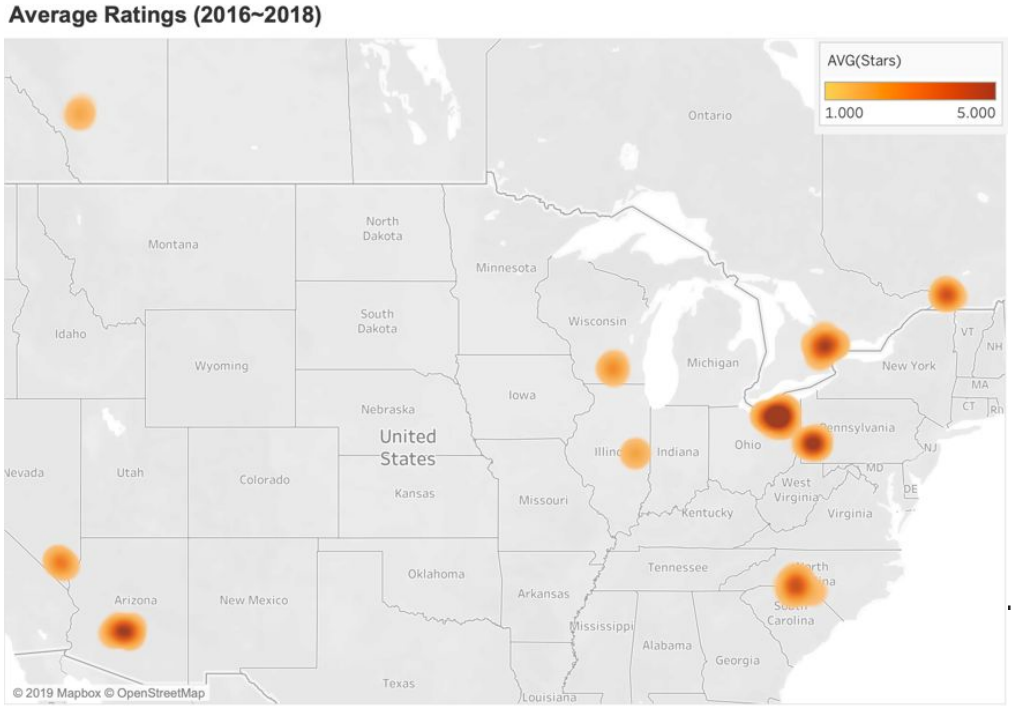
Yelp data analysis

Yuchen Zeng, Jingwen Yan, Chong Wei

Data selection

Our analysis focuses on cinemas that are still open in 2018. The selected dataset consists of 490 cinemas across 11 states in North America (3 in Canada, 8 in the US), 28747 reviews and 11356 tips from 24359 users.

Data Selection



Map based on Longitude (generated) and Latitude (generated). Color shows average of Stars (C Review.Csv). Details are shown for Nation, State and City. The data is filtered on Date and Is Open. The Date filter ranges from 2016/1/6 12:00:00 AM to 2018/11/14 4:26:44 PM. The Is Open filter ranges from 1 to 1.

- **The reviews are collected from users in 11 states:**

  Canada: Alberta, Ontario, Quebec

  USA: Arizona, Illinois, North Carolina, Nevada, Ohio, Pennsylvania, South Carolina, Wisconsin.

