Yelp

Analysis of the Impact of a Super-User

Impact of a super-user’s review on a restaurant’s average number of reviews posted after his/her visit

Agenda

- Overview
- Data Sources and Preparation
- Hypothesis testing
- Text mining analysis
- Conclusion & Limitations

Overview

Yelp is an **online search service** for restaurants and other businesses that publishes crowd-sourced reviews

Scope
Dataset

Original dataset retrieved from **Kaggle** (uploaded by **Yelp**)

- 5,200,000 user reviews
- 174,000 businesses
- Spanning 11 metropolitan areas

Scope -
Topic and Location

Our analysis focuses on **restaurants** in the **Las Vegas** area

Time period of the study was from **2010** to **2018**

Introduction

- Restaurants may want to consider how to improve the popularity of their business (to ultimately increase revenue):
    - Improving Yelp reviews
    - Improve information provided online
    - Etc.

Introduction

- Specifically, restaurants could consider:
    - Would super-user reviews and check-ins improve the restaurant’s popularity?
    - If so, they could consider incentivizing super-users to visit their restaurant in exchange for leaving a review on Yelp

Problem Statement

Does a super user review for a restaurant in Las Vegas lead to more reviews on average at that restaurant?

Our Problem Statement

Data preparation

Data source and initial choices

Dataset

We obtained the Yelp dataset from Kaggle, consists of Yelp data about business attributes, business hours, reviews, check-ins, tips, users for multiple cities in the US. The total data is about 5.14 GB in size (*174567 listings*) - which is quite hard to tackle using our laptops' processing power.

Choices

We chose the city of Las Vegas (*26775*) since it has the most number of businesses listed on Yelp.

Further we decided to choose the restaurants in Las Vegas (*5902*) for our analysis - by filtering based on the data present in business$categories attribute == "Restaurants"

Further narrowing of scope

We defined a metric for identifying a super-user within the list of people who have posted reviews for Las Vegas restaurants - it has to be someone who has posted more than 300 reviews in total, which results in a list of 15 such individuals in the city of Las Vegas.

Targeting a super-user for impact analysis

- To single-out our target super-user, we aggregated the reviews data, grouping by user_id and finding out count of total reviews posted - and chose the one with highest count for further analysis.

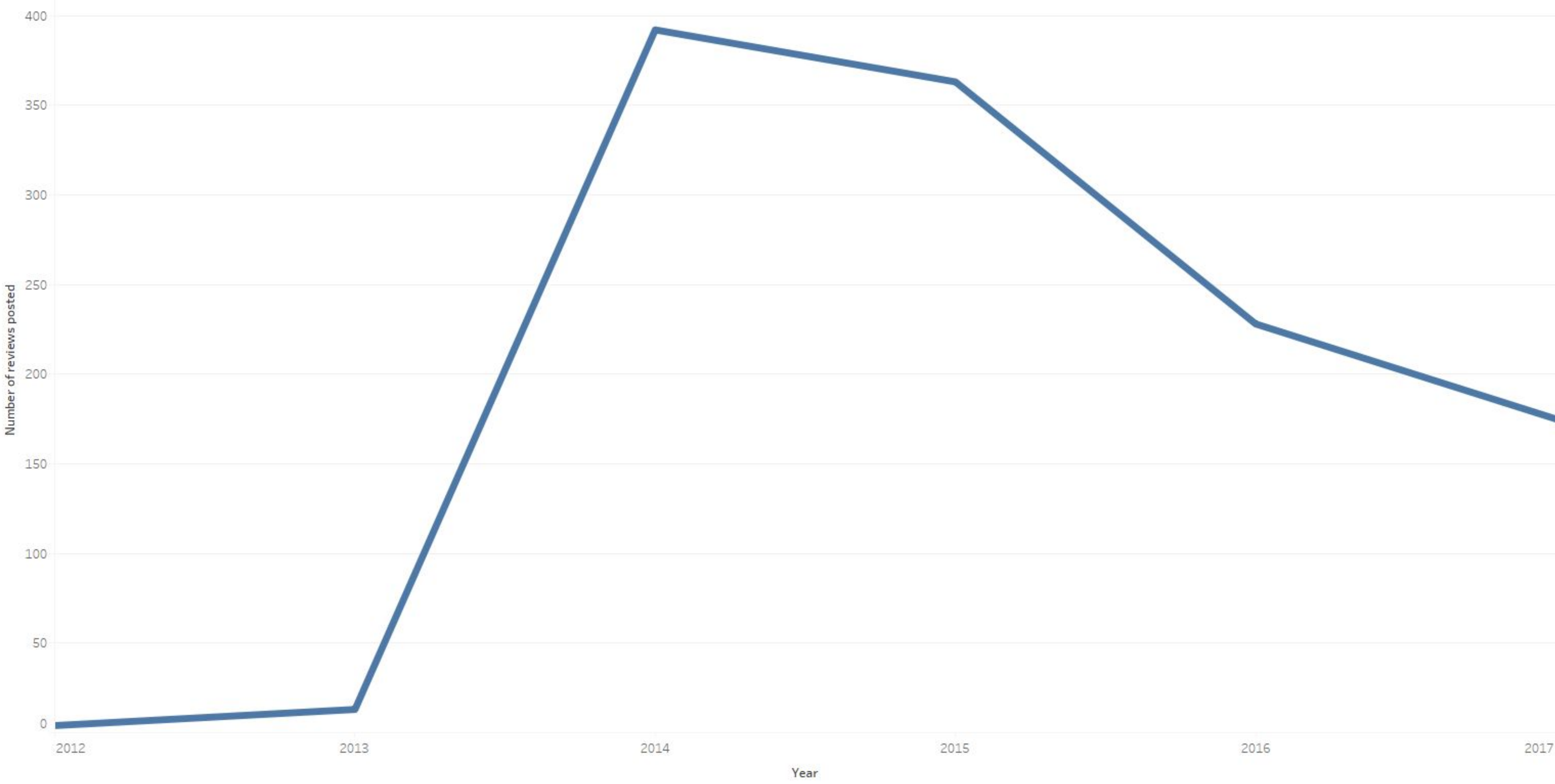
- Result : We chose *Stefany* who has posted 1175 reviews till date, which is the maximum by far - the next highest is 781 reviews. Choosing her allows us to have maximum data to test our hypothesis.

