Investigating Influential Features on Coffee Quality

Yuqi Pan，Jin-an Wu，Shuqi Cao，Blair Watson，and Shenghan Gao
School of Mathematics and Statistics, University of Glasgow, UK

Content

01 Introduction

02 Explanatory Data Analysis

03 Formal Analysis

04 Conclusion

05 Further Extension

06 Reference

PART.01 Introduction

1.1 Research question

What influence do different features of coffee have on whether the quality of a batch of coffee is classified as good or poor?

University of Glasgow

1.2 Data description

Overview of the dataset

	country_of_origin <chr>	aroma <dbl>	flavor <dbl>	acidity <dbl>	category_two_defects <int>	altitude_mean_meters <dbl>
1	Myanmar	7.25	7.42	7.50	4	1219.20
2	Uganda	8.33	7.92	7.92	1	1600.00
3	Ethiopia	8.42	8.00	8.00	7	1700.00
4	Mexico	7.17	7.08	7.25	3	1300.00
5	Burundi	7.75	7.67	7.50	5	1880.00
6	Tanzania, United Republic Of	7.92	7.75	7.75	0	1400.00
7	Colombia	7.92	7.83	7.67	1	NA
8	Colombia	7.83	7.67	7.58	2	1775.00
9	Guatemala	7.00	6.83	7.17	2	1310.64
10	Colombia	7.33	7.33	7.50	1	1900.00

1.2 Data description

- **Response variable：**
  - qualityclass

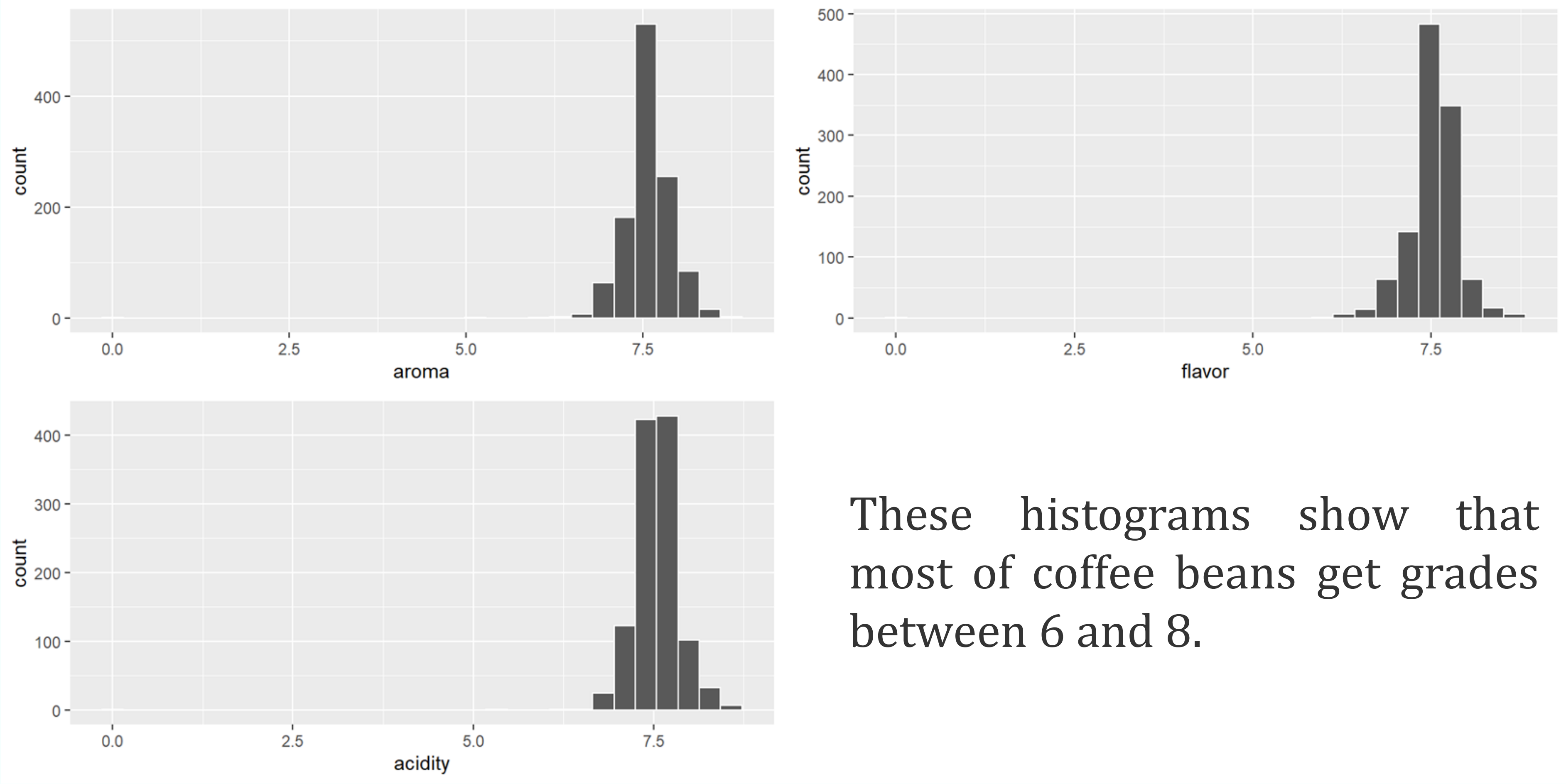
- **Explanatory variables：**
  - country_of_origin
  - aroma
  - flavor
  - acidity
  - category_two_defects
  - altitiude_mean_meters
  - harvested

