homework5

Alice Ding

2023-12-05

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

Overview

In this homework assignment, you will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Your objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided).

Below is a short description of the variables of interest in the data set:

- `INDEX`: Identification Variable (do not use) None
- `TARGET`: Number of Cases Purchased None
- `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 customes don't like the design. // Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.
- `ResidualSugar`: Residual Sugar of wine
- `STARS`: Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor // A high number of stars suggests high sales
- `Sulphates`: Sulfate conten of wine
- `TotalSulfurDioxide`: Total Sulfur Dioxide of Wine
- `VolatileAcidity`: Volatile Acid content of wine
- `pH`: pH of wine

	vars	n	mean	sd	median	trimmed	mad	min	max	range
INDEX	1	12795	8069.98	4656.91	8110.00	8071.03	5977.84	1.00	16129.00	16128.00
TARGET	2	12795	3.03	1.93	3.00	3.05	1.48	0.00	8.00	8.00
FixedAcidity	3	12795	7.08	6.32	6.90	7.07	3.26	-18.10	34.40	52.50
VolatileAcidity	4	12795	0.32	0.78	0.28	0.32	0.43	-2.79	3.68	6.47
CitricAcid	5	12795	0.31	0.86	0.31	0.31	0.42	-3.24	3.86	7.10
ResidualSugar	6	12179	5.42	33.75	3.90	5.58	15.72	-127.80	141.15	268.95
Chlorides	7	12157	0.05	0.32	0.05	0.05	0.13	-1.17	1.35	2.52
FreeSulfurDioxide	8	12148	30.85	148.71	30.00	30.93	56.34	-555.00	623.00	1178.00
TotalSulfurDioxide	9	12113	120.71	231.91	123.00	120.89	134.92	-823.00	1057.00	1880.00
Density	10	12795	0.99	0.03	0.99	0.99	0.01	0.89	1.10	0.21
pH	11	12400	3.21	0.68	3.20	3.21	0.39	0.48	6.13	5.65
Sulphates	12	11585	0.53	0.93	0.50	0.53	0.44	-3.13	4.24	7.37
Alcohol	13	12142	10.49	3.73	10.40	10.50	2.37	-4.70	26.50	31.20
LabelAppeal	14	12795	-0.01	0.89	0.00	-0.01	1.48	-2.00	2.00	4.00
AcidIndex	15	12795	7.77	1.32	8.00	7.64	1.48	4.00	17.00	13.00
STARS	16	9436	2.04	0.90	2.00	1.97	1.48	1.00	4.00	3.00

Data Exploration

First, we'll view the summary and then we'll check if there are data points missing. Then, we'll clean the fields up to make sure they're ready for analysis.

```
training <- read.csv('https://raw.githubusercontent.com/addsding/data621/main/homework5/wine-training-da
evaluation <- read.csv('https://raw.githubusercontent.com/addsding/data621/main/homework5/wine-evaluatio

summary <- as.data.frame(describe(training))
nulls <- 12795 - summary['n']
nulls_pct <- nulls / 12795
summary['nulls'] <- nulls
summary['nulls_pct'] <- nulls_pct
kable(summary, digits=2) |>
  kable_styling(c("striped", "scale_down")) |>
  scroll_box(width = "100%")
```

The data has 16 variables with 12,795 observations

It looks like the only fields with nulls are `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `pH`, `Sulphates`, `Alcohol`, and `STARS`. The last field has the most amount of nulls at 26% of the data, but luckily the rest of the fields have $< 10\%$ nulls so that's good to keep in mind.

At first glance, there don't seem to be too many skewed variables – medians and means all seem relatively close together which is nice to see.

Some interesting points are that a few of these fields go into the negatives – not entirely sure what for example a negative `Alcohol` value would mean.

