Build Models to Predict White Wine Quality by Data Mining from Physicochemical Properties of Wine Quality Data

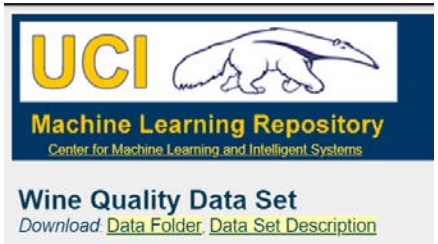
Critical Thinking Group1

Data 621 Final Project

Introduction

- Vinho Verde exclusively produced in the demarcated region of Vinho Verde in northwestern Portugal. It is only produced from the indigenous grape varieties of the region, preserving its typicity of aromas and flavors as unique in the world of wine.
- There are many psychochemical tests involved behind the quality of wine. Our goal is to predict the wine quality based on various psychochemical tests.

DATA Source



- Got from UCI machine learning repository
- (https://archive.ics.uci.edu/ml/datasets/Wine+Quality)
- Two datasets were built based on red and white vinho verde wine samples from the north of Portugal.

Attribute Information:

For more information, read [Cortez et al., 2009].
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Figure 1. No missing data in the dataset

Variables in Dataset

- (1) Fixed acidity: a measurement of the total concentration of titratable acids and free hydrogen ions present in the wine. (2)Volatile acidity: a measure of steam distillable acids present in a wine. (3) Citric acid: one of the many acids that are measured to obtained fixed acidity. (4) Residual sugar: measurement of any natural grape sugars that are leftover after fermentation ceases. (5) Chlorides: the amount of salt in the wine. (6)Free sulfuric dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; (7) Total sulfuric dioxide: amount of free and bound forms of SO2; (8) Density: measure of density of wine. (9)pH: value for pH. (10) Sulfates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant. (11) Alcohol: the percentage of alcohol present in the wine. (12) Quality: subjective measurement ranging from 1 to 10 (although the observed data ranges from 3 to 8).

Table 1. Summarize the variables of Data Set

Variable Name	Min	1st.Q	Median	3rd.Q	Max	Mean	SD
Fixed.acidity	3.80	6.30	6.80	7.30	14.20	6.86	0.844
Volatile.acidity	0.08	0.21	.026	0.32	1.10	0.28	0.101
Citric.acid	0.00	0.27	0.32	0.39	1.66	0.33	0.121
Residual.sugar	0.60	1.70	5.20	9.90	65.80	6.29	5.072
Chlorides	0.01	0.04	0.04	0.050	0.35	0.05	0.022
Free.sulfur.dioxide	2.00	23.00	34.00	46.00	289.00	35.31	17.01
Total.sulfur.dioxide	9.00	108.00	134.00	167.00	440.00	138.40	42.50
Density	0.99	0.99	0.99	1.00	1.04	0.99	0.003
PH	2.72	3.09	3.18	3.28	3.82	3.19	0.151
Sulphates	0.22	0.41	0.47	0.55	1.08	0.49	0.114
Alcohol	8.00	9.50	10.40	11.40	14.20	10.51	1.231

Table 2. Quantity of wines by quality of scores.

