NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Week 10 Assignment - Modeling: The Wine Study

MSDS 410

INTRODUCTION

This data set contains information on approximately 12,000 commercially available wines. A record can be considered the data associated with a bottle of wine. The variables are mostly related to the chemical properties of the wine being sold. But data comes from other sources as well. For example, the PURCHASE reflects whether or not a purchase was made of that wine any wine distribution companies after sampling a wine. The variable CASES then indicates the number of cases purchased. 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 a wine is to be sold at a high end restaurant. Similarly, each wine, when possible, was rated by a group of experts as to its quality (STARS).

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. But, it is also important to understand what influences purchase decisions as well as what contributes to the quality of the wine.

For this project, you may choose one of three response variables for which you will build a predictive model. The possible response variables are:

Purchase Decision (PURCHASE)
The rating of the wine (STARS)
The number of cases of wine sold (CASES)

Please note, the variable STARS can also be an explanatory variable for the PURCHASE and CASES variables. You can only use the variables given to you, or variable that you derive from the variables provided. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You are welcome to use OLS regression, Logistic regression, Poisson Regression, or Zero-Inflated Poisson Regression methods in fitting these models. Be sure you know which one fits your situation.

All topics that have been learned over the term should be employed in this data modeling assignment. This includes: EDA, use of categorical and continuous explanatory variables, model fit statistics and diagnostics, automated selection methods or some other approach to variable selection, validation, sample splitting, and model refinement and cleaning. From a statistical perspective, please note the size of the sample: n is approximately 12,000 records. You should immediately be thinking, "I have tons of statistical power." I have to be careful about statistical significance, as it is not the be all and end all. Again, you can think about randomly splitting the file into a 70% model development dataset, and into 30% validation data set.

As you proceed to the write up for this assignment, please keep in mind that you are writing this assignment for a manager or boss. What you should do for this assignment:

TASK 1: EXPLORATORY DATA ANALYSIS and DATA PREP

Use your data analysis knowledge to date, to conduct an Exploratory Data Analysis (EDA). Some suggestions for things that you could do are:

a. Means, standard deviations, minimum, maximum, median for all categorical and continuous variables

```
> mysummary(mydataW)
```

	Type	Nulls	Nulls_per	Min	Q_1st	StDev	Median	Mean	Q_3rd	Max	Range	Skewness	Kurtosis
Purchase	integer	0	0.0	0.0	1.0	0.4	1.0	0.8	1.0	1.0	1.0	-1.4	0.0
Cases	integer	0	0.0	0.0	2.0	1.9	3.0	3.0	4.0	8.0	8.0	-0.3	-0.9
STARS	integer	3359	26.3	1.0	1.0	0.9	2.0	2.0	3.0	4.0	3.0	0.4	-0.7
FixedAcidity	numeric	0	0.0	-18.1	5.2	6.3	6.9	7.1	9.5	34.4	52.5	0.0	1.7
VolatileAcidity	numeric	0	0.0	-2.8	0.1	0.8	0.3	0.3	0.6	3.7	6.5	0.0	1.8
CitricAcid	numeric	0	0.0	-3.2	0.0	0.9	0.3	0.3	0.6	3.9	7.1	-0.1	1.8
ResidualSugar	numeric	616	4.8	-127.8	-2.0	33.7	3.9	5.4	15.9	141.2	268.9	-0.1	1.9
Chlorides	numeric	638	5.0	-1.2	0.0	0.3	0.0	0.1	0.2	1.4	2.5	0.0	1.8
FreeSulfurDioxide	numeric	647	5.1	-555.0	0.0	148.7	30.0	30.8	70.0	623.0	1178.0	0.0	1.8
TotalSulfurDioxide	numeric	682	5.3	-823.0	27.0	231.9	123.0	120.7	208.0	1057.0	1880.0	0.0	1.7
Density	numeric	0	0.0	0.9	1.0	0.0	1.0	1.0	1.0	1.1	0.2	0.0	1.9
pH	numeric	395	3.1	0.5	3.0	0.7	3.2	3.2	3.5	6.1	5.7	0.0	1.6
Sulphates	numeric	1210	9.5	-3.1	0.3	0.9	0.5	0.5	0.9	4.2	7.4	0.0	1.8
Alcohol	numeric	653	5.1	-4.7	9.0	3.7	10.4	10.5	12.4	26.5	31.2	0.0	1.5
LabelAppeal	integer	0	0.0	-2.0	-1.0	0.9	0.0	0.0	1.0	2.0	4.0	0.0	-0.3
AcidIndex	integer	0	0.0	4.0	7.0	1.3	8.0	7.8	8.0	17.0	13.0	1.6	5.2

There are 12,795 wine sampling observations and 16 variables. There are 12 continuous variables, and they represent the chemical characteristics of the wine. There are five categorical variables and they come integer/ discrete form. The categorical variables have the marketing scores, such as the visual appeal of the label and wine rating information such as a number of stars. Our target variable can be predicting purchases, a number of cases purchased, or star rating of each wine. Let's start examining the integer/categorical variables first.

- Purchase: We are interested in the factors that influence whether a purchase decision made. The outcome (response) variable is binary (0/1); the wine was either purchased or not purchased by wine distributors after sampling them. By looking at the median and mean values, there are about 80% of purchased transactions than non-purchased; so, the skewness here is -1.4. This value implies that the distribution of the data is highly skewed to the left or negatively skewed. About 80% of the wine has been purchased and 20% has not been.
- Cases: The number of cases of wine sold ranges from 0 to 8. The mean and median are very close to each other but slightly skewed to the left. Negative excess kurtosis indicates a thin-tailed data distribution. The average is three cases per customer or transaction.
- Stars: 26% of the data is missing. The rating score of wine ranges from 1 to 4. Mean and median are very close to each other but slightly skewed to the right. Since skewness is positive, most of the star scores cluster around the lower end rating (ones and twos).
- Label Appeal: it ranges from -2 to 2. By looking at mean, median, and skewness, we can tell that it is normally distributed. According to the small kurtosis number, there should not be outliers. However, a negative score is not meaningful; we can relabel it from 1 to 5.
- Acid Index: it ranges from 4 to 17. Median is higher than mean; the data highly skewed to the right. We should see long tails on the right side of the distribution. Positive excess kurtosis indicates a fat-tailed distribution. There should be a lot of outliers. It has the largest/broadest range among the categorical variables

