Analysis of Coffee Quality Factors

THESIS

Group 11

Xingyu Du, Manni Wang, Xinyi Wang, Zhixin Li, Fangyi Zhou

content

- Introduction of data
- Data Wrangling
- Exploratory Data Analysis
- Formal Analysis
- Conclusion and Further task

content

Introduction of Data

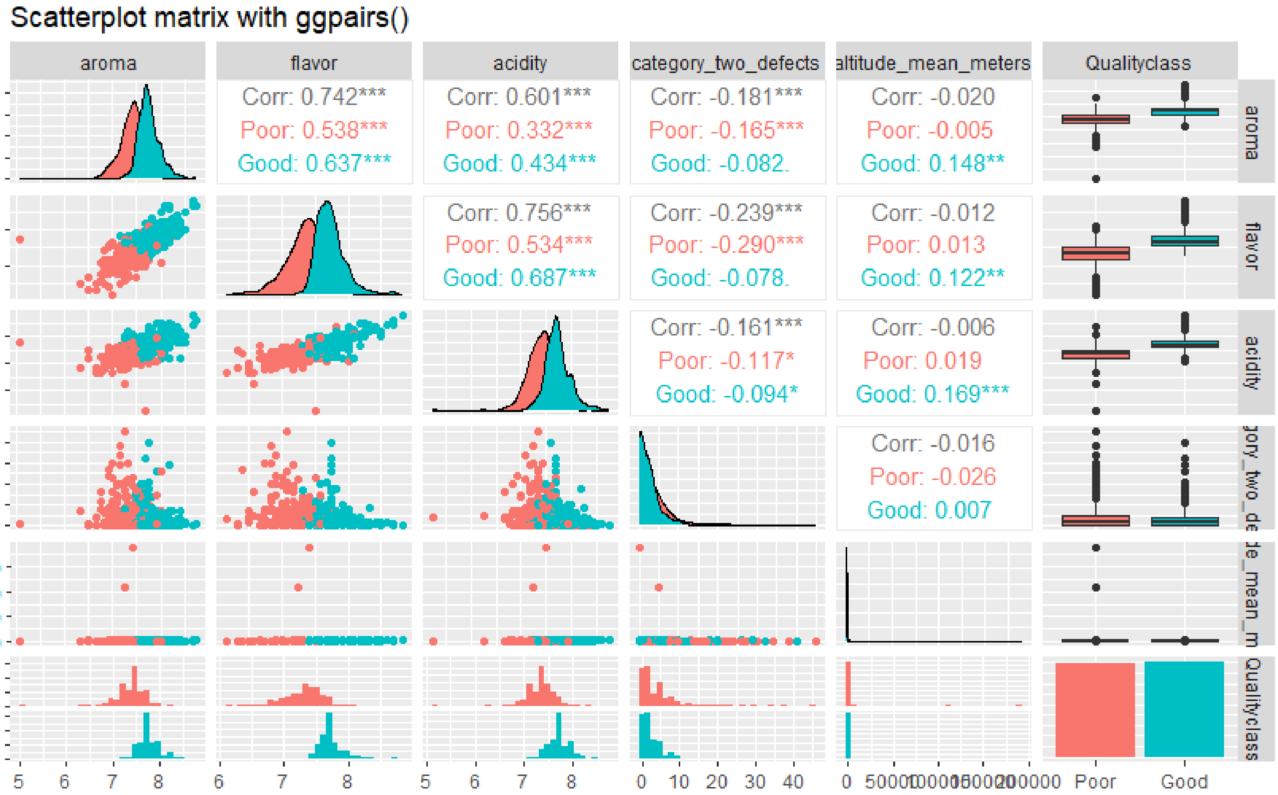
- Coffee is a highly popular beverage, it has been consumed for over 1000 years and today is consumed by about one-third of the world's population.

- In this project, we aim to explore the factors that affect the quality of coffee. We analyzed a dataset encompassing coffee bean information spanning from 2010 to 2018, including various coffee features and quality classifications (with beans sourced from different regions, such as Mexico and Colombia).

- Finally, through GLM analysis, we identified the varying degrees of influence that different coffee bean features have on coffee quality, generating an optimal model.

content

Data Wrangling



Remove outliers

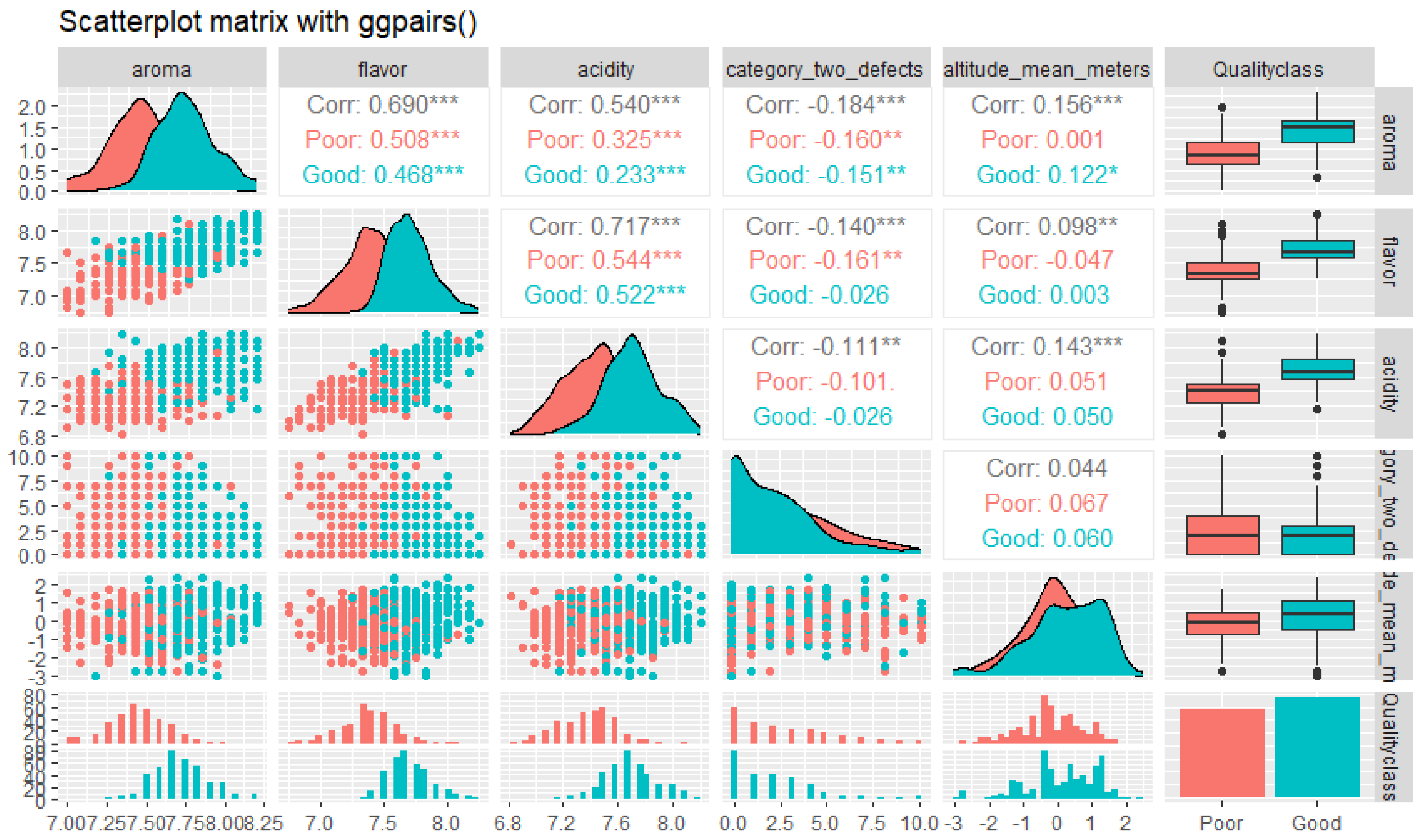
Remove outliers for each continuous variable (category_two_defects ,aroma, flavor, acidity,, altitude_mean_meters)

Standardization:

Standardize the 'altitude_mean_meters' variable to prevent it from having too much influence on the model due to its large numerical values, while keeping other variables equally important.

