Wine Sales

Biguzzi, Connin, Greenlee, Moscoe, Sooklall, Telab, and Wright

11/05/2021

Introduction

Using the information about sample wine orders by restaurants and wine stores after a tasting, how can we predict wine sales by various wine characteristics? Using the `wine` dataset with 12,000 entries and variables mostly related to the chemical properties of each wine, we will build a count regression model to predict the number of cases of wine that will be sold. In practice, if a wine manufacturer can predict which wines will lead to greater sales, they can choose to offer those at more tastings in restaurants and wine stores.

Using the training data set we will build:

- two different poisson regression models
- two different negative binomial regression models
- two different multiple linear regression models

In this report we will:

- explore the data
- transform the data to meet conditions of count modeling
- compare models
- select an optimal model
- generate predictions for the evaluation data set

Data Exploration

As part of our initial data exploration below, we find the following issues that will be handled in the Data Preparation section:

- large amounts of negative values for 8 of the chemical measures (which must be in error)
- 26.25% of cases have a missing `STARS` value
- `Label Appeal`, `AcidIndex`, `STARS` have imported as integers, but as could be interpreted as categorical variables due to lack of continuity
- many of the chemical variables could benefit from log transformations, after seeing the normality plots

First Look

Taking a look at the structure of the dataset we have 12,795 cases and 14 potential predictor variables. All variables are numeric. We'll remove the `INDEX` variable as it isn't needed in model building.