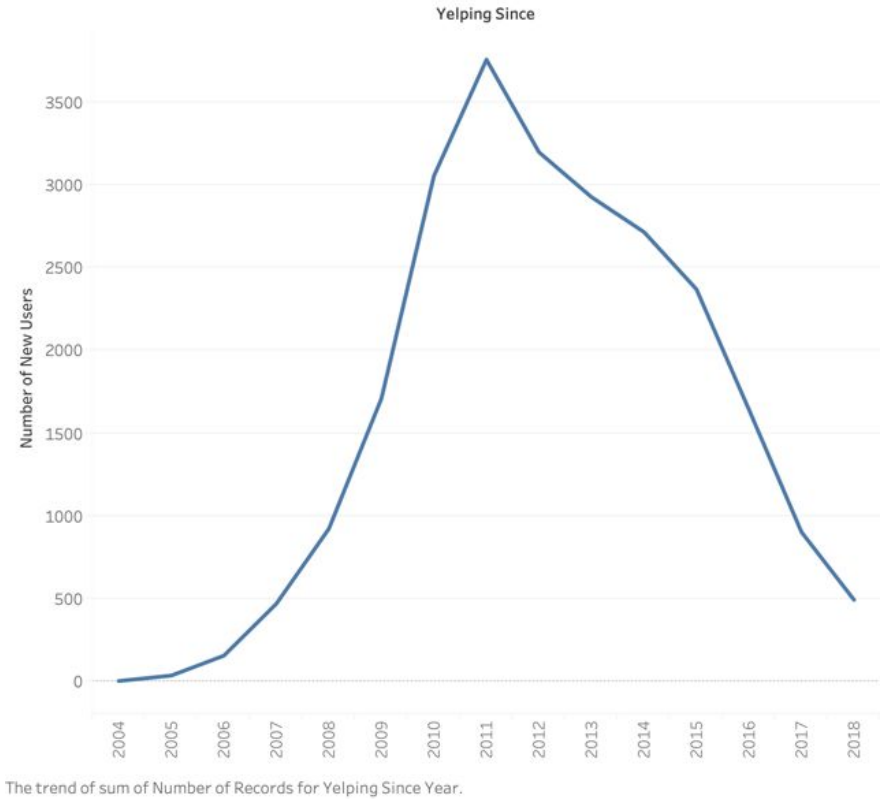
- **Average ratings in 2016 to 2018:**

  CA: 3.674

  USA: 3.401

Data Visualization

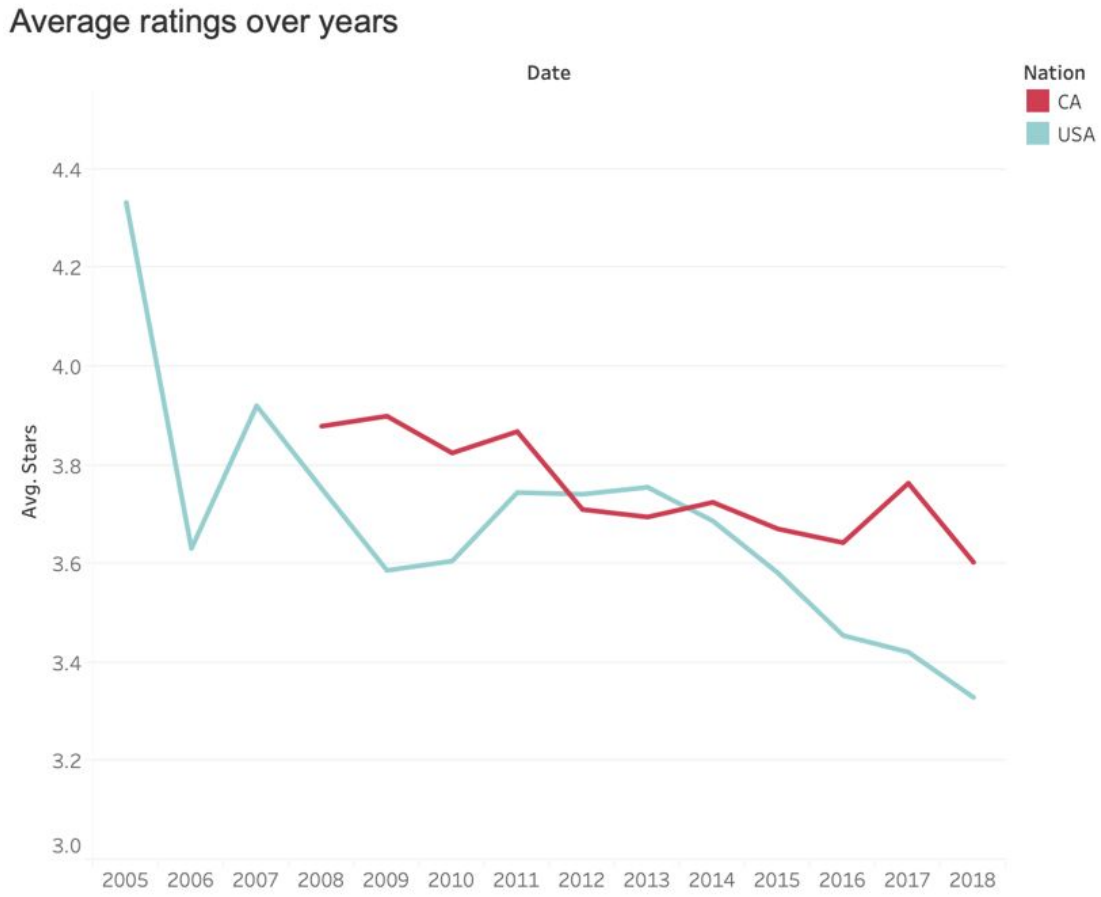
Data Visualization



The trend of sum of Number of Records for Yelping Since Year.

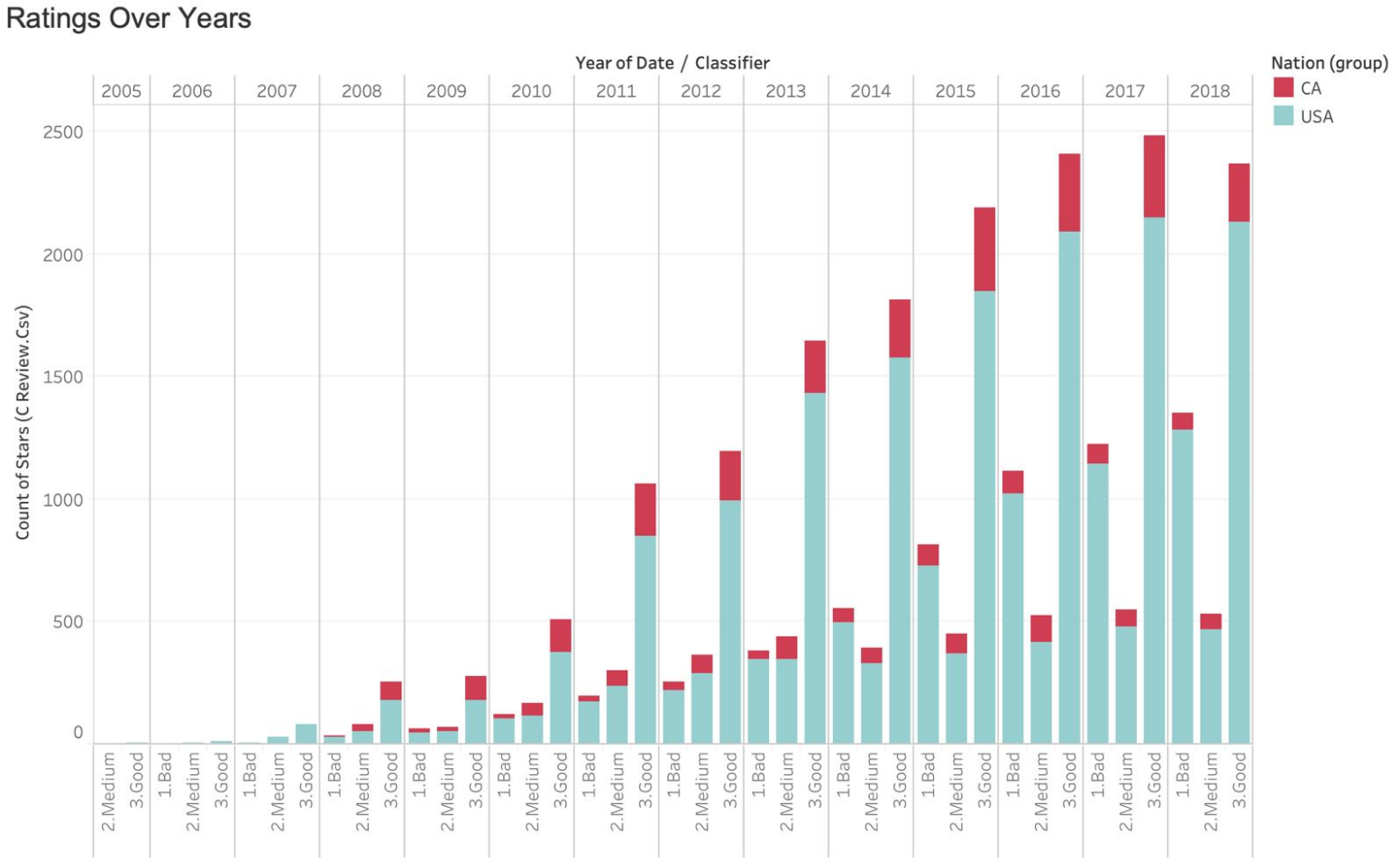
Yelp is a business directory service and crowd-sourced review forum founded in 2004. The number of new users in the selected dataset soared to its peak in 2011 and consistently decreased ever since.

