Data 621: Predicting Wine Sales

Aaron Grzasko

City University of New York

Abstract

In this report, we explore the relationship between wine sales and the unique characteristics of wine, including chemical composition, expert ratings, and label attractiveness. This relationship is examined using a variety of regression models, including Poisson, Negative Binomial, and Multiple Linear regression types. We conclude the report by listing important drivers of wine sales.

Keywords: Wine, Regression, Poisson, Negative Binomial

Data 621: Predicting Wine Sales

Introduction

A large wine manufacturer (LWM) is considering revisions to its current lineup of product offerings to improve total sales revenue. To aid in the decision-making process, LWM acquired a data set that identifies physical characteristics and ratings for more than 12,000 commercial wines. The data set set also quantifies the number of wine cases purchased by distributors after sampling each product.

LWM has contracted with the data analytics team (DAT) to build a model that that predicts the number of wine cases sold to distributors, given a set of unique wine characteristics. Because distributor purchases are strongly correlated with sales at restaurants and other retail outlets–where LWM derives the bulk of its revenue–the model output could help LWM determine wine qualities that are associated with strong sales.

Data Exploration

Overview

The training data set comprises 12,795 observations and 16 variables and is approximately 1.4 MB in size. The variable `INDEX` is used for observation identification purposes only. The other variables include 14 features and 1 response variable, `TARGET`.

Twelve of the features describe the chemical properties of wine. The remaining two predictors are rating variables: `LabelAppeal` refers to the perceived attractiveness of a wine's product label, while `STARS` is a subjective assessment of wine quality.

Finally, the output variable, `TARGET`, is a count measure indicating the number of wine case purchases by distributors.

For a detailed listing of all variable names and descriptions, please refer to Table 1, located at the end of this report. Subsequent tables and figures will also provided after the body of this document.

Categorical Features

There are three categorical features in the wine data set: LabelAppeal, STARS, and AcidIndex. The first two of these features are rating variables–see the Overview section and 1 for detailed descriptions. AcidIndex is related to the acidic quality of wine.

LabelAppeal scores range from -2 through 2. Based on the data set documentation, a negative score suggests a poor impression of the label design, a zero score is neutral, and a positive scores implies an attractive design. The distribution of scores in the training data are symmetrical and roughly bell-shaped.

STARS scores range from 1 through 4, with experts characterizing more wines as low quality (1-2) than high quality (3-4). A significant number of wines in the training data–over 25%–have no STARS score.

AcidIndex is a discrete variable with observed values ranging from 4 through 17. The data documentation describes the calculation of the index value using a proprietary method involving weighted averages. The values are clearly ordinal, but we have little additional information for understanding the meaning of this variable. For example, we cannot determine, based on the description provided, whether an index value of 10 reflects a wine with twice the acidic content of another wine with a value of 5. Unlike the other variables in the unadjusted data set, AcidIndex has a moderate, right skew.

Refer to Table 2 for a numerical summary of each categorical variable. Table 3 provides additional, descriptive statistics, and Figure 1 contains graphical summaries.

Discrete Features

We made a judgment call in classifying AcidIndex as a categorical variable–based on the limited information provided in the variable documentation. The variable is clearly discrete, with only 14 unique values in the training data. Another analyst may reasonably argue that this feature should be treated like a continuous variable for modeling purposes.

There are two additional, discrete predictors in the wine data set: `TotalSulfurDioxide` and `FreeSulfurDioxide`. Both variables measure similar chemical properties. Plots in Figure 1 indicate that the variables are symmetric, but have rather fat tails. We are concerned about the negative values exhibited by these variables: Sulfur Dioxide is typically measured in mg/l (or equivalently, ppm). We also note that both variables have missing values–see 3 for more details. We'll address these issues in subsequent sections.

For modeling purposes, we will treat the two sulfur dioxide features like any other continuous predictor. This treatment seems reasonable, given that both predictors take on a very wide array of discrete values.

Continuous Features

Nine predictors in the wine data set are continuous, with each variable corresponding to a particular chemical property: `FixedAcidity`, `VolatileAcidity`, `CitricAcid`, `ResidualSugar`, `Chlorides`, `Density`, `pH`, `Sulphates`, and `Alcohol`.

The three predictors related to acidity and the `Density` variable have a significant number of outlier observations–roughly 2% of each feature's observations are classified as outliers.

All nine predictors are roughly symmetrical and leptokurtic. Finally, all continuous features–with the exception of `pH` and `Density`–exhibit negative values. Based on our cursory research, none of these predictors are measured in units that would include negative values. We will address these issues in a later section.

Descriptive statistics and graphical summaries for the continuous features can be found in Table 3 and Figure 1, respectively.

Response Variable

The output variable `TARGET`, has a zero-inflated distribution, with more than 20% of observations indicating no sales. Counts in the training data range from 0 through 8, although the values are theoretically unbounded above. Finally, the distribution of cases sold, given at least one sale, is symmetric and approximately normal.

Refer to Table 3 and Figure 1 for additional details.

Relationship Between Features and Response

In Figure 2 we compare the three categorical variables to the `TARGET` variable. For each of these predictors, there appears to be a significant relationship between the ordered levels and the number of wine cases sold. Unit increases to the `LabelAppeal` metric are associated with consistent increases in the response. Similarly, increases in the `STARS` rating correspond with higher wine sales. Interestingly enough, wines that have not been rated tend to have low sales compared to their rated counterparts. Finally, `AcidIndex` score appear to be negatively correlated with the response.

The relationship between the `TARGET` variable and the continuous variables is less clear. We initially created scatterplots to explore these relationship. Unfortunately, there is significant variability in response values across narrow bands of similar predictor values; so patterns are not easily discerned. As an alternative we plotted Loess curves to get a general sense of the relationship between the predictors and the response–see Figure 3. The plotted curves are fairly complex but we noted several mild or moderate patterns:

- Higher Alcohol content is associated with higher sales.
- The concentration of Chlorides appears to be negatively associated with sales.
- Higher acidity tends to be associated with lower sales.
- Sulfur Dioxide concentration appears to have a positive association with sales.

Correlation Between Variables

In Figure 4, we produced a correlation matrix of all candidate variables. For computational purposes, we converted all categorical variables back to their original numerical form.

None of the predictors have a particularly strong correlation with the TARGET variable. The STARS predictor has a moderate positive association with response. LabelAppeal has a moderately weak positive correlation with TARGET, while AcidIndex has a fairly weak negative association with the response. All other predictors have extremely weak linear associations with TARGET.

Finally, the correlations between predictor variables are all weak. In other words, there there are no immediately obvious collinearity issues.

Data Preparation

Negative Values

As discussed previously, none of the predictors related to chemical composition should have negative values, as the units of measurement for these variables are always non-negative.

Based on the descriptive statistics and histogram plots in Figure 1 and Table 3, respectively, we see a similarity in the magnitudes of the negative and positive ranges for each of thee predictors of concern. In other words, the distributions are all symmetric.

It is possible that the minus signs in the data reflect data entry errors or some other identifying characteristic, but we have no data documentation to support these hypotheses. On the other hand, we found a resource online–see the References section–that analyzes the distributions of various wine chemical properties. In that resource, most of the wine distributions featured all positive values distributions and were mostly right skewed.

Based on this information, we decided to apply simple absolute value transformation for the predictors of concern. In total, we applied this transformation to nine predictors: FixedAcidity, VolatileACidity, CitricAcid, ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, Sulphate, and Alcohol.

Missing Values

Missing values impact roughly 50% of records in the training data. The values are present in varying degrees across eight of the predictor candidates: `STARS`, `Sulphates`, `TotalSulfurDioxide`, `Alcohol`, `FreeSulfurDioxide`, `Chlorides`, `ResidualSugar`, and `pH`. For a summary description of missing elements, refer to Figure 5.

The `STARS` variable has the highest number of missing values, with roughly one quarter of observations coded as NA. Given the large number of missing records, and the apparent significance of the NA records relationship with `TARGET` vis-a-vis the other variable levels, we decided to leave the NA records as is.

The `Sulphates` variable has missing values in approximately 10% of observations, while the six other predictors have missing values in the 3% - 5% range. Given the relatively small percentages of missing values for the remaining predictors, we will impute values using R's `MICE` package. The use of the `MICE` procedure results in less bias in model regression coefficients compared to simpler methods like mean substitution.

Refer to Figure 6 for a graphical depiction of the `MICE` imputation procedure. The blue curves represent the distribution of non-missing values, while the red curves are the imputed values. The red and blue distributions are very similar. This result suggests that the imputation procedure approximated the original distribution accurately.

Revised Descriptive Statistics and Graphical Summaries

We have made substantial revisions to our predictor variables; so we will review the revised variables based on the transformations applied thus far.

In Table 4, we recalculated basic descriptive statistics for all variables. We also generated new histograms, correlation plots, and loess curve plots–see Figures 7, 8, and 9, respectively.

We see that many of the predictor variables now have a moderate right skew. The majority of the predictors are still leptokurtic. Also, our transformed variables indicate a higher prevalence of outlier observations. The revised Loess curve plots indicate some significant changes in the relationship between the predictors and the `TARGET` variable. For instance, sulfur dioxide content now seems to be negatively correlated with wine sales. Interestingly, the loess curves are almost linear or slightly curved for typical predictor values. However, the curves are very complex and unstable for predictor values that are are sparsely populated. Finally, the revised correlation plot indicates no material changes to the original correlations.

Bin Acid Index Predictor

We previously decided to treat `AcidIndex` as an ordered, categorical variable. Because the index values are sparsely populated at the extreme low and high ends, we will bin the categories into fewer groupings. Specifically, we will merge values of 4 and 5 into the 6 index measure. We will also merge indices 11 and higher into the original 10 index value. This decision is also supported by the boxplot in Figure 2, where extreme low and high index values appear to have similar relationship with the `TARGET` variable.

Other Transformations Considered

We contemplated applying Box-Cox suggested power transformations to some of the predictor variables. Because many of the predictor values have a moderate right skew, we could reasonably apply log, square root, and/or quarter-root transformations. These transformations may make sense in a basic OLS model; however, we are building a variety of glm models in addition to OLS models. Most glms do not require such transformations. We ultimately decided to refrain from applying any power transformations at this point in the modeling process.

We also debated whether or not to winsorize some of the candidate predictors, given the prevalence of outlier observations. While this procedure can soften the influence of outlier observations, it can also introduce bias in the point estimates and inference mechanisms. We therefore decided to forgo winsorizing the predictors.

Finally, we noted that our client, LWM, requires an interpretable model. By refraining from applying complex predictor transformations, our model should remain relatively easy to explain.

Build Models

Multiple Linear Regression Model 1

For our first linear regression model, we decided to include the following four predictors: `AcidIndex`, `STARS`, `LabelAppeal`, and `Alcohol`. We included each of the categorical predictors because of their clear relationship with `TARGET`–see the boxplots in Figure 2. Also the correlation plot (Figure 8) indicated higher magnitude correlation values compared to the other candidate predictors.

We included the `Alcohol` variable for a variety of reasons. First, we recognize that our variable transformation involving the absolute value may not have been the best approach to handling negative values. Fortunately, the raw `Alcohol` values were all positive. Secondly, this variable is also not significantly skewed and has only mildly fat tails. We also observe fewer outliers with this variable compared to many other predictors. Finally, `Alochol's` relationship with `TARGET`–refer to Figure 9–appears to be more straightforward compared to other relationships.

See Table 5 for regression output. The model has an adjusted R^2 value of 0.54. All predictors and levels within each categorical predictor are significant at the 5% level. The model coefficients make intuitive sense. Sales of wine decrease with each increase in the `AcidIndex` value. Having a `STARS` rating results in more sales compared to no ratings, and higher `STARS` ratings are associated with higher sales. Finally, increases in alcohol content are associated with higher sales. This is consistent with observations from our EDA work.

Multiple Linear Regression Model 2

For our second multiple linear regression model, we we used stepwise regression with variable selection occurring in both directions. Model output can be found in Table 6. This second model has an adjusted R^2 of 0.54, and has 11 out of the 15 potential candidate predictors. Only `FixedAcidity`, `ResidualSugar`, and `FreeSulfureDioxide` excluded from the model.

Three predictors have p-values over 5%: `pH`, `CitricAcid`, and `Sulphates`. Let's create a reduced model that removes these variables–see Table 7 for regression output for this smaller model. The adjusted R^2 is still 0.54, but now all model p-values are below 5%.

Let's perform an ANOVA test to see if the full model is significantly different than the reduced model–refer to Table 8.

The p-value of the ANOVA test is greater than 5%, We proceed with the reduced Model 2.1.

All coefficients included in Model 2 that were also in Model 1 have similar signs and magnitudes as those in the prior model. Furthermore, the signs of the four additional predictors are consistent with results from our EDA. For instance, we expected the number of wine cases sold to decrease with increases in `VolatileAcidity`. We also expected some improvement in `TARGET` output with increases to `TotalSulfurDioxide`. Finally, our EDA–see Figure 9–indicated a general decrease in `TARGET` values with increases with `Chloride` and `Density`. This is consistent with the negative model coefficients ovserved in Model 2.1.

Poission Regression Model 3

A Poisson model is potentially more appropriate for our modeling problem because our output variable is count-based measure.

