Analyzing the influence of coffee characteristics on its quality classification

Hanwen Yuan, Zhujunyi Li, Chunyao Hou

School of Mathematics and Statistics, University of Glasgow, UK

Contents

1 Introduction

2 Exploratory Data Analysis

3 Formal Data Analysis

4 Conclusion

5 Future Work

1. Introduction

- Coffee quality assessment is a crucial aspect of the coffee industry, impacting both market value and consumer preferences. The study aims to explore the factors that influence whether a batch of coffee is classified as “Good” or “Poor”.

- Our 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 score threshold (82.5 as “Good”, <82.5 as “Poor”).

- The primary goal of this study is to analyze the relationship between coffee quality and factors mentioned above then decide which factors has significant influence in coffee quality classfication.

2. Exploratory Data Analysis

2.1 Data preparing and Cleaning

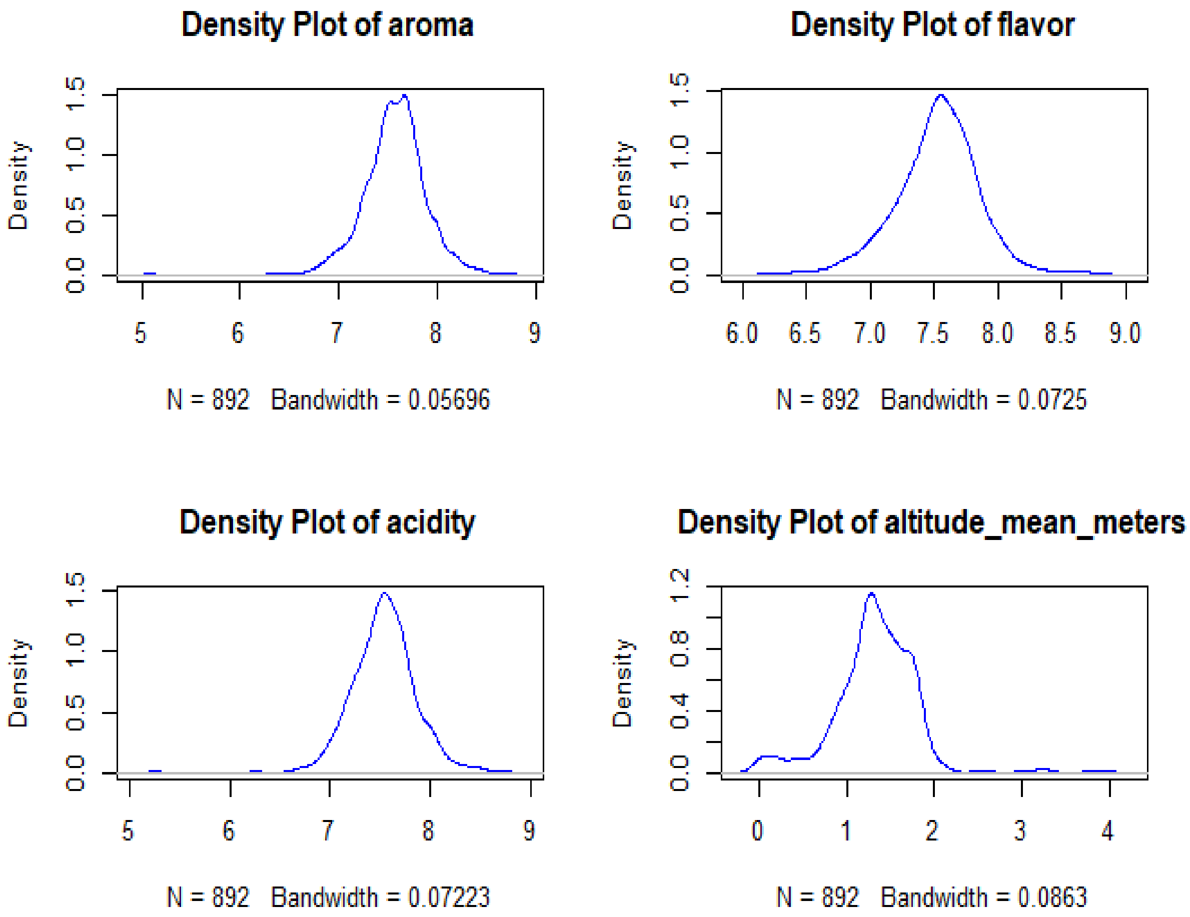


Figure 1: the density of explanatory variables

We check the data then found that data of altitude_mean_meters has **two outliers** which were higher than the highest mountain 8848 meters so we **remove** these two row.