Data Visualization



- Downward trends in both CA and USA
- Drop rapidly in USA from 2013

Data Visualization



Bad: 1-2 stars

Medium: 3 stars

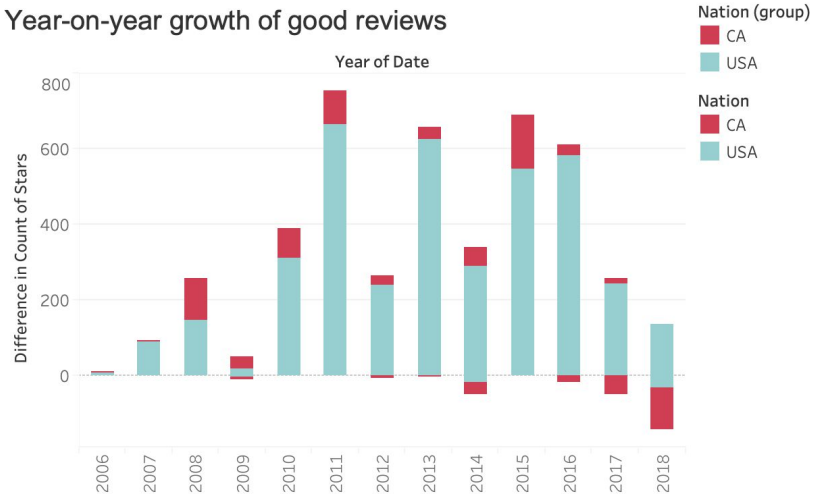
Good: 4-5 stars

Data Visualization

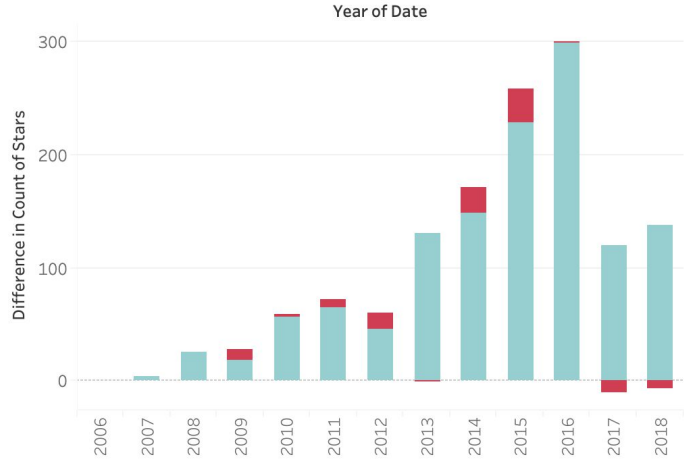
Average ratings over years

The trend of average of Stars (C Review.Csv) for Date Year. Color shows details about Nation. The data is filtered on Is Open, which ranges from 1 to 1.

