COFFEE SHOP REVIEWS:

A case study in Austin, Texas

Dickinson Department of Data Analytics
Ziwei Guo ‘24 & Tai Nguyen ’25
May 2nd, 2024

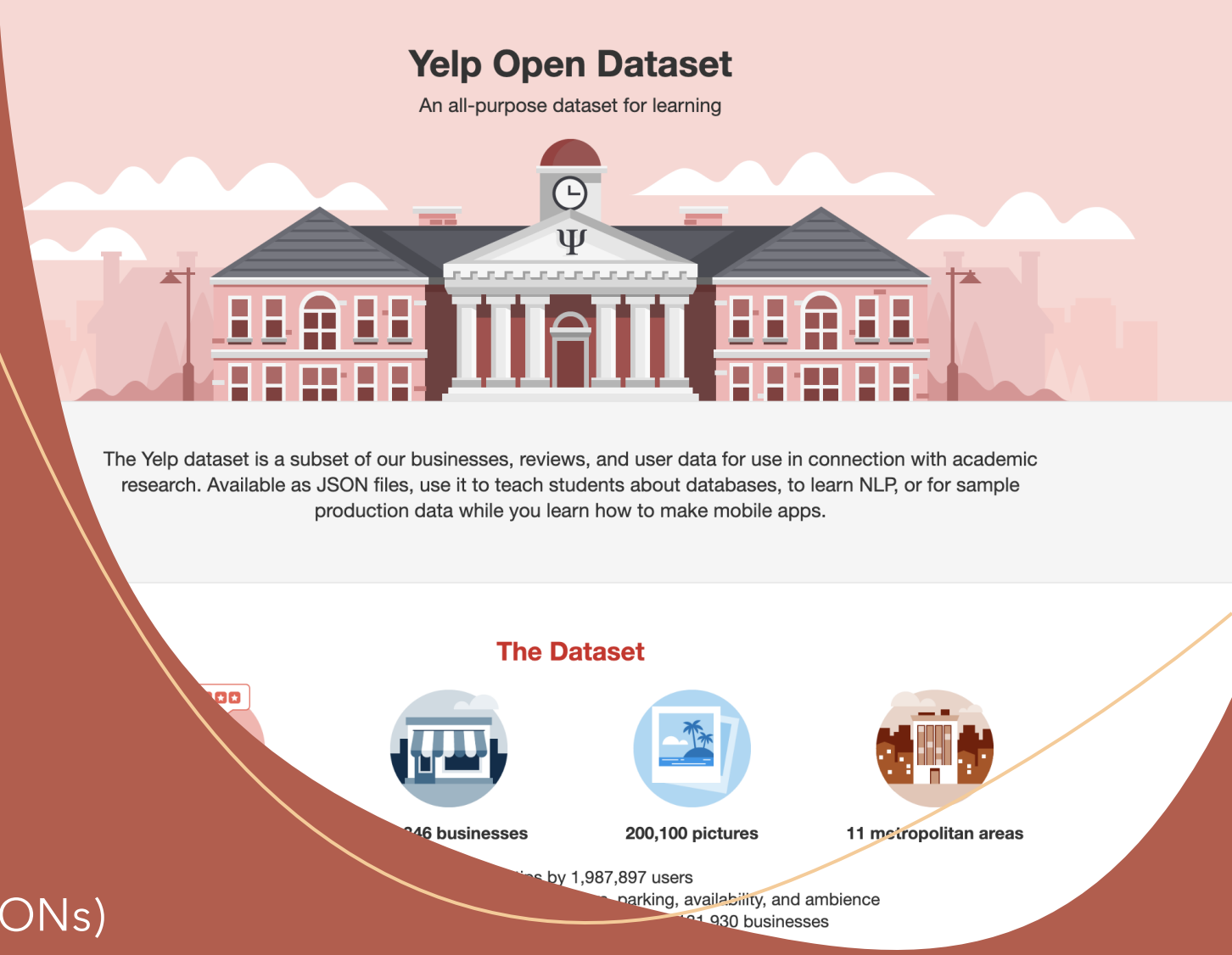
Agenda

I. Research Introduction & Objectives
II. Data Source & Description
III. Exploratory Data Analysis (EDA)
IV. Statistical Methodologies
V. Implications for Stakeholders
VI. Ethical, Legal and Social Implications
VII. QnA

I. Research Introduction & Objectives

- In an era where consumer opinions hold significant sway over business success, understanding customer sentiments and preferences is essential. This project delves into the rich resource of Yelp reviews to decode the nuances of customer experiences at coffee shops in Austin, TX or coffee businesses at large.
- By employing a variety of statistical techniques and modeling, we seek to understand factors that influence overall ratings and categorical ratings of coffee shops, thereby providing actionable insights through which coffee shop owners may adopt to optimize business strategies.

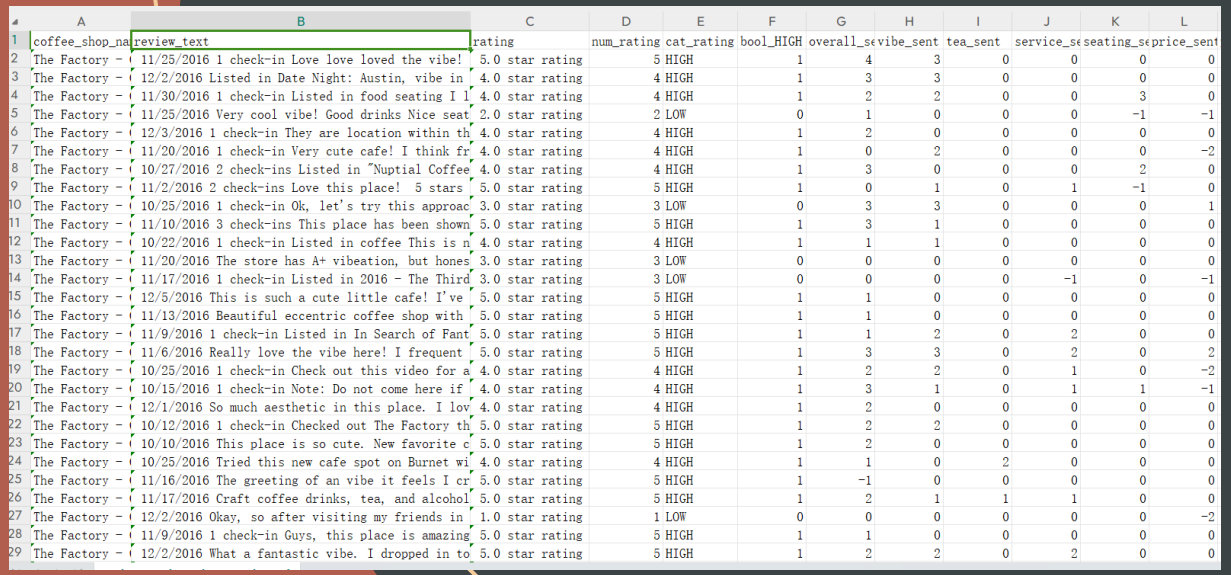
II. Data Source



- Original: Yelp's Open Dataset
- Problem: too large (4GB and all in JSONs)
- Selected for use: Kaggle's scraped version https://www.kaggle.com/datasets/sripaadsrinivasan/yelp-coffee-reviews/code

Data Description

- 3 datasets: raw reviews, reviews with ratings and sentiment scores, sentiment rates by shops
- 7800 rows of scraped individual reviews from 66 coffee shops
- 20 columns or variables consisted of 4 datatypes:
  - Numerical: ratings, various sentiment metrics such as wifi, location and seating
  - Categorical: 'high' for rating greater than 3, 'low' otherwise
  - String (Text): shop names, reviews
  - Date : date and time of review, extracted



