STAT 5630 - Statistical Machine Learning

due friday october 11, 2019

Homework2

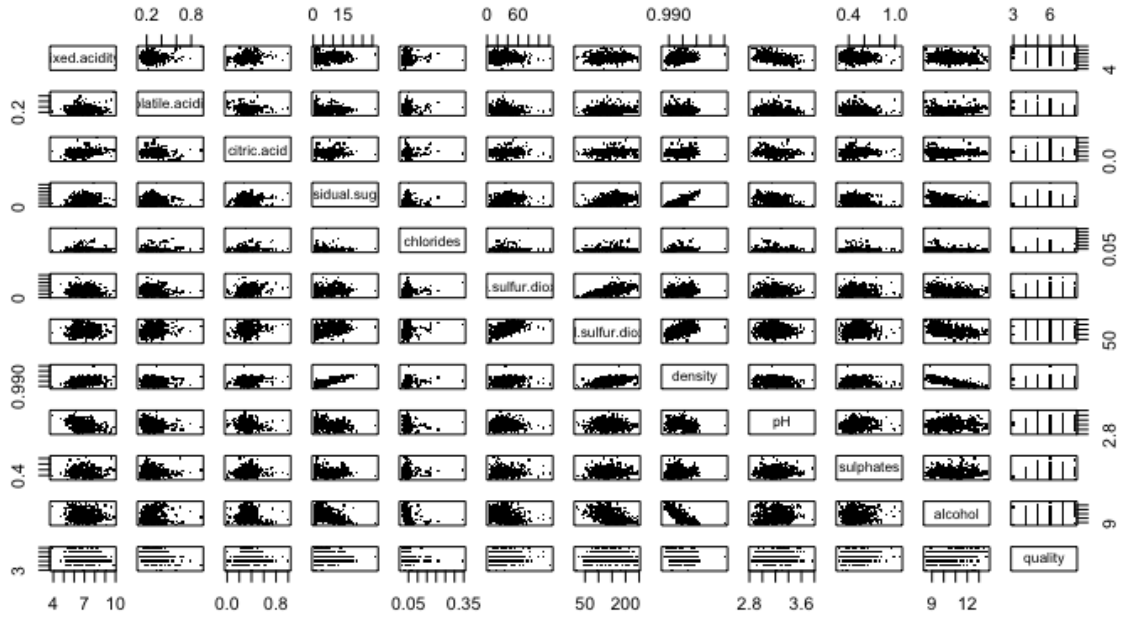
Max Ryoo (hr2ee)

