Yelp Customer Review Analysis

Analytics Team Member:
Yuyin Fan, Mengdie Ji, Saman Khan, Jiaxuan Wang

Statement of the Problem

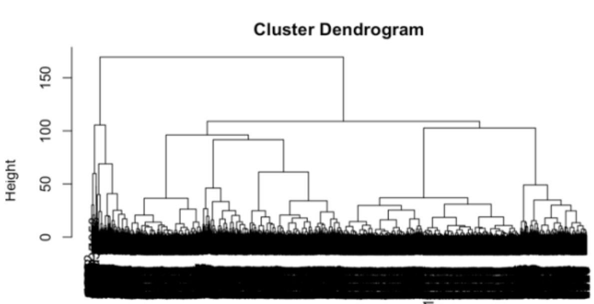
Yelp, a location-based social media platform, is integrated into many people's lives because it provides users a chance to not only share experience, write reviews, and grade the individual location using a one to five star rating system, but also achieve some recommendations and suggestions from others. Jeremy Stoppelman, Yelp's co-founder and chief executive, described it as "Yelp is about the review experience. It is like a blog with a little bit of structure." In addition, most local businesses joined Yelp and desired the site could bring impressions to attract more customers and boost product sales.

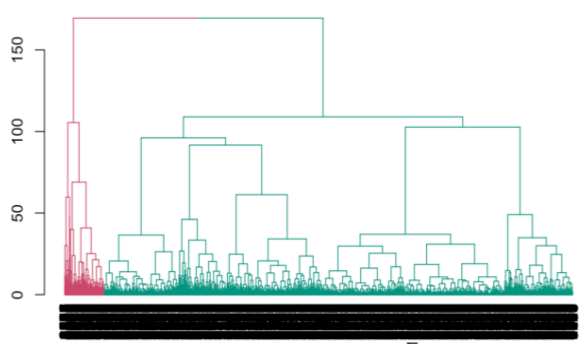
In this project our team chose the Yelp Reviews dataset which contains 10000 observations gathered from Yelp. In addition, we mainly focused on three analysis aspects: clustering analysis, sentiment analysis, and recommendation analysis.

Clustering analysis is important for businesses to do the market segments and targeting; Sentiment analysis is essential for businesses to better understand customers' attitudes toward services and different products in order to increase rating, make improvements on products, help customers make decisions, and build company reputations; and Recommendation analysis is useful to recommend items that may be of interest to the users. The ultimate goal is not only to help companies understand customers' feelings and opinions in order to make adjustments and improvements on products and services but also to find relationships between customer reviews and ratings, and attract more customers in order to increase business sales.

Clustering Analysis

1. Hierarchical cluster analysis
   Hierarchical clustering is a method that groups variables/observations based on similarities. We first defined similarity with Euclidean distance and got Cophenetic correlation coefficient (CPC) in 0.5268 which is a medium fit level of dendrogram matching the true distance metric. Based on the distance, either a two or seven cluster solution works. We then chose 2 clusters because it is more practical.



