Yelp Reviews of Ice Cream Shops

STAT 628
Group 3: Ke Chen, Nan Yan, Chen Hu, Richard Yang

Outline

- Introduction
  - Goals, study interests
  - Dataset information
- Study of Reviews
  - Overview of words
  - General findings
- Study of Users and Tips
  - Data processing of users' infor
  - Plans of tips
- Further Study
  - Proposed model
  - Expected result

Goals

- Analyze datasets to extract useful features
- Build model to predict yelp ratings by merged dataset
- Provide data-driven suggestions to ice-cream shop owners in Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison and Cleveland, to improve their Yelp ratings

Business selecting

- Provide category information, still open, review count > 10 72677
- Category contains “Ice cream” 1294

Business selecting

- Selecting target city 455

Las Vegas	173
Phoenix	118
Charlotte	57
Pittsburgh	51
Madison	28
Cleveland	24
Urbana-Champaign	4

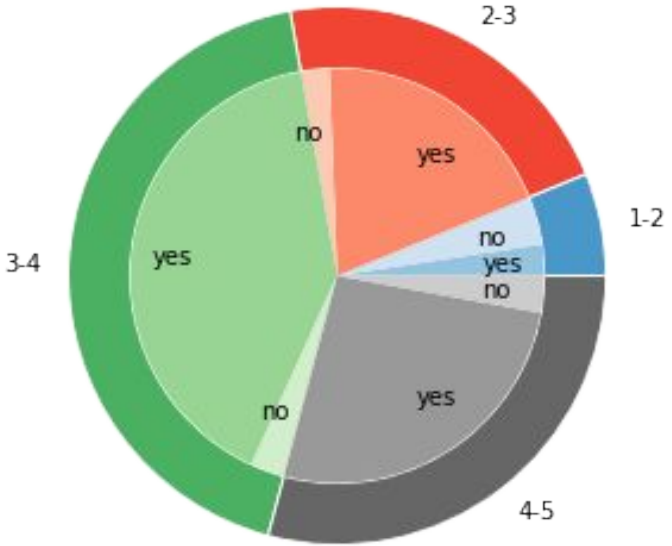
Business attributes

Most common attribute information : (total 32 attributes)

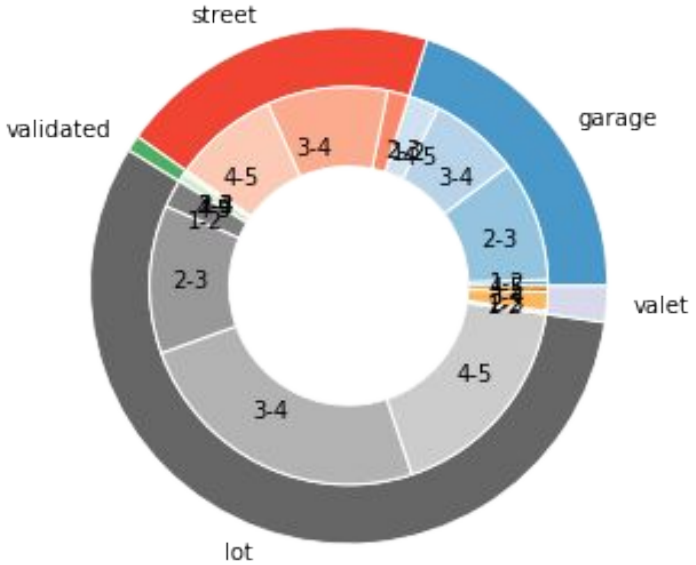
Attributes	Frequency
Credit card acceptance	451
Price range	448
Business parking	445
Take out	442
Bike parking	435
Wifi	426
Caters	365
...	...

Business attribute: parking

Group by star ratings and with/without parking



Group by parking place and star ratings

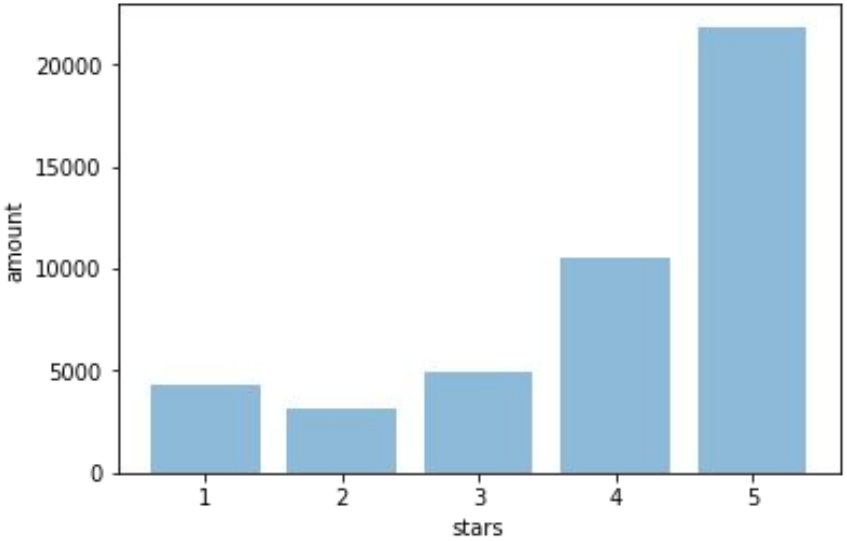


Restaurants provide parking may have higher ratings.

