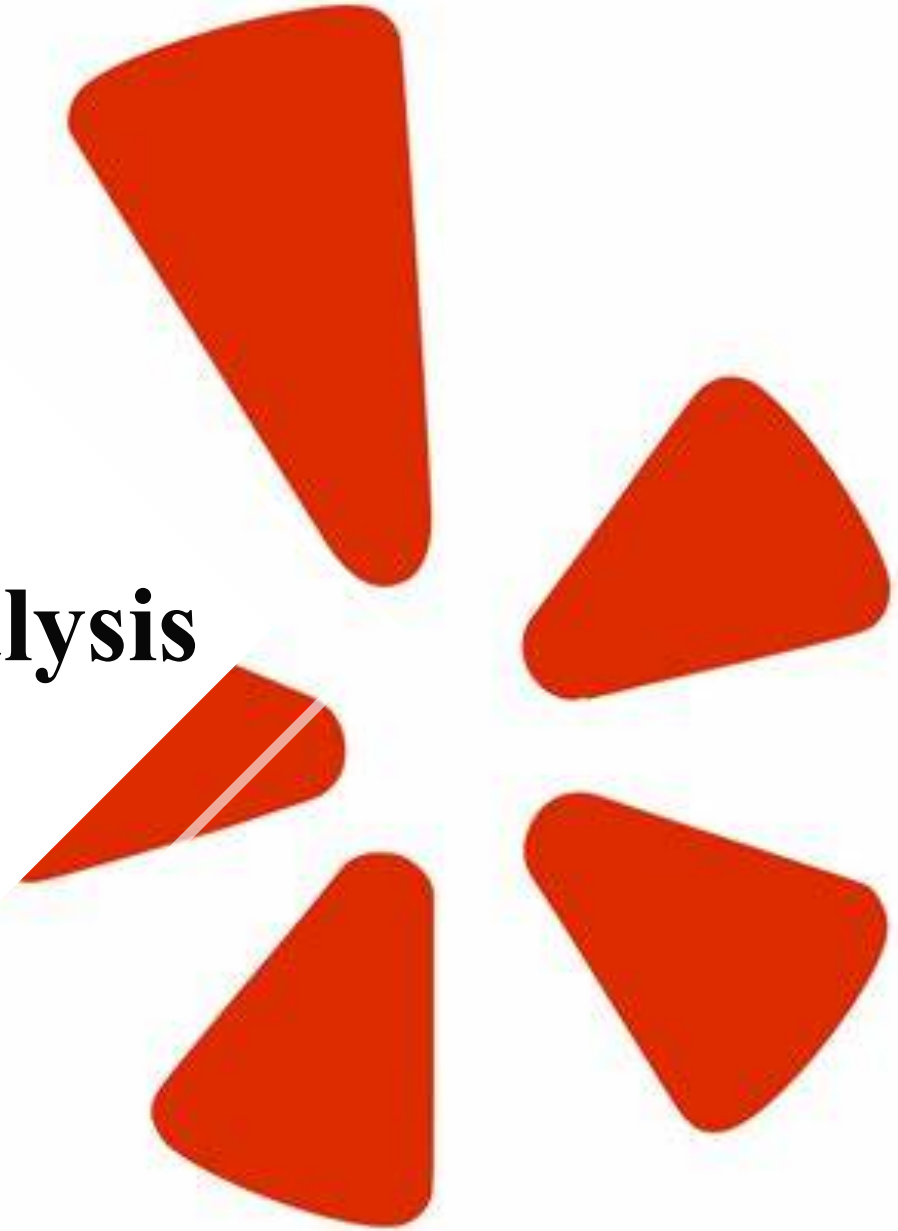
Yelp Data Analysis

Group 1



Overview of the raw data

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

Focus on the raw data

- business.json

business_id	name	address	city	state	postal_code	……	hours.Sunday
PE9uqAjdw0E4-8mjGl3wVA	Crossfit Terminus	1046 Memorial Dr SE	Atlanta	GA	30316		*NA*
……	……	……	……	……	……		……

- review.json

review_id	user_id	business_id	stars	useful	funny	……	date
WbZa7cFXLeIpooAXu77cyQ	xTr_EPGC46nfKBkBKjdKfA	1p2bfq4k-tY8Ea9JuolUmw	5	1	0		2013-07-12 21:58:59
……	……	……	……	……	……		……

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, wa s definitely worth the trip!! Can't wait to drive back and try out the new gym when it opens across th e river.	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

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.

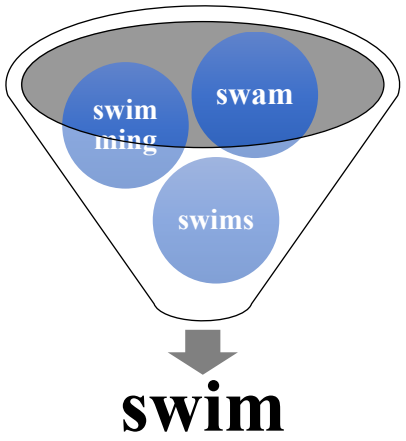
my visit from vancouver bc was d efinitely worth the trip can't wait to drive back and try out the new gym when it opens across the rive r

