Predicting Red Wine Preferences Based on Physicochemical Properties

Joyce Nguyen

December 20th, 2022

Introduction

The wine industry has been investing in multiple technologies to improve their wine making process. But wine classification is a difficult task because of the complex savoring process. This brings up a challenge for wine producers to determine which type of wine to target their consumers.

From customers' point of view, although a lot can be said from the extrinsic elements of wine such (Muller & Szolnoki, 2010), scientific reports are consistently showing that customers are into making decisions based on taste (Staub & Siegrist, 2022) as it is a useful indicator even when the consumer is not wine-savvy. By focusing on the intrinsic side of wine, Cortez et al. (2009) had analyzed the quality of wine based on its physicochemical properties, but the fuzzy techniques that were used were not rigorous enough as they were based on a lot of assumptions and might not be widely accepted.

In order to facilitate that, this report aims to investigate the linear relationship between consumers' red wine preferences and different physiocochemical properties. It will result in a linear regression model that wine producers can use to predict the trend in customers' preferences to improvise their wine-making and selling strategies.

Methods

The *Wine Quality Data Set* used for our analysis was collected from the UCI machine learning respiratory, containing concentration information of 11 physicochemical properties in Portuguese "Vinho Verde" red wine variants. Using the R and statistical methods, we will be able to pick out the best fit linear regression model to predict the dependent variable Quality.

Linear Regression Model is a linear approach for modeling the relationship between a response and independent variables, along with some random deviations. By understanding the underlying relationships, we will be able to use the model to predict the outcome.

Variable Selection

Firstly, we will randomly sample and split the data into training and testing datasets so we can validate the accuracy of the model later on. To make sure that a linear model is a good choice, we will go through preliminary and formal checks for assumptions and additional conditions of a linear model. If these visualizations confirm the violations, it is necessary to perform appropriate data transformations and refit the model. We can consider transforming both predictors and outcome simultaneously or individually, depending on where the patterns of violations exist.

After that, by using t-tests, partial F-tests and multicollinearity, it should be possible to determine which predictors are statistically significant and which ones can be removed from the model. To be more specific, we will base our judgment of removal on p-values in t-tests and partial F-tests, and variance inflation factors

in multicollinearity. Keeping in mind that after each reducing step, it is crucial to take the new fitted model and check for the assumptions and p-values all over again.

Finally, problematic points should also be identified and acknowledged.

Model Validation

The goal of building this linear regression model is to make predictions about data so it is necessary to validate the accuracy of the model. In order to do that, after building the model with the training set, we will apply the model onto the testing one to compare and validate whether the final model works properly. During this process, we should go through the whole process of checking for the coefficients, statistical significance of predictors, assumptions, multicollinearity, problematic points, and finally compare the results with the training dataset.

Model Violations and Diagnosis

After each time fitting a model, we should go through some histograms and scatterplots to preliminarily check for whether the assumptions and additional conditions of a linear model hold. If the graphs suggest any signs of violations, we will continue with the checking process, but this time using the fitted model to formally approach the assumptions. There are three types of scatterplots to make: one between fitted values and residuals, between predictors and residuals, and a normal Q-Q plot, which will help us in identifying the assumptions of linearity, constant variance, and normality.

When violations exist, we can consider transforming our variables. Although simple transformations are encouraged, we can still use Power Transformations or BoxCox methods of transformation to facilitate this modification step of data.

Finally, it is important to use different methods such as Cook's and DFFITS to diagnose any problematic points.

