Group_11_Analysis

Group11

2023-03-09

Aims of the Analysis

Exploratory Data Analysis

```
# load library
library(tidyr)
library(ggplot2)
library(skimr)
library(dplyr)
library(gridExtra)
library(GGally)
library(gmodels)
library(stats)
library(sjPlot)
library(jtools)
library(MASS)

# load the dataset for group 11
coffee <- read.csv('https://raw.githubusercontent.com/rrachelxi/DAS-Group-11/main/dataset11.csv')
```

```
# skim the dataset
skim_without_charts(coffee)
```

Table 1: Data summary

Name	coffee
Number of rows	1094
Number of columns	8
Column type frequency:	
character	2
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
country_of_origin	1	1	4	28	0	35	0
Qualityclass	0	1	4	4	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
aroma	0	1.00	7.57	0.32	5.08	7.42	7.58	7.75	8.75
flavor	0	1.00	7.52	0.34	6.08	7.33	7.58	7.75	8.83
acidity	0	1.00	7.54	0.32	5.25	7.33	7.58	7.75	8.75
category_two_defects	0	1.00	3.56	5.33	0.00	0.00	2.00	4.00	55.00
altitude_mean_meters	191	0.83	1649.82	7262.27	1.00	1100.00	1310.64	1600.00	190164.00
harvested	56	0.95	2013.69	1.81	2010.00	2012.00	2014.00	2015.00	2018.00

```
# There are missing values in altitude_mean_meters, country_of_origin and harvested
# For altitude_mean_meters:
# missing values are replaced with altitude_mean_meters mean values
# For country_of_origin and harvested:
# since observations with NA are in relatively small amount, they are dropped
```

```
# Calculate the mean value for altitude_mean_meters
mean_altitude <- mean(coffee$altitude_mean_meters,na.rm = TRUE)

# Perform data wrangling to the dataset
coffee_m <- coffee %>%
  mutate(
    # turn harvested into the factor type  for analysis
    harvested = as.factor(harvested),
    # turn country_of_origin into the factor type for analysis
    country_of_origin = as.factor(country_of_origin),
    # turn Qualityclass into the factor type for analysis
    Qualityclass = as.factor(Qualityclass),
    # Replace NA in altitude_mean_meters by the mean value)
    altitude_mean_meters = replace_na(altitude_mean_meters, mean_altitude)) %>%
  filter(
    !is.na(country_of_origin) & # drop observations with NA in country_of_origin
    !is.na(harvested)) # drop observations with NA in harvested

# Skim the dataset the second time to check if NA values still present
skim_without_charts(coffee_m)
```

Table 4: Data summary

Name	coffee_m
Number of rows	1038
Number of columns	8
Column type frequency:	
factor	3
numeric	5

Table 4: Data summary

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
country_of_origin	0	1	FALSE	34	Mex: 191, Gua: 151, Col: 140, Bra: 101
harvested	0	1	FALSE	9	201: 283, 201: 215, 201: 148, 201: 130
Qualityclass	0	1	FALSE	2	Goo: 527, Poo: 511

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
aroma	0	1	7.57	0.32	5.08	7.42	7.58	7.75	8.75
flavor	0	1	7.52	0.34	6.08	7.33	7.58	7.75	8.83
acidity	0	1	7.54	0.32	5.25	7.33	7.58	7.75	8.75
category_two_defects	0	1	3.55	5.15	0.00	0.00	2.00	4.00	45.00
altitude_mean_meters	0	1	1653.80	6772.82	1.00	1200.00	1400.00	1649.82	190164.00

```
# The dataset is now with no NA values
```

```
# Draw pair plot to visualize data
ggpairs(coffee_m[,c(-1,-7)],
        aes(color=Qualityclass,alpha=0.2),
        title = "Distribution between variables")
```

```
# Outliers are spotted for aroma, acidity and altitude_mean_meters
```

```
# Draw scatter plots to show outliers more clearly
# Draw scatter plots for aroma and acidity
ggplot(coffee_m,aes(x=aroma,y=acidity,color=Qualityclass)) +
  geom_point() +
  geom_jitter(width = 0.2, height = 0.2) +
  labs(x = "Aroma grade (ranging from 0-10)",
       y = "Acidity grade (ranging from 0-10)",
       title = "Aroma grade and acidity grade for coffee quality")
```

```
# Draw scatter plots for altitude_mean_meters
ggplot(coffee_m,aes(x=Qualityclass,y=altitude_mean_meters,color=Qualityclass)) +
  geom_point() +
  geom_jitter(width = 0.2, height = 0.2) +
  labs(x = "Quality Class",
       y = "Mean altitude of the growers farm (in metres)",
       title = "Mean altitude of the growers group by coffee quality") +
  theme(legend.position = 'none')
```

Figure 1: Pair plot of numeric variables and Qualityclass.