	A	B	C	D	E	F	G	H	I	J	K	L
1	coffee_shop_na	review_text	rating	num_rating	cat_rating	bool_HIGH	overall_se	vibe_sent	tea_sent	service_se	seating_se	price_sent
2	The Factory –	11/25/2016 1 check-in Love love loved the vibe!	5.0 star rating	5	HIGH	1	4	3	0	0	0	0
3	The Factory –	12/2/2016 Listed in Date Night: Austin, vibe in	4.0 star rating	4	HIGH	1	3	3	0	0	0	0
4	The Factory –	11/30/2016 1 check-in Listed in food seating I l	4.0 star rating	4	HIGH	1	2	2	0	0	3	0
5	The Factory –	11/25/2016 Very cool vibe! Good drinks Nice seat	2.0 star rating	2	LOW	0	1	0	0	0	-1	-1
6	The Factory –	12/3/2016 1 check-in They are location within th	4.0 star rating	4	HIGH	1	2	0	0	0	0	0
7	The Factory –	11/20/2016 1 check-in Very cute cafe! I think fr	4.0 star rating	4	HIGH	1	0	2	0	0	0	-2
8	The Factory –	10/27/2016 2 check-ins Listed in "Nuptial Coffee	4.0 star rating	4	HIGH	1	3	0	0	0	2	0
9	The Factory –	11/2/2016 2 check-ins Love this place! 5 stars	5.0 star rating	5	HIGH	1	0	1	0	1	-1	0
10	The Factory –	10/25/2016 1 check-in Ok, let's try this approac	3.0 star rating	3	LOW	0	3	3	0	0	0	1
11	The Factory –	11/10/2016 3 check-ins This place has been shown	5.0 star rating	5	HIGH	1	3	1	0	0	0	0
12	The Factory –	10/22/2016 1 check-in Listed in coffee This is n	4.0 star rating	4	HIGH	1	1	1	0	0	0	0
13	The Factory –	11/20/2016 The store has A+ vibeation, but hones	3.0 star rating	3	LOW	0	0	0	0	0	0	0
14	The Factory –	11/17/2016 1 check-in Listed in 2016 – The Third	3.0 star rating	3	LOW	0	0	0	0	-1	0	-1
15	The Factory –	12/5/2016 This is such a cute little cafe! I've	5.0 star rating	5	HIGH	1	1	0	0	0	0	0
16	The Factory –	11/13/2016 Beautiful eccentric coffee shop with	5.0 star rating	5	HIGH	1	1	0	0	0	0	0
17	The Factory –	11/9/2016 1 check-in Listed in In Search of Fant	5.0 star rating	5	HIGH	1	1	2	0	2	0	0
18	The Factory –	11/6/2016 Really love the vibe here! I frequent	5.0 star rating	5	HIGH	1	3	3	0	2	0	2
19	The Factory –	10/25/2016 1 check-in Check out this video for a	4.0 star rating	4	HIGH	1	2	2	0	1	0	-2
20	The Factory –	10/15/2016 1 check-in Note: Do not come here if	4.0 star rating	4	HIGH	1	3	1	0	1	1	-1
21	The Factory –	12/1/2016 So much aesthetic in this place. I lov	4.0 star rating	4	HIGH	1	2	0	0	0	0	0
22	The Factory –	10/12/2016 1 check-in Checked out The Factory th	5.0 star rating	5	HIGH	1	2	2	0	0	0	0
23	The Factory –	10/10/2016 This place is so cute. New favorite c	5.0 star rating	5	HIGH	1	2	0	0	0	0	0
24	The Factory –	10/25/2016 Tried this new cafe spot on Burnet wi	4.0 star rating	4	HIGH	1	1	0	2	0	0	0
25	The Factory –	11/16/2016 The greeting of an vibe it feels I cr	5.0 star rating	5	HIGH	1	-1	0	0	0	0	0
26	The Factory –	11/17/2016 Craft coffee drinks, tea, and alcohol	5.0 star rating	5	HIGH	1	2	1	1	1	0	0
27	The Factory –	12/2/2016 Okay, so after visiting my friends in	1.0 star rating	1	LOW	0	0	0	0	0	0	-2
28	The Factory –	11/9/2016 1 check-in Guys, this place is amazing	5.0 star rating	5	HIGH	1	1	0	0	0	0	0
29	The Factory –	12/2/2016 What a fantastic vibe. I dropped in to	5.0 star rating	5	HIGH	1	2	2	0	2	0	0

III. Exploratory Data Analysis (EDA)

Summary Statistics

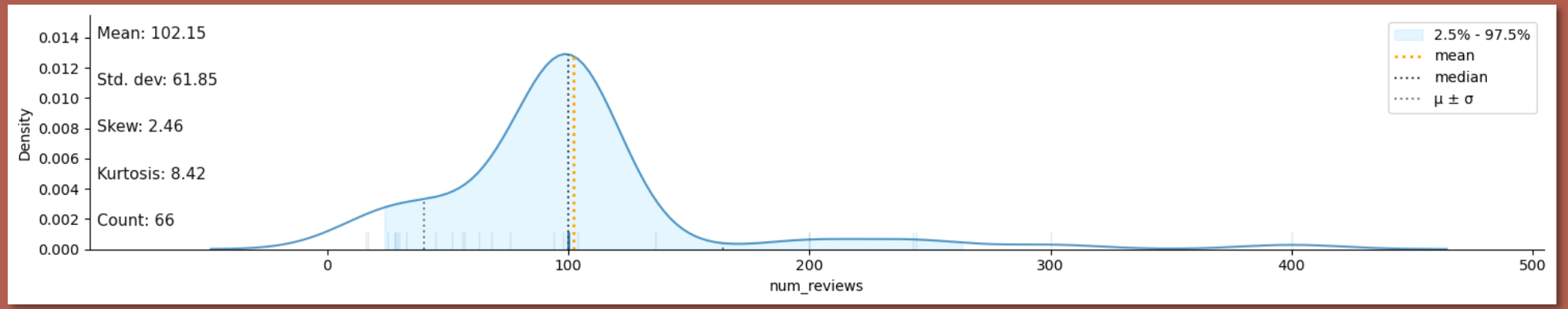
```
In [6]: ratesent.describe()
```

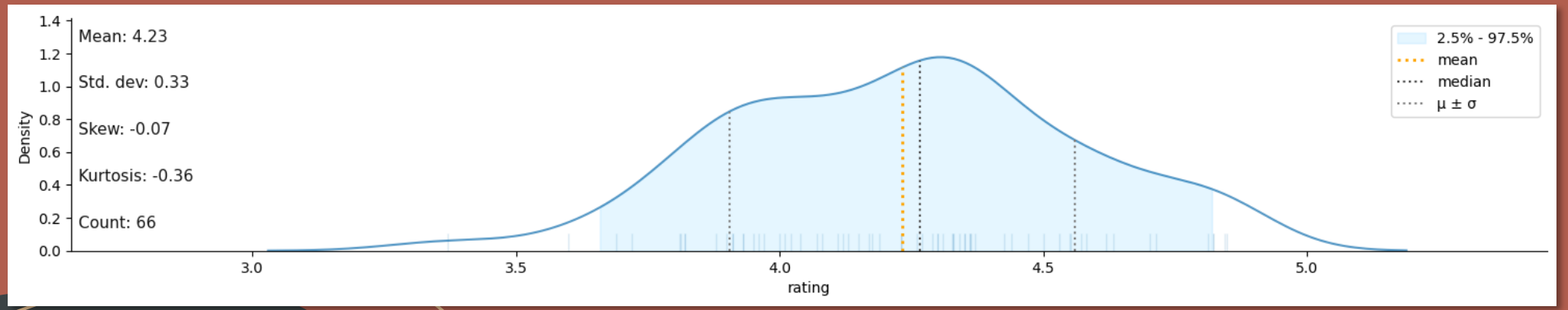
Out[6]:

	num_rating	bool_HIGH	overall_sent	tea_sent	service_sent	seating_sent	price_sent	location_sent	alcohol_sent	hours_sent	internet_sent	local_sent
count	7616.000000	7616.000000	7616.000000	7616.000000	7616.000000	7616.000000	7616.000000	7616.000000	7616.000000	7615.000000	7616.000000	7616.000000
mean	4.169118	0.806197	1.107537	0.047006	0.325105	0.124869	0.015362	0.074711	0.042936	0.031779	0.025210	0.035583
std	1.065311	0.395302	1.177984	0.330775	0.827549	0.521658	0.381999	0.395392	0.298598	0.274642	0.277679	0.271992
min	1.000000	0.000000	-4.000000	-3.000000	-4.000000	-3.000000	-3.000000	-4.000000	-3.000000	-3.000000	-3.000000	-1.000000
25%	4.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	4.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	5.000000	1.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	5.000000	1.000000	4.000000	4.000000	4.000000	4.000000	3.000000	4.000000	3.000000	3.000000	3.000000	4.000000

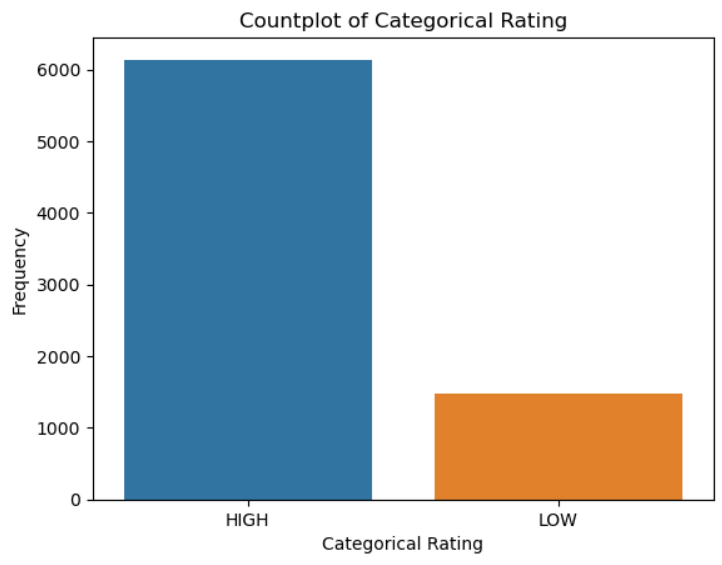
- Mean number of reviews per shop: 102
- Mean rating: 4.16 (on scale of 5)
- Various sentiment scores range from -4 or -3 to 3 or 4 with an average of:
  - 1.12 overall
  - 0.05 for tea drinks
  - 0.12 for seating
  - 0.01 for price
  - 0.07 for location
  - etc.

