Homework 5 Assignment

Talha Muhammad

May 5, 2018

Analysis of the Wine Sales Dataset

Dataset Description

The dataset is a wine sales and associated attributes dataset that contains the number of cases sold of a particular wine as the forecast variable. Other variables are the attributes of the particular wine related to taste or acidity. It also includes marketing related variables such as the attractiveness of the label and expert reviews of the wine.

variable	description	type
TARGET	Number of cases purchased	Original
FixedAcdity	Fixed Acidity of wine	Original
VolatileAcidity	Volatile Acidity of wine	Original
CitricAcid	Citric Acidity of wine	Original
ResidualSugar	Residual Sugar of wine	Original
Chlorides	Chloride content of wine	Original
FreeSulfurDioxide	Free Sulphur Dioxide content	Original
TotalSulfurDioxide	Total Sulphur Dioxide	Original
Density	Density of Wine	Original
pH	pH of wine	Original
Sulphates	Sulphate Content of Wine	Original
Alcohol	Alcohol content	Original
LabelAppeal	Marketing score of the label	Original
AcidIndex	Acid Index of wine	Original
Stars	Wine Rating by team of experts	Original
F_ACID_N_FLAG	Flag indicating original negative value converted to positive	Calculated
V_ACID_N_FLAG	Flag indicating original negative value converted to positive	Calculated
C_ACID_N_FLAG	Flag indicating original negative value converted to positive	Calculated
Rsugar_N_FLAG	Flag indicating original negative value converted to positive	Calculated
chl_N_FLAG	Flag indicating original negative value converted to positive	Calculated
free_sox_N_FLAG	Flag indicating original negative value converted to positive	Calculated
tot_sox_N_FLAG	Flag indicating original negative value converted to positive	Calculated
sulphates_N_FLAG	Flag indicating original negative value converted to positive	Calculated
alcohol_N_FLAG	Flag indicating original negative value converted to positive	Calculated
ResidualSugar_I_FLAG	Flag indicating imputed value	Calculated
chlorides_I_FLAG	Flag indicating imputed value	Calculated
free_sox_I_FLAG	Flag indicating imputed value	Calculated
tot_sox_I_FLAG	Flag indicating imputed value	Calculated
ph_I_FLAG	Flag indicating imputed value	Calculated
suphates_I_FLAG	Flag indicating imputed value	Calculated
alcohol_I_FLAG	Flag indicating imputed value	Calculated
stars_I_FLAG	Flag indicating imputed value	Calculated

The table below shows all the variable names in the dataset and also identifies whether the variables are original in the dataset or have been computed as part of the analysis. In particular, a number of variables

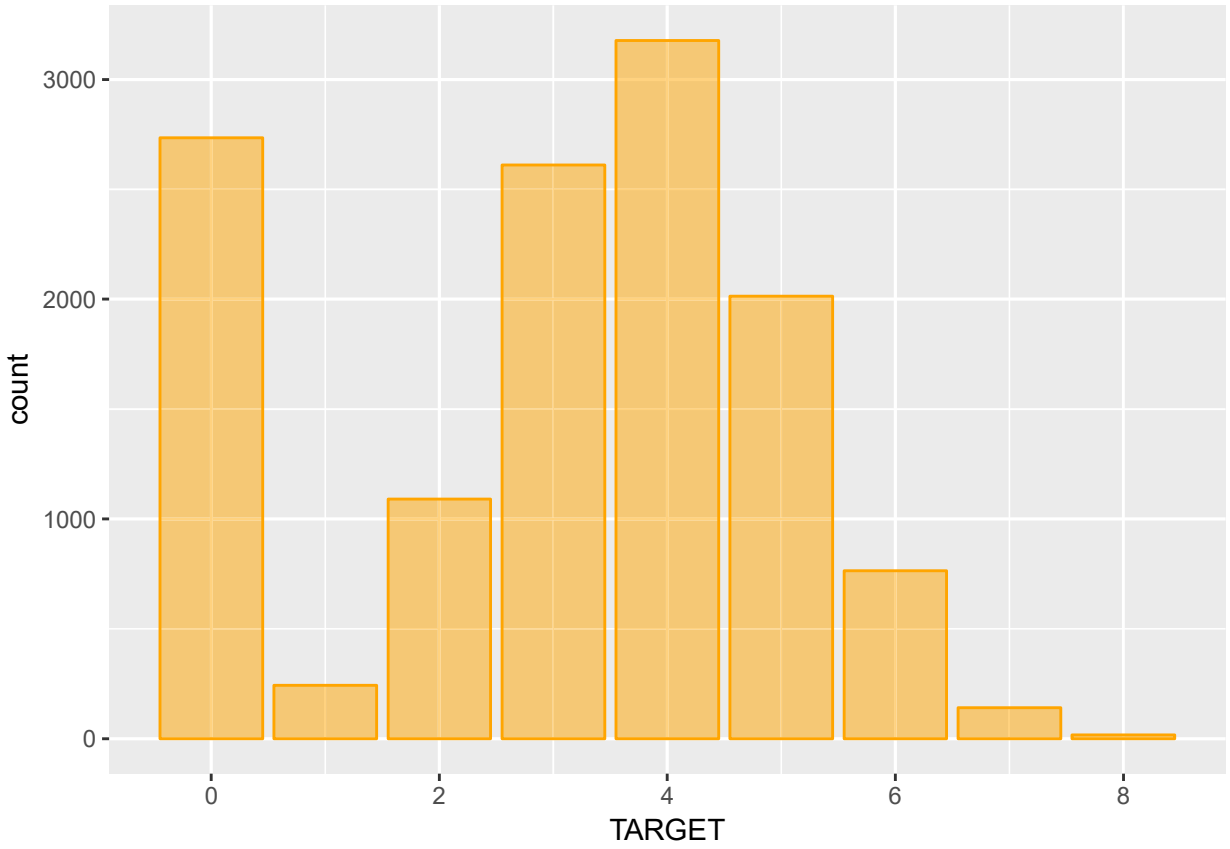
are computed which are not part of the original dataset. These variables are computed as flags and are used as identification variables, identifying which varables were imputed or changed.

1. DATA EXPLORATION (25 Points)

Describe the size and the variables in the wine training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas. a. Mean / Standard Deviation / Median b. Bar Chart or Box Plot of the data c. Is the data correlated to the target variable (or to other variables?) d. Are any of the variables missing and need to be imputed "fixed"?

Solution

The plot below shows the distribution of **TARGET**, the number of cases sold. The plot shows a bimodal distribution, with a high count of zeros, and then another mode at 4. Excluding the zero values the distribution does resemble a normal, but primarily consists of count variables. Additional, the target is composed entirely of integer values.



The tables below show the means and medians of the data. Means and medians are calculated such that any missing NA values are excluded when calculating the metrics. For a few variables there is a large difference between the mean values adn median values. These include **FreeSulfurdioxide**, **ResidualSugar**, **TotalSulfulurdioxide**, and **Sulphates**.

Means

TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
3.03	7.08	0.32	0.31	5.42	0.05	30.85	120.71	0.99	3.21	0.53	10.49	-0.01	7.77	2.04

Medians

TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
3	6.9	0.28	0.31	3.9	0.05	30	123	0.99	3.2	0.5	10.4	0	8	2

We do some further analysis on the data, to understand the missing values in the data. The tables below show the counts and proportions of missing values in the dataset. For most of the variables only about 5 percent of the data are missing. There are a few exceptions thought, for example for **Sulphates** about 10 percent of the data are missing and for rating of the wine, measured by **STARS** about 26 percent of the data are missing.

