DATA 621 Homework 5
Critical Thinking Group 3
Douglas Barley, Ethan Haley, Isabel Magnus, John Mazon, Vinayak Kamath, Arushi Arora

Data Exploration

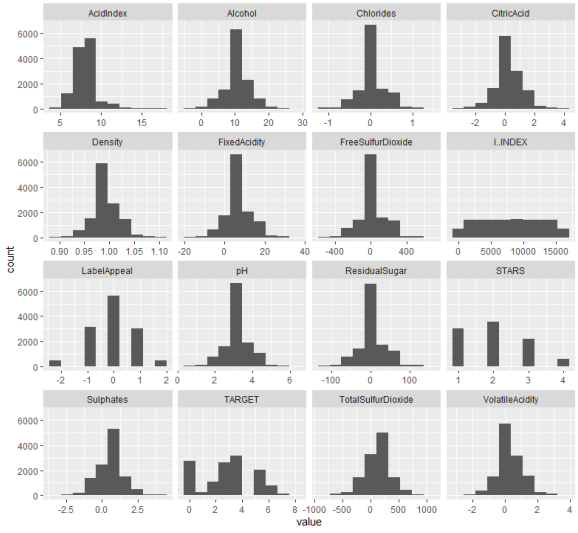
We are provided with two datasets on commercially available wine, one for the purpose of training our model with 12,796 observations of 16 variables and an evaluation dataset with 3,335 observations. The training dataset includes one response variable, TARGET, the number of cases purchased. TARGET is a continuous variable, with values between 0 and 8 in the training data. Because we are analyzing this data with the intention of maximizing sales for this wine manufacturer, we'll be creating models to better understand how much wine is ordered based on the wine's characteristics.

To begin our exploration of the data, we start with a broad look at the variables.

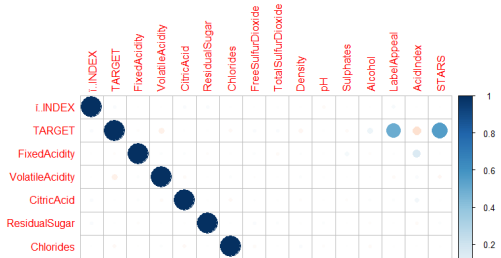
```
      I..INDEX         TARGET       FixedAcidity      VolatileAcidity    CitricAcid      ResidualSugar       Chlorides       FreeSulfurDioxide TotalSulfurDioxide    Density             pH          Sulphates
 Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0   Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
 1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0   1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
 Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0   Median :0.9945   Median :3.200   Median : 0.5000
 Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7   Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
 3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0   3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
 Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0   Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
                                                                     NA's   :616       NA's   :638       NA's   :647       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
```

There are missing observations in ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and STARS. STARS has the greatest rate of missingness with 26% null observations, while Sulphates has just under 10% null observations, followed by ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, and Alcohol all with about 5% null observations. Do observations with one null have multiple nulls across variables? Or are the nulls scattered throughout observations? Look at just the null values to understand them better. We will have to decide how to handle data missingness in the subsequent section.



Now, we can dig deeper into each variables' distribution to see if there are any integrity red flags or challenges with skewness. Strangely enough, we appear to have a dataset filled with variables that are normally distributed. Looking at these histograms, we might potentially be interested in a few transformations. AcidIndex, a proprietary method of testing total acidity of wine by using a weighted average, appears to be ever so slightly right-skewed, but it's a scaled composite variable so may be best left as is. STARS, with all it's missing observations, is also less than normally distributed. Depending on how we handle those missing observations, we'll be able to transform and/or handle the variable appropriately. It's also interesting to notice that our TARGET variable is the least normally distributed, with a large number of 0s and a conspicuous gap of observations between 4 and 5. Are there any other variables with



