Where should we take the kids?

Capstone Presentation
August 5, 2020
Michelle Levine

Main Objective

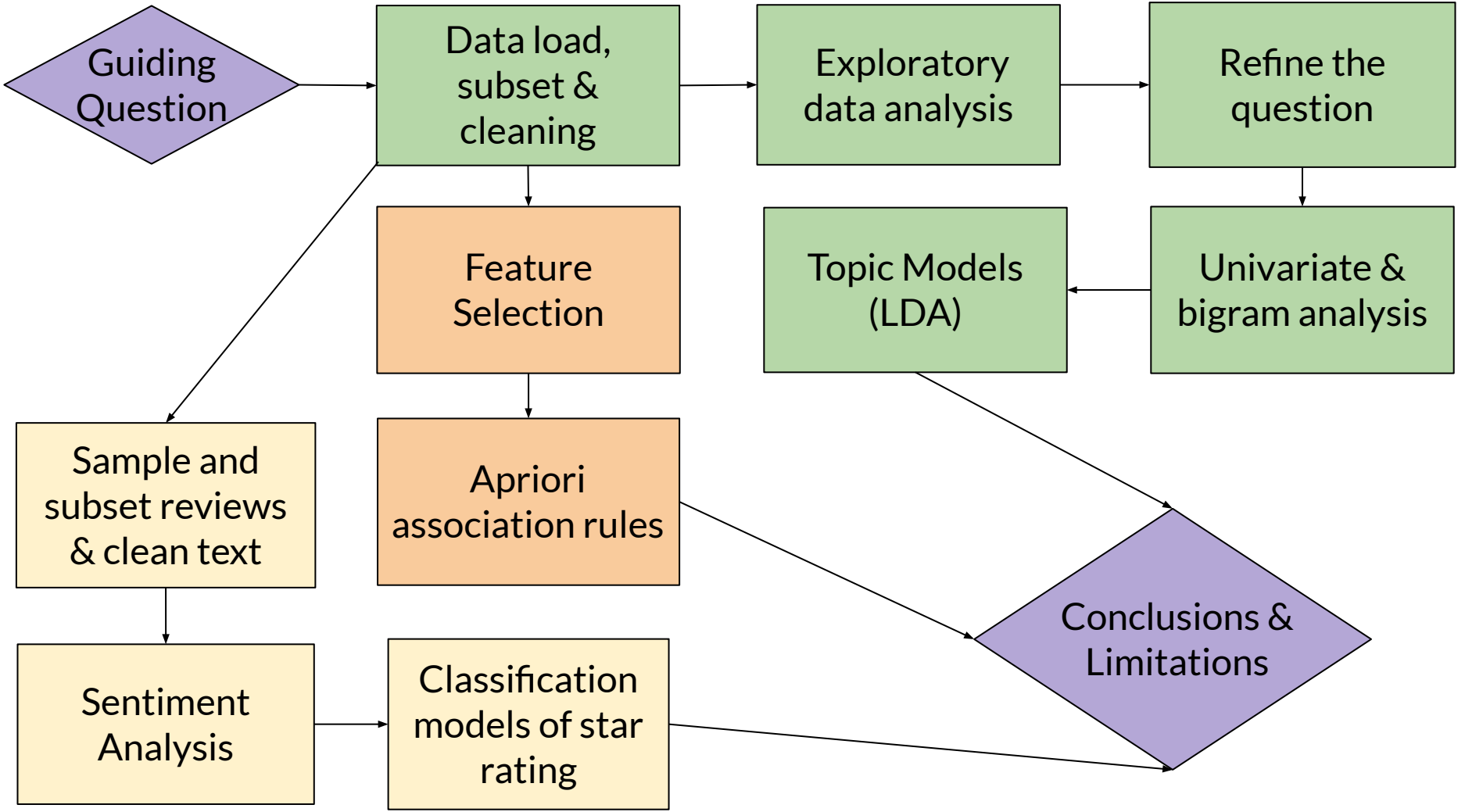
Exploratory analysis of family-friendly businesses using Yelp dataset.

Guiding Questions

What are the characteristics of family-friendly businesses?

Are there combinations of business features related to good ratings?