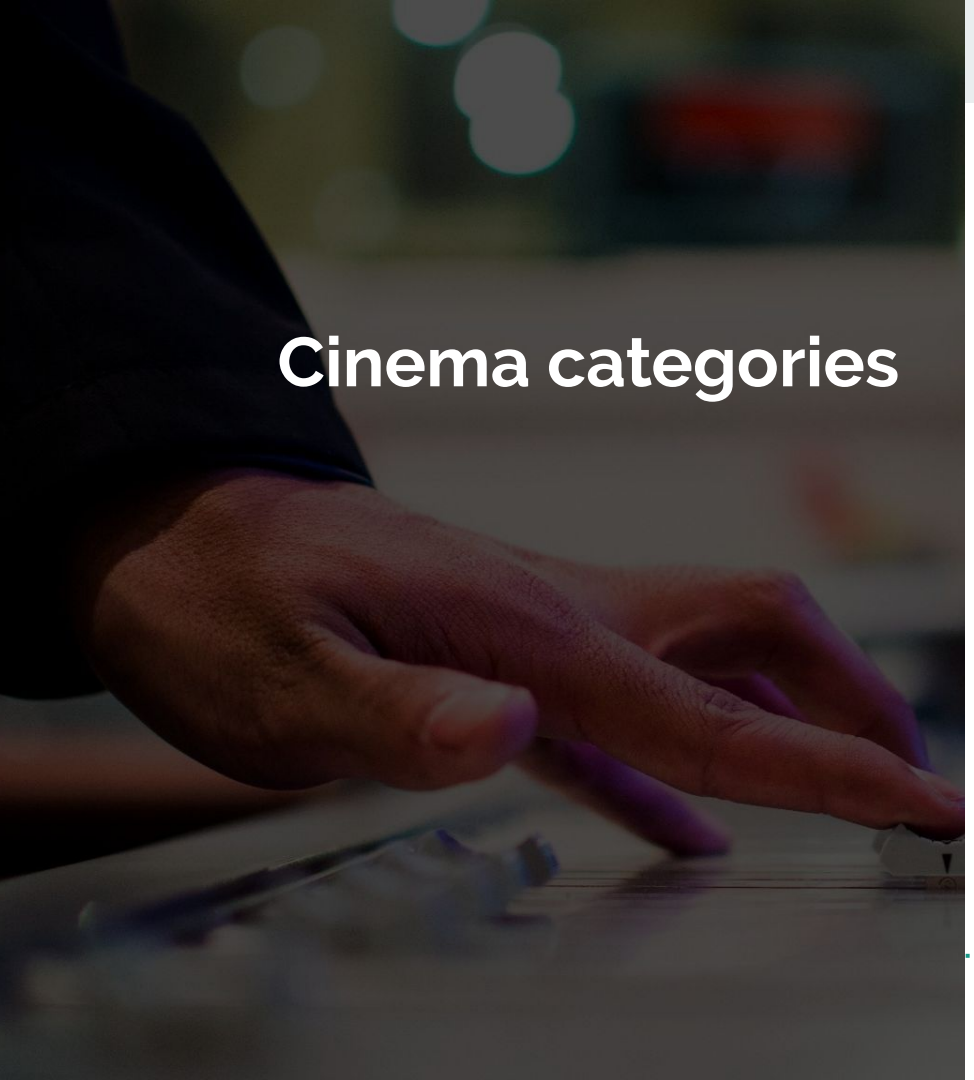
Year-on-year growth of good reviews



Year-on-year growth of bad reviews



Business Selection



Cinema categories

Theater

Drive-in Theater ❌
Dinner Theater ❌

Movie

Outdoor-movies ❌

Film

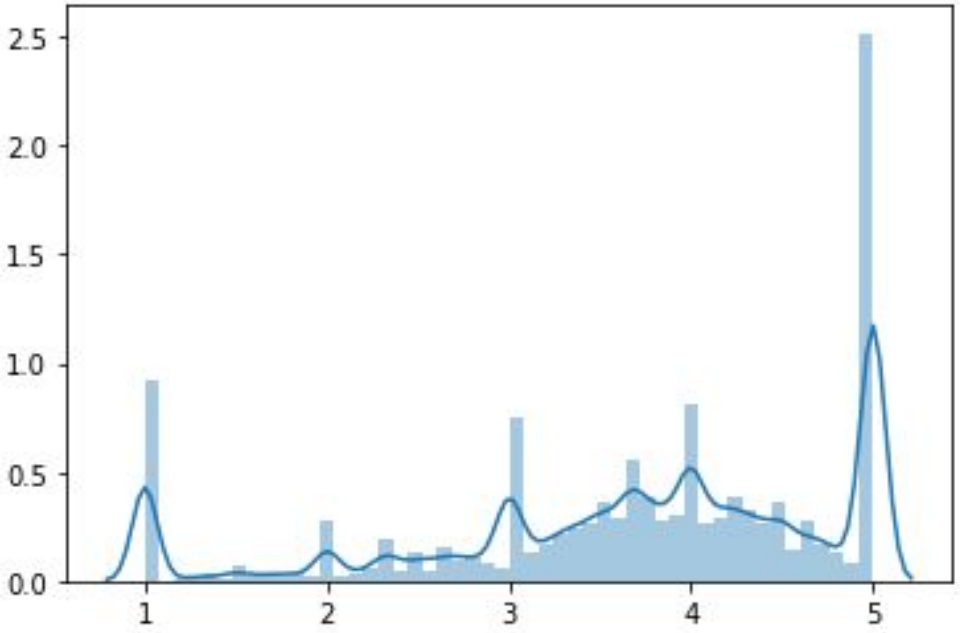
Video/Film Production ❌

Cinema

Cinema ✅

User Selection

User selection

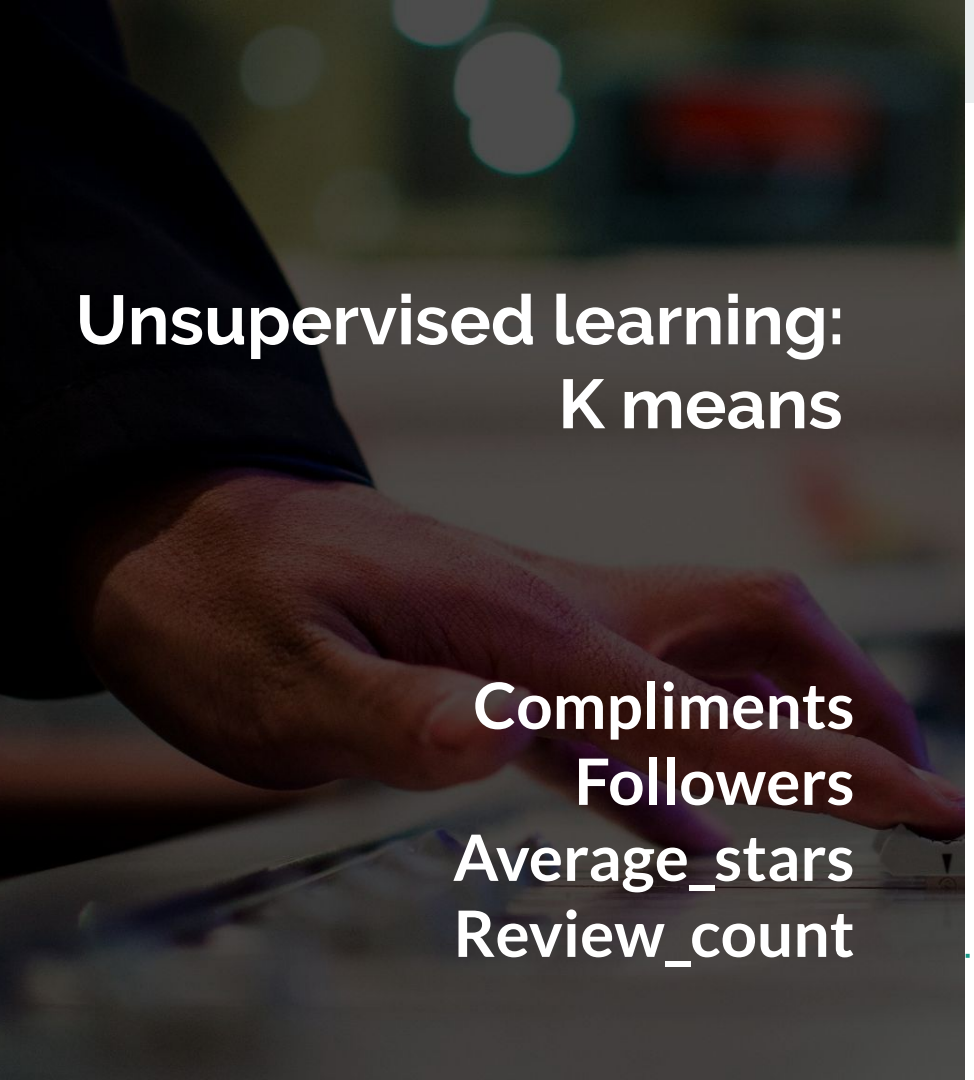


Average star distribution

Among 1,637,137 users, there exists a lot who have no friend, no fan, very few reviews and very extreme average star.

There are 1.2 million users who have no fan and 0.7 million users who have no friend.

We should get rid of these users.



Unsupervised learning: K means

Compliments
Followers
Average_stars
Review_count

Attributes

- Compliment_hot
- Compliment_more
- Compliment_cute
- .....
- Compliment_writer
- Compliment_photots
- Useful
- Funny
- Cool

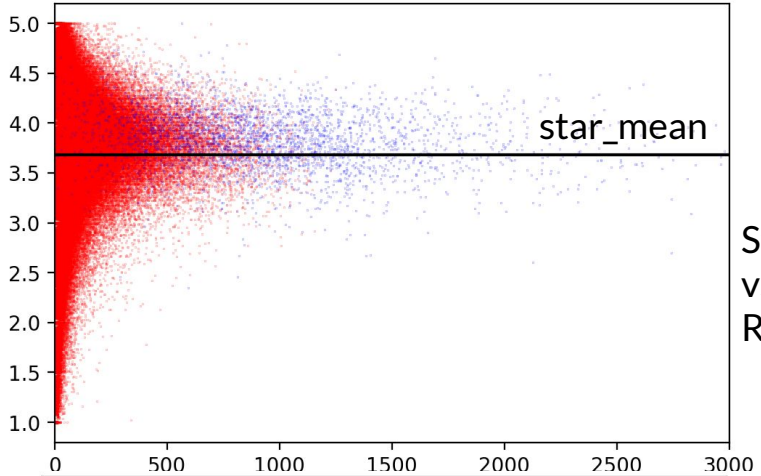
(Compliments)

- Fans
- Friends

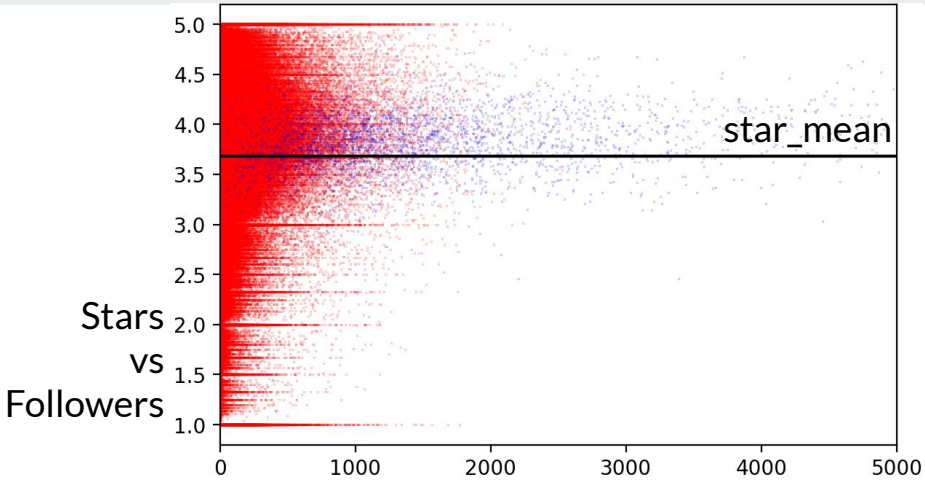
(Followers)