What types of fields are each of our variables?

```
summary(training)
```

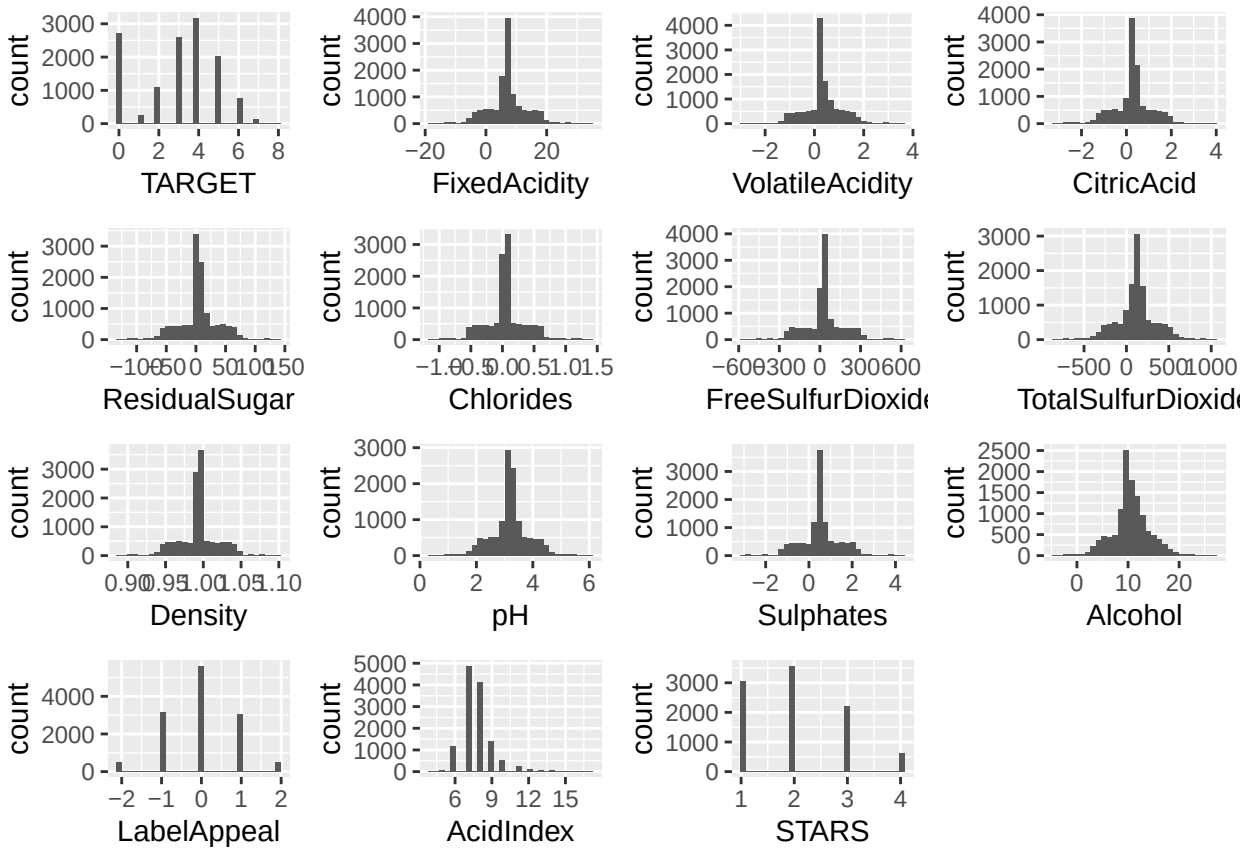
```
##      INDEX           TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
```

```
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar        Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616        NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
##     Alcohol       LabelAppeal          AcidIndex          STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

All of these variables are numeric and looks like they're formatted correctly which is great to see.

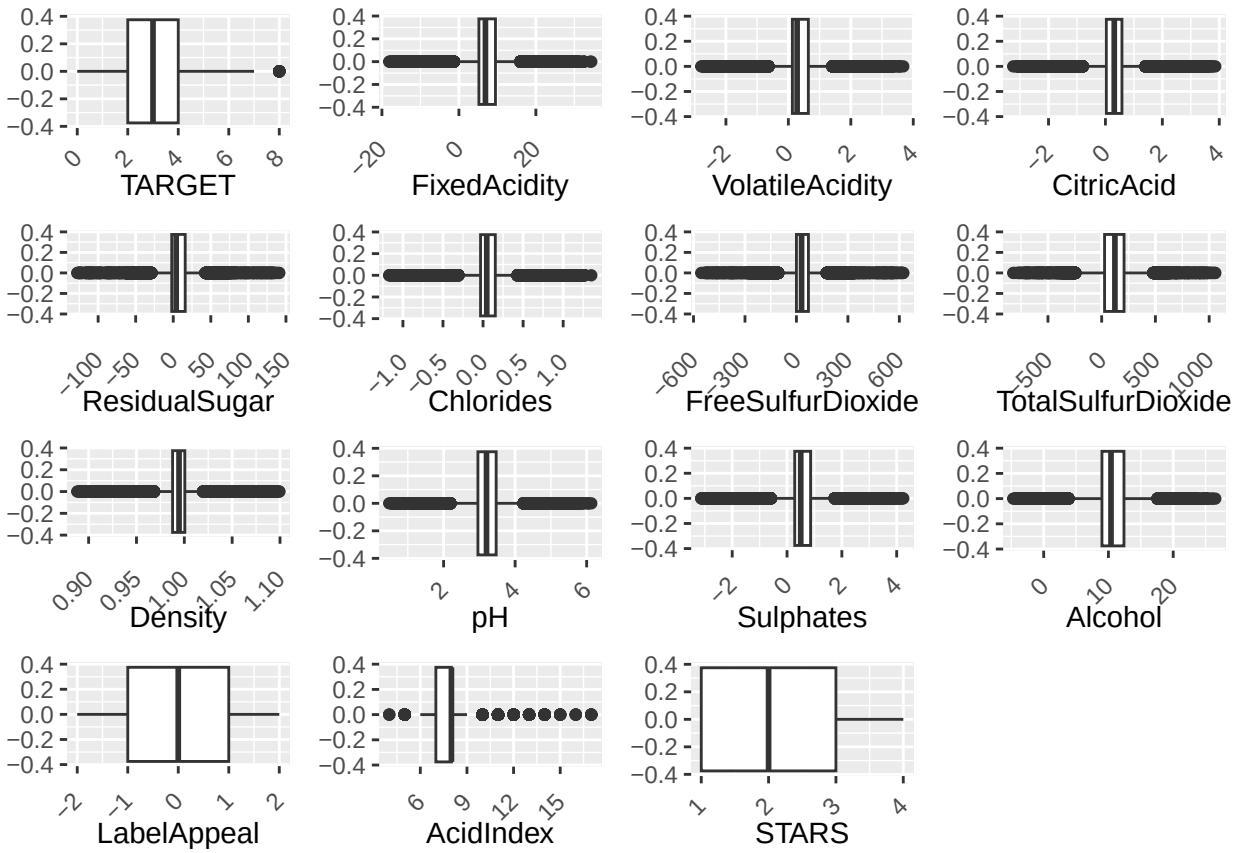
Distributions

Let's see what all of these fields look like distribution wise.



In line with the previous observation that medians and means all looked relatively close, this data all looks relatively normal. There are a few skews notably in the `Alcohol`, `CitricAcid`, `AcidIndex` and `STARS` fields, but compared to other datasets, this is all looking very solid. The peaks for the middle points are quite tall so there may be quite a few outliers for those on the edges of the curves.

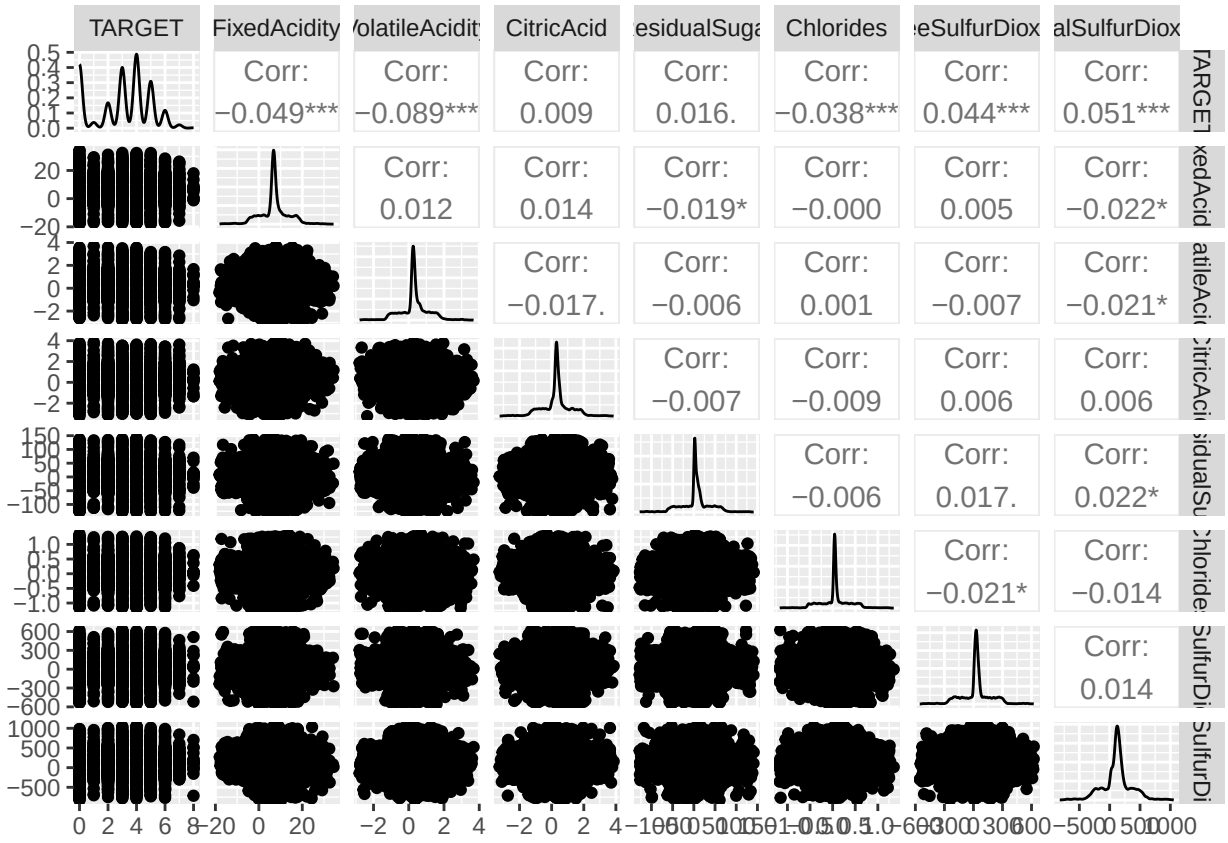
How do these look as boxplots?



As noted in the previous section, there are quite a few outliers in most of these fields due to such large peaks around the center for all of these. Just based on the nature of the data, I would likely not want to impute any of these outliers, but we'll see in a later section whether or not to do so.

Correlations

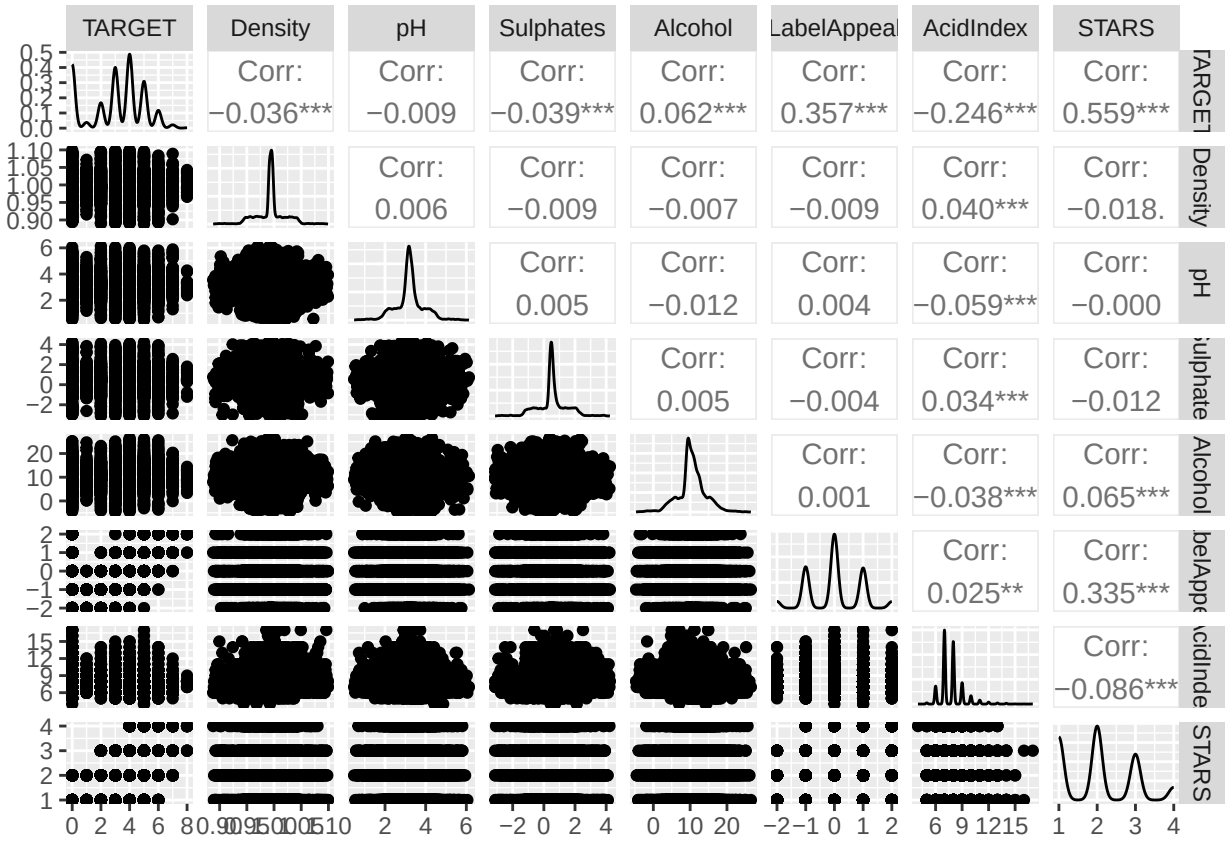
Let's see how each of the numerical fields correlate with `target` – we'll start with the first seven fields.



Interestingly, it seems that there aren't too many correlated fields again – only `Fixed Acidity` and `FreeSulfurDioxide` with the latter being much more significantly correlated. Implication wise, both are negatively correlated which means that the less acidity and less sulfur dioxide, the more cases of wine sold.

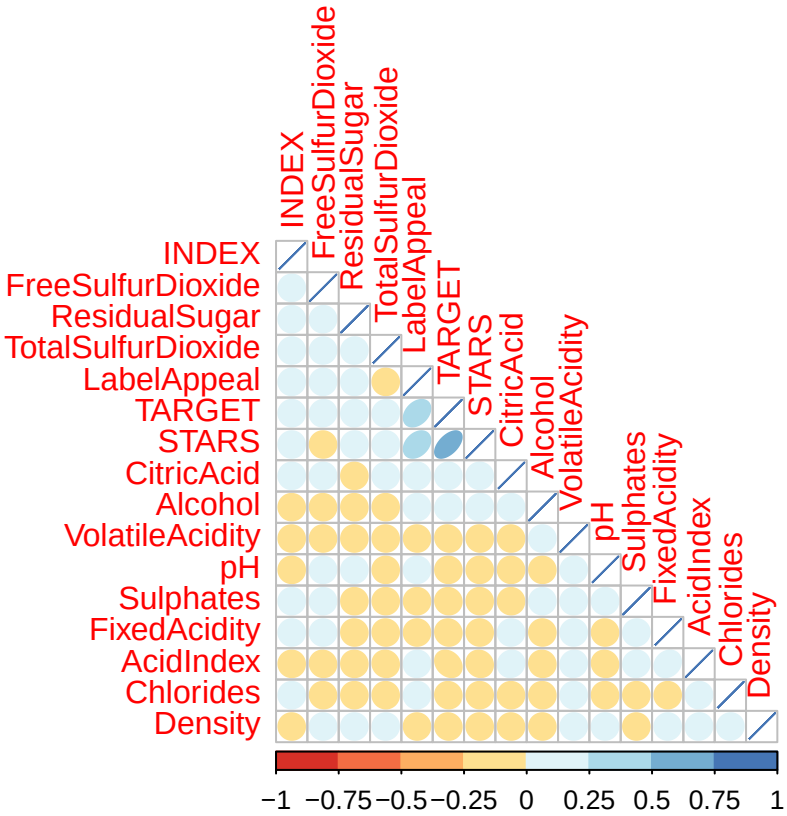
These fields aren't too correlated with one another interestingly enough.

What do the relationships look like for the rest of the fields?