```
## 'data.frame':    12795 obs. of  16 variables:
##  $ ï..INDEX          : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

The table below shows the range of the TARGET (# cases purchased) ranges from 0 - 8. We also see a large amount of negative values across some of the chemical variables. We will need to deal with these values and/or cases later as it isn't possible for wine to have negative values in any of these chemical measures.

variables	min	mean	median	max	zero	minus
TARGET	0.00	3.03	3.00	8.00	2,734	0
FixedAcidity	-18.10	7.08	6.90	34.40	39	1,621
VolatileAcidity	-2.79	0.32	0.28	3.68	18	2,827
CitricAcid	-3.24	0.31	0.31	3.86	115	2,966
ResidualSugar	-127.80	5.42	3.90	141.15	6	3,136
Chlorides	-1.17	0.05	0.05	1.35	5	3,197
FreeSulfurDioxide	-555.00	30.85	30.00	623.00	11	3,036
TotalSulfurDioxide	-823.00	120.71	123.00	1,057.00	7	2,504
Density	0.89	0.99	0.99	1.10	0	0
pH	0.48	3.21	3.20	6.13	0	0
Sulphates	-3.13	0.53	0.50	4.24	22	2,361
Alcohol	-4.70	10.49	10.40	26.50	2	118
LabelAppeal	-2.00	-0.01	0.00	2.00	5,617	3,640
AcidIndex	4.00	7.77	8.00	17.00	0	0
STARS	1.00	2.04	2.00	4.00	0	0

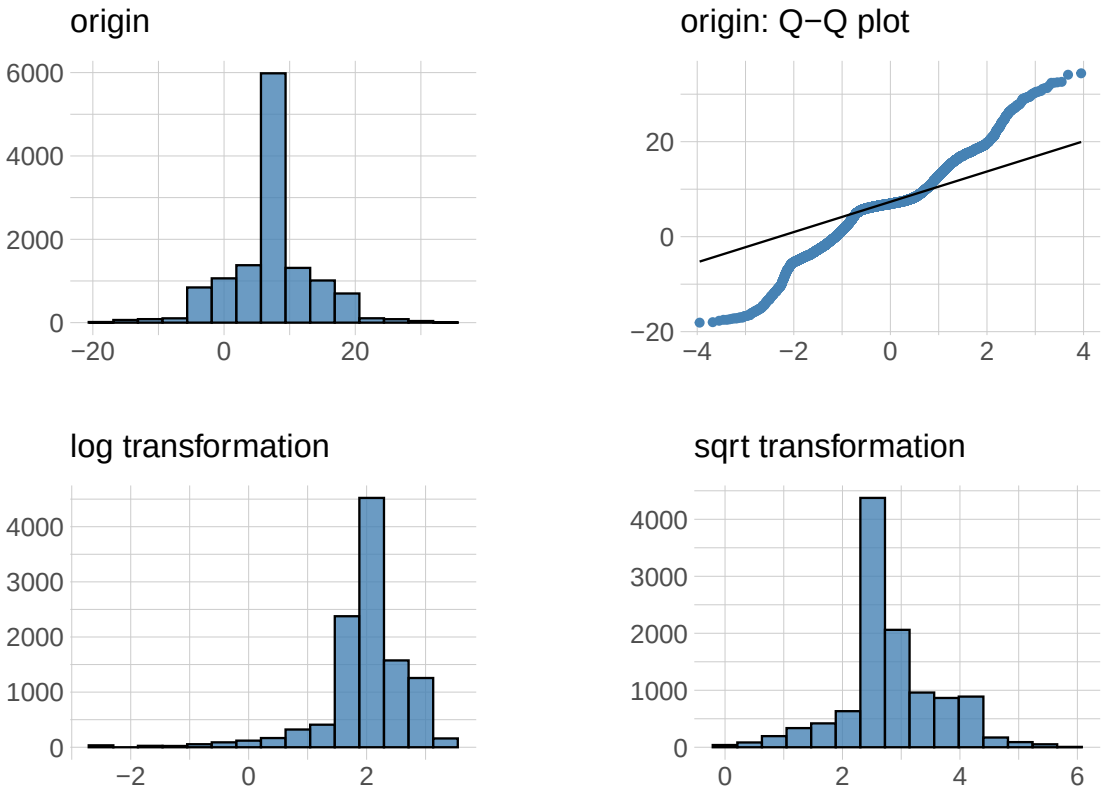
Checking for Normality

Using the Shapiro-Wilk normality test on each of the variables, we see all of the p-values are less than 0.05 which means none of our variables are normally distributed in their raw form.

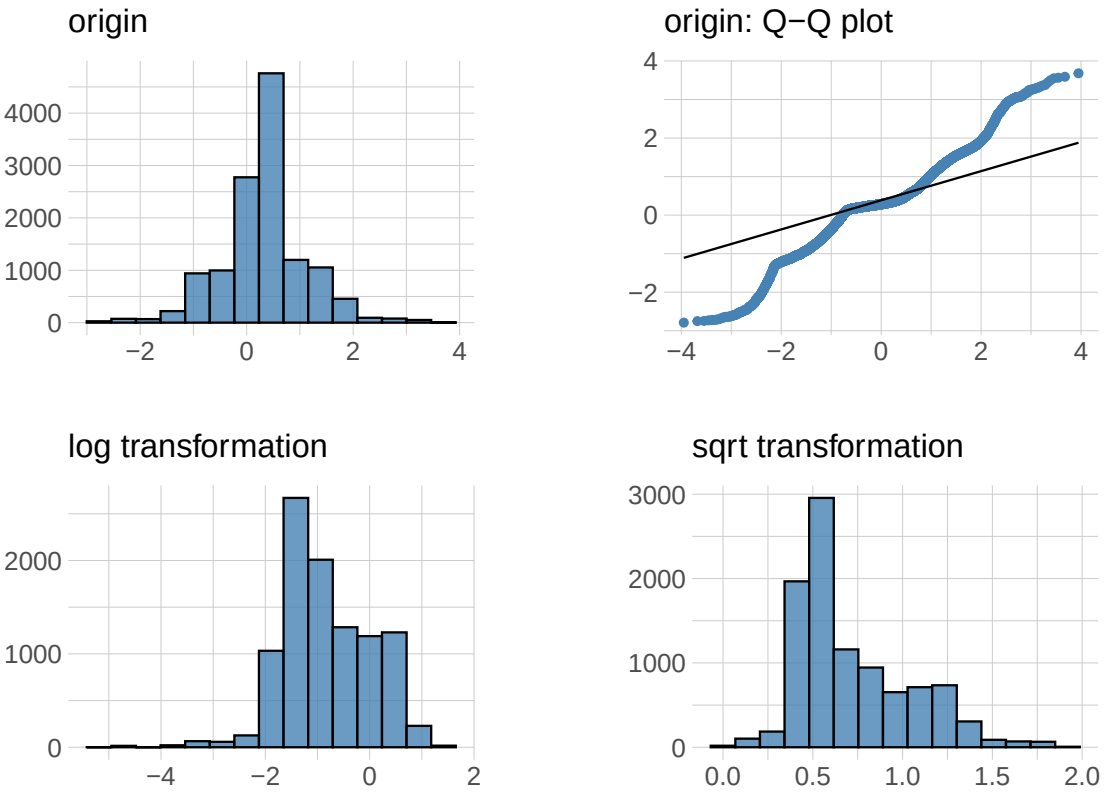
vars	statistic	p_value	sample
Alcohol	0.972	0.000	5,000.000
TotalSulfurDioxide	0.963	0.000	5,000.000
pH	0.963	0.000	5,000.000
Sulphates	0.949	0.000	5,000.000
VolatileAcidity	0.952	0.000	5,000.000
