Cater to your likes: Prediction on user ratings of Restaurants with Yelp Dataset

Krishna Vastare,
Lianming Shi,
Te Sun,
Tuoying Chen,
Yu Zhang,
Yizhou Chen,
Runfeng Jiang

Introduction

- Yelp is one of the largest platforms.
  - Over 145 million/ month unique visitors.
  - Over 102 million reviews to date.
- Main question: How a user would rate a restaurant before they even had food there?
- Our Goal: Identify what users care most and determine what restaurants are doing right and wrong to receive these ratings.

Datasets

Check in	Checkin_time (days of the week), business_id.
Business	business_id, Name, Address, Stars, review_count, open or not, attributes, hours, categories.
Review	Review_id, user_id, business_id, stars, text, useful, funny, cool.
User	user_id, name, review_count, yelping_since, friends..

Check-in

Check-ins on a business.

```
{
    // nested object of the day of the week with key of the hour (using a 24hr clock) with the count of checkins  for that hour (e.g. 14:00 - 14:59).
    "time": {
        "Wednesday": {
            "14:00": 2,
            "16:00": 1,
            "2:00": 1,
            "0:00": 1
        },
        "Sunday": {
            "16:00": 8,
            "14:00": 3,
            "15:00": 3,
            "13:00": 1,
        },
        "Friday": {
            "16:00": 1
        },
    },
    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg"
}
```

Check-in
-- Data Preprocessing

1. Aggregation checkins by business ID

```
length of checkin data frame 146350
length of business data frame 174567
length of contatenated data frame 174567
```

2. Drop the business with NaN check-in time
3. Throw useless columns, like 'address', 'latitude', 'longitude', 'neighborhood', 'is_open', and 'postal_code'

Check-in

-- Data Preprocessing

1. Number of total check-in : 16,648,352.
2. Number of total check-in hours: 168 (Monday - Sunday, 24 hour based).
3. Proportion of total check in times captured in top # check in hours:

#	%
Top 1	1.51%
Top 10	14.12%
Top 20	26.44%
Top 30	37.06%

Check-in

-- Data Exploratory

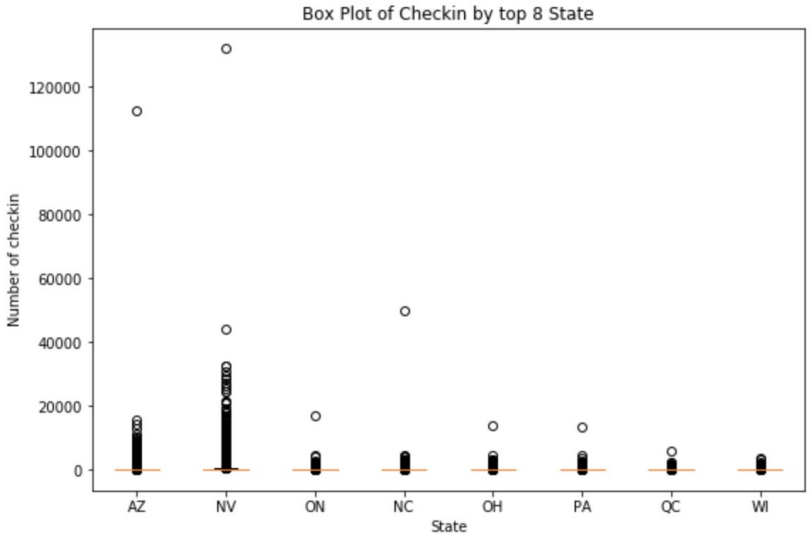
Top 10 Popular check-in hour:

