



PRÉDICTION DE LA QUALITÉ DU VIN ROUGE

Via l'utilisation d'algorithme de machine learning

Prepared by
Aourik khalid

Sommaire

Mise en context

preprocessing

EDA

Modeling

Evaluation / conclusion



Qui sommes nous ?

Nous sommes une entreprise dont la mission est de permettre aux entreprises d'améliorer leur productivité nous faisons des analyses et des prédictions affine de mieux les conseiller pour divers projets



Mise en context

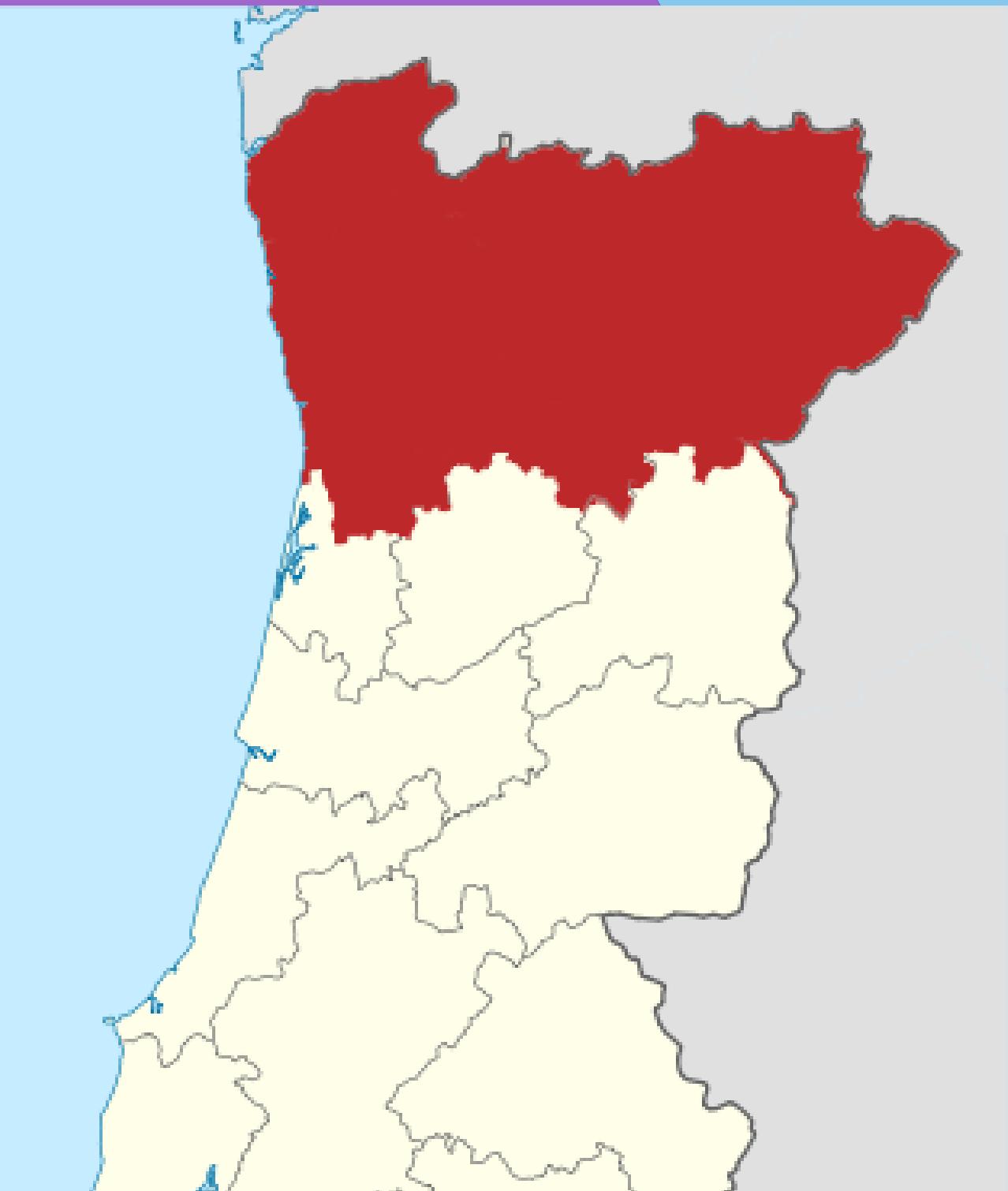


L'industrie du vin connaît une récente poussée de croissance en raison de l'augmentation de la consommation sociale. Le prix du vin dépend d'un concept plutôt abstrait d'appréciation du vin par les dégustateurs, dont l'opinion peut être très variable. Le prix du vin dépend dans une certaine mesure de ce facteur volatile. Les tests physico-chimiques sont un autre facteur essentiel de la certification et de l'évaluation de la qualité du vin. Ils sont effectués en laboratoire et prennent en compte des facteurs tels que l'acidité, le pH, le sucre et d'autres propriétés chimiques. Il serait intéressant pour le marché du vin que la qualité de la dégustation humaine puisse être liée aux propriétés chimiques du vin afin que les processus de certification et d'évaluation et d'assurance de la qualité soient mieux contrôlés.

Provenance des données

Les données ont été fournies par l'UCI :
<https://archive.ics.uci.edu/ml/datasets/wine+quality>

L'échantillon provient du Nord du Portugal
Il comporte 1599 observations.