Results

Data Description

Figure 1: Histograms of Wine Quality Dataset variables

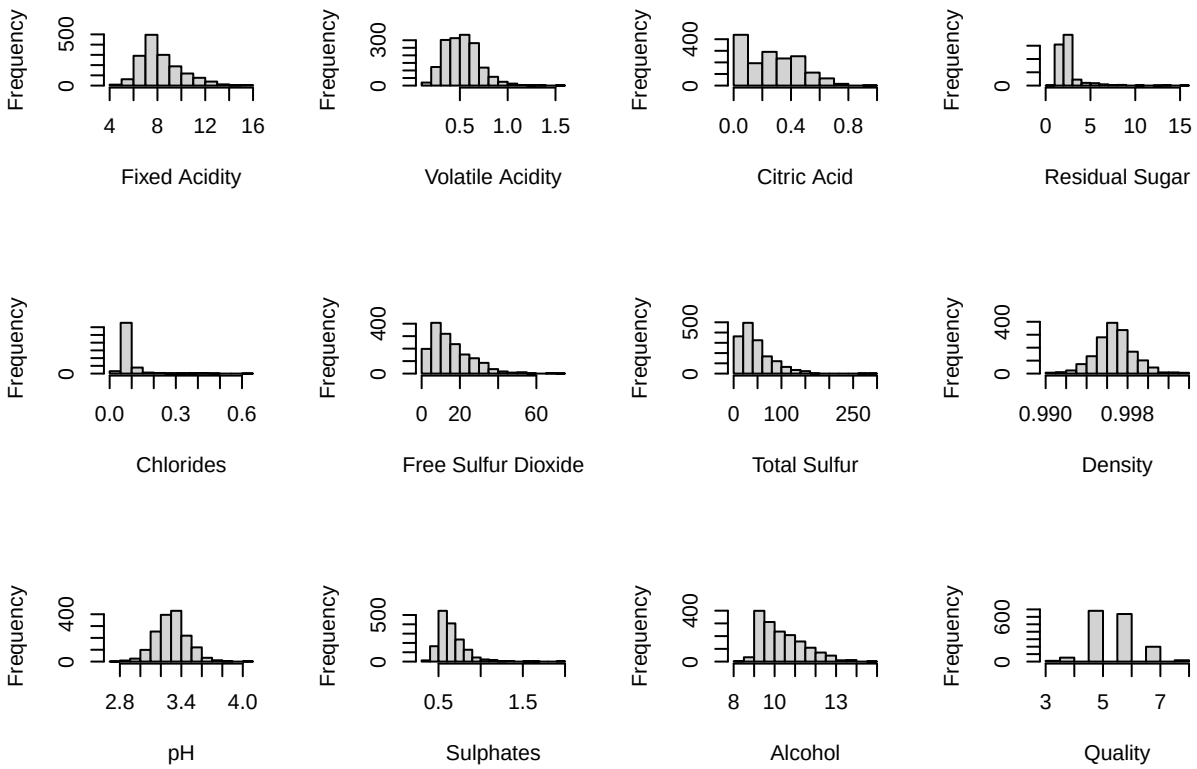
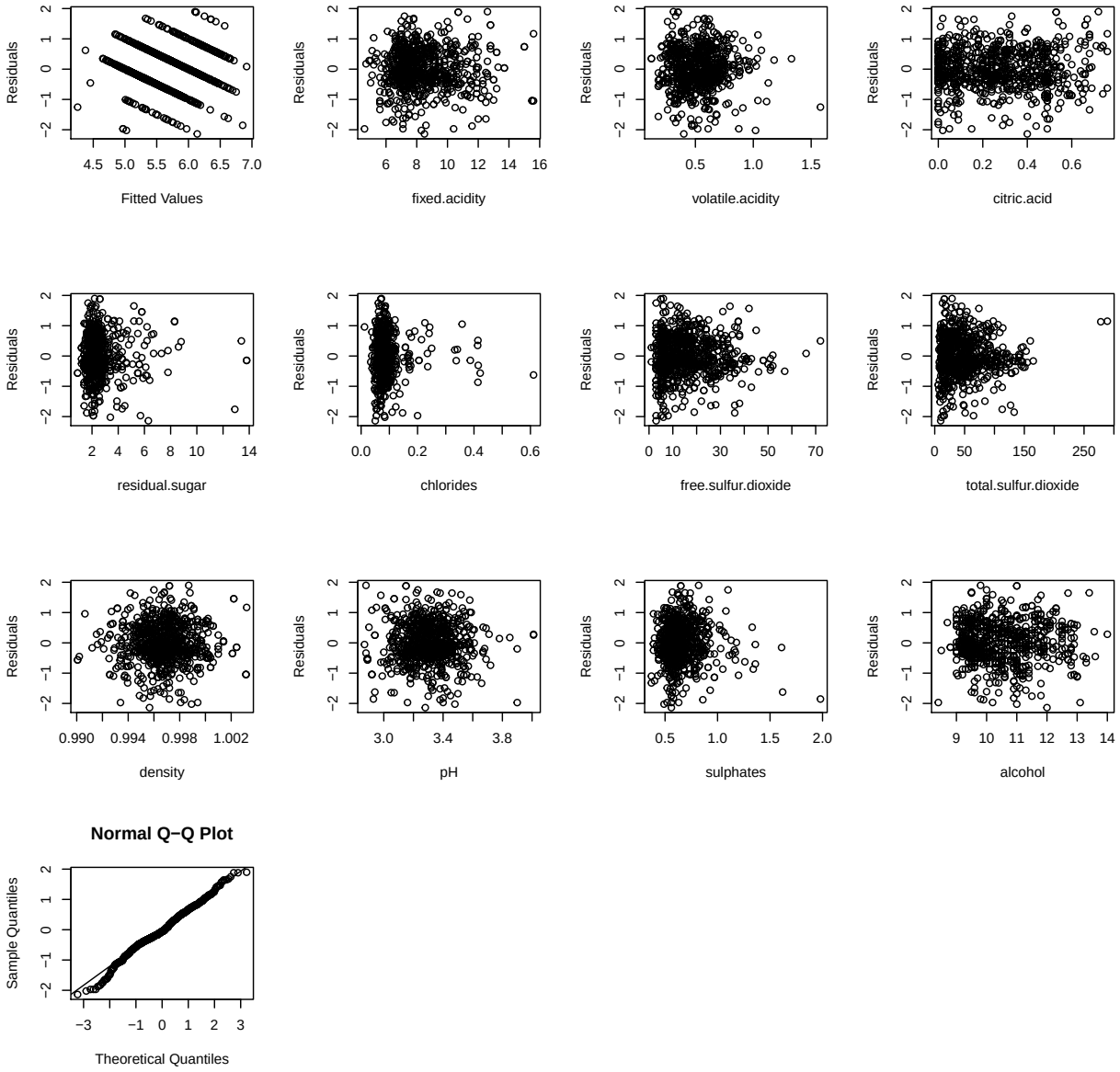


