Chelsea Clinger

411 Generalized Linear Methods 56

Unit 3

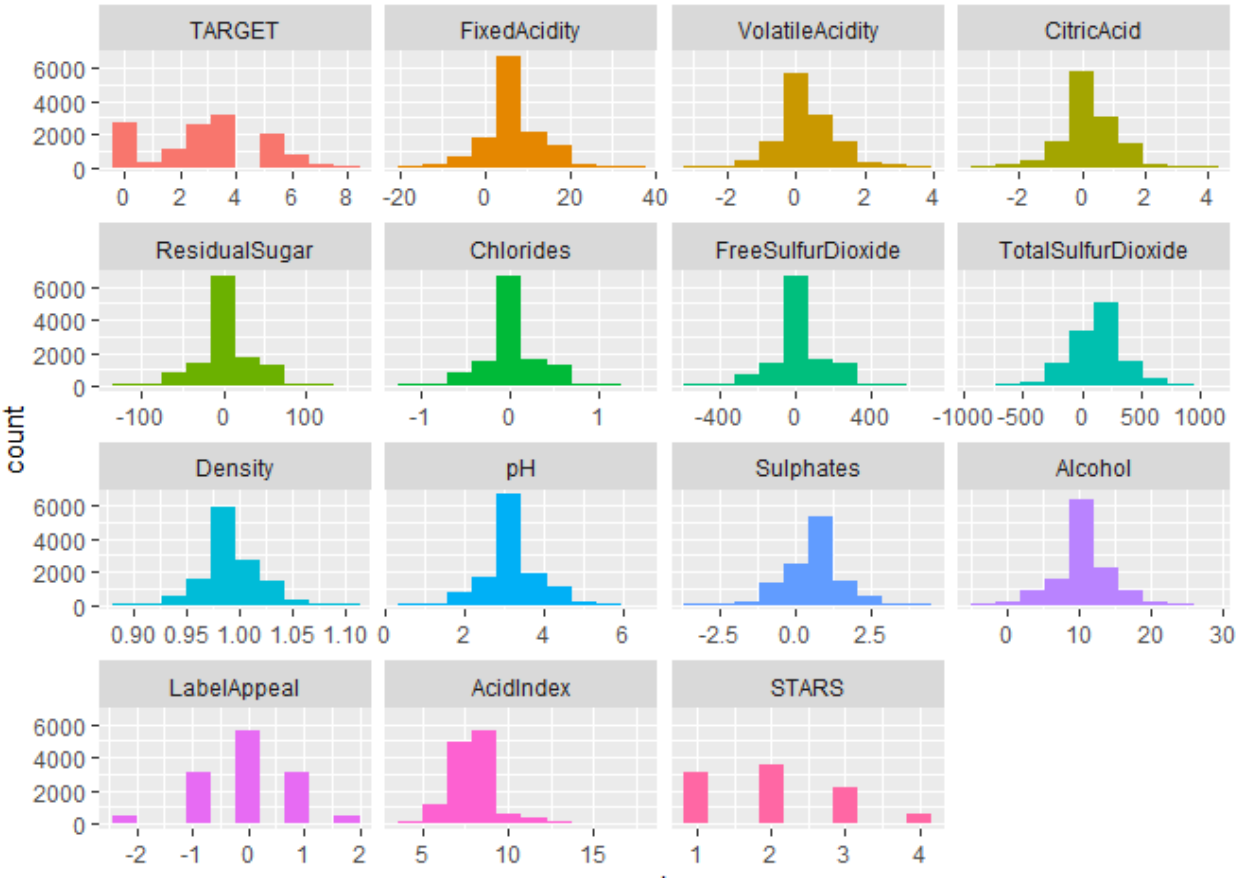
Introduction

The chemical makeup of wine can be very complex and small changes can greatly affect the taste and quality of wine. Thus, small changes in chemical make-up can adjust the probability of selling a particular case of wine. The purpose of this assignment is to utilize the various chemical components of different wines to predict the amount of boxes of wine sold.

