Group_9_Analysis

Brent Strong, Enyu Li, Haotian Wang, Honjin Ren, Mu He

3/7/2022

Exploratory Data Analysis

Table 1: Summary statistics of continuous variables in the data set.

Variable	Mean	SD	Min.	1st Q.	Median	3rd Q.	Max.
aroma	7.57	0.39	0	7.42	7.58	7.75	8.75
flavor	7.52	0.40	0	7.33	7.58	7.75	8.67
acidity	7.54	0.39	0	7.33	7.50	7.75	8.58
category__two__defects	3.67	5.41	0	0.00	2.00	5.00	55.00
altitude__mean__meters	1850.69	9392.09	1	1100.00	1310.64	1600.00	190164.00
harvested	2013.67	1.81	2010	2012.00	2014.00	2015.00	2018.00

The following boxplot is for good quality rates for each country, in which we can check if any countries have unusual high or low good quality rate.

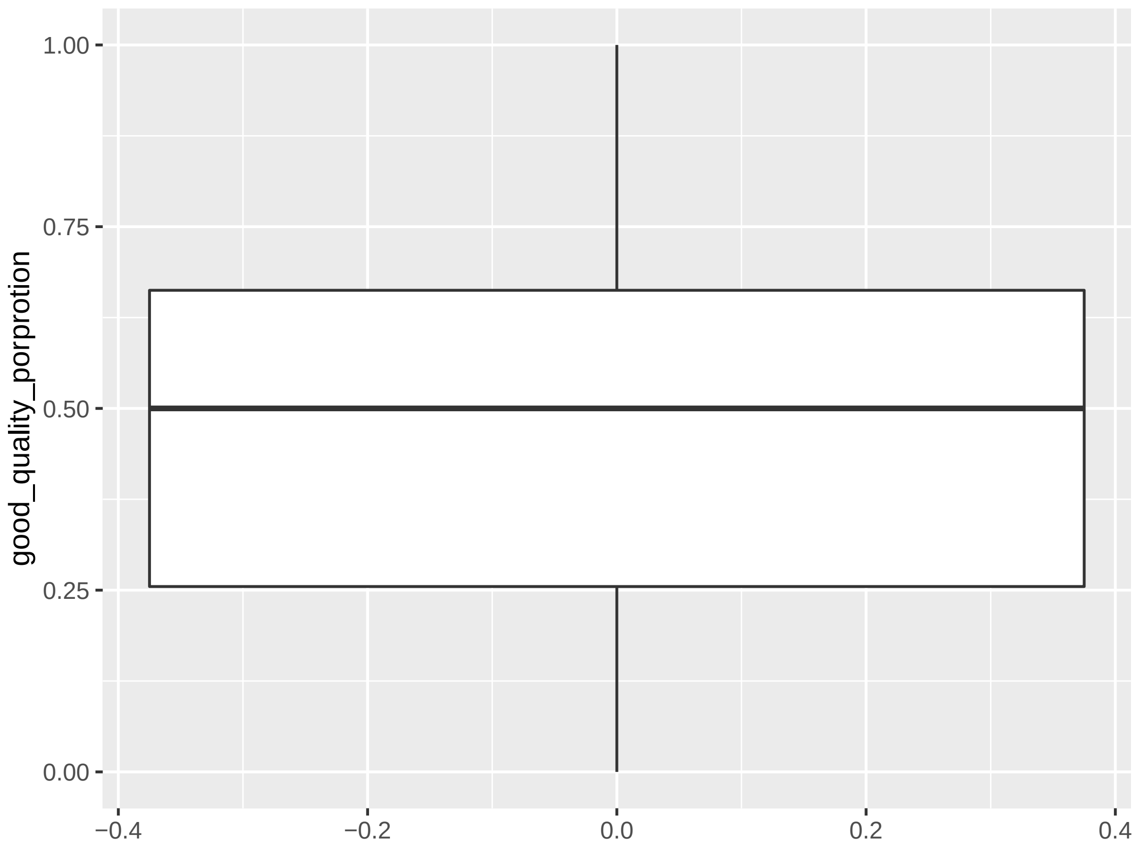


Figure 1: Boxplots of good quality rate for each country.

The following table filter countries and its number of batch with 20% good quality rate before and after, which provides more detailed information than the above boxplot. The number of batch can imply the reliability. For instance, Colombia has a relatively high good quality rate with large number of batch.

Table 2: Origins with twenty percent good quality rate before and after

country_of_origin	good_quality_porprotion	number_of_batch
Cote d?Ivoire	0.00	1
Laos	0.00	2
Mauritius	0.00	1
Myanmar	0.00	6
Zambia	0.00	1
Malawi	0.09	11
Haiti	0.20	5
El Salvador	0.70	20
Thailand	0.70	23
Panama	0.75	4
Uganda	0.78	32
Colombia	0.80	158
Ethiopia	0.92	38
Kenya	0.92	24
Japan	1.00	1

The following table is the distribution of features between coffee in good and poor quality. We can check if there is any obvious difference in some features.

Table 3: Summary statistics of the sepal length by species of irises

Variable	Qualityclass	n	Mean	SD	Min	Median	Max	IQR
aroma	Good	588	7.76	0.23	7.08	7.75	8.75	0.08
aroma	Poor	557	7.37	0.41	0.00	7.42	8.25	0.16
flavor	Good	588	7.74	0.23	7.00	7.67	8.67	0.16
flavor	Poor	557	7.29	0.42	0.00	7.33	8.08	0.17
acidity	Good	588	7.72	0.25	6.75	7.67	8.58	0.16
acidity	Poor	557	7.34	0.40	0.00	7.33	8.33	0.17
category__two__defects	Good	588	2.87	3.82	0.00	2.00	40.00	2.00
category__two__defects	Poor	557	4.52	6.60	0.00	2.00	55.00	4.00
altitude__mean__meters	Good	588	1431.04	629.05	1.00	1450.00	11000.00	255.16
altitude__mean__meters	Poor	557	2281.15	13346.02	1.00	1250.00	190164.00	200.00
harvested	Good	588	2013.74	1.90	2010.00	2014.00	2018.00	1.00
harvested	Poor	557	2013.59	1.71	2010.00	2013.00	2018.00	2.00