Scores	3	4	5	6	7	8	9
Quantity	20	163	1457	2198	880	175	5

Figure 2. Uneven distribution of observers by quality scores

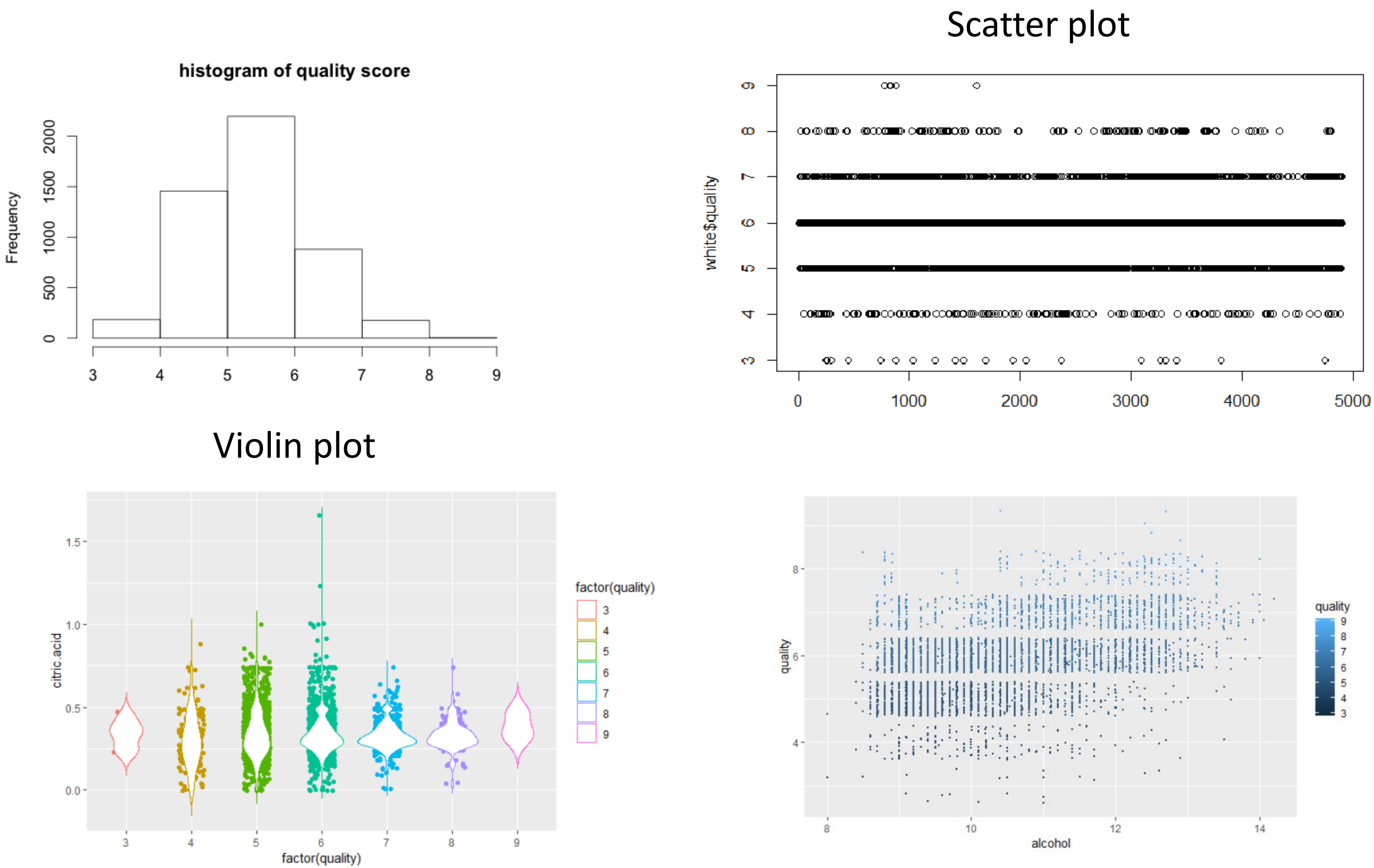


Figure 3. Density plots of variable by scores of quality

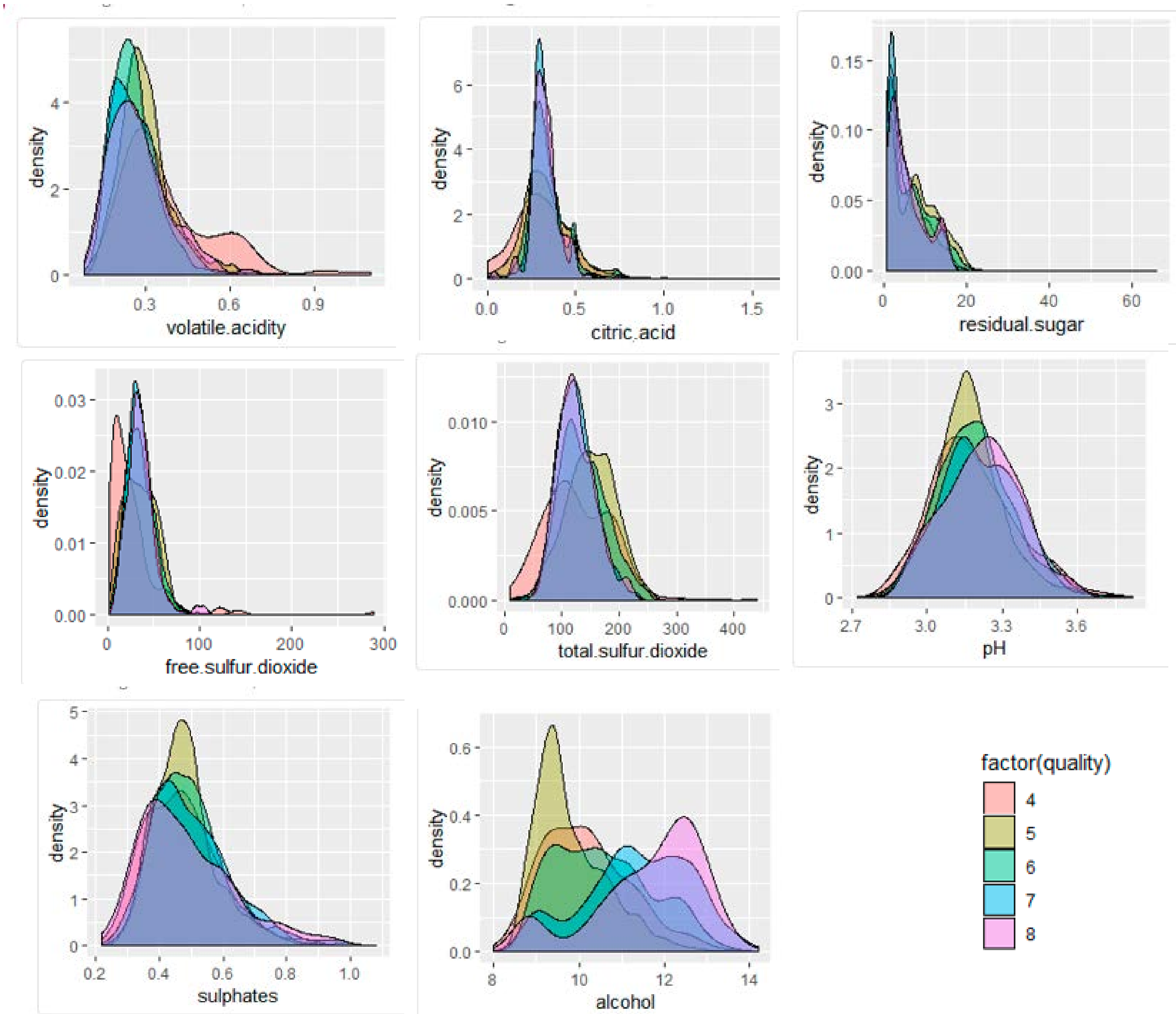
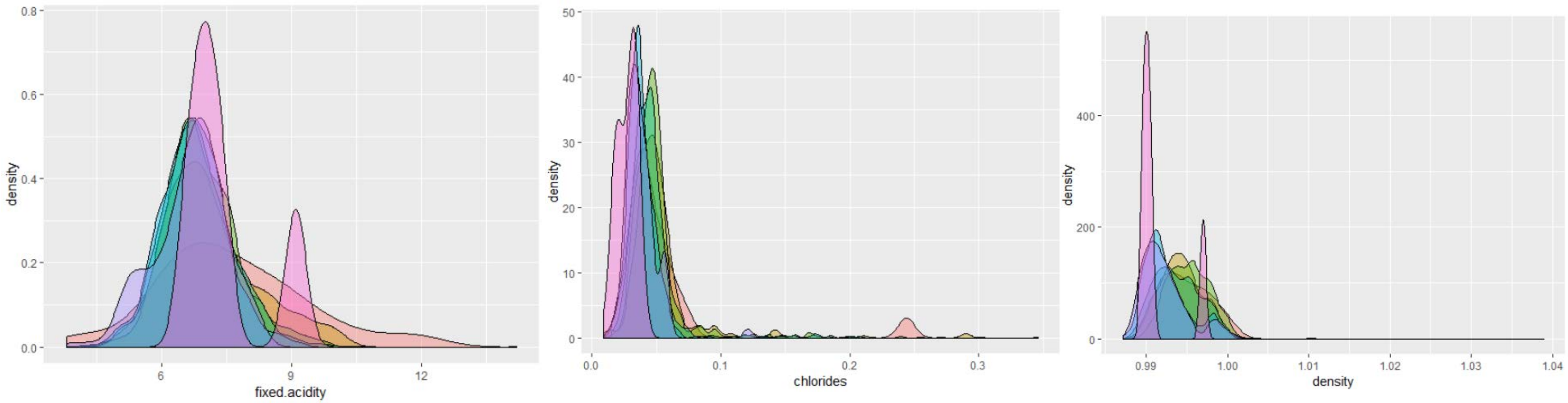