FixedAcidity	0.952	0.000	5,000.000
ResidualSugar	0.944	0.000	5,000.000
CitricAcid	0.946	0.000	5,000.000
FreeSulfurDioxide	0.932	0.000	5,000.000
Density	0.935	0.000	5,000.000
Chlorides	0.928	0.000	5,000.000
TARGET	0.903	0.000	5,000.000
LabelAppeal	0.895	0.000	5,000.000
STARS	0.851	0.000	5,000.000
AcidIndex	0.840	0.000	5,000.000

Below we visualize with a histogram and Q-Q plot for each variable for which we want to check for normality, excluding the `TARGET` variable and variables we will conver to factors later. It appears that all 11 of the variables would benefit from a log transformation, which will likely become even more necessary once we correct for the un-interpretable negative values many of these variables have in the original dataset.

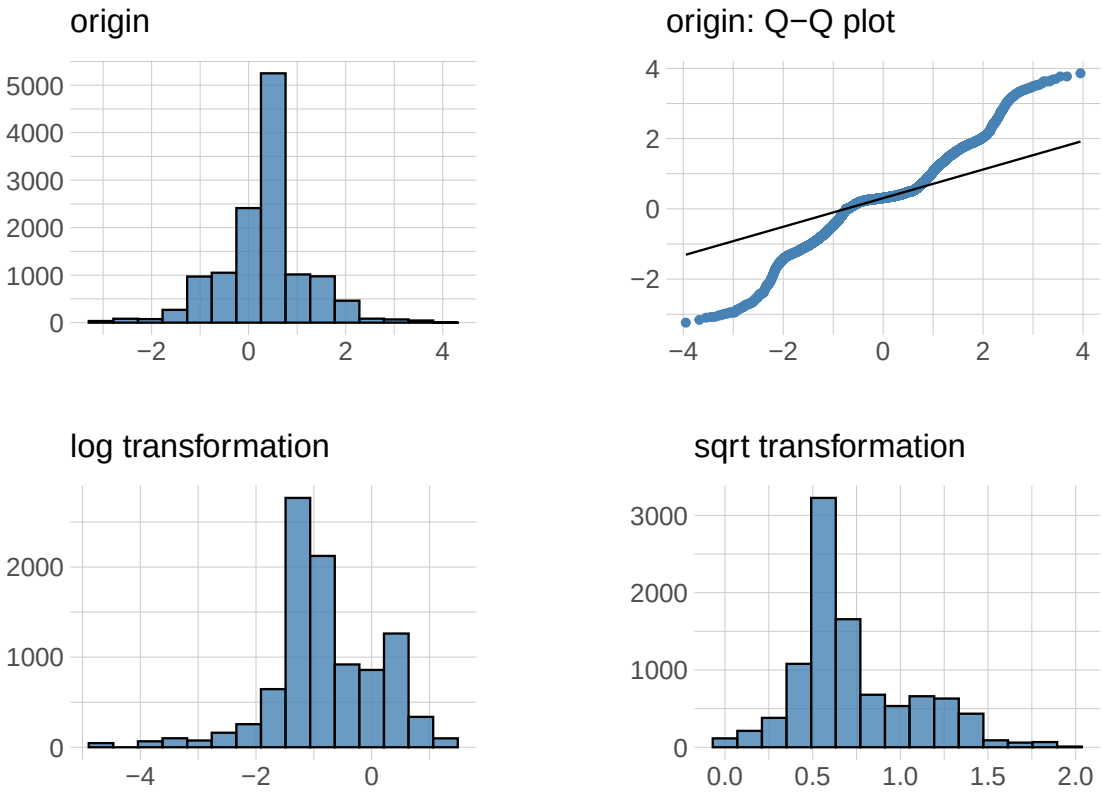
Normality Diagnosis Plot (FixedAcidity)



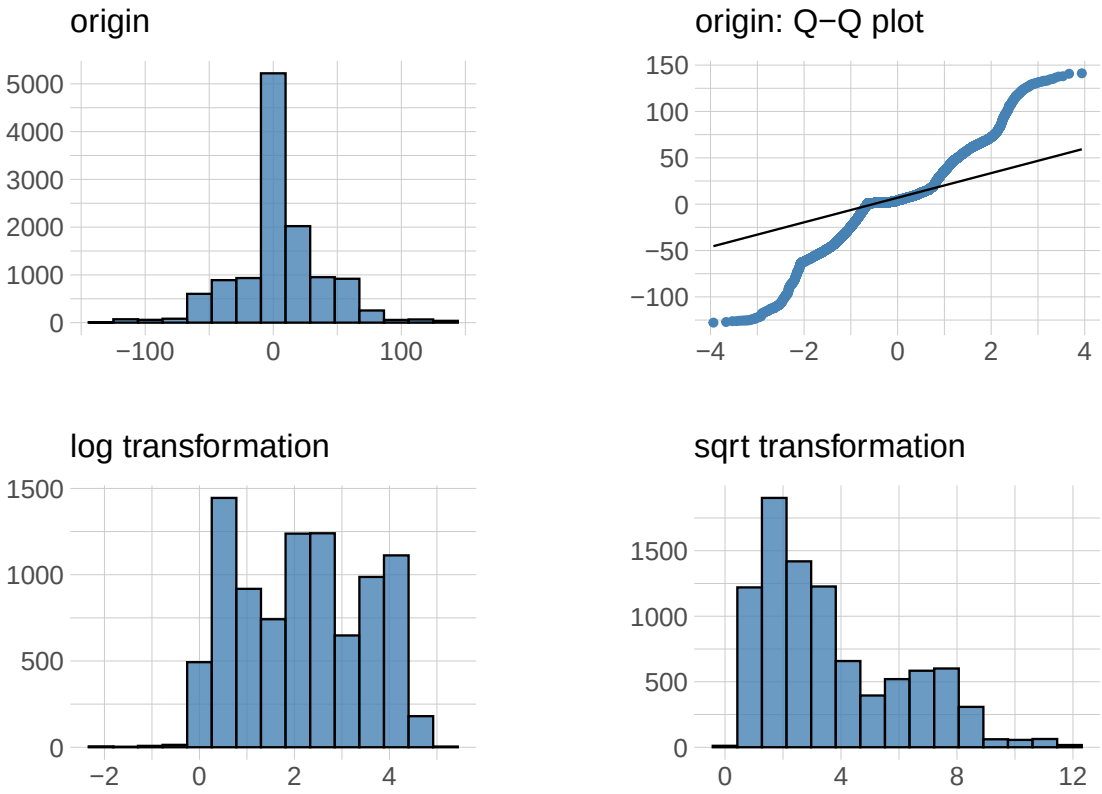
Normality Diagnosis Plot (VolatileAcidity)



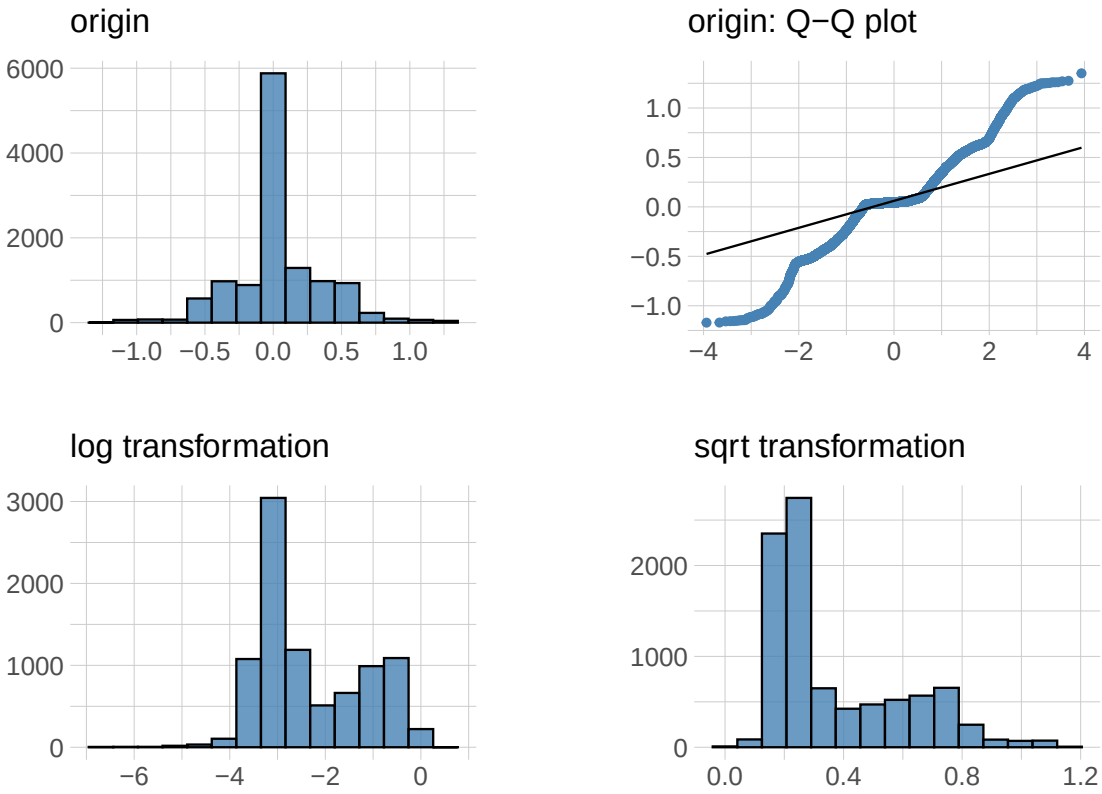
Normality Diagnosis Plot (CitricAcid)



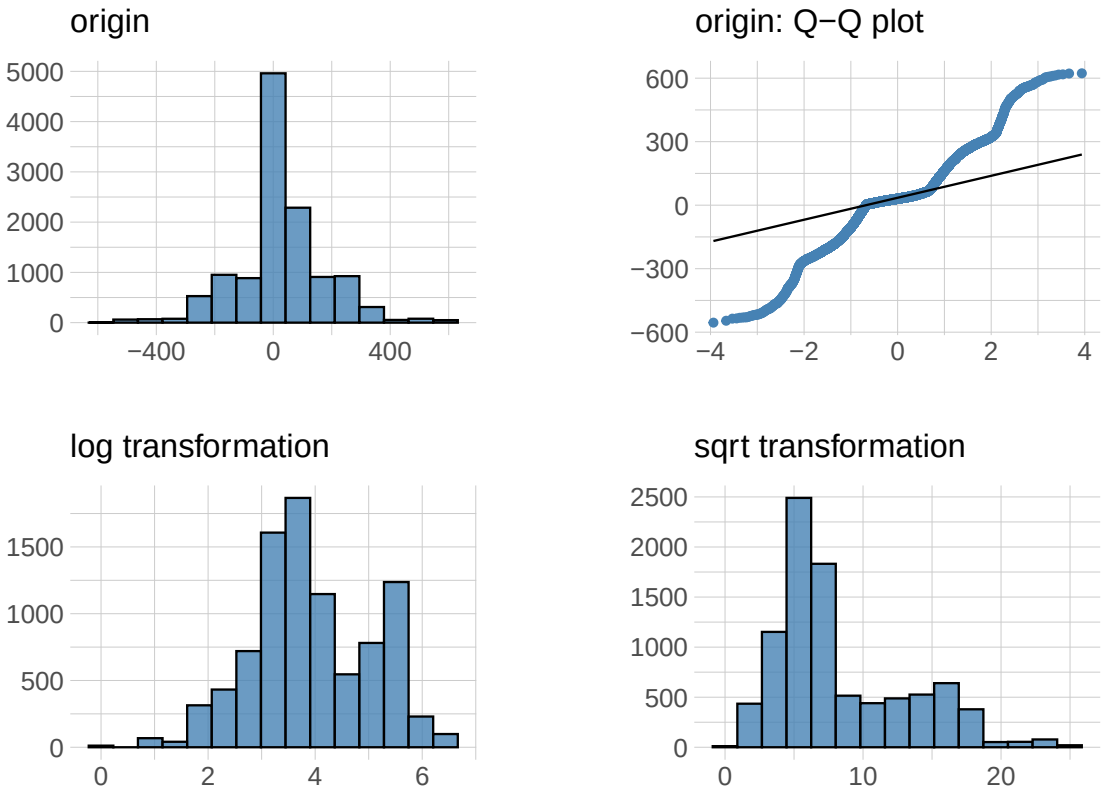
Normality Diagnosis Plot (ResidualSugar)



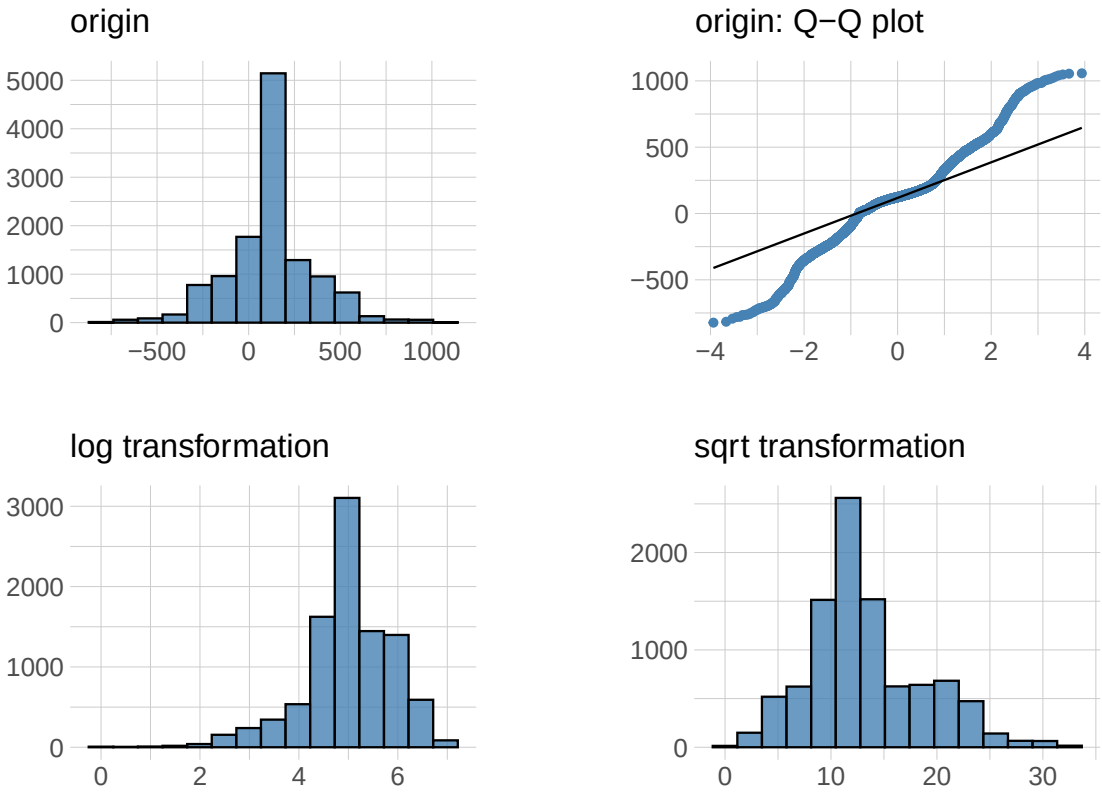
Normality Diagnosis Plot (Chlorides)



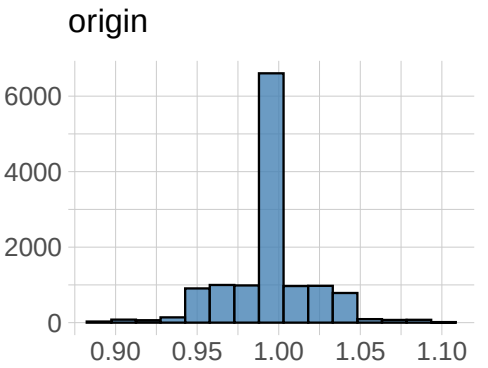
Normality Diagnosis Plot (FreeSulfurDioxide)

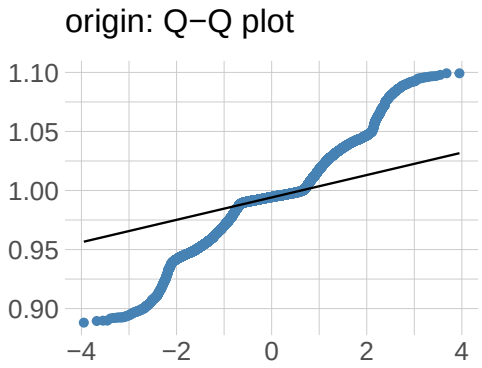


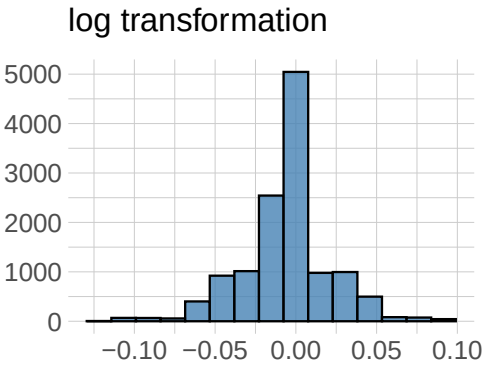
Normality Diagnosis Plot (TotalSulfurDioxide)

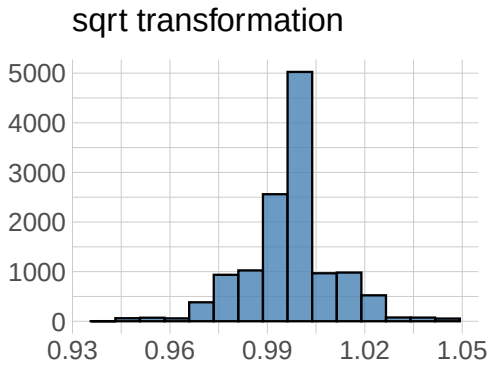


Normality Diagnosis Plot (Density)

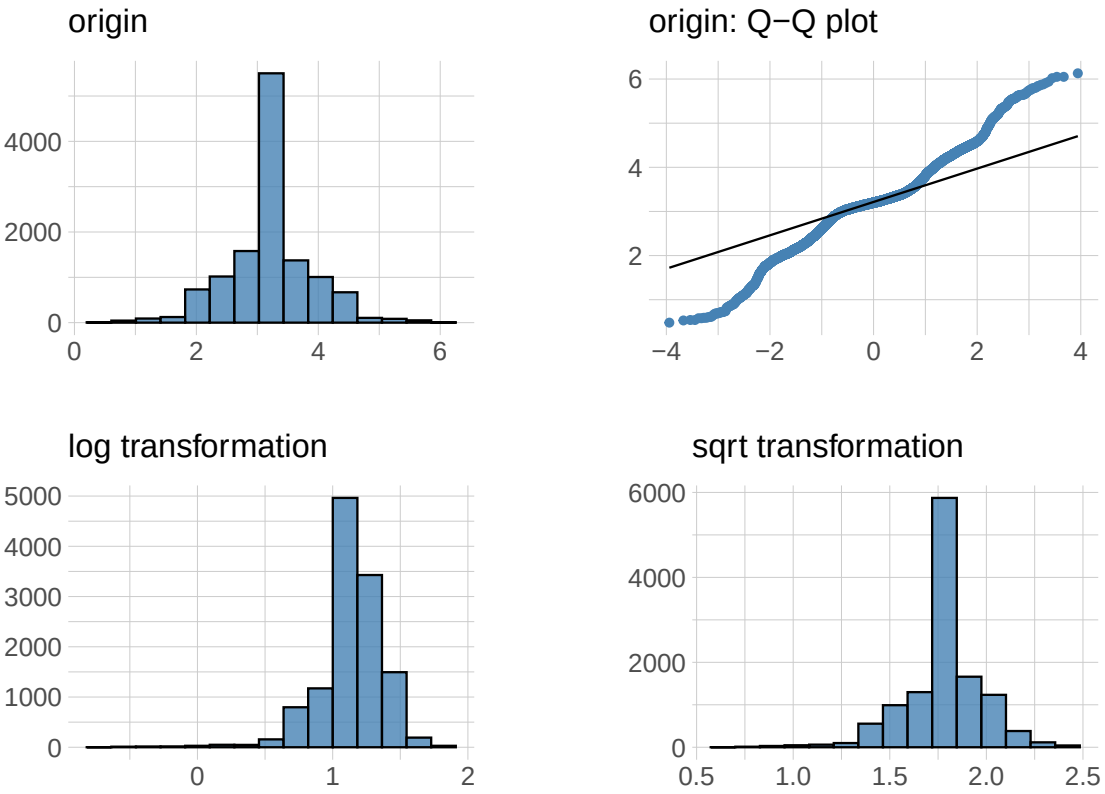




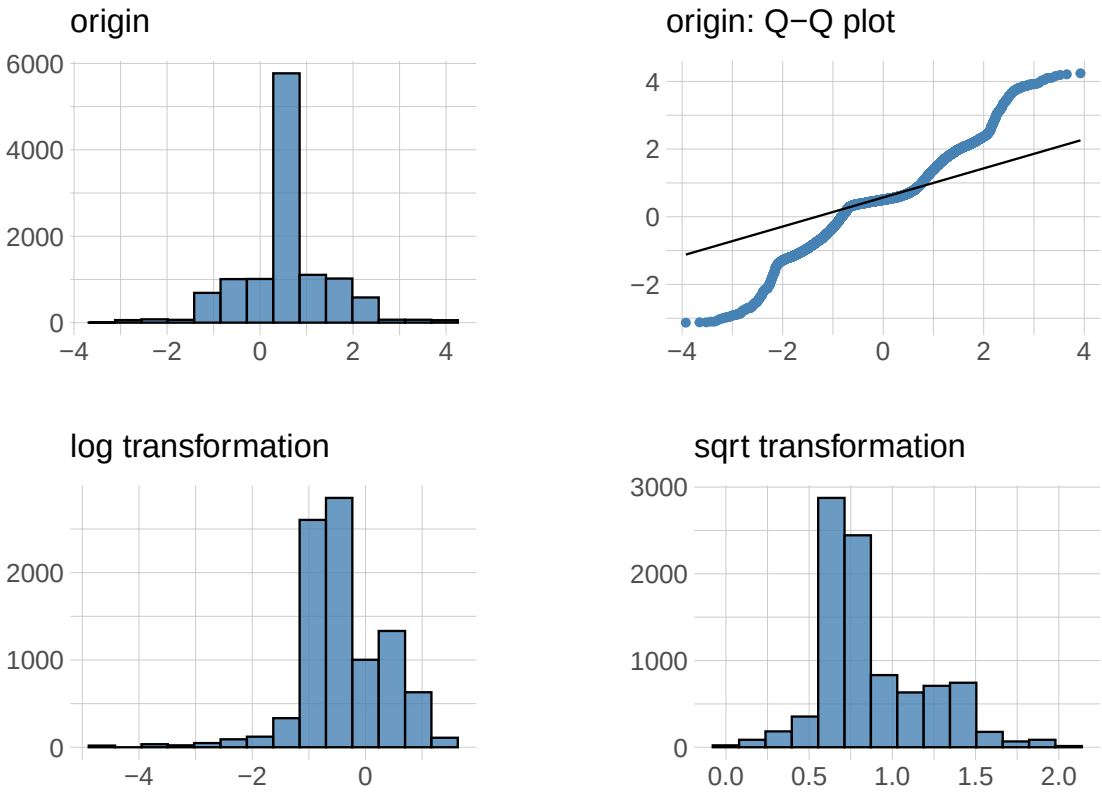


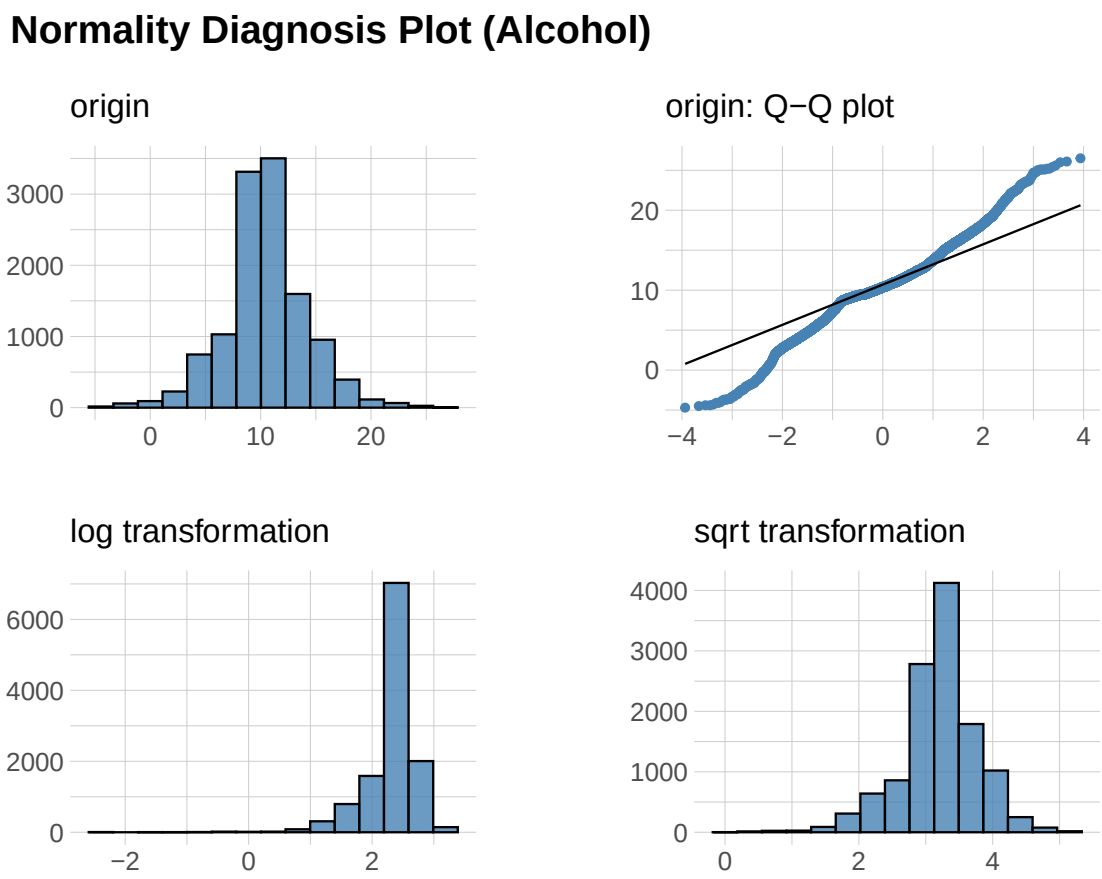


Normality Diagnosis Plot (pH)



Normality Diagnosis Plot (Sulphates)