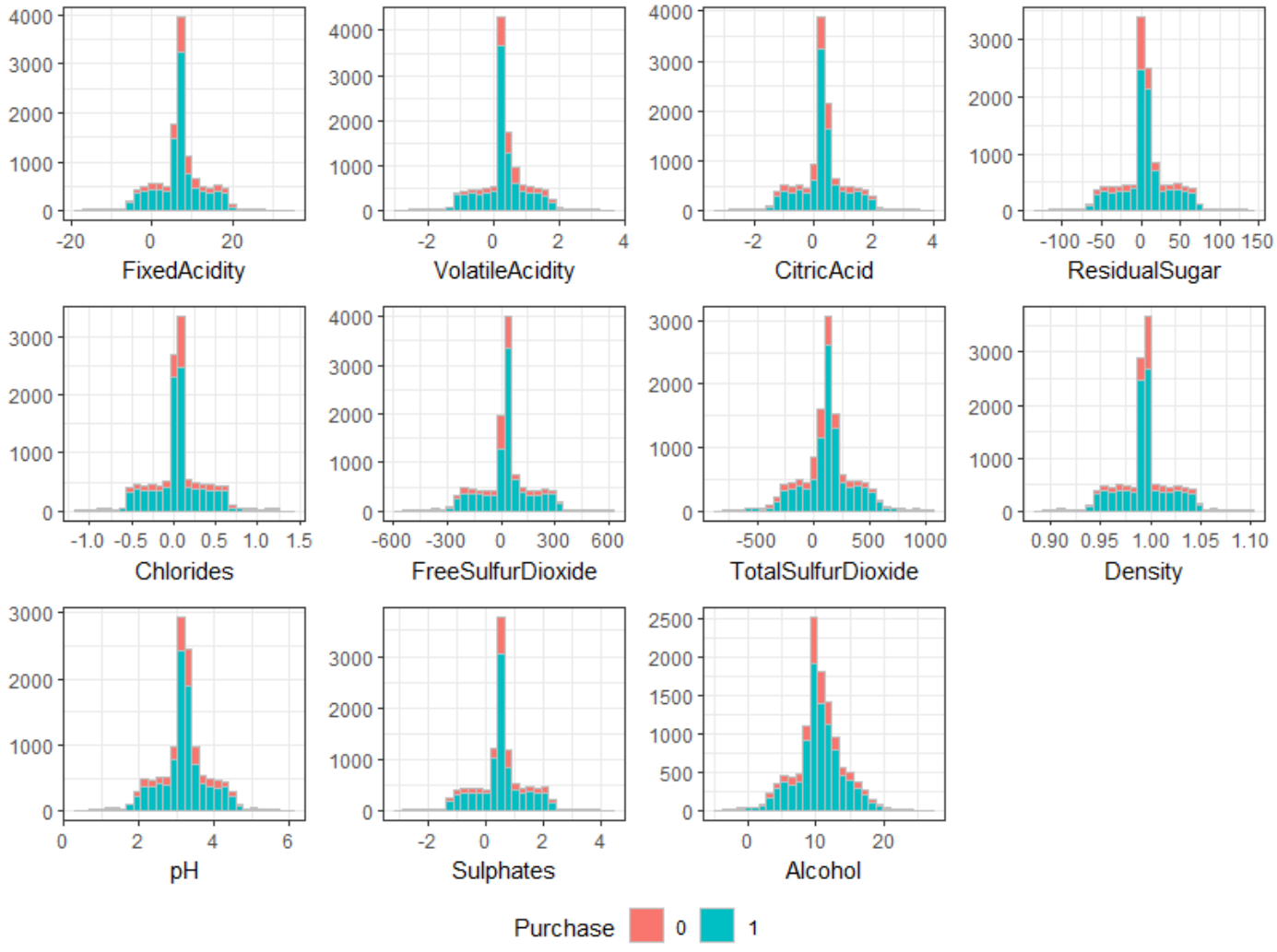
Let's examine continuous variables.

- There are some missing values: Residual Sugar, Chloride, Free Sulfur Dioxide, and Alcohol predictors are missing about 5% of data. Total Sulfur Dioxide, ph, and Sulphates are missing 5.3%, 3.1%, and 9.5% data, respectively. Interestingly some wine attributes are negative such as alcohol. Alcohol has 0.9% negative signs. According to the table below, ten attributes have large numbers of negative values (mostly between 12% and 25%, except alcohol), and they seem to occur randomly.
- The skewness of all continues data is around 0. However, the mean is larger than the median in Residual Sugar.
- All kurtoses are between positive 1.5 and 1.9 numbers. They have heavy tails or outliers.

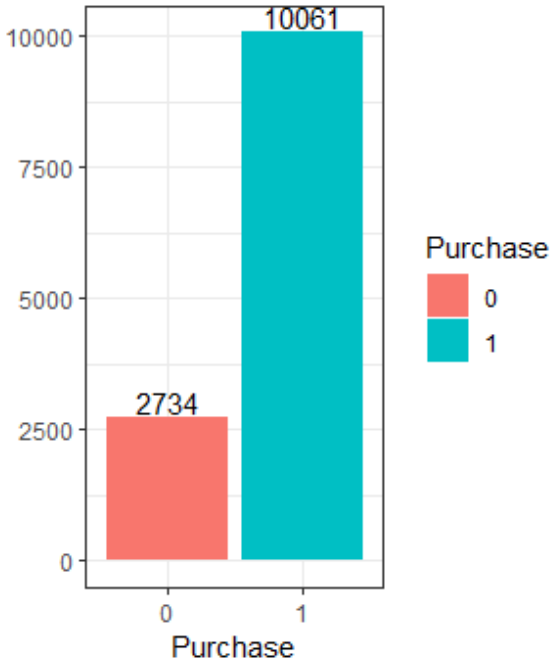
	negatives	negative_per
Purchase	0	0.0
Cases	0	0.0
STARS	0	0.0
FixedAcidity	1621	12.7
VolatileAcidity	2827	22.1
CitricAcid	2966	23.2
ResidualSugar	3136	24.5
Chlorides	3197	25.0
FreeSulfurDioxide	3036	23.7
TotalSulfurDioxide	2504	19.6
Density	0	0.0
pH	0	0.0
Sulphates	2361	18.5
Alcohol	118	0.9
LabelAppeal	3640	28.4
AcidIndex	0	0.0

b. Histograms for each continuous variable

We can expect the same results in histogram charts as we had seen their skewness and kurtoses. We can see that they all distributed symmetrically but have long tails on both sides. We can see that Alcohol has the lowest kurtosis, among others, and it has shorter or lighter tails on both sides. There are some outliers. We can do further analysis such as boxplot, the continuous variables.



c. Histograms for each categorical variable



- About 21% of customers did not purchase wines, and 79% of customers made a purchase. So, there were more sale transactions than no sale transactions.
- Wines sold with a number of cases ranged from 1 to 8. Obviously, non-purchased wines did not have cases, so it was zero. About 70% of the time wines were sold with cases from 2 to 5. Without zero cases, it is somewhat normally distributed. So, it is a zero-inflated Poisson distribution.
- 26% of star ratings were missing. About 39% of purchased items (1312/3359) did not have star ratings. As star ratings increase from 1 to 4, the sales decreases. There are some non-purchased wines in star rating 1, but it significantly decreases in star rating 2. Star 3 and 4 contain sales transactions only. The distribution is skewed to the right, where about 70% of our wines have a bad rating of one or two.
- Label appeal was distributed consistently (79% and 20%) among purchased and non-purchased items.
- We noticed that Acid Index had the highest kurtoses, which indicated long tails. It was also skewed right. The acid Index can be regrouped. Since there is a wide range (goes from 5 to 15), it would be a potentially good predictor. There is no potential predictor if we had a very small range or constant bars.

