Yelp Data Analysis

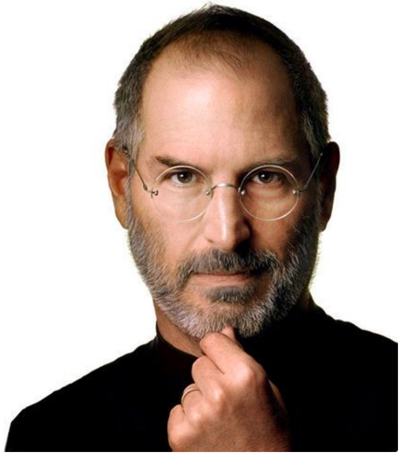
Machine Learning with R

Abhishek Choudhury
John Antony
Naveen Kumar
Vidyashree Ramu

Agenda

- Introduction
- Questions we answered
- Conclusion
- Question and answers

Introduction



The most common restaurant types in a state

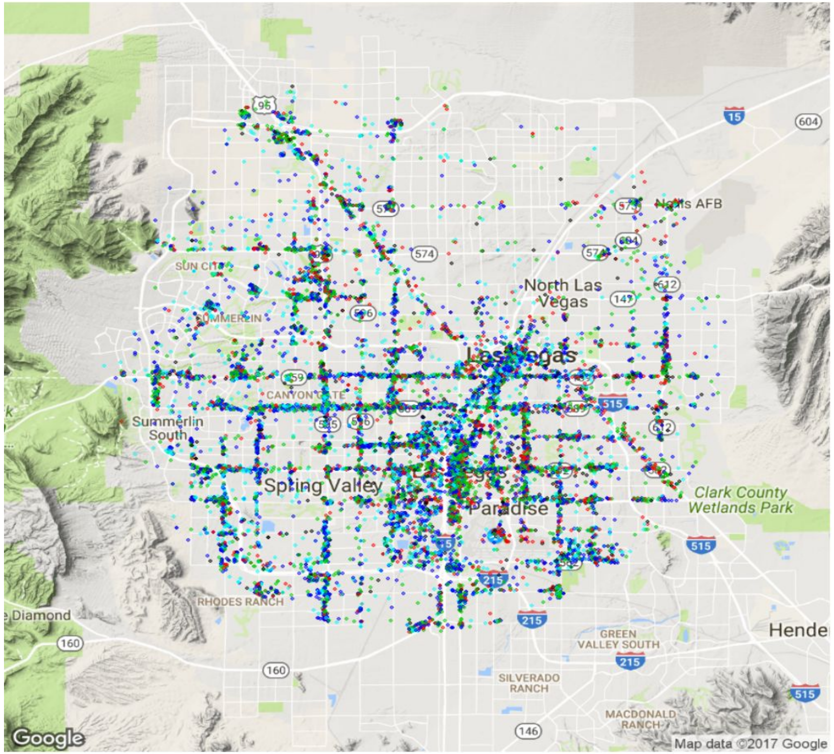
Show 25 entries

Search:

	state	categories	n
17	NV	American (Traditional)	853
20	NV	Mexican	821
22	NV	Nightlife	813
27	NV	Bars	767
31	NV	Sandwiches	724
32	NV	Pizza	699
34	NV	American (New)	648
43	NV	Burgers	571
53	NV	Chinese	471
56	NV	Italian	442
63	NV	Breakfast & Brunch	425
84	NV	Japanese	358
93	NV	Seafood	320