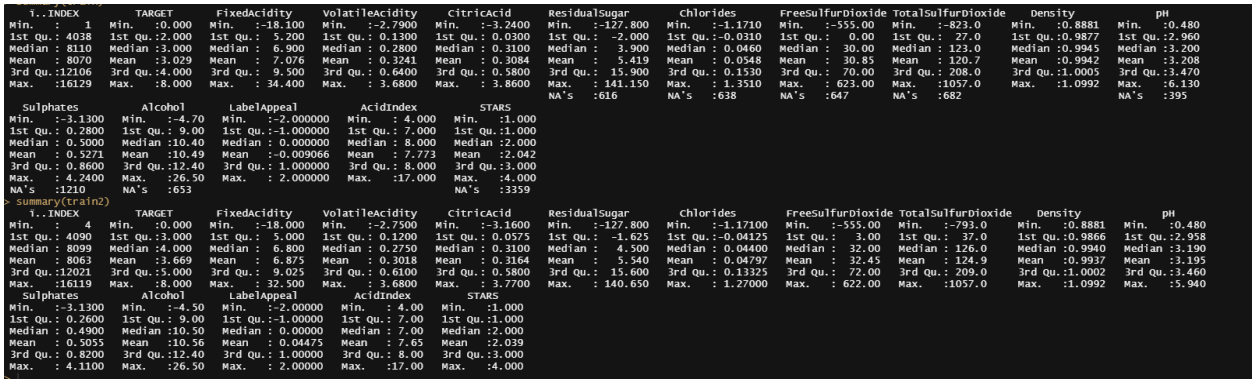
suspicious distributions /minimums /maximums/0s/means?

With a better understanding of individual variables, we can begin to look at how the variables are correlated. When we create a correlation table on complete observations, we see that there's not a lot of correlation between predictor variables. It will be interesting to look at correlation again once we've addressed nulls. There is the strongest positive correlation between STARS and TARGET, which theoretically makes sense given that STARS is the "wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor". LabelAppeal has the second strongest positive correlation with TARGET, and is another theoretically direct measurement of a customer's willingness to buy a specific wine. The strongest negative correlation is between AcidIndex and Target. Depending on how the index is set up, this may indicate that consumers have a strong preference for or against more acidic wines.

What else would we like to call out in our data exploration?

Data Preparation

In this section we will discuss how we prepared the data. Dropping missing data has important implications on a model's ability to predict on an evaluation dataset. When all nulls and missing data are removed, the means and medians of the variables change. With no nulls or missing data, there are 6,436 observations total. While this isn't a bad amount of sample, it does reduce the observations by half of the original dataset. Below, you can see how the removal has affected the medians and means of each variable.



```
   i..INDEX        TARGET       FixedAcidity      VolatileAcidity    CitricAcid      ResidualSugar       Chlorides        FreeSulfurDioxide TotalSulfurDioxide    Density             pH
Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0     Min.   :0.8881   Min.   :0.480
1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960
Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0     Median :0.9945   Median :3.200
Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7     Mean   :0.9942   Mean   :3.208
3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470
Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0     Max.   :1.0992   Max.   :6.130
                                                                                       NA's   :616        NA's   :638       NA's   :647       NA's   :682                         NA's   :395

   Sulphates         Alcohol       LabelAppeal         AcidIndex        STARS
Min.   :-3.1300   Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
1st Qu.: 0.2800   1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
Median : 0.5000   Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
Mean   : 0.5271   Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
3rd Qu.: 0.8600   3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
Max.   : 4.2400   Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
NA's   :1210      NA's   :653                                          NA's   :3359
> summary(train2)
   i..INDEX        TARGET       FixedAcidity      VolatileAcidity    CitricAcid      ResidualSugar       Chlorides        FreeSulfurDioxide TotalSulfurDioxide    Density             pH
Min.   :    4   Min.   :0.000   Min.   :-18.000   Min.   :-2.7500   Min.   :-3.1600   Min.   :-127.800   Min.   :-1.17100   Min.   :-555.00   Min.   :-793.0     Min.   :0.8881   Min.   :0.480
1st Qu.: 4090   1st Qu.:3.000   1st Qu.:  5.000   1st Qu.: 0.1200   1st Qu.: 0.0575   1st Qu.: -1.625   1st Qu.:-0.04125   1st Qu.:   3.00   1st Qu.:  37.0     1st Qu.:0.9866   1st Qu.:2.958
Median : 8099   Median :4.000   Median :  6.800   Median : 0.2750   Median : 0.3100   Median :  4.500   Median : 0.04400   Median :  32.00   Median : 126.0     Median :0.9940   Median :3.190
Mean   : 8063   Mean   :3.669   Mean   :  6.875   Mean   : 0.3018   Mean   : 0.3164   Mean   :  5.540   Mean   : 0.04797   Mean   :  32.45   Mean   : 124.9     Mean   :0.9937   Mean   :3.195
3rd Qu.:12021   3rd Qu.:5.000   3rd Qu.:  9.025   3rd Qu.: 0.6100   3rd Qu.: 0.5800   3rd Qu.:  15.600   3rd Qu.: 0.13325   3rd Qu.:  72.00   3rd Qu.: 209.0     3rd Qu.:1.0002   3rd Qu.:3.460
Max.   :16119   Max.   :8.000   Max.   : 32.500   Max.   : 3.6800   Max.   : 3.7700   Max.   : 140.650   Max.   : 1.27000   Max.   : 622.00   Max.   :1057.0     Max.   :1.0992   Max.   :5.940
   Sulphates         Alcohol       LabelAppeal         AcidIndex        STARS
Min.   :-3.1300   Min.   :-4.50   Min.   :-2.00000   Min.   : 4.00   Min.   :1.000
1st Qu.: 0.2600   1st Qu.: 9.00   1st Qu.:-1.00000   1st Qu.: 7.00   1st Qu.:1.000
Median : 0.4900   Median :10.50   Median : 0.00000   Median : 7.00   Median :2.000
Mean   : 0.5055   Mean   :10.56   Mean   : 0.04475   Mean   : 7.65   Mean   :2.039
3rd Qu.: 0.8200   3rd Qu.:12.40   3rd Qu.: 1.00000   3rd Qu.: 8.00   3rd Qu.:3.000
Max.   : 4.1100   Max.   :26.50   Max.   : 2.00000   Max.   :17.00   Max.   :4.000
```