Professor: Xiwei (Denny) Tang

problem 1 *Wine Quality*

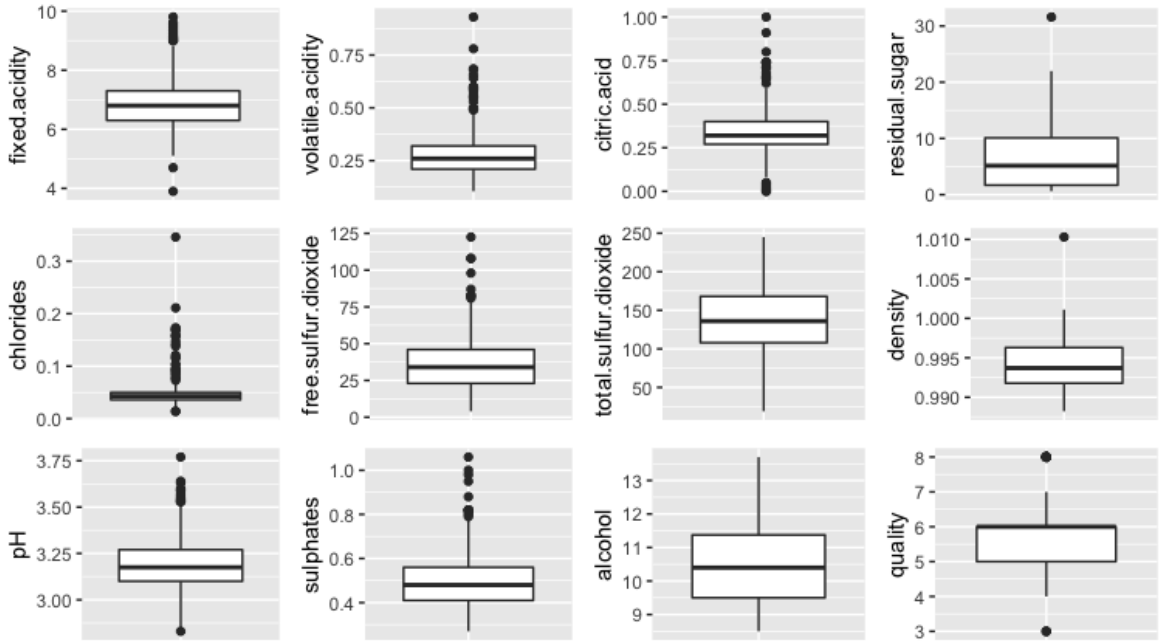
(a) EDA First 11 variables

(i) In the dataset there existed a total number of 32 variables with the final variable (column number 32) being the response variable of quality. The first 11 variables were summarized along with the response variable...



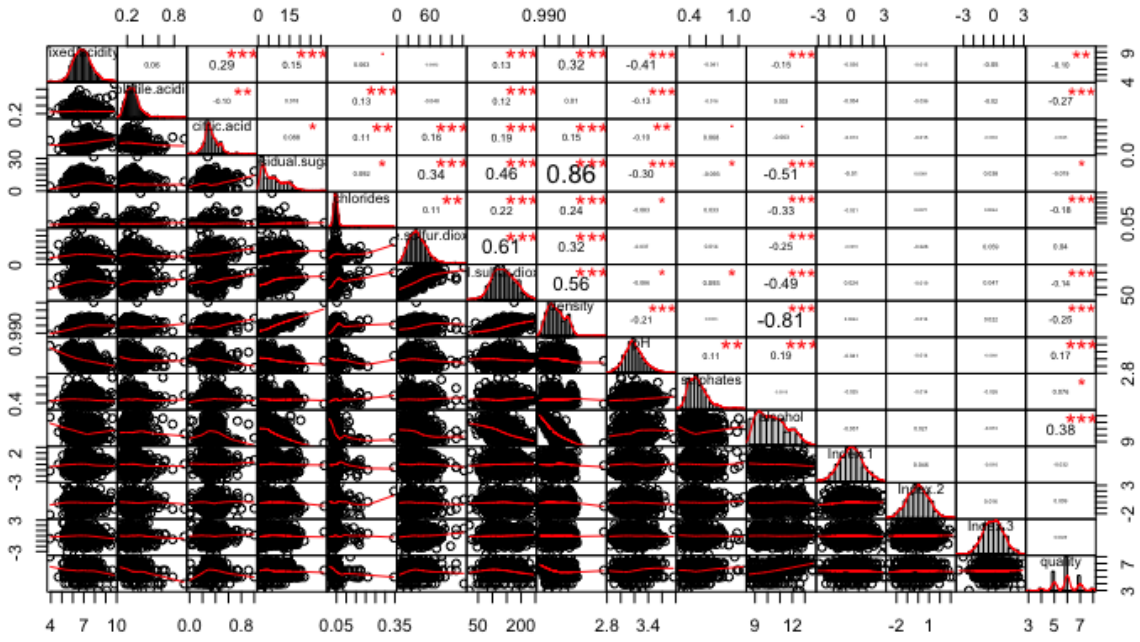
(See appendix.1 for bigger image)
As one can see the variables in association with one another it is evident that some variables have a linear correlation with one another, which hints in using a linear regression model. Also, because all the variables are continuous variables and most show a linear relationship with one another we can say that we are able to use a linear regression model to do analysis on this specific dataset.

Individually one can see the distributions as shown in the table below...



(See appendix.2 for bigger image)
In the boxplots provided one can see some skewness of the boxplots, but in linear regression this does not have such a big impact and will be safe to run linear regression.

(b) Pairwise association



(i)

(See appendix.3 for bigger image)
On the diagonal of the figure above one can see that many variables have a normal distribution and some slight skewness. Also, from the graph above one can see that there are many strong correlations between variables and signifi-

cant correlated variables within the dataset.