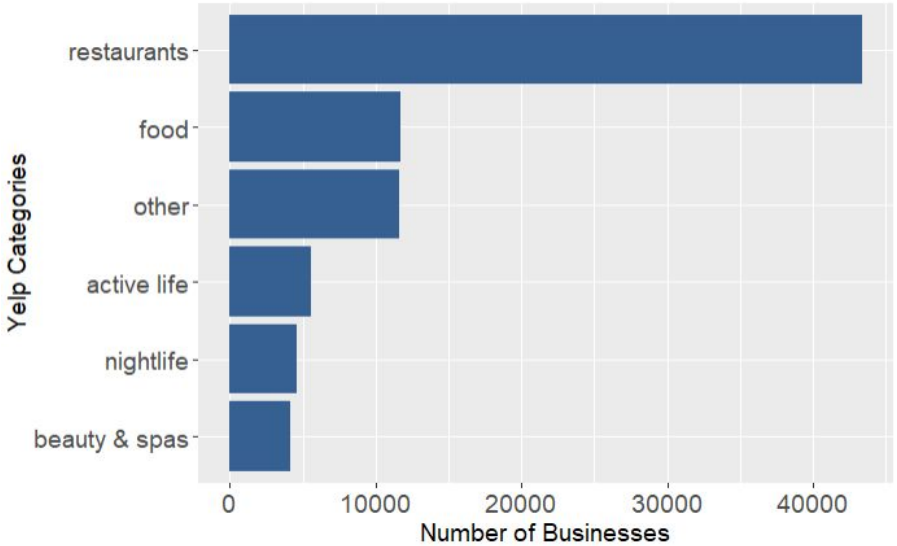
What matters most to customers that are satisfied or dissatisfied with their experience?


Guiding Question
Data load, subset & cleaning
Exploratory data analysis
Refine the question
Univariate & bigram analysis
Topic Models (LDA)
Feature Selection
Sample and subset reviews & clean text
Apriori association rules
Conclusions & Limitations
Sentiment Analysis
Classification models of star rating