Because data is rarely perfect and simply removing missing data introduces a certain type of bias, we can also try imputing missing data based on mean/grouping technique.

We first started by checking the variances of the variables with high ranges. Important to note that Standard Deviation is huge for 'FreesulfureDioxide' and 'TotalSulfureDioxide' and to a lesser extent 'ResidualSugar'. Also notice that 'FreesulfureDioxide' and 'TotalSulfureDioxide' appear to include negative values. 'STARS' has an exceptionally high number of missing values, as well as multiple chemical properties. It is very important to always note a LOG function cannot take negative values. We decided to use the KABLE function to provide us with a clear table of our variable stats. We decided to transform the 'FreesulfureDioxide' and 'TotalSulfureDioxide'. We also created NA's Flags in or to explore if they have an impact on 'TARGET'.

	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide
Stand dev	1.930000000	6.320000000	0.780000000	0.860000000	33.750000000	0.320000000	148.710000000
Mean	3.029073857	7.075717077	0.324103947	0.308412661	5.418733065	0.054822489	30.845571287
n	12795.000000000	12795.000000000	12795.000000000	12795.000000000	12795.000000000	12795.000000000	12795.000000000
Median	3.000000000	6.900000000	0.280000000	0.310000000	3.900000000	0.046000000	30.000000000
CoeffofVariation	0.635959475	0.892862644	2.419020945	2.795215292	NA	NA	NA
Minimum	0.000000000	-18.100000000	-2.790000000	-3.240000000	NA	NA	NA
Maximum	8.000000000	34.400000000	3.680000000	3.860000000	NA	NA	NA
Upper Quantile.100%	8.000000000	34.400000000	3.680000000	3.860000000	141.150000000	1.351000000	623.000000000
LowerQuartile.0%	0.000000000	-18.100000000	-2.790000000	-3.240000000	-127.800000000	-1.171000000	-555.000000000

