coffee

Shuqi Cao

1 introduction

1.1 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?

1.2 Data description

The dataset is collected from the Coffee Quality Database (CQD) of Coffee Quality Institute. As a non-profit organisation, the institute aims to improve the quality of coffee and the lives of farmers who produce the beans. The dataset contains information on features of coffee and its production, including an overall quality score.

Response variable

- `country_of_origin` – Country where the coffee bean originates from.

Explanatory variables

- `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 (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.

2 Data summarisation

We can load in our data set and see what it looks like by using the summary function.

```
##  country_of_origin      aroma           flavor         acidity
##  Length:1145        Min.   :0.000   Min.   :0.000   Min.   :0.000
##  Class :character   1st Qu.:7.420   1st Qu.:7.330   1st Qu.:7.330
##  Mode  :character   Median :7.580   Median :7.580   Median :7.500
##                     Mean   :7.571   Mean   :7.521   Mean   :7.536
##                     3rd Qu.:7.750   3rd Qu.:7.750   3rd Qu.:7.750
##                     Max.   :8.750   Max.   :8.670   Max.   :8.580
```

```
##
##  category_two_defects altitude_mean_meters   harvested    Qualityclass
##  Min.   : 0.000       Min.   :     1       Min.   :2010   Length:1145
##  1st Qu.: 0.000       1st Qu.:  1100       1st Qu.:2012   Class :character
##  Median : 2.000       Median :  1311       Median :2014   Mode  :character
##  Mean   : 3.673       Mean   :  1851       Mean   :2014
##  3rd Qu.: 5.000       3rd Qu.:  1600       3rd Qu.:2015
##  Max.   :55.000       Max.   :190164       Max.   :2018
##                       NA's   :201          NA's   :60
```

There are some missing values in numerical variables, 201 in `altitude_mean_meters` and 60 in `havested`. Curiously, the maximum of `altitude_mean_meters` is up to 190,164 meters; it is out of the question! It's also worth noting that some coffee beans get zero in the jugement of their features (`aroma`, `flavour`, `acidity`). We will plot histogram to show distributions of these features.

```
## `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`.
```

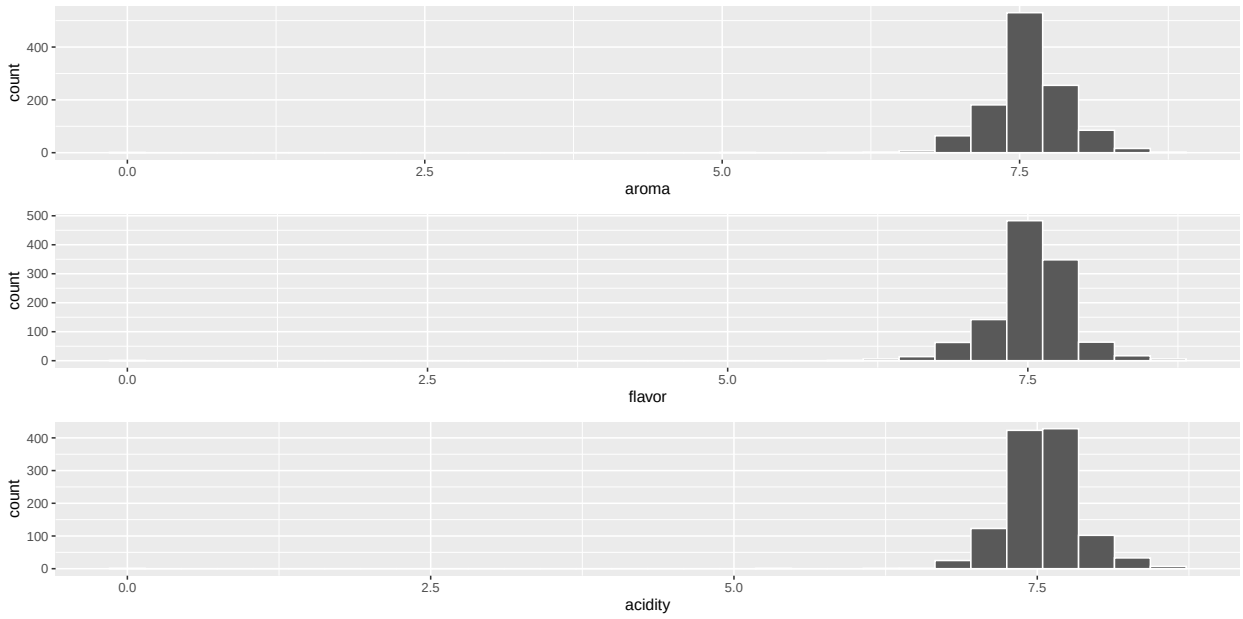
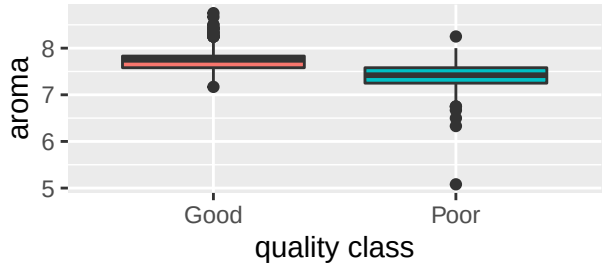
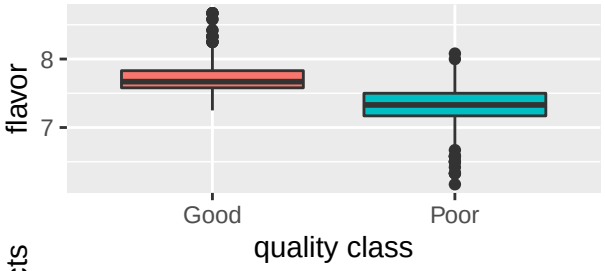