Figure 1. Continue



Tabel 3. means of different variable by score of wine quality

variables	4	5	6	7	8
fixed.acidity	7.18	6.93	6.83	6.73	6.67
volatile.acidity	0.37	0.30	0.26	0.26	0.27
citric.acid	0.30	0.33	0.33	0.32	0.32
residual.sugar	4.82	7.33	6.44	5.18	5.62
chlorides	0.05	0.05	0.04	0.03	0.03
free.sulfur.dioxide	26.63	36.43	35.65	34.12	36.62
total.sulfur.dioxide	130.23	150.90	137.04	125.1	125.88
density	0.99	0.99	0.99	0.99	0.99
pH	3.18	3.16	3.18	3.21	3.21
sulphates	0.47	0.48	0.49	0.50	0.48
alcohol	10.17	9.80	10.57	11.36	11.65

Figure 4. Boxplot of variables by score of quality

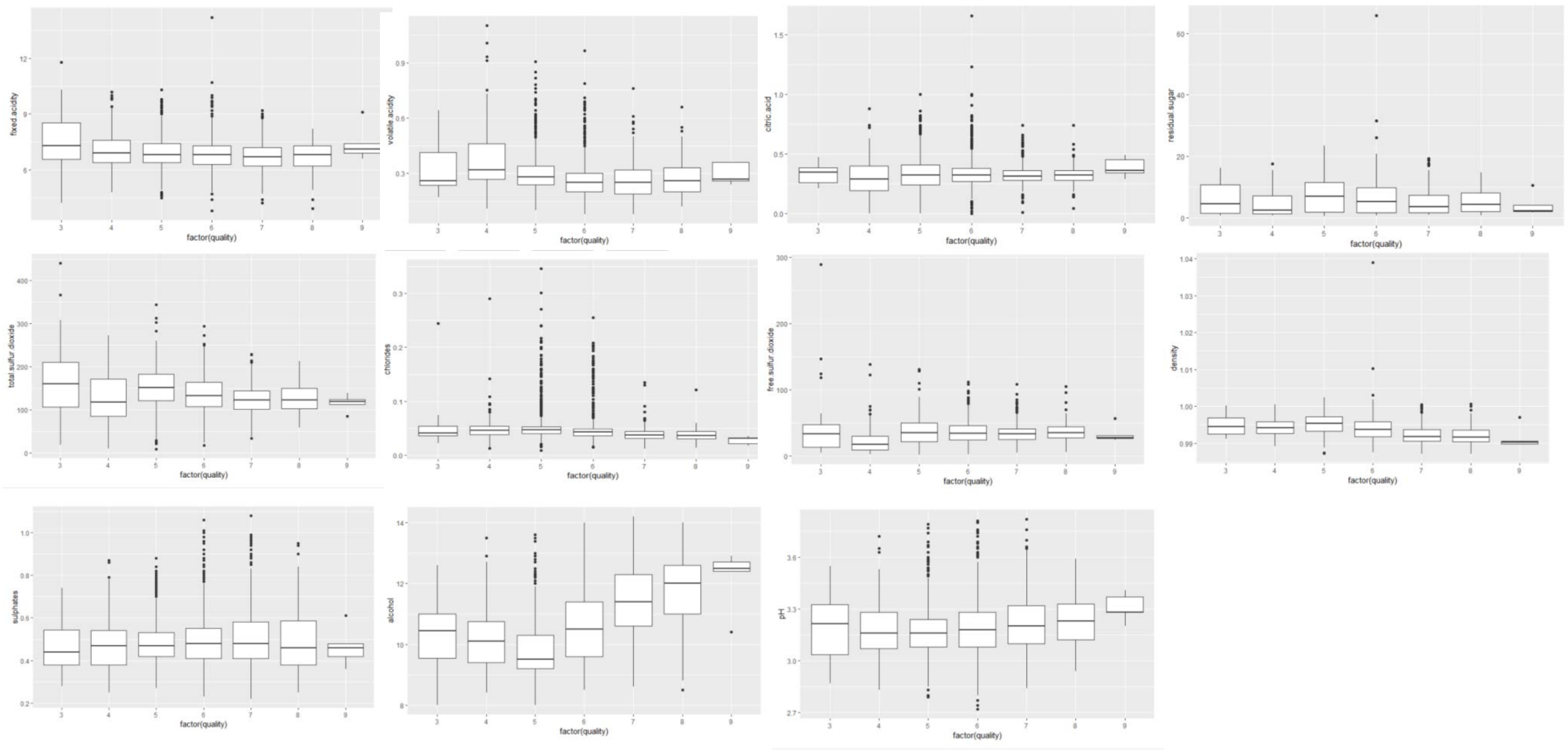


Figure 5. Correlations between variables

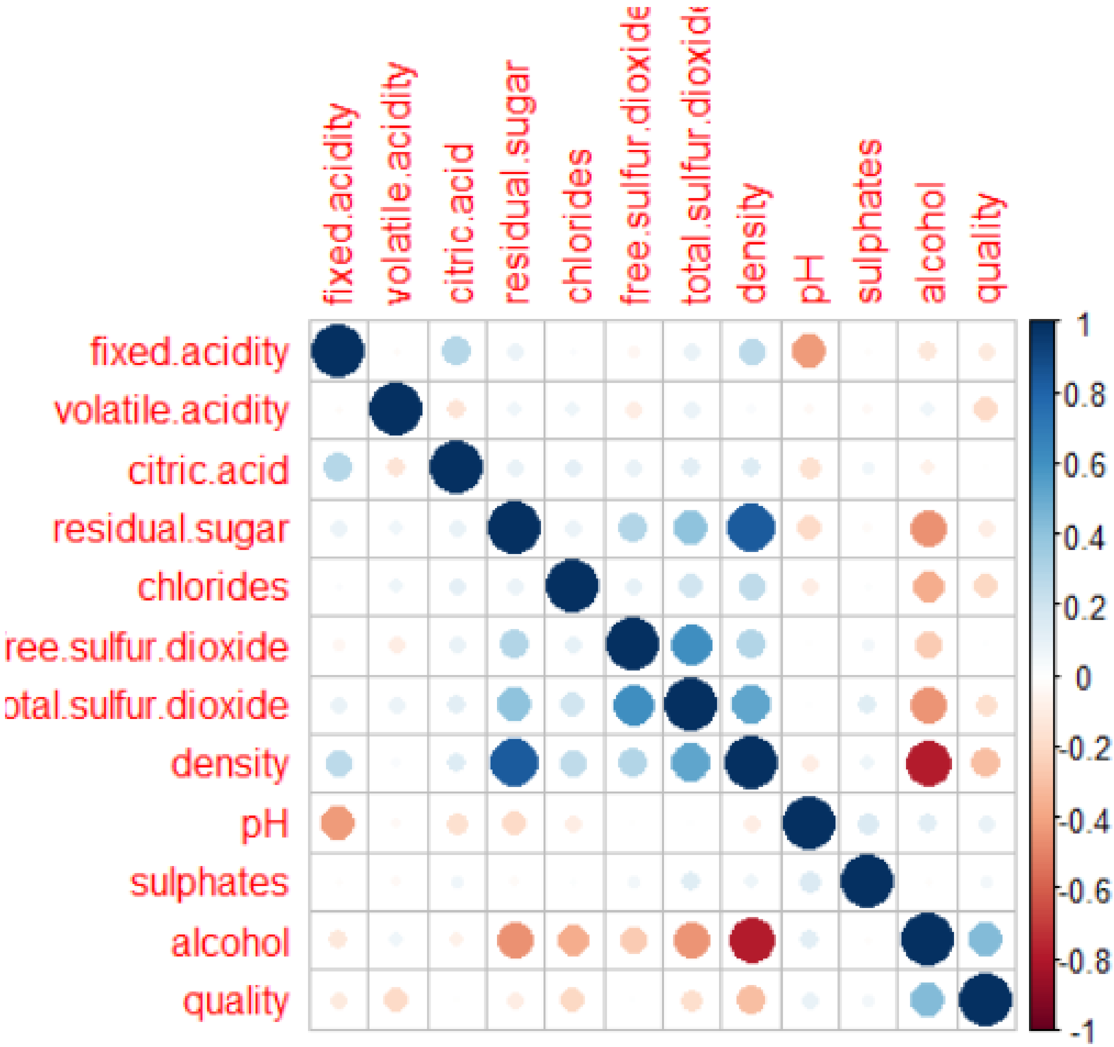


Figure 6. Summary of analysis of ordinal logistic regression model with all variables

```
Call:
polr(formula = quality ~ ., data = white, method = "logistic")

Coefficients:
                          Value Std. Error   t value
fixed.acidity         1.934e-01   0.041303    4.6815
volatile.acidity     -4.988e+00   0.335818  -14.8528
citric.acid           2.603e-01   0.266152    0.9780
residual.sugar        2.145e-01   0.007365   29.1324
chlorides            -4.814e-01   1.503450   -0.3202
free.sulfur.dioxide   1.253e-02   0.002437    5.1395
total.sulfur.dioxide -9.298e-04   0.001042   -0.8925