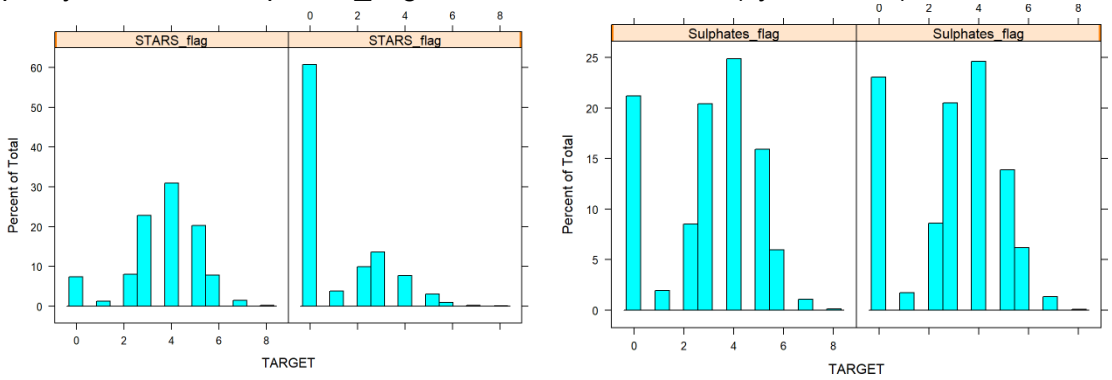
From the Data Exploration section you'll remember when running our correlation matrix that the correlation between 'STARS' and the target variable is high, the higher of a rating the more samples were requested by distribution companies for a specific brand. Very important to note, also when viewing our correlation matrix, in terms of our 'TARGET' variable and the complete variables, we also have 'LabelAppeal' appears to be a variable which is highly correlated with 'TARGET', 'AcidIndex' can be considered similarly so.

To recap, from the Data Exploration section you'll remember when running our correlation matrix that the correlation between 'STARS' and the target variable is high, the higher of a rating the more samples were requested by distribution companies for a specific brand. Very important to note, also when viewing our correlation matrix, in terms of our 'TARGET' variable and the complete variables, we also have 'LabelAppeal' appears to be a variable which is highly correlated with 'TARGET', 'AcidIndex' can be considered similarly so. # 'STARS', 'AcidIndex', 'LabelAppeal' are the ones with the highest correlation with the 'TARGET' variable. Basically, for every value of 'STARS', ie 'STARS=1 or =2 or =3 or =4, it shows the distribution of the 'TARGET' variable in each case. You'll notice, for example when 'STARS'=4, the target variable generally has more high value than low values. #'LabelAppeal' has a range from min -2 and max 2, this can possibly be thought of as categorical. Using 'LabelAppeal' we can think of -2 as "very Bad", 1 as "Bad", 0 as neutral, 1 as "good", 2 as "very good". Also quite important to note the higher the label appeal, then the higher the demand for specific brands.
Note 'AcidIndex' is NOT categorical, we cannot put one graph per every value of 'AcidIndex'. Instead we can switch and use 'TARGET' and have 'AcidIndex' as the histogram. Viewing the histogram, we can say that the ones with high 'TARGET' value should have lower 'AcidIndex'. As we go lower and lower according to the 'TARGET' variable then the 'AcidIndex' should get higher. This makes sense since 'AcidIndex' appeared to be negatively correlated (-0.25). We also decided to explore with Analysis of variance which is a collection of statistical models and their associated estimation procedures used to analyze the differences among means. From this we note that the acid index has a potential effect on the TARGET variable.

Before we continue the data preparation we witness randomness in some variables,these are ‘ResidualSugar’,’Chlorides’,’FreeSulfurDioxide’,’TotalSulfurDioxide’,’pH’,’Alcohol’.
 the 'missingness' may not have any indications, we impute using mean or conditional mean on target. We are creating groups based on TARGET, replacing missing values. We decided to utilize log transformation for the variables with high variance, the variables are translated first so that we get rid of the negative values so that the log function can handle them.
We transform 'y' like in last assignment, instead of log(y), we can do y + 1 - min(y)
For example, the minimum of 'FreeSulfurDioxide' is -555, this will transform all the values of 'FreeSulfurDioxide by 555, so the -555 is not in the negative range(will be positive) and log() will be able to handle. A common technique for handling negative values is to add a constant value

to the data prior to applying the log transform. The transformation is therefore log(Y+a) where a is the constant. Some people like to choose a so that min(Y+a) is a very small positive number (like 0.001). Others choose a so that min(Y+a) = 1. For the latter choice, you can show that a = b – min(Y), where b is either a small number or is 1.