(c) Mallow's Cp, AIC, BIC

(i) Step Wise

i. AIC

With running the code of ...

step(lmfit, direction="both", trace=0)

Where lmfit = lm(quality ., data=trainWine) ...

Variable Name	Coefficients
Intercept	$2.564255e+02$
fixed.acidity	$2.231203e-01$
volatile.acidity	$-2.528513e+00$
residual.sugar	$1.245873e-01$
free.sulfur.dioxide	$3.423202e-03$
density	$-2.598852e+02$
pH	$1.729615e+00$
sulphates	$1.080699e+00$
Index.9	$6.440924e-02$

ii. BIC

With running the code of ...

step(lmfit, direction="both", k=log(n), trace=0)

Where lmfit = lm(quality ., data=trainWine)
and n = nrow(trainWine) ...

Variable Name	Coefficients
Intercept	249.4539
fixed.acidity	0.2196
volatile.acidity	−2.5663
residual.sugar	0.1244
density	−252.6838
pH	1.7168
sulphates	1.0969

iii. Mallow's Cp

A. There is a wonderful proof done that Mallow's Cp is equivalent to AIC in terms of Guassian Linear Regression.
https://rstudio-pubs-static.s3.amazonaws.com/324771_0bd880964f064c53a70e757d5ef39669.html

Also, the Mallow's Cp was derived from AIC, making it a special kind of AIC, which doesn't apply to the model we are working with at the moment. Therefore we can say that the Mallow's Cp will be equivalent to AIC in doing model selection.

(ii) Forward Selection

i. AIC

With running the code of ...

step(lm(quality 1, data=trainWine), scope=list(upper=lmfit, lower= 1), direction="forward")

We start the model with lm(quality 1, which is considered a null model because in forward selection the model start with the null model and sequentially adds predictors until the score does not improve.

Variable Name	Coefficients
Intercept	−0.0299119
alcohol	0.318096
volatile.acidity	−2.746851
residual.sugar	0.028060
density	−252.6838
pH	0.833553
sulphates	0.721470
free.sulfur.dioxide	0.003995
Index.9	0.054131

ii. BIC
With running the code of ...

step(lm(quality 1, data=trainWine), k=log(n), scope=list(upper=lmfit, lower= 1), direction="forward")

This command is from combining the forward seelction process with the k=log(n) part oft he BIC process.

Variable Name	Coefficients
Intercept	0.143419
alcohol	0.30860
volatile.acidity	−2.78529
residual.sugar	0.03122
pH	0.84567
sulphates	0.75590

iii. Mallow's Cp

A. Again, regardless of forward or step wise the AIC and Mallow's Cp will hold the same result in the case of Gaussian Linear Regression therefore, the result will not be repeated for it is the same as AIC.

(iii) Model Comparison

i. AIC
In AIC the step wise model had a total of 8 parameters that were used to make a model. Interestingly enough the forward selection also selected a total of 8 parameters for its model building. However, the selected parameters were different. Fixed acidity was replaced by alcohol, which had a difference of only 0.1 between the two parameters.

ii. BIC
In BIC the step wise model had a total of 6 parameters that were used to a make a model. However, in forward selection of BIC method the model made had only 5 parameters. This doesn't mean that all the parameters in BIC forward were included in the step wise. Both procedures had different sets of parameters with only some overlapping parameters.

(d) Ridge vs Lasso

(a) Using the following the code a **ridge** regression was fit on the training data ...

```
cv.rigde<-cv.glmnet(x,y, nfolds = 5,alpha = 0)
cv.rigde$lambda.min
model.ridge <- glmnet(x, y, alpha = 0, lambda = cv.rigde$lambda.min)
coef(model.ridge)
```

Which resulted in the following ouput...

