Group_14_Analysis

Introduction

We are investigating what affects the quality of coffee. We will be using generalised linear models to analyse our data.

Here is the brief description of the data:

Variable Name	Description
country_of_origin	Country where the coffee bean originates from
aroma	Aroma grade (ranging from 0-10)
flavor	Flavour grade (ranging from 0-10)
acidity	Acidity grade (ranging from 0-10)
category_two_defects	Count of category 2 type defects in the batch of coffee beans tested
altitiude_mean_meters	Mean altitude of the growers farm (in metres)
harvested	Year the batch was harvested
Qualityclass	Quality score for the batch.

Note: About Qualityclass, 82.5 was selected as the cut off as this is the median score for all the batches tested.

We got a response variable, QualityClass and seven explanatory variables. We will perform various analysis and plots to see what the best model for this data is.

Summaries

We have cleaned the dataset and deleted all the empty (NA) values. We are left with 926 entries. *Table 2* below shows the distribution of all the numerical variables of the data we are using.

Table 2: Summary statistics of the coffee data

Variable	n	Mean	SD	Min	Median	Max	IQR
aroma	926	7.57	0.40	0	7.58	8.75	0.17
flavor	926	7.52	0.42	0	7.58	8.83	0.17
acidity	926	7.53	0.40	0	7.50	8.75	0.25
category_two_defects	926	3.52	5.24	0	2.00	47.00	2.00
altitude_mean_meters	926	1656.88	7180.23	1	1310.64	190164.00	289.36
harvested	926	2013.72	1.81	2010	2014.00	2018.00	1.00

We have plotted the explanatory variables to be able to see their distribution more clearly before starting the statistical models.

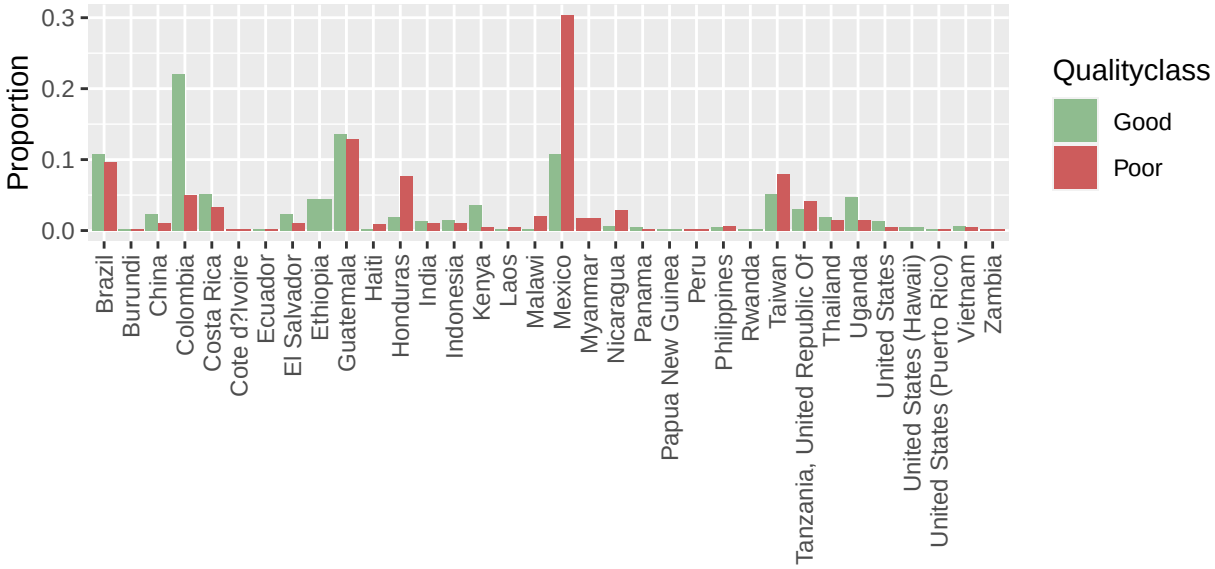

Figure 1: Barplot of country of origin

Figure 1 shows the relation between coffee quality and country of origin. It's easy to see than a part from Colombia, which has a lot of the good quality coffee, and Mexico which has a lot of low quality coffee, all the countries seem to have a similar proportion of both coffee qualities.