After inspecting you'll notice 'Sulphates' and 'STARS' need careful consideration as the percentage of N/A's is significant. We noticed some variables with "low-ish" N/A's, when we explored the data we saw that every single brand has an N/A in these properties. It may be that these N/A's are random. It may not be the case that specific brands are 'bad' or 'low price' since they have N/A's in certain properties. The N/A's we see dispersed among the different brands. We made the left equal to stars_flag=0 'no missing' and right side graph=1 which means 'missing data'. Our histogram on the left 'STARS_flag' has no missing data, the average is higher. The histogram on the right 'STARS_flag' has missing data, the average is generally lower. We might assume that those with missing data generally are less purchases or of lower quality. You'll see 'sulphates_flag' with 0 and 1 are similar(symmetrical).



For our corrective measures you'll notice using the mutate() function we perform action on 'Sulphates' such as if values have N/A it will be replaced with the mean. Next, 'STARS'(created stars merged) such as if there is a missing value it will equal to 0, if it has a value it will stay the same.
Running a 'table' function you'll see before 'STARS' had no zero column, following the corrective measure 'STARS_merged' now has a zero column. Additionally, another step we took for analysis was creating acid index clusters. k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. From this output we generated a histogram, we notice the higher the 'AcidIndex' the lower the target variable. Other clusters are almost identical in shape, generally even distributed It is better to utilize the original 'AcidIndex' variable.

Build Models

Results

Multiple Regression Models

I. We chose only the variables that were significantly related with the outcome in binary tests STARS, Label Appeal and Acid Index.

With an increase in STAR rating, the number of wine cases sold will increase, $p < .05$. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, $p< .05$. Acid Index was negatively predictive of wine case sale, $\beta = -0.21$.

An R-squared of .52 suggests the covariates predict 52% of the variance in the model.

II. For the second model, we added all the variables to the model to see which covariates predict the outcome .i.e. number of cases of wine that will be sold.

An increase in fixed acidity means resistance to microbial infection. According to the model, fixed acidity was not predictive of wine cases sale with $p=0.99$.

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, $\beta = -0.10$.

Excessive citric acid affects wine aroma negatively. In our model, the direction of the estimate was opposite of what should be expected but Citric Acid content and Residual Sugar were not significantly predictive of wine case sales.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sales.

Density of wine was not predictive of wine case sales. However, we saw an inverse trend of pH and a direct trend between alcohol content and wine sales, $\beta = 0.01$.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sales at $p< 0.05$.

Similar to what we observed in our selected model I, with an increase in STAR rating, the number of wine cases sold will increase, $p < .05$. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, $p< .05$. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, $\beta = -0.21$.

There was little difference in the R-Squared value (=0.53) compared to Model I (=0.52).

III. For the third model, we ran the forward selection method so that the model can identify an optimum model. The forward model ended up including all the variables that we originally added in the model

An increase in fixed acidity means resistance to microbial infection. According to the model, fixed acidity was not predictive of wine cases sales with p=0.99.

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, β = -0.10 similar to Model II.

Excessive citric acid affects wine aroma negatively. In our model, the direction of the estimate was opposite of what should be expected but Citric Acid content and Residual Sugar were not significantly predictive of wine case sale, P> 0.05.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sales.

Density of wine was not predictive of wine case sales. However, we saw an inverse trend of pH and a direct trend between alcohol content and wine sales, β = 0.01 similar to Model II.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sale at p< 0.05.

Similar to what we observed in our selected model and Model III, with an increase in STAR rating, the number of wine cases sold will increase, p < .05. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, p< .05. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, β = -0.21.

The R-Squared value was comparable to Model I and Model II (=0.53).

IV. For the fourth model, we used the backward selection method.

The backward model excluded a few variables such as Fixed Acidity, Acid Index and Residual Sugar.
The R-squared did not change for the model compared to other models (=0.53).

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, β = -0.10 similar to Model II and III.