density              -4.199e+02   0.499422 -840.6750
pH                    1.998e+00   0.228231    8.7560
sulphates             1.640e+00   0.266349    6.1570
alcohol               4.818e-01   0.034270   14.0602

Intercepts:
    Value     Std. Error t value
3|4 -409.7061    0.5081  -806.4238
4|5 -407.3984    0.5060  -805.1111
5|6 -404.3373    0.5106  -791.8207
6|7 -401.7462    0.5207  -771.6097
7|8 -399.4904    0.5306  -752.8607
8|9 -395.9682    0.6903  -573.6027

Residual Deviance: 9262.297
AIC: 9296.297
```

Figure 7. Comparing association of variables with quality

```
Analysis of Deviance Table (Type II tests)

Response: quality
                     LR Chisq Df Pr(>Chisq)
fixed.acidity           9.318  1   0.002269 **
volatile.acidity      225.882  1  < 2.2e-16 ***
citric.acid            10.793  1   0.001019 **
residual.sugar         87.896  1  < 2.2e-16 ***
chlorides               0.099  1   0.753501
free.sulfur.dioxide    26.019  1  3.381e-07 ***
total.sulfur.dioxide    0.753  1   0.385668
density                47.593  1  5.246e-12 ***
pH                     42.929  1  5.678e-11 ***
sulphates              34.279  1  4.776e-09 ***
alcohol                35.089  1  3.150e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8. Slightly increase of residual deviance and AIC values in significant variable model.