Distributions of number of reviews and ratings across 66 coffee shops

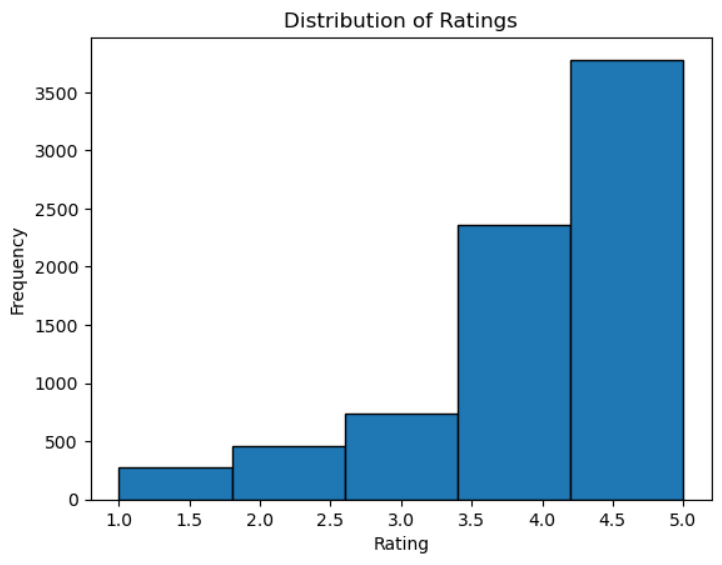




Distribution of Ratings



Categorical Rating

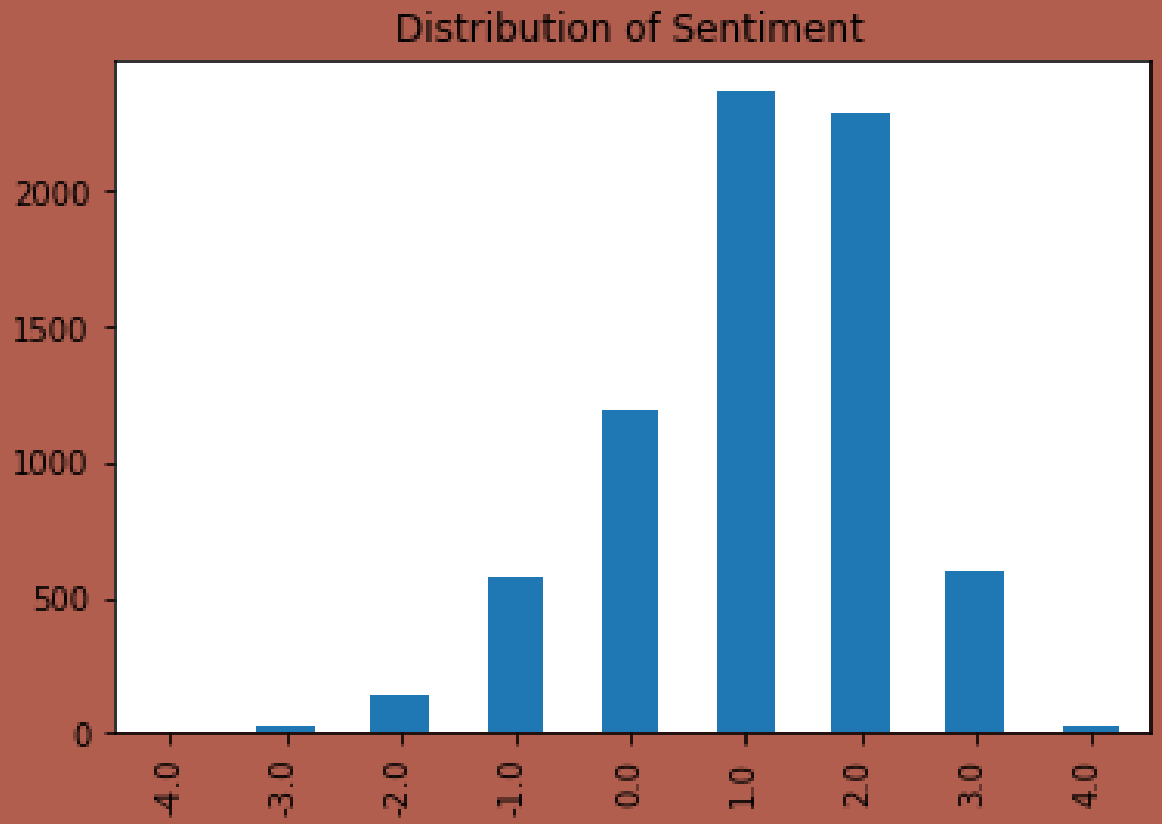


Numeric Rating

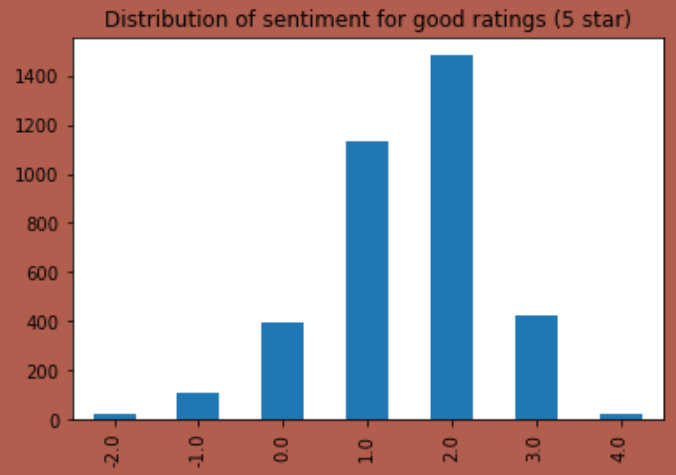
Average rating from all reviews of Austin coffee shops is 4.169 out of 5. The distribution above also shows that online reviews are heavily positively skewed, with over 80.7% of reviews being 4 or 5 stars.

Distribution of Sentiments

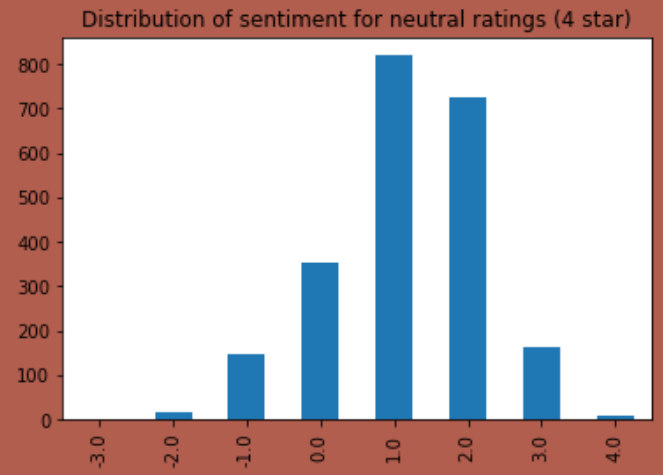
- The average overall sentiment of a review was 1.107, meaning slightly positive.
- 73.2% of reviews are positive (> 0 overall sentiment), as compared to 16.6% neutral and 10.2% negative.
- Review sentiments are rarely "extreme" with a sentiment greater than or equal to 3, or less than or equal to -3.
- This suggests that reviews on coffee shops are overall positive and comparatively moderate in their sentiment.



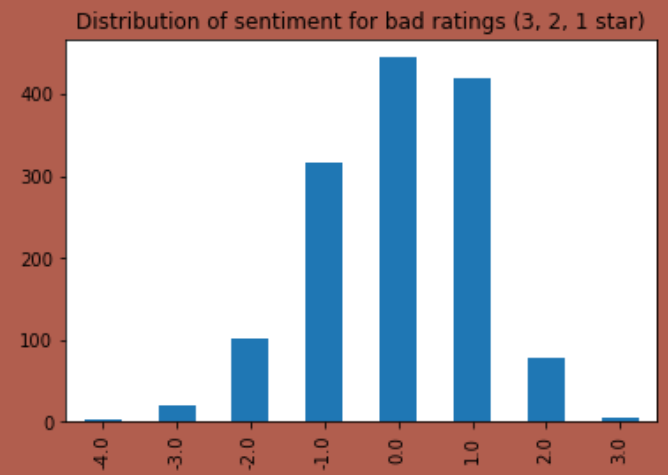
Does sentiment vary based on ‘good’, ‘bad’ and ‘neutral’ ratings?