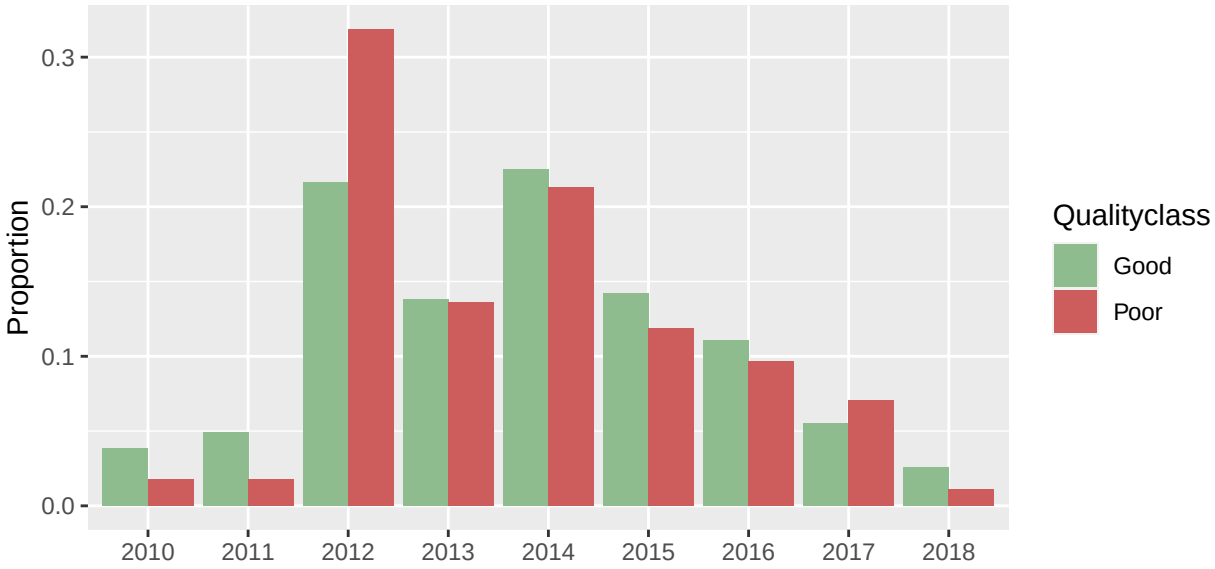

Figure 2: Barplot of year harvesteds

Figure 2 shows the proportion of coffee quality class in every year harvested. The does not seem to be a relation between they year and the quality if coffee a part from 2012 which has a big proportion of low quality coffee.

