Analysis of the Factors Affect the Quality Class Being Classified as Poor for Coffee Batch Using GLM

Group 13

```
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(gt)
library(MASS)
library(knitr)
library(broom)
library(GGally)
library(ggplot2)
library(dplyr)
library(broom)
```

1 Introduction

Given the coffee dataset with the features of coffee and overall quality scores from the Coffee Quality Institute, aiming to provide key information to the local coffee farmers and help them to improve coffee production with higher quality.

Research Question:

What factors will influence the coffee batch to be classified as good or poor quality, and how do these factors affect the quality score.

2 Exploratory Data Analysis

2.1 Description of Data

2.1.1 Data Source

The dataset utilized in this study originates from the Coffee Quality Institute, containing detailed attributes related to coffee quality. The primary objective of this dataset is to analyze the key factors influencing coffee quality scores and to provide insights that may assist local coffee producers in enhancing their production standards.

```
# Read dataset
coffee <- read.csv("dataset13.csv")
dim(coffee)
```

```
[1] 1145    8
```

The dataset is first loaded and examined to assess its structure and completeness, which comprises multiple variables that capture key characteristics of coffee, including sensory attributes (e.g., aroma, flavor, acidity), environmental factors (e.g., altitude, harvest year), and quality classification (good or poor).

2.1.2 Data Preprocessing

2.1.2.1 Remove missing value and transformation of Categorical Variables

```
coffee1 <- na.omit(coffee)
# Check the type of the binary response variable and change to factor type.
coffee1$Qualityclass <- as.factor(coffee1$Qualityclass)
```

Since the response variable Qualityclass represents a binary classification (good vs. poor), it is converted into a factor variable to facilitate appropriate statistical modeling. This conversion is particularly important for fitting a Generalized Linear Model (GLM), as it allows categorical outcomes to be appropriately modeled.

2.1.2.2 Detection and Removal of Outliers

```r
# Remove the outliers
numeric_cols <- sapply(coffee1, is.numeric)
keep_rows <- rep(TRUE, nrow(coffee1))

for (col in names(coffee1)[numeric_cols]) {
  Q1 <- quantile(coffee1[[col]], 0.15, na.rm = TRUE)
  Q3 <- quantile(coffee1[[col]], 0.85, na.rm = TRUE)

  IQR_value <- Q3 - Q1

  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value

  keep_rows <- keep_rows & (coffee1[[col]] >= lower_bound & coffee1[[col]] <= upper_bound)
}

coffee1 <- coffee1[keep_rows, ]
```

Extreme values in numerical variables may distort statistical inference. Therefore, outliers are detected and removed using the Interquartile Range (IQR) method.

2.1.2.3 Examination of Variable Distribution

```r
# Check type of harvested year.
length(unique(coffee1$country_of_origin))
```

```
[1] 33
```

```r
length(unique(coffee1$harvested))
```

```
[1] 9
```

```r
class(coffee1$harvested)
```

```
[1] "integer"
```

A preliminary assessment of the country of origin and harvest year variables is conducted to understand their diversity and distribution.

The results indicate: 33 unique countries of origin; 9 distinct harvest years.

The harvested year variable is stored as an integer.

2.1.3 Scaling the Variable

```
coffee1$altitude_mean_meters <- scale(coffee1$altitude_mean_meters)
```

Since the value of the variable, mean altitude is extremely larger than other continuous explanatory variables, this might cause the model instability and biased coefficient interpretation, which making it harder to interpret and compare with other variables. Therefore, the standardization or scaling of the variable is required.

2.1.4 Variable Definitions

```
coffee1$altitude_mean_meters <- as.numeric(scale(coffee1$altitude_mean_meters))

variable <- data.frame(
  variablename = c("Qualityclass", "Aroma", "Flavor", "Acidity",
                   "Category Two Defects", "Altitude Mean Meters",
                   "Harvested", "Country of Origin"),
  Description = c("Coffee quality classification (Good/Poor)",
                  "Sensory evaluation of coffee aroma",
                  "Sensory evaluation of coffee flavor",
                  "Sensory evaluation of coffee acidity",
                  "Number of defects affecting coffee quality",
                  "Mean altitude at which coffee is cultivated",
                  "Year in which the coffee was harvested",
                  "Country where the coffee was grown"),
  Type = c(sapply(coffee1[c("Qualityclass", "aroma", "flavor", "acidity",
                            "category_two_defects", "altitude_mean_meters",
                            "harvested", "country_of_origin")], class)),
  Unit = c("-", "Score (1-10)", "Score (1-10)", "Score (1-10)",
           "Count", "Meters", "Year", "-")
)

variable %>%
  gt() %>%
  tab_header(
    title = "Description of Variables"
  ) %>%
  tab_options(
    table.width = pct(100)
  )
```

Description of Variables

variablename	Description	Type	Unit
Qualityclass	Coffee quality classification (Good/Poor)	factor	-
Aroma	Sensory evaluation of coffee aroma	numeric	Score (1-10)
Flavor	Sensory evaluation of coffee flavor	numeric	Score (1-10)
Acidity	Sensory evaluation of coffee acidity	numeric	Score (1-10)
Category Two Defects	Number of defects affecting coffee quality	integer	Count
Altitude Mean Meters	Mean altitude at which coffee is cultivated	numeric	Meters
Harvested	Year in which the coffee was harvested	integer	Year
Country of Origin	Country where the coffee was grown	character	-

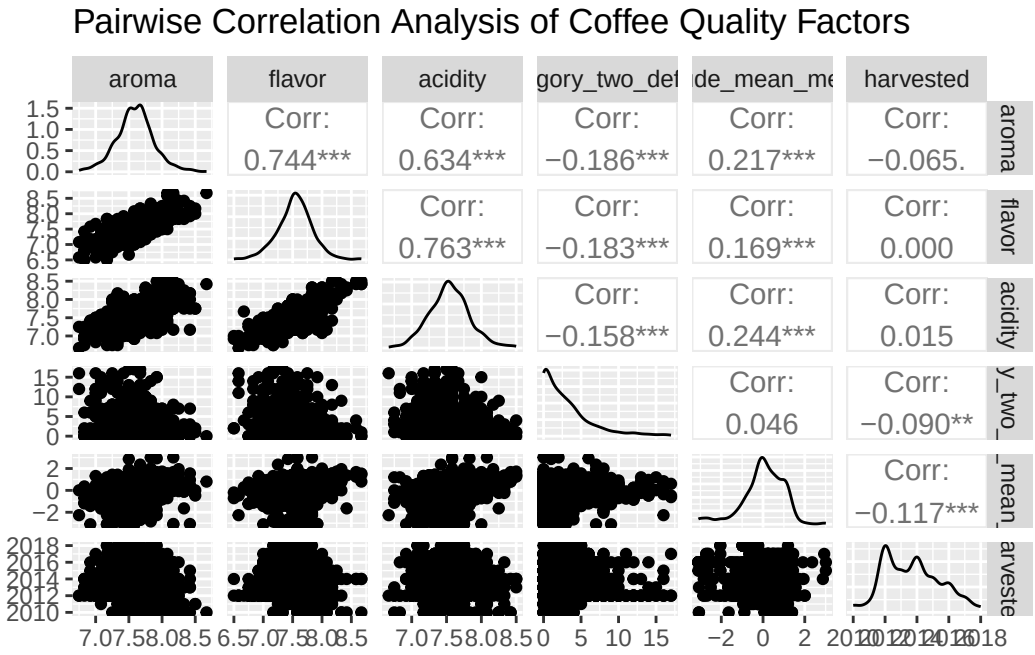
This table provides a structured overview of the dataset, ensuring clarity in variable interpretation.