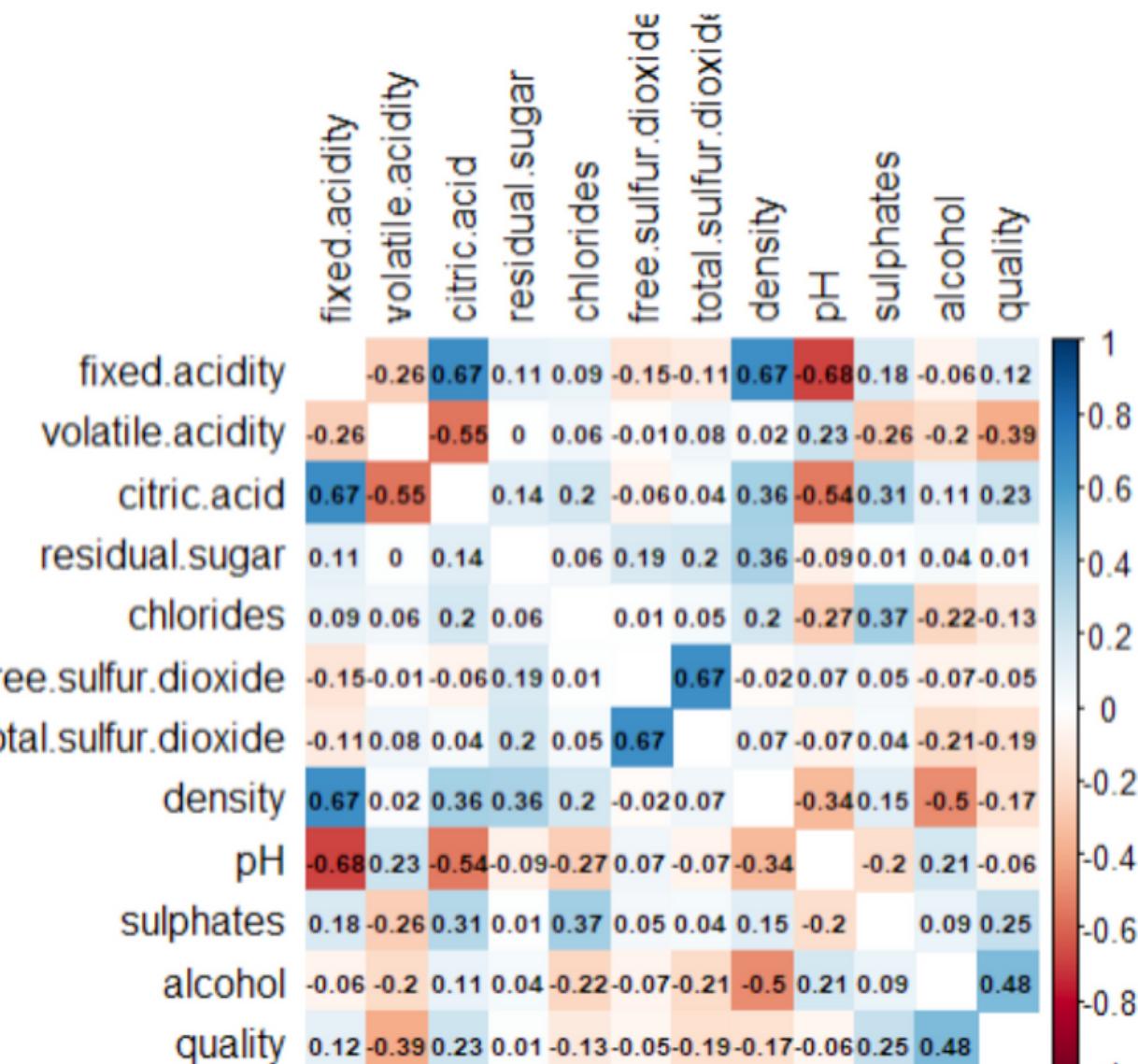
Preprocessing

Nettoyage des données :

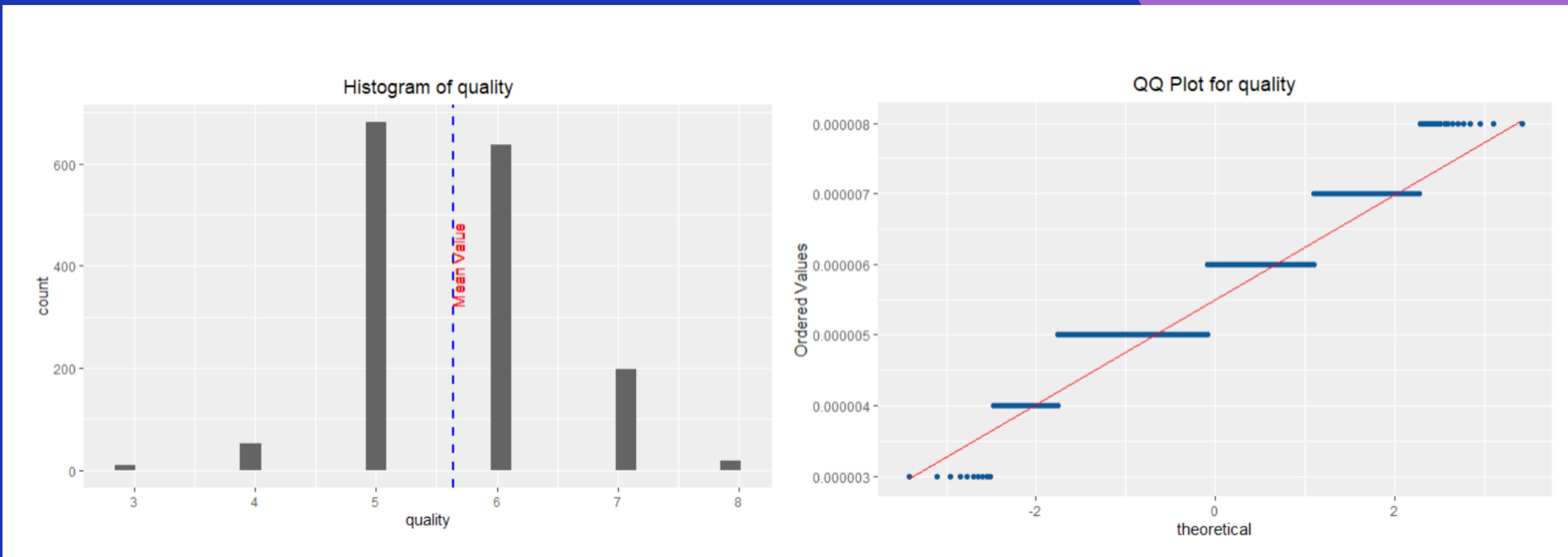
- vérification des types de donne
- identification des valeurs manquantes
- résumé statistique afin de vérifier les valeurs aberrante

EDA

Figure 1. Correlation Matrix



EDA



EDA

Figure 4.

