Group_11_Analysis

Group11

2023-03-09

Introduction

Based on the Coffee Quality Database (CQD), this report aims to investigate the features of coffee that influence its quality.

Step 1: Exploratory Data Analysis

Load the required packages and read the data.

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

coffee <- read.csv('https://raw.githubusercontent.com/rrachelxi/DAS-Group-11/main/dataset11.csv')
```

```
coffee <- coffee %>%
  # Harvested, country_of_origin and Qualityclass are turned into factor
  mutate(
    harvested = as.factor(harvested),
    country_of_origin = as.factor(country_of_origin),
    Qualityclass = as.factor(Qualityclass))
```

```r
# First skim of the dataset
skim_without_charts(coffee) %>%
  summary()
```

Table 1: Data summary

Name	coffee
Number of rows	1094
Number of columns	8
Column type frequency:	
factor	3
numeric	5
Group variables	None

```r
skim_without_charts(coffee) %>%
  yank("numeric") %>%
  kable(caption = '\\label{tab:c1} The numeric variable of the coffee data.', digits = 2)
```

Table 2: The numeric variable of the coffee data.

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

```r
skim_without_charts(coffee) %>%
  yank("factor") %>%
  kable(caption = '\\label{tab:c2} The vector variable of the coffee data.', digits = 2)
```

Table 3: The vector variable of the coffee data.

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
country_of_origin	1	1.00	FALSE	35	Mex: 191, Gua: 155, Col: 144, Bra: 111
harvested	56	0.95	FALSE	9	201: 283, 201: 215, 201: 148, 201: 130
Qualityclass	0	1.00	FALSE	2	Goo: 558, Poo: 536

From Table 1, there are 191 missing values in altitude_mean_meters, 56 missing values in harvested and 1 missing value in country_of_origin. Since altitude_mean_meters is a continuous variable and the amount of missing values is quite large, these missing values are replaced by the mean value. As for the harvested and country_of_origin, these missing values are deleted because of the small volumes.

```r
# The mean value for altitude_mean_meters is calculated
mean_altitude <- mean(coffee$altitude_mean_meters,na.rm = TRUE)

coffee_m <- coffee %>%
  # Missing values are replaced in altitude_mean_meters by its mean value
  mutate(
    altitude_mean_meters = replace_na(altitude_mean_meters, mean_altitude)) %>%

  # Missing values are dropped in country_of_origin and harvested
  filter(
    !is.na(country_of_origin) &
    !is.na(harvested))
```

```r
# data is skimmed after dropping missing values
skim_without_charts(coffee_m) %>%
  summary()
```

Table 4: Data summary

Name	coffee_m
Number of rows	1038
Number of columns	8
Column type frequency:	
factor	3
numeric	5
Group variables	None

```r
skim_without_charts(coffee_m) %>%
  yank("numeric") %>%
  kable(caption = '\\label{tab:m1} The numeric variable of the coffee_m data.', digits = 2)
```

Table 5: The numeric variable of the coffee_m data.

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

```
skim_without_charts(coffee_m) %>%
  yank("factor") %>%
  kable(caption = '\\label{tab:m2} The vector variable of the coffee_m data.', digits = 2)
```

Table 6: The vector variable of the coffee_m data.

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

The dataset is now with no missing value. Next, a pair plot is drawn to visualize data.