2.1.5 Correlation Analysis

Prior to model fitting, a pairwised correlation plot is computed to assess potential relationships between continuous explanatory variables.

```
coffee_numeric <- coffee1 %>%
    dplyr::select(aroma, flavor, acidity,category_two_defects,
                  altitude_mean_meters,harvested) %>%
    mutate(across(where(is.factor), as.numeric))

ggpairs(coffee_numeric,
        title = "Pairwise Correlation Analysis of Coffee Quality Factors")+
  theme(plot.background = element_rect(
    fill="transparent",
    colour = NA,
    size=1
  ))
```



Notice that there were high correlations found between the pairs aroma and flavor(0.83); aroma and acidity(0.75); flavor and acidity(0.84), suggesting that possible colinearity occurred.

2.2 Data Visualization

Boxplots of Explanatory Variables by Coffee Quality Class

```
ggplot(data = coffee1, aes(x=Qualityclass,
                           y = aroma,
                           fill=Qualityclass))+
  geom_boxplot()+
  labs(x = "Qualityclass", y = "Aroma")+
  theme(legend.position = "none")
```

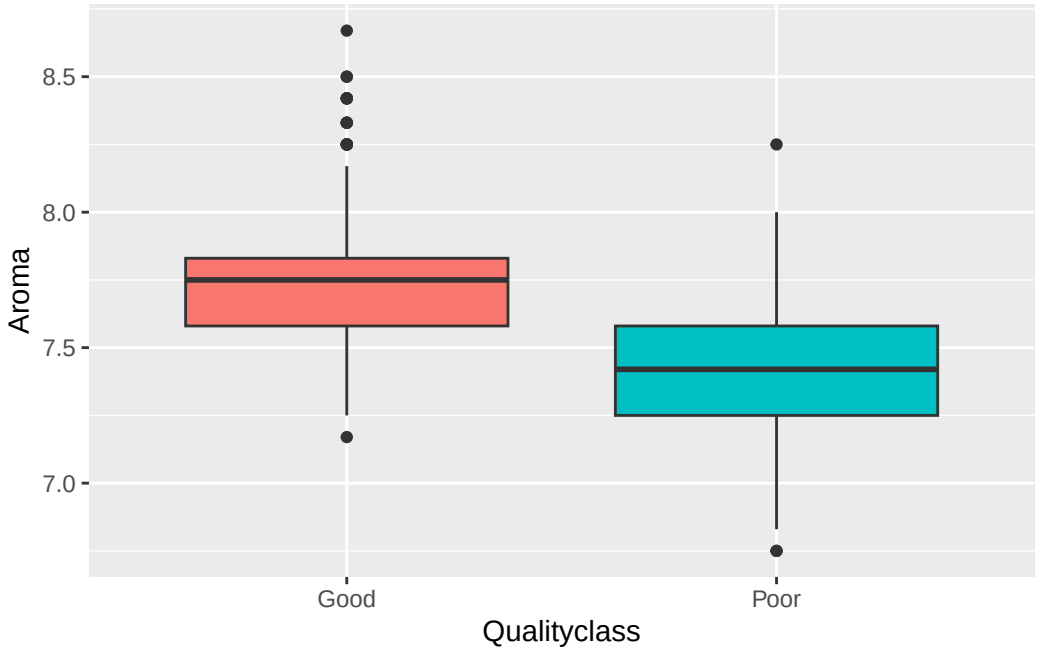


Figure 1: Aroma Grade by Coffee Quality Class

Figure 1 shows a boxplot of Aroma grade for coffee batch being classified as Good and Poor quality. The median aroma score is higher for the Good quality class, indicating better aroma ratings. Both classes have similar variability (IQR), but the Poor quality class shows a slightly wider overall spread.

```
ggplot(data = coffee1, aes(x=Qualityclass,
                           y = flavor,
                           fill=Qualityclass))+
  geom_boxplot()+
  labs(x = "Qualityclass", y = "Flavor")+
  theme(legend.position = "none")
```

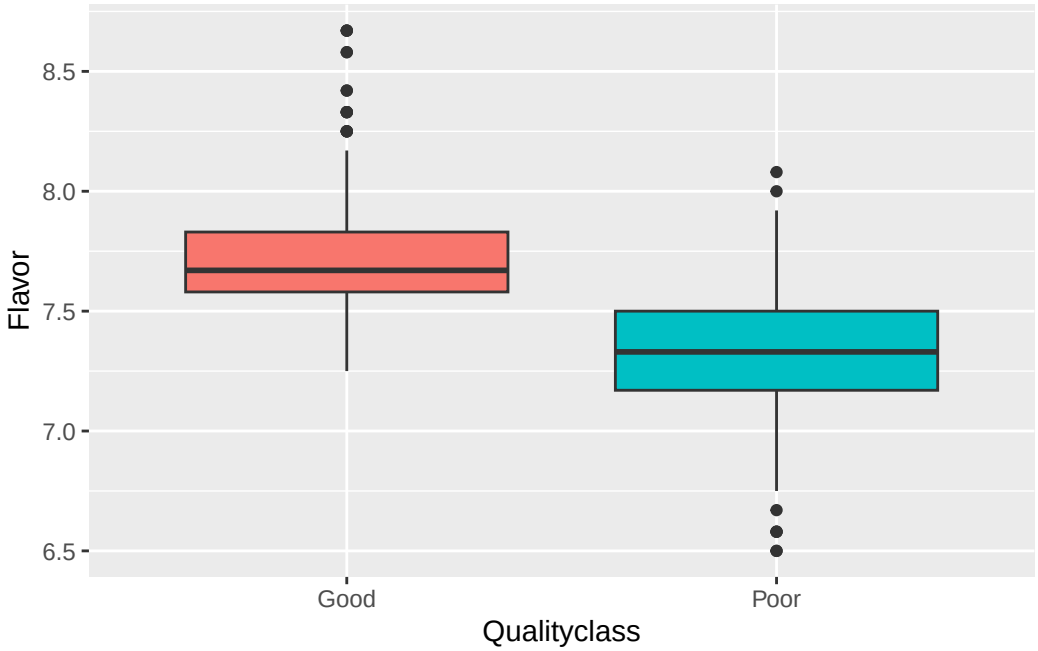


Figure 2: Flavor Grade by Coffee Quality Class

Figure 2 shows a boxplot of Flavor scores for coffee batch being classified as Good and Poor quality. The median flavor score for the Good quality class is noticeably higher than that of the Poor class. The IQR for Good quality coffee is slightly smaller than that of Poor quality coffee, but the Poor quality class has a wider spread in flavor grades.

```
ggplot(data = coffee1, aes(x=Qualityclass,
                           y = acidity,
                           fill=Qualityclass))+
  geom_boxplot()+
  labs(x = "Qualityclass", y = "Acidity")+
  theme(legend.position = "none")
```