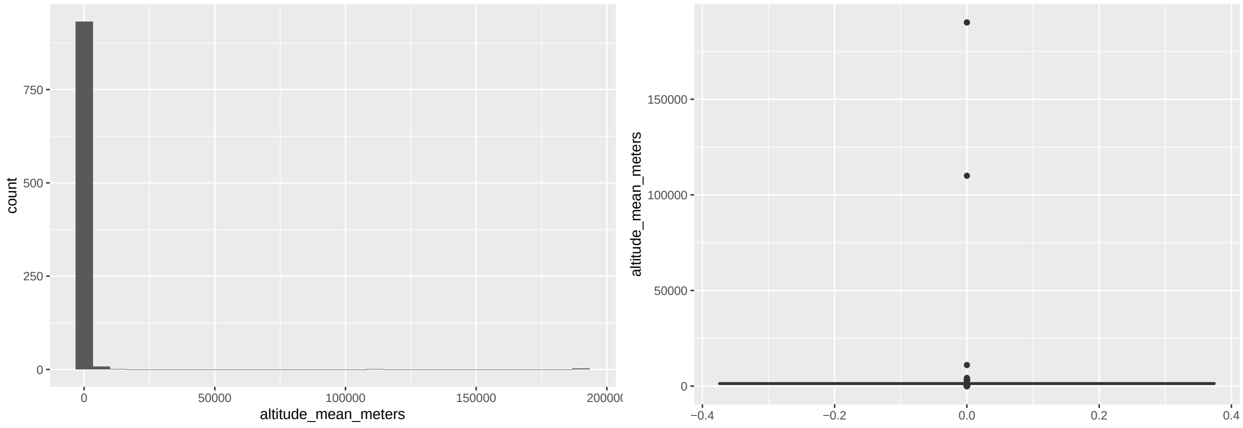


Figure 2: Histogram and boxplot for altitude.

There are several observations with extremly high altitude which are impossible. Hence, delete observations which have altitude higher than Mt. Everest.

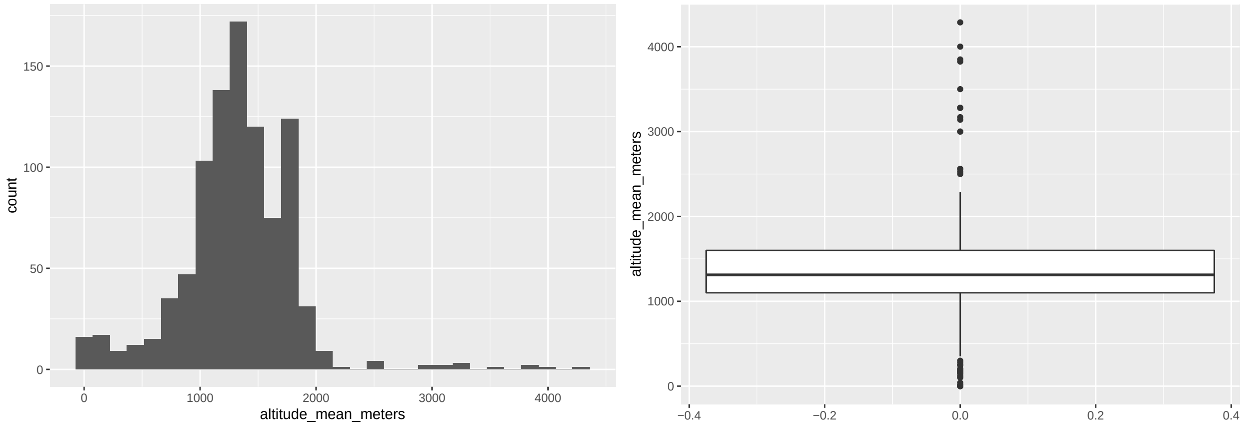


Figure 3: Histogram and boxplot for altitude after removing implausuble observations.

The following two histograms comparing distributions of altitude before and after removing implausible observations.

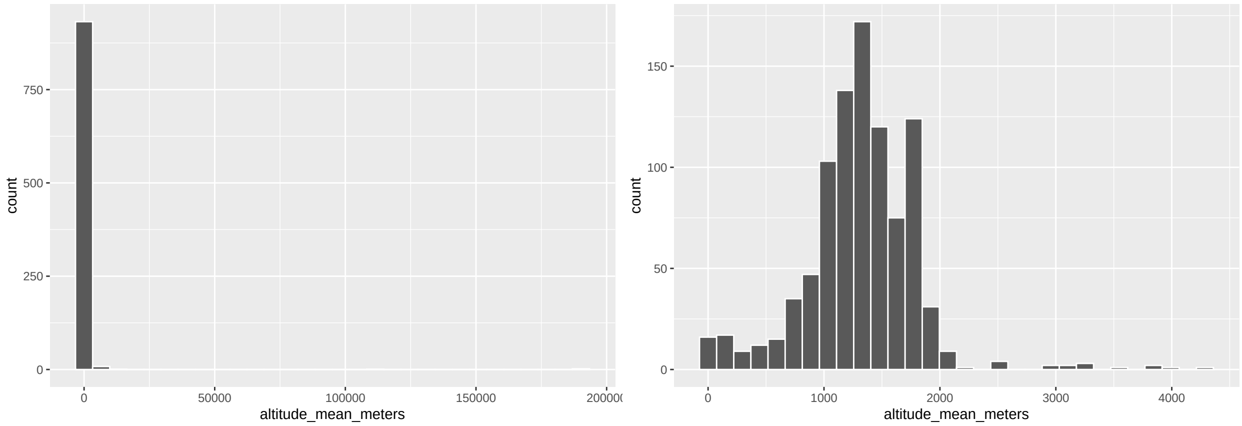


Figure 4: Histogram for altitude befor and after removing implausuble observations.

The following table is the distribution of featers between good and poor coffee. We can check if there is obvious difference in some features between good and poor coffee.