```
# A pair plot is drawn
ggpairs(coffee_m[,c(-1,-7)],
        aes(color=Qualityclass,alpha=0.2),
        title = "Distribution between variables")
```

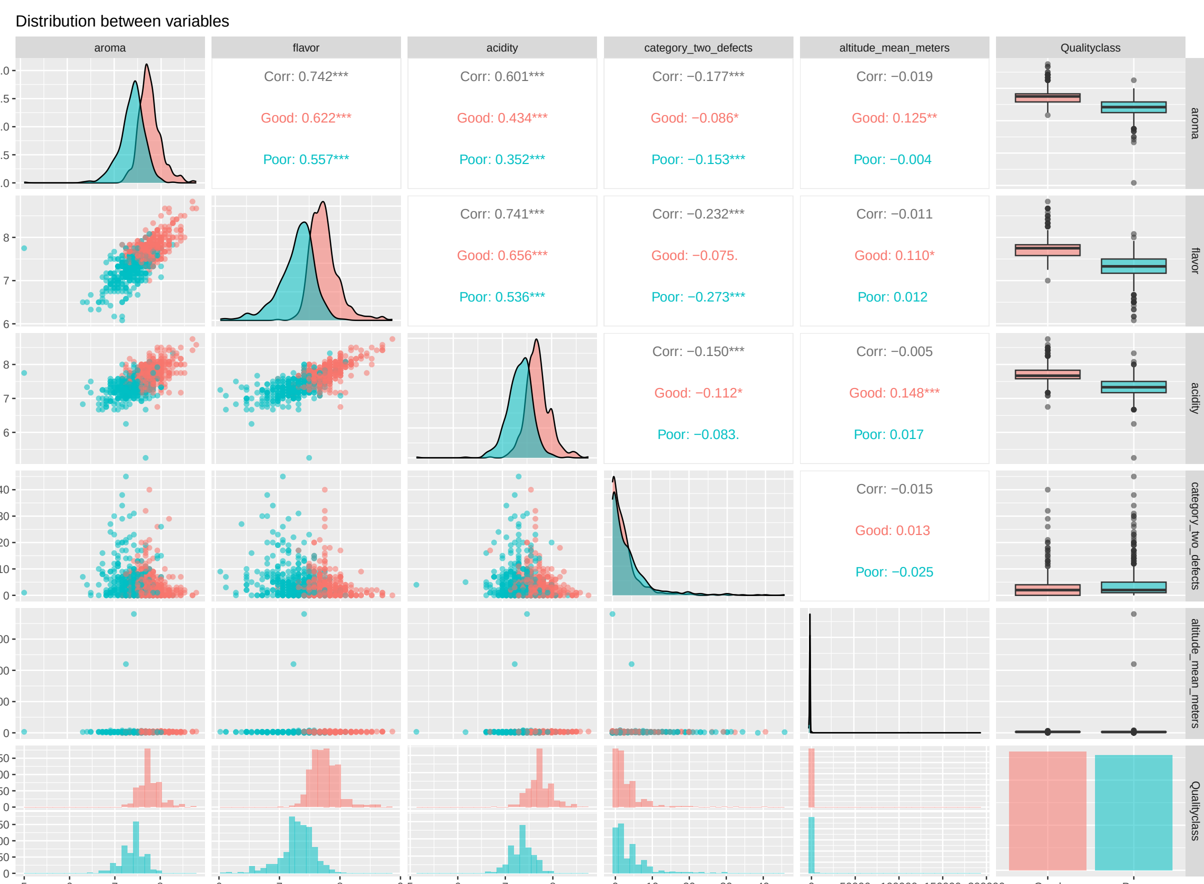

Figure 1: Pair plot of numeric variables and Qualityclass.

As several possible outliers are presented in figure 1, scatter plots are drawn to show outliers more clearly.

```r
# A scatter plot for aroma and acidity is drawn
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")
```

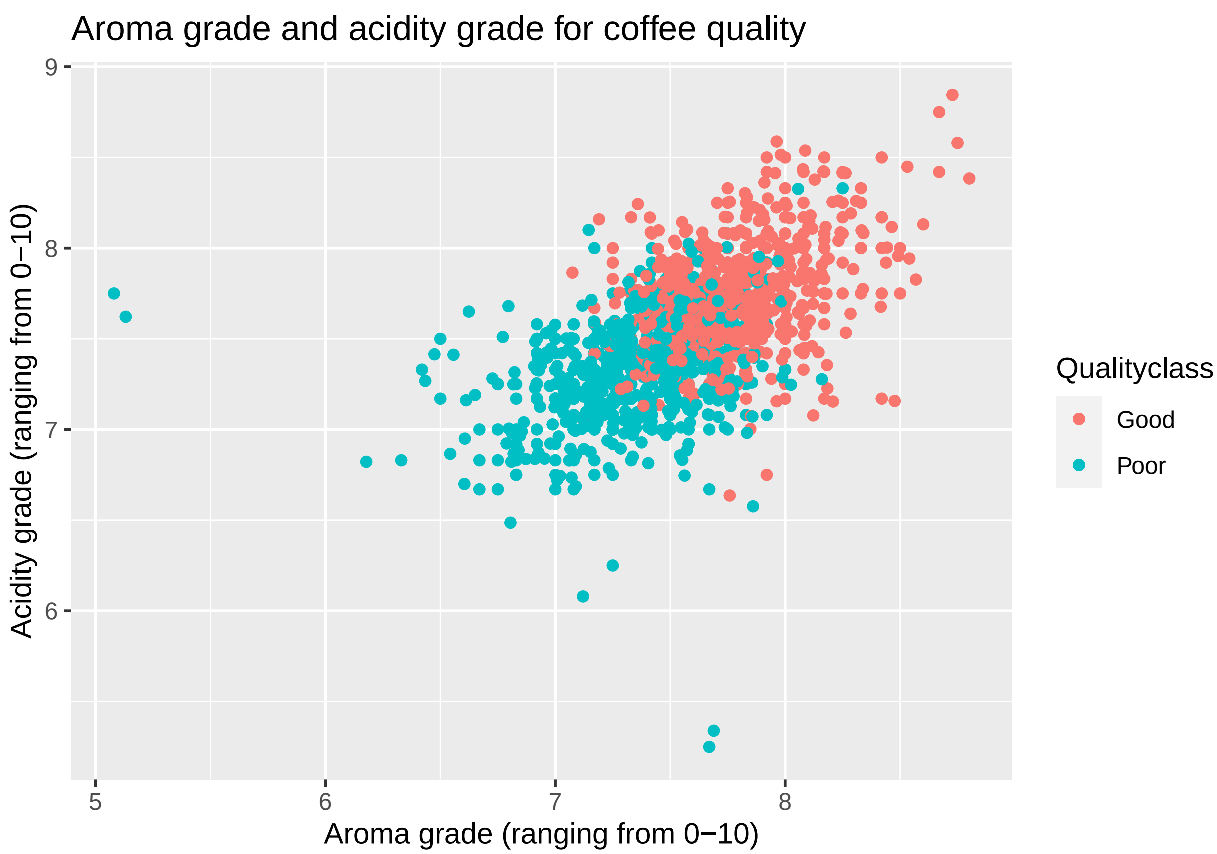

Figure 2: Scatter plot of aroma and acidity.

