region, country, winery, and the bottle's size in milliliters.

Before analyzing our data we will clean it and remove any records with missing information. Additionally, in order to gain more intuition on our dataset, we plan to use unsupervised methods like clustering and/or PCA. Furthermore, we will create a bi-variable plot and a histogram of price and rating. These will give us information on the interaction between our predictors and the distribution of our data.

With this dataset, we plan to predict the price and/or rating of a bottle of wine based on its characteristics. In doing so, we will learn what characterizes an expensive or popular bottle of wine. Both price and rating are continuous variables so we will need to perform regression analysis. In our dataset, there are categorical variables such as region which will become dummy variables in our regressions. We will begin with a linear regression, remove outliers as needed afterwards, and iterate upon it with LASSO/Ridge regularization. We will also try subset selection in order to find the best linear models with fewer predictors. We will try nonlinear models as well such as a GAM in case the effect of some predictors on our response variables is not linear. Every model that we create we will iterate upon and use to make predictions in order to track their effectiveness. In training and regularizing our models, we will plot cross-validation error against different parameters. As a result of wine culture and its international popularity, we expect that our analysis will find clear relationships between the predictors and price and rating.