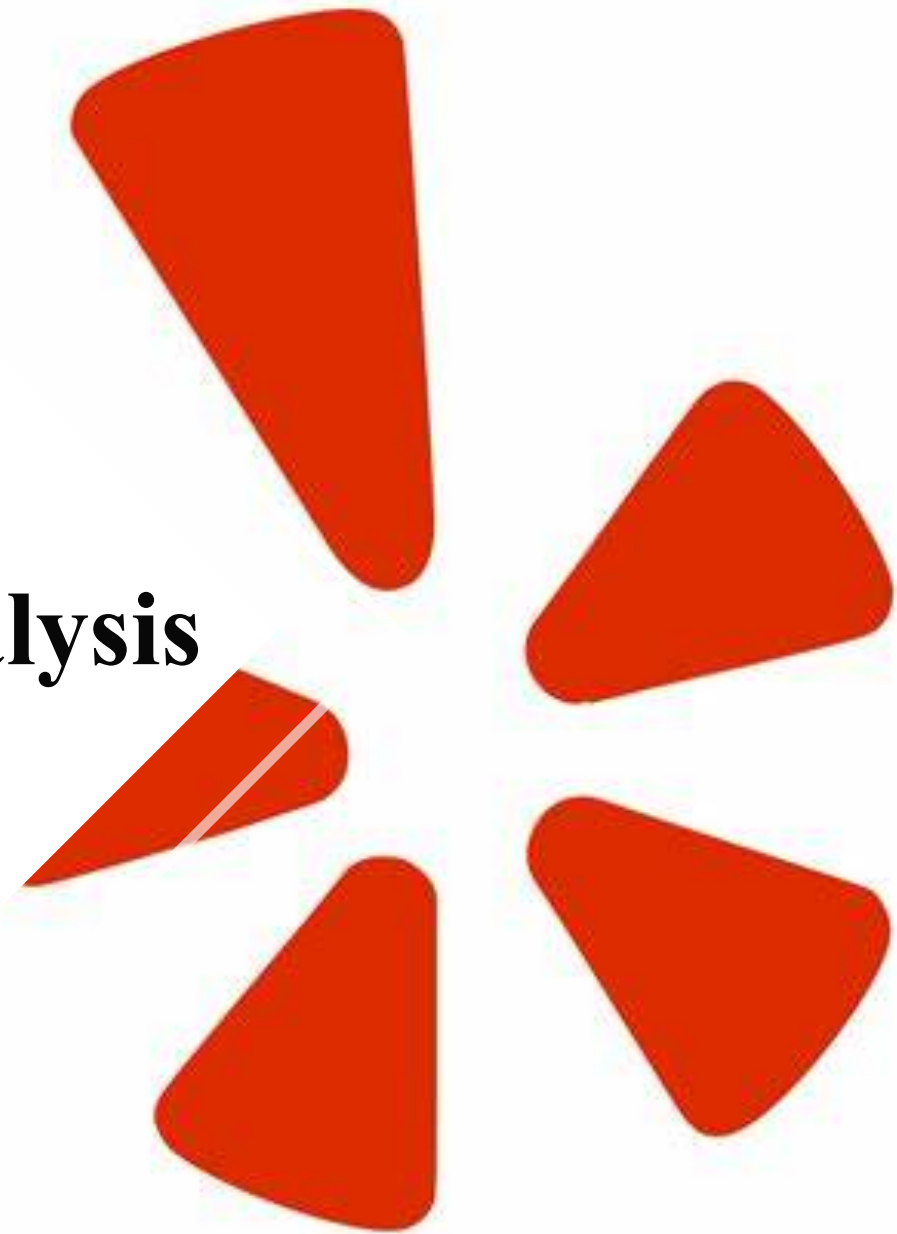
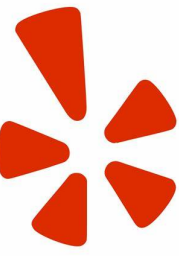




Yelp Data Analysis

Group 1

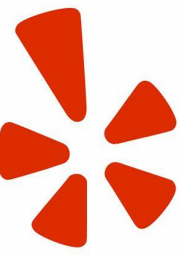




Overview of the raw data

- business.json
- covid.json
- review.json
- tip.json
- user.json

[illegible][illegible]



Data Extraction

- After taking a thorough look at the five datasets and considering their usability, “business.json” and “review.json” become the core of the following analysis.

business.json

covid.json

review.json

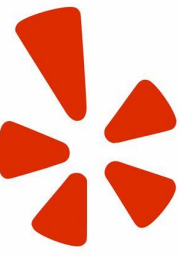
tip.json

user.json

- The common interests of our group is on gym. After considering the volume of dataset of gyms:

Dataset:	review	review on gyms
Length:	8635403	58459

we decide to move on with gyms.



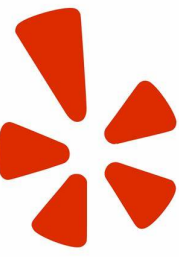
Data Extraction

Processed Dataset	Size
business_Gym.json	2052
review_Gym.json	58459
tip_Gym.json	9724
user_Gym.json	47124



Text Processing & Tokenization

- I. Remove reviews that is not written in English.



Text Processing & Tokenization

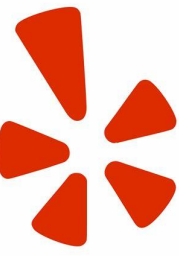
- I. Remove reviews that is not written in English.
- II. Remove all of punctuation except for *prime*.

e.g.

My visit from Vancouver, BC, was definitely worth the trip!! Can't wait to drive back and try out the new gym when it opens across the river.



My visit from Vancouver BC was definitely worth the trip Can't wait to drive back and try out the new gym when it opens across the river



Text Processing & Tokenization

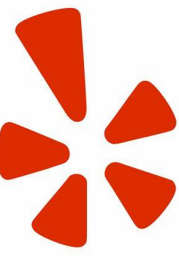
- I. Remove reviews that is not written in English.
- II. Remove all of punctuation except for *prime*.
- III. Turn each and every letter to lower-case letter.

e.g.

