Analysis of wine data to predict quality

A mock article for a Quarto course

Tiny van Boekel Jos Hageman

Introduction

This mock paper is used as an example for the WGS course Quarto. The goal is to show how to integrate text, calculations, import of data, data wrangling, calculations/data processing, graphic output and literature references.

This example is about the prediction of quality of wine using simple to measure variables. The original data are from Cortez et al. (2009). In this short note the data are characterized and a relation is investigated.

Material and Methods

This dataset is related to Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). The data has 1599 different sample of wine. Each wine sample has been characterized on 12 variables.

Results and discussion

Characterization of the data

As an example of descriptive statistics, Table 1 shows a summary of all the variables in the wine data set. Note that the median is displayed instead of the mean.

Boxplots are helpful to show the distributional aspects of the variables: see, for instance, Figure 1 for the variation in total SO_2 which shows the enormous variation in this variable; the figure also shows that the variation in alcohol content is not so large. Such plots could be made for every variable, of course.

Table 1: Statistical overview of the wine data set

Characteristic	**N = 1,599**
fixed.acidity	7.90 (7.10, 9.20)
volatile.acidity	0.52 (0.39, 0.64)
citric.acid	0.26 (0.09, 0.42)
residual.sugar	2.20 (1.90, 2.60)
chlorides	0.079 (0.070, 0.090)
free.sulfur.dioxide	14 (7, 21)
total.sulfur.dioxide	38 (22, 62)
density	0.9968 (0.9956, 0.9978)
pH	3.31 (3.21, 3.40)
sulphates	0.62 (0.55, 0.73)
alcohol	10.20 (9.50, 11.10)
quality	
3	10 (0.6%)
4	53 (3.3%)
5	681 (43%)
6	638 (40%)
7	199 (12%)
8	18 (1.1%)

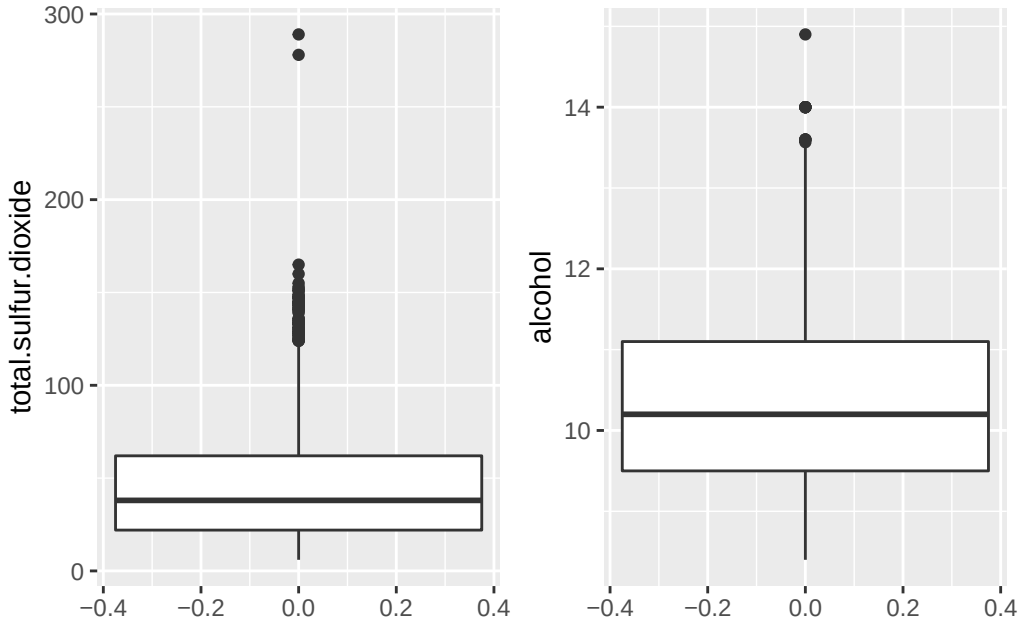

Figure 1: Boxplots showing the variation in total sulfur dioxide and alcohol in wine

An important variable is quality. It may be worthwhile to see its variation in a bar plot (quality is given as discrete values), see Figure 2. It seems as if the data are reasonably normally distributed, and so the wines are not rated as very bad or as very good, but in between.

