**Abstract**

I scraped data from <https://www.totalwine.com/> to analyze what features have the strongest correlation with price for bottles of red wine.

**Design**

The goal of this project is to understand which specific product features & values have a strong correlation with price.

**Data**

The dataset (csv file attached) contains 595 rows of data. Each row of data represents a single wine bottle and 14 pieces of information that were extracted from each product url.

**Algorithms**

I converted the csv file to a pandas dataframe and completed exploratory data analysis to understand the value counts for the categorical variables and outlier distribution for numeric variables. I filled null values for 4 categorical columns (imputed ‘missing’) and 2 numerical variables (imputed mean). I used the name field to extract the year of each wine (only 1 null value was imputed with mean).

I used feature engineering to generate dummy variable columns for 4 different feature variables.

I tested several different sklearn linear regression models.

With my final model, I used a train/validation/test split of 60/20/20. The R^2 performance of this model on the test data was 28%. I visualized my findings of actual price vs predicted price of the test data using seaborn.

**Tools**

HTML code for each product page was scraped with selenium and then converted to a Beautiful Soup so that specific pieces of information for target & feature variables could be extracted. To avoid getting flagged for bot activity, a new selenium driver had to be opened for each product url.

**Communication**

I have attached the PowerPoint slides which outline my linear regression model and practical conclusions.