Optimizing Coffee Selection: Data-Driven Insights for Coffee-Shop Owners

Team 12: Jiajun Fang, Henry Guo, Xin Wang

DS 5110

AGENDA

01 **Introduction**
- Background
- Game Plan
- Data Cleaning

02 **Linear Model - Jiajun Fang**
- Method
- Results

03 **Clustering - Xin Wang**
- Method
- Results

04 **Text Analysis - Henry Guo**
- Method
- Results

05 **Discussion**

Background

Challenge: Select the right coffee products that resonate with customers

Goal: Provide data-driven insights into which coffee features most influence coffee ratings

Dataset Overview

- Over 8000 reviews - Web Scraped from Coffee Review Website
- Ratings, sensory features (e.g., aroma, acidity, body, flavor), roast levels, price, and coffee origins
- Review text: blind assessment

Game Plan

1. Linear Model

- Discover if there are any linear relationships between features and coffee rating
- Serve as "baseline" to help customers interpret coffee ratings

2. Clustering

- Explore the relationships between coffee characteristics and their ratings by clustering coffees based on their features
- Examine if there exists different types of coffee based on combination of features

3. Text Analysis

- Analyze if written reviews have certain relationships with coffee ratings (sentiments, word frequency, etc)
- Looking for certain words/phrases that reviewers might use for higher rated coffee

Data Cleaning

Roast Level: Converted to an ordinal variable (Light → Very Dark).

Agtron: Split into **Agtron_whole** and **Agtron_ground**.

- **Agtron_whole**: Fixed 4 typos using official review pages (e.g., index 2050 corrected to 54).
- **Agtron_ground**: One extreme outlier (index 7309) corrected from 689 to 68.

Roaster:Merged similar roaster names for consistency.

Roaster Location: Format inconsistencies (e.g., “Toronto, Ontario, Canada” vs. “South Korea”).
- Combined with Roaster and used Google Maps API to extract clean location coordinates.

Est. Price: Extracted currency, price, quantity, and unit using regex.

- Standardized to **USD per 100g**, adjusted for **inflation to 2024** using CPI.

Review Date: Converted to datetime format.

Acidity: Merged “Acidity” and “Acidity/Structure” into a single column.

Coffee Origin: Summarized *to each cell’s country name (i.e. str_detect(coffee_origin, "yirgacheffe") ~ "ethiopia")*

Linear Model

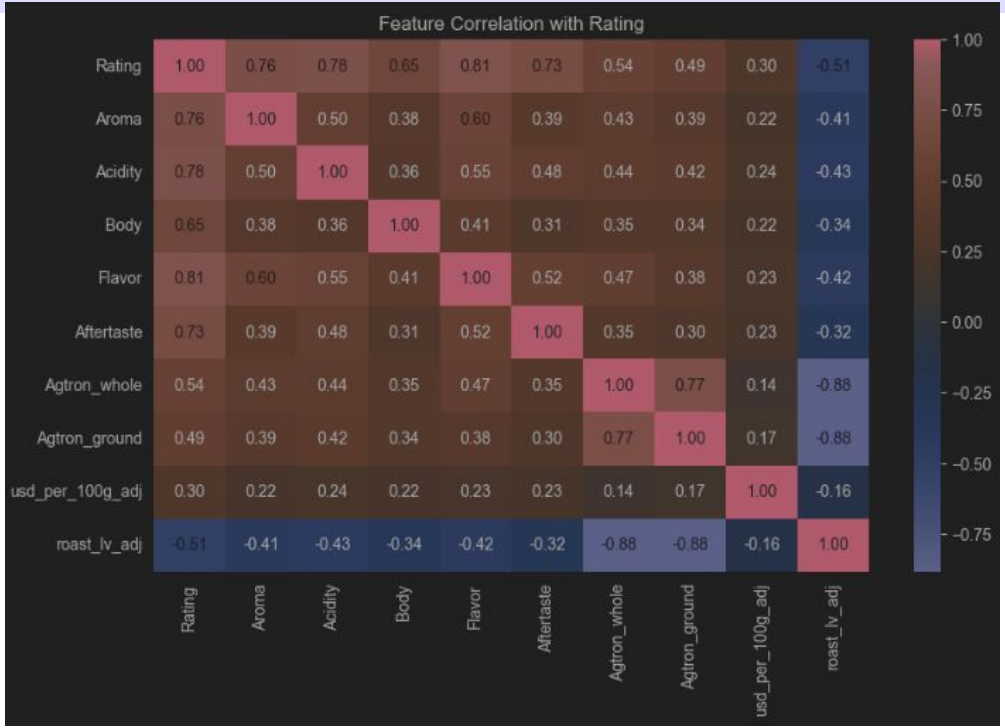
Jiajun Fang

Linear Model: EDA

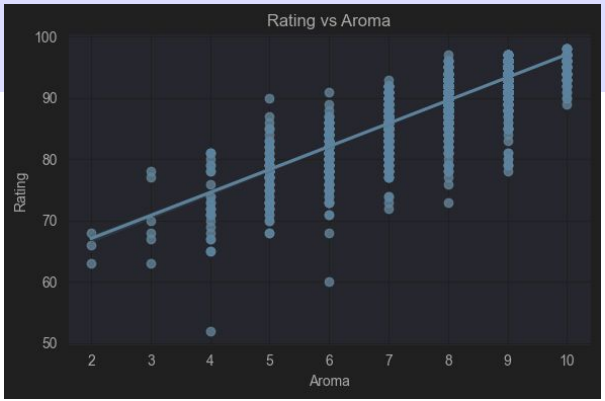
Design a linear model that can predict coffee ratings

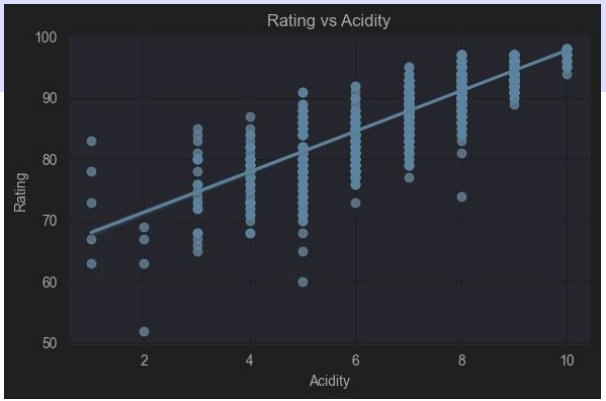
- **Features of interest:** Aroma, Acidity, Body, Flavor, Aftertaste, roast_lv_adj, Agtron_whole, Agtron_ground, usd_per_100g_adj
- From EDA, correlation heatmap, and feature importance, top 5 strongest predictors are Aroma, Acidity, Body, Flavor, and Aftertaste

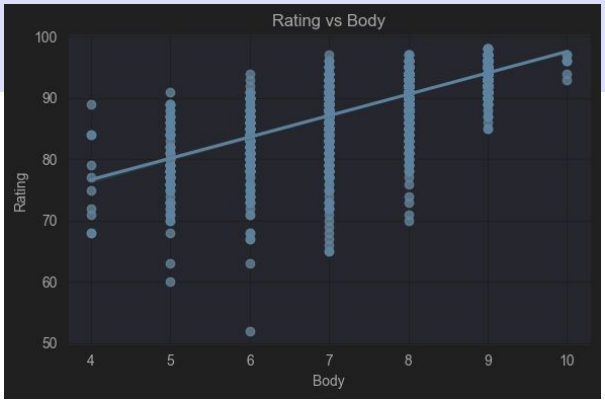
	R_squared
Flavor	0.663599
Acidity	0.613810
Aroma	0.574541
Aftertaste	0.528450
Body	0.417190
Agtron_whole	0.296180
roast_lv_adj	0.261282
Agtron_ground	0.235254
usd_per_100g_adj	0.092163

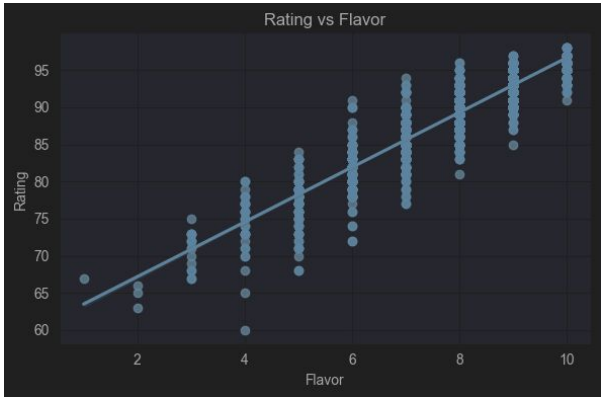


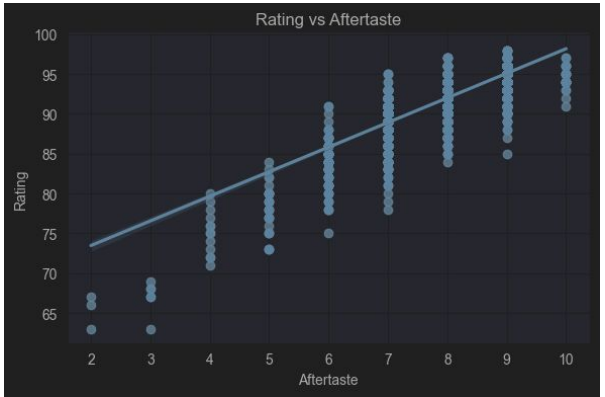
Linear Model: EDA

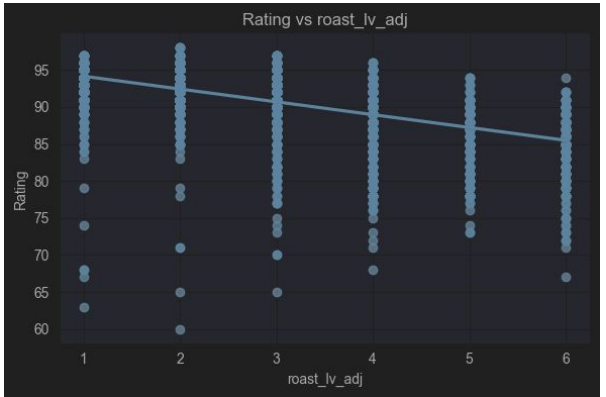












Linear Model
Baseline Model

Used Flavor as Baseline Model because it is the strongest predictor

- 0.658 R^2 and RMSE 1.256 on test set

Dep. Variable:	Rating	R-squared:	0.665
Model:	OLS	Adj. R-squared:	0.665
Method:	Least Squares	F-statistic:	7837.
Date:	Sun, 13 Apr 2025	Prob (F-statistic):	0.00
Time:	19:55:41	Log-Likelihood:	-6479.7
No. Observations:	3953	AIC:	1.296e+04
Df Residuals:	3951	BIC:	1.298e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	63.3516	0.331	191.453	0.000	62.703	64.000
Flavor	3.2910	0.037	88.524	0.000	3.218	3.364