d. Are variables correlated to the target variable (TARGET_WINS) or to other possible explanatory variables?

	cor_to_Cases	cor_to_Stars	cor_to_Purchase
Cases	1	0.55	0.67
Purchase	0.67	0.29	1
STARS	0.55	1	0.29
LabelAppeal	0.5	0.32	0.01
Alcohol	0.07	0.06	0
TotalSulfurDioxide	0.02	0.02	0.05
FreeSulfurDioxide	0.02	-0.02	0.02
ResidualSugar	0	0.02	0.03
CitricAcid	0	0.01	0
pH	0	0	-0.02
FixedAcidity	-0.01	0	-0.03
Sulphates	-0.02	-0.02	-0.03
Chlorides	-0.03	-0.01	-0.02
Density	-0.05	-0.03	-0.02
VolatileAcidity	-0.08	-0.04	-0.06
AcidIndex	-0.17	-0.1	-0.19

I used "complete.obs" to ignore NA or NULL values. We can see that Purchases, Cases, and Stars variables have a stronger positive correlation with each other. Label design appeal also correlates well with a number of cases and wine rating stars. Acid Index has a weaker negative correlation. Chemical attributes of wine have only a minimal relationship with the amount of that wine that is purchased or purchased in a number of cases except label appeal and acid index.

```
> table(mydataW$Purchase, mydataW$STARS)

      0    1    2    3    4
  0 2038  607   89    0    0
  1 1321 2435 3481 2212  612
> table(mydataW$Purchase, mydataW$Cases)

      0    1    2    3    4    5    6    7    8
  0 2734    0    0    0    0    0    0    0    0
  1    0  244 1091 2611 3177 2014  765  142   17
> table(mydataW$Cases, mydataW$STARS)

      0    1    2    3    4
  0 2038  607   89    0    0
  1  126   98   20    0    0
  2  335  469  253   34    0
  3  457  916  948  290    0
  4  260  716 1333  764  104
  5  101  214  716  750  233
  6   32   22  199  313  199
  7    8    0   12   57   65
  8    2    0    0    4   11
```

We can see there is a strong association between star and purchase and star and cases.
We can see that the higher stars lead to more purchases.
We can see that nonzero cases have purchased.
We can see that lower stars have lower cases and higher stars have higher cases.
So, stars can be a good predictor for purchases or a number of cases.

e. Are any of the variables with missing values that need to be imputed or “fixed”? Fix missing values (maybe with a Mean or Median value or use a decision tree). Are there variables with so many missing values that the entire variable should be eliminated from the analysis?

We can look at our summary table above to identify missing values and skewness to decide on imputation.

- 26% of star rating data is missing. We should expect missing ratings because not every customer leaves star ratings or reviews. We can reclassify NA to 0 so we can predict how many missing ratings we can have.
- Chloride, Free Sulfur Dioxide, and Alcohol variables with missing about 5% values can be imputed with mean or median values because their skewness is zero or distributed symmetrically. Most of the data is distributed around the mean data. Mean and median values are almost the same when skewness is zero.
- Total Sulfur Dioxide, ph, and Sulphates are missing 5.3%, 3.1%, and 9.5% data, respectively. Their skewness is also zero, which we can attribute them with mean or median values.
- Residual Sugar is slightly skewed to the left, -0.1. Since it is very close to zero and 4.8 of data was missing, we can impute it also with the mean value.

f. Do any of the variables have outliers or extreme values? Should these extreme values be replaced? Fix any extreme values that need fixing.

Earlier, we identified ten predictors that have negative values. The nine continuous negative variables might be the result of some sort of data entry error, and to fix this, we can replace the negative values with absolute value. Also, we can relabel the categorical design appeal marketing score from -2 – 2 to 1 – 5.

We can look at our summary table above to identify outliers by looking at kurtosis values. We can see more outliers or heavier tails if kurtosis values are away from zero.

- The heaviest tail is the Acid Index. Since it is a categorical variable, we can regroup the heaviest tails to one.
- The kurtoses range from 1.5 to 1.9 for all continuous variables. We can see them in a boxplot.

g. Do any of the variables need a mathematical transformation, such as log or square root? Create new variables with these transformations and add them to the end of the dataset.

h. Create any new variables that you are interested in.

We can regroup/lump the cases 1-2, 3-4, 5-6, 7-8 because there is a lot of noise. 0 means not bought and above zero means bought in a number of cases.

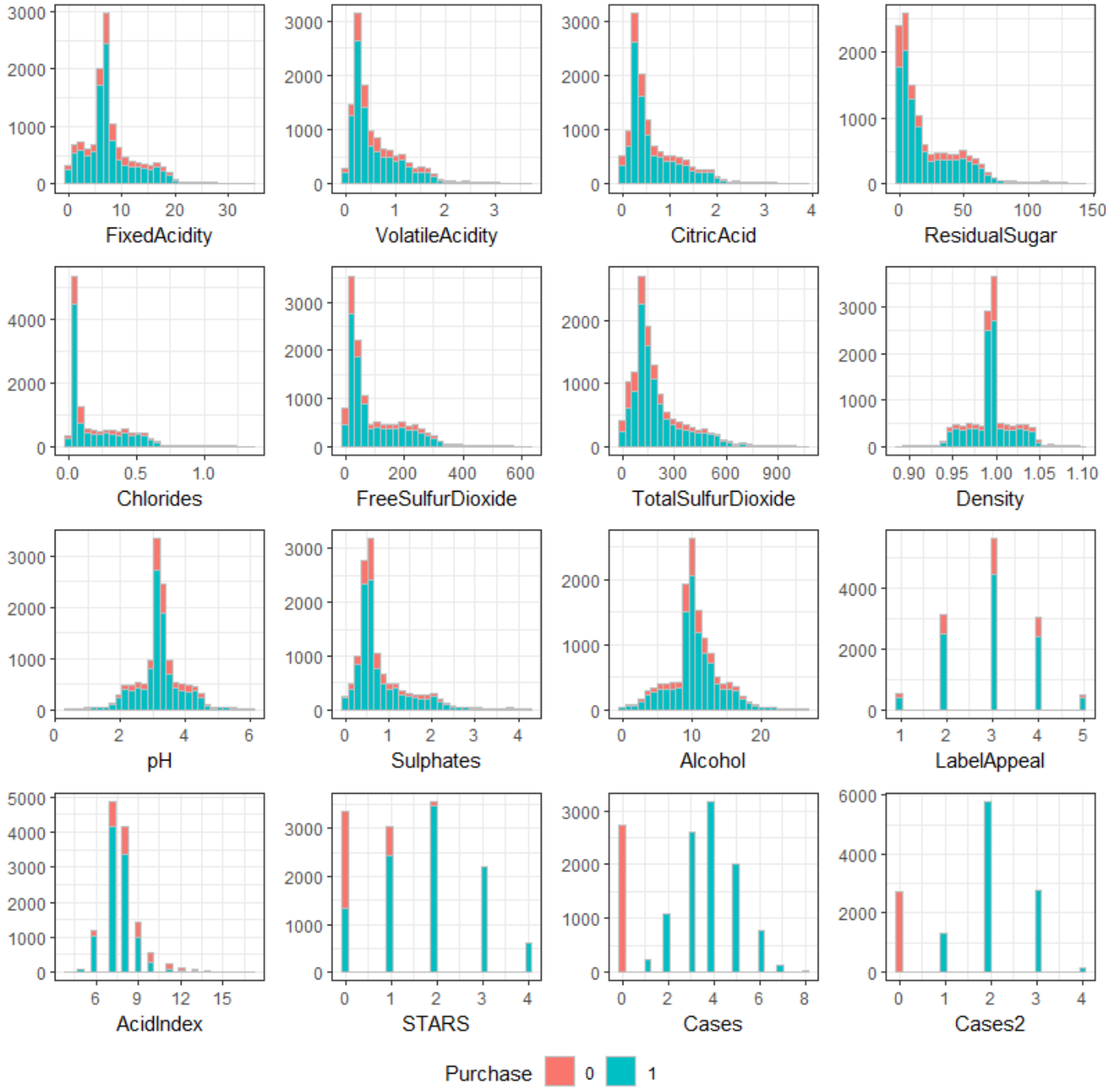
Stars		0	1	2	3	4
Cases	0	61%	20%	2%	0%	0%
	1	4%	3%	1%	0%	0%
	2	10%	15%	7%	2%	0%
	3	14%	30%	27%	13%	0%
	4	8%	24%	37%	35%	17%
	5	3%	7%	20%	34%	38%
	6	1%	1%	6%	14%	33%
	7	0%	0%	0%	3%	11%
	8	0%	0%	0%	0%	2%