2. Exploratory Data Analysis

2.2 Continuous Variables Summaries

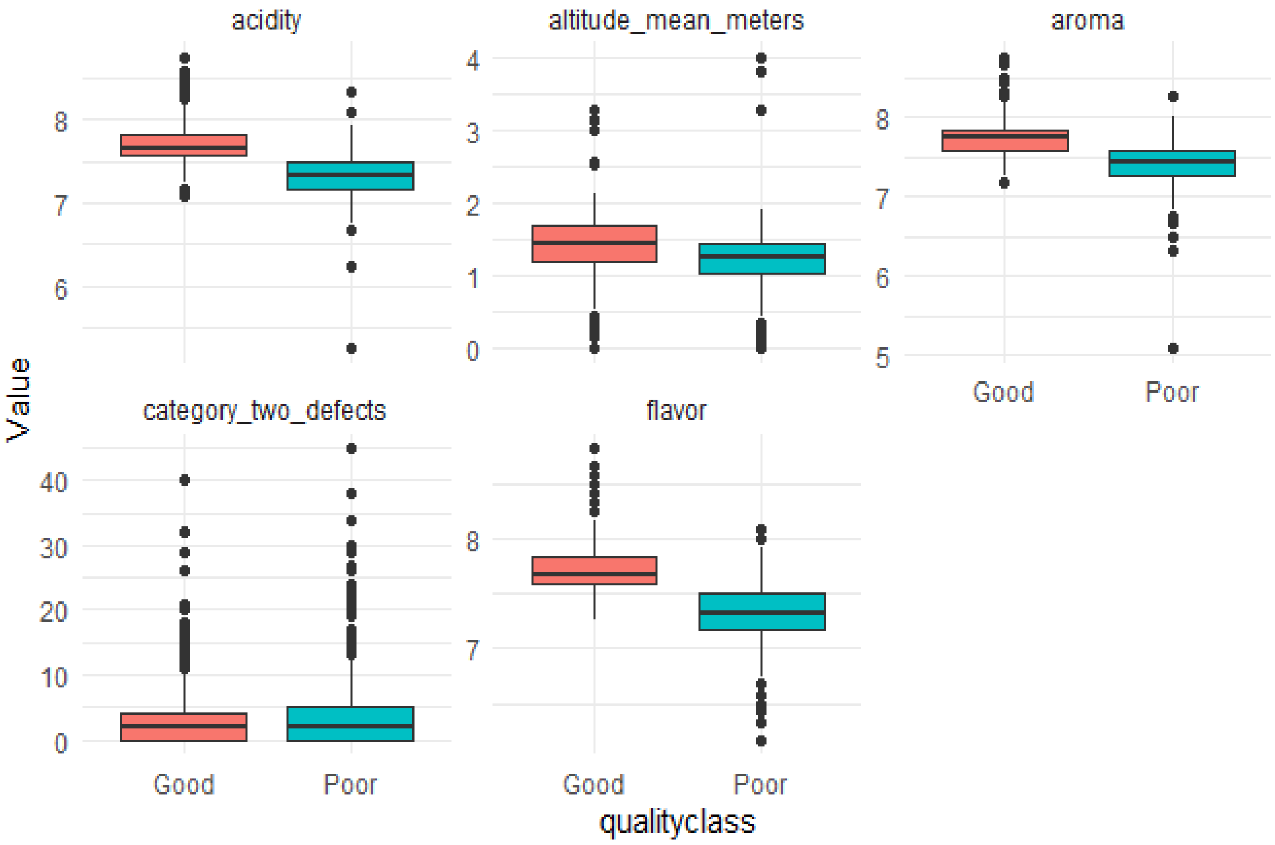


Figure 2: the boxplots of continuous explanatory variables

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

The boxplot and Table 1 show differences in **aroma**, **flavor**, **acidity, defects** and **altitude** between coffee quality classes from the graphic and digital side.

2. Exploratory Data Analysis

2.2 Categorical Variable Summaries

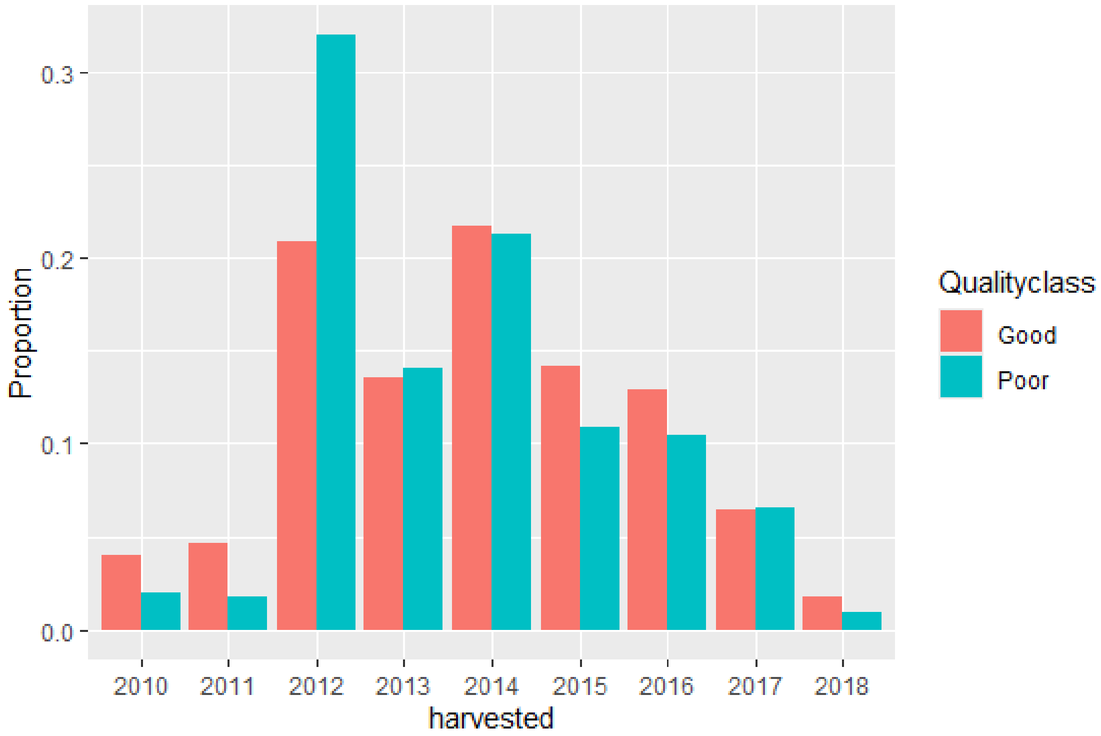
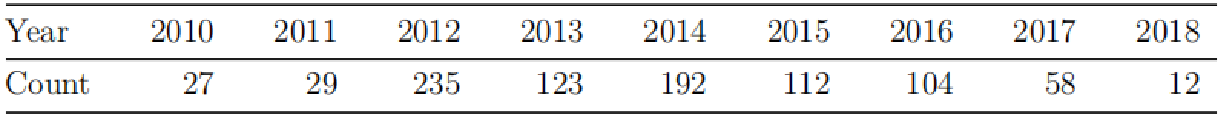


Figure 3: the barplot of categorical explanatory variable

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

- The barplot shows no clear trend, indicating that **harvested** might **not significantly affect** the classification.
- The table 2 shows the data of **harvested** are concentrated in **2012-2016**.

3. Formal Data Analysis

3.1 Model Creation

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

The function of harvested is:

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

$$\mathrm{I}_j(x)=\left\{\begin{matrix}1, & if\ group\ of\ harvested\ x\ is\ considered\ as\ j\ , \\ 0, & Otherwise\end{matrix}\right\}$$

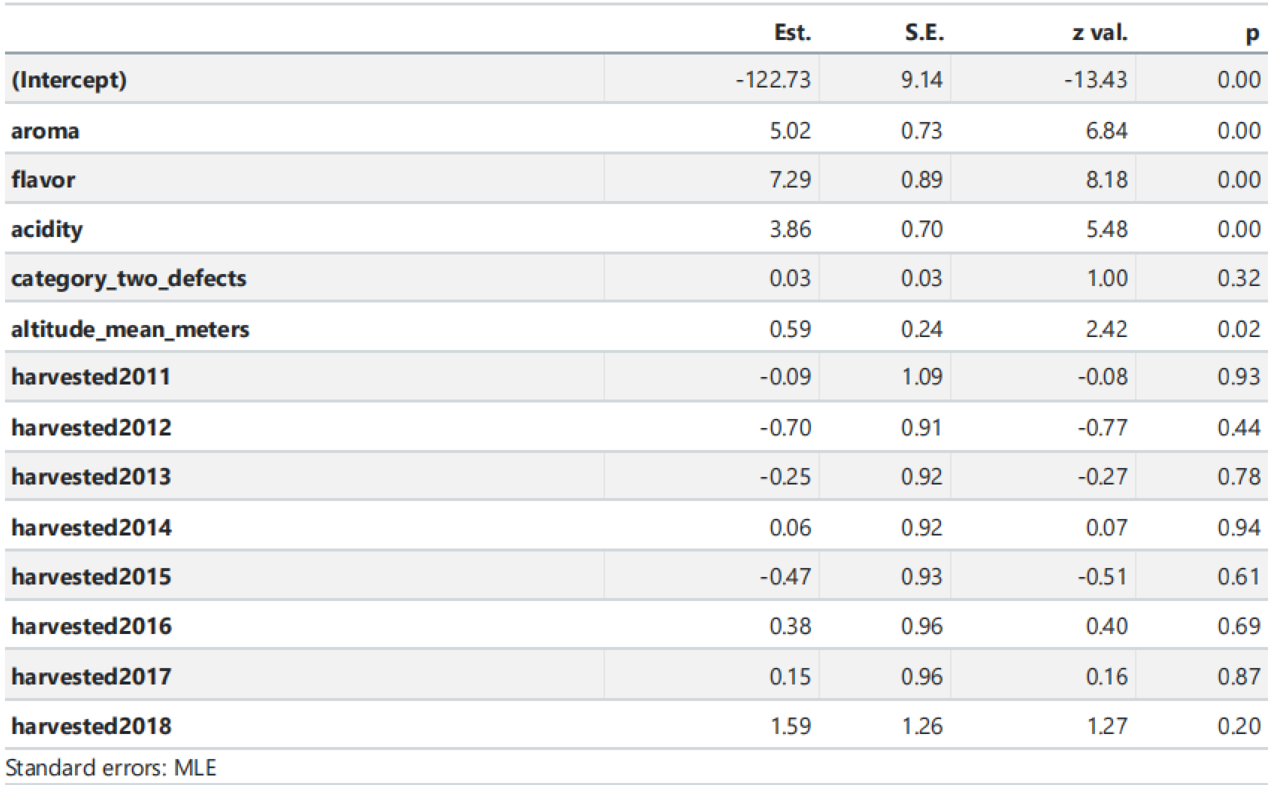
α: Represents the baseline log-odds when all explanatory variables are set to zero.