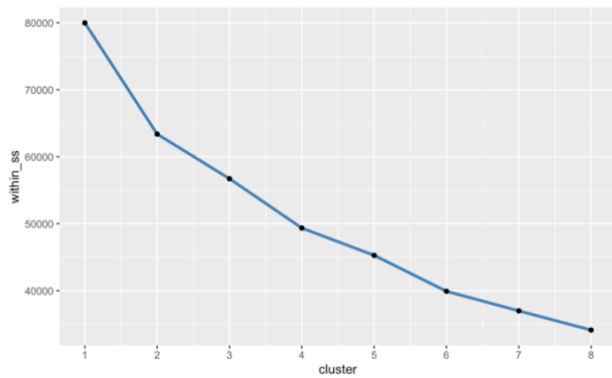


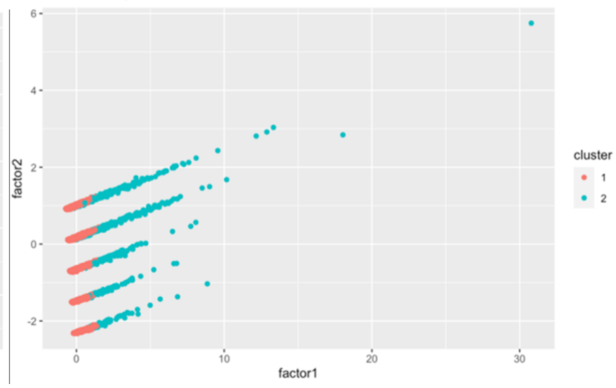
1.1. Although hierarchical clustering is clear, since we have large datasets that contain 10000 observations, this method then does not scale well.

2. K-means cluster analysis

K-means clustering attempts to find groups that are most compact and relies on Euclidean distance. We begin with an arbitrary assignment of observations to 2 cluster centroids

2.1. Interpret number of clusters with total within sum of squares plot. Clusters are inferred from a sudden change in the line graph. (Here we choose cluster of 2)



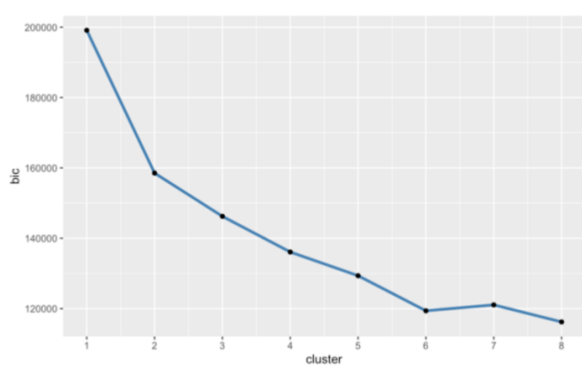


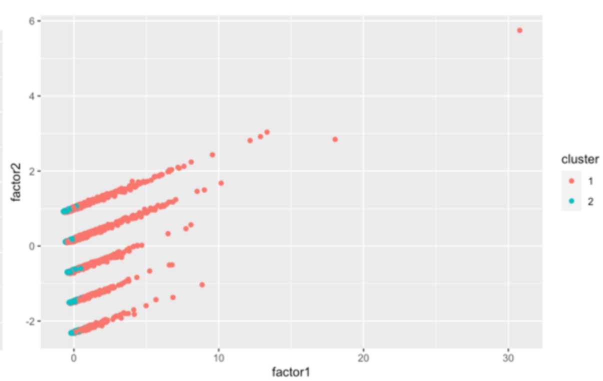
2.2. K-means is sensitive to starting values and lacks advantages of a dendrogram for inferring clusters. Therefore, it is not suitable for the Yelp datasets.

3. Model-based cluster analysis

Model-based clustering views the data as a mixture of groups sampled from different distributions and tries to find the best set of distributions to explain the observed data.

3.1. The optimal cluster solution is the one that performs best on BIC and log-likelihood. A plot of bic Interpret the number of clusters. We were looking for the lowest bic in the line graph, which is 8. 8 clusters is not practical to handle. Choose 2 clusters refer to hierarchical and k-means results.