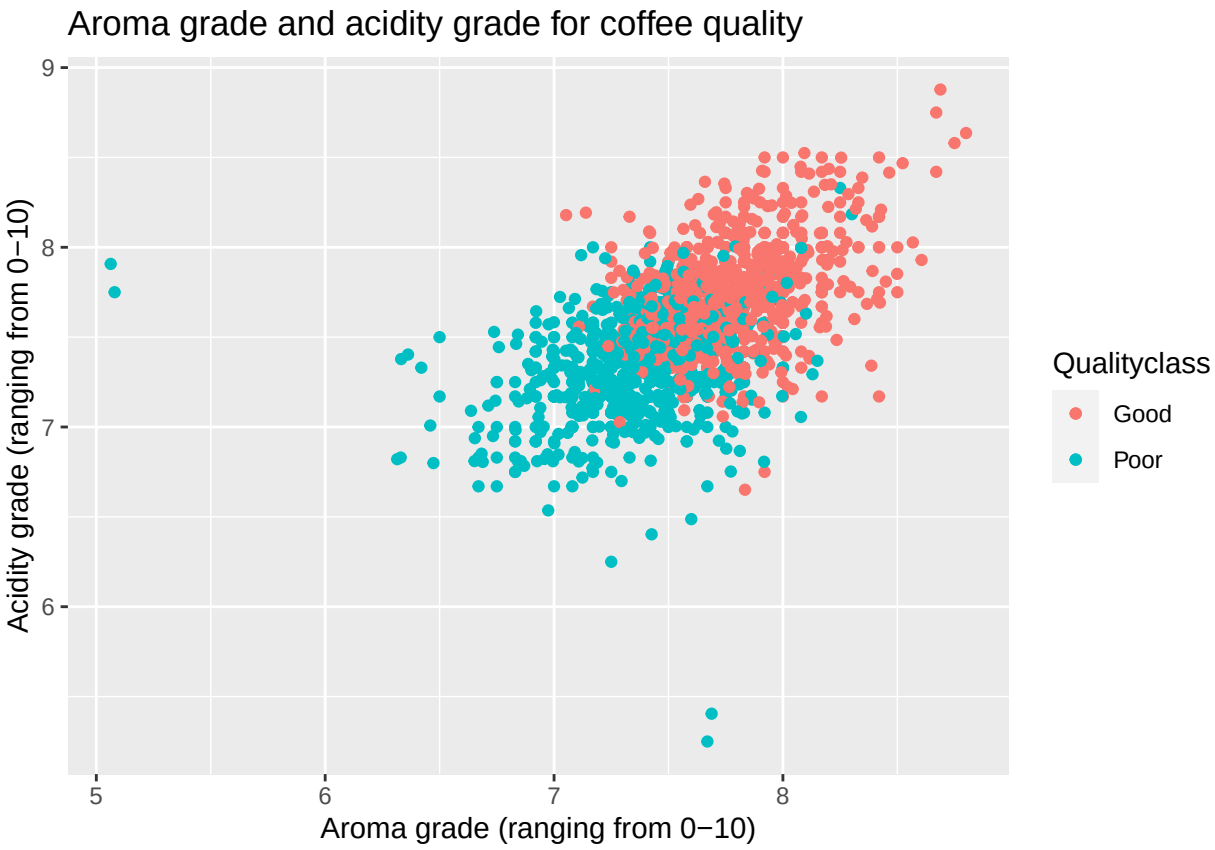

Figure 2: Dsitribution between aroma and acidity.

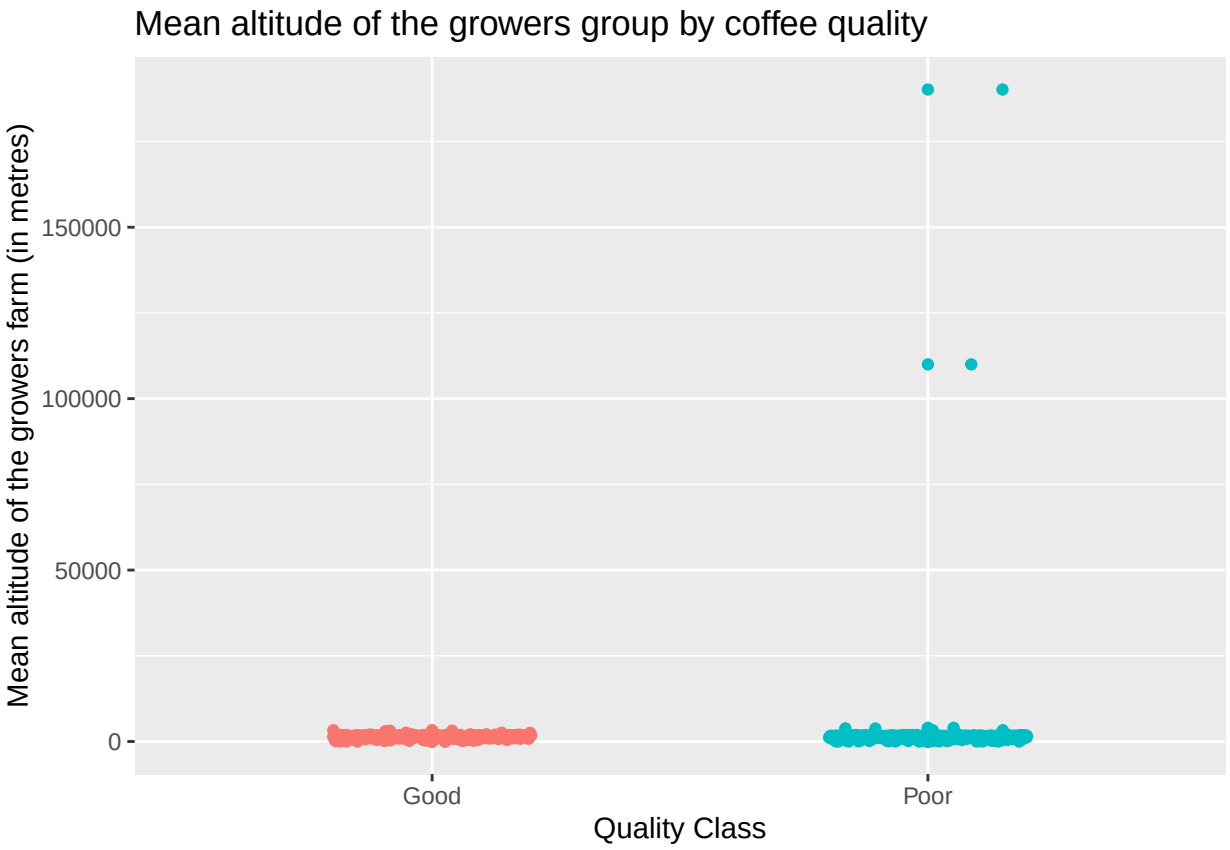

Figure 3: Scatter plot of altitude_mean_meters.

```r
# Drop outlier observations from the dataset
coffee_w <- coffee_m %>%
  filter(
    aroma > 6 &
    acidity > 6 &
    altitude_mean_meters < 100000)

# Skim the dataset the third time to give insights on variables
skim_without_charts(coffee_w)
```