In a Poisson-based model, we assume the dependent variable's mean is identical to its variance. Let's briefly check this assumption by comparing the unconditional mean and variance of TARGET variable. The variable's mean is 3.03 and the variance is 3.71. Because the variance is somewhat higher than the mean, there could possibly be an issue with overdispersion.

Let's start by building a Poisson glm using the same variables from Model 1. Model output is provided in Table 9. The AIC value is 45604. All coefficients are statistically significant, with signs in and magnitudes that are consistent with our understanding of the the predictor relationship with TARGET.

Poisson Regression Model 4

We will now fit a Poisson Regression model using stepwise regression. Model output can be found in Table 10. The coefficients in this model are identical to those in our reduced Model 2 (i.e. Model 2.1). The coefficient signs and magnitudes are consistent with our understanding from our earlier EDA work. However, we note that three variables in this variable have p-values above 5%: TotalSulfurDioxide, Chlorides, and Density.

We'll build a reduced model with these three variables removed. Output can be found in Table 11.

We'll now perform an analysis of deviance test to compare the full and reduced models-see Table 12. The p-value for this test roughly 3%, so we will proceed with the full model.

Zero-Inflated Poisson Regression Model 5

We noted earlier that the TARGET variable is zero-inflated. Therefore, it makes sense to attempt to fit a zero-inflated Poisson regression model.

We'll use the same set of predictors from Model 2.1 and 4. Model output is located in Table 13.

While the coefficients and signs of our variables are consistent with our understanding of these variables, we see four variables in the Poisson with log link component of the model with p-values greater than 5%.

Zero-Inflated Poisson Regression Model 6

Model 6 simply removes the insignificant predictors from model 5. Output can be found in Table 15.

All coefficients in the resulting model–with the exception of one dummy variable relating to `AcidIndex`–are significant.

Negative Binomial Model 7

In model 7, we build a negative binomial regression model. This model type seems appropriate given that the data is possibly overdispersed.

We'll begin by using the small subset of predictors using in Model 1. Output is located in Table 17.

All coefficients are significant. The values of the coefficients are consistent with previous models examined.

Negative Binomial Model 8

In the next model, we fit a negative binomial model using stepwise regression. Model output is in Table 18. All categorical predictors are included in the final model, as well as `VolatileAcidity`, `Alcohol`, `TotalSulfurDioxide`, `Chlorides`. and `Density`. Of these predictors, `TotalSulfurDioxide`, `Chlorides`. and `Density` are not significant at the 5% threshold.

We'll produce a reduced model with the three predictors removed. Output can be found in Table 19.

Now we'll do an analysis of deviance test to compare the full and reduced models–see Table 20 for details.

We proceed the full model, given the p-value of 3%.

Zero Inflated Negative Binomial Model 9

Next, we build a zero-inflated negative binomial model. We'll start with the set of candidate predictors used in model 8–refer to Table 21 for detailed output.

All variables are significant except `VolatileAcidity`, `TotalSulfurDioxed`, `Chlorides`, and `Density`.

Zero-inflated Negative Binomial Model 10

For our last model, we'll start with Model 9, but remove the predictors that were not statistically significant. Full model output is in Table 23.

All predictors are statistically significant in this model.

Select Models

Model Evaluations

We have 10 models in total to compare. We need a common set of statistics that we can use to compare each model. We are able to calculate BIC for 6 out of the 10 models, and AIC for all 10 models. Refer to Table 25 for a full model comparison summary.

Using AIC, we see that the multiple linear regression models (Model 1 and 2.1) outperform the basic Poisson (Models 3 and 4) and Negative Binomial Models (Models 7 and 8).

This result is somewhat surprising, as we suspected that the negative binomial models would be superior to the other model types. Interestingly, the basic Poisson and Negative Binomial performed about the same.

On the other hand, the zero inflated Poisson (Models 5 and 6) and Zero inflated Negative Binomial models performed the best. The models where we included in the statistically insignificant predictors (Models 5 and 9) performed the best.

While Models 5 and 9 have virtually identical AIC measures, we select Model 9, the zero inflated Negative Binomial model, as our model to use going forward, due to possible overdispersion issues with the data.

Make Predictions

We applied the same imputation and transformation procedures used for the training data on our evaluation data set.

Next, we predicted `TARGET` values using Model 9, one of our zero-inflated Negative Binomial models.

Finally we uploaded the prediction file to Github:

https://raw.githubusercontent.com/spitakiss/Data621/master/Homework5/test_data_with_predictions.csv

Compare Training TARGET and Predicted Values

As a final check, we'll compare the distributions of the training data `TARGET` variable and the predicted response values in the test data set–refer to Table 26 for the summary table.

The distributions are fairly similar. The means and medians are almost identical, and the kurtosis values are close. The standard deviation and range of the predicted `TARGET` values are slightly lower than the comparable statistics in the training data.

Conclusion

We produced an interpretable model that LWM can use for potentially modifying its current wine offerings.

Based on our modeling process, we've determined that high expert ratings and attractive label design are very important factors that are routinely associated with high wine sales. Wines that are less acidic tend to sell better, and higher alcohol content is associated above average sales.

Additional attributes that are common in high-selling wines include low concentrations of volatile acidic content, chlorides, and total sulfur dioxide. Finally, low density wines tend to sell better than high density wines.

References

We used R (Version 3.4.4; R Core Team, 2018) and the R-packages *apaTables* (Version 2.0.2; Stanley, 2018), *bindrcpp* (Version 0.2; Müller, 2017), *BiocInstaller* (Version 1.26.1; Tenenbaum & Team, 2017), *car* (Version 2.1.5; Fox & Weisberg, 2011), *DataExplorer* (Version 0.5.0; Cui, 2018), *devtools* (Version 1.13.3; Wickham & Chang, 2017), *dplyr* (Version 0.7.4; Wickham, Francois, Henry, & Müller, 2017), *faraway* (Version 1.0.7; Faraway, 2016), *geoR* (Version 1.7.5.2; Ribeiro Jr & Diggle, 2016), *GGally* (Version 1.3.2; Schloerke et al., 2017), *ggplot2* (Version 2.2.1; Wickham, 2009), *ggthemes* (Version 3.4.0; Arnold, 2017), *gridExtra* (Version 2.3; Auguie, 2017), *kableExtra* (Version 0.5.2; Zhu, 2017), *knitr* (Version 1.20; Xie, 2015), *lattice* (Version 0.20.35; Sarkar, 2008), *MASS* (Version 7.3.47; Venables & Ripley, 2002), *memisc* (Version 0.99.14.9; Elff, 2017), *mice* (Version 2.46.0; van Buuren & Groothuis-Oudshoorn, 2011), *moments* (Version 0.14; Komsta & Novomestky, 2015), *pacman* (Version 0.4.6; Rinker & Kurkiewicz, 2017), *papaja* (Version 0.1.0.9709; Aust & Barth, 2018), *pscl* (Version 1.5.2; Zeileis, Kleiber, & Jackman, 2008), *sme* (Version 1.0.2; Berk, 2018), *stargazer* (Version 5.2.1; Hlavac, 2018), *stringi* (Version 1.1.7; Gagolewski, 2018), *tidyr* (Version 0.7.1; Wickham & Henry, 2017), *tinytex* (Version 0.5; Xie, 2018), and *xtable* (Version 1.8.2; Dahl, 2016) for all our analyses.

Arnold, J. B. (2017). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=ggthemes

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics.* Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Berk, M. (2018). *Sme: Smoothing-splines mixed-effects models.* Retrieved from https://CRAN.R-project.org/package=sme

Cui, B. (2018). *DataExplorer: Data explorer.* Retrieved from https://CRAN.R-project.org/package=DataExplorer

Dahl, D. B. (2016). *Xtable: Export tables to latex or html*. Retrieved from https://CRAN.R-project.org/package=xtable

Elff, M. (2017). *Memisc: Management of survey data and presentation of analysis results*. Retrieved from https://CRAN.R-project.org/package=memisc

Faraway, J. (2016). *Faraway: Functions and datasets for books by julian faraway*. Retrieved from https://CRAN.R-project.org/package=faraway

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second.). Thousand Oaks CA: Sage. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Gagolewski, M. (2018). *R package stringi: Character string processing facilities*. Retrieved from http://www.gagolewski.com/software/stringi/

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). Retrieved from https://CRAN.R-project.org/package=stargazer

Komsta, L., & Novomestky, F. (2015). *Moments: Moments, cumulants, skewness, kurtosis and related tests*. Retrieved from https://CRAN.R-project.org/package=moments

Müller, K. (2017). *Bindrcpp: An 'rcpp' interface to active bindings*. Retrieved from https://CRAN.R-project.org/package=bindrcpp

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ribeiro Jr, P. J., & Diggle, P. J. (2016). *GeoR: Analysis of geostatistical data*. Retrieved from https://CRAN.R-project.org/package=geoR

Rinker, T. W., & Kurkiewicz, D. (2017). *pacman: Package management for R*. Buffalo, New York: University at Buffalo/SUNY. Retrieved from http://github.com/trinker/pacman

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., … Larmarange, J. (2017). *GGally: Extension to 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=GGally

Stanley, D. (2018). *ApaTables: Create american psychological association (apa) style tables*. Retrieved from https://CRAN.R-project.org/package=apaTables

Tenenbaum, D., & Team, B. (2017). *BiocInstaller: Install/update bioconductor, cran, and github packages*.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from http://www.jstatsoft.org/v45/i03/

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., & Chang, W. (2017). *Devtools: Tools to make developing r packages easier*. Retrieved from https://CRAN.R-project.org/package=devtools

Wickham, H., & Henry, L. (2017). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/

Xie, Y. (2018). *Tinytex: Helper functions to install and maintain 'tex live', and compile* *'latex' documents*. Retrieved from https://CRAN.R-project.org/package=tinytex

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8). Retrieved from http://www.jstatsoft.org/v27/i08/

Zhu, H. (2017). *KableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

Red Wine Analysis: https://rpubs.com/billa/redwine

Table 1

Variable Names and Descriptions in Wine Data Set

Variable Name	Definition	Theoretical Effect
INDEX	Identification only	
TARGET	Number of Wine Cases Purchased	
AcidIndex	Proprietary method of testing total acidity	
Alcohol	Alcohol Content	
Chlorides	Chloride Content	
Citric Acid	Citric Acid Content	
Density	Density of wine	
FixedAcidity	Fixed acidity of wine	
FreeSulfurDioxide	Sulfur Dioxide content of wine	
LabelAppeal	Marketing score of label attractiveness	High value suggests high sales
ResidualSugar	Residual Sugar of wine	
STARS	Expert Rating: 1=Poor through 4=Excellent	High value implies high sales
Sulphates	Sulfate content	
TotalSulfurDioxide	Total Sulfur Dioxide	
VolatileAcidity	Volatile acid content of wine	
pH	pH of wine	

Table 2

Categorical Features Summary

Variable	Levels	Count	Pct	Cumul.Pct
LabelAppeal	-2	504	3.9	3.9
	-1	3136	24.5	28.4
	0	5617	43.9	72.3
	1	3048	23.8	96.2
	2	490	3.8	100
STARS	1	3042	23.8	23.8
	2	3570	27.9	51.7
	3	2212	17.3	69
	4	612	4.8	73.7
		3359	26.3	100
AcidIndex	4	3	0	0
	5	75	0.6	0.6
	6	1197	9.4	10
	7	4878	38.1	48.1
	8	4142	32.4	80.5
	9	1427	11.2	91.6
	10	551	4.3	95.9
	11	258	2	97.9
	12	128	1	98.9
	13	69	0.5	99.5
	14	47	0.4	99.8
	15	8	0.1	99.9
	16	5	0	99.9
	17	7	0.1	100

Table 3

Descriptive Statistics Summary

	Min	Max	Range	Mean	Med	Stdev	Skew	Kurt	NAs	Outliers
TARGET	0.0	8.0	8.0	3.0	3.0	1.9	-0.3	2.1	0	0
FixedAcidity	-18.1	34.4	52.5	7.1	6.9	6.3	0.0	4.7	0	195
VolatileAcidity	-2.8	3.7	6.5	0.3	0.3	0.8	0.0	4.8	0	216
CitricAcid	-3.2	3.9	7.1	0.3	0.3	0.9	0.0	4.8	0	215
ResidualSugar	-127.8	141.2	268.9	5.4	3.9	33.7	-0.1	4.9	616	0
Chlorides	-1.2	1.4	2.5	0.1	0.0	0.3	0.0	4.8	638	0
FreeSulfurDioxide	-555.0	623.0	1,178.0	30.8	30.0	148.7	0.0	4.8	647	0
TotalSulfurDioxide	-823.0	1,057.0	1,880.0	120.7	123.0	231.9	0.0	4.7	682	0
Density	0.9	1.1	0.2	1.0	1.0	0.0	0.0	4.9	0	226
pH	0.5	6.1	5.7	3.2	3.2	0.7	0.0	4.6	395	0
Sulphates	-3.1	4.2	7.4	0.5	0.5	0.9	0.0	4.8	1,210	0
Alcohol	-4.7	26.5	31.2	10.5	10.4	3.7	0.0	4.5	653	0
LabelAppeal	-2.0	2.0	4.0	0.0	0.0	0.9	0.0	2.7	0	0
AcidIndex	4.0	17.0	13.0	7.8	8.0	1.3	1.6	8.2	0	264
STARS	1.0	4.0	3.0	2.0	2.0	0.9	0.4	2.3	3,359	0