Omnibus:	67.526	Durbin-Watson:	1.961
Prob(Omnibus):	0.000	Jarque-Bera (JB):	72.627
Skew:	-0.298	Prob(JB):	1.70e-16
Kurtosis:	3.295	Cond. No.	150.

Linear Model Stepwise Selection

Using Stepwise Selection: the final model includes Acidity, Aroma, Aftertaste, Flavor, and Body

- Test R^2: 0.9945 with RMSE of 0.158

Insight: Roast level and price of coffee does not affect coffee rating; order of influence on ratings is: flavor > acidity > aroma > aftertaste > body.

Dep. Variable:	Rating	R-squared:	0.993
Model:	OLS	Adj. R-squared:	0.993
Method:	Least Squares	F-statistic:	1.074e+05
Date:	Sun, 13 Apr 2025	Prob (F-statistic):	0.00
Time:	19:55:42	Log-Likelihood:	1084.5
No. Observations:	3953	AIC:	-2157.
Df Residuals:	3947	BIC:	-2119.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	49.9406	0.060	836.117	0.000	49.824	50.058
Acidity	0.9837	0.006	165.645	0.000	0.972	0.995
Aroma	1.0103	0.007	142.835	0.000	0.996	1.024
Aftertaste	1.0043	0.006	167.901	0.000	0.993	1.016
Flavor	1.0279	0.008	131.747	0.000	1.013	1.043
Body	0.9801	0.006	163.787	0.000	0.968	0.992

Omnibus:	2141.834	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3158086.446
Skew:	-1.005	Prob(JB):	0.00
Kurtosis:	141.455	Cond. No.	390.

Clustering

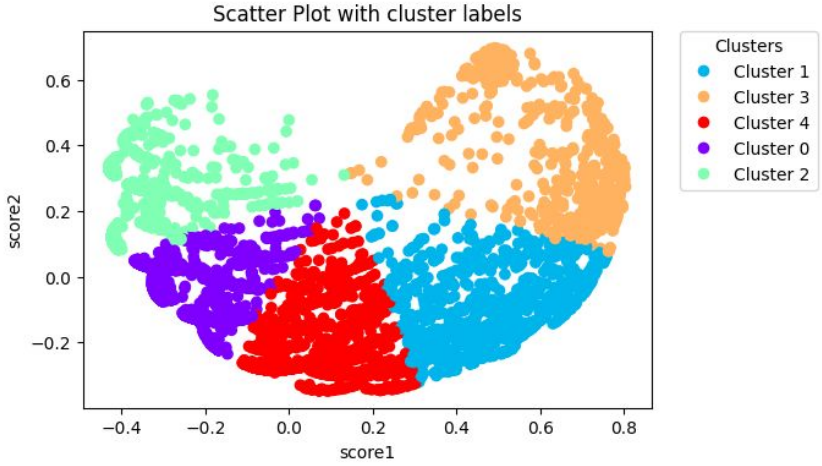
Xin Wang

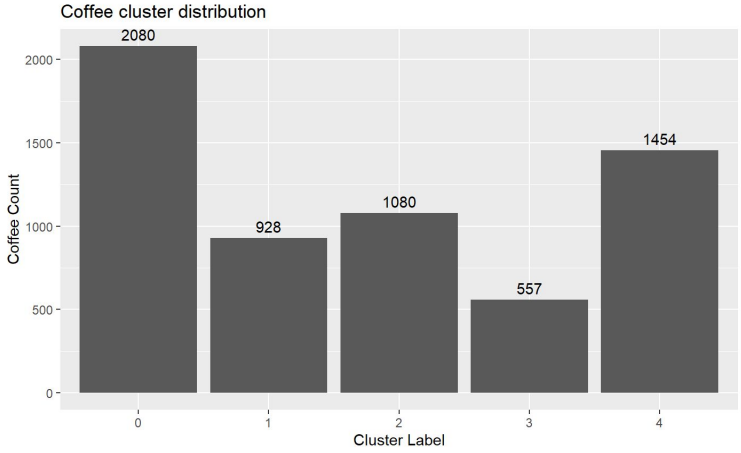
Features used for clustering

- **Roast Level:** Type of the roast
- **Agtron**: The smaller the number, the darker the roast
- **Aroma:** How intense and pleasurable when the nose first descends over the cup and is enveloped by fragrance
- **Acidity:** Higher - a lively, bright acidity
- **Body:** Higher - a richer, fuller mouthfeel
- **Flavor:** Encompasses all sensory experiences not covered by acidity, aroma, and body
- **Aftertaste:** The sensations that linger after the coffee has been swallowed (or spit out)

Dimensionality Reduction + KMeans

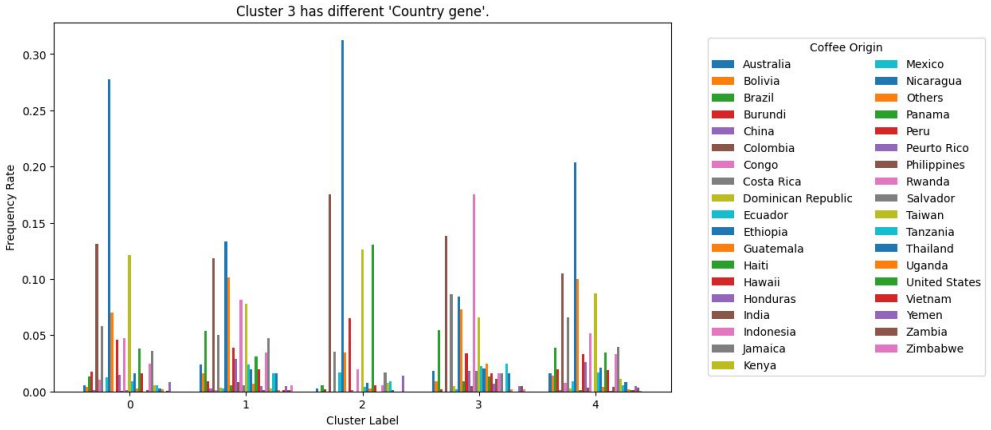
Kernel PCA-Silhouette Score: **0.4420**

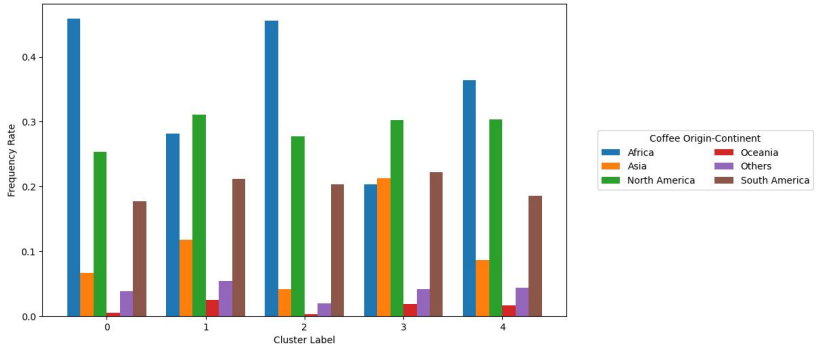




Cluster Analysis - Coffee Origin

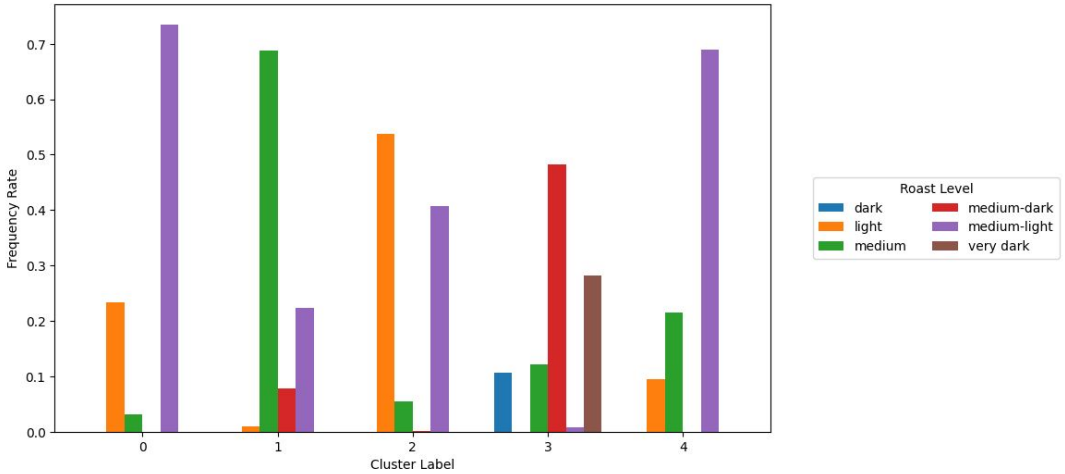
- **Cluster 0, 2, and 4** have similar Country distributions: High level proportion of Ethiopia coffee. But Cluster 2 has more American coffee
- **Clusters 1 and 3** tend to have more diverse distributions, but Cluster 3 has the highest proportion of coffee from Indonesia