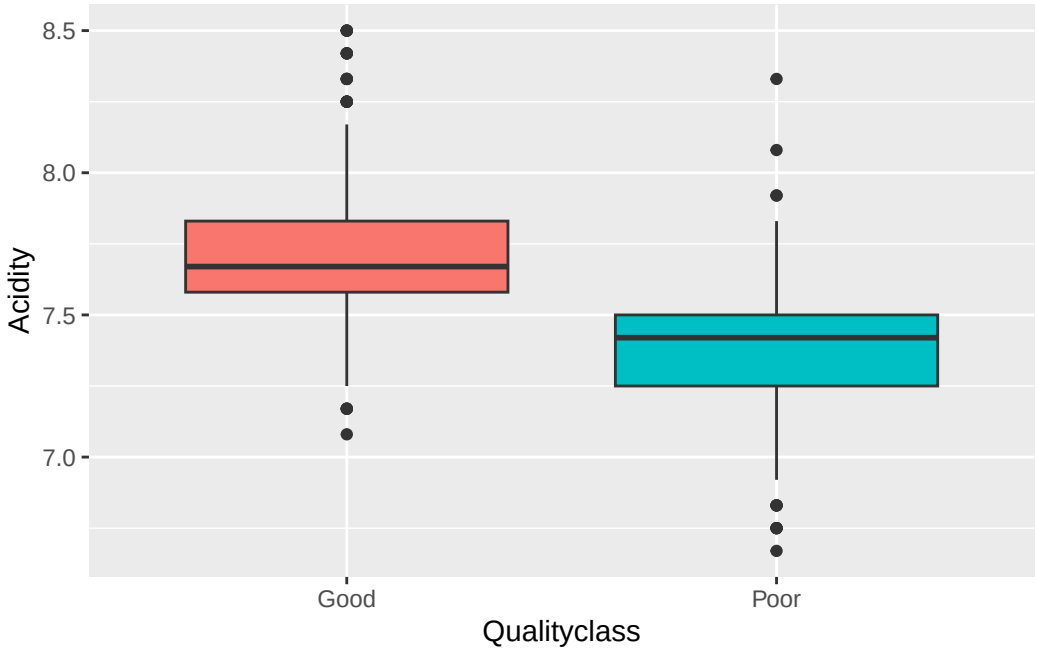


Figure 3: Acidity Grade by Coffee Quality Class

Figure 3 shows a boxplot of Acidity grades for coffee batch being classified as Good and Poor quality. The median acidity score for the Good quality class is higher than that of the Poor quality class, and the IQRs for two quality classes appears similar. However, Good quality coffee seems to have a slightly more compact distribution. Both classes have several outliers, but Good quality class has more outliers above 8.0, while Poor quality coffee has more outliers below 7.0.

```
ggplot(data = coffee1, aes(x=Qualityclass,
                           y = category_two_defects,
                           fill=Qualityclass))+
  geom_boxplot()+
  labs(x = "Qualityclass", y = "Category_two_defects")+
  theme(legend.position = "none")
```

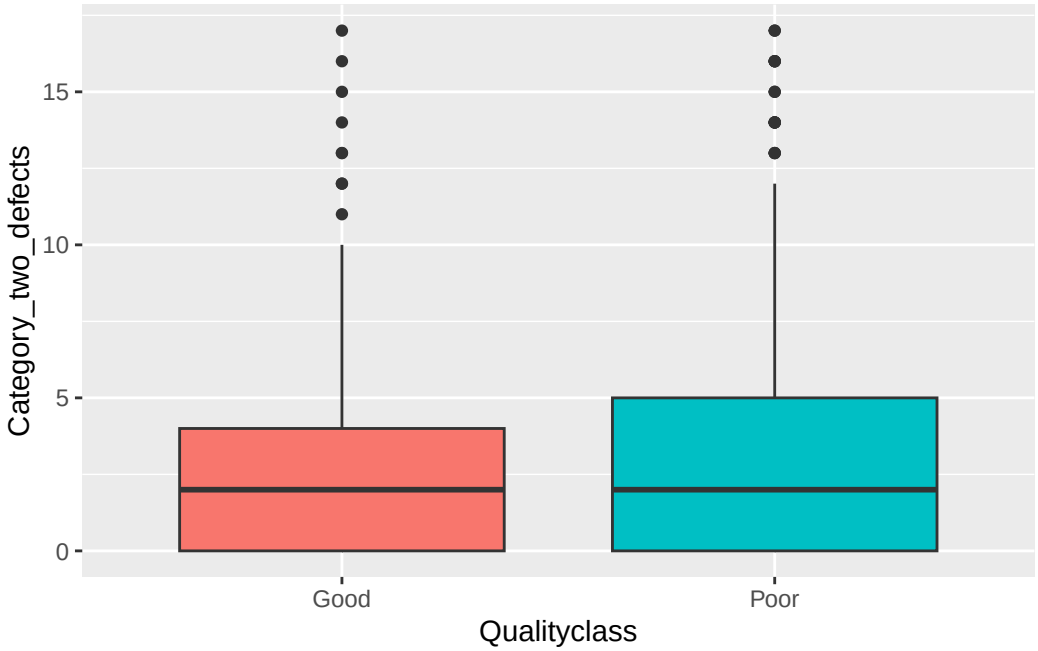


Figure 4: Count of two Category defects by Coffee Quality Class

Figure 4 is the boxplot of Category of two defects for coffee batch being classified as Good and Poor quality. The median number of defects are almost the same for both Poor and Good quality class. Both quality classes have a similar IQR, but a slightly wider spread is found for Poor quality class. Notice two quality classes contain large outliers, but Poor quality class has more extreme values, with defect counts exceeding 15.

```
ggplot(data = coffee1, aes(x=Qualityclass,
                           y = altitude_mean_meters,
                           fill=Qualityclass))+
  geom_boxplot()+
  labs(x = "Qualityclass", y = "Altitude (by meters)")+
  theme(legend.position = "none")
```

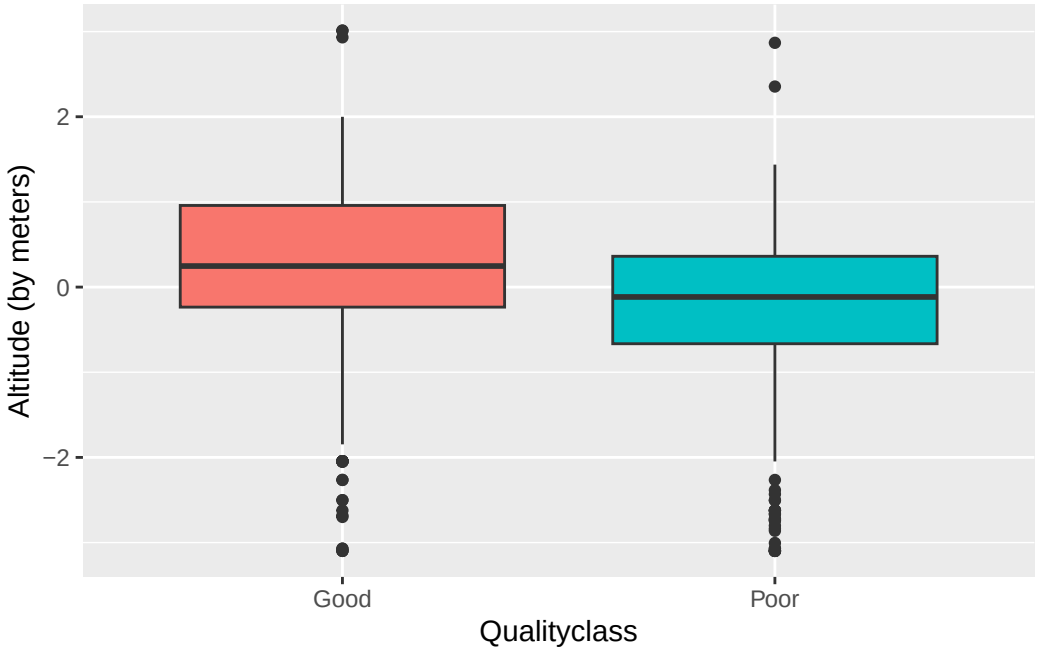


Figure 5: Mean Altitude (in meters) by Coffee Quality Class

Figure 5 shows a boxplot of scaled mean altitude(in meters) of coffee batch being classified as Good and Poor quality. The median altitude for the Good quality class is slightly higher than that of the Poor quality class and IQRs for two classes are similar. However, the Good quality class appears to have a higher overall distribution of altitude values. Both quality classes contain multiple outliers, but the Poor quality class has more extreme low-altitude values.

```
ggplot(data = coffee1, aes(x=Qualityclass,
                           y = harvested,
                           fill=Qualityclass))+
  geom_boxplot()+
  labs(x = "Qualityclass", y = "Harvested Year")+
  theme(legend.position = "none")
```

