Investigating 150,000 WineEnthusiast Reviews

William M.

January 27, 2018

Investigating 150,000 WineEnthusiast Reviews

Using an independently collected database of reviews on over 150,000 wines web scraped from winemag.com we will attempt to explore trends and predictable patterns in the features of the world's best wines.

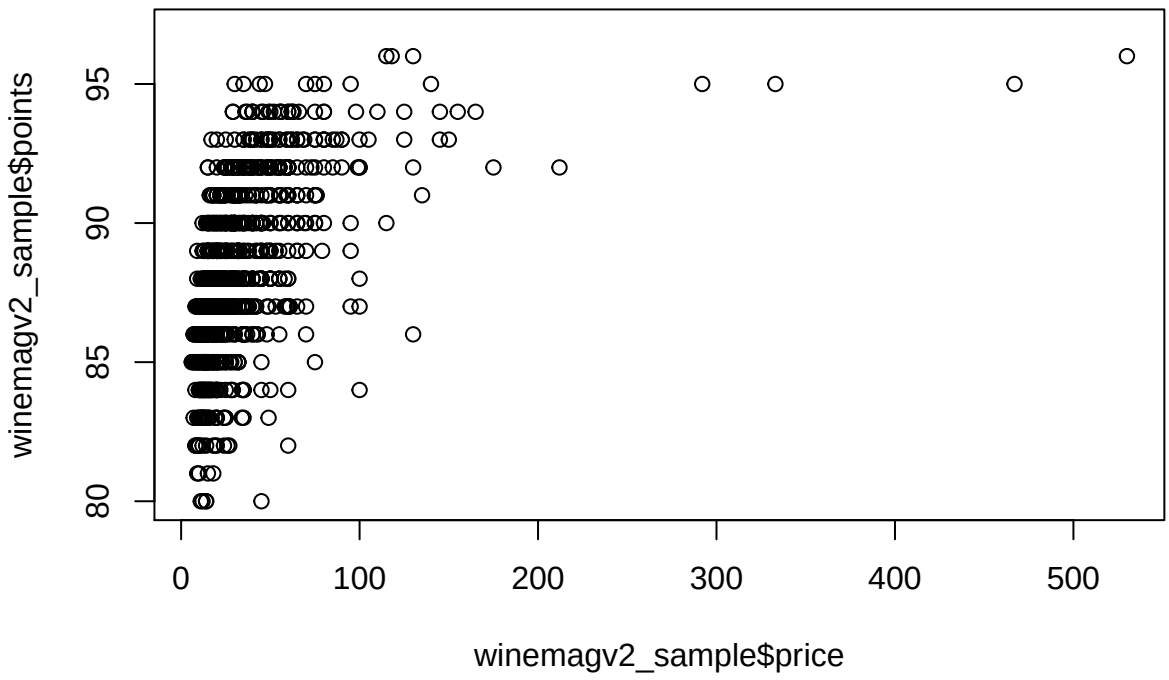
Acknowledgements

link: https://www.kaggle.com/zynicide/wine-reviews
user: "zackthoutt"