- **Cluster 0, 2, and 4** have similar Continent distributions with high level proportion of African coffees
- **Cluster 1 and 3** tend to have more North American and Asian coffees

Cluster Analysis - Roast Level



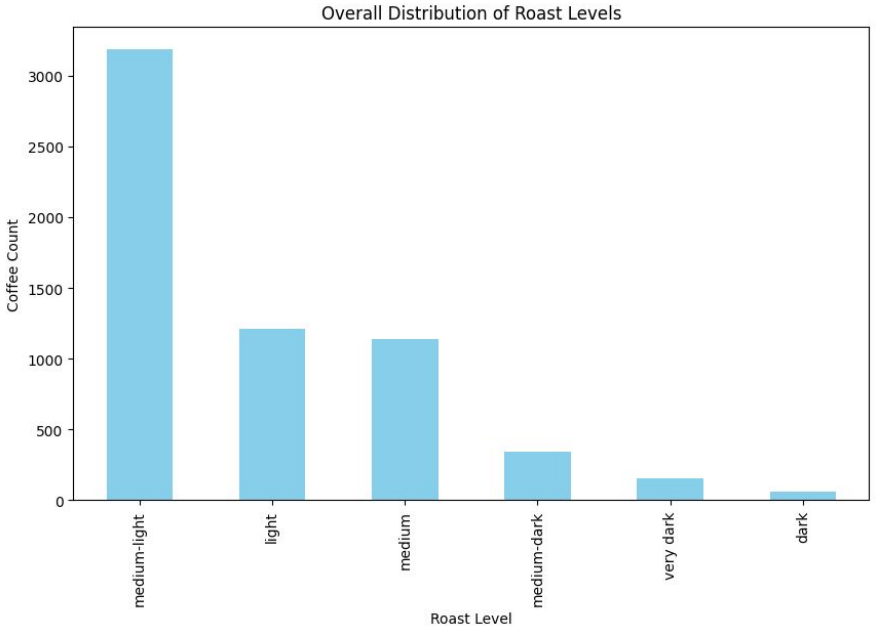
Cluster labels:

2 0 4 1 3

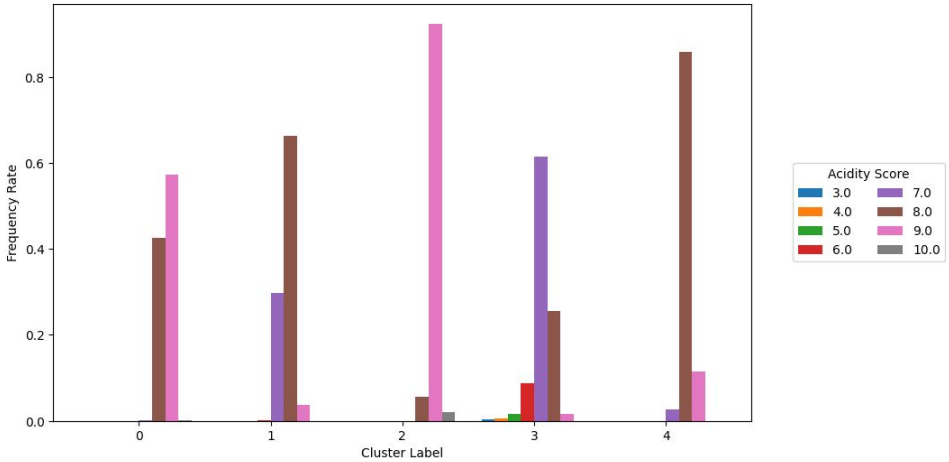
Light → Very dark

- **Clusters 0 & 4:** Predominantly medium-light roasts (purple), making up over 70%
- **Cluster 1:** Dominated by medium roasts (green), approximately 70%
- **Cluster 2:** Light roasts (orange) are predominant, over 50%
- **Cluster 3:** Most diverse roast profile, with medium-dark, very dark, and dark roasts all represented

Why not use roast level as cluster labels? – Imbalance!



Cluster Analysis-Acidity



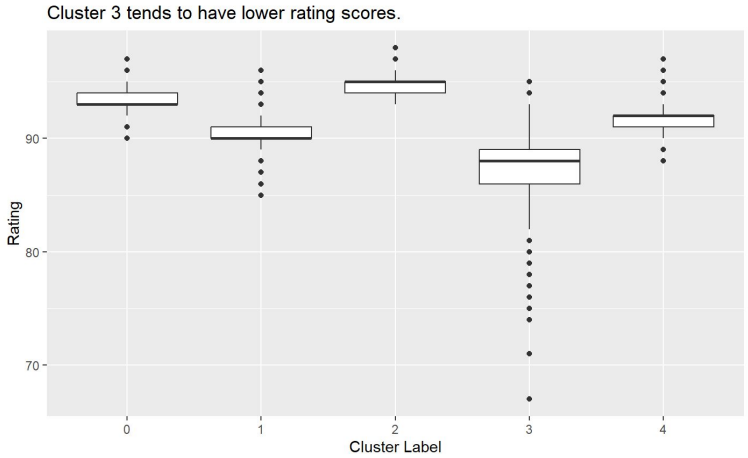
Cluster labels:

2 0 4 1 3

High acidity — Low acidity

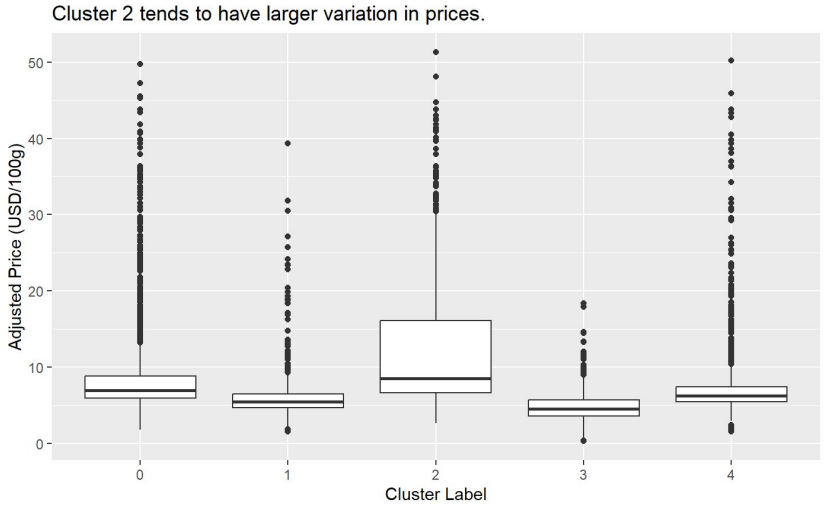
- **Cluster 0, 2, and 4:** 8.0 or 9.0
- **Cluster 1:** 7.0 / 8.0
- **Cluster 3:** 3.0-8.0

Cluster Analysis - Rating



- **Cluster 0 & 2** have similar higher ratings
- **Cluster 1 & 4** have similar ratings
- **Cluster 3** tends to have a lower rating

Cluster Analysis - Price



- **Cluster 2:** Largest price variation, with higher median and upper quartile than other clusters
- **Cluster 3:** Generally lower prices with minimal variation
- **Clusters 0, 1 & 4:** Moderate pricing, with Cluster 0 showing slightly wider distribution

Two Main Groupings

Premium group: Clusters 0, 2, and 4 (with 2 being somewhat distinctive in roast and price)

- High proportion of Ethiopian coffee
- African continent dominance
- Roast Level: Light - Medium
- Acidity score: 8.0/9.0

Difference: Cluster 2 tends to have higher ratings and higher price.

