Presentation Title: Factors Which Influence Wine Price and Quality

Researchers: Richard Hardis, Paulina Jane LoCicero, James Trawick

Date: 10/23/2019

Abstract:

Our goal is to evaluate factors that influence wine quality and price, based on a dataset collected from <https://www.kaggle.com/zynicide/wine-reviews>, which was scraped from WineEnthusiast in June of 2017, then again in November 2017 (<https://www.winemag.com/?s=&drink_type=wine>).  We augmented this dataset with several other geographical and textual predictors to help further explore the drivers of wine price and quality.  Winery latitude, longitude, and elevation were gathered from the the mapquest Open APIs. Temperature and Precipitation data for countries was gathered from the WorldBank’s open climate API. The foundational winemag.com data set contains information for 130,000 different wines, some of the factors include: country, description, and designation among others.  Length of review was also calculated for each wine to use as a potentially significant predictor. Due to API limitations, we are randomly sampling 2000 wines from the original 130,000 to use for data cleaning and then build our model. With this data, we hope to answer the following questions:

Under which conditions does higher wine rating correspond with higher reviews?  Specifically, with which predicting variables included in a model does that model show rating being a statistically significant predictor of price? How accurate is a multiple linear regression model built off of the winemag.com 2017 dataset at predicting the prices of the top 20 wines of 2019 according to totalwine.com? Do some regions produce wines with higher quality (how does region influence points awarded?). We intend to run an initial simple linear regression, a multiple linear regression, an ANOVA test, and will evaluate other models which may fit the data well based on our initial findings. Finally, our team will also explore robust regression methodologies (lasso, elastic net, etc…) in order to identify important attributes and reduce overfitting in for our predictive models.

Initial Tasks:

PJ/Will - SLR, then MLR with more predictors

Rich - Scraping Code to get totalwine 2019 data, geography data from APIs, using MLR to predict wine prices and compare against prices listed on totalwine.com.

PJ/Will - ANOVA difference in regions for wine quality