Interestingly, it seems that there aren't too many correlated fields – only `Fixed Acidity` and `FreeSulfurDioxide` with the latter being much more significantly correlated. Implication wise, both are negatively correlated which means that the less acidity and less sulfur dioxide, the more cases of wine sold.

There are more correlations here than in the previous set of variables, but let's see if they're also correlated with each other beyond just the seven displayed here.



These correlations between variables aren't too strong, ranging from -0.25 to 0.25 on average. `STARS` and `TARGET` are a bit stronger in correlation, however it isn't too high to be an issue. This was also expected based on our initial readings of the data.

Data Preparation

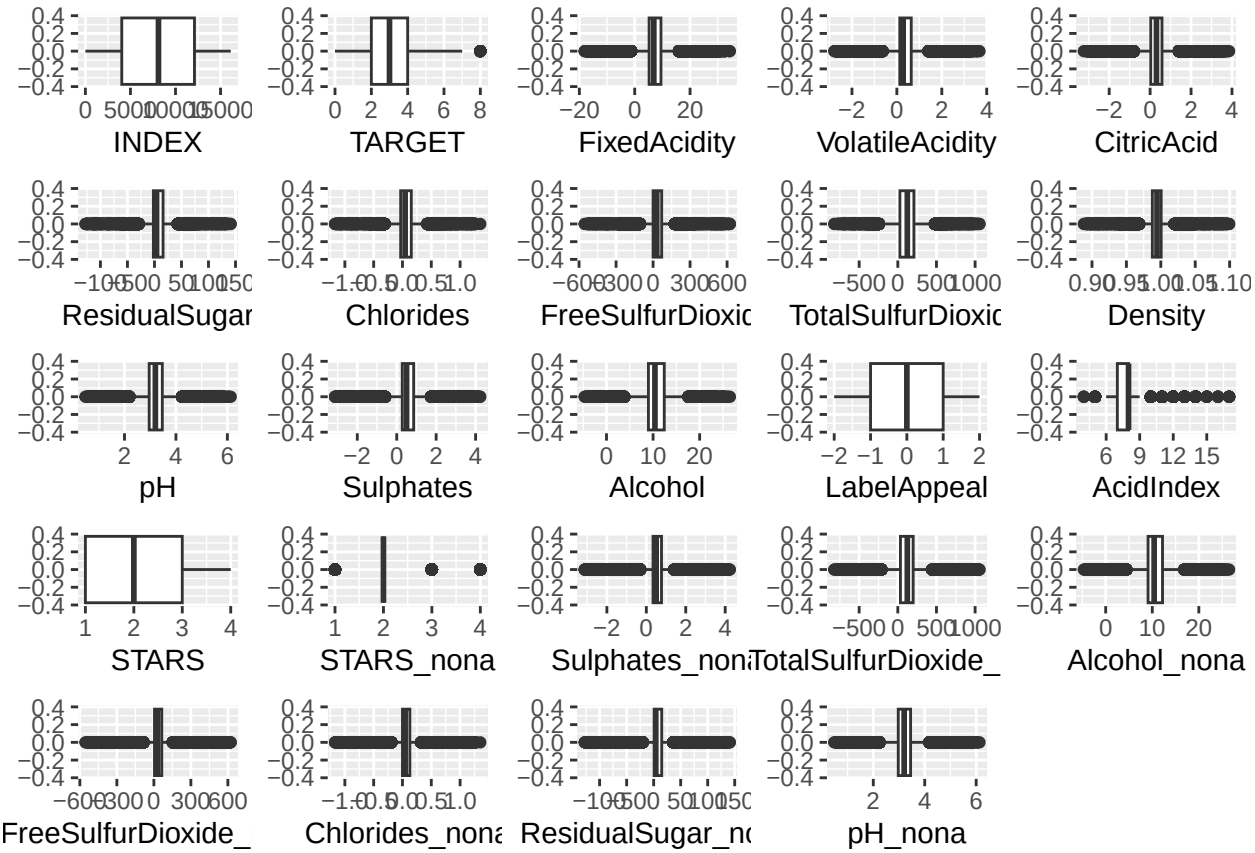
Imputing Values

A hint for this homework assignment originally was that sometimes the absence of a variable could be indicative of our target. Given this information, perhaps we actually leave the fields with nulls and create a new variable that imputes them so we have both versions of the field.

We usually also address outliers, however we're relying on the fact that:

- This data was all collected without error
- This data all truly represents natural variations in the population
- This data is not impacted by poor sampling

With these assumptions, we are erring on not imputing outliers and leaving the data as is, especially given the natural normalness of the curves.



Upon first glance, it looks like the amount of outliers has increased for each of these fields where we imputed nulls probably due to the increase in medians which typically are the peaks of the distribution.

Transform Non-Normal Variables

Given all of these fields are pretty normal (albeit a few skewed, but still overall in very good shape), we'll opt to not transform any of these fields.

Build Models

Before doing anything, we will split the data into training and test sets with a 70/30 split.

We'll go through two sets of models:

- Model 1: Poisson Regression
  - Using unadjusted variables
  - Using variables with no NAs
- Model 2: Negative Binomial Regression
  - Using unadjusted variables
  - Using variables with no NAs

Model 1A

This first model will use poisson regression and non-adjusted variables, then we'll refine it by looking at significant variables.