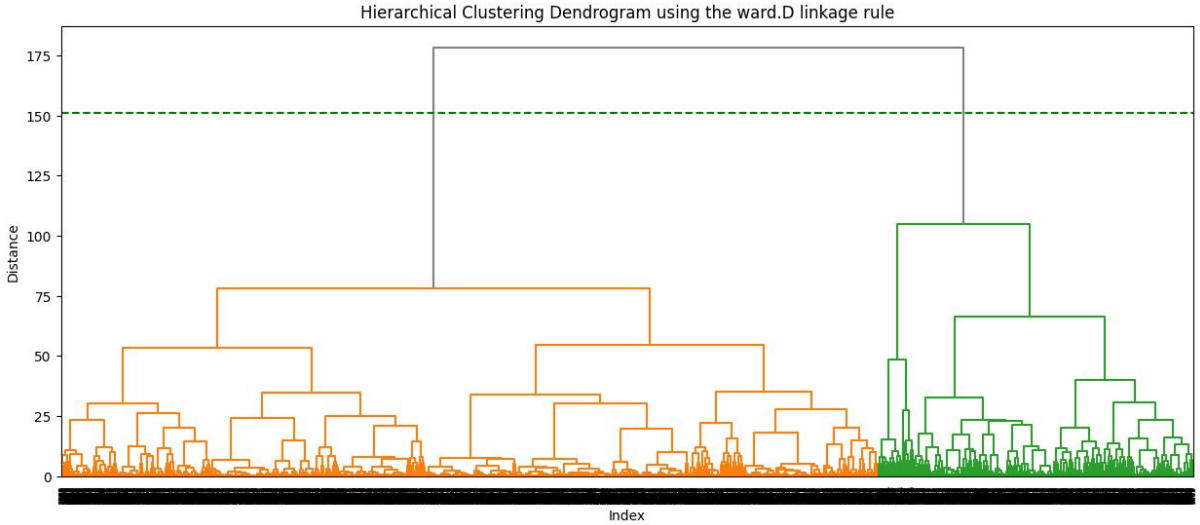
Diverse group: Clusters 1 and 3 (with 3 representing the lower-tier offering)

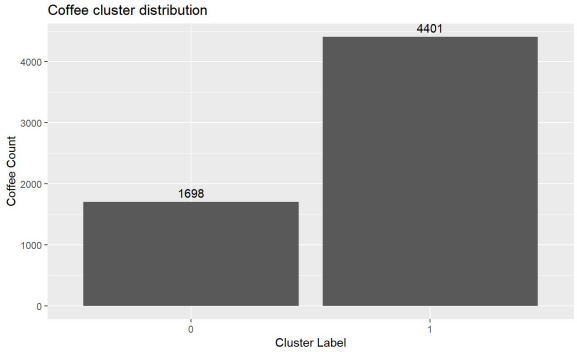
- More diverse country distributions
- Higher proportions of North American and Asian coffees
- Roast Level: Medium - Very dark
- Acidity score: 3.0-8.0

Difference: Cluster 3 has more Indonesia coffee and tends to be more dark roasted with lower ratings and slightly lower prices.

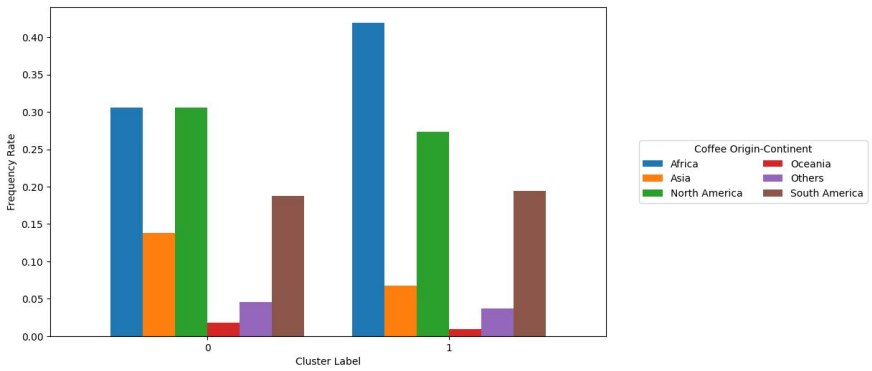
Hierarchical Clustering

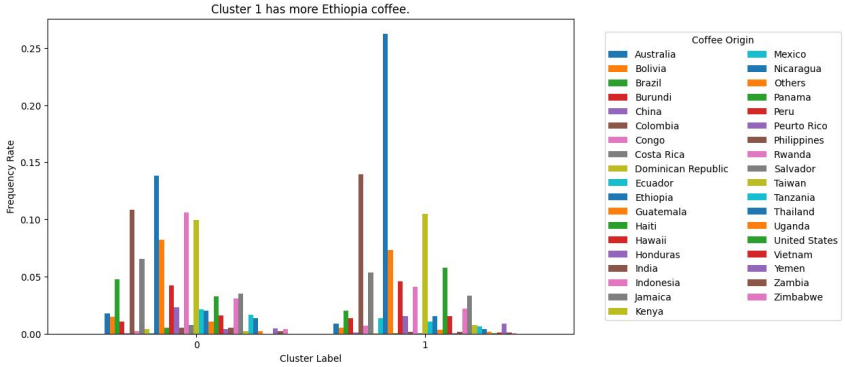
Largest height interval suggests #clusters = **2**





Cluster Analysis-Coffee Origin



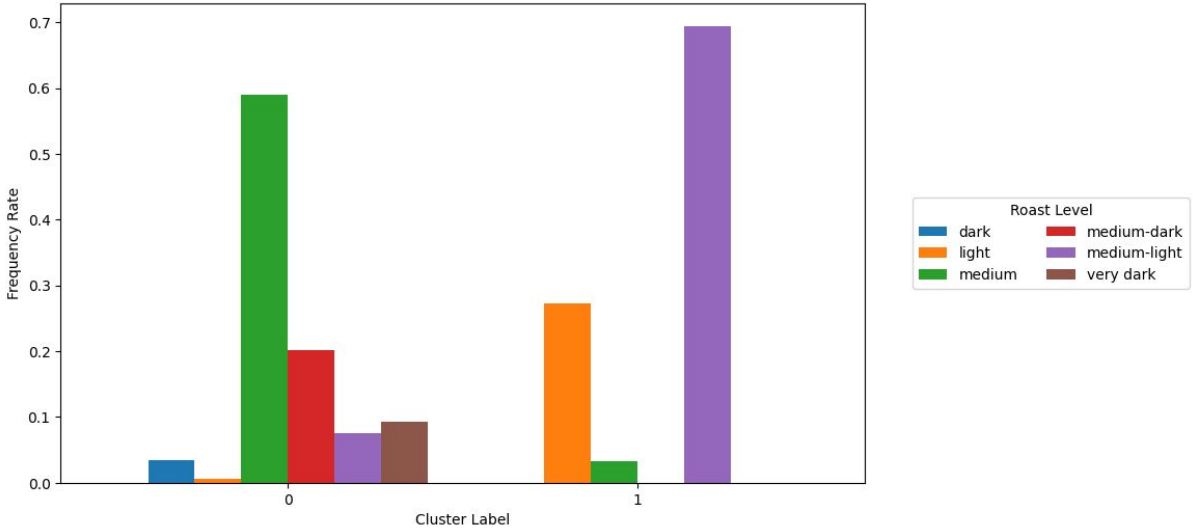


- **Cluster 0:** More balanced representation from other regions. South American coffees (like Brazil, Colombia) have a stronger presence, along with higher proportions from North American and Asian origins compared to Cluster 1.

- **Cluster 1:** Dominated by Ethiopian coffee (about 26%), with a higher overall proportion of African coffees (approximately 42%)

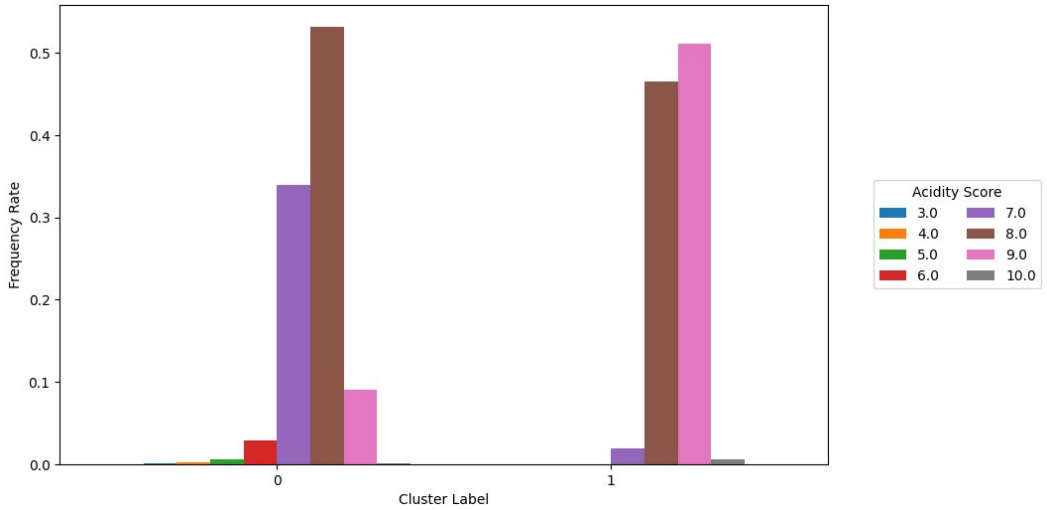
Cluster Analysis-Roast Level

- **Cluster 0:** Primarily medium roast (about 59%), followed by medium-dark (about 20%)
- **Cluster 1:** Predominantly medium-light roast (about 70%)

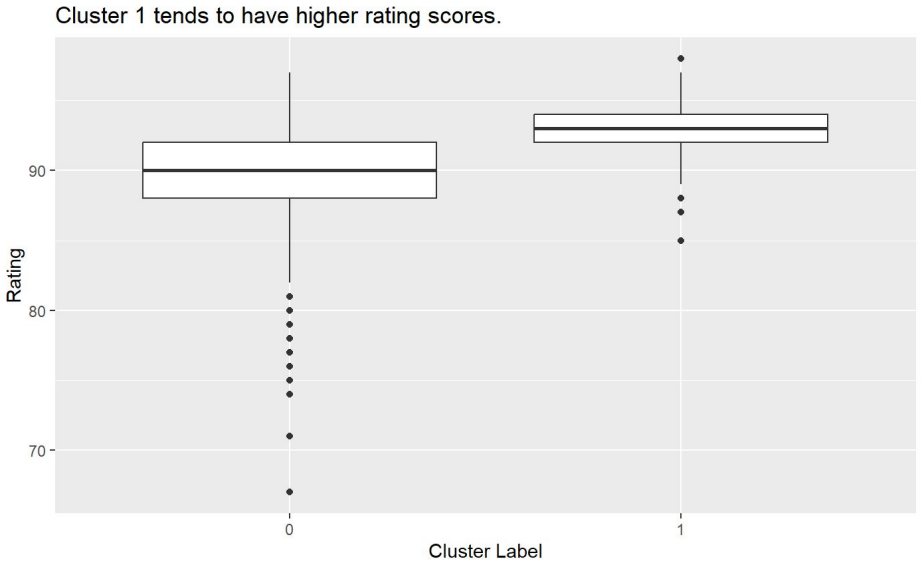


Cluster Analysis-Acidity

- **Cluster 0:** Lower acidity, mainly distributed in the 7.0-8.0 range
- **Cluster 1:** Higher acidity coffees, concentrated in the 8.0-9.0 range