$\{\beta_i\}, i=1,\ldots,5$, are the coefficients . Which means when X_i increase 1, the probability will change according to the β_i.

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

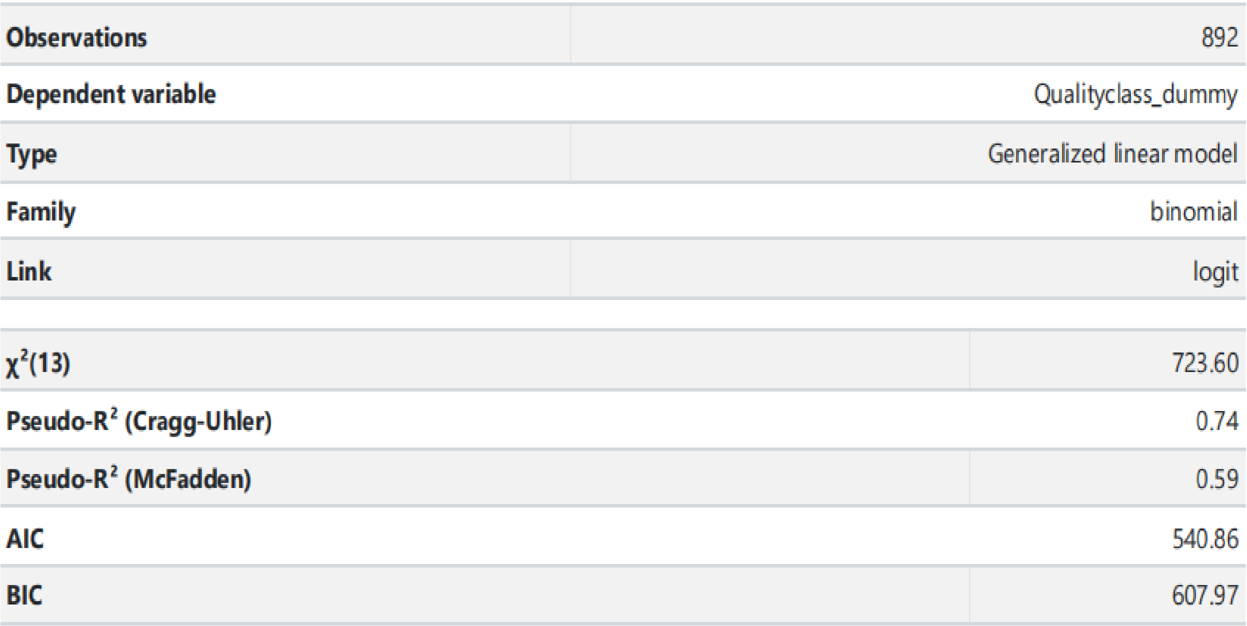
3. Formal Data Analysis

3.2 Model Fitted



	Est.	S.E.	z val.	p
(Intercept)	-122.73	9.14	-13.43	0.00
aroma	5.02	0.73	6.84	0.00
flavor	7.29	0.89	8.18	0.00
acidity	3.86	0.70	5.48	0.00
category_two_defects	0.03	0.03	1.00	0.32
altitude_mean_meters	0.59	0.24	2.42	0.02
harvested2011	-0.09	1.09	-0.08	0.93
harvested2012	-0.70	0.91	-0.77	0.44
harvested2013	-0.25	0.92	-0.27	0.78
harvested2014	0.06	0.92	0.07	0.94
harvested2015	-0.47	0.93	-0.51	0.61
harvested2016	0.38	0.96	0.40	0.69
harvested2017	0.15	0.96	0.16	0.87
harvested2018	1.59	1.26	1.27	0.20

Standard errors: MLE



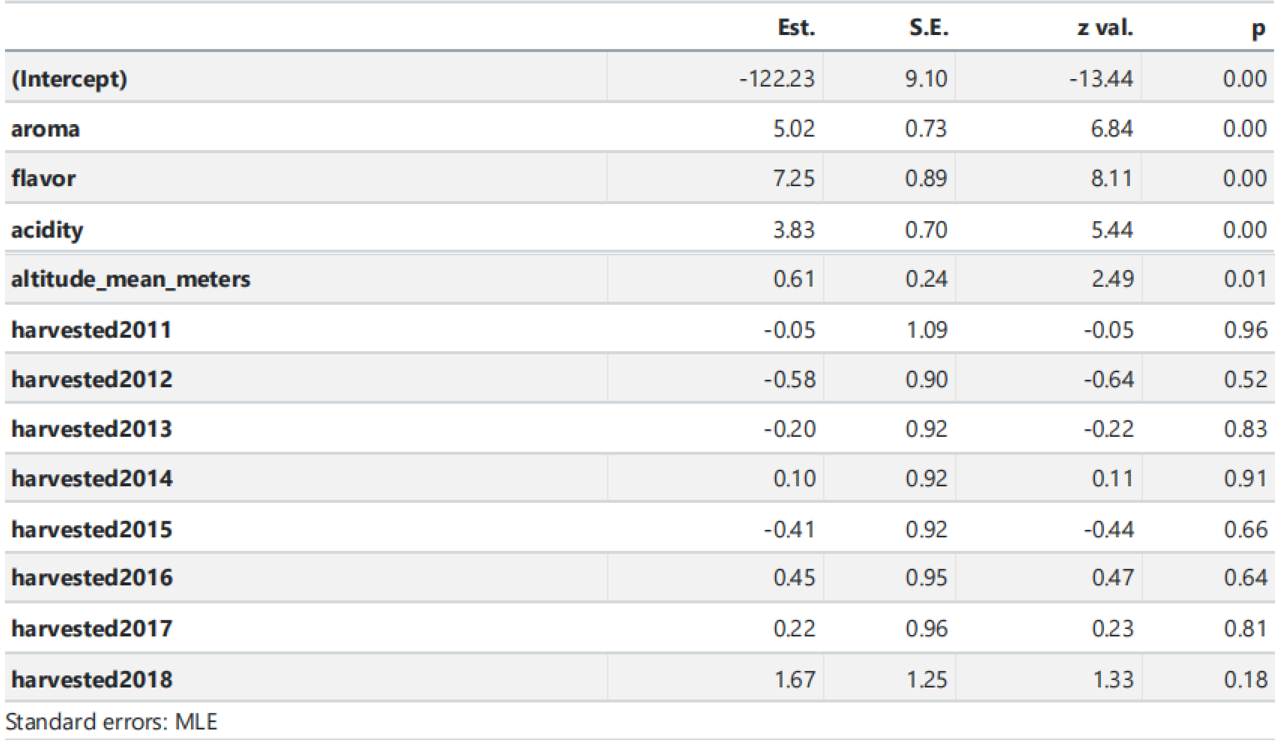
Observations	892
Dependent variable	Qualityclass_dummy
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(13)$	723.60
Pseudo-R^2 (Cragg-Uhler)	0.74
Pseudo-R^2 (McFadden)	0.59
AIC	540.86
BIC	607.97

