Test pro 2

Yuqi Pan

1 Data description

Research Question: What influence do different features of coffee have on whether the quality of a batch of coffee is classified as good or poor?

Response variable:

`Qualityclass`: Quality score for the batch (Good >=82.5, Poor <82.5). Note: 82.5 was selected as the cut off as this is the median score for all the batches tested.

Explanatory variables:

- `country_of_origin`: Country where the coffee bean originates from.
- `aroma`: Aroma grade(ranging from 1-10)
- `flavor`: Flavor grade(ranging from 1-10)
- `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

```
## Rows: 1,145
## Columns: 8
## $ country_of_origin    <chr> "Myanmar", "Uganda", "Ethiopia", "Mexico", "Burun~
## $ aroma                <dbl> 7.25, 8.33, 8.42, 7.17, 7.75, 7.92, 7.92, 7.83, 7~
## $ flavor               <dbl> 7.42, 7.92, 8.00, 7.08, 7.67, 7.75, 7.83, 7.67, 6~
## $ acidity              <dbl> 7.50, 7.92, 8.00, 7.25, 7.50, 7.75, 7.67, 7.58, 7~
## $ category_two_defects <int> 4, 1, 7, 3, 5, 0, 1, 2, 2, 1, 0, 8, 0, 2, 0, 0, 2~
## $ altitude_mean_meters <dbl> 1219.20, 1600.00, 1700.00, 1300.00, 1880.00, 1400~
## $ harvested            <int> 2015, 2013, 2014, 2012, 2012, 2014, NA, 2015, 201~
## $ Qualityclass         <chr> "Poor", "Good", "Good", "Poor", "Good", "Good", "~
```

2 Explanatory Data Analysis

The summary statistics are tabled below:

Table 1: Summary statistics on all numerical variables

Variables	Missing	Mean	SD	Min	Median	Max
aroma	0	7.57	0.39	0	7.58	8.75
flavor	0	7.52	0.40	0	7.58	8.67
acidity	0	7.54	0.39	0	7.50	8.58
category_two_defects	0	3.67	5.41	0	2.00	55.00
altitude_mean_meters	201	1850.69	9392.09	1	1310.64	190164.00
harvested	60	2013.67	1.81	2010	2014.00	2018.00

From the summary statistics, we noticed that there are many missing data in altitude mean and harvested year.

- Replace `altitude` with mean

```
## [1] 1450
```

```
##          skim_variable n_missing
## 1 altitude_mean_meters         0
```

- Replace `harvested` with median

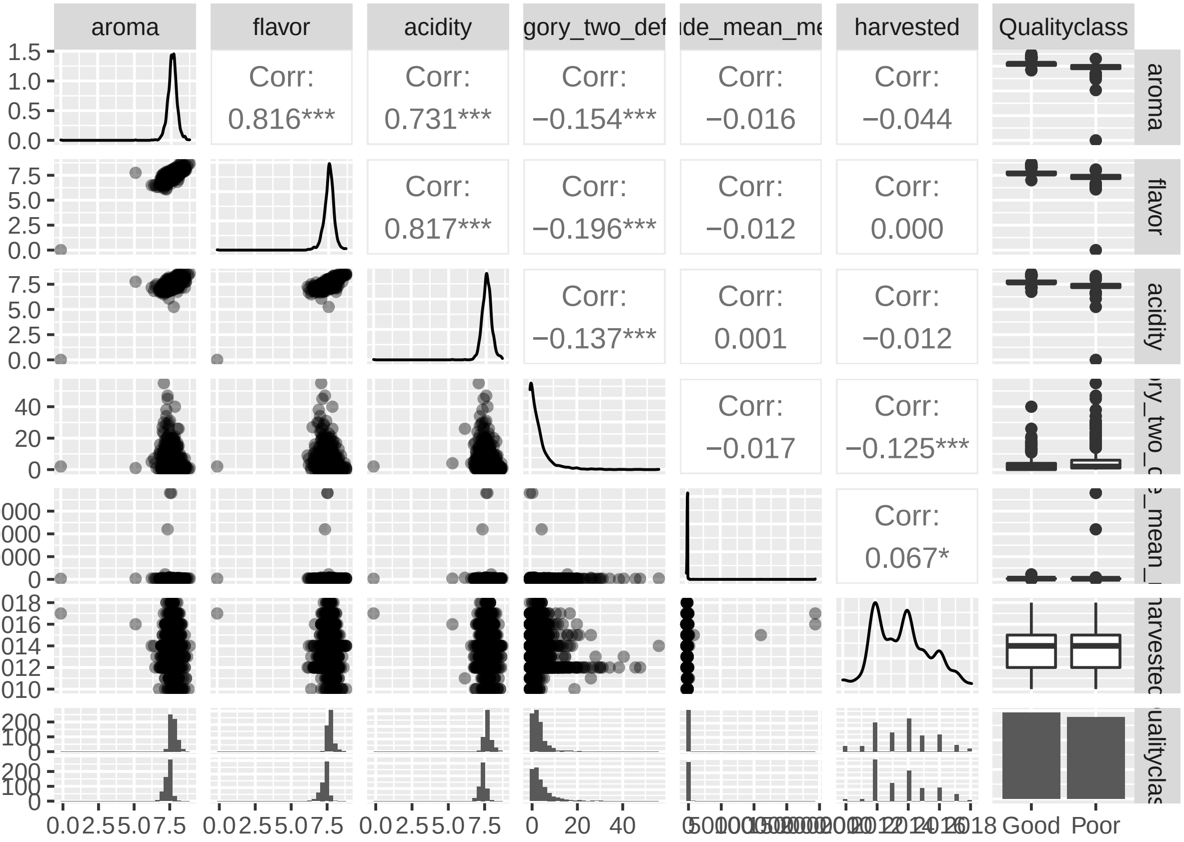
Since the mean and median of harvested is almost the same, which are 2013.67 and 2014 respectively. We replace the missing value of harvested with the median `2014`.

```
##   skim_variable n_missing
## 1     harvested         0
```