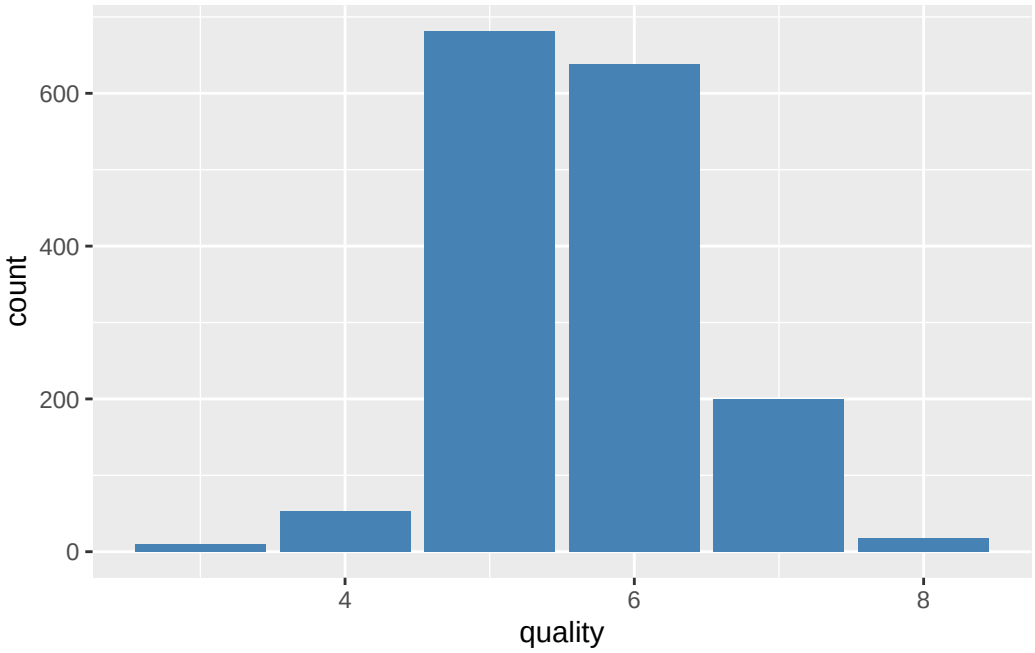

Figure 2: Bar plot of the variable quality in wines

Quality is what it is all about, so it could be interesting to see, for instance, whether or not quality is related to alcohol content. Quality rankings were discrete, so a boxplot can be made with quality as categorical factor and alcohol content as the variable: see Figure 3. There seems to be a trend that wine quality is rated higher when the alcohol content is higher than about 11%.

Analysis of the data

In the next step we would like to build a regression model that predicts wine quality using the other variables. The output in Table 2 gives the metabolites that are significantly related to quality rating. There was no check for co-linearity problems so some metabolites could be masked. This is an example of multiple regression.

Before putting some trust in the model, we need to check if we have colinearity problems, which might be a serious problem when doing multiple linear regression. One way of doing this is by calculating the VIF (Variance Inflation Factors) values for each variable, which can be done

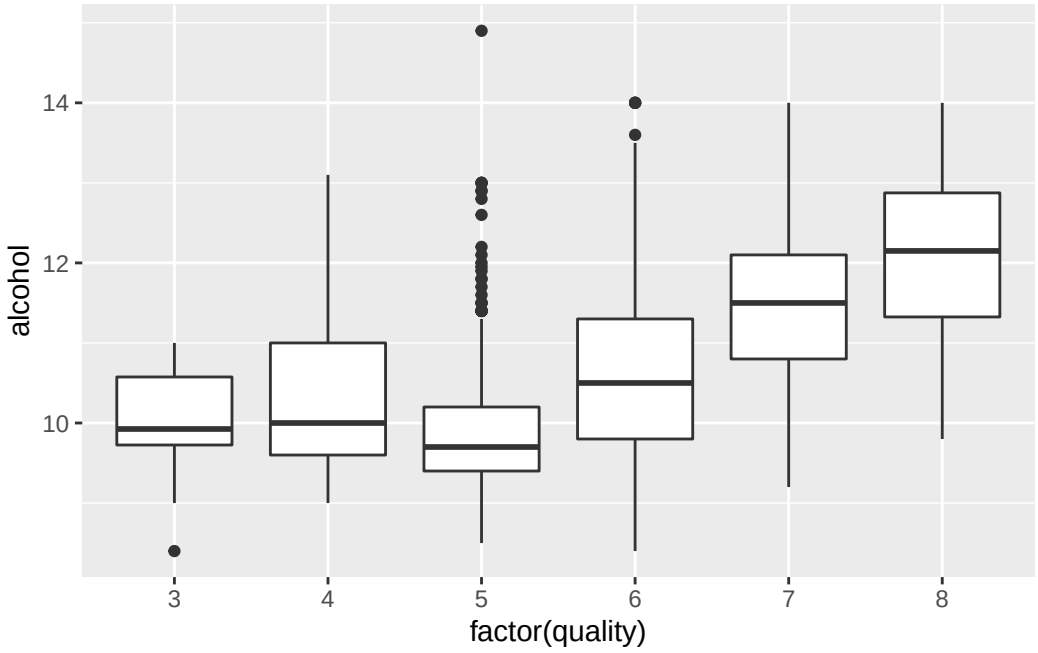

Figure 3: Boxplot of alcohol content in relation to quality gradings of wine

Table 2: output of multiple linear regression of quality rating versus predictor variables