```
Call:
polr(formula = quality ~ . - citric.acid - chlorides - total.sulfur.dioxide,
    data = white, method = "logistic")

Coefficients:
                          Value Std. Error  t value
fixed.acidity           0.03789   0.039465    0.960
volatile.acidity       -5.22554   0.322350  -16.211
residual.sugar          0.14034   0.007145   19.642
free.sulfur.dioxide     0.01132   0.001993    5.680
density              -214.09457   0.484513 -441.876
pH                      1.27784   0.224651    5.688
sulphates               1.31926   0.263501    5.007
alcohol                 0.73953   0.031510   23.470

Intercepts:
    Value     Std. Error t value
3|4 -206.4918    0.4913  -420.2695
4|5 -204.1862    0.4897  -417.0005
5|6 -201.1347    0.4949  -406.4061
6|7 -198.5616    0.5053  -392.9893
7|8 -196.3132    0.5154  -380.8615
8|9 -192.7926    0.6787  -284.0824

Residual Deviance: 9276.53
AIC: 9304.53
```

Figure 9. Significant association with quality scores

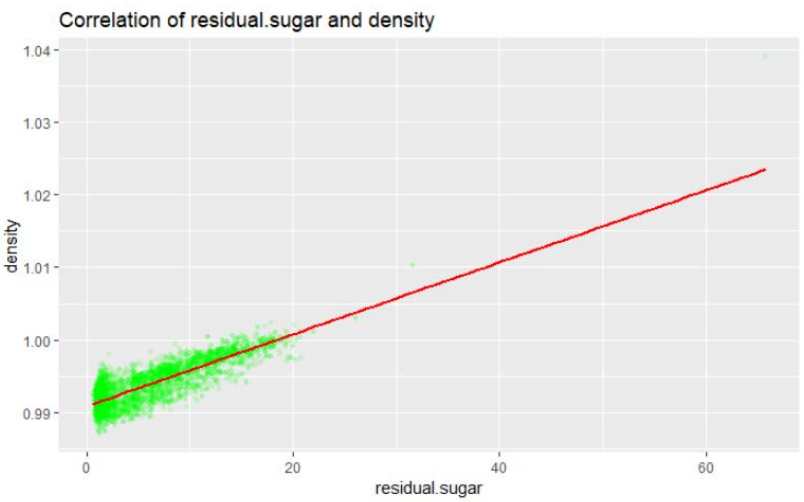
```
Analysis of Deviance Table (Type II tests)

Response: quality
                    LR Chisq Df Pr(>Chisq)
fixed.acidity         -0.059  1          1
volatile.acidity     249.421  1  < 2.2e-16 ***
residual.sugar        84.981  1  < 2.2e-16 ***
free.sulfur.dioxide   32.017  1  1.529e-08 ***
density               41.919  1  9.511e-11 ***
pH                    32.861  1  9.897e-09 ***
sulphates             31.408  1  2.091e-08 ***
alcohol               23.016  1  1.606e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10. Correlation matrix of wine quality data showing collinearity between variables

	fixed.acidity	volatile.acidity	residual.sugar	free.sulfur.dioxide
fixed.acidity	1.00000000	-0.01528755	0.08261316	-0.057891112
volatile.acidity	-0.01528755	1.00000000	0.07654702	-0.102564225
residual.sugar	0.08261316	0.07654702	1.00000000	0.295627045
free.sulfur.dioxide	-0.05789111	-0.10256423	0.29562705	1.000000000
density	0.25596307	0.03679084	0.84075694	0.291324526
pH	-0.41797426	-0.03372098	-0.19084394	0.006035243
sulphates	-0.02610350	-0.04875135	-0.03835415	0.060836551
alcohol	-0.11525555	0.06991631	-0.45192161	-0.252837543

	density	pH	sulphates	alcohol
fixed.acidity	0.25596307	-0.417974263	-0.026103499	-0.115255547
volatile.acidity	0.03679084	-0.033720982	-0.048751346	0.069916307
residual.sugar	0.84075694	-0.190843940	-0.038354146	-0.451921606
free.sulfur.dioxide	0.29132453	0.006035243	0.060836551	-0.252837543
density	1.00000000	-0.085761148	0.063213598	-0.777175970
pH	-0.08576115	1.000000000	0.163897891	0.115343593
sulphates	0.06321360	0.163897891	1.000000000	-0.009720521
alcohol	-0.77717597	0.115343593	-0.009720521	1.000000000

Figure 11. Scatter plot showing correlations between density and residual.sugar/alcohol



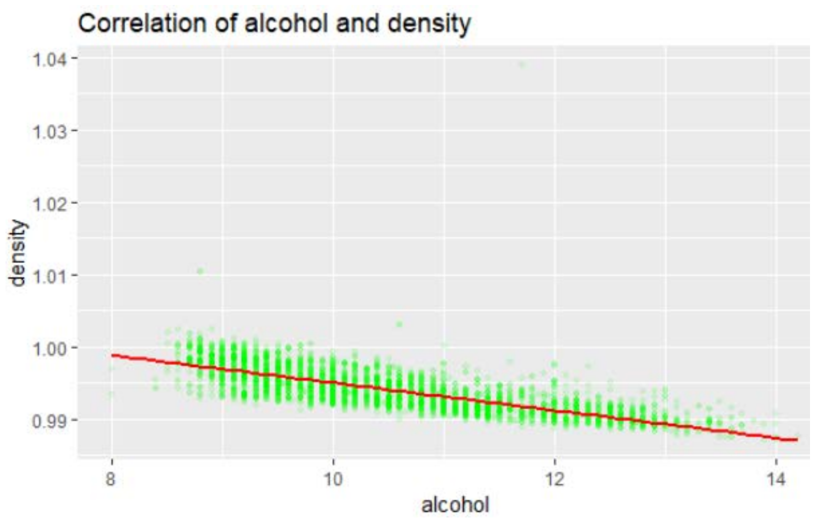


Figure 12. Justification collinearity of ordinal logistic model

```{r}
mod.fit.ord3 <- polr(formula = quality ~ .+fixed.acidity*pH+residual.sugar*alcohol+
residual.sugar*density+ density*alcohol-citric.acid-chlorides-total.sulfur.dioxide, data=white,
method= "logistic" )
summary(mod.fit.ord3)
```