'Good’ ratings



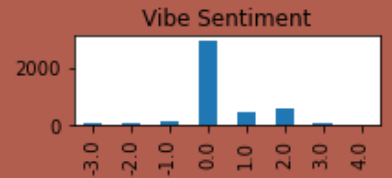
‘Neutral’ ratings

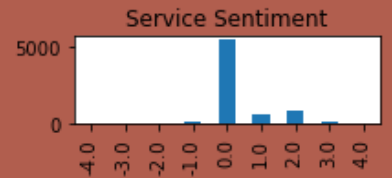


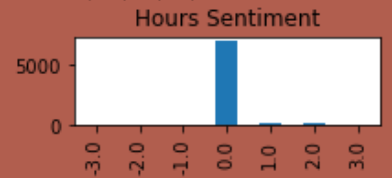
‘Bad’ ratings

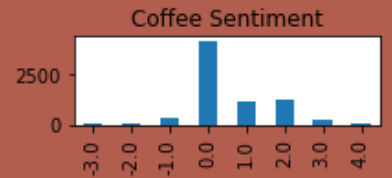
- Most good reviews, have a sentiment of 2, while neutral reviews have a sentiment of 1 and negative reviews have a sentiment of 0.
- This demonstrates how reviews are naturally positively skewed, either by human nature of not wanting to be overly negative in a public forum about a bad experience (since Yelp does tie your user profile to reviews), or because even when we are critiquing an experience, we tend to use less strong words and compliment the redeeming qualities.

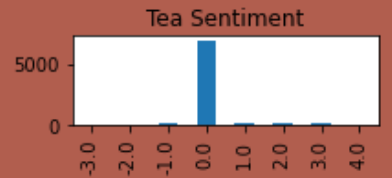
What does the distribution of sentiment look like around each attribute? Are there attributes that are more polarizing or elicit stronger sentiments than others?

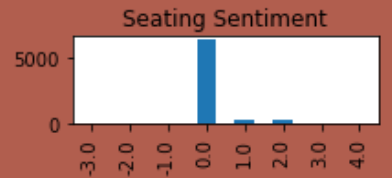


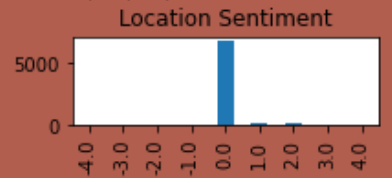


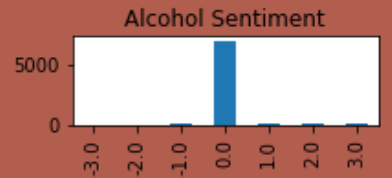


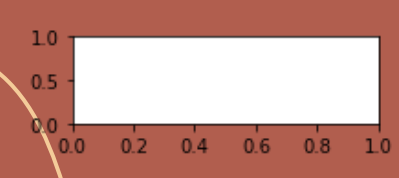


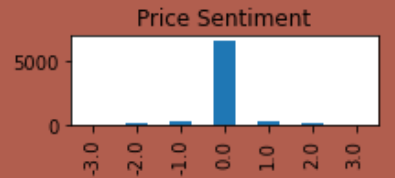


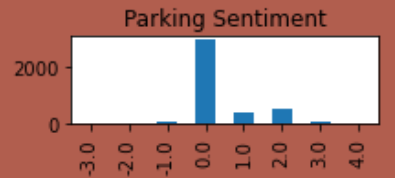


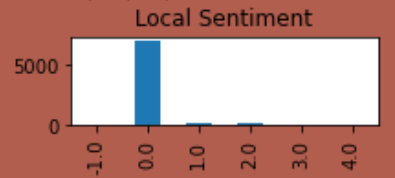


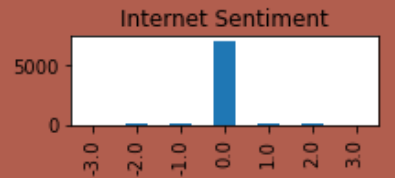


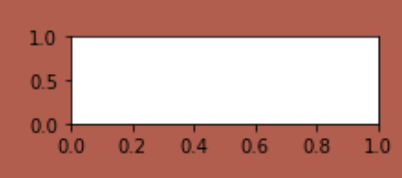










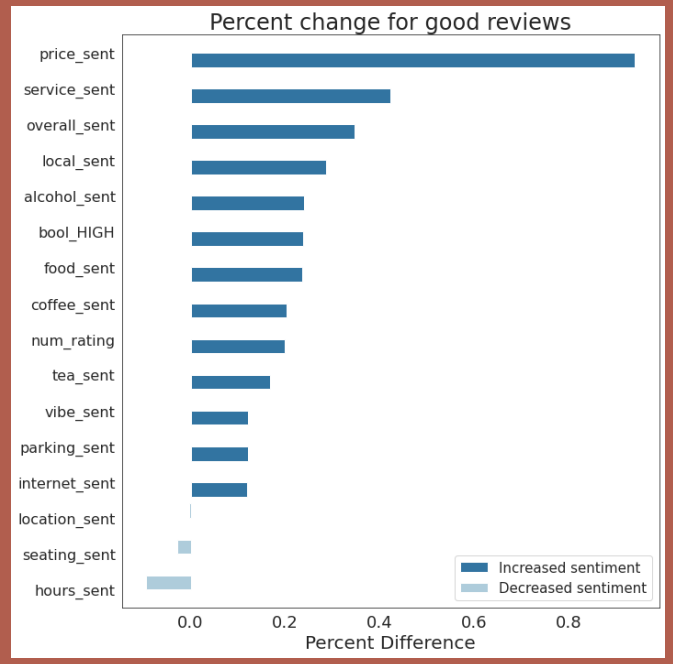


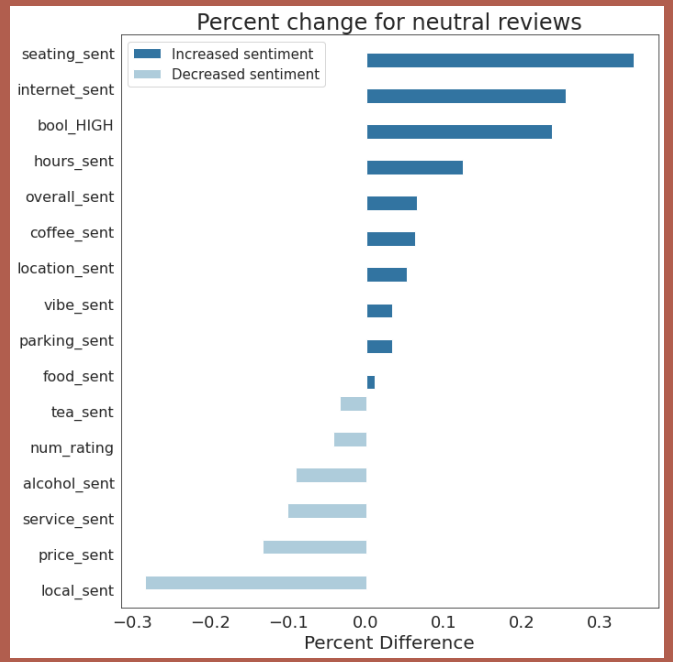
	mean	std_deviation
num_rating	4.173202	1.062846
bool_HIGH	0.807676	0.394153
overall_sent	1.097547	1.179282
vibe_sent	0.370100	0.835968
tea_sent	0.046280	0.330990
service_sent	0.326729	0.828535
seating_sent	0.122489	0.516593
price_sent	0.020091	0.373396
parking_sent	0.370100	0.835968
location_sent	0.075655	0.398635
alcohol_sent	0.041291	0.294465
coffee_sent	0.512749	0.990238
food_sent	0.355183	0.845408
hours_sent	0.031042	0.274347
internet_sent	0.025634	0.273116
local_sent	0.037412	0.277555