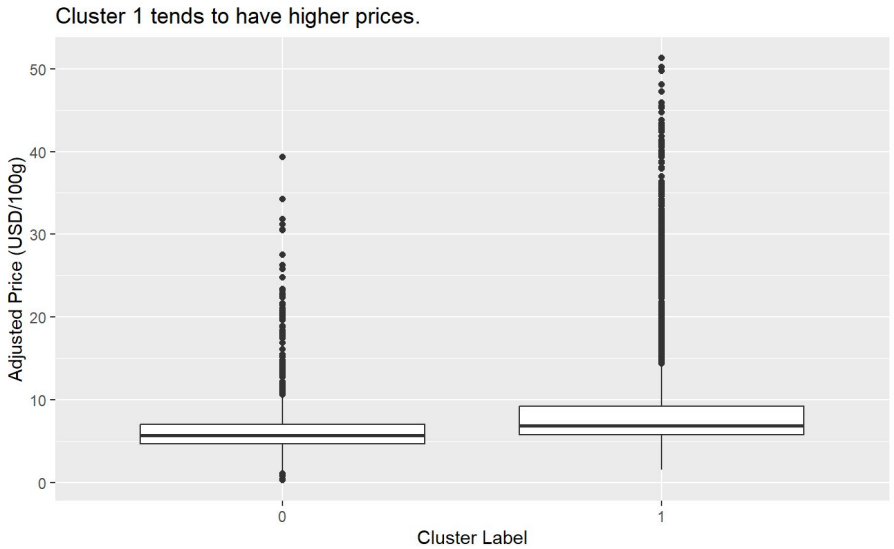


Cluster Analysis-Rating



- Cluster 0: Slightly lower ratings (median around 90 points), with more low-scoring outliers
- Cluster 1: Higher average ratings (median around 93 points), with a tighter distribution

Cluster Analysis-Price



- Cluster 0: Relatively lower prices, with a ceiling around 40 USD/100g
- Cluster 1: Higher average prices and higher price ceiling (up to 50 USD/100g)

Two Different Coffee Clusters

Cluster 0

- **Country:** More diverse distribution including Brazil, Colombia, and various origins
- **Continent:** Africa still significant (30%), but balanced with South America, North America, and Asia
- **Roast Level:** Medium - Very dark
- **Acidity:** Lower acidity, mainly in the 7.0-8.0 range
- **Rating:** Slightly lower ratings (median ~90 points) with more low-scoring outliers
- **Price:** Relatively lower prices, ceiling around 40 USD/100g

Cluster 1

- **Country:** Dominated by Ethiopian coffee
- **Continent:** Higher proportion of African coffees
- **Roast Level:** Light - Medium Light
- **Acidity:** Higher acidity, mainly in the 8.0-9.0 range
- **Rating:** Higher average ratings (median ~93 points) with tighter distribution
- **Price:** Higher average prices with ceiling up to 50 USD/100g