Variable	Coefficients
(Intercept)	$5.629959e+01$
fixed.acidity	$6.672364e-02$
volatile.acidity	$-2.693005e+00$
citric.acid	$-1.818547e-01$
residual.sugar	$4.521483e-02$
chlorides	$-1.374615e+00$
free.sulfur.dioxide	$3.552348e-03$
total.sulfur.dioxide	$3.680973e-04$
density	$-5.678471e+01$
pH	$9.824575e-01$
sulphates	$7.812738e-01$
alcohol	$2.341127e-01$
Index.1	$2.103638e-02$
Index.2	$-5.355639e-03$
Index.3	$3.628071e-03$
Index.4	$-2.267737e-02$
Index.5	$-3.811882e-04$
Index.6	$3.532345e-02$
Index.7	$-1.714039e-02$
Index.8	$9.269244e-03$
Index.9	$5.334085e-02$
Index.10	$2.532059e-02$
Index.11	$-1.639062e-02$
Index.12	$1.644660e-02$
Index.13	$-1.235275e-02$
Index.14	$3.104913e-02$
Index.15	$1.856300e-02$
Index.16	$-2.251833e-02$
Index.17	$1.087859e-02$
Index.18	$-3.235036e-03$
Index.19	$3.750384e-02$
Index.20	$1.790941e-02$

As one can see in Ridge Regression no variable was set to 0. This is because the nature of ridge regression is that the coefficients β^{bridge} are shrunken towards 0. Never will those values be exactly 0

(b) Using the following code a **lasso** regression was fit on the training data ...

```
cv.lasso<-cv.glmnet(x,y, nfolds = 5,alpha = 1)
cv.lasso$lambda.min
model.lasso <- glmnet(x, y, alpha = 1, lambda = cv.lasso$lambda.min)
coef(model.lasso)
```

Which resulted in the following ouput...

Variable	Coefficients
(Intercept)	1.719369989
fixed.acidity	.
volatile.acidity	−2.416625545
citric.acid	.
residual.sugar	0.012747418
chlorides	−0.659626653
free.sulfur.dioxide	0.002672462
total.sulfur.dioxide	.
density	.
pH	0.568850119
sulphates	0.418276682
alcohol	0.254437557
Index.1	.
Index.2	.
Index.3	.
Index.4	.
Index.5	.
Index.6	0.001421138
Index.7	.
Index.8	.
Index.9	0.013900479
Index.10	.
Index.11	.
Index.12	.
Index.13	.
Index.14	0.002326954
Index.15	.
Index.16	.
Index.17	.
Index.18	.
Index.19	0.005905421
Index.20	.

As one can see in Lasso regression table output, many variables were set to '.', which translates to the coefficient being 0. Ridge regression minimizes the coefficients, but Lasso has the power to eliminate the variable or bring the coefficient down to 0, which was shown clearly in the different output of Lasso and Ridge regression.

(c) Comparison of Models

i. **OLS**

```
testing<- as.data.frame(testWine)[,-32]
lmfit=lm(quality~., data=trainWine)
lmfit$coef
pred_ols <- predict(lmfit, testing)
mse_ols <- mean((testWine[,32]-pred_ols)^2)
mse_ols
```

The MSE of the OLS model equated to **0.6420006**

ii. **Mallow's Cp**

Mallow's Cp, as mentioned above, will result in the same result as AIC for Guassian Linear Regression...

iii. **AIC** (Step wise)
With running the code of ...

```
aic <- lm(quality~ fixed.acidity+
                   volatile.acidity+
                   residual.sugar+
                   free.sulfur.dioxide+
                   density+
                   pH+
                   sulphates+
                   Index.9, data=trainWine)
aicpred <- predict(aic, as.data.frame(testing))
mse_aic<- mean((testWine[,32]-aicpred)^2)
mse_aic
```

The MSE of the AIC Step Wise model selection model was **0.6132836**

iv. **BIC** (Step wise) With running the code of ...

```
bic <- lm(quality~ fixed.acidity+
                   volatile.acidity+
                   residual.sugar+
                   density+
                   pH+
                   sulphates, data=trainWine)
bicpred <- predict(bic, as.data.frame(testing))
mse_bic<- mean((testWine[,32]-bicpred)^2)
mse_bic
```