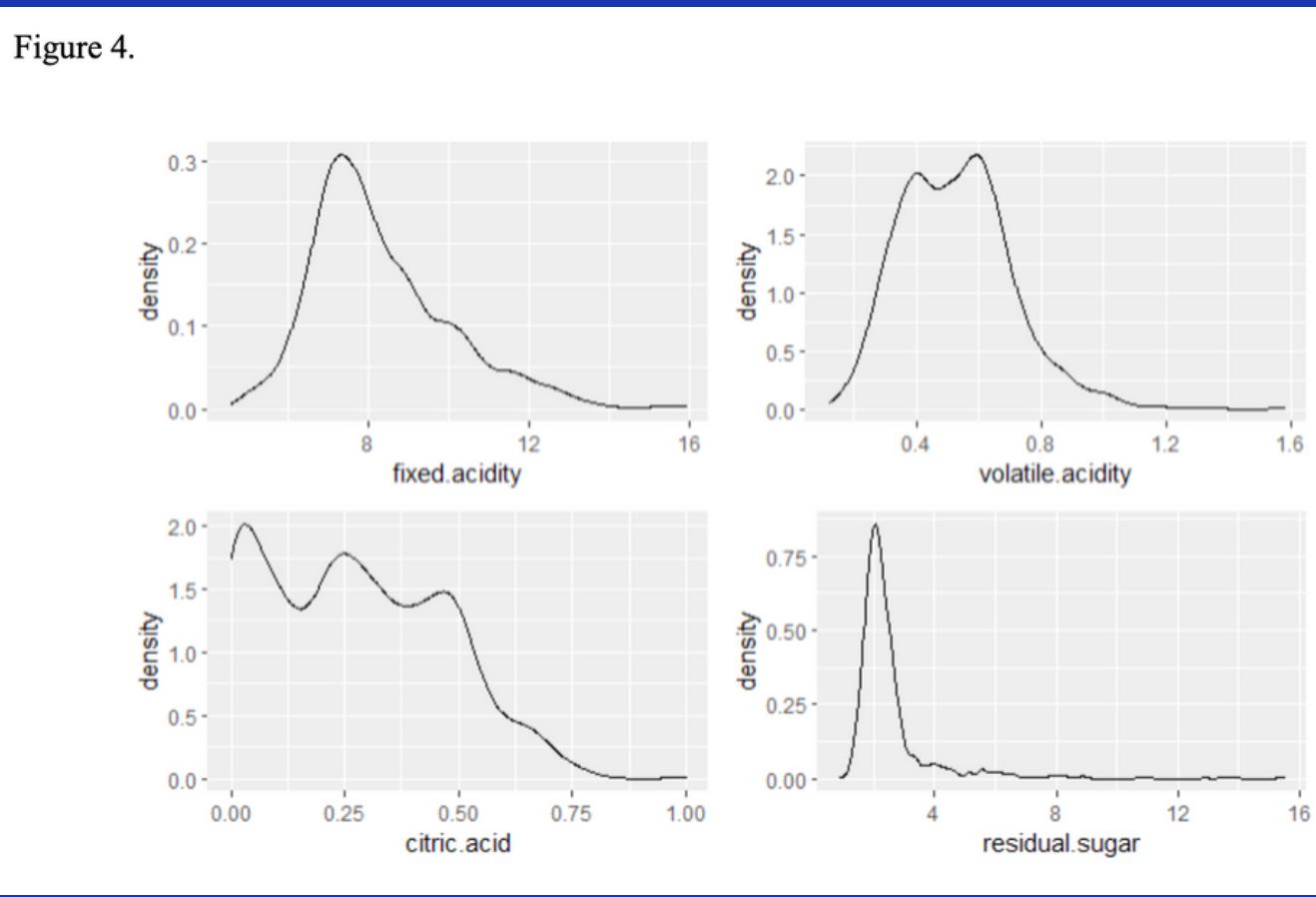


Figure 5.

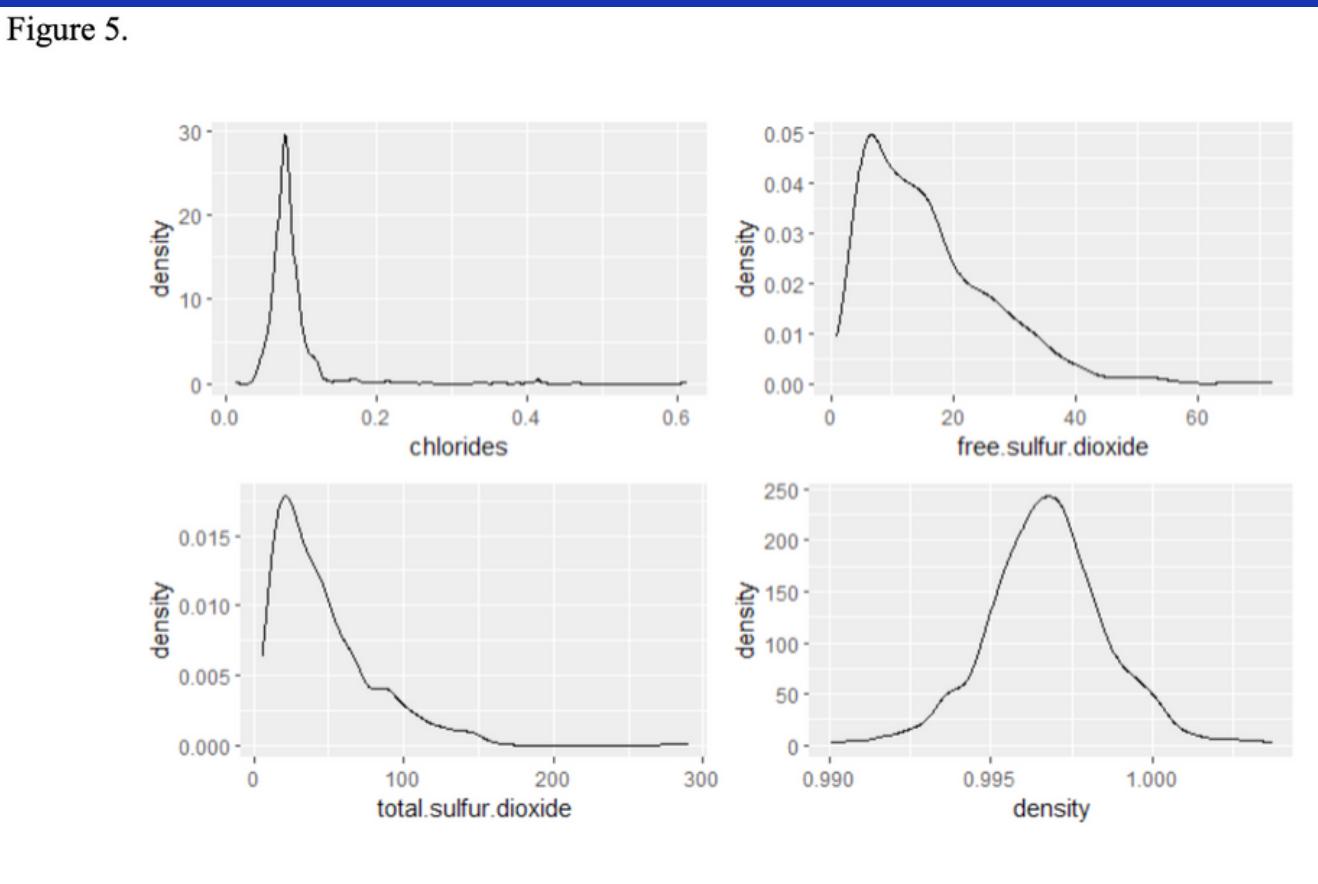
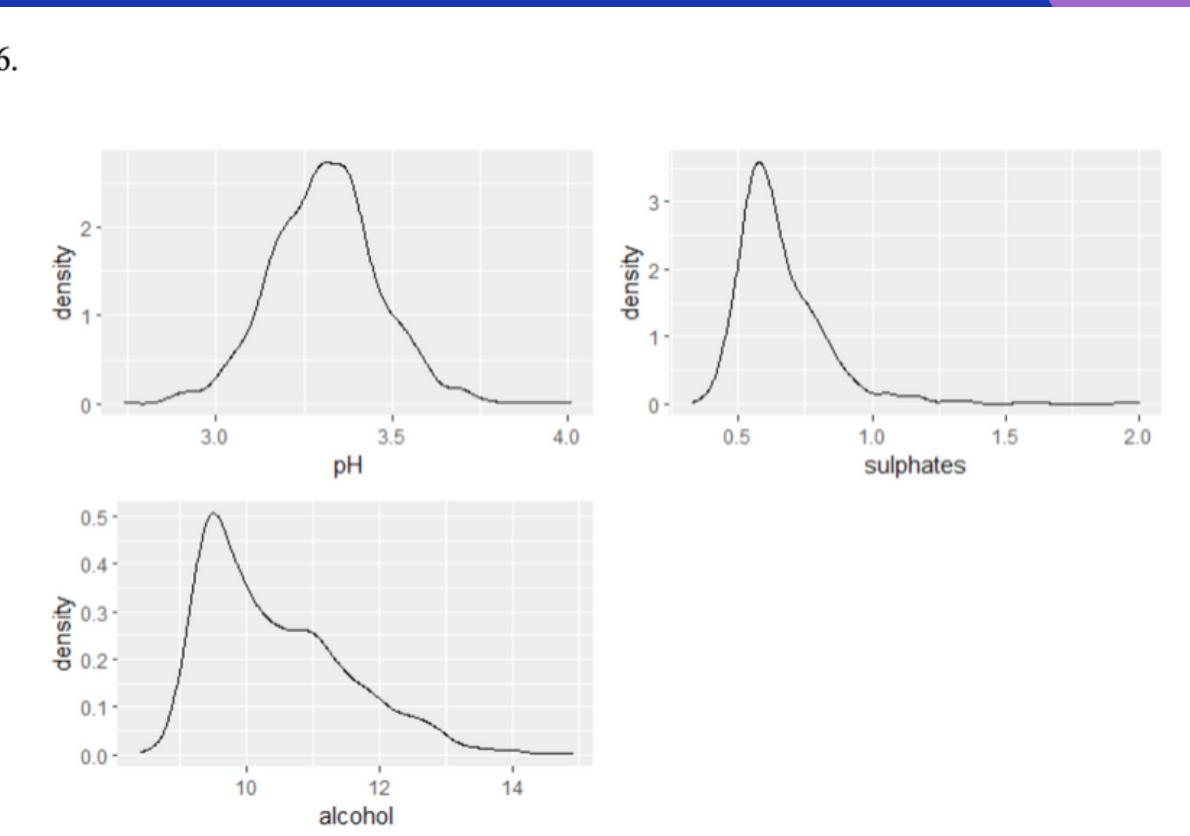