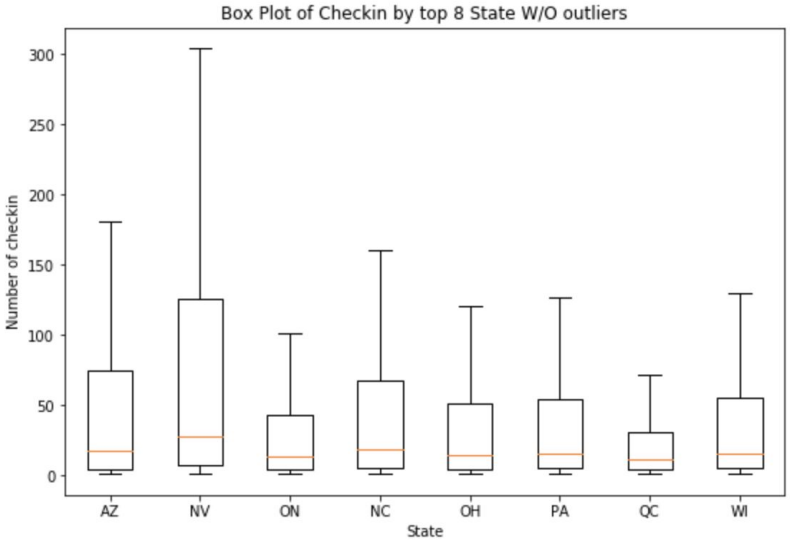
Rank	No. of times	Date	Time
1	251537	Saturday	19:00
2	246969	Saturday	20:00
3	245453	Saturday	1:00
4	236195	Saturday	2:00
5	235292	Sunday	1:00
6	229788	Saturday	21:00
7	229561	Sunday	19:00
8	225651	Saturday	18:00
9	225209	Sunday	0:00
10	224716	Sunday	2:00

Least 10 Popular check-in hour:

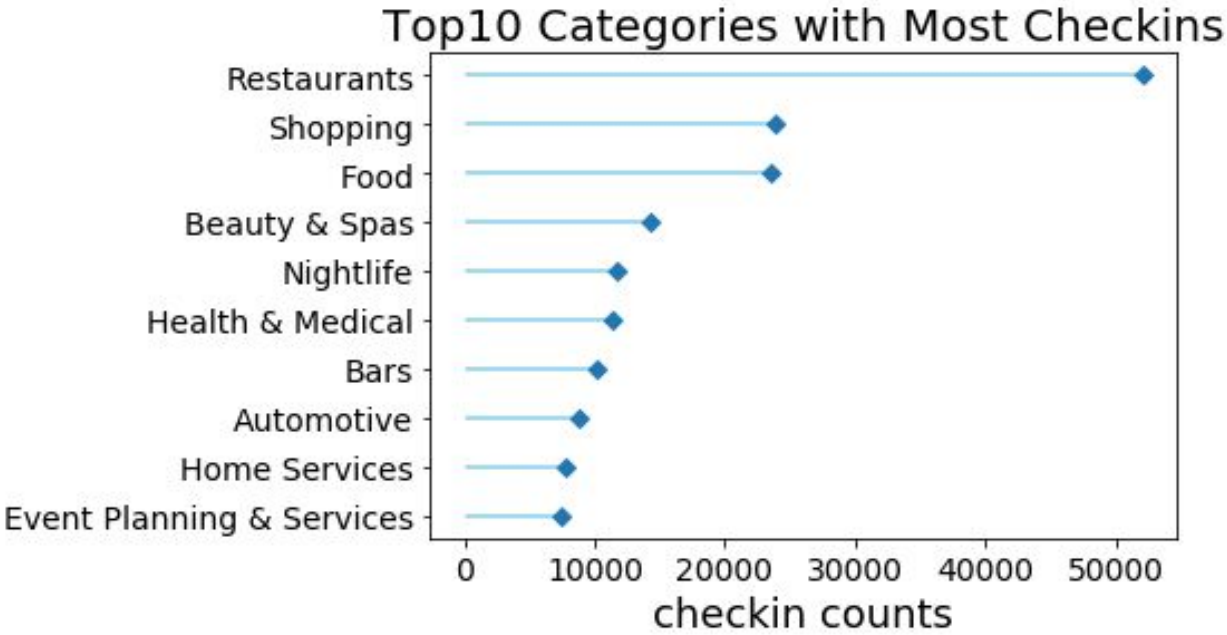
Rank	No. of times	Date	Time
1	8837	Tuesday	9:00
2	9200	Wednesday	10:00
3	9448	Tuesday	10:00
4	9776	THURSDAY	10:00
5	10058	Thursday	9:00
6	10432	Monday	10:00
7	11000	Monday	9:00
8	11837	Friday	10:00
9	12452	Tuesday	8:00
10	12484	Wednesday	8:00