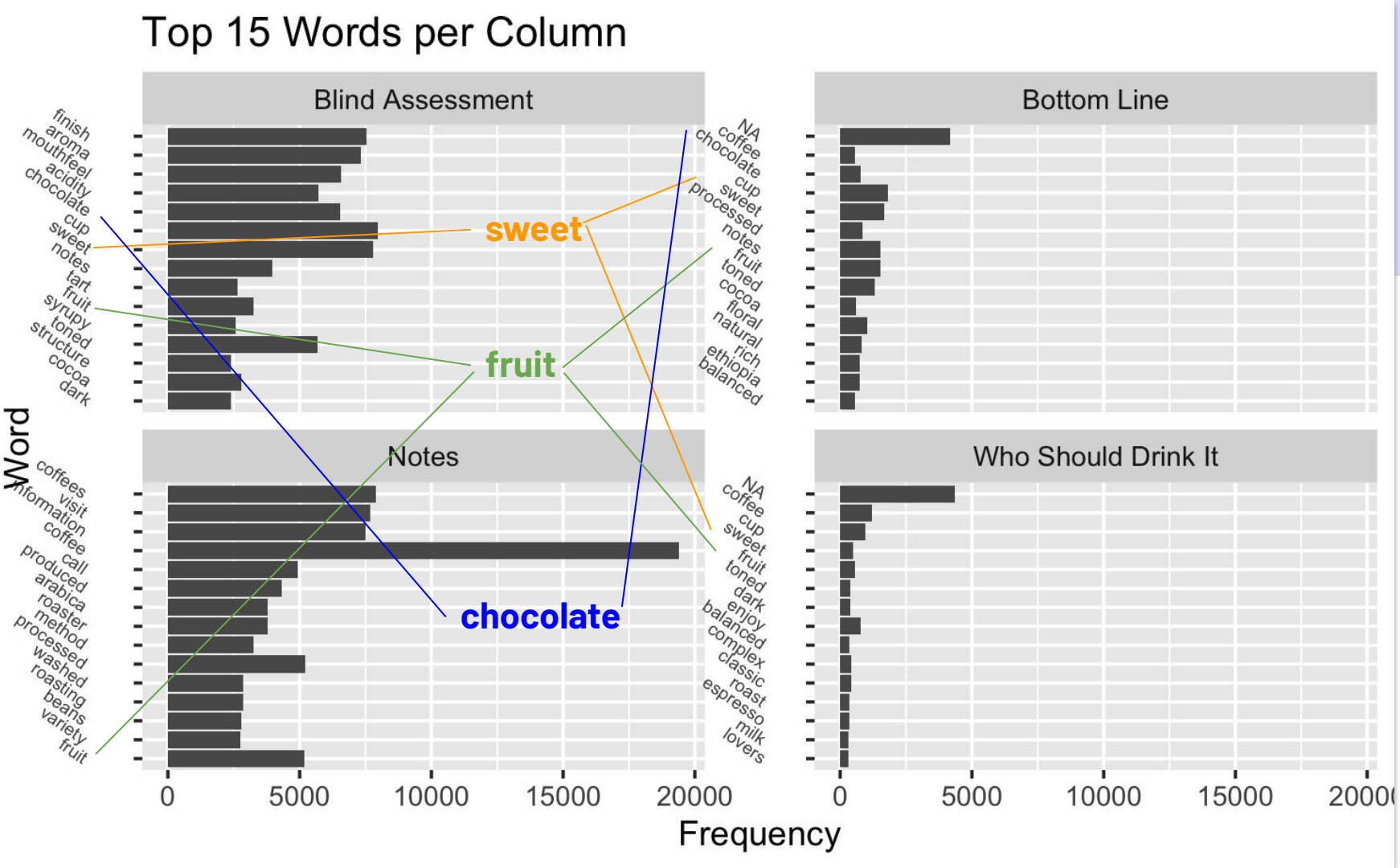
Text Mining & Sentiment Analysis

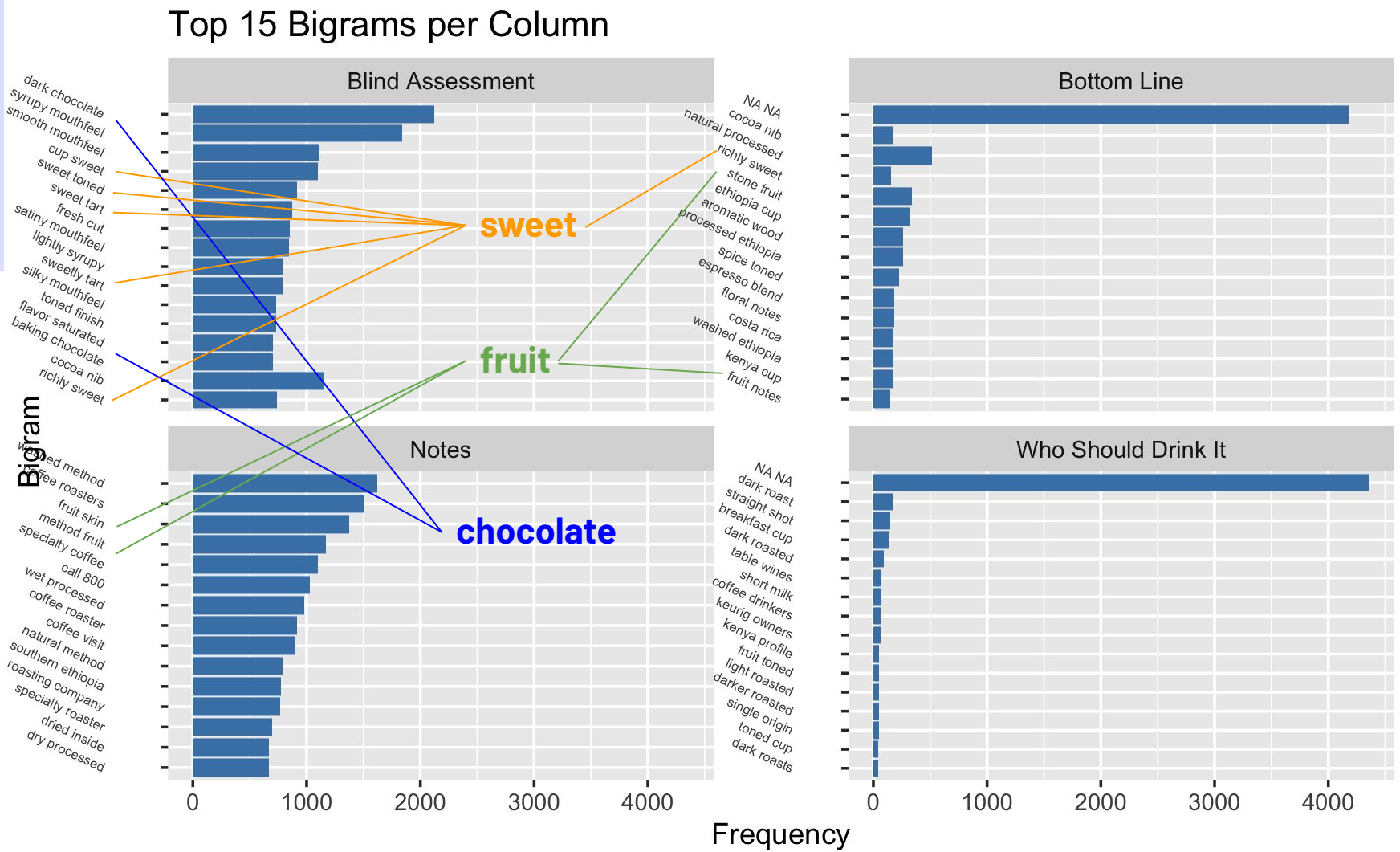
Henry Guo

Analyzing comments

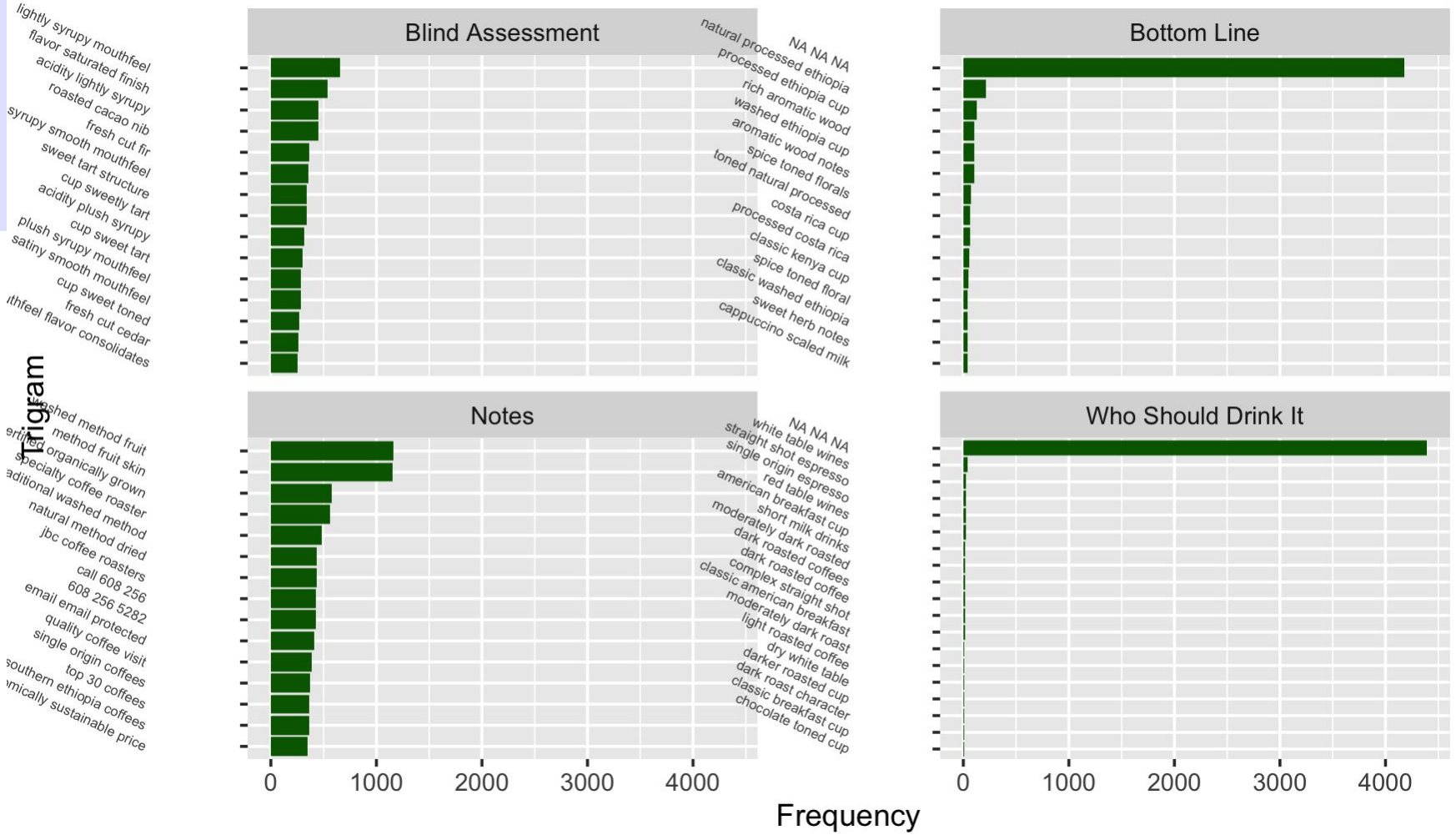
Total # Reviews = 8387

Column Name	Description	# Missing Val
`Blind Assessment`	An unbiased sensory evaluation of the coffee, describing its aroma, flavor, acidity, body, and finish without knowing its origin or brand.	0
`Notes`	Background information about the coffee, including its origin, variety, processing method, and roaster.	0
`Who should drink it`	Recommendations for the ideal audience based on the coffee's characteristics and appeal.	4360
`Bottom Line`	Summary of the coffee's key characteristics and overall impression	4181


Top 15 Words per Column
Blind Assessment
Bottom Line
Notes
Who Should Drink It
finish
aroma
mouthfeel
acidity
chocolate
cup
sweet
notes
tart
fruit
syrupy
toned
structure
cocoa
dark
NA
coffee
chocolate
cup
sweet
processed
notes
fruit
toned
cocoa
floral
natural
rich
ethiopia
balanced
sweet
fruit
chocolate
coffees
visit
information
coffee
call
produced
arabica
roaster
method
processed
washed
roasting
beans
variety
fruit
NA
coffee
cup
sweet
fruit
toned
dark
enjoy
balanced
complex
classic
roast
espresso
milk
lovers
Word
0
5000
10000
15000
20000
Frequency


Top 15 Bigrams per Column
Blind Assessment
Bottom Line
Notes
Who Should Drink It
Bigram
Frequency
0
1000
2000
3000
4000
dark chocolate
syrupy mouthfeel
smooth mouthfeel
cup sweet
sweet toned
sweet tart
fresh cut
satiny mouthfeel
lightly syrupy
sweetly tart
silky mouthfeel
toned finish
flavor saturated
baking chocolate
cocoa nib
richly sweet
sweet
fruit
chocolate
NA NA
cocoa nib
natural processed
richly sweet
stone fruit
ethiopia cup
aromatic wood
processed ethiopia
spice toned
espresso blend
floral notes
costa rica
washed ethiopia
kenya cup
fruit notes
fruit skin
method fruit
specialty coffee
call 800
wet processed
coffee roaster
coffee visit
natural method
southern ethiopia
roasting company
specialty roaster
dried inside
dry processed
NA NA
dark roast
straight shot
breakfast cup
dark roasted
table wines
short milk
coffee drinkers
keurig owners
kenya profile
fruit toned
light roasted
darker roasted
single origin
toned cup
dark roasts

Top 15 Trigrams per Column



LDA (k=5)

1	coffee	0.058322995
1	visit	0.019565391
1	information	0.017831364
1	coffees	0.016790953
1	call	0.016731829
1	certified	0.012290847
1	fair	0.009960473
1	roaster	0.009479369
1	roasting	0.009175254
1	grown	0.008295177

opic <int>	term <chr>	beta <dbl>
2	sweet	0.035670461
2	cup	0.035570792
2	finish	0.034687506
2	aroma	0.034013292
2	mouthfeel	0.032365094
2	chocolate	0.028662900
2	toned	0.026711227
2	acidity	0.026479072
2	notes	0.016751284
2	cocoa	0.014120424

3	coffee	0.055442978
3	coffees	0.018752386
3	call	0.018246416
3	visit	0.018228633
3	information	0.017587901
3	single	0.008993426
3	variety	0.008557345
3	roasters	0.008067268
3	green	0.007756272
3	roasting	0.007485296

Latent Dirichlet Allocation model on k topics
Identify latent optics in the reviews to understand common themes.

Beta: Probability of a word given a topic

Across k = 3,5,7, **sweet**, **chocolate**, and **fruit** are indeed the most occurred words.