The MSE of the BIC Step Wise model selection model was **0.6153804**

v. **Ridge** With running the code of ...

```
ridgetrain <- cv.glmnet(x,y, nfolds = 5,alpha = 0)
ridgepred <- predict(ridgetrain, as.matrix(testing))
mse_testing_ridge <- mean((testWine[,32]-ridgepred)^2)
mse_testing_ridge
```

The MSE of the Ridge Regression model was **0.6763519**

vi. **Lasso** With running the code of ...

```
lassotrain <- cv.glmnet(x,y, nfolds = 5,alpha = 1)
predlasso <- predict(lassotrain,as.matrix(testing))
mse_testing_lasso <- mean((testWine[,32]-predlasso)^2)
mse_testing_lasso
```

The MSE of the Lasso Regression model was **0.6583489**

vii. Comparison of models...

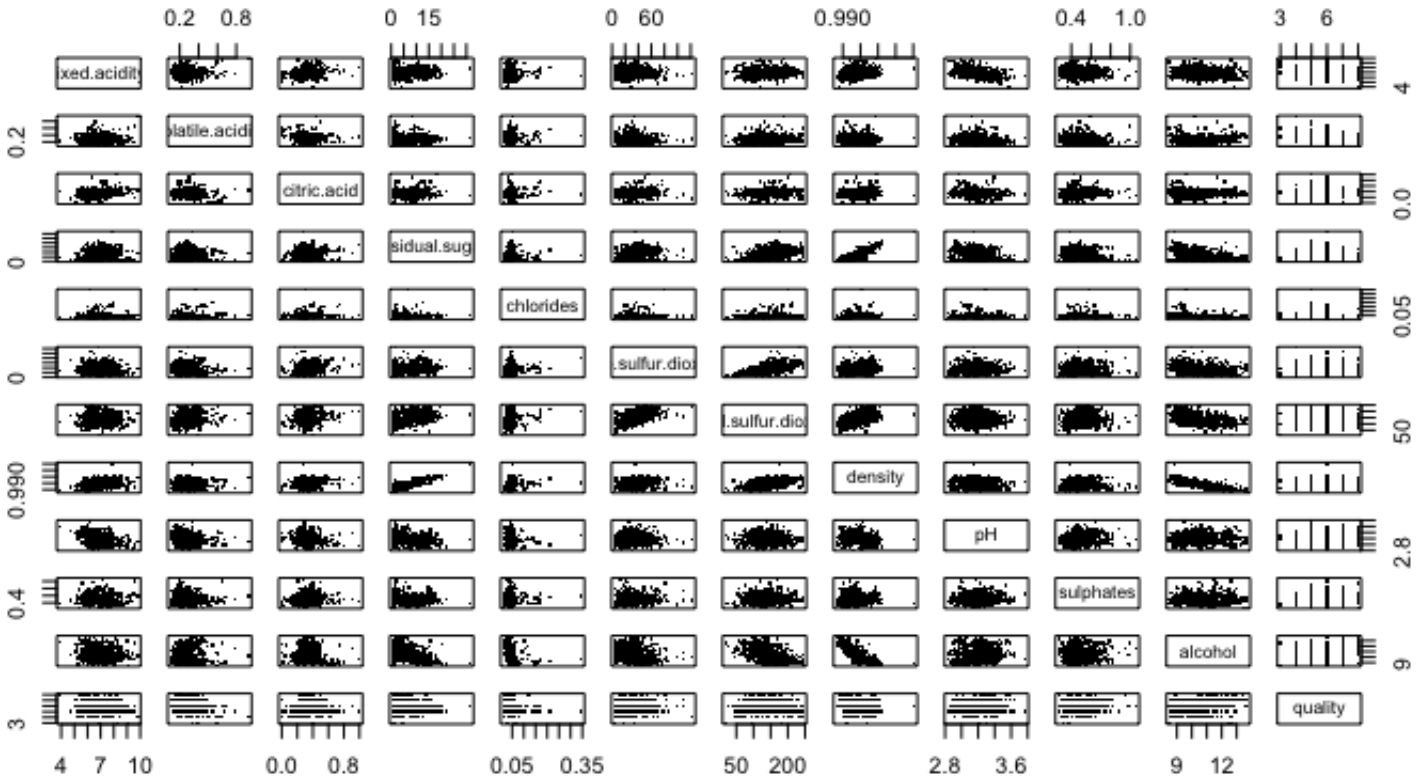
Models	MSE
OLS	0.6420006
Mallow's Cp	Same as AIC
AIC	0.6132836
BIC	0.6153804
Ridge	0.6763519
Lasso	0.6583489