Table 7: Data summary

Name	coffee_w
Number of rows	1034
Number of columns	8
Column type frequency:	
factor	3
numeric	5
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
country_of_origin	0	1	FALSE	34	Mex: 191, Gua: 150, Col: 138, Bra: 101
harvested	0	1	FALSE	9	201: 283, 201: 215, 201: 148, 201: 128
Qualityclass	0	1	FALSE	2	Goo: 527, Poo: 507

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
aroma	0	1	7.57	0.31	6.33	7.42	7.58	7.75	8.75
flavor	0	1	7.52	0.34	6.08	7.33	7.58	7.75	8.83
acidity	0	1	7.54	0.31	6.25	7.33	7.58	7.75	8.75
category_two_defects	0	1	3.56	5.15	0.00	0.00	2.00	4.00	45.00
altitude_mean_meters	0	1	1366.81	448.90	1.00	1200.00	1400.00	1649.82	4001.00

```r
# For numeric variables, pair plot, inclduing boxplots, is drawn to show distribution and correlation
ggpairs(coffee_w[c(-1,-7)],
        aes(color=Qualityclass,alpha=0.2),
        title = "Distribution between variables(without outliers)")
```

```r
# Collinearity is shown in pair plot, could be discussed in extension/further work
```

Figure 4: Pair plot of numeric variables and Qualityclass (without outliers).

```
# For categorical variables
# Draw a barplot to visualize the distribution of harvested and Qualityclass
ggplot(coffee_w, aes(x=harvested, y= after_stat(prop), group=Qualityclass, fill=Qualityclass )) +
  geom_bar(position = "dodge") +
  labs(x = "Year the batch was harvested",
       y = "Proportion",
       title = "Porportions of quality class by year the batch was harvested")
```