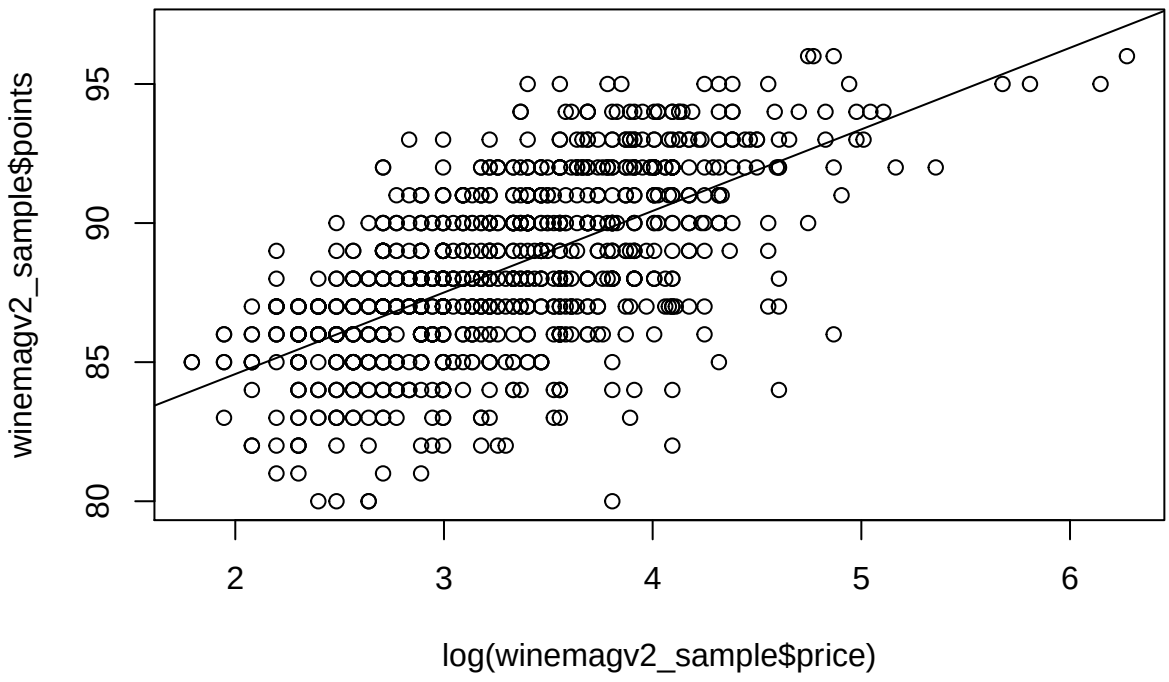
Trends in Price and Quality

We will seek to investigate the age-old question of whether or not our money is buying us discernible increases in quality when it comes to our wine. From the total data set we took a random sample of 1,000 listings and plotted price vs. points. The initial plot suggested an exponential relationship so we replotted log(price) vs. points and fit a linear model (by default, r uses base 10).

```
#Create sample set
winemagv2_sample<-winemag_data_130k_v2[sample(dim(winemag_data_130k_v2)[[1]], size = 1000, replace = FA
#Initial plot
plot(winemagv2_sample$price, winemagv2_sample$points)
```




```
#Log plot
plot(log(winemagv2_sample$price), winemagv2_sample$points)
abline(lm(points~log(price), data=winemagv2_sample))
```




```
lm_winev2<-lm(points~log(price), data=winemagv2_sample)
summary(lm_winev)
```

```
##
## Call:
## lm(formula = points ~ log(price), data = winemagv2_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8696 -1.4594  0.1949  1.5540  6.3193
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.7079     0.4015  196.03   <2e-16 ***
## log(price)    2.9321     0.1195   24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 927 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.3937, Adjusted R-squared:  0.393
## F-statistic: 601.9 on 1 and 927 DF,  p-value: < 2.2e-16
```

This implies an exponential relationship between price and quality where 34% of the response variance can be explained by the predictor given the equation:

$$points = 2.70 * \log_{10} price + 79.43$$

Quality by Country of Production

Using categorical linear regression, we can refine our original model to include country of production. This way we can investigate for systematic differences in quality based on price. We only include the top five most prolific countries in order to simplify our assessment.

```
wine_topcountry<-winemag_data_130k_v2[winemag_data_130k_v2$country %in% c("US", "France", "Italy", "Spa
lm_topcountry<-lm(points~log(price)+country, data=wine_topcountry, na.action=na.omit)
summary(lm_topcountry)
```

```
##
## Call:
## lm(formula = points ~ log(price) + country, data = wine_topcountry,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3802  -1.5110   0.0926   1.6776   9.4044
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     79.15788    0.04388 1803.85   <2e-16 ***
## log(price)       2.86636    0.01197  239.46   <2e-16 ***
## countryItaly    -0.35998    0.02588  -13.91   <2e-16 ***
## countryPortugal  0.71111    0.03921   18.13   <2e-16 ***
## countrySpain    -0.59012    0.03494  -16.89   <2e-16 ***
## countryUS       -0.40916    0.02083  -19.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.408 on 100397 degrees of freedom
##   (8070 observations deleted due to missingness)
## Multiple R-squared:  0.3712, Adjusted R-squared:  0.3712
## F-statistic: 1.186e+04 on 5 and 100397 DF,  p-value: < 2.2e-16
```

For a given price, a wine from Portugal is on average rated the highest by WineEnthusiast. Relative to a wine from Portugal, a wine from France is rated 0.71 points lower, a wine from Italy is rated 1.06 points lower, a wine from the US is rated 1.11 points lower and a wine from Spain is rated 1.30 points lower.

```
lm_total<-lm(points~log(price), data=winemag_data_130k_v2[winemag_data_130k_v2$country %in% c("US", "Fra
anova(lm_total, lm_topcountry, test="LRT")
```

```
## Analysis of Variance Table
##
## Model 1: points ~ log(price)
## Model 2: points ~ log(price) + country
##   Res.Df    RSS Df Sum of Sq  Pr(>Chi)
## 1 100401 589850
## 2 100397 582119  4    7731.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By performing an ANOVA test on the linear model of the subsetted dataframe with and without the categorical variable we can conclude that the addition of the variable is statistically significant at the p=0.001 level.

Varieties by Price and Points

We aggregate the entire data set by variety showing frequency of each on this website with both average listed price and average listed rating and order them from most to least frequent.

```
winemag_variety<-aggregate(winemag_data_130k_v2[,c("price", "points")], by=list(winemag_data_130k_v2$va
winemag_freq<-as.data.frame(table(winemag_data_130k_v2$variety))
winemag_finallist<-merge(x=winemag_variety, y=winemag_freq, by.x="Group.1", by.y="Var1")
winemag_finallist<-winemag_finallist[order(-winemag_finallist$Freq),]
winemag_finallist[1:25,]
```

```
##                       Group.1    price   points  Freq
## 443                Pinot Noir 47.52890 89.41147 13272
## 127                Chardonnay 34.52202 88.34008 11753
## 84          Cabernet Sauvignon 47.94002 88.60758  9472
## 475                 Red Blend 35.88119 88.38028  8946
## 64    Bordeaux-style Red Blend 47.21086 89.10644  6915
## 481                  Riesling 32.00040 89.45018  5189
## 518            Sauvignon Blanc 20.22852 87.42964  4967
## 564                      Syrah 39.13779 89.28658  4142
## 493                      Rosé 18.50644 86.84624  3564
## 328                    Merlot 29.54344 87.20858  3102
## 385                  Nebbiolo 65.60961 90.25107  2804
## 705                 Zinfandel 29.49225 87.82867  2714
## 509                Sangiovese 45.27934 88.55079  2707
## 282                    Malbec 29.92673 87.98303  2652
## 452             Portuguese Red 24.81922 88.81062  2466
## 692                White Blend 23.24079 87.35297  2360
## 557            Sparkling Blend 29.61125 88.04505  2153
## 587                Tempranillo 31.15092 87.51436  1810
## 477      Rhône-style Red Blend 34.92527 89.15364  1471
## 439                 Pinot Gris 23.07405 88.49622  1455
## 123            Champagne Blend 70.74484 89.66332  1396
## 75              Cabernet Franc 34.83678 88.15078  1353
## 240           Grüner Veltliner 27.78079 89.98067  1345
## 455           Portuguese White 15.34483 86.93097  1159
## 65  Bordeaux-style White Blend 34.72086 88.69043  1066
```