Scatterplot Matrix

After Standardization:



Summary Statistics

	aroma				flavor			
Qualityclass	ar.Mean	ar.Sd	ar.Min	ar.Max	fl.Mean	fl.Sd	fl.Min	fl.Max
Poor	7.44	0.19	7.00	8.00	7.36	0.21	6.75	8.08
Good	7.73	0.18	7.17	8.17	7.71	0.17	7.25	8.25

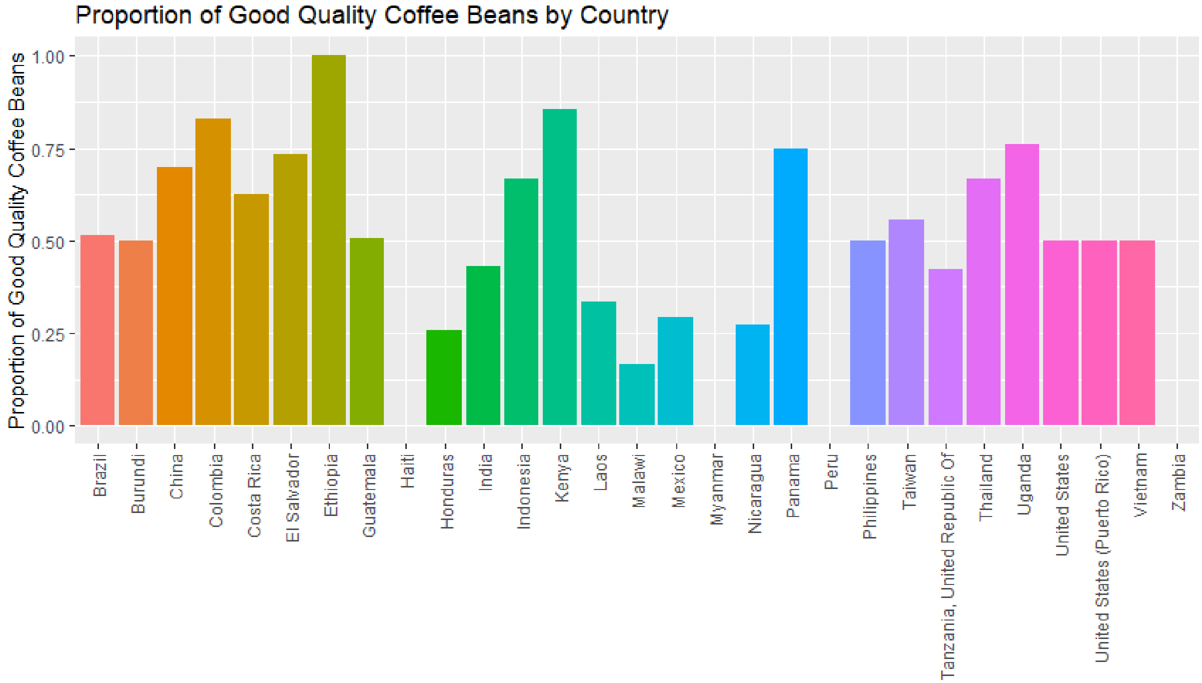
	acidity				Defects			
Qualityclass	ac.Mean	ac.Sd	ac.Min	ac.Max	C.Mean	C.Sd	C.Min	C.Max
Poor	7.38	0.20	6.83	8.08	2.75	2.64	0.00	10.00
Good	7.69	0.20	7.17	8.17	2.25	2.37	0.00	10.00

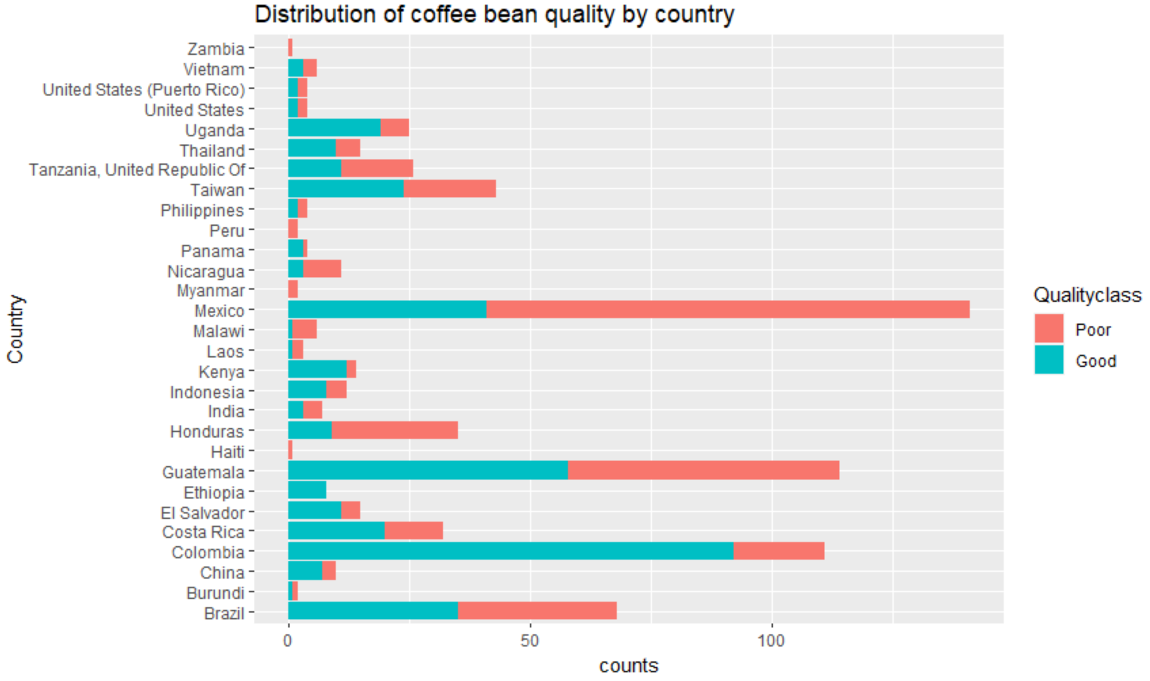
	Altitude mean meters			
Qualityclass	A.Mean	A.Sd	A.Min	A.Max
Poor	−0.18	0.91	−2.73	1.65
Good	0.16	1.05	−3.00	2.35

We observed that, in general, coffee beans classified as "good" quality tend to have higher values for aroma, flavor, and acidity compared to those classified as "poor" quality.