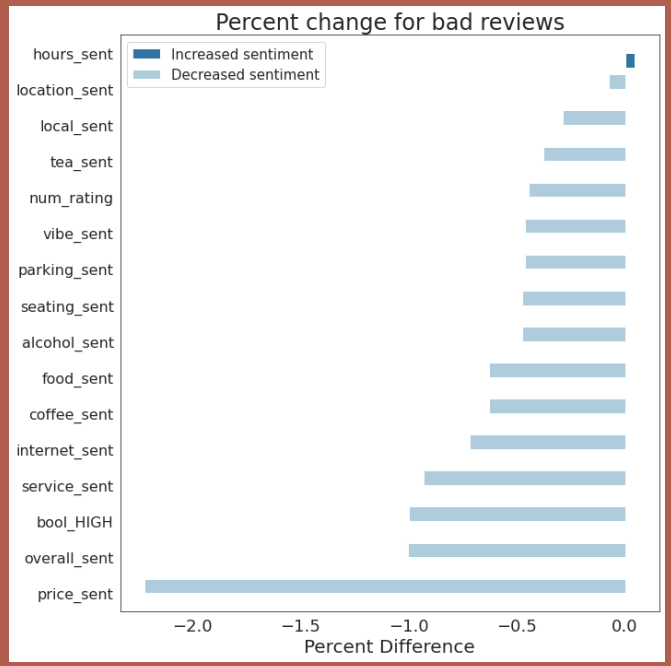
Average sentiment ratings by types of reviews

	all_reviews	good_reviews	neutral_reviews	bad_reviews
num_rating	4.173202	5.000000	4.000000	2.312680
bool_HIGH	0.807676	1.000000	1.000000	0.000000
overall_sent	1.097547	1.478563	1.168529	-0.002882
vibe_sent	0.370100	0.415025	0.382236	0.199095
tea_sent	0.046280	0.054009	0.044703	0.028818
service_sent	0.326729	0.464644	0.293697	0.023055
seating_sent	0.122489	0.118875	0.164506	0.064121
price_sent	0.020091	0.038976	0.017434	-0.024496
parking_sent	0.370100	0.415025	0.382236	0.199095
location_sent	0.075655	0.075445	0.079571	0.069885
alcohol_sent	0.041291	0.051225	0.037550	0.021614
coffee_sent	0.512749	0.617098	0.544926	0.190922
food_sent	0.355183	0.438875	0.358963	0.132565
hours_sent	0.031042	0.028118	0.034884	0.032421
internet_sent	0.025634	0.028675	0.032186	0.007205
local_sent	0.037412	0.048163	0.026822	0.026657

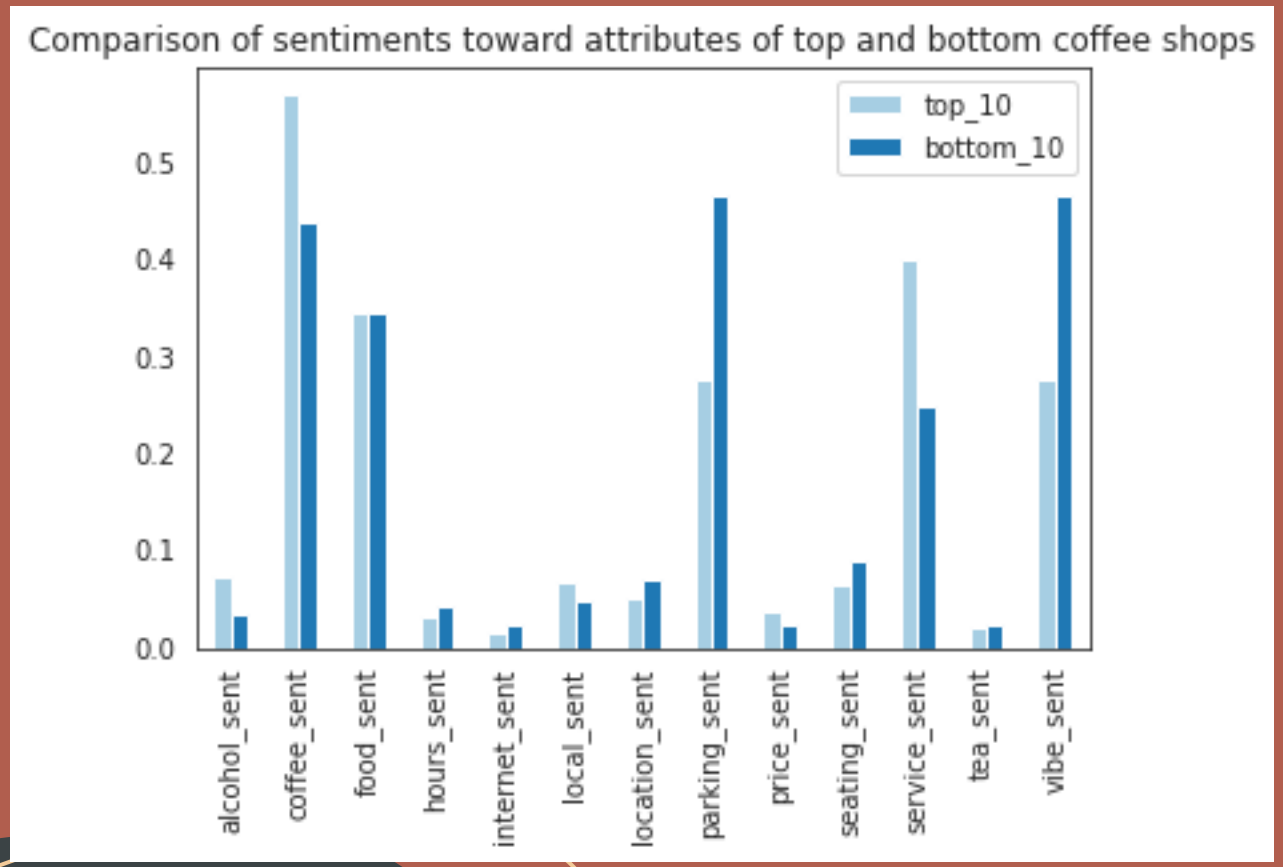
What does the distribution of sentiment look like around each attribute?







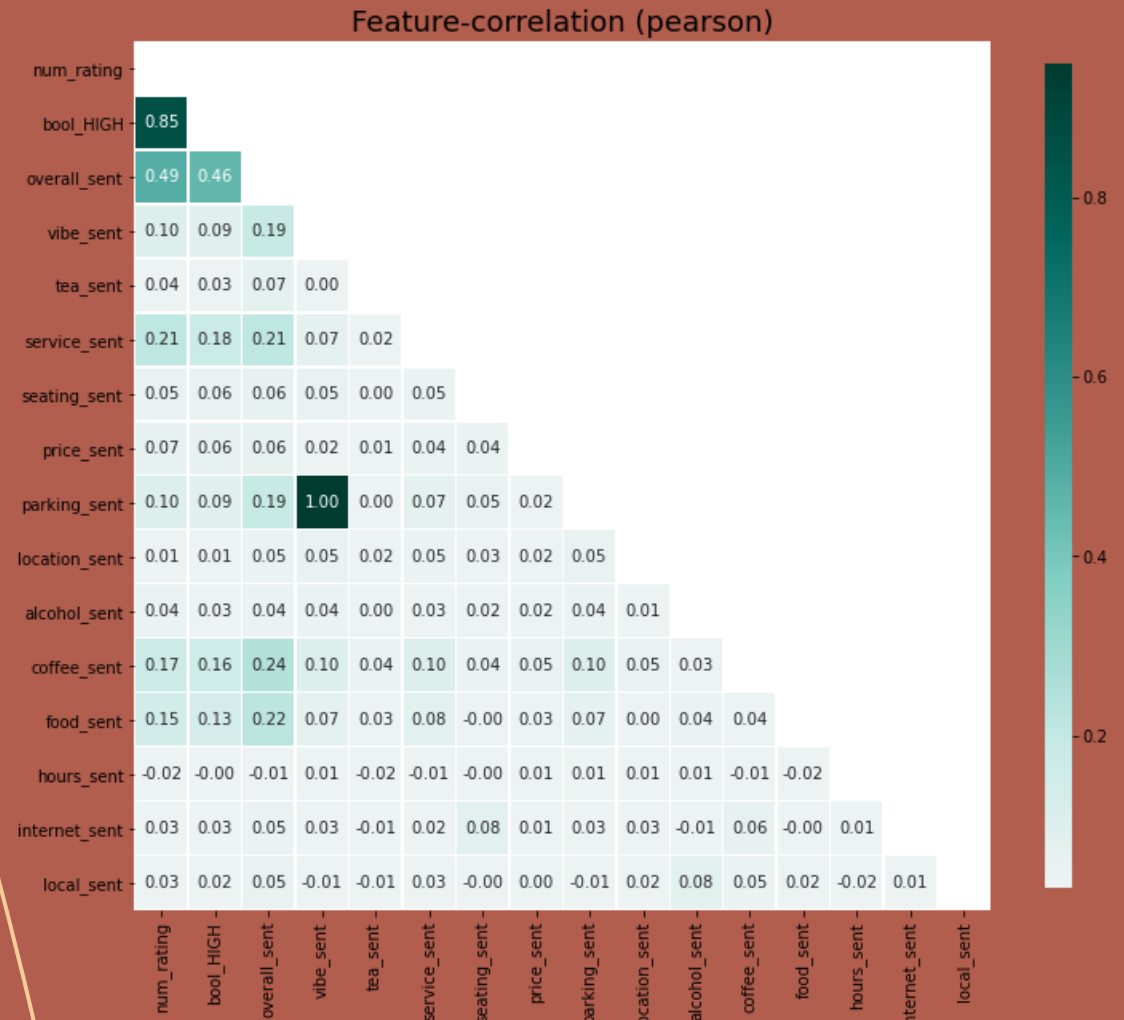
Becoming a top coffee shop - what are the major differences between the top 10 coffee shops and bottom 10 coffee shops?



