

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



University
of Glasgow

Content

01 Introduction

02 Explanatory Data Analysis

03 Formal Analysis

04 Conclusion

05 Further Extension

06 Reference



University
of Glasgow



PART.01

Introduction



University
of Glasgow

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

- **Response variable:**

- qualityclass

- **Explanatory variables:**

- country_of_origin
- aroma
- flavor
- acidity
- category_two_defects
- altitude_mean_meters
- harvested



University
of Glasgow



PART.02

Explanatory Data Analysis



University
of Glasgow

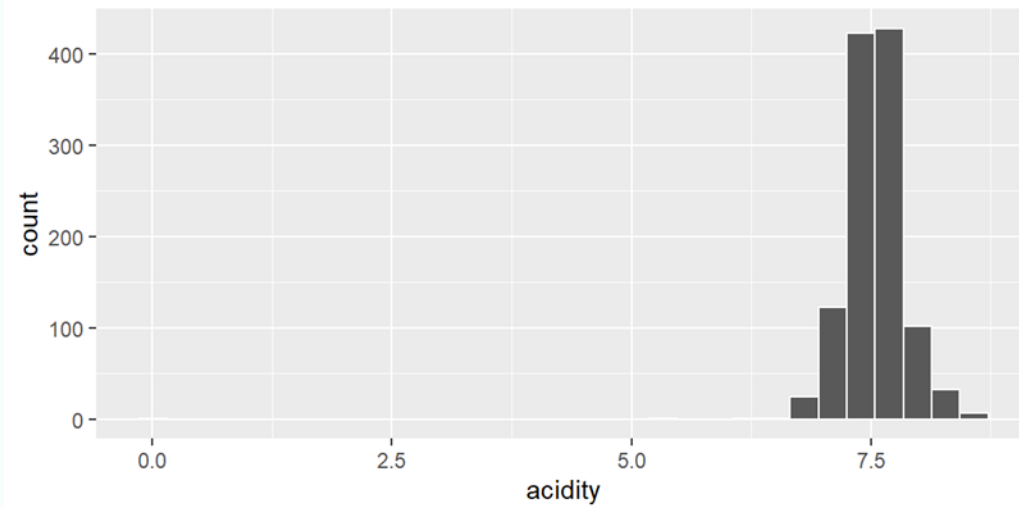
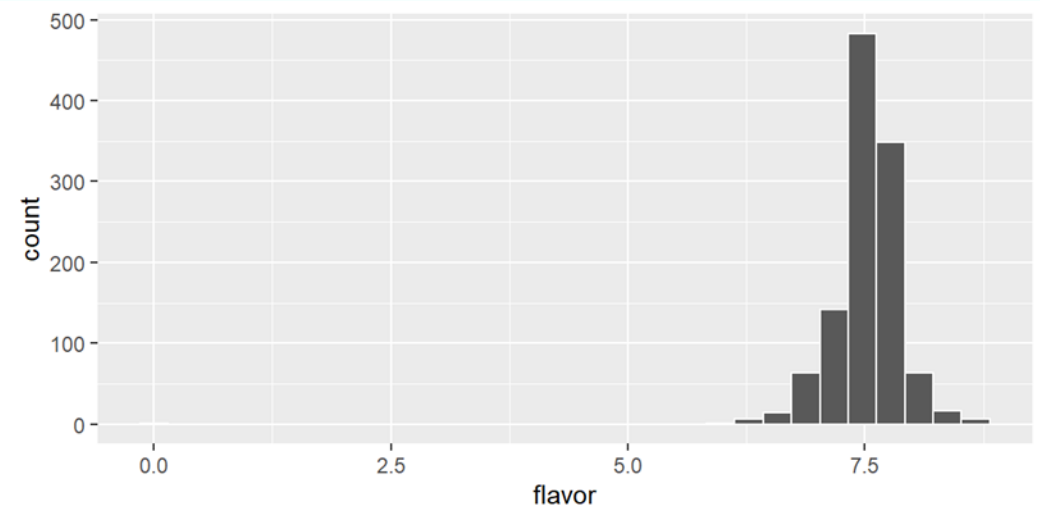
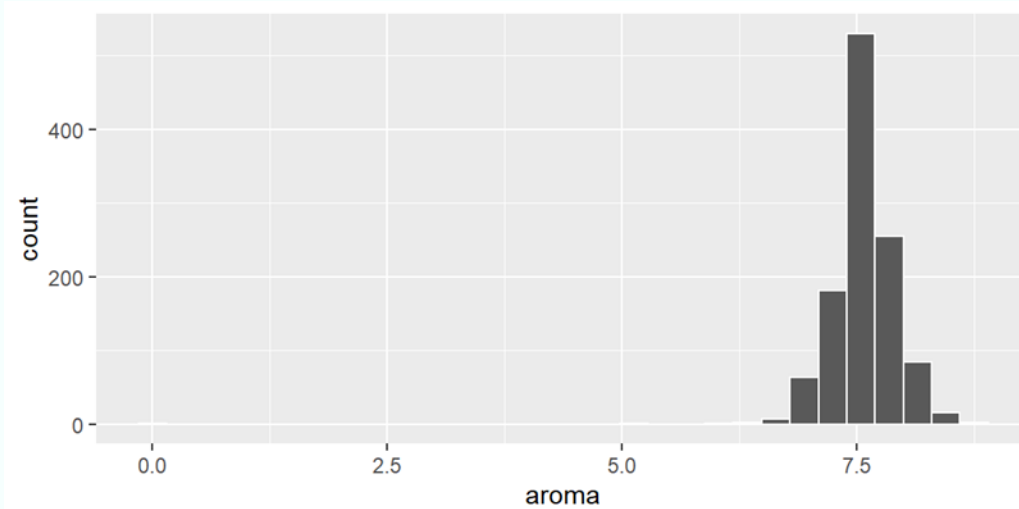
2.1 Data summarization

