Wines - Final Report

Richard Hardis, Paulina Jane LoCicero, James Trawick

12/1/2019

Table of Contents

# ISYE 6414 Final Project: Regression Analysis of Wines

## Summary

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.

## Background

### Study Motivations

We want to know more about wine…

### Study Expectations

Wine price and quality are related and can be explained better with the help of other variables

## Data

### Raw Data and Collection

Data scraped from winemag.com

### Data Exploration

Cook’s, VIF, etc.

## Modeling Analyses

### Model Approach 1: Multiple Linear Regression

MLR with variable selection and metric analyses (r^2, adj. r^2, assumptions, etc.)

### Model Approach 2: Poisson Regression

Poisson model with variable selection and metric analyses (r^2, adj. r^2, GOF, assumptions, etc.)

## Conclusions

### Implications

What did we find?

### Further Questions

What’s next?