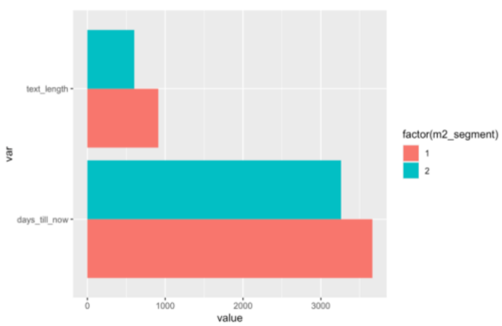


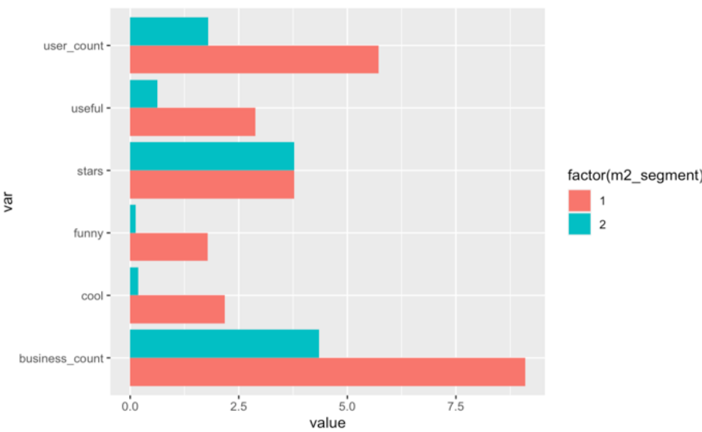
4. Contrast Result among three clustering analysis.

```
h2_segments
   1    2
9221  779
k2_segments
   1    2
9177  823
m2_segment
   1    2
3458 6542
```

Based on the result, Model-based clustering divides the observations in a more even level. Then we use the model-based clusters to inspect the characteristics in each cluster.

4.1. Combine segment membership in each cluster

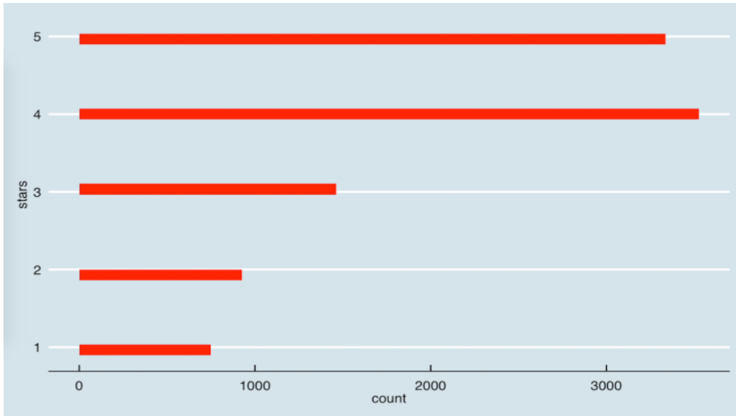




5. Conclusions in clustering analysis
    5.1. The second cluster has smaller value in each category except the category 'star'. Therefore, the star score may be irrelevant to customers' feelings on reviews, popularity of restaurants, and writers' motivations.
    Business impact: star scores may be determined by other factors. We could conduct further analysis combined with more data analysis to find out factors that affect the star scores.
    5.2. The reviews were written by less active users, described in shorter texts, or created for less popular restaurants, tend to be voted as less useful, funny, or cool.
    Business impact: High quality reviews will earn more votes on Yelp, which could help restaurants attain more attention. Restaurant managers could cooperate with active Yelp reviewers and provide incentives to encourage them to write both accurate and high quality reviews.