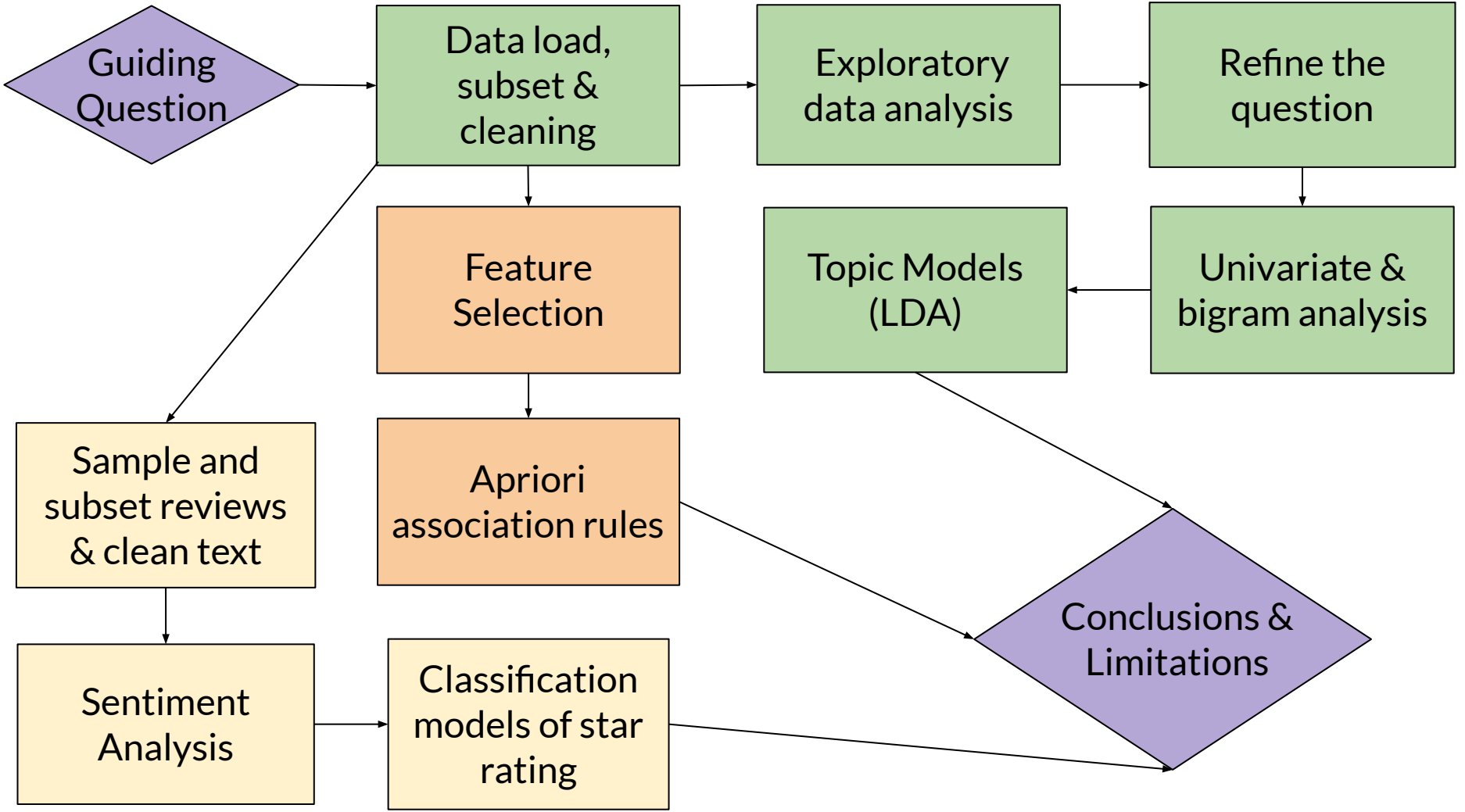
Load & Exploratory Data Analysis

- 26% of all businesses were “Good for Kids” (N=55,527)
- 78% were restaurants (N=43,368)
- Las Vegas, Toronto, Phoenix and surrounding areas
- Star reviews were positively skewed
- Many outliers in review counts

Number of businesses per category



Note: there is double counting of businesses as they can select more than 1 category.