2.1 Visualize the data

Let’s visualize the data first:

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



We can notice from the correlation coefficient that there is a strong relationship between acidity and flavor of 0.817. Harvested and altitude mean are slightly related with quality class. *Also aroma, flavor, acidity, category two defects, and altitude are skew*

3 Formal Analysis

3.1 Model 1

Firstly, we need to split the data as training data and test data to test how the model are working aiming to choose the model with the best performance.

```
##
## Call:
## glm(formula = Qualityclass ~ ., family = binomial(link = "logit"),
##     data = training.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4123  -0.4058   0.0010   0.3530   3.7518
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.868e+02  1.595e+02   1.798   0.0721 .
## aroma                -5.020e+00  8.540e-01  -5.878 4.16e-09 ***
## flavor              -6.935e+00  1.057e+00  -6.559 5.41e-11 ***
## acidity             -4.504e+00  7.750e-01  -5.812 6.17e-09 ***
## category_two_defects  2.036e-02  3.119e-02   0.653   0.5138
## altitude_mean_meters  9.141e-06  2.174e-05   0.420   0.6742
## harvested           -8.067e-02  7.849e-02  -1.028   0.3041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 792.85  on 571  degrees of freedom
## Residual deviance: 328.48  on 565  degrees of freedom
## AIC: 342.48
##
## Number of Fisher Scoring iterations: 7
```

```
## Waiting for profiling to be done...
```

	2.5 %	97.5 %
(Intercept)	-24.745	602.443
aroma	-6.769	-3.413
flavor	-9.115	-4.958
acidity	-6.070	-3.020
category_two_defects	-0.043	0.080
altitude_mean_meters	0.000	NA
harvested	-0.235	0.073

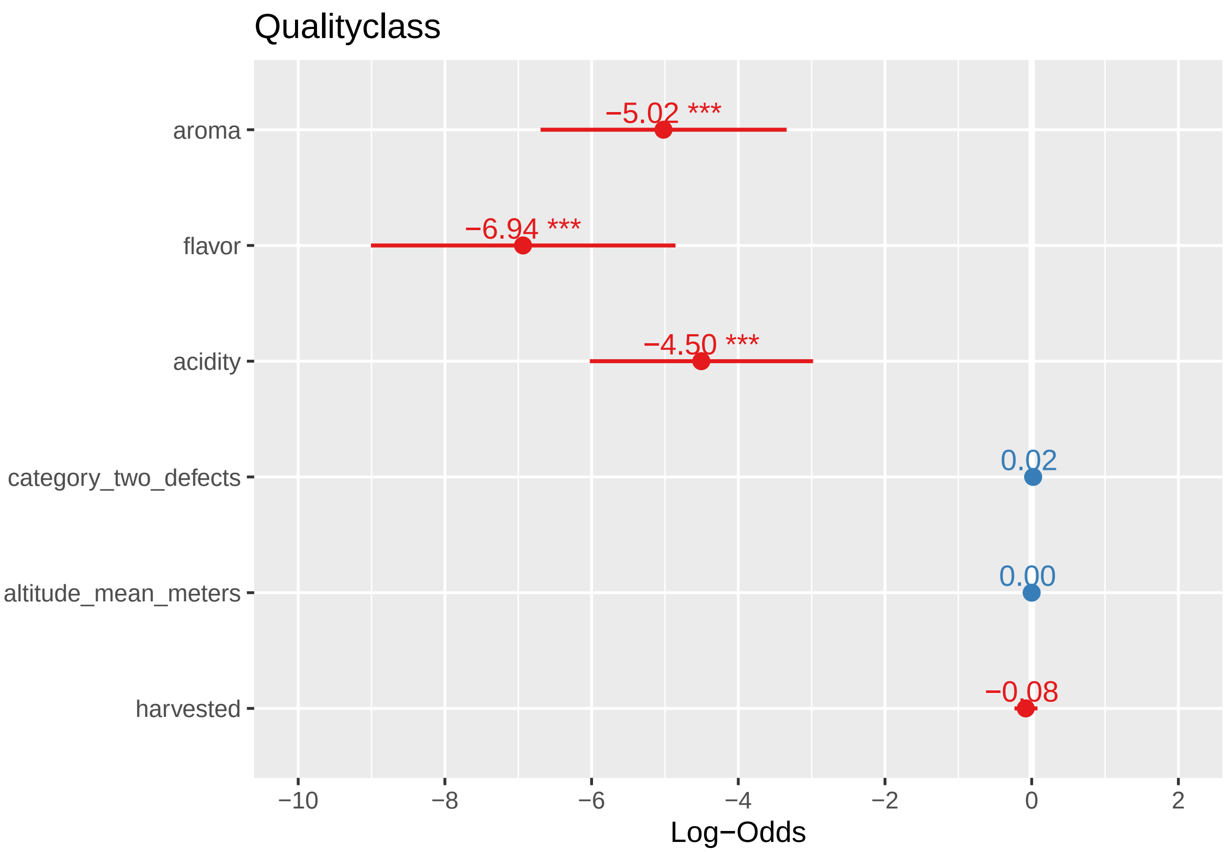


Table 2: Variance inflation factor (VIF)

	x
aroma	1.072
flavor	1.045
acidity	1.018
category_two_defects	1.020
altitude_mean_meters	1.003
harvested	1.039

Also from VIF table, we can tell that there is no multicolinearity problem between acidity and flavor.