- The average rating of Top 10 is 4.628545898644917 5
- The average sentiment of Top 10 is 1.2866932160910751
- The average rating of Bottom 10 is 3.742592039800994
- The average sentiment of Bottom 10 is 0.8192636815920398

Feature Coefficient Matrix/heatmap

Identify patterns and relationships between variables. High absolute values of correlation coefficients (close to 1 or -1) indicate strong (positive or negative) relationships, while values close to 0 suggest weak or no relationships.



IV. Statistical Methodology

- Logistic Regresion
- Feature Importance
- Sentiment Analysis
- Time Series
- Word Clouds

Logistic Regression

We built a logistic regression model to predict categorical rating (HIGH or LOW) based on review text and other features

Steps performed:

- Load the dataset containing the reviews and the target variable into a Pandas DataFrame.
- Define the features (review text) and the target variable (categorical rating).
- Split the dataset into training and testing sets using train_test_split() function.
- Use TF-IDF vectorizer to convert the text data into numerical features.
- Build a logistic regression model using LogisticRegression() class from scikit-learn.
- Make predictions on the test set using the trained model.
- Evaluate the performance of the model using accuracy score and classification report, which provides precision, recall, and F1-score for each class.

Accuracy: 0.8877952755905512

	precision	recall	f1-score	support
HIGH	0.89	0.99	0.93	1234
LOW	0.89	0.47	0.61	290
accuracy			0.89	1524
macro avg	0.89	0.73	0.77	1524
weighted avg	0.89	0.89	0.87	1524

Feature Selection through Random Forest

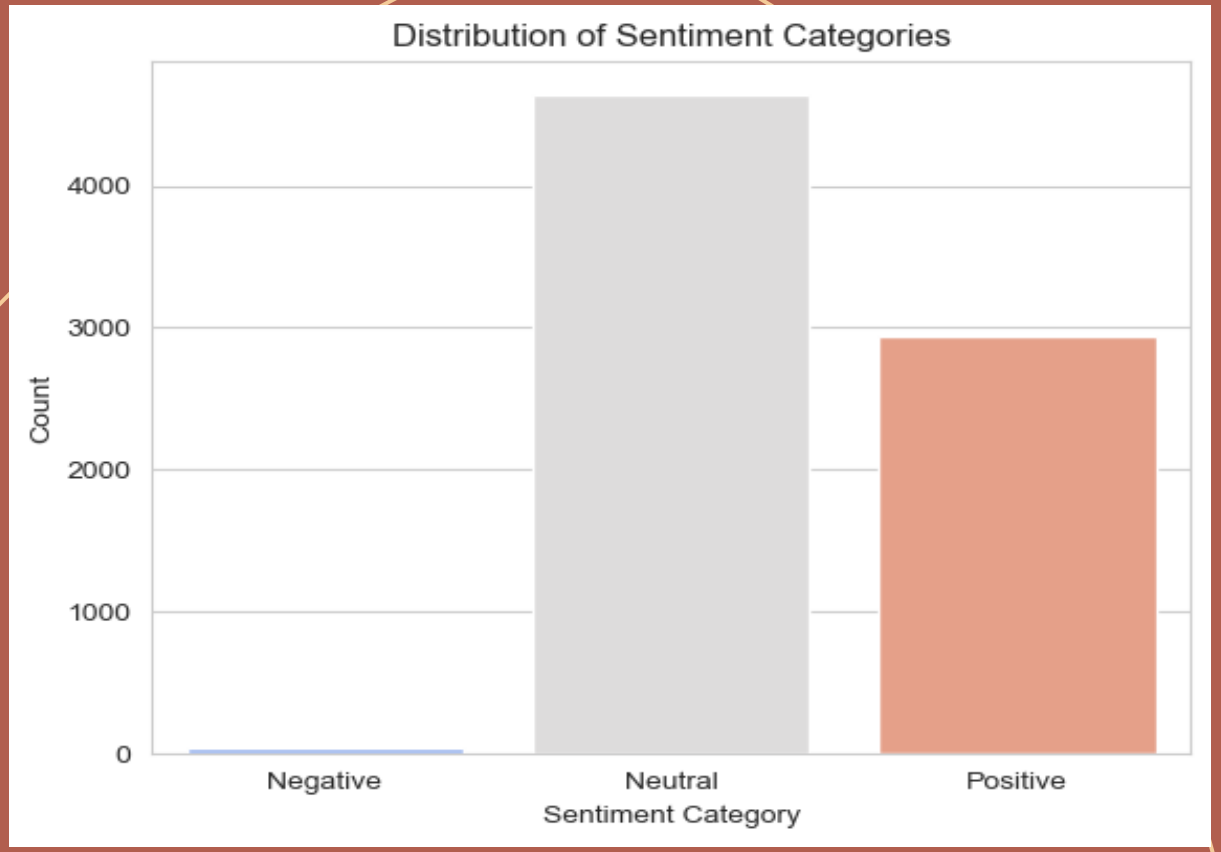
We built a model to identify the most important features (e.g., specific words in the review text, numerical ratings) that contribute to the overall rating or categorical rating of the coffee shop

```
Top 10 Most Important Features:
     Feature  Importance
720     rude    0.016019
598       ok    0.015608
365    great    0.013933
68       bad    0.011401
758  service    0.011225
599     okay    0.009153
55     asked    0.008926
453     just    0.008879
986    worst    0.008695
529    maybe    0.008587
```

Steps performed:

- Load the dataset containing the reviews and the target variable into a Pandas DataFrame.
- We define the features (review text) and the target variable (categorical rating).
- We split the dataset into training and testing sets using train_test_split() function.
- We use TF-IDF vectorizer to convert the text data into numerical features.
- For decision tree, we build a decision tree classifier using DecisionTreeClassifier() class from scikit-learn.
- For random forest, we build a random forest classifier using RandomForestClassifier() class from scikit-learn.
- We make predictions on the test set using the trained models.
- We evaluate the performance of the models using accuracy score and classification report, which provides precision, recall, and F1-score for each class.