Note. Outliers are defined as observations that are more than 3 standard deviations from the mean

Table 4

Revised Descriptive Statistics Summary

	Min	Max	Range	Mean	Med	Stdev	Skew	Kurt	NAs	Outliers
TARGET	0.0	8.0	8.0	3.0	3.0	1.9	-0.3	2.1	0	0
FixedAcidity	0.0	34.4	34.4	8.1	7.0	5.0	1.2	5.0	0	156
VolatileAcidity	0.0	3.7	3.7	0.6	0.4	0.6	1.7	6.1	0	239
CitricAcid	0.0	3.9	3.9	0.7	0.4	0.6	1.6	5.9	0	230
ResidualSugar	0.0	141.2	141.2	23.3	12.9	25.0	1.5	5.3	0	242
Chlorides	0.0	1.4	1.4	0.2	0.1	0.2	1.5	5.2	0	239
FreeSulfurDioxide	0.0	623.0	623.0	106.4	56.0	107.7	1.5	5.4	0	237
TotalSulfurDioxide	0.0	1,057.0	1,057.0	204.2	154.0	163.1	1.6	6.1	0	209
Density	0.9	1.1	0.2	1.0	1.0	0.0	0.0	4.9	0	226
pH	0.5	6.1	5.7	3.2	3.2	0.7	0.0	4.7	0	195
Sulphates	0.0	4.2	4.2	0.8	0.6	0.7	1.7	6.2	0	208
Alcohol	0.0	26.5	26.5	10.5	10.4	3.6	0.2	4.1	0	90
LabelAppeal	-2.0	2.0	4.0	0.0	0.0	0.9	0.0	2.7	0	0
AcidIndex	4.0	17.0	13.0	7.8	8.0	1.3	1.6	8.2	0	264
STARS	1.0	4.0	3.0	2.0	2.0	0.9	0.4	2.3	3,359	0

Note. Outliers are defined as observations that are more than 3 standard deviations from the mean

Table 5

Model 1 Multiple Regression Output

Predictor	b	95% CI	$t(12781)$	p
Intercept	0.61	$[0.46$, $0.76]$	7.75	< .001
AcidIndex7	-0.09	$[-0.17$, $-0.01]$	-2.12	.034
AcidIndex8	-0.20	$[-0.28$, $-0.11]$	-4.67	< .001
AcidIndex9	-0.51	$[-0.61$, $-0.41]$	-10.02	< .001
AcidIndex10	-1.06	$[-1.17$, $-0.95]$	-19.26	< .001
STARS1	1.36	$[1.30$, $1.43]$	41.28	< .001
STARS2	2.41	$[2.34$, $2.47]$	75.06	< .001
STARS3	2.97	$[2.90$, $3.04]$	79.93	< .001
STARS4	3.66	$[3.54$, $3.78]$	61.67	< .001
LabelAppeal-1	0.36	$[0.24$, $0.49]$	5.76	< .001
LabelAppeal0	0.83	$[0.71$, $0.95]$	13.50	< .001
LabelAppeal1	1.29	$[1.17$, $1.42]$	20.16	< .001
LabelAppeal2	1.88	$[1.71$, $2.05]$	22.26	< .001
Alcohol	0.01	$[0.01$, $0.02]$	4.31	< .001

Table 6

Model 2 Multiple Regression Output

Predictor	*b*	95% CI	*t*(12774)	*p*
Intercept	1.67	$[0.80$, $2.54]$	3.77	< .001
STARS1	1.36	$[1.29$, $1.42]$	41.14	< .001
STARS2	2.39	$[2.33$, $2.46]$	74.70	< .001
STARS3	2.96	$[2.88$, $3.03]$	79.65	< .001
STARS4	3.65	$[3.53$, $3.77]$	61.58	< .001
LabelAppeal-1	0.36	$[0.23$, $0.48]$	5.67	< .001
LabelAppeal0	0.82	$[0.70$, $0.94]$	13.44	< .001
LabelAppeal1	1.29	$[1.16$, $1.41]$	20.09	< .001
LabelAppeal2	1.88	$[1.71$, $2.04]$	22.24	< .001
AcidIndex7	-0.09	$[-0.17$, $-0.01]$	-2.25	.024
AcidIndex8	-0.20	$[-0.28$, $-0.12]$	-4.80	< .001
AcidIndex9	-0.51	$[-0.60$, $-0.41]$	-9.98	< .001
AcidIndex10	-1.05	$[-1.16$, $-0.94]$	-19.07	< .001
VolatileAcidity	-0.11	$[-0.15$, $-0.07]$	-5.25	< .001
Alcohol	0.01	$[0.01$, $0.02]$	4.44	< .001
TotalSulfurDioxide	0.00	$[0.00$, $0.00]$	2.74	.006
Density	-0.91	$[-1.77$, $-0.06]$	-2.10	.036
Chlorides	-0.10	$[-0.20$, $0.00]$	-2.00	.046
PH	-0.03	$[-0.06$, $0.01]$	-1.64	.101
CitricAcid	0.03	$[-0.01$, $0.07]$	1.49	.137
Sulphates	-0.03	$[-0.06$, $0.01]$	-1.43	.152

Table 7

Model 2.1 Multiple Regression Output

Predictor	*b*	95% CI	*t*(12777)	*p*
Intercept	1.60	$[0.73$, $2.46]$	3.63	< .001
STARS1	1.36	$[1.29$, $1.42]$	41.23	< .001
STARS2	2.40	$[2.33$, $2.46]$	74.78	< .001
STARS3	2.96	$[2.89$, $3.03]$	79.75	< .001
STARS4	3.65	$[3.54$, $3.77]$	61.62	< .001
LabelAppeal-1	0.36	$[0.24$, $0.48]$	5.70	< .001
LabelAppeal0	0.83	$[0.71$, $0.95]$	13.46	< .001
LabelAppeal1	1.29	$[1.16$, $1.41]$	20.12	< .001
LabelAppeal2	1.88	$[1.71$, $2.04]$	22.25	< .001
AcidIndex7	-0.09	$[-0.17$, $-0.01]$	-2.18	.029
AcidIndex8	-0.20	$[-0.28$, $-0.12]$	-4.70	< .001
AcidIndex9	-0.50	$[-0.60$, $-0.40]$	-9.91	< .001
AcidIndex10	-1.04	$[-1.15$, $-0.94]$	-19.02	< .001
VolatileAcidity	-0.11	$[-0.15$, $-0.07]$	-5.29	< .001
Alcohol	0.01	$[0.01$, $0.02]$	4.46	< .001
TotalSulfurDioxide	0.00	$[0.00$, $0.00]$	2.75	.006
Density	-0.93	$[-1.79$, $-0.08]$	-2.15	.032
Chlorides	-0.10	$[-0.20$, $0.00]$	-2.04	.041

Table 8

ANOVA of Model 2 and Model 2.1

Res.Df	RSS	Df	Sum.of.Sq	F	Pr..F.
12,777.00	21,800.20	NA	NA	NA	NA
12,774.00	21,788.37	3.00	11.83	2.31	0.07

Table 9

Model 3 Poisson Regression Output

Predictor	*b*	95% CI	*z*	*p*
Intercept	-0.15	$[-0.24$, $-0.06]$	-3.38	.001
AcidIndex7	-0.03	$[-0.06$, $0.00]$	-1.68	.092
AcidIndex8	-0.06	$[-0.10$, $-0.03]$	-3.45	.001
AcidIndex9	-0.17	$[-0.22$, $-0.13]$	-7.68	< .001
AcidIndex10	-0.48	$[-0.53$, $-0.42]$	-16.48	< .001
STARS1	0.76	$[0.72$, $0.80]$	38.96	< .001
STARS2	1.08	$[1.05$, $1.12]$	59.38	< .001
STARS3	1.20	$[1.16$, $1.24]$	62.56	< .001
STARS4	1.32	$[1.28$, $1.37]$	54.44	< .001
LabelAppeal-1	0.24	$[0.16$, $0.31]$	6.27	< .001
LabelAppeal0	0.43	$[0.36$, $0.50]$	11.56	< .001
LabelAppeal1	0.56	$[0.49$, $0.64]$	14.89	< .001
LabelAppeal2	0.70	$[0.61$, $0.78]$	16.44	< .001
Alcohol	0.00	$[0.00$, $0.01]$	3.09	.002

Table 10

Model 4 Poisson Regression Output

Predictor	*b*	95% CI	*z*	*p*
Intercept	0.18	$[-0.20$, $0.56]$	0.92	.358
STARS1	0.76	$[0.72$, $0.80]$	38.89	< .001
STARS2	1.08	$[1.04$, $1.12]$	59.16	< .001
STARS3	1.20	$[1.16$, $1.24]$	62.35	< .001
STARS4	1.32	$[1.27$, $1.37]$	54.33	< .001
LabelAppeal-1	0.24	$[0.16$, $0.31]$	6.24	< .001
LabelAppeal0	0.43	$[0.35$, $0.50]$	11.52	< .001
LabelAppeal1	0.56	$[0.49$, $0.63]$	14.83	< .001
LabelAppeal2	0.70	$[0.61$, $0.78]$	16.40	< .001
AcidIndex7	-0.03	$[-0.06$, $0.00]$	-1.72	.085
AcidIndex8	-0.06	$[-0.10$, $-0.03]$	-3.46	.001
AcidIndex9	-0.17	$[-0.21$, $-0.13]$	-7.58	< .001
AcidIndex10	-0.47	$[-0.53$, $-0.42]$	-16.30	< .001
VolatileAcidity	-0.04	$[-0.06$, $-0.02]$	-3.89	< .001
Alcohol	0.00	$[0.00$, $0.01]$	3.19	.001
TotalSulfurDioxide	0.00	$[0.00$, $0.00]$	1.91	.056
Chlorides	-0.04	$[-0.08$, $0.01]$	-1.67	.094
Density	-0.31	$[-0.69$, $0.07]$	-1.61	.107

Table 11

Model 4.1 Poisson Regression Output

Predictor	*b*	95% CI	*z*	*p*
Intercept	-0.12	$[-0.21$, $-0.04]$	-2.76	.006
STARS1	0.76	$[0.72$, $0.80]$	38.93	< .001
STARS2	1.08	$[1.05$, $1.12]$	59.23	< .001
STARS3	1.20	$[1.16$, $1.24]$	62.44	< .001
STARS4	1.32	$[1.27$, $1.37]$	54.34	< .001
LabelAppeal-1	0.24	$[0.16$, $0.31]$	6.23	< .001
LabelAppeal0	0.43	$[0.35$, $0.50]$	11.51	< .001
LabelAppeal1	0.56	$[0.49$, $0.63]$	14.84	< .001
LabelAppeal2	0.70	$[0.61$, $0.78]$	16.39	< .001
AcidIndex7	-0.03	$[-0.06$, $0.00]$	-1.74	.081
AcidIndex8	-0.06	$[-0.10$, $-0.03]$	-3.49	< .001
AcidIndex9	-0.17	$[-0.22$, $-0.13]$	-7.66	< .001
AcidIndex10	-0.48	$[-0.53$, $-0.42]$	-16.45	< .001
VolatileAcidity	-0.04	$[-0.06$, $-0.02]$	-3.93	< .001
Alcohol	0.00	$[0.00$, $0.01]$	3.16	.002

Table 12

Analysis of Deviance of Models 4 and Model 4.1

Resid..Df	Resid..Dev	Df	Deviance	Pr..Chi.
12,780.00	13,618.32	NA	NA	NA
12,777.00	13,609.28	3.00	9.04	0.03

Table 13

Model 5 Zero Inflated Poisson Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.680	0.202	3.367	0.001
AcidIndex7	-0.024	0.017	-1.387	0.165
AcidIndex8	-0.039	0.018	-2.203	0.028
AcidIndex9	-0.072	0.023	-3.150	0.002
AcidIndex10	-0.120	0.030	-3.994	0.0001
STARS1	0.060	0.021	2.848	0.004
STARS2	0.181	0.020	9.176	0
STARS3	0.278	0.021	13.465	0
STARS4	0.376	0.026	14.708	0
LabelAppeal-1	0.439	0.041	10.645	0
LabelAppeal0	0.727	0.040	18.026	0
LabelAppeal1	0.917	0.041	22.364	0
LabelAppeal2	1.076	0.046	23.636	0
Alcohol	0.007	0.001	4.903	0.00000
VolatileAcidity	-0.013	0.010	-1.375	0.169
TotalSulfurDioxide	-0.00003	0.00003	-0.922	0.357
Chlorides	-0.022	0.022	-0.973	0.330
Density	-0.262	0.197	-1.325	0.185

Table 14

Model 5 Zero Inflated Poisson Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.270	1.338	-2.445	0.014
AcidIndex7	0.105	0.133	0.788	0.431
AcidIndex8	0.319	0.133	2.410	0.016
AcidIndex9	1.004	0.148	6.767	0
AcidIndex10	1.914	0.156	12.254	0
STARS1	-2.064	0.075	-27.365	0
STARS2	-5.836	0.359	-16.253	0
STARS3	-20.192	349.627	-0.058	0.954
STARS4	-20.391	648.154	-0.031	0.975
LabelAppeal-1	1.473	0.318	4.635	0.00000
LabelAppeal0	2.202	0.315	6.989	0
LabelAppeal1	2.917	0.321	9.098	0
LabelAppeal2	3.343	0.374	8.935	0
Alcohol	0.023	0.009	2.487	0.013
VolatileAcidity	0.204	0.059	3.442	0.001
TotalSulfurDioxide	-0.001	0.0002	-4.868	0.00000
Chlorides	0.104	0.145	0.716	0.474
Density	0.850	1.292	0.658	0.511