The parking place doesn't matter

Study of Reviews

- Stars distribution



	Ice cream	Total
Mean	3.95	3.72
Median	4	4
Mode	5	0, 5
Std	1.32	1.46

The majority of people would give high score for ice cream shops.

Study of Reviews

- Overview of Words

Noun

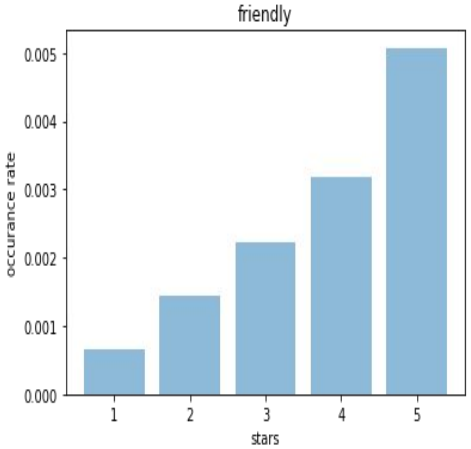
Verb

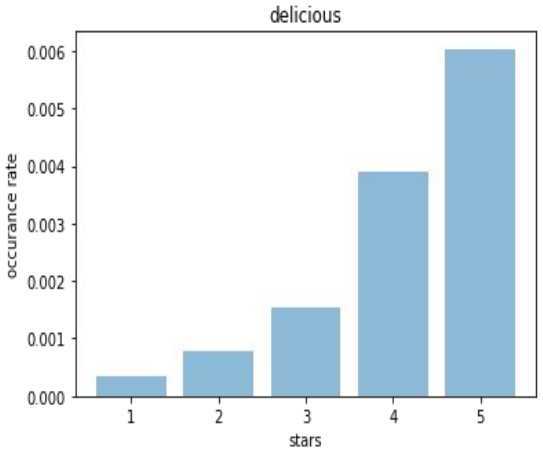
Adj/Adv

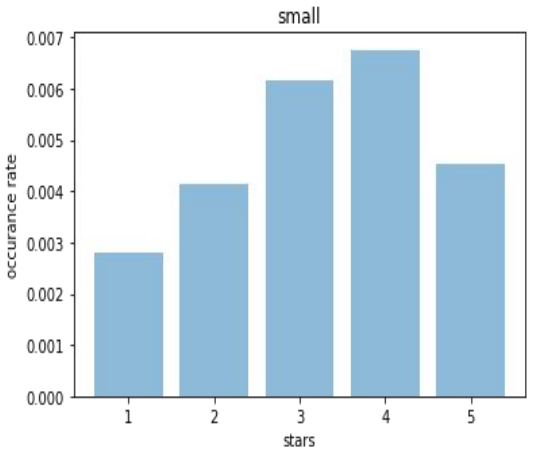
- Categories (main concerns)
  - Food: cookie, cone, topping, flavor, etc
  - Alternative food: coffee, sandwiches, etc
  - External: service, location, ambience, waiting time, etc

General Findings

- **Words Frequency v.s. Rating**
  - Due to unbalanced sample size, we use **occurrence rate** (number of the word/number of samples)
  - Possible factors: Service quality, food taste, food size (not necessarily)







Users

- We would like to use the table -- Users mainly by subjectively assigning weights to different reviews.
    - We cannot really determine whether a review is important or not based on the info of users, so the weights shouldn’t vary too much for different reviews.
- And specific for our model, we just sum the votes or compliments users sent and received separately.
- Using the simplified data, we can turn to some related research or models, and find out what model suits us best.

Tips

- We are not clear how to treat the text in Tips table differently, so we actually appended tips to the review, and analyze reviews text together with tips.
- The times of certain business or users shows up in the Tips table can also provide us with some useful information to explain the rating.
- In a nutshell, we mainly use these two tables in assigning weights to the reviews in the following model.

Future model

An ordinary regression model will be constructed based on the dataset.

- Target variables: Yelp ratings
- Predictors:
  - Locations (from which city);
  - Concerned attributes selected above (like price, open hours etc.) from business;
  - Some words appearing with high frequency (customers preferred like strawberry, vanilla flavor) selected from reviews;
  - Users analysis.

Suggestions may be provided

- We will provide multiple choices containing the predictors used in the linear model for merchant who want to get suggestions.
- Yelp score will be provided based on the information from merchant.
- According to the predicted score above, corresponding suggestions constructed from text analysis progress will be shown as well.

Thanks