Data Visulization

Data Visulization

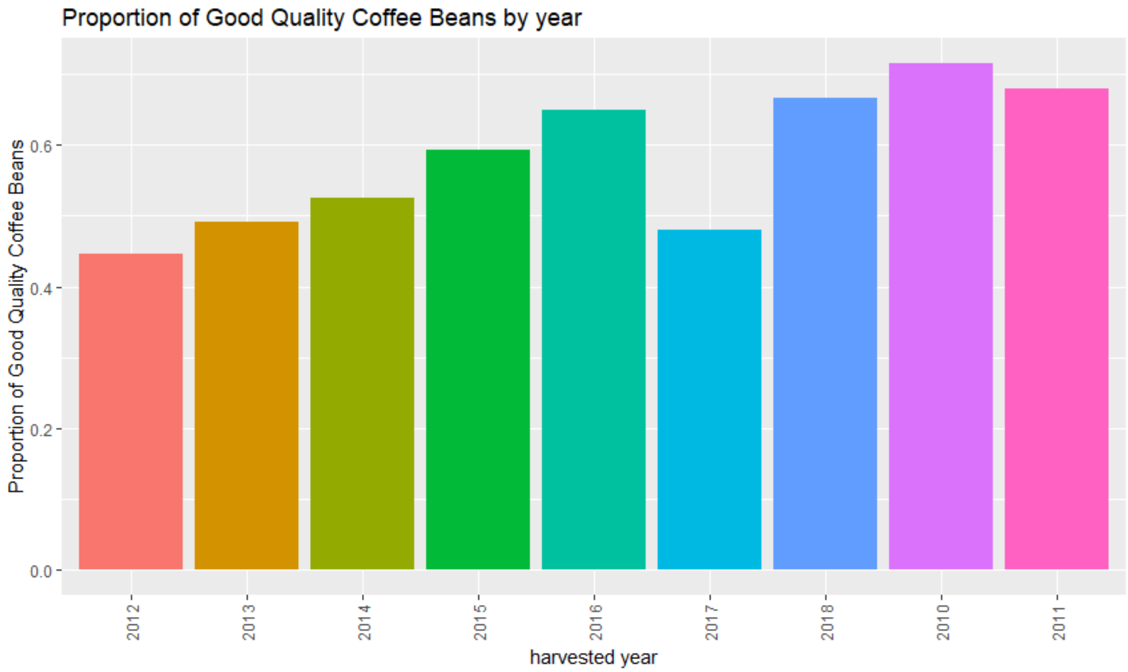


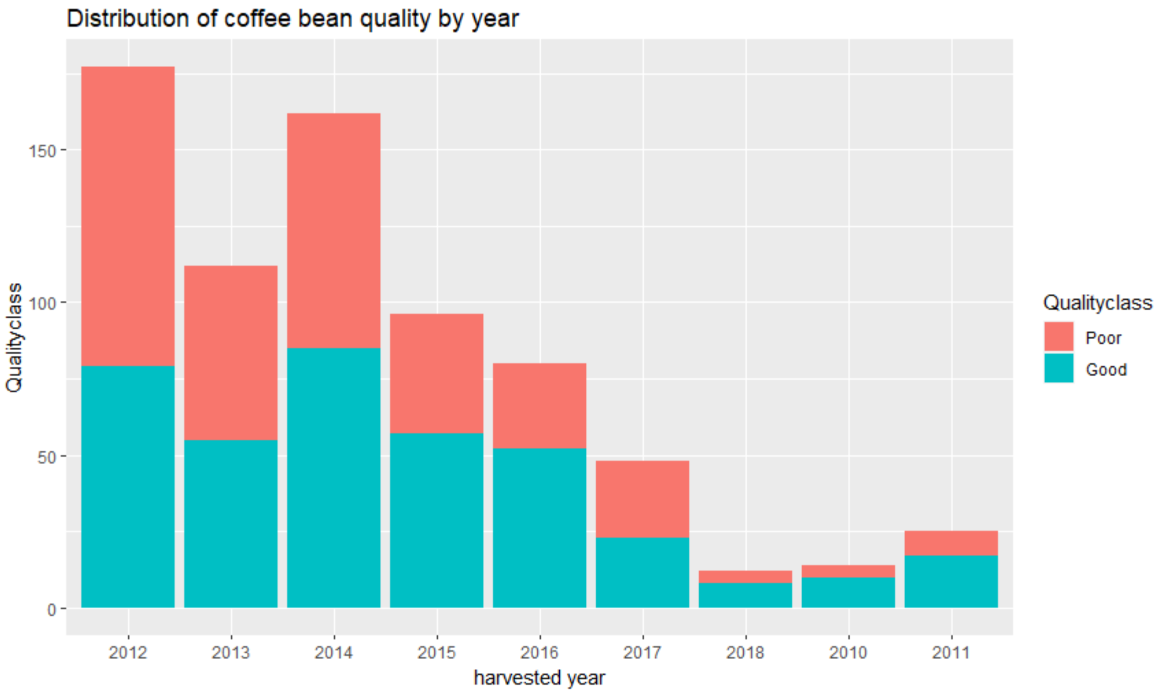


We can observe that certain countries, such as Brazil and Colombia, have a higher proportion of "good" quality coffee beans. In contrast, other countries like Honduras and Haiti have a lower proportion of "good" quality coffee beans, indicating relatively fewer high-quality coffee beans. Some countries, such as Mexico and Colombia, have more bar plot data points, which may indicate a larger sample size of coffee bean samples from these countries in the dataset.

Data Visulization

Data Visulization



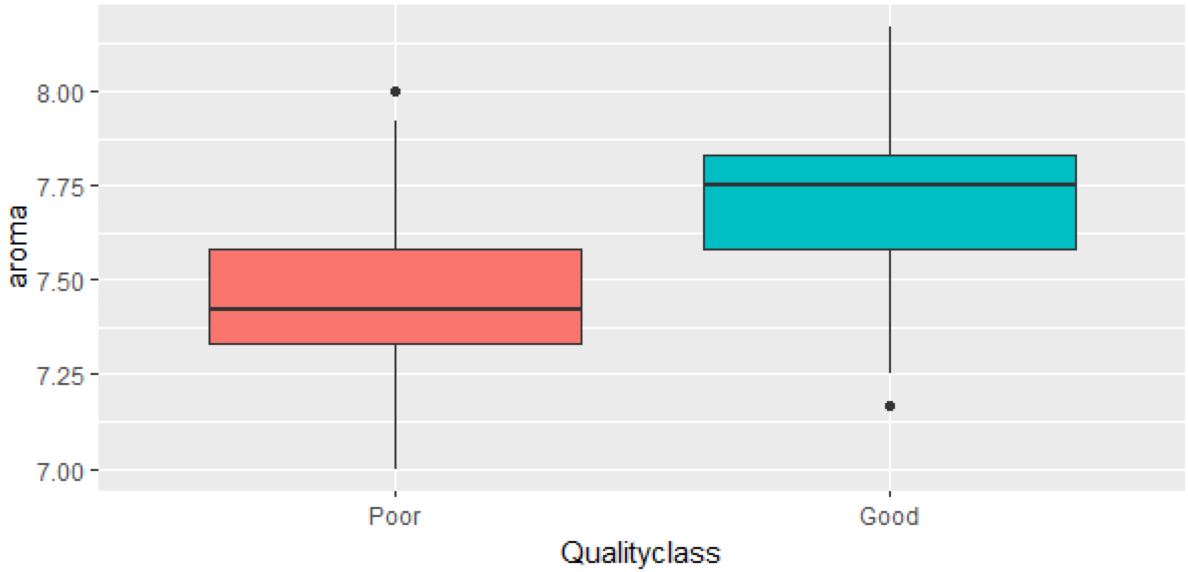


The analysis reveals a fluctuating trend in the proportion of good coffee beans over the years. Specifically, the proportion was at its peak in 2010, reaching approximately 70%. Subsequently, it experienced a decline, hitting its lowest point in 2012 at around 45%. From 2012 to 2016, there was a gradual increase observed, followed by a sharp decline in 2017 to approximately 47%. However, in 2018, there was another increase, with the proportion rising to about 67%.

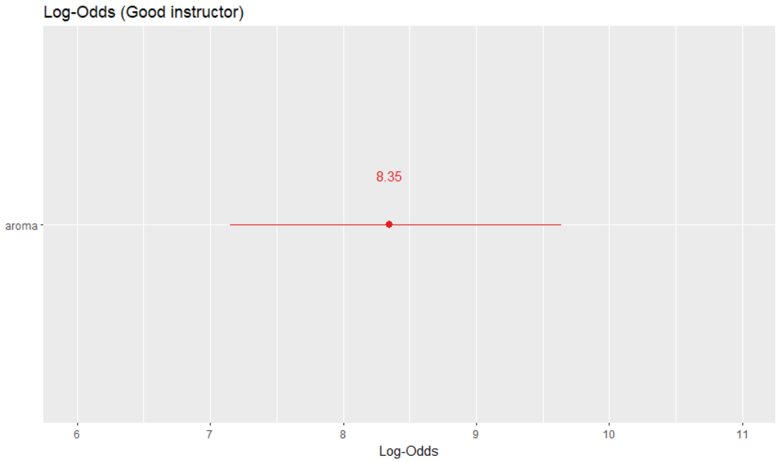
content

Exploratory Data Analysis

Aroma & Qualityclass



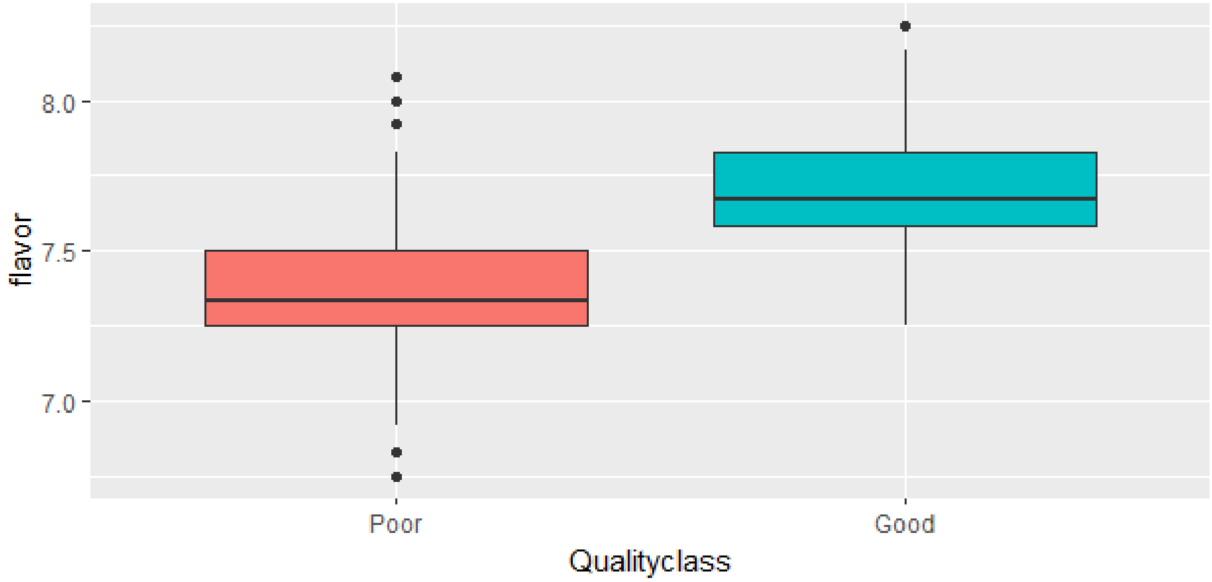
$$Log\text{-}odds:\ ln\left(\frac{p}{1-p}\right) = \alpha + \beta \times aroma$$