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

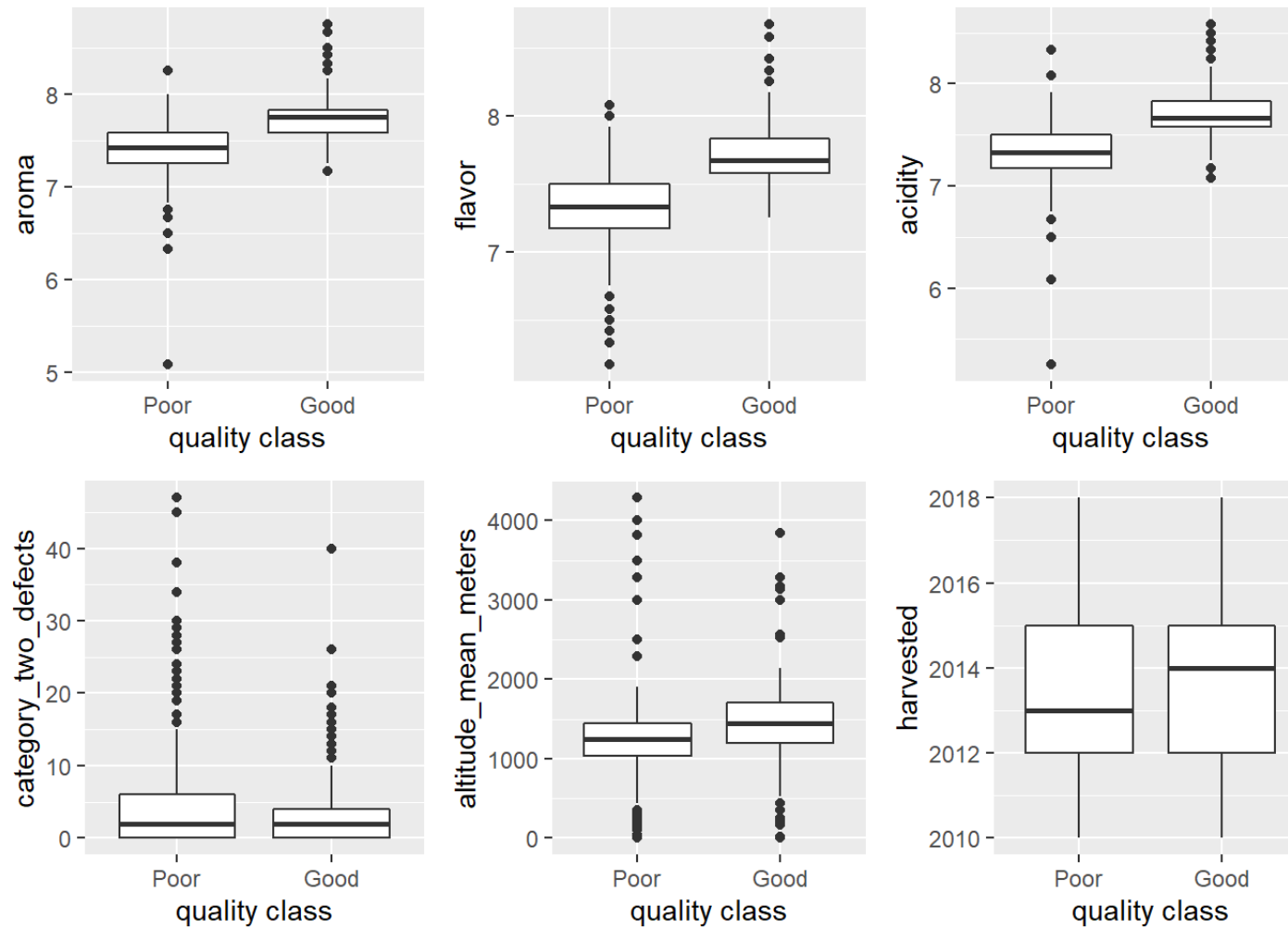


University
of Glasgow

2.1 Data summarization

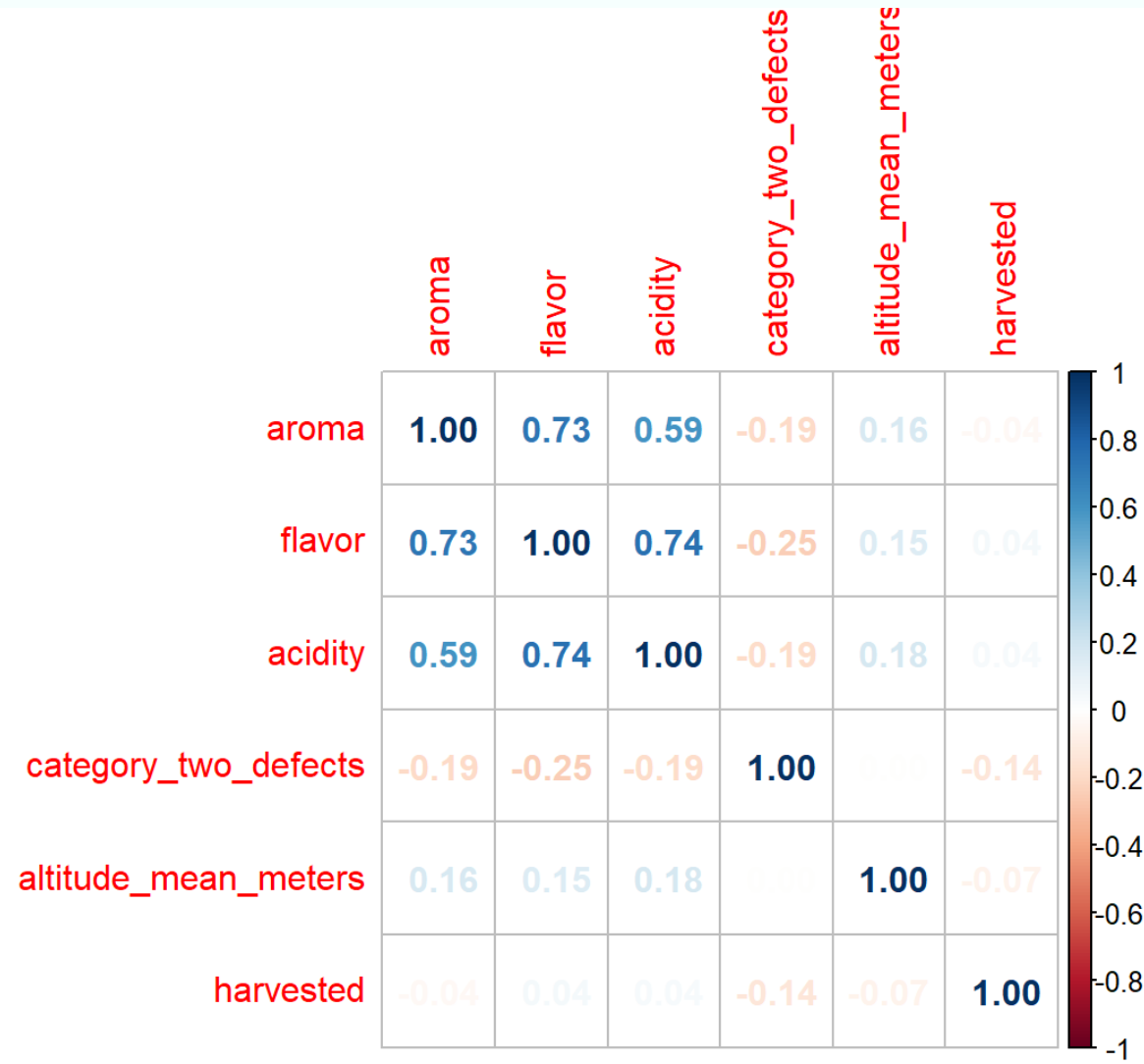


2.2 Data visualization



University
of Glasgow

2.2 Data visualization





PART.03

Formal Analysis



University
of Glasgow

3.1 Multicollinearity

Variance inflation factor (VIF)