Excessive citric acid affects wine aroma negatively. In our model, the direction of the estimate was opposite of what should be expected but it was not significantly predictive of wine case sales, P> 0.05.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sales.

Density of wine was not predictive of wine case sales. However, we saw an inverse trend of pH and a direct trend between alcohol content and wine sales, $\beta = 0.01$ similar to Model II and III.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sales at $p< 0.05$.

Similar to what we observed in our selected model and Model III, with an increase in STAR rating, the number of wine cases sold will increase, $p < .05$. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, $p< .05$. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, $\beta = -0.21$.

Poisson Regression Models

I. Similar to Model I in Multiple Regression (MR), we chose only the variables that were significantly related with the outcome in binary tests STARS, Label Appeal and Acid Index.

We saw a similar trend as we did with MR and PR model I. With an increase in STAR rating, the number of wine cases sold will increase, $p < .05$, except the β estimate was smaller at 0.31 compared to MR 0.99 unit increase in Sales. A similar trend was seen with Label Appeal, $\beta_{poisson} = 0.13$ vs $\beta_{multiple} = 0.43$. A higher marketing score on Label design indicated higher sales of wine cases, $p< .05$. Acid Index was negatively predictive of wine cases sales, $\beta_{poisson} = -0.21$ vs $\beta_{multiple} = -0.09$.

This can be attributed to the assumptions regarding the distributions of the two methods where one considers the data to be normally distributed while this regression methods the Poisson distribution which accounts for correlation between rows of observations.

AIC indicates model fit. In this case, the AIC for our model was 46754.

II. For the second model, we added all the variables to the model to see which covariates predict the sale of wine cases similar to Model II of Multiple Regression (MR).

An increase in fixed acidity means resistance to microbial infection. According to the model, fixed acidity was not predictive of wine cases sale with $p=0.99$ but the direction of the effect was accurate.

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, $\beta_{poisson} = -0.03$ compared to $\beta_{multiple} = -0.10$.

Excessive citric acid affects wine aroma negatively. In our model, the direction of the estimate was opposite of what should be expected but Citric Acid content and Residual Sugar were not significantly predictive of wine case sales similar to MR models.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sale, $\beta_{poisson} = -0.04$ compared to $\beta_{multiple} = -0.12$.

Density of wine was not significantly predictive of wine case sales however, like MR, in our poisson model, pH was significantly predictive of wine sales such that a unit decrease predicted an increase in sales while controlling for other covariates, $\beta_{poisson} = -0.02$ vs $\beta_{multiple} = -0.12$

An opposite trend is observed in case of alcohol content and wine sales, $\beta = 0.01$ similar to MR models but alcohol was not significantly predicted in our model, $p > .05$.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales, $\beta_{poisson} = -0.01$ vs $\beta_{multiple} = -0.03$.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sales at $p< 0.05$.

Similar to what we observed in our MR models, with an increase in STAR rating, the number of wine cases sold will increase, $p < .05$. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, $p< .05$. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, $\beta_{poisson} = -0.08$ vs $\beta_{multiple} = -0.10$.

The AIC for our model was 46706 which is lower than Model I which had lesser variables selected based on bivariate tests.

III. For the third model, we ran the forward selection regression. All the variables were included in the model like Model III of MR.

An increase in fixed acidity means resistance to microbial infection. According to the model, fixed acidity was not predictive of wine cases sale with $p=0.99$ but the direction of the effect was accurate.

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, $\beta_{poisson} = -0.03$ compared to $\beta_{multiple} = -0.10$.

Excessive citric acid affects wine aroma negatively. In our model, the direction of the estimate was opposite of what should be expected but Citric Acid content and Residual Sugar were not significantly predictive of wine case sales similar to MR models.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sale, βpoisson = -0.04 compared to βmultiple = -0.12.

Density of wine was not significantly predictive of wine case sales however, like MR, in our poisson model, pH was significantly predictive of wine sales such that a unit decrease predicted an increase in sales while controlling for other covariates, βpoisson = -0.02 vs βmultiple = -0.12

An opposite trend is observed in case of alcohol content and wine sales, β = 0.01 similar to MR models but alcohol was not significantly predicted in our model, $p > .05$.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales, βpoisson = -0.01 vs βmultiple = -0.03.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sales at $p< 0.05$.