Section 1 – Data Exploration

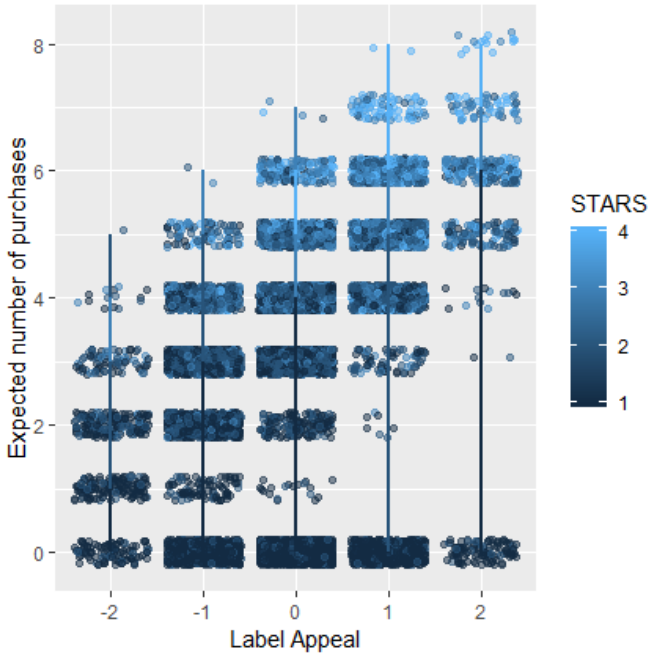
The dataset utilized contains 14 variables describing the makeup of a specific wine. Some of the variables are part of the chemical makeup of the wine while others such as label appeal and stars indicate how the wine is perceived by the drinker.

VARIABLE NAME	DESCRIPTION
ACIDINDEX	Proprietary method of testing total acidity of wine by using a weighted average
ALCOHOL	Alcohol Content
CHLORIDES	Chloride content of wine
CITRICACID	Citric Acid Content
DENSITY	Density of Wine
FIXEDACIDITY	Fixed Acidity of Wine
FREESULFURDIOXIDE	Sulfur Dioxide content of wine
LABELAPPEAL	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customers don't like the design.
RESIDUALSUGAR	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
SULPHATES	Sulfate content of wine
TOTALSULFURDIOXIDE	Total Sulfur Dioxide of Wine
VOLATILEACIDITY	Volatile Acid content of wine, an unpleasant characteristic of a wine.
PH	pH of wine



The target value can be influenced by a number of predictor values. On the right the target value is compared with the rating given by the group of experts and the associated appeal of the label. This graph shows us a few things:

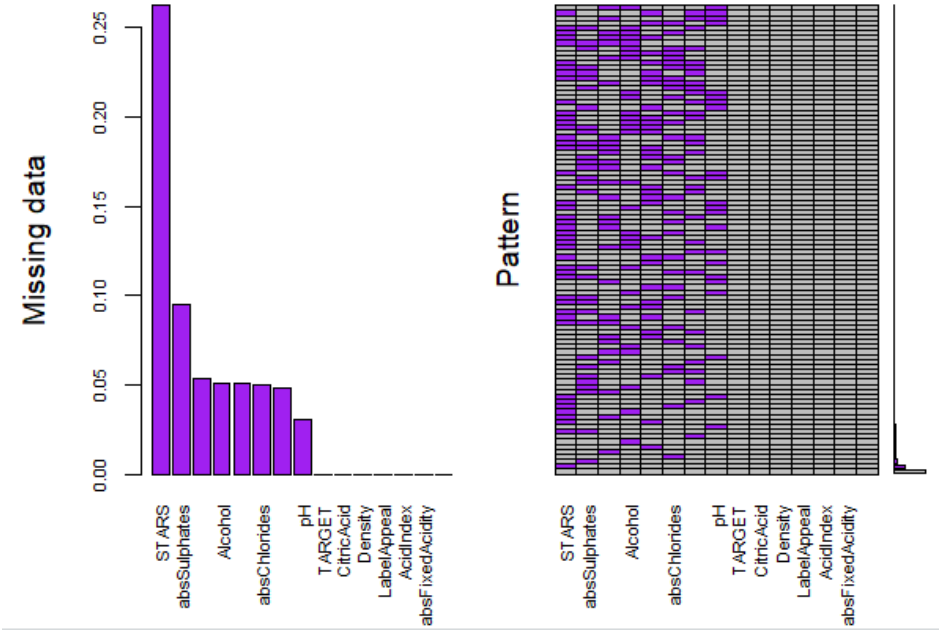
- There are many instances where no purchase was made
- If a purchase is made of a wine with an expert rating, more than 3 purchases will be made
- Both the number of stars and the appeal of the label have some amount of influence over the number of wine purchases.