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



University
of Glasgow

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



University
of Glasgow

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



University
of Glasgow

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



University
of Glasgow

3.2 Log-odds

95% confidence interval for log-odds

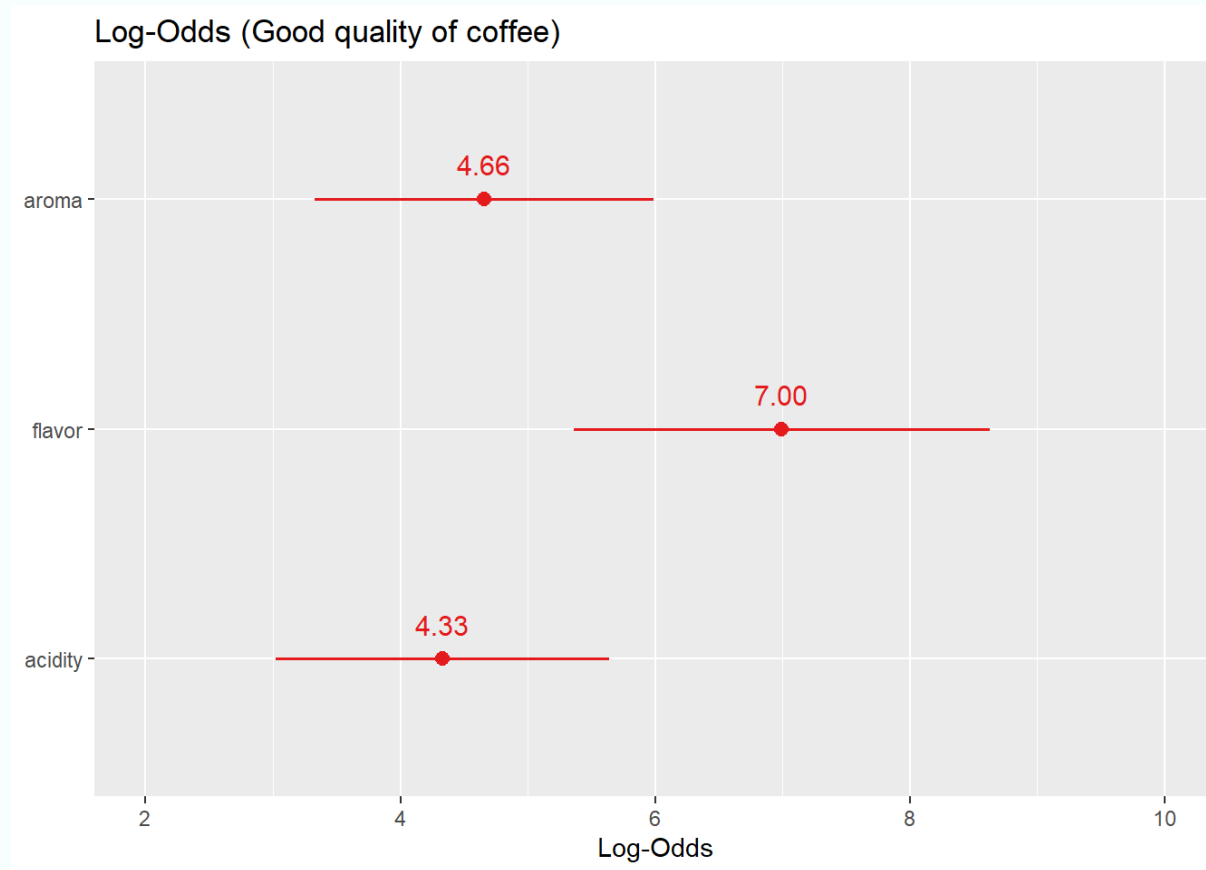
	2.5 %	97.5 %
(Intercept)	-139.331	-105.328
aroma	3.238	5.906
flavor	5.587	8.914
acidity	2.971	5.617
altitude_mean_meters	0.000	0.001