Table 4: Summary statistics of features of good and poor coffee

Variable	Qualityclass	n	Mean	SD	Min	Median	Max	IQR
aroma	Good	477	7.76	0.23	7.17	7.75	8.75	0.08
aroma	Poor	463	7.38	0.43	0.00	7.42	8.25	0.16
flavor	Good	477	7.74	0.22	7.25	7.67	8.67	0.16
flavor	Poor	463	7.30	0.43	0.00	7.33	8.08	0.17
acidity	Good	477	7.72	0.24	7.08	7.67	8.58	0.16
acidity	Poor	463	7.33	0.43	0.00	7.33	8.33	0.17
category_two_defects	Good	477	2.83	3.84	0.00	2.00	40.00	2.00
category_two_defects	Poor	463	4.43	6.43	0.00	2.00	47.00	4.00
altitude_mean_meters	Good	477	1410.98	451.40	1.00	1450.00	3850.00	250.00
altitude_mean_meters	Poor	463	1236.91	500.90	1.00	1250.00	4287.00	200.00
harvested	Good	477	2013.76	1.90	2010.00	2014.00	2018.00	1.00
harvested	Poor	463	2013.63	1.72	2010.00	2013.00	2018.00	2.00

Here is 6 box-plots comparing features distribution between good and poor coffee.

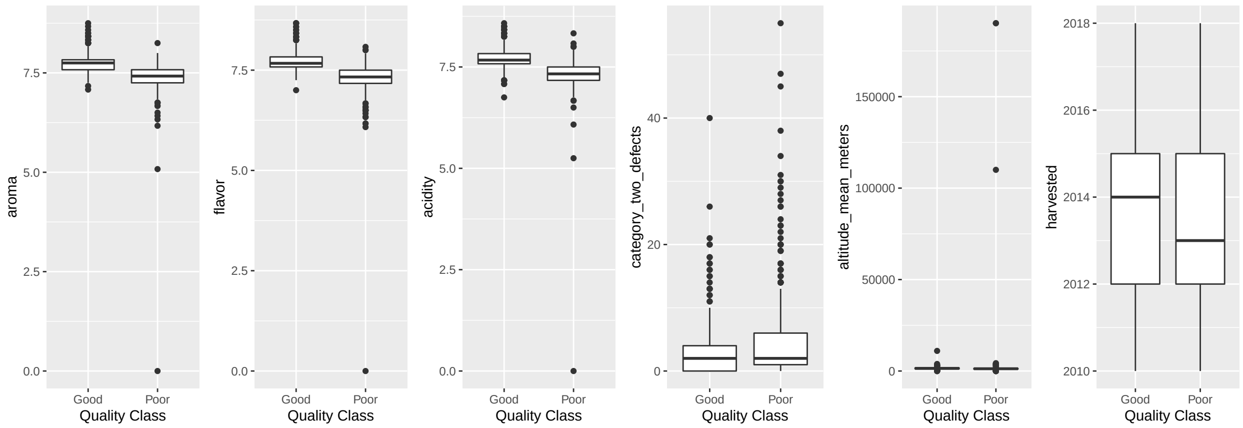


Figure 5: Boxplots2 of countinous features on different quality class.

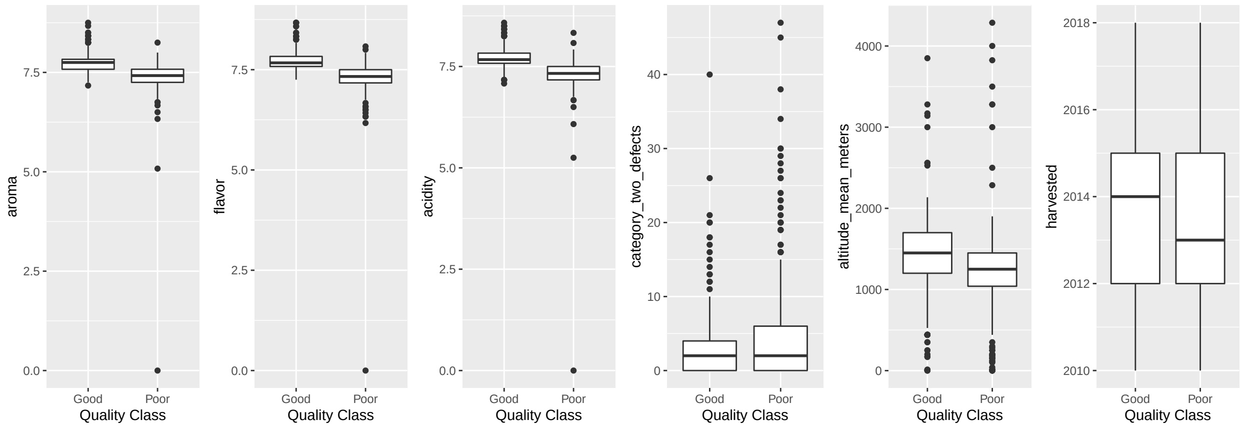


Figure 6: Boxplots of countinous features on different quality class after data cleaning.

Formal Analysis Using Logistic Regression

```
Call:
glm(formula = Qualityclass ~ level - 1, family = binomial(link = "logit"),
    data = coffee_final)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.286  -1.286   1.073   1.073   1.369

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
level1 -0.43891    0.18321  -2.396  0.01659 *
level2 -0.40968    0.14513  -2.823  0.00476 **
level3  0.25131    0.08175   3.074  0.00211 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.6  on 931  degrees of freedom
Residual deviance: 1267.1  on 928  degrees of freedom
AIC: 1273.1

Number of Fisher Scoring iterations: 4
```

If the level of altitude is the only explanatory variable in the model, the effect of three levels are all statistically significant. In detail, high altitude has a positive influence on the quality of coffee.

```
Call:
glm(formula = Qualityclass ~ year - 1, family = binomial(link = "logit"),
    data = coffee_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7125  -1.1774   0.7244   1.1146   1.3683

Coefficients:
           Estimate Std. Error z value Pr(>|z|)
year2010  1.204e+00  4.655e-01   2.587 0.009694 **
year2011  1.012e+00  4.129e-01   2.450 0.014277 *
year2012 -4.383e-01  1.283e-01  -3.417 0.000634 ***
year2013  1.495e-01  1.733e-01   0.863 0.388102
year2014  2.742e-15  1.436e-01   0.000 1.000000
year2015  1.699e-01  1.848e-01   0.919 0.357851
year2016  2.144e-01  1.982e-01   1.082 0.279346
year2017 -1.133e-01  2.752e-01  -0.412 0.680441
year2018  9.555e-01  5.262e-01   1.816 0.069408 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.6  on 931  degrees of freedom
Residual deviance: 1257.3  on 922  degrees of freedom
AIC: 1275.3

Number of Fisher Scoring iterations: 4
```