Missingness

With regards to missingness, we have 6 variables with data for every case. This leaves 8 variables that have some degree of missing data, with the worst being the `STARS` variable with 26.25% of cases missing a value for `STARS` (the wine rating by experts). None of the remaining 7 variables have more than 10% missing, so we will not consider dropping or imputing missing values for them.

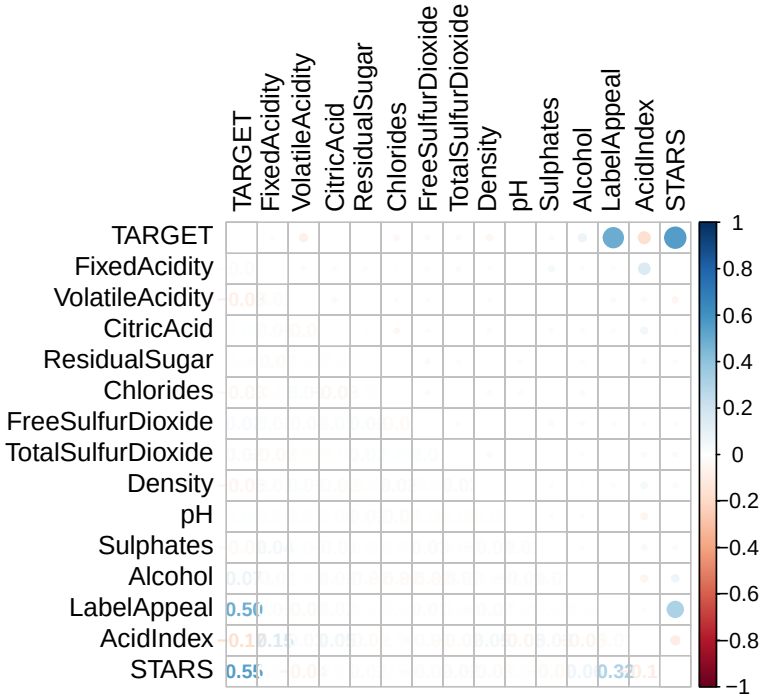
variables	types	missing_count	missing_percent
STARS	integer	3,359	26.25
Sulphates	numeric	1,210	9.46
TotalSulfurDioxide	numeric	682	5.33
Alcohol	numeric	653	5.10
FreeSulfurDioxide	numeric	647	5.06
Chlorides	numeric	638	4.99
ResidualSugar	numeric	616	4.81
pH	numeric	395	3.09

Looking for patterns in missingness can be telling. Below the 8 variables with at least some missingness are

plotted for the top 30 most frequent combinations of missing values.

We don't see any major relationships of missingness that affect a large number of cases in the dataset.

Correlation



Data Preparation

To address the issues in the data we discovered in the data exploration section, we will:

- **CHANGE DATA TYPE** to factors for `Label Appeal`, `AcidIndex`, `STARS`, interpreting as categorical variables due to lack of continuity
- **TAKE ABSOLUTE VALUE of all negative values** for 8 of the chemical measures (which must be in error)
- **IMPUTE missing values** for 26.25% of cases that have a missing `STARS` value
- **LOG transformation* for all of the numeric chemical variables that were not normally distributed

Change Some Variables to Factors