```
##
## Call:
## glm(formula = TARGET ~ AcidIndex + Alcohol + Chlorides + CitricAcid +
##     Density + FixedAcidity + FreeSulfurDioxide + LabelAppeal +
##     ResidualSugar + STARS + Sulphates + TotalSulfurDioxide +
##     VolatileAcidity + pH, family = "poisson", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.06204  -0.28073   0.06838   0.37915   1.68764
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.441e+00  2.949e-01   4.885 1.03e-06 ***
## AcidIndex          -5.212e-02  7.042e-03  -7.401 1.35e-13 ***
## Alcohol             3.110e-03  2.145e-03   1.450 0.147101
## Chlorides          -3.401e-02  2.473e-02  -1.375 0.169055
## CitricAcid         -6.329e-03  9.026e-03  -0.701 0.483200
## Density            -1.749e-01  2.896e-01  -0.604 0.545779
## FixedAcidity        2.956e-04  1.260e-03   0.235 0.814523
## FreeSulfurDioxide   7.522e-05  5.330e-05   1.411 0.158142
## LabelAppeal         1.745e-01  9.529e-03  18.313  < 2e-16 ***
## ResidualSugar      -8.358e-05  2.327e-04  -0.359 0.719434
## STARS               1.862e-01  8.945e-03  20.820  < 2e-16 ***
## Sulphates          -4.426e-03  8.482e-03  -0.522 0.601805
## TotalSulfurDioxide  4.036e-05  3.407e-05   1.184 0.236235
## VolatileAcidity    -3.434e-02  1.006e-02  -3.415 0.000639 ***
## pH                 -6.492e-03  1.154e-02  -0.563 0.573774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4064.4  on 4490  degrees of freedom
## Residual deviance: 2812.0  on 4476  degrees of freedom
##   (4467 observations deleted due to missingness)
## AIC: 16172
##
## Number of Fisher Scoring iterations: 5
```

Coefficient Evaluation

Looking at the model's coefficients, these had negative values:

- AcidIndex
- Chlorides
- CitricAcid
- Density
- ResidualSugar

- Sulphates
- VolatileAcidity
- pH

These negative coefficients imply that the higher these values, the lower amount of cases will be sold. For example, just using some of the variables, we can interpret here that higher acidity, sugar, and density would make for a lower performing wine.

For the positive values:

- Alcohol
- FixedAcidity
- FreeSulfurDioxide
- LabelAppeal
- STARS
- TotalSulfurDioxide
- Intercept

This would imply that for these fields, if they are higher in value, this would mean that more units would be sold. For example, just using some of the variables, wine with a higher alcohol content and more label appeal would have higher sales than those with lower values for those fields.

Significance Evaluation & Performance

Only a few fields were statistically significant:

- Intercept
- AcidIndex
- LabelAppeal
- STARS
- VolatileAcidity

The theoretical effect table had alluded to LabelAppeal and STARS being good indicators so that's not too surprising, but acidity having an effect is interesting.

With an AIC of 16172 and residual deviance of 2812, we'll use this as a baseline to compare to as we iterate on the model.

Model 1B