Similar to what we observed in our MR models, with an increase in STAR rating, the number of wine cases sold will increase, $p < .05$. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, $p< .05$. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, βpoisson = -0.08 vs βmultiple = -0.10.

The AIC for our model was 46706 similar to Model II since both the models have the same predictors.

IV. For the fourth model, we used the backward selection method. The backward model excluded a few variables such as Fixed Acidity Citric Acid, and Residual Sugar.

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, βpoisson = -0.03 compared to βmultiple = -0.10.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sale, βpoisson = -0.04 compared to βmultiple = -0.12.

Density of wine was not significantly predictive of wine case sales however, like MR, in our poisson model, pH was significantly predictive of wine sales such that a unit decrease predicted an increase in sales while controlling for other covariates, βpoisson = -0.02 vs βmultiple = -0.12

An opposite trend is observed in case of alcohol content and wine sales, β = 0.01 similar to MR models but alcohol was not significantly predicted in our model, $p > .05$.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales, βpoisson = -0.01 vs βmultiple = -0.03.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sales at $p < 0.05$.

Similar to what we observed in our MR models, with an increase in STAR rating, the number of wine cases sold will increase, $p < .05$. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, $p < .05$. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, βpoisson = -0.08 vs βmultiple = -0.10.

The AIC for our model was 46706 similar to Model II and III even though we have lesser variables

Negative Binomial Regression Models

Negative binomial regression is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model. The traditional negative binomial regression model is based on the Poisson-gamma mixture distribution. As can be seen from the results, the estimate sizes did not change when using the negative binomial regression method compared with Poisson.

I. Similar to Model I in Multiple Regression and Poisson Regression, we chose only the variables that were significantly related with the outcome in binary tests STARS, Label Appeal and Acid Index.

We saw a similar trend as we did with multiple regression (MR) model I and Poisson Regression. The Beta estimate sizes are the same as that of the Poisson model.

With an increase in STAR rating, the number of wine cases sold will increase, $p < .05$, except the β estimate was smaller at 0.31 compared to 0.99 unit increase in Sales. A similar trend was seen with Label Appeal, βpoisson/-ve = 0.13 vs βmultiple = 0.43. A higher marketing score on Label design indicated higher sales of wine cases, $p < .05$. Acid Index was negatively predictive of wine cases sales, βpoisson/-ve = -0.21 vs βmultiple = -0.09.

II. For the second model, we added all the variables to the model to see which covariates predict the sale of wine cases similar to Model II of Multiple Regression (MR).

An increase in fixed acidity means resistance to microbial infection. According to the model, fixed acidity was not predictive of wine cases sales with p=0.99 but the direction of the effect was accurate.

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, βpoisson/-ve = -0.03 compared to βmultiple = -0.10.

Excessive citric acid affects wine aroma negatively. In our model, the direction of the estimate was opposite of what should be expected but Citric Acid content and Residual Sugar were not significantly predictive of wine case sales similar to MR models.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sale, βpoisson/-ve = -0.04 compared to βmultiple = -0.12.

Density of wine was not significantly predictive of wine case sales however, like MR, in our poisson model, pH was significantly predictive of wine sales such that a unit decrease predicted an increase in sales while controlling for other covariates, βpoisson/-ve = -0.02 vs βmultiple = -0.12

An opposite trend is observed in case of alcohol content and wine sales, β = 0.01 similar to MR models but alcohol was not significantly predicted in our model, $p > .05$.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales, βpoisson/-ve = -0.01 vs βmultiple = -0.03.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sales at $p< 0.05$.

Similar to what we observed in our MR models, with an increase in STAR rating, the number of wine cases sold will increase, $p < .05$. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, $p< .05$. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, βpoisson/-ve = -0.08 vs βmultiple = -0.10.

The AIC for our model was 46708 which is lower than Model I which had lesser variables selected based on bivariate tests.

III. For the third model, we ran the forward selection regression. All the variables were included in the model like Model III of MR.

An increase in fixed acidity means resistance to microbial infection. According to the model, fixed acidity was not predictive of wine cases sale with p=0.99 but the direction of the effect was accurate.