We set the 3 variables identified earlier (`STARS`, `AcidIndex`, `LabelAppeal`) as factors. While they have numeric data, it is not continuous and represents a rating.

Fix Negative Values

With 21,766 negative values across many of the chemical variables, for which it doesn't make sense to have a negative value, we convert these to positive values so we retain some value (as opposed to omitting these measures). Variables include: `FixedAcidity`, `VolatileAcidity`, `CitricAcid`, `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `Sulphates`, `Alcohol`.

Impute Missing Values

For the 7 variables that had <10% missing values, we will impute the missing values with the median for that variable. Variables include: `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `pH`, `Sulphates`, `Alcohol`.

Flag STARS Missing Values & Set to 1-Star Rating

With 26.25% of cases missing a `STARS` rating, we need to consider if we should drop the variable from modeling, impute the missing values, or flag these cases. There are some guidelines in data science that if there are beyond 30% missing values, the variable may be best dropped. Theoretically, the `STARS` variable should represent something close to our `TARGET` variable as we'd expect a relationship between high expert reviews and high case sales. Since this seems like too valuable of a variable to drop, and considering that depending on how the `STARS` data was obtained, a missing value could indicate a lesser quality or less popular wine (such that it hasn't been rated by experts), we choose to create a new variable where a '1' indicates a missing `STARS` value. Further, we fill in a '1' `STARS` rating for the NA values (so these cases will be included in the modeling).