Figure 6.



EDA

Figure 7.

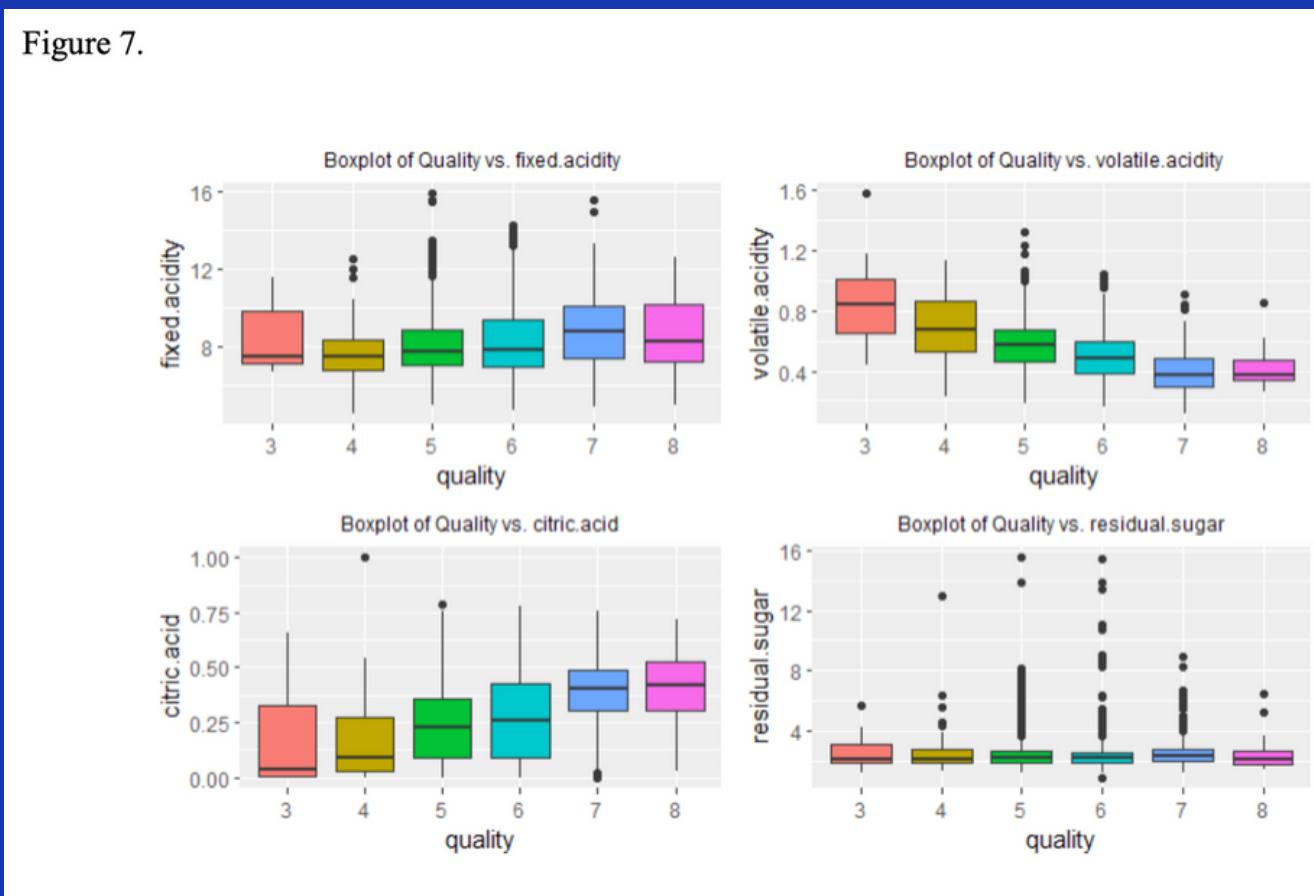


Figure 8.