- *Stefany* has been a Yelp Elite member since 2012 through 2017 - which is a yearly squad of people Yelp recognizes to be role models on and off the site - based on well-written reviews, high quality tips, detailed profile, active voting and complementing record, playing well with others and so on.

- All of the required data manipulation was done using packages : *dplyr* and *sqldf*

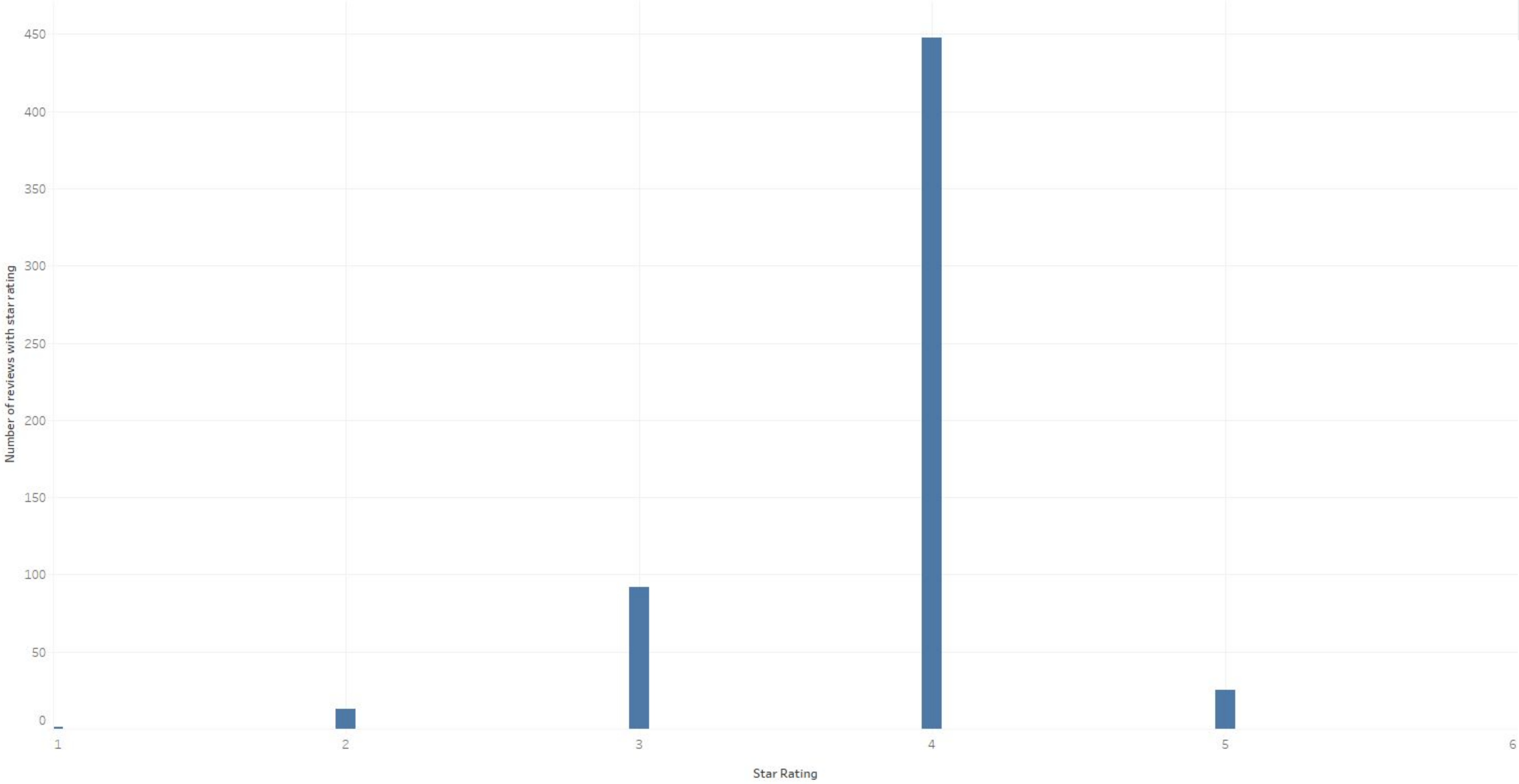
- Also, used *Tableau* for visualizing data in general.