Section 2 – Data Preparation

When initially individually viewing each of the variables it becomes apparent that there are many negative values that are not correct. For the purpose of this assignment, the absolute value of these variables will be utilized.

There is one other quality concern with the data that must be addressed, there is a large amount of missing data. There are eight fields that contain missing data. Stars is missing roughly 30% of its data. This could possibly be due to the human input factor of rating that not all surveyed purchasers responded.

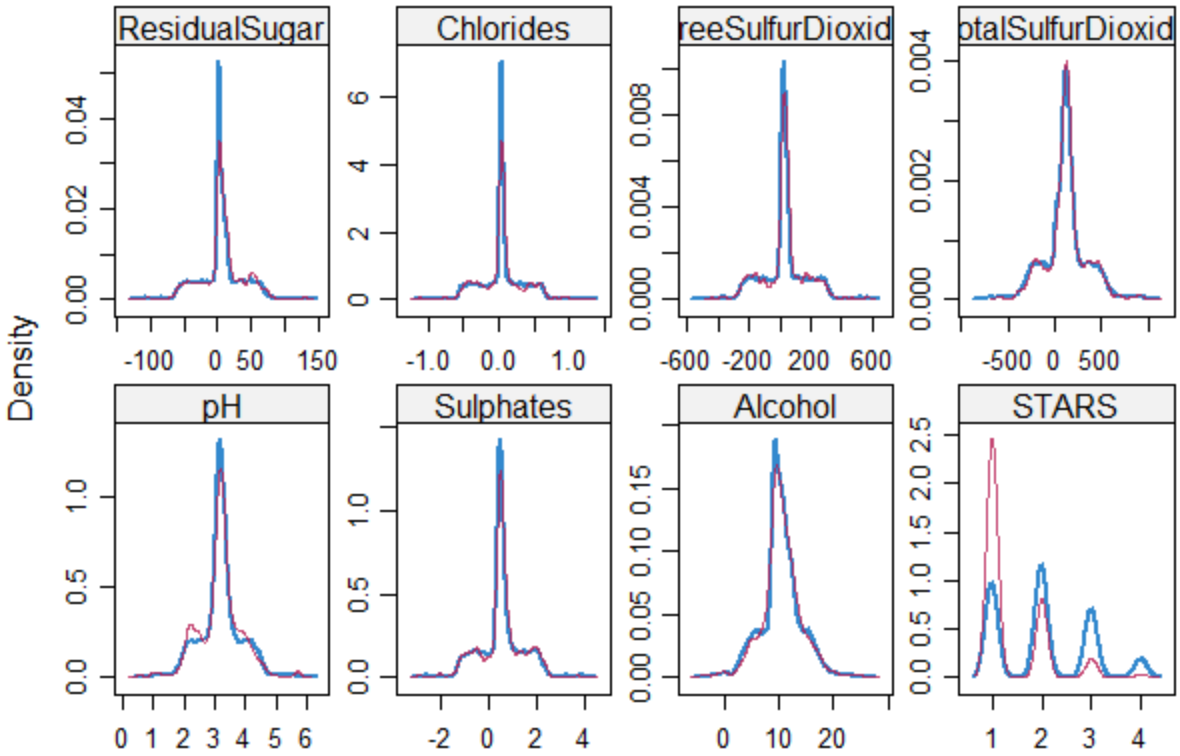


Two approaches will be utilized to determine the most accurate way to utilize the missing data. A flag will be created for each of the eight variables indicating if a value is missing or not.