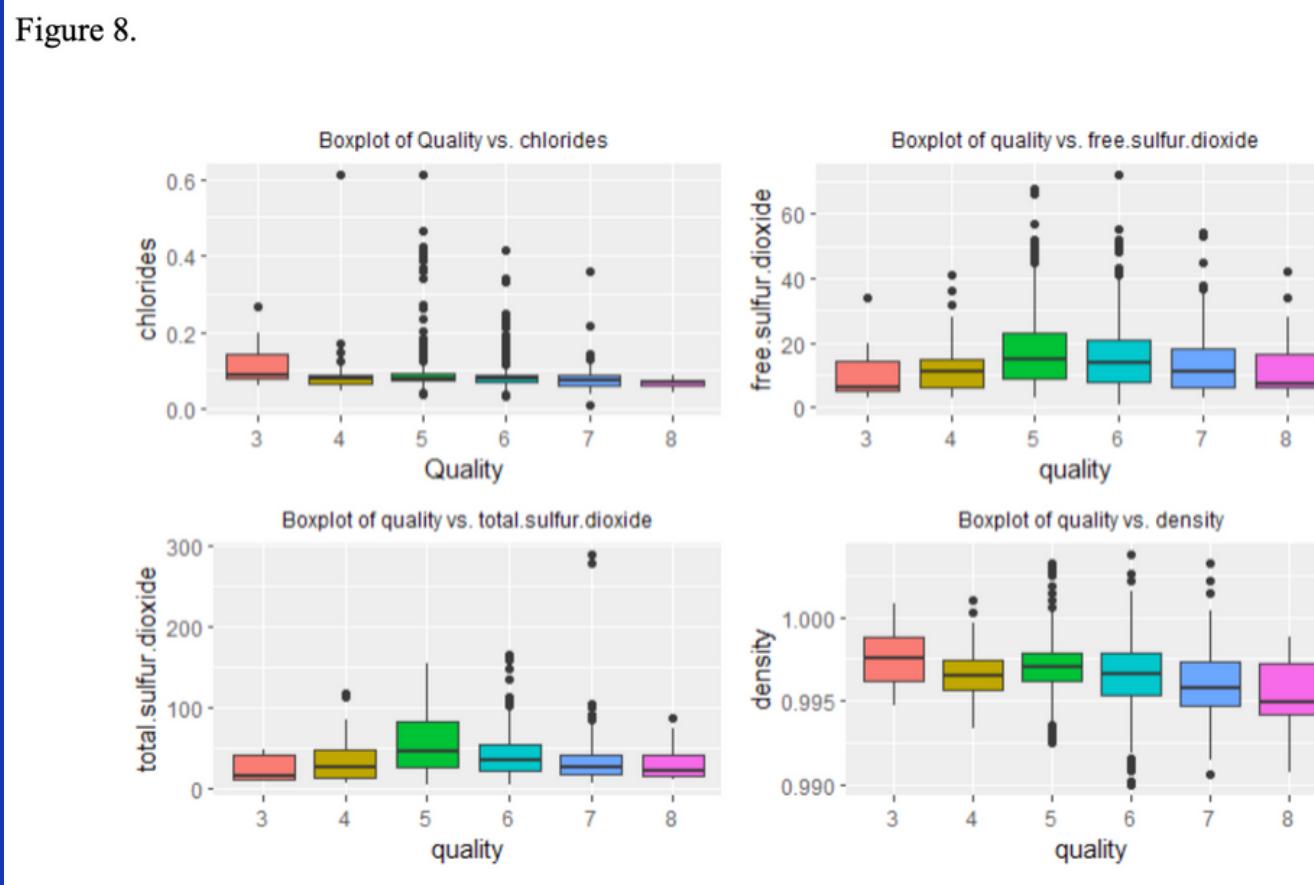
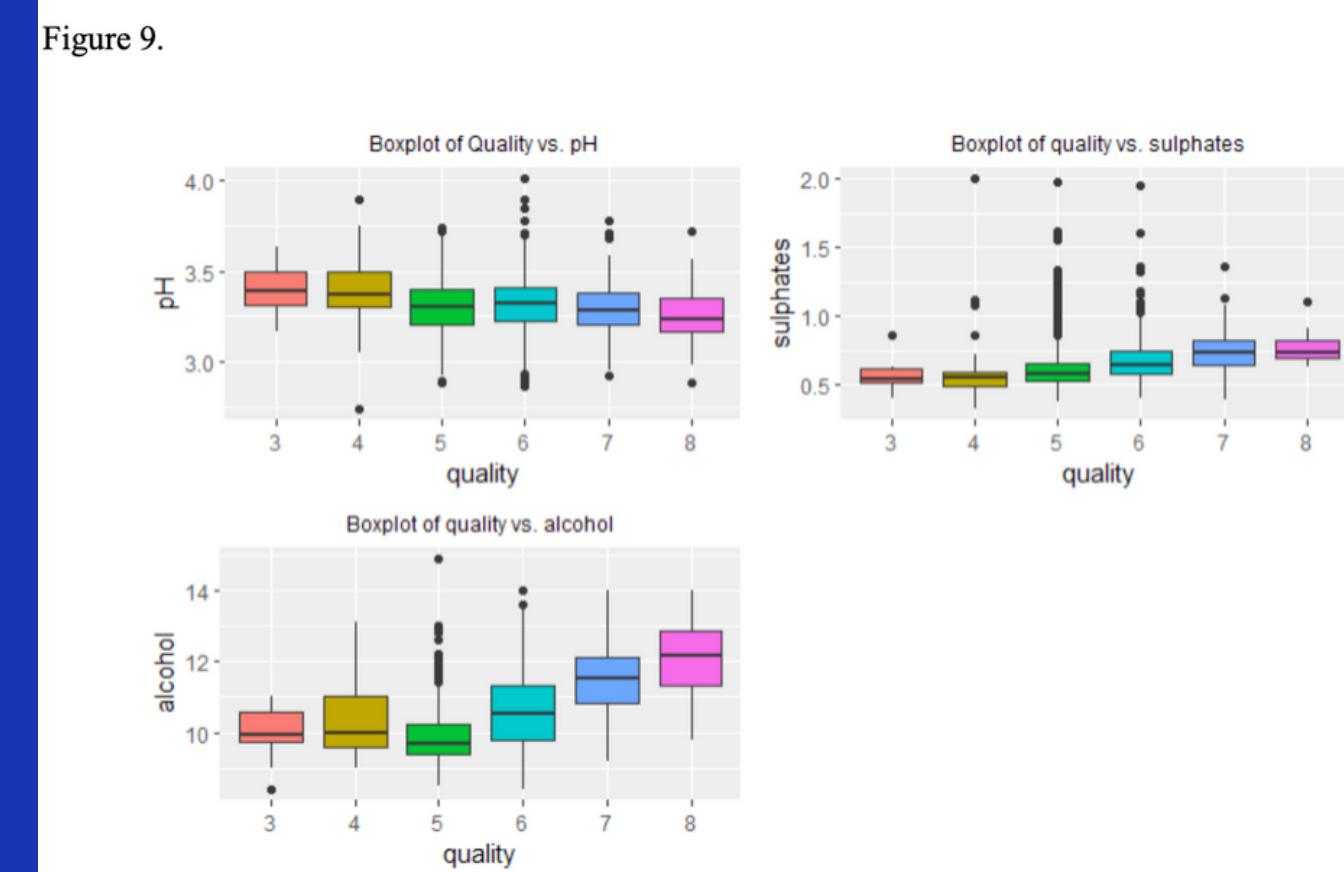


Figure 9.



EDA

Figure 10.