Characteristic	**Beta**	**95% CI**	**p-value**
fixed.acidity	0.02	-0.03, 0.08	0.3
volatile.acidity	-1.1	-1.3, -0.85	<0.001
citric.acid	-0.18	-0.47, 0.11	0.2
residual.sugar	0.02	-0.01, 0.05	0.3
chlorides	-1.9	-2.7, -1.1	<0.001
free.sulfur.dioxide	0.00	0.00, 0.01	0.044
total.sulfur.dioxide	0.00	0.00, 0.00	<0.001
density	-18	-60, 25	0.4
pH	-0.41	-0.79, -0.04	0.031
sulphates	0.92	0.69, 1.1	<0.001
alcohol	0.28	0.22, 0.33	<0.001

Table 3: VIF values for the predictor values for quality ratings of wine

	x
fixed.acidity	7.767825
volatile.acidity	1.789396
citric.acid	3.127999
residual.sugar	1.702744
chlorides	1.481956
free.sulfur.dioxide	1.964930
total.sulfur.dioxide	2.188749
density	6.343991
pH	3.329862
sulphates	1.429417
alcohol	3.031168

Table 4: Table showing which predictor variables have a significant influence on wine quality rating

	Estimate	Std. Error	t value	Pr(>|t|)
volatile.acidity	-1.0833957	0.1211000	-8.946291	0.0000000
chlorides	-1.8745843	0.4192812	-4.470947	0.0000083
free.sulfur.dioxide	0.0043843	0.0021723	2.018291	0.0437290
total.sulfur.dioxide	-0.0032702	0.0007290	-4.485745	0.0000078
pH	-0.4139764	0.1915987	-2.160643	0.0308722
sulphates	0.9162862	0.1143354	8.014023	0.0000000
alcohol	0.2762029	0.0264833	10.429330	0.0000000

with a function from the R package `car`. The results are shown in Table 3. As long as VIF values are below 10, we do not have a serious co-linearity problem, though some researchers put the limit already at a value of 5.

Not all predictor variables in the model are significant. Table 4 shows which ones are significant, i.e., with a p-value < 0.05.

To get an impression whether or not the data are approximately normal (actually the residuals), a normal distribution can be imposed on the data: see Figure 4. It shows that there is a small deviation from normality at the tail of the distribution but not disturbingly so.

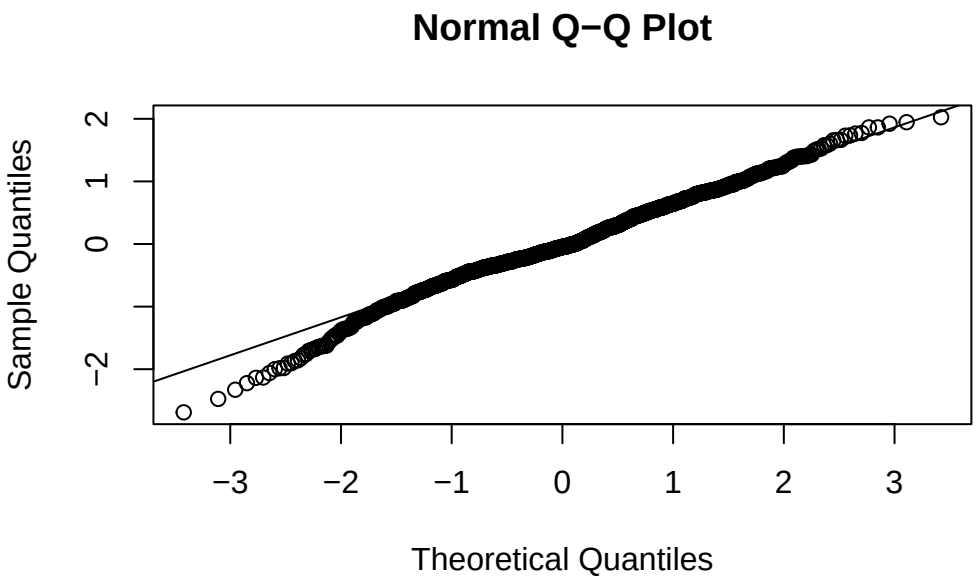

Figure 4: Normal Q-Q plot of the residuals

Conclusion

In this short note, a model was created that predicts quality for wines. A multiple linear regression model was created to predict quality. This model has an R^2 of 0.36. This shows that Taste Sweet can be explained to a certain extent using the 12 parameters measured on these wines. R^2 is not that high that it may be a useful thing to apply this model. The QQ plot showed that the assumption of normal distributed residuals was approximately fulfilled. Not all predictor variables appeared to be significant, but some interesting variables were pinpointed that may be interesting for follow up research.

References

Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." *Decision Support Systems* 47 (4): 547–53. https://doi.org/10.1016/j.dss.2009.05.016.