My visit from Vancouver BC was
definitely worth the trip Can't w
ait to drive back and try out the n
ew gym when it opens across the
river



my visit from vancouver bc was d
efinitely worth the trip can't wait
to drive back and try out the new
gym when it opens across the rive
r



Text Processing & Tokenization

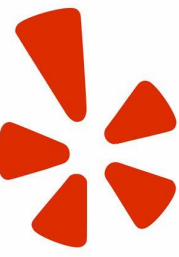
- I. Remove reviews that is not written in English.
- II. Remove all of punctuation except for *prime*.
- III. Turn each and every letter to lower-case letter.
- IV. Stopword removal.

e.g.

my visit from vancouver bc was d
efinitely worth the trip can't wait
to drive back and try out the new
gym when it opens across the rive
r



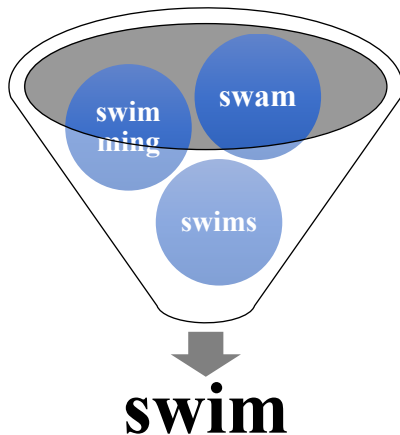
my visit from vancouver bc was d
efinitely worth the trip can't wait
to drive back and try out the new
gym when it opens across the rive
r



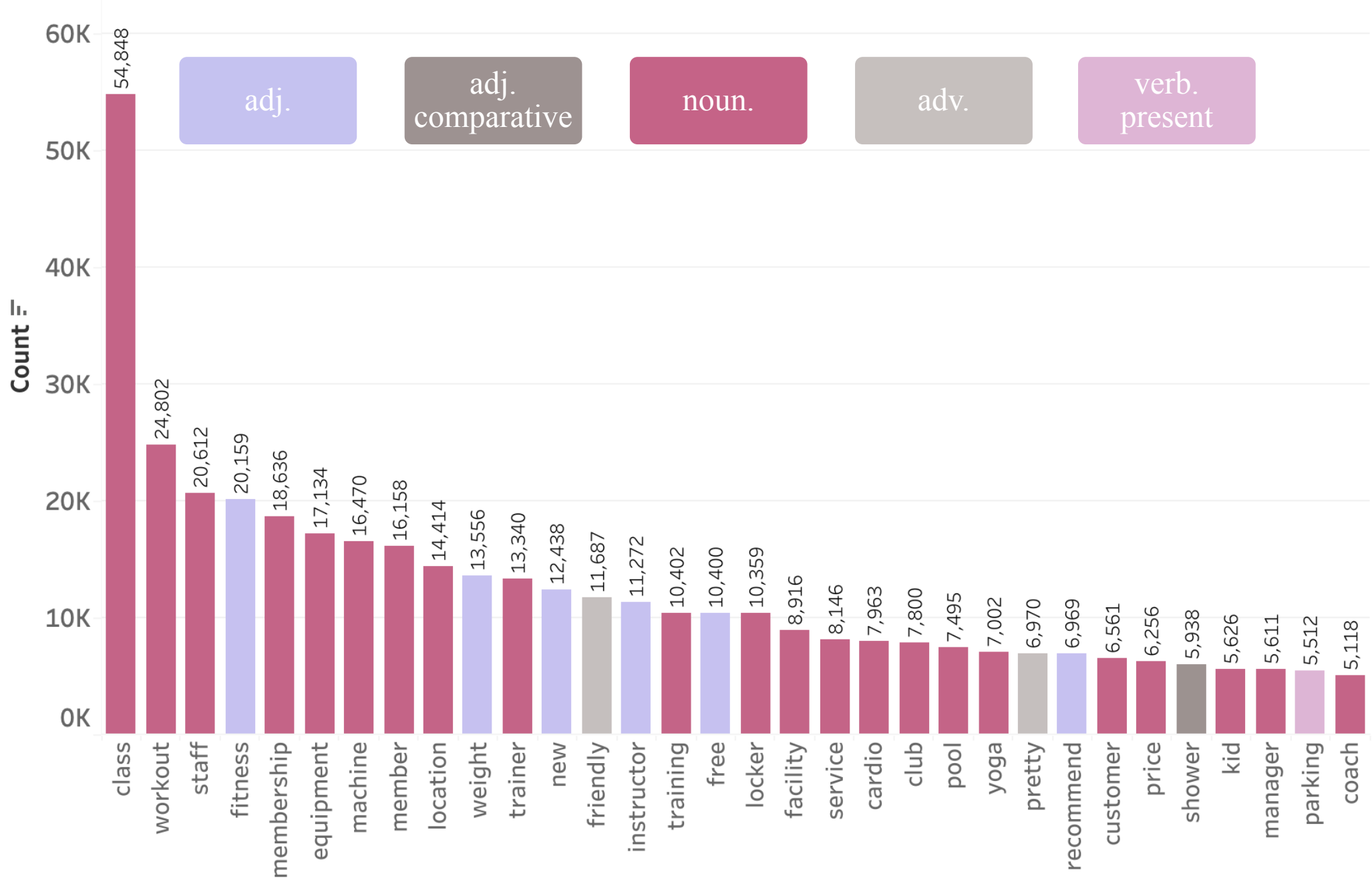
Text Processing & Tokenization

- I. Remove reviews that is not written in English.
- II. Remove all of punctuation except for *prime*.
- III. Turn each and every letter to lower-case letter.
- IV. Stopword removal.
- V. Lemmatization.

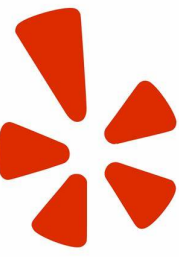
e.g.