- Average_stars
- review_count

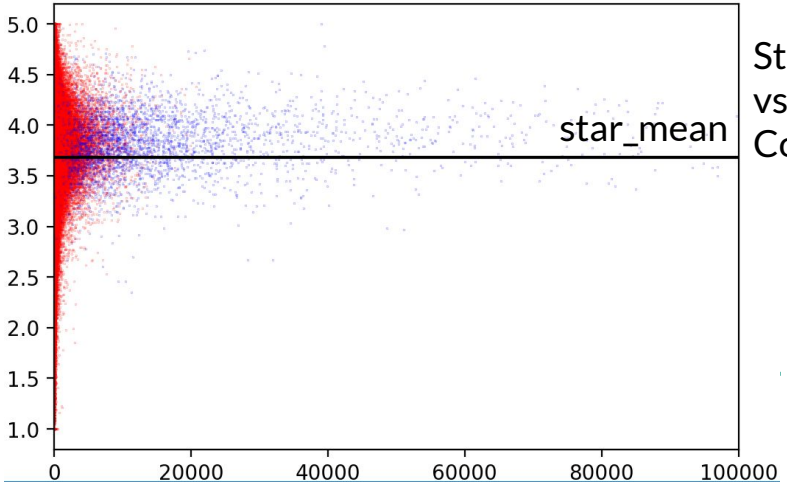
User selection n-clusters=2



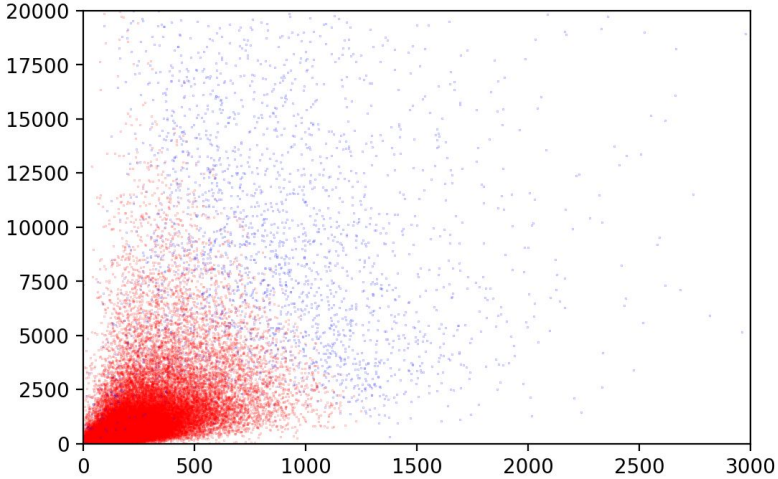
Stars
vs
Reviews



Stars
vs
Followers

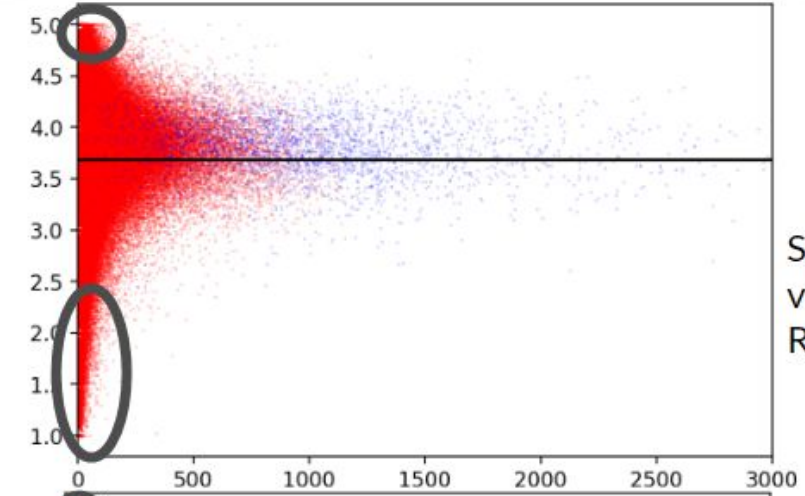


Stars
vs
Compliments

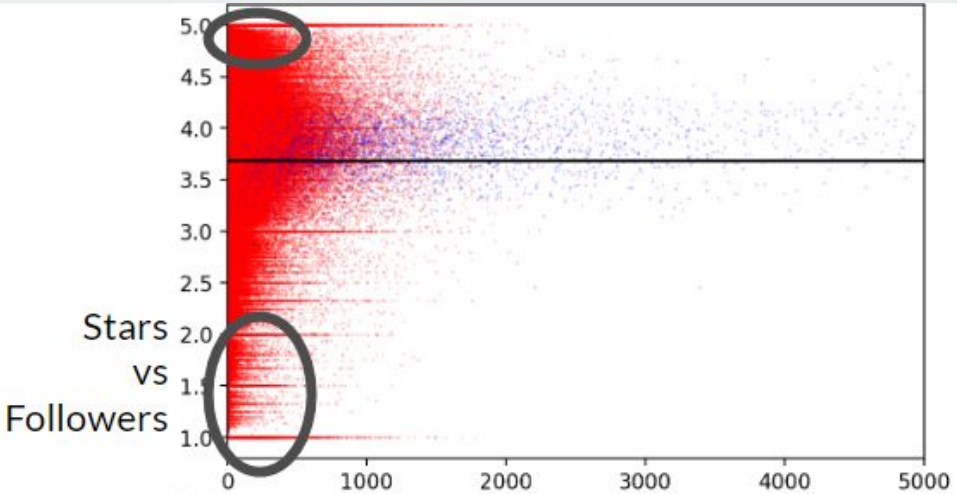


Compliments
vs
Reviews

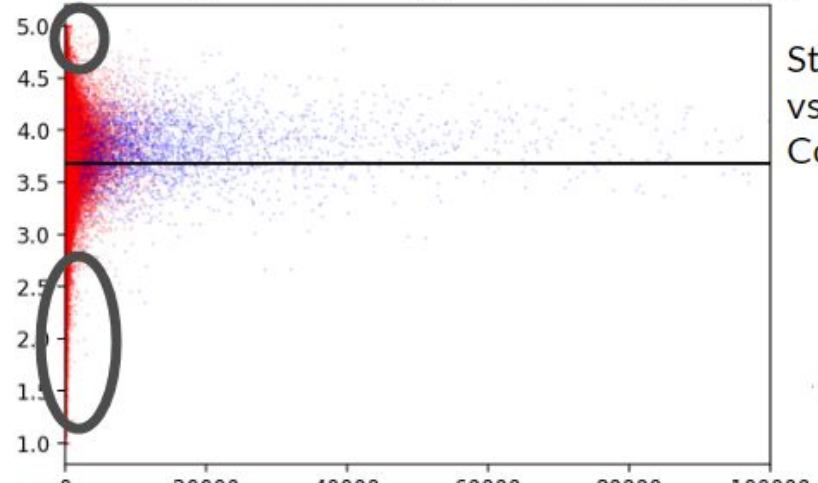
User selection n-clusters=2