University
of Glasgow

3.2 Log-odds

95% confidence interval for log-odds



University
of Glasgow

3.2 Log-odds

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



University
of Glasgow

3.2 Log-odds

95% confidence interval for log-odds

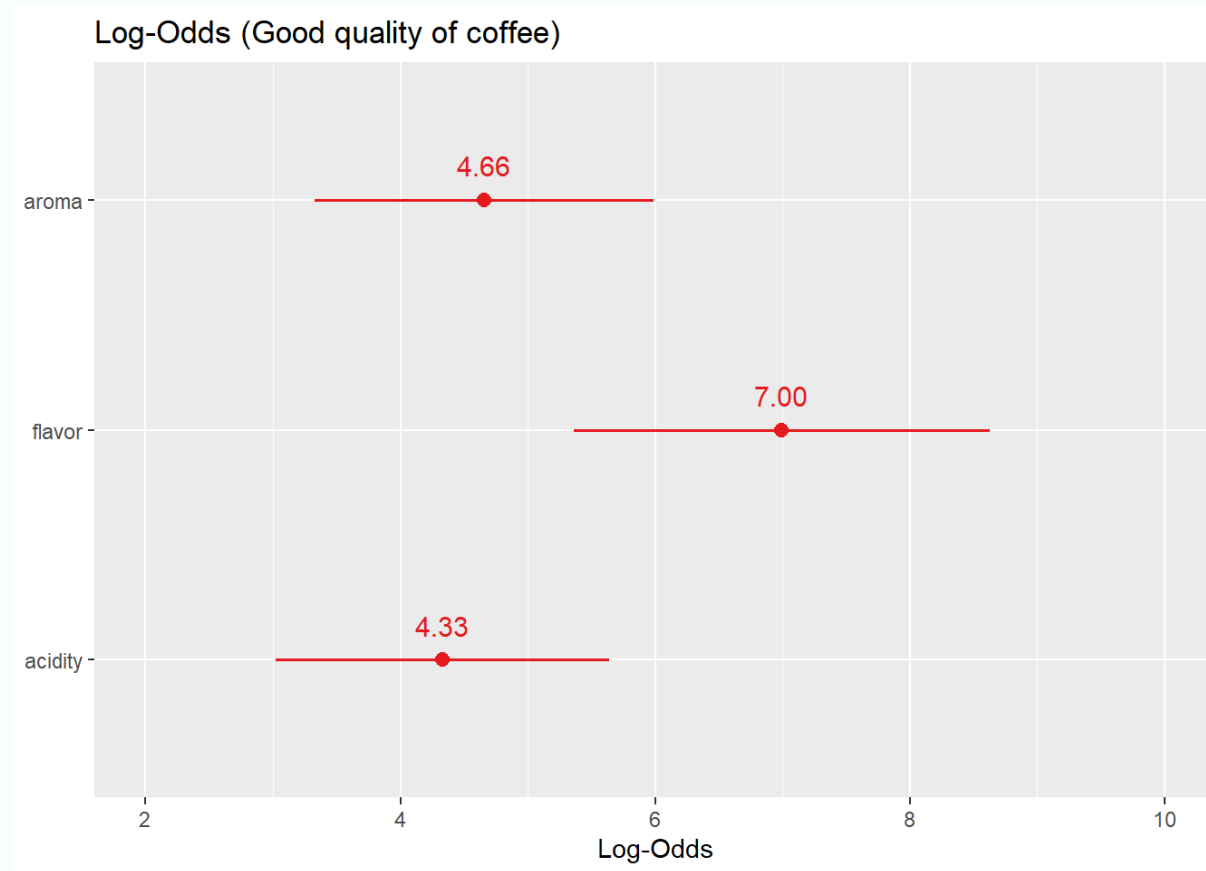
	2.5 %	97.5 %
(Intercept)	-138.309	-104.773
aroma	3.363	6.025
flavor	5.418	8.692
acidity	3.050	5.667