Table 15

Model 6 Zero Inflated Poisson Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.408	0.047	8.633	0
AcidIndex7	-0.025	0.017	-1.452	0.146
AcidIndex8	-0.041	0.018	-2.280	0.023
AcidIndex9	-0.073	0.023	-3.200	0.001
AcidIndex10	-0.121	0.030	-4.038	0.0001
STARS1	0.061	0.021	2.902	0.004
STARS2	0.181	0.020	9.193	0
STARS3	0.279	0.021	13.499	0
STARS4	0.378	0.026	14.765	0
LabelAppeal-1	0.439	0.041	10.656	0
LabelAppeal0	0.727	0.040	18.052	0
LabelAppeal1	0.918	0.041	22.396	0
LabelAppeal2	1.076	0.045	23.650	0
Alcohol	0.007	0.001	4.946	0.00000
VolatileAcidity	-0.013	0.010	-1.313	0.189

Table 16

Model 6 Zero Inflated Poisson Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.629	0.351	-7.492	0
AcidIndex7	0.093	0.132	0.702	0.483
AcidIndex8	0.306	0.132	2.315	0.021
AcidIndex9	1.010	0.148	6.832	0
AcidIndex10	1.922	0.156	12.339	0
STARS1	-2.055	0.075	-27.386	0
STARS2	-5.859	0.370	-15.836	0
STARS3	-20.191	351.218	-0.057	0.954
STARS4	-20.373	652.299	-0.031	0.975
LabelAppeal-1	1.471	0.315	4.664	0.00000
LabelAppeal0	2.205	0.313	7.055	0
LabelAppeal1	2.912	0.318	9.158	0
LabelAppeal2	3.328	0.371	8.961	0
Alcohol	0.025	0.009	2.688	0.007
VolatileAcidity	0.214	0.059	3.635	0.0003

Table 17

Model 7 Negative Binomial Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.149	0.044	-3.376	0.001
AcidIndex7	-0.029	0.017	-1.684	0.092
AcidIndex8	-0.061	0.018	-3.452	0.001
AcidIndex9	-0.173	0.022	-7.681	0
AcidIndex10	-0.477	0.029	-16.480	0
STARS1	0.762	0.020	38.956	0
STARS2	1.083	0.018	59.376	0
STARS3	1.202	0.019	62.558	0
STARS4	1.323	0.024	54.442	0
LabelAppeal-1	0.238	0.038	6.275	0
LabelAppeal0	0.428	0.037	11.560	0
LabelAppeal1	0.561	0.038	14.885	0
LabelAppeal2	0.698	0.042	16.441	0
Alcohol	0.004	0.001	3.087	0.002

Table 18

Model 8 Negative Binomial Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.180	0.196	0.919	0.358
STARS1	0.761	0.020	38.894	0
STARS2	1.080	0.018	59.157	0
STARS3	1.199	0.019	62.352	0
STARS4	1.321	0.024	54.330	0
LabelAppeal-1	0.237	0.038	6.235	0
LabelAppeal0	0.427	0.037	11.517	0
LabelAppeal1	0.559	0.038	14.833	0
LabelAppeal2	0.696	0.042	16.402	0
AcidIndex7	-0.029	0.017	-1.725	0.085
AcidIndex8	-0.061	0.018	-3.461	0.001
AcidIndex9	-0.171	0.022	-7.582	0
AcidIndex10	-0.473	0.029	-16.302	0
VolatileAcidity	-0.037	0.009	-3.885	0.0001
Alcohol	0.005	0.001	3.192	0.001
TotalSulfurDioxide	0.0001	0.00003	1.910	0.056
Chlorides	-0.037	0.022	-1.673	0.094
Density	-0.310	0.192	-1.614	0.107

Table 19

Model 8.1 Negative Binomial Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.123	0.044	-2.763	0.006
STARS1	0.761	0.020	38.932	0
STARS2	1.081	0.018	59.224	0
STARS3	1.200	0.019	62.436	0
STARS4	1.321	0.024	54.343	0
LabelAppeal-1	0.237	0.038	6.234	0
LabelAppeal0	0.427	0.037	11.513	0
LabelAppeal1	0.559	0.038	14.835	0
LabelAppeal2	0.695	0.042	16.385	0
AcidIndex7	-0.030	0.017	-1.744	0.081
AcidIndex8	-0.061	0.018	-3.490	0.0005
AcidIndex9	-0.172	0.022	-7.662	0
AcidIndex10	-0.476	0.029	-16.445	0
VolatileAcidity	-0.037	0.009	-3.931	0.0001
Alcohol	0.004	0.001	3.158	0.002

Table 20

Analysis of Deviance of Models 8 and Model 8.1

theta	Resid..df	X...2.x.log.lik.	Test	X...df	LR.stat.	Pr.Chi.
40,743.88	12,780.00	-45,560.76		NA	NA	NA
40,788.65	12,777.00	-45,551.72	1 vs 2	3.00	9.04	0.03

Table 21

Model 9 Zero Inflated Negative Binomial Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.680	0.202	3.367	0.001
AcidIndex7	-0.024	0.017	-1.386	0.166
AcidIndex8	-0.039	0.018	-2.203	0.028
AcidIndex9	-0.072	0.023	-3.150	0.002
AcidIndex10	-0.120	0.030	-3.993	0.0001
STARS1	0.060	0.021	2.848	0.004
STARS2	0.181	0.020	9.177	0
STARS3	0.278	0.021	13.465	0
STARS4	0.377	0.026	14.710	0
LabelAppeal-1	0.439	0.041	10.645	0
LabelAppeal0	0.727	0.040	18.026	0
LabelAppeal1	0.917	0.041	22.364	0
LabelAppeal2	1.076	0.046	23.636	0
Alcohol	0.007	0.001	4.902	0.00000
VolatileAcidity	-0.013	0.010	-1.376	0.169
TotalSulfurDioxide	-0.00003	0.00003	-0.922	0.356
Chlorides	-0.022	0.022	-0.973	0.330
Density	-0.262	0.197	-1.325	0.185
Log(theta)	17.951	2.052	8.750	0

Table 22

Model 9 Zero Inflated Negative Binomial Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.270	1.338	-2.445	0.014
AcidIndex7	0.105	0.133	0.788	0.431
AcidIndex8	0.319	0.133	2.410	0.016
AcidIndex9	1.004	0.148	6.767	0
AcidIndex10	1.914	0.156	12.253	0
STARS1	-2.065	0.075	-27.366	0
STARS2	-5.835	0.359	-16.263	0
STARS3	-20.188	348.951	-0.058	0.954
STARS4	-20.388	647.423	-0.031	0.975
LabelAppeal-1	1.474	0.318	4.635	0.00000
LabelAppeal0	2.202	0.315	6.988	0
LabelAppeal1	2.917	0.321	9.097	0
LabelAppeal2	3.343	0.374	8.935	0
Alcohol	0.023	0.009	2.487	0.013
VolatileAcidity	0.204	0.059	3.442	0.001
TotalSulfurDioxide	-0.001	0.0002	-4.869	0.00000
Chlorides	0.104	0.145	0.716	0.474
Density	0.850	1.292	0.658	0.511

Table 23

Model 10 Zero Inflated Negative Binomial Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.399	0.047	8.528	0
AcidIndex7	-0.025	0.017	-1.439	0.150
AcidIndex8	-0.041	0.018	-2.274	0.023
AcidIndex9	-0.074	0.023	-3.215	0.001
AcidIndex10	-0.122	0.030	-4.064	0.00005
STARS1	0.061	0.021	2.893	0.004
STARS2	0.181	0.020	9.198	0
STARS3	0.279	0.021	13.502	0
STARS4	0.378	0.026	14.769	0
LabelAppeal-1	0.440	0.041	10.682	0
LabelAppeal0	0.728	0.040	18.081	0
LabelAppeal1	0.919	0.041	22.429	0
LabelAppeal2	1.077	0.045	23.685	0
Alcohol	0.007	0.001	4.935	0.00000
Log(theta)	17.761			

Table 24

Model 10 Zero Inflated Negative Binomial Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.474	0.347	-7.137	0
AcidIndex7	0.079	0.132	0.601	0.548
AcidIndex8	0.291	0.132	2.209	0.027
AcidIndex9	1.002	0.147	6.796	0
AcidIndex10	1.917	0.155	12.339	0
STARS1	-2.050	0.075	-27.402	0
STARS2	-5.893	0.376	-15.654	0
STARS3	-20.202	351.964	-0.057	0.954
STARS4	-20.388	652.258	-0.031	0.975
LabelAppeal-1	1.460	0.314	4.645	0.00000
LabelAppeal0	2.196	0.311	7.054	0
LabelAppeal1	2.906	0.317	9.174	0
LabelAppeal2	3.323	0.370	8.970	0
Alcohol	0.026	0.009	2.773	0.006

Table 25

Compare Models

	AIC	BIC	loglik
mod1_out	43,204.08	43,315.93	-21,587.04
mod2.1_out	43,166.64	43,308.32	-21,564.32
mod3_out	45,603.97	45,708.37	-22,787.99
mod4_out	45,587.30	45,721.52	-22,775.65
mod5_out	40,814.95	NA	-20,371.48
mod6_out	40,832.20	NA	-20,386.10
mod7_out	45,606.40	45,718.25	-22,788.20
mod8_out	45,589.72	45,731.40	-22,775.86
mod9_out	40,816.95	NA	-20,371.48
mod10_out	40,845.92	NA	-20,393.96

Table 26

Compare Training and Test TARGET Variable

	Min	Max	Range	Mean	Med	Stdev	Skew	Kurt	NAs	Outliers
train.TARGET	0.00	8.00	8.00	3.03	3.00	1.93	-0.33	2.12	0.00	0.00
test.TARGET	0.14	7.09	6.94	3.06	3.02	1.47	0.11	2.24	0.00	0.00