PART.02 Explanatory Data Analysis

2.1 Data summarization

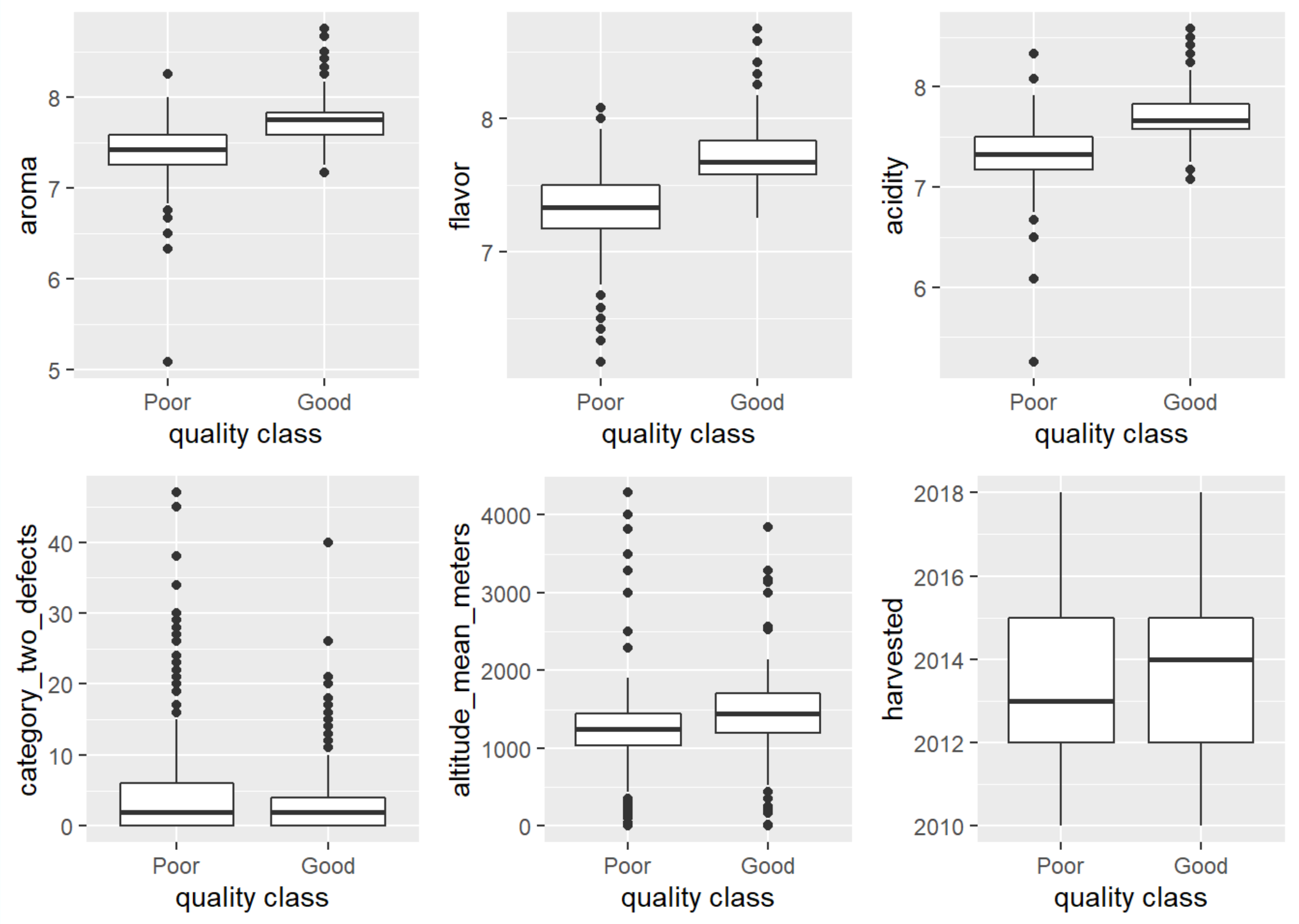
Variables	Missing	Mean	SD	Min	Median	Max
aroma	0	7.57	0.39	0	7.58	8.75
flavor	0	7.52	0.40	0	7.58	8.67
acidity	0	7.54	0.39	0	7.50	8.58
category_two_defects	0	3.67	5.41	0	2.00	55.00
altitude_mean_meters	201	1850.69	9392.09	1	1310.64	190164.00
harvested	60	2013.67	1.81	2010	2014.00	2018.00