If harvested year is the only explanatory variable in the model, the effects of year 2010, 2011 and 2012 are statistically significant. Coffee harvested in year 2012 has a higher odds ratio. Coffee harvested in year 2010 and 2011 has a lower odds ratio.

```
Call:
glm(formula = Qualityclass ~ country_of_origin - 1, family = binomial(link = "logit"),
    data = coffee_final)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.14597  -1.01655   0.00036   1.08424   2.18993
```

```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
country_of_originBrazil         6.596e-02  2.098e-01   0.314  0.75320
country_of_originBurundi        0.000e+00  1.414e+00   0.000  1.00000
country_of_originChina          5.878e-01  5.578e-01   1.054  0.29197
country_of_originColombia       1.563e+00  2.345e-01   6.666 2.64e-11 ***
country_of_originCosta Rica     2.231e-01  3.354e-01   0.665  0.50587
country_of_originCote d?Ivoire -1.657e+01  2.400e+03  -0.007  0.99449
country_of_originEcuador       -1.570e-16  1.414e+00   0.000  1.00000
country_of_originEl Salvador    9.555e-01  5.262e-01   1.816  0.06941 .
country_of_originEthiopia       1.657e+01  5.003e+02   0.033  0.97359
country_of_originGuatemala      7.878e-02  1.776e-01   0.444  0.65736
country_of_originHaiti         -1.386e+00  1.118e+00  -1.240  0.21500
country_of_originHawaii         1.657e+01  2.400e+03   0.007  0.99449
country_of_originHonduras      -1.070e+00  3.345e-01  -3.200  0.00137 **
country_of_originIndia          0.000e+00  6.325e-01   0.000  1.00000
country_of_originIndonesia      2.877e-01  5.401e-01   0.533  0.59425
country_of_originKenya          2.197e+00  7.454e-01   2.948  0.00320 **
country_of_originLaos          -1.657e+01  1.697e+03  -0.010  0.99221
country_of_originMalawi        -2.303e+00  1.049e+00  -2.195  0.02813 *
country_of_originMauritius     -1.657e+01  2.400e+03  -0.007  0.99449
country_of_originMexico        -1.046e+00  1.612e-01  -6.488 8.68e-11 ***
country_of_originMyanmar       -1.657e+01  9.796e+02  -0.017  0.98651
country_of_originNicaragua     -1.204e+00  6.583e-01  -1.829  0.06740 .
country_of_originPanama         1.099e+00  1.155e+00   0.951  0.34139
country_of_originPeru          -1.657e+01  2.400e+03  -0.007  0.99449
country_of_originPhilippines   -4.055e-01  9.129e-01  -0.444  0.65692
country_of_originPuerto Rico   -6.931e-01  1.225e+00  -0.566  0.57143
country_of_originTaiwan        -3.909e-01  2.700e-01  -1.448  0.14769
country_of_originTanzania      -6.899e-02  3.716e-01  -0.186  0.85271
country_of_originThailand       2.877e-01  5.401e-01   0.533  0.59425
country_of_originUganda         1.190e+00  4.317e-01   2.756  0.00585 **
country_of_originUnited States  6.931e-01  7.071e-01   0.980  0.32696
country_of_originVietnam        2.877e-01  7.638e-01   0.377  0.70642
country_of_originZambia        -1.657e+01  2.400e+03  -0.007  0.99449
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.6  on 931  degrees of freedom
Residual deviance: 1072.7  on 898  degrees of freedom
AIC: 1138.7

Number of Fisher Scoring iterations: 15
```

If the country of origin is the only explanatory variable, Colombia, Mexico, Honduras, Kenya, Malawi, Uganda have statistically significant effect on the odds ratio.

```
Call:
glm(formula = Qualityclass ~ Colombia + Mexico + Honduras + Kenya +
    Malawi + Uganda - 1, family = binomial(link = "logit"), data = coffee_final)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.146  -1.177   0.459   1.177   2.190

Coefficients:
         Estimate Std. Error z value Pr(>|z|)
Colombia   1.5629     0.2345   6.666 2.64e-11 ***
Mexico    -1.0460     0.1612  -6.488 8.68e-11 ***
Honduras  -1.0704     0.3345  -3.200  0.00137 **
Kenya      2.1972     0.7453   2.948  0.00320 **
Malawi    -2.3026     1.0486  -2.196  0.02810 *
Uganda     1.1896     0.4317   2.756  0.00585 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.6  on 931  degrees of freedom
Residual deviance: 1139.6  on 925  degrees of freedom
AIC: 1151.6

Number of Fisher Scoring iterations: 4
```