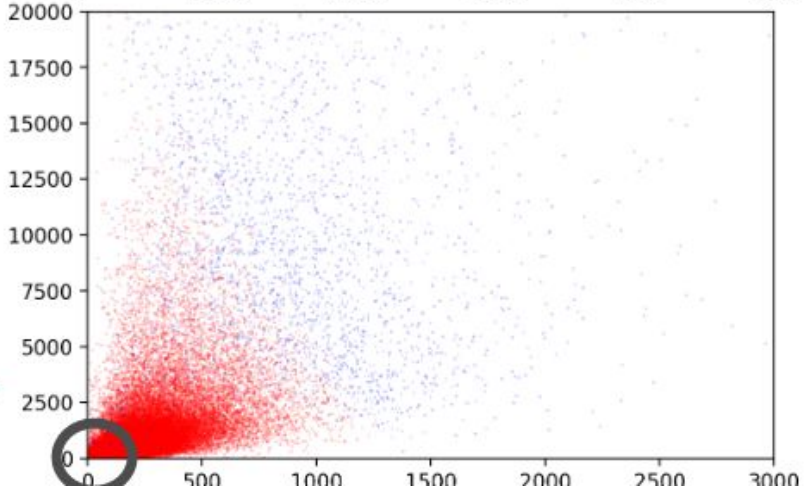
Stars vs Reviews



Stars vs Followers

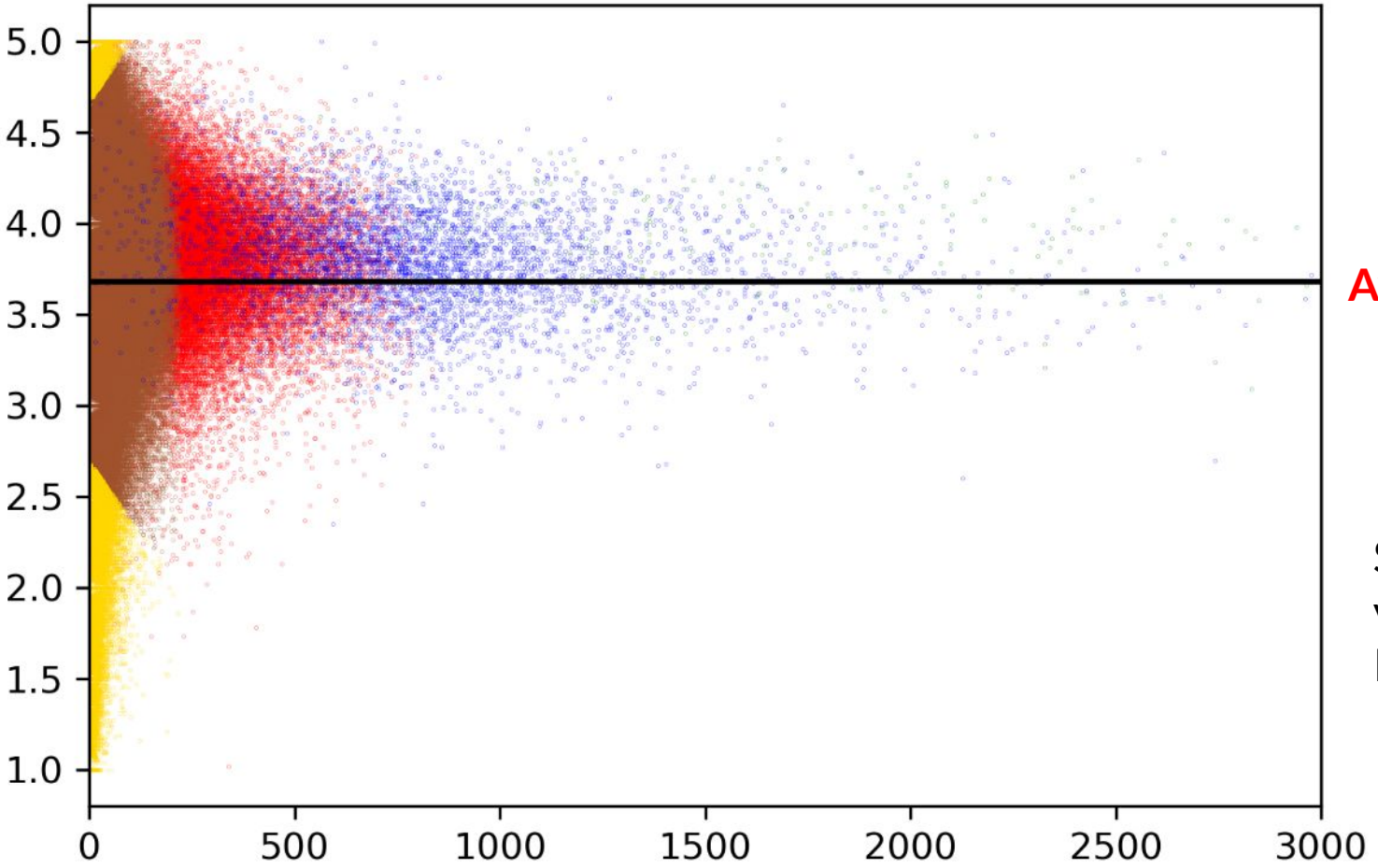


Stars vs Compliments



Compliments vs Reviews

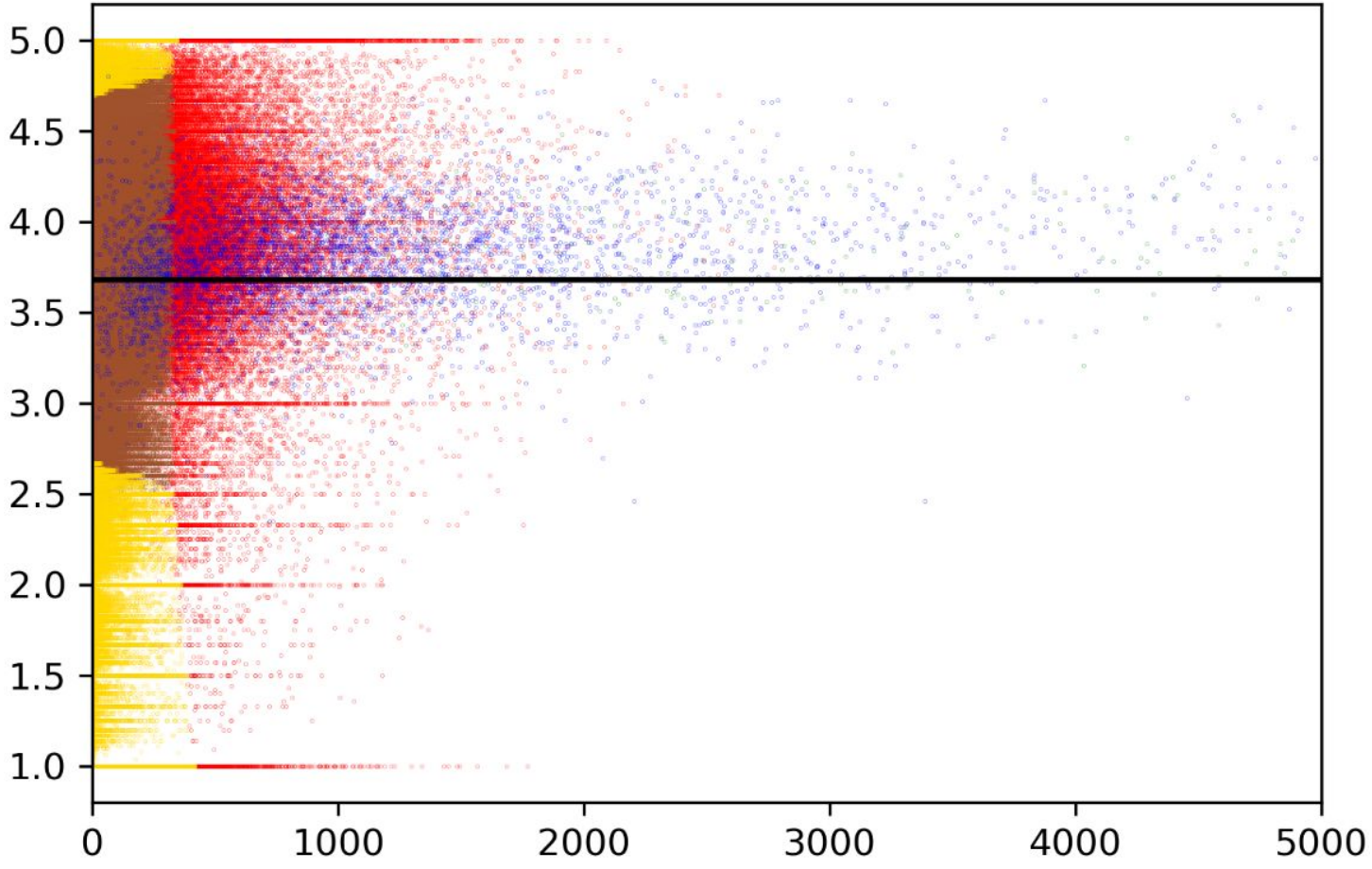
n-clusters=5 spam user: yellow points (0.67 million)



Average_stars

Stars
vs
Reviews

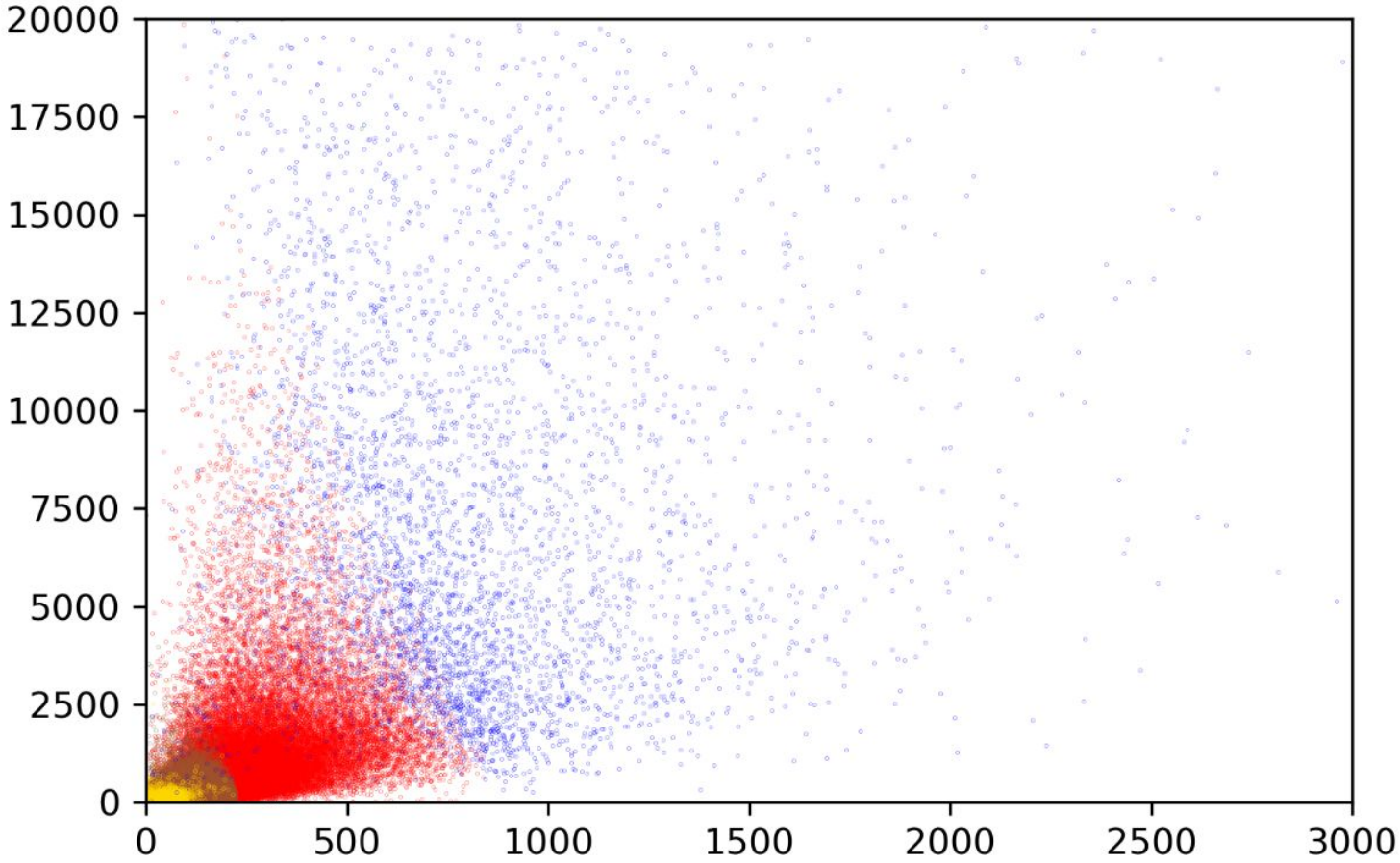
n-clusters=5 spam user: yellow points (0.67 million)