Super user activity over years



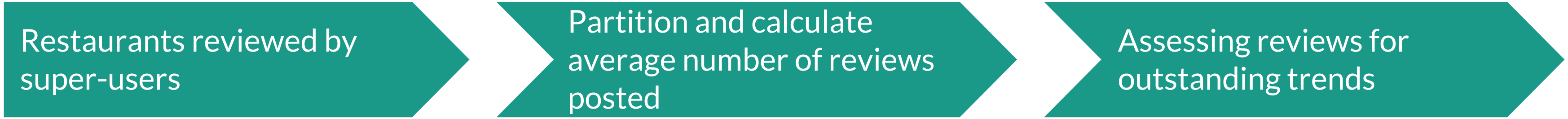
The trend of count of User Id for Date Year. The data is filtered on User Id, which keeps bLbSNkLggFnqwNNzzq-Ijw.

Super user review trend



The plot of count of Stars for Stars. The data is filtered on User Id, which keeps bLbSNkLggFnqwNNzzq-Ijw. The view is filtered on Stars, which keeps non-Null values only.

Procedure to prepare for analysis



Restaurants reviewed by super-users

We subset our reviews dataset for restaurants that have been visited by these super-users : by applying a filter on the business_ids that matter.

Partition and calculate average number of reviews posted

We subset data into two parts - one has reviews that are posted before Stefany visited those restaurants, while the other has reviews posted after. We calculate the time elapsed within the two parts, to find our mean statistic

Assessing reviews for outstanding trends

We subset review_text data for qualitative analysis : whether the reviews posted before and after Stefany’s review differ in terms of overall content and underlying sentiment.

Hypothesis Testing

Null Hypothesis(H_0)*: There is no difference in the average number of reviews after a super user has reviewed a restaurant.* $\mu_0 = \mu_A$

Alternate Hypothesis(H_A)*: The average number of reviews increases after a super user has reviewed a restaurant.* $\mu_0 < \mu_A$

Overview of the samples

Sample 1: Average number of reviews per day before a super user has reviewed the restaurant.

Sample 2: Average number of reviews per day after a super user has reviewed the restaurant.