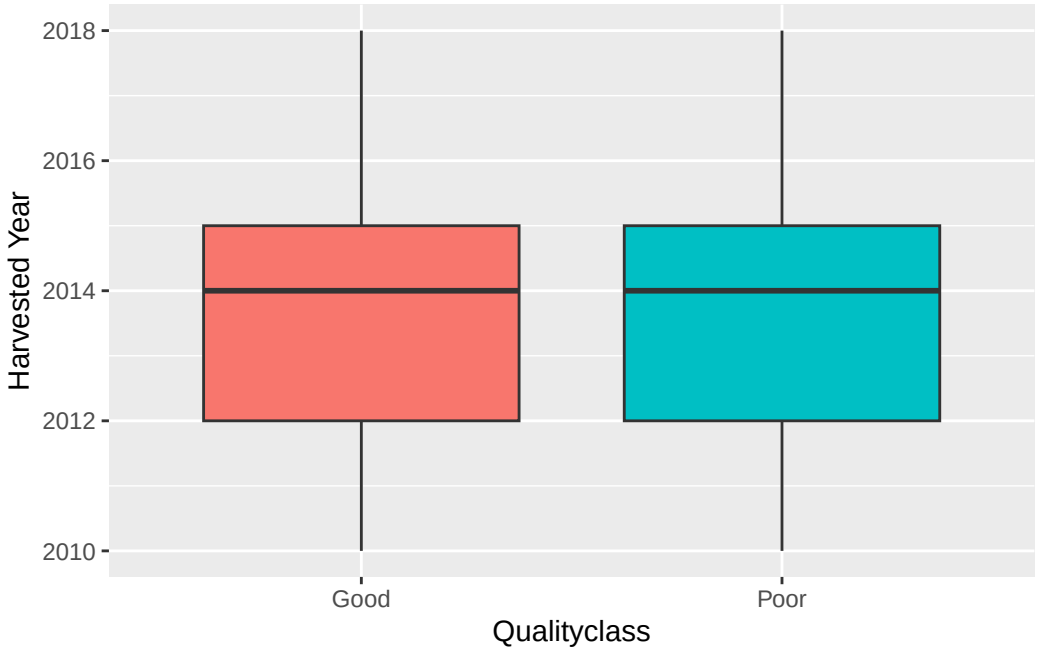


Figure 6: Harvested Year by Coffee Quality Class

Figure 6 is the boxplot of harvested year for coffee batch being classified as Good and Poor quality. The median harvested year for both Good and Poor quality classes appears to be similar and the IQRs are nearly the same. No significant outliers are found in this boxplot.

3 Formal Analysis

3.1 Model Fitting

```
# Fit GLM model with 6 variables exclude country.
model1 <- glm(formula=Qualityclass~ aroma + flavor + acidity +
                category_two_defects +
                altitude_mean_meters +
                harvested,
              data = coffee1,
              family = binomial(link = "logit"))
```

```
# check for the baseline category response variable
levels(coffee1$Qualityclass)
```

```
[1] "Good" "Poor"
```

Note in this example, the baseline category for binary response (Qualityclass) is *Good*.

```
# Summary statistics for the first model.
summ(model1)
```

```
MODEL INFO:
Observations: 882
Dependent Variable: Qualityclass
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
²(6) = 688.21, p = 0.00
Pseudo-R² (Cragg-Uhler) = 0.72
Pseudo-R² (McFadden) = 0.56
AIC = 547.34, BIC = 580.82

Standard errors:MLE
------------------------------------------------------------
                              Est.     S.E.   z val.      p
------------------------- -------- -------- -------- ------
(Intercept)                312.11   120.84     2.58   0.01
aroma                       -4.64     0.72    -6.46   0.00
flavor                      -7.09     0.87    -8.17   0.00
acidity                     -4.13     0.68    -6.09   0.00
category_two_defects         0.00     0.03     0.00   1.00
altitude_mean_meters        -0.35     0.11    -3.13   0.00
harvested                   -0.10     0.06    -1.61   0.11
------------------------------------------------------------
```

The Model 1 suggests that Aroma, Flavor, Acidity grades and mean altitude are the most important factors in determining coffee quality class. Harvested Year is not significant ($p = 0.11$), suggesting that coffee quality classification does not strongly depend on the year it was harvested. Note that the explanatory variable category_two_defects has a p-value of 1, which is not significant at all. Consequently, it should not be considered for the further analysis and should be removed initially.

```r
# Model with one variable (category_two_defects) dropped.
model2 <- glm(formula = Qualityclass~ aroma + flavor + acidity +
                altitude_mean_meters +
                harvested,
              data = coffee1,
              family = binomial(link = "logit"))

summ(model2)
```

```
MODEL INFO:
Observations: 882
Dependent Variable: Qualityclass
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
²(5) = 688.21, p = 0.00
Pseudo-R² (Cragg-Uhler) = 0.72
Pseudo-R² (McFadden) = 0.56
AIC = 545.34, BIC = 574.04

Standard errors:MLE
------------------------------------------------------------
                             Est.     S.E.   z val.      p
------------------------- -------- -------- -------- ------
(Intercept)                312.11   120.35     2.59   0.01
aroma                       -4.64     0.72    -6.46   0.00
flavor                      -7.09     0.87    -8.17   0.00
acidity                     -4.13     0.68    -6.09   0.00
altitude_mean_meters        -0.35     0.11    -3.14   0.00
harvested                   -0.10     0.06    -1.61   0.11
------------------------------------------------------------
```

The Model 2 shows that the variables Aroma, Flavor, Acidity grades and mean altitude are significant predictors of coffee quality. Altitude is statistically significant but has a small effect, meaning it may not be a primary determinant of quality.