Here is the table after regrouping the cases.

Stars		0	1	2	3	4
Cases	0	61%	20%	2%	0%	0%
	1	14%	19%	8%	2%	0%
	2	21%	54%	64%	48%	17%
	3	4%	8%	26%	48%	71%
	4	0%	0%	0%	3%	12%

i. Rerun the tables and graphs after cleaning and restructuring the data.

	Type	Nulls	Nulls_per	Min	Q_1st	StDev	Median	Mean	Q_3rd	Max	Range	Skewness	Kurtosis
Purchase	integer	0	0	0.0	1.0	0.4	1.0	0.8	1.0	1.0	1.0	-1.4	0.0
Cases	integer	0	0	0.0	2.0	1.9	3.0	3.0	4.0	8.0	8.0	-0.3	-0.9
STARS	integer	0	0	0.0	0.0	1.2	1.0	1.5	2.0	4.0	4.0	0.3	-0.9
FixedAcidity	numeric	0	0	0.0	5.6	5.0	7.0	8.1	9.8	34.4	34.4	1.2	2.0
VolatileAcidity	numeric	0	0	0.0	0.2	0.6	0.4	0.6	0.9	3.7	3.7	1.7	3.1
CitricAcid	numeric	0	0	0.0	0.3	0.6	0.4	0.7	1.0	3.9	3.9	1.6	2.9
ResidualSugar	numeric	0	0	0.0	4.0	24.6	11.8	22.5	37.2	141.2	141.2	1.5	2.5
Chlorides	numeric	0	0	0.0	0.0	0.2	0.1	0.2	0.4	1.4	1.4	1.6	2.4
FreeSulfurDioxide	numeric	0	0	0.0	29.0	106.6	51.0	102.8	164.0	623.0	623.0	1.6	2.7
TotalSulfurDioxide	numeric	0	0	0.0	102.0	159.8	147.5	199.9	251.0	1057.0	1057.0	1.7	3.4
Density	numeric	0	0	0.9	1.0	0.0	1.0	1.0	1.0	1.1	0.2	0.0	1.9
pH	numeric	0	0	0.5	3.0	0.7	3.2	3.2	3.5	6.1	5.7	0.0	1.8
Sulphates	numeric	0	0	0.0	0.4	0.6	0.6	0.8	1.0	4.2	4.2	1.9	3.9
Alcohol	numeric	0	0	0.0	9.1	3.5	10.5	10.5	12.2	26.5	26.5	0.2	1.3
LabelAppeal	numeric	0	0	1.0	2.0	0.9	3.0	3.0	4.0	5.0	4.0	0.0	-0.3
AcidIndex	integer	0	0	4.0	7.0	1.3	8.0	7.8	8.0	17.0	13.0	1.6	5.2
Cases2	numeric	0	0	0.0	1.0	1.1	2.0	1.7	2.0	4.0	4.0	-0.4	-0.9



We know that zero purchases also mean a zero number of cases. Nonzero purchases can be split into 1-8 number of cases. The distributions which used to be symmetric, now it is not symmetric because skewness

deviates from 0. Kurtosis is higher after imputing NAs and converting zero signs into positive ones. We can explore the boxplot to see the outliers.

	cor_to_Cases	cor_to_Cases2	cor_to_Stars	cor_to_Purchase
Cases	1	0.98	0.69	0.82
Cases2	0.98	1	0.67	0.83
Purchase	0.82	0.83	0.54	1
STARS	0.69	0.67	1	0.54
LabelAppeal	0.36	0.32	0.26	-0.01
Alcohol	0.06	0.06	0.06	0.01
TotalSulfurDioxide	0.03	0.03	0.02	0.06
FreeSulfurDioxide	0.02	0.02	0.01	0.03
CitricAcid	0.01	0.01	0.01	0.01
ResidualSugar	0	0	0	0.01
pH	-0.01	-0.01	-0.01	-0.03
Chlorides	-0.03	-0.03	-0.01	-0.02
Sulphates	-0.03	-0.03	-0.02	-0.05
Density	-0.04	-0.04	-0.03	-0.02
FixedAcidity	-0.05	-0.06	-0.04	-0.06
VolatileAcidity	-0.07	-0.07	-0.05	-0.06
AcidIndex	-0.25	-0.24	-0.17	-0.27

When we compare this correlation table to the previous one, we can see that correlation numbers improved significantly.