Since the 95% confidence interval of altitude is not clear, we will calculate it in another way:

```
## [1] -3.347303e-05  5.175463e-05
```

Variables of category two defects, altitude and harvested include zero. Drop them.

And AIC of model 1 is: 342.478081

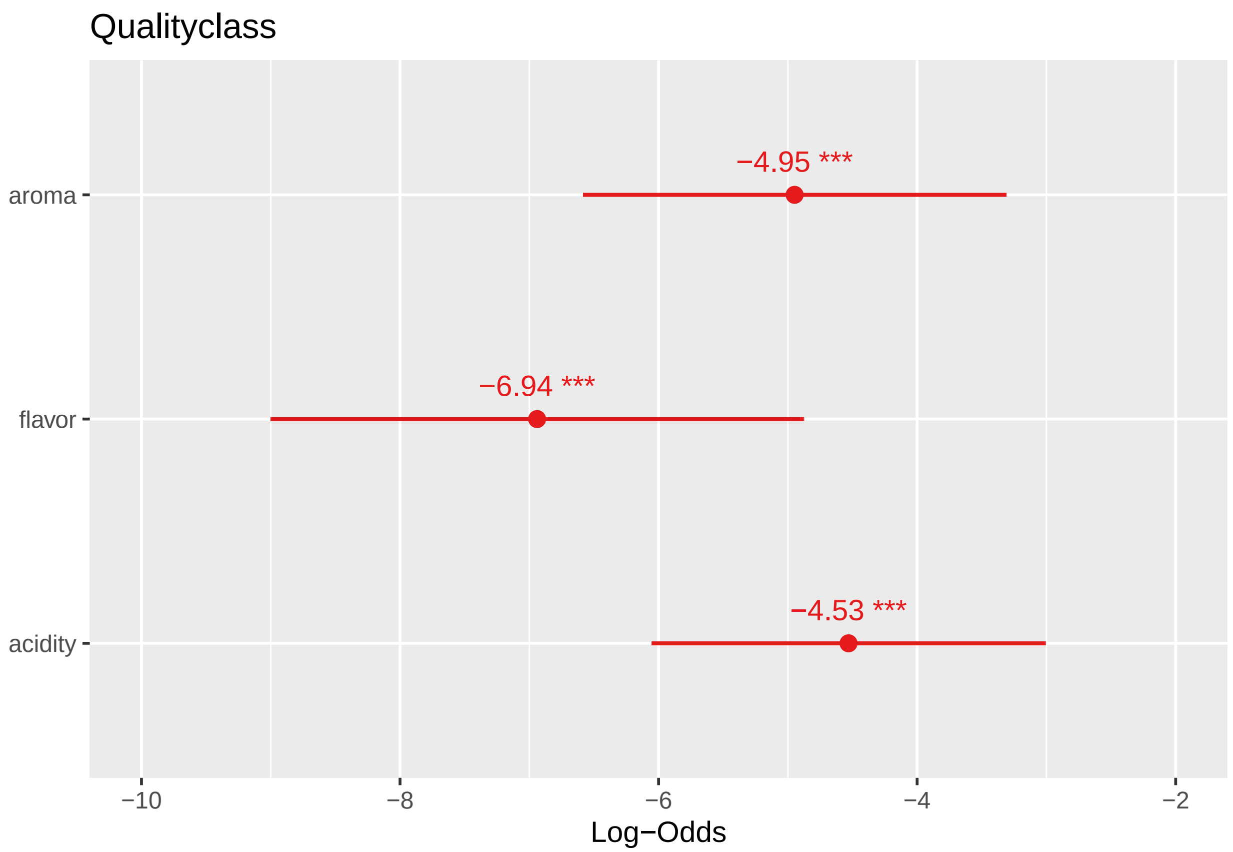
3.2 Model 2

```
##
```

```
## Call:
## glm(formula = Qualityclass ~ ., family = binomial, data = training.data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4250  -0.4087   0.0012   0.3467   3.7082
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 124.1516    10.9945  11.292  < 2e-16 ***
## aroma        -4.9477     0.8349  -5.926 3.11e-09 ***
## flavor       -6.9398     1.0522  -6.596 4.23e-11 ***
## acidity      -4.5306     0.7771  -5.830 5.53e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 792.85  on 571  degrees of freedom
## Residual deviance: 330.34  on 568  degrees of freedom
## AIC: 338.34
##
## Number of Fisher Scoring iterations: 8
```

```
## Waiting for profiling to be done...
```

	2.5 %	97.5 %
(Intercept)	103.987	147.199
aroma	-6.656	-3.374
flavor	-9.107	-4.972
acidity	-6.098	-3.042