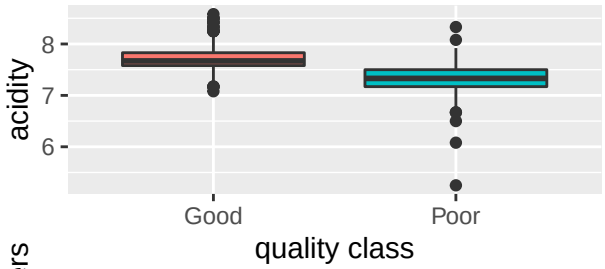
Figure 1: Boxplot and histogram of variables

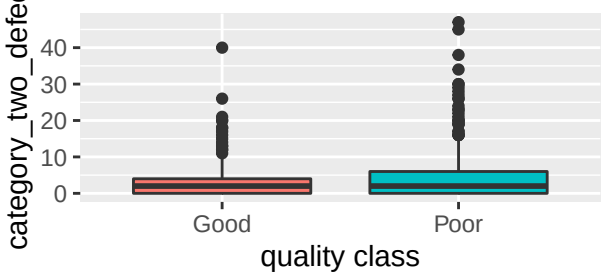
These boxplots show that most of coffee beans get grades between 6 and 8, so we can delete the observation with zero grade. Meanwhile, as mentioned earlier, we will remove outliers from our analysis.

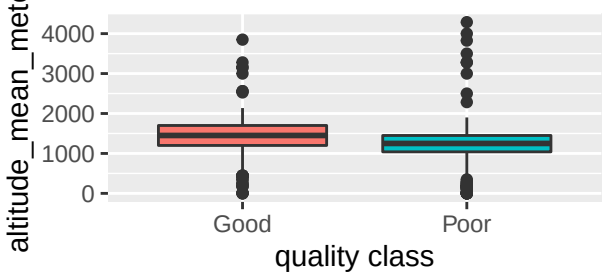
After cleaning the data, we will plot boxplots of `Qualityclass` by other features of coffee.