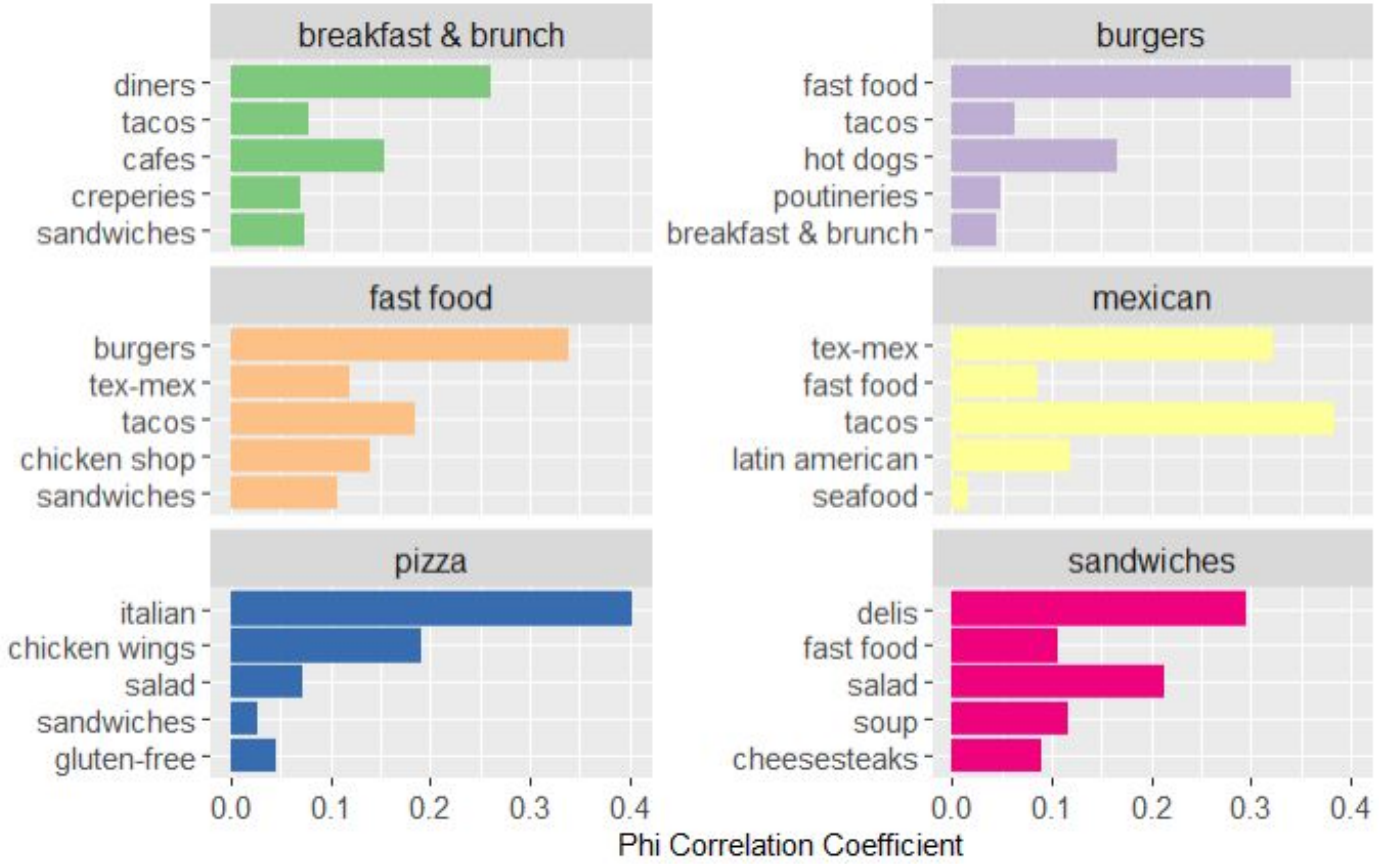

Guiding Question
Data load, subset & cleaning
Exploratory data analysis
Refine the question
Univariate & bigram analysis
Topic Models (LDA)
Feature Selection
Sample and subset reviews & clean text
Apriori association rules
Conclusions & Limitations
Sentiment Analysis
Classification models of star rating

Revised question 1 and methods

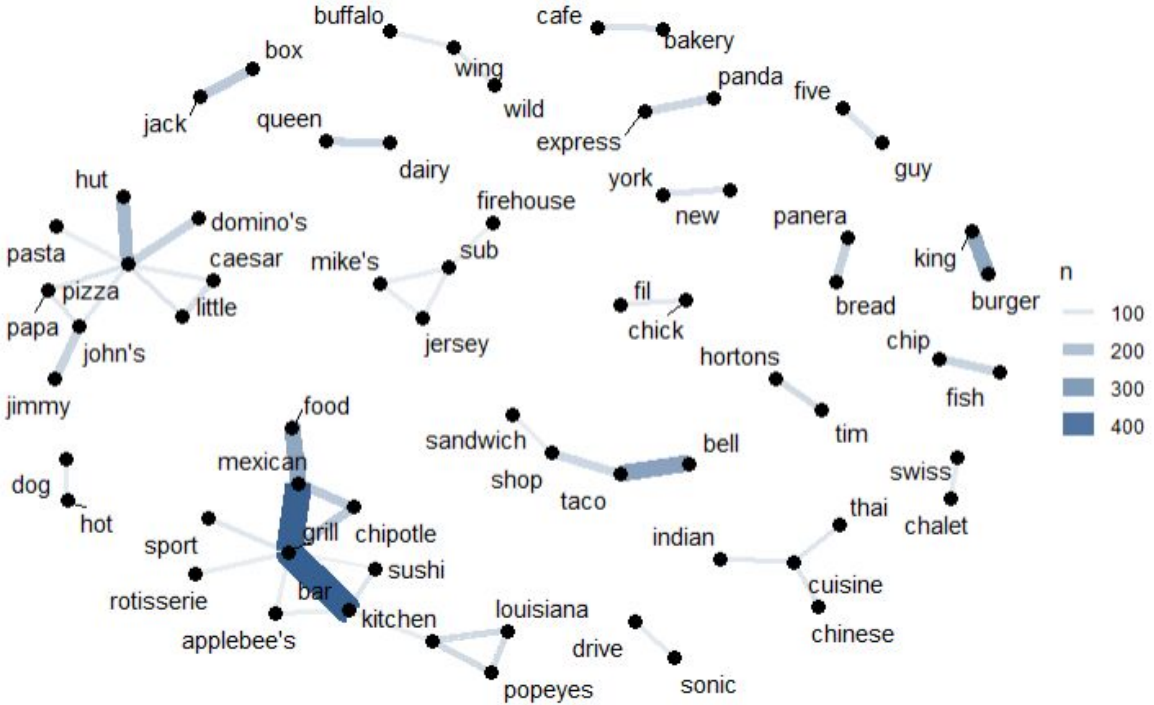
What are the characteristics of family-friendly **restaurants**?

- Univariate analysis of Yelp categories and subcategories
- Bigram analysis of Yelp subcategories
- Bigram analysis of restaurant titles
- Topic modelling of restaurant titles

Top subcategory combinations



Network diagram of restaurant title bigram frequency



Latent Dirichlet Allocation

- Unsupervised machine learning algorithm used to infer latent or underlying themes in restaurant titles
- Gibbs Sampling
  - 6000 total iterations
  - 4000 burnin
  - Sample every 500th of the remaining 2000 iterations
- Tested k values from 5 to 20
- Analysis of β and γ did not add clear or useful information to the exploratory analysis.

Revised question 2 and methods

Are there combinations of **restaurant** features related to good ratings?