Paired Observations

Two sets of observations are paired if each observation in one set has a special correspondence with exactly one observation in the other data set

Paired Samples T-test

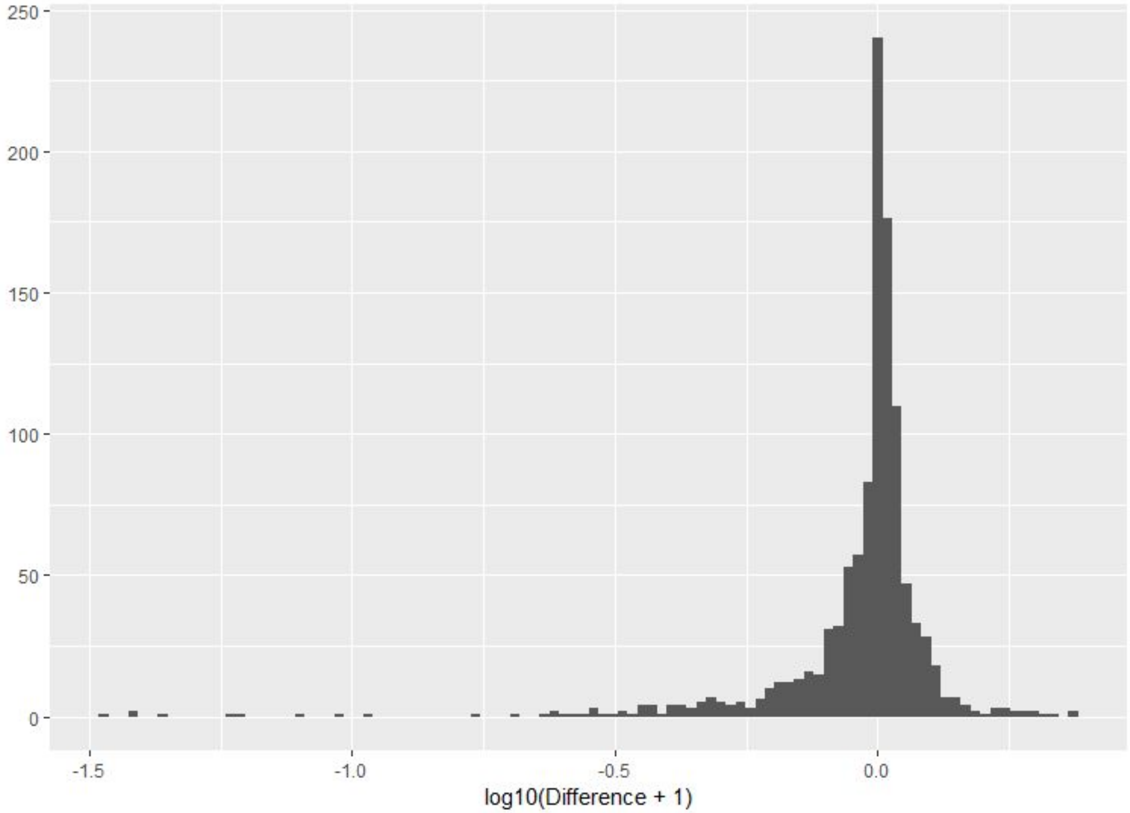
The Paired Samples T-test is used to compare the means between two related groups of samples.

- Before and after observations on the same subjects.
- Comparisons of measurements/treatments applied to the same subjects.

Conditions for Paired T-test

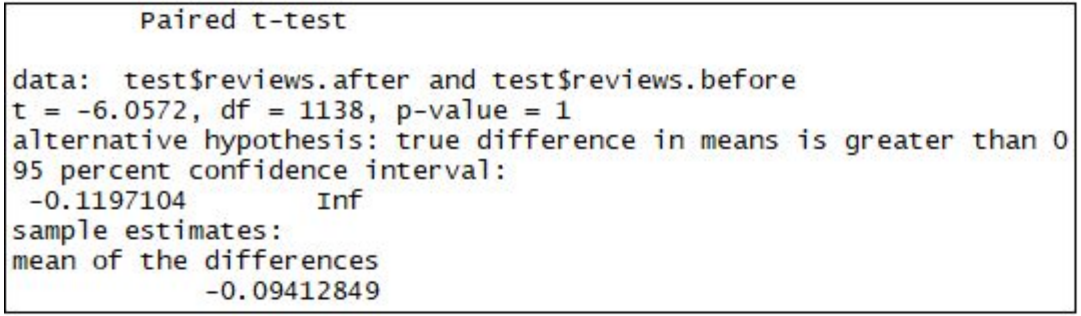
1. Paired Observations
2. Small Sample Size: Our dataset has 1139 restaurants reviewed by super user. The T distribution will approximate to a normal model with increase in sample size.
3. Normality of paired difference sample.

