Analyzing the influence of coffee characteristics on its quality classification

Group 11

1 Introduction

Coffee quality assessment is a crucial aspect of the coffee industry, impacting both market value and consumer preferences. This research aims to explore the factors that influence whether a batch of coffee is classified as "Good" or "Poor", using data from the Coffee Quality Database (CQD).

This dataset includes more than 1000 coffee samples from different countries and records sensory attributes (aroma, flavor, and acidity), production characteristics (harvest year, altitude), and defect counts. The coffee batches are assigned a quality classification based on an overall score threshold (82.5 as "Good", <82.5 as "Poor").

The primary goal of this research is to analyze how the factors in this dataset affect coffee quality, especially the sensory attributes(aroma, flavor, and acidity).

2 Exploratory Data Analysis

2.1 Data preparing & Cleaning

```
#data cleaning
data<-read.csv("dataset11.csv") #read the data
data<-na.omit(data) #remove the NA value
data$Qualityclass_dummy<-ifelse(data$Qualityclass=="Good",1,0) #given the
 ↪  value to Qualityclass for "Good"=1 "Poor"=0
data$Qualityclass <- as.factor(data$Qualityclass)
data$harvested <- as.factor(data$harvested) #consider variables Qualityclass
 ↪  and harvested as categorical variables
#Outlier in altitude_mean_meters
```

```r
sum(data$altitude_mean_meters>8848)+sum(data$altitude_mean_meters<0)
data=data%>%
  filter(data$altitude_mean_meters<8848 & data$altitude_mean_meters>0)
#standarized for altitude_mean_meters
data$altitude_mean_meters=data$altitude_mean_meters/1000 #change the unit of
 ↪  variable altitude_mean_meters
```

After handling missing values and filtering altitude outliers, the data cleaning process resulted in a robust dataset suitable for modeling. The altitude variable (altitude_mean_meters) initially showed unrealistic values, such as extreme altitudes over 8,848 meters (the height of Mount Everest), which are highly unlikely for coffee cultivation.

Filtering these anomalies ensures subsequent analyses reflect realistic conditions.

```r
#visualization of the density for 4 continuous variables
par(mfrow=c(2,2))
plot(density(data$aroma), col = "blue", main = "Density Plot of aroma")
plot(density(data$flavor), col = "blue", main = "Density Plot of flavor")
plot(density(data$acidity), col = "blue", main = "Density Plot of acidity")
plot(density(data$altitude_mean_meters), col = "blue", main = "Density Plot
 ↪  of altitude_mean_meters")
```

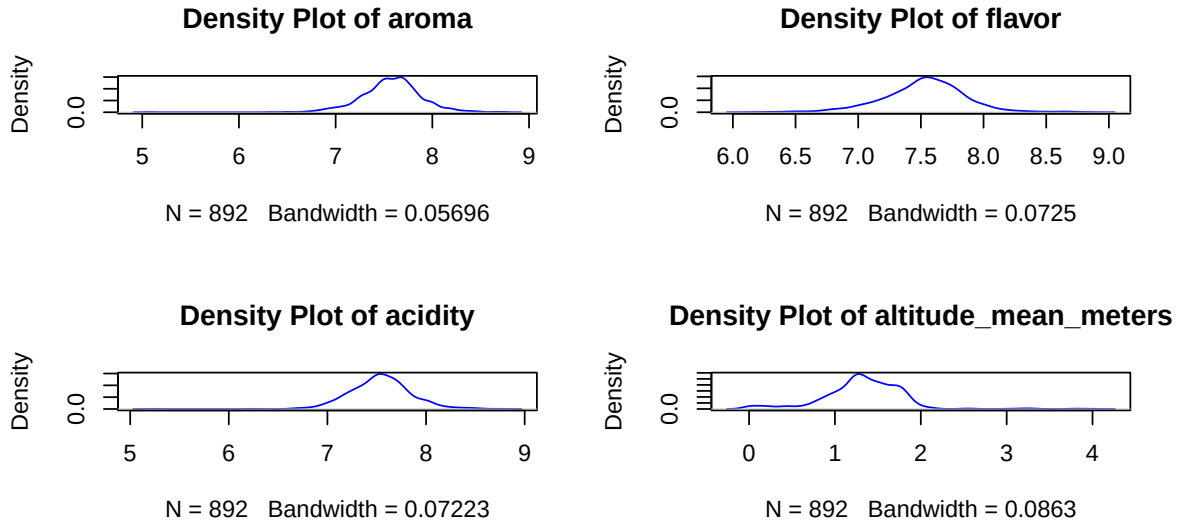

Figure 1: the density of explanatory variables

In this dataset, aroma,flavor,acidity and altitude and category_two_defects are continuous variables, Qualityclass and harvested are categorical variables.

According to Figure 1, it indicate that aroma, flavor, and acidity follow approximate normal distributions with a central peak around a score of 7.5.

These sensory characteristics appear consistently rated across samples. In contrast, the distribution of altitude_mean_meters exhibits strong right-skewness, indicating that while most coffees originate from relatively moderate altitudes, a few originate from significantly higher altitudes.