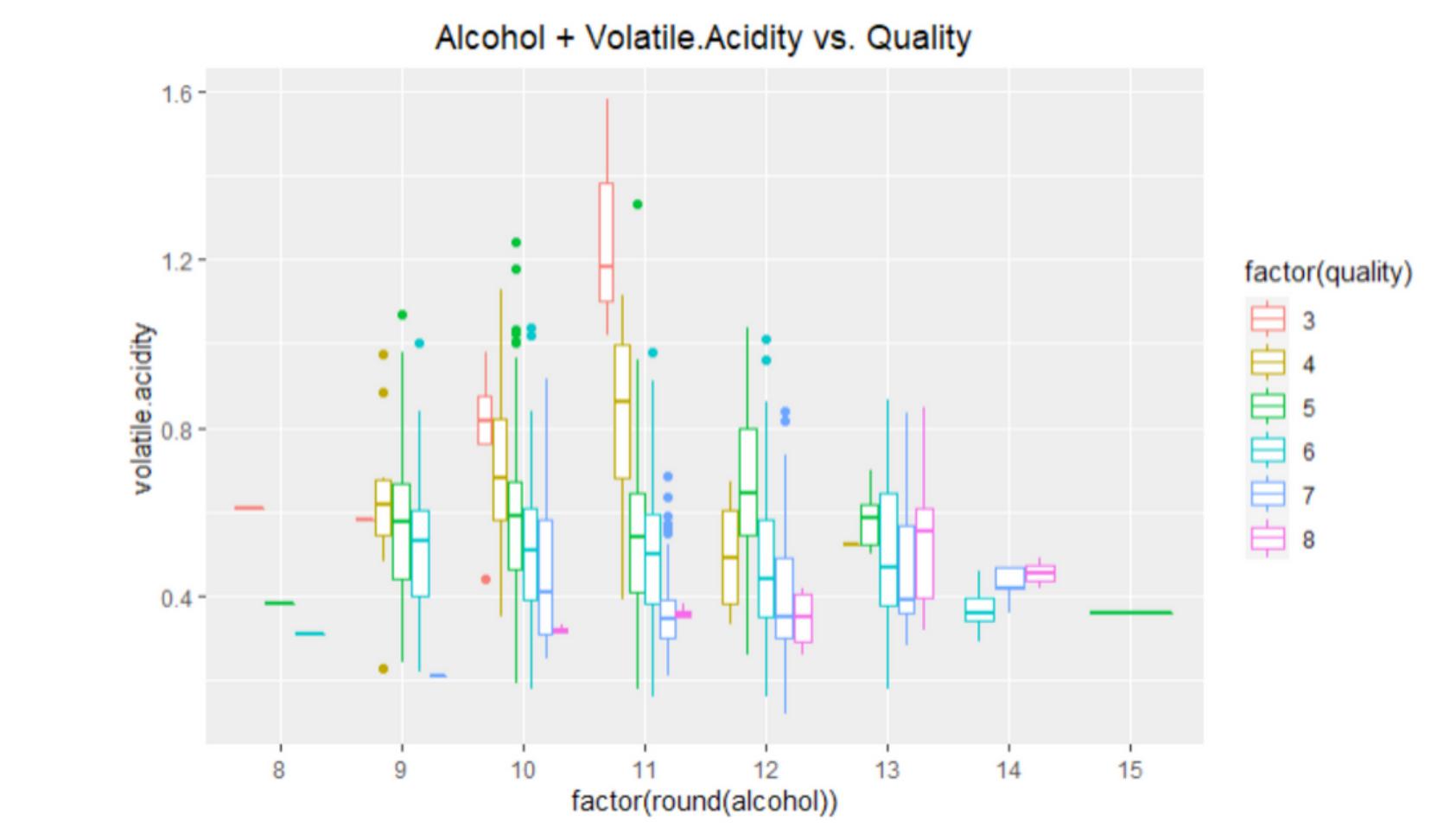
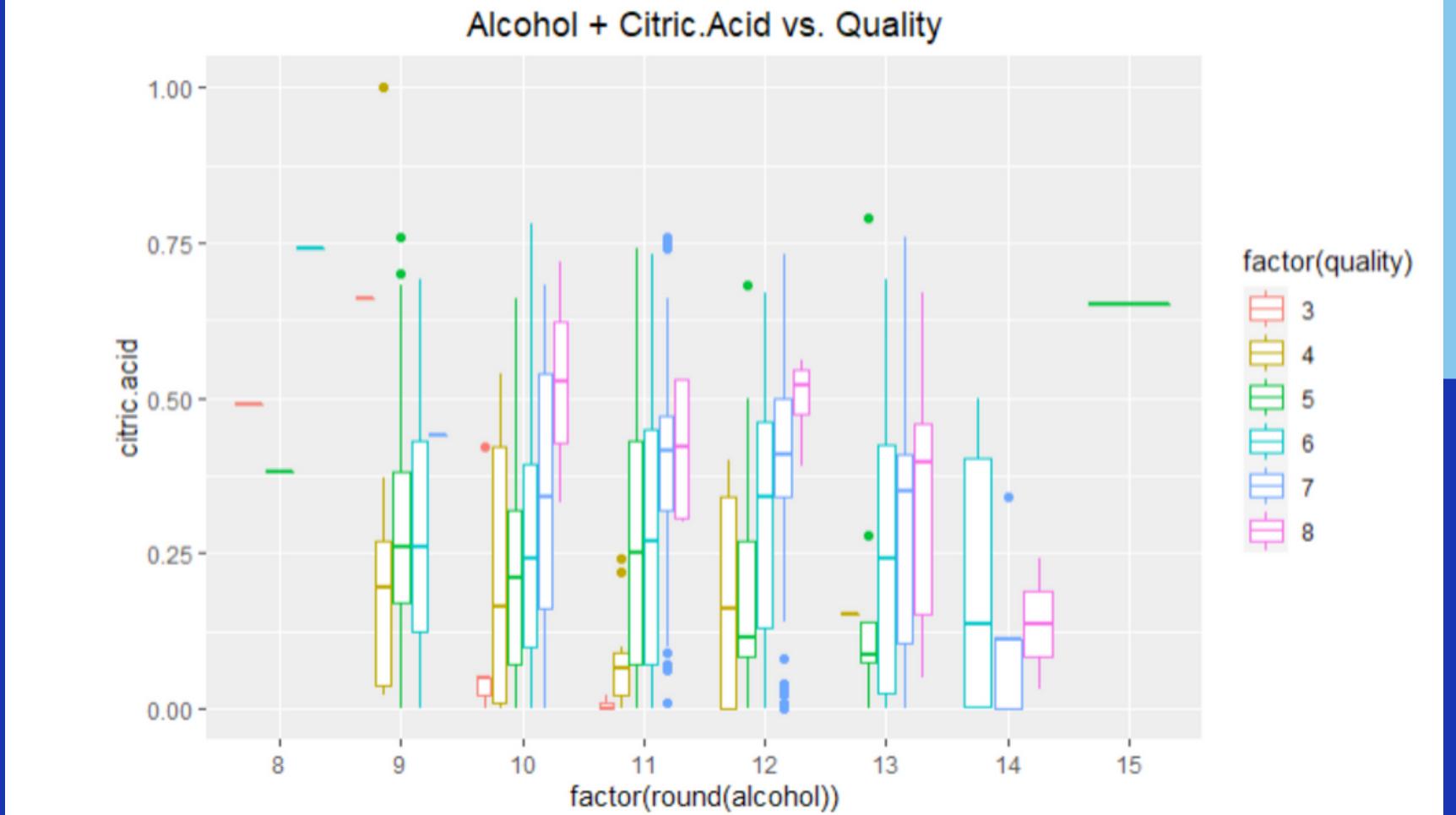
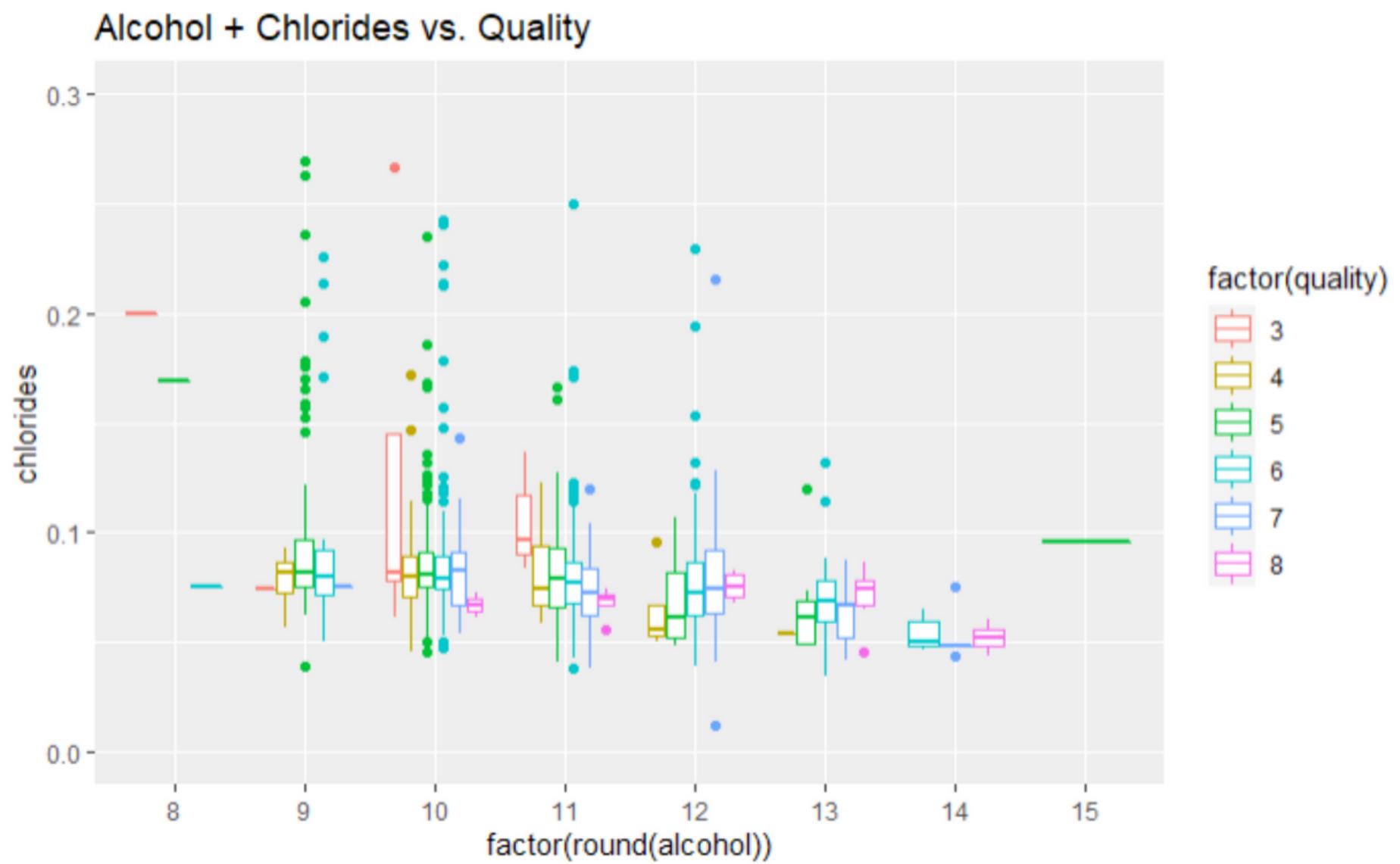


