Factor affecting the quality of coffee beans

Library packages.

```
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(dplyr)
library(knitr)
library(janitor)
library(MASS)
library(caret)
library(GGally)
library(tables)
```

1 Exploratory data analysis

Read data and convert Qualityclass and harvested year into categorical variables.

```
data = read.csv('dataset14.csv')
data <- na.omit(data)
data$Qualityclass = as.factor(data$Qualityclass)
data$harvested = as.factor(data$harvested)
```

Deleting invalid observations there.

```
#remove Cote d?Ivoire
data <- data %>%
```

```
  filter(country_of_origin != "Cote d?Ivoire")
```

Here we map countriy_of_origin to corresponding continent.

```
unique(data$country_of_origin)
```

```
 [1] "China"                          "Mexico"
 [3] "Brazil"                         "Guatemala"
 [5] "Taiwan"                         "Uganda"
 [7] "Vietnam"                        "Thailand"
 [9] "Colombia"                       "Kenya"
[11] "Costa Rica"                     "Haiti"
[13] "Honduras"                       "Philippines"
[15] "El Salvador"                    "Indonesia"
[17] "Ethiopia"                       "Tanzania, United Republic Of"
[19] "Nicaragua"                      "Malawi"
[21] "United States"                  "Peru"
[23] "Rwanda"                         "India"
[25] "Myanmar"                        "Papua New Guinea"
[27] "Laos"                           "Panama"
[29] "Burundi"                        "United States (Puerto Rico)"
[31] "United States (Hawaii)"         "Zambia"
[33] "Ecuador"
```

```
continent_mapping <- c("China" = "Asia", "Mexico" = "North America",
 ↪  "Brazil" = "South America",
                        "Guatemala" = "North America", "Taiwan" = "Asia",
                         ↪  "Uganda" = "Africa",
                        "Vietnam" = "Asia", "Thailand" = "Asia",
                         ↪  "Colombia" = "South America",
                        "Kenya" = "Africa", "Costa Rica" = "North
                         ↪  America", "Haiti" = "North America",
                        "Honduras" = "North America", "Philippines" =
                         ↪  "Asia", "El Salvador" = "North America",
                        "Indonesia" = "Asia", "Ethiopia" = "Africa",
                        "Tanzania, United Republic Of" = "Africa",
                         ↪  "Nicaragua" = "North America", "Malawi" =
                         ↪  "Africa",
                        "United States" = "North America", "Peru" =
                         ↪  "South America", "Rwanda" = "Africa",
```

```
                        "India" = "Asia", "Myanmar" = "Asia", "Papua New
                         ↪  Guinea" = "Oceania", "Laos" = "Asia",
                        "Panama" = "North America", "Burundi" = "Africa",
                        "United States (Puerto Rico)" = "North America",
                         ↪  "United States (Hawaii)" = "North America",
                        "Cote d'Ivoire" = "Africa", "Zambia" = "Africa",
                         ↪  "Ecuador" = "South America")

# map countries to its continent
data$continent <- continent_mapping[data$country_of_origin]
table(data$continent)
```