Popular Location and Star Ratings

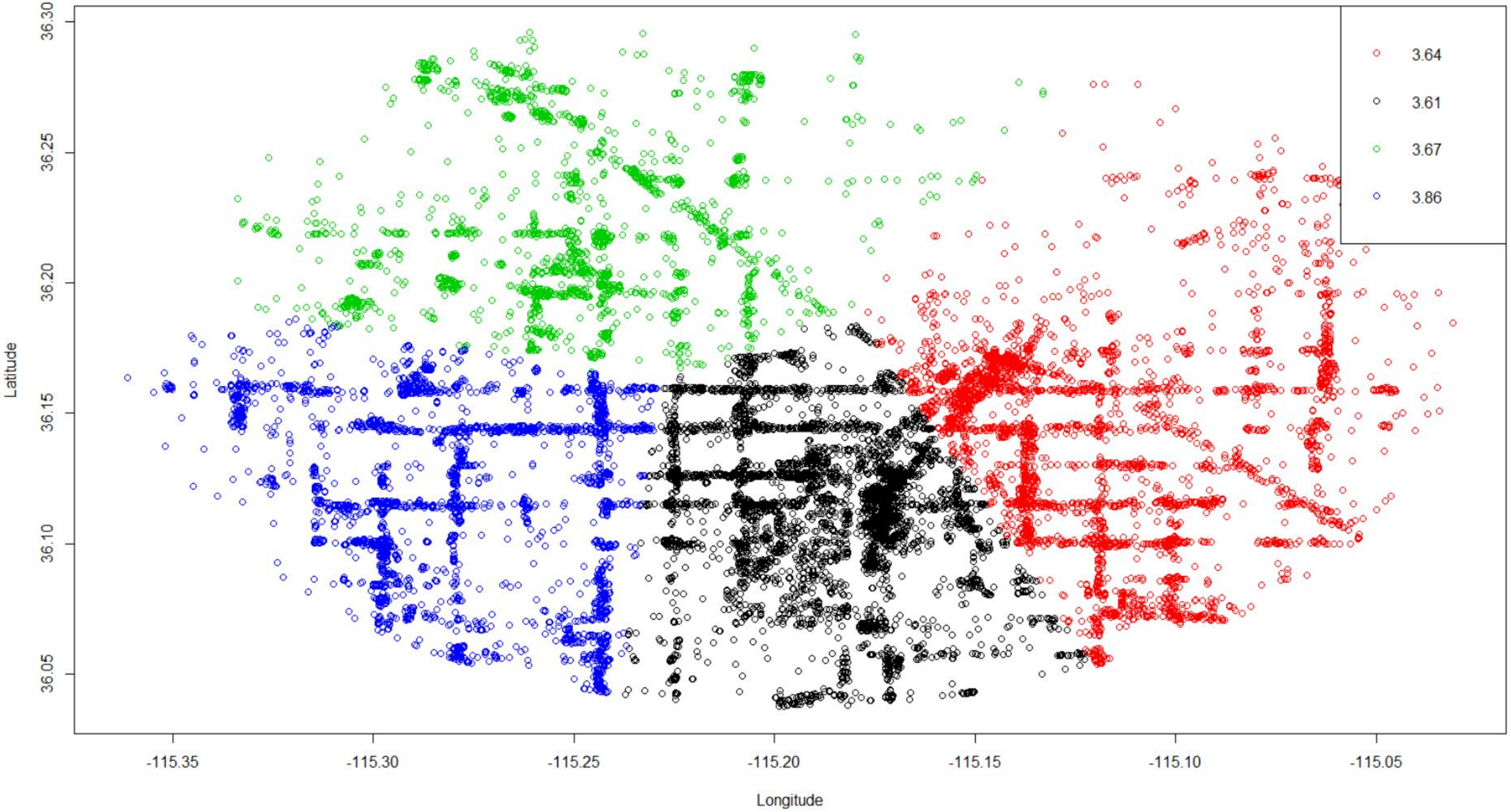
- The star rating of a business does not really depend on the location of business.
- The distance is a bit more correlated to the number of reviews though as compared to the star rating, but that figure too is very small and not really effective.



Average Ratings

- In the previous map, we noticed the difference in density of businesses across the city.
- This affects the intensity of competition in various parts of the city.
- We tried finding out the average ratings across different regions of the city.
- Thus we divided the city into 4 primary clusters.
- For the comparison, we calculate the average rating per cluster.
- Following is the outcome which we see:

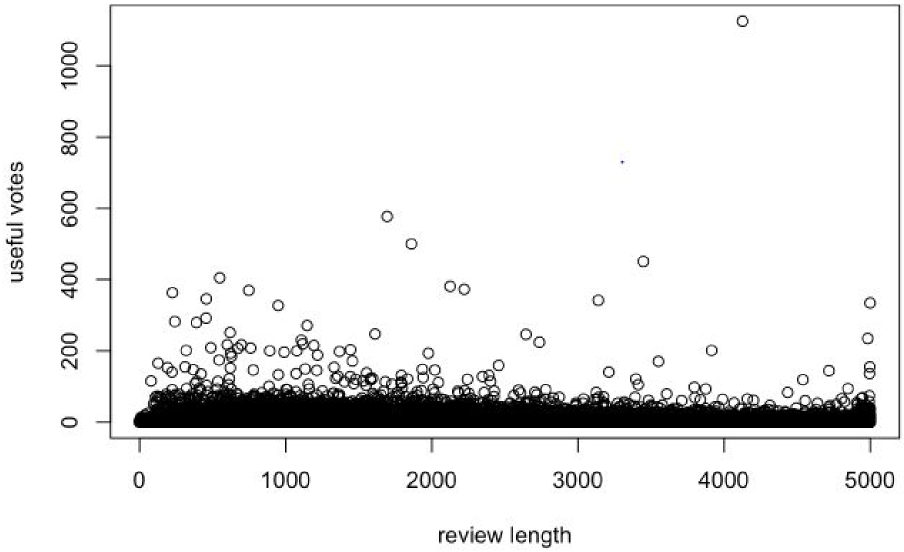
Average Ratings



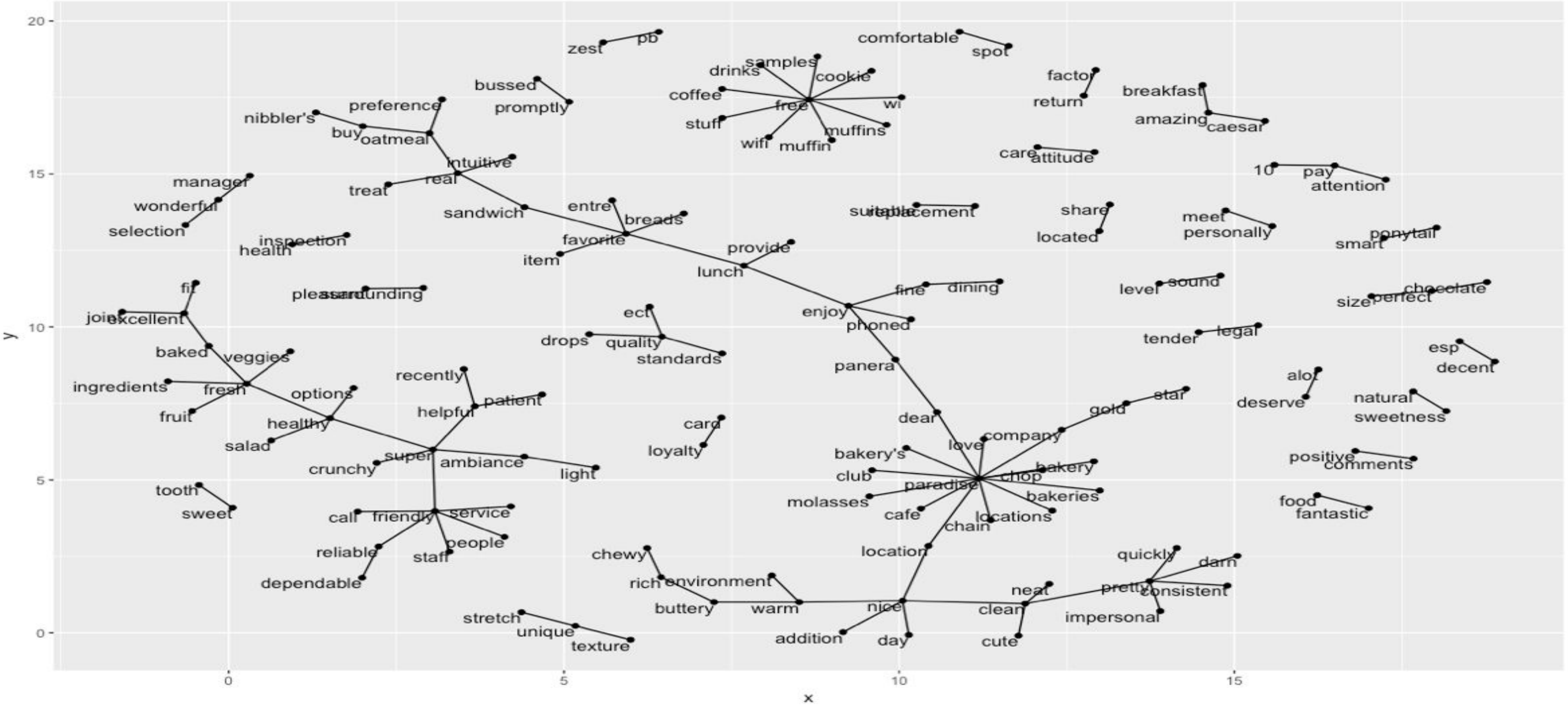
Regression (Review Sentiment and Rest Rating)

```
Call:
lm(formula = business_score_rating$SentimentScore ~ business_score_rating$stars)

Residuals:
    Min      1Q  Median      3Q     Max 
-109.29  -21.40  -10.18    8.82  605.77 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  11.6247     1.4733    7.89 3.32e-15 ***
business_score_rating$stars   5.8879     0.3938   14.95  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.28 on 10275 degrees of freedom
  (4399 observations deleted due to missingness)
Multiple R-squared:  0.02129,	Adjusted R-squared:  0.02119 
F-statistic: 223.5 on 1 and 10275 DF,  p-value: < 2.2e-16
```