Category	Counts							
Complete / Not Missing	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	pH	Sulphates	Alcohol	STARS
FALSE	616	638	647	682	395	1210	653	3359
TRUE	12179	12157	12148	12113	12400	11585	12142	9436

Proportions of Missing							
ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	pH	Sulphates	Alcohol	STARS
0.048	0.05	0.051	0.053	0.031	0.095	0.051	0.263
0.952	0.95	0.949	0.947	0.969	0.905	0.949	0.737

The difference between median and mean values of the data invites some further investigation. It turns out that a number of variables have negative values. Given, many of these represent real wine qualities, such **CitricAcid** it is not concievable that these quantities can actually be negative. Two possibilities exist,
1) The data are actually positive and have a negative number by mistake or through data errors
2) The negative values are not reliable and need to be treated as missing values

The tables below show the absolute number and proportion of negative values in the dataset. For a number of the variables, the proportion of negative values is quite high, i.e. above 20 percent. For only **Alcohol** and **FixedAcidity** are the proportions of negative values less than 13 percent.

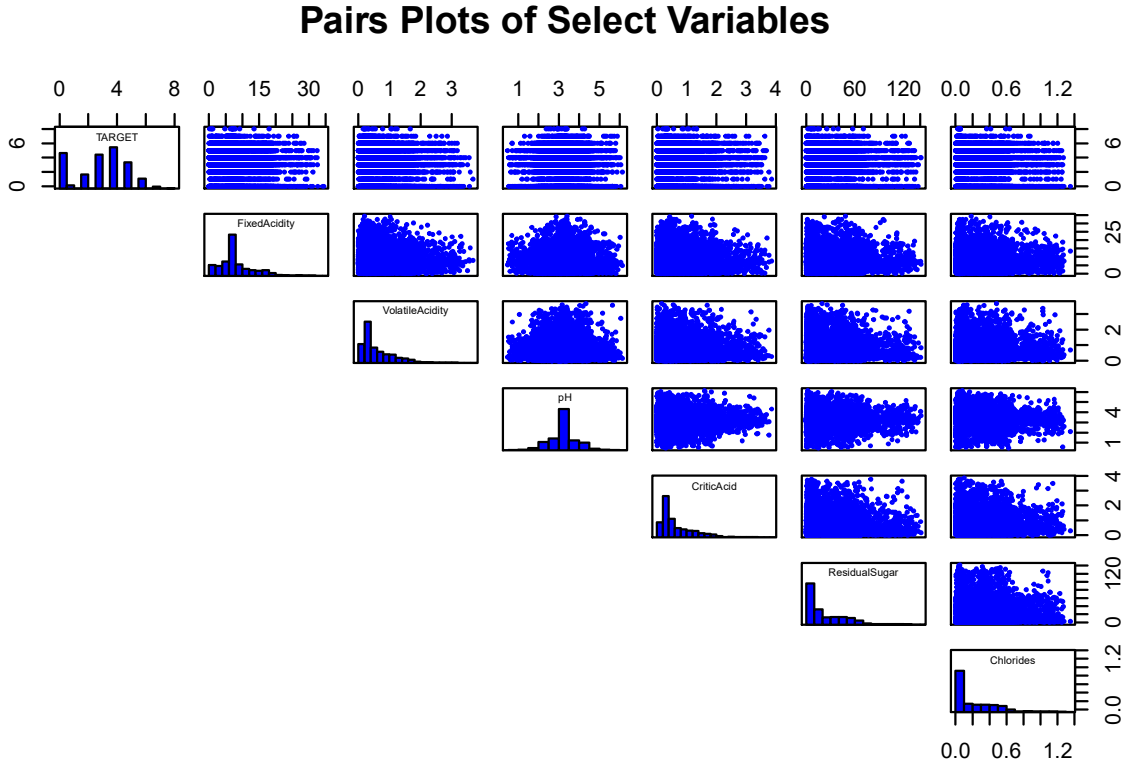
Using the approach in **2)** above would require potentially losing information, since the absolute values are likely indicative. Therefore I use the approach in **1)** to deal with the missing values.

Category	Counts of Negative Values								
Negative	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulphurdioxide	TotalSulphurdioxide	Sulphates	Alcohol
FALSE	11174	9968	9829	9043	8960	9112	9609	9224	12024
TRUE	1621	2827	2966	3136	3197	3036	2504	2361	118

Proportions of Negative Values								
FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulphurdioxide	TotalSulphurdioxide	Sulphates	Alcohol
0.873	0.779	0.768	0.743	0.737	0.75	0.793	0.796	0.99
0.127	0.221	0.232	0.257	0.263	0.25	0.207	0.204	0.01

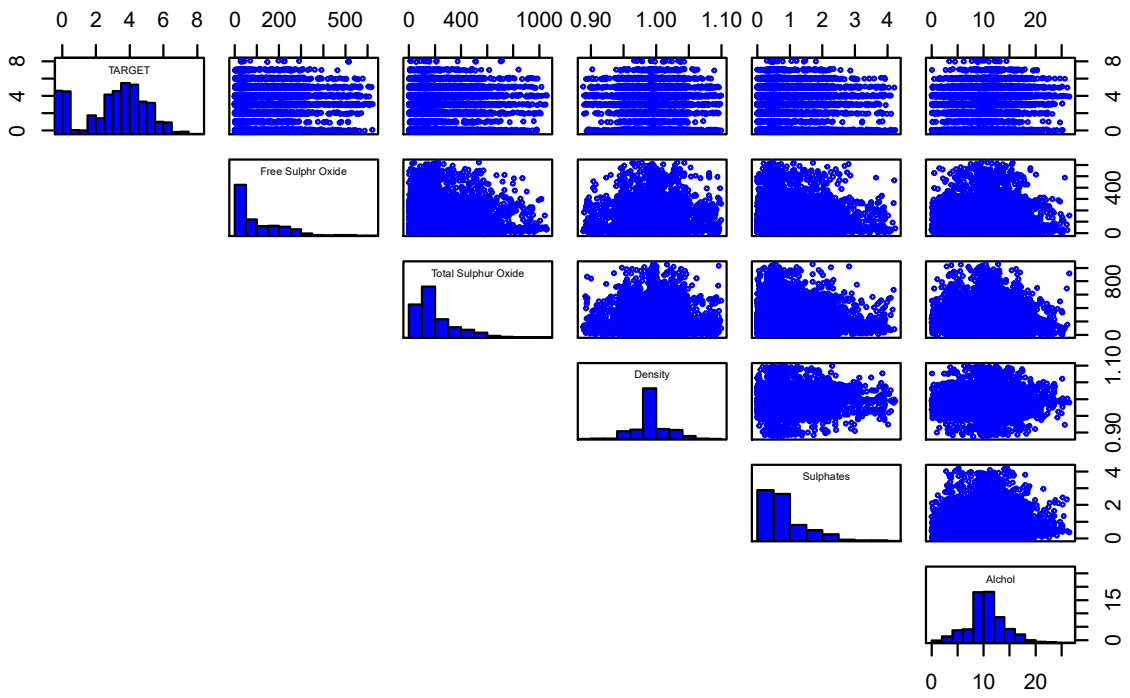
The figure below shows pair plots of the wine database. Only the original variables are shown, without any imputation of missing values. However, variables that have negative values are converted to absolute values.