Sentiment Analysis

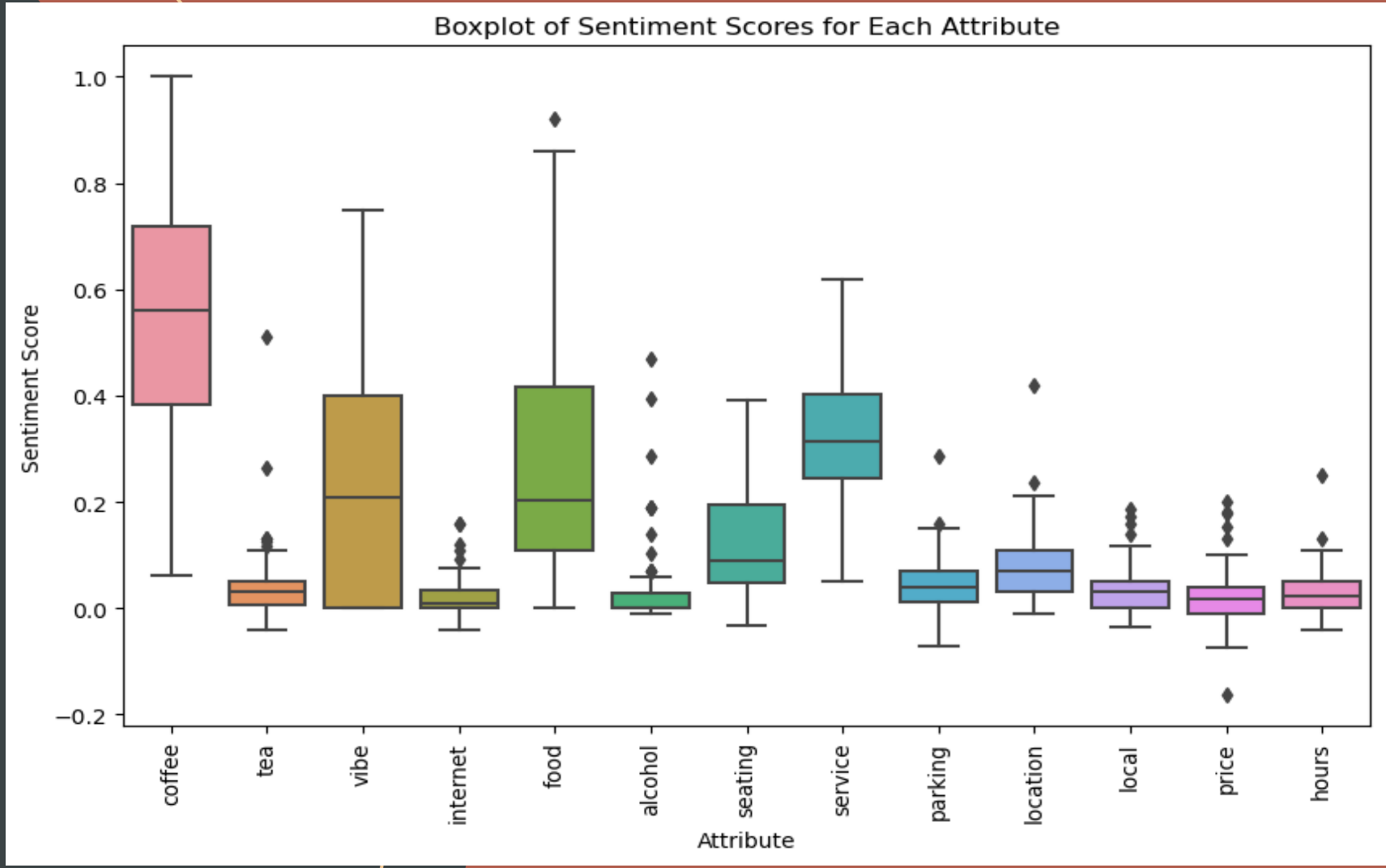


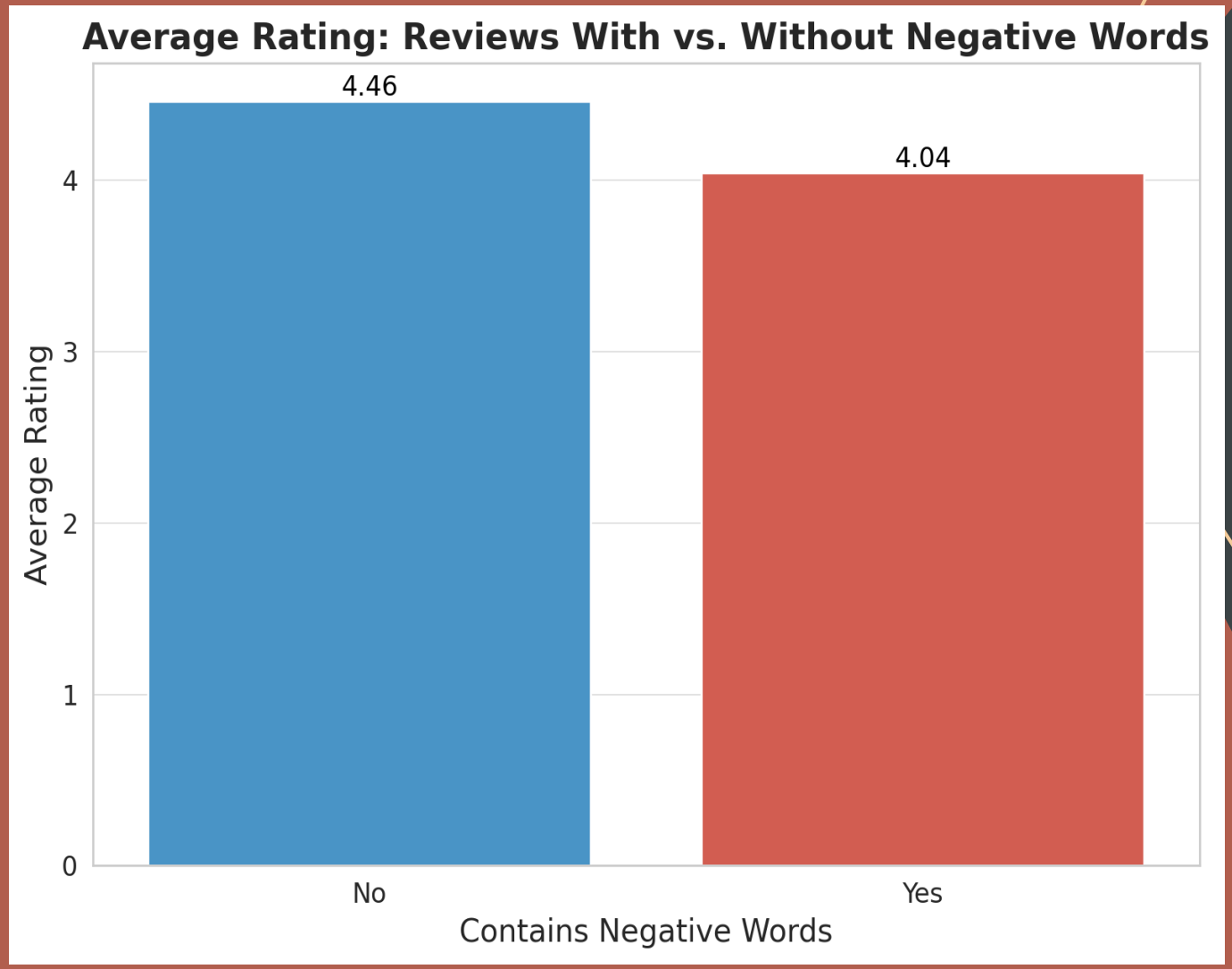
Correlation table

	review_length	Polarity
review_length	1.000000	-0.288783
Polarity	-0.288783	1.000000

Polarity score within the range [-1, 1], where -1 means a negative sentiment, 1 means a positive sentiment, and 0 represents a neutral sentiment.