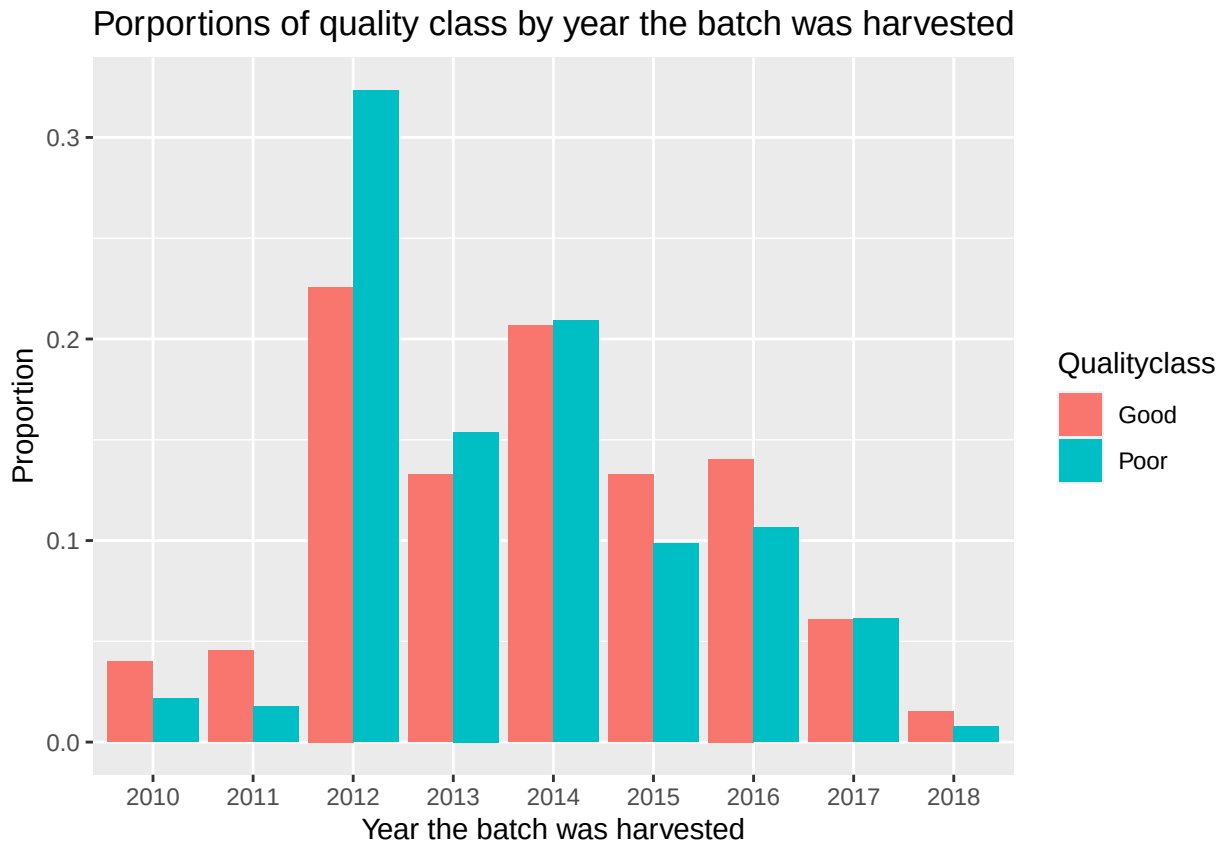

Figure 5: Bar plot of quality class by harvested year.

```
# Build a table of proportions to support the barplot of harvested and Qualityclass
prop.table(table(coffee_w$Qualityclass,coffee_w$harvested),2) %>%
  round(digits=2)
```

```
##
##         2010 2011 2012 2013 2014 2015 2016 2017 2018
##   Good  0.66 0.73 0.42 0.47 0.51 0.58 0.58 0.51 0.67
##   Poor  0.34 0.27 0.58 0.53 0.49 0.42 0.42 0.49 0.33
```

```
# Draw a barplot to visualize the distribution of country_of_origin and Qualityclass
ggplot(coffee_w, aes(x=country_of_origin, group=Qualityclass )) +
  geom_bar(aes(y=after_stat(prop), fill = Qualityclass),  position = "dodge") +
  labs(x = "Country of origin",
```

```
      y = "Proportion",
      title = "Porportions of quality class by country of origin") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

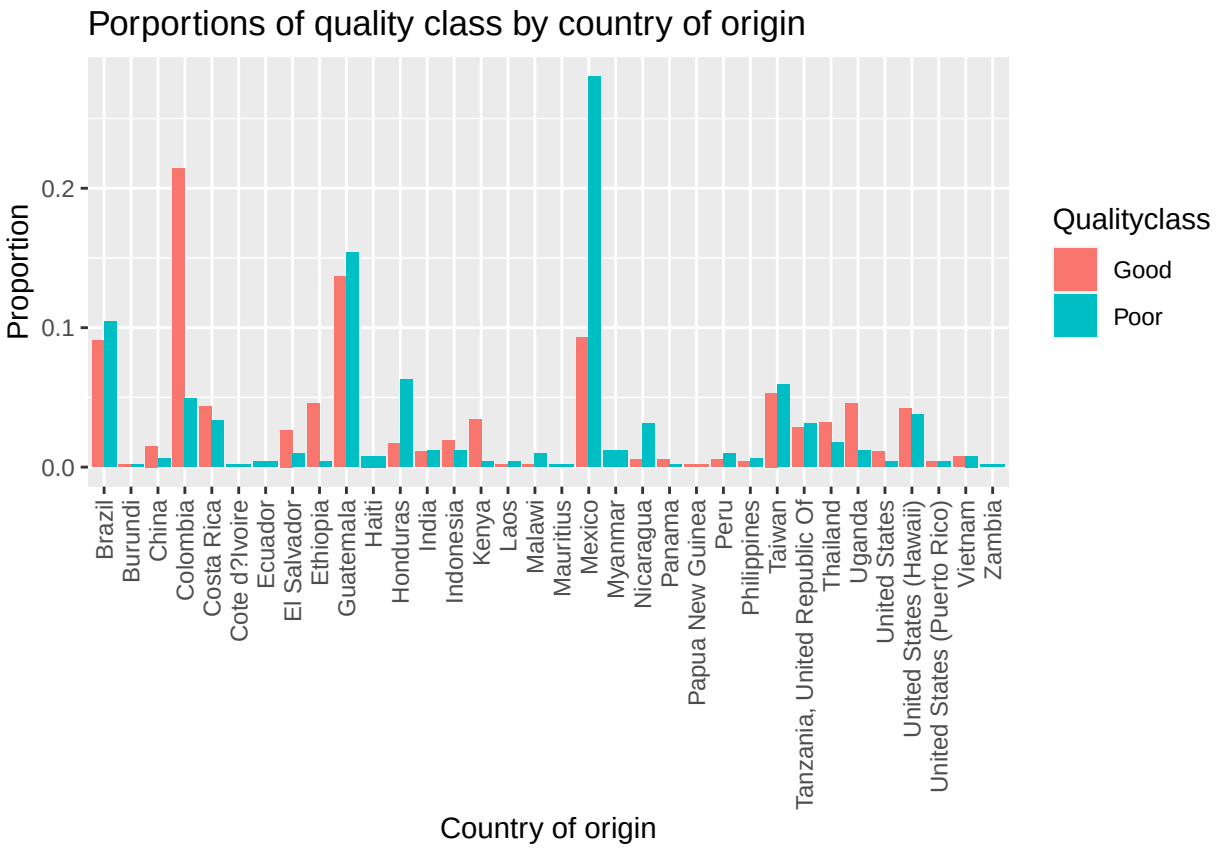

Figure 6: Barplot of quality class by country of origin.

```
# Buil-d a table of proportions to support the barplot of country_of_origin and Qualityclass
prop.table(table(coffee_w$Qualityclass,coffee_w$country_of_origin),2) %>%
  round(digits=2)
```

```
##
##        Brazil Burundi China Colombia Costa Rica Cote d?Ivoire Ecuador
##   Good   0.48    0.50  0.73     0.82       0.58          0.00    0.00
##   Poor   0.52    0.50  0.27     0.18       0.42          1.00    1.00
##
##        El Salvador Ethiopia Guatemala Haiti Honduras India Indonesia Japan
##   Good        0.74     0.92      0.48  0.00     0.22  0.50      0.62
##   Poor        0.26     0.08      0.52  1.00     0.78  0.50      0.38
##
##        Kenya Laos Malawi Mauritius Mexico Myanmar Nicaragua Panama
##   Good  0.90 0.33   0.17      0.00   0.26    0.00      0.16   0.75
##   Poor  0.10 0.67   0.83      1.00   0.74    1.00      0.84   0.25
##
##        Papua New Guinea Peru Philippines Taiwan Tanzania, United Republic Of
```

```
##   Good             1.00 0.38        0.40   0.48                          0.48
##   Poor             0.00 0.62        0.60   0.52                          0.52
##
##        Thailand Uganda United States United States (Hawaii)
##   Good     0.65   0.80          0.75                   0.54
##   Poor     0.35   0.20          0.25                   0.46
##
##        United States (Puerto Rico) Vietnam Zambia
##   Good                        0.50    0.50   0.00
##   Poor                        0.50    0.50   1.00
```

Statistical Modelling