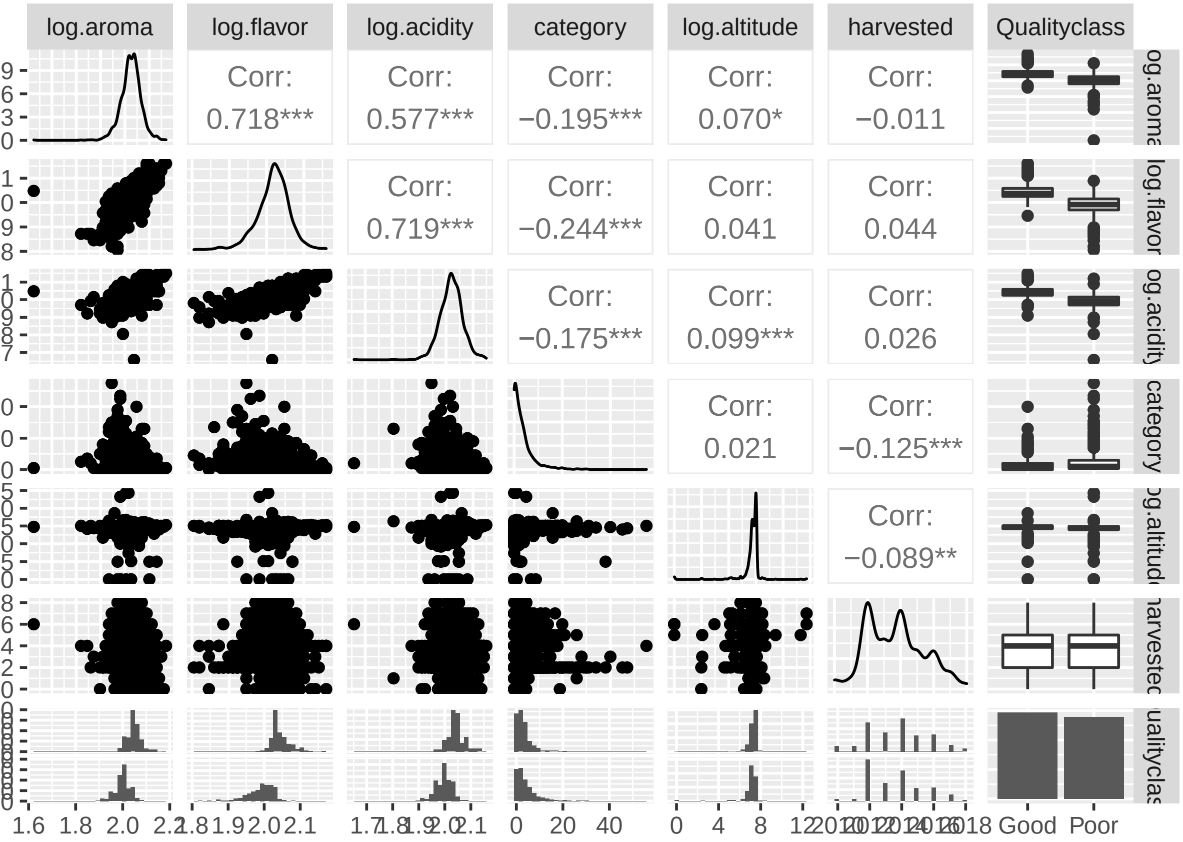


All the 95% confidence interval of explanatory variables does not include zero and the AIC of model 2 is 338.339023, which is smaller than AIC of model 1. We can consider this model as our final model.

3.3 Model 3 (with log transformation)

Since the distribution of almost every variable is slightly skewed, we can take log transformation to solve this problem.

