Analysis on Yelp Data set

Group 7

Qiaoyu Wang, Chaoran Wang, Lu Li, Yuhan Meng

Outline

- Determination of interested topic
- Data exploration

- Internal Factors: Business & User
  1. Data cleaning
  2. Data analysis
- External Factors: Review
  1. Data cleaning
  2. Data analysis

- Future Work

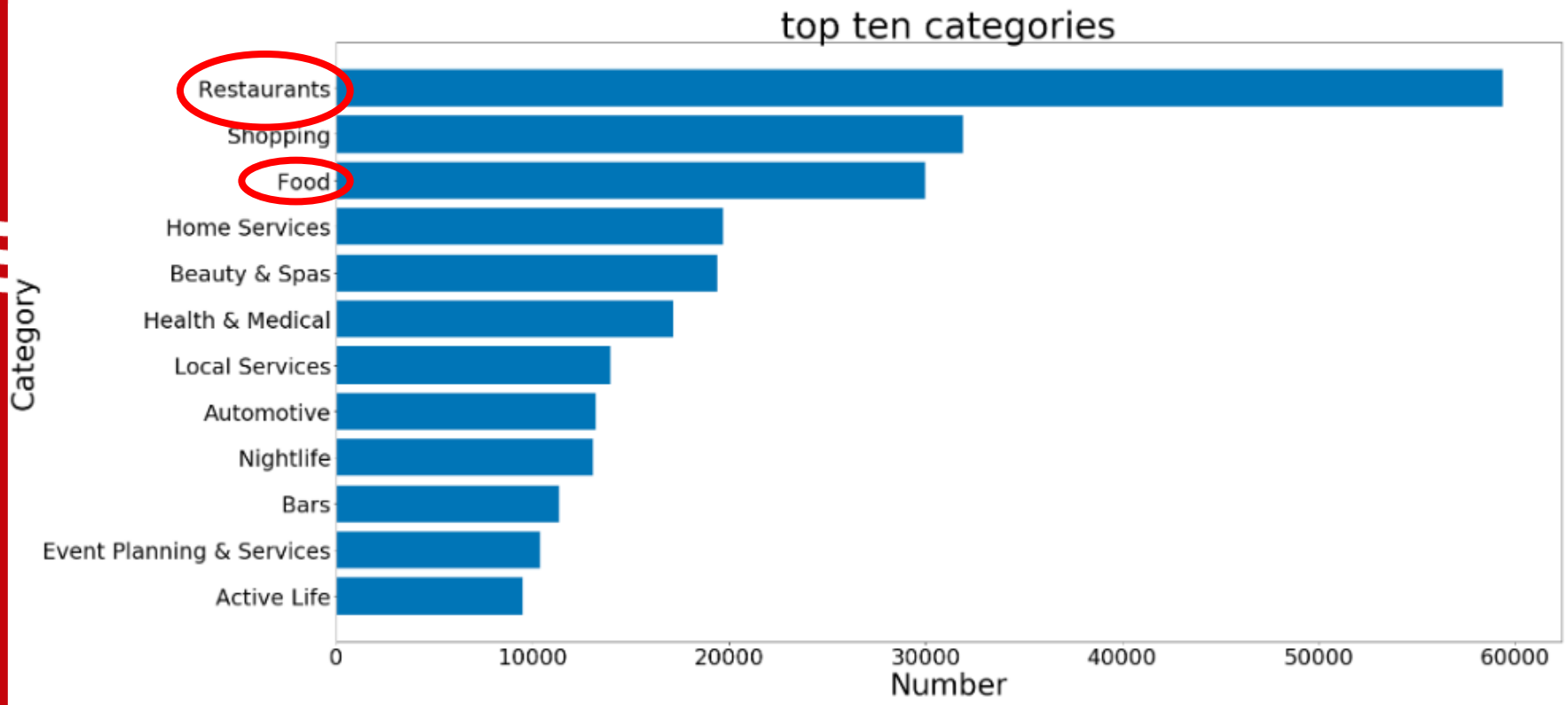
- Determination of interested topic

What we care about in business.json?

- 'business_id'
- 'name'
- 'address'
- 'city'
- 'state'
- 'postal_code'
- 'latitude'&'longitude'
- 'stars'
- 'review_count'
- 'is_open'
- 'attributes'
- 'categories'
- 'hours'