Preform Log Transformations

The normality plots earlier identified the following variables could benefit from log transformations, which is even more true after having taken the absolute value of many of these variables (as a result of un-interpretable negative chemical values). We also add 1 to the new value in order to avoid any values of zero for which there is no defined log value.

am I doing this log transformation right? I'm not confident -Rachel

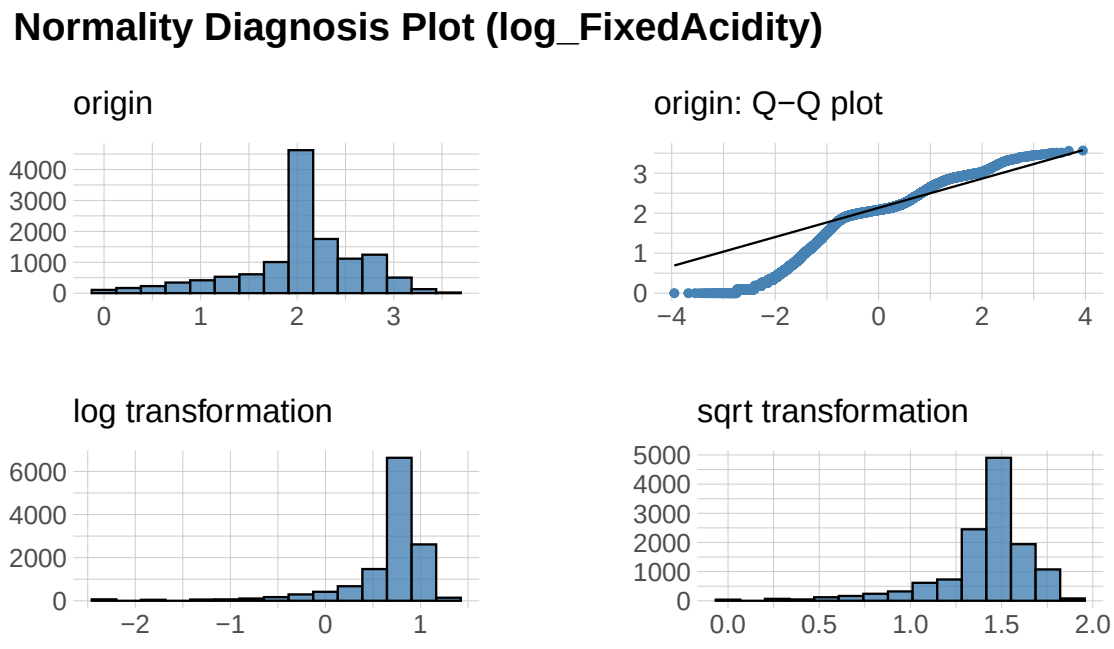
Re-Run Normality & Correlation Plots

After all of the above changes, we check the dataset below. We no longer have negative values and the log_ variables are present.

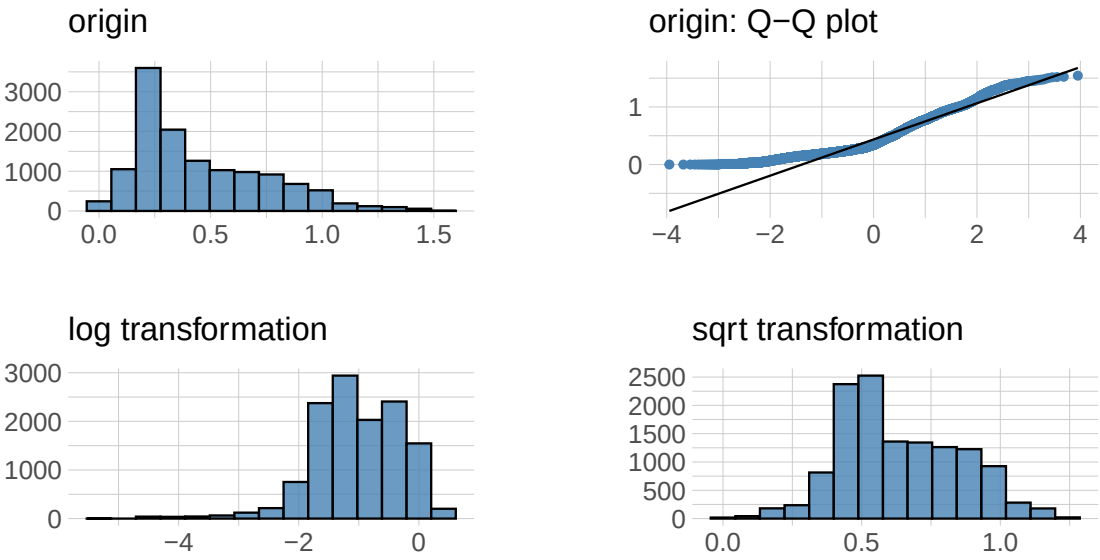
variables	min	mean	median	max	zero	minus
TARGET	0.00	3.03	3.00	8.00	2,734	0
FixedAcidity	0.00	8.06	7.00	34.40	39	0
VolatileAcidity	0.00	0.64	0.41	3.68	18	0
CitricAcid	0.00	0.69	0.44	3.86	115	0
ResidualSugar	0.00	22.86	12.90	141.15	6	0
Chlorides	0.00	0.22	0.10	1.35	5	0
FreeSulfurDioxide	0.00	104.12	56.00	623.00	11	0
TotalSulfurDioxide	0.00	5.00	5.04	6.96	7	0
Density	0.89	0.99	0.99	1.10	0	0
pH	0.48	3.21	3.20	6.13	0	0
Sulphates	0.00	0.82	0.59	4.24	22	0
Alcohol	0.00	10.52	10.40	26.50	2	0

variables	min	mean	median	max	zero	minus
log_FixedAcidity	0.00	2.04	2.08	3.57	39	0
log_VolatileAcidity	0.00	0.45	0.34	1.54	18	0
log_CitricAcid	0.00	0.47	0.36	1.58	115	0
log_ResidualSugar	0.00	2.60	2.63	4.96	6	0
log_Chlorides	0.00	0.18	0.09	0.85	5	0
log_FreeSulfurDioxide	0.00	4.14	4.04	6.44	11	0
log_Density	0.64	0.69	0.69	0.74	0	0
log_PH	0.39	1.42	1.44	1.96	0	0
log_Sulphates	0.00	0.55	0.46	1.66	22	0
log_Alcohol	0.00	2.39	2.43	3.31	2	0

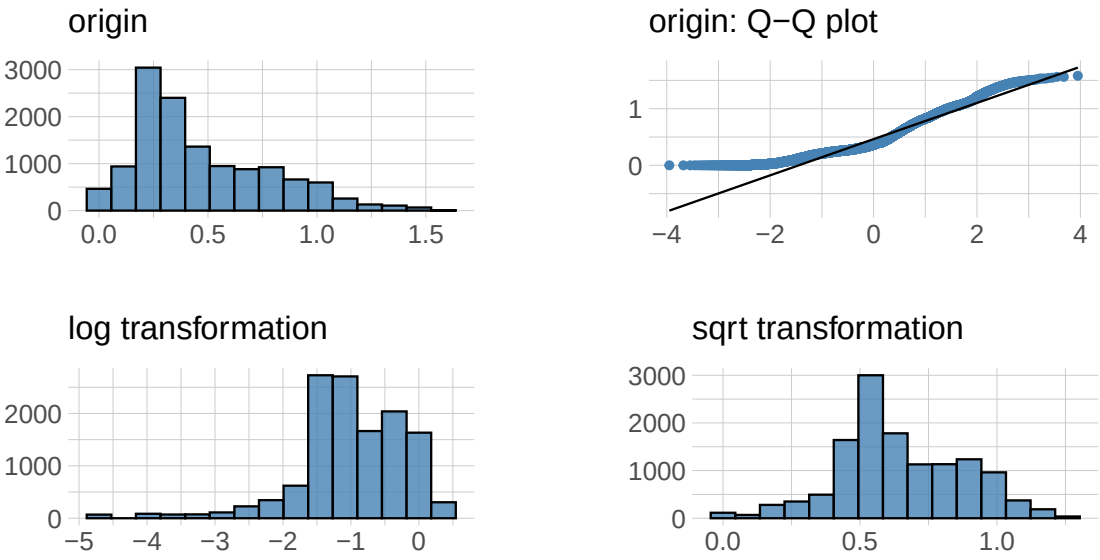