Figure 1 highlights some problems we might face with a model. Most of the predictors are skewed, showing the potential to see maybe linearity problems or just poorly fitting models. Now we formally check for assumptions.

Figure 2: Residual plots for accessing the assumptions of models



These plots show that the assumptions of linearity and normality are adequately satisfied. However, we observe one main problem: fanning of the residuals which tells us constant variance is violated. So transformations should be attempted to mitigate these problems. Since we seem to observe a pattern with only a few variables, we will only take the logarithm of those that need transformations, and re-check the assumptions to ensure they were corrected.

After transforming the data and checking again for the newly fitted model, the problem of fanning pattern in variance has significantly improved. We then split the data into training and testing datasets after randomly sampling and splitting equally, and here is the table of the characteristics of the two datasets.

Table 1: Summary of charateristics for Training (n = 800) and Testing (n = 799) data sets.

Variable	Training Set	Testing Set
Fixed Acidity	2.1 (0.2)	2.1 (0.2)
Volatile Acidity	-0.7 (0.34)	-0.7 (0.36)
Citric Acid	0.27 (0.19)	0.27 (0.2)
Residual Sugar	0.84 (0.35)	0.86 (0.36)
Chlorides	-2.5 (0.33)	-2.51 (0.33)
Free Sulfur Dioxide	2.55 (0.7)	2.55 (0.68)
Total Sulfur Dioxide	3.61 (0.72)	3.59 (0.69)
Density	1 (0)	1 (0)
pH	3.31 (0.16)	3.31 (0.15)
Sulphates	-0.46 (0.22)	-0.43 (0.23)
Alcohol	2.34 (0.1)	2.34 (0.1)
Quality	5.62 (0.8)	5.65 (0.81)

All of the variables in two datasets are numerical, summarized as means (standard deviations). From the table, it is clear that the values are quite similar in both datasets.

Analysing Process and Results

By fitting the linear model for the transformed data, we can now evaluate them and consider reducing to only having a few key indicators, i.e. using the backward elimination.

We fit a linear model for Quality that included 11 predictors respectively to 11 physicochemical properties in the dataset, but there are only 4 properties that are significantly related to the quality (p-value of t-test on slope < 0.0001). So we conducted a partial F-test to compare the linear model involving only these four predictors to the initial model. The test failed to reject the null hypothesis that all the removed predictors were not necessary, so we got our first reduced model. The remaining predictors for this model are: Fixed Acidity, Volatile Acidity, Sulphates, and Alcohol.

Now looking at multicollinearity of the full model, we see that Fixed Acidity and Density have VIF > 5. By removing these two variables in addition to the four removed ones, we got our second reduced model. The remaining predictors for this model are: Volatile Acidity, Sulphates, and Alcohol.

After removing all insignificant terms, we are now left with two potential models.

Goodness of Final Model

Here we implement the corrected AIC with the penalty term, BIC, adjusted R squared, as well as applying testing data to the models and looking at some problematic points to find out which one of the two candidate models has the best fit.