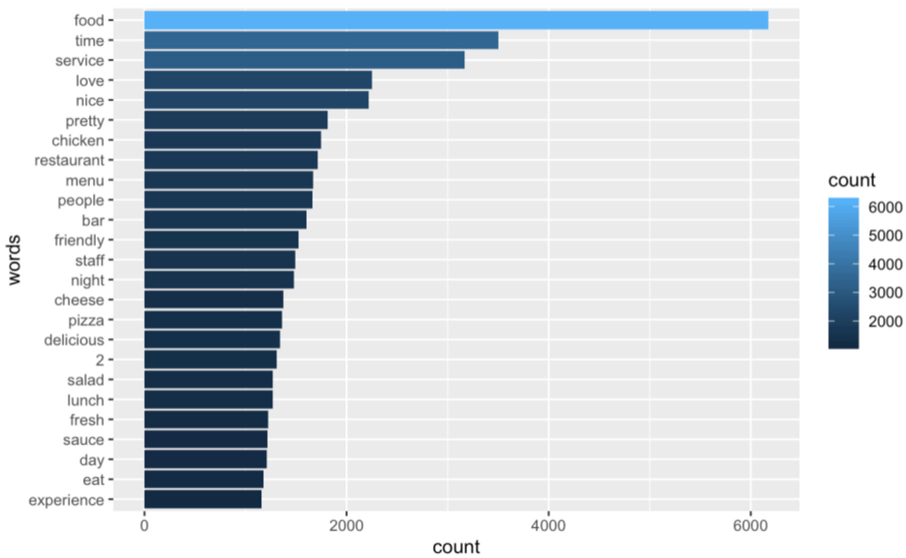
Sentiment Analysis

1. Distribution of review ratings



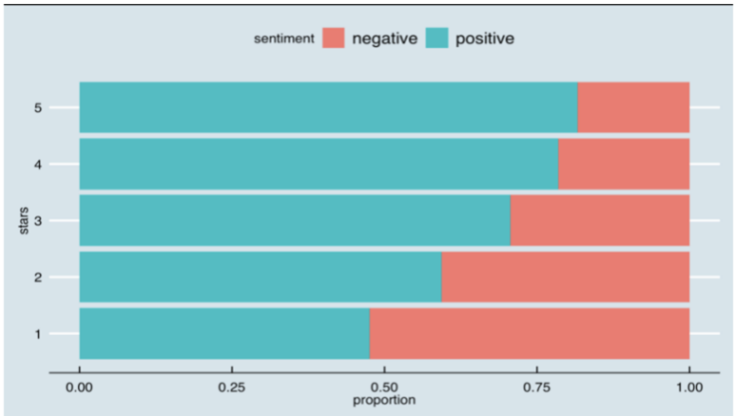
Most ratings are distributed in star 4 and star 5, and star 1 has the least amount of reviews. Which means users in this Yelp dataset are inclined to give positive ratings and they are generally satisfied with food and services.

2. Most common words



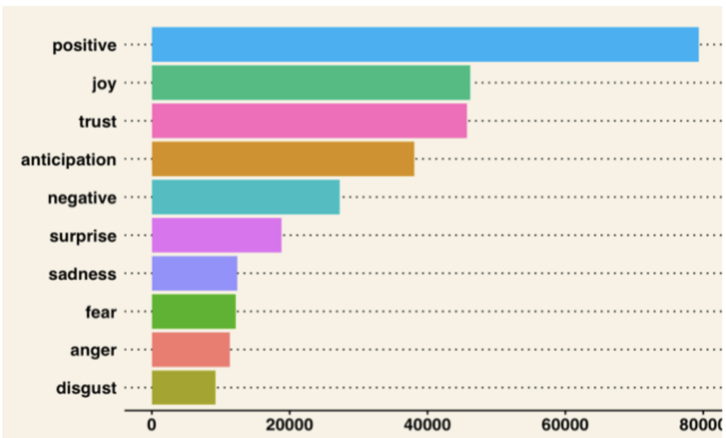
The top 30 common words in the Yelp datasets are shown in the bar chart above, we can see “food”, “time” and “service” are the most common words in the review, which means that customers care about the quality of the food and service the most, and they also value the time spent on waiting and serving foods. In addition, we could find out that chicken is probably the most popular food on the menu, and the second one is pizza. The customers are more likely to get happiness if staff are friendly to them when serving.

3. Sentiment analysis for positive reviews



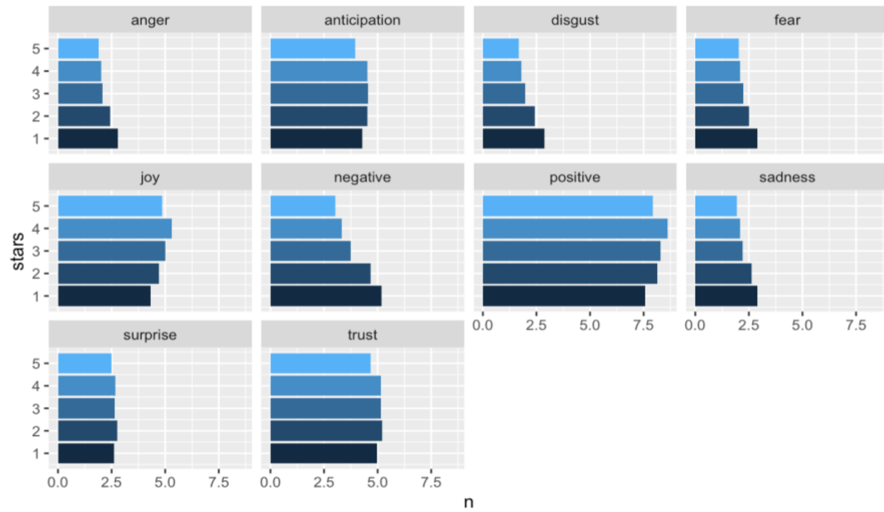
The bar chart of sentiment analysis for reviews generally shows that the higher the rating star is, the more positive words there are in the reviews. In other words, reviews that have a lot of positive words are rated as helpful.