Looking at the normality plots for the 10 variables for which we converted negative values into positive values and then performed a log transformation, they are more normal than they were previously, but some are still seriously skewed.



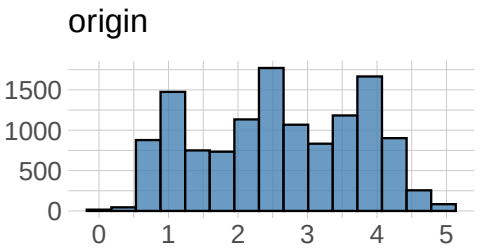
Normality Diagnosis Plot (log_VolatileAcidity)

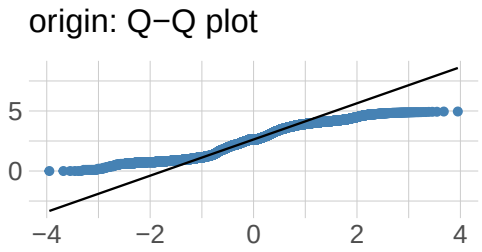


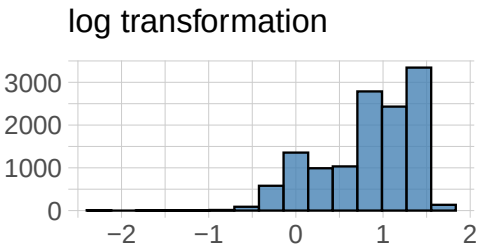
Normality Diagnosis Plot (log_CitricAcid)

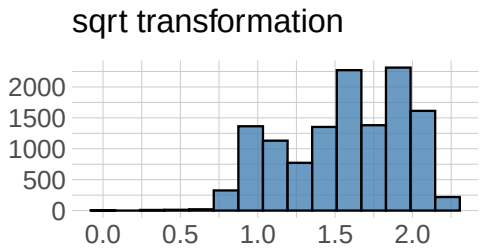


Normality Diagnosis Plot (log_ResidualSugar)

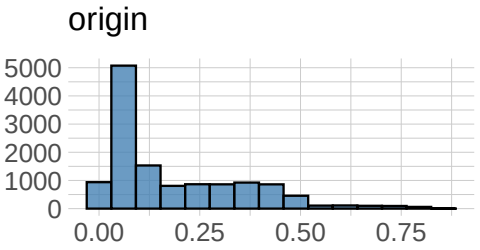


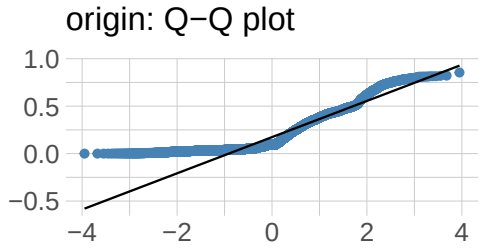


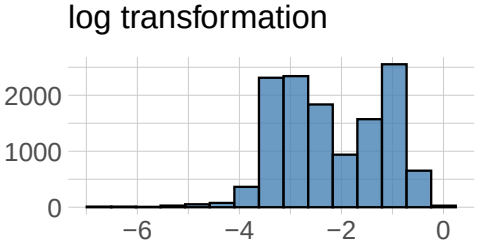


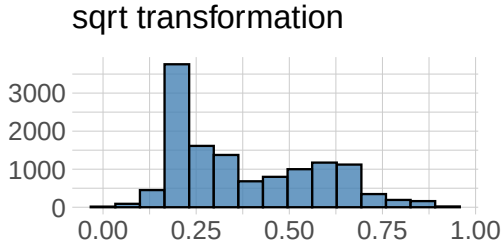


Normality Diagnosis Plot (log_Chlorides)

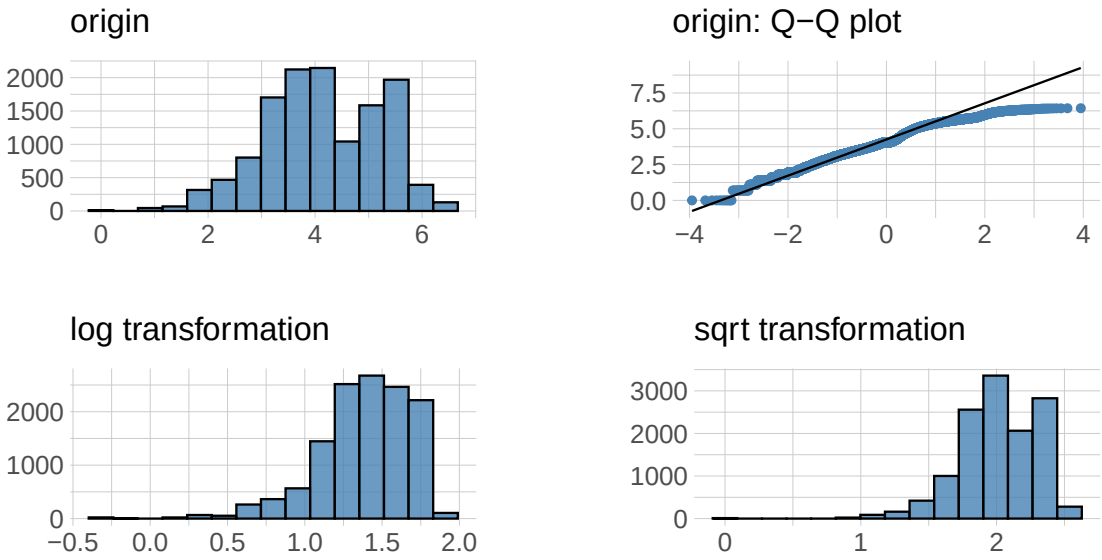




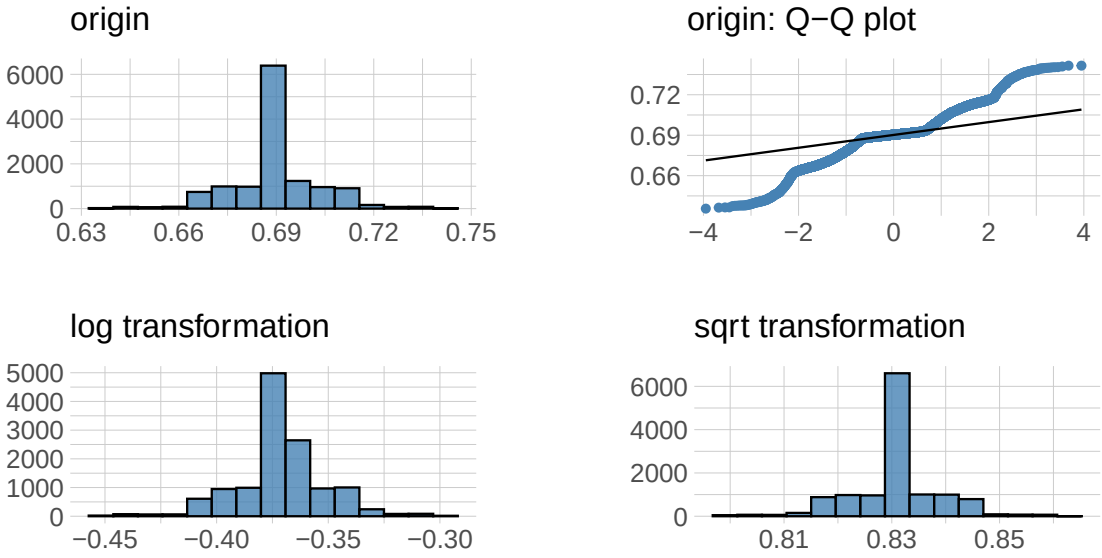




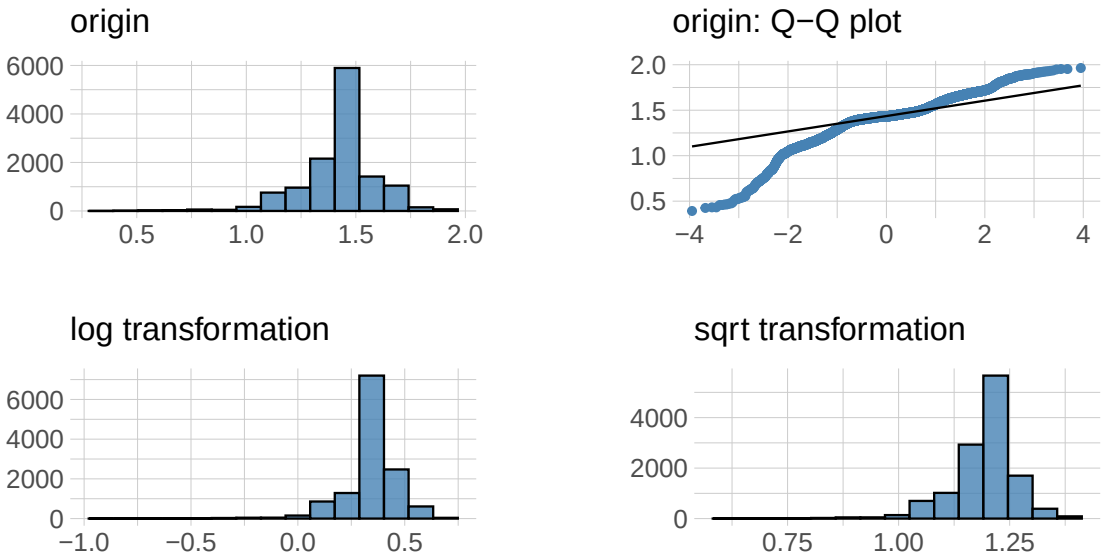
Normality Diagnosis Plot (log_FreeSulfurDioxide)



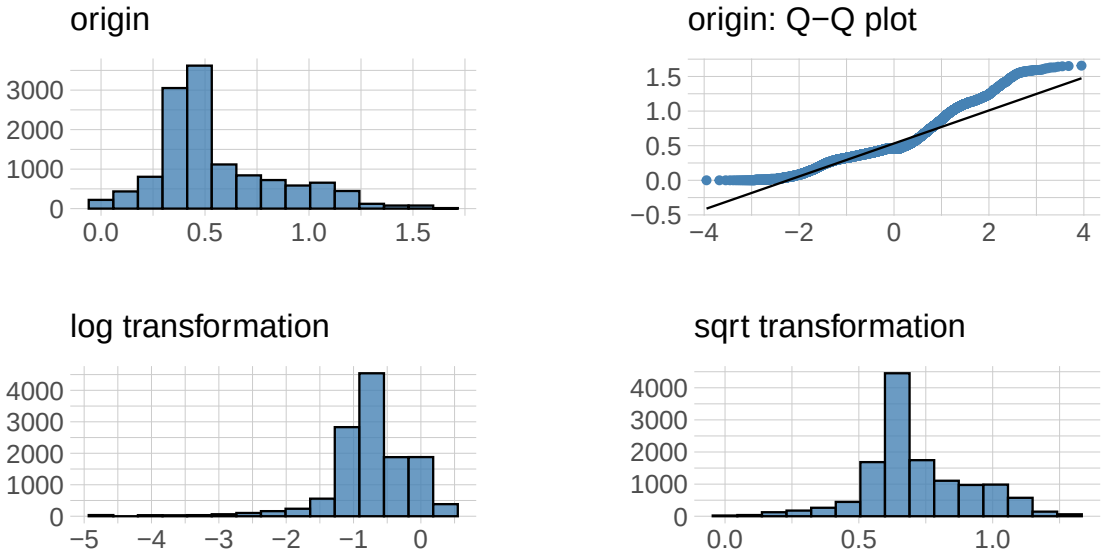
Normality Diagnosis Plot (log_Density)

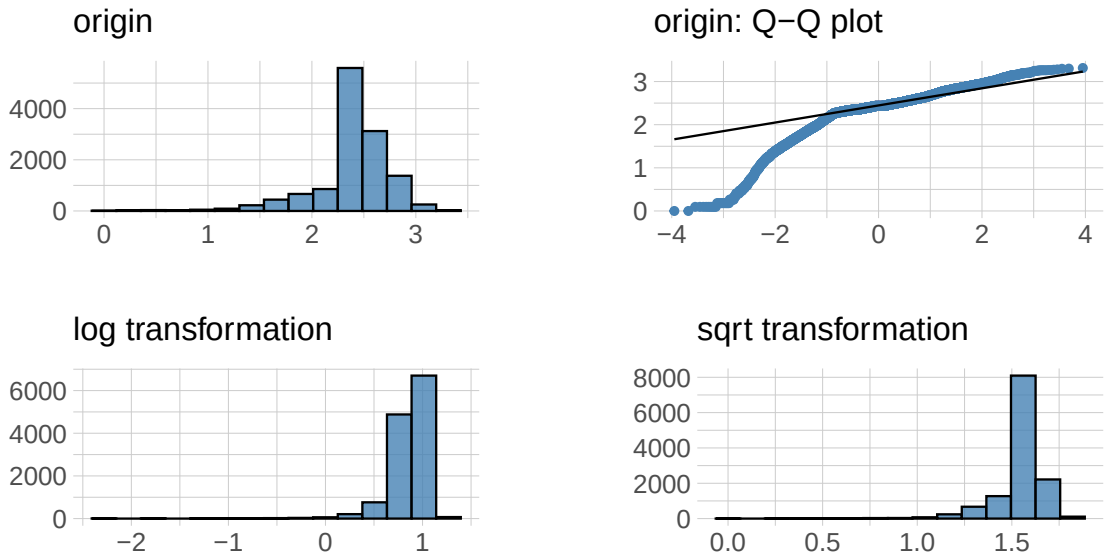


Normality Diagnosis Plot (log_PH)

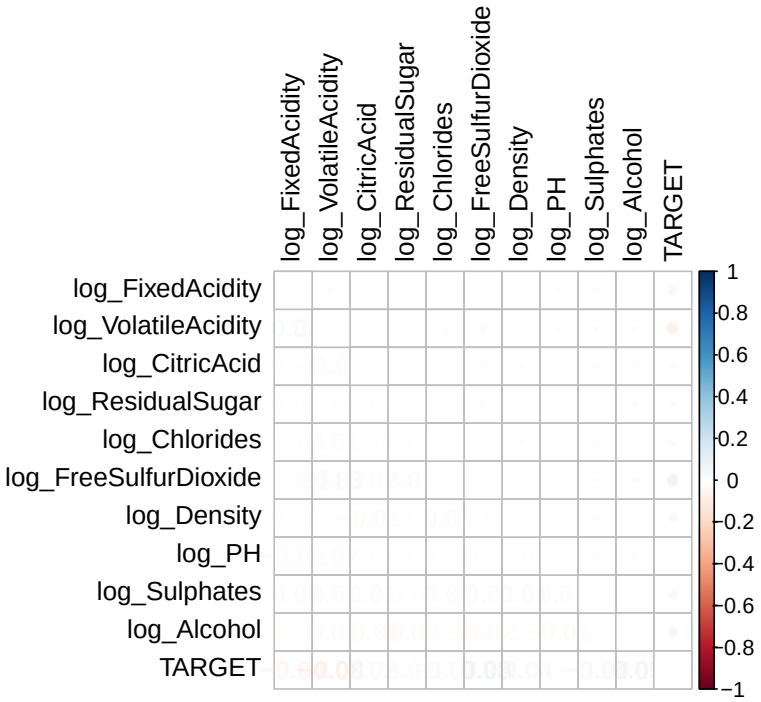


Normality Diagnosis Plot (log_Sulphates)





Creating a new correlation plot with our transformed variables it appears there is very little correlation between our `TARGET` variable and any of these chemical measures.



Build Models

Our prompt asks us to build at least two of three different models: *poisson*, *negative binomial*, and *multiple linear regression*. As there appears to be no correlation from the logged chemical variables, we won't include those in any of the following models.

The Poissons

Need to explain quasi- choice.