The following is the model considering all possible expalntatory variables.

```
Call:
glm(formula = Qualityclass ~ aroma + flavor + acidity + country_of_origin +
    category_two_defects + level + year, family = binomial(link = "logit"),
    data = coffee_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6259  -0.2422   0.0000   0.2902   3.5656

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.453e+02  1.172e+01 -12.392  < 2e-16 ***
aroma                           5.177e+00  8.458e-01   6.121 9.29e-10 ***
flavor                          8.627e+00  1.071e+00   8.053 8.08e-16 ***
acidity                         5.255e+00  8.273e-01   6.353 2.12e-10 ***
country_of_originBurundi        1.851e+00  4.782e+00   0.387  0.69869
country_of_originChina          4.662e-01  1.077e+00   0.433  0.66498
country_of_originColombia       1.828e+00  5.741e-01   3.185  0.00145 **
country_of_originCosta Rica     2.729e-01  7.598e-01   0.359  0.71948
country_of_originCote d?Ivoire -1.203e+01  6.523e+03  -0.002  0.99853
country_of_originEcuador       -1.204e+00  1.523e+00  -0.791  0.42911
country_of_originEl Salvador    2.819e-01  9.666e-01   0.292  0.77058
country_of_originEthiopia       1.338e+01  9.449e+02   0.014  0.98870
country_of_originGuatemala     -7.471e-01  5.761e-01  -1.297  0.19467
country_of_originHaiti          2.231e+00  2.048e+00   1.089  0.27609
country_of_originHawaii         4.587e+00  6.523e+03   0.001  0.99944
country_of_originHonduras      -6.526e-01  7.090e-01  -0.920  0.35735
country_of_originIndia         -2.746e+00  1.068e+00  -2.570  0.01016 *
country_of_originIndonesia     -3.673e-01  1.010e+00  -0.364  0.71609
country_of_originKenya          5.348e-01  1.574e+00   0.340  0.73396
```

```
country_of_originLaos          -1.544e+01  4.511e+03  -0.003  0.99727
country_of_originMalawi        -8.054e-01  1.302e+00  -0.619  0.53606
country_of_originMauritius     -1.194e+01  6.523e+03  -0.002  0.99854
country_of_originMexico        -7.950e-01  5.221e-01  -1.523  0.12785
country_of_originMyanmar       -1.555e+01  2.378e+03  -0.007  0.99478
country_of_originNicaragua      5.363e-01  2.028e+00   0.264  0.79144
country_of_originPanama         3.390e+00  1.799e+00   1.884  0.05951 .
country_of_originPeru          -1.438e+01  6.523e+03  -0.002  0.99824
country_of_originPhilippines    2.925e+00  2.668e+00   1.096  0.27286
country_of_originPuerto Rico   -2.774e+00  1.751e+00  -1.584  0.11310
country_of_originTaiwan         1.032e+00  6.963e-01   1.482  0.13835
country_of_originTanzania       9.536e-01  7.594e-01   1.256  0.20921
country_of_originThailand       2.751e+00  9.978e-01   2.757  0.00583 **
country_of_originUganda        -1.591e+00  7.933e-01  -2.006  0.04490 *
country_of_originUnited States  1.358e-01  1.565e+00   0.087  0.93085
country_of_originVietnam        2.118e+00  1.163e+00   1.822  0.06847 .
country_of_originZambia        -1.385e+01  6.523e+03  -0.002  0.99831
category_two_defects            5.568e-02  3.464e-02   1.607  0.10797
level2                          5.416e-01  4.846e-01   1.117  0.26379
level3                          1.053e+00  4.823e-01   2.184  0.02898 *
year2011                       -1.212e-01  1.120e+00  -0.108  0.91384
year2012                        1.298e-01  9.710e-01   0.134  0.89370
year2013                        6.002e-01  9.776e-01   0.614  0.53925
year2014                        3.728e-03  9.862e-01   0.004  0.99698
year2015                       -4.363e-02  9.760e-01  -0.045  0.96434
year2016                        8.617e-01  1.029e+00   0.838  0.40220
year2017                        5.787e-01  1.029e+00   0.563  0.57369
year2018                        2.541e+00  1.311e+00   1.938  0.05257 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.55  on 930  degrees of freedom
Residual deviance:  449.73  on 884  degrees of freedom
AIC: 543.73

Number of Fisher Scoring iterations: 17


Call:
glm(formula = Qualityclass ~ aroma + flavor + acidity + Colombia +
    Mexico + Honduras + Kenya + Malawi + Uganda + category_two_defects +
    level + year2010 + year2011 + year2012, family = binomial(link = "logit"),
    data = coffee_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2233  -0.3110   0.0010   0.3332   3.4913

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -124.52106    9.39969 -13.247  < 2e-16 ***
aroma                   4.42037    0.73018   6.054 1.41e-09 ***
flavor                  7.21914    0.87975   8.206 2.29e-16 ***
```

```
acidity                 4.80821    0.72879   6.598 4.18e-11 ***
Colombia                1.77971    0.39212   4.539 5.66e-06 ***
Mexico                 -0.82994    0.34382  -2.414   0.0158 *
Honduras               -0.58919    0.53161  -1.108   0.2677
Kenya                   0.99751    1.34387   0.742   0.4579
Malawi                 -1.26603    1.15543  -1.096   0.2732
Uganda                 -1.29548    0.62048  -2.088   0.0368 *
category_two_defects    0.05244    0.02894   1.812   0.0700 .
level2                  0.13831    0.39437   0.351   0.7258
level3                  0.39276    0.33774   1.163   0.2449
year2010               -0.40857    0.86051  -0.475   0.6349
year2011               -0.47132    0.65362  -0.721   0.4708
year2012               -0.22952    0.31593  -0.726   0.4675
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.6  on 930  degrees of freedom
Residual deviance:  499.0  on 915  degrees of freedom
AIC: 531

Number of Fisher Scoring iterations: 7


Start:  AIC=531
Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
    Honduras + Kenya + Malawi + Uganda + category_two_defects +
    level + year2010 + year2011 + year2012

                       Df Deviance    AIC
- level                 2   500.68 528.68
- year2010              1   499.22 529.22
- year2011              1   499.51 529.51
- year2012              1   499.53 529.53
- Kenya                 1   499.62 529.62
- Honduras              1   500.28 530.28
- Malawi                1   500.51 530.51
<none>                      499.00 531.00
- category_two_defects  1   502.16 532.16
- Uganda                1   503.06 533.06
- Mexico                1   504.93 534.93
- Colombia              1   522.71 552.71
- aroma                 1   550.65 580.65
- acidity               1   552.26 582.26
- flavor                1   589.47 619.47

Step:  AIC=528.68
Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
    Honduras + Kenya + Malawi + Uganda + category_two_defects +
    year2010 + year2011 + year2012

                       Df Deviance    AIC
- year2010              1   500.81 526.81
- year2011              1   501.21 527.21
- year2012             1   501.31 527.31
- Honduras             1   501.45 527.45
- Kenya                1   501.46 527.46
- Malawi               1   502.00 528.00
<none>                     500.68 528.68
- category_two_defects  1   504.11 530.11
- Uganda               1   504.17 530.17
+ level                2   499.00 531.00
- Mexico               1   506.09 532.09
- Colombia             1   528.97 554.97
- aroma                1   554.14 580.14
- acidity              1   556.27 582.27
- flavor               1   589.52 615.52

Step:  AIC=526.81
Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico + 
    Honduras + Kenya + Malawi + Uganda + category_two_defects + 
    year2011 + year2012

                       Df Deviance    AIC
- year2011              1   501.31 525.31
- year2012              1   501.40 525.40
- Honduras              1   501.56 525.56
- Kenya                 1   501.60 525.60
- Malawi                1   502.12 526.12
<none>                      500.81 526.81
- Uganda                1   504.25 528.25
- category_two_defects  1   504.30 528.30
+ year2010              1   500.68 528.68
+ level                 2   499.22 529.22
- Mexico                1   506.23 530.23
- Colombia              1   529.18 553.18
- aroma                 1   554.21 578.21
- acidity               1   556.49 580.49
- flavor                1   589.60 613.60

Step:  AIC=525.31
Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico + 
    Honduras + Kenya + Malawi + Uganda + category_two_defects + 
    year2012

                       Df Deviance    AIC
- year2012              1   501.78 523.78
- Honduras              1   502.01 524.01
- Kenya                 1   502.12 524.12
- Malawi                1   502.58 524.58
<none>                      501.31 525.31
- Uganda                1   504.68 526.68
- category_two_defects  1   504.80 526.80
+ year2011              1   500.81 526.81
+ year2010              1   501.21 527.21
+ level                 2   499.68 527.68
- Mexico                1   506.85 528.85
- Colombia              1   529.36 551.36
- aroma                 1   555.00 577.00
- acidity               1   556.77 578.77
- flavor                1   589.70 611.70

Step:  AIC=523.78
Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico + 
    Honduras + Kenya + Malawi + Uganda + category_two_defects

                       Df Deviance    AIC
- Honduras              1   502.38 522.38
- Kenya                 1   502.61 522.61
- Malawi                1   503.00 523.00
<none>                      501.78 523.78
- Uganda                1   505.00 525.00
- category_two_defects  1   505.00 525.00
+ year2012              1   501.31 525.31
+ year2011              1   501.40 525.40
+ year2010              1   501.71 525.71
+ level                 2   500.04 526.04
- Mexico                1   511.24 531.24
- Colombia              1   529.38 549.38
- aroma                 1   555.13 575.13
- acidity               1   557.73 577.73
- flavor                1   590.18 610.18

Step:  AIC=522.38
Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico + 
    Kenya + Malawi + Uganda + category_two_defects

                       Df Deviance    AIC
- Kenya                 1   503.26 521.26
- Malawi                1   503.52 521.52
<none>                      502.38 522.38
- Uganda                1   505.47 523.47
- category_two_defects  1   505.49 523.49
+ Honduras              1   501.78 523.78
+ year2012              1   502.01 524.01
+ year2011              1   502.03 524.03
+ year2010              1   502.32 524.32
+ level                 2   501.13 525.13
- Mexico                1   511.31 529.31
- Colombia              1   531.79 549.79
- aroma                 1   556.16 574.16
- acidity               1   559.22 577.22
- flavor                1   591.11 609.11

Step:  AIC=521.26
Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico + 
    Malawi + Uganda + category_two_defects

                       Df Deviance    AIC
- Malawi                1   504.42 520.42
<none>                      503.26 521.26
+ Kenya                 1   502.38 522.38
```

```
- Uganda                1   506.40 522.40
- category_two_defects  1   506.40 522.40
+ Honduras              1   502.61 522.61
+ year2012              1   502.89 522.89
+ year2011              1   502.89 522.89
+ year2010              1   503.20 523.20
+ level                 2   501.87 523.87
- Mexico                1   512.59 528.59
- Colombia              1   532.34 548.34
- aroma                 1   556.51 572.51
- acidity               1   561.04 577.04
- flavor                1   593.35 609.35

Step:  AIC=520.42
Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
    Uganda + category_two_defects

                       Df Deviance    AIC
<none>                      504.42 520.42
+ Malawi                1   503.26 521.26
- Uganda                1   507.44 521.44
+ Kenya                 1   503.52 521.52
+ Honduras              1   503.85 521.85
- category_two_defects  1   507.92 521.92
+ year2011              1   504.06 522.06
+ year2012              1   504.09 522.09
+ year2010              1   504.36 522.36
+ level                 2   503.14 523.14
- Mexico                1   513.43 527.43
- Colombia              1   534.72 548.72
- aroma                 1   557.63 571.63
- acidity               1   562.84 576.84
- flavor                1   596.95 610.95


Call:
glm(formula = Qualityclass ~ aroma + flavor + acidity + Colombia +
    Mexico + Uganda + category_two_defects, family = binomial(link = "logit"),
    data = coffee_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1969  -0.3208   0.0010   0.3370   3.4697

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -125.42332    9.27395 -13.524  < 2e-16 ***
aroma                   4.44405    0.71829   6.187 6.13e-10 ***
flavor                  7.16176    0.86051   8.323  < 2e-16 ***
acidity                 4.98081    0.72085   6.910 4.86e-12 ***
Colombia                1.83828    0.36232   5.074 3.90e-07 ***
Mexico                 -0.87447    0.29601  -2.954  0.00313 **
Uganda                 -1.09151    0.60860  -1.793  0.07290 .
category_two_defects    0.05394    0.02831   1.905  0.05672 .
---
```

```
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.55  on 930  degrees of freedom
Residual deviance:  504.42  on 923  degrees of freedom
AIC: 520.42

Number of Fisher Scoring iterations: 7
```