2.1 Data summarization

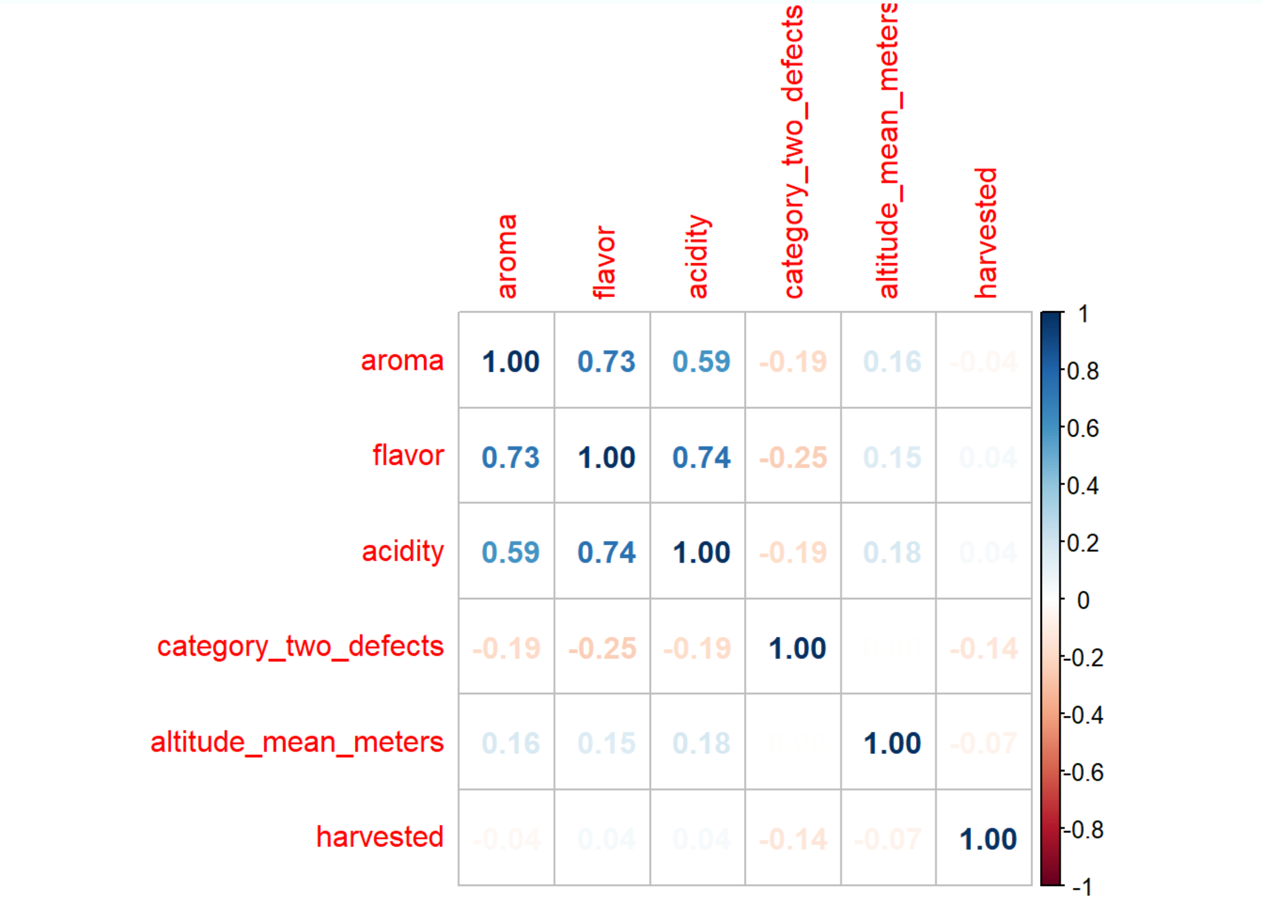


These histograms show that most of coffee beans get grades between 6 and 8.

Boxplots of Quality class against the other variables



Correlation plot of all numerical variables



	aroma	flavor	acidity	category_two_defects	altitude_mean_meters	harvested
aroma	1.00	0.73	0.59	-0.19	0.16	-0.04
flavor	0.73	1.00	0.74	-0.25	0.15	0.04
acidity	0.59	0.74	1.00	-0.19	0.18	0.04
category_two_defects	-0.19	-0.25	-0.19	1.00	0.00	-0.14
altitude_mean_meters	0.16	0.15	0.18	0.00	1.00	-0.07
harvested	-0.04	0.04	0.04	-0.14	-0.07	1.00

PART.03 Formal Analysis

3 Formal Analysis

Generalized linear model

$$y_i \sim Bin(1, p_i)$$

$$g(p_i) = log(\frac{pi}{1-pi}) = \alpha + \sum_{i=1}^{n} \beta_i x_i$$

3.1 Multicollinearity

Variance inflation factor (VIF>10)

	VIF
aroma	1.042
flavor	1.067
acidity	1.033
category_two_defects	1.012
altitude_mean_meters	1.037
harvested	1.053

3.1 Multicollinearity

Variance inflation factor (VIF>10)

	VIF
aroma	1.042
flavor	1.067
acidity	1.033
category_two_defects	1.012
altitude_mean_meters	1.037
harvested	1.053

3.2 Log-odds

model 1

	Est.	S.E.	z val.	p
(Intercept)	-283.59	118.31	-2.40	0.02
aroma	4.66	0.69	6.77	0.00
flavor	7.20	0.85	8.47	0.00
acidity	4.21	0.67	6.25	0.00
category_two_defects	0.00	0.03	0.14	0.89
altitude_mean_meters	0.00	0.00	2.68	0.01
harvested	0.08	0.06	1.38	0.17