- Dimensionality reduction (incl. feature selection)
- Apriori Association analysis for feature sets related to high star ratings

Dimensionality reduction

- 70 total features
- 42 removed due to more than 30% NAs
- “Good for kids” feature removed due to low variance
- 27 remaining features
    - Ordinal features converted to rank
    - Nominal features converted to dummy variables
    - Star rating converted to 2 categories (‘3.5 or less’ & ‘4 or more’)

Top features by feature selection method

Information Gain	Chi-square	Stepwise Regression
Business Parking Street †	Business Parking Street †	Noise Level 2 †
Noise Level †	Noise Level †	Noise Level 3 †
Ambience Trendy †	Ambience Trendy †	Noise Level 4 †
Ambience Classy †	Has TV †	Has TV †
Ambience Hipster ✦	Ambience Classy †	Ambience Casual †
Ambience Casual †	Ambience Hipster ✦	Ambience Classy †
Has TV †	Ambience Casual †	Ambience Trendy †
Bike Parking ✦	Bike Parking ✦	Business Parking Street †
WiFi Free	Alcohol Full Bar ✦	Alcohol Full Bar ✦
WiFi No	Ambience Intimate	--

† 3 feature selection techniques;

✦ 2 feature selection techniques

Apriori Association Rules Analysis

295 total rules

Support = 0.05; Confidence = 0.40; and RHS = ‘4 or more’ stars

196 rules

Limited LHS to have at least one feature present

Subset rules

{Street Parking = 1} → {4+ stars}
{Ambience casual = 1} → {4+ stars}
{Bike Parking = 1} → {4+ stars}
{Noise Level = 1, Has TV = 1} → {4+stars}

Revised question 3 and methods

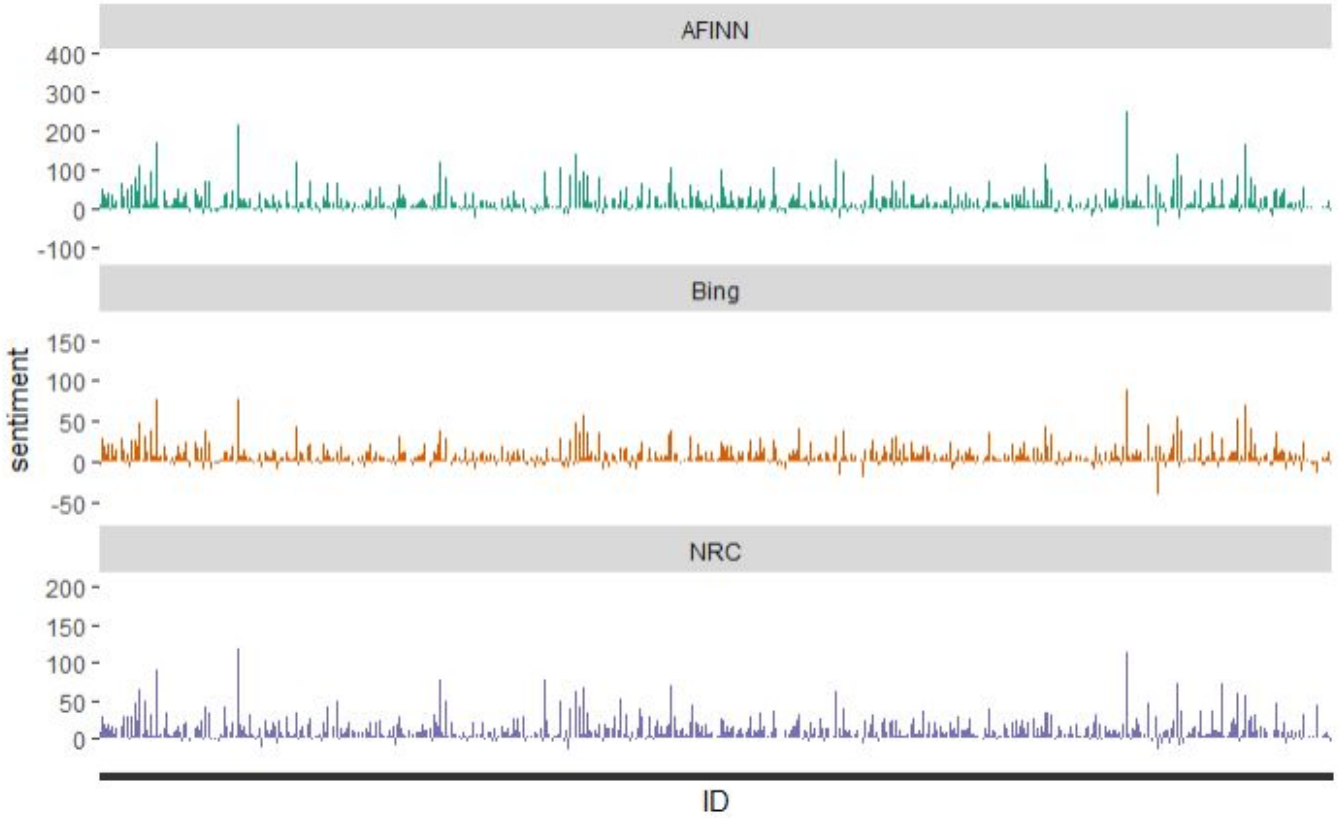
What matters most to customers that are satisfied or dissatisfied with their experience?