The first plot shows a relationship between **FixedAcidity**, **VolatileAcidity**, **pH**, **Citric Acid**, **Residual Sugar** and **Chlorides**. The diagonals show histograms of the data variables. Except for **pH** which has the most normal-like distribution, the other variables have rather skewed distributions. There none very strong patterns in the data however, it does seem that high sale wines tend to have higher acidity - i.e. lower pH; although most wines have pH in the range of 2-4.

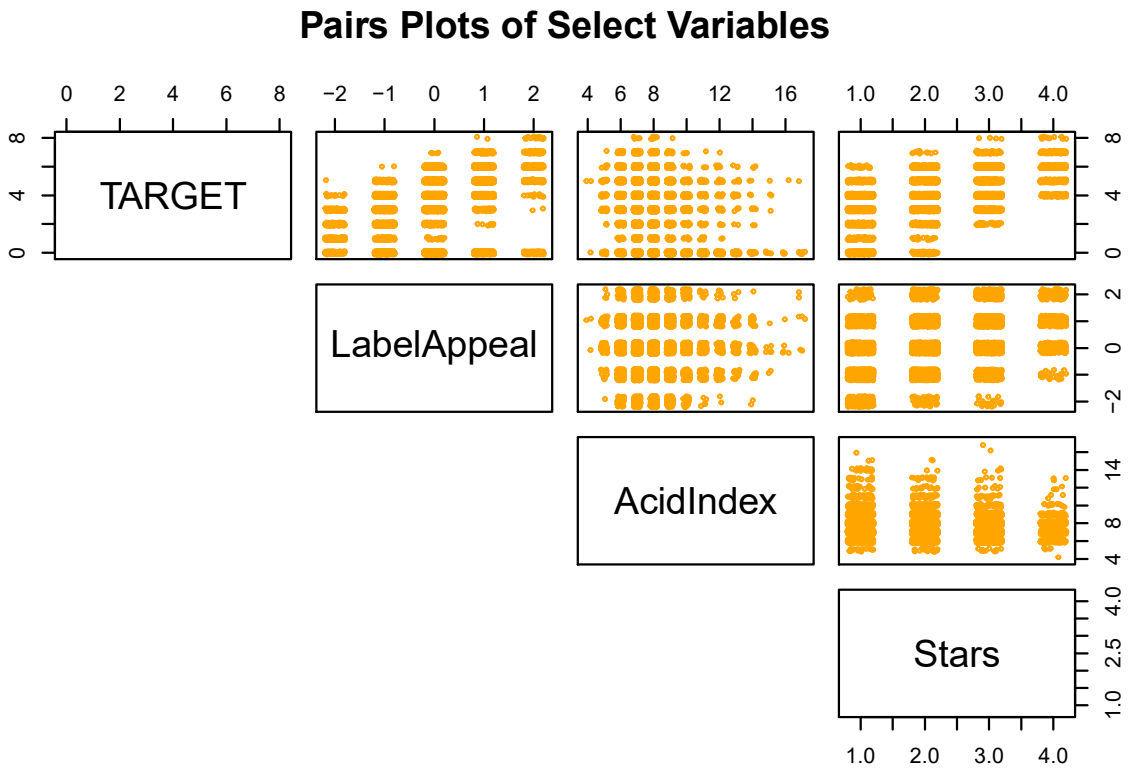


The pair plots below show the relationship between **TARGET** and additonal variablesi in the dataset, including variables capturing the sulphur content of the wine. It does seem that low sulpher content wines are more popular but the relationship does seem weak. Alcohol content also does seem to play an interesting role with some very high alcohol content wines being relatively popular, but most wines with medium alcohol content being most popular.

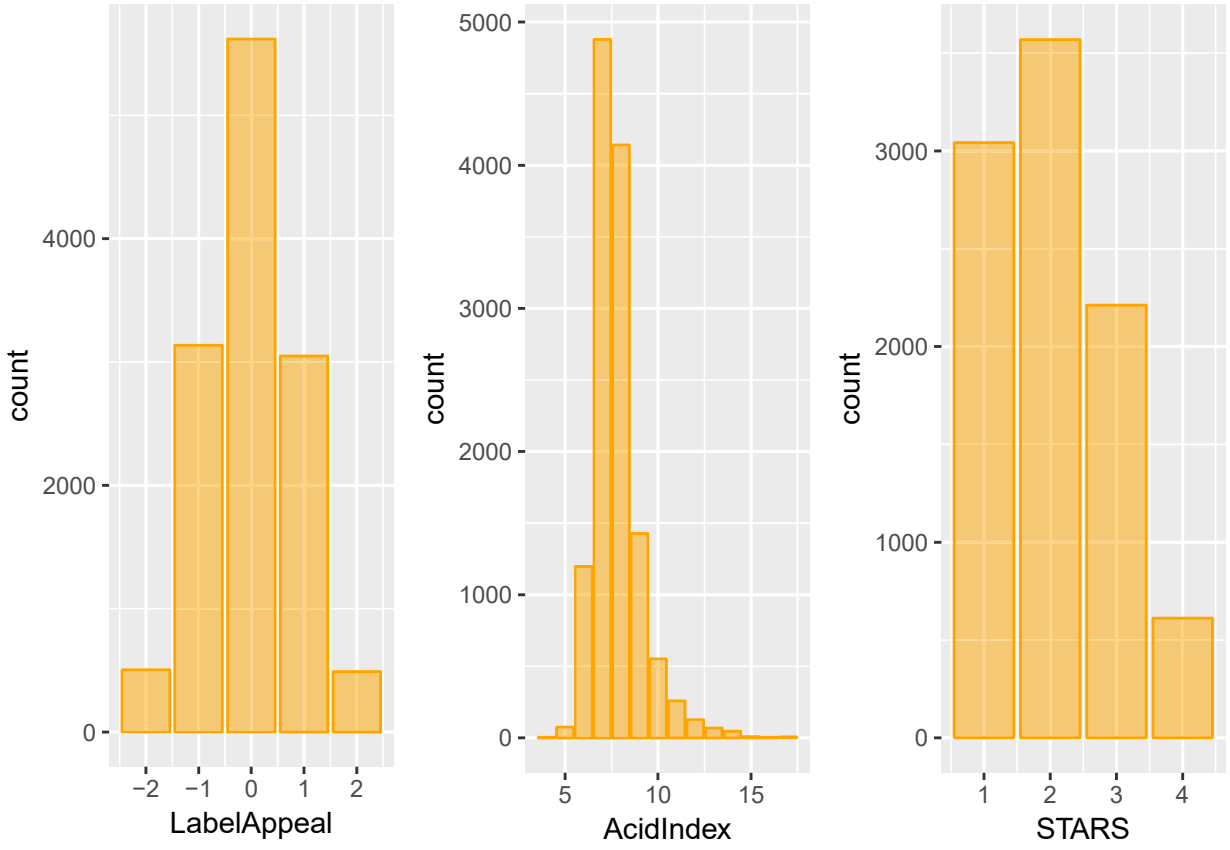
Pairs Plots of Select Variables



The plots below some addtional plots of select variables. From the plots it is clear that amongst the strongest predictors of wine sales are the label appeal and the star rating of the wine. The star rating accounts for the full taste of the wine, with highest rated wines generally accounting for the greatest number of sales.



Bar plots of **LabelAppead**, **AcidIndex** and **STARS** are shown below. Missing values are excluded from the analysis. Most wines seem to have **AcidIndex** of 7 or 8. Label appeal of most wines is average with only a few wines being rated 4-star.