Check-in

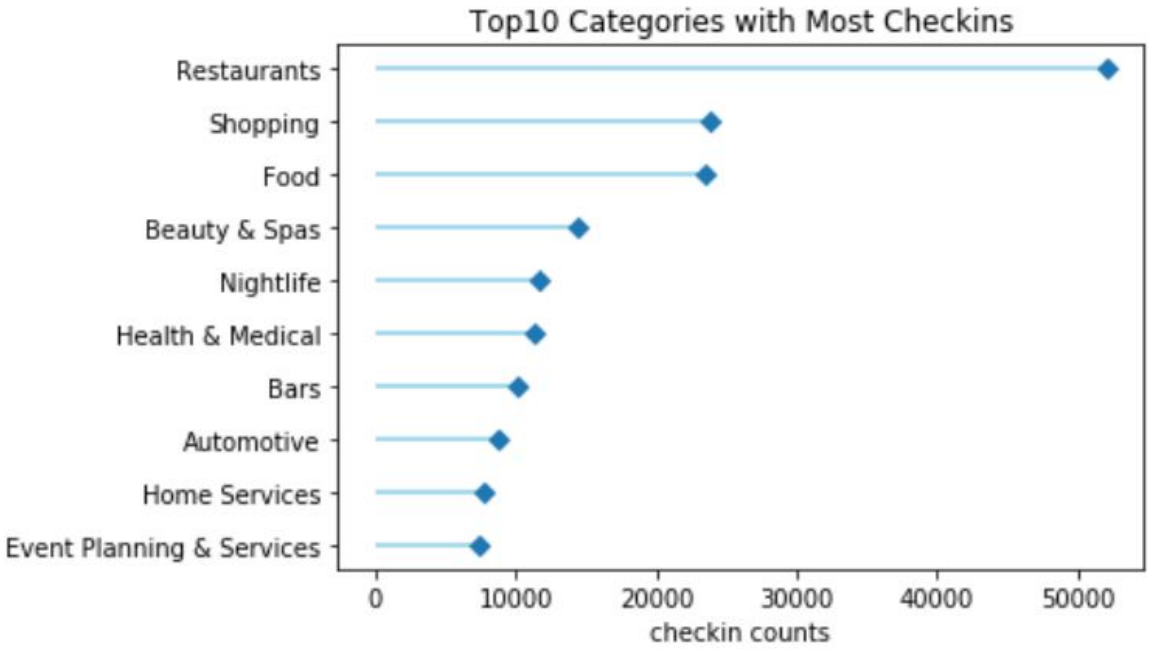

Box Plot of Checkin by top 8 State
Number of checkin
120000
100000
80000
60000
40000
20000
0
AZ
NV
ON
NC
OH
PA
QC
WI
State


Box Plot of Checkin by top 8 State W/O outliers
Number of checkin
300
250
200
150
100
50
0
AZ
NV
ON
NC
OH
PA
QC
WI
State

Check-in


Top10 Categories with Most Checkins
Restaurants
Shopping
Food
Beauty & Spas
Nightlife
Health & Medical
Bars
Automotive
Home Services
Event Planning & Services
0
10000
20000
30000
40000
50000
checkin counts

Check-in


Top10 Categories with Most Checkins
Restaurants
Shopping
Food
Beauty & Spas
Nightlife
Health & Medical
Bars
Automotive
Home Services
Event Planning & Services
0
10000
20000
30000
40000
50000
checkin counts