Average_stars

Stars
vs
Followers

n-clusters=5 spam user: yellow points (0.67 million)



compliments
vs
Reviews

Text Processing

Goal: extract useful words from a sentence.

Text Processing

1. Change all letters into lower case
2. Delete stop words
3. Split the sentences
4. Delete punctuation

```
clean("Very happy about the food!".lower().split(' '))
```

```
['happy', 'food!']
```

```
word_process('Nice!')
```

```
'nice'
```

Text Processing

1. Turn the tense into the present tense
2. Correct the spelling error

```
word_process('happz!')
```

'happy'

```
word_process('went')
```

'go'

```
word_process('gooood')
```

'good'

Words Counting

Word Counting

Count separately on the reviews with different ratings.

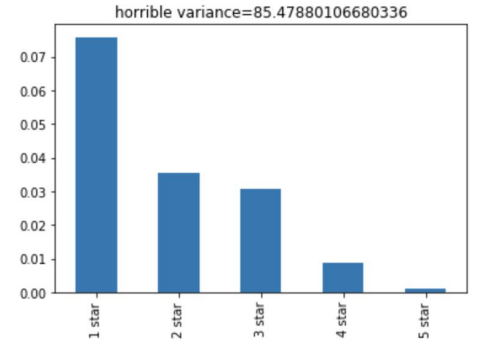
Advantages: get the distribution of different words.

Word Counting

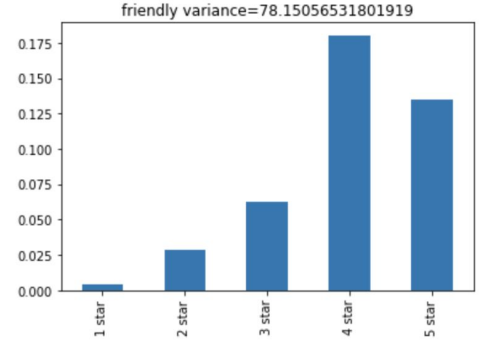
High variance indicate the adjective contains some kind of sentiment.

Compare the variance: adjective

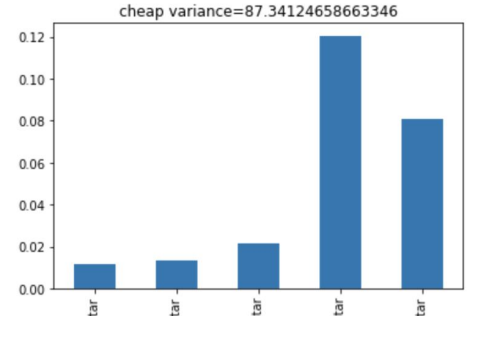
word_weight_bar('horrible')



word_weight_bar('friendly')



word_weight_bar('cheap')

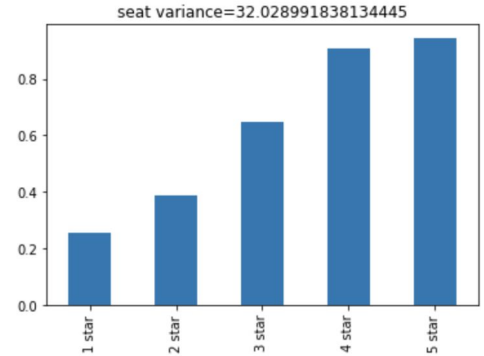


Word Counting

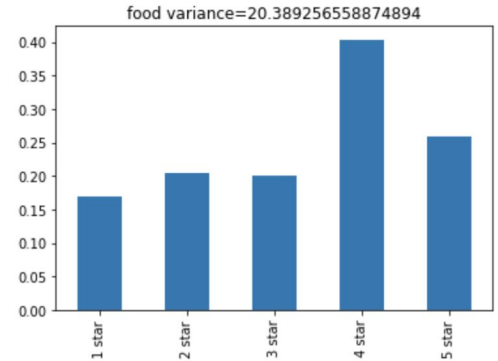
Low variance of a high-frequency noun implies it can be seen as an indicator for customers to write reviews and tips.

Compare the variance：noun

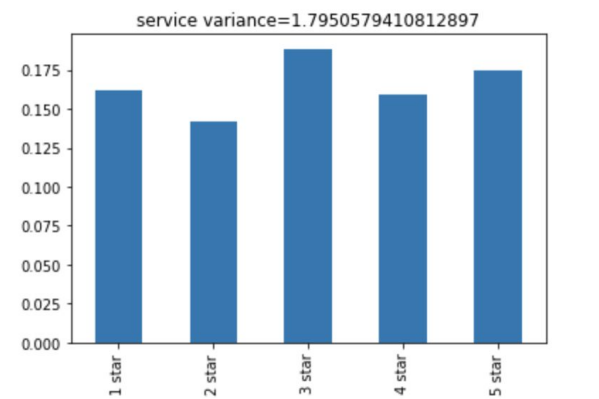
word_weight_bar('seat')



word_weight_bar('food')



word_weight_bar('service')



Word Counting

Compare the variance : noun

Note : When processing one certain theater, if those indicators show high variance, then we can pay attention to those indicators and give corresponding advice.

Word Counting

Put weights when counting the words

1. The amount of complement in the review or tip (funny, useful, cool)
2. The time of the review or tip

	Unnamed: 0	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	188	gQl8mh2FlksNO3_An2IRxQ	gElrEf488lkHwCoVOYd3vA	8WBHKj2davW6hhZWlUQ7DA	5.0	2	0	1	Great venue. Every seat in the house is awesom...	2016-02-11 23:45:29
1	410	1ue9DOGcYUPXU-NlJLLW7w	lZmfLRHJTCxxkq8wlKONlA	pJQSdbrtQQVstMUUIICSsA	3.0	0	0	0	Reclining chairs with arm rests that rise are ...	2011-01-09 02:59:52
2	666	RKEFZH9NBpFJhmqwHlR_4w	wR8GcbfgzhS4C1oAEqcn7w	f4mh1Y0rnvbJRfQ3jPkqzQ	4.0	0	0	0	There just isn't any better way to watch a mov...	2016-12-19 08:50:55
3	761	23w32IUzXwO-93uyAK8aWQ	WvM3GdNwgY5lK_8iXjv7Ag	jobP3ywRd3QNZ_GCoPG2DQ	1.0	0	0	0	Cheap tickets, awful staff, probably a good id...	2018-07-17 00:33:09

Word Counting

Put weights when counting the words

1. “should/would/not+adj”
2. “never+verb”