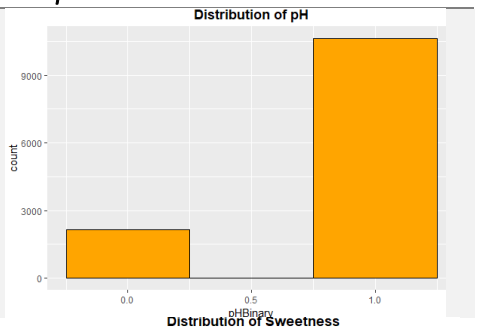
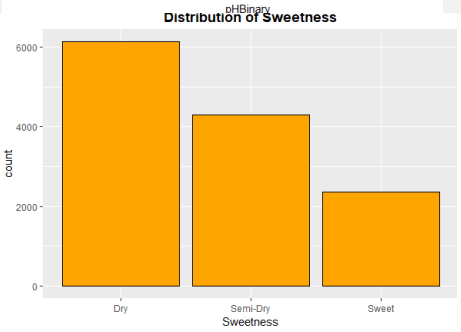
Once each of the flags are created, they will be compared to the target value. In the below graph it is evident that the only flag that has a significant correlation to the target value is the Star flag. Missing Star values could indicate that the wine is new and has not yet been reviewed by experts or that it simply is not a good enough wine to be on the radar for the experts to consider reviewing.

Variable Name	Correlation to Target
TARGET	1
ResidualSugar_Flag	0.0112
Chlorides_Flag	-0.00269
FreeSulfurDioxide_Flag	0.00015
TotalSulfurDioxide_Flag	-0.00617
pH_Flag	0.009965
Sulphates_Flag	0.012504
Alcohol_Flag	-0.00148
STARS_Flag	0.571579

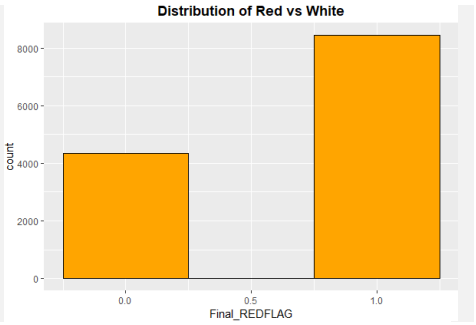
Additionally, an inputted variable will be created to attempt to predict what value would have been attributed to the wine had it not been excluded. The below histograms illustrate the original and inputted values. All values appear to be extremely similar with the only exception being Stars which appears to be heavily adding 1 star ratings. This indicates that perhaps the latter reasoning for a lack of stars may be the case, that the wines without stars are not as good of wines and therefore have not been rated.

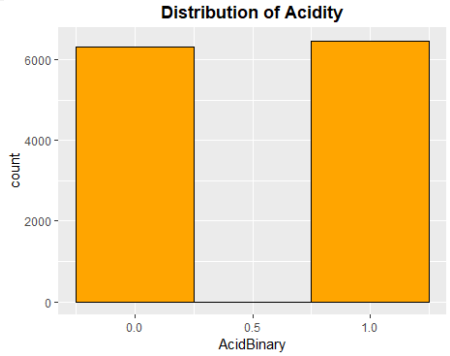


In addition to correcting for missing values a few more variable adjustments will be conducted.

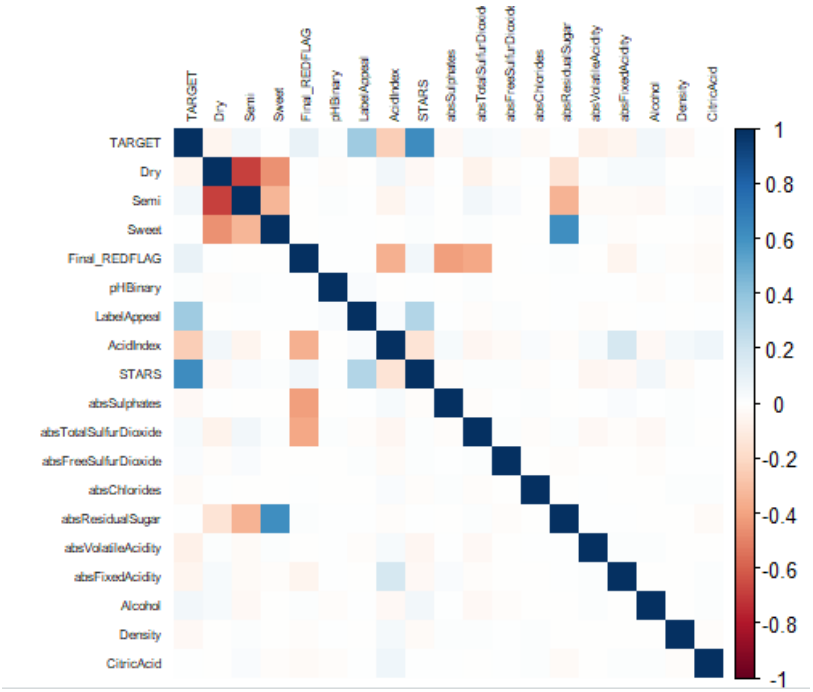
Variable Name	*Adjustment Made*	*Graph*
pHBinary	According to the website WineFolly (Puckette, 2018), the pH of most wine is between 2.5 and 4.5. A binary variable will be created to indicate if the actual amount is inside or outside this range. (See diagram in appendix)	
Sweetness	Wine can be classified as sweet or dry regardless of color. According to Wine Turtle, "Dry wines have the lowest content of sugar. Typically, you can expect dry wines to have between 0.1 - 0.3 percent residual sugar in them (or 1 - 3 grams per liter)." Wine Turtle goes on to describe semi-dry as between 10-30 grams and sweet as greater than 30. (Edison, 2019)	