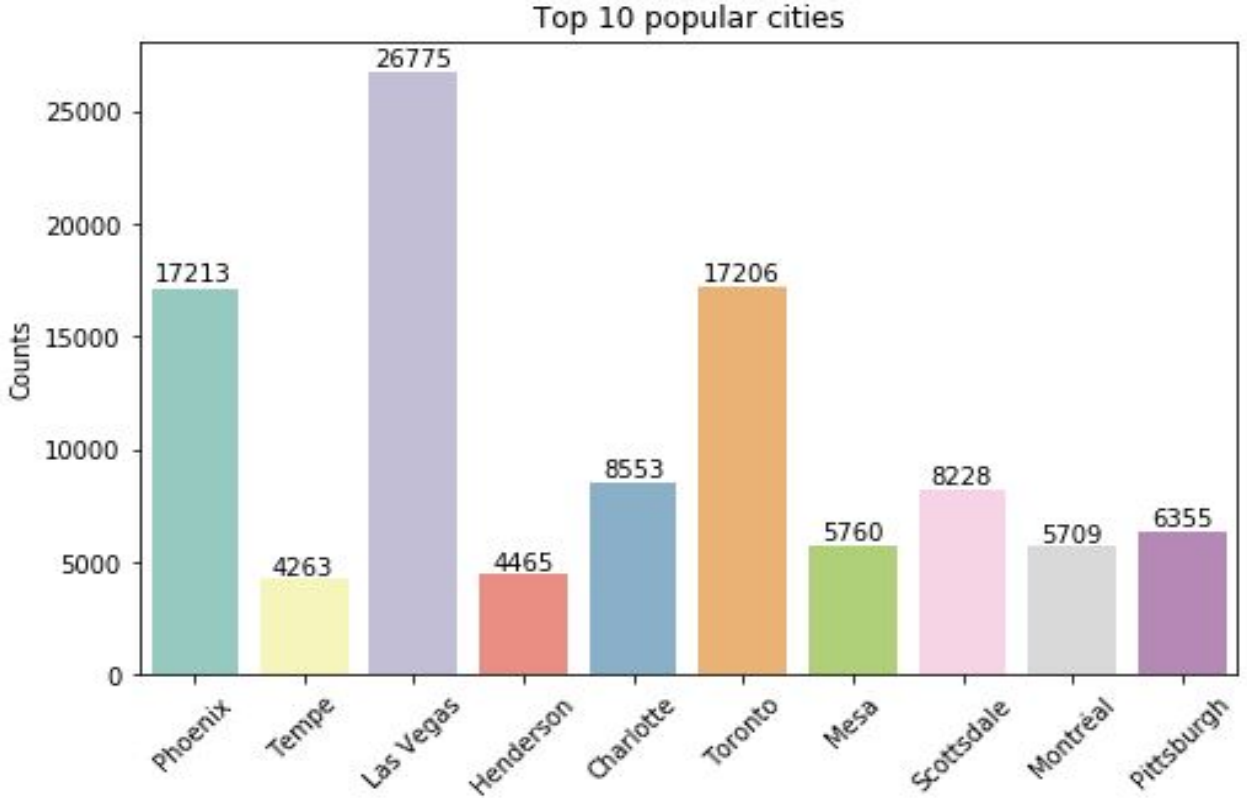
Business

Contains business data including:

- Location data
- Attributes
- Categories
- Hours

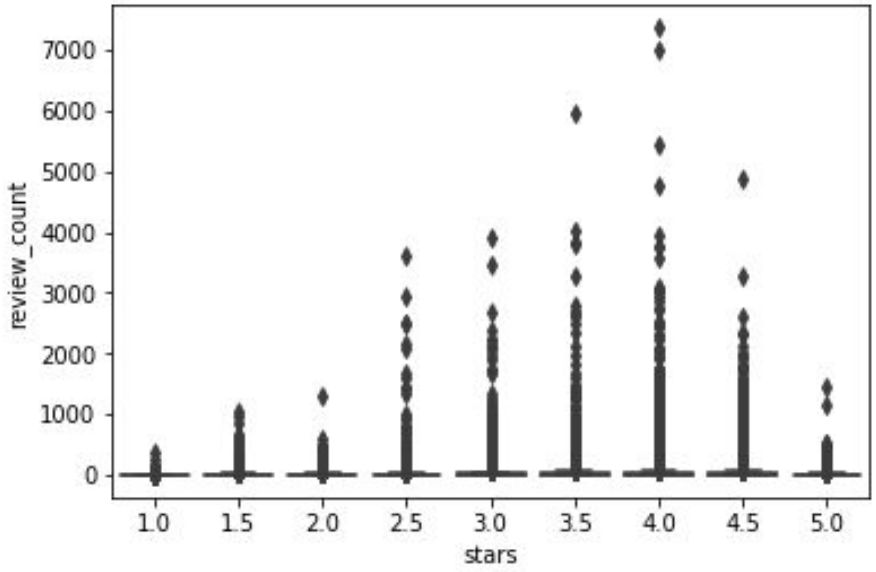

```
{
    // string, 22 character unique string business id
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",
    // string, the business's name
    "name": "Garaje",
    // string, the city
    "city": "San Francisco",
    // string, 2 character state code, if applicable
    "state": "CA",
    // float, latitude
    "latitude": 37.7817529521,
    // float, longitude
    "longitude": -122.39612197,
    // float, star rating, rounded to half-stars
    "stars": 4.5,
    // interger, number of reviews
    "review_count": 1198,
    // an array of strings of business categories
    "categories": [
        "Mexican",
        "Burgers",
        "Gastropubs"
    ]
}
```