```
# Build the full glm model
model_full <- glm(
  Qualityclass ~
               aroma +
               flavor +
               acidity +
               category_two_defects +
               altitude_mean_meters +
               harvested +
               country_of_origin,
  data = coffee_w,
  family = binomial(link="logit"))

# Conduct stepwise feature selection by AIC
model_step<-model_full %>%
  stepAIC(trace=TRUE)
```

```
## Start:  AIC=644.83
## Qualityclass ~ aroma + flavor + acidity + category_two_defects +
##     altitude_mean_meters + harvested + country_of_origin
##
##                         Df Deviance    AIC
## - harvested              8   560.72 638.72
## <none>                       550.83 644.83
## - category_two_defects   1   553.31 645.31
## - altitude_mean_meters   1   553.93 645.93
## - country_of_origin     33   628.97 656.97
## - acidity                1   586.48 678.48
## - aroma                  1   595.92 687.92
## - flavor                 1   660.94 752.94
##
## Step:  AIC=638.72
## Qualityclass ~ aroma + flavor + acidity + category_two_defects +
##     altitude_mean_meters + country_of_origin
##
##                         Df Deviance    AIC
## <none>                       560.72 638.72
## - altitude_mean_meters   1   562.85 638.85
```

```
## - category_two_defects  1   563.16 639.16
## - country_of_origin    33   641.40 653.40
## - acidity               1   601.19 677.19
## - aroma                 1   603.56 679.56
## - flavor                1   670.49 746.49
```

```
# Pull summary of model chosen by stepwise selection
summary(model_step)
```

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity + category_two_defects +
##     altitude_mean_meters + country_of_origin, family = binomial(link = "logit"),
##     data = coffee_w)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2694  -0.3475  -0.0022   0.3085   4.0541
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                   1.210e+02  8.793e+00  13.757
## aroma                                        -4.238e+00  6.910e-01  -6.132
## flavor                                       -7.613e+00  8.576e-01  -8.878
## acidity                                      -4.090e+00  6.711e-01  -6.094
## category_two_defects                         -4.385e-02  2.821e-02  -1.554
## altitude_mean_meters                         -4.205e-04  2.883e-04  -1.458
## country_of_originBurundi                     -1.551e+00  4.237e+00  -0.366
## country_of_originChina                       -2.767e-01  1.054e+00  -0.262
## country_of_originColombia                    -1.116e+00  4.447e-01  -2.510
## country_of_originCosta Rica                   1.509e-01  6.275e-01   0.241
## country_of_originCote d?Ivoire                1.105e+01  2.400e+03   0.005
## country_of_originEcuador                      1.771e+01  1.691e+03   0.010
## country_of_originEl Salvador                 -4.135e-01  8.003e-01  -0.517
## country_of_originEthiopia                     1.735e-01  1.279e+00   0.136
## country_of_originGuatemala                    4.247e-01  4.320e-01   0.983
## country_of_originHaiti                        1.145e+01  8.366e+02   0.014
## country_of_originHonduras                     6.932e-01  6.343e-01   1.093
## country_of_originIndia                        2.484e+00  9.133e-01   2.720
## country_of_originIndonesia                    2.390e-01  8.842e-01   0.270
## country_of_originKenya                       -9.692e-01  1.346e+00  -0.720
## country_of_originLaos                        -4.380e-01  1.713e+00  -0.256
## country_of_originMalawi                       8.971e-01  1.261e+00   0.711
## country_of_originMauritius                    1.097e+01  2.400e+03   0.005
## country_of_originMexico                       1.020e+00  4.056e-01   2.515
## country_of_originMyanmar                      1.507e+01  9.198e+02   0.016
## country_of_originNicaragua                    1.746e-01  1.407e+00   0.124
## country_of_originPanama                      -2.478e+00  1.670e+00  -1.484
## country_of_originPapua New Guinea            -3.292e+00  2.400e+03  -0.001
## country_of_originPeru                         3.309e+00  1.495e+00   2.213
## country_of_originPhilippines                 -2.009e+00  2.207e+00  -0.911
## country_of_originTaiwan                      -6.576e-01  5.897e-01  -1.115
## country_of_originTanzania, United Republic Of -1.059e+00  6.780e-01  -1.563
## country_of_originThailand                    -1.720e+00  7.113e-01  -2.418
```

```
## country_of_originUganda                          1.089e+00  6.877e-01   1.583
## country_of_originUnited States                  -1.427e+00  1.654e+00  -0.862
## country_of_originUnited States (Hawaii)          2.051e-01  6.525e-01   0.314
## country_of_originUnited States (Puerto Rico)     1.284e+00  1.387e+00   0.925
## country_of_originVietnam                        -1.686e+00  1.081e+00  -1.560
## country_of_originZambia                          1.265e+01  2.400e+03   0.005
##                                                 Pr(>|z|)
## (Intercept)                                      < 2e-16 ***
## aroma                                           8.65e-10 ***
## flavor                                           < 2e-16 ***
## acidity                                         1.10e-09 ***
## category_two_defects                             0.12007
## altitude_mean_meters                             0.14475
## country_of_originBurundi                         0.71430
## country_of_originChina                           0.79301
## country_of_originColombia                        0.01207 *
## country_of_originCosta Rica                      0.80994
## country_of_originCote d?Ivoire                   0.99633
## country_of_originEcuador                         0.99164
## country_of_originEl Salvador                     0.60533
## country_of_originEthiopia                        0.89212
## country_of_originGuatemala                       0.32551
## country_of_originHaiti                           0.98909
## country_of_originHonduras                        0.27445
## country_of_originIndia                           0.00652 **
## country_of_originIndonesia                       0.78691
## country_of_originKenya                           0.47155
## country_of_originLaos                            0.79814
## country_of_originMalawi                          0.47692
## country_of_originMauritius                       0.99635
## country_of_originMexico                          0.01190 *
## country_of_originMyanmar                         0.98692
## country_of_originNicaragua                       0.90127
## country_of_originPanama                          0.13775
## country_of_originPapua New Guinea                0.99891
## country_of_originPeru                            0.02691 *
## country_of_originPhilippines                     0.36252
## country_of_originTaiwan                          0.26486
## country_of_originTanzania, United Republic Of    0.11813
## country_of_originThailand                        0.01562 *
## country_of_originUganda                          0.11335
## country_of_originUnited States                   0.38857
## country_of_originUnited States (Hawaii)          0.75324
## country_of_originUnited States (Puerto Rico)     0.35475
## country_of_originVietnam                         0.11884
## country_of_originZambia                          0.99579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1433.04  on 1033  degrees of freedom
## Residual deviance:  560.72  on  995  degrees of freedom
## AIC: 638.72
```

```
##
## Number of Fisher Scoring iterations: 15
```

```
# Adding predict results from model_step to dataset
coffee_step <- coffee_w %>%
  mutate(logodds_good = predict(model_step),
         probs_good = fitted(model_step)) %>%
  mutate(odd_good = exp(logodds_good),
         class_pred = ifelse(probs_good>0.5,"Poor","Good"))

# Check accuracy for model_step
sum(coffee_step$class_pred == coffee_step$Qualityclass)/nrow(coffee_step)
```

```
## [1] 0.8955513
```

```
# Build a model with major terms with significant coefficients from model_step
model_1 <- glm(
  Qualityclass ~
               aroma +
               flavor +
               acidity,
  data = coffee_w,
  family = binomial(link="logit"))

# Pull summary of model_1
summ(model_1)
```