Distribution of paired difference sample



Paired T-test to determine P-value

```
t.test(test$reviews.after,test$reviews.before, alternative = "greater", mu = 0,
paired = TRUE, var.equal = FALSE,conf.level = 0.95)
```



```
        Paired t-test

data:  test$reviews.after and test$reviews.before
t = -6.0572, df = 1138, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1197104        Inf
sample estimates:
mean of the differences
            -0.09412849
```

Results

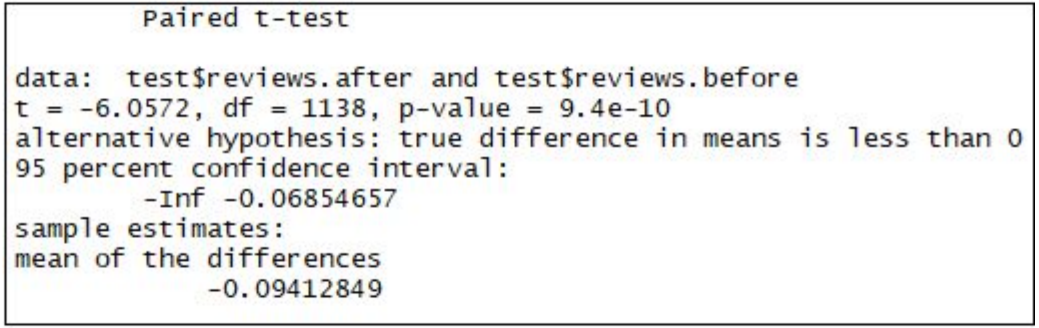
- P-value is 1.
- No evidence to show that there is an increase in the average number of reviews.
- We fail to reject null hypothesis.

Further Analysis and Findings

- Test to see if the average number of reviews has reduced after a super user has reviewed a restaurant.
- Paired Sample T-test to observe the lower tail side of the distribution.
- $\mu_A < \mu_0$

Paired T-test to determine P-value

```
t.test(test$reviews.after,test$reviews.before, alternative = "less", mu = 0,
paired = TRUE, var.equal = FALSE,conf.level = 0.95)
```



```
        Paired t-test

data:  test$reviews.after and test$reviews.before
t = -6.0572, df = 1138, p-value = 9.4e-10
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.06854657
sample estimates:
mean of the differences
            -0.09412849
```

Results

- P-value is 9.4e-10.
- Overwhelming evidence to show that there is a decrease in the average number of reviews.
- We will reject null hypothesis in this case.

Statistics

- Mean of paired difference sample : **-0.094128**
- Standard Deviation: **0.524462**
- Standard Error: **0.015540**
- 95% Confidence Interval: **[-0.1245, -0.0636]**

Text Mining Analysis

Bag of words approach

Creating a Corpus

Data Cleaning

Word Frequencies

Packages:
‘tm' and ‘quanteda’

- **Corpus** - collection of words before and after super user reviews
- Selected a random sample of 10,000 words

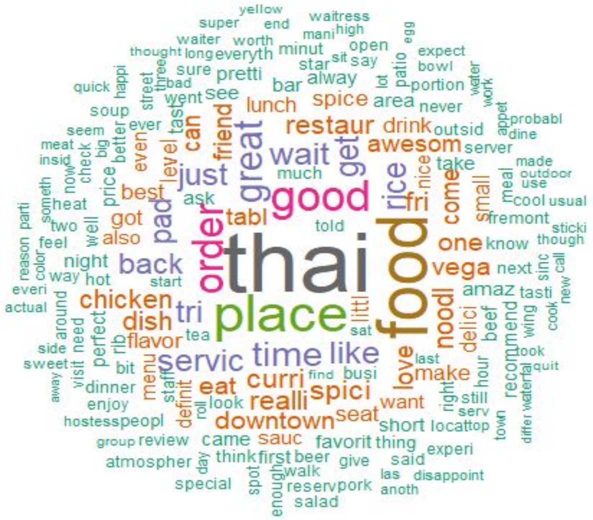
Packages:
‘tm' and ‘SnowballC’

- Converted to **lowercase**
- Removed **numbers, punctuations**
- Removed **stopwords**
- Stripped **whitespaces**
- **Stemming**