The model 1 with variables **harvested, aroma**, **flavor**, **acidity, defects** and **altitude** shows the p-value of **defects** was **higher than 0.05**.

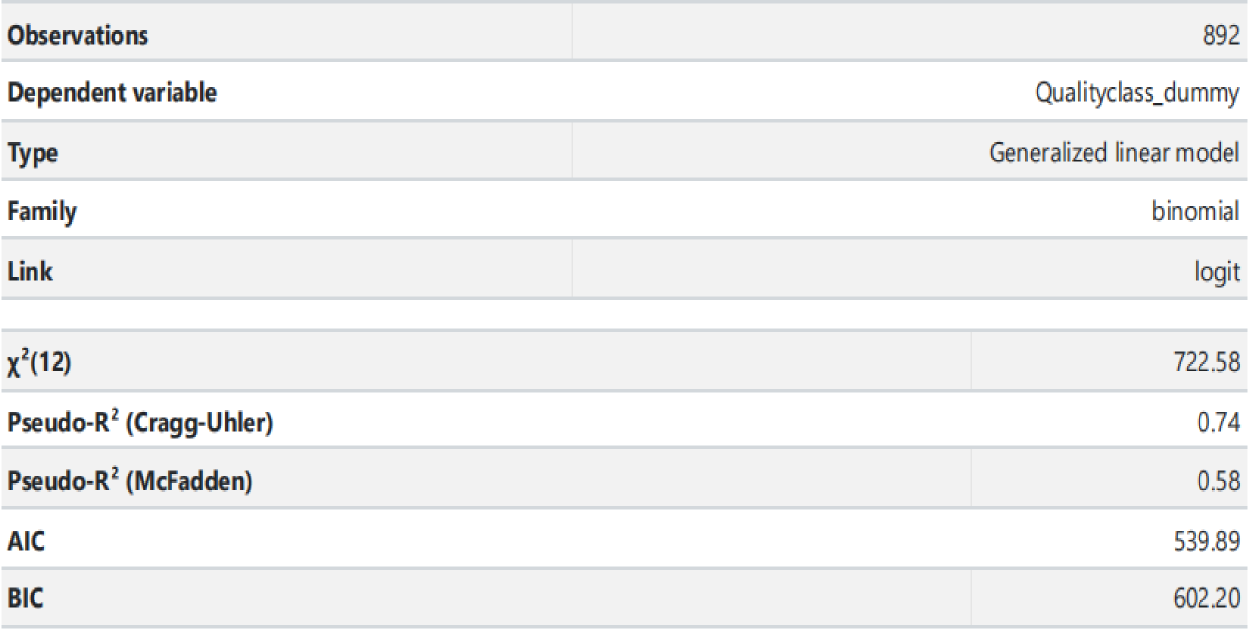
3. Formal Data Analysis

3.2 Model Fitted



	Est.	S.E.	z val.	p
(Intercept)	-122.23	9.10	-13.44	0.00
aroma	5.02	0.73	6.84	0.00
flavor	7.25	0.89	8.11	0.00
acidity	3.83	0.70	5.44	0.00
altitude_mean_meters	0.61	0.24	2.49	0.01
harvested2011	-0.05	1.09	-0.05	0.96
harvested2012	-0.58	0.90	-0.64	0.52
harvested2013	-0.20	0.92	-0.22	0.83
harvested2014	0.10	0.92	0.11	0.91
harvested2015	-0.41	0.92	-0.44	0.66
harvested2016	0.45	0.95	0.47	0.64
harvested2017	0.22	0.96	0.23	0.81
harvested2018	1.67	1.25	1.33	0.18

Standard errors: MLE



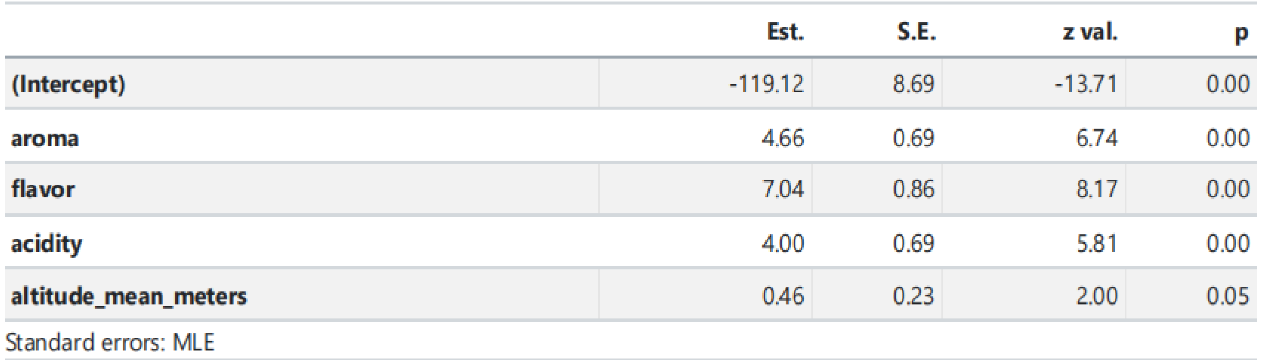
Observations	892
Dependent variable	Qualityclass_dummy
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(12)$	722.58
Pseudo-R^2 (Cragg-Uhler)	0.74
Pseudo-R^2 (McFadden)	0.58
AIC	539.89
BIC	602.20

The model 2 with 5 variables **harvested, aroma**, **flavor**, **acidity** and **altitude** shows the p-value of **harvested** was higher than 0.05. **AIC and BIC decreased.**