4. Emotions in reviews



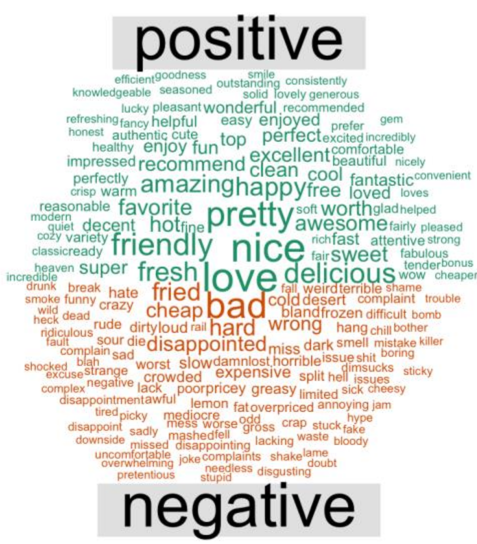
The emotions in reviews are summarized in the bar chart. The most common emotions expressed in the reviews are positive, such as joy, trust, anticipation and surprise. However, a small fraction of reviews are negative, including sadness, fear, anger and disgust. But more negative the words are, the less reviews there are.

5. Ratings of all Reviews based on Emotion Expressed



Here is the relationship between number of words and review rating for each emotion. It is clear that positive reviews are not as sensitive to ratings as negative reviews to ratings. For words like "anticipation", "joy", "surprise" ,and "trust", there is not much difference for the distribution among all rating levels. By contrast, the number of negative words increase as the rating level goes down. For example, 1 star ratings have the highest number of negative words such as "disgust", "fear", "anger" and "sadness".

6. Word cloud



Word clouds tend to be good at capturing interest in non-technical audiences.
Here is a comparison cloud to contrast positive and negative words in the reviews. The results are similar to what most common words in reviews. More customers tend to give positive reviews to services, and they care about food flavor and service quality. The emotions connected with positive reviews are "pretty", "nice and friendly service",

“delicious” ,and “fresh food”. Among all the negative reviews, they are easily disappointed with “fried greasy food”, “cold frozen food” ,and “overpricing”.

Recommendation Analysis

There are multiple types of recommendations. However, we have chosen to proceed with collaborative filtering.

- The data set we used had more number of integer values such as stars, cool review rating, useful review rating. Also the columns such as user_id and business_id beautifully fit into the matrix for performing recommendations

1. Collaborative Filtering
   User Based Collaborative Filtering has been used to identify identical users. In this section, we first converted Dataframe into matrix and used Pearson correlation in the first model. Pearson’s correlation coefficient identifies how closely variables are correlated. Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.

   1.1 Item to Item based Recommendation
   It is a form of collaborative filtering for recommender systems based on the similarity between items calculated using people's ratings of those items
   Below matrix describes similarity among items. As it is evident that item 1 is closely related to item 2 as compared to item 3 and item 4. This means If a user likes item 1, the same user is bound to like item2.

	Item1	Item2	Item3	Item4
Item1	1	0.502	0.12	0.201
Item2	0.502	1	0.75	0.8
Item3	0.1	0.75	1	
Item4	0.2	0.802	0.3	1

2. User to Item Based similarity
   It uses logic to recommend items by finding similar users to the active user. Predicts the rating that user will give to all items the k neighbors have consumed but has not.
   Below matrix describes similarity among users. If user1 and user3 seem to have the same pattern of rating. This means if we know which items are liked by user1 we can recommend similar items to user 3.