Table 2: Summary of goodness measures for models fit to Quality.

Model	Adjusted R^2	Corrected AIC	BIC
Full model	0.33	-670	-605
Reduced Model 1	0.33	-670	-638
Reduced Model 2	0.32	-666	-638

R squared values in both models do not change much from the full model ones, indicating little information was lost by removing those predictors. Although results from the BIC for both models are the same, AIC

tells us that model 2 seems to be a better fit.

Table 3: Summary of characteristics of two candidate models in the training and test datasets.

Characteristic	Model 1 (Train)	Model 1 (Test)	Model 2 (Train)	Model 2 (Test)
Largest VIF value	1.1935517	1.2438849	1.1313264	1.1497188
# Cook's D	0	0	0	0
# DFFITS	51	53	46	53
Violations	none	none	none	none
Intercept	-1.878 ± 0.66 (*)	-3.472 ± 0.671 (*)	-1.03 ± 0.573	-2.995 ± 0.555
Fixed Acidity	0.31 ± 0.121	0.159 ± 0.125	-	-
Volatile Acidity	-0.556 ± 0.074 (*)	-0.503 ± 0.071 (*)	-0.599 ± 0.072 (*)	-0.528 ± 0.068 (*)
Sulphates	0.797 ± 0.113 (*)	0.474 ± 0.107(*)	0.835 ± 0.113 (*)	0.496 ± 0.105 (*)
Alcohol	2.92 ± 0.247 (*)	3.692 ± 0.243(*)	NA ± 0.245 (*)	3.627 ± 0.238 (*)

Model 1 uses log(Fixed Acidity), log(Volatile Acidity), log(Sulphates), and log(Alcohol) as predictors, while Model 2 uses log(Volatile Acidity), log(Sulphates), and log(Alcohol) as predictors. Response is Quality in both models. Coefficients are presented as estimate ± SE (* = significant t-test at $\alpha = 0.05$).

The Cook's D measurement shows there are no observations that were identified as influential on the entire regression surface. Results from DFFITS show that Model 2 holds a better fit since we identified only 46 who influenced their own fitted values when fitting with the training set, rather than 51 in Model 1. We also proceeded to check the assumptions for each linear model, and found that there are not any major issues with them.

Considering results from both Table 2 and 3, we can conclude that Model 2 would be the best fit linear regression model for our prediction on wine quality.

Discussion

Final Model Description and Importance

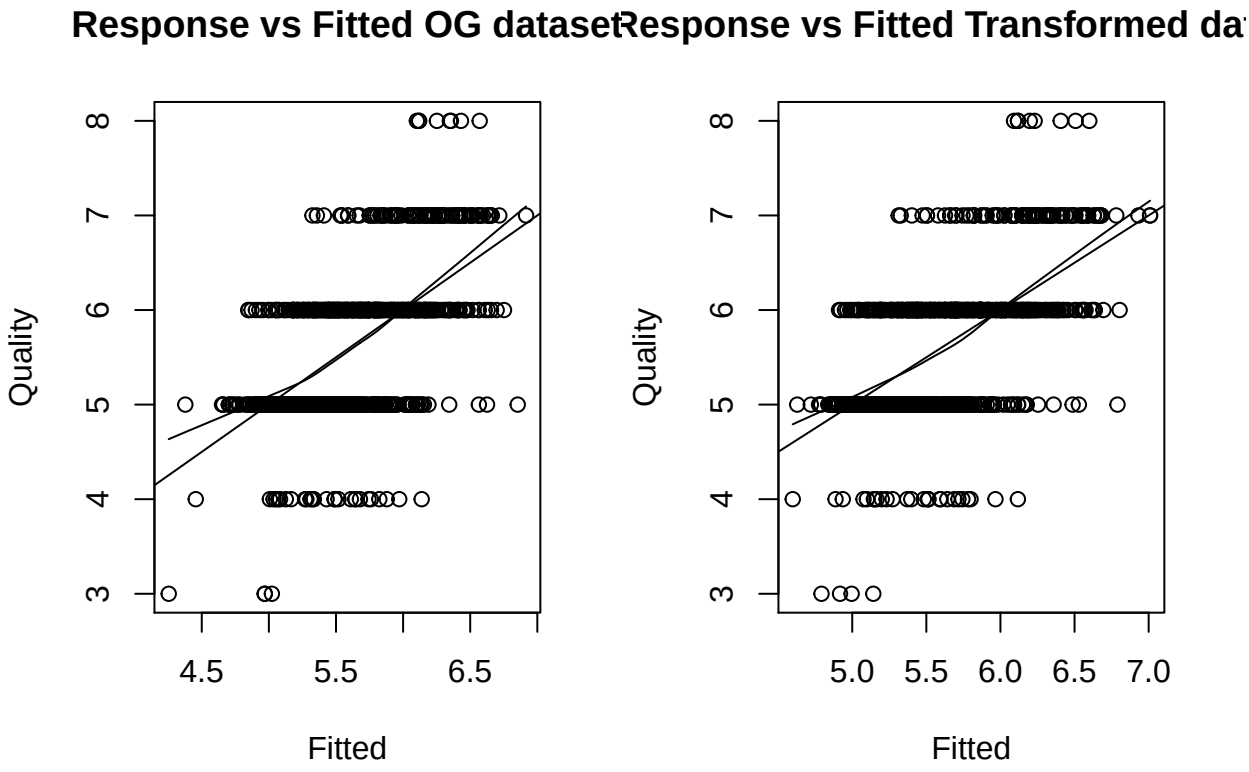
We have found the final model for our prediction on wine quality, which contains log(Volatile Acidity), log(Sulphates), and log(Alcohol) as predictors. The model adequately meets all the assumptions and additional conditions for the linear model, all predictors are statistically significant (p-value on slope of t-test < 0.0001), little information was lost by reducing the predictors from 11 to 3, and no major problems occurred with problematic points. The remaining predictors also give us a hint about how customers like their wine to be: Volatile Acidity gives a fruity-smelling, raspberry, passion fruit, or cherry-like flavors, Sulphates either offers a citrus-like smells or cooked egg-like smells depending on its concentration, with a hint of bitterness from Alcohol. These are important insights that wine producers can depend on and explore new strategies to target their customers.