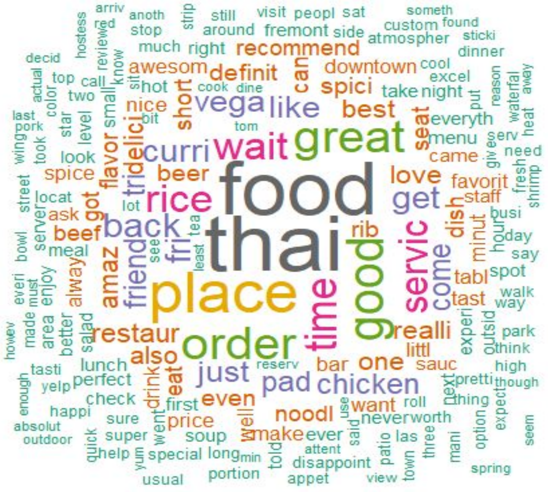
Packages:
‘tm', ‘wordcloud’, ‘RColorBrewer’

- Created a **Term Document Matrix**
- **Word cloud** visualization

Word cloud visualizations



Before super-user reviews



After super-user reviews

Conclusion

Impact

- Yelp reviews can have a large impact on how well a restaurant performs
    - A previous study found that a **one-star** rating increase can lead to **5-9%** increase in revenue

Impact

- However, we fail to reject the null hypothesis:
    - *There is no difference in the average number of reviews after a super user has reviewed a restaurant*

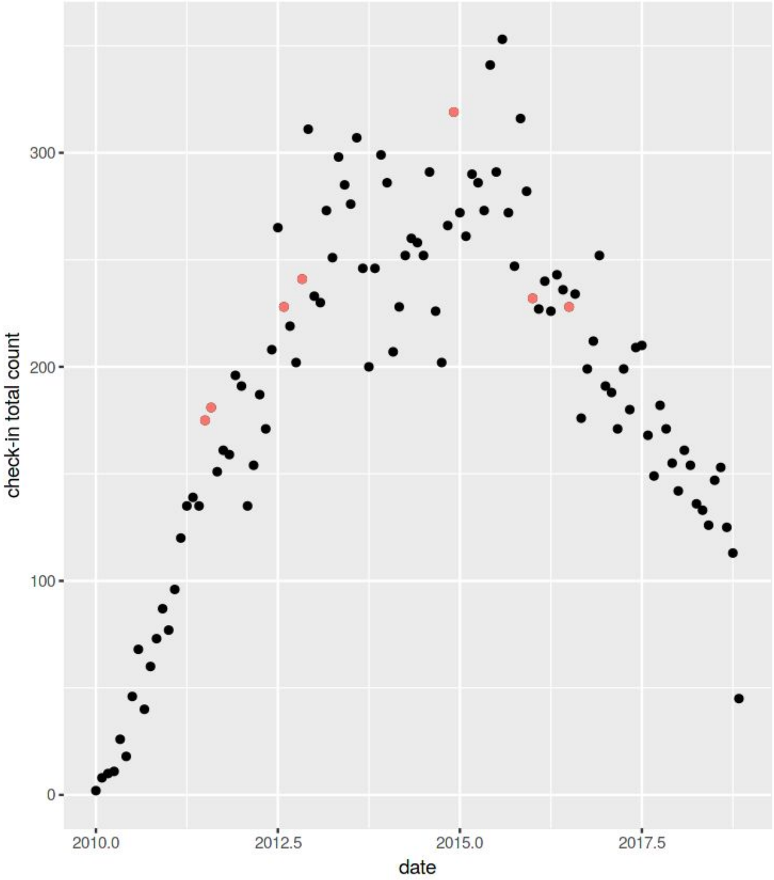
- Restaurants may want to consider investing in other ways to promote their business besides utilizing super-user reviews

Limitations

- There may be other several factors that we haven’t accounted for that may have lead to a lack of improvement of reviews for restaurants:
  - Restaurant quality may have decreased over time, resulting in less visitors and less reviews
  - If a super-user left a low-rated review, it may be that the restaurant had under-performed that day

Limitations

- Other limitations:
  - There could be less overall Yelp users who leave a review at popular restaurants
  - Popularity of a restaurant may be decreasing
- Need to account for the day of a review
  - I.e., people tend to visit restaurants more on weekends



Questions?