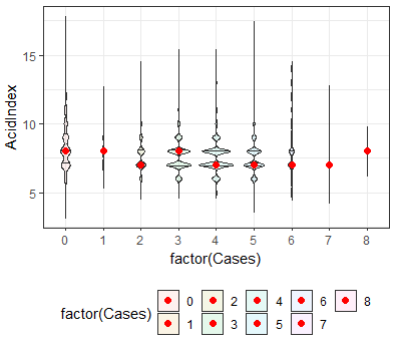
Boxplot/Violin plot

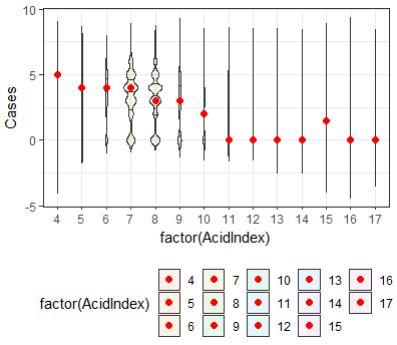
Purchases:

- We can also see that almost all continuous variables have very thin tails, which are outliers. There are more outliers in non-purchased transactions than purchased. There are median and interquartile lines. Since we are trying to compare the median values, redpoint introduced.
- Let's compare the median of each variable of purchased and non-purchased. For example, is the median of FixedAcidity different for the sold wines than non-sold ones? Not really. It seems the median values for purchased and non-purchased are identical. So, FixedAcidity doesn't impact the purchases, and it is not an important predictor. We can see that most of the predictors are not good predictors because the median values are the same except Acid Index, Stars, Cases, and Cases2.

Number of Cases

- We can rerun the same violin plot with a number of cases instead of purchases.
- Now we can see the differences in label appeal. When the LabelAppeal goes up, the average number of cases goes up. So, there is an association between LabelAppeal and the number of cases sold.
- We see the differences. When the STARS rating goes up, the average number of cases goes up. So, there is also an association between STARS and cases.





- We can see that Acid Index should correlate negatively. So, to see it, we need to rerun the violin plot with the x-axis as Acid Index.
- When the acid index is from 5 to 7, the median cases is around 4. When the median number of the acid index goes up, the number of cases goes down. So, there is a relationship between the Acid Index and a number of cases.

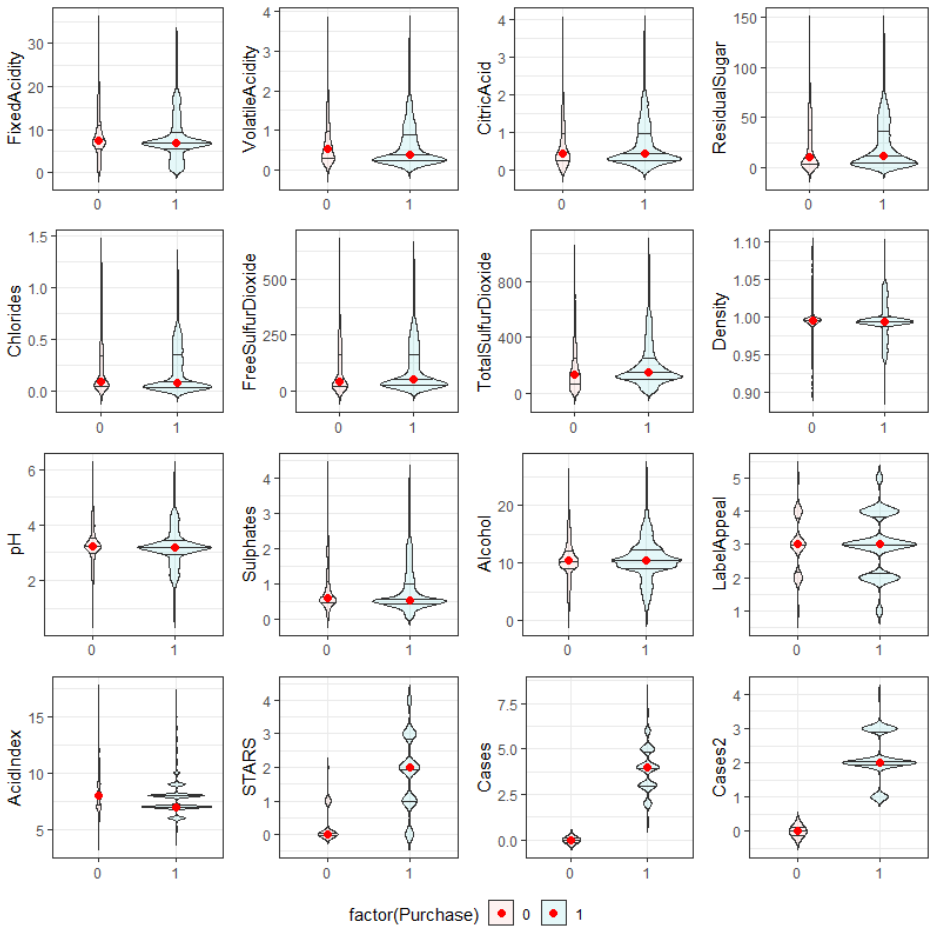
Star Rating

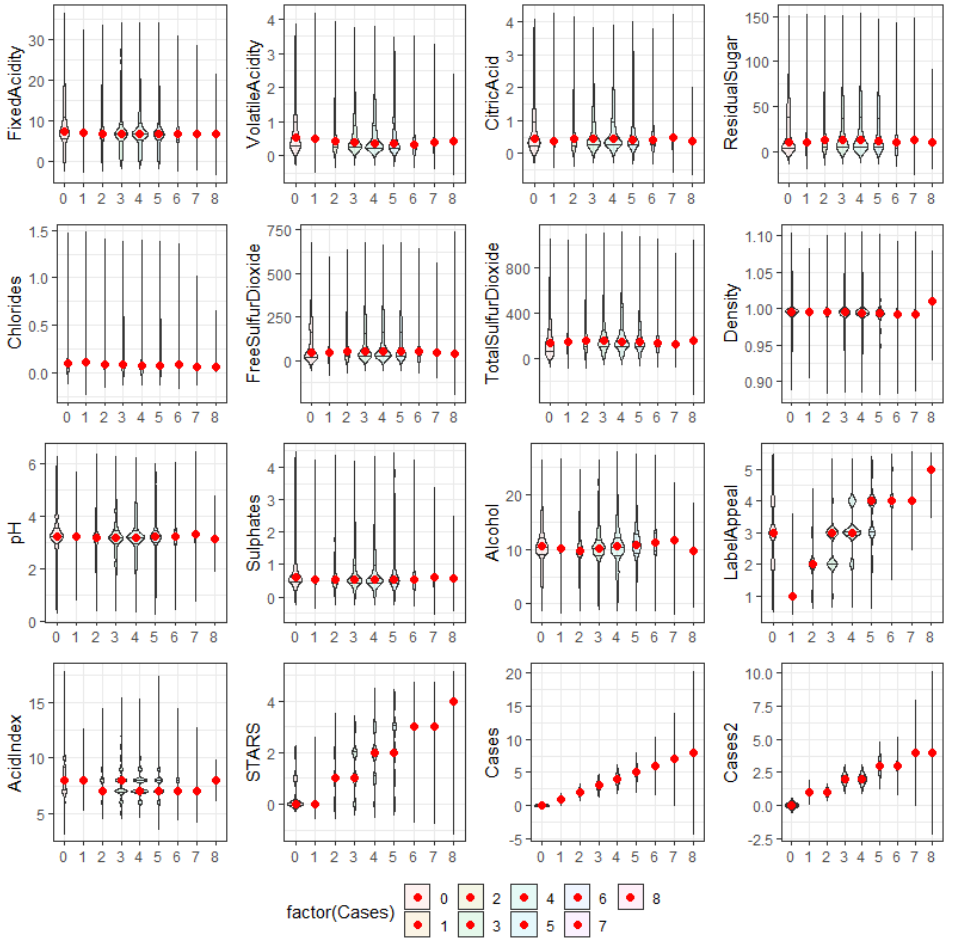
- We can see that Label Appeal and Acid Index are not good predictors for the star rating.