Red/White	Wine can be classified into two categories, red or white. This new variable attempts to classify the wines as such utilizing the values of sulphates, sulfur dioxide and the acid index.
Acid Binary	In quantities of 0.2 to 0.4 g/L, volatile acidity does not affect a wine's quality. At higher levels, however, VA can give wine a sharp, vinegary tactile sensation, which is caused by acetic acid. Extreme volatile acidity signifies a seriously faulty wine, and can be referred to as volatile. (Hubble, 2017)





The existing and created variables will now be compared by their relationship with the target value.



An interesting point to note is that dry wines have a negative correlation to the target value while semi-dry wines have a positive and sweet wines have no correlation. This could be due to dry and sweet wines having smaller, more particular, groups of drinkers while semi-dry may be enjoyable by everyone.

Another notable observation is that the two variables not in relation to the chemical make up of the wine, label appeal and stars, appear to have the largest correlation with the target value. This could indicate that potential buyers care less about the chemical make-up of wine that "should" affect the taste, and instead are drawn more to wine by the design of the label and if a rating is given to the wine.

Section 3 – Model Creation

Now that the data quality issues have been resolved and additional variables have been created the models can be created.

Five models will be created utilizing the same predictor values. The predictor values were selected by having the highest positive or negative correlation with the target value. The values selected are Sweetness, AcidIndex, LabelAppeal and STAR_FLAG.

The five models that will be created are:

- Multiple Linear Regression Model
- Poisson Model
- Zero-Inflated Poisson Model
- Negative Binomial Regression Model
- Zero-Inflated Negative Binomial Regression

Model 1 - Multiple Linear Regression Model

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.038305   0.103715  19.653  < 2e-16 ***
SweetnessSemi-Dry  0.109374   0.028653   3.817 0.000136 ***
SweetnessSweet     0.073971   0.034835   2.123 0.033731 *
AcidIndex         -0.241643   0.009758 -24.763  < 2e-16 ***
LabelAppeal-1      0.503029   0.069050   7.285 3.41e-13 ***
LabelAppeal0       1.155580   0.067026  17.241  < 2e-16 ***
LabelAppeal1       1.846575   0.069397  26.609  < 2e-16 ***
LabelAppeal2       2.599599   0.091443  28.429  < 2e-16 ***
STAR_FLAG          2.235297   0.029525  75.708  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.437 on 12786 degrees of freedom
Multiple R-squared:  0.4437,    Adjusted R-squared:  0.4434
F-statistic:  1275 on 8 and 12786 DF,  p-value: < 2.2e-16
```

The outcome of the model further proves the previous assumptions. Initially, that stars have a major influence on the purchases of wine, for each star the experts give to a wine, 2 additional purchases are made. We can also see that the appeal of the label significantly increases the sale of wine by each additional score.

Model 2 - Poisson Model

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.39432    0.05250   7.511 5.87e-14 ***
SweetnessSemi-Dry   0.03909    0.01137   3.438 0.000585 ***
SweetnessSweet      0.02712    0.01396   1.943 0.051999 .
AcidIndex          -0.09383    0.00447 -20.993  < 2e-16 ***
LabelAppeal-1       0.29882    0.03790   7.884 3.17e-15 ***
LabelAppeal0        0.55102    0.03681  14.969  < 2e-16 ***
LabelAppeal1        0.74760    0.03720  20.097  < 2e-16 ***
LabelAppeal2        0.92848    0.04177  22.227  < 2e-16 ***
STAR_FLAG           1.04137    0.01691  61.581  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22861  on 12794  degrees of freedom
Residual deviance: 14770  on 12786  degrees of freedom
AIC: 46730
```