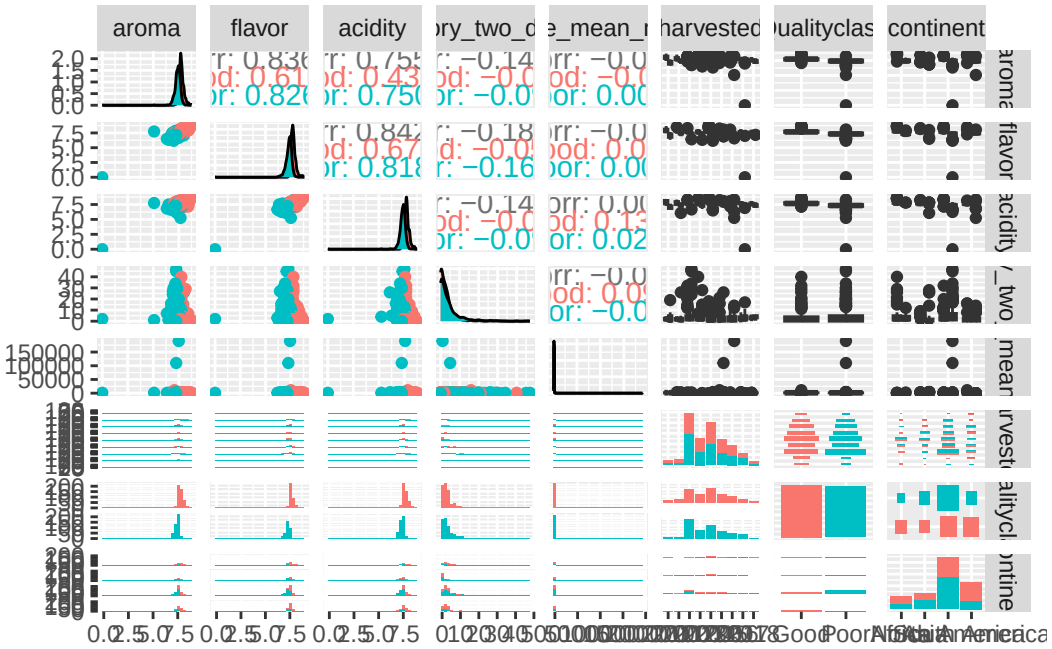
```
    Africa       Asia North America    Oceania South America
       116        136            447          1            225
```

Delete sole observation in Oceania.

```
data = data %>% filter(continent != 'Oceania')
```

Pairwise plot for variables:

```
ggpairs(data[,-1],mapping = aes(color = Qualityclass))
```



Summary statistics for numerical variables with respect to different quality classes:

```
table = tabular(aroma + flavor + acidity + category_two_defects +
 ↪  altitude_mean_meters ~ Qualityclass * (mean+sd+min+max+IQR),
 ↪  data=data)
print(table, type = 'latex')
```

	Qualityclass					
	Good					Poor
	mean	sd	min	max	IQR	mean
aroma	7.761	0.2329	7.08	8.75	0.25	7.373
flavor	7.744	0.2268	7.25	8.83	0.25	7.288
acidity	7.723	0.2419	7.17	8.75	0.25	7.325
category_two_defects	2.806	4.0931	0.00	40.00	4.00	4.271
altitude_mean_meters	1418.261	637.0004	1.00	11000.00	500.00	1907.025

sd	min	max	IQR
4.408e-01	0	8.25	0.33
4.396e-01	0	8.08	0.33

```
4.359e-01 0         8.33   0.33
6.125e+00 0        47.00   5.00
1.023e+04 1    190164.00 450.00
```

Mapping good quality and poor quality into 1 and 0 respectively.

```r
data$Qualityclass = as.numeric(data$Qualityclass)
data$Qualityclass[data$Qualityclass == 2] = 0
```

2 Model selection

Here we have continent as our new categorical explanatory variable to fit a full logistic regression model:

```r
#Here we have continent as our new categorical explanatory variable
full_model = glm(Qualityclass ~ . - country_of_origin, family =
 ↪  binomial(link = 'logit'), data = data)
```

```r
summary(full_model)
```

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

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.270e+02  9.381e+00 -13.543  < 2e-16 ***
aroma                    4.635e+00  7.216e-01   6.424 1.33e-10 ***
flavor                   7.511e+00  8.995e-01   8.349  < 2e-16 ***
acidity                  4.630e+00  7.085e-01   6.535 6.37e-11 ***
category_two_defects     6.422e-02  3.042e-02   2.111   0.0348 *
altitude_mean_meters    -3.760e-06  1.888e-05  -0.199   0.8421
harvested2011           -1.861e-01  1.113e+00  -0.167   0.8672
harvested2012           -3.086e-01  9.383e-01  -0.329   0.7422
harvested2013            1.775e-01  9.514e-01   0.187   0.8520
harvested2014            1.906e-01  9.514e-01   0.200   0.8412
harvested2015           -2.182e-01  9.608e-01  -0.227   0.8203
```

