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

Our raw data started from a root dataset of wines taken from <https://www.kaggle.com/zynicide/wine-reviews> as part of their machine learning competition. This dataset contains wine data for over 130,000 distinct wines and was scraped from the search results pages of winemag.com in June and November of 2017. The original dataset contained country, description, designation, points (quality score), price, province (state in the US or province in Europe), region\_1, region\_2 (subregion of the province or region\_1), taster\_name, taster\_twitter\_handle, title, variety, and winery. This dataset had many data points, but did not have many useful predictors for regression analysis, so we chose to scrape further webpages to gather more factors. Two datasets were developed with additional factors. The first large dataset has predictors review length (word count), year produced, abv (alcohol by volume), bottle\_size, category (red, white, rose), and importer scraped from the individual wine page on winemag.com. This large dataset has 73,256 individual wines. The second, smaller dataset contains all of the features of the large dataset and also includes average country temperature, standard deviation of country temperature, average country precipitation, standard deviation of country precipitation, lattitude, longitude, and elevation. There are 1123 unique wines in this smaller dataset. 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 large quantity of data requires detailed exploration to ensure its quality and suitability for regression analysis.

### Data Exploration

The first step in the data exploration is cleaning. This involved removing any wines that did not have data for price, quality, year,temperature, precipitation, lattitude, longitude, elevation, review length, abv, bottle size, or importer. Because of the large quantity of data available, this approach of removing wines with missing data is preferable to imputing using averages or other means. Next, data exploration of the data can commence.

The first step in exploring the data is checking how the data are distrubted and how each factor varies with all other factors. Upon basic exploration it can be found that the variable Bottle Size does not vary as all wines were judged in 750ml bottles. This variable is removed because it does not vary. Next, outliers need to be examined and dealt with.

We start with the large dataset. First, box plots will be used to find outliers. Outliers are then removed. For example, wines with alcohol contents above 50%, prices above $200, year produced outside of the years 2000-2020,

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

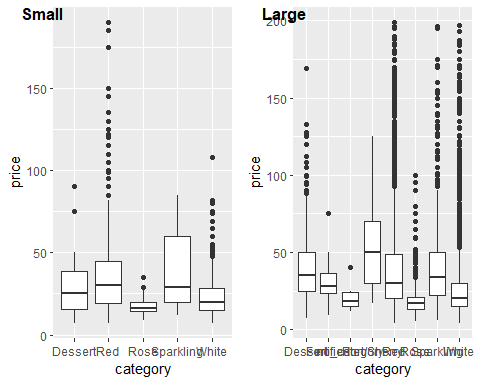
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

## Note: As of version 1.0.0, cowplot does not change the

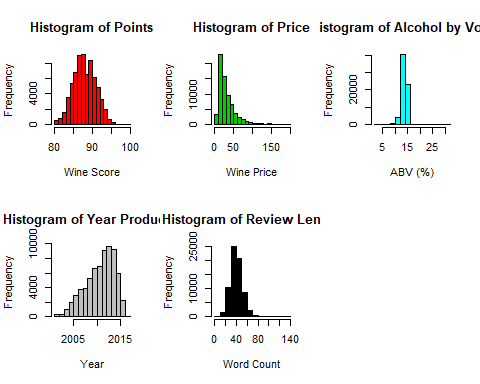
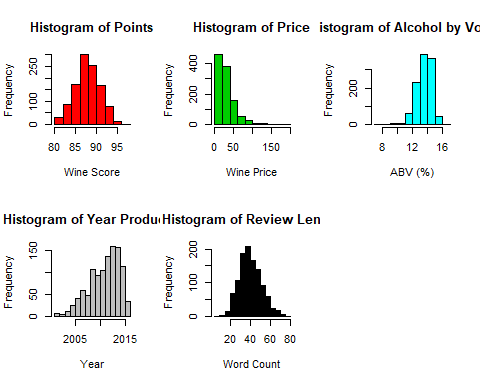
## default ggplot2 theme anymore. To recover the previous

## behavior, execute:  
## theme\_set(theme\_cowplot())

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*



## The following objects are masked from small\_data:  
##   
## abv, category, country, importer, points, price, province,  
## review\_length, year



## 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?