

# STAT 6340 Statistical Machine Learning

## Mini Project 2

*Author:*

Suarez Rodes, Randy

*Supervisor:*

Choudhary, Pankaj Ph.D.

February 22, 2021



# 1 Answers

## Question 1.a

We first explore the correlations of our variables, both predictors and response on Figure 1. Visualizing the correlation matrix we can see that predictors Oakiness and Clarity have a very low correlation with Quality. We infer that these variables are not much relevant to predict Quality and are the most likely to be removed from future models. On the other hand predictors Aroma, Flavor and Region has a high correlation with Quality, making them good candidates for our models.

Another point to notice is that variable Body presents some correlation with Quality, although, it is not as high as other predictors, therefore we require further analysis to determine its relevance for the response variable. Finally, predictors Aroma, Body, Flavor and Region are highly correlated. This can cause overfitting issues, since the data would be over explained, so with further analysis one or some of this predictors might be dropped since they can be explained by other predictors.

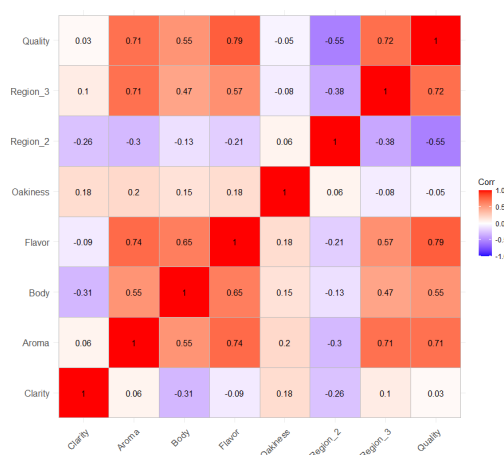


Figure 1: Correlation Matrix

We present the scatterplot of our variables on Figure 2. As expected from the correlation matrix we can notice that the predictors Aroma, Body and Flavor exhibit a positive linear relation with Quality.

On Figure 3 we present the Quality boxplots by Region. Pay attention to the clear distribution of Quality by Region (high on 3, medium on 1, low on 2). This fact evidences the importance of this predictor for determining the Quality of a wine.