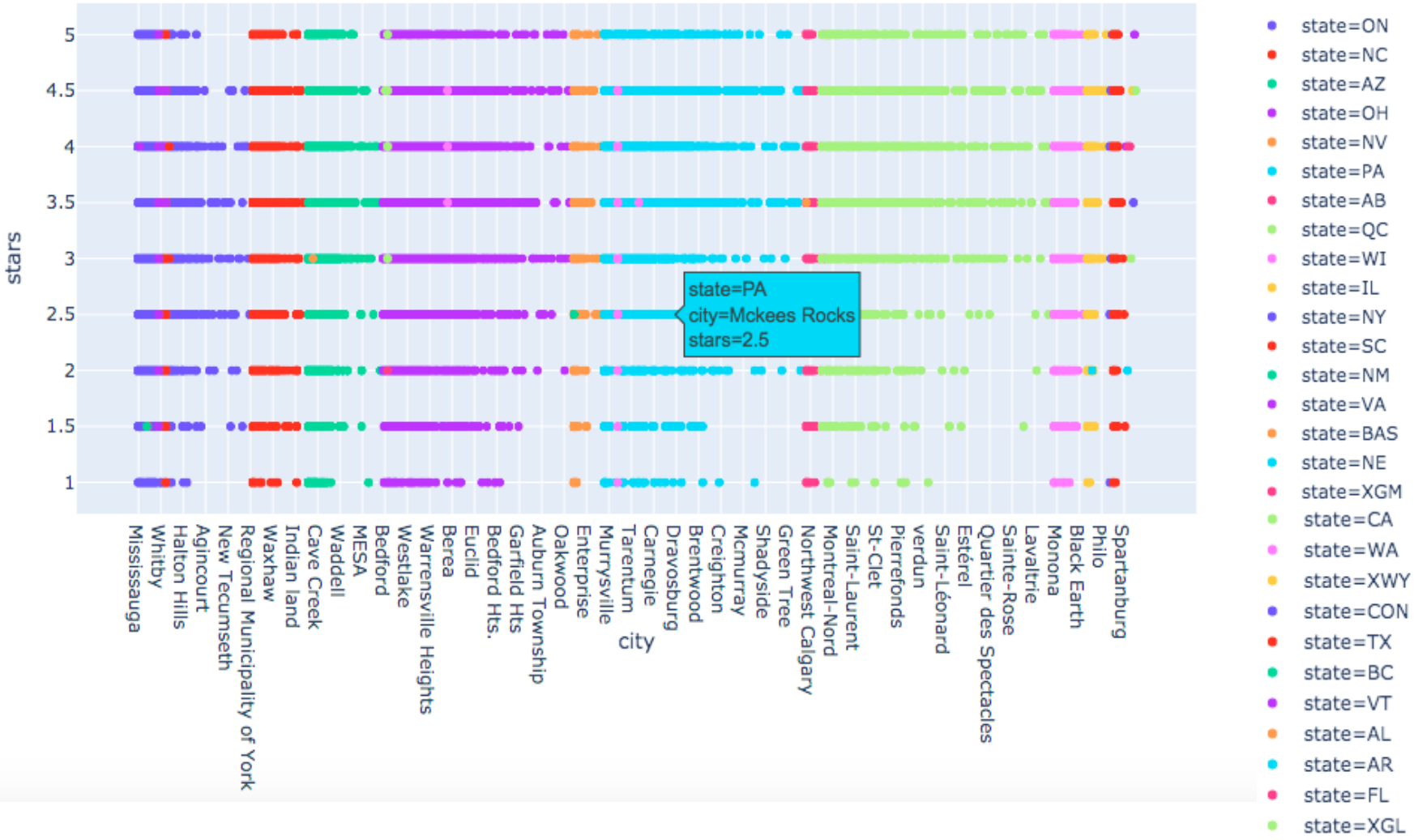
- Determination of interested topic

Choose topic:



- Determination of interested topic

Choose state:



Outline

- Determination of interested topic
- Data exploration
  - Internal Factors: Business & User
    1. Data cleaning
    2. Data analysis
  - External Factors: Review
    1. Data cleaning
    2. Data analysis
- Future Work

- Data exploration: Business & User

  - Review counts for each business and user >=3 ?

Data processing:

  - Remain 4524 business.

business_id	categories	stars
s-lwOqEEWb_peWh8DhhWUg	Food, Grocery	4.0
9sb2IZIYc3KnotJ2dM0dNQ	Convenience Stores, Automotive, Food, Gas Station	3.0
vgGijxITEbgF44fkG-lGJw	Beer Tours, Hotels & Travel, Bar Crawl, Tours	5.0
w43yHlJzoCEqUVNRezo_7A	Specialty Food, Fruits & Veggies, Grocery, Restaurant	3.5
L0DJ7-GUDMLIlIR-7vykvQ	Flowers & Gifts, Gift Shops, Shopping, Italian, Cooking Classes, Restaurants, Arts & Crafts	4.5

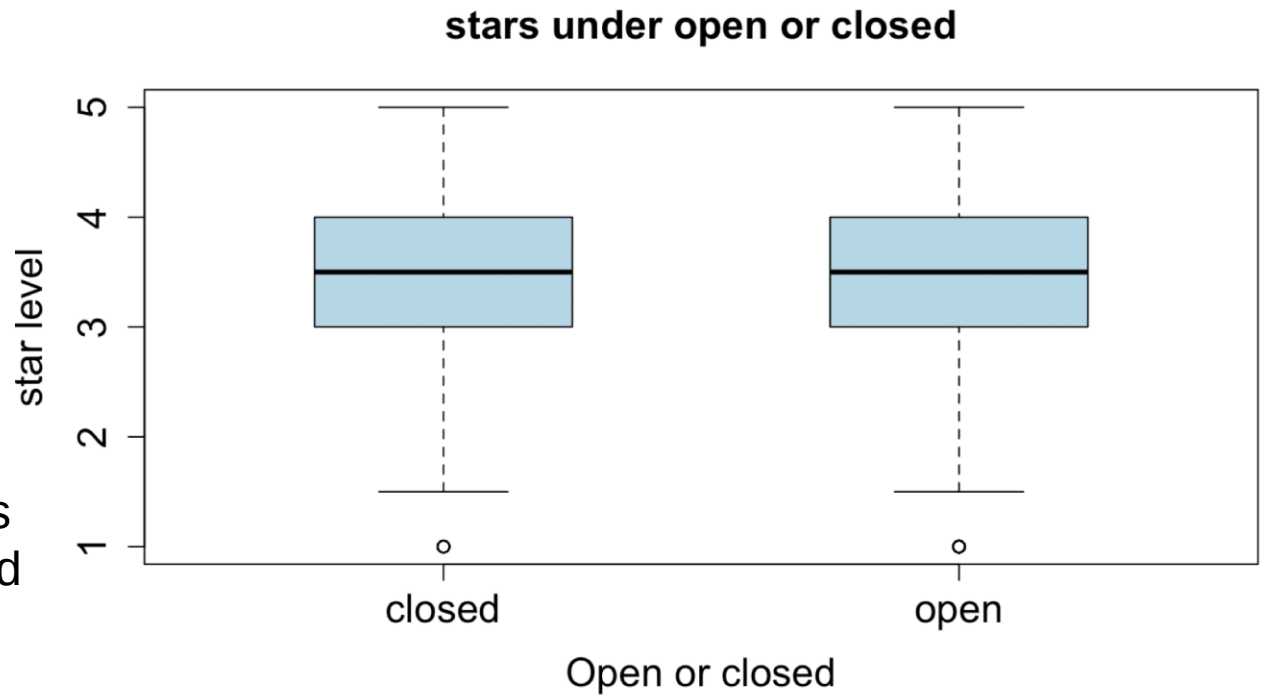
● Data exploration: Business & User

- Is_open

  ➢ Kendall correlation between is_open and stars:

  0.04897973

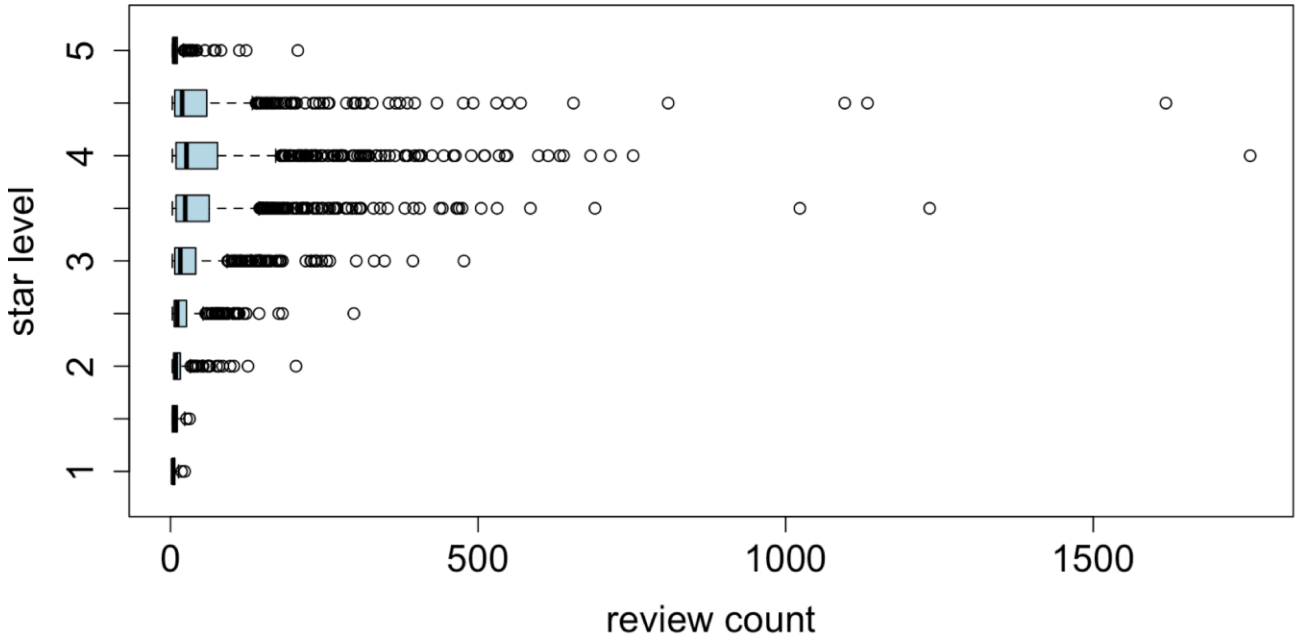
  ➢ Boxplot for the stars under open or closed status



● Data exploration: Business & User

- Review_count

  - ➢ Significantly correlated with stars by doing ordinal regression and anova (p-value < 0.05)
  - ➢ boxplot for the review count under different star levels

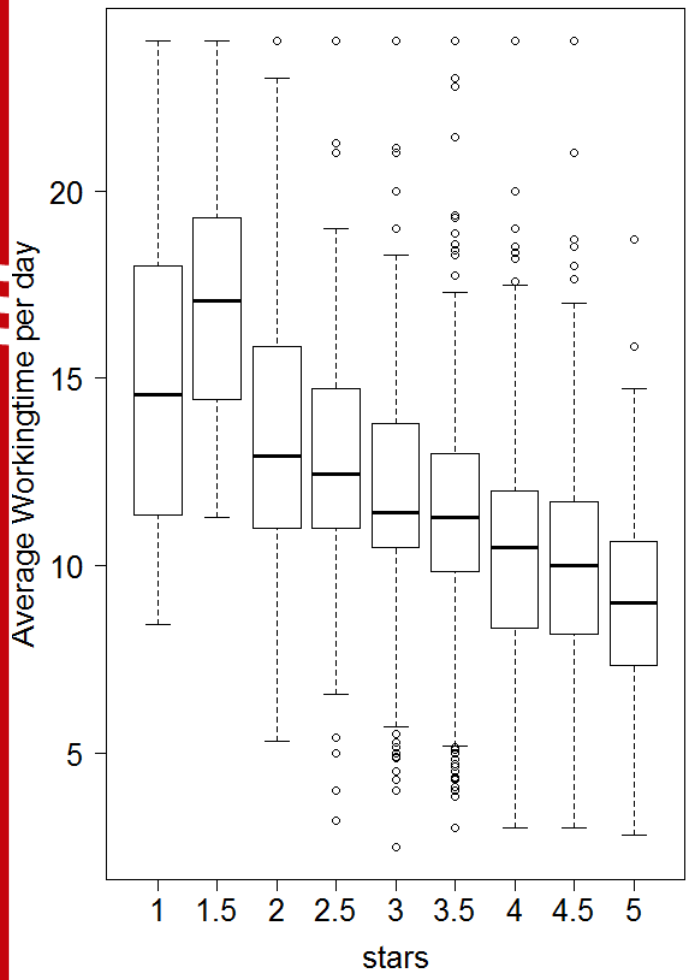


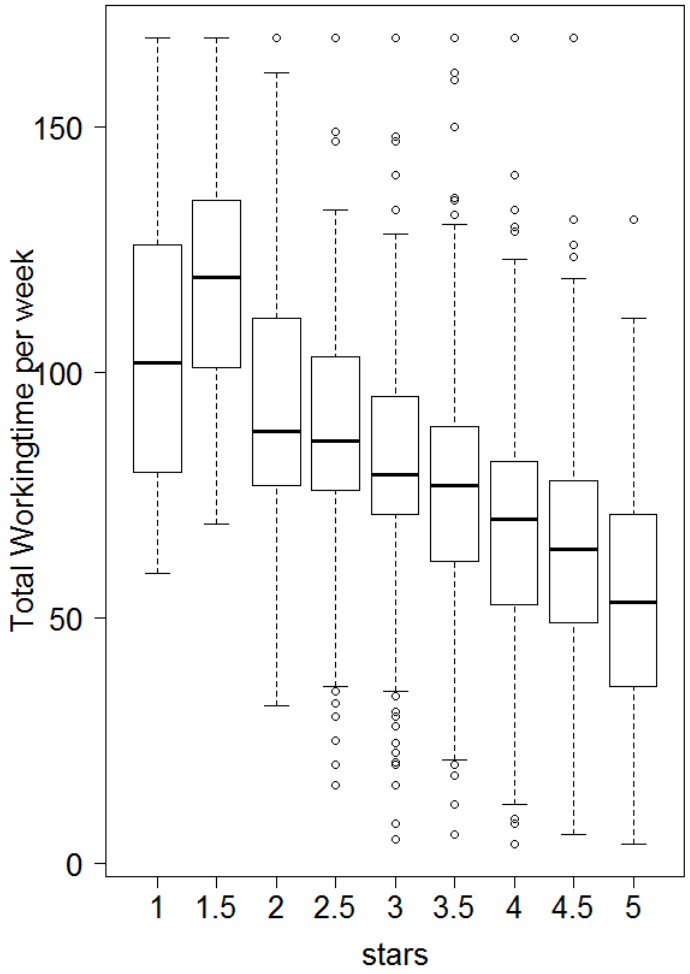
Data exploration: Business & User

```
'Monday': '0:0-0:0'
'Tuesday': '16:0-21:0'
'Wednesday': '16:0-21:0'
'Thursday': '16:0-22:0'
'Friday': '16:0-0:0'
'Saturday': '14:0-0:0'
```