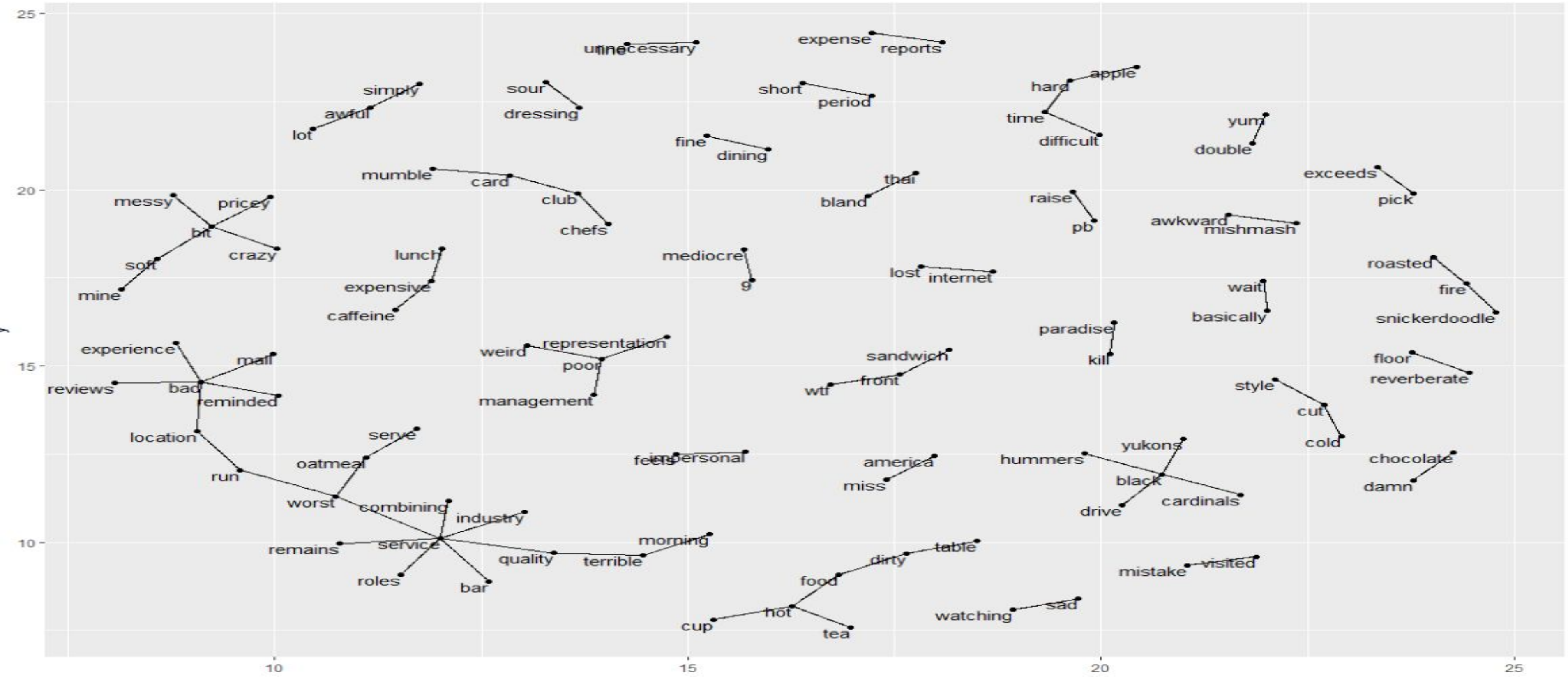
Relation between review length and usefulness