Business: Top 10 popular cities

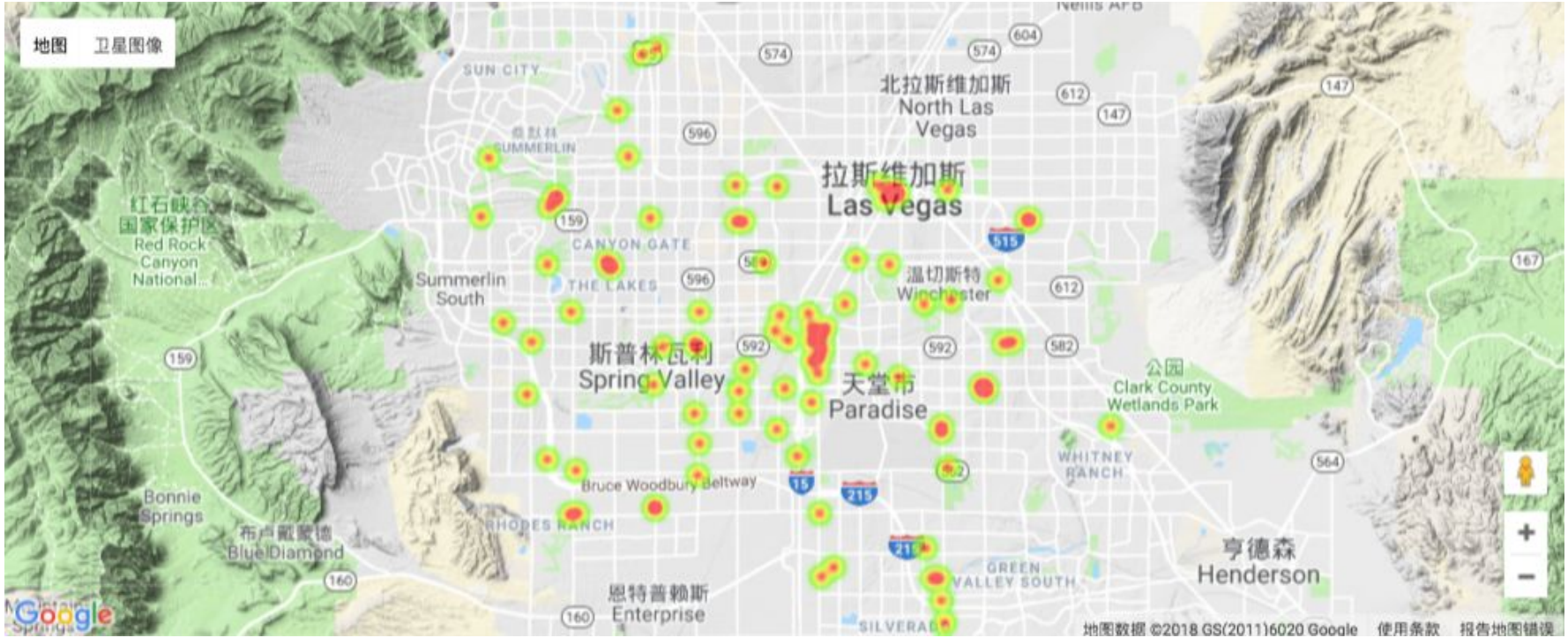


Business: Review count and Stars

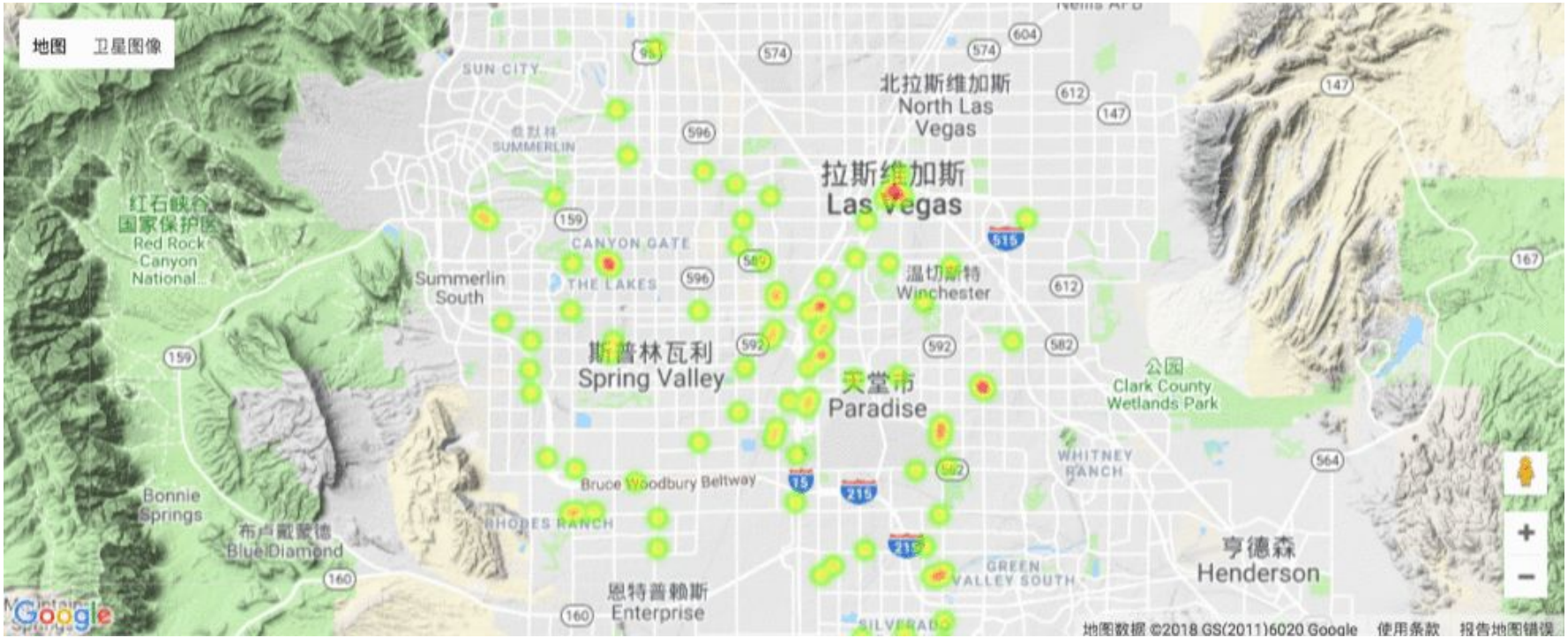
Restaurants with 3.5 stars to 4.5 stars have more review counts



Business: Las Vegas heatmap



Business: Dynamic heatmap



Review

Contains full review text data including:

- user_id that wrote the review
- business_id that the review is written for


```
{
    // string, 22 character unique review id
    "review_id": "zdSx_SD6obEhz9VrW9uAWA",
    // string, 22 character unique user id, maps to the user in user.json
    "user_id": "Ha3iJu77CxlrFm-vQRs_8g",
    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",
    // integer, star rating
    "stars": 4,
    // string, date formatted YYYY-MM-DD
    "date": "2016-03-09",
    // string, the review itself
    "text": "Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks.",
    // integer, number of useful votes received
    "useful": 0,
    // integer, number of funny votes received
    "funny": 0,
    // integer, number of cool votes received
    "cool": 0
}
```