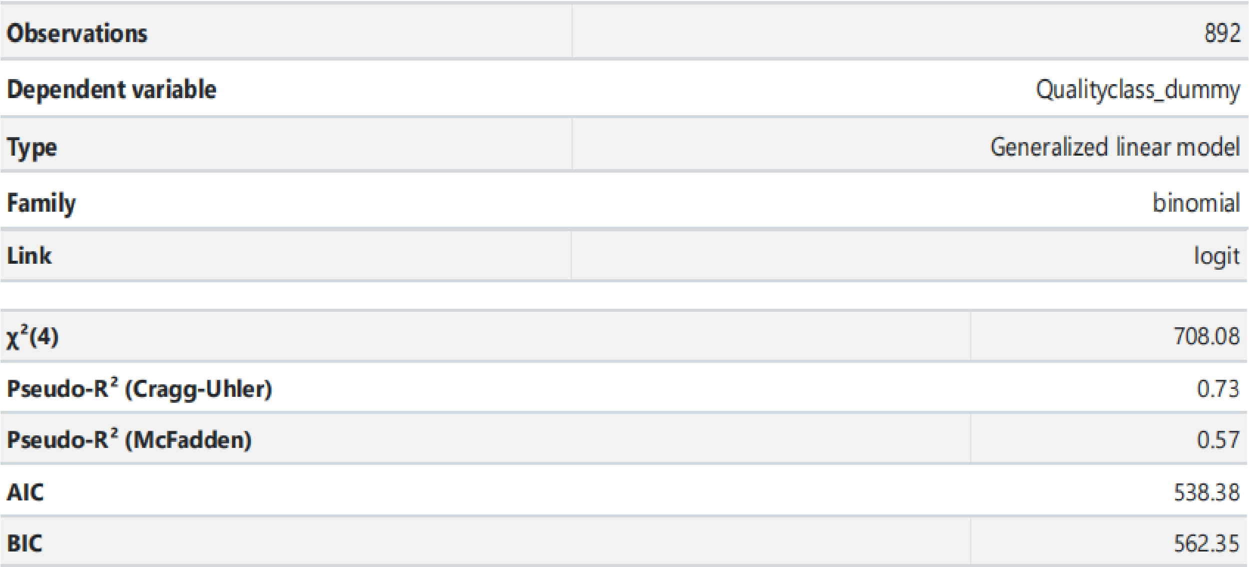
3. Formal Data Analysis

3.2 Model Fitted



	Est.	S.E.	z val.	p
(Intercept)	-119.12	8.69	-13.71	0.00
aroma	4.66	0.69	6.74	0.00
flavor	7.04	0.86	8.17	0.00
acidity	4.00	0.69	5.81	0.00
altitude_mean_meters	0.46	0.23	2.00	0.05

Standard errors: MLE



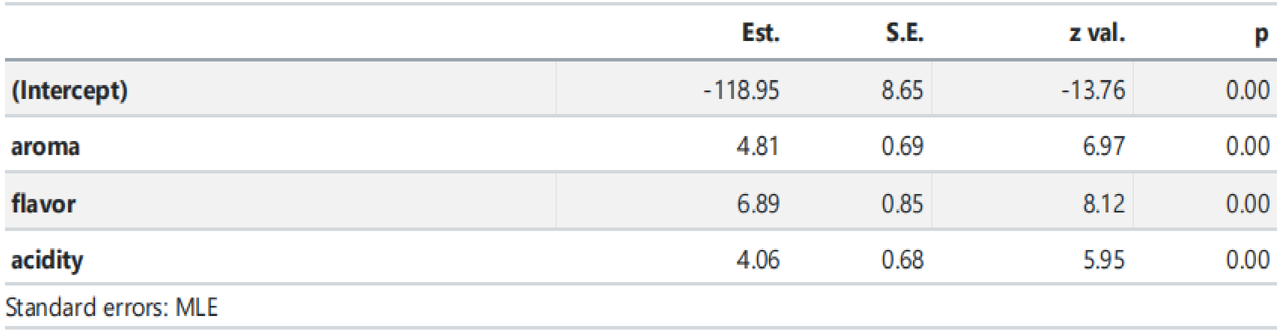
Observations	892
Dependent variable	Qualityclass_dummy
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(4)$	708.08
Pseudo-R^2 (Cragg-Uhler)	0.73
Pseudo-R^2 (McFadden)	0.57
AIC	538.38
BIC	562.35

The model 3 with 4 variables **aroma**, **flavor**, **acidity** and **altitude**
shows the p-value of all variables were **lower than 0.05**. **AIC and BIC decreased**.

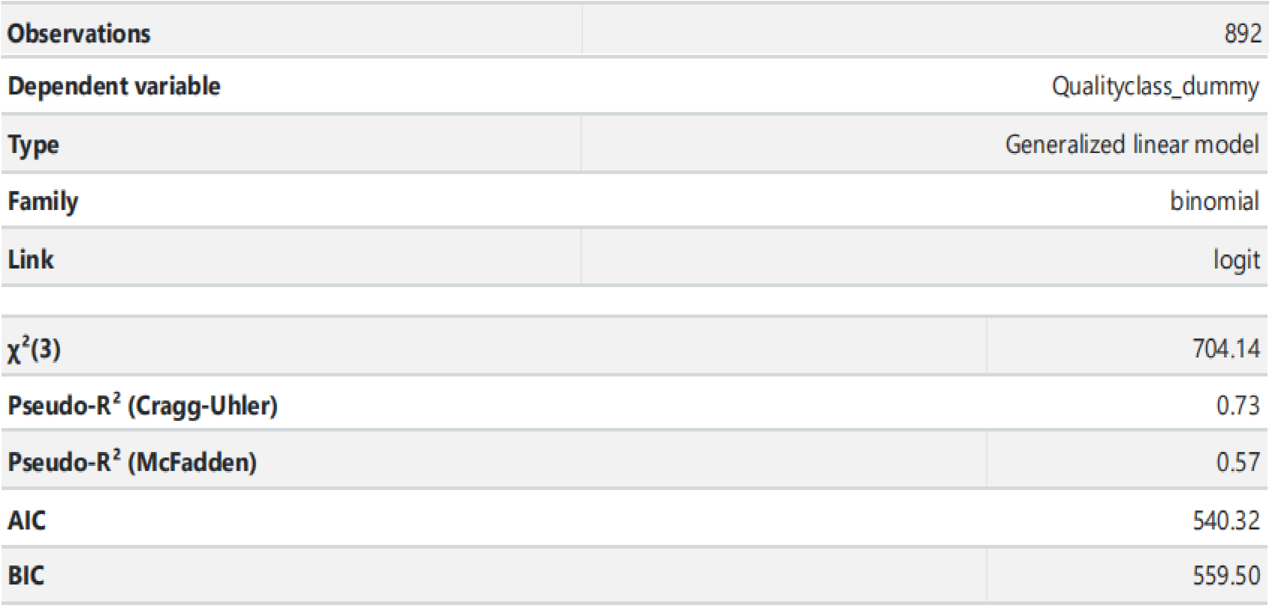
3. Formal Data Analysis

3.2 Model Fitted



	Est.	S.E.	z val.	p
(Intercept)	-118.95	8.65	-13.76	0.00
aroma	4.81	0.69	6.97	0.00
flavor	6.89	0.85	8.12	0.00
acidity	4.06	0.68	5.95	0.00