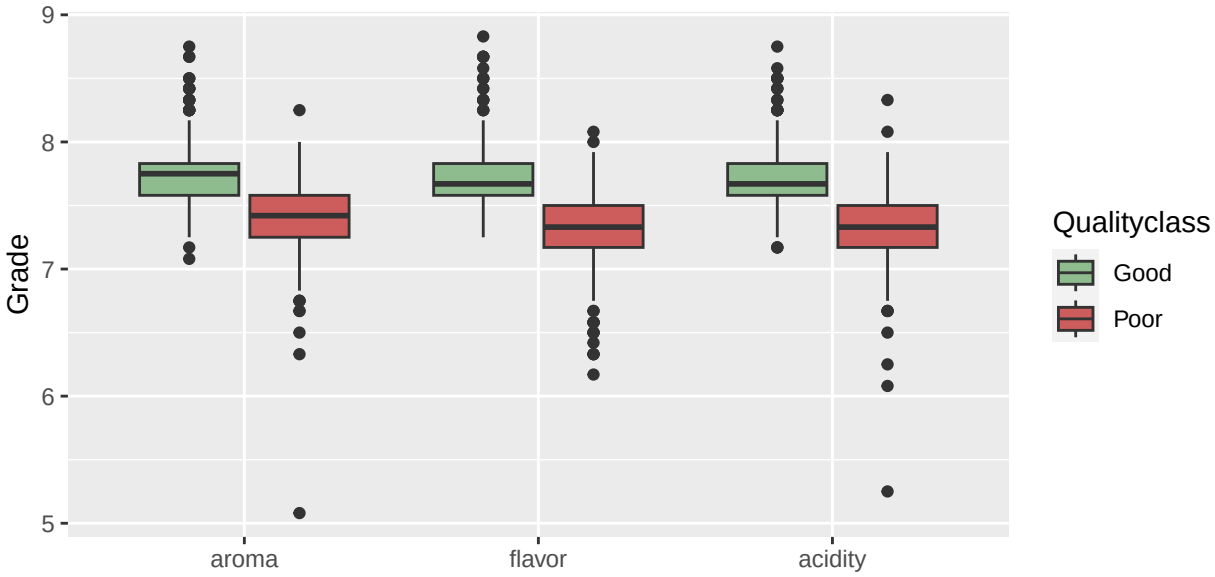

Figure 3: Boxplots of coffee qualities

Figure 3 show three boxplots, one for each of the graded variables, aroma, flavour and acidity; it looks like all the variables have plenty of outliers, this has to be had in account for the analysis. There seems to be a very obvious relation between the grade of the variables and the coffee quality, with all the variables having higher grades in the good quality coffee than in the poor quality coffee.

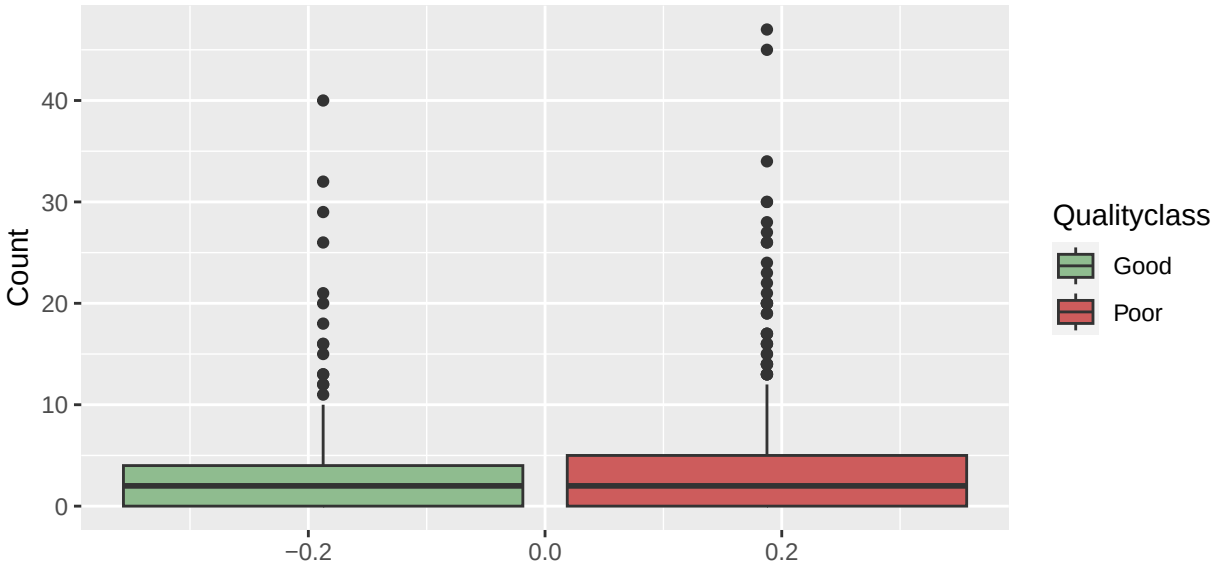

Figure 4: Boxplot of cateogry two defects

Figure 4 shows a boxplot comparing the number of category two defects by quality class. It doesn't look to be any relationship between the number of category two defects and quality of coffee. This variable has a lot of outliers in both of the quality classes, this is tobe had into account when performing the formal analysis.