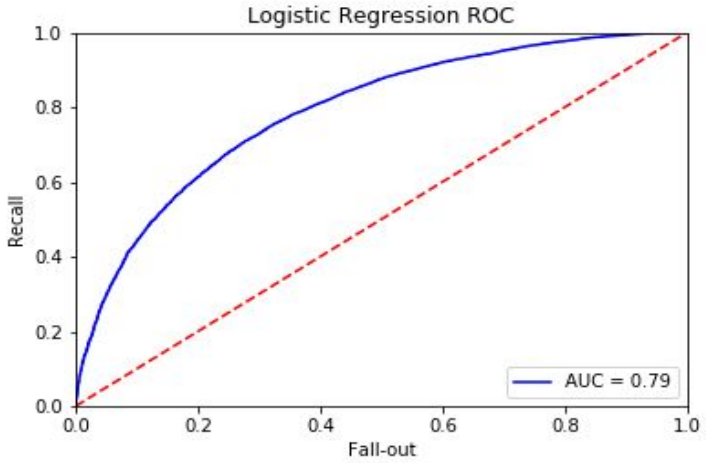
Review

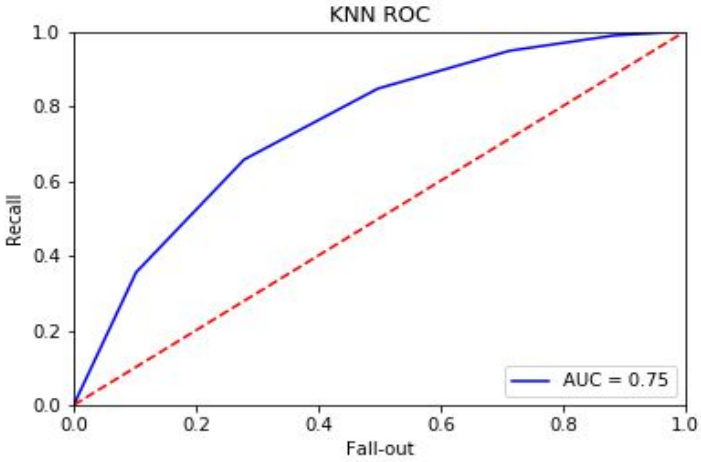
Review text topic analysis:

TF-IDF + K-Means

Cluster 1	Cluster 2	Cluster 3	
ice thai cream pad chicken nice tea place good food	sushi rolls roll place good fresh great service food sashimi	chicken good food fried rice burger sauce great salad ordered	...