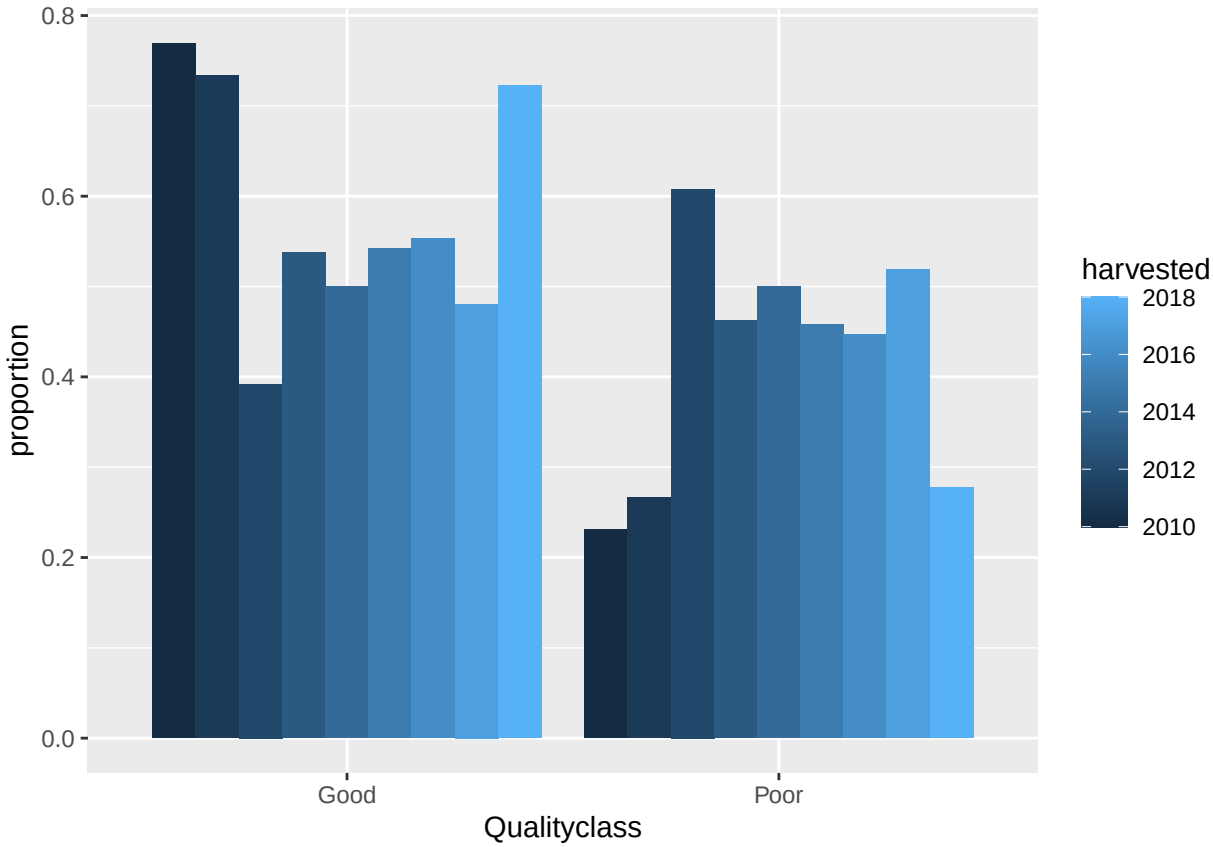

aroma
8
7
6
5
Good
Poor
quality class


flavor
8
7
Good
Poor
quality class


acidity
8
7
6
Good
Poor
quality class


category_two_defects
40
30
20
10
0
Good
Poor
quality class


altitude_mean_meters
4000
3000
2000
1000
0
Good
Poor
quality class



3 Methods

3.1 Log-odds

model 1

Firstly, We fit the logistic regression model with `Qualityclass` as the response and others as the explanatory variable. Let's explore the significance of the coefficients.

```
##
## Call:
## glm(formula = Qualityclass ~ ., family = binomial(link = "logit"),
##     data = dat1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2530  -0.3195   0.0010   0.4101   3.6122
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.232e+02  9.055e+00 -13.606  < 2e-16 ***
## aroma                4.762e+00  7.186e-01   6.627 3.42e-11 ***
## flavor               7.413e+00  8.759e-01   8.463  < 2e-16 ***
## acidity              4.074e+00  6.896e-01   5.907 3.48e-09 ***
```

```
## category_two_defects  1.398e-02  2.913e-02   0.480   0.6312
## altitude_mean_meters  6.940e-04  2.314e-04   2.999   0.0027 **
## harvested2011        -1.248e-01  1.071e+00  -0.117   0.9073
## harvested2012        -7.214e-01  9.066e-01  -0.796   0.4262
## harvested2013        -2.878e-01  9.220e-01  -0.312   0.7549
## harvested2014        -3.937e-01  9.139e-01  -0.431   0.6666
## harvested2015        -8.723e-01  9.197e-01  -0.948   0.3429
## harvested2016         9.376e-02  9.488e-01   0.099   0.9213
## harvested2017        -5.777e-01  9.701e-01  -0.596   0.5515
## harvested2018         1.344e+00  1.146e+00   1.173   0.2408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.15  on 929  degrees of freedom
## Residual deviance:  535.71  on 916  degrees of freedom
## AIC: 563.71
##
## Number of Fisher Scoring iterations: 7
```

model 2

Remove the variable `harvested`.

```
##
## Call:
## glm(formula = Qualityclass ~ ., family = binomial(link = "logit"),
##     data = dat2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1256  -0.3491   0.0011   0.4112   3.4973
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.214e+02  8.676e+00 -13.997  < 2e-16 ***
## aroma                 4.533e+00  6.794e-01   6.673 2.51e-11 ***
## flavor                7.192e+00  8.483e-01   8.478  < 2e-16 ***
## acidity               4.267e+00  6.739e-01   6.332 2.42e-10 ***
## category_two_defects  1.089e-03  2.701e-02   0.040   0.9678
## altitude_mean_meters  5.476e-04  2.168e-04   2.526   0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.15  on 929  degrees of freedom
## Residual deviance:  550.26  on 924  degrees of freedom
## AIC: 562.26
##
## Number of Fisher Scoring iterations: 7
```