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

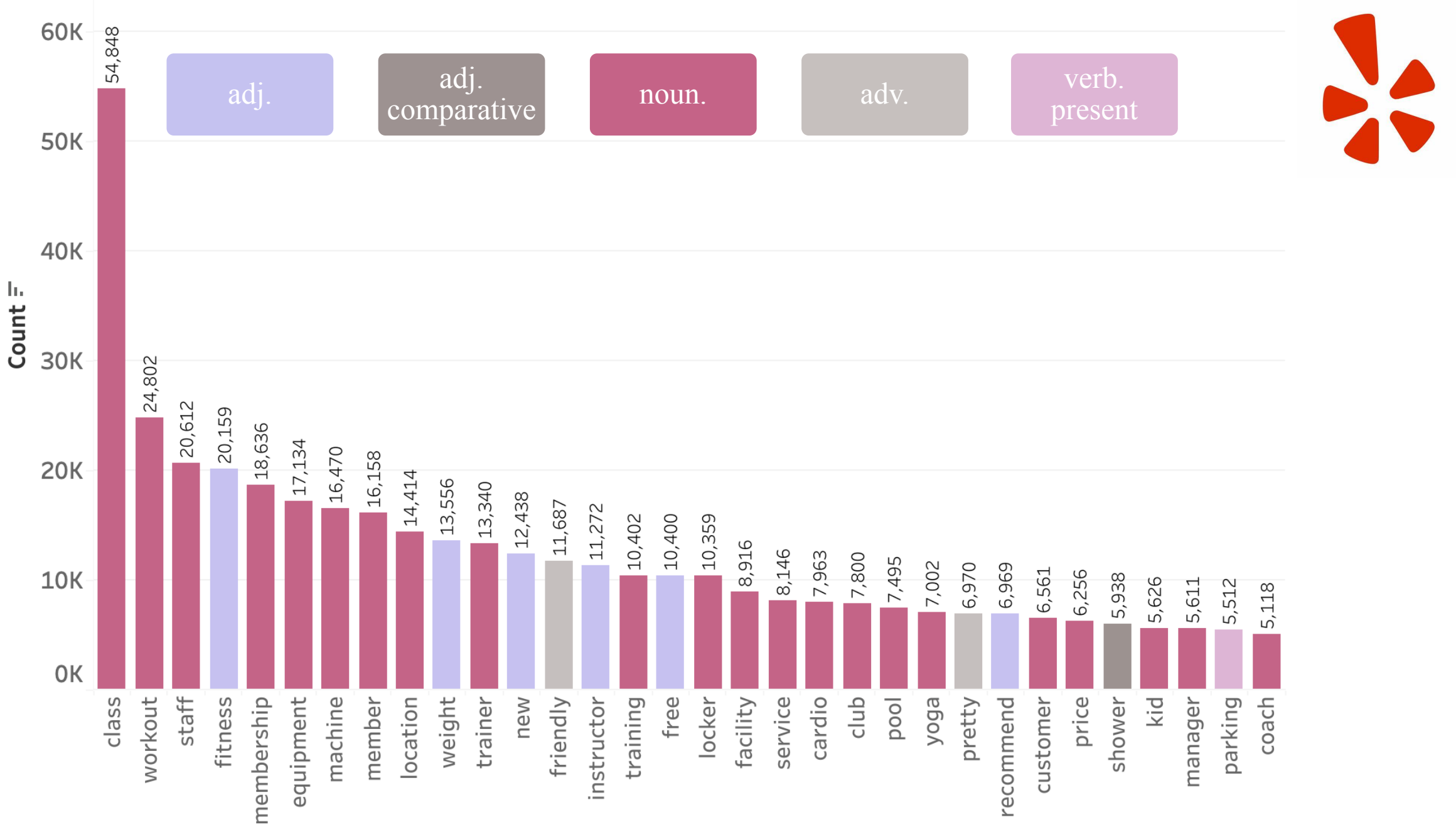
Text Processing & Tokenization

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.