- Sentiment analysis of customer reviews
- Develop classification models

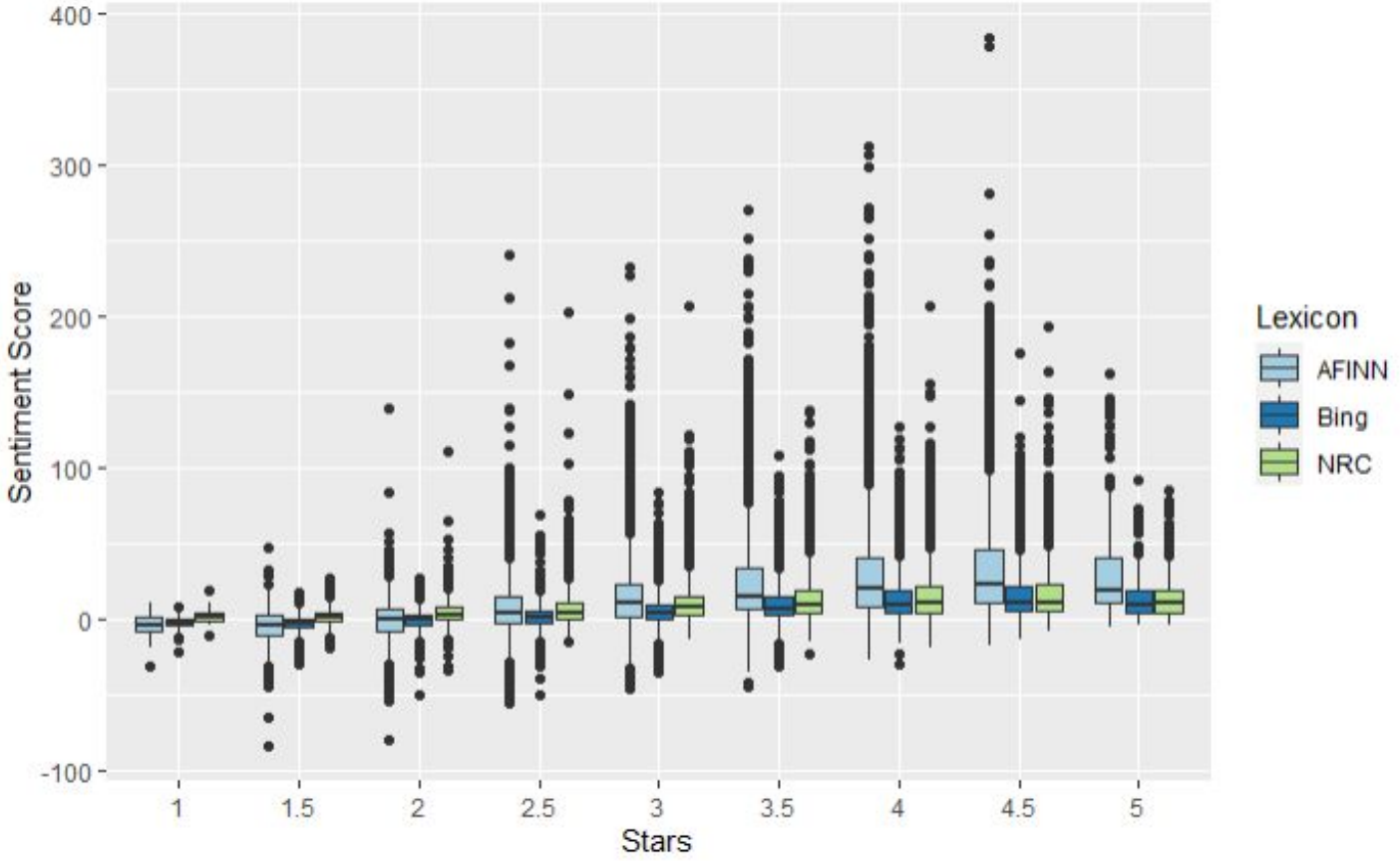
Data Preparation for Sentiment Analysis

1. Select a random sample (25%) of the 4M reviews for family-friendly restaurants
2. 95% of restaurants are represented in the 993,038 sampled reviews
3. Remove outliers
    a. 10% of businesses with high review counts
    b. 554,103 of reviews were for outliers
4. Identify child- or family-specific reviews
    a. Custom lexicon (e.g., kids, daughter, niece, grandson, baby, family, mom)
    b. 64,806 reviews (14.76%) mentioned children or families
5. Text cleaning and normalization
    a. Remove numbers, special characters, words less than 2 characters long, stopwords
    b. Lemmatize
    c. Remove most (10) and least frequent words (27,869 that appear once)