Observations	1034
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(3)$	787.22
Pseudo-R^2 (Cragg-Uhler)	0.71
Pseudo-R^2 (McFadden)	0.55
AIC	653.82
BIC	673.59

	Est.	S.E.	z val.	p
(Intercept)	110.02	7.33	15.01	0.00
aroma	-4.40	0.60	-7.38	0.00
flavor	-6.88	0.74	-9.28	0.00
acidity	-3.29	0.57	-5.76	0.00

Standard errors: MLE

```
# Adding predict results from model_1 to dataset
coffee_1 <- coffee_w %>%
  mutate(logodds_good = predict(model_1),
```

```
         probs_good = fitted(model_1)) %>%
  mutate(odd_good = exp(logodds_good),
         class_pred = ifelse(probs_good>0.5,"Poor","Good"))

# Check accuracy for model_1
sum(coffee_1$class_pred == coffee_1$Qualityclass)/nrow(coffee_1)
```

```
## [1] 0.8762089
```

model_1 has slightly higher AIC and less accuracy than model_step, but all coefficients are significant, thus model_1 is chosen as the final model.

```
# Plot point estimate and confidence intervals of model_step
plot_model(model_1, show.values = TRUE, title = "Log-Odds (Poor quality)")
```

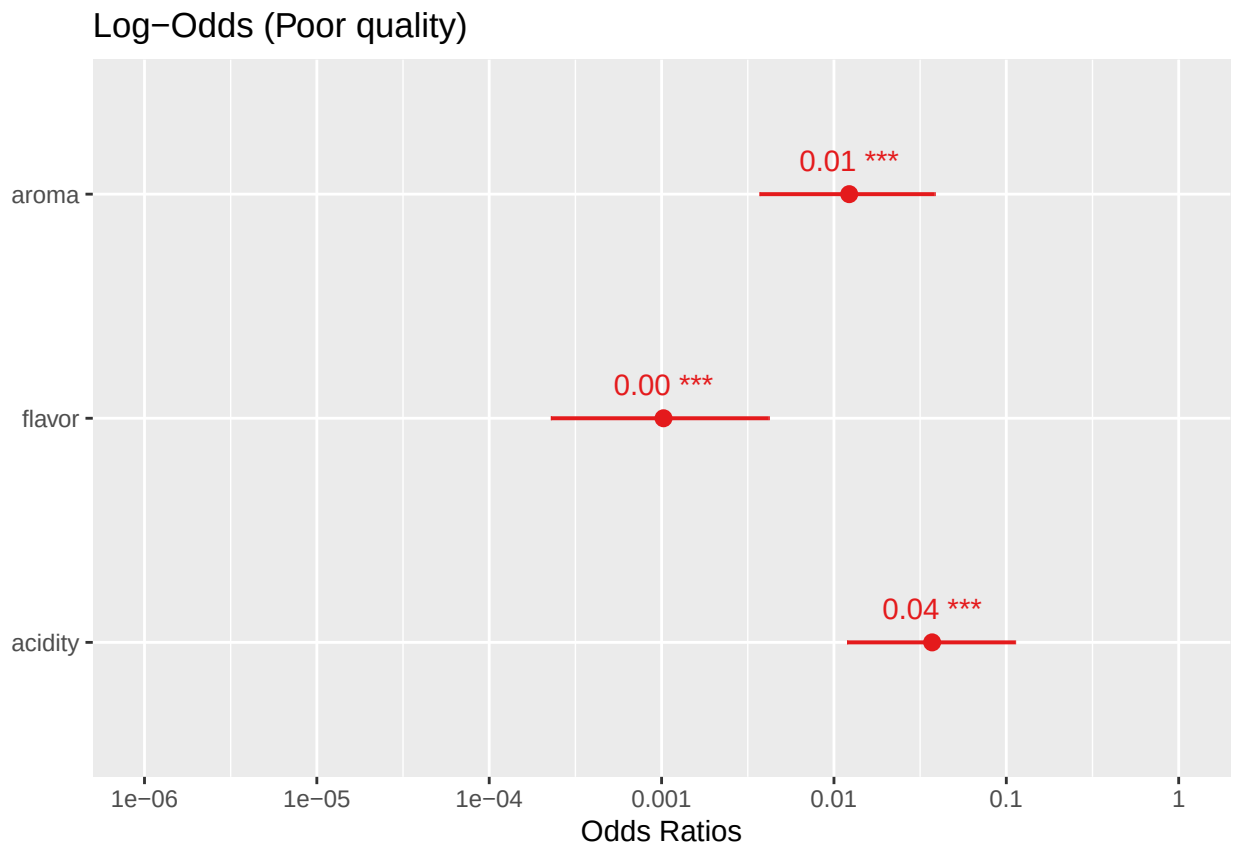

Figure 7: Point estimates and confidence intervals.

```
# Plot estimate possibilities by each explanatory variables in model_1
plot_model(model_1, type="pred", terms="aroma [all]",
           axis.title = "Probability",
           title = "Probability of being poor quality by aroma")
```