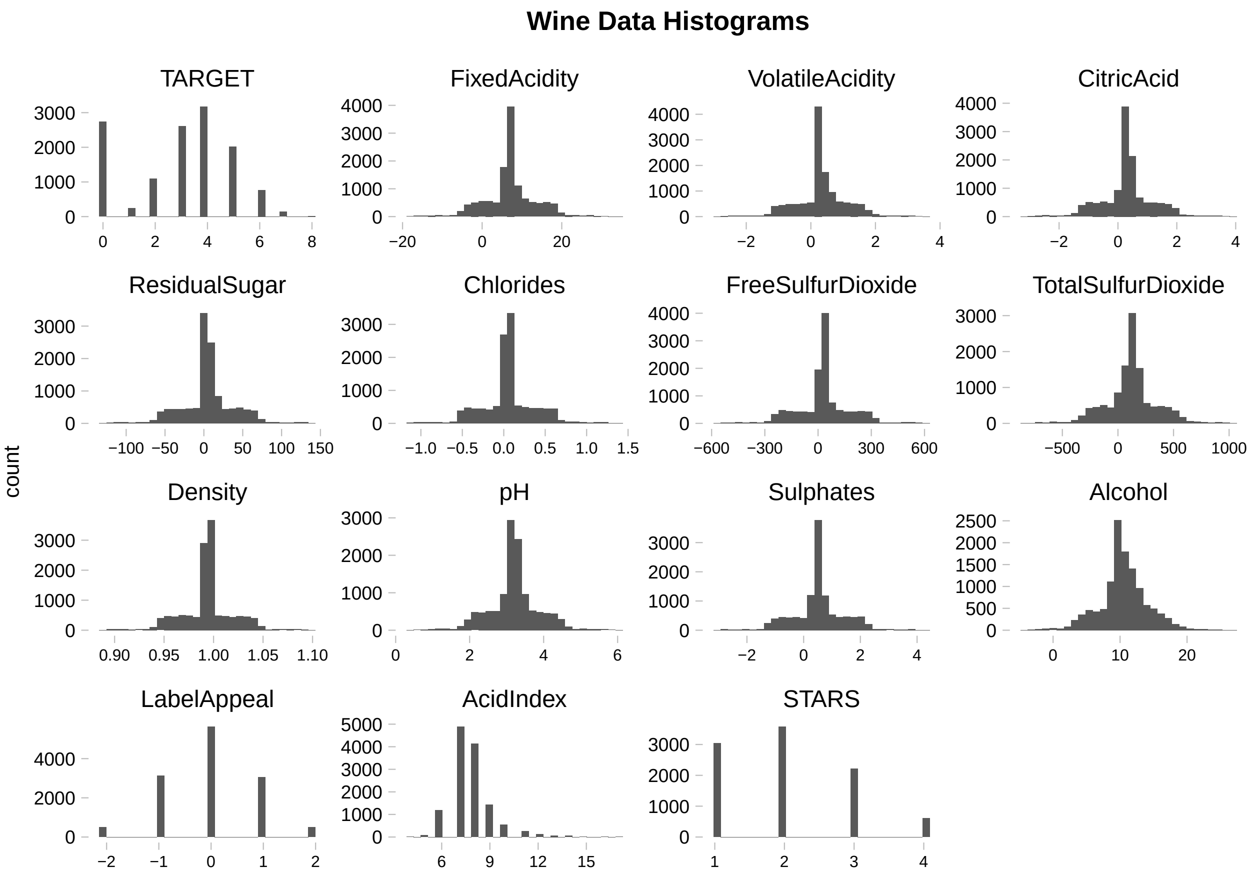


Figure 1. Histogram of Wine Training Data Variables

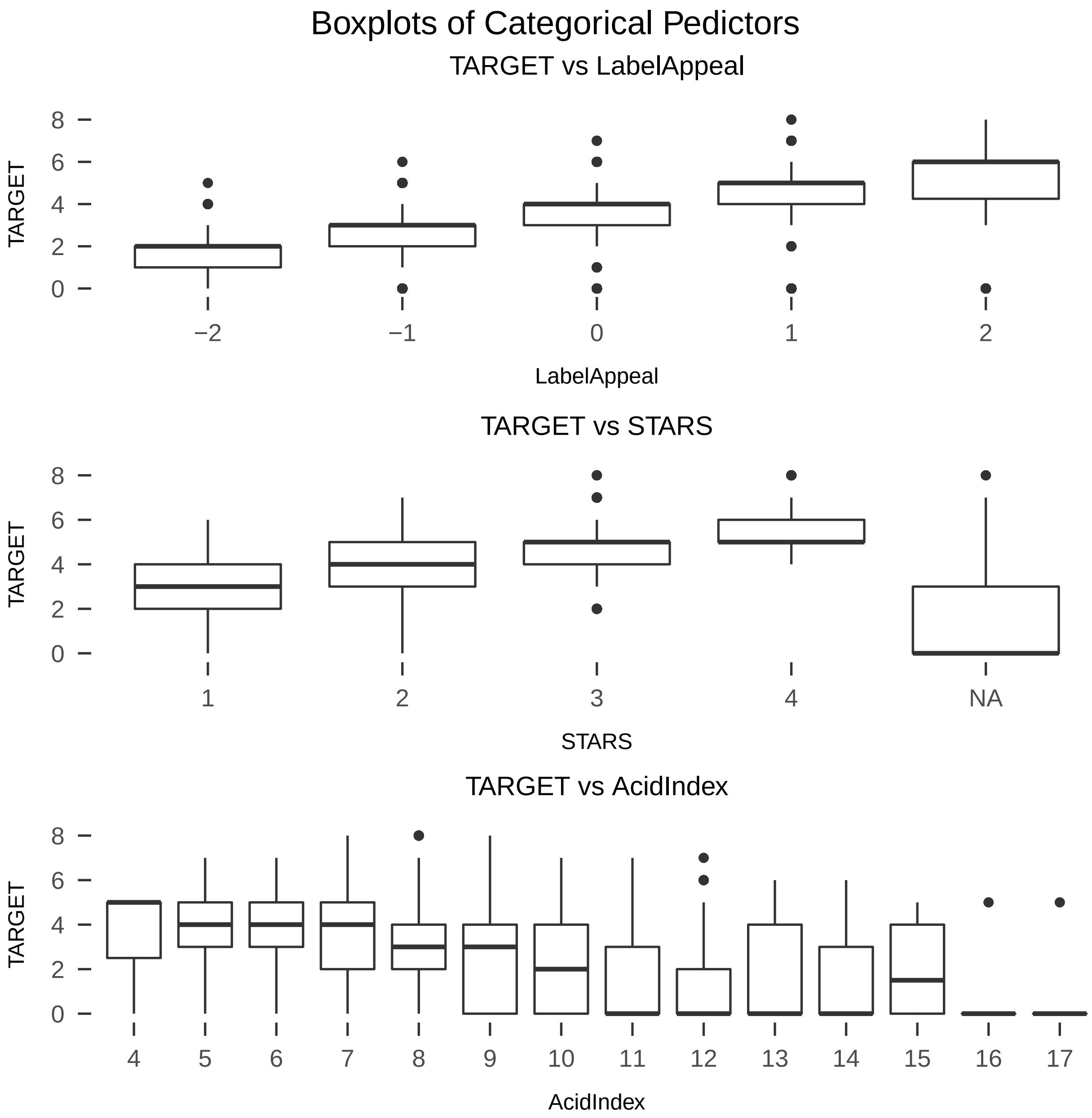


Figure 2. Boxplots of TARGET vs. Categorical Variables

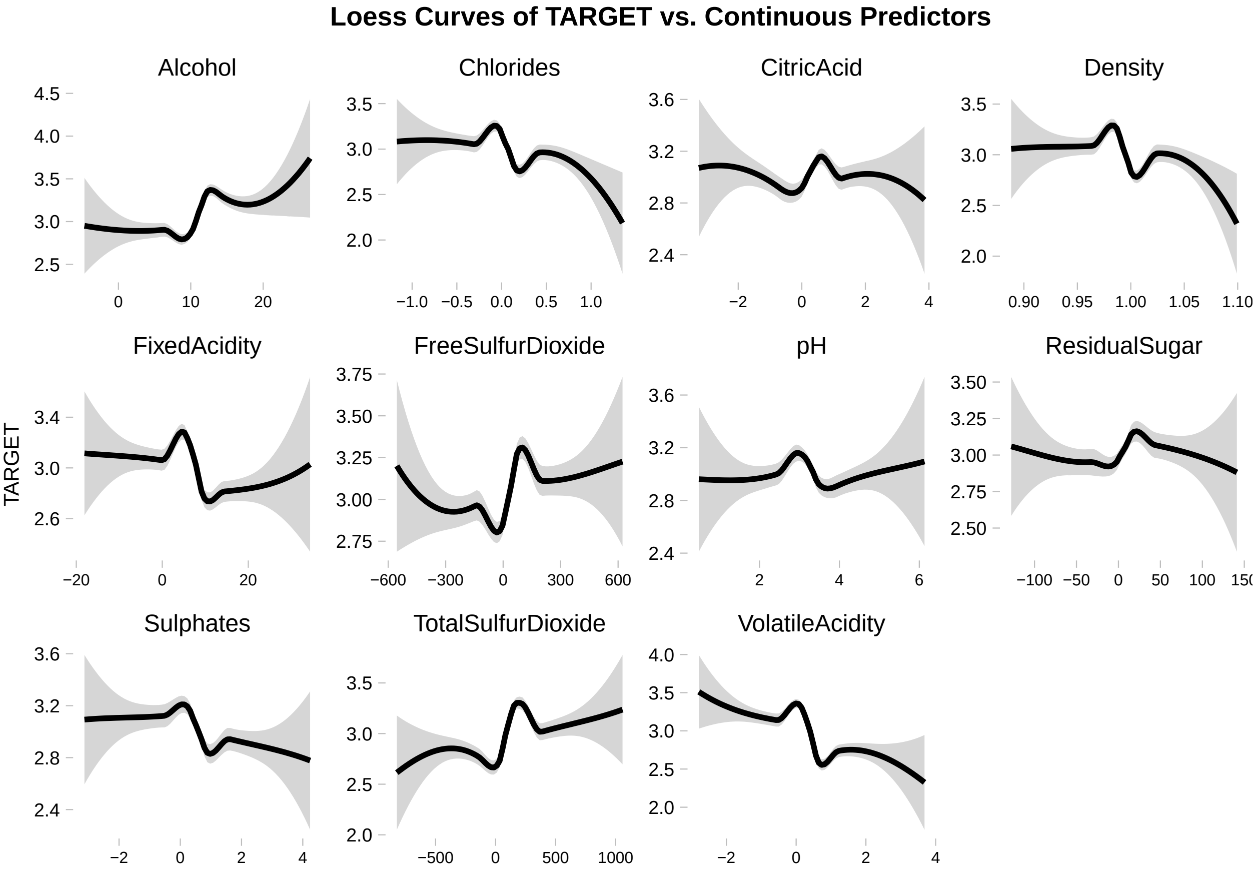
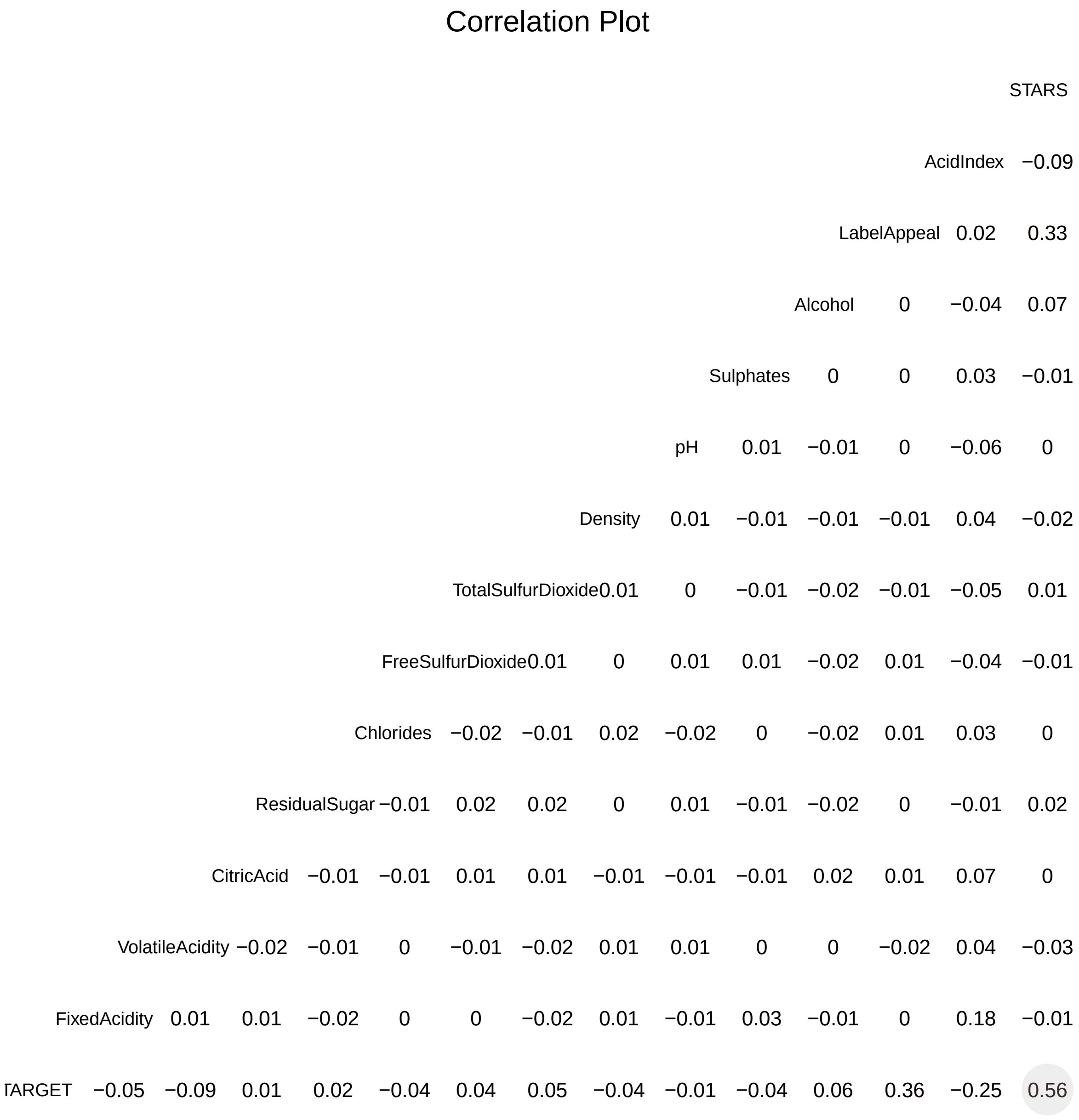


Figure 3. Loess Curves of TARGET vs. Continuous Variables



Correlation Plot

	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
STARS														
AcidIndex														−0.09
LabelAppeal													0.02	0.33
Alcohol												0	−0.04	0.07
Sulphates											0	0	0.03	−0.01
pH										0.01	−0.01	0	−0.06	0
Density									0.01	−0.01	−0.01	−0.01	0.04	−0.02
TotalSulfurDioxide								0.01	0	−0.01	−0.02	−0.01	−0.05	0.01
FreeSulfurDioxide							0.01	0	0.01	0.01	−0.02	0.01	−0.04	−0.01
Chlorides						−0.02	−0.01	0.02	−0.02	0	−0.02	0.01	0.03	0
ResidualSugar					−0.01	0.02	0.02	0	0.01	−0.01	−0.02	0	−0.01	0.02
CitricAcid				−0.01	−0.01	0.01	0.01	−0.01	−0.01	−0.01	0.02	0.01	0.07	0
VolatileAcidity			−0.02	−0.01	0	−0.01	−0.02	0.01	0.01	0	0	−0.02	0.04	−0.03
FixedAcidity		0.01	0.01	−0.02	0	0	−0.02	0.01	−0.01	0.03	−0.01	0	0.18	−0.01
TARGET	−0.05	−0.09	0.01	0.02	−0.04	0.04	0.05	−0.04	−0.01	−0.04	0.06	0.36	−0.25	0.56

