Histogram of Points for wine Reviewed in WineEnthusiast
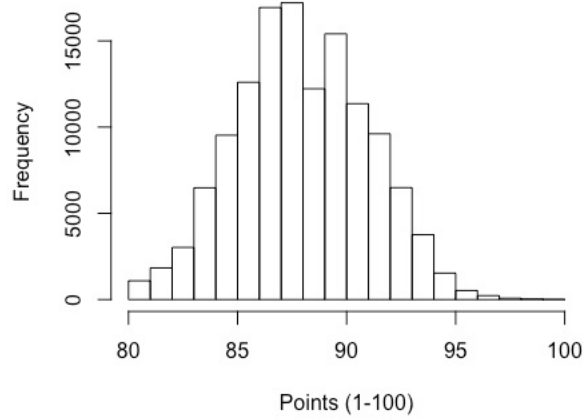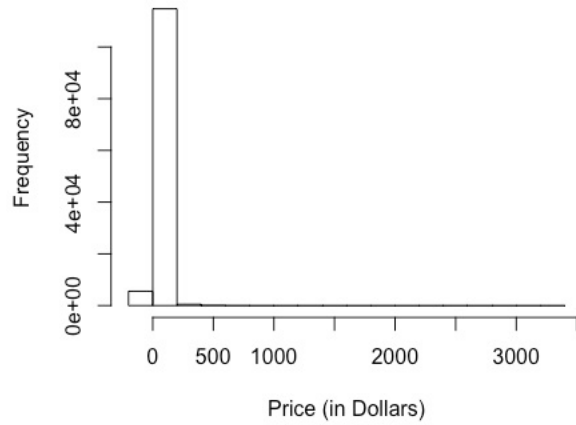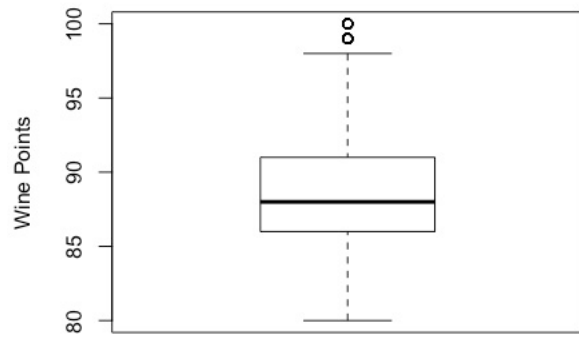

We can see from the graph on the left that points for wine has a normal frequency distribution.
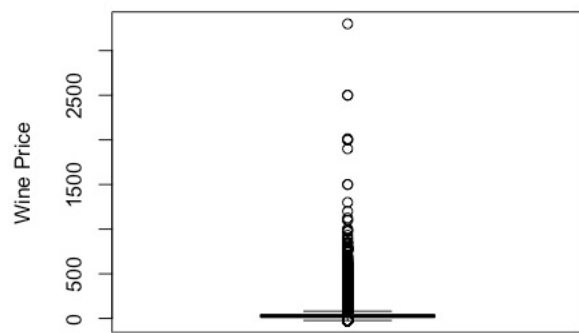
Histogram of Prices of Wine Reviewed in Wine Enthusiast

This graph on the left of price of wine has a slightly skewed distribution.
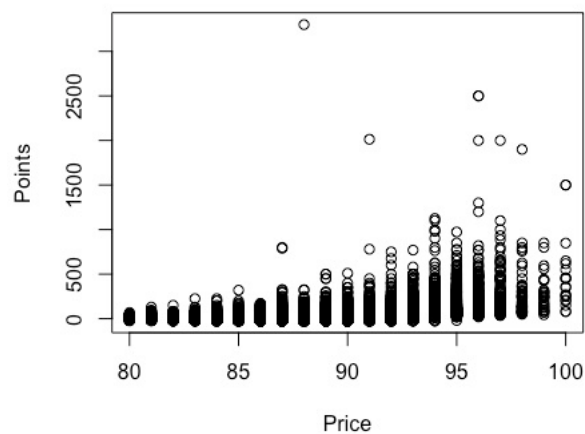
Boxplot of Wine Points

Checking the box plots we can see that points for wine has a normal distribution.