3.2 Log-odds

model 2

	Est.	S.E.	z val.	p
(Intercept)	-282.39	117.98	-2.39	0.02
aroma	4.66	0.69	6.77	0.00
flavor	7.20	0.85	8.48	0.00
acidity	4.20	0.67	6.25	0.00
altitude_mean_meters	0.00	0.00	2.69	0.01
harvested	0.08	0.06	1.37	0.17

3.2 Log-odds

model 3

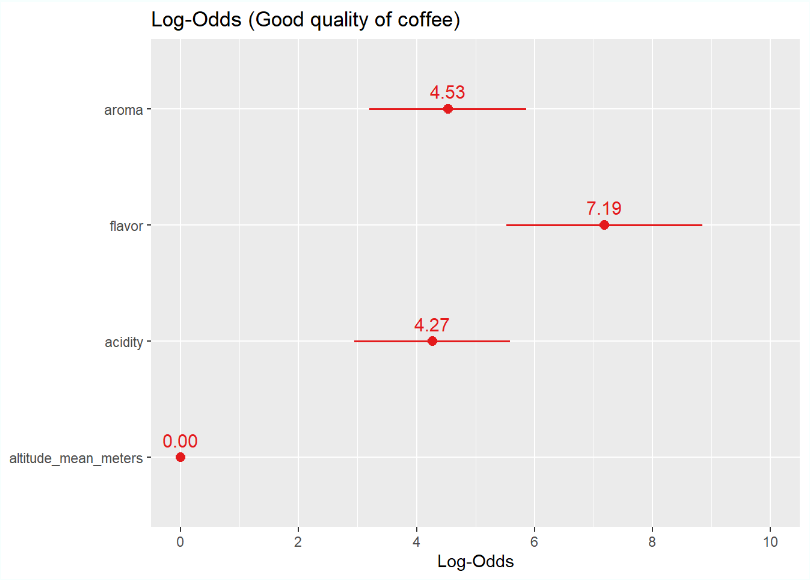
	Est.	S.E.	z val.	p
(Intercept)	-121.42	8.66	-14.02	0.00
aroma	4.53	0.68	6.67	0.00
flavor	7.19	0.85	8.48	0.00
acidity	4.27	0.67	6.33	0.00
altitude_mean_meters	0.00	0.00	2.53	0.01

The bound of 95% confidence interval for altitude_mean_meters is almost zero

3.2 Log-odds

95% confidence interval for log-odds

	2.5 %	97.5 %
(Intercept)	-139.331	-105.328
aroma	3.238	5.906
flavor	5.587	8.914
acidity	2.971	5.617
altitude_mean_meters	0.000	0.001



The bound of 95% confidence interval for altitude_mean_meters is almost zero.

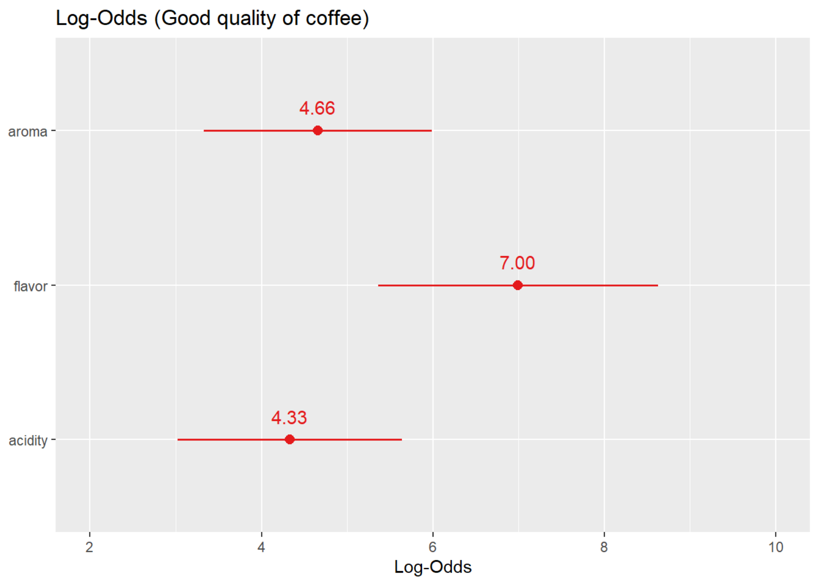
model 4

	Est.	**S.E.**	**z val.**	**p**
(Intercept)	-120.66	8.54	-14.12	0.00
aroma	4.66	0.68	6.87	0.00
flavor	7.00	0.83	8.39	0.00
acidity	4.33	0.67	6.50	0.00