```
##
## Call:
## glm(formula = TARGET ~ AcidIndex + STARS + STARS_imputed + LabelAppeal,
##     family = quasipoisson, data = raw)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2236  -0.6803   0.0149   0.4358   3.7225
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.77110    0.29915   2.578  0.00996 **
## AcidIndex5     -0.14440    0.30256  -0.477  0.63318
## AcidIndex6     -0.10874    0.29744  -0.366  0.71469
## AcidIndex7     -0.14228    0.29717  -0.479  0.63209
## AcidIndex8     -0.17476    0.29719  -0.588  0.55651
## AcidIndex9     -0.28691    0.29748  -0.964  0.33482
## AcidIndex10    -0.45008    0.29850  -1.508  0.13163
## AcidIndex11    -0.81344    0.30185  -2.695  0.00705 **
## AcidIndex12    -0.82983    0.30717  -2.702  0.00691 **
## AcidIndex13    -0.66115    0.30989  -2.133  0.03290 *
## AcidIndex14    -0.76768    0.32166  -2.387  0.01702 *
## AcidIndex15    -0.31630    0.37862  -0.835  0.40351
## AcidIndex16    -0.98375    0.51431  -1.913  0.05580 .
## AcidIndex17    -1.23734    0.51442  -2.405  0.01617 *
## STARS2          0.32069    0.01347  23.809  < 2e-16 ***
## STARS3          0.44161    0.01464  30.174  < 2e-16 ***
## STARS4          0.56356    0.02029  27.769  < 2e-16 ***
## STARS_imputed1 -0.75936    0.01836 -41.359  < 2e-16 ***
## LabelAppeal-1   0.23968    0.03566   6.721 1.88e-11 ***
## LabelAppeal0    0.42933    0.03478  12.344  < 2e-16 ***
## LabelAppeal1    0.56284    0.03538  15.908  < 2e-16 ***
## LabelAppeal2    0.69834    0.03983  17.532  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.8811094)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 13591  on 12773  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
##
## Call:
## glm(formula = TARGET ~ STARS + STARS_imputed, family = quasipoisson,
##     data = raw)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.7562  -0.7729   0.1026   0.5877   4.1159
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.94800    0.01123   84.38   <2e-16 ***
## STARS2          0.38656    0.01412   27.38   <2e-16 ***
## STARS3          0.56598    0.01499   37.75   <2e-16 ***
## STARS4          0.74239    0.02061   36.02   <2e-16 ***
## STARS_imputed1 -0.77937    0.01938  -40.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.9907687)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 14691  on 12790  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

The Negative Binomials

The Multiple Linears

Model Selection