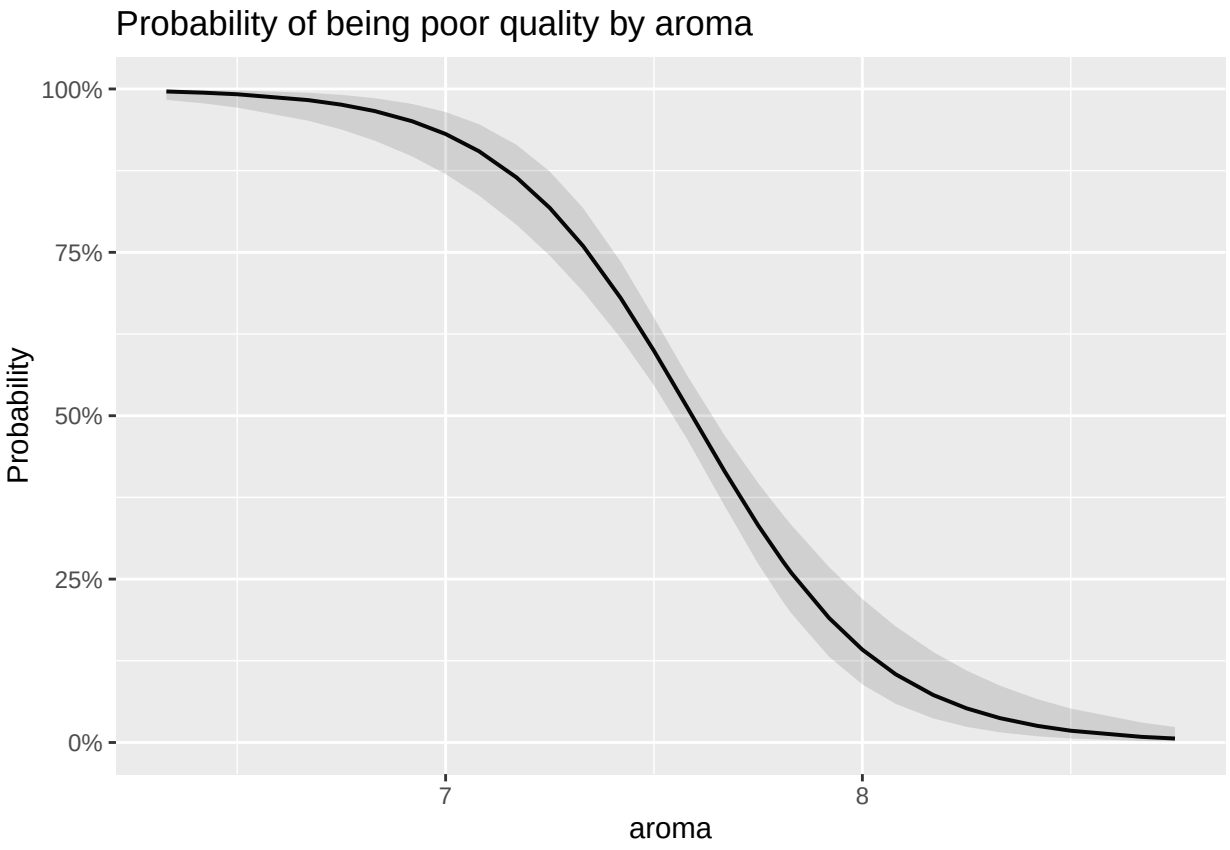

Figure 8: Probability of being poor quality by aroma.

```
# Plot estimate possibilities by each explanatory variables in model_1
plot_model(model_1, type="pred", terms="flavor [all]",
           axis.title = "Probability",
           title = "Probability of being poor quality by flavor")
```