Boxplot of Wine Price

Wine price is skewed from this graph

Points vs Price of Wine

This graph show the price of wine vs the points received. There does not seem to be a strong relationship between the two.

cor(winemag_data_pandata$price, winemag_data_pandata$points, use ="complete.obs")
[1] 0.4198094

> summary(winepricepointmodel)

Call:
lm(formula = points ~ price, data = WinePricePointsDF)

Residuals:
   Min   1Q Median   3Q   Max
-99.178 -1.852 0.027 1.997 10.272

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.740e+01 1.018e-02 8585.9   <2e-16 ***
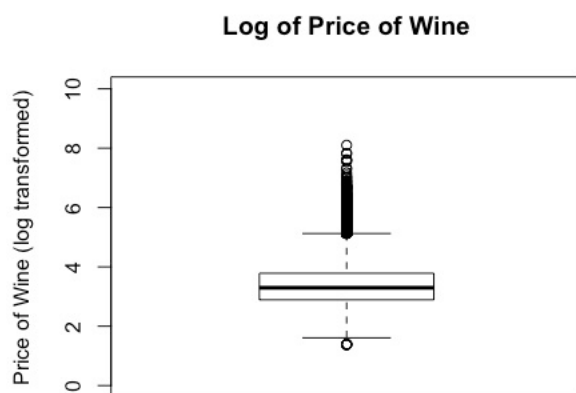price   3.024e-02 1.879e-04   160.9   <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

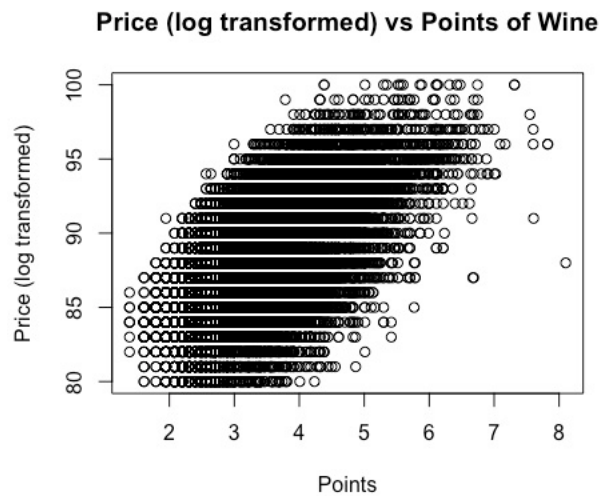Residual standard error: 2.763 on 120973 degrees of freedom
  (8996 observations deleted due to missingness)
Multiple R-squared: 0.1762, Adjusted R-squared: 0.1762
F-statistic: 2.588e+04 on 1 and 120973 DF, p-value: < 2.2e-16

Log of Price of Wine

The log transformation of price seemed to normalize the data.

Price (log transformed) vs Points of Wine

cor.test(winemag_data_pandata$price, winemag_data_pandata$LogPrice, use="complete.obs")

        Pearson's product-moment correlation

data: winemag_data_pandata$price and winemag_data_pandata$LogPrice
t = 362.85, df = 115430, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7272576 0.7326476
sample estimates:
    cor
0.7299639

summary(WineLogPriceModel)

Call:
lm(formula = points ~ LogPrice, data = WineLogPriceDF)

Residuals:
    Min   1Q Median   3Q   Max
-14.0333 -1.5163   0.1448   1.7084   9.2420

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) 78.99904   0.03698 2136.2 <2e-16 ***
LogPrice   2.84315   0.01086 261.8 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.416 on 115426 degrees of freedom
  (14543 observations deleted due to missingness)
Multiple R-squared: 0.3725, Adjusted R-squared: 0.3725
F-statistic: 6.853e+04 on 1 and 115426 DF, p-value: < 2.2e-16

Price (log transformed) vs Points of Wine
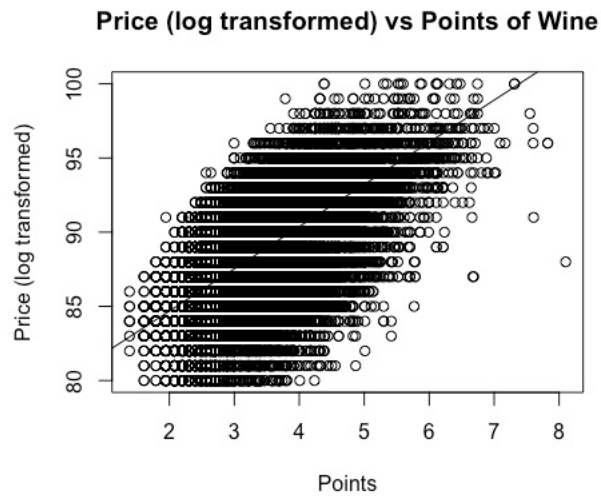
You can see from this model that the log transformation is a better predictor of points that are awarded to a wine.