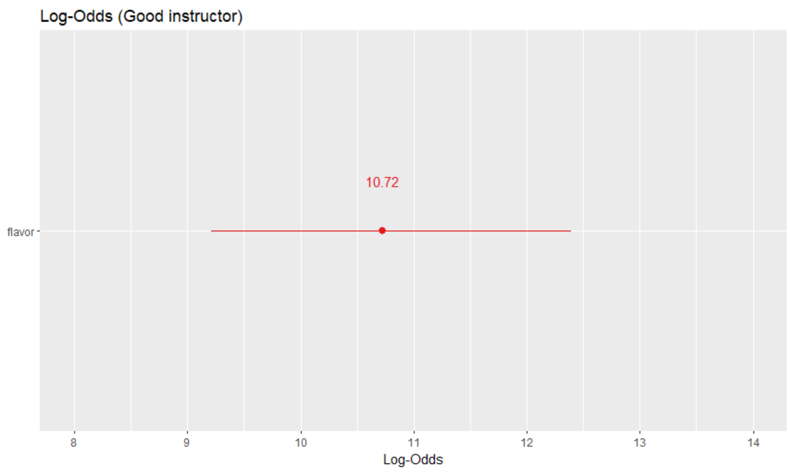
(7.10 , 9.59)

When the quality of coffee beans is poor, the aroma tends to be low, while for good quality beans, the aroma is higher. The 95% CI of β not including 0 indicates its significance. A positive coefficient suggests that as the aroma grade increases, the probability of coffee beans being classified as good also increases. Therefore, aroma grade is an important factor in measuring coffee bean quality.

Flavor & Qualityclass



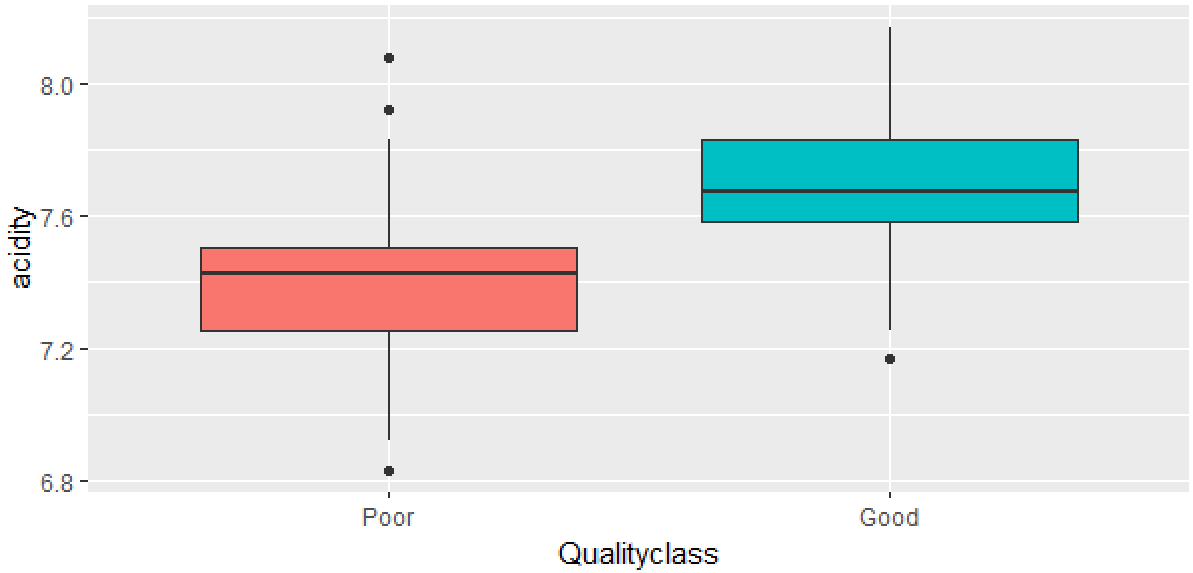
$$Log\text{-}odds:\ ln(\frac{p}{1-p}) = \alpha + \beta \times falvor$$



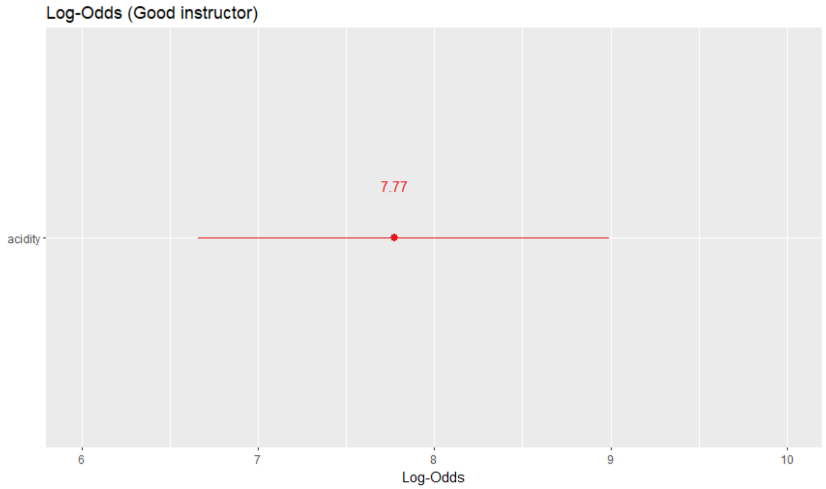
(9.14 , 12.31)

When the quality of coffee beans is poor, the flavor tends to be low**,** **w**hile for good quality beans, the flavor is higher. The 95% CI of β not including 0 indicates its significance. A positive coefficient suggests that as the flavor grade increases, the probability of coffee beans being classified as good also increases. Therefore, flavor grade is an important factor in measuring coffee bean quality.

Acidity & Qualityclass



Log-odds: $\boldsymbol{ln(\frac{p}{1-p}) = \alpha + \beta \times acidity}$

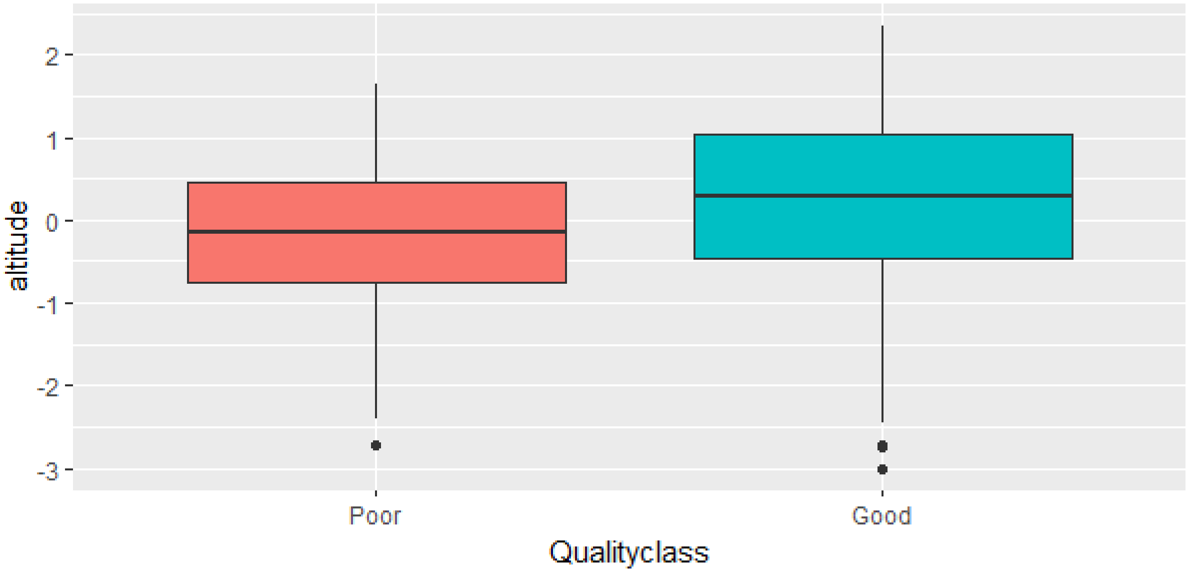


(6.61 , 8.94)