Review





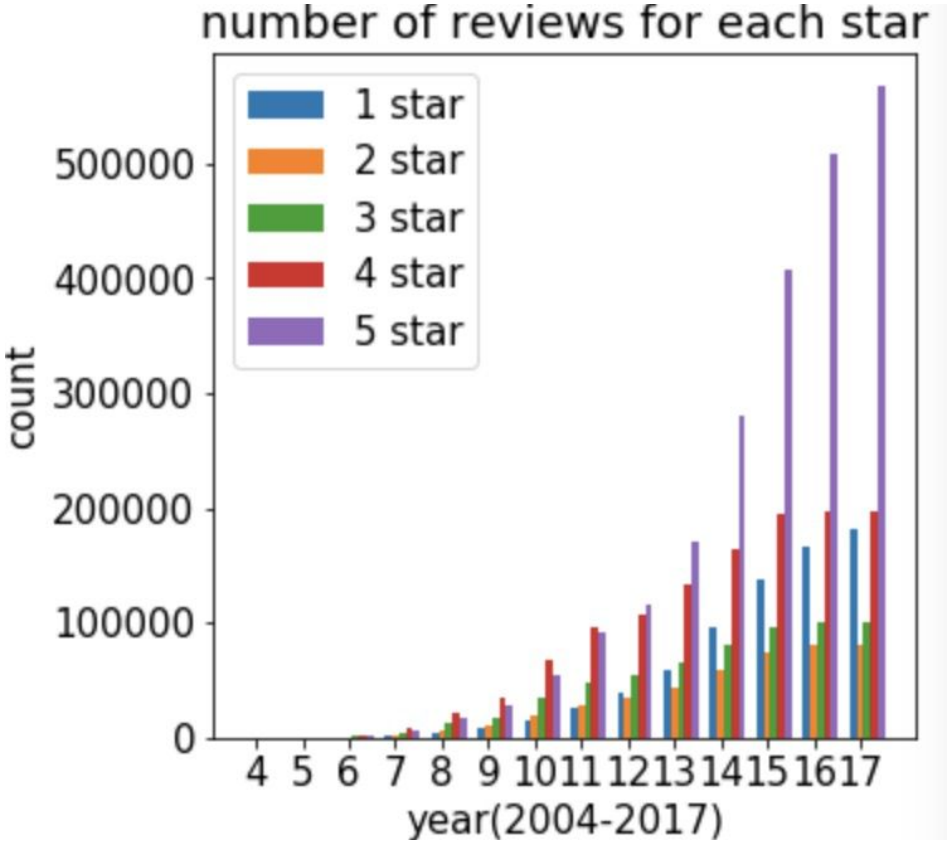
Logistic Regression Accuracy: 70.16%

KNN Accuracy: 75.09%

Review

Basic statistics

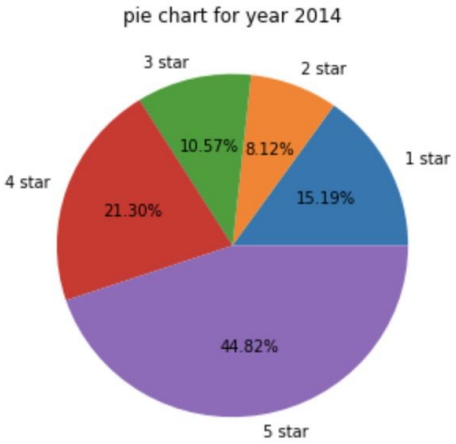
- All ratings increase rapidly with year
- 5-star ratings have the highest increasing rate
- Might want to take a look into percentage of each star level

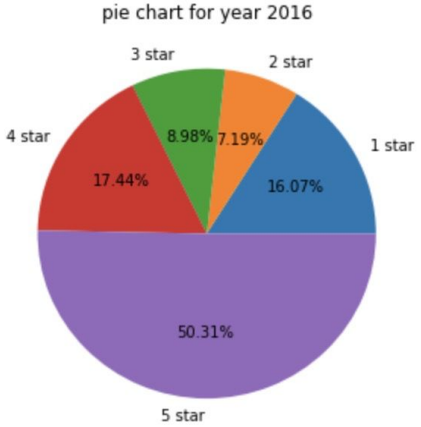


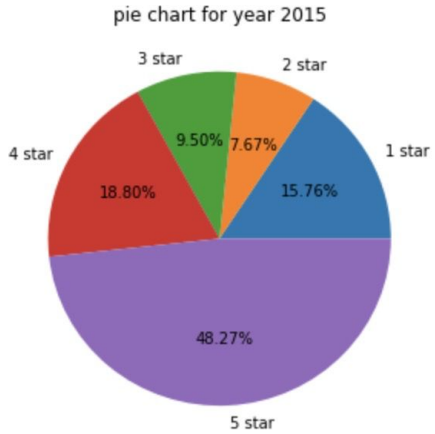
Review

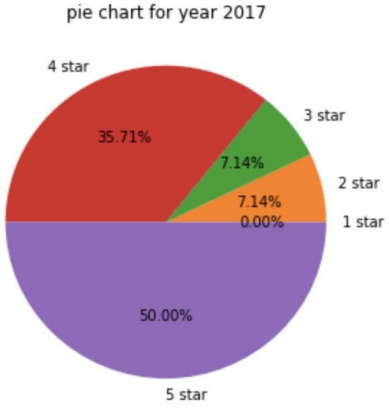
Basic statistic.

- Percentage of high star level(>3) is high.
- Percentage of high star level increase with year.
- Indicating a bias shift of user's average rating or a change in yelp's user interface.









Predictions

Prediction by collaborative filtering

- For a user-restaurant pair
- Find k-most similar users that visited the restaurant
- Use a weighted average to predict the ratings

Find users that behaves similar to user1!

	Re 1	Re 2	Re 3	Re 4	Re 5	Re 6
user1	5			4	4	1
user2	4		3	3	3	
user3	1			5		4
user4	1	1		5		4
user5			3	4		

Conclusion

- Data exploration showed us that there were over 16 million unique entries.
- Implemented TF - IDF + K-means to cluster review text topics.
- With 10D sentimental features:
  - Logistic Regression Accuracy: 70.16%
  - KNN Accuracy: 75.09%
- Finally implemented Collaborative filtering to perform prediction of the rating system:
  - RMSE around 0.91

Thank you!