```
harvested2016           6.395e-02  9.757e-01   0.066   0.9477
harvested2017           2.482e-01  9.951e-01   0.249   0.8031
harvested2018           8.726e-01  1.189e+00   0.734   0.4632
continentAsia           1.921e-01  4.945e-01   0.388   0.6977
continentNorth America -3.333e-01  4.288e-01  -0.777   0.4370
continentSouth America  1.049e+00  4.512e-01   2.326   0.0200 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1280.66  on 923  degrees of freedom
Residual deviance:  502.89  on 907  degrees of freedom
AIC: 536.89

Number of Fisher Scoring iterations: 7
```

It is shown that the 'harvested' and 'altitude_mean_meters' variables are insignificant.

Then we use stepAIC to measure goodness of fit.

```
#direction = 'backward'
stepAIC(full_model)
```

```
Start:  AIC=536.89
Qualityclass ~ (country_of_origin + aroma + flavor + acidity +
    category_two_defects + altitude_mean_meters + harvested +
    continent) - country_of_origin

                      Df Deviance    AIC
- harvested            8   507.83 525.83
- altitude_mean_meters  1   502.94 534.94
<none>                      502.89 536.89
- category_two_defects  1   507.06 539.06
- continent             3   525.92 553.92
- acidity               1   553.33 585.33
- aroma                 1   556.62 588.62
- flavor                1   598.95 630.95

Step:  AIC=525.83
Qualityclass ~ aroma + flavor + acidity + category_two_defects +
    altitude_mean_meters + continent
```

```
                       Df Deviance    AIC
- altitude_mean_meters  1   507.88 523.88
<none>                      507.83 525.83
- category_two_defects  1   510.91 526.91
- continent             3   537.01 549.01
- acidity               1   562.25 578.25
- aroma                 1   564.58 580.58
- flavor                1   603.65 619.65

Step:  AIC=523.88
Qualityclass ~ aroma + flavor + acidity + category_two_defects +
    continent

                       Df Deviance    AIC
<none>                      507.88 523.88
- category_two_defects  1   510.97 524.97
- continent             3   537.19 547.19
- acidity               1   562.25 576.25
- aroma                 1   564.73 578.73
- flavor                1   604.01 618.01


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

Coefficients:
           (Intercept)                   aroma                  flavor
            -125.45755                 4.53950                 7.33458
               acidity    category_two_defects           continentAsia
               4.71371                 0.05364                 0.05621
continentNorth America  continentSouth America
              -0.50889                 0.92507

Degrees of Freedom: 923 Total (i.e. Null);  916 Residual
Null Deviance:      1281
Residual Deviance: 507.9    AIC: 523.9
```

Both indicated that to obtain a better glm model, 'harvested' and 'altitude_mean_meters' variable should be removed.

Therefore we obtained a better glm model.

```r
model1 = glm(formula = Qualityclass ~ aroma + flavor + acidity +
               category_two_defects + continent,
             family = binomial(link = "logit"), data = data)
summary(model1)
```

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

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -125.45755    9.07730 -13.821  < 2e-16 ***
aroma                      4.53950    0.68859   6.592 4.33e-11 ***
flavor                     7.33458    0.87497   8.383  < 2e-16 ***
acidity                    4.71371    0.69499   6.782 1.18e-11 ***
category_two_defects       0.05364    0.02973   1.804   0.0712 .
continentAsia              0.05621    0.48188   0.117   0.9071
continentNorth America    -0.50889    0.40399  -1.260   0.2078
continentSouth America     0.92507    0.42427   2.180   0.0292 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1280.66  on 923  degrees of freedom
Residual deviance:  507.88  on 916  degrees of freedom
AIC: 523.88

Number of Fisher Scoring iterations: 7
```