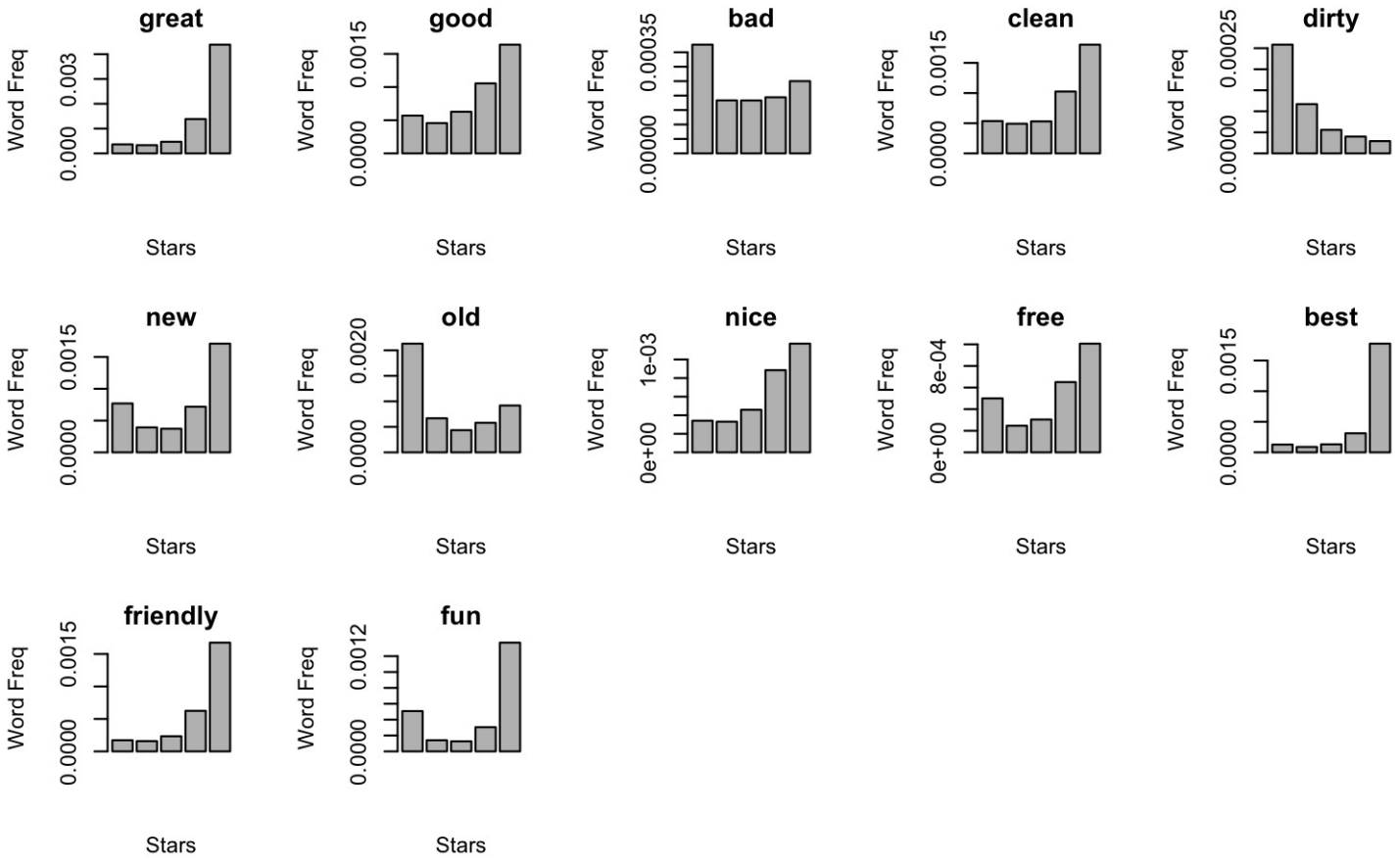


Visualization



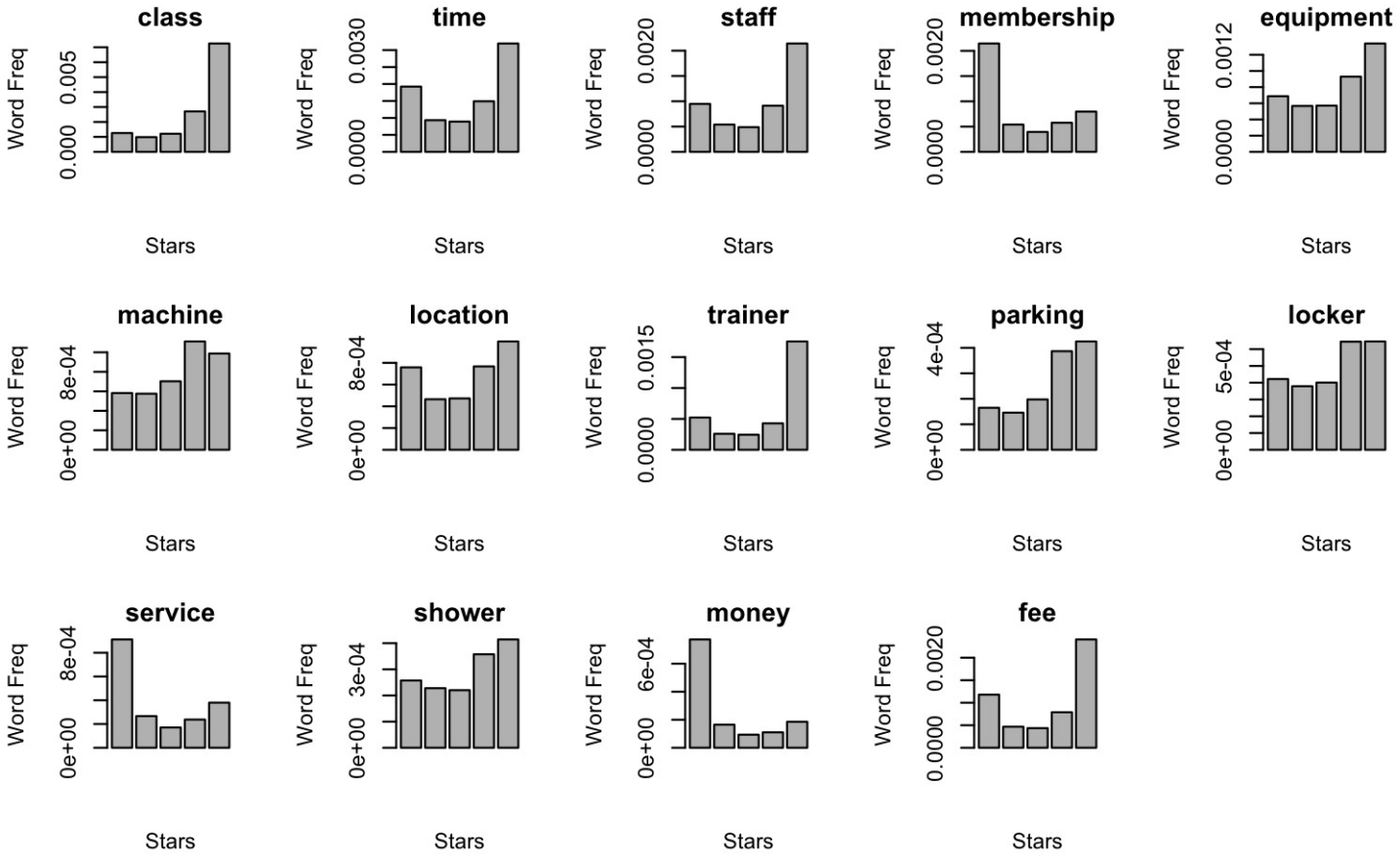
Word count of interesting words
(threshold >5000)