Although the values of BIC and AIC were very close 0.6153804, 0.6132836 respectively AIC yeilded a slighlty better MSE, which thus made AIC the best performing model. AIC and BIC account for overfitting penalty, which may lead to a better MSE in comparison to OLS since OLS uses the most

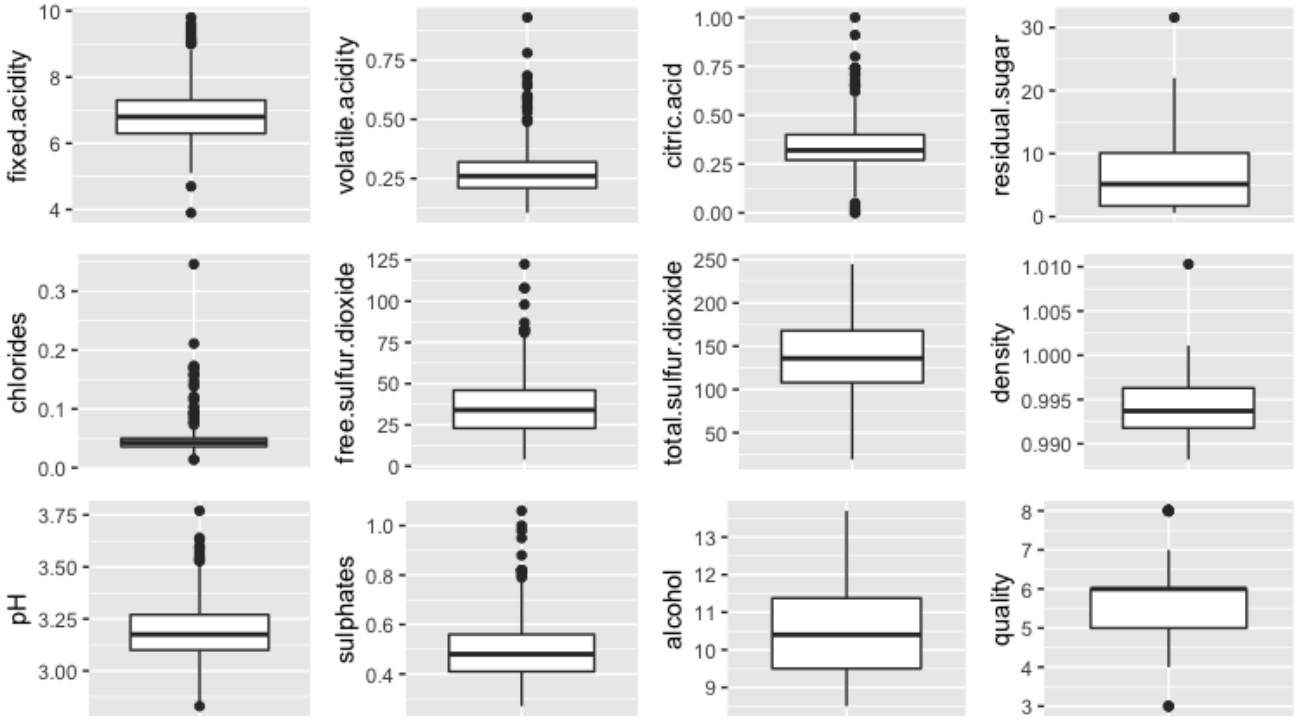
complex model. Ridge and Lasso depend on a value of lambda, which is the hyper parameter to do optimization. Therefore, the default value may not be the optimal value to evaluate the model, which may lead to a higher MSE in comparison to AIC and BIC. A way to combat this issue is to do more cross validation (¿5) to reach a more optimal lambda value for penalization.