However, in the model with lowest AIC 523.88, variable category_two_defects is not so significant.

```r
model2 = glm(formula = Qualityclass ~ aroma + flavor + acidity +
      continent, family = binomial(link = "logit"), data = data)
summary(model2)
```

```
Call:
```

```
glm(formula = Qualityclass ~ aroma + flavor + acidity + continent,
    family = binomial(link = "logit"), data = data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -123.55003    8.90391 -13.876  < 2e-16 ***
aroma                      4.47083    0.69093   6.471 9.75e-11 ***
flavor                     7.27551    0.88012   8.266  < 2e-16 ***
acidity                    4.61051    0.69151   6.667 2.61e-11 ***
continentAsia             -0.04023    0.47417  -0.085   0.9324
continentNorth America    -0.45974    0.39457  -1.165   0.2440
continentSouth America     0.87870    0.41706   2.107   0.0351 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1280.66  on 923  degrees of freedom
Residual deviance:  510.97  on 917  degrees of freedom
AIC: 524.97

Number of Fisher Scoring iterations: 7
```

The model dropping category_two_defects has AIC 524.97, which is slightly higher than 523.88. But all the variables are significant.

So we want to compare this two models by predictive power.

3 Predictive power

To compare predictive power of these two models, mean accuracy rate of 10-fold cross validation is used there.

```
set.seed(14)
folds <- createFolds(y=data$Qualityclass,k=10)
```

```
total.n = 0
right1.n = 0
right2.n = 0
for(i in 1:10){
```

```
  fold.val = data[folds[[i]],]
  fold.train = data[-folds[[i]],]
  model1.val = glm(Qualityclass ~ aroma + flavor + acidity +
              category_two_defects + continent,
            family = binomial(link = "logit"), data = fold.train)
  predict1.val = ifelse(predict(model1.val,
 ↪  type='response',newdata=fold.val)>0.5, 1, 0)
  model2.val = glm(Qualityclass ~ aroma + flavor + acidity + continent,
            family = binomial(link = "logit"), data = fold.train)
  predict2.val = ifelse(predict(model2.val,
 ↪  type='response',newdata=fold.val)>0.5, 1, 0)
  total.n = total.n + nrow(fold.val)
  right1.n = right1.n + sum(predict1.val == fold.val$Qualityclass)
  right2.n = right2.n + sum(predict2.val == fold.val$Qualityclass)
}
cat('Mean accuracy rate for model including category_two_defects: ',
 ↪  right1.n/total.n, "\n",'Mean accuracy rate for model dropping
 ↪  category_two_defects: ', right2.n/total.n,"\n")
```

```
Mean accuracy rate for model including category_two_defects:  0.8744589
 Mean accuracy rate for model dropping category_two_defects:  0.8722944
```

Including variable category_two_defects cause a slightly higher accuracy rate.

4 Result and Conclusion

We start with full_model whose AIC=536.89, Residual deviance=502.89.
By removing 'harvested' and 'altitude_mean_meters' variable which were insignificant: AIC=523.88, Residual Deviance=507.88.