```
##
## Call:
## glm(formula = TARGET ~ AcidIndex + LabelAppeal + STARS + VolatileAcidity,
##     family = "poisson", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.13014  -0.27152   0.06431   0.37281   1.61364
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.274863   0.048152  26.476  < 2e-16 ***
## AcidIndex       -0.050627   0.005750  -8.805  < 2e-16 ***
```

```
## LabelAppeal      0.179195   0.007786  23.015  < 2e-16 ***
## STARS            0.184462   0.007307  25.243  < 2e-16 ***
## VolatileAcidity -0.029122   0.008197  -3.553 0.000381 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6001.6  on 6616  degrees of freedom
## Residual deviance: 4094.4  on 6612  degrees of freedom
##   (2341 observations deleted due to missingness)
## AIC: 23789
##
## Number of Fisher Scoring iterations: 5
```

Coefficient Evaluation

Between this and Model 1A, the coefficients have only changed in magnitude and not so much direction, however not by too much.

Significance Evaluation & Performance

They've all increased in significance which is good to see.

Performance wise though, AIC and residual deviance have both increased sadly signalling this model is not a better fit than our previous iteration.

Let's try a poisson model now using our non-NA fields.

Model 1C

```
##
## Call:
## glm(formula = TARGET ~ AcidIndex + Alcohol_nona + Chlorides_nona +
##     CitricAcid + Density + FixedAcidity + FreeSulfurDioxide_nona +
##     LabelAppeal + ResidualSugar_nona + STARS_nona + Sulphates_nona +
##     TotalSulfurDioxide_nona + VolatileAcidity + pH_nona, family = "poisson",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5192  -0.5162   0.2026   0.6307   2.5769
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.007e+00  2.322e-01   8.642  < 2e-16 ***
## AcidIndex               -1.238e-01  5.356e-03 -23.117  < 2e-16 ***
## Alcohol_nona             4.901e-03  1.710e-03   2.866  0.00416 **
## Chlorides_nona          -4.925e-02  1.970e-02  -2.500  0.01240 *
## CitricAcid               6.344e-03  7.000e-03   0.906  0.36478
## Density                 -3.936e-01  2.273e-01  -1.731  0.08342 .
## FixedAcidity            -3.551e-04  9.848e-04  -0.361  0.71842
## FreeSulfurDioxide_nona   1.660e-04  4.224e-05   3.929 8.53e-05 ***
## LabelAppeal              1.938e-01  7.167e-03  27.037  < 2e-16 ***
## ResidualSugar_nona       2.925e-04  1.854e-04   1.578  0.11468
```