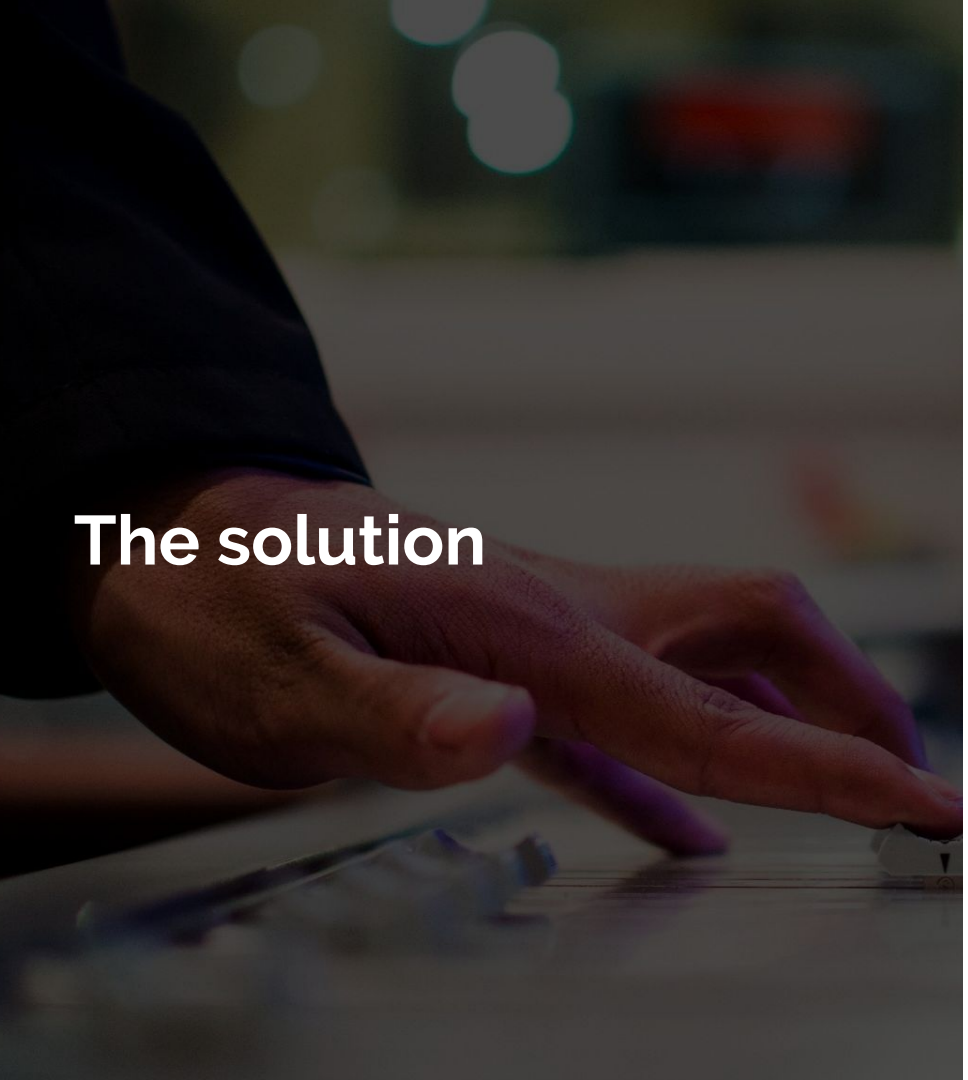
should be count as one word instead of two.

Should/would be more enjoyable
Not bad
Never come again

Word Counting

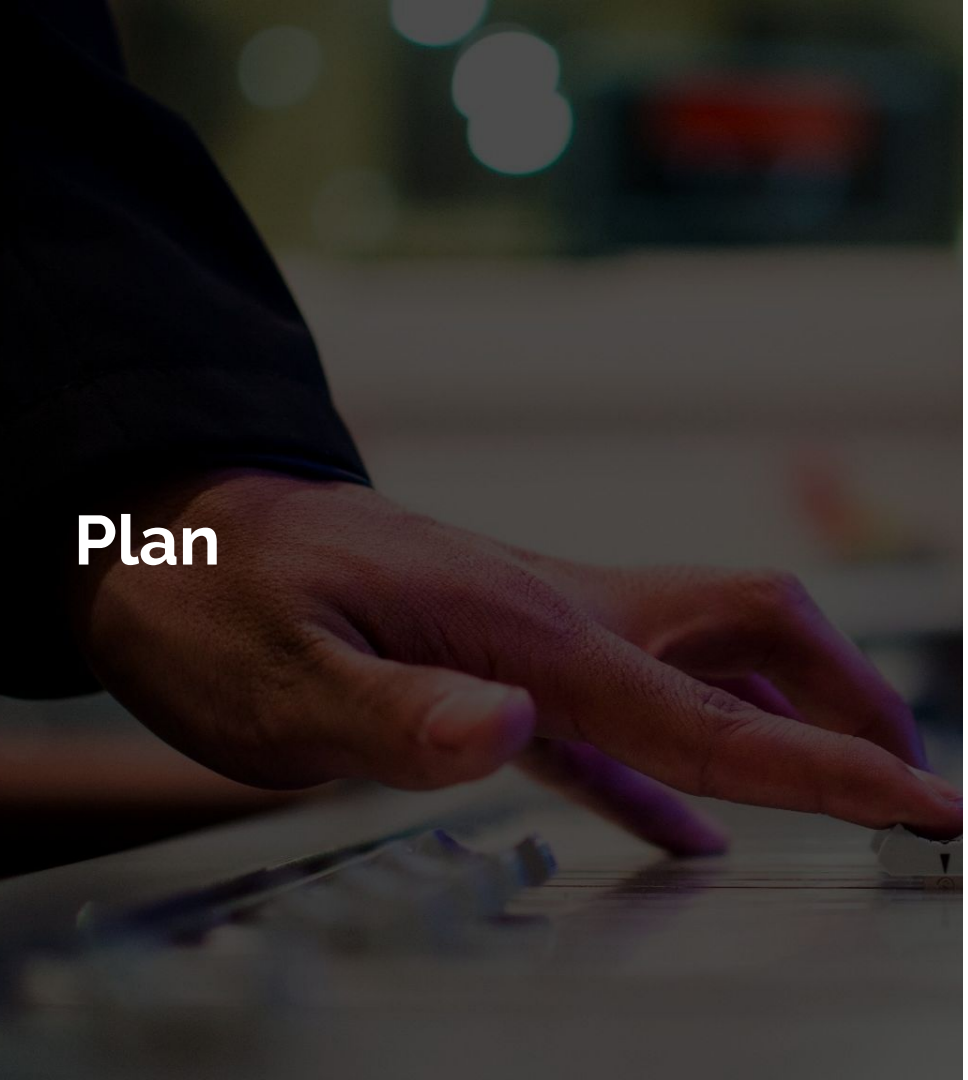
Scale the count

word	review	1 star	2 star	3 star	4 star	5 star	tip
movie	3.954783	0.495554	0.423895	1.161218	1.123802	0.912456	0.141361
seat	3.179458	0.255545	0.386398	0.646151	0.908490	0.944495	0.112067
theater	3.164761	0.301601	0.330253	0.741574	1.013628	0.697048	0.081249
time	1.883074	0.260189	0.261905	0.454396	0.640864	0.357353	0.066488
like	1.785714	0.160298	0.184206	0.423253	0.559443	0.352116	0.036226
ticket	1.754751	0.187773	0.173906	0.373384	0.475358	0.281377	0.067474
place	1.563950	0.170512	0.172299	0.265018	0.568014	0.546835	0.066664



The solution

Combine the adjective and the noun to rate the theatre in different aspects such as seat, food, screen… and give corresponding suggestions.



Plan

For a certain theater, we might use some classification tools to filter the features need to be improved.

Thank you for your attention!