Standard errors: MLE



Observations	892
Dependent variable	Qualityclass_dummy
Type	Generalized linear model
Family	binomial
Link	logit

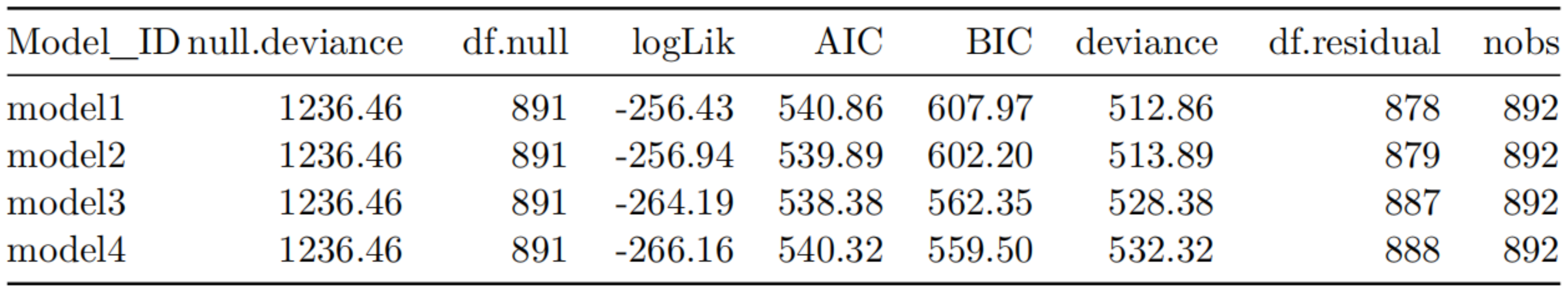
$\chi^2(3)$	704.14
Pseudo-R^2 (Cragg-Uhler)	0.73
Pseudo-R^2 (McFadden)	0.57
AIC	540.32
BIC	559.50

The model 4 with 3 variables **aroma**, **flavor** and **acidity**
shows the p-value of **harvested** was higher than 0.05. **AIC and BIC decreased.**

3. Formal Data Analysis

3.2 Model Fitted

Table 3: Comparison of all model



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 using hypothesis testing, we choose **Model 3**.

Assumption:

M0=model4, M1=model 3;

D0=532.32, q=3; D1=528.38, p=4;

D0-D1=532.32-528.38=3.94;

Chi-square(p-q=4-3=1)=3.84; It has 3.94>3.84, then reject H0.

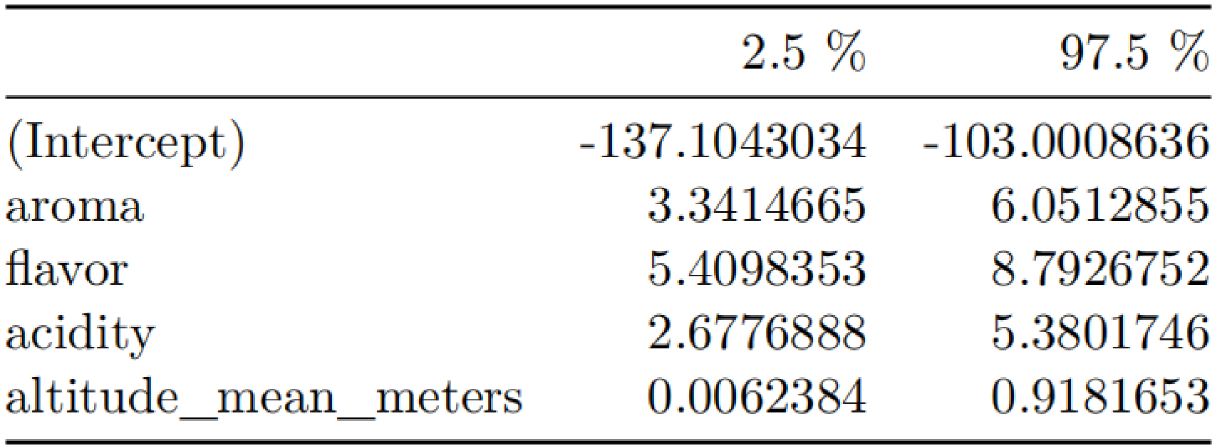
3. Formal Data Analysis

3.2 Model Fitted – Model 3

The function of Model 3 is:

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

Table 4: 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

According CI, it shows **not include 0**. All variables in this model is significant for coffee quality.

3. Formal Data Analysis

3.3 Log-odds

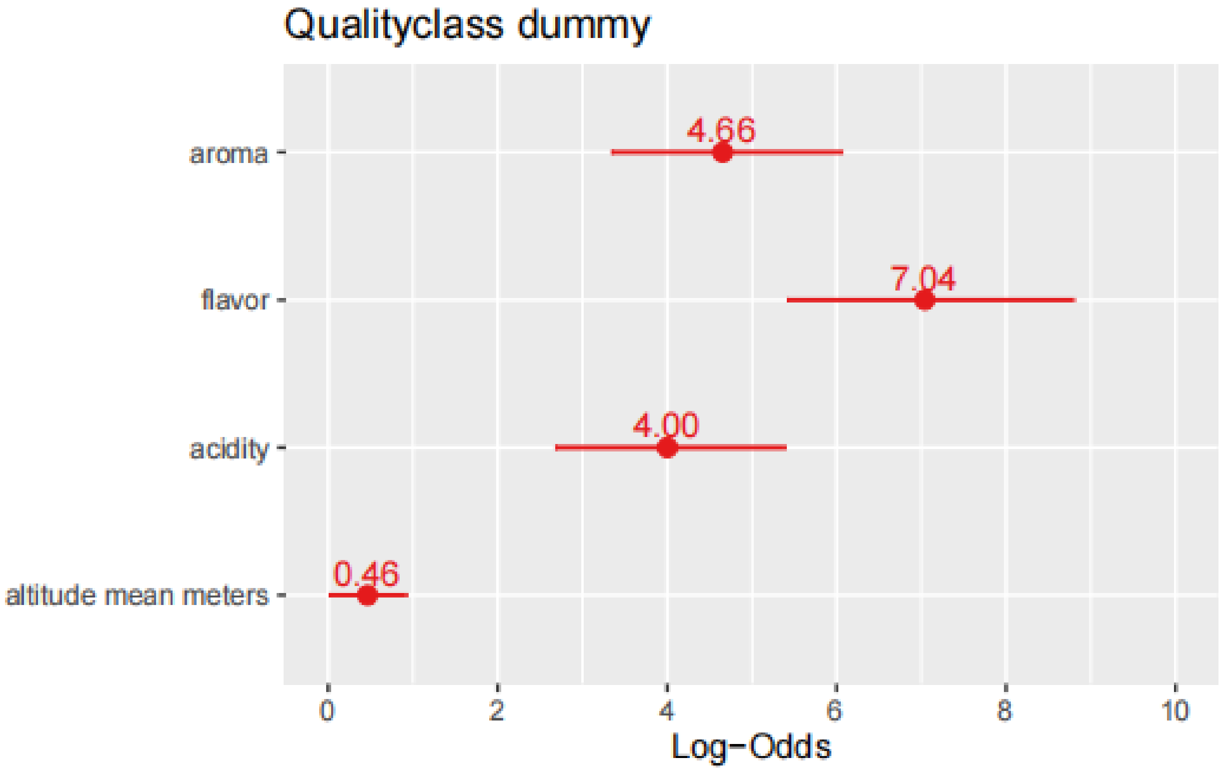


Figure 3: log-odds of explanatory variables for quality good

According to the Log-odds plot:

- All four variables retained had a **significant positive effect.**
- The increase in **flavor** leads to a significant increase in the probability that the coffee quality to be “good”.
- **Altitude** has a small positive effect on coffee quality to be “good”.