Firstly, we conduct a model with all significant explanatory variables and use step_AIC to select variables. In the selected model, two terms are not significant. Then, we try to delete term Uganda which has the highest p-value.

```
Call:
glm(formula = Qualityclass ~ aroma + flavor + acidity + Colombia +
    Mexico + category_two_defects, family = binomial(link = "logit"),
    data = coffee_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1273  -0.3217   0.0012   0.3439   3.4487

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -122.68887    9.00647 -13.622  < 2e-16 ***
aroma                   4.16837    0.69442   6.003 1.94e-09 ***
flavor                  7.11890    0.85662   8.310  < 2e-16 ***
acidity                 4.93107    0.71480   6.899 5.25e-12 ***
Colombia                1.89169    0.35907   5.268 1.38e-07 ***
Mexico                 -0.81385    0.29179  -2.789  0.00528 **
category_two_defects    0.05398    0.02817   1.916  0.05534 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.55  on 930  degrees of freedom
Residual deviance:  507.44  on 924  degrees of freedom
AIC: 521.44

Number of Fisher Scoring iterations: 7
```

```
Call:
glm(formula = Qualityclass ~ aroma + flavor + acidity + Colombia +
    Mexico, family = binomial(link = "logit"), data = coffee_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1419  -0.3215   0.0013   0.3473   3.3870

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -121.1845     8.8906 -13.631  < 2e-16 ***
aroma          4.1755     0.6982   5.980 2.23e-09 ***
flavor         7.0057     0.8582   8.163 3.27e-16 ***
acidity        4.8571     0.7118   6.824 8.86e-12 ***
Colombia       1.8308     0.3557   5.147 2.64e-07 ***
Mexico        -0.6596     0.2780  -2.372   0.0177 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.55  on 930  degrees of freedom
Residual deviance:  510.96  on 925  degrees of freedom
AIC: 522.96

Number of Fisher Scoring iterations: 7

Analysis of Deviance Table

Model 1: Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
    Uganda + category_two_defects
Model 2: Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
    category_two_defects
Model 3: Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico
  Resid. Df Resid. Dev Df Deviance
1       923     504.42
2       924     507.44 -1  -3.0223
3       925     510.96 -1  -3.5196

[1] 3.841459
```