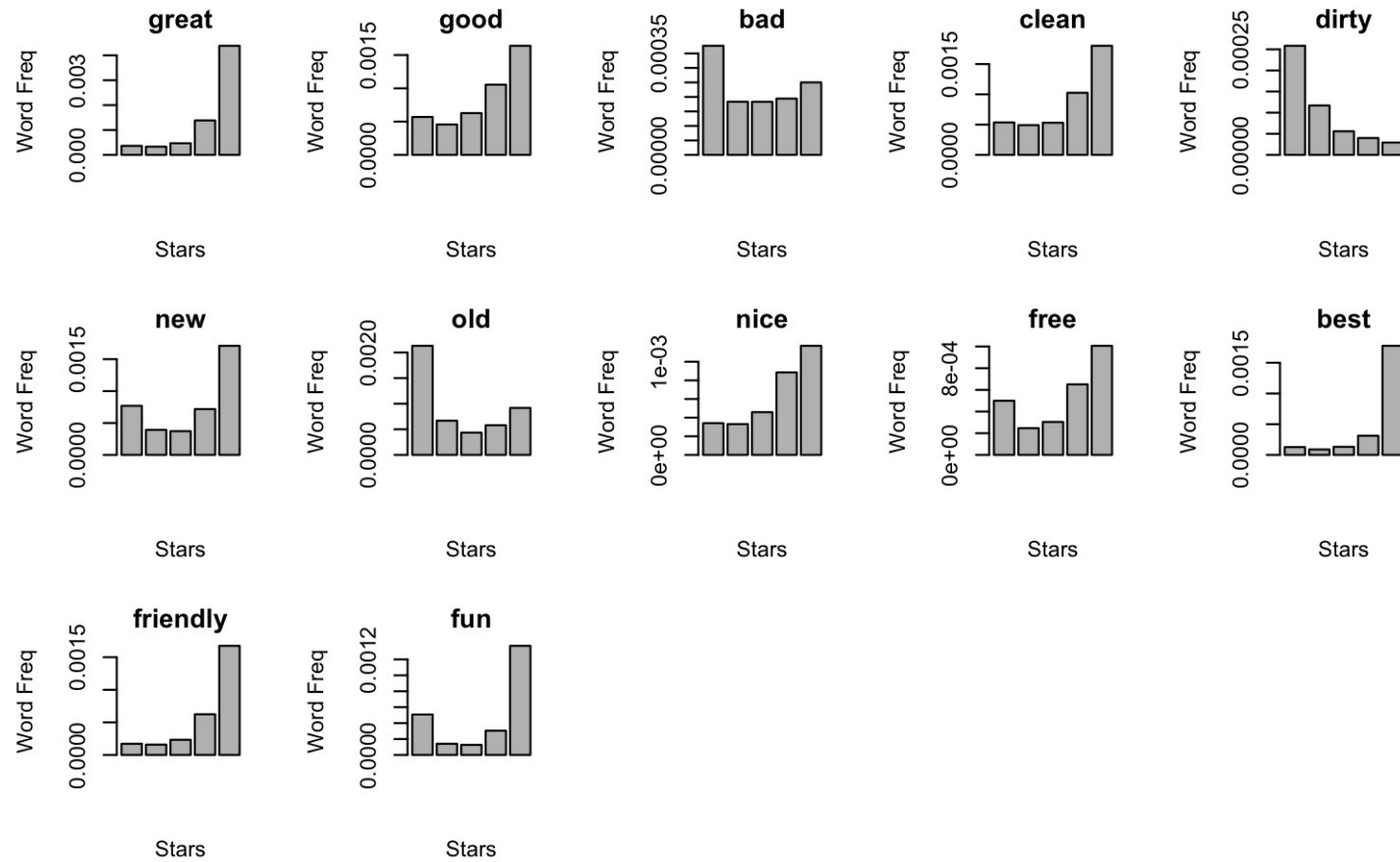
Visualization



Word count of interesting words
(threshold >5000)