Limitations

One limitation of the model is that since quality is an ordinal categorical variable, it is quite hard to see whether the linearity in the additional condition of the linear model is hold. This problem might be solved by collecting more datapoints from a larger sample. The second problem identified is that although there are no observations influencing the entire regression, there are still around 50 that influenced their own fitted values. Since it is unethical to modify data to make our model "perfect", it is important to acknowledge its existence so users of the model can keep these in mind.

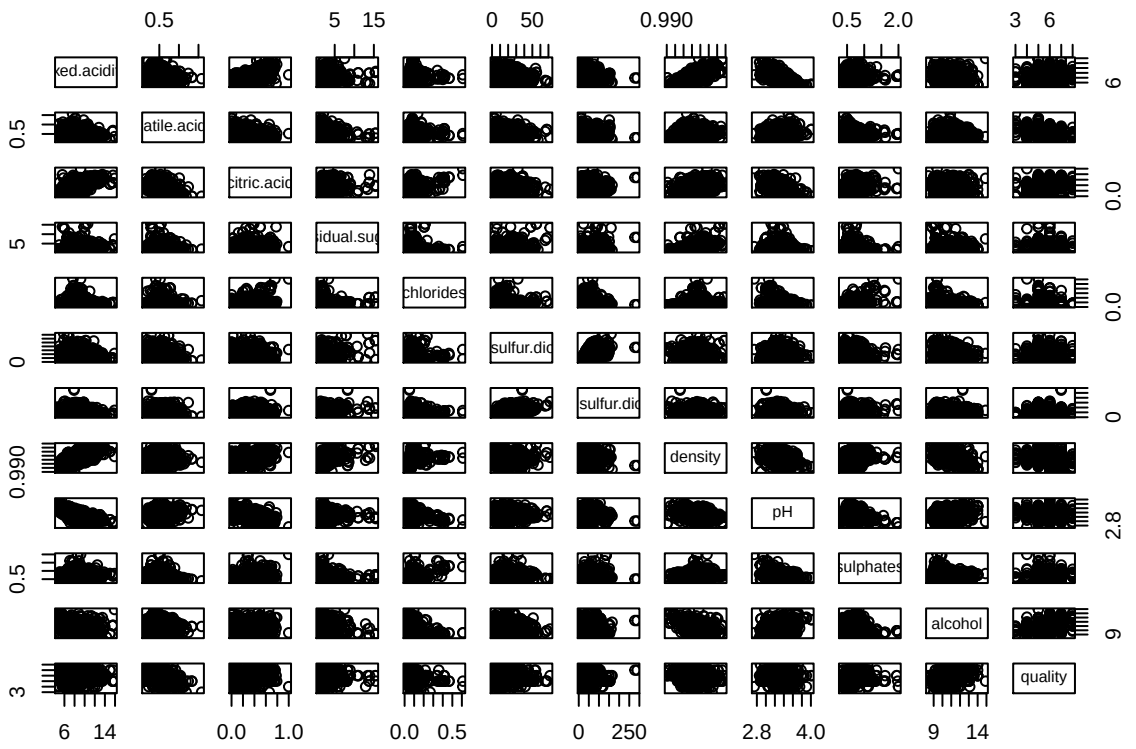
Appendix

Apendix 1: Check for additional condition 1 of original dataset and transformed one