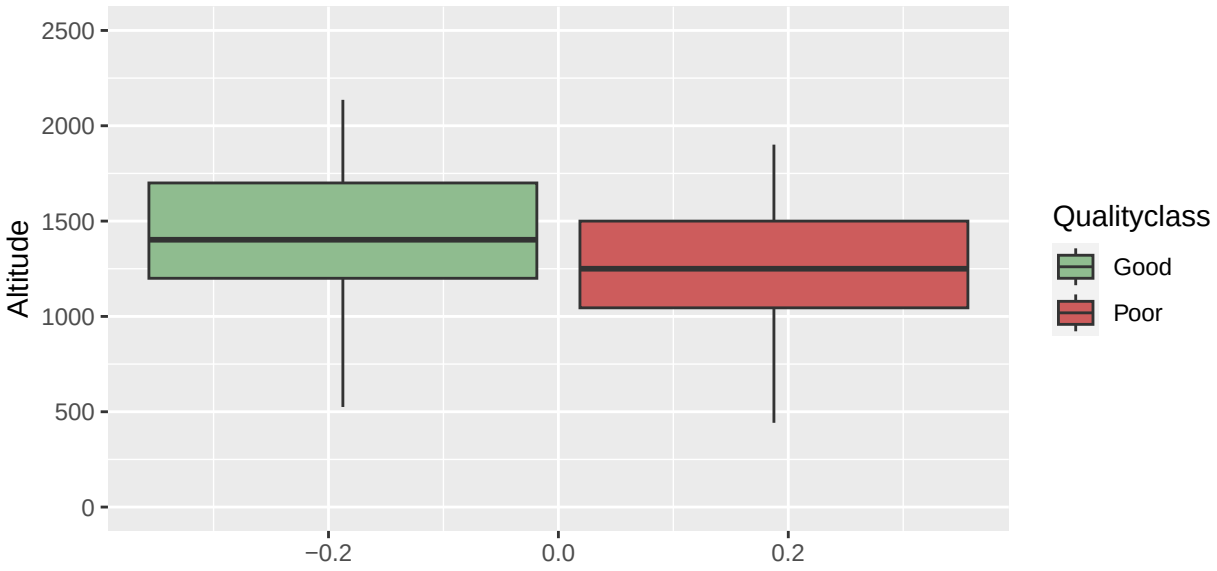

Figure 5: Boxplot of altitude

Figure 5 is the last plot of the explanatory variables, it compares the altitude of the coffee farms by quality class. It looks like there is not a big difference between both boxplots, but good quality coffee has slightly higher values than poor quality coffee.

Correlation analysis

Next we perform a correlation matrix, shown in *Figure 6* on the numerical variables to see how they are related to each and be able to make a better decision about what models are best to fit. We have not used the country of origin as from the plots and summaries there didn't seem to be a relation between this variable and quality of coffee.