To keep all variables under the same metric scale. We scale the altitude variable by changing the unit from meters to kilometers.

2.2 Graphical Summaries

```
#change formula into long formula
data_long <- data %>%
  pivot_longer(cols = c(aroma,flavor,acidity,category_two_defects,
                        altitude_mean_meters),
               names_to = "Variable",
               values_to = "Value")
```

```
#the boxplots of continuous explanatory variables
ggplot(data = data_long, aes(x = Qualityclass, y = Value, fill =
 ↪  Qualityclass))+
  geom_boxplot() +
  facet_wrap(~Variable, scales = "free_y") +
  theme_minimal() +
  labs(x = "qualityclass",
       y = "Value") +
  theme(legend.position = "none")
```

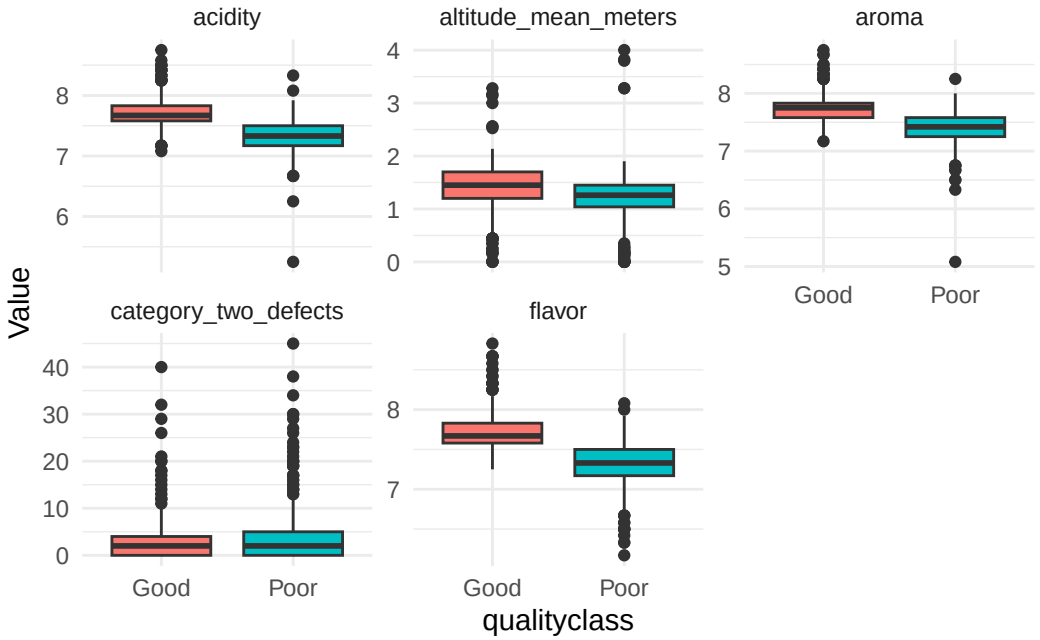

Figure 2: the boxplots of continuous explanatory variables

Figure 2 reveal clear differences between the Good and Poor coffee classes.

Coffees classified as "Good" consistently have higher median scores in sensory characteristics (aroma, flavor, and acidity) compared to those classified as "Poor".

Notably, flavor and aroma exhibit minimal overlap between categories, highlighting their crucial roles in quality assessment. Additionally, the distribution of category_two_defects demonstrates that lower defect counts correlate strongly with higher quality classifications, suggesting defect management is essential in improving coffee quality.

The altitude variable shows less pronounced differences between classes, implying altitude alone may not be a decisive factor for coffee quality.

```
#the barplot of categorical explanatory variable
ggplot(data, aes(x=harvested ,  y = ..prop.., group=Qualityclass,
                 fill=Qualityclass)) +
    geom_bar(position="dodge", stat="count") +
    labs(y = "Proportion")
```