```r
# A scatter plot for altitude_mean_meters is drawn
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')
```

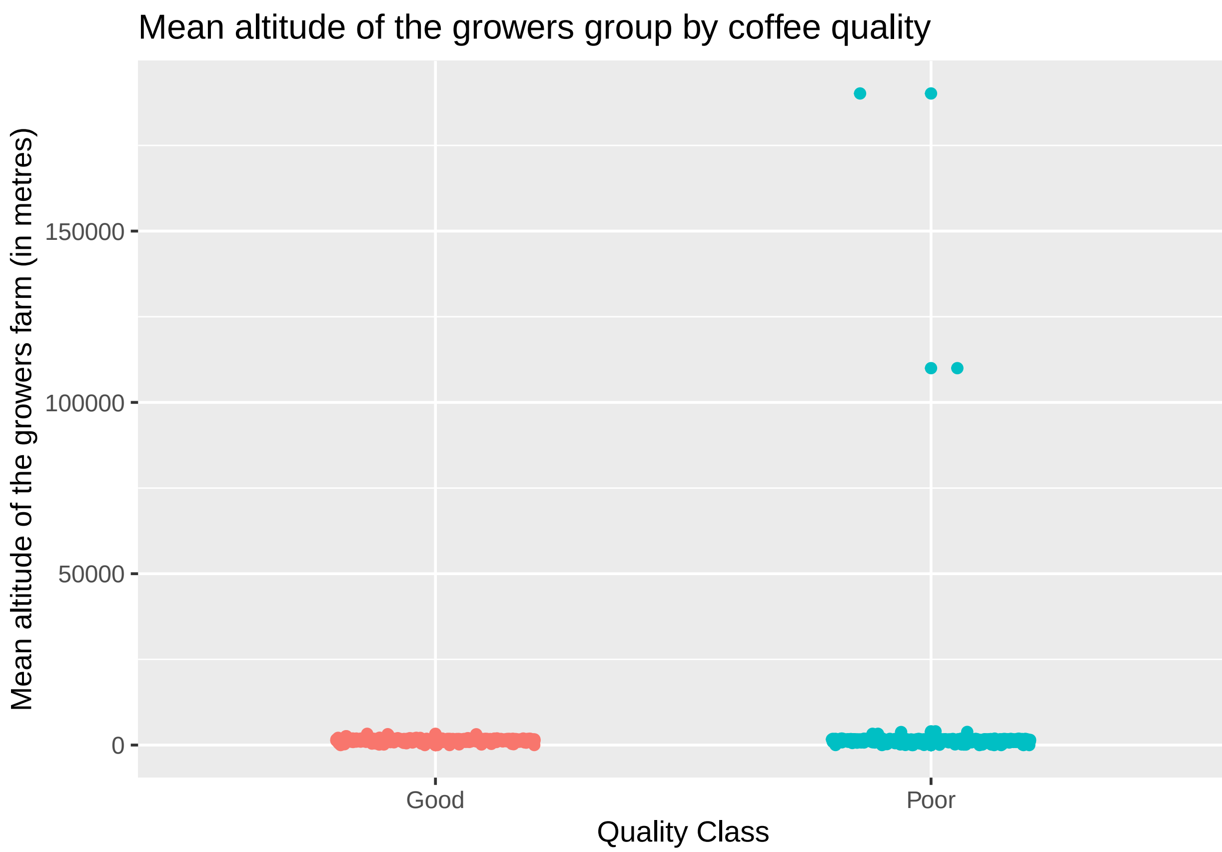

Figure 3: Scatter plot of altitude_mean_meters.

It can be seen from Figure 1, 2 and 3 that there are many outilers in aroma, acidity and altitude_mean_meters. At the mean time, it can also be noticed that good-quality coffee tend to have a higher level of acidity and aroma, comparing with the poor-quality coffee.