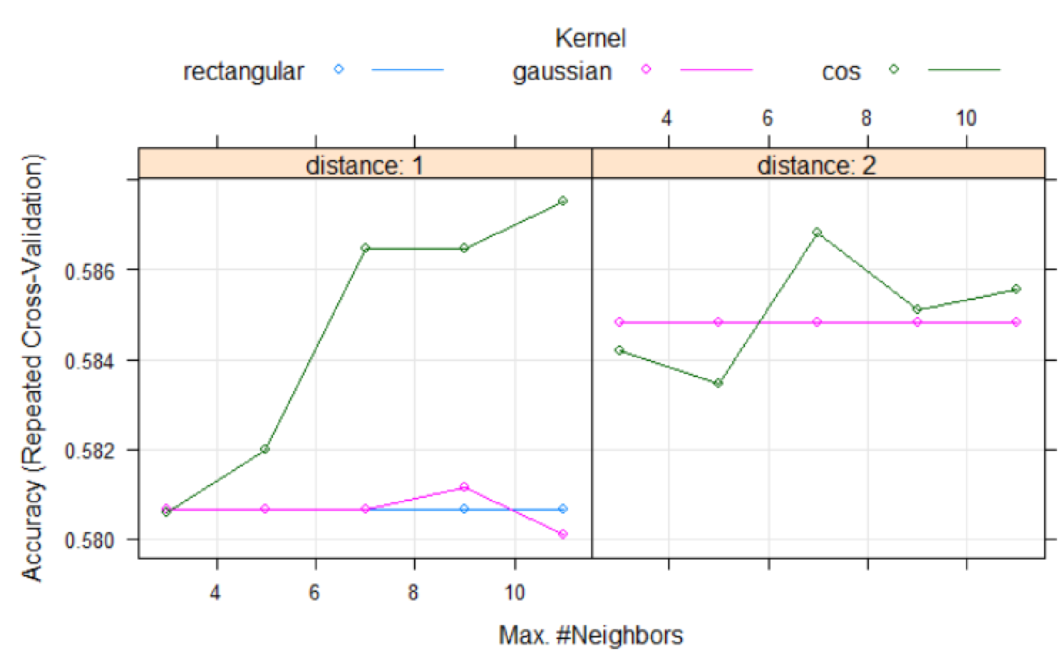
```
Re-fitting to get Hessian

Call:
polr(formula = quality ~ . + fixed.acidity * pH + residual.sugar *
    alcohol + residual.sugar * density + density * alcohol -
    citric.acid - chlorides - total.sulfur.dioxide, data = white,
    method = "logistic")

Coefficients:
                           Value Std. Error   t value
fixed.acidity           -1.43715   0.119787   -11.998
volatile.acidity        -5.67279   0.324432   -17.485
residual.sugar          -4.75987   0.646935    -7.358
free.sulfur.dioxide      0.01069   0.002006     5.330
density               1110.39960   0.064548 17202.625
pH                      -1.46549   0.161107    -9.096
sulphates                1.74953   0.265528     6.589
alcohol                146.59435   0.068390  2143.503
fixed.acidity:pH         0.52403   0.043548    12.033
residual.sugar:alcohol   0.07548   0.005636    13.394
residual.sugar:density   4.19085   0.646702     6.480
density:alcohol       -147.59519   0.069515 -2123.216

Intercepts:
    Value      Std. Error t value
3|4 1094.3246     0.0633 17282.1168
4|5 1096.6449     0.2205  4973.9804
5|6 1099.6763     0.2343  4693.2641
6|7 1102.2792     0.2431  4533.7568
7|8 1104.6236     0.2593  4259.5299
8|9 1108.1687     0.5117  2165.5535

Residual Deviance: 9259.631
AIC: 9295.631
```

Figure 13. Confusion Matrix of Model 3

```
Confusion Matrix and Statistics

    quality
pred    3    4    5    6    7    8    9
   3    0    1    0    0    0    0    0
   4    0    2    0    0    0    0    0
   5    5   79  590  321   31    8    0
   6   11   48  629 1390  537   86    1
   7    1    2   10  165  175   54    4
   8    0    0    0    0   12    1    0
   9    0    0    0    0    0    0    0

Overall Statistics

               Accuracy : 0.5184
                 95% CI : (0.5031, 0.5337)
    No Information Rate : 0.4506
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2183

 Mcnemar's Test P-Value : NA
```

```
                     Class: 3  Class: 4 Class: 5 Class: 6 Class: 7  Class: 8 Class: 9
Sensitivity          0.0000000 0.0151515   0.4801   0.7409  0.23179 0.0067114 0.000000
Specificity          0.9997588 1.0000000   0.8487   0.4263  0.93075 0.9970105 1.000000
Pos Pred Value       0.0000000 1.0000000   0.5706   0.5144  0.42579 0.0769231      NaN
Neg Pred Value       0.9959154 0.9687575   0.7958   0.6674  0.84542 0.9643373 0.998799
Prevalence           0.0040836 0.0317079   0.2952   0.4506  0.18136 0.0357915 0.001201
Detection Rate       0.0000000 0.0004804   0.1417   0.3339  0.04204 0.0002402 0.000000
Detection Prevalence 0.0002402 0.0004804   0.2484   0.6491  0.09873 0.0031227 0.000000
Balanced Accuracy    0.4998794 0.5075758   0.6644   0.5836  0.58127 0.5018609 0.500000
```