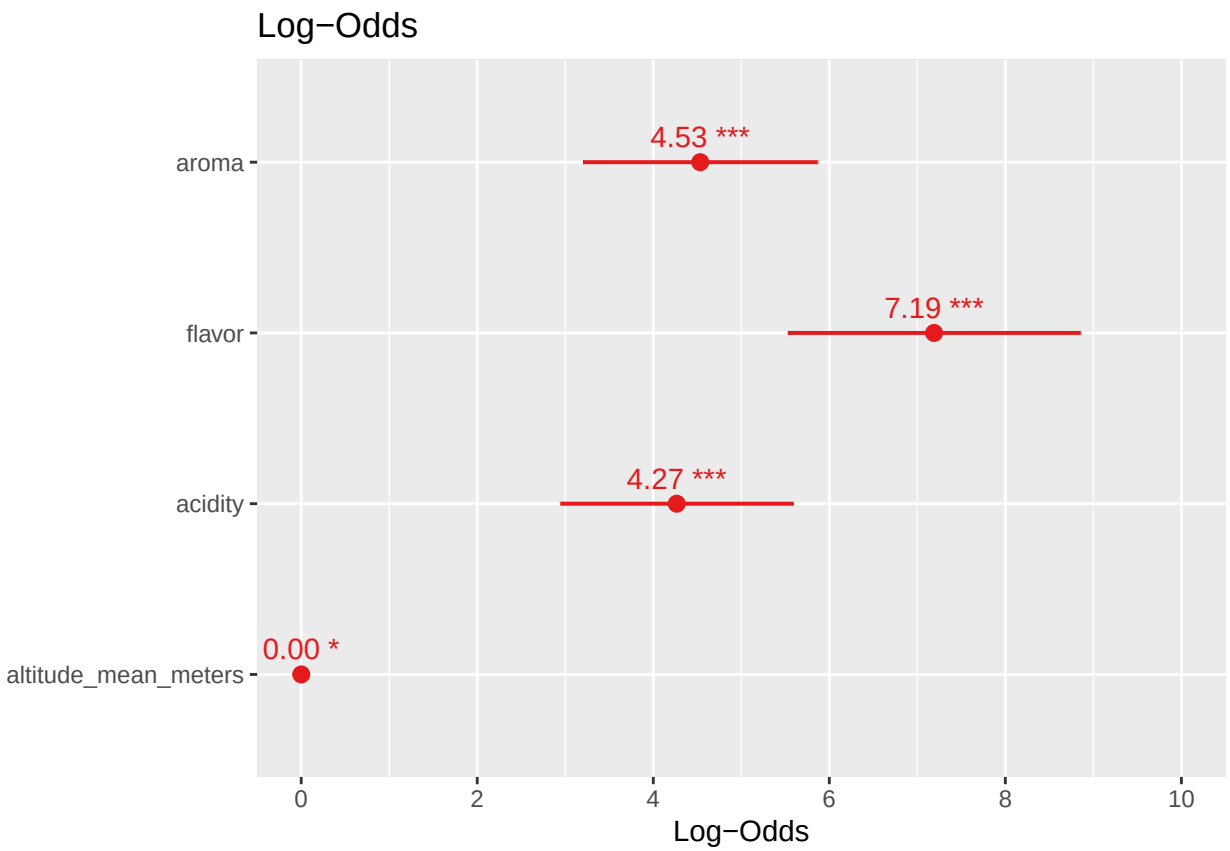
model 3

Remove the variable `category_two_defects`.

```
##
## Call:
## glm(formula = Qualityclass ~ ., family = binomial(link = "logit"),
##     data = dat3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1262  -0.3496   0.0011   0.4114   3.4962
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.214e+02  8.661e+00 -14.019  < 2e-16 ***
## aroma                 4.533e+00  6.795e-01   6.672 2.52e-11 ***
## flavor                7.190e+00  8.475e-01   8.485  < 2e-16 ***
## acidity               4.266e+00  6.738e-01   6.332 2.42e-10 ***
## altitude_mean_meters  5.480e-04  2.166e-04   2.530   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.15  on 929  degrees of freedom
## Residual deviance:  550.26  on 925  degrees of freedom
## AIC: 560.26
##
## Number of Fisher Scoring iterations: 7
```

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

	2.5 %	97.5 %
(Intercept)	-139.3313334	-105.3280386
aroma	3.2384211	5.9056861
flavor	5.5871773	8.9142372
acidity	2.9710042	5.6169236
altitude_mean_meters	0.0001228	0.0009765