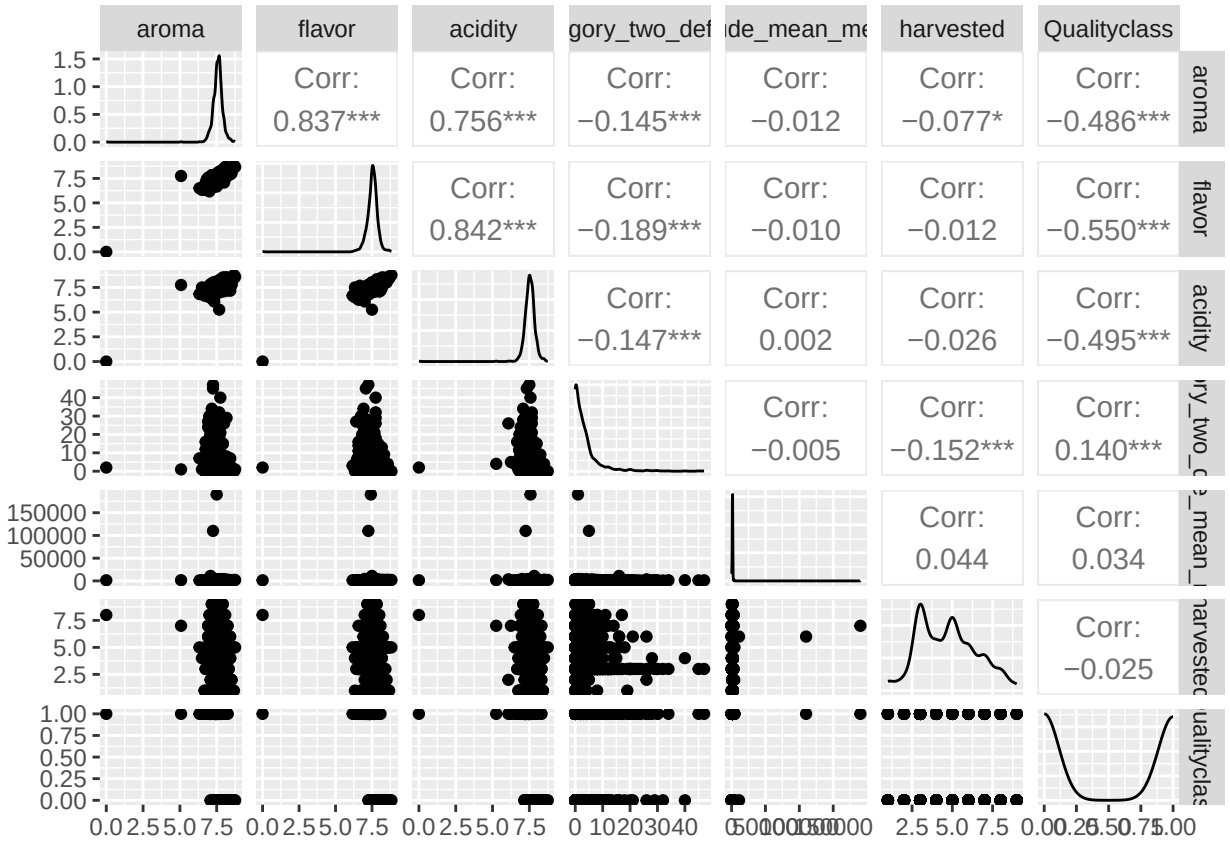

Figure 6: Correlation between each variables

Figure 6 shows there is strong correlation between the first three variables: aroma, flavor and acidity. In addition, aroma, flavor and acidity have medium correlation with Quality Class, category is in weak correlation. However, the correlation of the last two variables are strongly weak.

Formal Analysis

GLM Part 1

The strong correlation shows that we need can do further data cleaning. For instance, deleting rows with null values and converting quality of coffee from Poor and Good to a numerical variable with 0 and 1. We have also separated the dataset into two categories, the first one with the variables with high correlation values we have called this dataset coffee1_1 and the second with the uncorrelated variables, called coffee1_2.

Model 1 : The first three variables with principal component analysis and 3 variables.

We performed a principal components analysis on coffee1_1.

```
## Importance of components:
##                            Comp.1    Comp.2     Comp.3
## Standard deviation     1.6198441 0.4944852 0.36275257
## Proportion of Variance 0.8746317 0.0815052 0.04386314
## Cumulative Proportion  0.8746317 0.9561369 1.00000000
```

As we can see from the table above, the cumulative proportion of the first component is 0.8746, so we can only choose the first component.