3.2 Log-odds

95% confidence interval for log-odds

	2.5 %	97.5 %
(Intercept)	-138.309	-104.773
aroma	3.363	6.025
flavor	5.418	8.692
acidity	3.050	5.667



3.3 Model selection

Model comparison values for different models

model	AIC	BIC
GLM1	562.357	596.203
GLM2	560.376	589.387
GLM3	560.263	584.439
GLM4	564.616	583.957

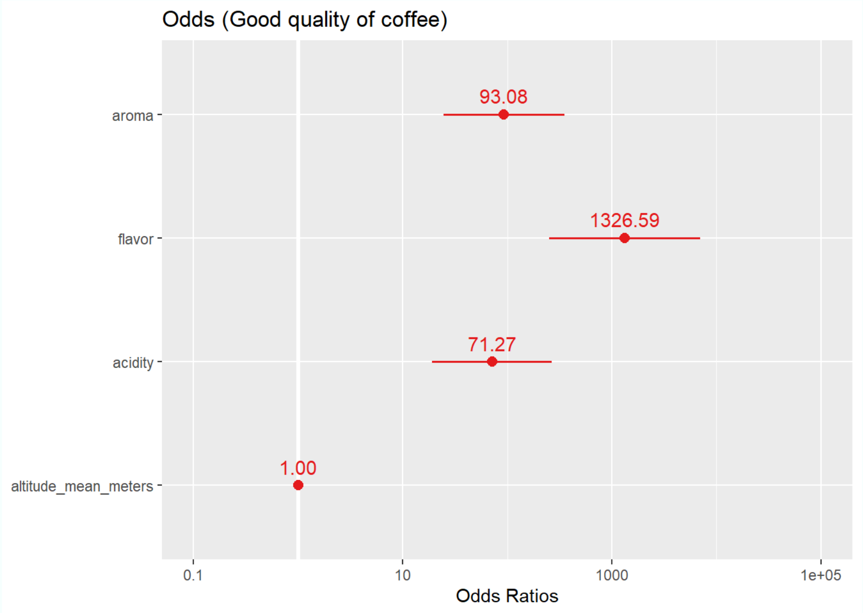
Final model on the log-odds scale

$$\log(\frac{p}{1-p}) = -121.42 + 4.53 \cdot aroma + 7.19 \cdot flavor + 4.27 \cdot acidity + 0.0005 \cdot altitude$$

3.4 Odds

Odds (Good quality of coffee)

	Odds
(Intercept)	0.000
aroma	93.075
flavor	1326.589
acidity	71.268
altitude_mean_meters	1.001



$$\frac{p}{1-p} = \exp(-121.42 + 4.53 \bullet aroma + 7.19 \bullet flavor + 4.27 \bullet acidity + 0.0005 \bullet altitude)$$