```r
# Outlies in aroma, acidity and altitude_mean_meters are dropped from the dataset
coffee_w <- coffee_m %>%
  filter(
    aroma > 6 &
    acidity > 6 &
    altitude_mean_meters < 100000)
```

```r
# The dataset is skimmed after dropping outliers
skim_without_charts(coffee_w) %>%
  summary()
```

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

```r
skim_without_charts(coffee_w) %>%
  yank("numeric") %>%
  kable(caption = '\\label{tab:w1} The numeric variable of the coffee_w data.', digits = 2)
```

Table 8: The numeric variable of the coffee_w data.

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
skim_without_charts(coffee_w) %>%
  yank("factor") %>%
  kable(caption = '\\label{tab:w2} The vector variable of the coffee_w data.', digits = 2)
```

Table 9: The vector variable of the coffee_w data.

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

```r
# a pair plot without outliers is drawn to visualize data
ggpairs(coffee_w[c(-1,-7)],
        aes(color=Qualityclass,alpha=0.2),
        title = "Distribution between variables(without outliers)")
```

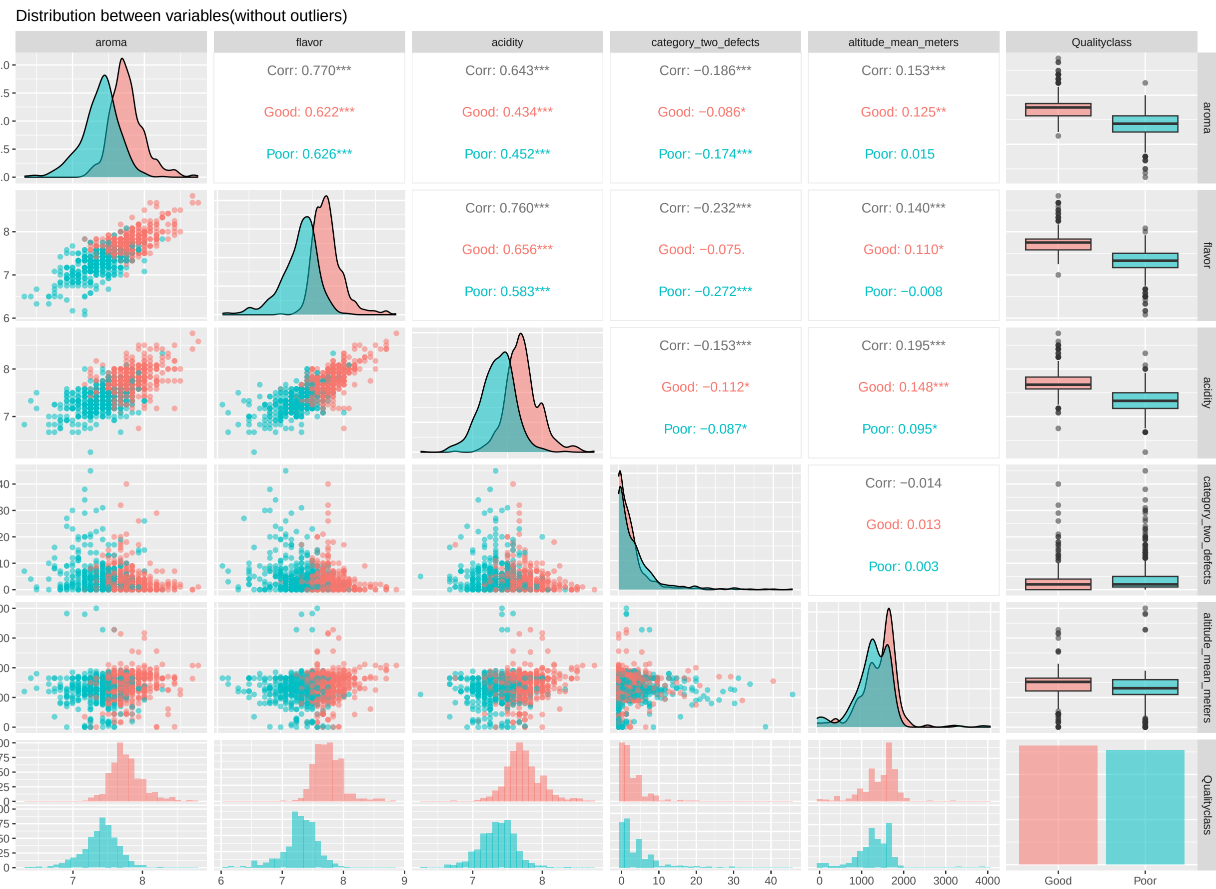

Figure 4: Pair plot of numeric variables and Qualityclass (without outliers).