```
## STARS_nona               2.175e-01  7.735e-03  28.124  < 2e-16 ***
## Sulphates_nona          -1.225e-02  6.888e-03  -1.779  0.07527 .
## TotalSulfurDioxide_nona  1.319e-04  2.714e-05   4.861 1.17e-06 ***
## VolatileAcidity         -5.620e-02  7.733e-03  -7.267 3.68e-13 ***
## pH_nona                 -2.752e-02  9.136e-03  -3.012  0.00259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 15938  on 8957  degrees of freedom
## Residual deviance: 12838  on 8943  degrees of freedom
## AIC: 35232
##
## Number of Fisher Scoring iterations: 5
```

Coefficient Evaluation

Looking at the model's coefficients, these had negative values:

- `AcidIndex`
- `Chlorides_nona`
- `Density`
- `FixedAcidity`
- `Sulphates`
- `VolatileAcidity`
- `pH_nona`

Interestingly, compared to Model 1A, `CitricAcid`, `FixedAcidity`, and `ResidualSugar` seem to have flipped; `FixedAcidity` is now negative despite being positive in impact previously.

For the positive values:

- `Alcohol_nona`
- `CitricAcid`
- `FreeSulfurDioxide_nona`
- `LabelAppeal`
- `ResidualSugar_nona`
- `STARS_nona`
- `TotalSulfurDioxide_nona`
- Intercept

Significance Evaluation & Performance

Compared to Model 1A, many more fields are significant now:

- Intercept
- `Alcohol_nona`
- `AcidIndex`
- `Chlorides_nona`
- `Density`
- `FreeSulfurDioxide_nona`
- `LabelAppeal`
- `STARS_nona`

- Sulphates_nona
- TotalSulfurDioxide_nona
- VolatileAcidity
- pH_nona

The additional significant fields seem to be a good indicator.

With an AIC of 35232 and residual deviance of 12838, we'll use this as a baseline to compare to as we iterate on the model. Note that this is much higher (worse) than the first iterations of the model though.

Model 1D

```
##
## Call:
## glm(formula = TARGET ~ AcidIndex + Alcohol_nona + Chlorides_nona +
##     Density + FreeSulfurDioxide_nona + LabelAppeal + STARS_nona +
##     Sulphates_nona + TotalSulfurDioxide_nona + VolatileAcidity +
##     pH_nona, family = "poisson", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5268  -0.5132   0.2017   0.6332   2.5912
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               2.008e+00  2.322e-01   8.647  < 2e-16 ***
## AcidIndex                -1.238e-01  5.275e-03 -23.479  < 2e-16 ***
## Alcohol_nona              4.868e-03  1.709e-03   2.848  0.00439 **
## Chlorides_nona           -4.926e-02  1.969e-02  -2.502  0.01235 *
## Density                  -3.947e-01  2.273e-01  -1.737  0.08245 .
## FreeSulfurDioxide_nona    1.678e-04  4.222e-05   3.973 7.10e-05 ***
## LabelAppeal               1.938e-01  7.169e-03  27.041  < 2e-16 ***
## STARS_nona                2.178e-01  7.732e-03  28.173  < 2e-16 ***
## Sulphates_nona           -1.260e-02  6.886e-03  -1.830  0.06727 .
## TotalSulfurDioxide_nona   1.331e-04  2.713e-05   4.907 9.24e-07 ***
## VolatileAcidity          -5.659e-02  7.731e-03  -7.320 2.48e-13 ***
## pH_nona                  -2.709e-02  9.133e-03  -2.966  0.00301 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 15938  on 8957  degrees of freedom
## Residual deviance: 12842  on 8946  degrees of freedom
## AIC: 35229
##
## Number of Fisher Scoring iterations: 5
```

Coefficient Evaluation

Nothing has flipped from positve to negative or vice versa and the magnitude for each variable is similar to Model 1C.

Significance Evaluation & Performance

Interestingly, nothing has changed in terms of significance – they all have the same levels.

We do see similar AIC and residual deviance as well when compared to Model 1C.

Model 2A

We'll now move onto negative binomial regression models now.

```
##
## Call:
## glm.nb(formula = TARGET ~ AcidIndex + Alcohol + Chlorides + CitricAcid +
##     Density + FixedAcidity + FreeSulfurDioxide + LabelAppeal +
##     ResidualSugar + STARS + Sulphates + TotalSulfurDioxide +
##     VolatileAcidity + pH, data = train, init.theta = 138528.0004,
##     link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.06202  -0.28073   0.06838   0.37914   1.68762
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.441e+00  2.950e-01   4.885 1.03e-06 ***
## AcidIndex           -5.212e-02  7.042e-03  -7.401 1.35e-13 ***
## Alcohol              3.110e-03  2.145e-03   1.450 0.147107
## Chlorides           -3.401e-02  2.473e-02  -1.375 0.169058
## CitricAcid          -6.329e-03  9.026e-03  -0.701 0.483207
## Density             -1.749e-01  2.896e-01  -0.604 0.545781
## FixedAcidity         2.957e-04  1.260e-03   0.235 0.814519
## FreeSulfurDioxide    7.522e-05  5.330e-05   1.411 0.158145
## LabelAppeal          1.745e-01  9.529e-03  18.313  < 2e-16 ***
## ResidualSugar       -8.358e-05  2.327e-04  -0.359 0.719441
## STARS                1.862e-01  8.945e-03  20.820  < 2e-16 ***
## Sulphates           -4.426e-03  8.482e-03  -0.522 0.601804
## TotalSulfurDioxide   4.036e-05  3.407e-05   1.184 0.236238
## VolatileAcidity     -3.434e-02  1.006e-02  -3.415 0.000639 ***
## pH                  -6.492e-03  1.154e-02  -0.563 0.573775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(138528) family taken to be 1)
##
##     Null deviance: 4064.3  on 4490  degrees of freedom
## Residual deviance: 2812.0  on 4476  degrees of freedom
##   (4467 observations deleted due to missingness)
## AIC: 16174
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  138528
##           Std. Err.:  277344
## Warning while fitting theta: iteration limit reached
```