AIC and BIC both decreased, indicating that removing insignificant variables does not impact predictive power. Assuming a p-value of 0.05, harvested Year ($p = 0.11$) remains not statistically significant, meaning coffee quality is not strongly dependent on the year of harvest, an attempt was made to remove the only non-significant variable, harvested.

```r
# Model with one more variable dropped.
model3 <- glm(formula = Qualityclass~ aroma + flavor + acidity +
                altitude_mean_meters,
              data = coffee1,
              family = binomial(link = "logit"))

summ(model3)
```

```
MODEL INFO:
Observations: 882
Dependent Variable: Qualityclass
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
²(4) = 685.59, p = 0.00
Pseudo-R² (Cragg-Uhler) = 0.72
Pseudo-R² (McFadden) = 0.56
AIC = 545.96, BIC = 569.87

Standard errors:MLE
------------------------------------------------------------
                              Est.   S.E.   z val.      p
------------------------- -------- ------ -------- ------
(Intercept)                 119.10   8.71    13.68   0.00
aroma                        -4.50   0.71    -6.36   0.00
flavor                       -7.08   0.87    -8.18   0.00
acidity                      -4.20   0.68    -6.19   0.00
altitude_mean_meters         -0.32   0.11    -2.92   0.00
------------------------------------------------------------
```

All Model 2 shows that the variables Aroma, Flavor, Acidity grades and mean altitude are significant. The AIC of Model 3 is observed to increase a little following the removal of the variable harvested, but the BIC decrease as expected. It's a bit hard to determine either Model 2 or Model 3 should be used, thus a further model comparison using deviance is required.

```r
# Fit GLM model including all variables.
coffee1$country_of_origin <- as.factor(coffee1$country_of_origin)
glm_model <- glm(Qualityclass ~ .,
                 data = coffee1,
```

```
                  family = binomial(link = "logit"))
summ(glm_model)
```

```
MODEL INFO:
Observations: 882
Dependent Variable: Qualityclass
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
²(38) = 774.22, p = 0.00
Pseudo-R² (Cragg-Uhler) = 0.78
Pseudo-R² (McFadden) = 0.63
AIC = 525.33, BIC = 711.84

Standard errors:MLE
------------------------------------------------------------------------
                                    Est.      S.E.   z val.       p
---------------------------------- -------- --------- -------- ------
(Intercept)                       417.09    155.75     2.68   0.01
country_of_originBurundi           -1.74      4.94    -0.35   0.72
country_of_originChina              0.20      1.02     0.19   0.85
country_of_originColombia          -1.68      0.56    -3.02   0.00
country_of_originCosta              0.03      0.71     0.05   0.96
Rica
country_of_originCote              11.18   3956.18     0.00   1.00
d?Ivoire
country_of_originEcuador            1.20      1.48     0.81   0.42
country_of_originEl                -0.26      0.94    -0.27   0.78
Salvador
country_of_originEthiopia         -10.89    646.51    -0.02   0.99
country_of_originGuatemala          0.82      0.51     1.61   0.11
country_of_originHaiti             -2.10      1.87    -1.12   0.26
country_of_originHonduras           0.79      0.65     1.21   0.23
country_of_originIndia              3.82      1.22     3.14   0.00
country_of_originIndonesia          0.55      0.91     0.61   0.54
country_of_originKenya             -0.61      1.45    -0.42   0.67
country_of_originLaos              14.68   2674.02     0.01   1.00
country_of_originMalawi             1.35      1.23     1.09   0.27
country_of_originMauritius         11.12   3956.18     0.00   1.00
country_of_originMexico             0.87      0.48     1.83   0.07
```

```
country_of_originMyanmar              14.94   2797.24     0.01   1.00
country_of_originNicaragua            -0.13      1.82    -0.07   0.94
country_of_originPanama               -2.92      1.72    -1.70   0.09
country_of_originPeru                 13.57   3956.18     0.00   1.00
country_of_originPhilippines          -2.29      2.82    -0.81   0.42
country_of_originTaiwan               -0.50      0.62    -0.81   0.42
country_of_originTanzania,            -0.71      0.73    -0.97   0.33
United Republic Of
country_of_originThailand             -2.17      0.96    -2.26   0.02
country_of_originUganda                1.66      0.76     2.17   0.03
country_of_originUnited               -1.67      1.77    -0.94   0.35
States
country_of_originUnited               -3.73   3956.18    -0.00   1.00
States (Hawaii)
country_of_originUnited                2.91      1.66     1.75   0.08
States (Puerto Rico)
country_of_originVietnam              -1.71      1.10    -1.55   0.12
country_of_originZambia               13.40   3956.18     0.00   1.00
aroma                                 -5.30      0.85    -6.25   0.00
flavor                                -7.91      1.02    -7.72   0.00
acidity                               -5.37      0.81    -6.60   0.00
category_two_defects                  -0.06      0.04    -1.49   0.14
altitude_mean_meters                  -0.27      0.16    -1.69   0.09
harvested                             -0.14      0.08    -1.80   0.07
-------------------------------------------------------------------------
```

In the full model, Aroma, Flavor, and Acidity grades remain highly significant, confirming their strong impact on coffee quality, and several countries of origin (like Colombia, India and Thailand) are also statistically significant, indicating that origin may influence coffee quality. However, some countres of origin (like China, Ethiopia and Vietnam) are not significant, meaning they do not strongly predict coffee quality. Moreover, mean altitude, harvest Year, and Category two defects are not significant in this case.

Following the incorporation of all variables into the model, there is an enhancement in R square. That is to say an improvement in interpretability was observed, associated with a reduction in AIC. However, a substantial increase was observed in the BIC, which may not be considered a valuable gain due to the cost of adding a large number of degrees of freedom.

```
Models <- c('model1','model2','model3','glm_model')
model.comp.values.model1 <- glance(model1)
model.comp.values.model2 <- glance(model2)
model.comp.values.model3 <- glance(model3)
model.comp.values.model <- glance(glm_model)
```

```r
bind_rows(model.comp.values.model1,
          model.comp.values.model2,
          model.comp.values.model3,
          model.comp.values.model,.id="Model") %>%
  dplyr::select(Model,null.deviance,df.null,deviance,df.residual,AIC,BIC) %>%
  mutate(Model=Models) %>%
  kable(
     digits = 2
  )
```

Table 1: Model comparison values for different models

Model	null.deviance	df.null	deviance	df.residual	AIC	BIC
model1	1221.55	881	533.34	875	547.34	580.82
model2	1221.55	881	533.34	876	545.34	574.04
model3	1221.55	881	535.96	877	545.96	569.87
glm_model	1221.55	881	447.33	843	525.33	711.84

```r
anova(model3,model2,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: Qualityclass ~ aroma + flavor + acidity + altitude_mean_meters
Model 2: Qualityclass ~ aroma + flavor + acidity + altitude_mean_meters +
    harvested
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       877     535.96
2       876     533.34  1   2.6138   0.1059
```

As Table 1 shows that Model 2 seems to have the smallest AIC values (534.34) followed by Model 3(545.96), while Model 3 has the lowest BIC values(574.04). To determine the selection of the model, we perform the hypothesis test using residual deviance. As the result (2.61) is less than the critical value of $\chi^2(1)$ which is 3.84, so that we're not able to reject the null hypothesis and there's no evidence to show that model 2 (with 5 variables) is better.

Therefore, Model 3 (with 4 variables) is considered as the most suitable model.

3.2 log-odds

```
# Get the coefficents for the intercept and explanatory variables.
# Uese the code for latex expression.
mod1coefs <- round(coef(model3),4)
```

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \cdot \text{aroma} + \beta_2 \cdot \text{flavor} + \beta_3 \cdot \text{acidicity} + \beta_4 \cdot \text{altitude\_mean\_meters}$$
$$= 119.1046 - 4.5035 \cdot \text{aroma} - 7.0773 \cdot \text{flavor} - 4.1992 \cdot \text{acidity} - 0.3219 \cdot \text{altitude\_mean\_meters}$$

where p = Prob (Quality score is **Poor**) and 1 - p = Prob(Quality score is **Good**) as we already check and confirmed the baseline category response is Good in the previous step.

The intercept is 119, meaning that when all explanatory variables are zero, the log-odds of being classified as poor quality is 119. This is extreme condition not normally happened in the real world for coffee batches.