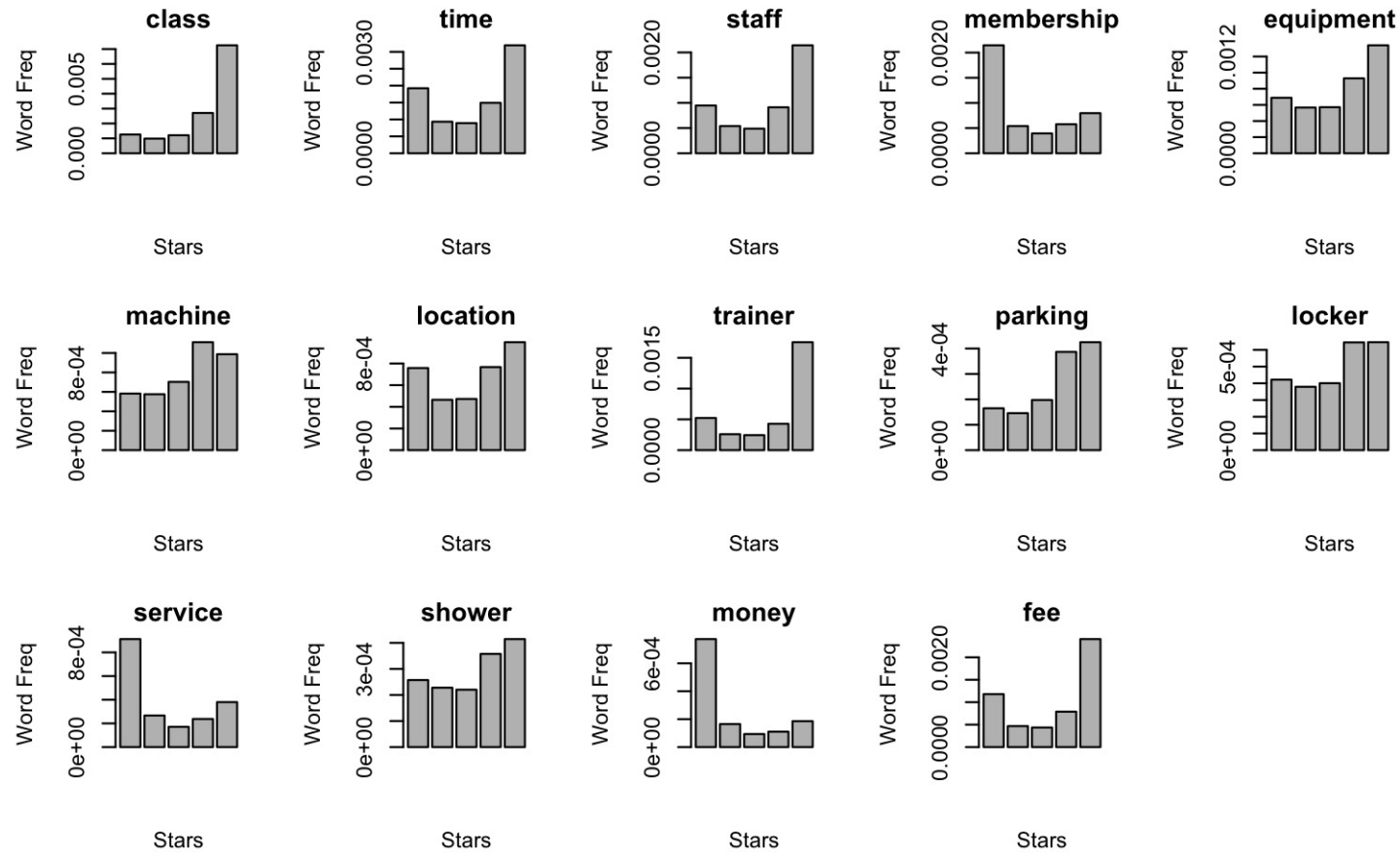
High-frequency words



Adjective

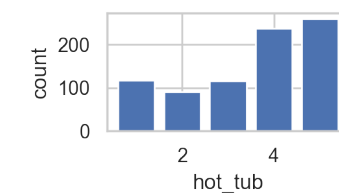
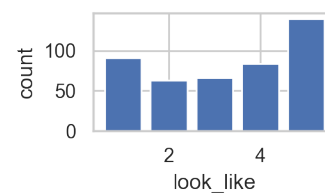
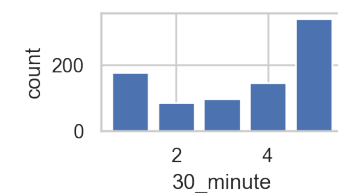
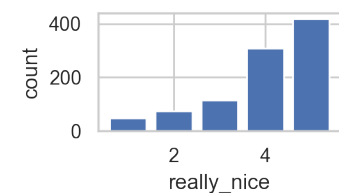
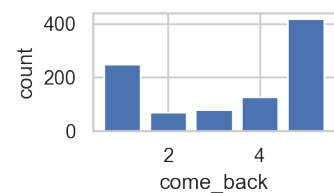
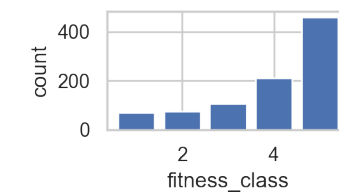
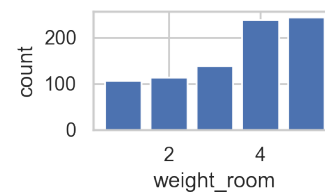
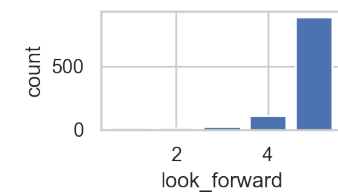
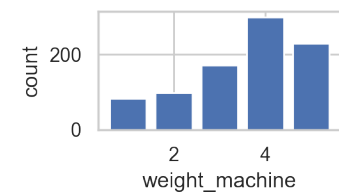
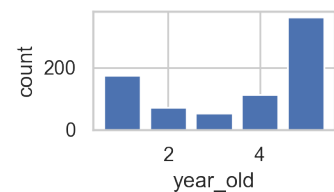
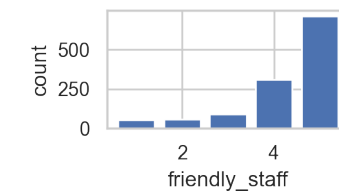
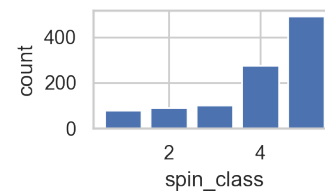
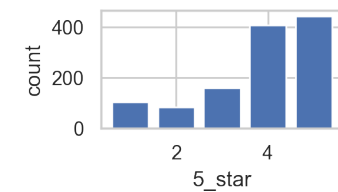
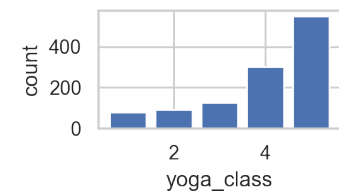
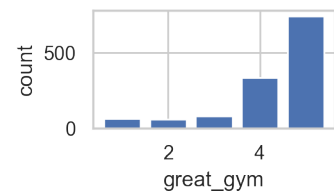
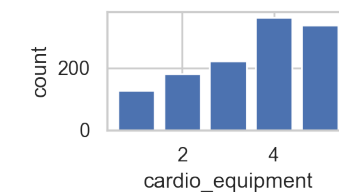
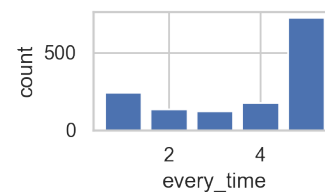
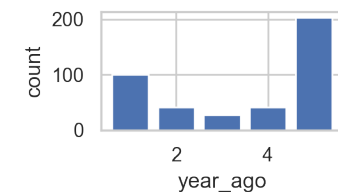
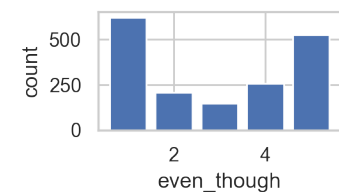
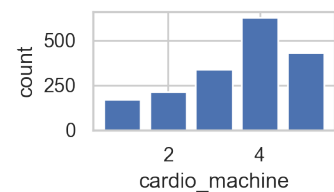
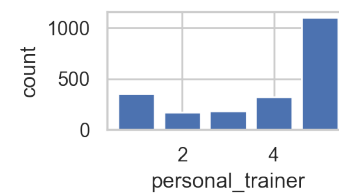
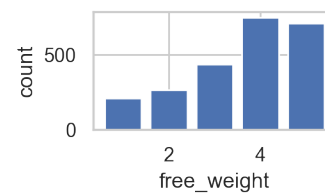
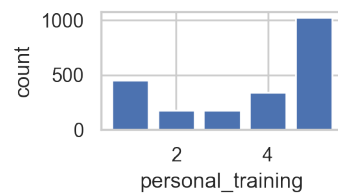
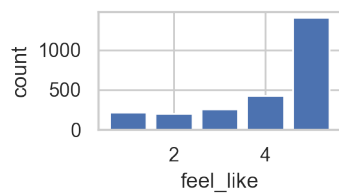
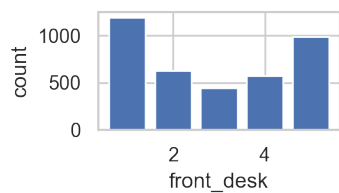