Figure 11.



EDA

Figure 12.



Modeling

Model 1

Multi-linear regression

Model 2

LASSO method

Model 3

Random Forest

Model 1: Multi-lineair

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.72716 -0.38486 -0.06503  0.44980  2.13257 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.8258128  0.2006892 14.081 < 0.0000000000000002 *** 
alcohol       0.2953105  0.0160331 18.419 < 0.0000000000000002 *** 
volatile.acidity -1.1985632  0.0966011 -12.407 < 0.0000000000000002 *** 
sulphates     0.7121396  0.1005146   7.085  0.00000000000208 *** 
total.sulfur.dioxide -0.0022354  0.0005108  -4.376  0.00001284518270 *** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.655 on 1594 degrees of freedom

1599 samples
 4 predictor

No pre-processing
Resampling: cross-validated (10 fold)
Summary of sample sizes: 1438, 1439, 1440, 1438, 1439, 1439, ...
Resampling results:

RMSE      Rsquared      MAE
0.6549281  0.3479475  0.5092899
```

Model 2 : LASSO Method

```
Residuals:
    Min      1Q   Median      3Q     Max 
-2.70812 -0.37181 -0.06238  0.45933  1.99472 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.7365412  0.2325021 11.770 < 0.000000000000002 *** 
fixed.acidity 0.0236576  0.0099187  2.385   0.0172 *    
volatile.acidity -1.0856214  0.0996323 -10.896 < 0.000000000000002 *** 
chlorides     -1.7376885  0.3913566 -4.440  0.00000960779597327 *** 
total.sulfur.dioxide -0.0021460  0.0005121 -4.191  0.00002933553690691 *** 
sulphates      0.8846921  0.1108310  7.982  0.0000000000000272 *** 
alcohol        0.2825603  0.0166180 17.003 < 0.000000000000002 *** 
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.6504 on 1592 degrees of freedom

1599 samples
  6 predictor

No pre-processing
Resampling: Cross-validated (10 fold)
Summary of sample sizes: 1438, 1439, 1440, 1438, 1439, 1439, ...
Resampling results:

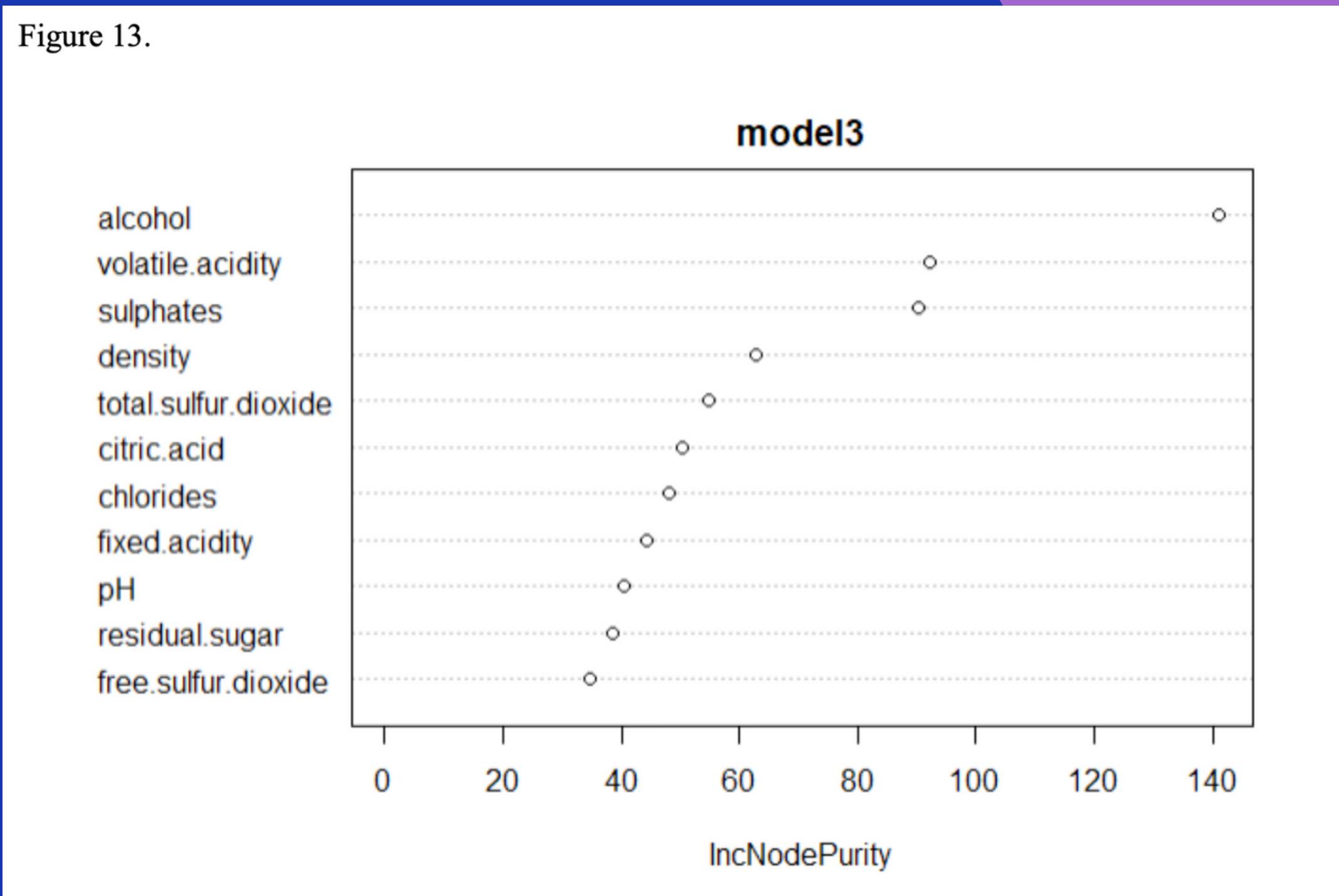
  RMSE      Rsquared      MAE  
  0.6515146  0.3546232  0.5063893
```

Model 3 : Random Forest

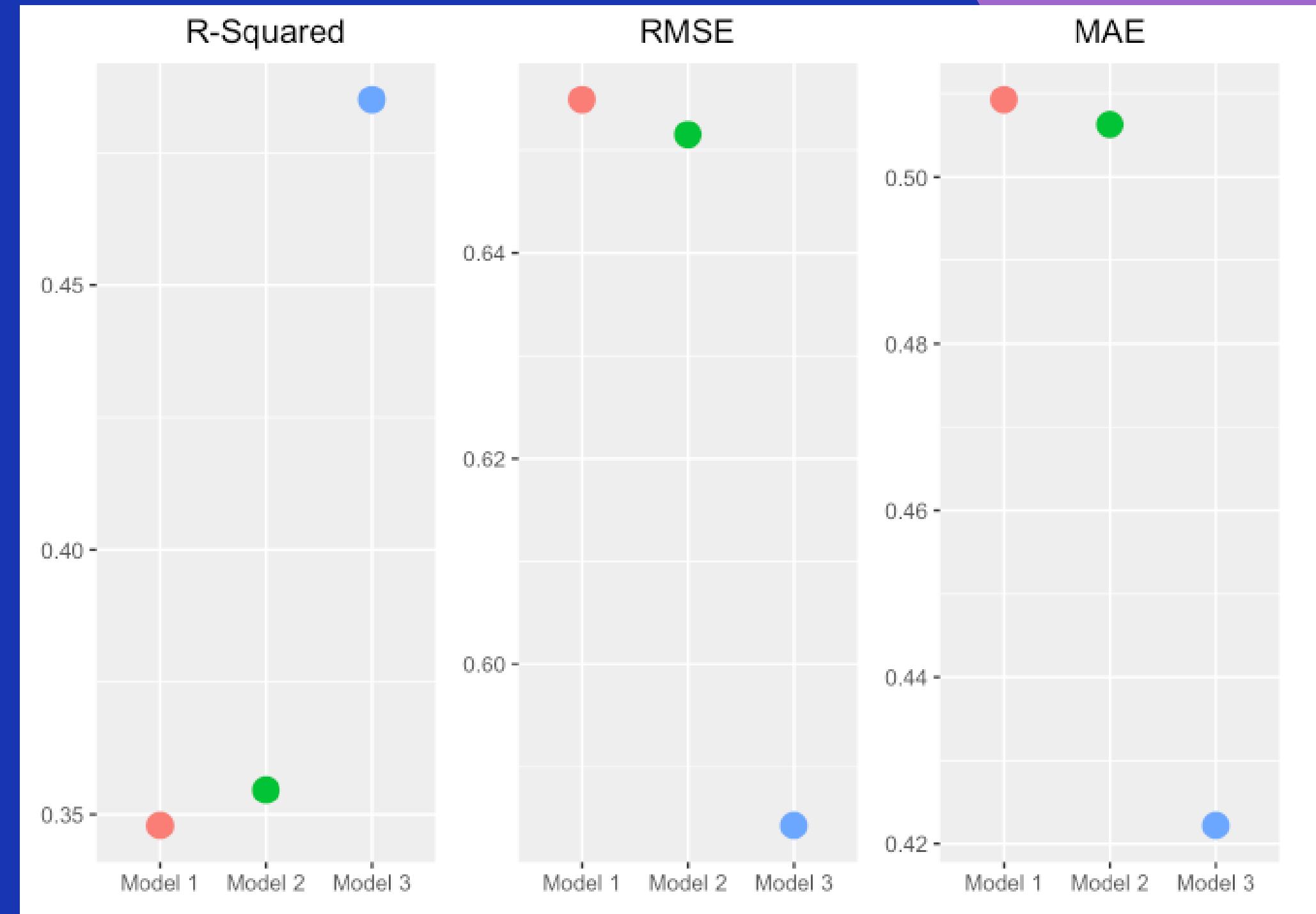
```
call:  
  randomForest(formula = quality ~ ., data = training, mtry = 3,           importance = TRUE,  
na.action = na.omit)  
    Type of random forest: regression  
    Number of trees: 500  
No. of variables tried at each split: 3  
  
  Mean of squared residuals: 0.3414535  
  % var explained: 48.5
```

Model 3

Figure 13.



Evaluation



conclusion

en utilisant le modèle 3 comme meilleur modèle de prédiction, j'ai déterminé que quatre des caractéristiques étaient les plus influentes : l'acidité volatile, l'acide citrique, les sulfates et l'alcool.



Merci !