2. DATA PREPARATION (25 Points)

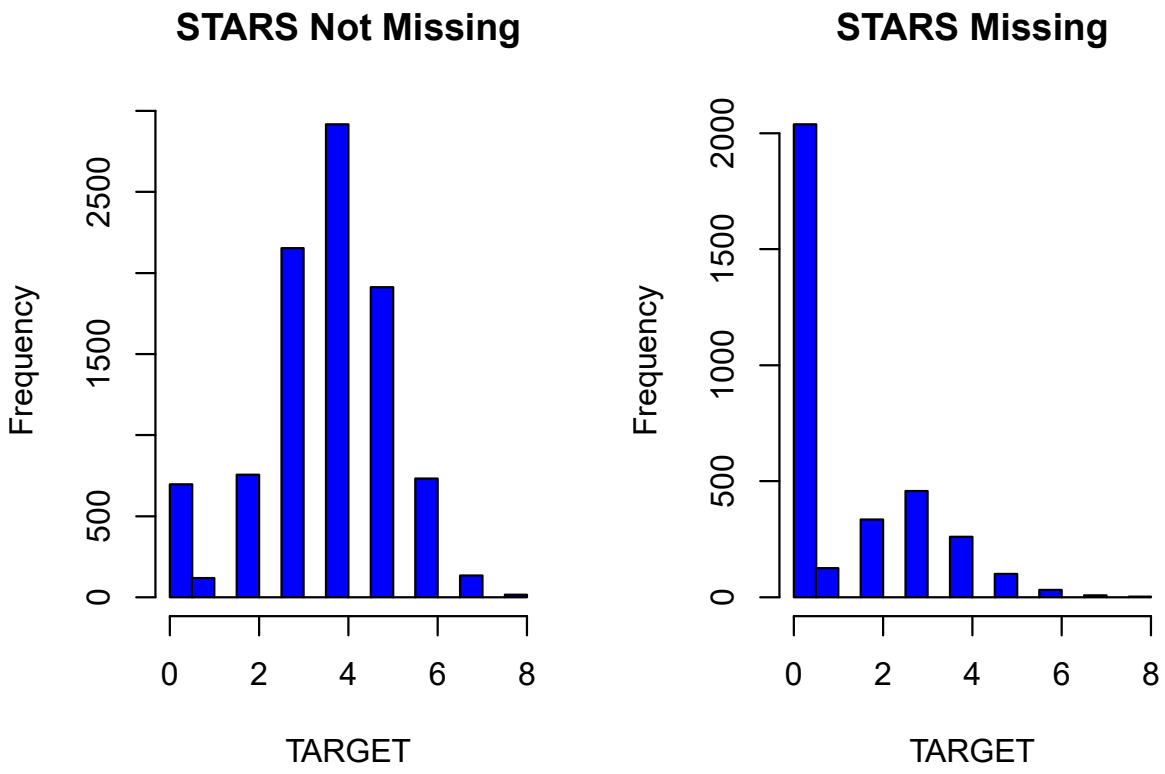
Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

Solution

As a first step we convert the negative values in the data as discussed above, to absolute values. We also generate flags for the variables that have been transformed. Code for this analysis can be found in the appendix.

As a second step, we impute missing values in the dataset. For most of the variables shown below we impute the missing values as the median of the dataset. The medians are calculated after transformations of the negative values in the data. The code can be found in the **appendix**.

While for most of the variables using the median to impute the data is sufficient in the case of the **STARS** variable, distribution of sales for missing STARS data and non-misssing stars data is significantly different as shown in the table below. Stars not missing data is almost normally distributed relative to sales. For missing stars data, a number of wines have zero sales - this indicates that these wines are not well known which is why the expert panel did not rate them.



Imputation code for the stars variable is shown below. We first generate a flag variable to track all missing stars data. To impute missing stars variable, we conduct a two step process. We first look at the median value of all sales i.e. **TARGET** that have missing values. The median value of **TARGET** is zero. For the missing values, we impute the median number of stars of the **TARGET** with zero sales.

```
#STARS
wine$stars_I_FLAG<-0
wine$stars_I_FLAG[is.na(wine$STARS)]<-1
#median TARGET for NA / Missing stars
median(wine$TARGET[is.na(wine$STARS)])
```

```
## [1] 0
```

```
#Impute median for missing stars
wine$STARS[is.na(wine$STARS)]<-median(wine$STARS[wine$TARGET==0],na.omit(TRUE))
```

3. Build Models (25 Points)

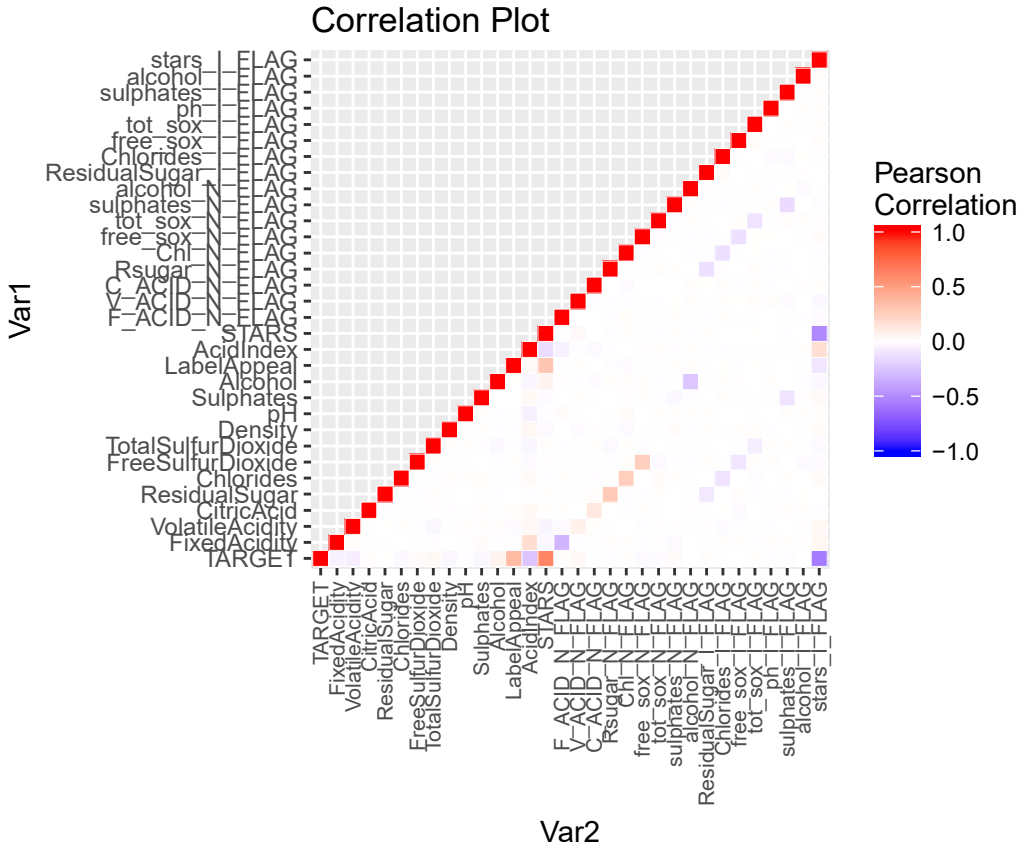
Using the training data set, build at least two different poisson regression models, at least two different negative binomial regression models, and at least two multiple linear regression models, using different variables (or the same variables with different transformations). Sometimes poisson and negative binomial regression models give the same results. If that is the case, comment on that. Consider changing the input variables if that occurs so that you get different models. Although not covered in class, you may also want to consider building zero-inflated poisson and negative binomial regression models. You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models,

do they make sense? In this case, about the only thing you can comment on is the number of stars and the wine label appeal. However, you might comment on the coefficient and magnitude of variables and how they are similar or different from model to model. For example, you might say "pH seems to have a major positive impact in my poisson regression model, but a negative effect in my multiple linear regression model". Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

Solution

The plot below shows a correlation plot of variables included in the analysis. The variables included are all the original variables as well as all the computed variables. It is clear that most of the variables do not exhibit high correlation with the target variable and do some very low correlation between themselves i.e multicollinearity does not seem to be a big problem for this database.