Trying: 1st glm function

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = Qualityclass ~ PC1 + category_two_defects + altitude_mean_meters +
##     harvested, family = binomial(link = "cloglog"), data = final_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7770  -0.4827  -0.0138   0.2034   3.7660
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.016e-01  2.132e-01  -0.945    0.344
## PC1                  -4.237e+00  2.833e-01 -14.955   <2e-16 ***
## category_two_defects -7.312e-03  1.787e-02  -0.409    0.682
## altitude_mean_meters  6.180e-06  7.953e-06   0.777    0.437
## harvested            -3.712e-02  3.868e-02  -0.960    0.337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1283.4  on 925  degrees of freedom
## Residual deviance:  533.6  on 921  degrees of freedom
## AIC: 543.6
##
## Number of Fisher Scoring iterations: 9
```

Model 2 : All variables with principal component analysis.

We performed a principal components analysis on coffee1_3.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.749350 1.0734284 0.9976865 0.9042951 0.77722534
## Proportion of Variance 0.437175 0.1646069 0.1421969 0.1168214 0.08629703
## Cumulative Proportion  0.437175 0.6017819 0.7439788 0.8608002 0.94709722
##                             Comp.6     Comp.7
## Standard deviation     0.49245507 0.35750169
## Proportion of Variance 0.03464457 0.01825821
## Cumulative Proportion  0.98174179 1.00000000
```

As we can see from the table above, the cumulative proportion of the fourth component is 0.8608002, so we can only choose the first four component.

Trying: 2st glm function

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = V5 ~ RC1 + RC2 + RC4 + RC3, family = binomial(link = "logit"),
##     data = final_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3701  -0.4274  -0.0028   0.3495   4.3489
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.18990    0.11127   1.707   0.0879 .
## RC1         -6.07330    0.42747 -14.208   <2e-16 ***
## RC2         -0.14806    0.10444  -1.418   0.1563
## RC4         -0.10265    0.14475  -0.709   0.4782
## RC3          0.06929    0.14167   0.489   0.6248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1283.43  on 925  degrees of freedom
## Residual deviance:  545.25  on 921  degrees of freedom
## AIC: 555.25
##
## Number of Fisher Scoring iterations: 7
```

All of these two models do not fit well, and the correlation figure shows that the last three variables have really weak relationship with Quality Class, so in this model, we try to fit the third model by deleting harvested, category and altitude.

Model 3 : First 3 variables with principal component analysis.

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3
## Standard deviation     1.6198441 0.4944852 0.36275257
## Proportion of Variance 0.8746317 0.0815052 0.04386314
## Cumulative Proportion  0.8746317 0.9561369 1.00000000
```

As we can see from the table above, the cumulative proportion of the second component is 0.9561369, so we can only choose the first two components.

Trying: 3rd glm function

```
##
## Call:
## glm(formula = V3 ~ RC1 + RC2, family = binomial(link = "logit"),
##     data = final_3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2960  -0.4380  -0.0026   0.3756   4.3049
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1857     0.1100   1.689   0.0912 .
## RC1          -4.1619     0.3364 -12.372   <2e-16 ***
## RC2          -2.2199     0.2550  -8.704   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1283.43  on 925  degrees of freedom
## Residual deviance:  547.78  on 923  degrees of freedom
## AIC: 553.78
##
## Number of Fisher Scoring iterations: 7
```

The p-values presented in the table above are all less than 0.05, indicating that the third model fits well and that the first three variables are the most important factors influencing the goodness of the fit.

Model 4 : First 3 variables and last one with principal component analysis.

As seen in the boxplots and the correlation matrix there is a relationship between altitude and Quality Class, therefore we have performes a PCA using aroma, flavour, acidity and altitude and fitted a linear model with its components.