	User1	User2	User3
Item1	5	2	5
Item2	2	3	1
Item3	4	2	5
Item4	1	5	1
Item5	2	5	1

3. Examples

   3.1 Recommended restaurants for user7 based on rating

```
44   --65q1FpAL_UQtVZ2PTGew _1TVDGhUKCpTp4qHVQODTA    4.4
45   --65q1FpAL_UQtVZ2PTGew _ISHH8x8Qe6IoVTQWegbpQ    4.4
46   --65q1FpAL_UQtVZ2PTGew _IwgdSjrIMwAcQdeLNhcZQ    4.4
47   --65q1FpAL_UQtVZ2PTGew _jbeTDw6XSuvLaR1V_z8jw    4.4
48   --65q1FpAL_UQtVZ2PTGew _kL4evOcRHHEquKnndUTkg    4.4
49   --65q1FpAL_UQtVZ2PTGew _L5kScT8_QCmFoHV6fmrqg    4.4
50   --65q1FpAL_UQtVZ2PTGew _lanRtzQGRkbhd2b6pQleQ    4.4
```

   The above information represents one user tends to recommend similar services and provides rating based on previous experience.

   3.2 Restaurants that user might like based on previous rating

```
170 -494b8jvjlI07v9HheHhrw 18X496TOlKunSiAnm6z1dg      4
171 -494b8jvjlI07v9HheHhrw 1AMncE6Lxdr6J9PeqZi4nA      4
172 -494b8jvjlI07v9HheHhrw 1Ap6ZNCvyLLKHP0wvCk9yA      4
173 -494b8jvjlI07v9HheHhrw 1CNKe3H07sLu6rigOpAYsg      4
174 -494b8jvjlI07v9HheHhrw 1crzPdwnlm2zHdQ2nP4n7w      4
```

4. <u>Conclusions in recommendation analysis</u>

   There is a close relationship between real and predicted ratings. In addition, reviews and ratings are aligned with each other. In most cases, reviews are influenced by rating, showing that the more positive reviews users usually provided the higher rating they graded. However, there are very few cases where the rating given has been high and reviews are negative or neutral.

Conclusion and Recommendation

There is no better way to understand customers' thoughts and attitudes toward businesses other than read their words, hear their opinions, and communicate and respond to them directly. Our team by using Yelp review datasets to analyze customer satisfaction on local businesses. These results could bring benefits to companies knowing feedback in order to improve services and quality of product, and to boost business sales and revenue.

In the Clustering analysis, the star has a relatively weak relationship among customers' feelings on reviews, popularity of restaurants, and writers' motivations. Based on this situation, we could conduct further analysis to figure out factors that are related to the star grading. In addition, it would be helpful if restaurants could encourage customers to write higher quality reviews that have more credibility. For example, businesses could cooperate with active Yelp reviewers inviting them to the places and ask them to share comments to the public. The reason to do so is popular and high quality words will not only earn more attention but also are more useful, funny and trustful. Similarly in the recommendation analysis, it helps to prove that there is a positive correlation between real and predicted rating. It is important to achieve more positive reviews and higher quality contexts. Higher rating is a factor for customers to consider when they make decisions.

Furthermore, in order to better understand customer satisfaction, it is helpful to do the sentiment analysis. From the top 30 common words in the Yelp datasets, “food”, “time” and “service” are the most common words in the review, showing that customers value the most on the quality of the food and service, and time spent on waiting and serving foods. Also, “chicken” and “pizza” appear frequently in the data. Therefore, in order to attract more customers, restaurants are responsible for having ideas on how to save time while providing high quality products and friendly environments for people to enjoy and relax. There is a relationship between the number of words and review rating for each emotion. Positive reviews are less sensitive to ratings than negative reviews do. Catering to the needs of customers and giving the impression that businesses are constantly innovating could bring positive influences to people. For example, regularly making improvements and changes on menus may make customers more excited about the places.