3.5 Probabilities

Probability formula

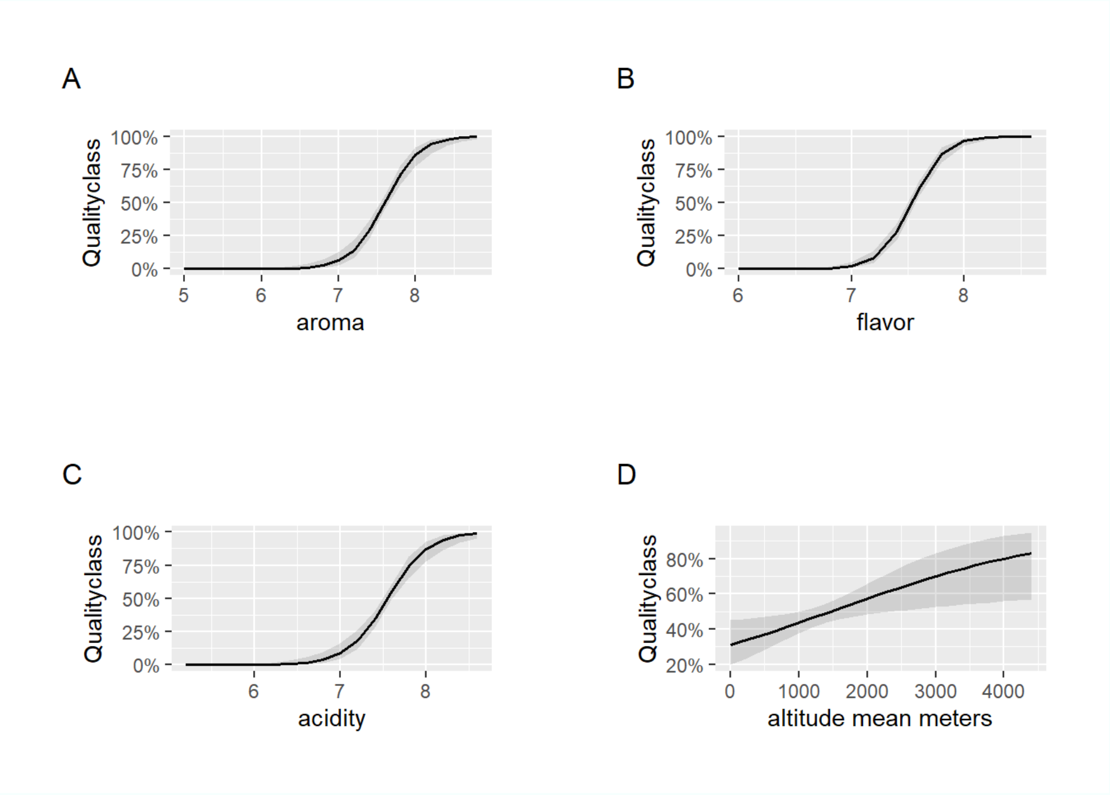
$$p=\frac{\exp(-121.42+4.53\bullet \mathit{aroma}+7.19\bullet \mathit{flavor}+4.27\bullet \mathit{acidity}+0.0005\bullet \mathit{altitude})}{1+\exp(-121.42+4.53\bullet \mathit{aroma}+7.19\bullet \mathit{flavor}+4.27\bullet \mathit{acidity}+0.0005\bullet \mathit{altitude})}$$

An example

$$p=\frac{\mathit{exp}(-121.42+4.53\bullet 8.3+7.19\bullet 7.9+4.27\bullet 7.3+0.0005\bullet 1700)}{1+\mathit{exp}(-121.42+4.53\bullet 8.3+7.19\bullet 7.9+4.27\bullet 7.3+0.0005\bullet 1700)}=0.993.$$

3.5 Probabilities

Probability of being good quality of coffee beans



The probability approaches 100% the larger the explanatory variables get, and approaches 0% the smaller the explanatory variables get.

PART.04 Conclusion

5 Conclusion

- Choose model 3 as the final model.
- The main three factors affecting the quality of coffee are aroma, flavor and acidity .
- Flavor is the most influential factor.

PART.05 Further Extension

6 Further Extension

- Delve into the causes of missing values

- Further work of this data

e.g., looking at the PH of the soil and how tall the plant grew.

PART.06 Reference

6 Reference

[1] Kutner, M. H.; Nachtsheim, C. J.; Neter, J. (2004). Applied Linear Regression Models (4th ed.). McGraw-Hill Irwin.

[2] ccs-amsterdam/r-course-material. GitHub. (2021). Retrieved 17 July 2021, from https://github.com/ccs-amsterdam/r-course-material/blob/master/tutorials/advanced_modeling.md#multilevel-models-or-mixed-effects-models.