4	coffee	0.047061686
4	processed	0.031079041
4	fruit	0.027313058
4	coffees	0.025798850
4	information	0.022592470
4	visit	0.022218978
4	produced	0.020213618
4	method	0.019159300
4	arabica	0.018688732
4	washed	0.017263998

5	*NA*	0.060529066
5	cup	0.024901507
5	fruit	0.022262029
5	sweet	0.019511643
5	coffee	0.017280381
5	notes	0.016879431
5	toned	0.015483137
5	chocolate	0.012463450
5	floral	0.012337909
5	rich	0.009432718

LDA (k=3)

1	*NA*	0.073165620
1	coffee	0.035681519
1	cup	0.010752950
1	blend	0.008878924
1	roast	0.008755833
1	fruit	0.008512989
1	roasted	0.007525728
1	natural	0.007292869
1	enjoy	0.007277225
1	espresso	0.007227489

2	cup	0.036722686
2	sweet	0.036236890
2	chocolate	0.028880374
2	finish	0.028652386
2	aroma	0.028257745
2	toned	0.027385958
2	mouthfeel	0.025216036
2	acidity	0.022217868
2	notes	0.020722100
2	fruit	0.016639285

3	coffee	0.052895916
3	coffees	0.022508872
3	visit	0.022433444
3	information	0.021874529
3	processed	0.015765577
3	fruit	0.014869401
3	call	0.014310327
3	produced	0.012866657
3	arabica	0.011182221
3	roaster	0.011052381

LDA (k = 7)

1	chocolate	0.027599779
1	cup	0.023513686
1	sweet	0.021446956
1	fruit	0.019427648
1	coffee	0.017421989
1	finish	0.016156720
1	aroma	0.015260836
1	notes	0.014553785
1	dark	0.012470983
1	roast	0.011586867

2	mouthfeel	0.036141430
2	sweet	0.035382379
2	finish	0.035230659
2	cup	0.035224402
2	aroma	0.035088710
2	acidity	0.028809926
2	toned	0.027392608
2	chocolate	0.025156329
2	notes	0.015813707
2	cocoa	0.015242742

3	coffee	0.046741818
3	visit	0.019233672
3	call	0.017828209
3	coffees	0.017251504
3	information	0.016226151
3	single	0.010419575
3	green	0.009829647
3	roasters	0.008741348
3	specialty	0.008721936
3	roasted	0.008098557

4	*NA*	0.139490688
4	coffee	0.036510280
4	information	0.020448523
4	visit	0.019780354
4	fair	0.018813961
4	certified	0.017115042
4	email	0.014037103
4	roaster	0.013879376
4	sustainable	0.013604761
4	trade	0.011446577

5	cup	0.036094694
5	fruit	0.030248532
5	sweet	0.028707878
5	toned	0.027159685
5	notes	0.024897203
5	floral	0.020916108
5	natural	0.013534491
5	ethiopia	0.013448574
5	classic	0.013232593
5	balanced	0.012656137

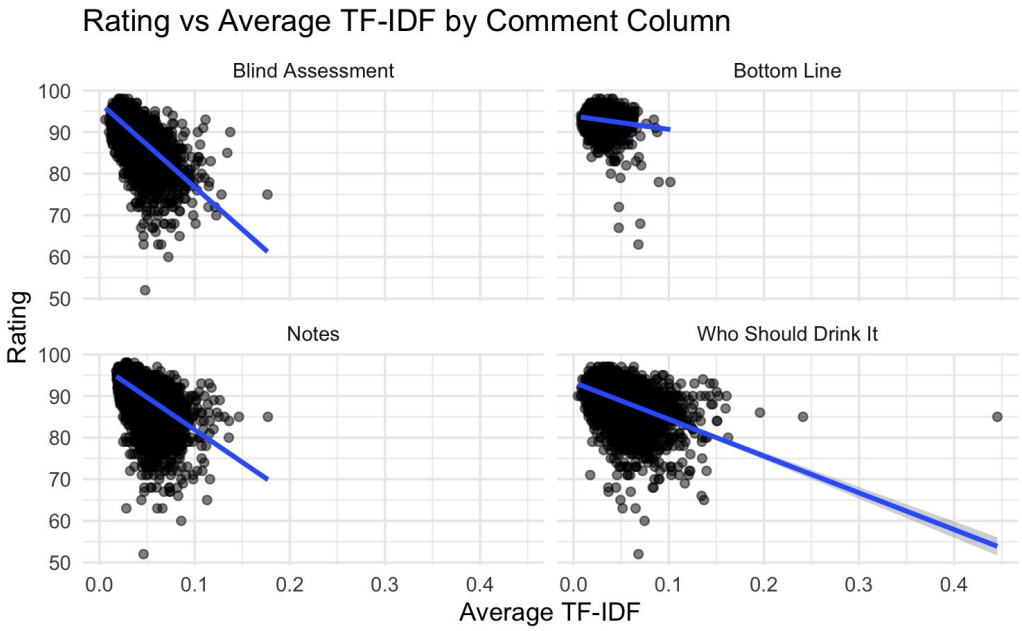
6	coffee	0.050286925
6	processed	0.029570726
6	fruit	0.026972527
6	coffees	0.023727769
6	information	0.020436669
6	visit	0.020422277
6	produced	0.019175609
6	arabica	0.018719897
6	method	0.017506645
6	washed	0.014812287

7	coffee	0.066490583
7	coffees	0.027676726
7	information	0.019372278
7	visit	0.019254265
7	call	0.018153658
7	roasting	0.011664256
7	produced	0.009794678
7	fruit	0.009396136
7	arabica	0.008875703
7	roaster	0.008820992

TF-IDF

Term Frequency–Inverse Document Frequency scores for each word represents each word's rarity.

High score means that a word is not only frequent in a particular review but also relatively rare elsewhere. This makes such words more **distinctive in its column**.

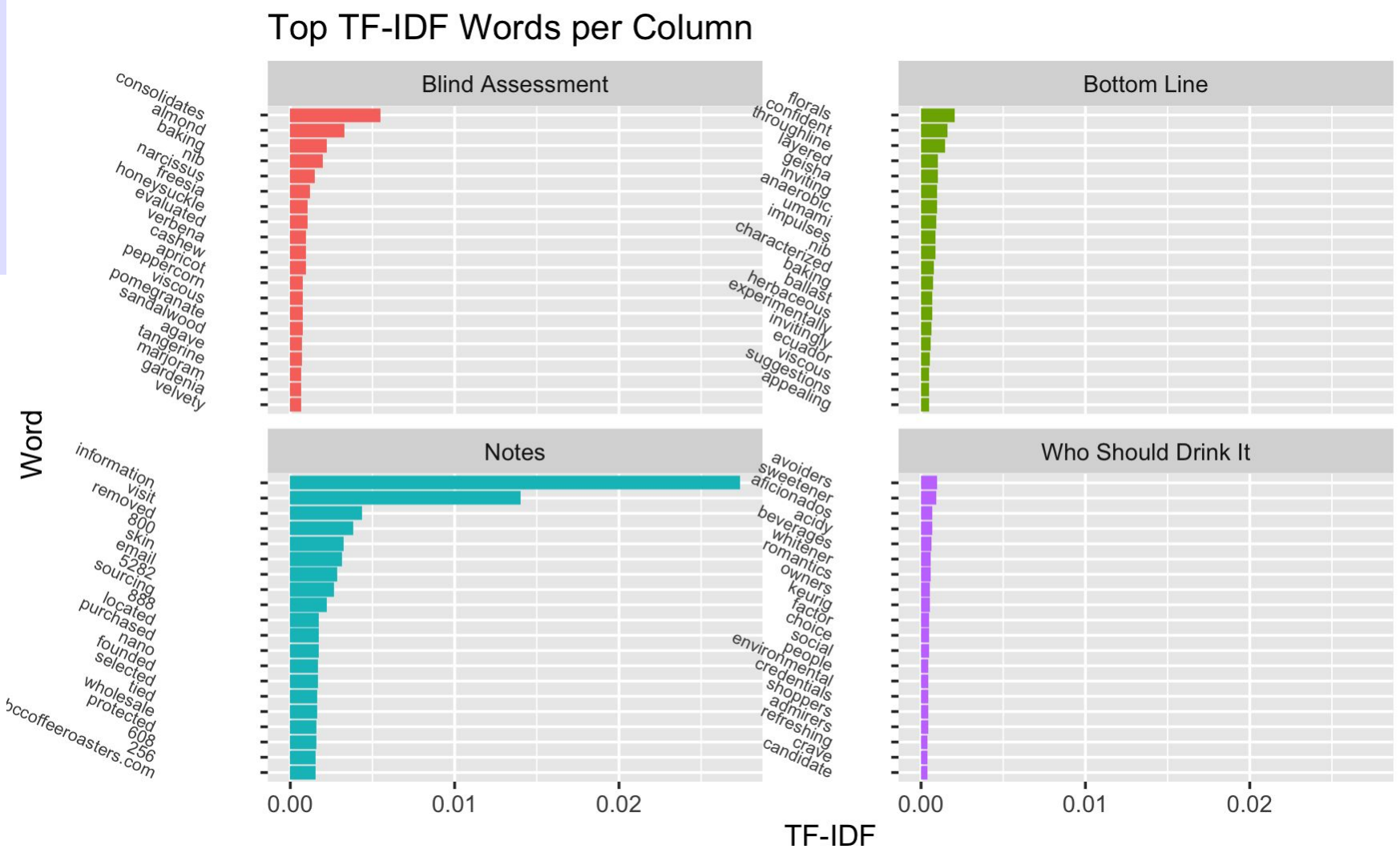


A **higher average TF-IDF** could indicate that the review contains more unique or less common language.

Negative correlation with Rating: Reviewers tend to give lower ratings when using distinctive comments.