Only a few variables have strong correlations with the **TARGET** variable. These include **LabelAppeal**, **AcidIndex** and **Stars**. It does seem that **STARS** and **LabelAppeal** have some correlation (0.3) but not extremely high to pose a multicollinearity problem.



Sorted correlations are shown in the table below. **AcidIndex** and the imputed stars variables **STARS_I_FLAG** have the highest negative correlations while **Alcohol**, **LabelAppeal** and **STARS** have the highest positive correlations.

Sorted Correlations	
	x
stars_I_FLAG	-0.5716
AcidIndex	-0.2460
VolatileAcidity	-0.0702
FixedAcidity	-0.0530
Density	-0.0355
Sulphates	-0.0318
Chlorides	-0.0278
free_sox_N_FLAG	-0.0206
tot_sox_N_FLAG	-0.0186
sulphates_I_FLAG	-0.0125
ph_I_FLAG	-0.0100
Rsugar_N_FLAG	-0.0097
pH	-0.0093
alcohol_N_FLAG	-0.0002
free_sox_I_FLAG	-0.0002
alcohol_I_FLAG	0.0015
ResidualSugar	0.0025
Chlorides_I_FLAG	0.0027
C_ACID_N_FLAG	0.0032
Chl_N_FLAG	0.0060
tot_sox_I_FLAG	0.0062
ResidualSugar_I_FLAG	0.0112
F_ACID_N_FLAG	0.0124
CitricAcid	0.0140
sulphates_N_FLAG	0.0204
FreeSulfurDioxide	0.0238
TotalSulfurDioxide	0.0338
V_ACID_N_FLAG	0.0346
Alcohol	0.0617
LabelAppeal	0.3565
STARS	0.6238
TARGET	1.0000

Simple Poisson Model We develop a simple poission model with **STARS**, **LabelAppeal**, **STARS_I_FLAG**, **Alcohol** and **AcidIndex**. Code to generate the model is shown below. A more expanded model is also included that **Volatile Acidity** and **TotalSulfurDioxide**.

```
#Develop a simple Poisson Model
m1<-glm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex,family=poisson,wi
m2<-glm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidity+T
```

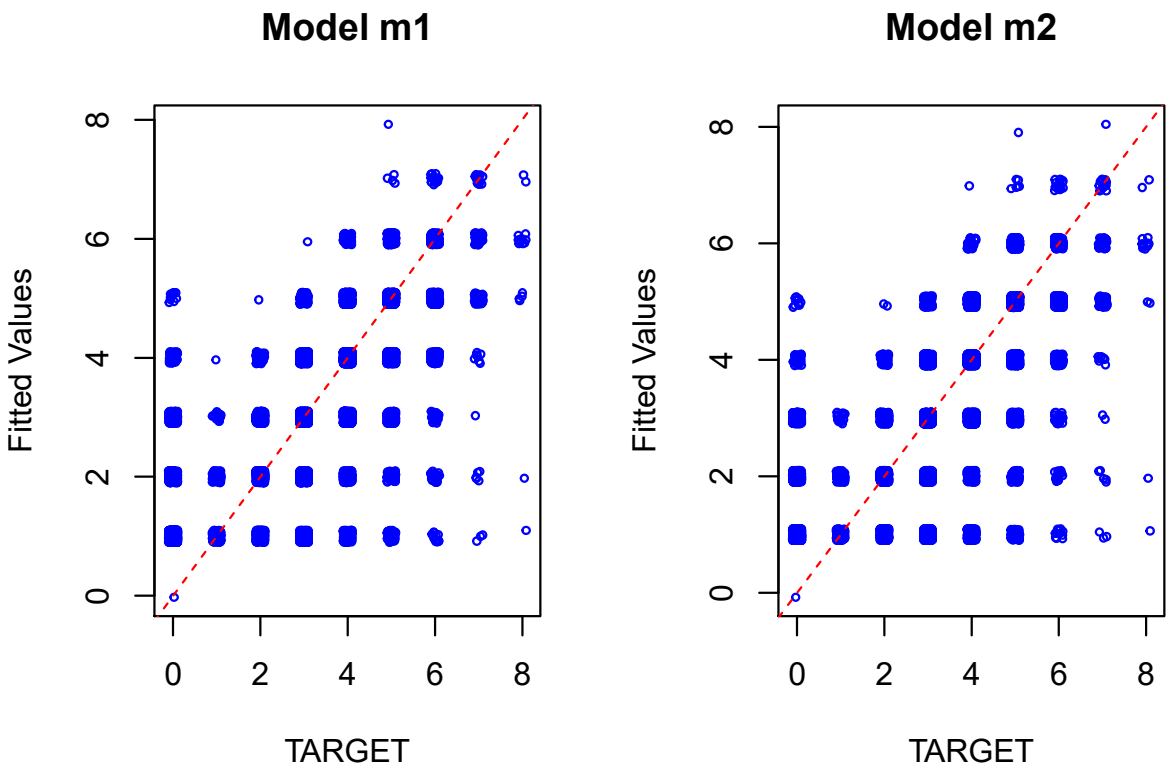
As can be seen in the tables below the model results are significant and standard errors low. Variance inflation factors are reasonable - with only label appeal showing slightly high VIF's .

Variables	Coefficient Statistics Model M2			
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1614733	0.0528065	21.994880	0.0000000
as.factor(STARS)2	0.3212627	0.0143484	22.390124	0.0000000
as.factor(STARS)3	0.4421153	0.0156030	28.335321	0.0000000
as.factor(STARS)4	0.5614331	0.0216440	25.939449	0.0000000
as.factor(LabelAppeal)-1	0.2373171	0.0379821	6.248130	0.0000000
as.factor(LabelAppeal)0	0.4268241	0.0370460	11.521475	0.0000000
as.factor(LabelAppeal)1	0.5598454	0.0376818	14.857174	0.0000000
as.factor(LabelAppeal)2	0.6966733	0.0424210	16.422845	0.0000000
stars_I_FLAG	-0.7709995	0.0195296	-39.478426	0.0000000
Alcohol	0.0038842	0.0014439	2.690075	0.0071436
AcidIndex	-0.0808400	0.0044918	-17.997052	0.0000000