High-frequency words



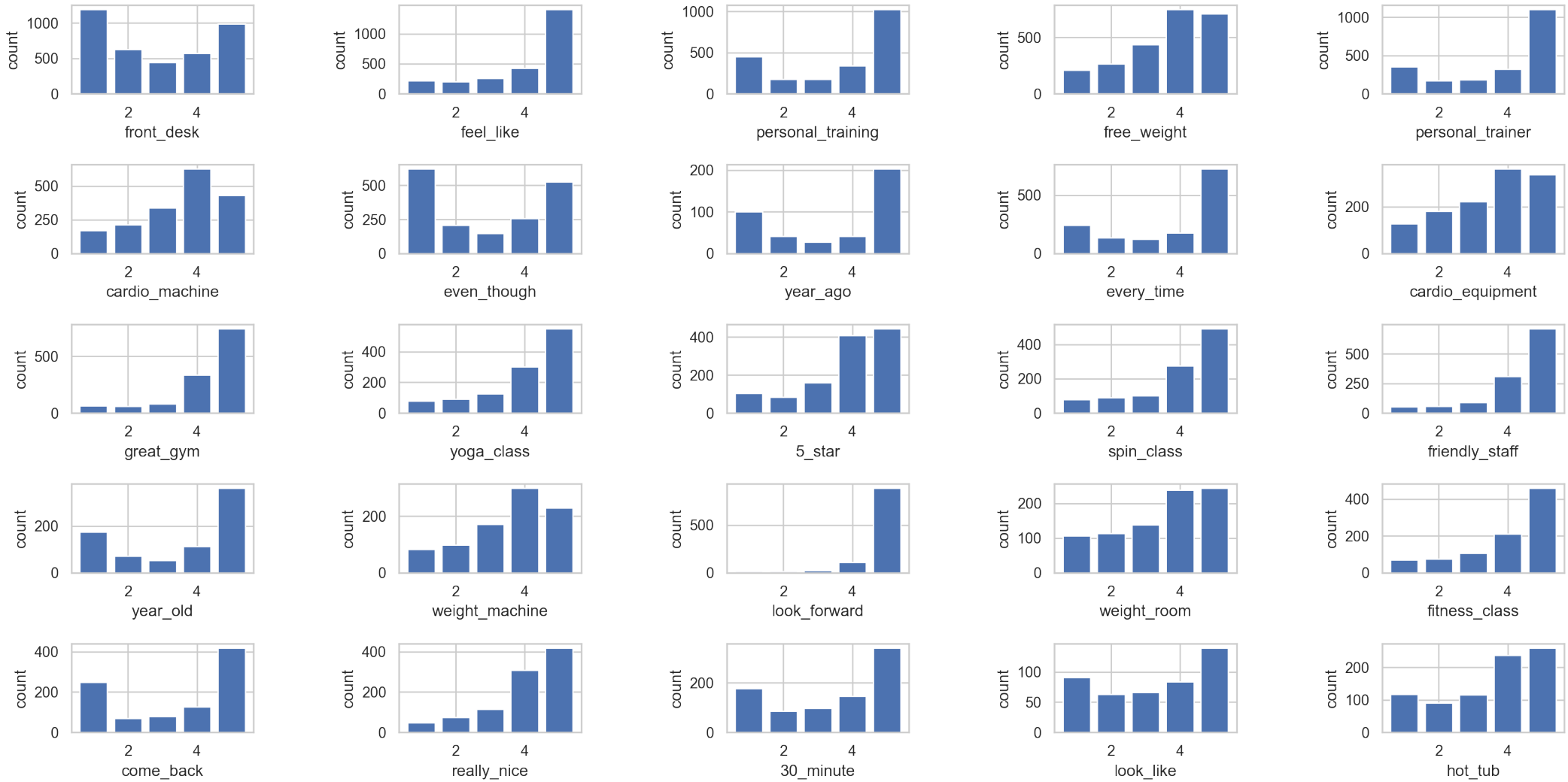
Adjective

High-frequency words

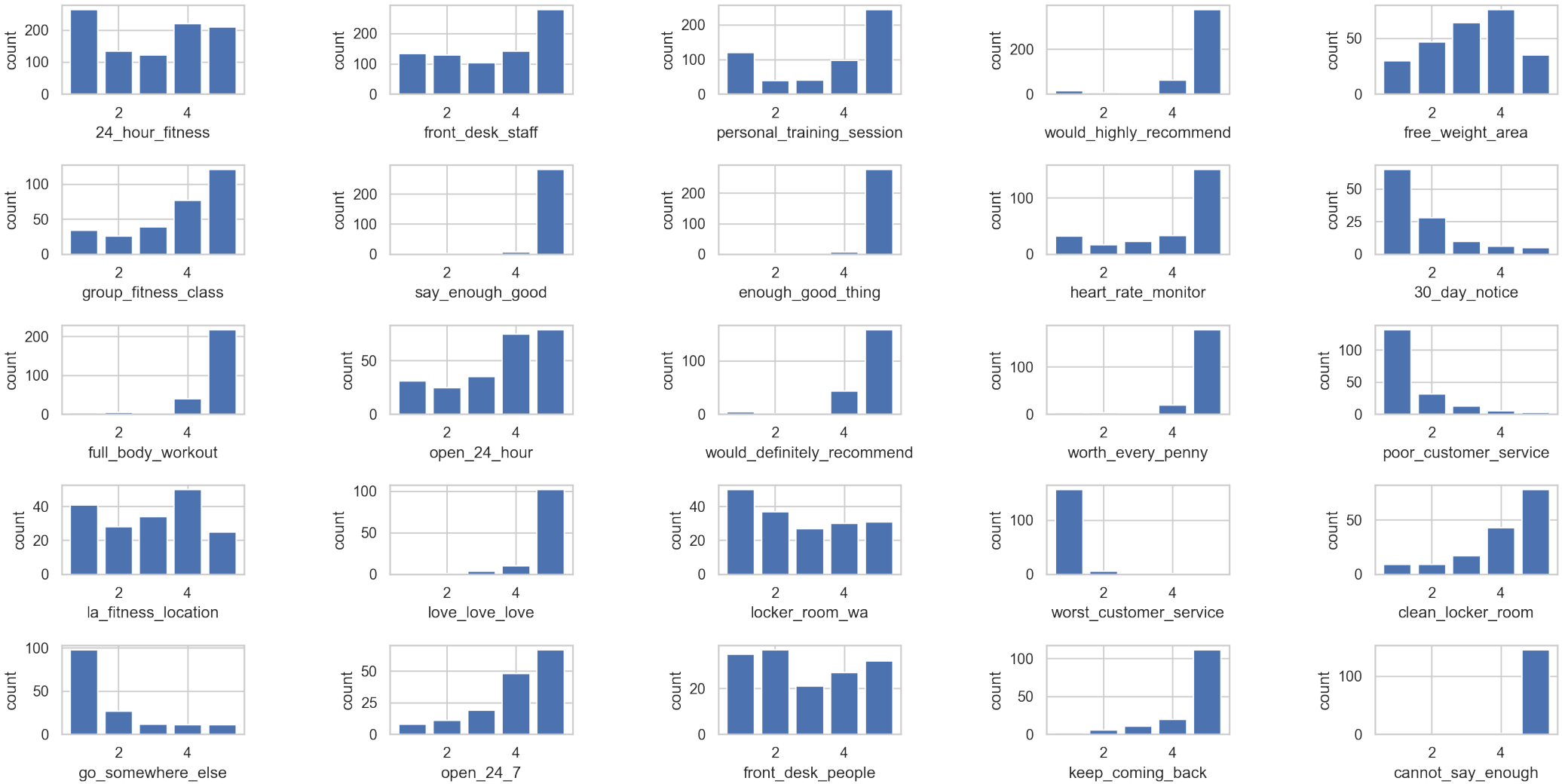


Noun

2_grams


front_desk
feel_like
personal_training
free_weight
personal_trainer
cardio_machine
even_though
year_ago
every_time
cardio_equipment
great_gym
yoga_class
5_star
spin_class
friendly_staff
year_old
weight_machine
look_forward
weight_room
fitness_class
come_back
really_nice
30_minute
look_like
hot_tub
count

3_grams


count
24_hour_fitness
front_desk_staff
personal_training_session
would_highly_recommend
free_weight_area
group_fitness_class
say_enough_good
enough_good_thing
heart_rate_monitor
30_day_notice
full_body_workout
open_24_hour
would_definitely_recommend
worth_every_penny
poor_customer_service
la_fitness_location
love_love_love
locker_room_wa
worst_customer_service
clean_locker_room
go_somewhere_else
open_24_7
front_desk_people
keep_coming_back
cannot_say_enough