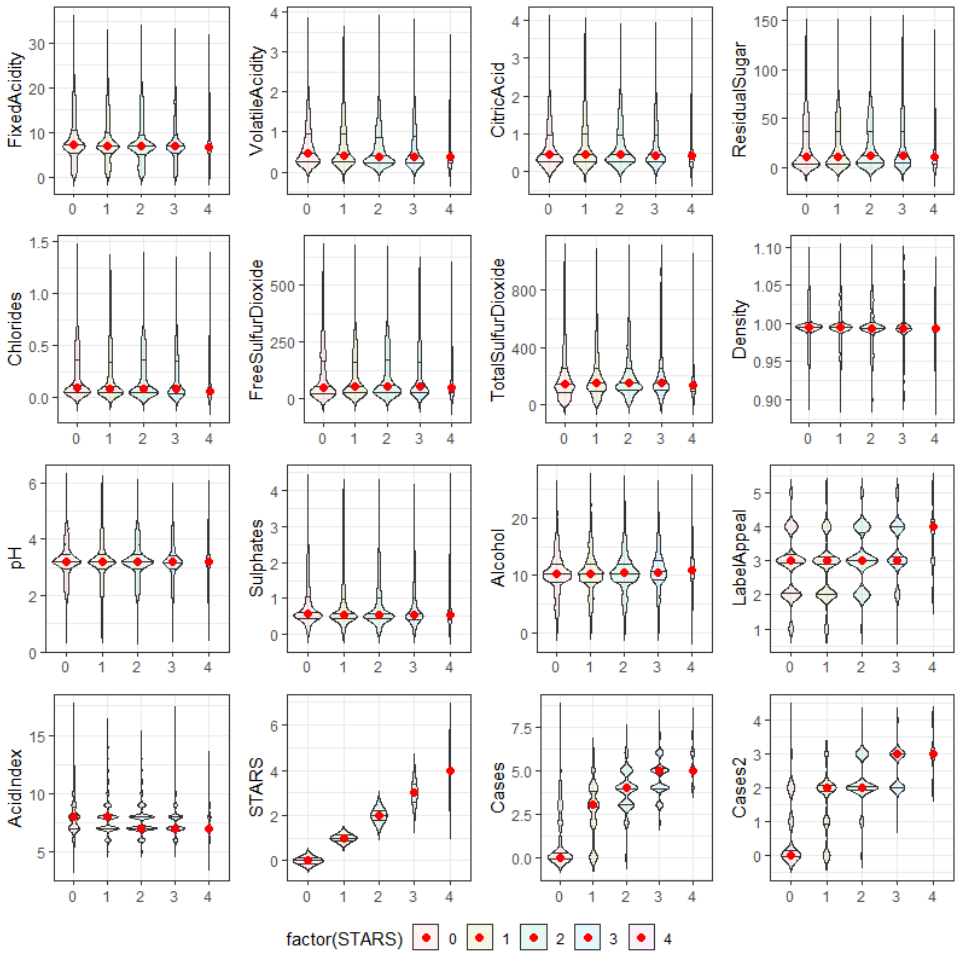
We can choose a number of cases as our target variable because more variables have some association with a number of cases than purchases and star ratings. While we also considered the possibility of transforming one or more of the predictor variables that have skewed distributions, we chose not to apply any such transforms prior to the model building since normal distributions aren't necessarily required for Poisson or ZIP regression modeling.

We can use the following variables to predict the number of cases:

- Purchase
- STARS
- Label Appeal
- Acid Index




FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH
Sulphates
Alcohol
LabelAppeal
AcidIndex
STARS
Cases
Cases2
factor(Cases)
0 1 2 3 4 5 6 7 8


FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH
Sulphates
Alcohol
LabelAppeal
AcidIndex
STARS
Cases
Cases2
factor(STARS)
0 1 2 3 4

Please do NOT treat this as a check list of things you must do to complete the assignment. The EDA is your responsibility. You should have your own thoughts about this step based on your prior experiences this class so far.

Write a description of what you did in performing your EDA and data cleaning. Describe what you did and what you found so that a manager can understand it. Consider that too much detail will cause a manager to lose interest. DO NOT DATA DUMP! If you include a graph, you must describe and discuss that graph! Similarly, too little detail will make the manager consider that you aren't doing your job or that you were not careful. Your reputation is at stake! This description is to be included in your final project write-up.

TASK 2: MODELING

There is not one perfectly correct way to approach model building. You are now charged with the task of producing your best predictive model for the WINE data and the response variable that you chose. Be sure that you are selecting the appropriate method for the variable you chose. This is an open-ended task where you are free to do whatever you wish in modeling this data. You may select the variables manually, use an approach such as Forward or Stepwise, do the variable selection by hand, or try to use an automated procedure. You have enough data, so you should very seriously consider taking a validation approach to this modeling endeavor.

You need to be sure you can interpret your models, have evidence on goodness of fit, and check on assumptions via diagnostics. What criteria are you going to use select your "best" model?

Write of description of the technique you used to decide on your final model. DO NOT DATA DUMP. A manager does not need to see everything that you produce, if though you have to produce the graphs and models to come to your final decision. If you include something in this report like a graph, table, or anything else, you must write about it! If you manually selected a variable for inclusion into the model or exclusion into the model, indicate how this was done.

Write up your final model. Report the model. Discuss the coefficients in the model, do they make sense? Report on goodness of fit and model diagnostics.

As you put together your "Final" prediction model for this project, remember that you can do the above and implement validation modeling like you did for Modeling Assignment #3. You can fit your final model on the Development Data and use the model to make predictions in the Validation dataset. You could fit your model on the Development Data and then refit your model on the Total dataset (development plus validation data) or just the validation dataset. Look to see how much the coefficients change, as a means of checking the stability of the model, as well assess its potential exporting success.
This isn't totally rigorous, but it's a nice check. Not required. Doesn't even have to be reported. Just reminding you, that you know how to do this. Whatever validation work you do and want to report, put the evidence in a table so that comparisons can be easily made and discussed.

Let's test Poisson distribution if the mean and the variance of target variables are the same. The table below shows that all means and variances are not equal and they violate Poisson distribution rule. However, Star grouping the means and variances somewhat closer but not the same. Since cases have zeros (or previously known NA's or missed values), we can explore the Zero Inflated Poisson (ZIP) rule.