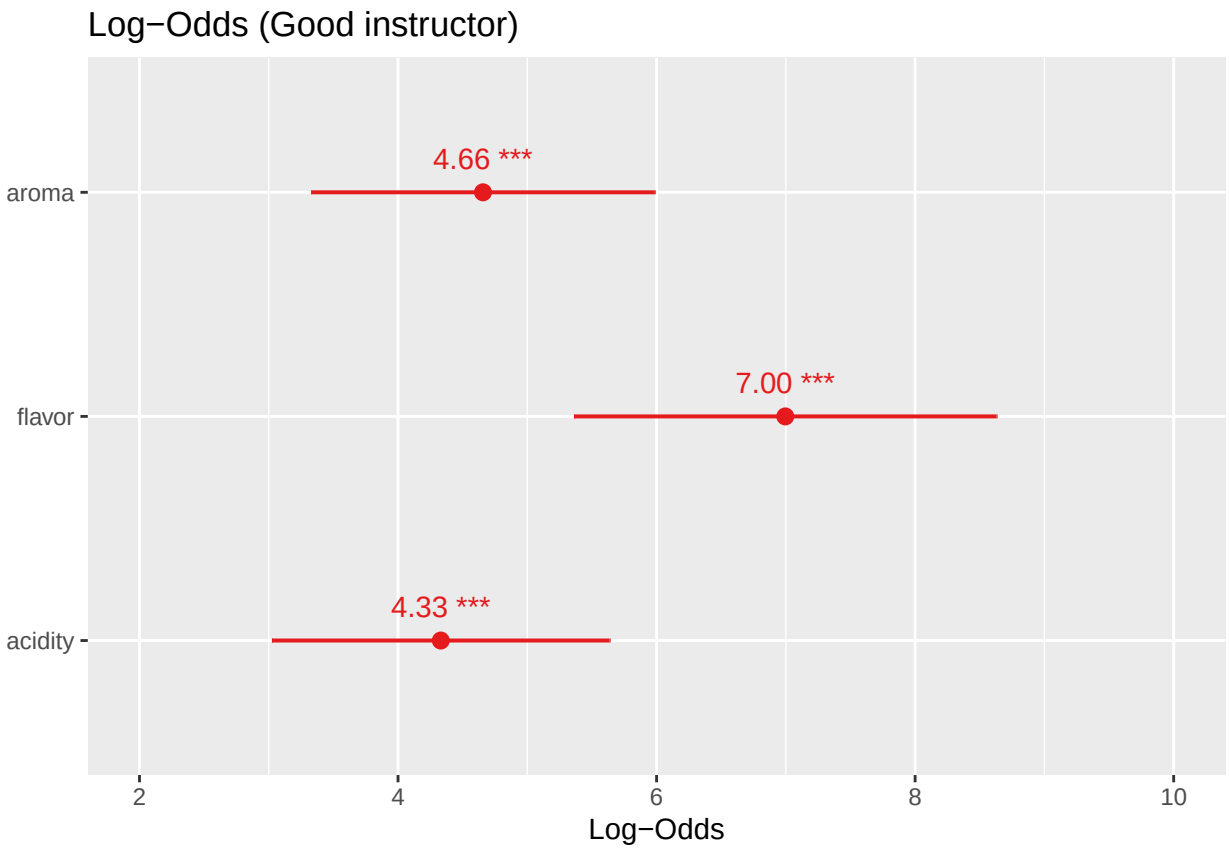
model 4

Remove the variable `altitude_mean_meters`.

```
##
## Call:
## glm(formula = Qualityclass ~ ., family = binomial(link = "logit"),
##     data = dat4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1514  -0.3597   0.0014   0.4323   3.3213
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -120.6560     8.5427 -14.124  < 2e-16 ***
## aroma          4.6572     0.6780   6.869 6.49e-12 ***
## flavor         6.9955     0.8339   8.389  < 2e-16 ***
## acidity        4.3308     0.6664   6.499 8.11e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.15  on 929  degrees of freedom
## Residual deviance:  556.62  on 926  degrees of freedom
## AIC: 564.62
```

