## Define

I’m a winemaker looking to make a new wine. I want this wine to be make with the intention of earning a high score from a wine critic. Since wine scores are a great predictor of wine sales, I want to make a wine that will score high and result in high sales. The best way to achieve this is to build a model that will predict wine scores based on wine variety and region the grapes are grown in. Defining what variety of wine to make and where to source the grapes will allow me to make a wine that will with some certainty score high and generate sales.

## Gather

On Kaggle I found a robust set of Wine Magazine data that looked at nearly **130,000** wines and their score. Wine scores were given by critics and came with a description of the wine. The key pieces I wanted to look at in the data set was variety, the regional data (county,province,region) and price.

## Explore

The first items I explored in the dataset was for NaN values. Especially numeric NaN values. Luckily points had no NaN values, but the price had over **8,000** NaN values. Do I replace NaN values with the mean price of **$35** or drop all the rows with NaN price values. Given the significant size of **8,000** plus NaN values and Price having the strongest correlation to points. I chose to drop all NaN price rows. That brought the working model to a little over **120,000** rows.

I then set out to find the Null Prediction value. The mean points was **88.** I found the value count of points == **88** and divided it by total count to get **.13** or **13%**. If I were to guest a wine would score an **88** (on a scale of **80-100**) I’d be right **13%** of the time. I need to build a model with an R2 score greater than **.13**.

Price has a range from min=**$4** to max = **$3,300**. To help normalize the price difference I created a new column Log\_Price which takes the log of the original price.

## Model

The initial model I wanted to build was a model to address the defined problem of using Variety, Region and Price to predict scores. With Price being the only numeric value, I needed to create dummy variables for Variety, Country, Province, and Region. I made a value counter for each string column to see where I should make a cut off of unique features and ‘Other’ features.

I then concatenated all my dummy DataFrames into a single Wine\_Model dataframe that had Log\_Price, Variety, Country, Province, Region features and Points to predict. While the **80-100** point scale is not infinite values perfect for linear regression, it’s neither a Y/N or 1/0 of a categorical logistic regression. I chose to run with a Ridge Regression model that incorporates a standard scaler.

## Evaluate

Running the Regional model on Ridge regression returned a **.43** R2 score. Over **3** times the Null Prediction of **.13**. Not too bad!

I ran the RidgeReg Coefficients to see if there was a theme in the coefficients that I could narrow a follow up model to. The strongest coefficients were a good mix of every feature category (Country, Province, Region and Variety). Such a good mix that there wasn’t a real strong case to eliminate one or the other from a follow up model.

To compare the model performance to another I chose to perform a similar Ridge model on a more inherently bias set of features. Winery and Price. Winery is inherently bias in the way a winery can gain acclaim over the years of high scoring wine which critics are bias towards. Also Winery is a variable that is out of our control when making wine, because we are our own winery and can’t be another winery.

Similar to the dummy column function I ran on the Regions model, I used a value\_count method function to select unique values. There were over **5,000** unique wineries and when concatenating a data frame into **120,000 x 5,000** my computer nearly melted. I throttled the value\_count / dummy columns down to 500 to start. Run thru the same Ridge model, it scored **.40**. I bumped the value\_counts / dummy columns to **1,000**, it scored **.43**. Again I bumped it to **2,000** and it scored **.46**, but **2,000** was at the brink of CPU my computer could handle.

## Answer

The original Regions model I conceived was a really solid predictor compared to other ways of predicting with the dataset. A score of .**43** compared to an incremental **.46** + of the bais winery model holds a lot of weight in the grand scheme of it all. While there is a certain level of inherent bias to Varieties grow in specific Regions. Those are features you can control as a winemaker when predicting what score you want to achieve at the right price.