The log-odds of the quality score for the batch is Poor decrease by 4.50 for every unit increase in aroma grade when hold other variables kept unchanged.

Similarly, the log-odds of being classified as poor quality for the batch will decrease by 7.07 and 4.20 for one unit increase in flavor and acidity grades respectively when keep all other variables constant.

Mean altitude has a relative small negative coefficient compared to the previous three variables, so the log-odds of a coffee batch being Poor slightly decrease by 0.32 as one unit increase in mean altitude when hold other variables kept unchanged.

```
# 95% Confidence interval for the log-odds by different explanatory variable.
confint(model3)|>
  kable()
```

Table 2: 95% Confidence Interval for the Log-odds

	2.5 %	97.5 %
(Intercept)	102.9293992	137.1055482
aroma	-5.9335595	-3.1506386
flavor	-8.8354161	-5.4369193
acidity	-5.5576558	-2.8939166
altitude_mean_meters	-0.5408238	-0.1073586

For Table 2, since all the intervals($[-8.83, -5.44]$ for flavor grade, $[-5.93, -3.15]$ for aroma grade, $[-5.56, -2.89]$ for acidity grade, $[-0.54, -0.11]$ for mean altitude) don't contain zero thus indicates that the all explanatory variables in Model 3 are significant.

3.2.1 95% Confidence Interval Plot For Log-Odds

```
# 95% Confidence Interval Plot for Log-Odds.
plot_model(model3, show.values = TRUE, transform = NULL,
           title = "Log-Odds (Poor Coffee Batch Qualtiy Score)", show.p = FALSE)
```

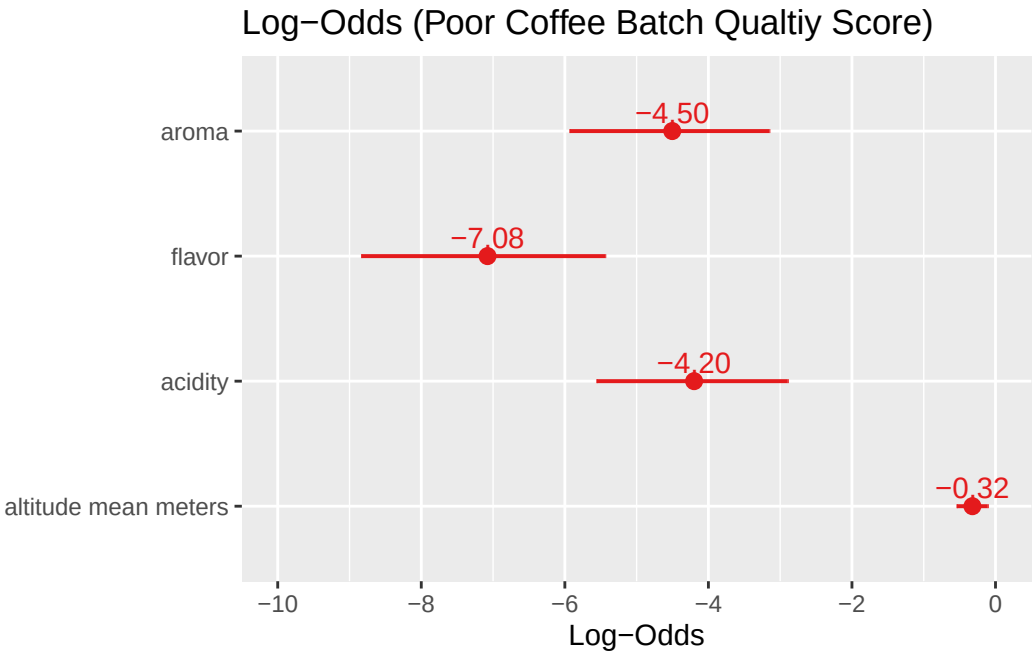


Figure 7: 95% Confidence Interval Plot for Log-Odds

The Figure 8 again confirmed the significance of the explanatory variables visually. Notice that flavor has the widest range of negative values whereas mean altitude has the narrowest confidence range.

For more straightforward interpretation, we use Odds ratio scale.

3.3 Odds Ratio

$$\frac{p}{1-p} = \exp(\alpha + \beta_1 \cdot \text{aroma} + \beta_2 \cdot \text{flavor} + \beta_3 \cdot \text{acidicity} + \beta_4 \cdot \text{mean altitude})$$

Table 3: Summary Table on the Odds Scale

Variable	exp(coef(model3))
(Intercept)	5.326930e+51
aroma	1.106983e-02
flavor	8.440891e-04
acidity	1.500796e-02
altitude_mean_meters	7.247855e-01

```
model3 |>
  coef()|>
  exp()|>
  as.data.frame()|>
  rownames_to_column(var = "Variable")|>
  gt()
```

the value of the intercept (1.44×10^{52}) gives the odds of a quality class being poor given all explanatory variables equal to zero, but in reality, it is highly unlikely for a batch of coffee have the grades of aroma, flavor and acidity all approach to zero. It could still be interpreted as when chance of a batch of coffee be classified as poor quality are (1.44×10^{52}) % greater than them being classified as good quality when all the explanatory variable equals to zero.

For aroma grade, we have an odds of 1.11×10^{-2}, which indicates that for every 1 unit increase in aroma grade, the odds of the quality class being poor decrease by $(1 - 1.11 \times 10^{-2}) \times 100$ % = 98.9% when other variables unchanged. Thus higher aroma grades strongly reduce the likelihood of being classified as poor quality.

Similarly, the odds for flavor grade and acidity grades are 8.44×10^{-4} and for 1.50×10^{-2} respectively, which both less than 1. This means that odds of being classified as poor quality will reduce by 99.9% and 98.5% respectively for every 1 unit increase in flavor grade or acidity grade, keeping all other variables constant.

Finally, the odds for mean altitude(meters) is 7.25×10^{-1}, nearly approach to 1, indicating that for every 1 unit increase in mean altitude (meters), it decreases the odds of being classified as poor quality by 27.5% when hold all other variables unchanged. Notice that compared to other explanatory variables, mean altitude has a relative small effect on coffee quality classification.

```
# 95% Confidence interval for the odds by different explanatory variable.
confint(model3) |>
  exp() |>
  kable()
```

Table 4: 95% Confidence Interval for the Odds

	2.5 %	97.5 %
(Intercept)	5.031183e+44	3.500927e+59
aroma	2.649000e-03	4.282480e-02
flavor	1.455000e-04	4.352900e-03
acidity	3.857800e-03	5.535900e-02
altitude_mean_meters	5.822684e-01	8.982035e-01

```
plot_model(model3, show.values = TRUE,
           title = "Odds (Poor Quality Score)", show.p = FALSE)
```