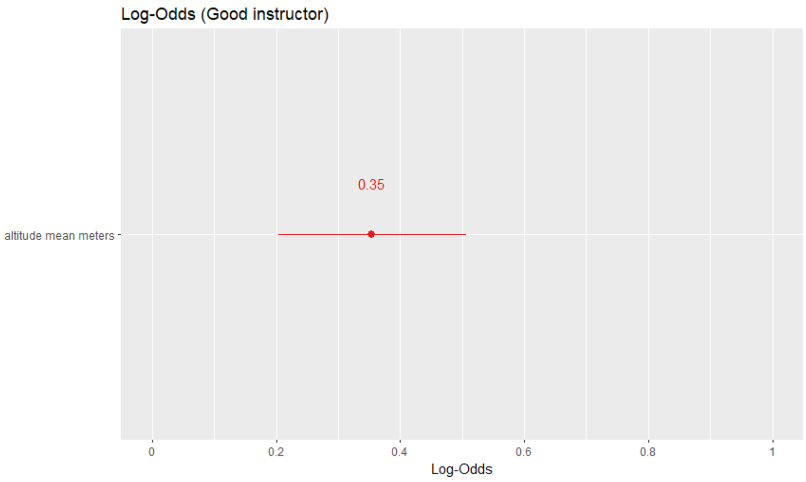
When the quality of coffee beans is poor, the acidity tends to be low, while for good quality beans, the acidity is higher. The 95% CI of β not including 0 indicates its significance. A positive coefficient suggests that as the acidity grade increases, the probability of coffee beans being classified as good also increases. Therefore, acidity grade is an important factor in measuring coffee bean quality.

Exploratory Data Analysis

Altitude mean meters & Qualityclass



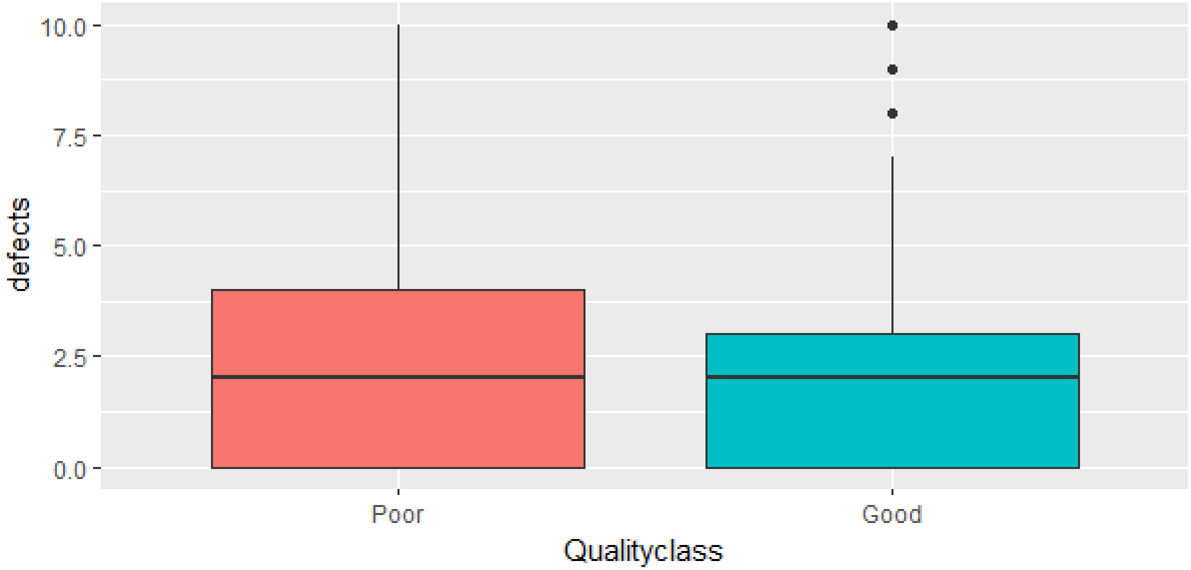
$$Log\text{-}odds: (\ln \frac{p}{1-p}) = \alpha + \beta \times altitude$$



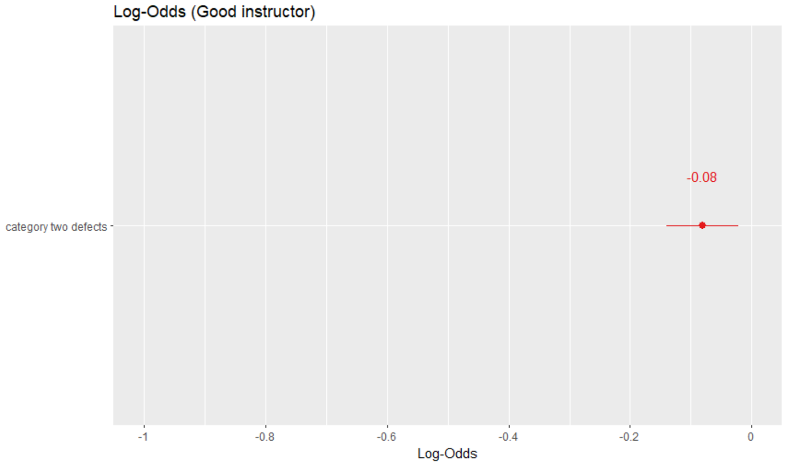
(0.20 , 0.50)

When the quality of coffee beans is poor, the altitude tends to be low, while for good quality beans, the altitude is higher. The 95% CI of β not including 0 indicates its significance. A positive coefficient suggests that as the altitude increases, the probability of coffee beans being classified as good also increases. Therefore altitude is an important factor in measuring coffee bean quality.

Category 2 type defects & Qualityclass



Log-odds: $ln(\frac{p}{1-p}) = \alpha + \beta \times defects$



(-0.13 , -0.02)

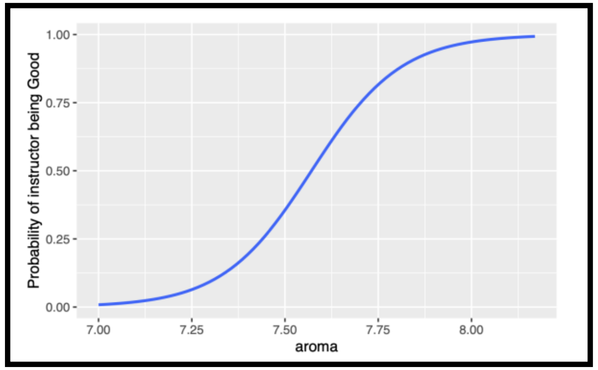
The 95% CI of β not including 0 indicates its significance. However, β is close to zero, indicating that the impact of the number of defects on coffee bean quality is relatively small compared to other variables.

A negative coefficient suggests that as the number of defects increases, the probability of coffee beans being classified as good decreases.

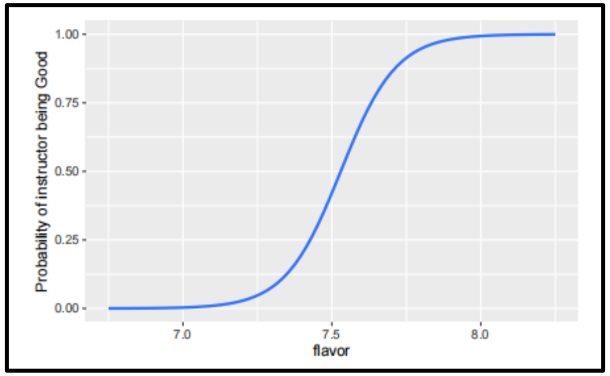
Overall, the number of defects is an important factor in measuring coffee bean quality.