Figure 4. Correlation of Wine Data Variables

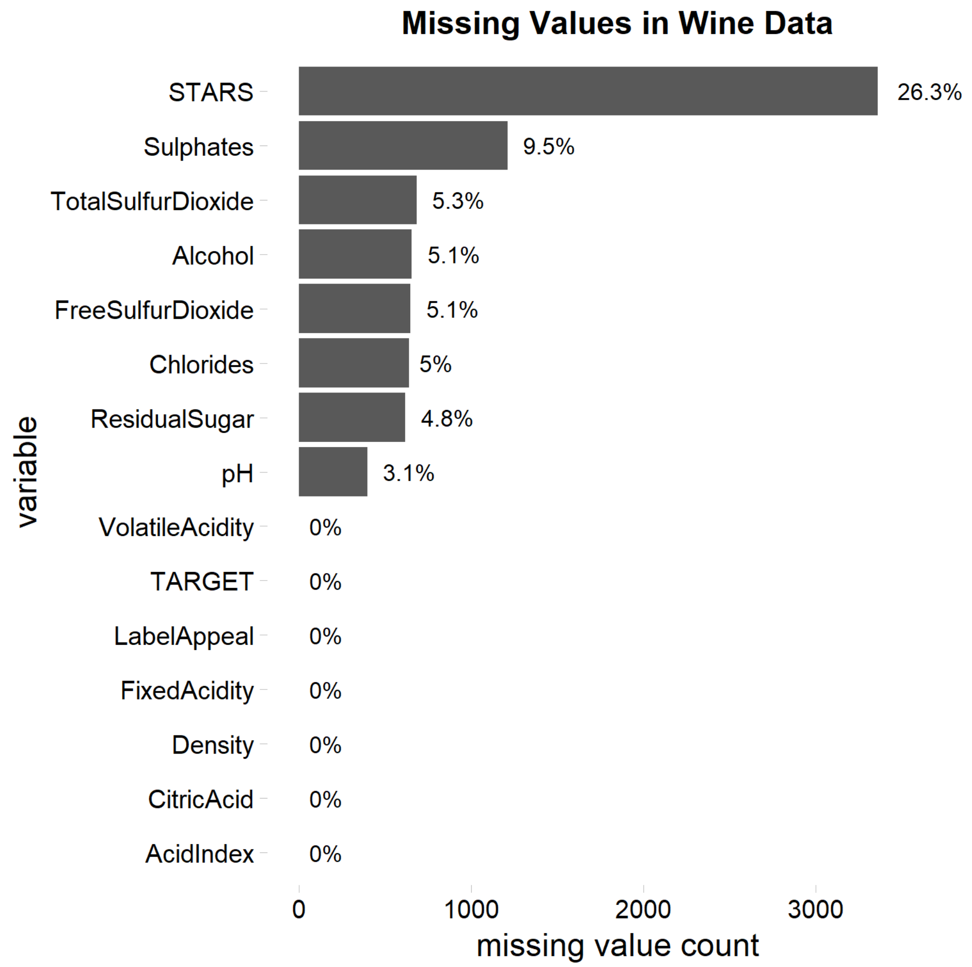


Figure 5. Plot of Missing Observations in Wine Training Data

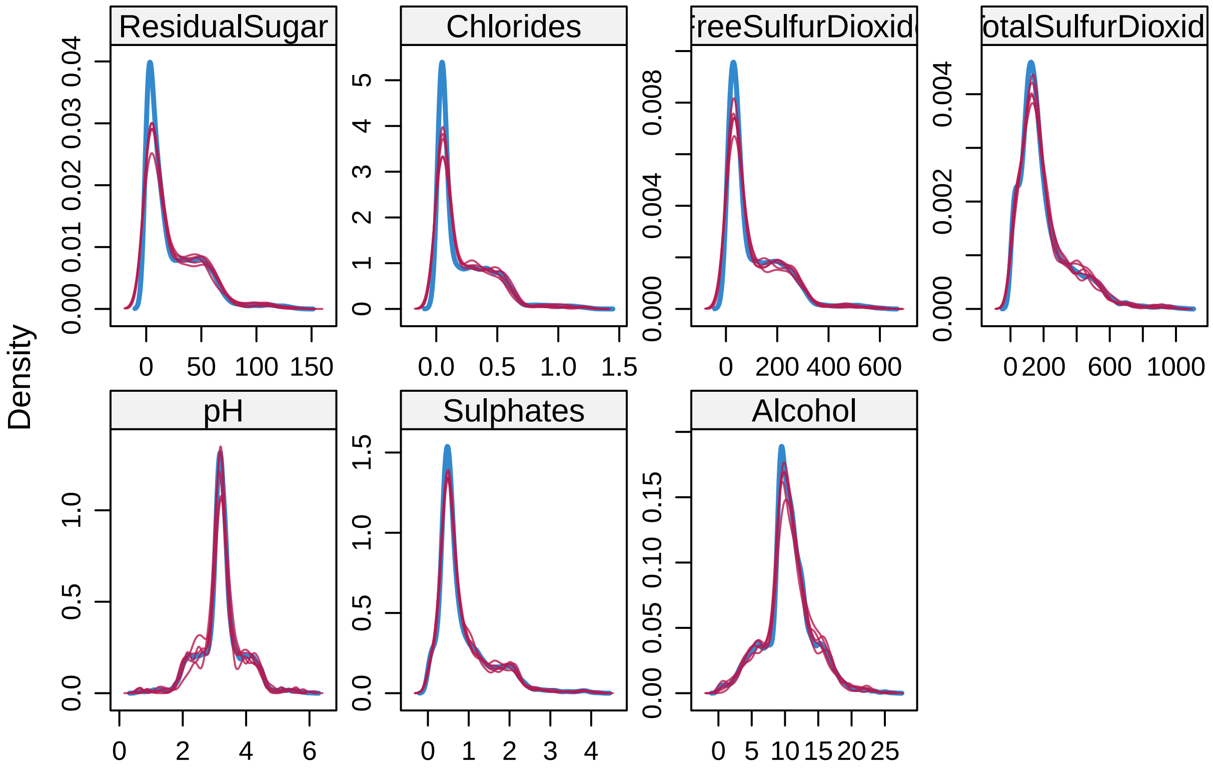


Figure 6. Graphical Summary of MICE Imputation Procedure

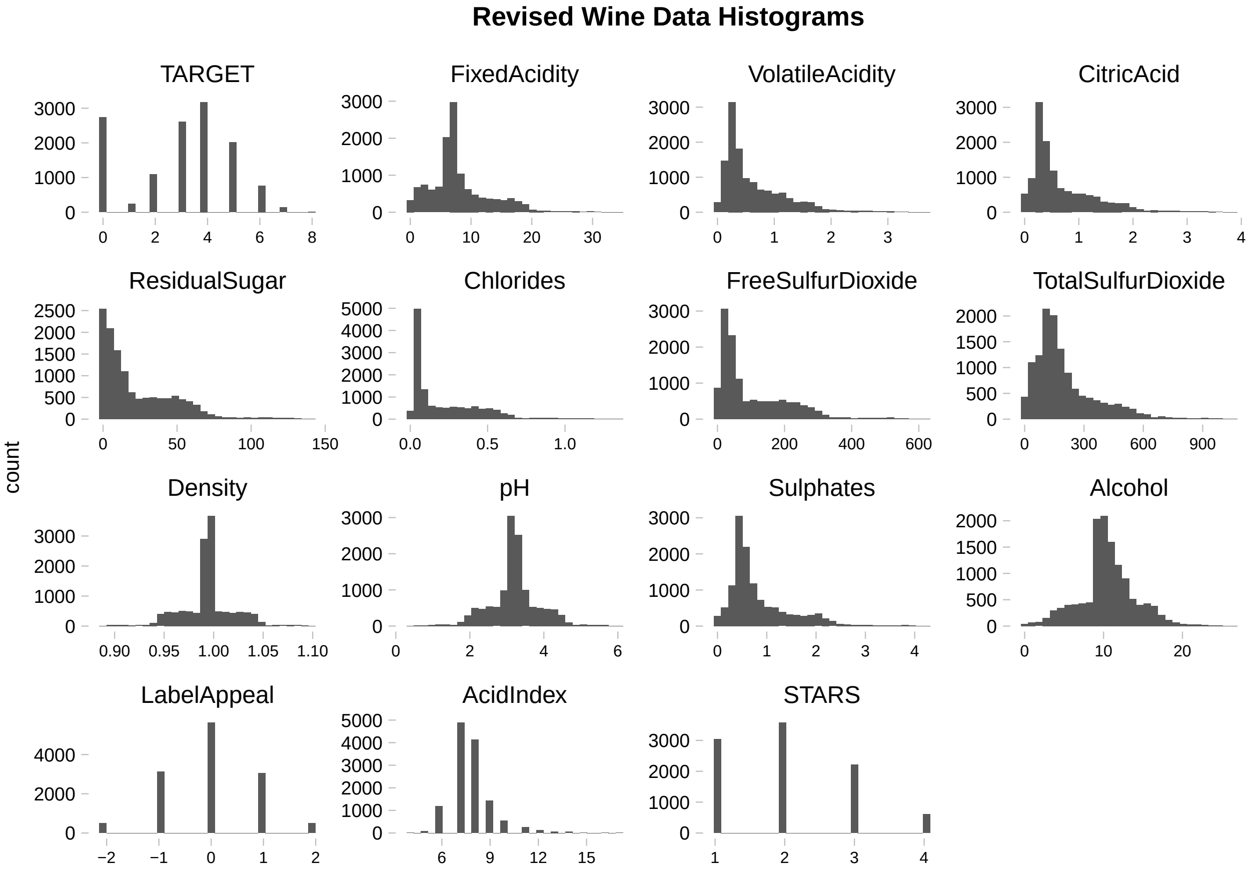
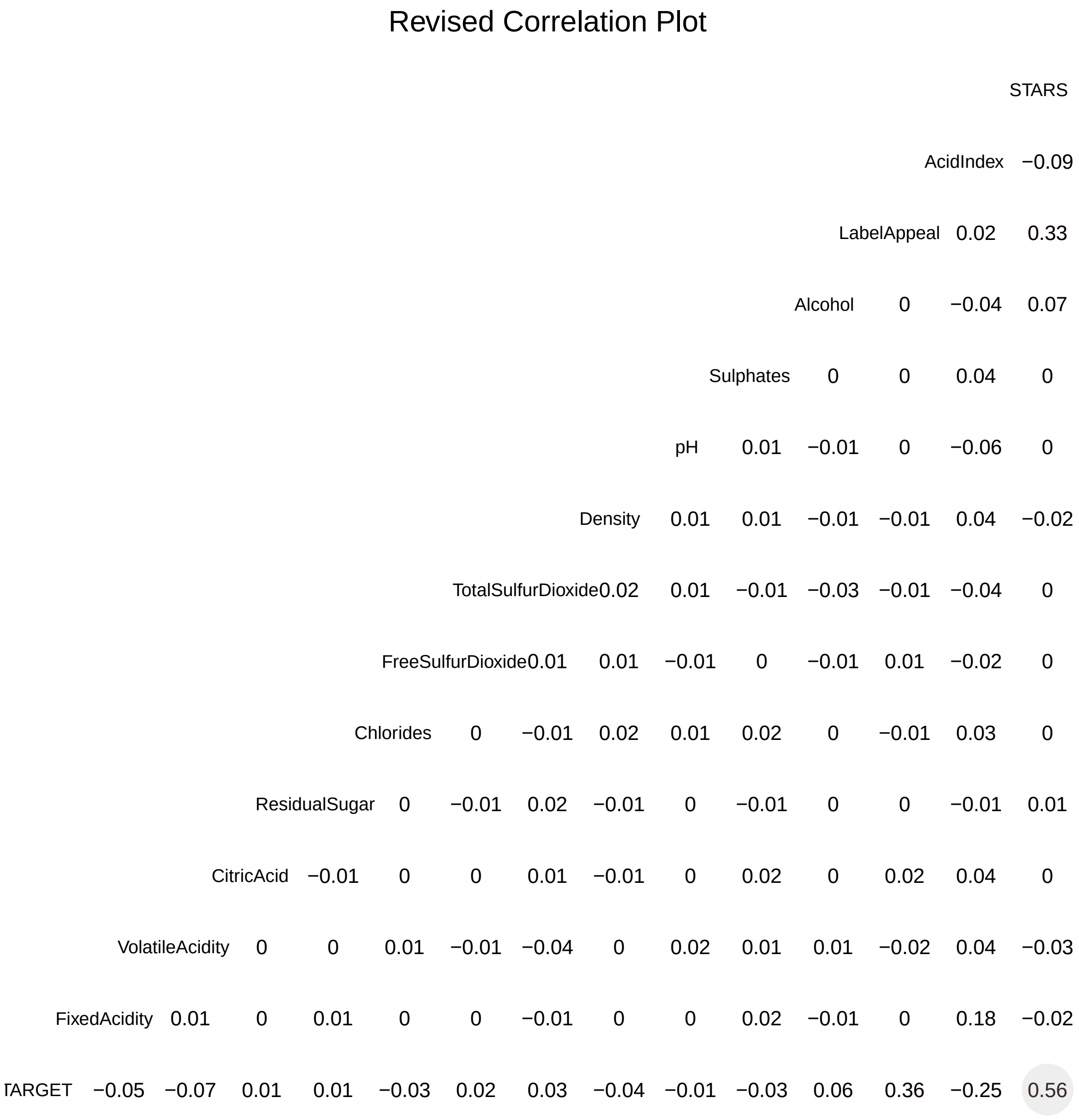