University
of Glasgow

3.2 Log-odds

95% confidence interval for log-odds



University
of Glasgow

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

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



University
of Glasgow

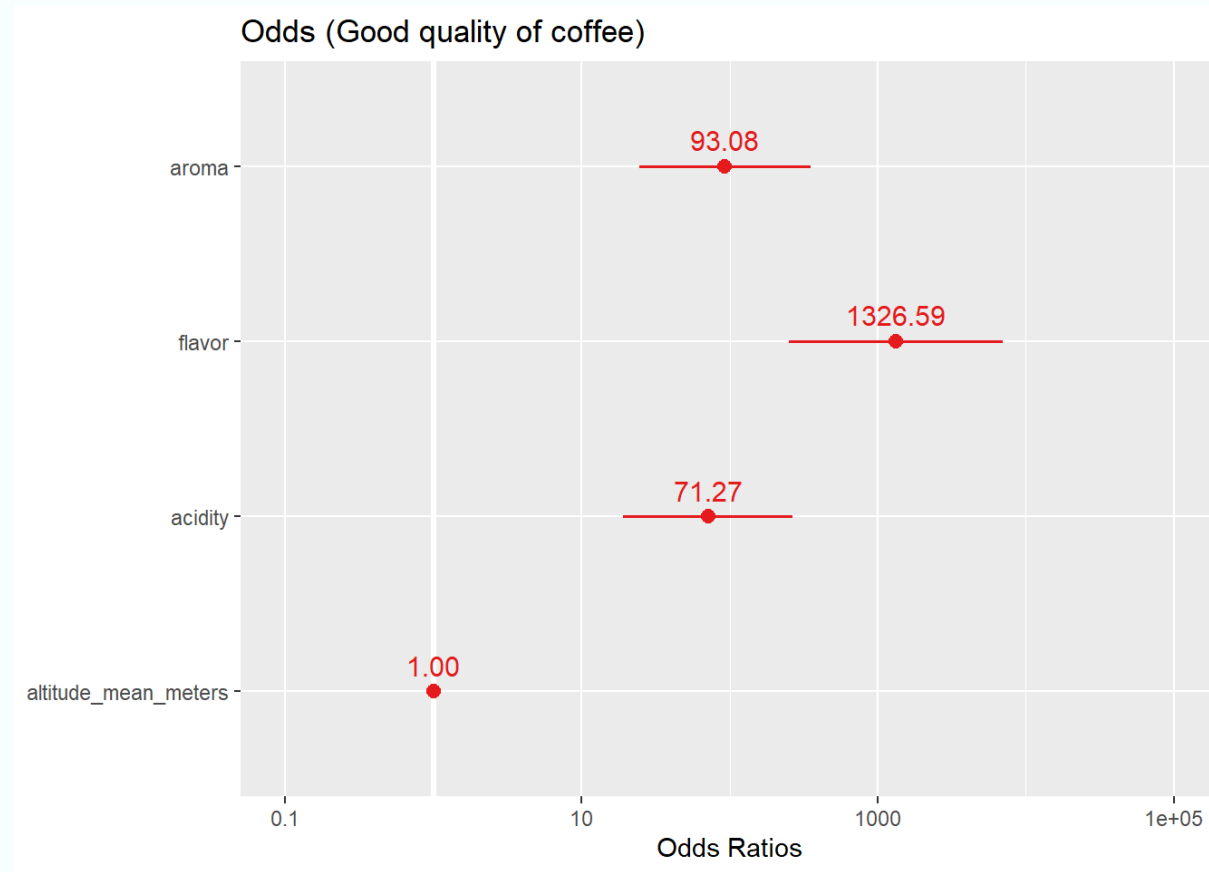
Odds (Good quality of coffee)

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

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



3.4 Odds

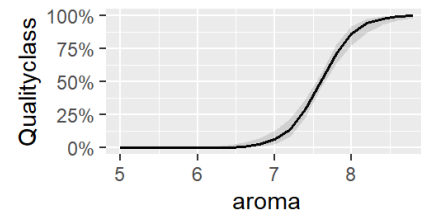


University
of Glasgow

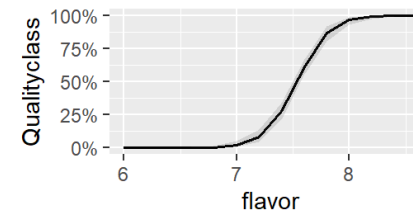
3.5 Probabilities

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

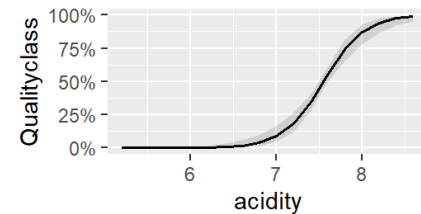
A



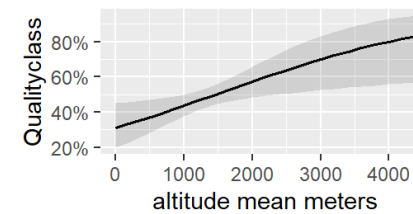
B



C



D



University
of Glasgow



PART.04

Conclusion



University
of Glasgow

5 Conclusion



University
of Glasgow



PART.05

Further Extension



University
of Glasgow

6 Further Extension

- Delve into the causes of missing values
- Further work of this data



University
of Glasgow



PART.06

Reference



University
of Glasgow

- [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.