```
##
## Number of Fisher Scoring iterations: 7

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

	2.5 %	97.5 %
(Intercept)	-138.309306	-104.773331
aroma	3.363159	6.024777
flavor	5.417795	8.691704
acidity	3.050303	5.667201



3.2 Model selection

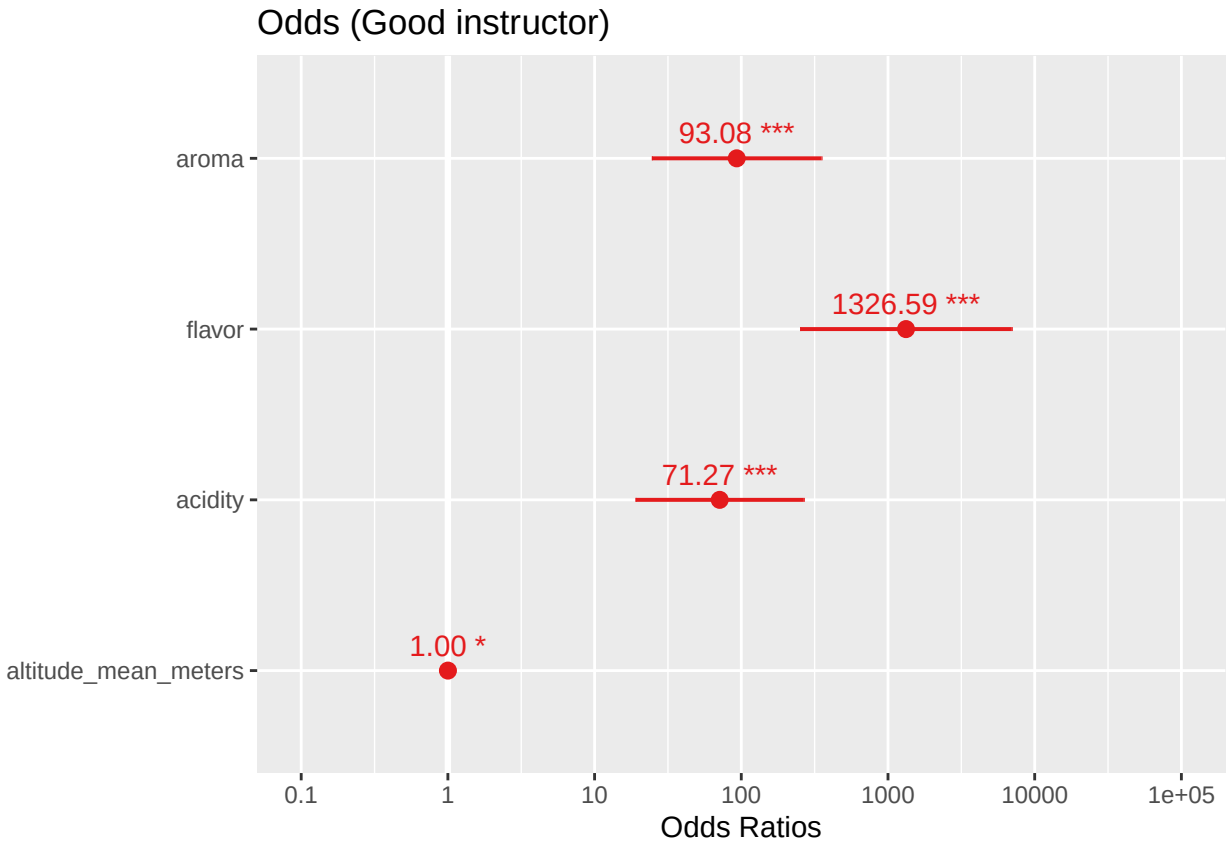
Table 1: Model comparison values for different models

model	AIC	BIC
GLM1	563.708	631.400
GLM2	562.261	591.272
GLM3	560.263	584.439
GLM4	564.616	583.957

3.3 Odds

Table 2: Odds-ratio

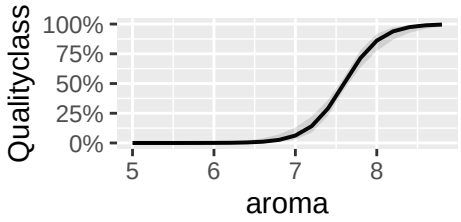
	Odds
(Intercept)	0.000
aroma	93.075
flavor	1326.589
acidity	71.268
altitude_mean_meters	1.001



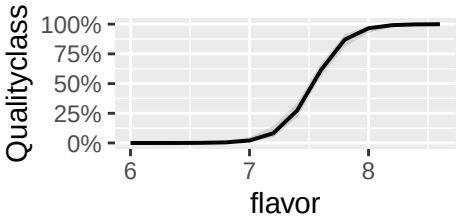
3.4 Probabilities