From Figure 4, there is a relatively strong correlation between aroma and flavor (correlation: 0.770), similar to acidity and aroma (correlation: 0.643). And box plots of aroma, flavor and acidity show significant differences between classes of quality.

```r
# A bar plot 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 = "Proportions of quality class by year the batch was harvested")
```

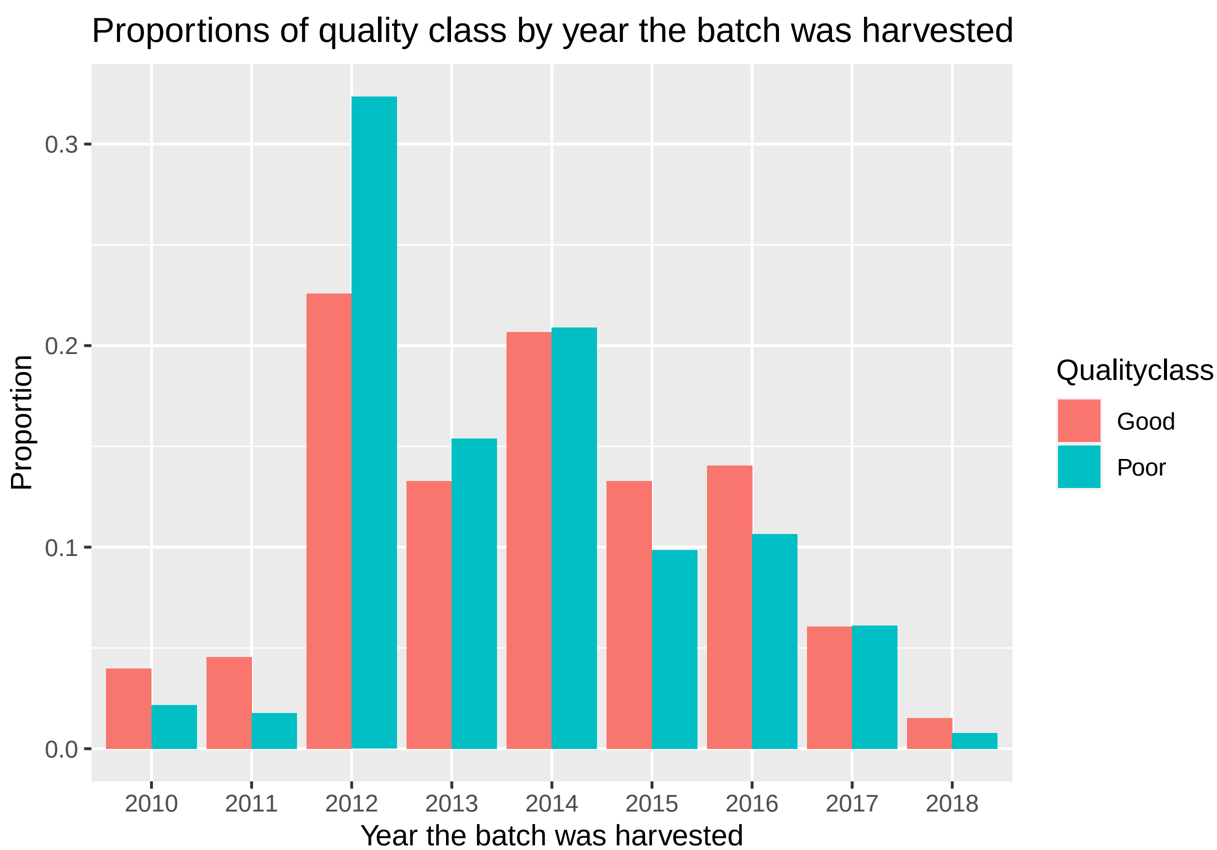

Figure 5: Bar plot of quality class by harvested year.