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4
## Standard deviation     1.6198697 1.0000405 0.49434177 0.3627220
## Proportion of Variance 0.6559945 0.2500203 0.06109345 0.0328918
## Cumulative Proportion  0.6559945 0.9060147 0.96710820 1.0000000
```

The PCA has cumulative proportion of 0.967 in the third component so we have fitted a GLM wiht the first 3 components.

The GLM has low p-values in all the components but one so it is not the model we have fitted.

GLM Part 2

Even though the correlation matrix showed a high correlation between the variables aroma, flavor and acidity, we have perfored GLMs with them without performing PCA.

We have done this given the quote by Applied Linear Statistical Models, p289, 4th Edition; *"The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new observations."*

We have removed all the outliers of our dataframe, as there looked to be a lot of them in the boxplots.

Model 1

We have fitted a model with aroma, flavor, acidity, category two defects, year harevested and altitude as explanatory variables

Observations	738
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(6)$	567.76
Pseudo-R^2 (Cragg-Uhler)	0.72
Pseudo-R^2 (McFadden)	0.56
AIC	463.05
BIC	495.28

	Est.	S.E.	z val.	p
(Intercept)	130.35	10.13	12.87	0.00
aroma	-5.26	0.85	-6.21	0.00
flavor	-7.73	0.98	-7.87	0.00
acidity	-4.04	0.73	-5.52	0.00
category_two_defects	-0.05	0.05	-0.86	0.39
harvested	-0.13	0.06	-2.07	0.04
altitude_mean_meters	-0.00	0.00	-2.28	0.02