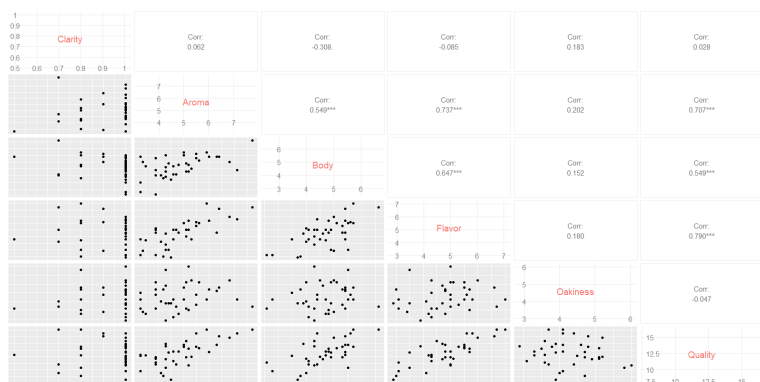


Figure 2: Scatter Plots

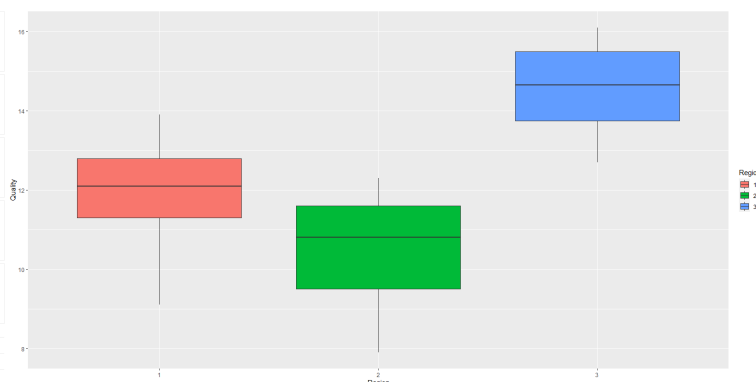


Figure 3: Quality by Region

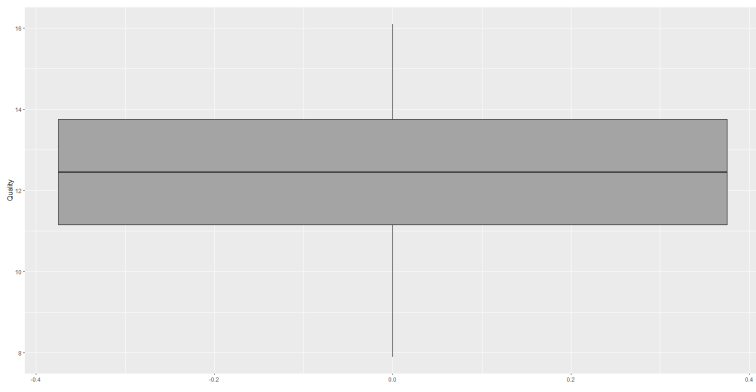


Figure 4: Quality Boxplot

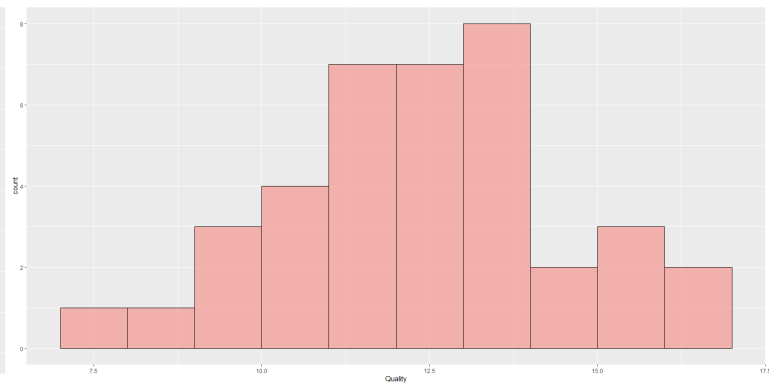


Figure 5: Quality Histogram

### Question 1.b

We explore the variable Quality using boxplots to identify possible outliers (Figure 4). We did not detect any outlier. We also explore the underlying distribution of Quality using histograms (Figure 5). We can appreciate that data distribution looks approximately normal. At this moment we consider that transformation is not required and we will check model assumptions once we build an appropriate model.

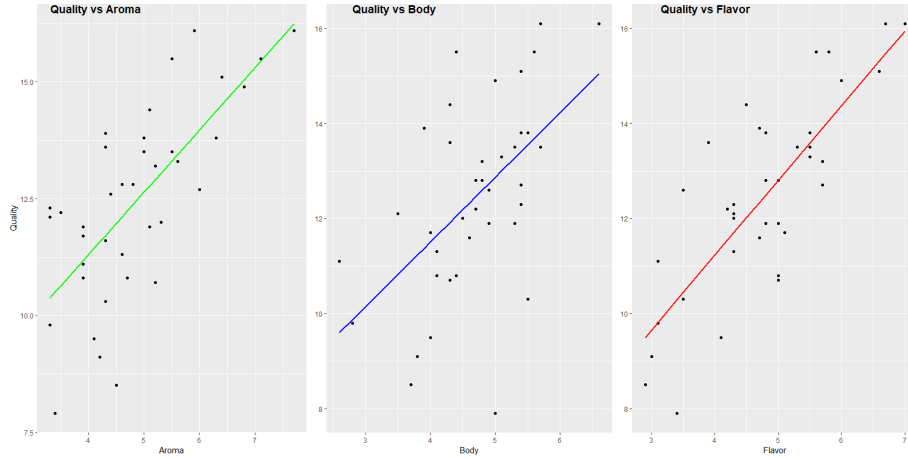


Figure 6: Quality vs Continuous Predictors

**Question 1.c** We fitted a simple linear regression model for every predictor (Tables 1 to 6). Taking a look to the p-values of every model we can affirm that there is a statistically significant association between the predictor and the response variable on all models except for the models with predictors Clarity (p-value = 0.865) and Oakiness (p-value = 0.779). This is consistent with our previous exploration. We can see on Figure 6 how there is a clear linear relation for predictors Flavor, Aroma and Body. Also we discussed on question 1.a how Region is relevant to predict Quality (Figure 3).