For each comment column, Pearson's correlation coefficient between average TF-IDF values and Rating

document <chr>	correlation <dbl>
Blind Assessment	-0.6529444
Bottom Line	-0.1595291
Notes	-0.5541167
Who Should Drink It	-0.4484817

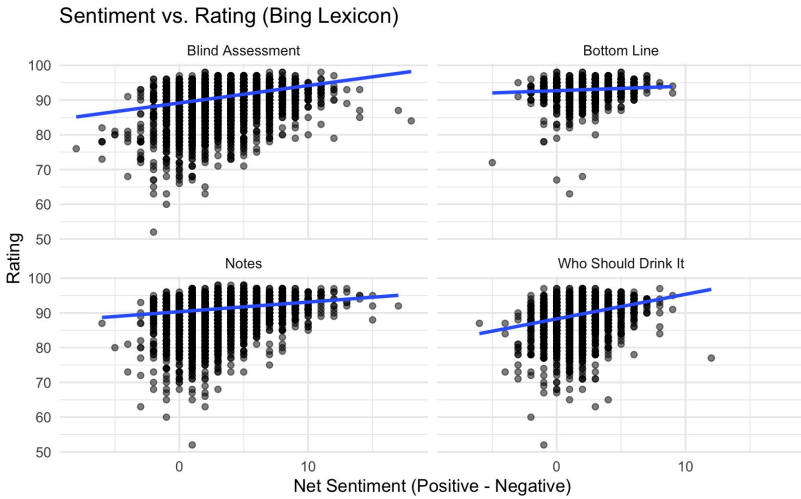

Top TF-IDF Words per Column
Blind Assessment
consolidates
almond
baking
nib
narcissus
freesia
honeysuckle
evaluated
verbena
cashew
apricot
peppercorn
viscous
pomegranate
sandalwood
agave
tangerine
marjoram
gardenia
velvety
florals
confident
throughline
layered
geisha
inviting
anaerobic
umami
impulses
nib
characterized
baking
ballast
herbaceous
experimentally
invitingly
ecuador
viscous
suggestions
appealing
Bottom Line
Notes
information
visit
removed
800
skin
email
5282
sourcing
888
located
purchased
nano
founded
selected
tied
wholesale
protected
608
256
bccoffeeroasters.com
avoiders
sweetener
aficionados
acidy
beverages
whitener
romantics
owners
keurig
factor
choice
social
people
environmental
credentials
shoppers
admirers
refreshing
crave
candidate
Who Should Drink It
Word
0.00
0.01
0.02
TF-IDF

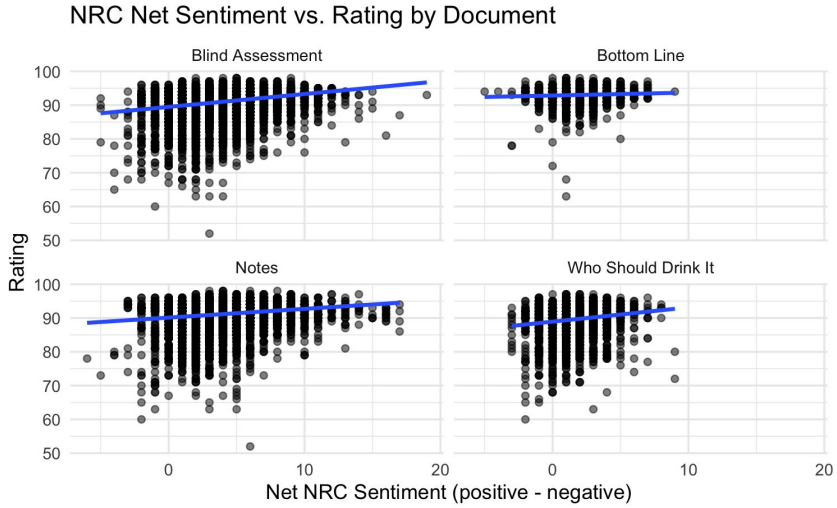
Sentiment Analysis

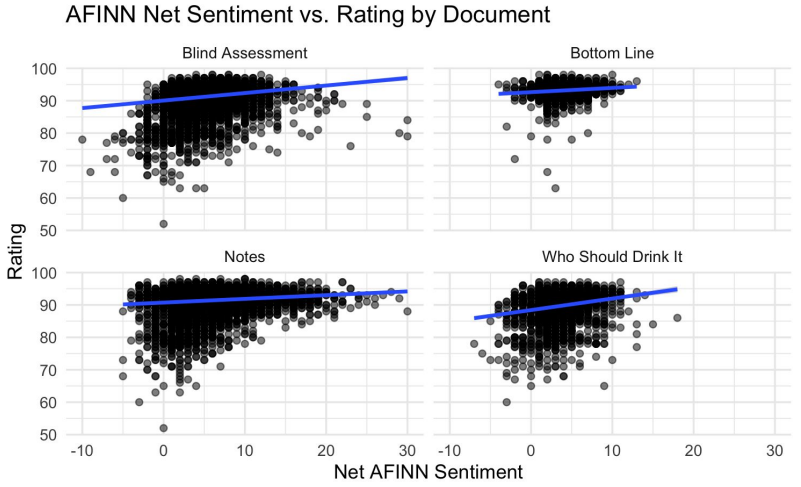
A positive net sentiment means that, overall, the words in that review contribute positive values and vice versa.

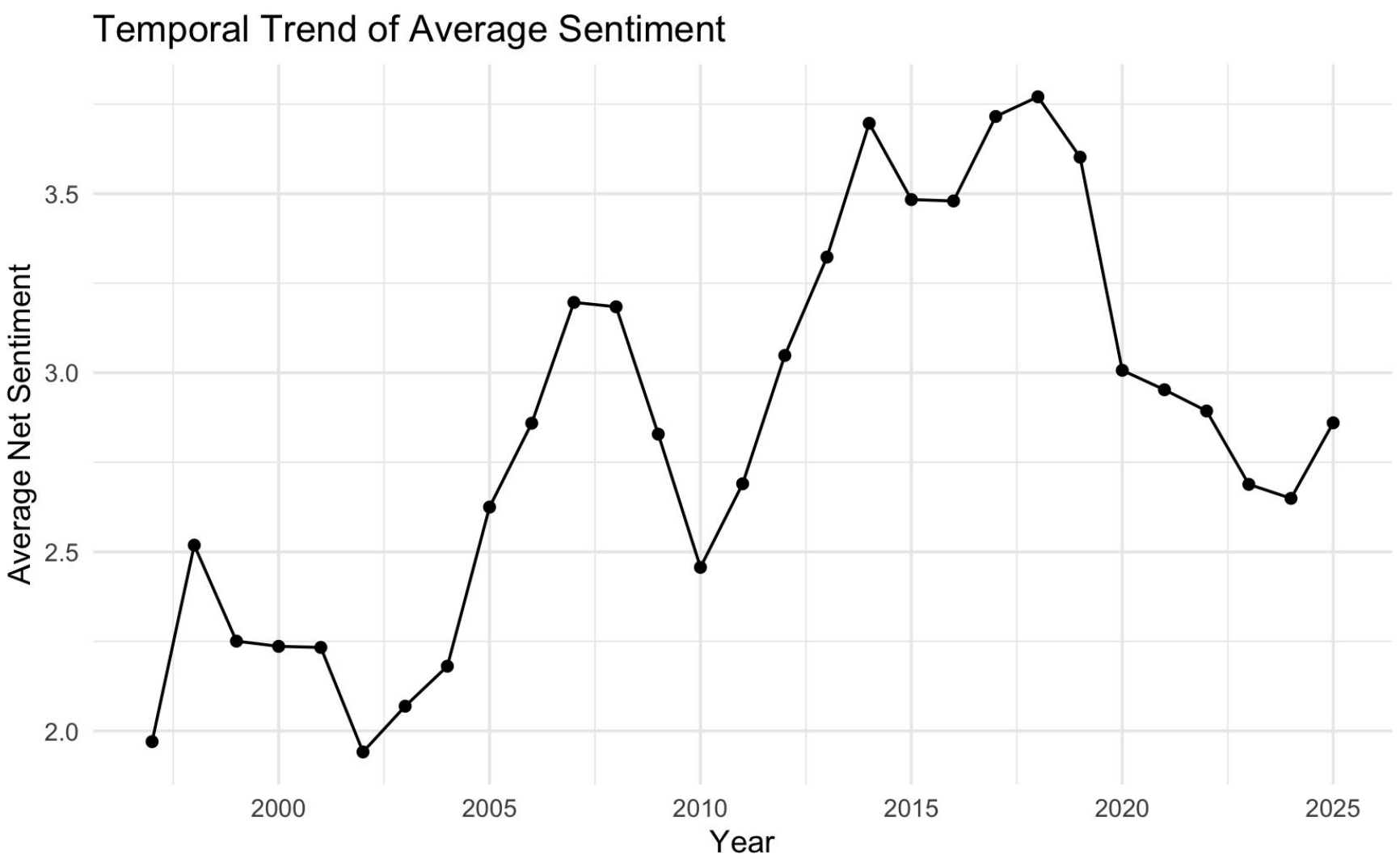
Lexicon	Method	Notes
Bing	Word <- positive / negative	Net sentiment = (total # positive words - total # negative words)
AFINN	Provides numeric sentiment scores (weight) for words	Net sentiment = sum(word scores in a cell)
NRC	Assigns words to multiple emotion categories, explore the mix of emotions present in the reviews.	Similar to Bing

Sentiment & Rating has a positive correlation








Temporal Trend of Average Sentiment
Average Net Sentiment
3.5
3.0
2.5
2.0
2000
2005
2010
2015
2020
2025
Year

Discussion

- To achieve higher coffee scores, it is crucial to optimize both flavor and acidity.

- Specifically, enhancing the coffee's natural sweetness, highlighting its fruit-forward notes, and incorporating chocolate nuances are key elements for success.