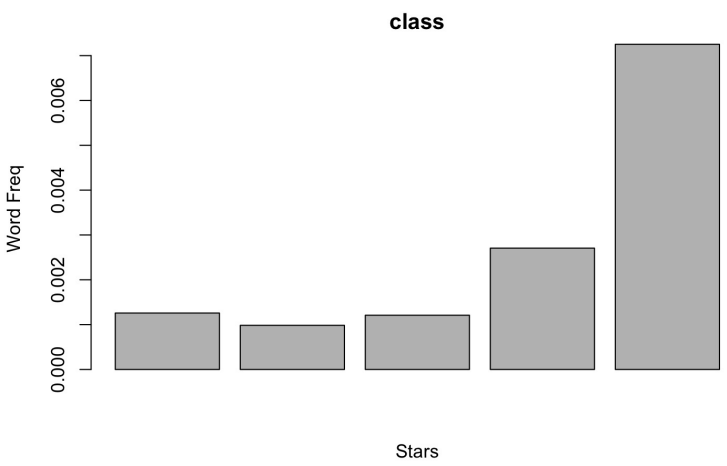
Statistical Analysis

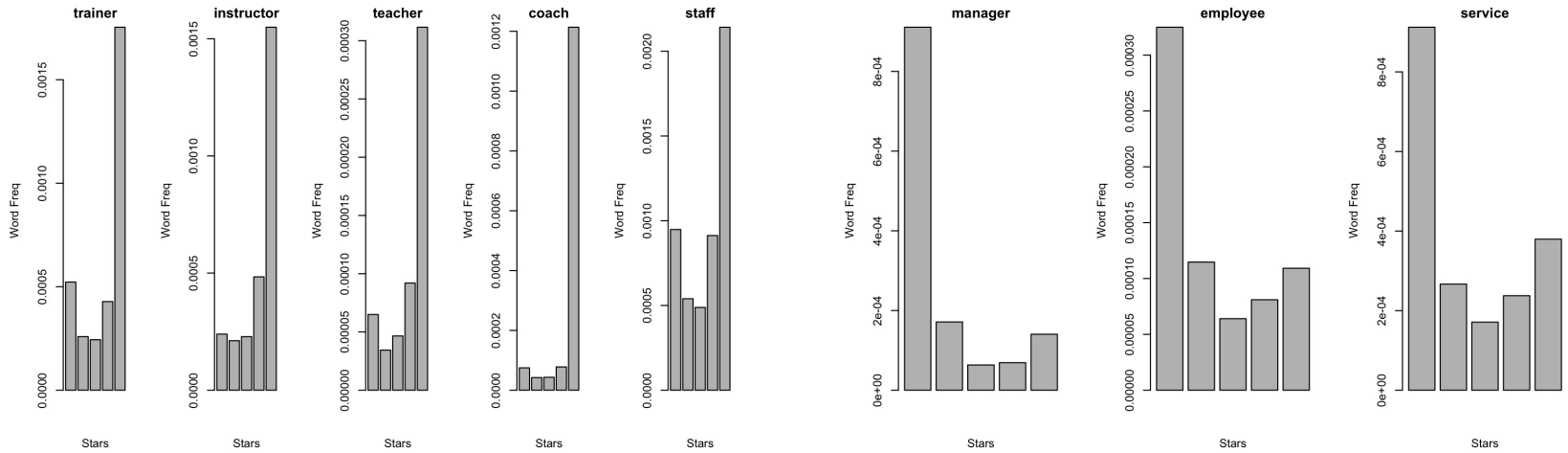
Linear Model Results

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

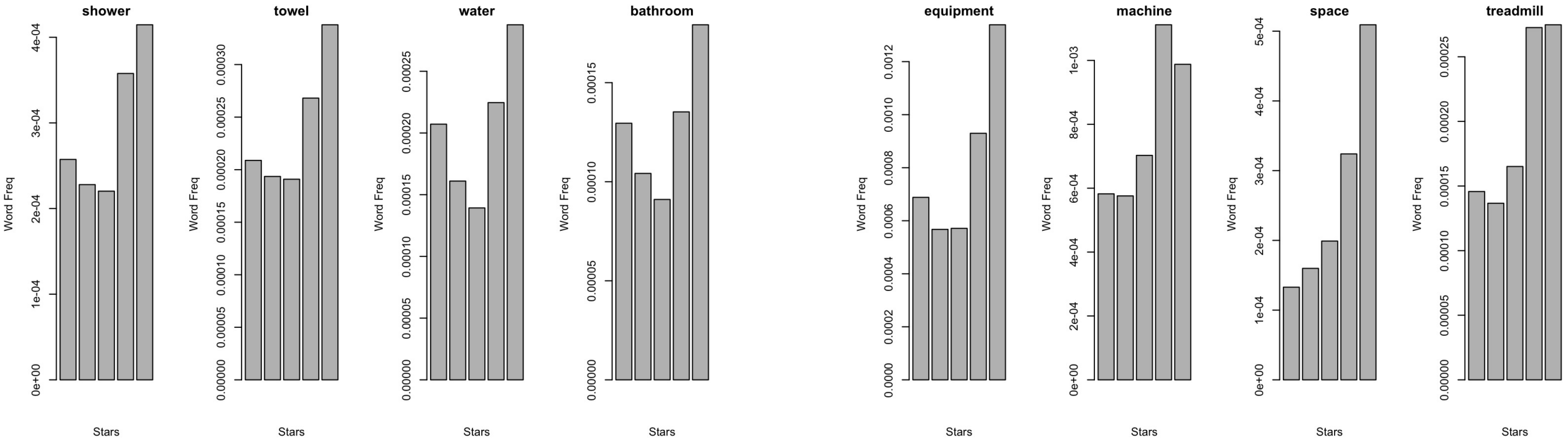
*Note that parking space available or not is not important to gyms located in smaller towns.