Exploratory Data Analysis

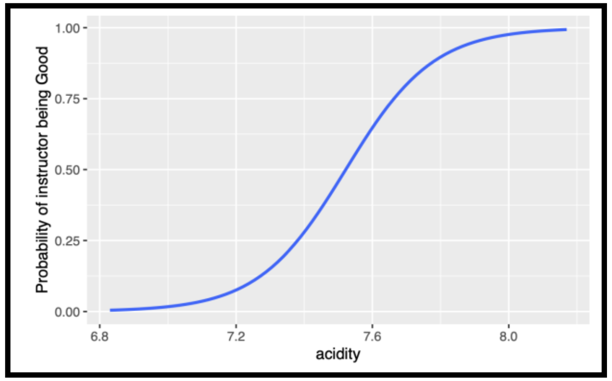
Predicted Probabilities



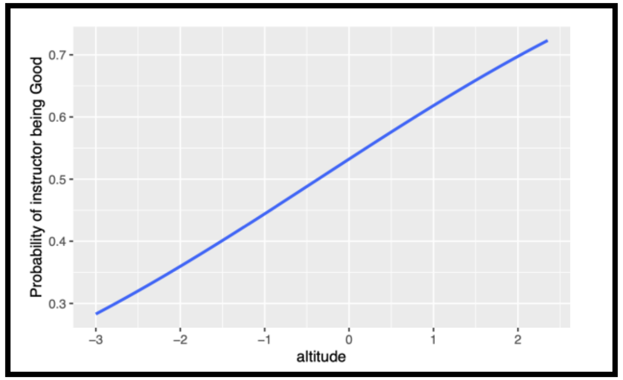
Aroma



Flavor

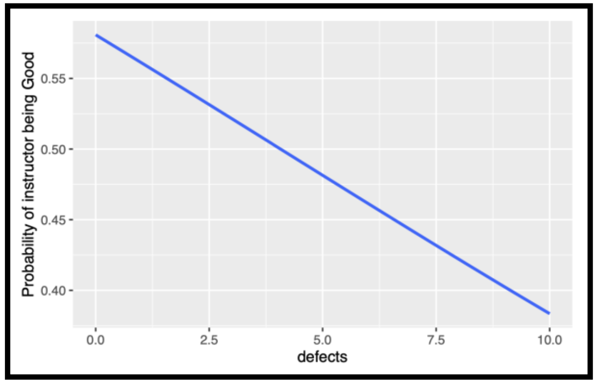


Acidity



Altitude

The upward trend in the curve indicates that as the explanatory variables increase, the probability of coffee beans being classified as "good" also increases. This implies a positive correlation between these explanatory variables and the quality of the coffee beans.



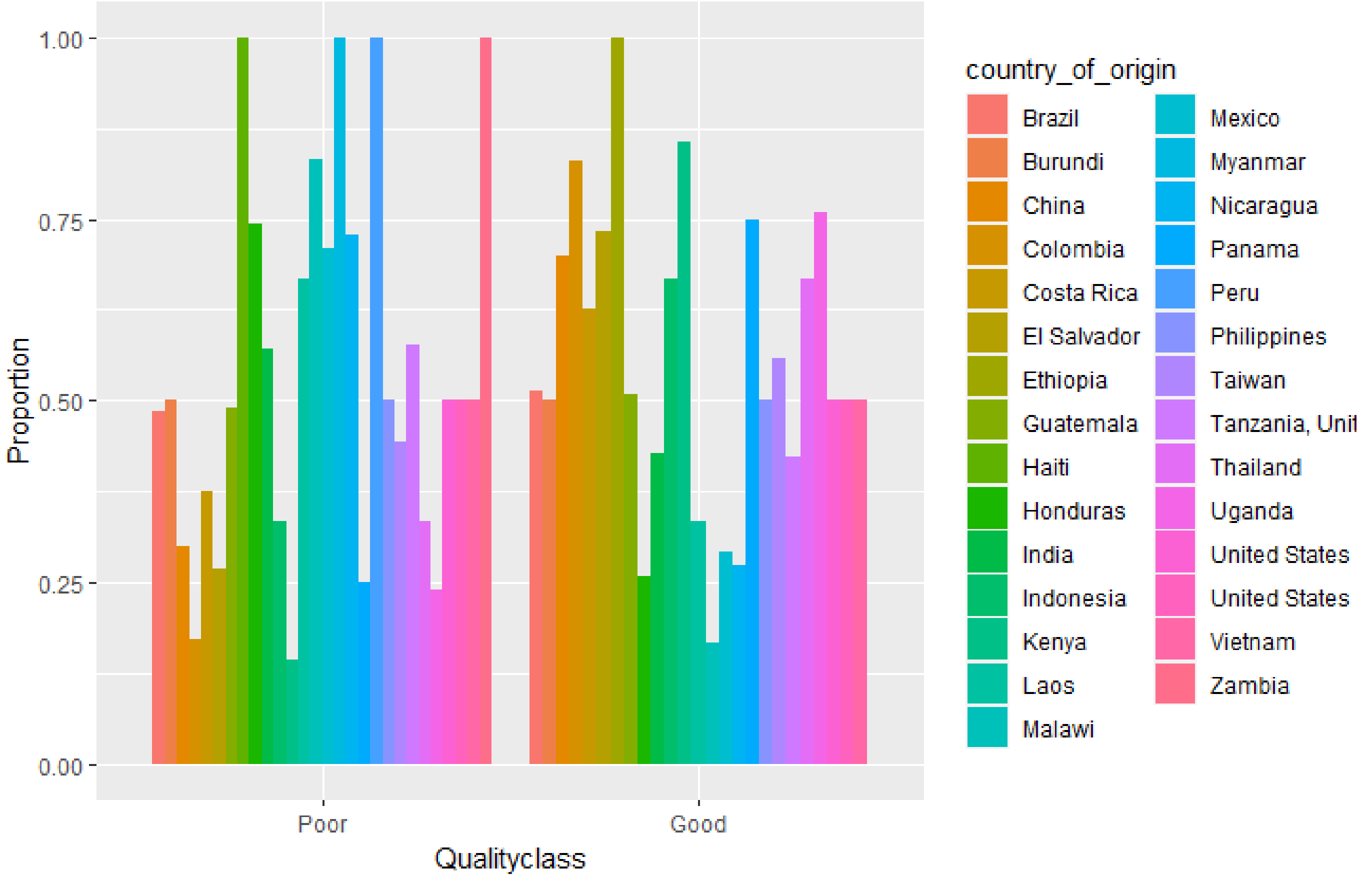
Defects

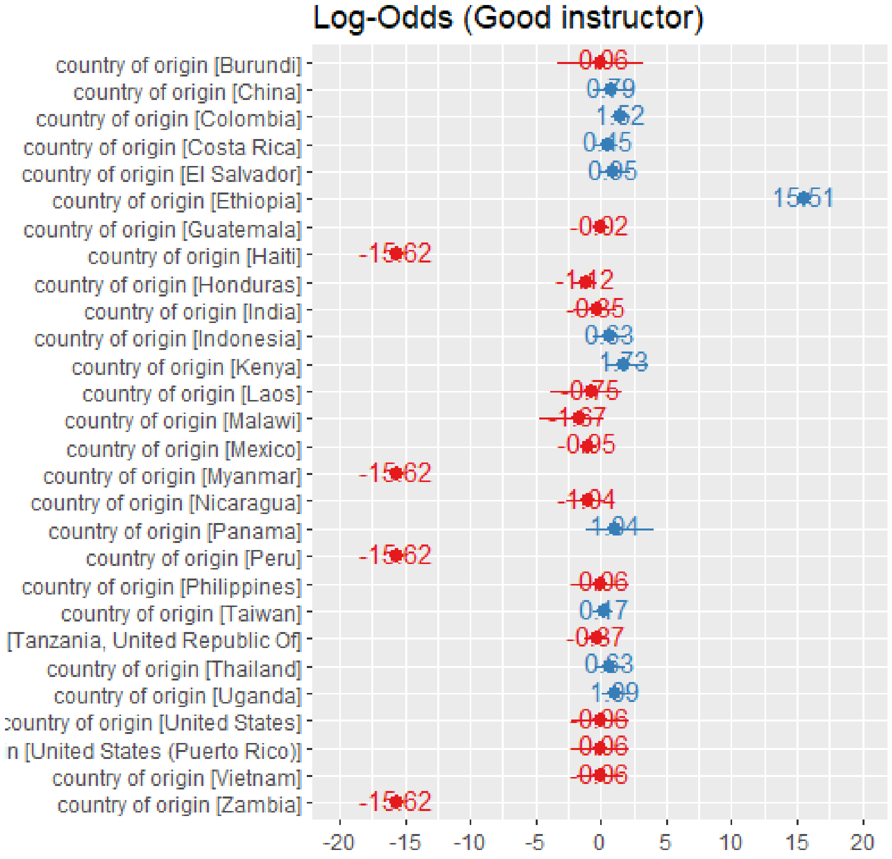
The downward trend in the curve indicates that as the number of defects increase, the probability of coffee beans being classified as "good" decreases. This implies a negative correlation between the number of defects and the quality of the coffee beans.