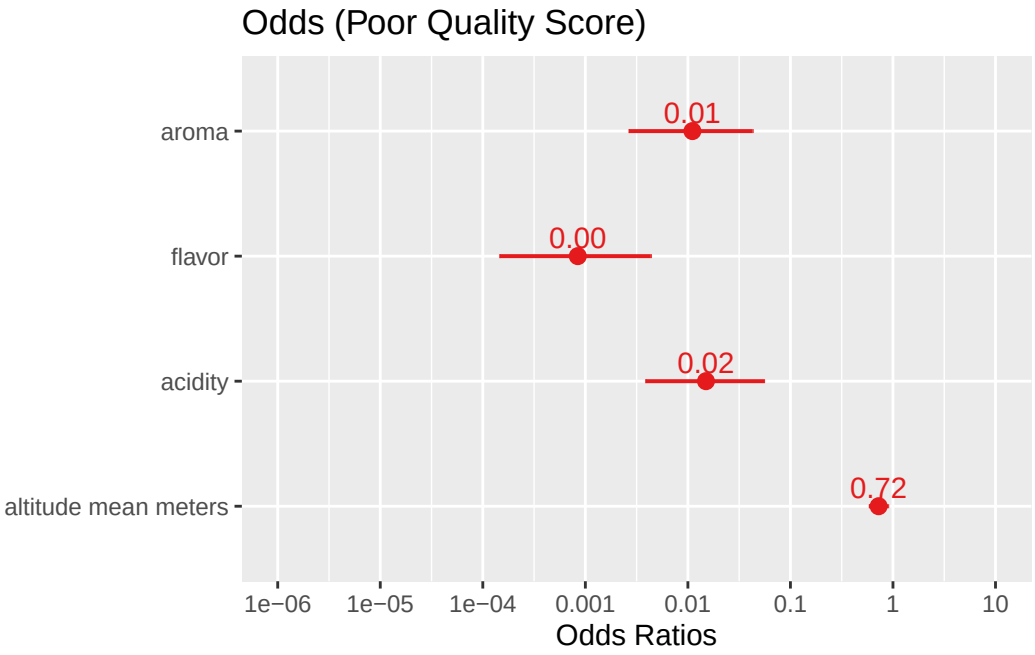


Figure 8: 95% Confidence Interval Plot for Odds Ratio

From Table 4, most of the intervals($[0.00015, 0.0044]$ for flavor grade, $[0.0026, 0.0428]$ for aroma grade, $[0.0039, 0.0554]$ for acidity grade, $[0.582, 0.898]$ for mean altitude) don't contain one thus indicate that the explanatory variables in model 3 are significant.

The Figure 8 shows the 95% confidence intervals for three explanatory variable, again confirmed the significance of the explanatory variables visually.

3.4 Probability

As we could obtain the probability $p = \text{Prob}$ (Quality score is **Poor**) using:

$$p = \frac{\exp(\alpha + \beta_1 \cdot \text{aroma} + \beta_2 \cdot \text{flavor} + \beta_3 \cdot \text{acidicity} + \beta_4 \cdot \text{mean altitude})}{1 + \exp(\alpha + \beta_1 \cdot \text{aroma} + \beta_2 \cdot \text{flavor} + \beta_3 \cdot \text{acidicity} + \beta_4 \cdot \text{mean altitude})}$$

```
# Add probability to the dataset.
coffee1 <- coffee1 |>
  mutate(probs.poor = fitted(model3))
```

```
ggplot(data = coffee1, aes(x = aroma, y = probs.poor)) +
  geom_smooth(method="glm",
              method.args = list(family="binomial"),
              se = FALSE) +
  labs(x = "Aroma Grade",
       y = "Probability of Quality Score is Poor by Aroma Grade")
```

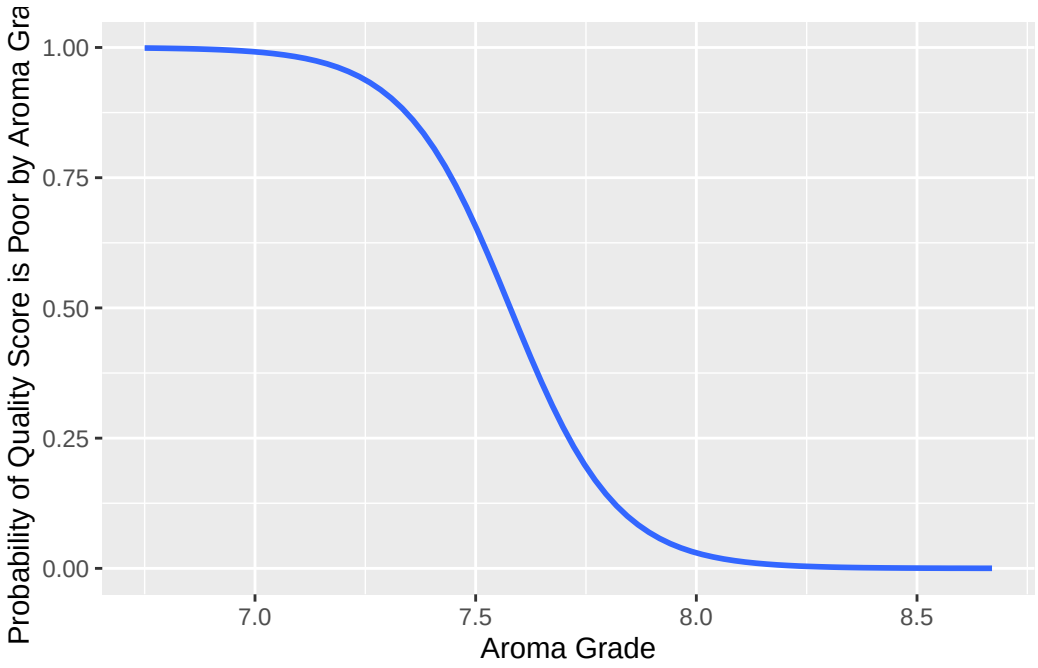


Figure 9: Probability of Quality Score is Poor by Aroma Grade.

```
ggplot(data = coffee1, aes(x = flavor, y = probs.poor)) +
  geom_smooth(method="glm",
              method.args = list(family="binomial"),
              se = FALSE) +
  labs(x = "Flavor Grade",
       y = "Probability of Quality Score is Poor by Flavor Grade")
```

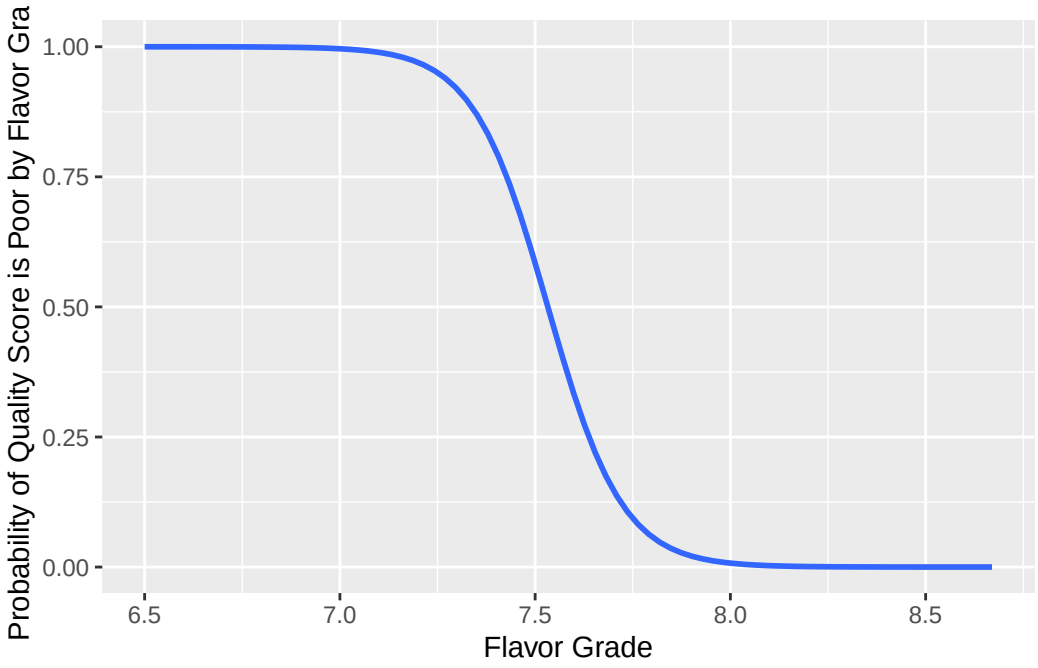


Figure 10: Probability of Quality Score is Poor by Flavor Grade.

```
ggplot(data = coffee1, aes(x = acidity, y = probs.poor)) +
  geom_smooth(method="glm",
              method.args = list(family="binomial"),
              se = FALSE) +
  labs(x = "Acidity Grade",
       y = "Probability of Quality Score is Poor by Acidity Grade")
```