From Figure 5, the batch of the year 2012 has the highest proportion of good-quality coffee, with the batch of the year 2014 following behind. And the situation is the same for poor-quality coffee.

```
# A bar plot 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 = "Proportions of quality class by country of origin") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

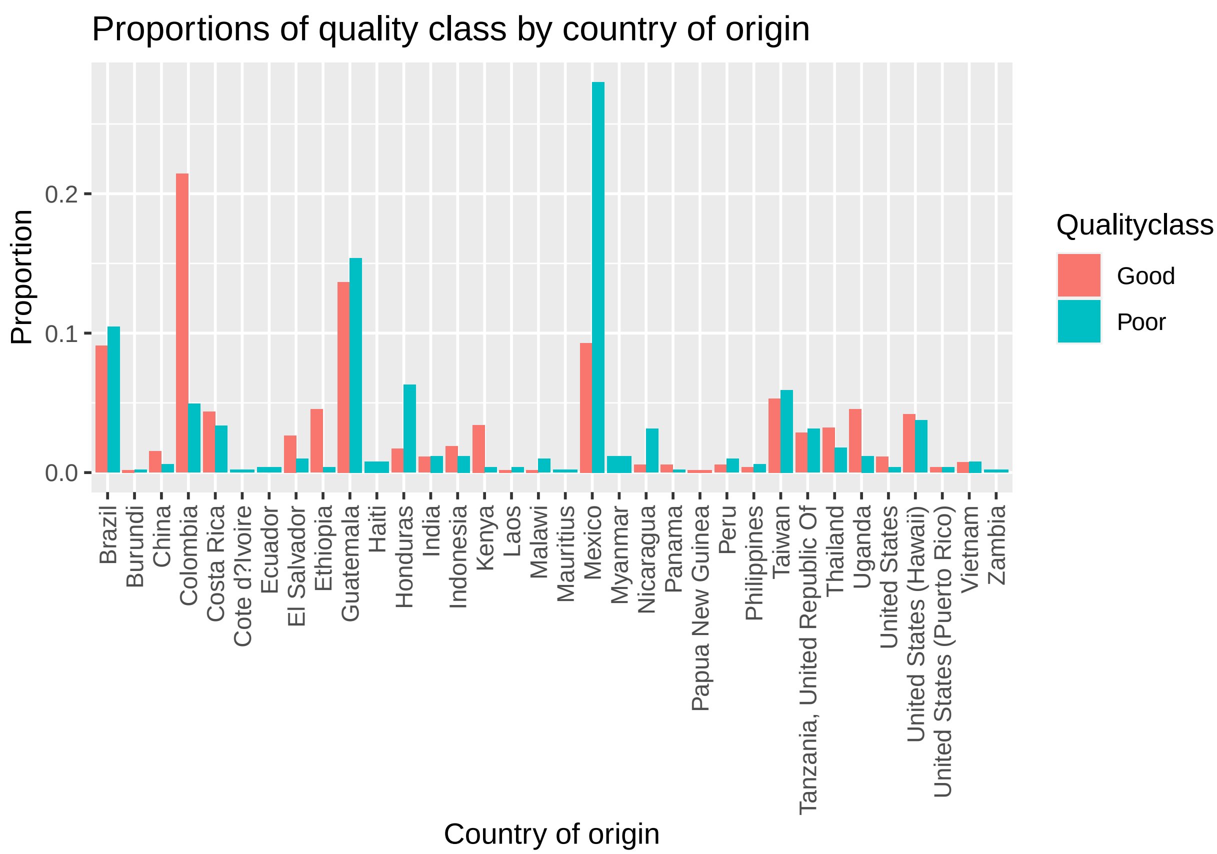

Figure 6: Barplot of quality class by country of origin.

From Figure 6, Mexico has the highest proportion of poor-quality coffee while Colombia has the highest proportion of good-quality coffee.

Step 2: Statistical Modelling

A full general linear model is built and feature selection is conducted stepwise by AIC.

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

# Stepwise 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
```

Based on the AIC table, the model with variable of aroma, flavor, acidity, category_two_defects, altitude_mean_meters and country_of_region would be chosen (without harvested).

```r
# A summary of the 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
```

```r
# Predict results from model_step are added to the dataset
coffee_step <- coffee_w %>%
  mutate(logodds_poor = predict(model_step),
         probs_poor = fitted(model_step)) %>%
  mutate(odd_poor = exp(logodds_poor),
         class_pred = ifelse(probs_poor>0.5,"Poor","Good"))

# Accuracy for model_step is calculated
sum(coffee_step$class_pred == coffee_step$Qualityclass)/nrow(coffee_step)
```

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

The p values of aroma, flavor and acidity are smaller than 0.05, which indicates these variables have a significant relationship with the quality class of coffee. Then, a new model called model_1 with these significant coefficients is built.

```r
# Model with significant coefficients
model_1 <- glm(
  Qualityclass ~
               aroma +
               flavor +
               acidity,
  data = coffee_w,
  family = binomial(link="logit"))
```

```r
# Summary model_1
summary(model_1)
```

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity, family = binomial(link = "logit"),
##     data = coffee_w)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1342  -0.4739  -0.0044   0.3752   3.9002
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 110.0224     7.3317  15.006  < 2e-16 ***
## aroma        -4.4010     0.5963  -7.380 1.58e-13 ***
## flavor       -6.8819     0.7420  -9.275  < 2e-16 ***
## acidity      -3.2934     0.5717  -5.761 8.38e-09 ***
## ---
## 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:  645.82  on 1030  degrees of freedom
## AIC: 653.82
##
## Number of Fisher Scoring iterations: 7
```