Exploratory Data Analysis

Country & Qualityclass

$$Log\text{-}odds:\ ln\left(\frac{p}{1-p}\right) = \alpha + \beta \times country$$



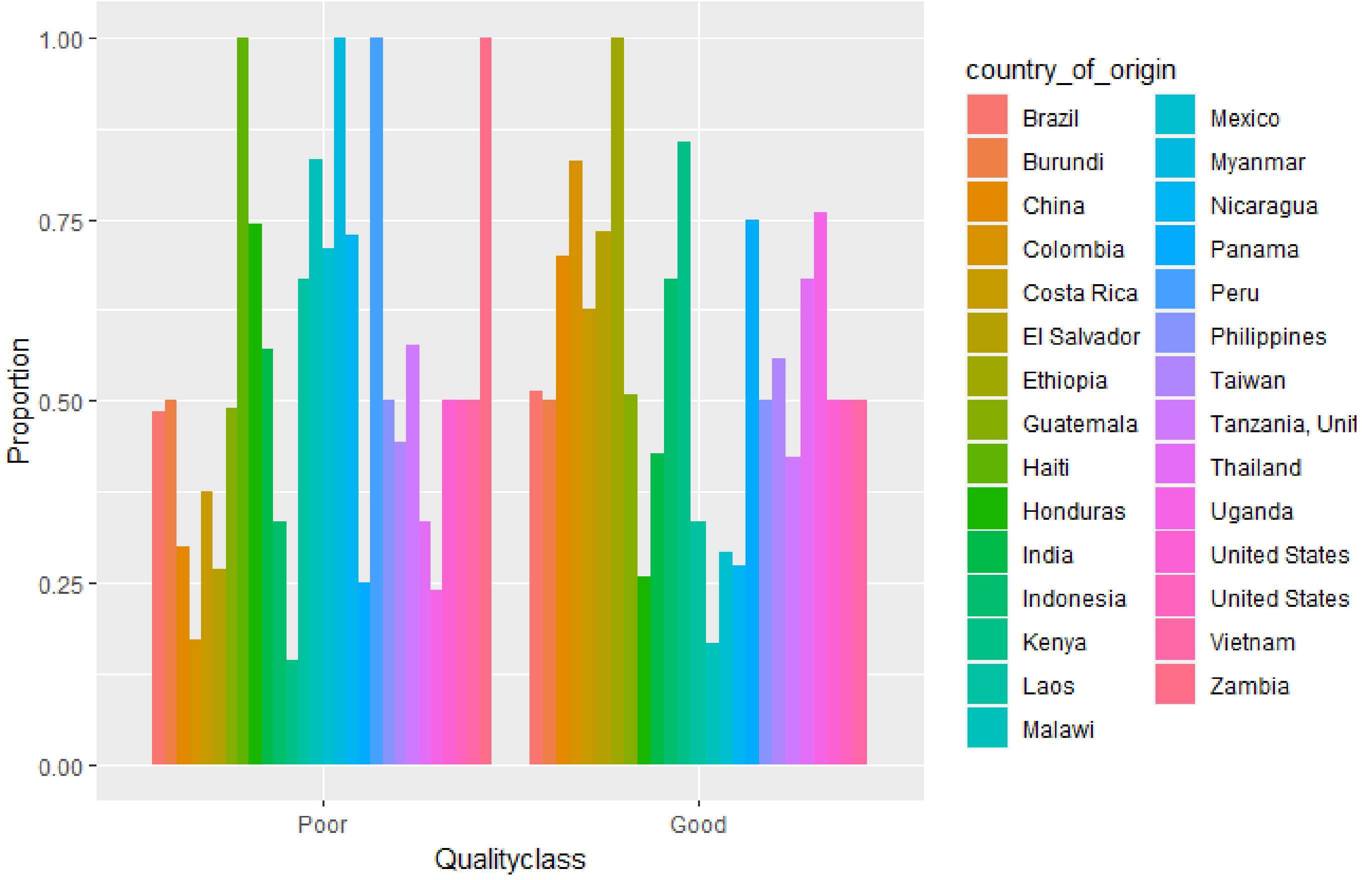


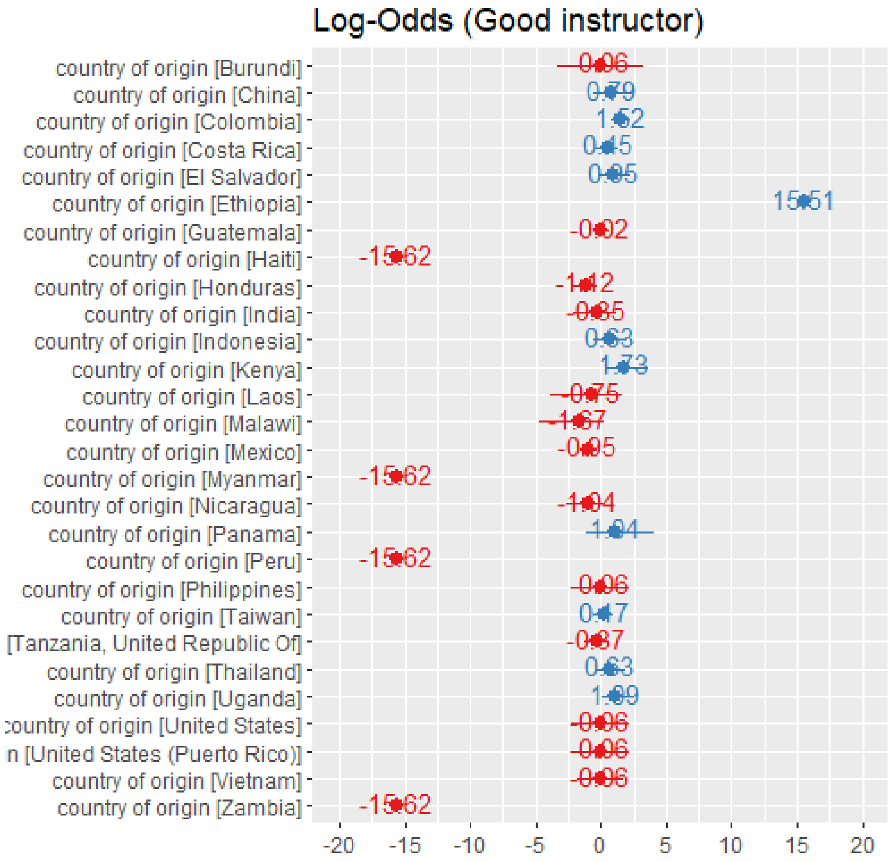
Compared to the baseline, countries with positive coefficients indicate a positive influence on the probability of coffee beans being classified as good, while countries with negative coefficients indicate a negative influence on the probability of the coffee beans being classified as good.

Exploratory Data Analysis

Country & Qualityclass

$$\textit{Log-odds: } \ln\left(\frac{p}{1-p}\right) = \alpha + \beta \times country$$