After deleting Uganda, category_two_defects is still not significant. Hence, it was deleted. And we use anova to compare three models. There isn’t statistically significant difference among them. Hence, it is reasonable to delete them and get a simple model.

Final Model

```
Call:
glm(formula = Qualityclass ~ aroma + flavor + acidity + Colombia +
    Mexico, family = binomial(link = "logit"), data = coffee_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1419  -0.3215   0.0013   0.3473   3.3870

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -121.1845     8.8906 -13.631  < 2e-16 ***
aroma          4.1755     0.6982   5.980 2.23e-09 ***
flavor         7.0057     0.8582   8.163 3.27e-16 ***
acidity        4.8571     0.7118   6.824 8.86e-12 ***
Colombia       1.8308     0.3557   5.147 2.64e-07 ***
```

```
Mexico        -0.6596     0.2780  -2.372   0.0177 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1290.55  on 930  degrees of freedom
Residual deviance:  510.96  on 925  degrees of freedom
AIC: 522.96

Number of Fisher Scoring iterations: 7
```

$$\ln(\frac{p_i}{1-p_i}) = \alpha + \beta_1 \cdot aroma_i + \beta_2 \cdot flavor_i + \beta_3 \cdot acidity_i + \beta_4 \cdot \mathbb{I}_{\text{Colombia}}(x) + \beta_5 \cdot \mathbb{I}_{\text{Mexico}}(x)$$

$$\mathbb{I}_{\text{Colombia}}(x) = \begin{cases} 1 & \text{if Country of region of } x\text{th observation is Colombia,} \\ 0 & \text{Otherwise.} \end{cases}$$

$$\mathbb{I}_{\text{Mexico}}(x) = \begin{cases} 1 & \text{if Country of region of } x\text{th observation is Mexico,} \\ 0 & \text{Otherwise.} \end{cases}$$

The following is the fitted model.

$$\ln(\frac{p_i}{1-p_i}) = -121.18 + 4.18 \cdot aroma_i + 7.01 \cdot flavor_i + 4.86 \cdot acidity_i + 1.83 \cdot \mathbb{I}_{\text{Colombia}}(x) - 0.66 \cdot \mathbb{I}_{\text{Mexico}}(x)$$

Generate a summary table containing confidence intervals of estimated parameters of final model.

Table 5: confidence interval of estimated parameters

	estimate	std_error	p_value	lower_ci	upper_ci	est_exp	lower_ci_exp	upper_ci_exp
(Intercept)	-121.18	8.89	0.00	-139.59	-104.68	0.00	0.00	0.00
aroma	4.18	0.70	0.00	2.85	5.58	65.37	17.34	266.06
flavor	7.01	0.86	0.00	5.38	8.75	1107.65	217.66	6332.18
acidity	4.86	0.71	0.00	3.49	6.29	129.02	32.91	538.71
Colombia	1.83	0.36	0.00	1.16	2.56	6.23	3.18	12.89
Mexico	-0.66	0.28	0.02	-1.21	-0.12	0.52	0.30	0.89

Based on the model we built, we try to use 10-folds cross validation to test the validity of our final model. In the validation we prefer three criteria: accuracy, sensitivity and specificity.

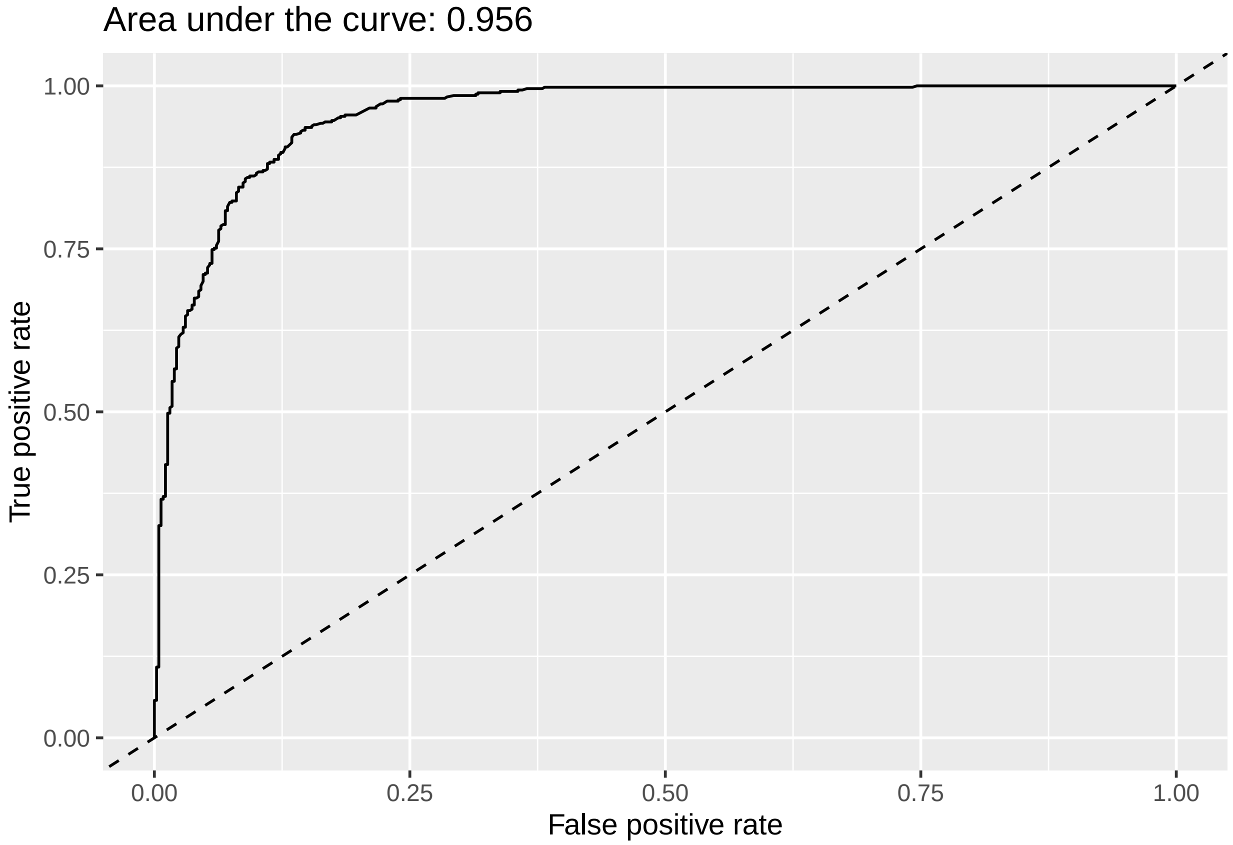
```
[1] 0.8839625
```

```
[1] 0.8931425
```

```
[1] 0.8766441
```