Comparison of Sentiment Analysis Lexicons



Sentiment Score by Star Rating and Lexicon



Classification Models to Predict Star Ratings

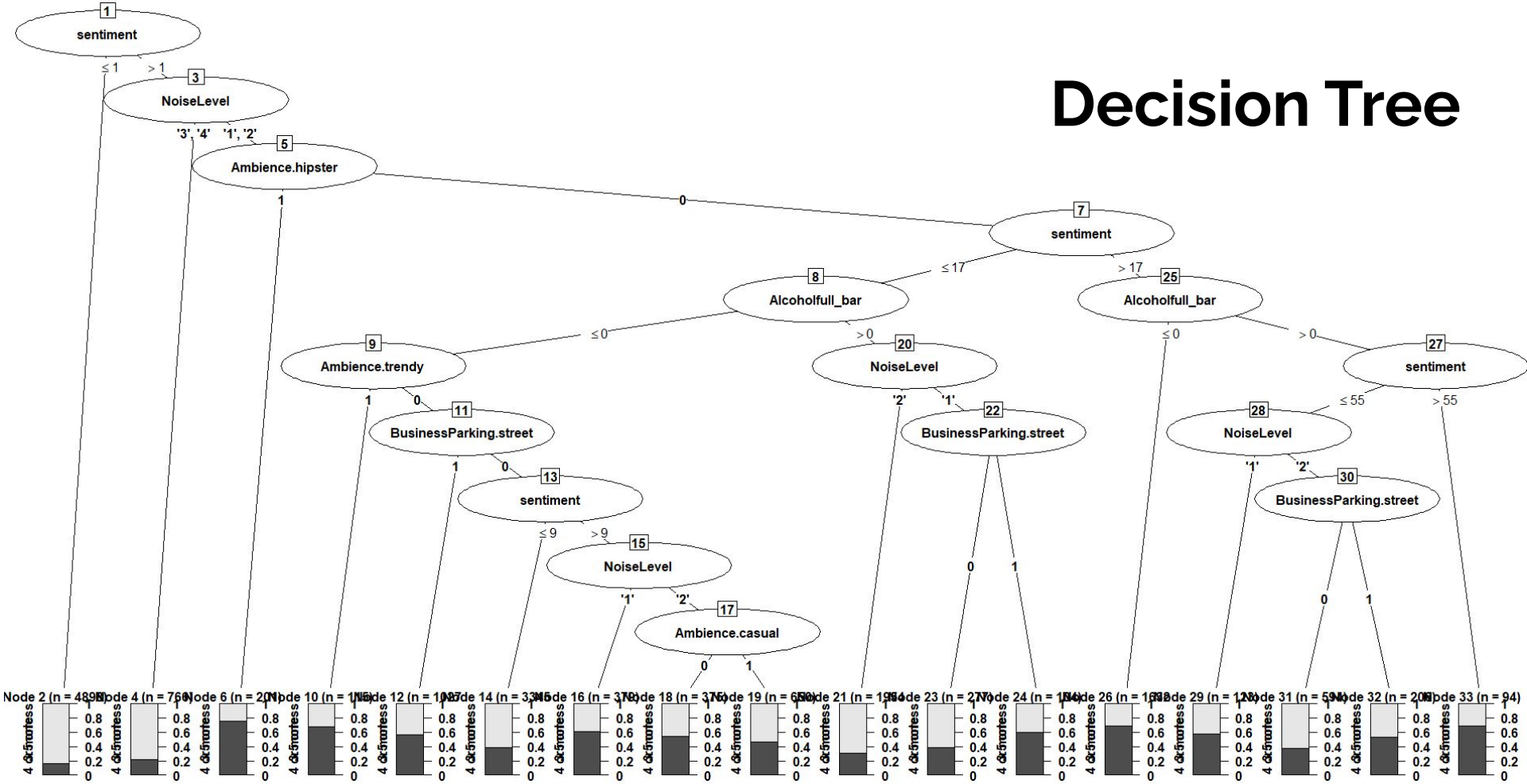
- Total records 22,319
- Train/Test split of 75/25 used for both models
- Binary dependent class split 63% (3.5 or less stars) and 37% (4 or more stars)
- Independent variables included those from feature selection and sentiment analysis using Bing lexicon
- k-Nearest Neighbors used to impute one or more missing values for 6,651 records
- Compared binary logistic regression and C5.0 decision tree
- Models were trained using 10-fold cross validation