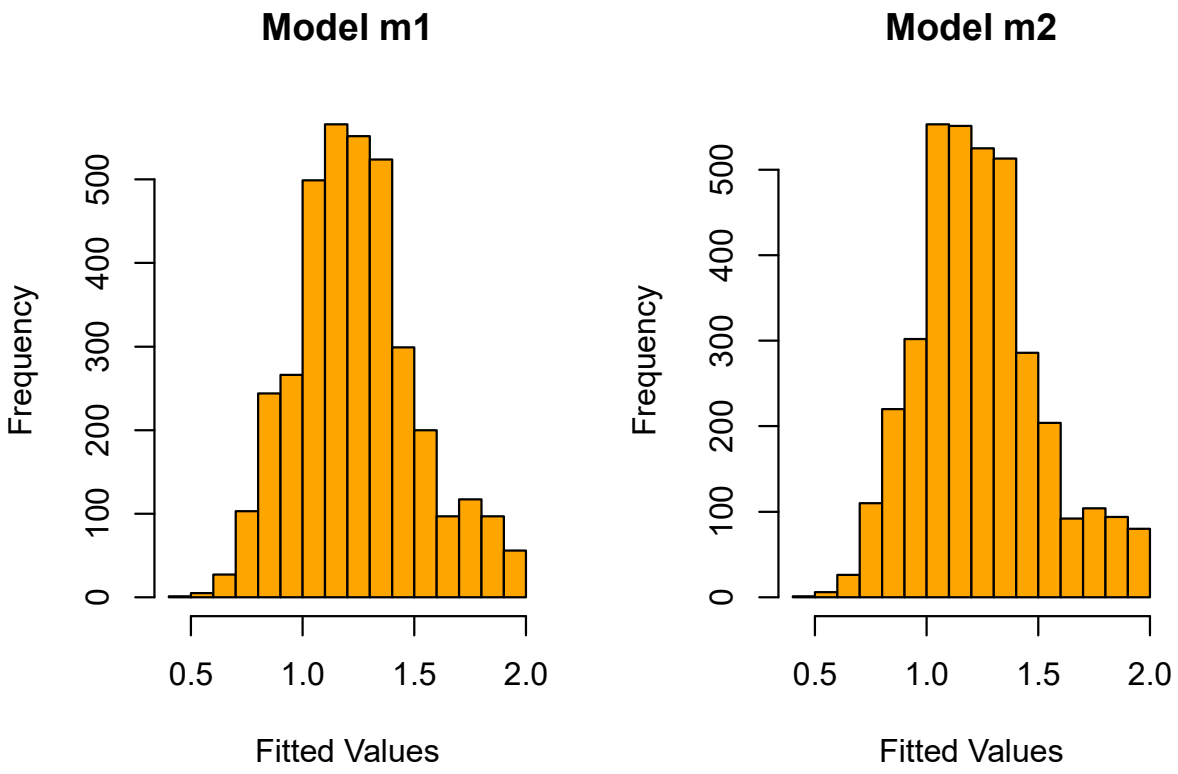
Variables	Coefficient Statistics Model M2			
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1653342	0.0537317	21.688039	0.0000000
as.factor(STARS)2	0.3188292	0.0143589	22.204281	0.0000000
as.factor(STARS)3	0.4400111	0.0156129	28.182481	0.0000000
as.factor(STARS)4	0.5604057	0.0216513	25.883198	0.0000000
as.factor(LabelAppeal)-1	0.2360802	0.0379857	6.214969	0.0000000
as.factor(LabelAppeal)0	0.4256523	0.0370518	11.488025	0.0000000
as.factor(LabelAppeal)1	0.5587523	0.0376899	14.824991	0.0000000
as.factor(LabelAppeal)2	0.6949136	0.0424294	16.378103	0.0000000
stars_I_FLAG	-0.7702682	0.0195303	-39.439561	0.0000000
Alcohol	0.0040887	0.0014447	2.830124	0.0046530
AcidIndex	-0.0803010	0.0044972	-17.855966	0.0000000
VolatileAcidity	-0.0371364	0.0093962	-3.952287	0.0000774
TotalSulfurDioxide	0.0000762	0.0000320	2.381333	0.0172501

	x
as.factor(STARS)2	0.5299098
as.factor(STARS)3	0.4454188
as.factor(STARS)4	0.2729857
as.factor(LabelAppeal)-1	3.4152763
as.factor(LabelAppeal)0	4.3246336
as.factor(LabelAppeal)1	3.2969278
as.factor(LabelAppeal)2	0.8480060
stars_I_FLAG	0.9448136
Alcohol	0.3333916
AcidIndex	0.4524636

Shown below are comparison plots of the fitted values compared to the target values. Fitted values have been rounded and then a jitter plot put together for comparison. The redline shows what a perfect fit would like. In other words plotted points along the red diagonal would only be populated.



It is clear that model does a very poor job of estimating zero sales based on wine characteristics. Focusing on this a little further plots of the fitted values less than 2 are shown in the model below.