Figure 7. Revised Histogram of Wine Training Data Variables



Revised Correlation Plot

	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
STARS														
AcidIndex														−0.09
LabelAppeal													0.02	0.33
Alcohol												0	−0.04	0.07
Sulphates											0	0	0.04	0
pH										0.01	−0.01	0	−0.06	0
Density									0.01	0.01	−0.01	−0.01	0.04	−0.02
TotalSulfurDioxide								0.02	0.01	−0.01	−0.03	−0.01	−0.04	0
FreeSulfurDioxide							0.01	0.01	−0.01	0	−0.01	0.01	−0.02	0
Chlorides						0	−0.01	0.02	0.01	0.02	0	−0.01	0.03	0
ResidualSugar					0	−0.01	0.02	−0.01	0	−0.01	0	0	−0.01	0.01
CitricAcid				−0.01	0	0	0.01	−0.01	0	0.02	0	0.02	0.04	0
VolatileAcidity			0	0	0.01	−0.01	−0.04	0	0.02	0.01	0.01	−0.02	0.04	−0.03
FixedAcidity		0.01	0	0.01	0	0	−0.01	0	0	0.02	−0.01	0	0.18	−0.02
TARGET	−0.05	−0.07	0.01	0.01	−0.03	0.02	0.03	−0.04	−0.01	−0.03	0.06	0.36	−0.25	0.56