The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed. The deviance of this model is relatively low and the p value is as well indicating that this is a significant model.

```
  (Intercept) SweetnessSemi-Dry     SweetnessSweet          AcidIndex      LabelAppeal-1
    1.4833745         1.0398623          1.0274911          0.9104367          1.3482661
 LabelAppeal0      LabelAppeal1       LabelAppeal2          STAR_FLAG
    1.7350191         2.1119341          2.5306510          2.8331070
```

In the above model we can see that Dry has been made the base for the Sweetness and -2 has been made the base for the Label Appeal as the appeal of the label increases, the possibility for purchasing increasing by roughly 0.2 each time.

The baseline number of wine purchases is 1.48. Increasing the sweetness and acid index all increase the number of purchases by 1. Increasing the label appeal from -2 to -1 increases by 1, increasing further will increase by 2 and then 3. For each star added by a wine expert, the number of wine purchases is increased by 3.

Model 3 - Zero-Inflated Poisson Model

```
Count model coefficients (poisson with log link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.582735   0.055679  10.466   <2e-16 ***
SweetnessSemi-Dry -0.006954   0.011731  -0.593    0.553
SweetnessSweet     0.001761   0.014384   0.122    0.903
AcidIndex         -0.026145   0.004908  -5.327    1e-07 ***
LabelAppeal-1      0.458229   0.040612  11.283   <2e-16 ***
LabelAppeal0       0.788282   0.039484  19.964   <2e-16 ***
LabelAppeal1       1.025366   0.039882  25.710   <2e-16 ***
LabelAppeal2       1.218916   0.044229  27.559   <2e-16 ***
STAR_FLAG          0.175032   0.018474   9.474   <2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.19826    0.40700 -12.772  < 2e-16 ***
SweetnessSemi-Dry -0.44344    0.07855  -5.645 1.65e-08 ***
SweetnessSweet    -0.23772    0.09169  -2.593  0.00952 **
AcidIndex          0.44791    0.02423  18.482  < 2e-16 ***
LabelAppeal-1      1.57521    0.35778   4.403 1.07e-05 ***
LabelAppeal0       2.35708    0.35468   6.646 3.02e-11 ***
LabelAppeal1       2.78800    0.35864   7.774 7.61e-15 ***
LabelAppeal2       2.94279    0.38570   7.630 2.35e-14 ***
STAR_FLAG         -3.59230    0.08046 -44.648  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 27
Log-likelihood: -2.103e+04 on 18 Df
```

This model will take into account zeros. For example, there are some bottles of wine that will not be bought for one reason or another.

The baseline odds of people not buying a bottle of wine is .006. These odds are decreased by increasing the sweetness of the wine to semi-dry by 0.64. The variables that decrease the odds of not purchasing the most are the variables associated with a very eye catching label.

	Count_model	Zero_inflation_model
Intercept	1.7909304	0.005526188
Semi-Dry	0.9930704	0.641823175
Sweet	1.0017622	0.788420167
Acid Index	0.9741943	1.565040747
Label -1	1.5812719	4.831741192
Label 0	2.1996138	10.560023466
Label 1	2.7881169	16.248479777
Label 2	3.3835177	18.968648660
Star Flag	1.1912845	0.027534836