Positive Bigrams

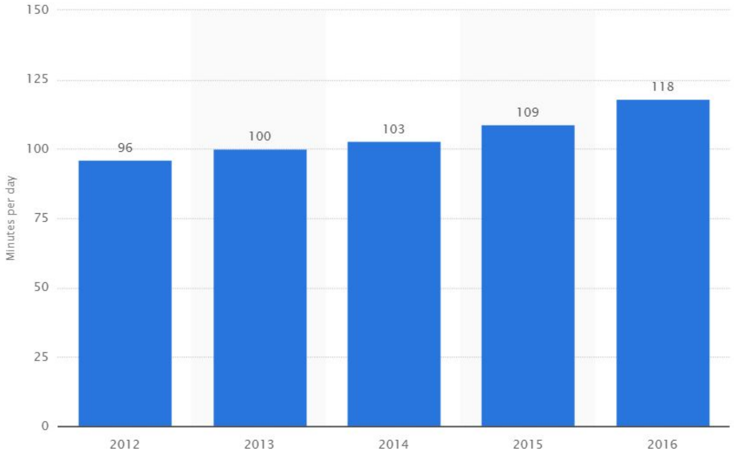


Negative Bigrams



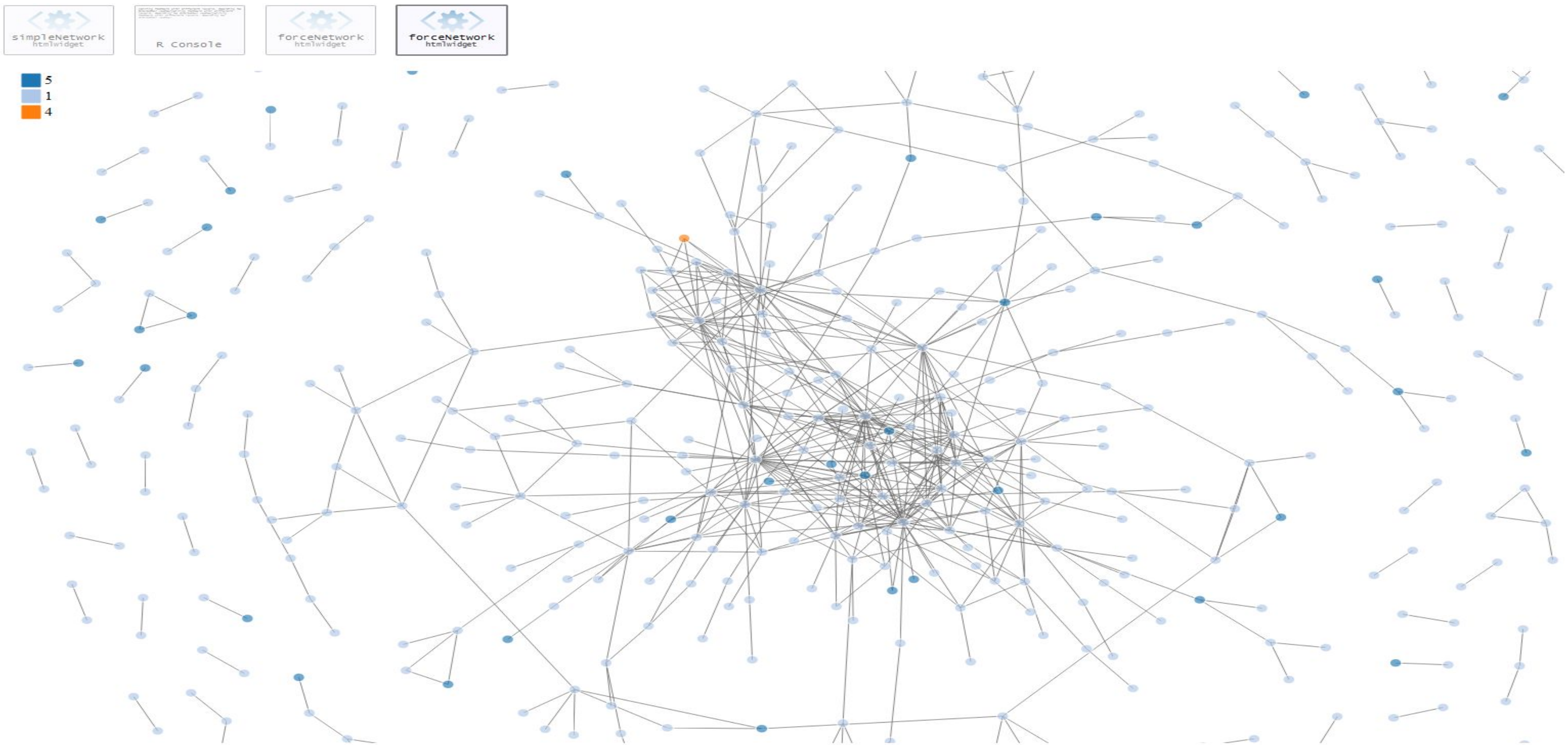
Identify influential customers

- In a world powered by the Internet, we spend more time on social networks – 1.7 hours per day, on average.[1]
- There are many customers who review a business, finding the most influential from them and increasing engagement with them can increase their revenue and social media footprint.
- Building a network of reviewers and their friends will enable a business owner to easily identify the influential reviewers.