Figure 14. knn full model




```{r}
kknn.train$bestTune
```

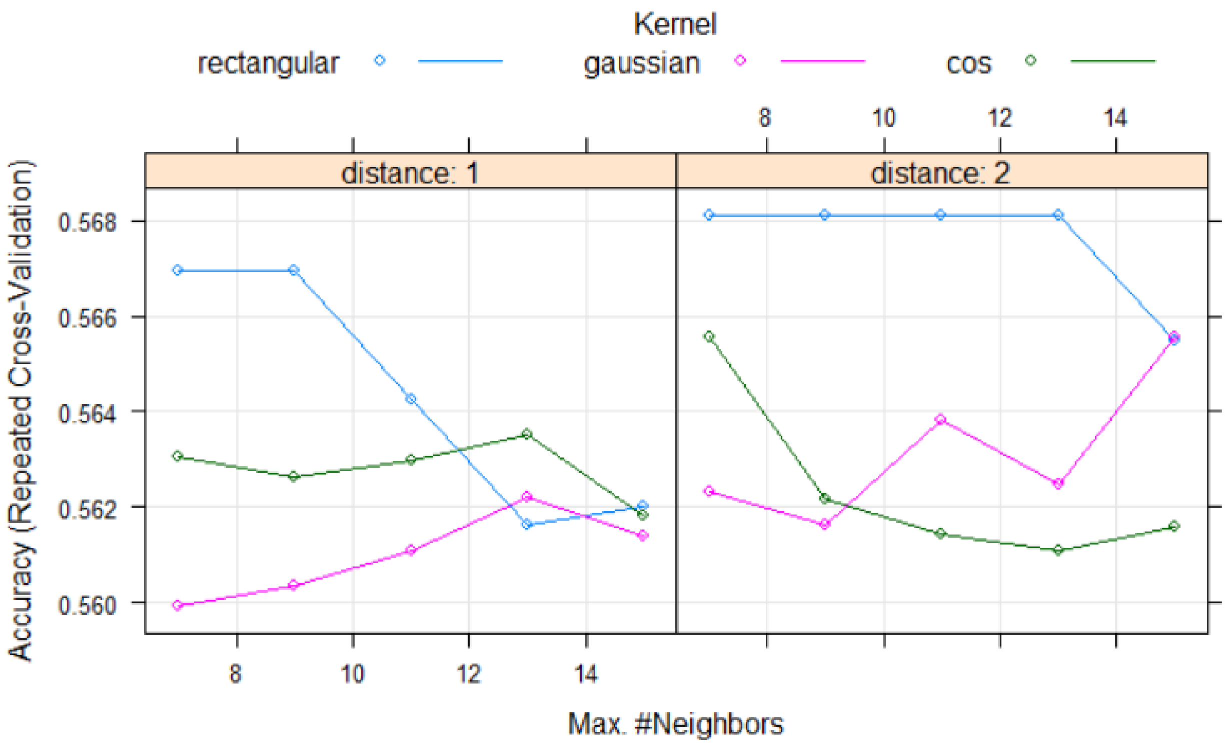
	kmax	distance	kernel
	<dbl>	<dbl>	<fctr>
27	11	1	cos

1 row

Figure 15. Confusion matrix and statistics of k-NN full model

```
Confusion Matrix and Statistics

          Reference
Prediction   3   4   5   6   7   8   9
         3   0   0   0   0   0   0   0
         4   0  11  12  10   0   0   0
         5   5  29 312 138  15   3   1
         6   1  14 149 500 102  14   0
         7   0   0  12  74 160  18   0
         8   0   0   0  10  16  23   0
         9   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.6176
                 95% CI : (0.5935, 0.6412)
    No Information Rate : 0.4494
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4223
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8  Class: 9
Sensitivity          0.000000 0.203704   0.6433   0.6831  0.54608  0.39655 0.0000000
Specificity          1.000000 0.986032   0.8330   0.6878  0.92216  0.98345 1.0000000
Pos Pred Value            NaN 0.333333   0.6203   0.6410  0.60606  0.46939       NaN
Neg Pred Value       0.996317 0.973058   0.8464   0.7267  0.90256  0.97785 0.9993861
Prevalence           0.003683 0.033149   0.2977   0.4494  0.17986  0.03560 0.0006139
Detection Rate       0.000000 0.006753   0.1915   0.3069  0.09822  0.01412 0.0000000
Detection Prevalence 0.000000 0.020258   0.3088   0.4788  0.16206  0.03008 0.0000000
Balanced Accuracy    0.500000 0.594868   0.7382   0.6855  0.73412  0.69000 0.5000000
```

Figure 16. knn reduced model




```{r}
kknn.train$bestTune
```

	kmax <dbl>	distance <dbl>	kernel <fctr>
22	13	2	rectangular

Table 18A. Compare major parameters of different models

	KNN full model	KNN reduced	Ordered log. Full
Accuracy	0.62	0.60	0.53
95% CI	0.59-0.64	0.58-0.62	0.50-0.55
P Value	<2.2e-16	<2.2e-16	<3.86e-9
Kappa value	0.42	0.40	0.23

Table 18B. Compare statistics by classes in different models

Statistics by Class:

KNN Full Model

	Class: 3	Class: 4	Class: 5	Class: 6	Class: 7	Class: 8	Class: 9
Sensitivity	0.000000	0.203704	0.6433	0.6831	0.54608	0.39655	0.0000000
Specificity	1.000000	0.986032	0.8330	0.6878	0.92216	0.98345	1.0000000
Pos Pred Value	NaN	0.333333	0.6203	0.6410	0.60606	0.46939	NaN
Neg Pred Value	0.996317	0.973058	0.8464	0.7267	0.90256	0.97785	0.9993861
Prevalence	0.003683	0.033149	0.2977	0.4494	0.17986	0.03560	0.0006139
Detection Rate	0.000000	0.006753	0.1915	0.3069	0.09822	0.01412	0.0000000
Detection Prevalence	0.000000	0.020258	0.3088	0.4788	0.16206	0.03008	0.0000000
Balanced Accuracy	0.500000	0.594868	0.7382	0.6855	0.73412	0.69000	0.5000000