PROBLEM 2 *Appendix*

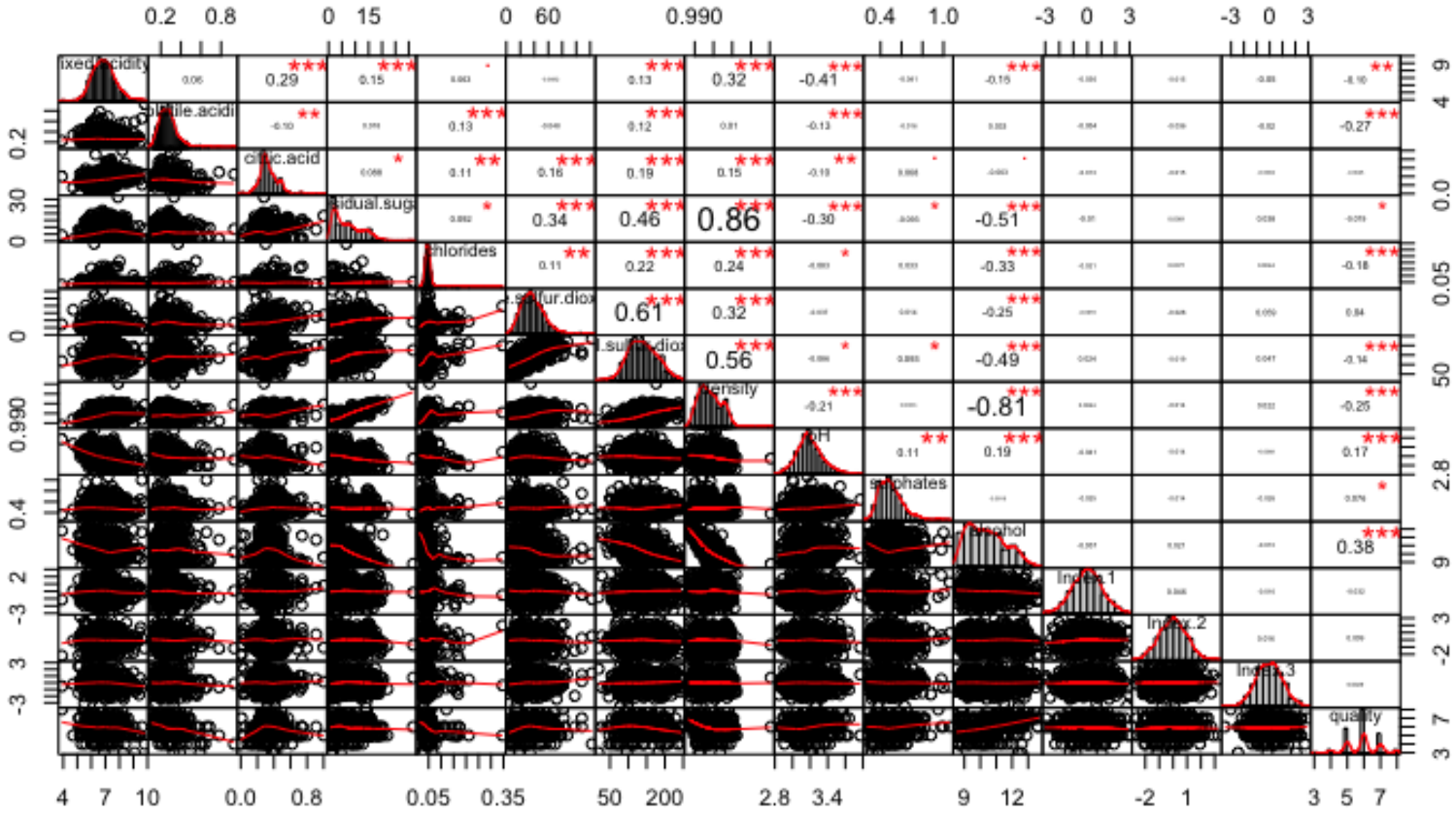
1.



2.


fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality

3.


fixed.acidity
volatile.acidi
citric.acid
idual.sug
hlorides
ur.diox
ensity
pH
sulphates
lcohol
Index.1
Index.2
Index.3
quality