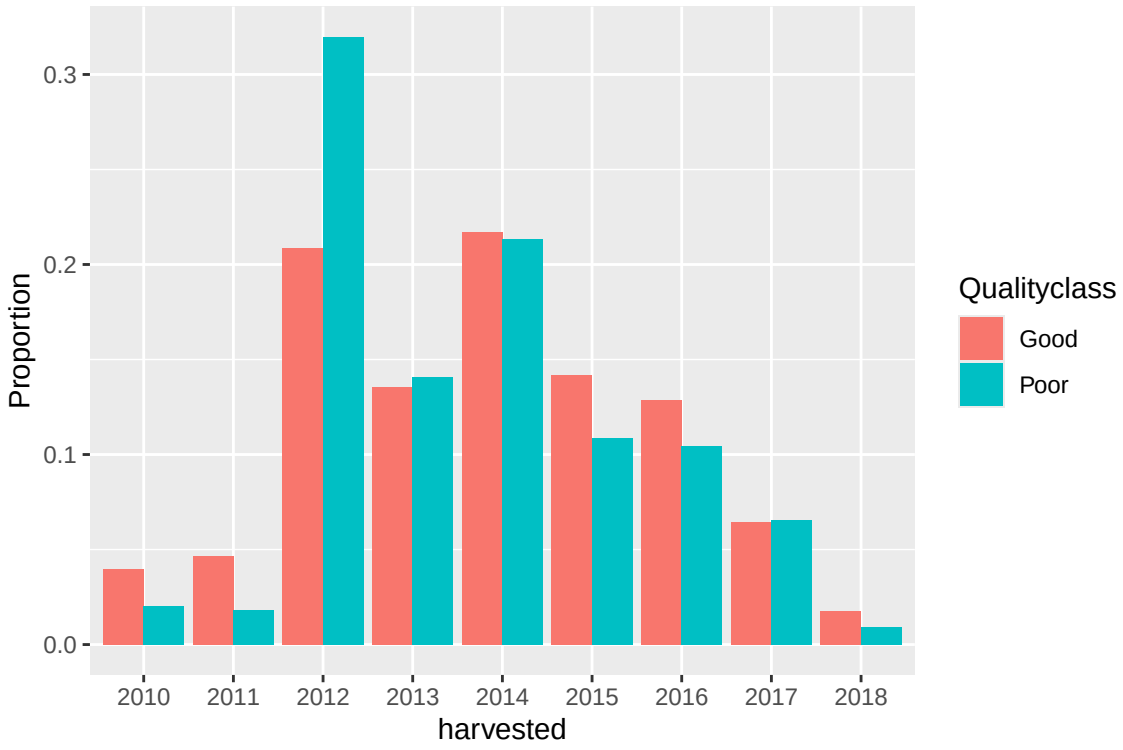

Figure 3: the barplot of categorical explanatory variable

Figure 3 illustrates yearly variations in coffee quality from 2010 to 2018.

Notably, higher proportions of "Good" coffee occur during the harvest years of 2014 to 2016, suggesting potential improvements in growing conditions or processing methods during this period.

Conversely, the years 2012 and 2013 exhibit higher proportions of "Poor" quality coffee, indicating possible unfavorable conditions or practices.

These temporal fluctuations highlight the potential impact of external factors such as climatic conditions, agricultural practices, or technological advancements on coffee quality.

2.3 Numerical Summaries

```
#create a table for Summary Statistics of continuous factors
library(gt)
data_long |>
summarize('Mean' = mean(Value),
```

```
'Median' = median(Value),
'St.Dev' = sd(Value),
'Min' = min(Value),
'Max' = max(Value),
'IQR' = quantile(Value,0.75)-quantile(Value,0.25),
.by = Variable) |>
gt() |>
fmt_number(decimals=2) |>
cols_label(
Mean = html("Mean"),
Median = html("Median"),
St.Dev = html("Std. Dev"),
Min = html("Minimum"),
Max = html("Maximum"),
IQR = html("IQR"),
)
```

Table 1: Summary Statistics of continuous factors

Variable	Mean	Median	Std. Dev	Minimum	Maximum	IQR
aroma	7.57	7.58	0.32	5.08	8.75	0.33
flavor	7.53	7.58	0.33	6.17	8.83	0.42
acidity	7.54	7.50	0.31	5.25	8.75	0.42
category_two_defects	3.50	2.00	5.21	0.00	45.00	4.00
altitude_mean_meters	1.32	1.31	0.47	0.00	4.00	0.50

According to Table 1, the continuous variables Aroma, Flavor, Acidity have similar distributions, which mean is around 7.5, ranging from 5.08 to 8.83.The number of category_two_defects ranges from 0 to 45, which mean is 3.5, indicating some variation in defect levels.The average altitude of coffee cultivation is approximately 1.32 kilometers, ranging from 1 meter to 4,001 meters.

```
# create a table for Summary Statistics of harvested
data |>
  count(harvested, name = "Count") |> #count the number of each year
  pivot_wider(
    names_from = harvested,
    values_from = Count,
  ) |>
  mutate(Row = "Count") |>
  relocate(Row) |>
```

```
gt()|>
cols_label(Row = "Year") #make a table
```

Table 2: Summary Statistics of harvested

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
Count	27	29	235	123	192	112	104	58	12

Table 2 covers coffee harvested from 2010 to 2018. The data are mainly concentrated in the period 2012-2016. The year 2010,2011,2017,2018 have data less than 100.

3 Formal Data Analysis

3.1 Model Creation

The logistic regression model is given by:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \cdot X_{\text{aroma}} + \beta_2 \cdot X_{\text{flavor}} + \beta_3 \cdot X_{\text{acidity}} + \beta_4 \cdot X_{\text{defects}} + \beta_5 \cdot X_{\text{altitudes}} + \text{harvested}$$

Where:

- α is the intercept of the model.
- $\{\beta_i\}, i = 1, \ldots, 5$are the coefficients. Which means when X_i increases 1, the probability will change according to the β_i.
- harvested is considered as a categorical variable. And we consider 2010 as the baseline.So the harvested like the function below:

  $$\text{harvested} = \mathbb{I}_{2011}(x) + \mathbb{I}_{2012}(x) + \mathbb{I}_{2013}(x) + \mathbb{I}_{2014}(x) + \mathbb{I}_{2015}(x) + \mathbb{I}_{2016}(x) + \mathbb{I}_{2017}(x) + \mathbb{I}_{2018}(x)$$

- The$\mathbb{I}_{\text{j}}(x)$ is an indicator function of group $j = 2011, 2012, \ldots, 2018$ like

  $$\mathbb{I}_{\text{j}}(x) = \begin{cases} 1 & \text{if group of harvested } x \text{ is considered as j,} \\ 0 & \text{Otherwise.} \end{cases}$$

- $p = Prob(good)$ represent the probability of coffee quality being good.