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

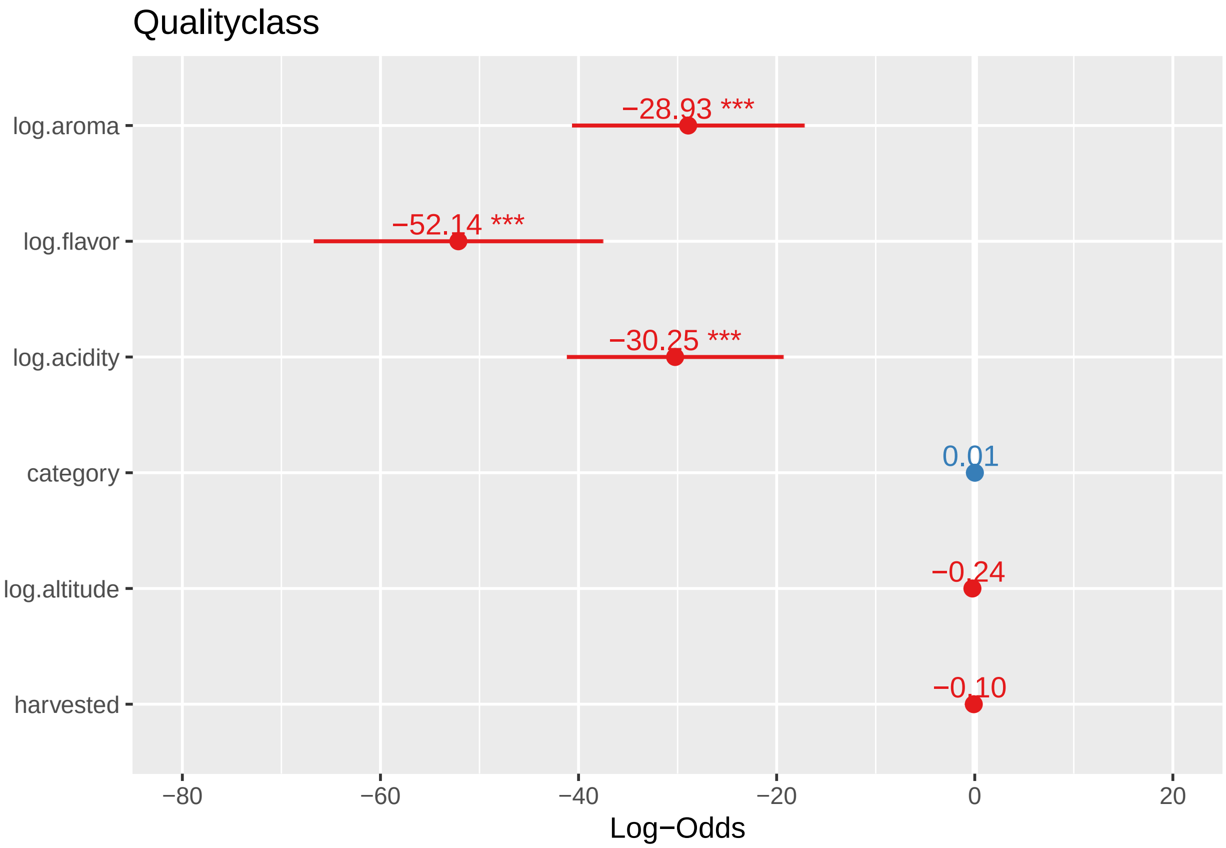


Fit model 3:

```
##
## Call:
## glm(formula = Qualityclass ~ ., family = binomial, data = training.data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1890  -0.4944  -0.0273   0.3385   3.9820
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  423.62092  152.72512   2.774  0.00554 **
## log.aroma    -28.93169    5.98015  -4.838 1.31e-06 ***
## log.flavor   -52.13526    7.44339  -7.004 2.48e-12 ***
## log.acidity  -30.25314    5.57225  -5.429 5.66e-08 ***
## category       0.01213    0.03062   0.396  0.69204
## log.altitude  -0.23539    0.13843  -1.700  0.08906 .
## harvested     -0.09791    0.07438  -1.316  0.18803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 790.43  on 571  degrees of freedom
## Residual deviance: 353.75  on 565  degrees of freedom
## AIC: 367.75
##
## Number of Fisher Scoring iterations: 7
```

```
## Waiting for profiling to be done...
```

	2.5 %	97.5 %
(Intercept)	127.194	727.570
log.aroma	-41.021	-17.510
log.flavor	-67.539	-38.280
log.acidity	-41.419	-19.490
category	-0.047	0.071
log.altitude	-0.528	0.026
harvested	-0.245	0.048



AIC of model 3 is 367.7506892. Since 95% confidence interval of category, log altitude and harvested include zero, drop these three variables.

Fit model 4.

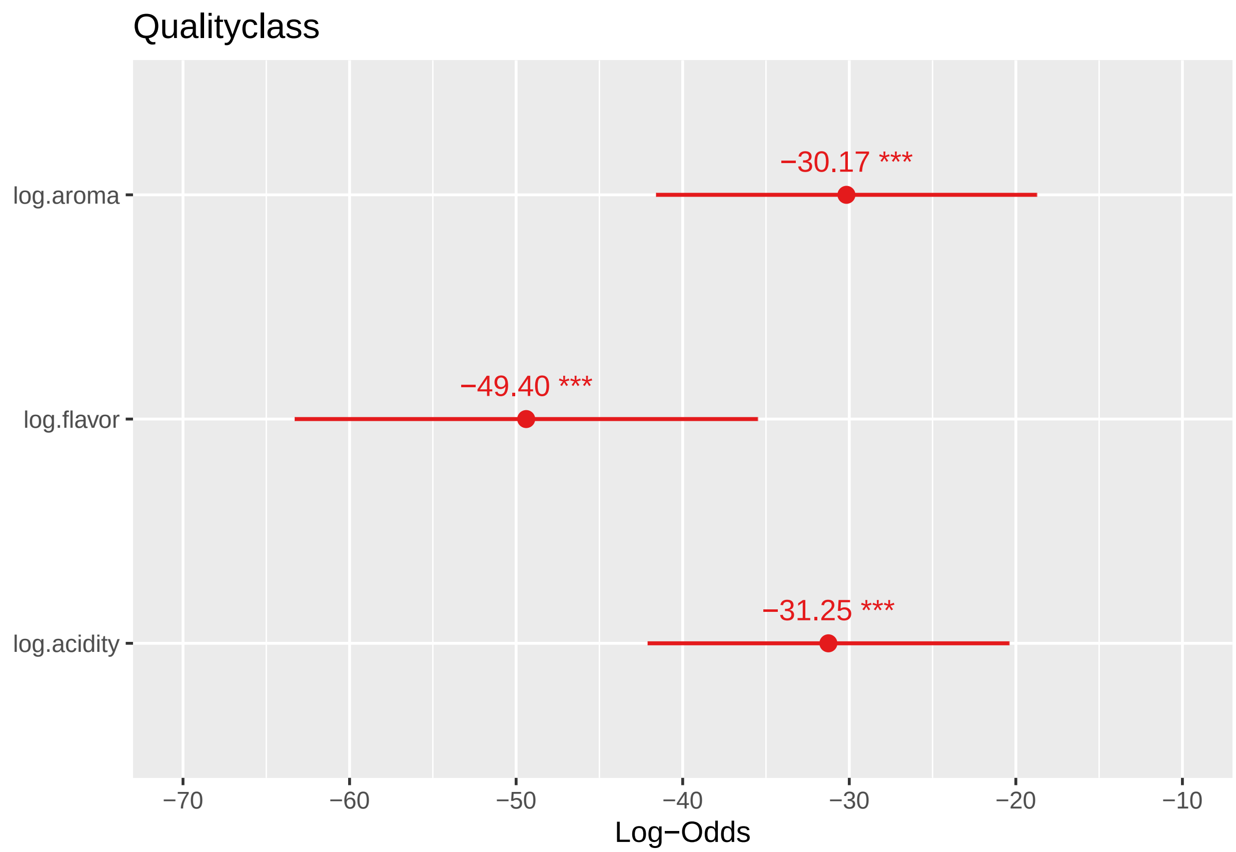
3.4 Model 4 (drop category, log altitude and harvested)

```
##
## Call:
```

```
## glm(formula = Qualityclass ~ ., family = binomial, data = training.data4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1721  -0.5018  -0.0274   0.3472   3.9389
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  223.818     20.219  11.070  < 2e-16 ***
## log.aroma    -30.170      5.829  -5.176 2.27e-07 ***
## log.flavor   -49.396      7.088  -6.969 3.19e-12 ***
## log.acidity  -31.254      5.536  -5.646 1.65e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 790.43  on 571  degrees of freedom
## Residual deviance: 358.19  on 568  degrees of freedom
## AIC: 366.19
##
## Number of Fisher Scoring iterations: 7
```

```
## Waiting for profiling to be done...
```