Table 1:

	Estimate	$\Pr(> t )$
(Intercept)	12.0034	3.89e-5
Clarity	0.4692	0.865

Table 2:

	Estimate	$\Pr(> t )$
(Intercept)	5.9583	4.51e-6
Aroma	1.3365	6.87e-7

Table 3:

	Estimate	$\Pr(> t )$
(Intercept)	6.0580	0.0007
Body	1.3618	0.0004

Table 4:

	Estimate	$\Pr(> t )$
(Intercept)	4.9414	1.57e-5
Flavor	1.5719	3.68e-9

Table 5:

	Estimate	$\Pr(> t )$
(Intercept)	12.9916	1.4e-7
Oakiness	-0.1304	0.779

Table 6:

	Estimate	$\Pr(> t )$
(Intercept)	11.9765	2e-16
Region2	-1.5320	0.00757
Region3	2.6069	7.01e-6

### Question 1.d

From the summary of the full model we can reject the null hypothesis for the

Flavor and Region predictors, meaning that this predictors are relevant in order to predict Quality.

Table 7: Full Model

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.8144	1.9694	3.97	0.0004
Clarity	0.0171	1.4563	0.01	0.9907
Aroma	0.0890	0.2525	0.35	0.7269
Body	0.0797	0.2677	0.30	0.7681
Flavor	1.1172	0.2403	4.65	6.25e-5
Oakiness	-0.3464	0.2330	-1.49	0.1475
Region2	-1.5129	0.3923	-3.86	0.0006
Region3	0.9726	0.5102	1.91	0.0662

Table 8: Resulting Model

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.0943	0.7912	8.97	1.76e-10
Flavor	1.1155	0.1738	6.42	2.49e-07
Region2	-1.5335	0.3688	-4.16	0.0002
Region3	1.2234	0.4003	3.06	0.0043

### Question 1.e

To start building our model, we will proceed to drop one variable at a time from the full model until all predictors remaining are relevant for our response variable. Starting from the highest p-value predictor and rechecking the test hypothesis results for every new model.

After this process we ended up with the model  $\text{Quality} \sim \text{Flavor} + \text{Region}$  (Table 8) where all variables are . The anova test (Table 9) between the full model and the current model confirms our findings. With a p-value of 0.6528, we fail to reject the null hypothesis, this is all extra predictors are 0.

Table 9: Anova with Full Model

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	34	27.21				
2	30	25.14	4	2.07	0.62	0.6528

Table 10: Anova vs Interactions

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	34	27.21				
2	32	25.43	2	1.78	1.12	0.3378

Now we will proceed to explore interactions between these two predictors.

We will compare the model  $\text{Quality} \sim \text{Flavor} + \text{Region}$  with the model  $\text{Quality} \sim \text{Flavor} + \text{Region} + \text{Flavor}:\text{Region}$ . By carrying out an anova test we obtained a p-value of 0.3378, meaning that we fail to reject the null hypothesis, this is all additional predictors are 0. Therefore the interactions between Flavor and Region are not meaningful for our model.

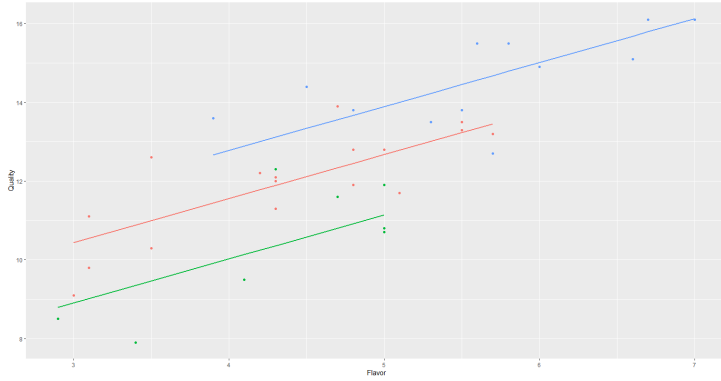


Figure 7: Model Assumptions

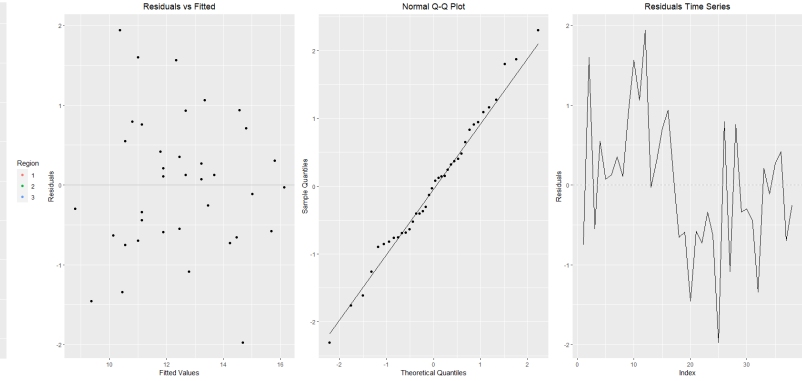


Figure 8: Model Assumptions

We can appreciate the three resulting regression lines of our model for each region on Figure 7. Clearly our model captures the linear trend on each region. Also we present several diagnostics plots on Figure 8 in order to check for model assumptions.