3.2 Model Comparison

```
#model fitted for original
model1 <- glm(Qualityclass_dummy ~
 ↪  aroma+flavor+acidity+category_two_defects+altitude_mean_meters+harvested,
 ↪  data = data, family = binomial(link = "logit"))
model1 %>%
  tidy() %>%                  # make a table of information of model1
  gt()
```

Table 3: model information of model1

term	estimate	std.error	statistic	p.value
(Intercept)	-122.73398126	9.13818803	-13.43088814	3.986021e-41
aroma	5.01919349	0.73348682	6.84292256	7.759361e-12
flavor	7.28685055	0.89073490	8.18071749	2.821587e-16
acidity	3.85847808	0.70390741	5.48151366	4.217021e-08
category_two_defects	0.03012466	0.03010589	1.00062349	3.170089e-01
altitude_mean_meters	0.59344524	0.24493464	2.42287183	1.539836e-02
harvested2011	-0.09248562	1.09081913	-0.08478548	9.324319e-01
harvested2012	-0.70231732	0.91315658	-0.76910942	4.418284e-01
harvested2013	-0.25318140	0.92487705	-0.27374601	7.842798e-01
harvested2014	0.06465252	0.92130736	0.07017476	9.440546e-01
harvested2015	-0.47105899	0.92750507	-0.50787754	6.115392e-01
harvested2016	0.38056668	0.95734052	0.39752488	6.909804e-01
harvested2017	0.15493486	0.96122888	0.16118415	8.719484e-01
harvested2018	1.59093134	1.25506421	1.26760952	2.049374e-01

```
summ(model1)
```

Table 3 shows p-value of category_two_defects and harvested in model1 are higher than 0.05.

```
#model fitted for original
#del category_two_defects variable
model2 <- glm(Qualityclass_dummy ~
 ↪  aroma+flavor+acidity+altitude_mean_meters+harvested, data = data, family
 ↪  = binomial(link = "logit"))
model2 %>%
  tidy() %>%                  # make a table of information of model2
  gt()
```

Table 4: model information of model2

term	estimate	std.error	statistic	p.value
(Intercept)	-122.23497549	9.0972939	-13.43641046	3.699515e-41
aroma	5.02279752	0.7346423	6.83706540	8.083185e-12
flavor	7.24651911	0.8934024	8.11114824	5.014364e-16
acidity	3.82890946	0.7040998	5.43802095	5.387564e-08
altitude_mean_meters	0.60857298	0.2443429	2.49065178	1.275090e-02
harvested2011	-0.05273292	1.0888969	-0.04842784	9.613753e-01
harvested2012	-0.58226694	0.9034899	-0.64446424	5.192744e-01
harvested2013	-0.19931567	0.9215781	-0.21627649	8.287722e-01
harvested2014	0.10122429	0.9188511	0.11016398	9.122793e-01
harvested2015	-0.40803666	0.9234916	-0.44184121	6.586041e-01
harvested2016	0.44625604	0.9534571	0.46803996	6.397560e-01
harvested2017	0.22479967	0.9567970	0.23495023	8.142474e-01
harvested2018	1.66547070	1.2502808	1.33207736	1.828348e-01

```
summ(model2)
```

In model2, it remove the variable category_two_defects from the model1. According the Table 4, the p-value of harvested in model2 are still higher than 0.05. The AIC & BIC are decreased compared with model1.

```
#del harvested variable
model3 <- glm(Qualityclass_dummy ~ aroma+flavor+acidity+altitude_mean_meters,
 ↪  data = data, family = binomial(link = "logit"))
model3 %>%
  tidy() %>%                    # make a table of information of model3
  gt()
```

```
summ(model3)
```

In model3, it remove the variable category_two_defects from the model1. According the Table 5, the p-value of harvested in model3 are all lower than 0.05. The AIC & BIC are decreased compared with model2. But in this model, altitude shows a bit significant. The variable altitude needs to be confirmed as reserved.

Table 5: model information of model3

term	estimate	std.error	statistic	p.value
(Intercept)	-119.1157483	8.6855970	-13.714169	8.351738e-43
aroma	4.6554798	0.6903270	6.743876	1.542166e-11
flavor	7.0402422	0.8616280	8.170860	3.061979e-16
acidity	4.0018227	0.6881917	5.814982	6.064028e-09
altitude_mean_meters	0.4634647	0.2317218	2.000091	4.549041e-02

```
#del altitude_mean_meters variable
model4 <- glm(Qualityclass_dummy ~ aroma+flavor+acidity, data = data, family
 ↪  = binomial(link = "logit"))
model4 %>%
  tidy() %>%                    # make a table of information of model3
  gt()
```