```r
# Predict results from model_1 are added to the dataset
coffee_1 <- coffee_w %>%
  mutate(logodds_poor = predict(model_1),
         probs_poor = fitted(model_1)) %>%
  mutate(odd_poor = exp(logodds_poor),
         class_pred = ifelse(probs_poor>0.5,"Poor","Good"))

# Accuracy for model_1 is calculated
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 in model_1 are significant, thus model_1 is chosen as the final model.

```r
# Plot point estimates and confidence intervals of model_1
plot_model(model_1, show.values = TRUE,
           title = "Odds (Poor quality)",
           show.p = FALSE, digits=3)
```

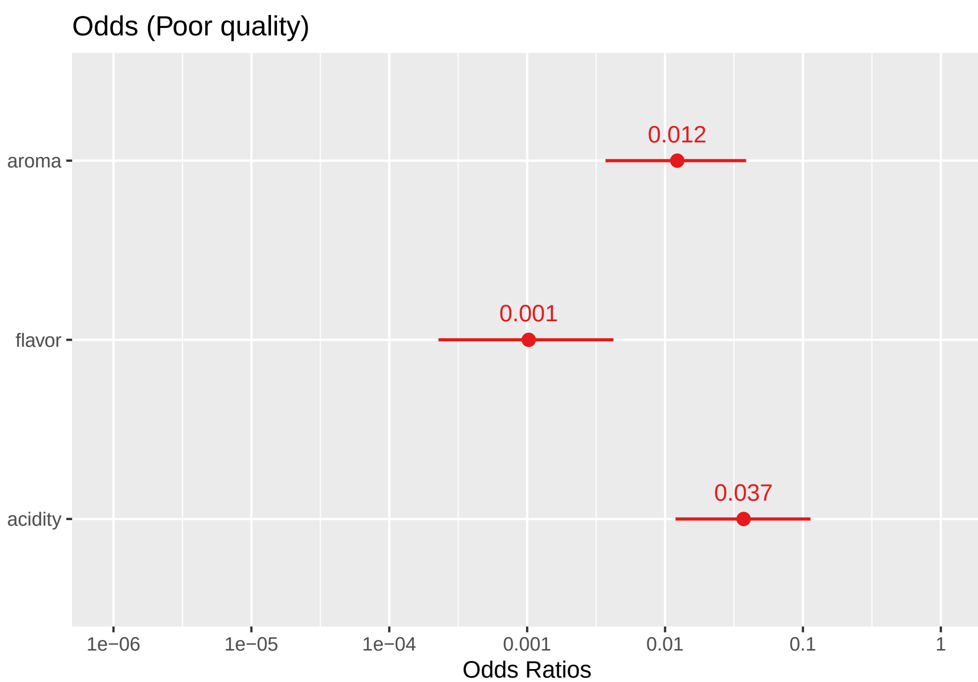

Figure 7: Point estimates and confidence intervals for odds.

From Figure 7, it can be seen that for every 0.1 unit increase in aroma, the probability of being poor coffee becomes 0.012 times that of before. For every 0.1 increase in flavor, the probability of being poor coffee becomes 0.001 times that of before and for every 0.1 unit increase in acidity, the probability of being poor coffee quality becomes 0.037 times that of before.

```r
# Possibilities of being poor-quality by aroma in model_1 is visualized
plot_model(model_1, type="pred", terms="aroma [all]",
           axis.title = "Probability",
           title = "Probability of being poor quality by aroma")
```

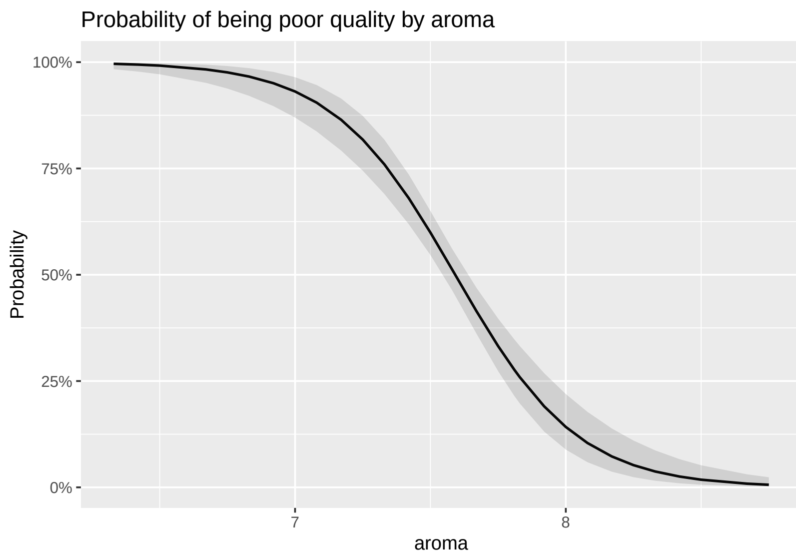

Figure 8: Probability of being poor quality by aroma.

```r
# Possibilities of being poor-quality by flavor in model_1 is visualized
plot_model(model_1, type="pred", terms="flavor [all]",
           axis.title = "Probability",
           title = "Probability of being poor quality by flavor")
```