Logistic Regression Results

	Estimate	Odds Ratio	z Value	p Value
Sentiment	0.05	1.05	33.551	< 2e-16***
Bike Parking	0.14	1.16	3.374	0.000741
Noise Level 2	-0.41	0.66	-10.049	< 2e-16***
Noise Level 3	-1.23	0.29	-12.619	< 2e-16***
Noise Level 4	-1.55	0.21	-8.70	< 2e-16***
TV	-0.20	0.82	-4.807	1.53e-6***
Ambience Casual	-0.7	0.93	-1.837	0.066161 .
Ambience Classy	0.22	1.25	3.255	0.001136**
Ambience Hipster	0.98	2.67	6.462	1.04e-10***
Ambience Trendy	0.34	1.41	3.41	0.000647***
Street Parking	0.61	1.84	13.60	< 2e-16***
Full Bar	-0.64	0.53	-15.09	< 2e-16***

Signif. codes:
*** 0.001
** 0.01
* 0.05
. 0.1
1

Decision Tree

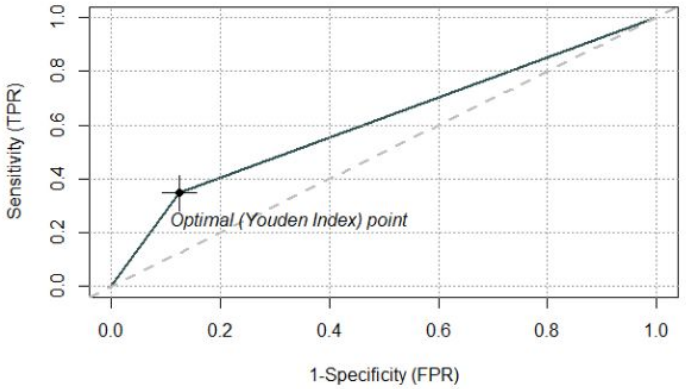


Logistic Regression

	Reference	
Prediction	3.5 or less	4 or more
3.5 or less	3090	1339
4 or more	441	709

Accuracy: 68%

ROC Curve

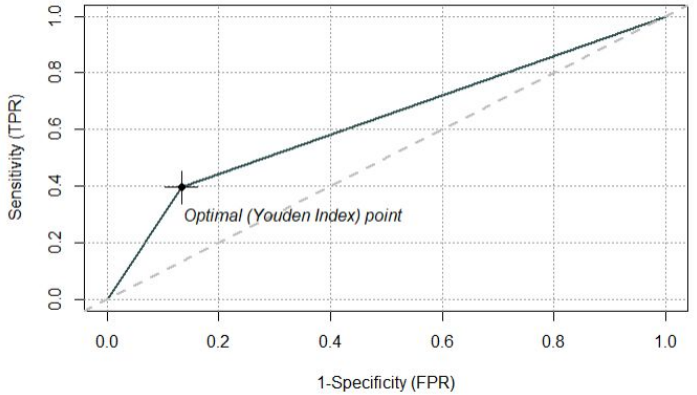


Decision Tree

	Reference	
Prediction	3.5 or less	4 or more
3.5 or less	3061	1239
4 or more	470	809

Accuracy: 69%

ROC Curve



Conclusions

- Identified restaurants types and cuisines popular at family friendly restaurants
  - Breakfast & brunch, Burgers, Fast food, Mexican, Pizza, and Sandwiches.
  - Sushi, Indian, Thai and Chinese
- Found restaurant features related to star ratings
  - Bike parking, Noise level, Has a TV, Ambience (casual, classy, hipster, trendy), Parking (street, bike), and Has a full bar.
- Found 4 feature sets related to high star ratings
- Reviews were overwhelmingly positive and related to star ratings
- Classification models that correctly classified two-thirds of restaurants as having high or low star ratings

Limitations

- Analysis focused exclusively on family-friendly restaurants
- Additional text cleaning for topic modelling and sentiment analysis
- LDA is better to suited to sparse vectors, try Word2Vec instead
- Spurious association rules
- Feature selection separate from association rules
- User reviews likely violate assumption of document-level sentiment analysis
- Imbalanced dependent class for modelling