Table 6: model information of model4

term	estimate	std.error	statistic	p.value
(Intercept)	-118.953719	8.6476512	-13.755610	4.712871e-43
aroma	4.808518	0.6897372	6.971522	3.135294e-12
flavor	6.890154	0.8482201	8.123074	4.545221e-16
acidity	4.057666	0.6822451	5.947518	2.722381e-09

```
summ(model4)
```

In model4, AIC increased but BIC decreased, so the variable altitude has a bit obvious significance to this model.

```
# the comparison of all models
Models<-c('model1','model2','model3','model4')
model_summaries <- bind_rows(
  glance(model1),
  glance(model2),
  glance(model3),
  glance(model4),
  .id = "Model_ID"
)
```

```r
model_summaries$Model_ID <- Models
kable(model_summaries, digits = 2, caption = "Comparison of All Models")
```

Table 7: Comparison of All Models

Model_ID	null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
model1	1236.46	891	-256.43	540.86	607.97	512.86	878	892
model2	1236.46	891	-256.94	539.89	602.20	513.89	879	892
model3	1236.46	891	-264.19	538.38	562.35	528.38	887	892
model4	1236.46	891	-266.16	540.32	559.50	532.32	888	892

By compared the AIC, BIC, log-likelihood, and deviance values. Model3 shows the lowest AIC(538.38). Although model4 has lowest BIC and highest logLik, more relevant variables need to be considered(e.g altitude). Finally, model3 is selected.

So the final logistic regression model is given by:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \cdot X_{\text{aroma}} + \beta_2 \cdot X_{\text{flavor}} + \beta_3 \cdot X_{\text{acidity}} + \beta_4 \cdot X_{\text{altitudes}}$$

Where:

- α is the intercept of the model.
- $\{\beta_i\}, i = 1, \ldots, 5$are the coefficients. Which means when X_i increases 1, the probability will change according to the β_i.
- $p = Prob(good)$ represent the probability of coffee quality being good.

```r
levels(data$Qualityclass) #base on "good"
```

```r
# CI for optimization model
mod1coefs3 <- round(coef(model3), 2)
library(knitr)
confint(model3) %>%
  kable(caption = "Confidence Interval of model-3")
```

Table 8: Confidence Interval of model-3

	2.5 %	97.5 %
(Intercept)	-137.1043034	-103.0008636
aroma	3.3414665	6.0512855
flavor	5.4098353	8.7926752
acidity	2.6776888	5.3801746
altitude_mean_meters	0.0062384	0.9181653

Through the 95% confidence intervals for all variables, the CI of Aroma, flavor, and acidity are both positive and 0 is not included in CI. Which means these three variables have strong positive effect to the qualityclass.The variable altitude is also positive and not include 0.But the lower bound of CI is near zero.

3.3 Log-odds

```
mod.coef.logodds<-model3 %>%
                  summary() %>%
                  coef()
data<- data%>%
      mutate(logodds.good = predict(model3))
```

```
plot_model(model3, show.values = TRUE, transform = NULL,
           show.p = FALSE)
```

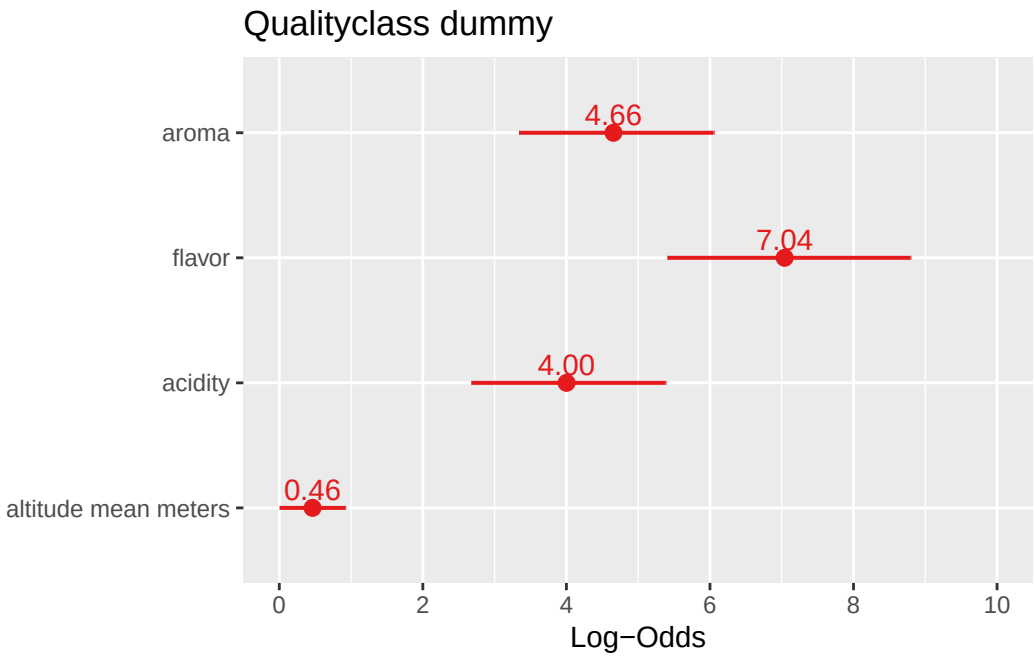

Figure 4: the log-odds of explanatory variables for quality good

The Figure 4 indicates that variables like aroma, flavor, acidity have higher scores significantly increase the likelihood of the quality of coffee being classified as "Good." Specifically, the variable flavor has the largest and positive log-odds(7.04). Which means flavor has the strong impact to the classified of coffee quality. The second factor has positive influence is aroma with log-odds(4.66). The third one is acidity which has a bit lower impact than aroma.The altitude has the smaller coefficient indicates that altitude's direct effect on classified of coffee quality is comparatively minor.