	Mean	Variances	Diff	exp(-Mean)	Actual	Diff
Purchase	0.79	0.17	0.62	45.6%	21.4%	24.2%
STARS	1.51	1.41	0.10	22.2%	26.3%	-4.1%
Cases	3.03	3.71	(0.68)	4.8%	21.4%	-16.5%
Cases (regrouped)	1.71	1.14	0.57	18.1%	21.4%	-3.3%

Let's interpret the second half of the table:
Poisson prob (number of cases = 0) = exp(-mean)

- It is predicting 45.6% (probability) of times that we are going to have zero purchases; however, 21.4% of the transactions have got zero purchases

- It is predicting 22.2% (probability) of times that we are going to have zero-star ratings; however, 26.3% of the transactions have got zero-star ratings
- It is predicting 4.8% (probability) of times that we are going to have zero number of cases; however, 21.4% of the transactions have got zero number of cases
- It is predicting 18.1% (probability) of times that we are going to have zero number of regrouped cases; however, 21.4% of the transactions have got zero number of regrouped cases

We realized that we have more zeros than what Poisson would have predicted, and that's a red flag. Poison regression is not a good model to predict; we need to explore the zero-inflated Poisson (zip) model.

After doing the EDA, we chose the number of cases purchased as a target response variable.

Poisson Model

I know Poisson for cases doesn't do good because we checked it earlier. However, let's see what we get when we run the Poisson regression.

```
Call:
glm(formula = Cases ~ LabelAppeal + AcidIndex + STARS + Purchase,
    family = poisson(link = "log"), data = mydataW1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.69999  -0.21617  -0.00006   0.14876   1.79169

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept) -20.842510 294.602248  -0.071             0.943598
LabelAppeal   0.218358   0.006170  35.391 < 0.0000000000000002
AcidIndex    -0.016833   0.004657  -3.614             0.000301
STARS         0.087371   0.004991  17.506 < 0.0000000000000002
Purchase     21.475175 294.602245   0.073             0.941889

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22860.9  on 12794  degrees of freedom
Residual deviance:  1854.4  on 12790  degrees of freedom
AIC: 33806

Number of Fisher Scoring iterations: 18
```

```
Call:
glm(formula = Cases ~ LabelAppeal + AcidIndex + STARS, family
= poisson(link = "log"),
    data = mydataW1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9872  -0.7168   0.0485   0.5527   3.2791

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  0.824618   0.039124   21.08 <0.0000000000000002
LabelAppeal  0.132978   0.006060   21.95 <0.0000000000000002
AcidIndex   -0.088835   0.004462  -19.91 <0.0000000000000002
STARS        0.313946   0.004507   69.65 <0.0000000000000002

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22861  on 12794  degrees of freedom
Residual deviance: 14804  on 12791  degrees of freedom
AIC: 46754

Number of Fisher Scoring iterations: 5
```

- `Y_hat_Poisson_wP = -20.843 + 0.218*LabelAppeal - 0.017*AcidIndex + 0.087*STARS + 21.475*Purchase.` However, the purchase varaible is not significant according to the p-value.
- `Y_hat_Poisson_withoutP = 0.825 + 0.133*LabelAppeal - 0.089*AcidIndex + 0.314*STARS ## excluding purchase variable`
- Goodness Fit:
  - `Res. Dev/df = 1854.4/12790 = 0.145` => it is a good fit with purchase predictor because the value is less than 1
  - `Res. Dev/df = 14804.0/12791 = 1.157` => it is a bad fit without purchase predictor because the value is more than 1
  - If the model goodness fit number is close to one, then this model is a good fit. In this case, it is bigger than 1, which is bad, and it might have of missing predictor, missing data, or overdispersion.
- Chi-square test :
  - `pvalue <- 1 - pchisq((fit1$null.deviance - fit1deviance), (fit1$df.null - fit1$df.residual)) ## with purchase predictor`
  - `pvalue = 0`
  - `pvalue <- 1 - pchisq((fit2$null.deviance - fit2deviance), (fit2$df.null - fit2$df.residual)) ## without purchase predictor`
  - `pvalue = 0`
  - There is no F test when we deal with count data /categorical data then it's all about chi squared value. Since the p-value is 0, the model is significantly better than the intercept. So overall model is significant.
- Perfect model test:

- The p-value is very small for both models with purchased and without purchase. p-value small means that it's not close to the prefect model; the model is significantly worse compared to the perfect model.

- Here is a sample of interpretation without purchased variable:
  - Estimated event rate when all predictors are 0, the median number of cases will be exp(0.825)= 2.281; on average, we got the 2.281 (intercept) cases. Note: it is multiplicative.
  - A one-unit increase in Label Appeal increases the number of cases by exp(0.133) = 1.142. Every time Label Appeal goes up by one unit, we've got 14.2% more number of cases – gaining cases
  - A one-unit increase in Acid Index increases the number of cases by exp(-0.089) = 0.915. Every time Acid Index goes up by one unit, we've got (1 - 0.915) 8.5% less number of cases – losing cases
  - A one-unit increase in star rating increases the number of cases by exp(0.314) = 1.369. Every time star rating goes up by one unit, we've got 36.9% more number of cases – gaining cases.
  - Summary: Wines that have greater label appeal and higher star ratings will likely sell in larger volumes than wines with lower levels of label appeal and lower star ratings. Wines that have higher levels of acidity will likely sell in smaller quantities than wines with relatively lower levels of acidity.
- AIC and BIC numbers are higher (bad) in the model excluding the purchase variable than including it. However, since cases derived from purchases, 0 - no cases sold and 1 - cases sold with a number of cases, we can exclude the purchase variable for future models. However, as we tested earlier, the Poisson model is significantly worse compared to the perfect model. We should explore the Zero Inflated Poisson (ZIP) model.

	RR	2.50%	97.50%
(Intercept)	-	-	-
LabelAppeal	1.244	1.229	1.259
AcidIndex	0.983	0.974	0.992
STARS	1.091	1.081	1.102
Purchase	2,121,046,388.781	Inf	Inf

	RR	2.50%	97.50%
(Intercept)	2.281	2.113	2.463
LabelAppeal	1.142	1.129	1.156
AcidIndex	0.915	0.907	0.923
STARS	1.369	1.357	1.381

ZIP Model

```
Call:
zeroinfl(formula = Cases ~ LabelAppeal + AcidIndex + STARS, data = mydataW1)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-2.1638 -0.4175 -0.0157  0.3836  6.2383

Count model coefficients (poisson with log link):
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  0.559953   0.041368  13.536 < 0.0000000000000002
LabelAppeal  0.233108   0.006298  37.015 < 0.0000000000000002
AcidIndex   -0.020628   0.004830  -4.271              0.0000195
STARS        0.103214   0.005184  19.909 < 0.0000000000000002

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value            Pr(>|z|)
(Intercept) -5.13774    0.24180  -21.25 <0.0000000000000002
LabelAppeal  0.71718    0.04234   16.94 <0.0000000000000002
AcidIndex    0.43713    0.02501   17.48 <0.0000000000000002
STARS       -2.36862    0.05956  -39.77 <0.0000000000000002

Number of iterations in BFGS optimization: 14
Log-likelihood: -2.048e+04 on 8 Df
```

A ZIP model automatically generates two separate models: logistic regression (cases = 0) and Passion (cases > 0). Logistic distribution is used to get the probability of distribution one or two. Poisson is used to modeling the count if the observation comes from distribution two.
After running the ZIP model, we can compare the AIC values of the Poisson model (excluding the purchase variable) and the ZIP model. The Poisson model AIC was 46754 and the ZIP model AIC was 40977. The ZIP model achieved an improved AIC score relative to that of the Poisson model.