High-frequency words

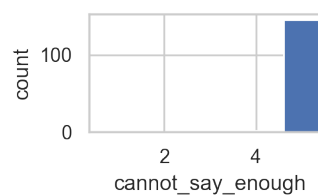
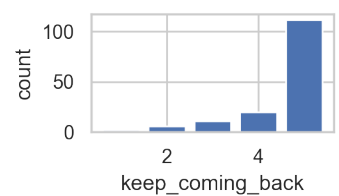
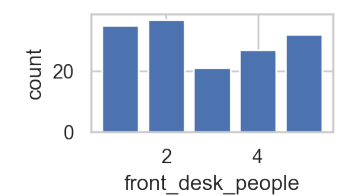
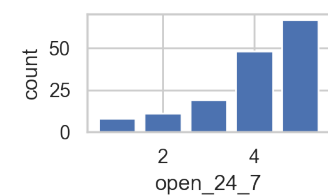
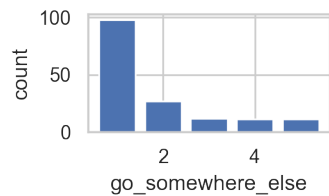
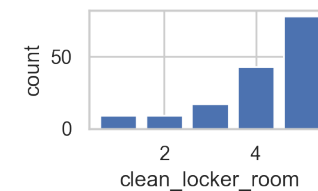
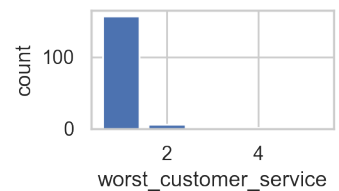
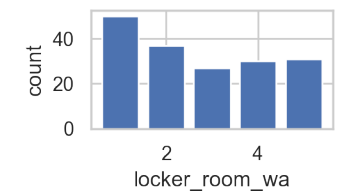
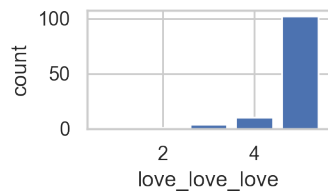
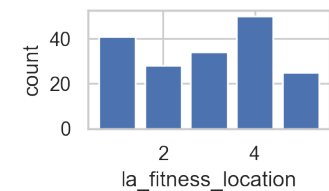
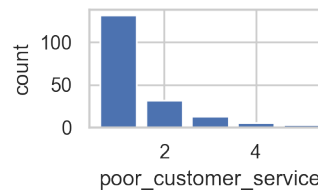
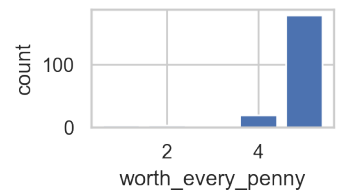
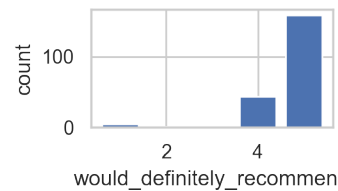
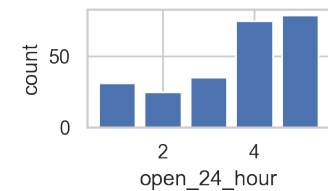
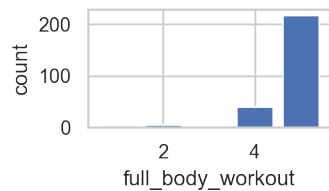
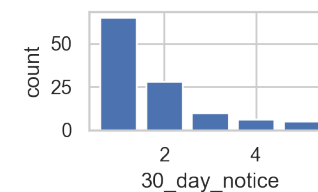
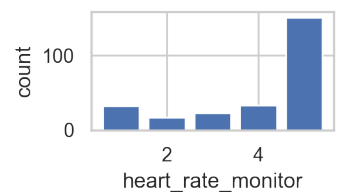
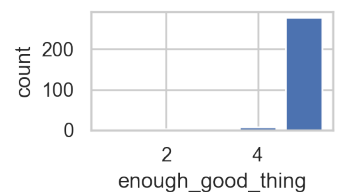
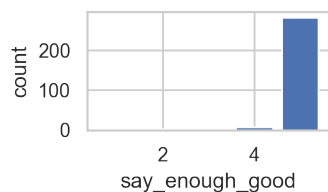
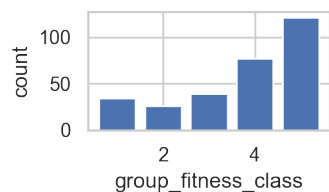
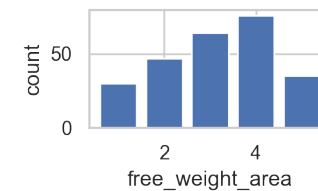
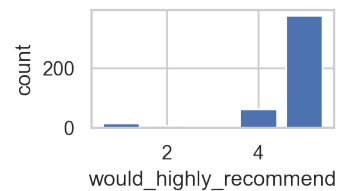
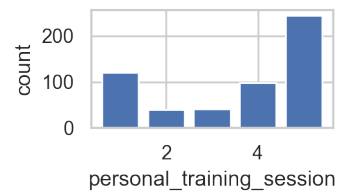
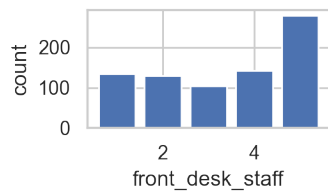
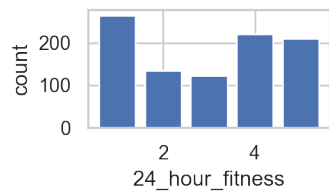


Noun

2_grams



3_grams



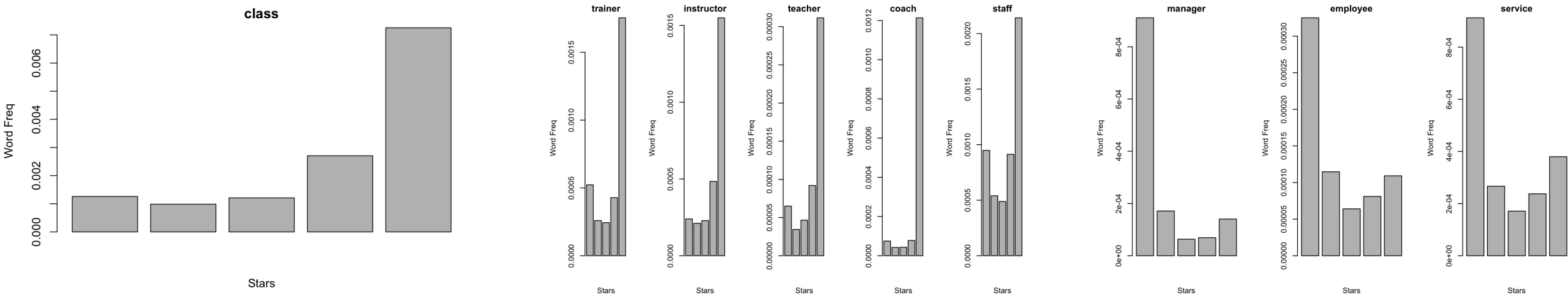
Statistical Analysis

Linear Model Results

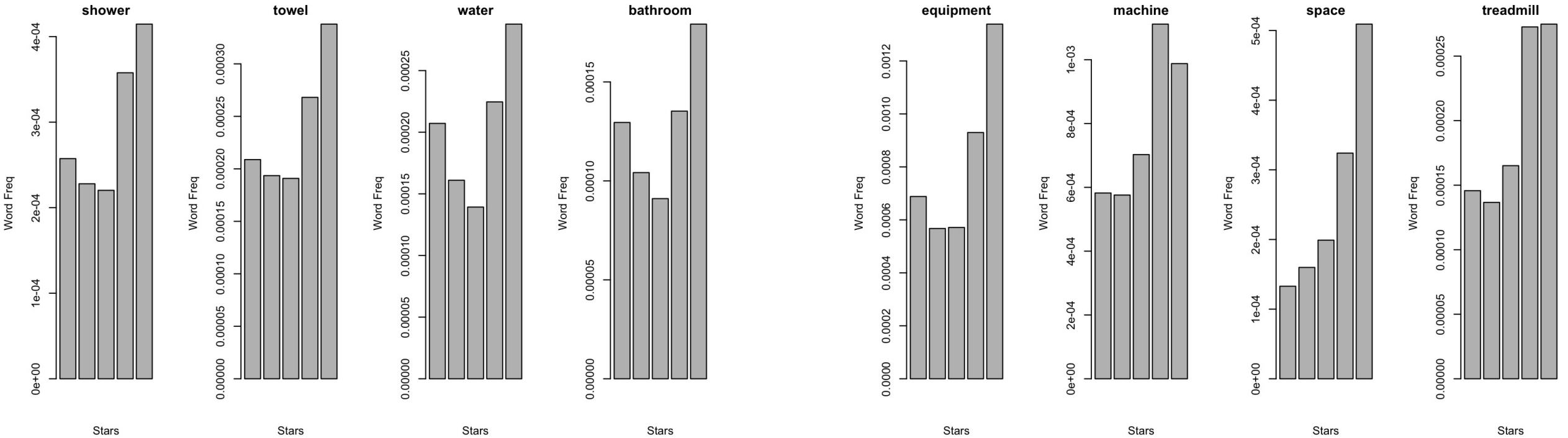
Attribute	Effect	P-value
Wi-Fi (<i>TRUE</i>)	Positive Effect	0.0123
Dog-friendly(<i>TRUE</i>)	Positive Effect	0.000341
Parking Space in big cities(<i>TRUE</i>)	Positive Effect	0.0447