Model 4 - Negative Binomial Regression Model

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.39434    0.05250   7.511 5.87e-14 ***
SweetnessSemi-Dry  0.03909    0.01137   3.438 0.000585 ***
SweetnessSweet     0.02712    0.01396   1.943 0.052005 .
AcidIndex         -0.09383    0.00447 -20.993  < 2e-16 ***
LabelAppeal-1      0.29882    0.03790   7.884 3.18e-15 ***
LabelAppeal0       0.55102    0.03681  14.969  < 2e-16 ***
LabelAppeal1       0.74760    0.03720  20.097  < 2e-16 ***
LabelAppeal2       0.92847    0.04177  22.226  < 2e-16 ***
STAR_FLAG          1.04137    0.01691  61.579  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(36631.54) family taken to be 1)

    Null deviance: 22860  on 12794  degrees of freedom
Residual deviance: 14770  on 12786  degrees of freedom
AIC: 46733

Number of Fisher Scoring iterations: 1


              Theta:  36632
          Std. Err.:  33106
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -46712.64


  (Intercept) SweetnessSemi-Dry    SweetnessSweet         AcidIndex     LabelAppeal-1
    1.4834028        1.0398641        1.0274918         0.9104348         1.3482642
LabelAppeal0     LabelAppeal1     LabelAppeal2         STAR_FLAG
    1.7350140        2.1119238        2.5306385         2.8331035
```

The baseline number of wine bottles purchased is 1.43. This is increased by 1 by increasing the sweetness of the wine, the acid index or a slightly better label. If the label is further made more appealing it can increase the odds of a purchase by 3. The same odds increased is observed when the wine is rated by an expert. These coefficients are extremely similar to those calculated with the Poisson model.

Model 5 - Zero-Inflated Negative Binomial Regression

```
Count model coefficients (negbin with log link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.582771   0.055679  10.467  < 2e-16 ***
SweetnessSemi-Dry -0.006954   0.011732  -0.593 0.553362
SweetnessSweet     0.001761   0.014384   0.122 0.902566
AcidIndex         -0.026145   0.004908  -5.327    1e-07 ***
LabelAppeal-1      0.458197   0.040611  11.283  < 2e-16 ***
LabelAppeal0       0.788250   0.039483  19.964  < 2e-16 ***
LabelAppeal1       1.025336   0.039881  25.710  < 2e-16 ***
LabelAppeal2       1.218888   0.044229  27.559  < 2e-16 ***
STAR_FLAG          0.175026   0.018474   9.474  < 2e-16 ***
Log(theta)        12.131910   3.663443   3.312 0.000928 ***

Zero-inflation model coefficients (binomial with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.19752    0.40682 -12.776  < 2e-16 ***
SweetnessSemi-Dry -0.44345    0.07855  -5.645 1.65e-08 ***
SweetnessSweet    -0.23773    0.09169  -2.593  0.00952 **
AcidIndex          0.44791    0.02424  18.482  < 2e-16 ***
LabelAppeal-1      1.57444    0.35757   4.403 1.07e-05 ***
LabelAppeal0       2.35634    0.35447   6.647 2.98e-11 ***
LabelAppeal1       2.78729    0.35843   7.776 7.46e-15 ***
LabelAppeal2       2.94208    0.38551   7.632 2.32e-14 ***
STAR_FLAG         -3.59237    0.08046 -44.647  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 185704.0454
Number of iterations in BFGS optimization: 1
Log-likelihood: -2.103e+04 on 19 Df
```

With this model, the baseline number of wine bottles purchased is 1.79 and the baseline of bottles not purchased is .006. This appears to be identical to the Zero-Inflated Poisson Model results. The addition of additional sweetness increases the odds of not selling wine but becomes less significant the sweeter the wine becomes.

	Count_model	Zero_inflation_model
Intercept	1.7909304	0.005526188
Semi-Dry	0.9930704	0.641823175
Sweet	1.0017622	0.788420167
Acid Index	0.9741943	1.565040747
Label -1	1.5812719	4.831741192
Label 0	2.1996138	10.560023466
Label 1	2.7881169	16.248479777
Label 2	3.3835177	18.968648660
Star Flag	1.1912845	0.027534836

Model Comparison

To compare the Poisson model with the Zero-Inflated Poisson model, we can use the Vuong test.

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                  -42.84576 model2 > model1 < 2.22e-16
AIC-corrected        -42.67970 model2 > model1 < 2.22e-16
BIC-corrected        -42.06056 model2 > model1 < 2.22e-16
```

In the above output we can see that model 2 (the ordinary Poisson regression model) is significant. This indicates that the zero-inflated model is a better model than the ordinary Poisson model.