MERGE KEYWORDS

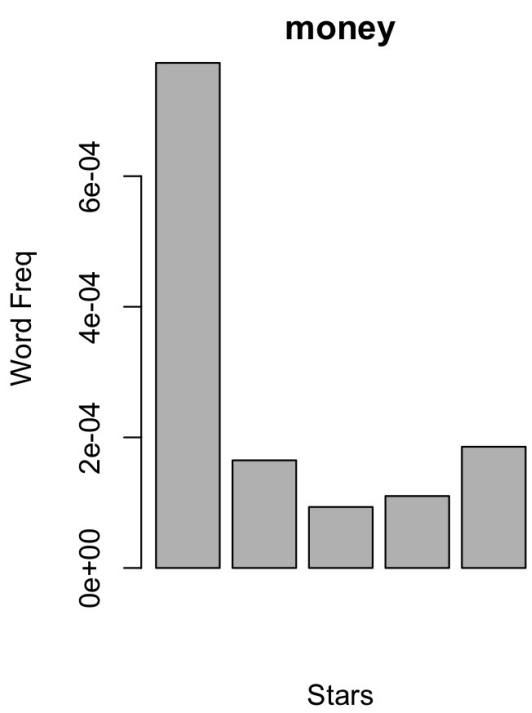


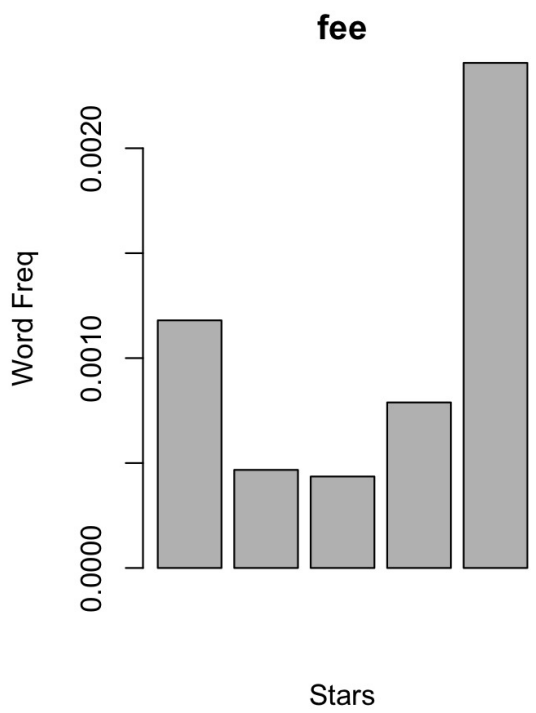


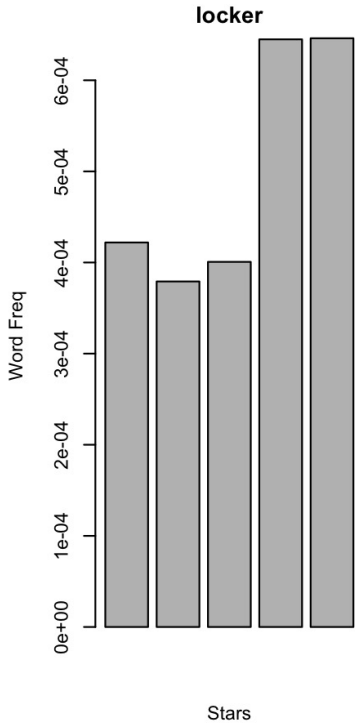
MERGE KEYWORDS

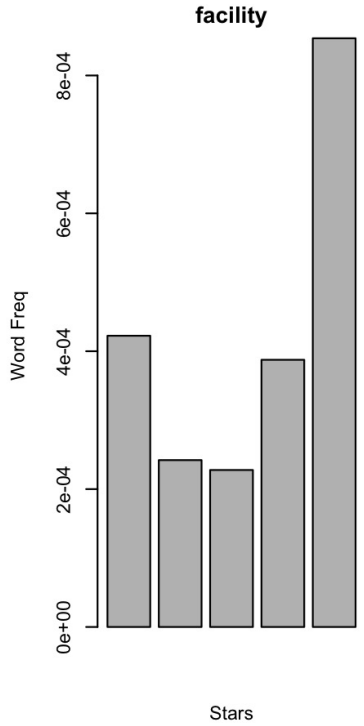


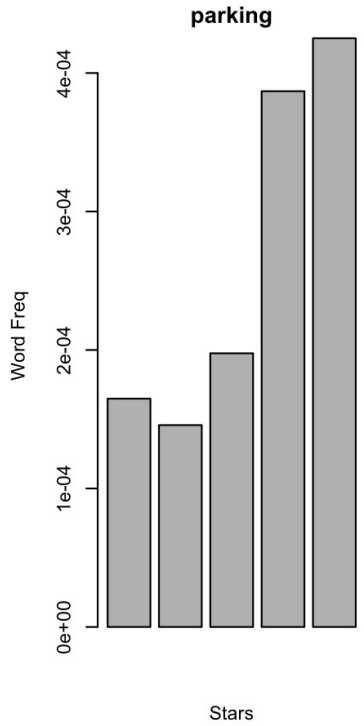
MERGE KEYWORDS







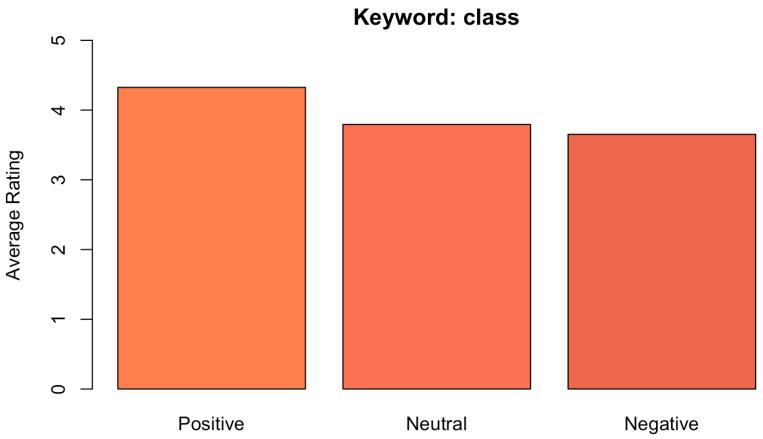


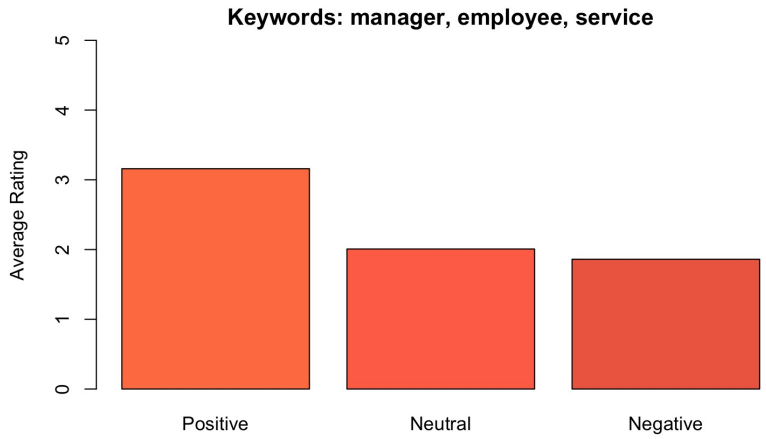


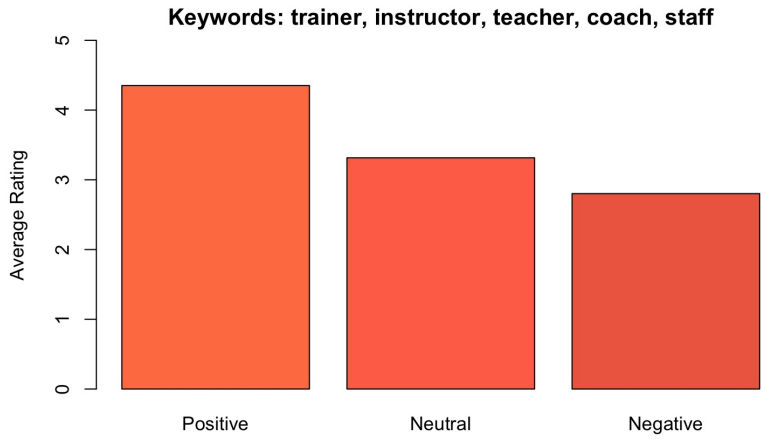
MERGE KEYWORDS

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	30.67%