3. Formal Data Analysis

3.3 Odds

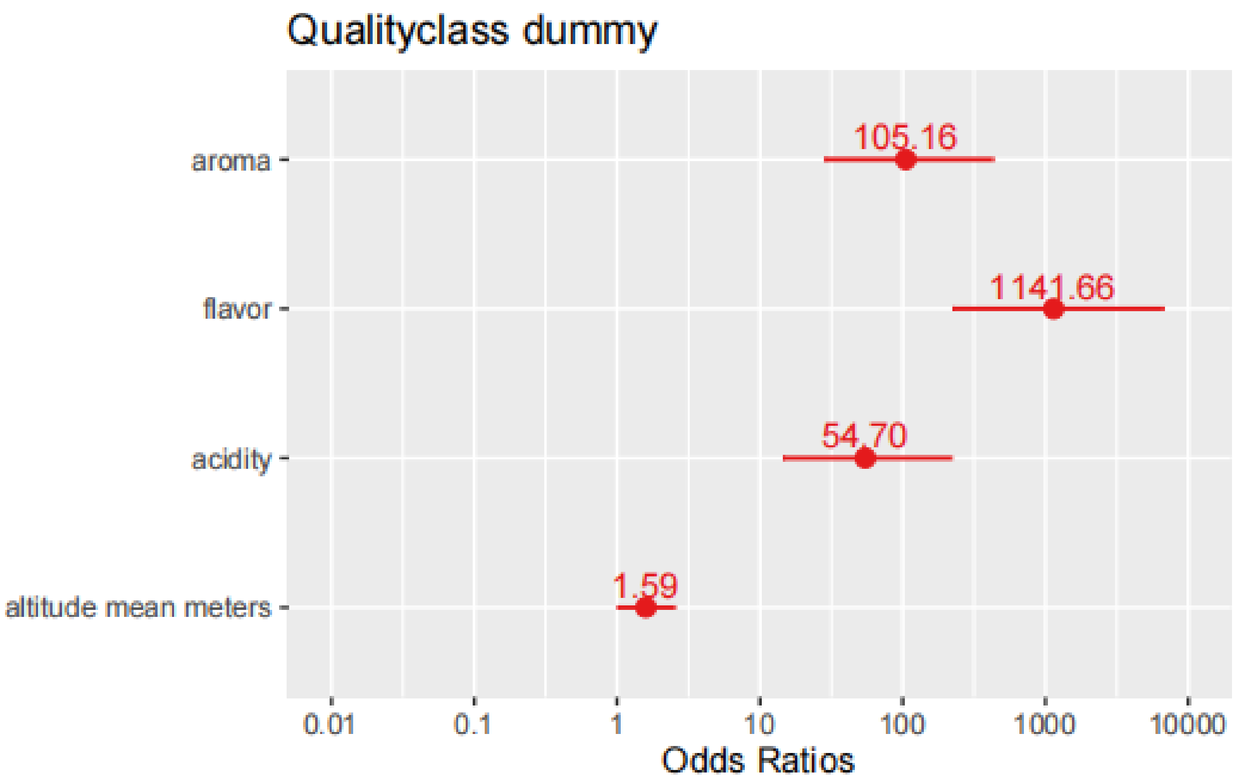


Figure 4: the odds of explanatory variables for quality good

According to the Odds plot:

- Each unit increase in **flavor** and **aroma** increases the chances of a coffee being rated “good” by hundreds or even thousands of times.
- **Acidity** and **altitude** having a slightly weaker but not negligible effect than flavor.