3.4 Odds

```
model3 %>%
 coef() %>%
  exp() %>%
  enframe(name = "Variable", value = "Odds Ratio") %>%
  kable(digits = 2, caption = "Odds Ratios from Model3")
```

Table 9: Odds Ratios from Model3

Variable	Odds Ratio
(Intercept)	0.00
aroma	105.16
flavor	1141.66
acidity	54.70
altitude_mean_meters	1.59

```
#check value
exp(coef(model3))
```

```
#add a column to data
data <- data %>%
  mutate(
    odds.good = exp(logodds.good),
    prob.good = fitted(model3)
  )
```

```
#odd ratio for quality good
plot_model(model3, show.values = TRUE, axis.lim=c(0,10000),
           show.p = FALSE)
```

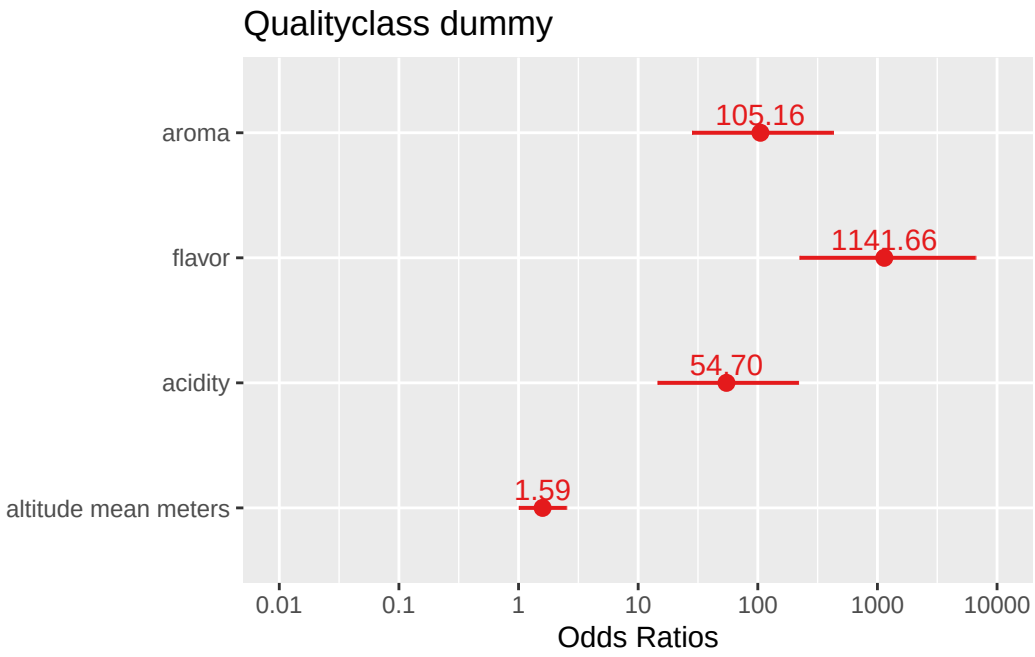

Figure 5: the odds of explanatory variables for quality good

Figure 5 highlights that among all explanatory variables, flavor emerges as the strongest predictor of coffee quality, where flavor increase 1(point) the probability of coffee quality considered as good will multiply 1141.66. Aroma and acidity also positively influence coffee quality, where aroma increase 1(point) the probability of coffee quality considered as good will multiply 105.16 and acidity increase 1(point) the probability of coffee quality considered as good will multiply 54.7. The odds ratio for altitude is close to 1, reinforcing earlier insights that altitude alone minimally influences quality classification. Therefore, it can be concluded that the most important factor influencing the classification of coffee quality is flavor followed by aroma and acidity. altitude has almost no effect on the classification of coffee quality.

3.5 Probabilities

These probability curves further detail the predictive relationships between explanatory variables and coffee quality.