```
##
##  2 x log-likelihood:  -16142.19
```

Coefficient Evaluation
Looking at the model's coefficients, these had negative values:

- AcidIndex
- Chlorides
- CitricAcid
- Density
- ResidualSugar
- Sulphates
- VolatileAcidity
- pH

This is identical to Model 1A.

For the positive values:

- Alcohol
- FixedAcidity
- FreeSulfurDioxide
- LabelAppeal
- STARS
- TotalSulfurDioxide
- Intercept

Significance Evaluation & Performance
Only a few fields were statistically significant:

- Intercept
- AcidIndex
- LabelAppeal
- STARS
- VolatileAcidity

Again, identical to Model 1A.

With an AIC of 16174 and residual deviance of 2812, we'll use this as a baseline to compare to as we iterate on the model.

Model 2B

```
##
## Call:
## glm.nb(formula = TARGET ~ AcidIndex + LabelAppeal + STARS + VolatileAcidity,
##     data = train, init.theta = 141013.4602, link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.13011  -0.27152   0.06431   0.37280   1.61362
##
```

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.274864   0.048153  26.475  < 2e-16 ***
## AcidIndex        -0.050627   0.005750  -8.805  < 2e-16 ***
## LabelAppeal       0.179195   0.007786  23.015  < 2e-16 ***
## STARS             0.184462   0.007307  25.243  < 2e-16 ***
## VolatileAcidity -0.029122   0.008197  -3.553 0.000381 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(141013.5) family taken to be 1)
##
##     Null deviance: 6001.5  on 6616  degrees of freedom
## Residual deviance: 4094.3  on 6612  degrees of freedom
##   (2341 observations deleted due to missingness)
## AIC: 23792
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  141013
##           Std. Err.:  232464
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -23779.6
```

Coefficient Evaluation

Between this and Model 2A, the coefficients have only changed in magnitude and not so much direction, however not by too much.

Significance Evaluation & Performance

They're all at the same level of significance, but the actual numbers themselves have gotten stronger.

Performance wise though, AIC and residual deviance have both increased sadly signalling this model is not a better fit than our previous iteration.

Let's try a negative binomial model now using our non-NA fields.

Model 2C

```
##
## Call:
## glm.nb(formula = TARGET ~ AcidIndex + Alcohol_nona + Chlorides_nona +
##     CitricAcid + Density + FixedAcidity + FreeSulfurDioxide_nona +
##     LabelAppeal + ResidualSugar_nona + STARS_nona + Sulphates_nona +
##     TotalSulfurDioxide_nona + VolatileAcidity + pH_nona, data = train,
##     init.theta = 39858.05452, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5190  -0.5162   0.2026   0.6306   2.5768
##
## Coefficients:
```

```
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               2.007e+00  2.322e-01   8.642  < 2e-16 ***
## AcidIndex                -1.238e-01  5.356e-03 -23.116  < 2e-16 ***
## Alcohol_nona              4.901e-03  1.710e-03   2.866  0.00416 **
## Chlorides_nona           -4.925e-02  1.970e-02  -2.500  0.01241 *
## CitricAcid                6.344e-03  7.000e-03   0.906  0.36480
## Density                  -3.936e-01  2.273e-01  -1.731  0.08343 .
## FixedAcidity             -3.551e-04  9.848e-04  -0.361  0.71844
## FreeSulfurDioxide_nona    1.660e-04  4.225e-05   3.929 8.53e-05 ***
## LabelAppeal               1.938e-01  7.168e-03  27.036  < 2e-16 ***
## ResidualSugar_nona        2.925e-04  1.855e-04   1.577  0.11468
## STARS_nona                2.175e-01  7.735e-03  28.122  < 2e-16 ***
## Sulphates_nona           -1.225e-02  6.889e-03  -1.779  0.07527 .
## TotalSulfurDioxide_nona   1.319e-04  2.714e-05   4.861 1.17e-06 ***
## VolatileAcidity          -5.620e-02  7.734e-03  -7.267 3.69e-13 ***
## pH_nona                  -2.752e-02  9.136e-03  -3.012  0.00259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(39858.05) family taken to be 1)
##
##     Null deviance: 15937  on 8957  degrees of freedom
## Residual deviance: 12838  on 8943  degrees of freedom
## AIC: 35234
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  39858
##           Std. Err.:  71092
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -35201.9
```

Coefficient Evaluation

Looking at the model's coefficients, these had negative values:

- `AcidIndex`
- `Chlorides_nona`
- `Density`
- `FixedAcidity`
- `Sulphates_nona`
- `VolatileAcidity`
- `pH_nona`

Interestingly, compared to Model 1A and Model 2A, `CitricAcid`, `FixedAcidity`, and `ResidualSugar` seem to have flipped; `FixedAcidity` is now negative despite being positive in impact previously.

For the positive values:

- `Alcohol_nona`
- `CitricAcid`
- `FreeSulfurDioxide_nona`

- LabelAppeal
- ResidualSugar_nona
- STARS_nona
- TotalSulfurDioxide_nona
- Intercept

Significance Evaluation & Performance

Compared to Model 2A, many more fields are significant now:

- Intercept
- Alcohol_nona
- AcidIndex
- Chlorides_nona
- Density
- FreeSulfurDioxide_nona
- LabelAppeal
- STARS_nona
- Sulphates_nona
- TotalSulfurDioxide_nona
- VolatileAcidity
- pH_nona

The additional significant fields seem to be a good indicator.

With an AIC of 35234 and residual deviance of 12838, we'll use this as a baseline to compare to as we iterate on the model. Note that this is much higher (worse) than the first iterations of the model though.