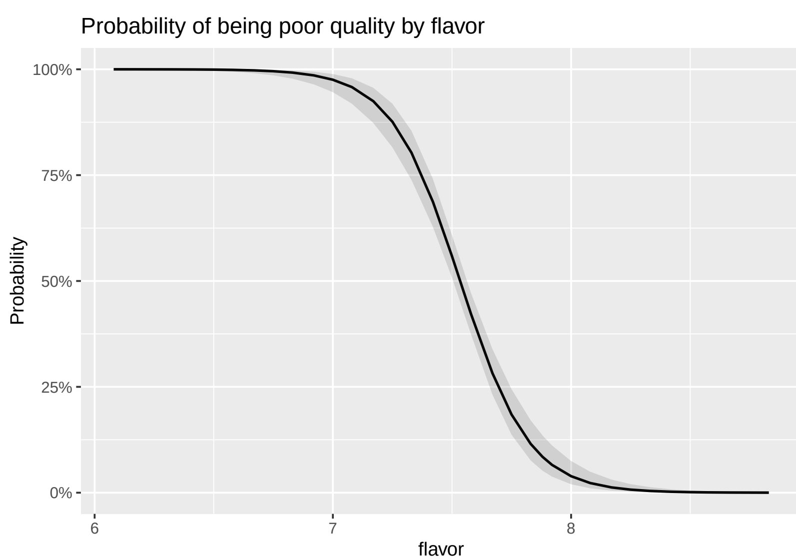

Figure 9: Probability of being poor quality by flavor.

```r
# Possibilities of being poor quality by acidity in model_1 is visualized
plot_model(model_1, type="pred", terms="acidity [all]",
           axis.title = "Probability",
           title = "Probability of being poor quality by acidity")
```

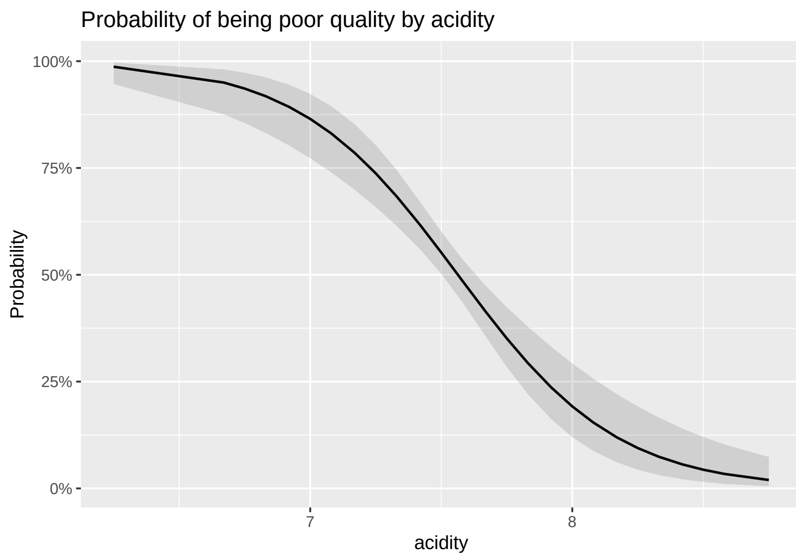

Figure 10: Probability of being poor quality by acidity

Step 3: Summary and further task

Summary: Based on these plots above, high values of aroma, flavor and acidity all indicate a low probability of being poor coffee quality. In particular, the flavor has the most significant influence on the quality of coffee.

Further work:

1. DARA SEPARATION
   Since the dataset did not be separated into the train, valid and test sets, the model may be overfitting. This problem can be alleviated by data splitting.

2. HIGH CORRELATION
   It can be noticed that there are high correlations between flavor and aroma, acidity and aroma, and acidity and flavor, which may lead to collinearity. Variance Inflation Factor(VIF) can be used to measure multicollinearity further. If it exists, transformation can be considered to make variables less correlated but still maintain their feature. Another method to solve this problem is Principal Component Analysis, which will reduce the dimensions of data by decomposing data into several independent factors.