Figure 8. Revised Correlation of Wine Data Variables

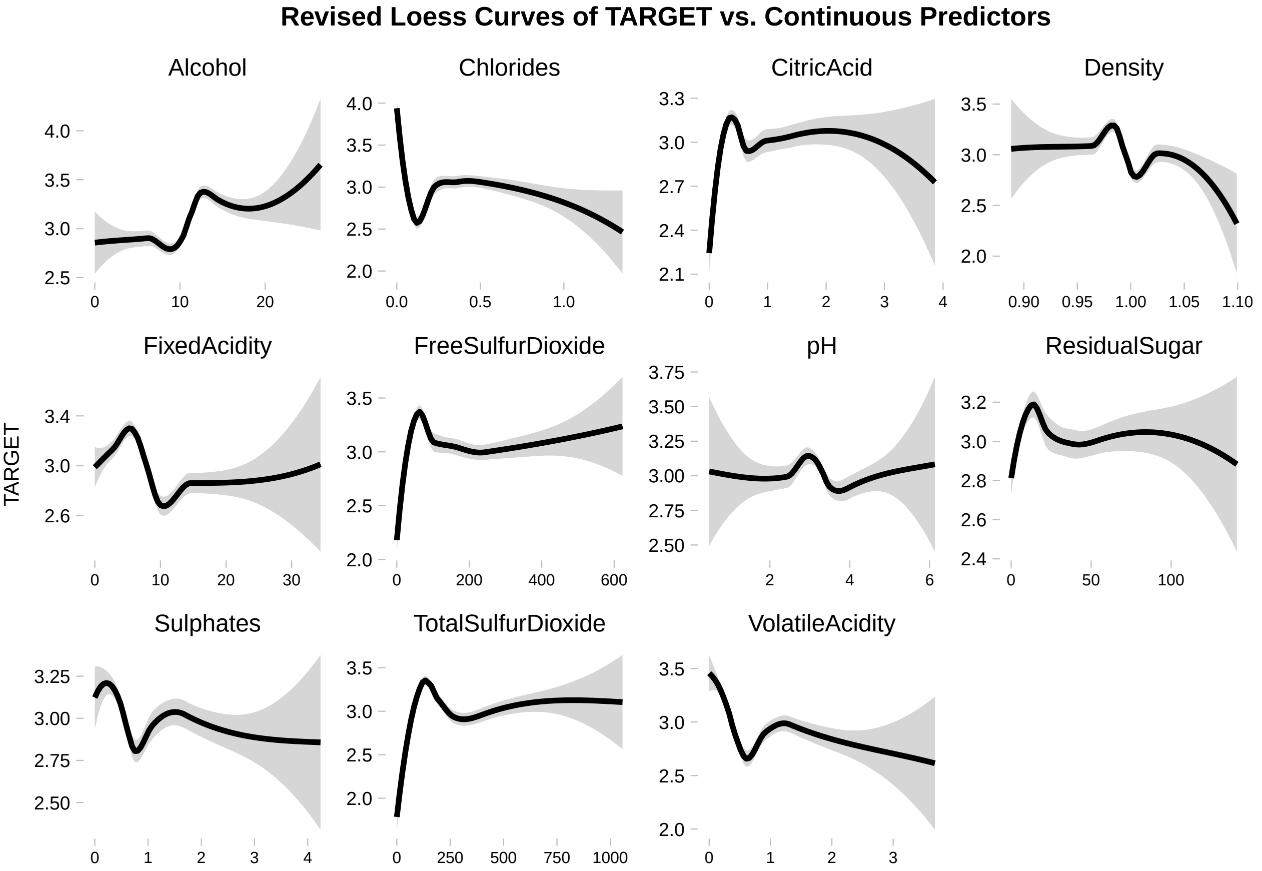


Figure 9. Revised Loess Curves of TARGET vs. Continuous Variables

Appendix A

*

Below are the scripts used to produce this paper:

```
# must have following three packages insalled for rest of code to work
if (!require(pacman)) install.packages('pacman'); library(pacman)
if(!"devtools" %in% rownames(installed.packages())) install.packages("devtools")
if(!"papaja" %in% rownames(installed.packages())) devtools::install_github("crsh/papaja")

p_load(knitr, dplyr, papaja, DataExplorer, moments, ggplot2, ggthemes, tidyr,
       grid,gridExtra, GGally, mice, car, pscl, stargazer,memisc, MASS, sme)
opts_chunk$set(warning=FALSE, message=FALSE, cache=TRUE)

# custom function for altering code chunk output size
def.chunk.hook  <- knitr::knit_hooks$get("chunk")
knitr::knit_hooks$set(chunk = function(x, options) {
  x <- def.chunk.hook(x, options)
  ifelse(options$size != "normalsize", paste0("\\", options$size,"\n\n", x, "\n\n \\normalsize"), x)
})

# read in data
train_url <- 'https://raw.githubusercontent.com/spitakiss/Data621/master/Homework5/wine-training-data.csv'
train <- read.csv(train_url)
names(train)[1] <- "INDEX"
# print variable descriptions
var_url <- 'https://raw.githubusercontent.com/spitakiss/Data621/master/Homework5/VariableDefinitions.csv'
vars <- read.csv(var_url, stringsAsFactors = F)
apa_table(vars,caption = "Variable Names and Descriptions in Wine Data Set",small=T,
          col.names=c("Variable Name","Definition", "Theoretical Effect"))
# print table of categorical features
tblFun <- function(x, lab){
    tbl <- table(x, useNA = "ifany")
    tbl <- cbind(tbl,prop.table(tbl)*100)
    tbl <- cbind(tbl, cumsum(tbl[,2]))
    tbl[,2] <- round(tbl[,2],1)
    tbl[,3] <- round(tbl[,3],1)
    colnames(tbl) <- c('Count','Pct', 'Cumul Pct')
    data.frame(cbind(Variable = lab,Levels=row.names(tbl), tbl), row.names=NULL)
}
```

```r
a <- tblFun(train$LabelAppeal, 'LabelAppeal')
b <- tblFun(train$STARS, 'STARS')
c <- tblFun(train$AcidIndex, 'AcidIndex')
tbl <- data.frame(rbind(a,b,c))
tbl[2:5,1] <- ""
tbl[7:10,1] <- ""
tbl[12:24,1] <- ""

options(knitr.kable.NA = '')
apa_table(tbl,midrules = c(5,10), format.args = list(big.mark=","),
          small=TRUE, align = c("lrrrr"),
          caption = "Categorical Features Summary",
          longtable=TRUE)

#  function to create descriptive stats
stats = function (dataframe) {
  Min = sapply(dataframe, function(x) {min(x, na.rm=TRUE)})
  Max = sapply(dataframe, function(x) {max(x, na.rm=TRUE)})
  Range = Max - Min
  Mean = sapply(dataframe, function(x) {round(mean(x, na.rm=TRUE),3)})
  Med = sapply(dataframe, function(x) {round(median(x, na.rm=TRUE),3)})
  Stdev = sapply(dataframe, function(x) {round(sd(x, na.rm=TRUE),3)})
  Skew = sapply(dataframe, function(x) {round(skewness(x, na.rm=TRUE),3)})
  Kurt = sapply(dataframe, function(x) {round(kurtosis(x, na.rm=TRUE),3)})
  NAs = sapply(dataframe, function(x) {round(sum(length(which(is.na(x)))),3)})
  Outliers = sapply(dataframe, function(x) {
    round(length(which(x>mean(x)+ 3*sd(x) | x<mean(x)- 3*sd(x))),3)
    })
    return(cbind(Min, Max, Range, Mean, Med, Stdev, Skew, Kurt, NAs, Outliers))
}
# print table of stats
stats_df <- data.frame(stats(train[,2:ncol(train)]))
apa_table(stats_df,small=T,
          caption= "Descriptive Statistics Summary",
          format.args = list(digits = c(rep(1,8),rep(0,2)), margin = 2),
          align = rep("r",10),
          note="Outliers are defined as observations
          that are more than 3 standard deviations from the mean",
          placement="tbp")
```

```r
# plot histograms of training data variables
mylevels <- names(train[,2:16])
hplot <- train %>%
  select(-INDEX) %>%
  gather() %>%
  mutate(facet = factor(key, levels=mylevels)) %>%
  ggplot(aes(value)) +
  facet_wrap(~ facet, scales = "free") +
  geom_histogram()  + theme_pander()  +
  theme(axis.text.y = element_text(size=7),
        strip.text.x = element_text(size= 9),
        axis.text.x = element_text(size=6),
        plot.title = element_text(hjust = 0.5, size=10),
        axis.title.y = element_text(size=8)) +
  labs(x=NULL, title="Wine Data Histograms")
hplot
# convert categorical features to factors
train$LabelAppeal <- ordered(train$LabelAppeal, levels = c(-2,-1,0,1,2))
train$STARS <- ordered(train$STARS, levels = c(1,2,3,4))
train$AcidIndex <- ordered(train$AcidIndex, levels = seq(4,17))

# print boxplots for categorical variables
bp1 <- ggplot(train, aes(LabelAppeal,TARGET)) + geom_boxplot() + theme_apa() +
  theme(axis.title = element_text(size=10),
        plot.title = element_text(hjust= 0.5, size = 12)) +
  labs(title = 'TARGET vs LabelAppeal')

bp2 <- ggplot(train, aes(STARS,TARGET)) + geom_boxplot() + theme_apa() +
  theme(axis.title = element_text(size=10),
        plot.title = element_text(hjust= 0.5, size = 12)) +
  labs(title = 'TARGET vs STARS')

bp3 <- ggplot(train, aes(AcidIndex,TARGET)) + geom_boxplot() + theme_apa() +
  theme(axis.title = element_text(size=10),
        plot.title = element_text(hjust= 0.5, size = 12)) +
  labs(title = 'TARGET vs AcidIndex')

grid.arrange(bp1, bp2, bp3, ncol = 1,
             top = textGrob("Boxplots of Categorical Pedictors",
                            gp=gpar(fontsize=15)))
```

```
lcplot <- train %>%
  select(-LabelAppeal, -STARS, -INDEX, -AcidIndex) %>%
  gather(key="key", value="value",-TARGET) %>%
  ggplot(aes(value, TARGET)) +
  facet_wrap(~ key, scales = "free") +
  theme_pander()  +
  theme(axis.text.y = element_text(size=7),
        strip.text.x = element_text(size= 9),
        axis.text.x = element_text(size=6),
        plot.title = element_text(hjust = 0.5, size=10),
        axis.title.y = element_text(size=8)) +
  labs(x=NULL, title="Loess Curves of TARGET vs. Continuous Predictors") +
  geom_smooth(method="loess", color="black")

lcplot

# convert categorical variables back to integers; used for correlation purposes only
train$LabelAppeal <- as.integer(as.character(train$LabelAppeal))
train$STARS <- as.integer(as.character(train$STARS))
train$AcidIndex <- as.integer(as.character(train$AcidIndex))

ggcorr(train[,2:16], method=c("pairwise","pearson"), label=T,
       label_round = 2, size=2.9, hjust=0.65, label_size=3, geom="text") +
  geom_point(size = 10, color="grey",aes(alpha = abs(coefficient) > 0.5)) +
  scale_alpha_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +
  guides(color = FALSE, alpha = FALSE) + labs(title = "Correlation Plot") +
  theme(plot.title = element_text(hjust=0.5))

# absolute value transformations to predictors with negative vals
train$FixedAcidity <- abs(train$FixedAcidity)
train$VolatileAcidity <- abs(train$VolatileAcidity)
train$CitricAcid <- abs(train$CitricAcid)
train$ResidualSugar <- abs(train$ResidualSugar)
train$Chlorides <- abs(train$Chlorides)
train$FreeSulfurDioxide <- abs(train$FreeSulfurDioxide)
train$TotalSulfurDioxide <- abs(train$TotalSulfurDioxide)
train$Sulphates <- abs(train$Sulphates)
train$Alcohol <- abs(train$Alcohol)
```

```
# plot missing
missing_ct <- data.frame(variable=row.names(stats_df), NAs=stats_df$NAs,
                         Pct=round(stats_df$NAs/nrow(train)*100,1))
miss_plot <- ggplot(missing_ct, aes(reorder(variable,NAs), NAs)) +
  geom_col() + coord_flip() + theme_pander() +
  geom_text(aes(y = NAs, label=paste0(Pct,"%")), hjust=-0.3, size=3) +
  ylim(c(0,3700)) + labs(x="variable",y="missing value count") +
  labs(title="Missing Values in Wine Data") +
  theme(plot.title = element_text(hjust = 0.5, size=12))

miss_plot
# convert STARS back to factor
train$STARS <- ifelse(is.na(train$STARS),"NA",train$STARS)
train$STARS <- ordered(train$STARS, levels = c("NA","1","2","3","4"))

# fix missing values with mice
tmp_data <- mice(train,maxit=3, method='pmm',seed=20, print=F)
train <- complete(tmp_data,1)

# plot mice
densityplot(tmp_data)

# print table of revised desc stats - post fixing negative values and missing
train$STARS <- as.integer(as.character(train$STARS))
stats_df <- data.frame(stats(train[,2:ncol(train)]))
apa_table(stats_df,small=T,
          caption= "Revised Descriptive Statistics Summary",
          format.args = list(digits = c(rep(1,8),rep(0,2)), margin = 2),
          align = rep("r",10),
          note="Outliers are defined as observations
          that are more than 3 standard deviations from the mean",
          placement="tbp")

# plot histograms of revised training dat
mylevels <- names(train[,2:16])
hplot <- train %>%
  select(-INDEX) %>%
  gather() %>%
  mutate(facet = factor(key, levels=mylevels)) %>%
  ggplot(aes(value)) +
  facet_wrap(~ facet, scales = "free") +
  geom_histogram()  + theme_pander()  +
  theme(axis.text.y = element_text(size=7),
        strip.text.x = element_text(size= 9),
        axis.text.x = element_text(size=6),
        plot.title = element_text(hjust = 0.5, size=10),
        axis.title.y = element_text(size=8)) +
  labs(x=NULL, title="Revised Wine Data Histograms")
hplot
# revised correlation plots
ggcorr(train[,2:16], method=c("pairwise","pearson"), label=T,
       label_round = 2, size=2.9, hjust=0.65, label_size=3, geom="text") +
  geom_point(size = 10, color="grey",aes(alpha = abs(coefficient) > 0.5)) +
  scale_alpha_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +
  guides(color = FALSE, alpha = FALSE) + labs(title = "Revised Correlation Plot") +
  theme(plot.title = element_text(hjust=0.5))

lcplot <- train %>%
  select(-LabelAppeal, -STARS, -INDEX, -AcidIndex) %>%
  gather(key="key", value="value",-TARGET) %>%
  ggplot(aes(value, TARGET)) +
  facet_wrap(~ key, scales = "free") +
  theme_pander()  +
  theme(axis.text.y = element_text(size=7),
        strip.text.x = element_text(size= 9),
        axis.text.x = element_text(size=6),
        plot.title = element_text(hjust = 0.5, size=10),
        axis.title.y = element_text(size=8)) +
  labs(x=NULL, title="Revised Loess Curves of TARGET vs. Continuous Predictors") +
  geom_smooth(method="loess", color="black")

lcplot

# bin transformation to AcidIndex
train$AcidIndex <- ifelse(train$AcidIndex < 6,6,
                          ifelse(train$AcidIndex > 10,
                                 10,train$AcidIndex))

train$AcidIndex <- ordered(train$AcidIndex, levels = seq(6,10))

# convert other categories back to factors
train$STARS <- ifelse(is.na(train$STARS),"NA",train$STARS)
```

```r
train$STARS <- ordered(train$STARS, levels = c("NA","1","2","3","4"))
train$LabelAppeal <- ordered(train$LabelAppeal, levels = c(-2,-1,0,1,2))
# model 1; linear model
train$LabelAppeal <- factor(train$LabelAppeal, ordered=F)
train$STARS <- factor(train$STARS, ordered=F)
train$AcidIndex <- factor(train$AcidIndex, ordered=F)

mod1 <- lm(TARGET ~ AcidIndex + STARS + LabelAppeal + Alcohol, data=train)
apa.mod1 <- apa_print(mod1)
apa_table(apa.mod1$table, caption="Model 1 Multiple Regression Output")
# model 2: multiple linear regression
model.upper <- lm(TARGET ~ ., data=train)
model.null <- lm(TARGET ~ 1, data=train)
mod2 <- step(model.null,scope=list(upper=model.upper, lower=model.null),
           trace= 0, direction='both')
apa.mod2 <- apa_print(mod2)
apa_table(apa.mod2$table, caption="Model 2 Multiple Regression Output")

mod2.1 <- update(mod2, . ~ . -pH -CitricAcid -Sulphates)
apa.mod2.1 <- apa_print(mod2.1)
apa_table(apa.mod2.1$table, caption="Model 2.1 Multiple Regression Output")
# anova
anovtab1 <- anova(mod2.1, mod2)
apa_table(data.frame(anovtab1), caption= "ANOVA of Model 2 and Model 2.1")
mod3 <- glm(TARGET ~ AcidIndex + STARS + LabelAppeal + Alcohol,
            data=train, family="poisson")
apa.mod3 <- apa_print(mod3)
apa_table(apa.mod3$table, caption="Model 3 Poisson Regression Output")

model.upper <- glm(TARGET ~ ., data=train, family="poisson")
model.null <- glm(TARGET ~ 1, data=train, family="poisson")
mod4 <- step(model.null,scope=list(upper=model.upper, lower=model.null),
           trace= 0, direction='both')

apa.mod4 <- apa_print(mod4)
apa_table(apa.mod4$table, caption="Model 4 Poisson Regression Output")
mod4.1 <- update(mod4, . ~ . - Density - Chlorides -TotalSulfurDioxide)
apa.mod41 <- apa_print(mod4.1)
apa_table(apa.mod41$table, caption="Model 4.1 Poisson Regression Output")
# anova model 4: full and reduced
anovtab2 <- anova(mod4.1, mod4, test="Chisq")
```

```
apa_table(data.frame(anovtab2), caption= "Analysis of Deviance of Models 4 and Model 4.1")
mod5 <- zeroinfl(TARGET ~ AcidIndex + STARS + LabelAppeal + Alcohol +
                   VolatileAcidity + TotalSulfurDioxide + Chlorides + Density,
                 data=train, dist="poisson")

t <- summary(mod5)[1]

stargazer(t, title="Model 5 Zero Inflated Poisson Regression Output",
          header=FALSE)

mod6 <- update(mod5, . ~ . - Chlorides - Density - TotalSulfurDioxide - Density)
t <- summary(mod6)[1]
stargazer(t, title="Model 6 Zero Inflated Poisson Regression Output",
          header=FALSE)

mod7 <- glm.nb(TARGET ~ AcidIndex + STARS + LabelAppeal + Alcohol, data=train)
t <- summary(mod7)
stargazer(t$coefficients, title="Model 7 Negative Binomial Regression Output",
          header=FALSE)

model.upper <- glm.nb(TARGET ~ ., data=train)
model.null <- glm.nb(TARGET ~ 1, data=train)
mod8 <- step(model.null,scope=list(upper=model.upper, lower=model.null),
           trace= 0, direction='both')

t <- summary(mod8)
stargazer(t$coefficients, title="Model 8 Negative Binomial Regression Output",
          header=FALSE)
mod8.1 <- update(mod8, . ~ . -TotalSulfurDioxide - Chlorides -Density)
t <- summary(mod8.1)
stargazer(t$coefficients, title = "Model 8.1 Negative Binomial Regression Output",
          header=FALSE)
anovtab3 <- anova(mod8.1, mod8, test="Chisq")
apa_table(data.frame(anovtab3[,2:8]), caption= "Analysis of Deviance of Models 8 and Model 8.1", small=T)
mod9 <- zeroinfl(TARGET ~ AcidIndex + STARS + LabelAppeal + Alcohol +
                   VolatileAcidity + TotalSulfurDioxide + Chlorides + Density,
                 data=train, dist="negbin")
t <- summary(mod9)[1]
stargazer(t, title="Model 9 Zero Inflated Negative Binomial Regression Output",
          header=FALSE)
```

```
mod10 <- update(mod9, . ~ . -VolatileAcidity - TotalSulfurDioxide -Chlorides - Density)

t <- summary(mod10)[1]
stargazer(t, title="Model 10 Zero Inflated Negative Binomial Regression Output",
          header=FALSE)

mod1_out <- cbind(AIC=AIC(mod1), BIC = BIC(mod1), loglik=logLik(mod1))
mod2.1_out <- cbind(AIC=AIC(mod2.1),BIC = BIC(mod2.1), loglik=logLik(mod2.1))
mod3_out <- cbind(AIC=AIC(mod3),BIC = BIC(mod3), loglik=logLik(mod3))
mod4_out <- cbind(AIC=AIC(mod4), BIC = BIC(mod4), loglik=logLik(mod4))
mod5_out <- cbind(AIC=AIC(mod5), BIC = BIC(mod5), loglik=logLik(mod5))
mod6_out <- cbind(AIC=AIC(mod6), BIC = BIC(mod6), loglik=logLik(mod6))
mod7_out <- cbind(AIC=AIC(mod7), BIC = BIC(mod7), loglik=logLik(mod7))
mod8_out <- cbind(AIC=AIC(mod8), BIC = BIC(mod8), loglik=logLik(mod8))
mod9_out <- cbind(AIC=AIC(mod9), BIC = BIC(mod9), loglik=logLik(mod9))
mod10_out <- cbind(AIC=AIC(mod10), BIC = BIC(mod10), loglik=logLik(mod10))

model_comp <- rbind(mod1_out, mod2.1_out,mod3_out,mod4_out,
                          mod5_out,mod6_out,mod7_out,mod8_out,
                          mod9_out, mod10_out)



rownames(model_comp) <- c("mod1_out","mod2.1_out","mod3_out","mod4_out",
                          "mod5_out","mod6_out","mod7_out","mod8_out",
                          "mod9_out", "mod10_out")



apa_table(model_comp, caption="Compare Models")
# read in test data
test_url <- 'https://raw.githubusercontent.com/spitakiss/Data621/master/Homework5/wine-evaluation-data.csv'
test <- read.csv(test_url)
names(test)[1] <- "INDEX"

# absolute value transformations to test predictors with negative vals
test$FixedAcidity <- abs(test$FixedAcidity)
test$VolatileAcidity <- abs(test$VolatileAcidity)
test$CitricAcid <- abs(test$CitricAcid)
test$ResidualSugar <- abs(test$ResidualSugar)
test$Chlorides <- abs(test$Chlorides)
test$FreeSulfurDioxide <- abs(test$FreeSulfurDioxide)
test$TotalSulfurDioxide <- abs(test$TotalSulfurDioxide)
```

```r
test$Sulphates <- abs(test$Sulphates)
test$Alcohol <- abs(test$Alcohol)

test$STARS <- ifelse(is.na(test$STARS),"NA",test$STARS)
test$STARS <- ordered(test$STARS, levels = c("NA","1","2","3","4"))

# fix missing values with mice
tmp_data <- mice(test,maxit=3, method='pmm',seed=20, print=F)
test <- complete(tmp_data,1)

# bin transformation to AcidIndex
test$AcidIndex <- ifelse(test$AcidIndex < 6,6,
                          ifelse(test$AcidIndex > 10,
                                 10,test$AcidIndex))

test$AcidIndex <- ordered(test$AcidIndex, levels = seq(6,10))

# convert other categories back to factors
test$LabelAppeal <- ordered(test$LabelAppeal, levels = c(-2,-1,0,1,2))
test$LabelAppeal <- factor(test$LabelAppeal, ordered=F)
test$STARS <- factor(test$STARS, ordered=F)
test$AcidIndex <- factor(test$AcidIndex, ordered=F)

mypreds <- predict(mod9,test)

test$TARGET <- mypreds

#write.csv(test, "test_data_with_predictions.csv")
# compare disitributions of training target and predicted test target
fnlcmp <- rbind(stats(data.frame(train$TARGET)), stats(data.frame(test$TARGET)))
apa_table(fnlcmp, caption="Compare Training and Test TARGET Variable")
r_refs(file = "r-references.bib")
```

Appendix B

*

Table captions

Table 25. Compare Models

Table 26. Compare Training and Test TARGET Variable

Appendix C

*

Figure captions

Figure 1. Histogram of Wine Training Data Variables

Figure 2. Boxplots of TARGET vs. Categorical Variables

Figure 3. Loess Curves of TARGET vs. Continuous Variables

Figure 4. Correlation of Wine Data Variables

Figure 5. Plot of Missing Observations in Wine Training Data

Figure 6. Graphical Summary of MICE Imputation Procedure

Figure 7. Revised Histogram of Wine Training Data Variables

Figure 8. Revised Correlation of Wine Data Variables

Figure 9. Revised Loess Curves of TARGET vs. Continuous Variables