Sentiment Scores For Each Attribute

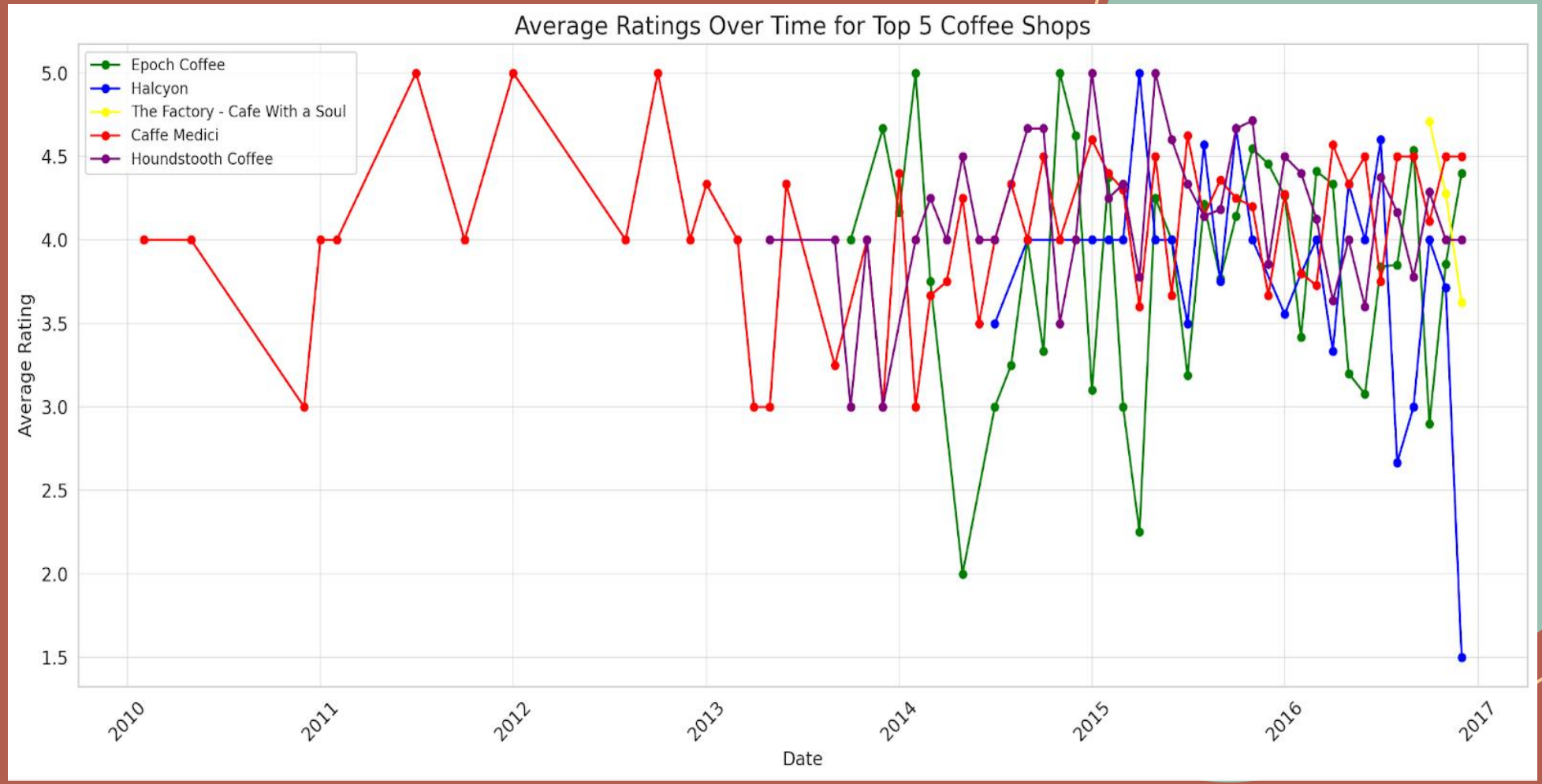




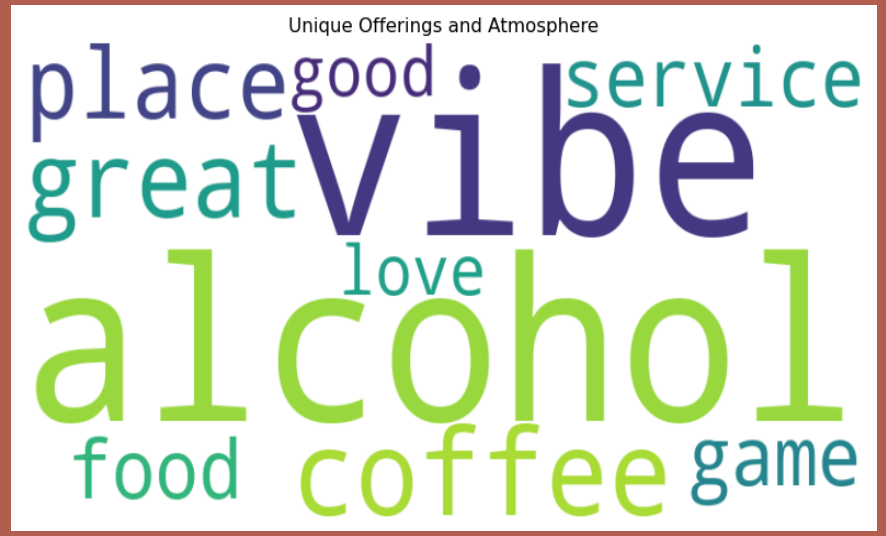
Sentiment Analysis

Negative words like 'not', "don't", "doesn't", 'never', 'no', 'cannot', "can't", 'lack', 'without'

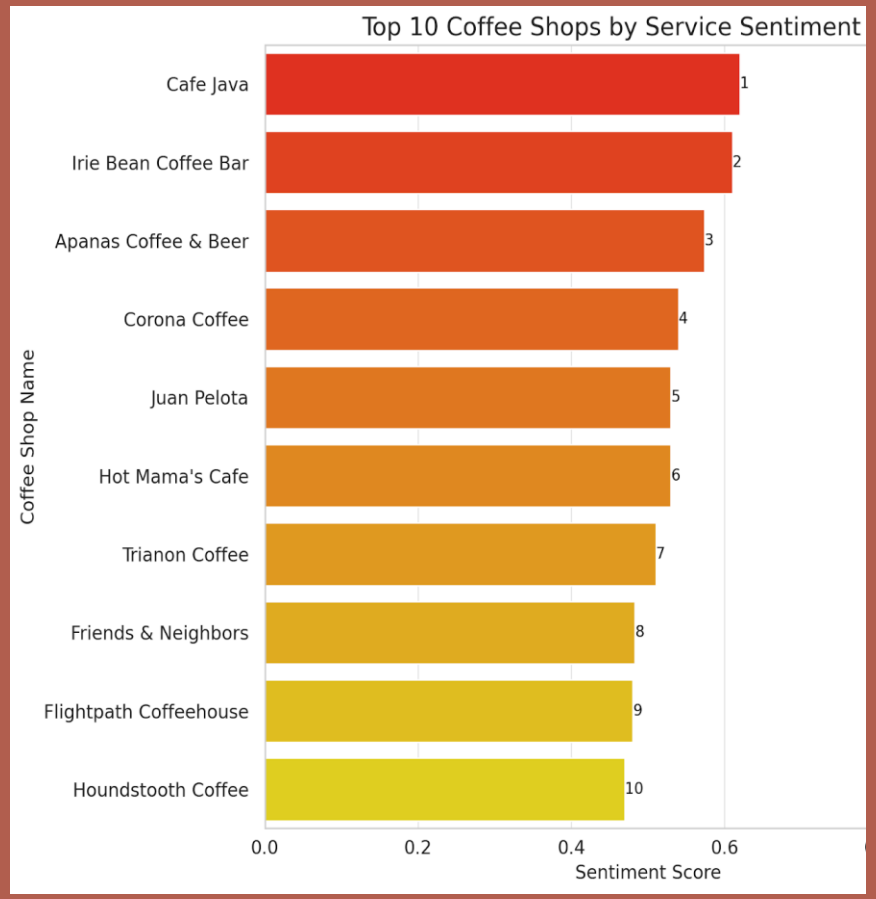
Sentiment Analysis – A Time Perspective

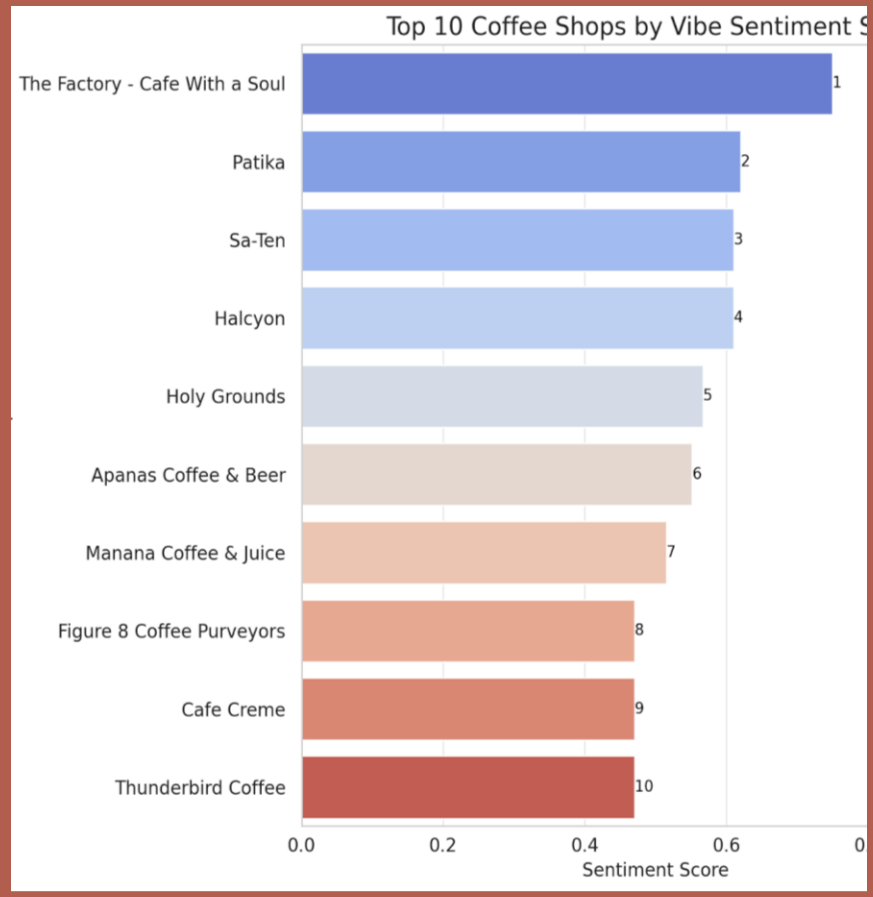


Word cloud under topics



Ranking of top 10 coffee shops by attributes





VI. Implication for Stakeholders

Implications for stakeholders

- Enhanced Customer Satisfaction: Understand factors for positive experiences, leading to happier customers.
- Improved Business Strategies: Informed decisions on offerings, pricing, and marketing for better outcomes.
- Targeted Marketing: Tailor messaging and promotions to resonate with the audience.
- Operational Efficiency: Identify areas for improvement, streamlining processes for cost savings.
- Competitive Advantage: Stand out in the market by adapting to customer preferences effectively.

VII. Ethical, legal and societal implications

Ethical, legal and societal implications

- Privacy: Uphold customer privacy rights by safeguarding their personal information and ensuring anonymity in data analysis.
- Representation: Strive for a balanced representation of diverse perspectives and experiences within the dataset to avoid skewing analysis outcomes.
- Transparency: Maintain transparency in data collection, preprocessing, and analysis methods to foster trust and accountability in research practices.
- Security: Implement robust data security measures to protect the dataset from unauthorized access, ensuring the confidentiality and integrity of sensitive information.
- Bias: Recognize and mitigate biases that may arise from factors like demographics or cultural backgrounds to provide impartial and accurate insights.
- Business Impact: Consider the potential effects of analysis findings on coffee shop owners, including competitive dynamics and market trends, to inform responsible decision-making.
- Responsibility: Embrace the ethical responsibility of researchers to utilize data ethically and responsibly, considering the broader societal implications of their analysis.

Thank you for listening!
Q&A session

All images source: unsplash.com