3.5.1 Continuous variables

```
#aroma/acidity/flavor prob
data_long1 <- data %>%
 pivot_longer(cols = c(aroma, flavor,acidity), names_to = "Type", values_to =
  ↪  "Value")
```

In Figure 6 the curves for variables (aroma, flavor, and acidity) exhibit clear upward trends, indicating a steep increase in the probability of achieving "Good" quality with higher sensory scores. For instance, coffees scoring above approximately 7.7-8 in flavor have probabilities exceeding 90% of being classified as "Good."

```
ggplot(data = data_long1, aes(x =Value, y =prob.good, color = Type)) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
  ↪  FALSE) +
 labs(x = "score", y = "Probability of quality being good", color =
  ↪  "character") +
 theme_minimal()
```

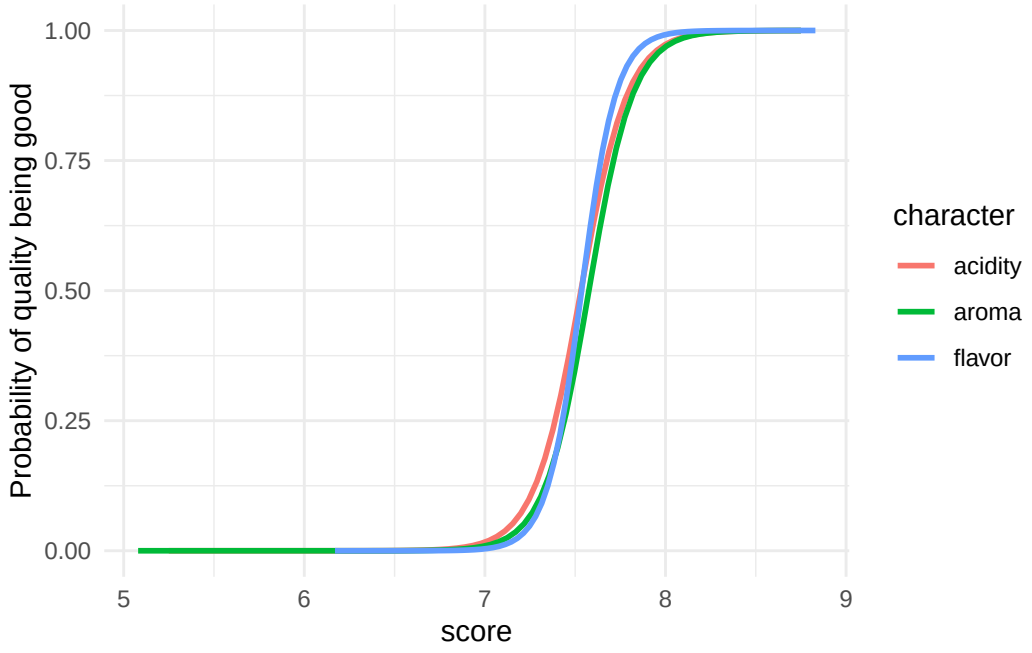

Figure 6: the prob of aroma/acidity/flavor for quality good

3.5.2 Categorical variable

```
ggplot(data = data, aes(x =altitude_mean_meters, y =prob.good)) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
  ↪  FALSE) +
 labs(x = "kilometers", y = "Probability of quality being good", color =
  ↪  "character") +
 theme_minimal()
```

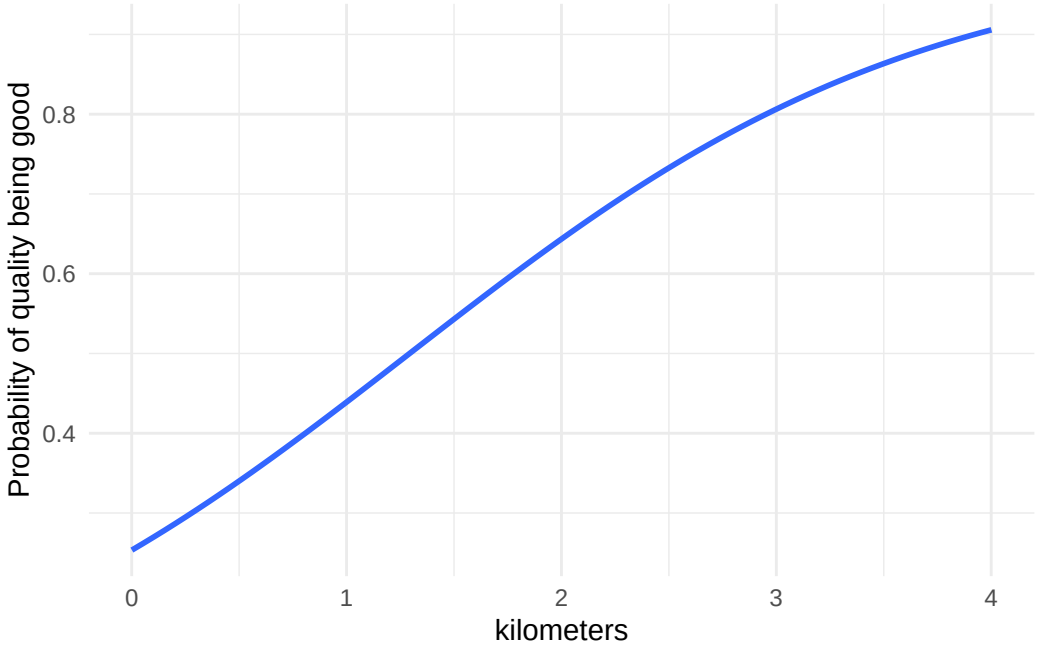

Figure 7: the prob of altitude_mean_meters for quality good

In Figure 7, the curve for altitude appears relatively flat, suggesting altitude plays a limited role in predicting coffee quality in this model.

```
p1=plot_model(model3, type = "pred",terms = "aroma" ,title = "Aroma",
            axis.title = c("aroma", "Prob. of quality being good"))
p2=plot_model(model3, type = "pred", terms="flavor",title = "Flavor",
            axis.title = c("flavor", "Prob. of quality being good"))
p3=plot_model(model3, type = "pred",terms = "acidity", title = "Acidity",
            axis.title = c("acidity", "Prob. of quality being good"))
p4=plot_model(model3, type = "pred",terms = "altitude_mean_meters", title =
 ↪  "altitude_mean_meters",
```

```
            axis.title = c("altitude_mean_meters", "Prob. of quality being
             ↪  good"))
#merge
grid.arrange(p1,p2,p3,p4,nrow=2)
```