→ Total working time per week

Average Working time per day



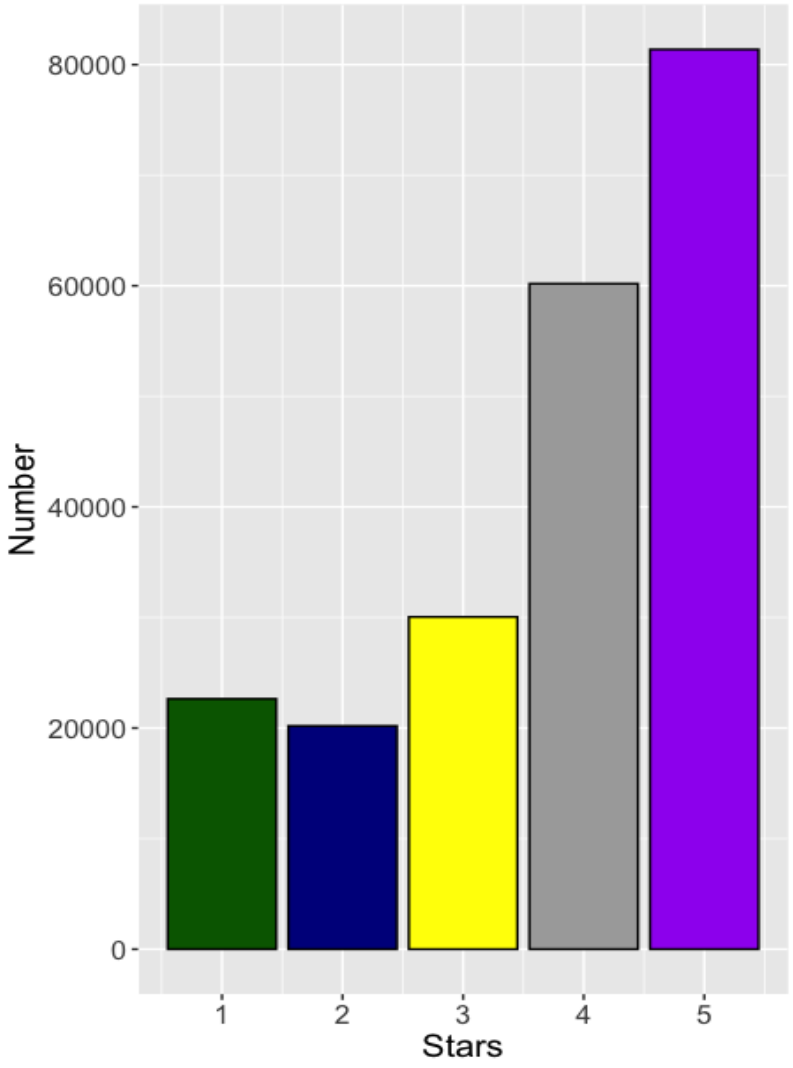


Check p-value of Ordered logistic regression --> Significant

Outline

- Determination of interested topic
- Data exploration
  - Internal Factors: Business & User
    1. Data cleaning
    2. Data analysis
  - External Factors: Review
    1. Data cleaning
    2. Data analysis
- Future Work

- Data exploration: Review

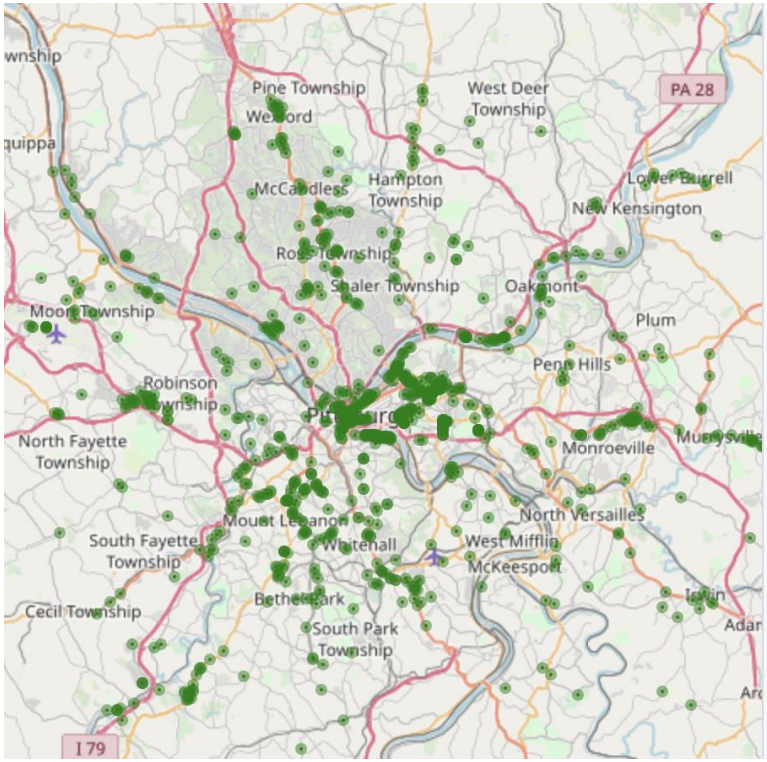


What's the distribution of stars?