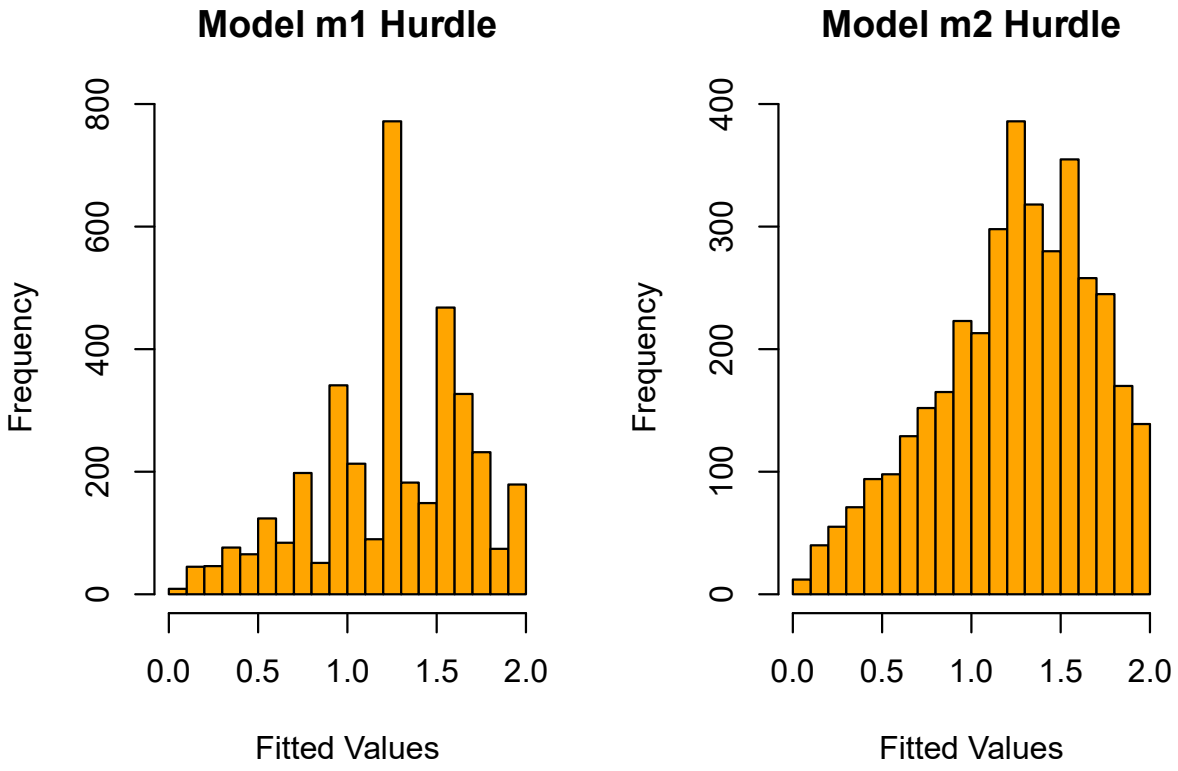


Hurdle Rate Model

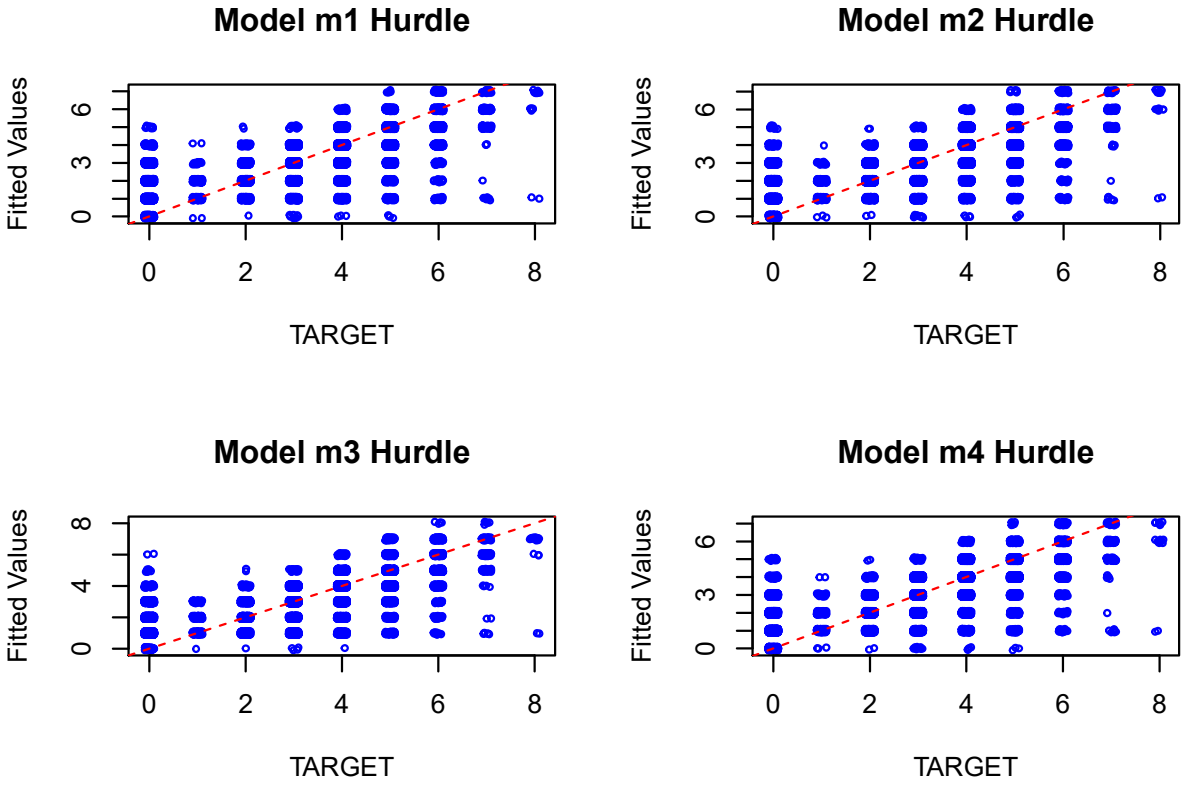
Hurdle rate models are shown below. Hurdle rate models account for the zero portion of the model separately and include two portions of the model. Four different models are developed and include some different variables and different distributional assumptions of the zero hurdle rate model, including poisson and binomial assumptions.

```
#develop hurdle rate models
m1h<-hurdle(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex,dist = c("pois

m2h<-hurdle(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidi

m3h<-hurdle(TARGET~as.factor(STARS)+LabelAppeal+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidity+TotalSul

m4h<-hurdle(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidi
```

Below are shown plots similar to those shown above regarding fitted values less than 2, to see how the model does for zero sales. It does sem that now a much higher number of wines are expected to have zero or near zero sales.



Jitter plots for the hurdle rate models are shown below comparing the target variable to the model fitted values. The red line shows the ideal fit. The models perform similarly and visually it is hard to see which ones are clearly better but model m3 and model m4 do seem to be better.

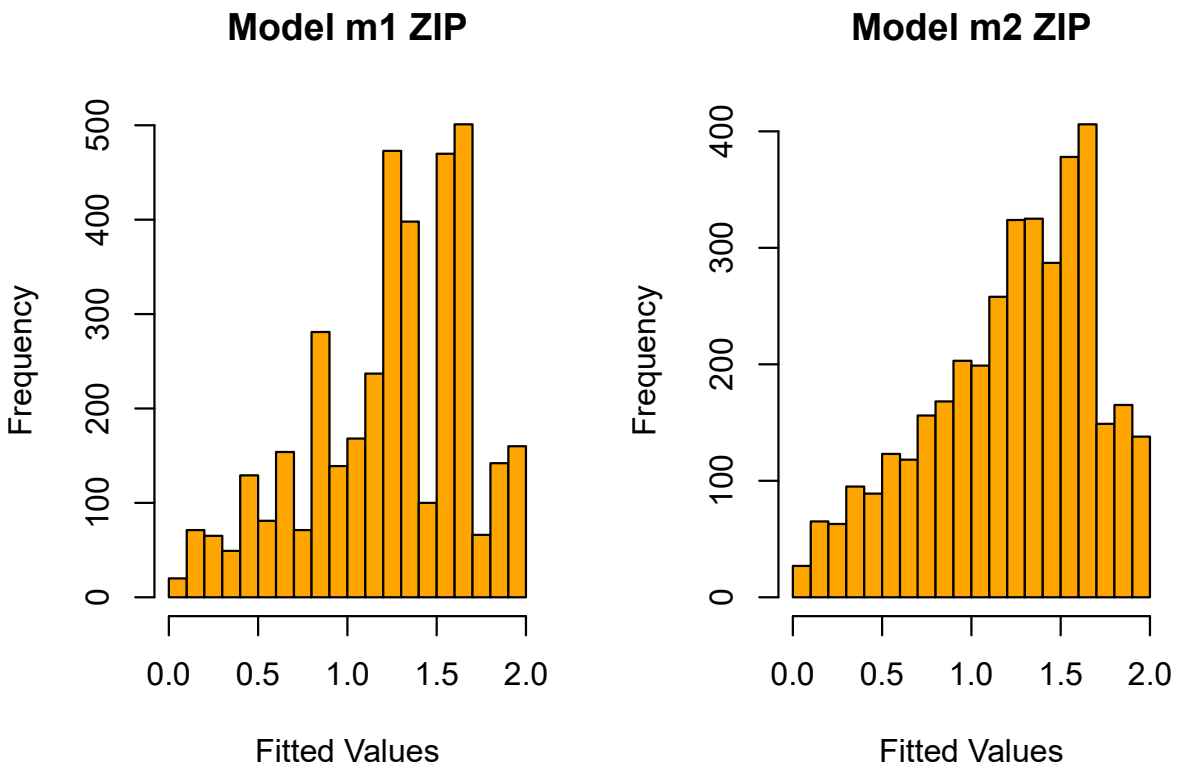


Zero Inflated Poisson Models

Two zero inflated poisson models are shown below.The models utilize the same variables as those in model m1 and model m2.