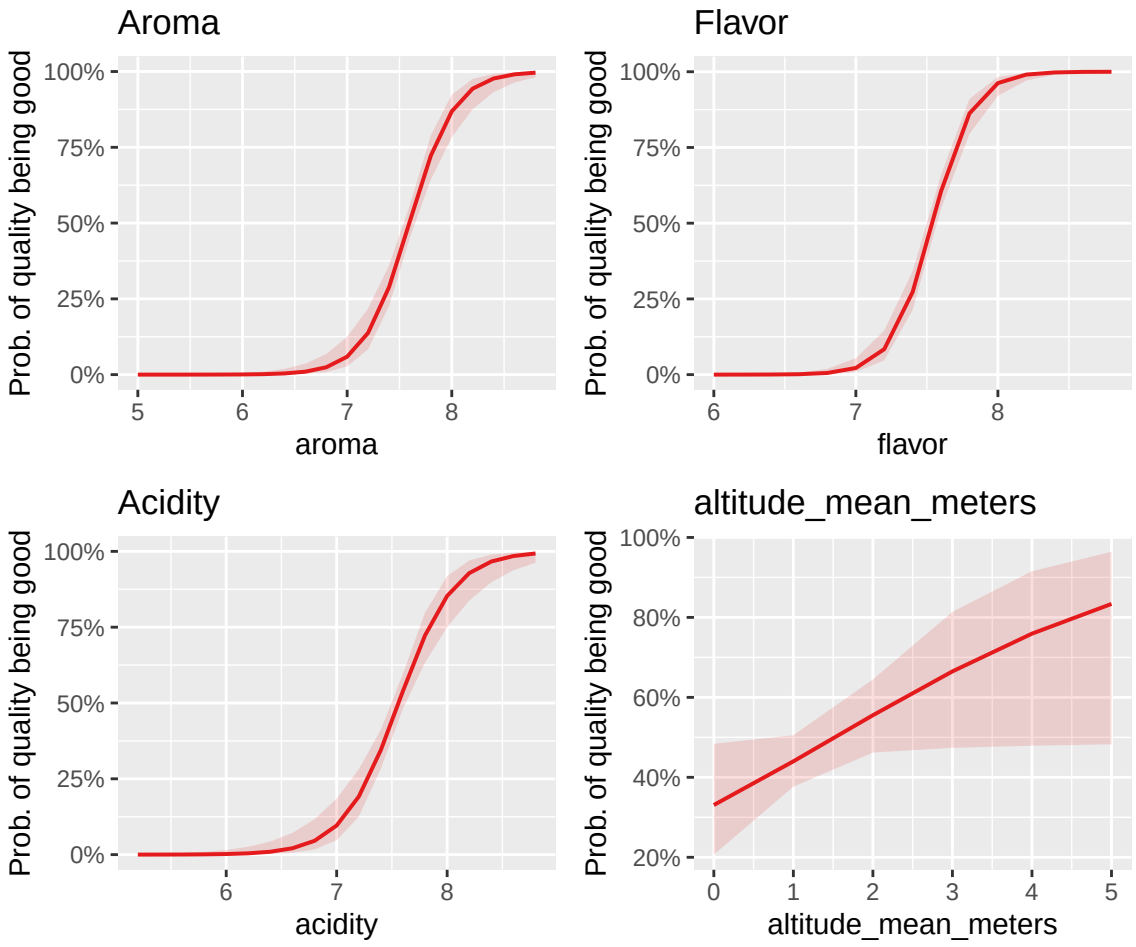

Figure 8: Probability of quality being good

The Figure 8 produce the estimated probabilities of coffee quality being good by aroma, flavor, acidity and altitude. In this four variables, except altitude have significant S-shaped nonlinear increase in predictive probability, with flavor having the most significant effect. Altitude had a weak linear positive correlation with predictive probability, but with wide confidence intervals, suggesting that its effect was unstable or with high uncertainty.

4 Conclusions

Among all evaluated factors, flavor, followed closely by aroma and acidity—are the strongest factors of coffee quality.These findings suggest prioritizing improvements in these areas could substantially enhance coffee quality.

The number of category-two defects and harvested almost have no effect on the quality classified of coffee in this model.

The altitude at which coffee is grown has a relatively modest effect on quality classifications in this model. In some ways, altitude may influence other factors like flavor, aroma and acidity, but it does not have a obvious influence on these factors. Other data are needed for this suspect.