Since the ZIP regression model produces two tables, we will try to interoperate each sub model's coefficients separately. The count model coefficients listed above apply when a given wine is likely to accrue the sale of at least one case, while those of the logit model applies when a wine is unlikely to accrue the sale of even a single case.
Count Model Coefficients: Wines that have greater label appeal and higher star ratings will likely sell in larger volumes than wines with lower levels of label appeal and lower star ratings. Wines that have higher levels of acidity will likely sell in lower volumes than wines with relatively lower levels of acidity.
Zero-Inflated Logit Model Coefficients: Wines having higher levels of label appeal and acidity are more likely to fail to accrue the sale of a single case than are wines with a relatively lower level of label appeal and acidity. Wines having higher ratings of the star are less likely to fail to accrue the sale of a single case than are wines with a relatively lower star rating.

Furthermore, after running the ZIP model, the average of actual number (y) and expected numbers (fitted value) are very close to each other: 3.029 vs. 3.028. Before, we had 5% and 21% (which was terrible). Note that the prob(0) is 2.2% of zero cases (Poisson and logistic regression), which is close to the observed 0 cases 2.1%.

```
> head(resdata)
            pr       mu    probof0 y1 fitted_values stars AcidIndex expcount
1 0.0143939099 3.671786 0.03945886  3      3.618935     2         8 3.618935
2 0.0004308247 3.291729 0.03760428  3      3.290311     3         7 3.290311
3 0.0006668674 3.224522 0.04041513  5      3.222372     3         8 3.222372
4 0.0307910310 2.733582 0.09377612  3      2.649412     1         6 2.649412
5 0.0221109041 3.596819 0.04891560  4      3.517290     2         9 3.517290
6 0.8608268856 2.807698 0.86922511  0      0.390756     0        11 0.390756
```

Here are the first 5 rows of predicted values

- pr is the probability that the number of cases equal to zero
- mu is the probability that the number of cases equal to non-zero
- probof0 is a weighted probability; by using the binomial/logistic regression, the number of cases could be either 0 or 1

```
   fitted
y1     0    1    2    3    4    5    6    7
  0  310 1461  480  388   77   14    4    0
  1    1   71  137   34    1    0    0    0
  2    1  194  419  437   38    2    0    0
  3   10  323  391 1228  630   29    0    0
  4    8  214   96  900 1439  491   29    0
  5    6   84   27  215  694  783  172   33
  6    1   27    7   11  127  322  191   79
  7    2    6    0    0    3   36   60   35
  8    0    2    0    0    0    0    4   11
```

- The accuracy of our confusion matrix is 35%, which is very low.
- Let's try to use regrouped cases as the target variable.
  - AIC is 33541 which is better than previous 40977
  - the average of actual number (y) and expected numbers (fitted value) are almost the same: 1.7103 vs. 1.710 4.
  - The accuracy improved from 35% to 54%

	fitted				
y1	0	1	2	3	4
0	669	1645	416	4	0
1	17	986	331	1	0
2	72	1260	4144	312	0
3	36	133	1381	1115	114
4	8	2	6	96	47

TASK 3: CONCLUSIONS AND REFLECTIONS

What conclusions do you draw from having conducted this analysis? What did you learn about the wine world through your modeling endeavor? What actions can you recommend to anyone involved in this field? How did your perspective on modeling change? Discuss anything else you wish to discuss.

I learned how to use Zero Inflated Poisson. The EDA was somewhat challenging because there was less information about the chemical properties of wine data. Since I had less knowledge about the chemical properties, I was not able to derive new features. EDA helped me to see what regression model I can use visually. I learned that the wine predictors should be positive values and how to impute the missing values. In real problems like this, we will deal with cleaning and preparing the data. I learned that star rating, label design appeal, and acid index are the indicators of how well the wine sells and in what quantity. The wine sellers should invest more money in designing and labeling the wine bottles and collect more higher star ratings. They also need to observe carefully on Acidity Index.

EXTRA CREDIT

If you would like to gain extra credit or simply experience the joy of model fitting once again, please feel free to create a predictive model for either or both of the remaining response variables. Most likely, you would not need to do the EDA again, except for the logistic regression model looking for good explanatory variables, so the work would be centered around model fitting. I hope you take advantage of this opportunity. Each variable requires a different type of modeling.

We can explore logistic regression for the purchase target variable. If we recall our EDA boxplot purchase data distribution, we notice that there is a very high likelihood of a wine getting purchased. According to the EDA bar chart, we can see that any wine with over two stars bought 100%.
We can see that AIC number 7791 reduced significant compare to the ZIP regression model (AIC = 33541).
Interpretation: the intercept term is an obvious placeholder value for the rest of the model, and the intercept or baseline probability of exp(3.934) = 51.11%.
The coefficient reflects the negative impact label appeal and acid index have on the purchase decision, which are denoted by 1-exp(-0.45347) =1-0.635 = 0.365 and 1 - exp(-0.39510) = 1-0.674 = 0.326, which are a 36.5% and 32.5% decrease per unit. The stars rating has a positive impact on the purchase decision, which increases the probability of a wine being purchases by exp(2.0448) = 7.39% per level increase in stars rating the wine has.

```
Call:
glm(formula = Purchase ~ LabelAppeal + AcidIndex + STARS, family = "binomial",
    data = mydataW1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.10255   0.03983   0.19471   0.44575   2.45281

Coefficients:
            Estimate Std. Error z value             Pr(>|z|)
(Intercept)  3.93432    0.19207   20.48 <0.0000000000000002
LabelAppeal -0.45347    0.03309  -13.71 <0.0000000000000002
AcidIndex   -0.39510    0.02104  -18.78 <0.0000000000000002
STARS        2.04480    0.04263   47.97 <0.0000000000000002

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13275.8  on 12794  degrees of freedom
Residual deviance:  7782.8  on 12791  degrees of freedom
AIC: 7790.8

Number of Fisher Scoring iterations: 6

> confint(glm1)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept)  3.5599636  4.3129798
LabelAppeal -0.5185951 -0.3888723
AcidIndex   -0.4365906 -0.3541087
STARS        1.9622781  2.1293991
```

The following confusion table gave me 86% accuracy

```
> table(round(glm1$fitted), mydataW1[,"Purchase"])

      0    1
  0 1802  929
  1  932 9132
```

CONGRATULATIONS! You are done with Modeling Assignment 4. You are home free! Well done!