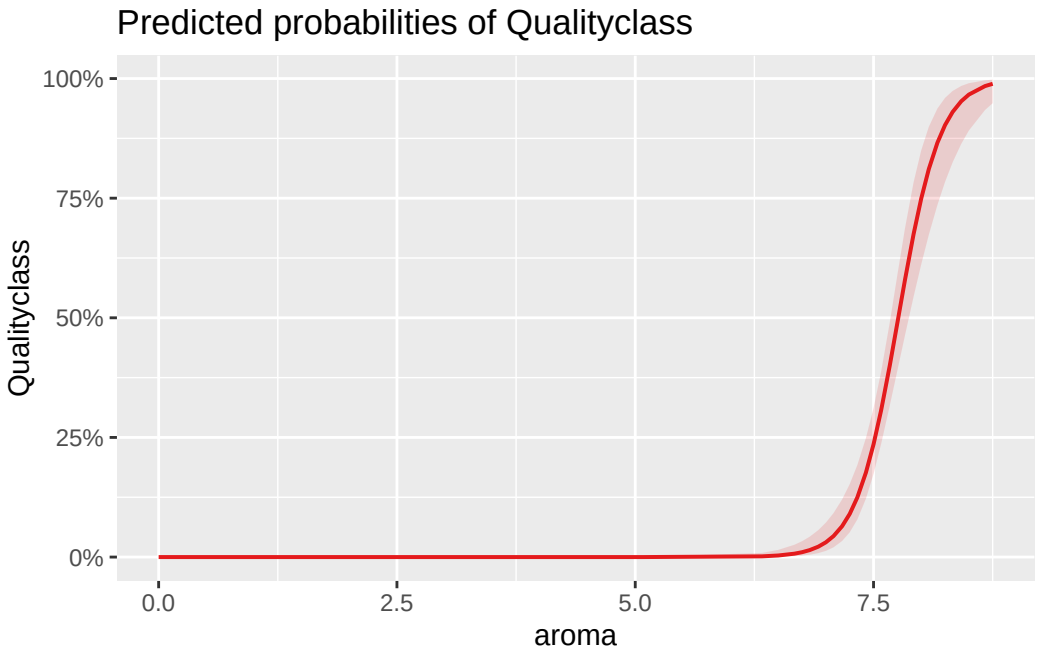
Finally, we choose the binomial generalized linear model with aroma, flavor, acidity, category_two_defects and continent to fit and predict quality class of coffee.

5 Visualization for model

Based on the model we choose, estimated probabilities for a good quality class with respect to each variable:

1. aroma

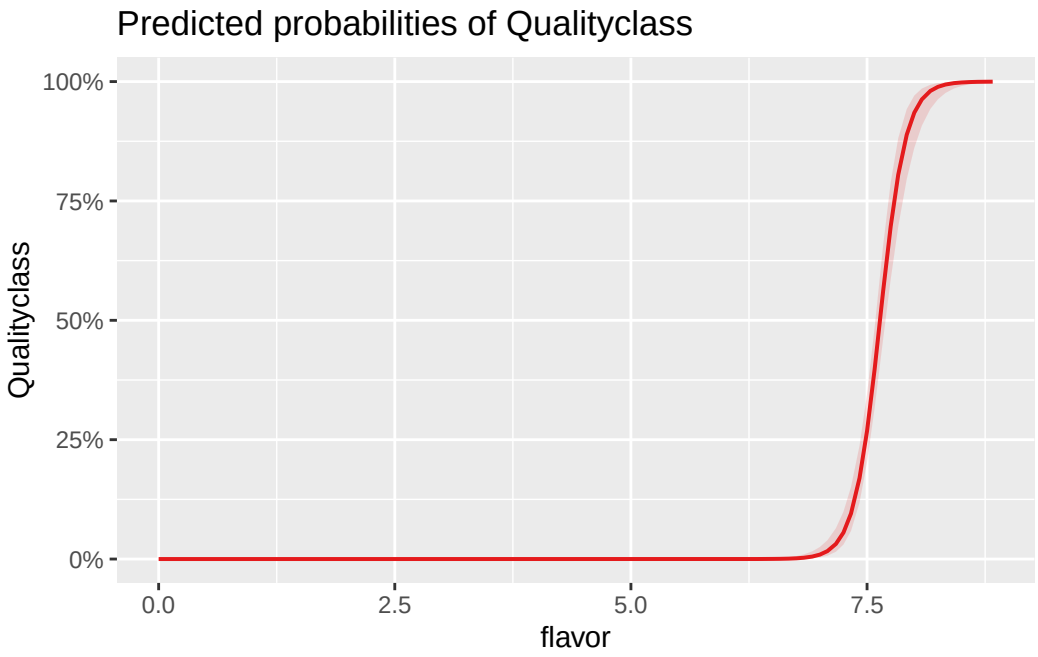
```
plot_model(model1, type = "pred", terms = "aroma[all]")
```



The predicted probability of Qualityclass increases sharply as the aroma score increases. This suggests a strong positive association between the quality of coffee and its aroma.