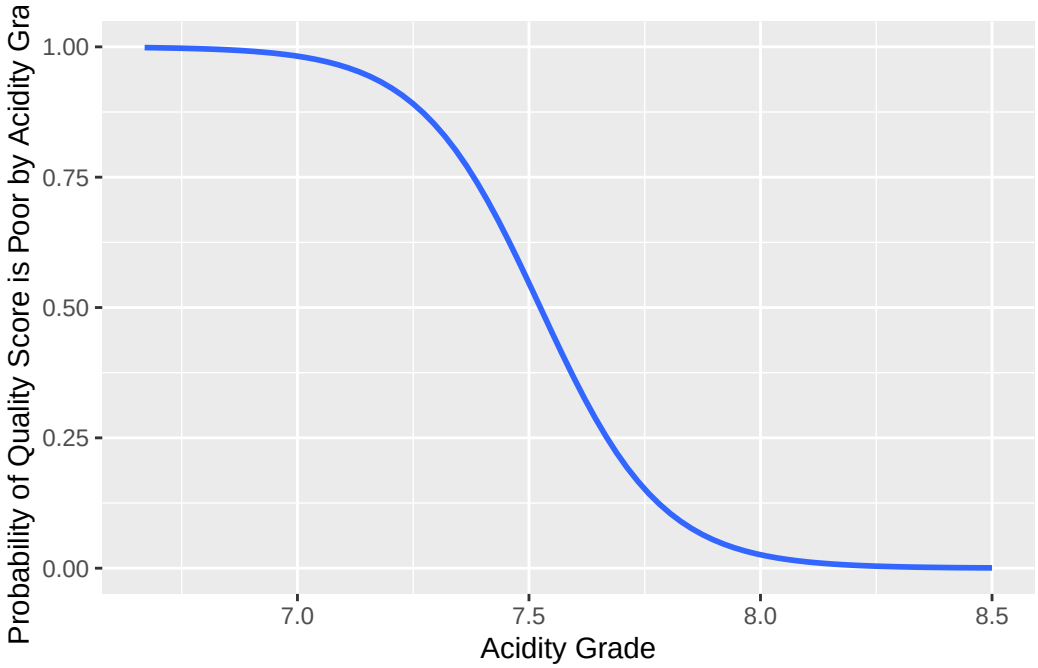


Figure 11: Probability of Quality Score is Poor by Acidity Grade.

```
ggplot(data = coffee1, aes(x = altitude_mean_meters, y = probs.poor)) +
  geom_smooth(method="glm",
              method.args = list(family="binomial"),
              se = FALSE) +
  labs(x = "Mean Altitude(in meters)",
       y = "Probability of Quality Score is Poor by Mean Altitude(in meters)")
```

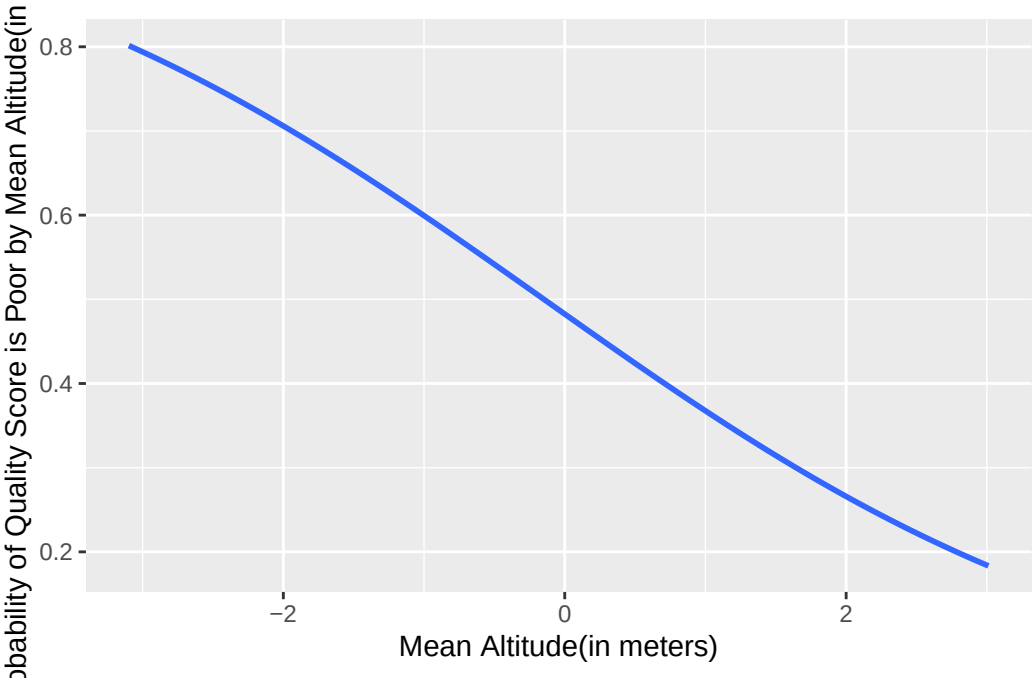


Figure 12: Probability of Quality Score is Poor by altitude_mean_meters Grade.

As Figure 9 shows the probability of the quality score being classified as poor by aroma grade, the less the aroma grade , the higher probability being classified as poor quality. It's reasonable to have such shape as the median aroma grade is round 7.2 (recalled from the Figure 1) so the outliers (aroma grade equals to zero) will be directly classified as poor quality with the probability of 1.

Similar interpretations will used for Figure 10, Figure 11. These two plots indicate the probabilities of the quality score being classified as poor by flavor and acidity grade. Again, the higher score of the flavor grade, acidity grade, the less probability being classified as poor quality. The probabilities decrease gradually at first, then sharply around a certain point, and finally gradually approach to zero.

As Figure 12 shows the probability of the quality score being classified as poor by mean altitude, lower altitudes correspond to a higher probability of being classified as poor quality, aligning with Figure 5, where the median altitude of Poor quality class has more low-altitude samples. The probability curve declines with a general constant rate, indicating better quality at higher altitudes.

4 Conclusion

Sensory attributes like Aroma, Flavor, Acidity grades are the the most influential variables for predicting the quality score of coffee batch: Figure 9, Figure 10, and Figure 11 indicate that higher aroma, flavor, and acidity scores significantly reduce the probability of being classified as Poor quality. Notice that flavor has the largest impact, followed by aroma and acidity. Boxplots (Figure 1, Figure 2, Figure 3) further show that good quality coffee has consistently higher sensory scores. Then, mean altitude has a moderate but still significant effect on the quality class of coffee batch.

Therefore, by focusing on improving flavor, aroma, and acidity grades (or choose coffee with higher aroma, flavor and acidity grades), and growing coffee at higher altitudes, the local coffee farmers will have a greater chance that their coffee being classified as high quality (Good).