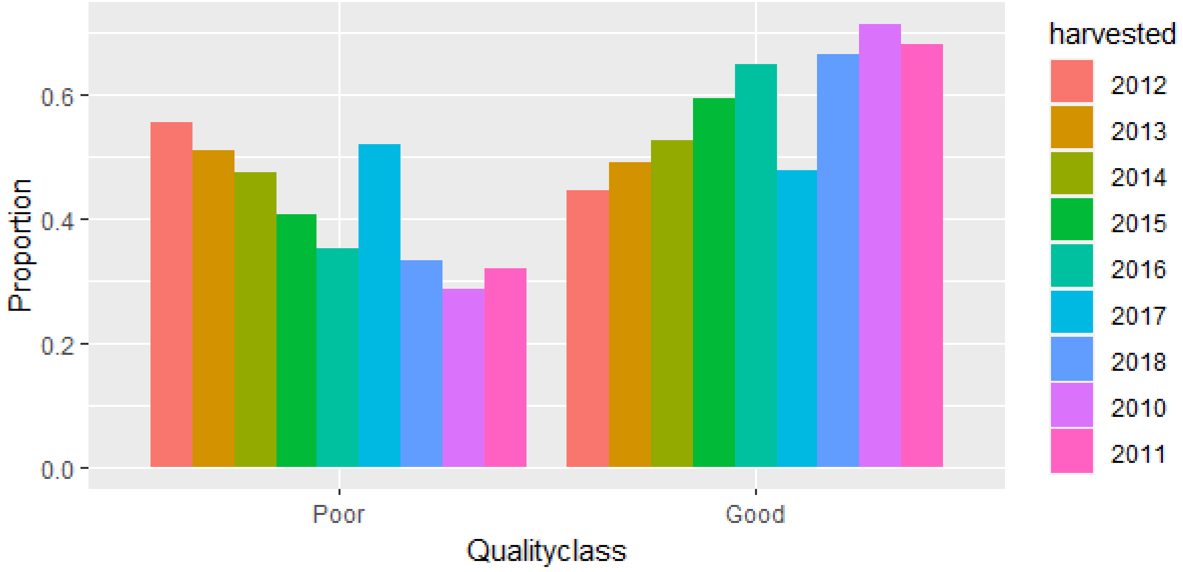


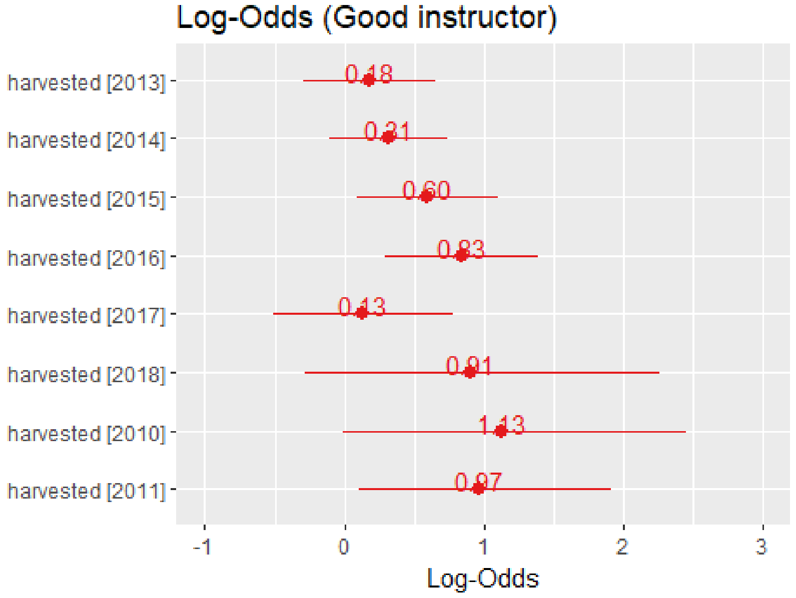


Relative to the baseline, when the coefficient is close to 0, it indicates that the country has a minimal impact on the probability of coffee beans being classified as good or bad.

Harvested & Qualityclass

$$\textit{Log-odds: } \ln\left(\frac{p}{1-p}\right) = \alpha + \beta \times harvested$$





Compared to the baseline, harvested years with positive coefficients indicate a positive influence on the probability of coffee beans being classified as good.

content

- introduction of data
- Data Wrangling
- Exploratory Data Analysis
- Formal Analysis
- Conclusion and Further task

Principle Component Analysis

	aroma	flavor	acidity
aroma	1.00	0.69	0.54
flavor	0.69	1.00	0.72
acidity	0.54	0.72	1.00

	PC1	PC2	PC3
sd.	1.5170	0.6790	0.48747
variance prop.	0.7671	0.1537	0.07921
cumulative prop.	0.7671	0.9208	1.0000

Based on the correlation matrix, it is evident that these three variables exhibit high correlation. Therefore, we adopt PCA to help address multicollinearity, thereby enhancing the stability and interpretability of the model.

The cumulative proportion of three variables adds up to 1, indicating that these three principal components fully explain the variability in the original data without losing information. Therefore, adopting principal component analysis is justified. We choose PC1 and PC2 as the combination.

Model Selection

- We initiate the modeling process with the full model and apply AIC for stepwise model selection.

$$P=Prob(\,Qualityclass = \text{“Good”}\,)$$

- ***Full model(PCA)*** : $ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \cdot x_{PC1} + \beta_2 \cdot x_{PC2} + \beta_{country} + \beta_3 \cdot x_{defects} + \beta_4 \cdot x_{altitude} + \beta_{harvested}$

	AIC
Full model	446.01
-harvested	441.55
-altitude	440.57

- ***Optimal model*** : $ln\left(\frac{p}{1-p}\right) = -0.042 + 0.715 \cdot x_{PC1} - 0.073 \cdot x_{PC2} + \hat{\beta}_{country} + 0.119 \cdot x_{defects}$

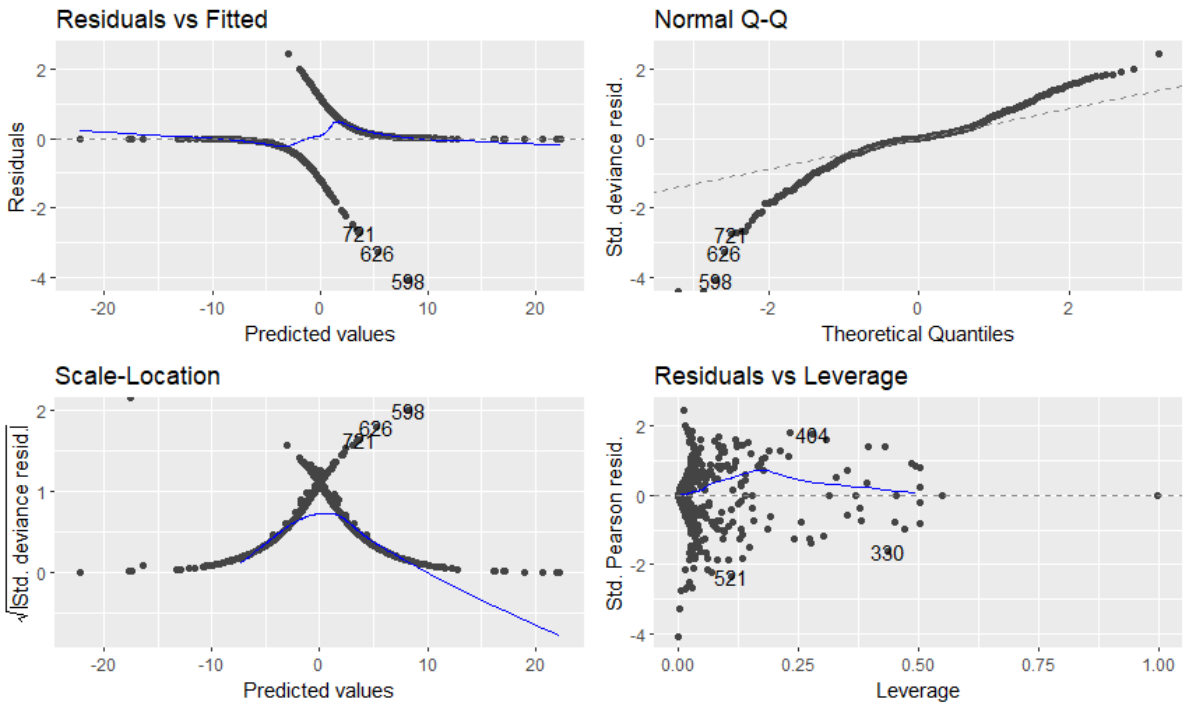
$$\hat{\beta}_{country} = \begin{cases} 0, & Country = Brazil(baseline) \\ 1.678, & Country = Colombia \\ -3.350, & Country = India \\ -1.324, & Country = Mexico \\ 1.819, & Country = Thiland \\ -1.434, & Country = Uganda \\ -1.550, & Country = other\ country\ (average) \end{cases}$$

Formal Analysis

Assumption Check & Cross-validation

- **Assumption Check**

Residual vs Fitted:
Appearing some non-random pattern, non-linear correlation might exist



Normal QQ:
The tail of the residual has some deviation

Scale-Location:
As the fit value increases, so does the variation

Residuals vs Leverage:
Some possible outliers might exist

- Through cross-validation, we can further assess the generalization ability of the model.

accuracy	kappa	poor	good
0.8343	0.6675	340	386

Accuracy: model correctly predicted the class labels for around 83.44% of the samples.

content

Conclusion and Further Task

Conclusion:

Based on the model selection, four potential exploratory variables are suggested: **PC1**, **PC2** (a linear combination of **aroma**, **flavor**, and **acidity**), **country_of_origin**, and **category_two_defects**. This indicates that changes in these four variables could affect the quality of a batch of coffee. Additionally, the AIC of the optimal model is 440.57.

Further Task:

1. Check nonlinearity: Try to add polynomial terms or interaction terms for predictors to capture nonlinear relationships.

2. Data stratification: If there are many levels of classification variables such as **country_of_origin**, consider whether some rare classes need to be combined to reduce model complexity and prevent overfitting.

THANKS

Thank you for listening

Group 11