*Note that parking space available or not is not important to gyms located in smaller towns.

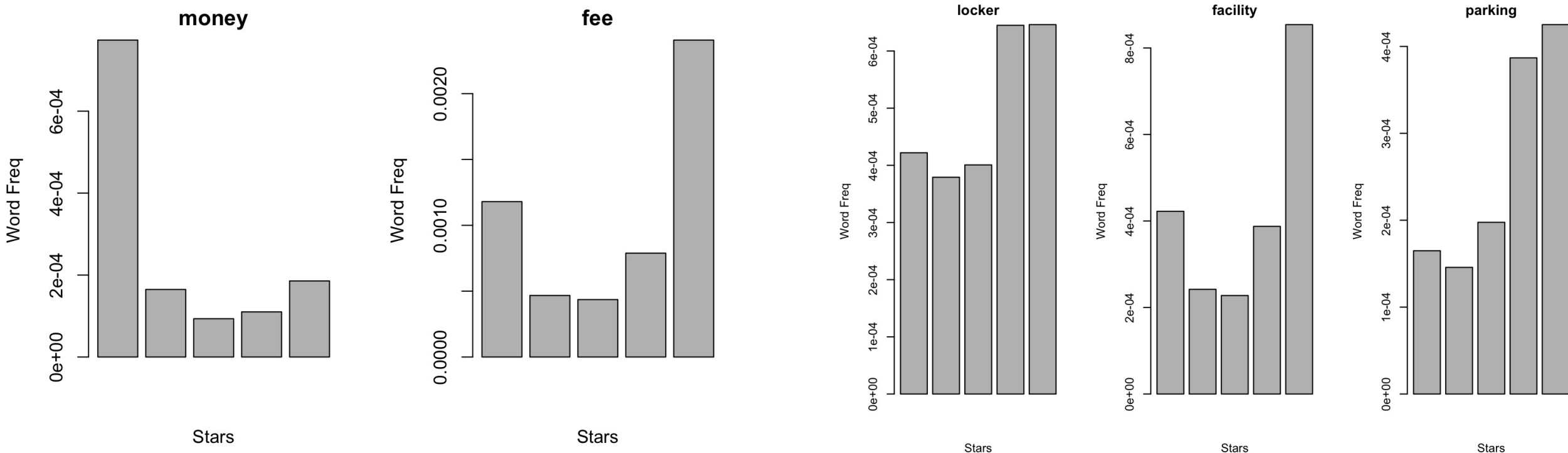
MERGE KEYWORDS



MERGE KEYWORDS



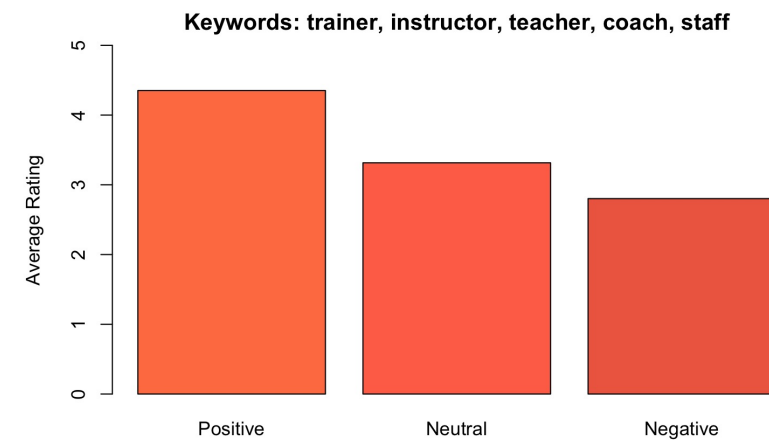
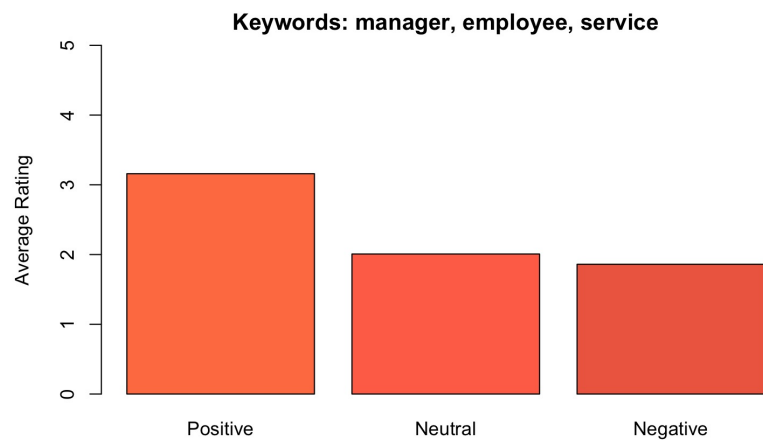
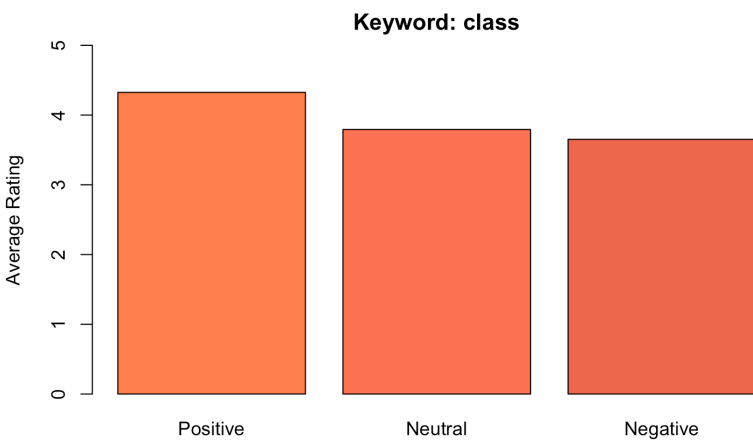
MERGE KEYWORDS