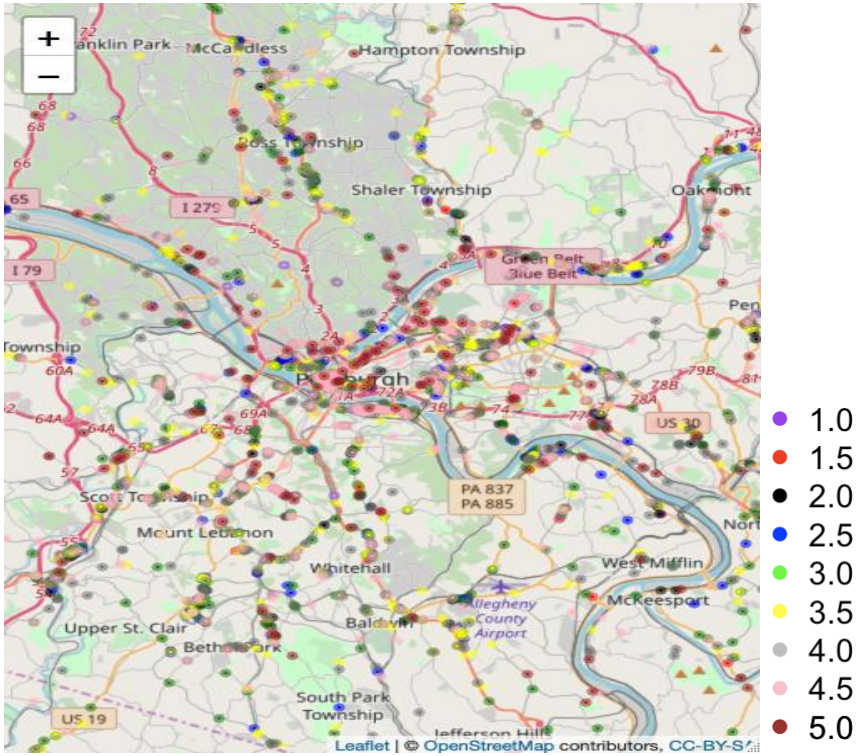
```
> table(PAreview$stars)

    1     2     3     4     5
22640 20190 30048 60189 81376
```

- Data exploration: Review

The distribution of 3.5 star



The distribution of various stars



- Data exploration: Review

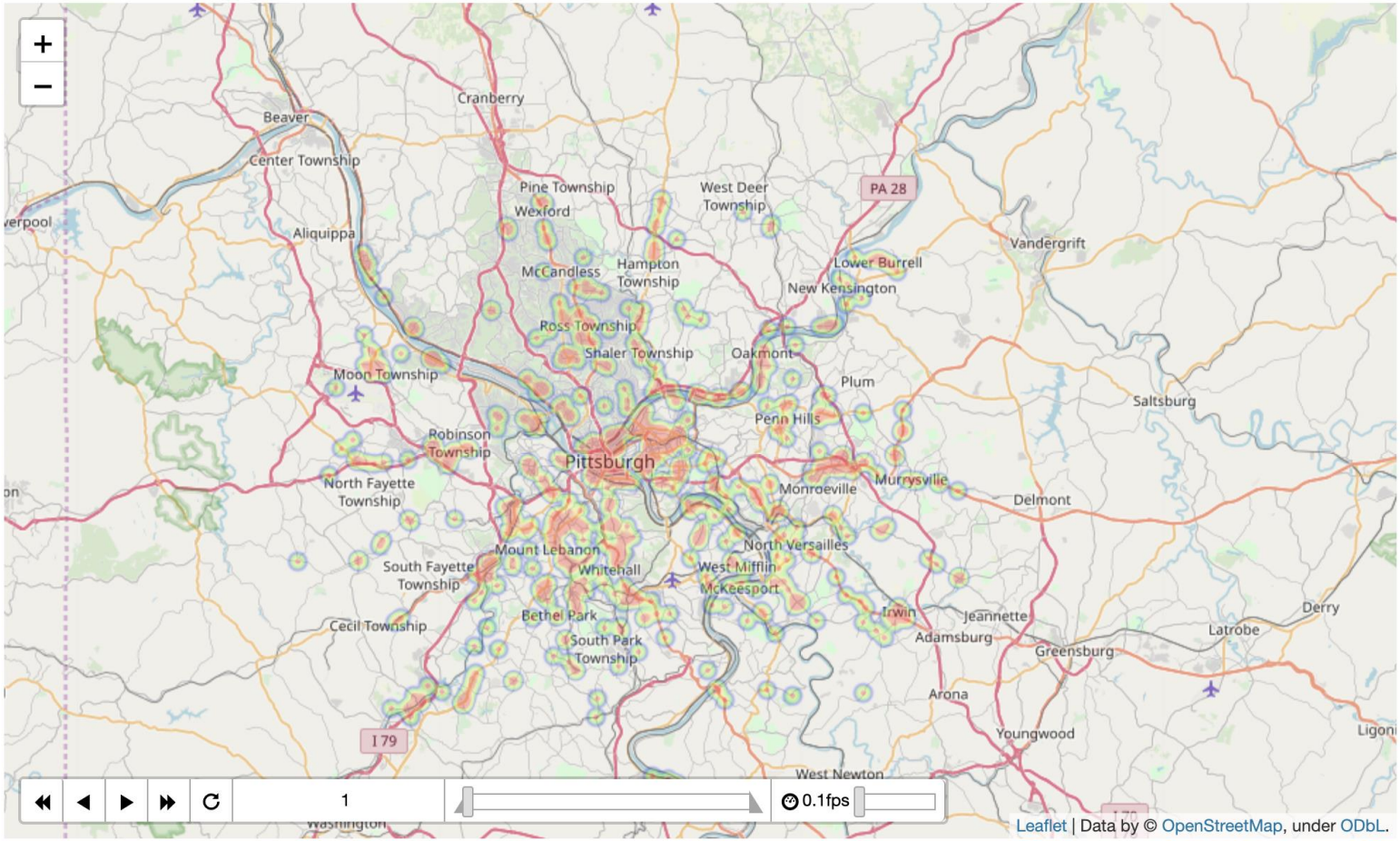
What's the classification under main category 'food' and 'restaurant' ?