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, βpoisson/-ve = -0.03 compared to βmultiple = -0.10.

Excessive citric acid affects wine aroma negatively. In our model, the direction of the estimate was opposite of what should be expected but Citric Acid content and Residual Sugar were not significantly predictive of wine case sales similar to MR models.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sale, $\beta_{poisson/-ve} = -0.04$ compared to $\beta_{multiple} = -0.12$.

Density of wine was not significantly predictive of wine case sales however, like MR, in our poisson model, pH was significantly predictive of wine sales such that a unit decrease predicted an increase in sales while controlling for other covariates, $\beta_{poisson/-ve} = -0.02$ vs $\beta_{multiple} = -0.12$

An opposite trend is observed in case of alcohol content and wine sales, $\beta = 0.01$ similar to MR models but alcohol was not significantly predicted in our model, $p > .05$.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales, $\beta_{poisson/-ve} = -0.01$ vs $\beta_{multiple} = -0.03$.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sales at $p< 0.05$.

Similar to what we observed in our MR models, with an increase in STAR rating, the number of wine cases sold will increase, $p < .05$. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, $p< .05$. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, $\beta_{poisson/-ve} = -0.08$ vs $\beta_{multiple} = -0.10$.

The AIC for our model was 46708 similar to Model II since both the models have the same predictors.

IV. For the fourth model, we used the backward selection method. The backward model excluded a few variables such as Fixed Acidity Citric Acid, and Residual Sugar.

An increase in volatile acidity means spoilage of wine. In our model, a decrease in volatile acidity was significantly predictive of wine sales, $\beta_{poisson/-ve} = -0.03$ compared to $\beta_{multiple} = -0.10$.

Sodium chloride adds to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Our model indicated similar findings such that a decrease in Chloride content of wine was predictive of increased wine cases sale, $\beta_{poisson/-ve} = -0.04$ compared to $\beta_{multiple} = -0.12$.

Density of wine was not significantly predictive of wine case sales however, like MR, in our poisson model, pH was significantly predictive of wine sales such that a unit decrease predicted an increase in sales while controlling for other covariates, $\beta_{poisson/-ve} = -0.02$ vs $\beta_{multiple} = -0.12$

An opposite trend is observed in case of alcohol content and wine sales, β = 0.01 similar to MR models but alcohol was not significantly predicted in our model, p > .05.

Sulfates in wine protect against oxidation, which can affect the color and taste of wine and prevent the growth of unwanted microorganisms. A decrease in sulfates was predictive of increased wine sales, βpoisson/-ve = -0.01 vs βmultiple = -0.03.

Sulfur dioxide, the main preservative used in wine, is produced naturally through fermentation. An increase in log of Free Sulfur Dioxide and log of Total Sulfur Dioxide was predictive of increased wine cases sales at p< 0.05.

Similar to what we observed in our MR models, with an increase in STAR rating, the number of wine cases sold will increase, p < .05. A similar trend was seen with Label Appeal. A higher marketing score on Label design suggests higher sale of wine cases, p< .05. The estimate size was similar for the Acid Index variable that negatively predicted sale of wine cases, βpoisson/-ve = -0.08 vs βmultiple = -0.10.

The AIC for our model was 46704 similar to Model II and III even though we have lesser variables

Select Models

Multiple Regression Models

Model II with ONLY the variables that are significantly associated with wine sales because we're interested in prediction over estimation. So we use all the variables that are predictive.

Poisson Regression Models

Model II with ONLY the variables that are significantly associated with wine sales because we're interested in prediction over estimation. So we use all the variables that are predictive.

Negative Binomial Regression Models

Model II with ONLY the variables that are significantly associated with wine sales because we're interested in prediction over estimation. So we use all the variables that are predictive.

Appendix

Github Link

.RMD - https://github.com/johnm1990/msds-621/blob/main/Assignment5_fin.Rmd
.PDF(Knitted) - https://github.com/johnm1990/msds-621/blob/main/Assignment5_fin.pdf
Predictions - https://github.com/johnm1990/msds-621/blob/main/Predictions_5/predict1.csv