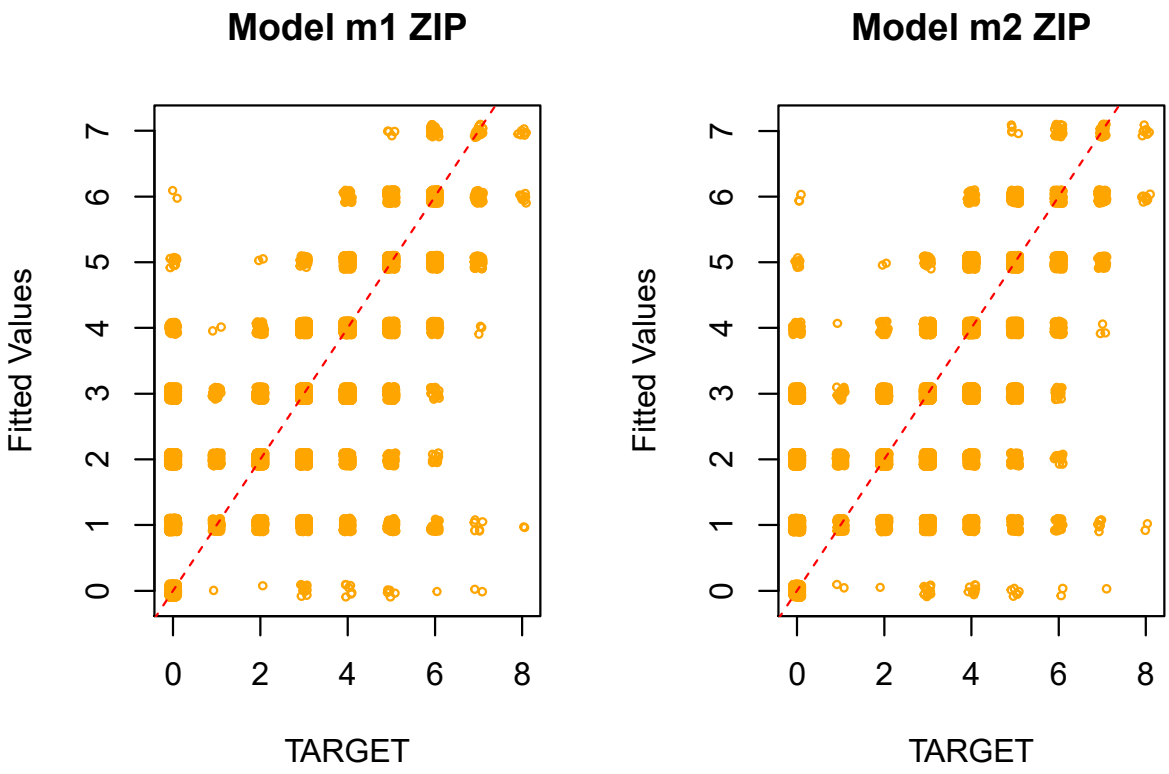
```
#develop ZIP models
m1zip<-zeroinfl(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex,dist = c(
m2zip<-zeroinfl(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileA
```

Histograms of values less than 2 in fitted value are shown below. For the zero portion, the results are quite similar to the hurdle rate models.



Fitted values compared to the target for the zero inflated poisson models are shown below. The ZIP models do seem to do better in some respects than other models but for example have no sales that reach 8, while some of the hurdle models capture this. Overall, the zip models do seem to make errors that are of lower magnitude.

```
par(mfrow=c(1,2))
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m1zip),0),amount=0.1),cex=0.5,col="orange",xlab=
abline(0,1,lty=2,col="red")
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m2zip),0),amount=0.1),cex=0.5,col="orange",xlab=
abline(0,1,lty=2,col="red")
```

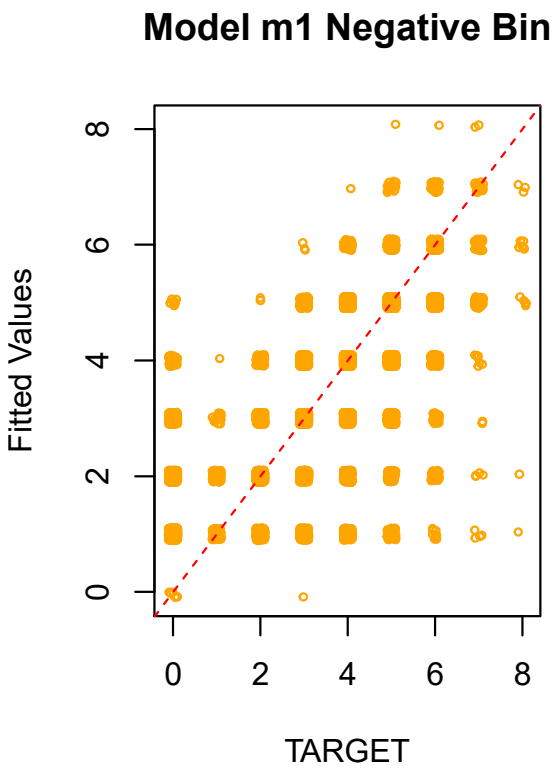


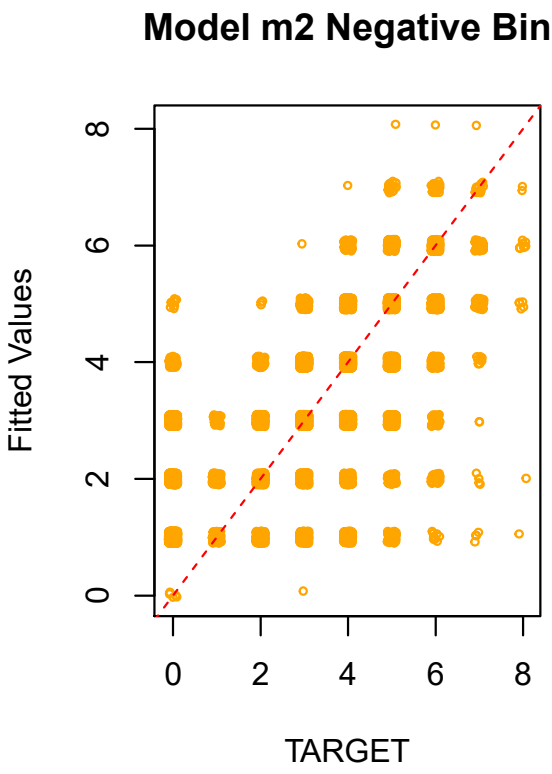
Negative Binomial Models

Negative bionomial models are shown below. A dispersion factor of 2 is fixed.

```
#negative binomial models
nb1<-glm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex, negative.binomi
nb2<-glm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidity+
```

Jitter plots for the Negative binomail models are shown below comparing the target variable to the model fitted values. The red line shows the ideal fit. It is hard to tell how much worse the models are however, the small number of zero forecasts are indicative that model can not capture the large number of no sale wines.





Finally, a simple linear regression model is developed as a comparison.

```
l1<-lm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex,wine)
summary(l1)$coefficients %>% kable("latex",booktabs=T) %>% kable_styling(full_width=F)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.3309778	0.0983954	33.852971	0.00e+00
as.factor(STARS)2	1.0416526	0.0326904	31.864151	0.00e+00
as.factor(STARS)3	1.6111490	0.0377670	42.660175	0.00e+00
as.factor(STARS)4	2.2962485	0.0599319	38.314279	0.00e+00
as.factor(LabelAppeal)-1	0.3634081	0.0630624	5.762677	0.00e+00
as.factor(LabelAppeal)0	0.8289754	0.0614908	13.481286	0.00e+00
as.factor(LabelAppeal)1	1.2935158	0.0642224	20.141193	0.00e+00
as.factor(LabelAppeal)2	1.8781253	0.0846105	22.197320	0.00e+00
stars_I_FLAG	-1.3771624	0.0329978	-41.735016	0.00e+00
Alcohol	0.0129347	0.0032863	3.935903	8.33e-05
AcidIndex	-0.2043639	0.0089217	-22.906338	0.00e+00

Jitter plot of the regression model shows that there are some negative value forecasts as well which is problematic given that we know the output variable to zero or strictly greater than zero.