Model 2D

```
##
## Call:
## glm.nb(formula = TARGET ~ AcidIndex + Alcohol_nona + Chlorides_nona +
##     Density + FreeSulfurDioxide_nona + LabelAppeal + STARS_nona +
##     Sulphates_nona + TotalSulfurDioxide_nona + VolatileAcidity +
##     pH_nona, data = train, init.theta = 39847.56819, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5267  -0.5132   0.2017   0.6332   2.5911
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.008e+00  2.322e-01   8.647  < 2e-16 ***
## AcidIndex               -1.238e-01  5.275e-03 -23.479  < 2e-16 ***
## Alcohol_nona             4.868e-03  1.709e-03   2.848  0.00439 **
## Chlorides_nona          -4.926e-02  1.969e-02  -2.502  0.01235 *
## Density                 -3.947e-01  2.273e-01  -1.737  0.08246 .
## FreeSulfurDioxide_nona   1.678e-04  4.223e-05   3.973 7.10e-05 ***
## LabelAppeal              1.938e-01  7.169e-03  27.040  < 2e-16 ***
## STARS_nona               2.178e-01  7.733e-03  28.171  < 2e-16 ***
## Sulphates_nona          -1.260e-02  6.886e-03  -1.830  0.06727 .
## TotalSulfurDioxide_nona  1.331e-04  2.713e-05   4.907 9.25e-07 ***
## VolatileAcidity         -5.659e-02  7.731e-03  -7.320 2.48e-13 ***
```

```
## pH_nona                 -2.709e-02  9.133e-03  -2.966  0.00301 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(39847.57) family taken to be 1)
##
##     Null deviance: 15937  on 8957  degrees of freedom
## Residual deviance: 12841  on 8946  degrees of freedom
## AIC: 35231
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  39848
##           Std. Err.:  71113
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -35205.32
```

Coefficient Evaluation

Nothing has flipped from positve to negative or vice versa and the magnitude for each variable is similar to Model 2C.

Significance Evaluation & Performance

Interestingly, nothing has changed in terms of significance – they all have the same levels.

We do see similar AIC and residual deviance as well when compared to Model 2C.

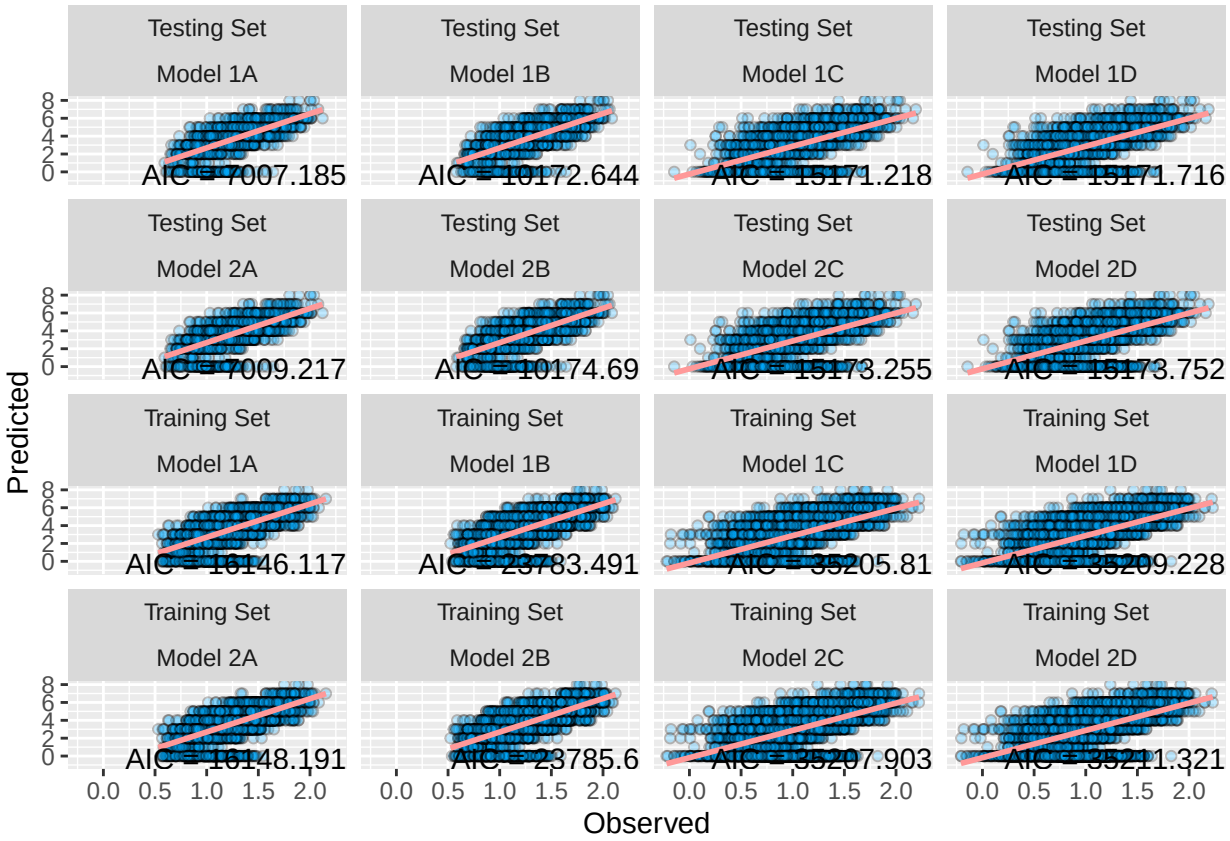
Select Models

```
##                 MSE      AIC
## Model 1A 0.1408798 16172.12
## Model 1B 0.1385659 23789.49
## Model 1C 0.3853824 35231.81
## Model 1D 0.3853259 35229.23
## Model 2A 0.1408800 16174.19
## Model 2B 0.1385660 23791.60
## Model 2C 0.3853835 35233.90
## Model 2D 0.3853270 35231.32
```

Based on the above output, Model B had the best MSEs with Model A close behind. For AIC however, Model A performed the best byfar with Model B performing the best next, but still not too good. Models C and D for both iterations were pretty poor.

In terms of distinction between Models 1 and 2, there doesn't seem to be much so it's safe to say that poisson and negative binomial regression perform relatively similarly.

Next, we will compare how these models do with the test dataset and compare residuals.



Interestingly, it seems that the testing sets performed better than the training sets with AICs of under half the training set iterations. The trends hold true though where Models A outperform the rest with Models B as second best.

Final Selection

Based on all of the factors shown above, it seems like Model 1A is just slightly better than Model 2A which both outperform the rest. Given the lowest AICs and second-best MSEs, they seem to capture the target value of wine purchases the best out of all 8 models we created.