Chinese, Japanese, Italian, Steak, Mexican, Bars, Pizza, American, Bakeries

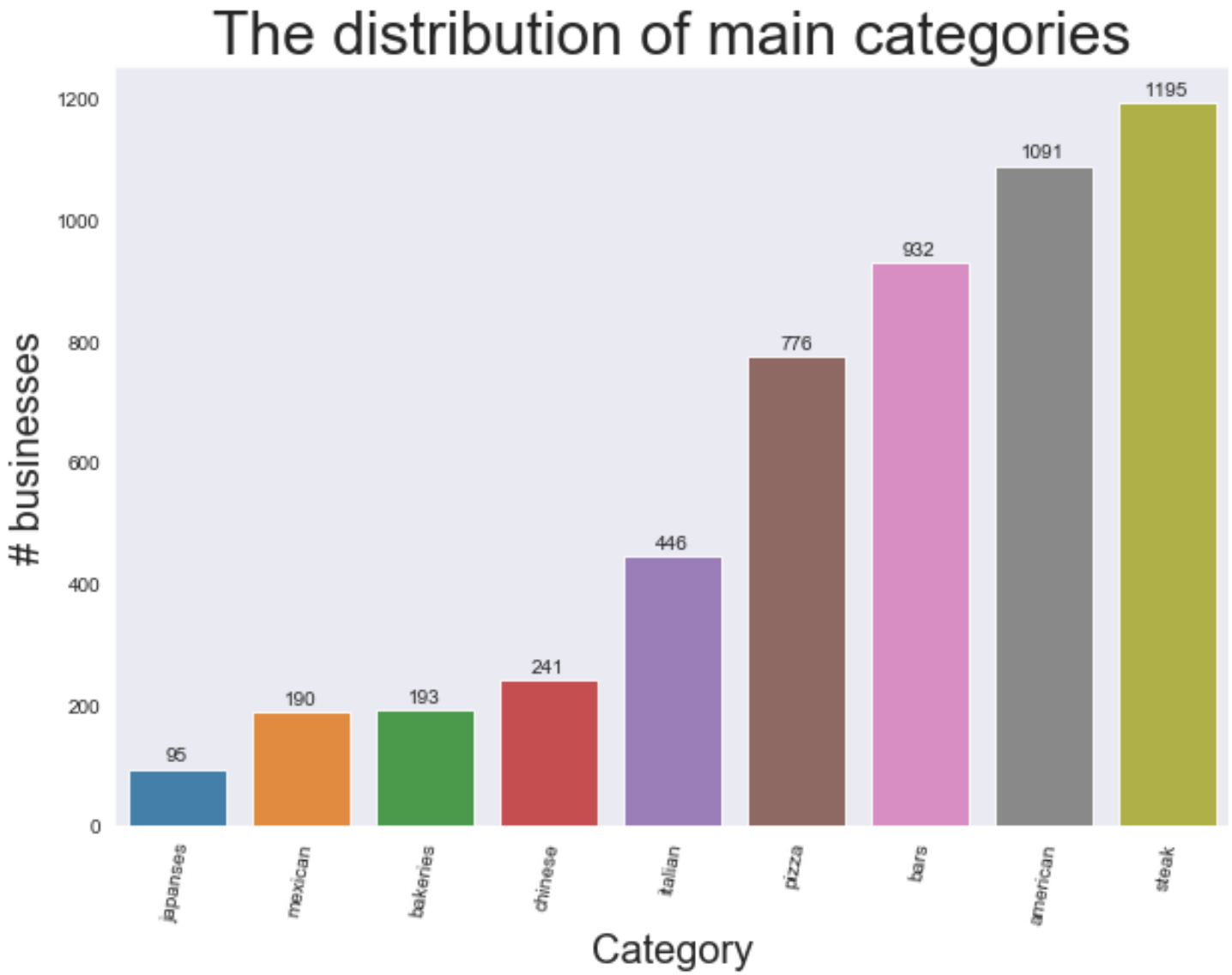
- Data exploration: Review

Pennsylvania Review heatmap by categories

Pennsylvania Review heatmap Animation by categories



- Data exploration: Review



- Data exploration: Review

Example `'stars': '2.0',`

```
Very busy at lunch. Almost like they
weren't ready for a Friday lunch.
Slow service, mediocre food.
```

Step1:
Change into lower case
and Expand contractions

```
very busy at lunch. almost like they
were not ready for a friday lunch.
slow service, mediocre food.
```

Step2:
Break paragraphs to sentences

```
very busy at lunch
almost like they were not ready for a
friday lunch
slow service
mediocre food
```

- Data exploration: Review

Example

```
very busy at lunch
almost like they were not ready for a
friday lunch
slow service
mediocre food
```

Step3:
Remove stop words

```
busy lunch
almost like not ready friday lunch
slow service
mediocre food
```

Step4:
Add NOT mark after negative words

```
busy lunch
almost like NOTready NOTfriday NOTlunch
slow service
mediocre food
```

- Data exploration: Review

Example

```
busy lunch
almost like NOTready NOTfriday
NOTlunch
slow service
mediocre food
```