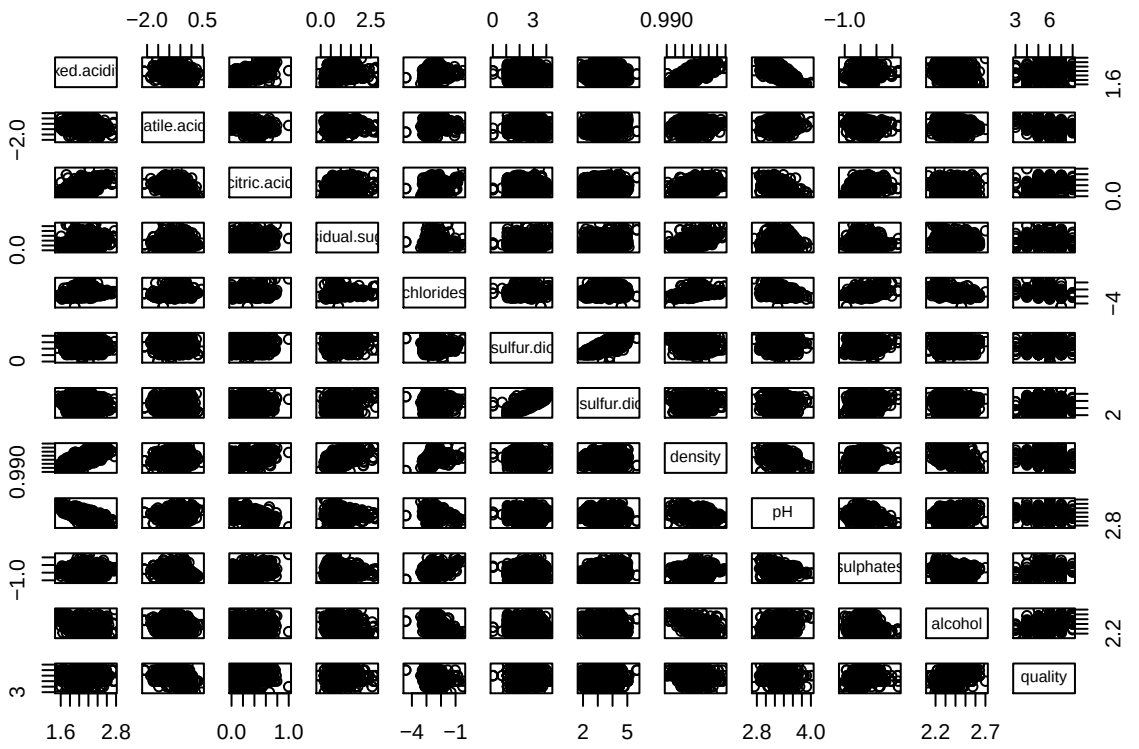
Despite the ordinal categorical nature of the variable, we can see a slight trend of linearity here in both datasets.

Appendix 2: Check for additional condition 2 of original dataset



We can tell from the original dataset that there exists some multicollinearity between variables that need to be accessed.

Appendix 3: Check for additional condition 2 of transformed dataset



After transforming the data, there still exists some multicollinearity, which indicates we should try to remove some with high VIF.

References

Cortez, P.,Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. Retrieved October 22, 2022, from https://www.scitepress.org/Papers/2015/55519/55519.pdf

Mueller and Szolnoki (2010). S. Mueller, G. Szolnoki. The relative influence of packaging, labelling, branding and sensory attributes on liking and purchase intent: Consumers differ in their responsiveness. Food Quality and Preference, 21 (7) (2010), pp. 774-783. https://doi.org/10.1016/j.foodqual.2010.07.011

Staub, C., & Siegrist, M. (2022, February 3). Rethinking the wine list: Restaurant customers' preference for listing wines according to wine style. International Journal of Wine Business Research. Retrieved October 22, 2022, from https://www.emerald.com/insight/content/doi/10.1108/IJWBR-06-2021-0034/full/html