3. Formal Data Analysis

3.4 Probability Trends

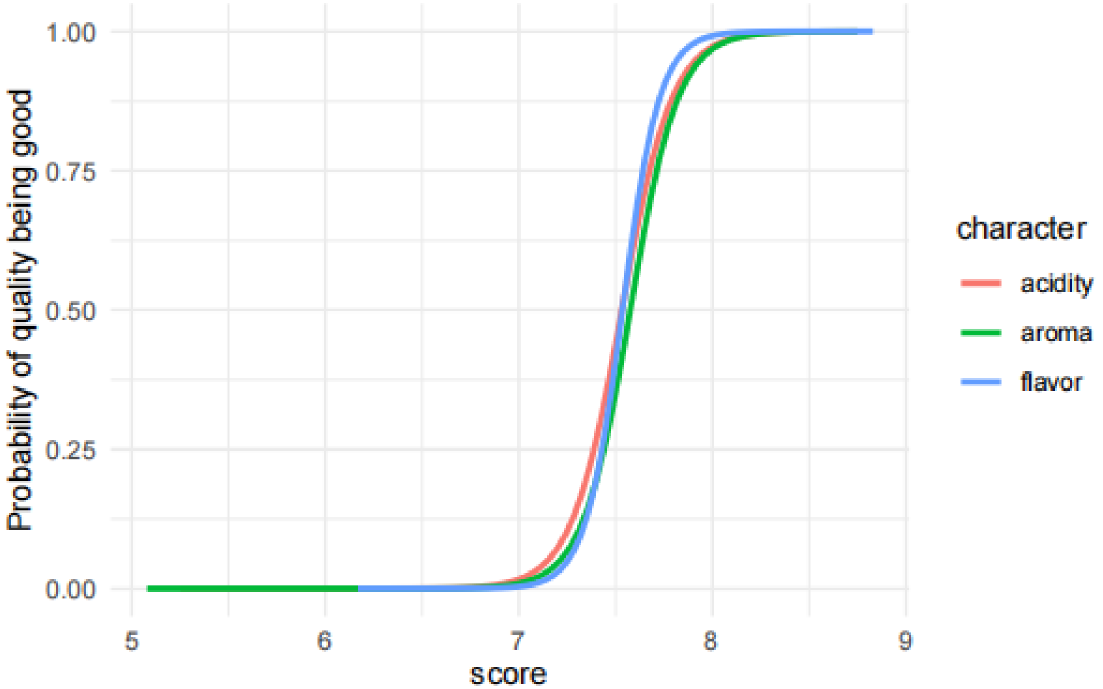


Figure 5: the prob of aroma/acidity/flavor for quality good

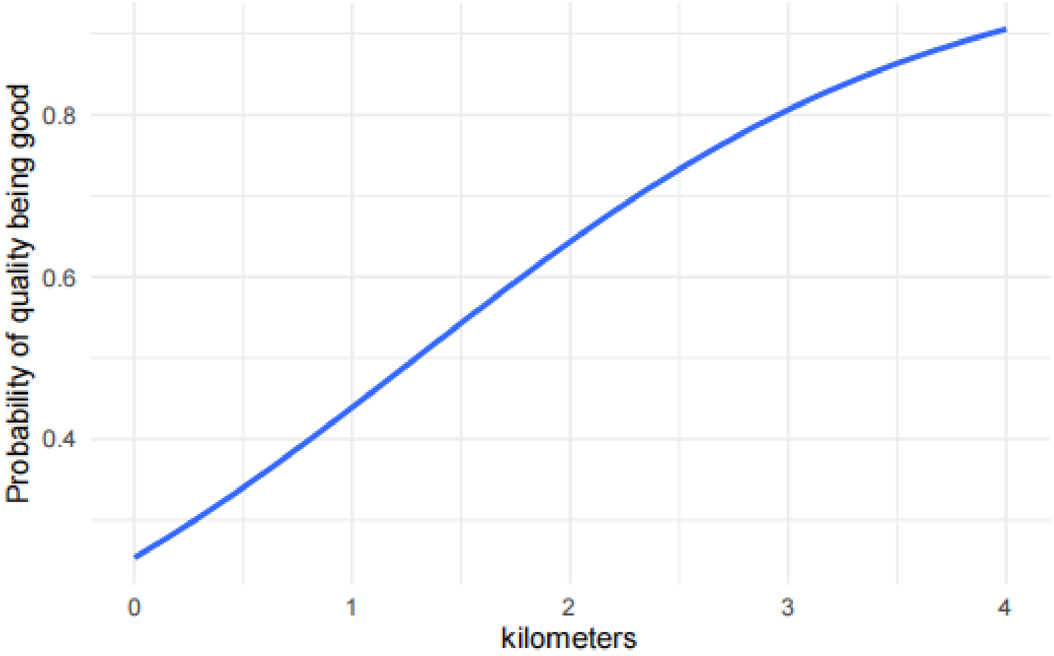


Figure 6: the prob of altitude for quality good

Figure 5 shows:

- all continuous variables positively impact the probability of coffee being classified as ‘Good’.
- **flavor has the steepest curve**, indicating that improvements in flavor score have the most substantial effect on quality classification.

Figure 6 illustrates:

- higher **altitudes** are associated with a higher probability of being classified as ‘Good’.
- The positive trend becomes more pronounced between **1km and 3km.**

3. Formal Data Analysis

3.4 Probability Trends

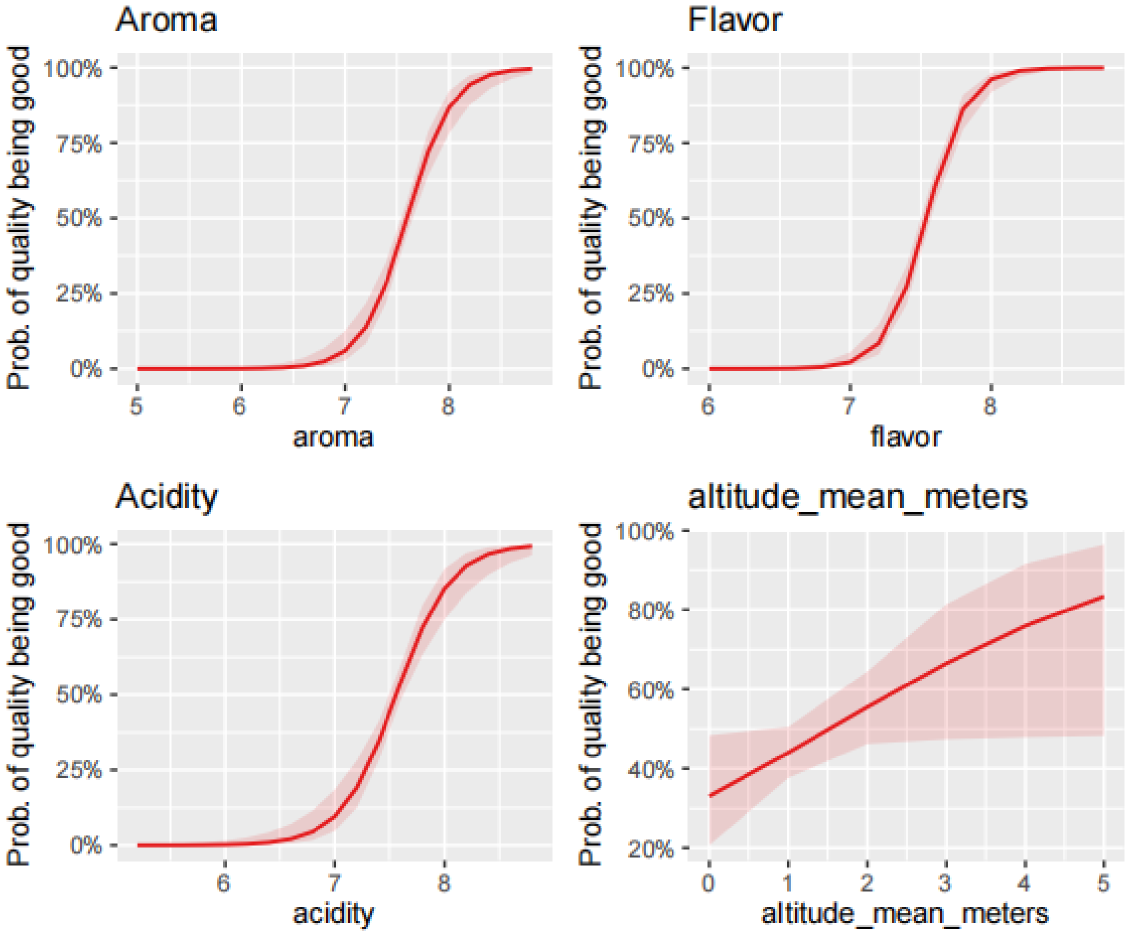


Figure 7: Probability of quality being good

- **aroma**, **flavor** and **acidity** have significant S-shaped nonlinear increase in predictive probability.
- **Altitudes** has large confidence intervals indicating unstable or high uncertainty in its effects.

4. Conclusion

- 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 **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.

5. Future Work

- **Limited Variable Scope**: This study focused only on a subset of available variables, including sensory scores (aroma, flavor, acidity), average altitude, defect counts, and harvest year. However, other potentially influential factors such as **coffee variety, processing methods,** or **climatic conditions** were not included, which may lead to different results.

- Another potential limitation is **multicollinearity** among the sensory attributes (aroma, flavor, acidity), which could affect the stability of coefficient estimates.

- Additionally, this study did not account for **potential interaction effects** between variables. For example, the combined influence of aroma and flavor might have a synergistic effect on coffee quality, which could be explored in future models by including interaction terms.

Thank you

Group 11 members:

Hanwen Yuan, Zhujunyi Li, Chunyao Hou, Wei Li, Congle Wang