2. flavor

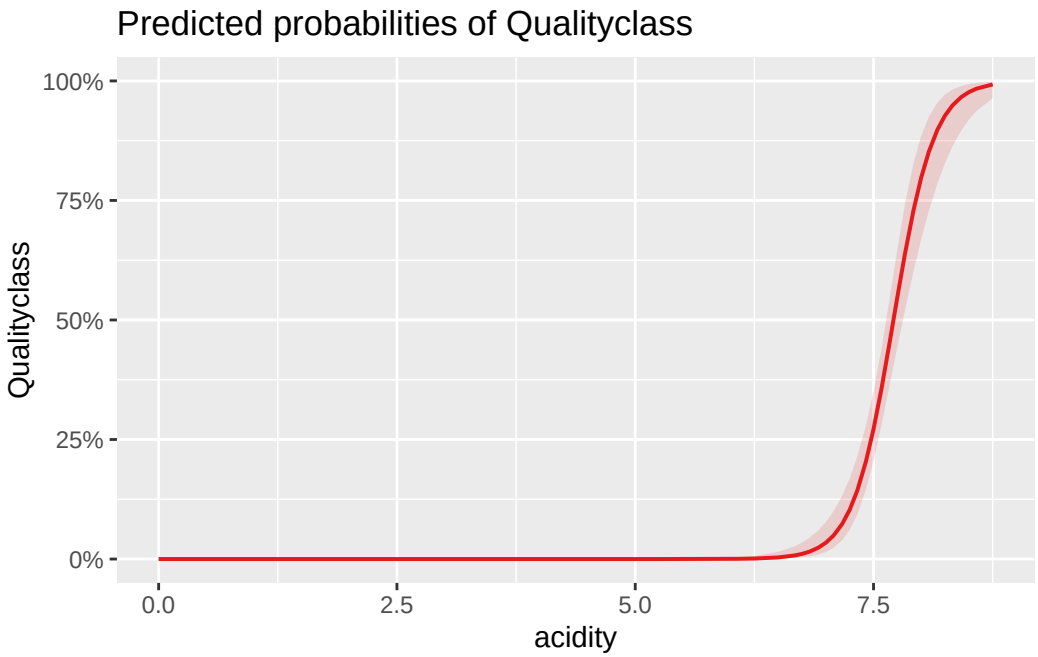
```
plot_model(model1, type = "pred", terms = "flavor[all]")
```



Similar to aroma, the probability of a higher Qualityclass increases sharply with the flavor score. Flavor appears to be a significant predictor of coffee quality.

3. acidity

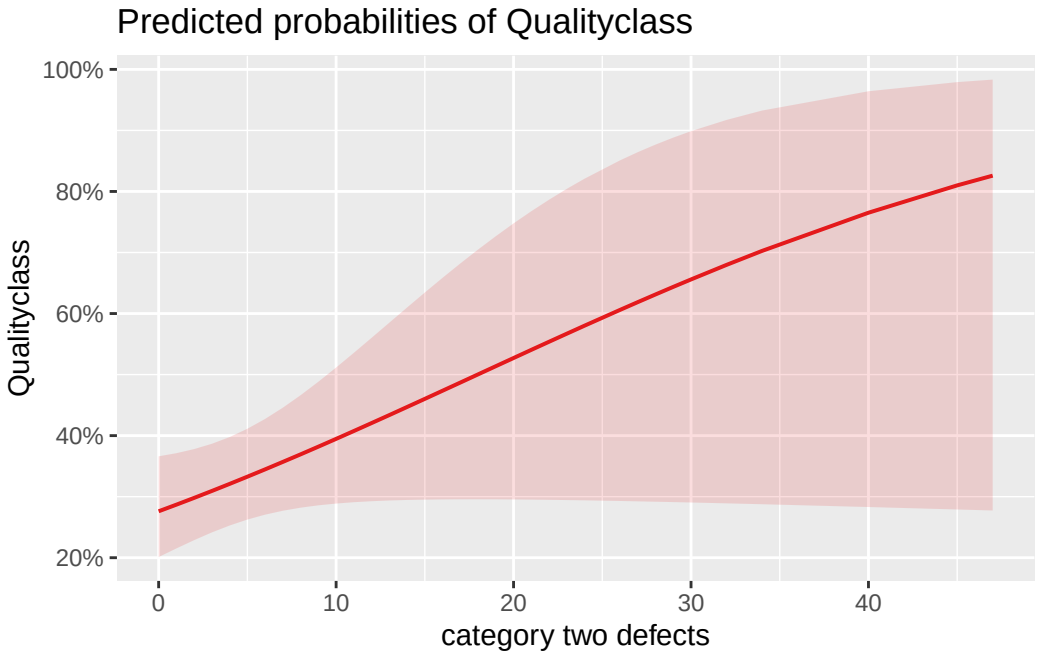
```
plot_model(model1, type = "pred", terms = "acidity[all]")
```