Time spent on social networking - from Statista

Identify influential customers



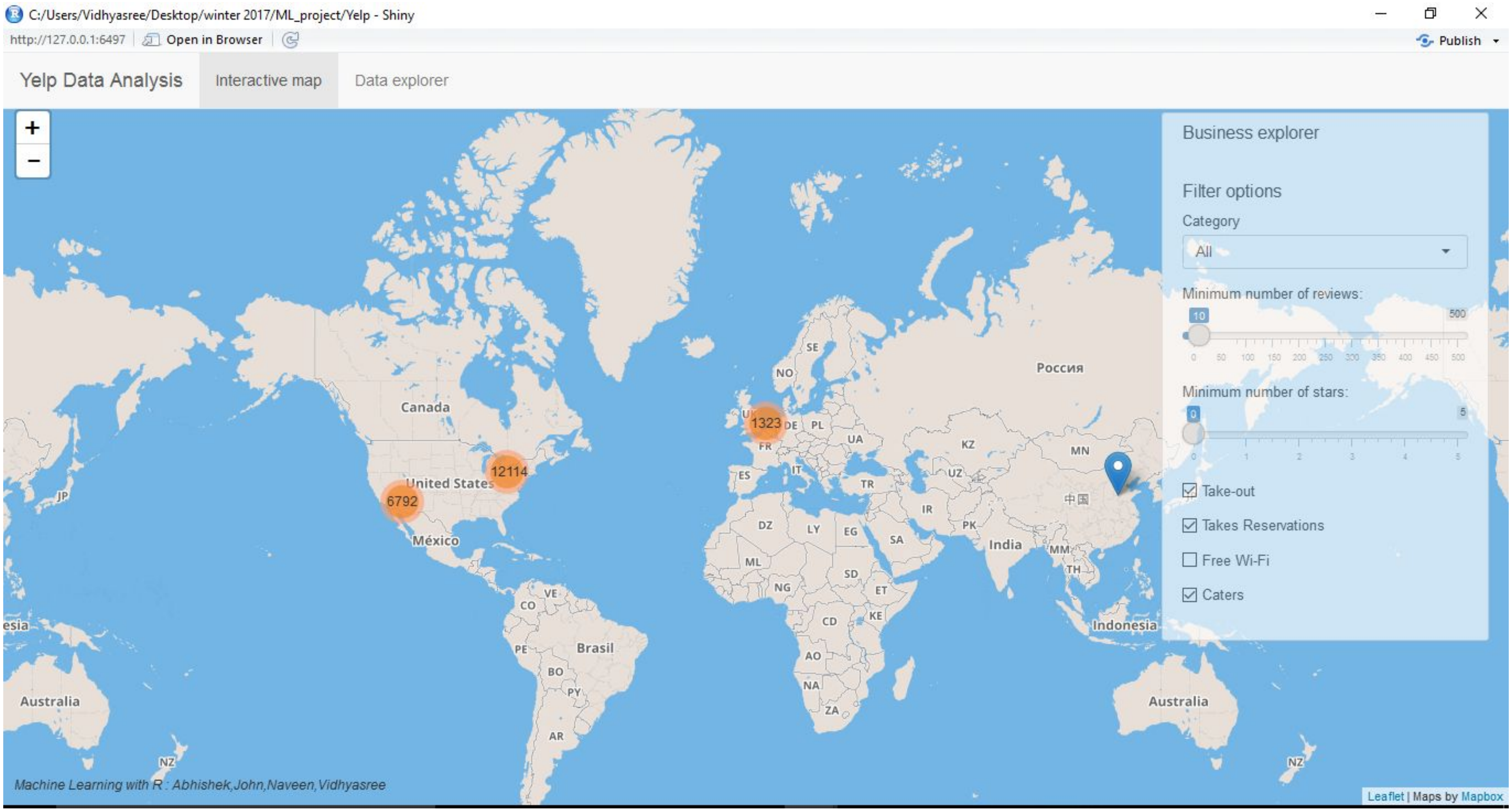
Recommendation for Customers

- Reviewers share their opinion and ratings for a business. We used that and built a matrix with customers and business.
- We built a recommendation system, which will recommend customers with a set of business.
- This recommendation system results are based on past reviewer ratings.
- We used ALS and built a recommendation system.

```
"List of Business recommended based on users past rating"
 61 107  50  99 137  69
```

```
Iteration (opt. S): 1, RSS: 16238228, RD: 0.1111018
Iteration (opt. C): 2, RSS: 49159.49, RD: 0.9969726
Iteration (opt. S): 3, RSS: 8187.716, RD: 0.8334459
Iteration (opt. C): 4, RSS: 8187.629, RD: 1.068255e-05
Initial RSS / Final RSS = 18267814 / 8187.629 = 2231.148
```

Shiny Time!


C:/Users/Vidhyasree/Desktop/winter 2017/ML_project/Yelp - Shiny
http://127.0.0.1:6497
Open in Browser
Publish
Yelp Data Analysis
Interactive map
Data explorer
Business explorer
Filter options
Category
All
Minimum number of reviews:
10
500
Minimum number of stars:
0
5
Take-out
Takes Reservations
Free Wi-Fi
Caters
12114
6792
1323
Canada
United States
México
Россия
India
Brasil
Australia
Indonesia
中国
Machine Learning with R : Abhishek,John,Naveen,Vidhyasree
Leaflet | Maps by Mapbox

Data Explorer

C:/Users/Vidhyasree/Desktop/winter 2017/ML_project/Yelp - Shiny