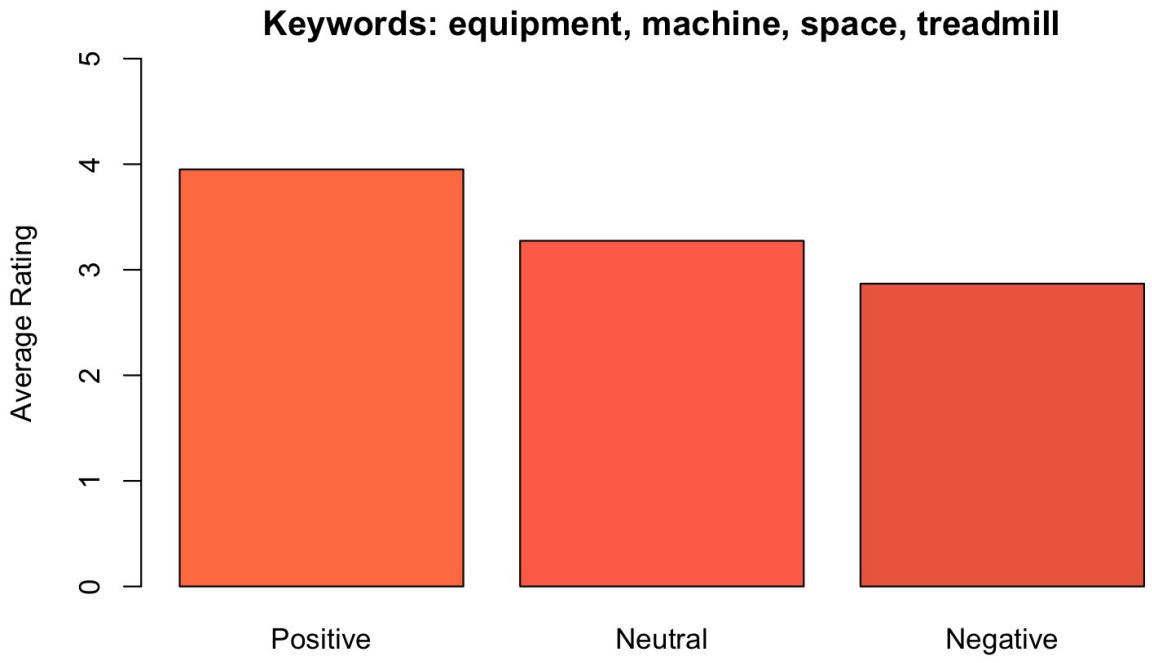
Non-attributes statistical analysis

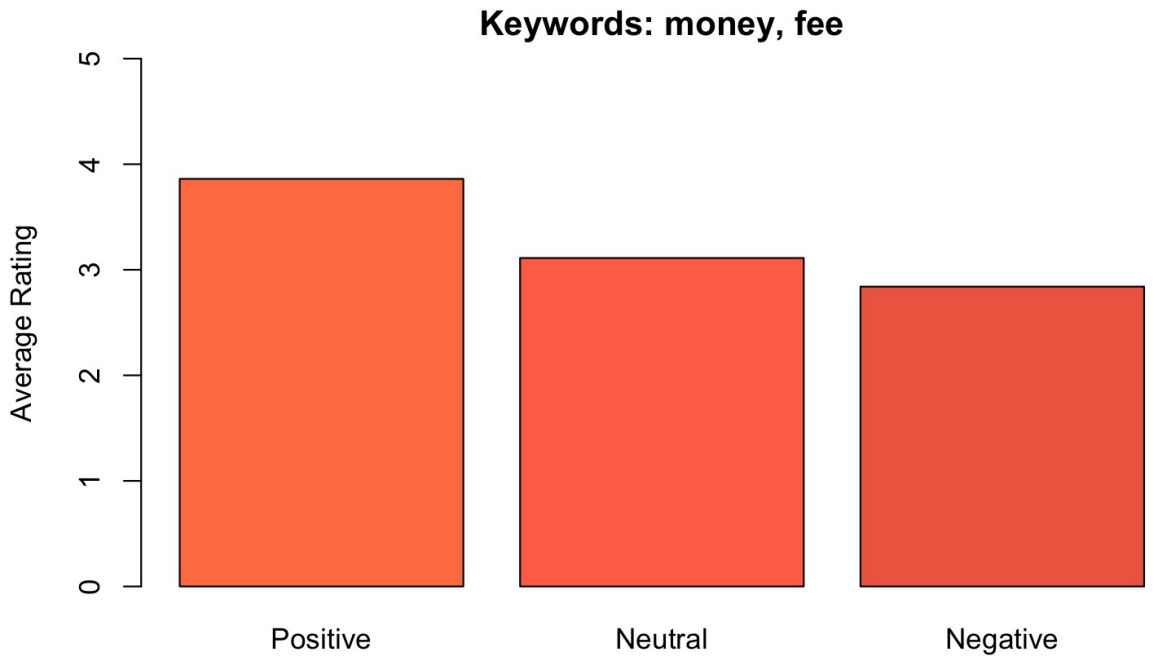




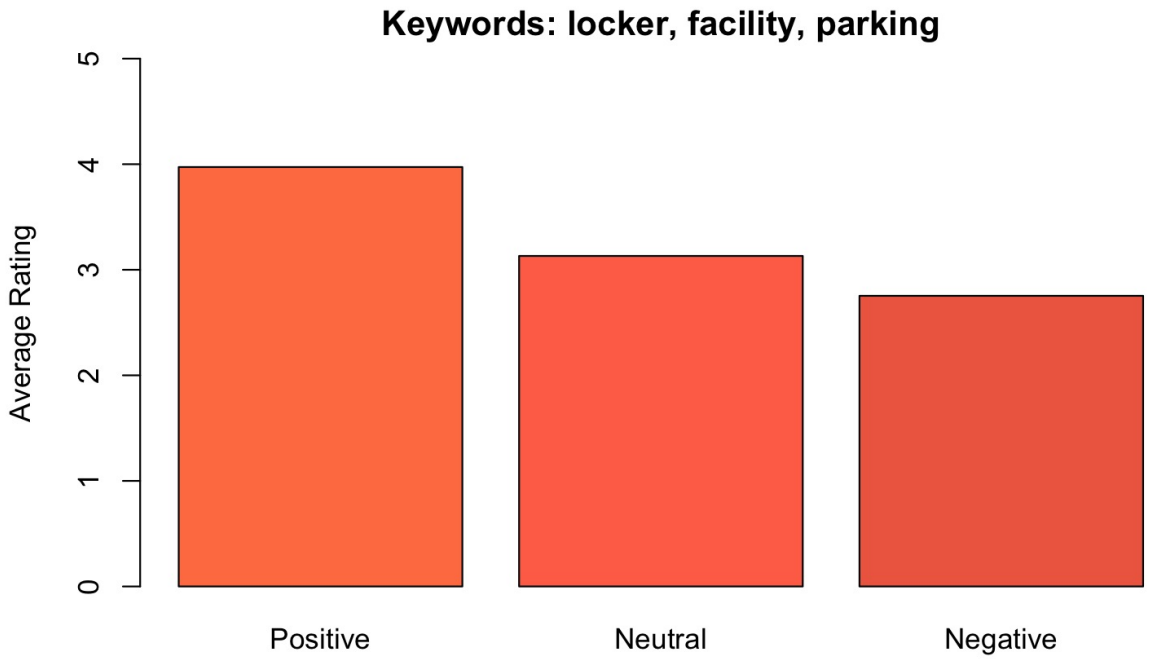


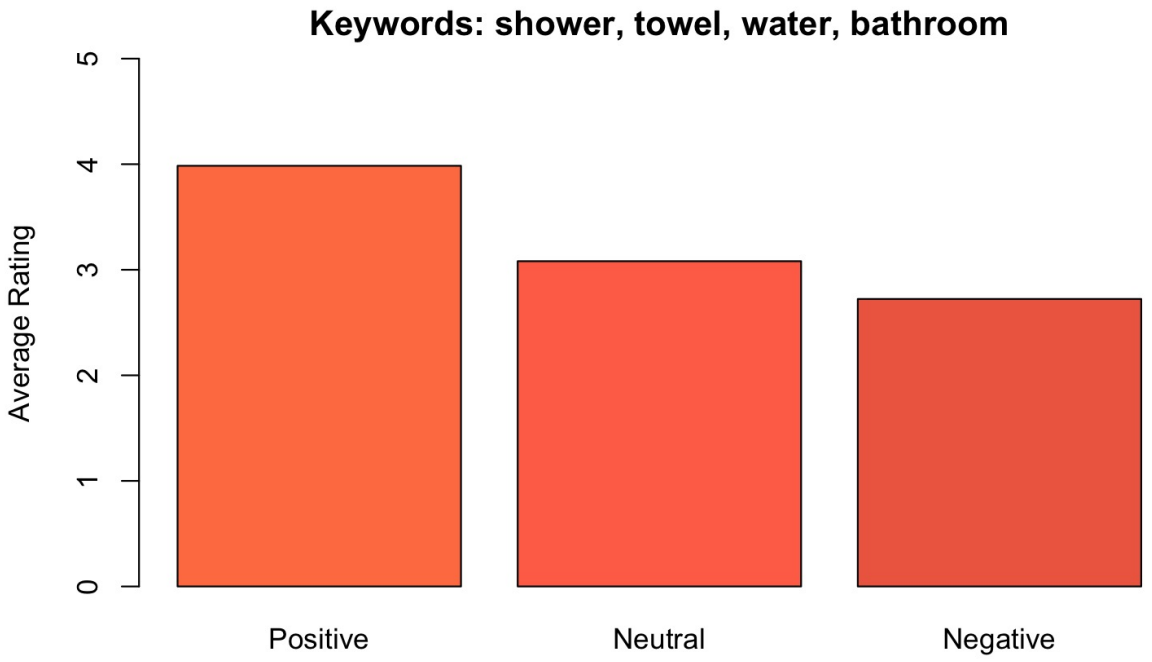
Non-attributes statistical analysis





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

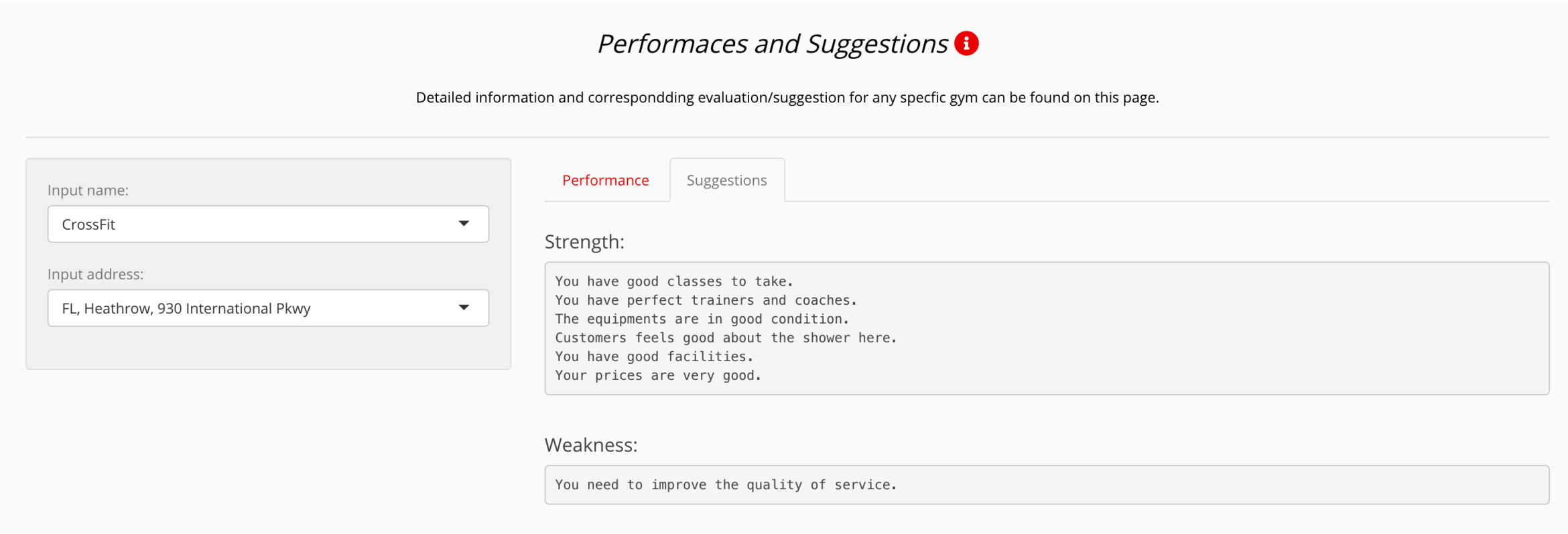
Keywords Group: locker, facility, parking		
Negative vs Neutral	Negative vs Positive	Neutral vs Positive
P-value < 2.2e-16	P-value < 2.2e-16	P-value < 2.2e-16

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:



R SHINY APP

Review Analysis on GYMs

FREQUENT WORDS MAPS EVALUATIONS

Performaces and Suggestions

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

Input name:

CrossFit

Input address:

FL, Heathrow, 930 International Pkwy

Performance Suggestions

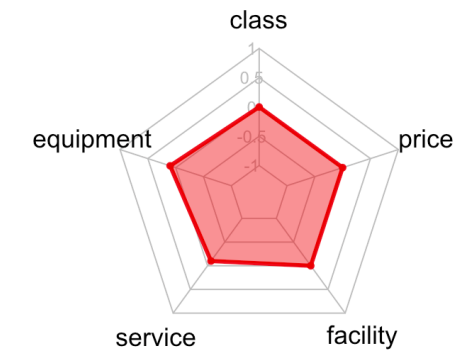
Yelp Rating:

Popularity:

Attributes:

```
parking : False
```

Nonattributes



THANK YOU