Dickinson

Dept. of Data Analytics

COFFEE SHOPS Review Analysis

Ziwei Guo ’24 &
Tai Nguyen ‘25

INTRODUCTION

This project focuses on analyzing Yelp reviews of coffee shops in Austin, TX aiming to uncover insights into customer perceptions and preferences by employing statistical techniques such as sentiment analysis, classification modeling, feature importance analysis, topic modeling and regression analysis.

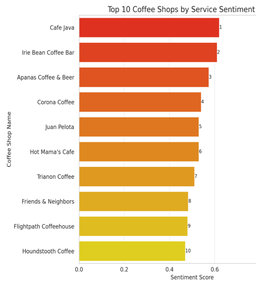
Objective:

We seek to understand factors that influence overall ratings and categorical ratings of coffee shops and provide actionable insights to coffee shop owners in Austin, Texas or coffee business owners in general, helping them enhance customer satisfaction and optimize business strategies.

DATA RETRIEVAL

- 3 datasets: Ratings and Sentiments, Raw Reviews, Sentiments by Shop with nearly 7000 *scraped* reviews from yelp.com of 78 coffee shops in Austin, TX with various sentiment parameters
- Includes 20 variables consisted of 4 types: numerical, categorical, string/text, date/time, including but not limited to coffee shop names, review text, rating, etc.

EXPLORATORY DATA ANALYSIS

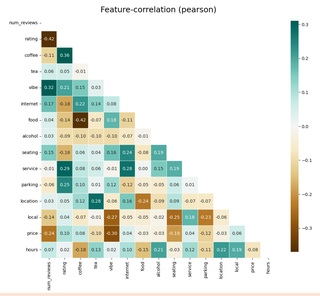


	review_length	Polarity
review_length	1.000000	-0.288783
Polarity	-0.288783	1.000000

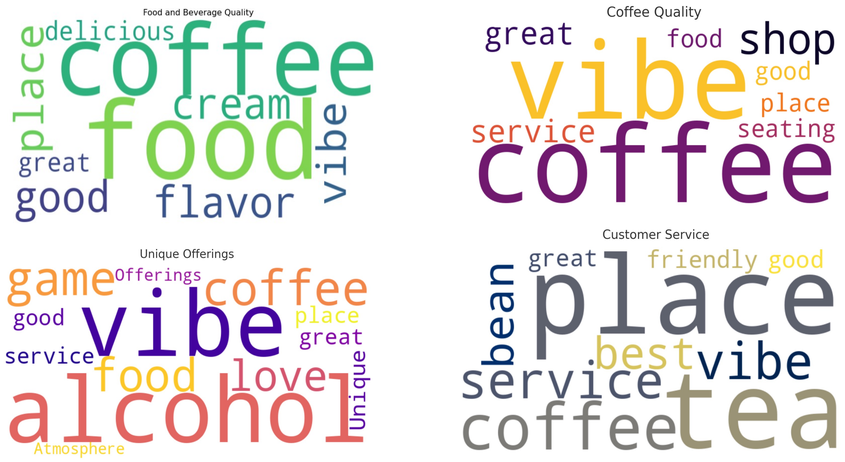
Polarity is the range [-1, 1], where -1 means a negative sentiment, 1 means a positive sentiment, and 0 represents a neutral sentiment.

Correlation Table/Heatmap

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



WORD CLOUD BY TOPICS



The size of each word indicates its frequency or importance in the text reviews.

STATISTICAL MODELING

- **Feature selection through Random Forest:** 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
```

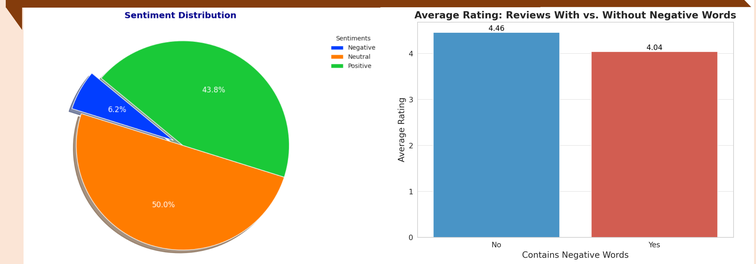
- **Logistic Regression**: Prediction model of categorical rating (HIGH or LOW) based on review text and other features

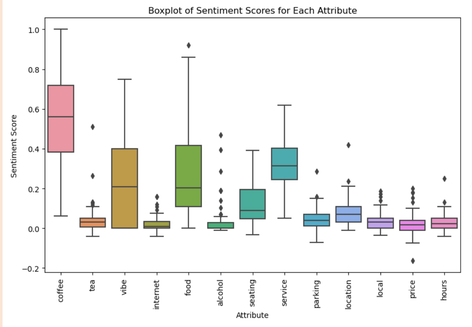
```
Accuracy: 0.8877952755905512
              precision    recall  f1-score   support

        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
```

- **Topic Modelling**: Utilizing algorithms like Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF), to uncover latent topics or themes within the review text and provide a deeper understanding of the key aspects discussed by customers

```
Topic 1:
coffee vibe seating place great good check service 2016 parking

Topic 2:
coffee great service place love 2016 good check vibe austin

Topic 3:
coffee service vibe like good just 2016 place time really

Topic 4:
food foods place vibe coffee good 2016 like just service

Topic 5:
food tea cream chocolate 2016 place flavors check like good
```

SENTIMENT ANALYSIS

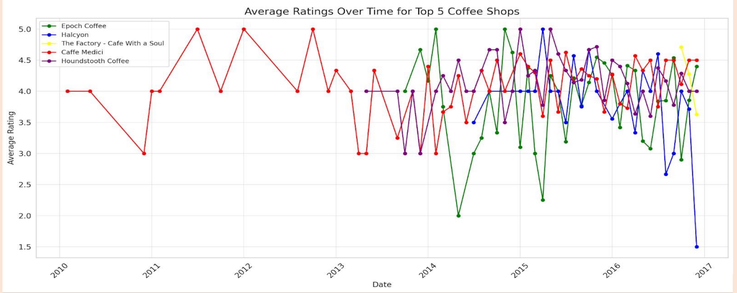




Negative words like 'not', "don't", "doesn't", 'never', 'no', 'cannot', "can't", 'lack', 'without'

Attribute with high ratings : Coffee, vibe service
Attribute with low ratings: Parking, internet, tea

Stable? Fluctuate? Drop?



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.