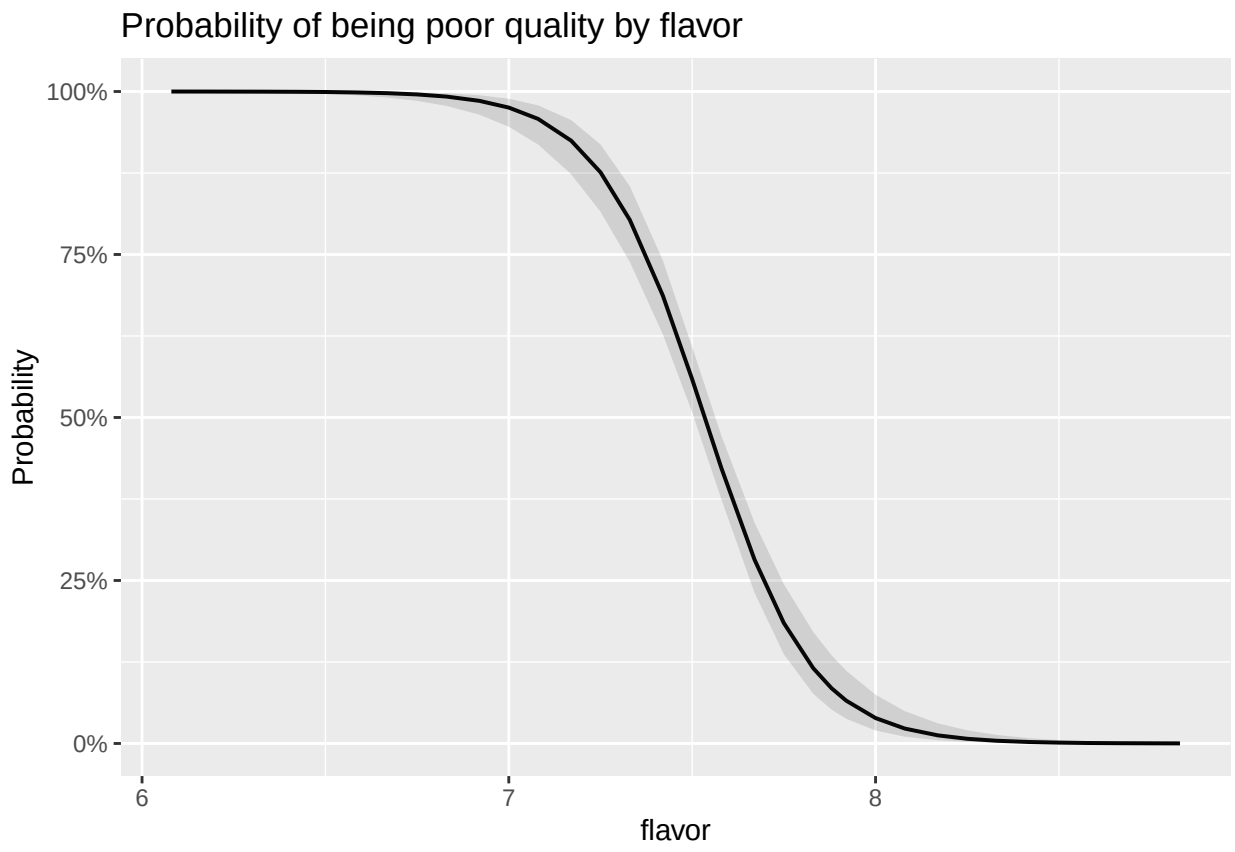

Figure 9: Probability of being poor quality by flavor.

```
# Plot estimate possibilities by each explanatory variables in model_1
plot_model(model_1, type="pred", terms="acidity [all]",
           axis.title = "Probability",
           title = "Probability of being poor quality by acidity")
```

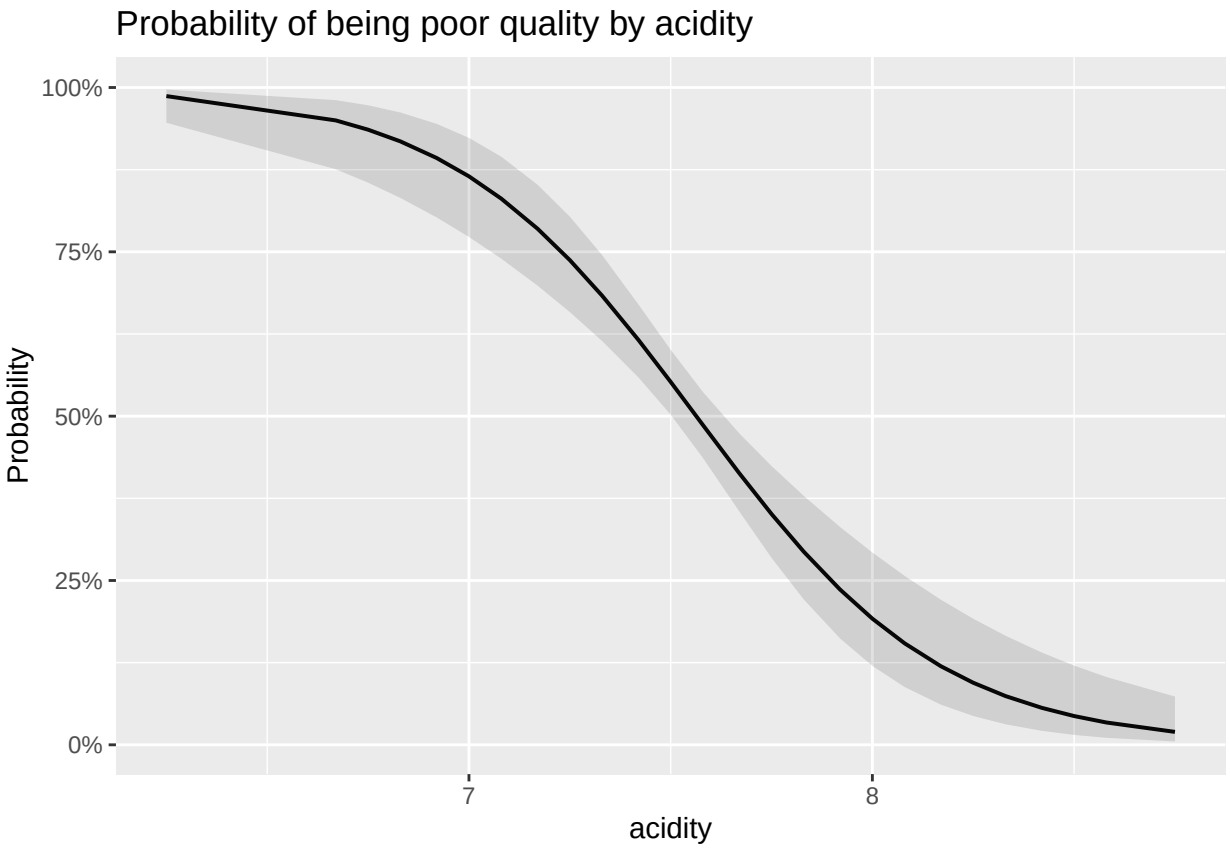

Figure 10: Probability of being poor quality by acidity