We can see in the Residuals vs Fitted plot how the points are scattered around zero and there is no a clear pattern, verifying that our errors have an approximate zero mean and constant variance. For the Q-Q Plot, observe how for an exception of a few points the errors are close to be normally distributed. Finally

checking on the Residuals Time Series, there is not a clear dependence of the residuals, verifying the independence assumption.

#### Question 1.f

Final model:

$$\text{Quality} = 7.0943 + 1.1155(\text{Flavor}) - 1.5335\mathbb{1}_{(\text{Region}2)} + 1.2234\mathbb{1}_{(\text{Region}3)}$$

We can interpret our model as follows: Approximately every level change in Flavor will directly change the Quality by one level. Region 1 wine has base quality of 7, Region 3 offers an additional level (quality base around 8) and Region 2 a decrease of 1.5 (quality base approximately of 5.5).

#### Question 1.g

Our prediction of Quality for a wine from Region 1 with a mean Flavor is: 12.41371.

We obtained a 95% confidence interval of [11.95152, 12.8759], this is that given an average Flavor wine and assuming it is from Region 1, 95% of intervals of this form will contain the true value of the wine Quality.

Also we obtained a 95% prediction interval of [10.53775, 14.28967], meaning that we can be 95% confident that the Quality of an average Flavor wine from Region 1 will lie in this range.

#### Question 2.a

On Figure 9 we present several plots to explore how the GMAT and GPA can help to predict the group of an applicant. We can observe in the graph of GMAT vs GPA, how the GPA predictor can clearly separate the three groups. Moreover in the GPA boxplot we have a clear distribution of the GPA by group, evidence of how much this predictor contributes to identify the correct group. For the GMAT boxplot there is some confusion for applicants of group 2 and 3, but a clear distinction for applicants of group 1, meaning that GMAT also helps us characterizes the possible group of an applicant.

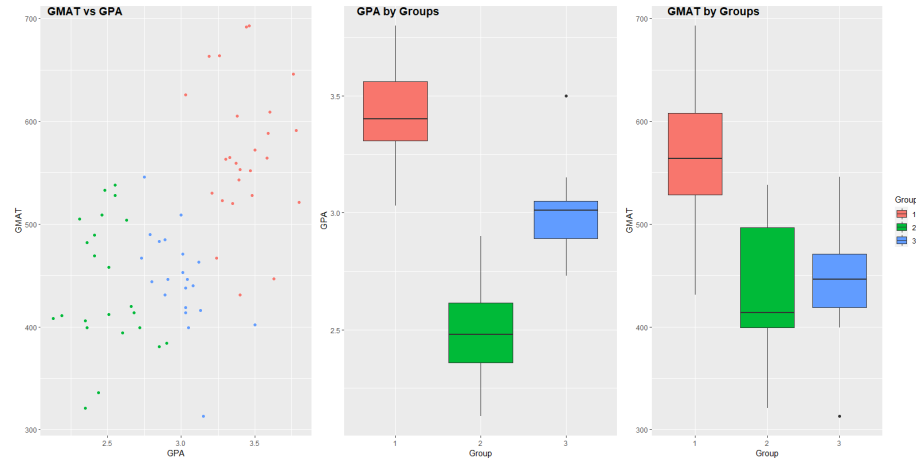


Figure 9:

#### Question 2.b

We present the confusion matrix for training and testing sets on Tables 11 and 12 obtained from the LDA.

#### Question 2.c

#### Question 2.d

Table 11: LDA Training  
Actual

		1	2	3
	1	24	0	1
	2	0	21	1
Predicted	3	2	2	19

Class Error Rate: 0.08571

Table 12: LDA Testing  
Actual

		1	2	3
	1	2	0	0
	2	0	5	0
Predicted	3	3	0	5

Class Error Rate: 0.2

Table 13: QDA Training  
Actual

		1	2	3
	1	26	0	1
	2	0	22	0
Predicted	3	0	1	20

Class Error Rate: 0.02857

Table 14: QDA Testing  
Actual

		1	2	3
	1	4	0	0
	2	0	5	0
Predicted	3	1	0	5

Class Error Rate: 0.06667

**Question 3.a**

**Question 3.b**

**Question 3.c**

**Question 3.d**

## 2 Code