Standard errors: MLE

Model 2

As the p-value for category two defects in the model is greater than 0.05, we have fitted a new GLm without it as an explanory variable.

Observations	738
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(5)$	567.01
Pseudo-R^2 (Cragg-Uhler)	0.72
Pseudo-R^2 (McFadden)	0.56
AIC	461.80
BIC	489.42

	Est.	S.E.	z val.	p
(Intercept)	129.67	10.07	12.88	0.00
aroma	-5.19	0.84	-6.18	0.00
flavor	-7.73	0.98	-7.87	0.00
acidity	-4.03	0.73	-5.51	0.00
harvested	-0.13	0.06	-2.08	0.04
altitude_mean_meters	-0.00	0.00	-2.30	0.02

Standard errors: MLE

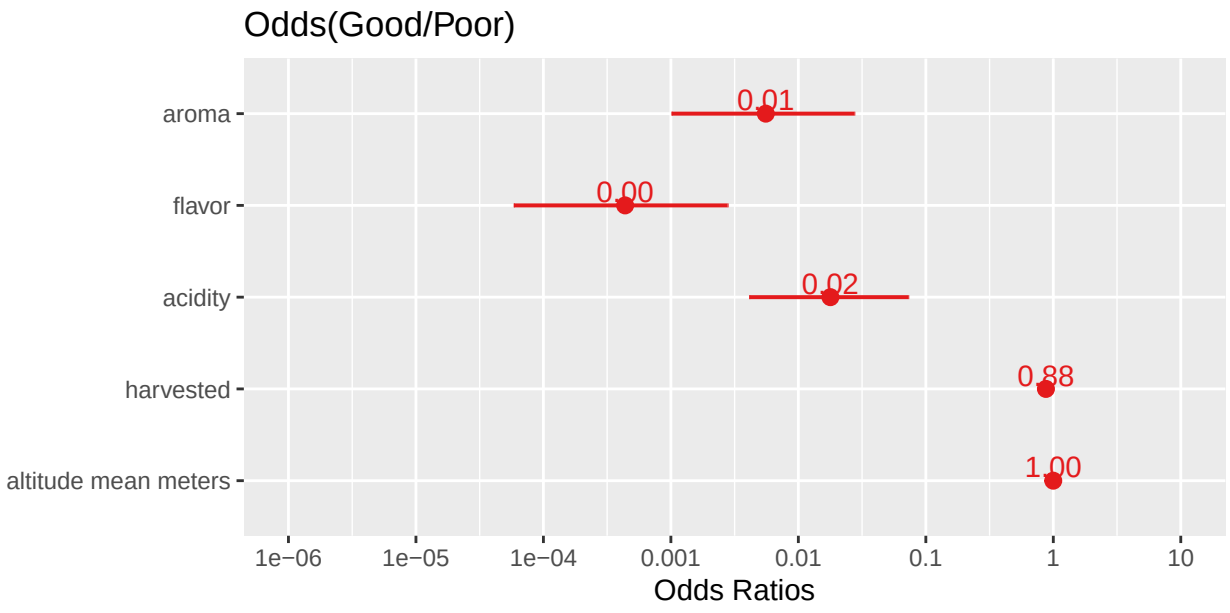

Figure 7: Odds ratio for Model 2

Model 3

As the p-value for category two defects in the model is greater than 0.05, we have fitted a new GLm without it as an explanory variable.

Observations	738
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(4)$	562.64
Pseudo-R^2 (Cragg-Uhler)	0.71
Pseudo-R^2 (McFadden)	0.55
AIC	464.17
BIC	487.19

	Est.	S.E.	z val.	p
(Intercept)	127.50	9.90	12.88	0.00
aroma	-5.00	0.83	-6.03	0.00
flavor	-7.72	0.98	-7.88	0.00
acidity	-4.05	0.73	-5.54	0.00
altitude_mean_meters	-0.00	0.00	-2.12	0.03

Standard errors: MLE

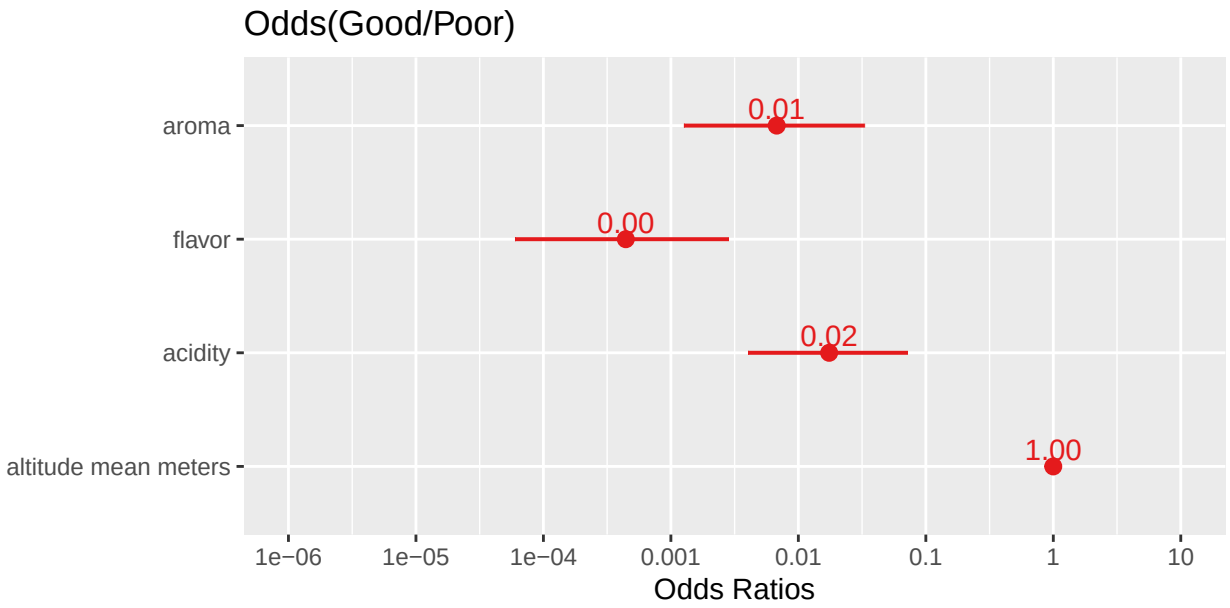

Figure 8: Odds ratio for Model 3

This last model has all p-values under 0.05. If we compare the AIC and BIC of the models, model 2 has a lower AIC but model 3 has a lower BIC. This suggests that both models are good and that we can use either of them.