Step5:
Remove punctuations and
Normalize the words

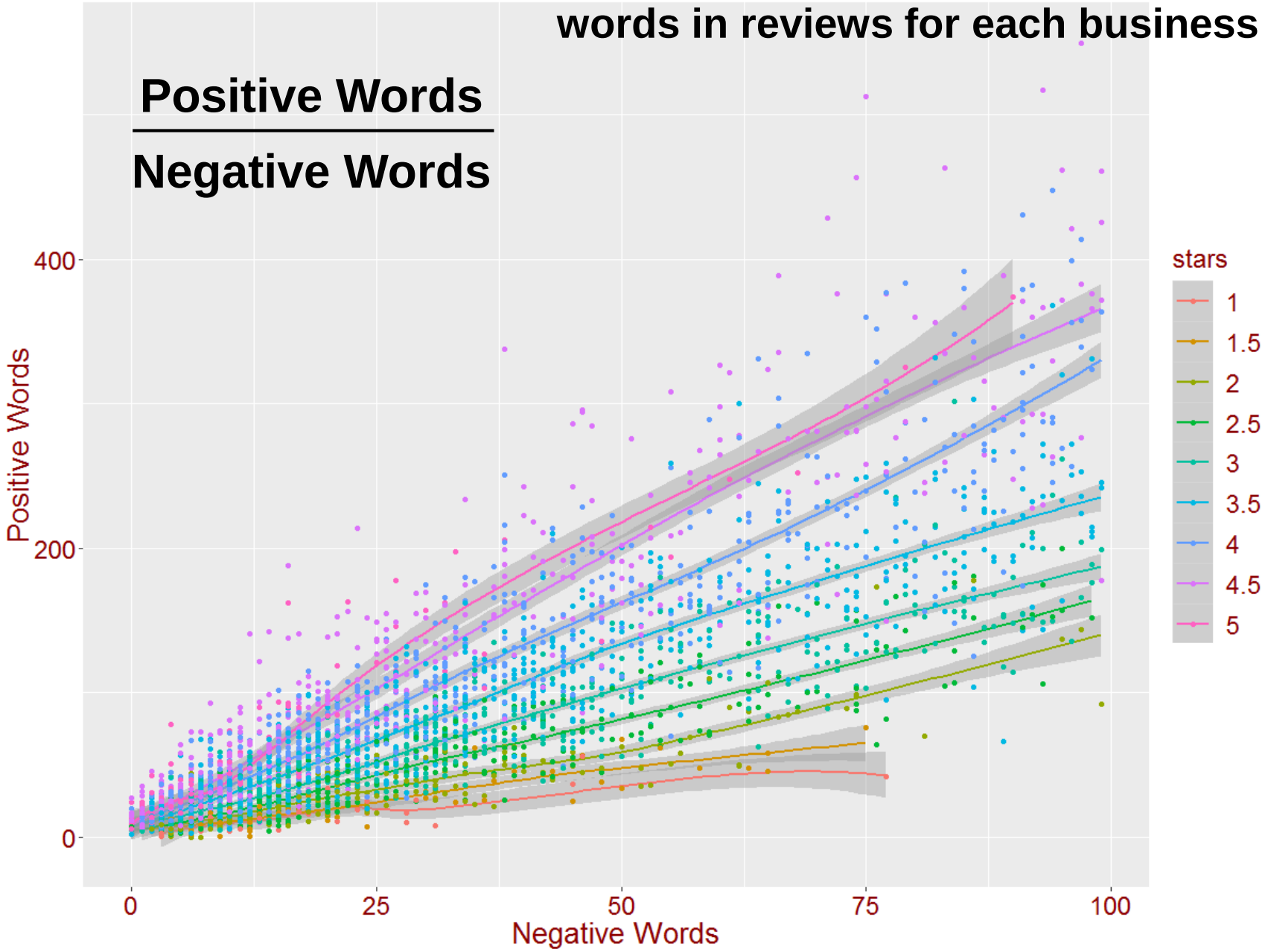
```
busi lunch
almost like NOTreadi NOTfriday
NOTlunch
slow servic
mediocr food
```