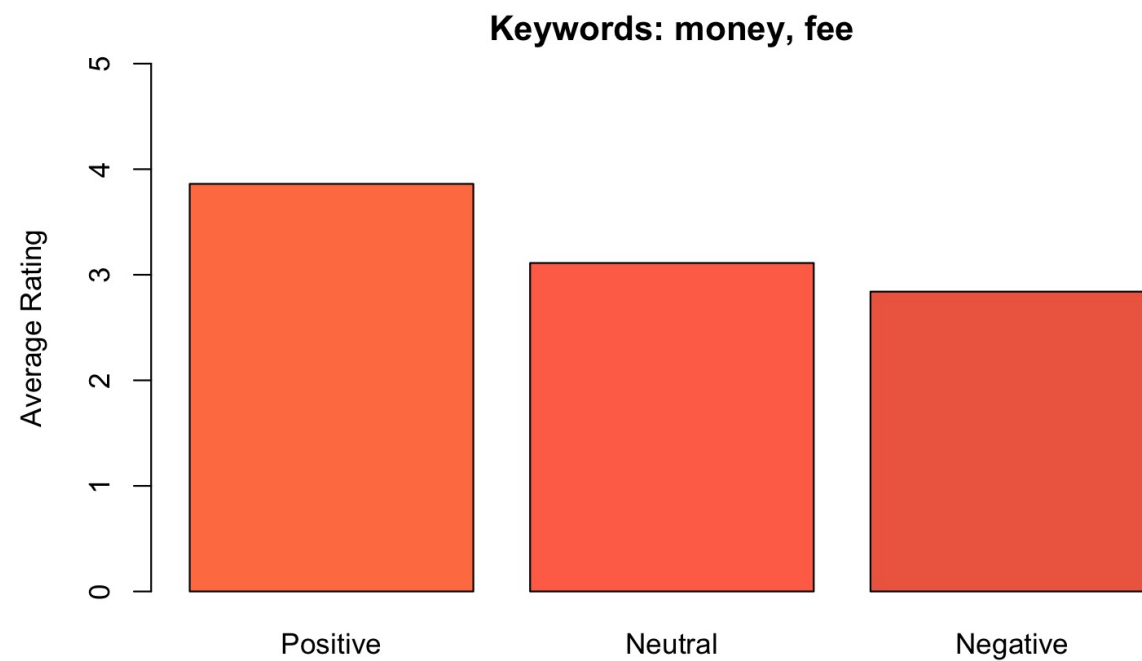
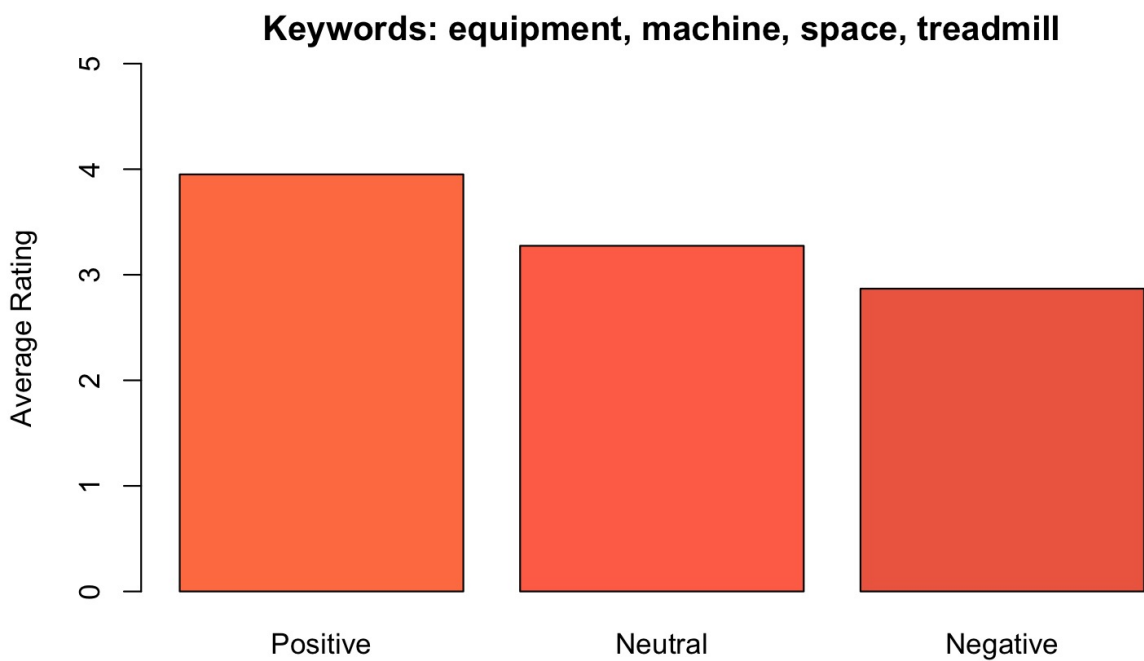
MERGE KEYWORDS

Keywords Groups	Number of relative reviews	Percentage of mentioning in reviews
class	21972	37.59%
trainer, instructor, teacher, coach, staff	30622	52.38%
manager, employee, service	10197	17.44%
equipment, machine, space, treadmill	21319	36.47%
shower, towel, water, bathroom	9320	15.94%
locker, facility, parking	13160	22.51%
money, fee	17932/58459	30.67%