The plot shows that higher acidity scores are associated with a higher probability of Qualityclass. There is a positive relationship between acidity and coffee quality, though it seems to have a threshold effect, with probability rising more sharply after a certain point.

4. category_two_defects

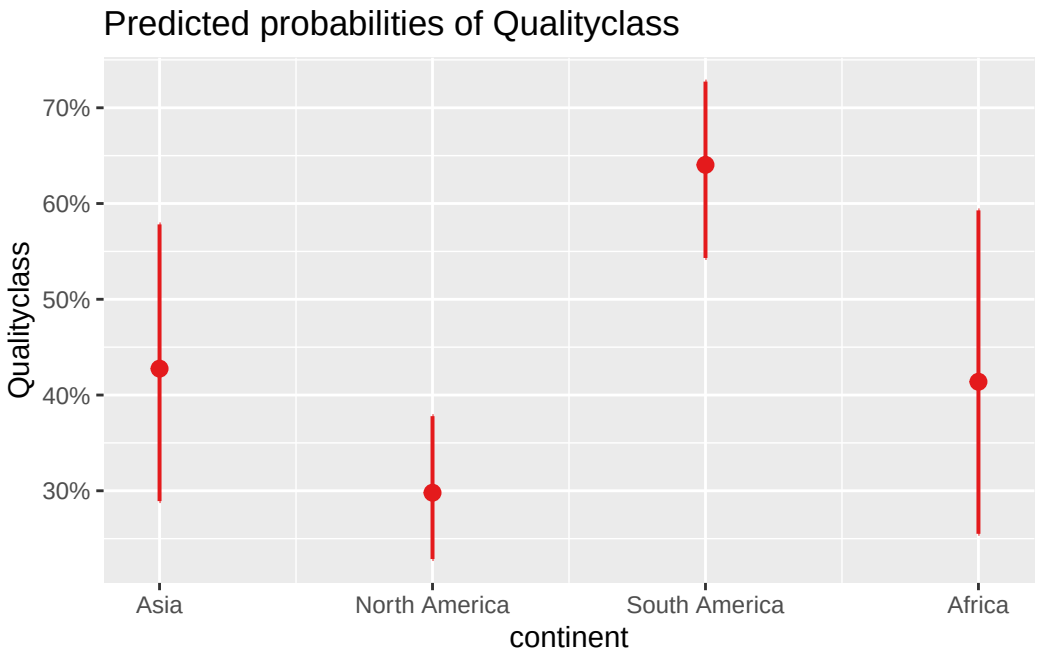
```
plot_model(model1, type = "pred", terms = "category_two_defects[all]")
```



As the number of defects increases, the probability of a higher Qualityclass increases. This is somewhat counterintuitive, as one might expect more defects to decrease quality. However, the confidence band is wide, suggesting there may be high variability or that this variable interacts with others in complex ways.

5. continent

```
plot_model(model1, type = "pred", terms = "continent")
```



There appears to be variation among continents, with some showing higher probabilities than others, suggesting geographical variation in coffee quality.