http://127.0.0.1:6497 | Open in Browser | Publish

Yelp Data Analysis | Interactive map | Data explorer

States: Arizona

Cities:

Zipcodes:

Show 10 entries

Search:

	City	State	Zipcode	Stars	Lat	Long	Action
1	Glendale	AZ	85305	3.5	33.538157	-112.2599654	
2	Tempe	AZ	85281	3.5	33.4259376526	-111.940246582	
3	Glendale	AZ	85302	4.5	33.5675383	-112.1607266	
4	Phoenix	AZ	85051	3	33.5735301	-112.1257408	
5	Mesa	AZ	85209	4	33.3794495268	-111.600837708	
6	Avondale	AZ	85392	3.5	33.464349	-112.276461	
7	Phoenix	AZ	85003	4.5	33.45496	-112.079908	
8	Glendale	AZ	85310	2	33.6985	-112.14	
9	Phoenix	AZ	85004	4.5	33.4485961	-112.0721686	
10	Fountain Hills	AZ	85268	4	33.6095378	-111.7253089	

Showing 1 to 10 of 4,119 entries

Previous 1 2 3 4 5 … 412 Next

Learnings and Challenges

- Advantage of having demographic information
- Plotting and calculating distances using ggmap package
- Identify and filter relevant information

- Simple Recommendation system using ALS
- The JSON formatted entries had to be joined into a complete JSON object in order to be read into R and converted to a dataframe.

- Bigram Sentiment Analysis
- Regression may not be as expected
- RSentiment - tough package
- Restricting production data on laptops
- Coordination of each view to create an integrate and reactive design in Shiny
- None of the predictor variables of interest (category, attributes) was in the proper format for analysis. It was in the **nested list**

Thank you!