The accuracy of our final model is 0.88. The sensitivity of our final model is 0.89. The specificity of our final model is 0.88.

Classfication boundry

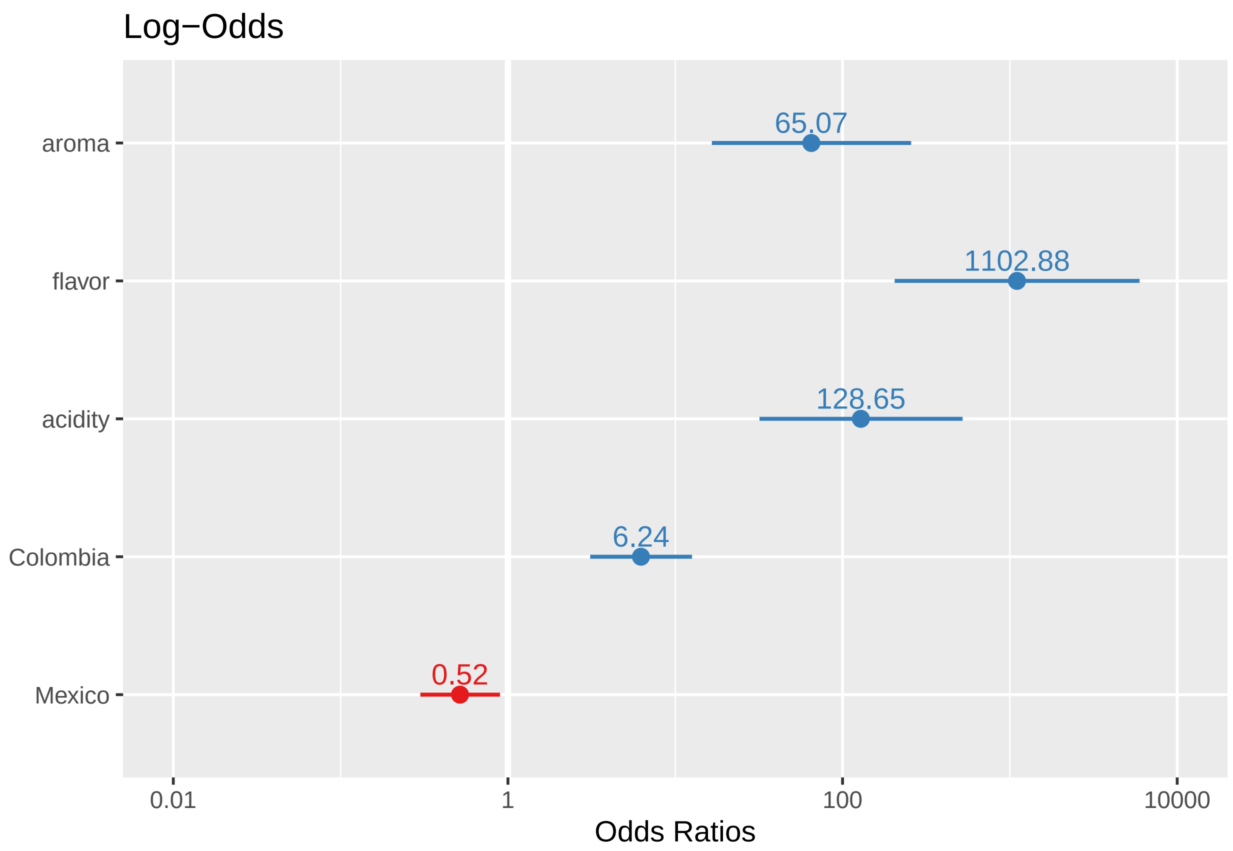


```
(Intercept)       aroma      flavor    acidity   Colombia     Mexico 
    -121.18        4.18        7.01       4.86       1.83      -0.66 

(Intercept)       aroma      flavor    acidity   Colombia     Mexico 
       0.00       65.07     1102.88     128.65       6.24       0.52 
```

Table 6: Regression coefficients and exponentiated coefficients.

	(Intercept)	aroma	flavor	acidity	Colombia	Mexico
coefficients	-121.18	4.18	7.01	4.86	1.83	-0.66
exp(coefficients)	0.00	65.07	1102.88	128.65	6.24	0.52



```
[1] 0.8871768
```

```
[1] 0.9177465
```

```
[1] 0.8571202
```

After adjusting the classification boundary. The accuracy of our final model is 0.89. The sensitivity of our final model is 0.92. The specificity of our final model is 0.86.