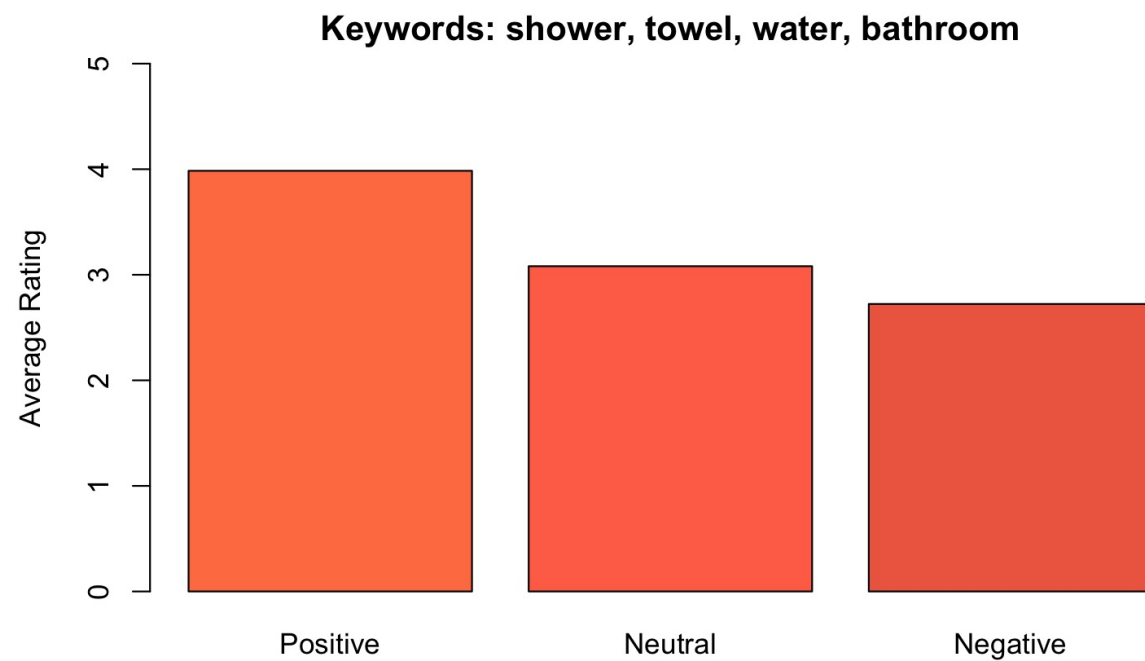
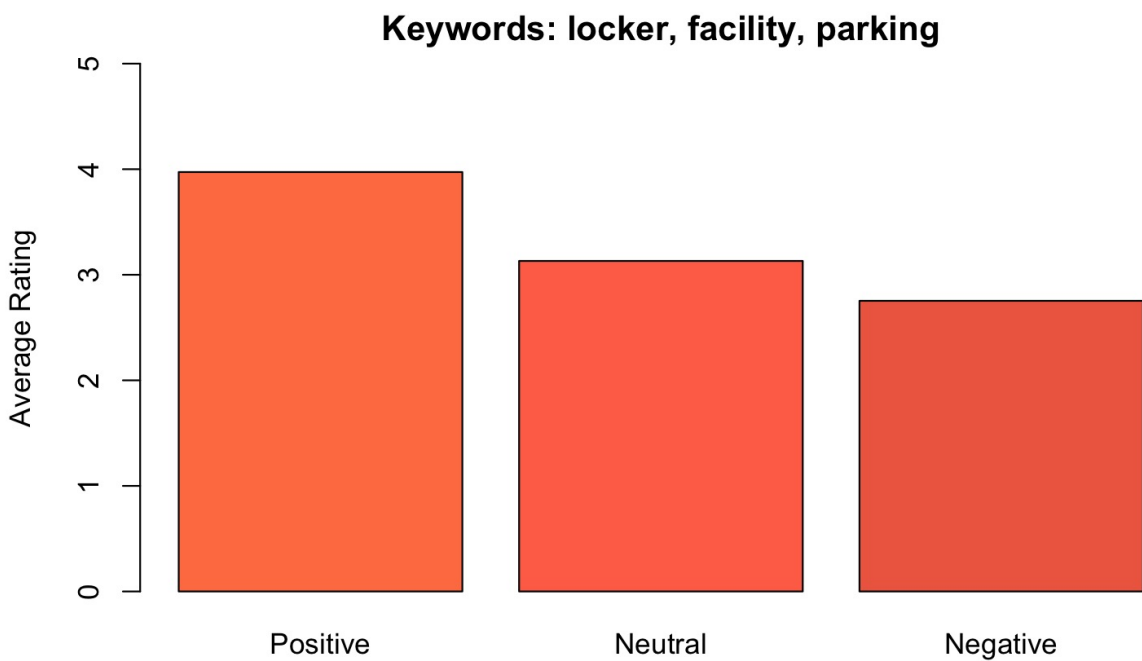
Non-attributes statistical analysis



Non-attributes statistical analysis



Non-attributes statistical analysis



Non-attributes statistical analysis

Keywords Group: class		
Negative vs Neutral	Negative vs Positive	Neutral vs Positive
P-value = 2.2e-05	P-value < 2.2e-16	P-value < 2.2e-16

Keywords Group: trainer, instructor, teacher, coach, staff		
Negative vs Neutral	Negative vs Positive	Neutral vs Positive
P-value < 2.2e-16	P-value < 2.2e-16	P-value < 2.2e-16

Keywords Group: manager, employee, service		
Negative vs Neutral	Negative vs Positive	Neutral vs Positive
P-value < 4.394e-05	P-value < 2.2e-16	P-value < 2.2e-16

Non-attributes statistical analysis

Keywords Group: equipment, machine, space, treadmill		
Negative vs Neutral	Negative vs Positive	Neutral vs Positive
P-value < 2.2e-16	P-value < 2.2e-16	P-value < 2.2e-16

Keywords Group: shower, towel, water, bathroom		
Negative vs Neutral	Negative vs Positive	Neutral vs Positive
P-value < 2.2e-16	P-value < 2.2e-16	P-value < 2.2e-16

Keywords Group: locker, facility, parking		
Negative vs Neutral	Negative vs Positive	Neutral vs Positive
P-value < 2.2e-16	P-value < 2.2e-16	P-value < 2.2e-16

Non-attributes statistical analysis

Keywords Group: money, fee		
Negative vs Neutral	Negative vs Positive	Neutral vs Positive
P-value = 9.921e-15	P-value < 2.2e-16	P-value < 2.2e-16

Suggestions

- **EXAMPLE:**

Performaces and Suggestions

Detailed information and corresponding evaluation/suggestion for any specific gym can be found on this page.

Input name:

CrossFit

Input address:

FL, Heathrow, 930 International Pkwy

Performance

Suggestions

Strength:

You have good classes to take.
You have perfect trainers and coaches.
The equipments are in good condition.
Customers feels good about the shower here.
You have good facilities.
Your prices are very good.

Weakness:

You need to improve the quality of service.

R SHINY APP



Review Analysis on GYM's



FREQUENT WORDS

MAPS

EVALUATIONS

Performances and Suggestions

Detailed information and corresponding evaluation/suggestion for any specific gym can be found on this page.

Input name:

CrossFit

Input address:

FL, Heathrow, 930 International Pkwy

Performance

Suggestions

Yelp Rating:



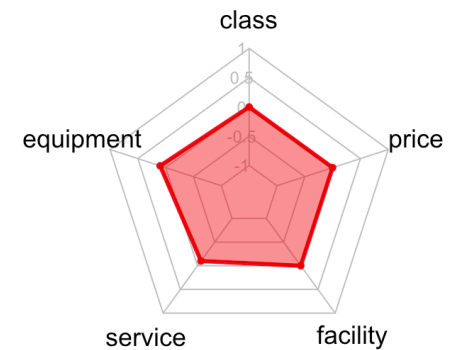
Popularity:



Attributes:

parking : False

Nonattributes





THANK YOU