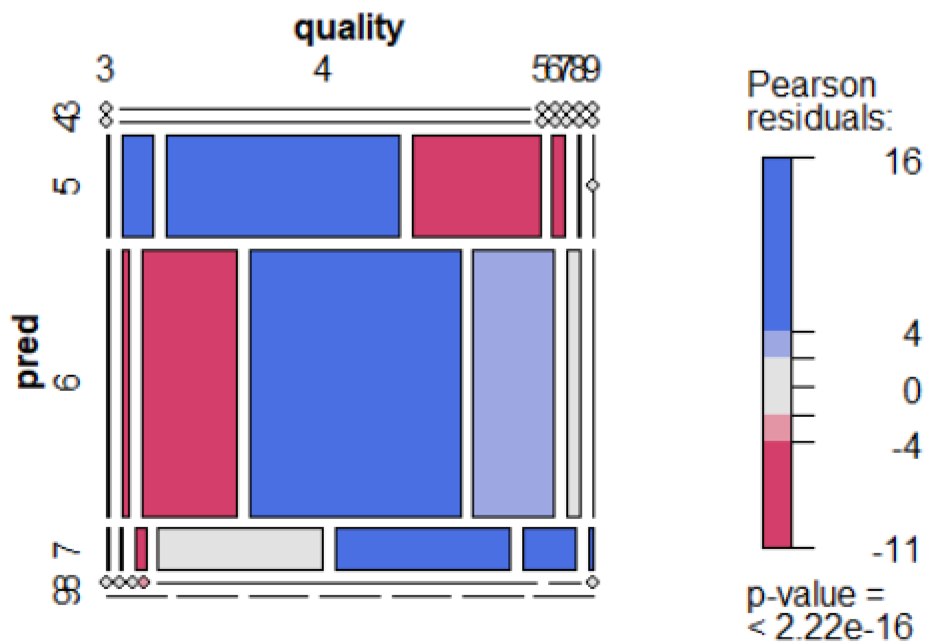
KNN Reduced Model

	Class: 3	Class: 4	Class: 5	Class: 6	Class: 7	Class: 8	Class: 9
Sensitivity	0.000000	0.33333	0.6103	0.6462	0.5563	0.46552	0.0000000
Specificity	0.996919	0.97714	0.8295	0.6979	0.9169	0.97836	1.0000000
Pos Pred Value	0.000000	0.33333	0.6029	0.6358	0.5949	0.44262	NaN
Neg Pred Value	0.996305	0.97714	0.8339	0.7073	0.9041	0.98023	0.9993861
Prevalence	0.003683	0.03315	0.2977	0.4494	0.1799	0.03560	0.0006139
Detection Rate	0.000000	0.01105	0.1817	0.2904	0.1001	0.01657	0.0000000
Detection Prevalence	0.003069	0.03315	0.3014	0.4567	0.1682	0.03745	0.0000000
Balanced Accuracy	0.498460	0.65524	0.7199	0.6720	0.7366	0.72194	0.5000000

Ordered Logistic Regression

	Class: 3	Class: 4	Class: 5	Class: 6	Class: 7	Class: 8	Class: 9
Sensitivity	0.000000	0.0208333	0.5058	0.7545	0.20677	0.00000	0.000000
Specificity	1.000000	0.9985935	0.8364	0.4243	0.95100	1.00000	1.000000
Pos Pred Value	NaN	0.3333333	0.5619	0.5192	0.48246	NaN	NaN
Neg Pred Value	0.995238	0.9679618	0.8031	0.6772	0.84440	0.96463	0.998639
Prevalence	0.004762	0.0326531	0.2932	0.4517	0.18095	0.03537	0.001361
Detection Rate	0.000000	0.0006803	0.1483	0.3408	0.03741	0.00000	0.000000
Detection Prevalence	0.000000	0.0020408	0.2639	0.6565	0.07755	0.00000	0.000000
Balanced Accuracy	0.500000	0.5097134	0.6711	0.5894	0.57888	0.50000	0.500000

Figure 19. Mosaic plot of model 3



Conclusion

- Quality of wine can be predicted by the following variables: "alcohol", "residual.sugar", "pH", "fixed.acidity", "volatile.acidity" and "free.sulfur.dioxide".
- Outlinears and collinearity are needed to be justified for ordinal logistic regression models but not knn models.
- K-nn models performs better than ordinal logistic regression models because most of the independent variables are not linear related to the response variable.
- All the ordinal regression models and knn models are failed to predict wine quality scores of "3" and "9" due to lack of cases fit into that two category.
- Although collapsed categories can improve prediction accuracy, it loses prediction power. It is not wise to collapsed the categories in this study.
- A large discrepancy between observed and expected values is due to the blind spots of models. There are still room for improve the knn models. But the computational cost will be increased significantly.
-

Review of Literature

- Review other study using the same database, Lemionet used knn, weighted linear regression, additive logistic regression, they found additive logistic regression had least test error. They believed that additive logistic regression does better at leveraging the ordinal structure of the data and hence produces better results. As for weighted linear regression, they noted that weighted linear regression performed well when the number of predictors was small. In the case of 10 variables, the predictor space may be too sparse to generate good results. (This can also be explained by the curse of dimensionality). Because they did not use the same methods such as overall accuracy, sensitivity and specificity to evaluate their model, we could not make comparison with the models.

- Uniyall et al reported using machine learning algorithm to build a linear regression model based on this database. We feel that they used the wrong model for their study. How could they use a linear regression model for an ordinal response variable? Also, we could not find they had treat the outliers and collinearity between the lines of their paper. They neither provided ROC curve, nor provided any detailed predictor between each quality scores. We felt that it was a poor written manuscript.

- Cortez et al had spent significant amount effort to use neural network and small vector model to build a couple of prediction models on wine database. The overall accuracy were slight higher than our knn models. However, they combined 8/9 of wine quality score together might be the reason outperform accuracy than our models. We tried some other model such as random forest, which got similar results as our knn. Because those model were quite time consuming, we did not dig deeper this time. We were failed to perform SVM because our computer stopped running after a couple hours computation. In his later part of his report, he claimed that he had improved accuracy around 90%. But he was failed to provide detailed information for us to repeat his model.

- Based on our preliminary analysis of the white wine dataset and review of the literature, we feel: (1) small vector model, random forest, knn models are better predict the wine quality because most the variables are not linear correlated; (2) the computer expense are enormous because the nature of model; (3) those three models may be more practical on red wine dataset due to the number of observers in the dataset.