```
## Data were 'prettified'. Consider using 'terms="aroma [all]"' to get smooth plots.

## Data were 'prettified'. Consider using 'terms="flavor [all]"' to get smooth plots.

## Data were 'prettified'. Consider using 'terms="acidity [all]"' to get smooth plots.

## Data were 'prettified'. Consider using 'terms="altitude_mean_meters [all]"' to get smooth plots.
```

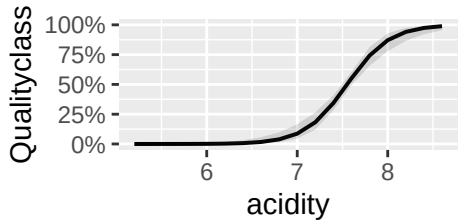
A


Qualityclass
100%
75%
50%
25%
0%
5
6
7
8
aroma

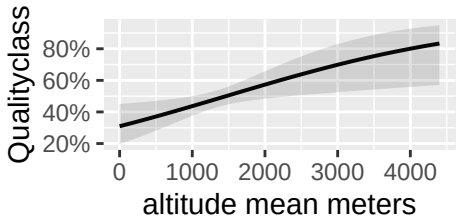
B


Qualityclass
100%
75%
50%
25%
0%
6
7
8
flavor

C


Qualityclass
100%
75%
50%
25%
0%
6
7
8
acidity

D


Qualityclass
80%
60%
40%
20%
0
1000
2000
3000
4000
altitude mean meters