Step6:
Calculate positive and negative words

```
Good_num=0
Bad_num=3
```

- Data exploration: Review

Calculate positive and negative words in reviews for each business



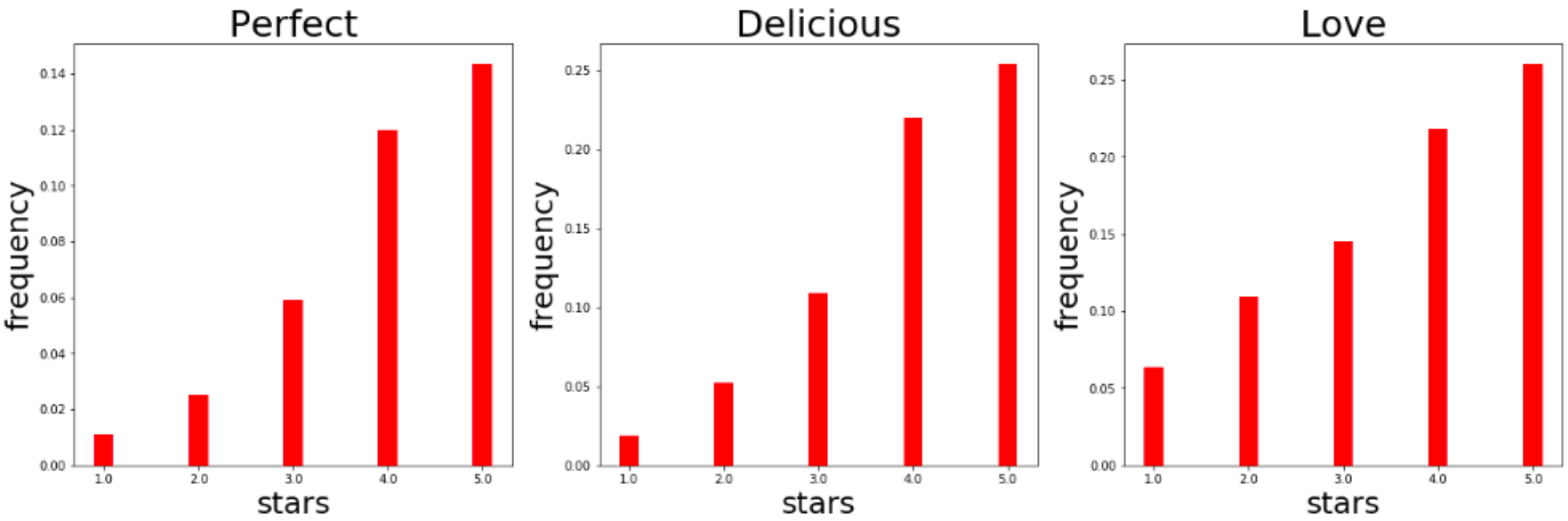
- Data exploration: Review

- 5 star

- 1 star

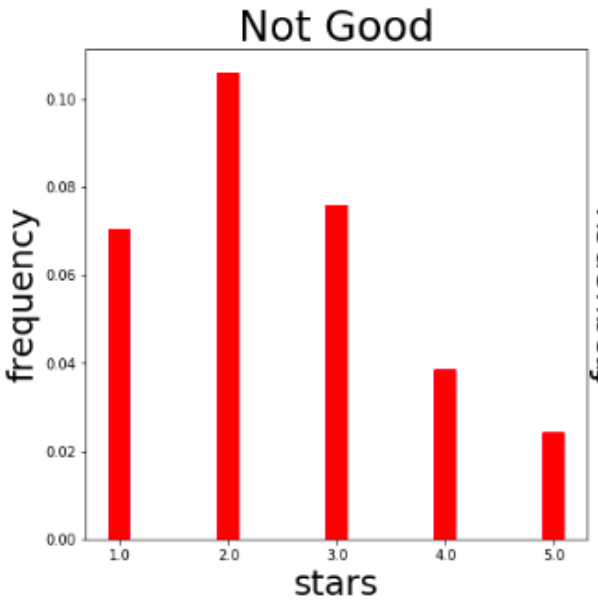
- Data exploration: Review

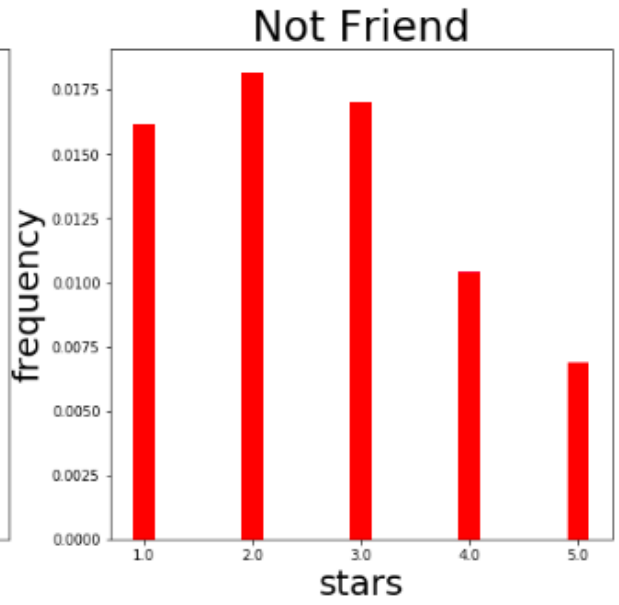
Words distribution over stars

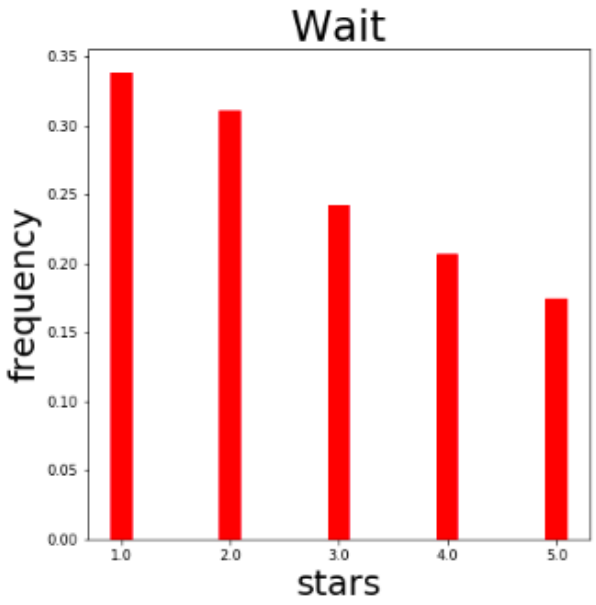


- Data exploration: Review

Words distribution over stars

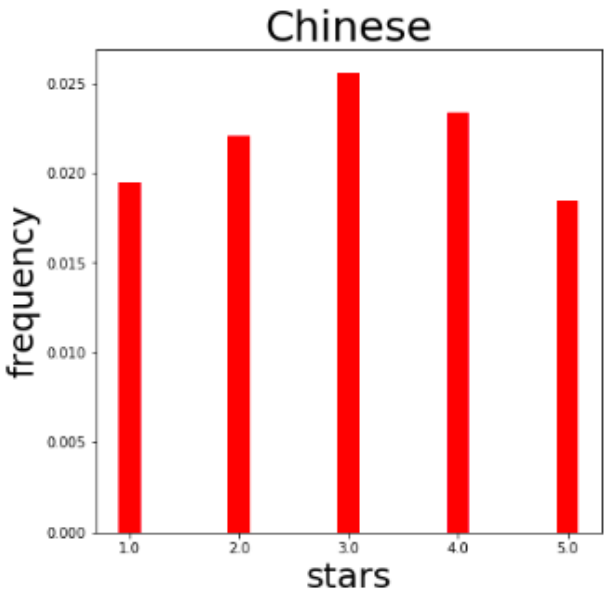


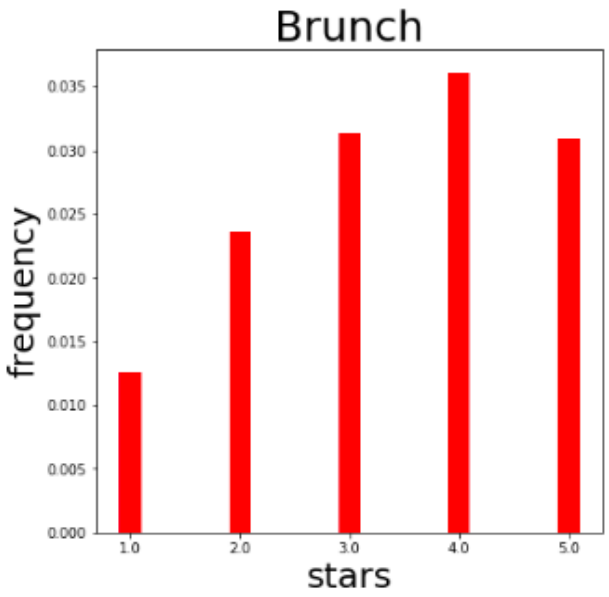


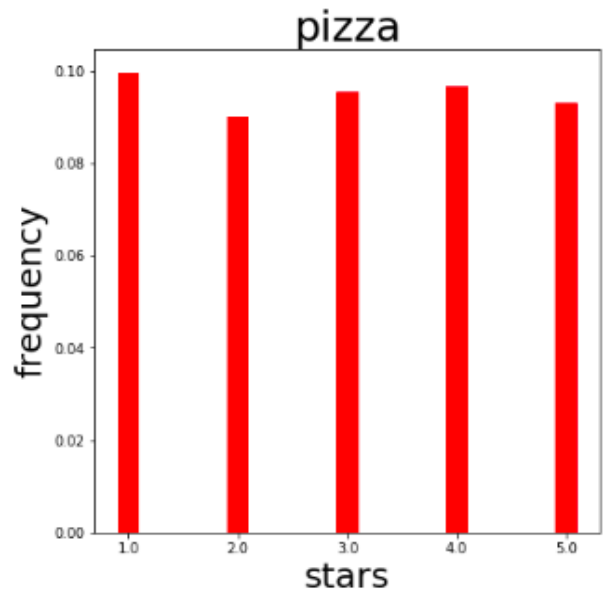


- Data exploration: Review

Words distribution over stars







- Future work

➢ Text process:
- Lemma
- Tf-idf
- Ngrams

➢ Predicting the stars
- Linear regression
- Ordinal regression
- Logistic regression
- Random forest

➢ Recommendation system :
- Classification after using tf-idf: LDA…
- Feature importance within each cluster : random forest …