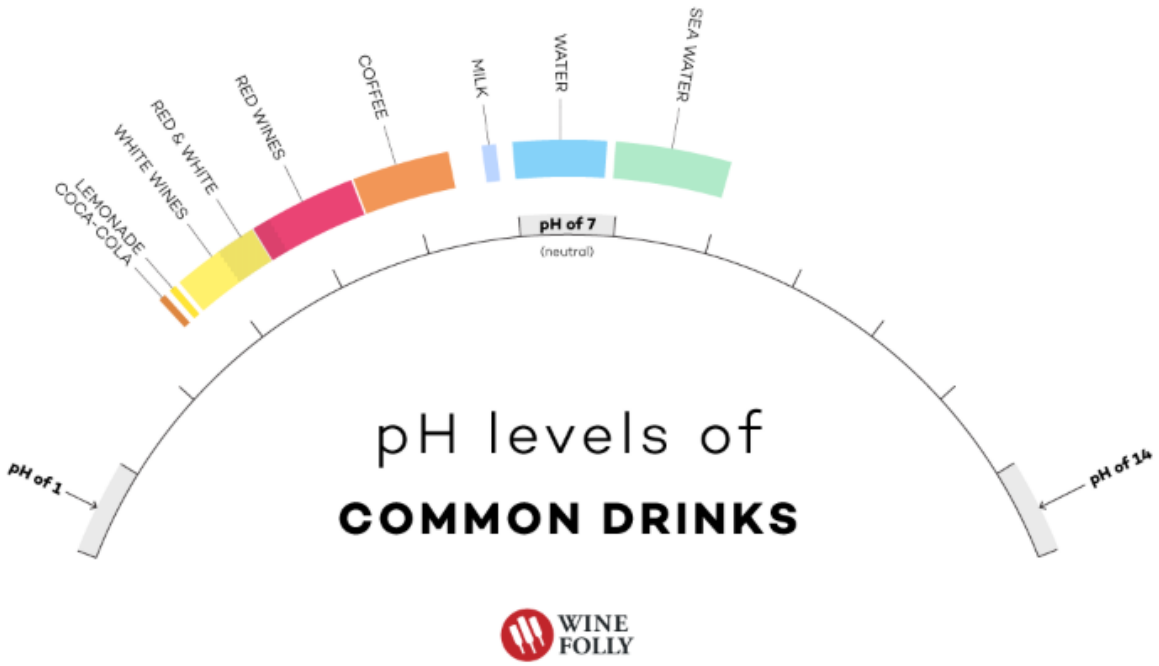
Method	*Model 1*	*Model 2*	*Model 3*	*Model 4*	*Model 5*
AIC	45603.45	46730.28	42104.04	46732.64	42106.19
ROC	.5596	.5584	.5913	.5584	.5913
R Squared	.444	.354		.354	
Log Likelihood	-22792	-23356	-21034	-23356	-21034

Given the above statistics and the fact that the target value has so many zero values, the 3rd model, the Zero-Inflated Poisson Model will be utilized.

Conclusion

I've heard it said in the past to like the type of wine you like and to not let "expert" opinions influence what wine you think you should like. This experiment proved that the opposite of this tends to be the case, that wine purchases are often made by other people's preferences in wine. The results of this study are also interesting because they would be particularly helpful in the marketing of the wine. Simply ensuring that a competent graphic designer is hired for the label design and the wine garners enough influence to be rated by an expert can guarantee sales.

Appendix



Works Cited

Puckette, M. (2018). *Understanding Acidity in Wine* | *Wine Folly*. [online] Wine Folly. Available at: https://winefolly.com/review/understanding-acidity-in-wine/.

Edison, T. (2019). *Does White Wine Have More Sugar Than Red? Or Vice Versa?* [online] Wine Turtle. Available at: https://www.wineturtle.com/white-red-more-sugar/.

Hubble, G. (2017). *Volatile Acidity* [online] Wine Guy. Available at: https://www.wineguy.co.nz/index.php/glossary-articles-hidden/482-volatile-acidity.