	2.5 %	97.5 %
(Intercept)	186.743	266.173
log.aroma	-41.962	-19.049
log.flavor	-64.034	-36.174
log.acidity	-42.344	-20.562



AIC of all the four models are listed below:

Table 3: Model comparison values for different models

Model	null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
Mod1	792.85	571	-164.24	342.48	372.92	328.48	565	572
Mod2	792.85	571	-165.17	338.34	355.74	330.34	568	572
Mod3	790.43	571	-176.88	367.75	398.19	353.75	565	572
Mod4	790.43	571	-179.09	366.19	383.59	358.19	568	572

4 Assessing model fit

Plot the AUC value of models:

From the table of AIC and BIC we can conclude that model 2 is the model with the best performance since it has the lowest AIC and BIC. And from this model we can get the conclusion that acidity, aroma and flavor have strong influence to the quality of a coffee from a certain country.

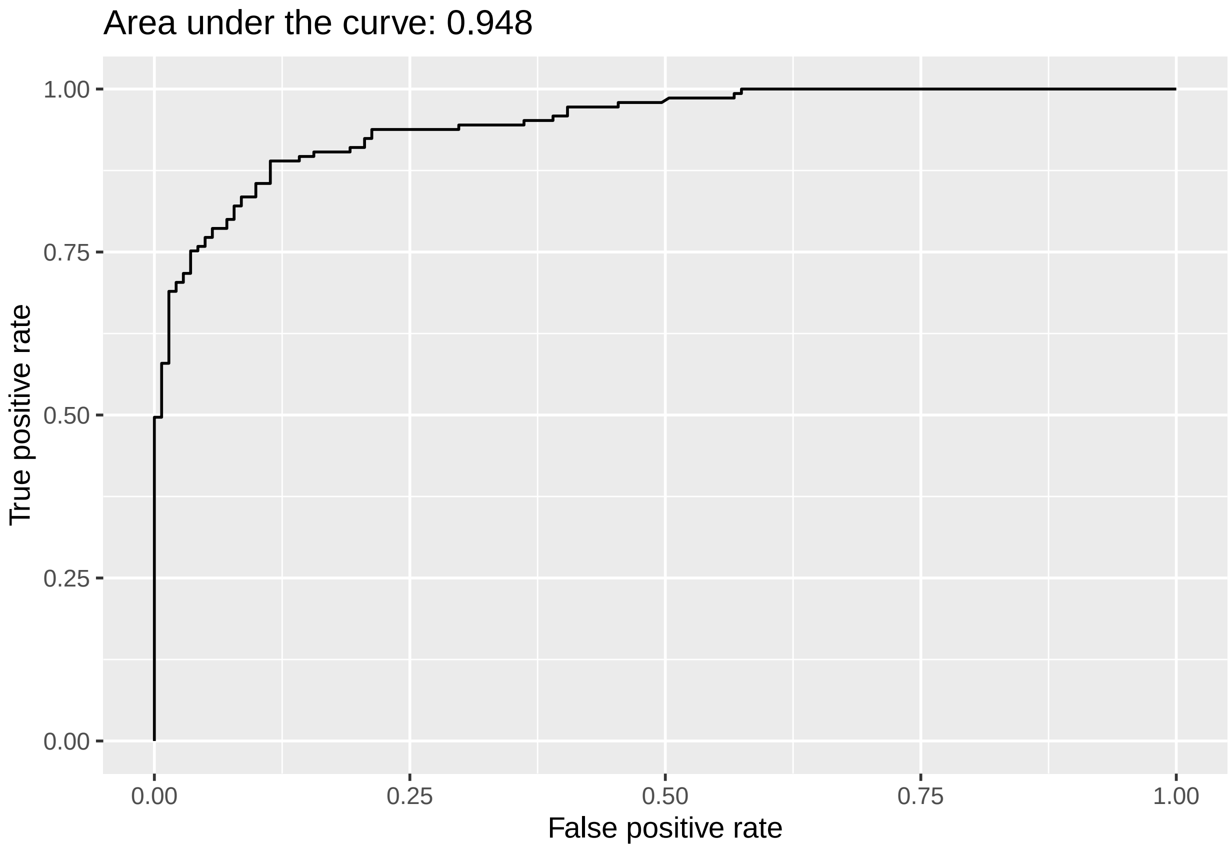


Figure 1: AUC of model 1

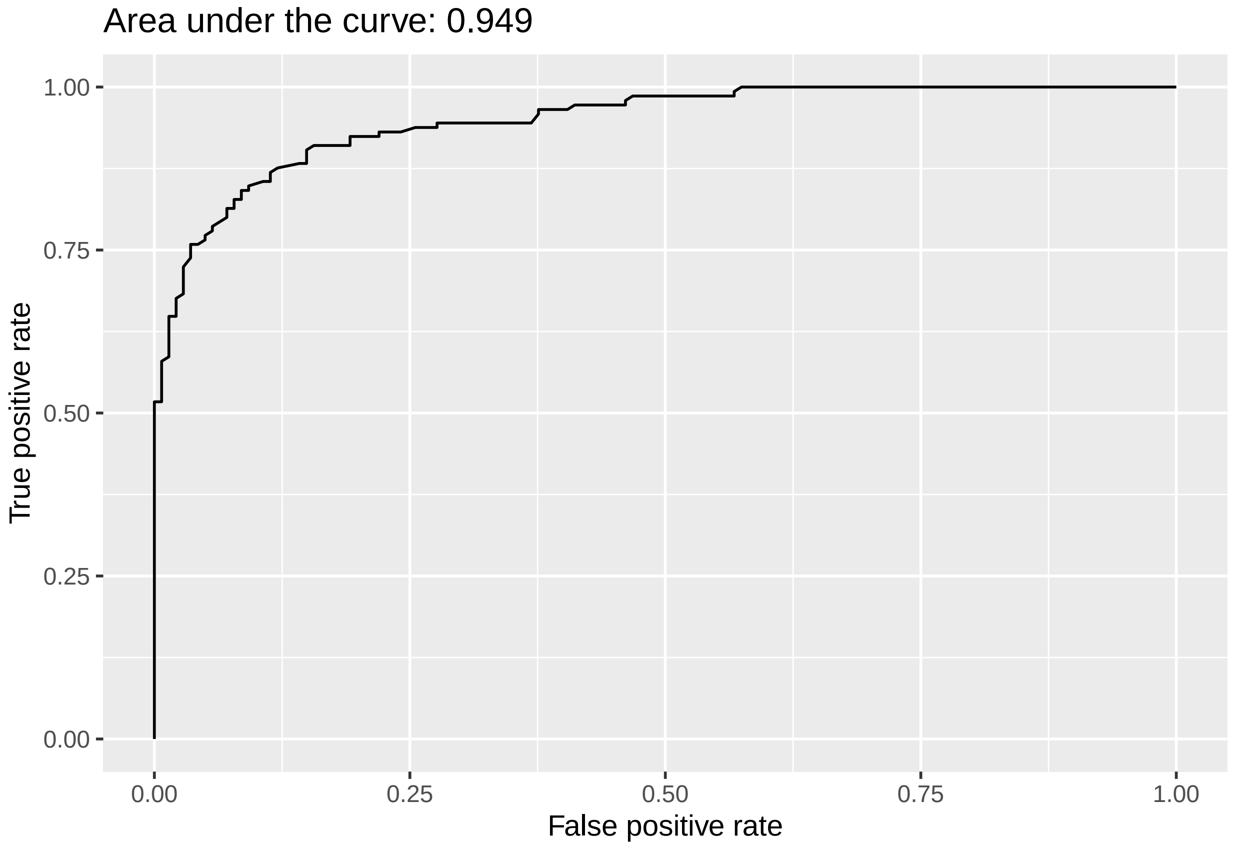


Figure 2: AUC of model 2

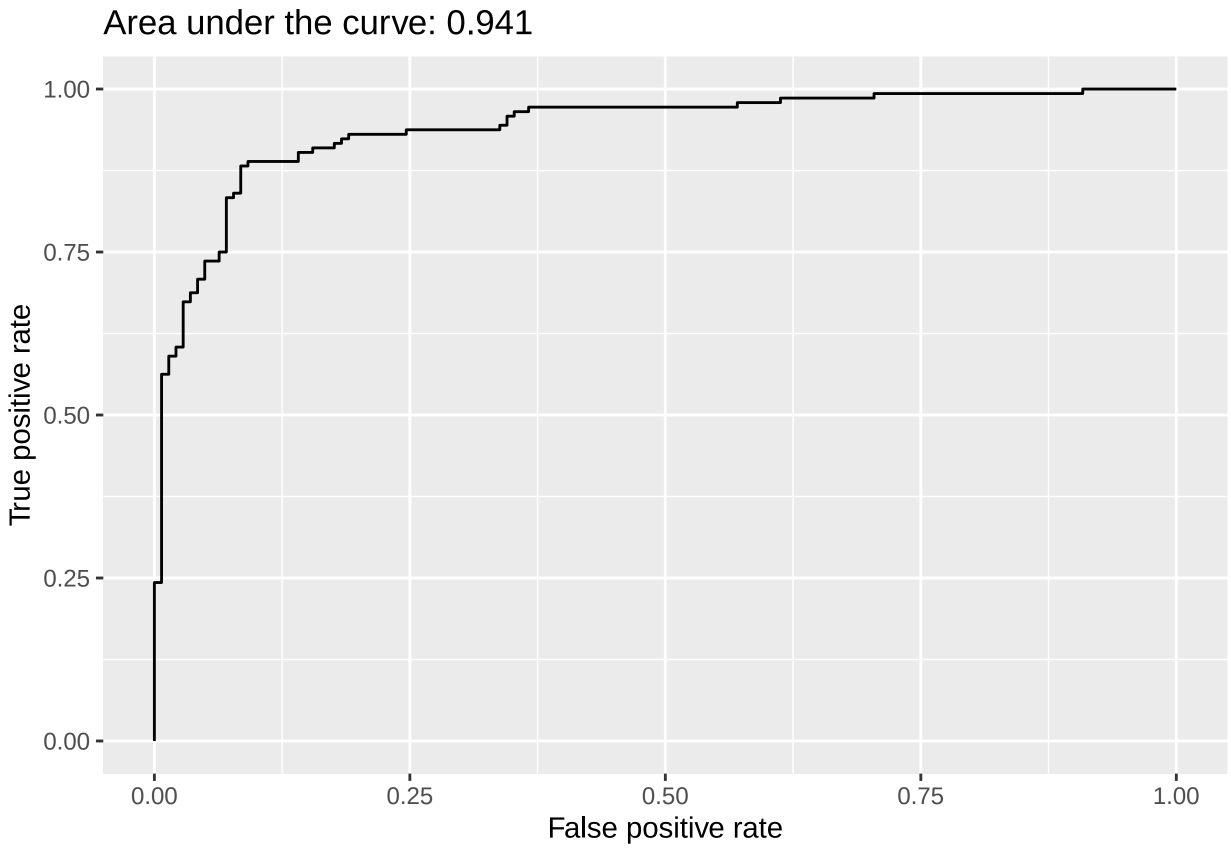


Figure 3: AUC of model 3

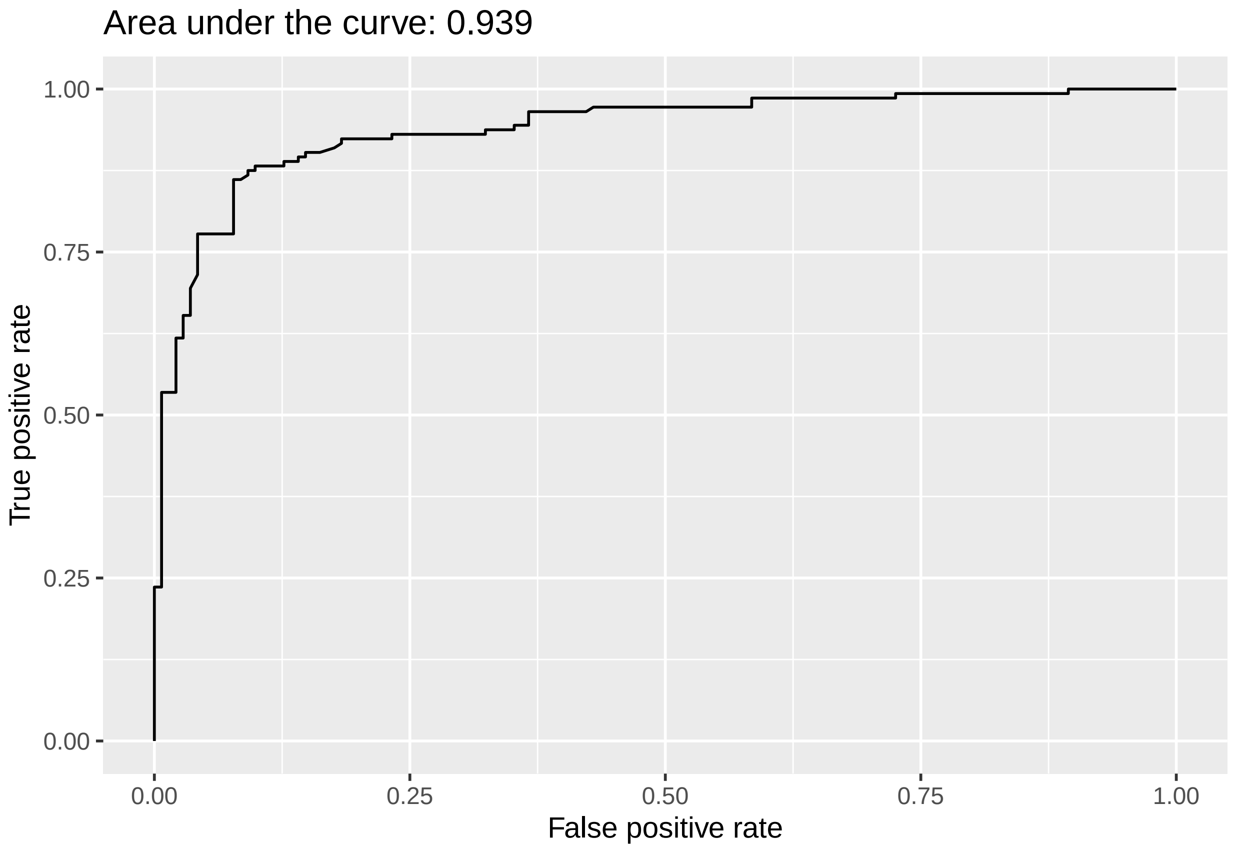


Figure 4: AUC of model 4