Yelp Analysis

Collaborators

Name	SID
黄景娟	12112847
胡强	12111214
邱俊杰	12111831

1 Introduction

Yelp Inc. is an American company that develops the Yelp.com website and the Yelp mobile app, which publish crowd-sourced reviews about businesses.[1] It is of considerable value to analysis its data and find out whether restaurant performance is indeed dictated by some factors like their location, density, sentiment and many other components. The purpose of the project is to extract features from the data to summarize some important features for future star-rating prediction if the necessary messages are provided. The result from further analysis may provide a solution for making best decision for opening a shop.

2 Dataset Inspect

The dataset used is retrieved from Yelp Dataset. We have adhered to the Yelp dataset protocol and are not using the analysis for any commercial purposes.

The Yelp **business** data includes the the business id & name, position (state, city, neighborhood, longitude, latitude, zip code, etc.), working time, number of reviews, average star rating, status of business, and category of cuisine offered for a variety of restaurants.

The Yelp **review** data maintains the star rating on a scale of 1 to 5. The necessary information of business & user, review stars, review text are provided in this table. The review can be rated as **cool**, **funny** or **useful**. In this project, reviews of *Reading Terminal Market* are extracted from Yelp dataset to analyze some specific features making graphs like **world cloud diagram**.

The Yelp **user** data contains the user's basic information, average stars as well as the **elite status** of user, which can help distinguish qualified data by extracting elite user's record. When conducting our research, we also used the user data to find the relationship between the stars elite user commented on one popular shop and the stars they averagely give.

3 Data Analysis

Topics We Are Going to Discuss:

- Broad Analysis

  1. Distribution Pattern of Restaurants
  2. Adjacent Relationship of Restaurants
  3. How to Evaluate the Popularity of a Restaurant
  4. Choose a Best Restaurant etc.

- Specific Analysis for a Specific Restaurant:

    1. Review Words Analysis (NLP)
    2. Stars Distribution Analysis
    3. Correlation Between Star and Other Variables
    4. Elites' Impact On Friends etc.

After an initial screening of the Yelp data, we selected the most popular restaurant industries as the research subjects.

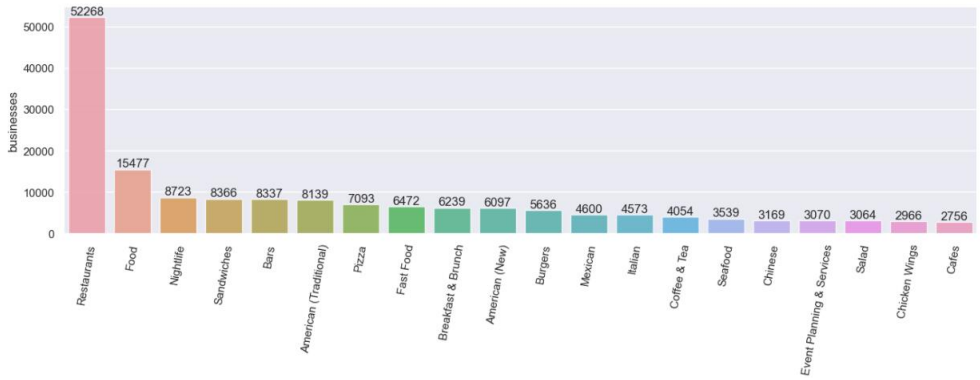

Figure 1: The Rank of Industrial Category

Figure 1 shows the business number of different kind of industrial category from Yelp dataset. The y-axis is the exact number of each category.

To better describe the popularity of a restaurant, we used the definition of popularity that is widely used in Kaggle

$$popularity = review\ count \times average\ stars$$

3.1 Broad Analysis

3.1.1 The Distribution for Restaurants in Each State

We group table business by states and count the sum of each restaurant's popularity. And then show it on the map using **pyecharts.charts.Map**.

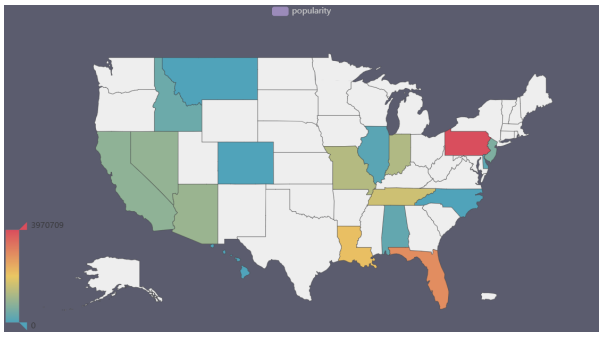

Figure 2: The Restaurants Business Distribution of Yelp Dataset

Besides, we also plot of scatterplot to found the possible relation between **count** and **average popularity** in each states, and found that there's no apparent relation between count and average popularity(correlation coefficient is only **0.21**).

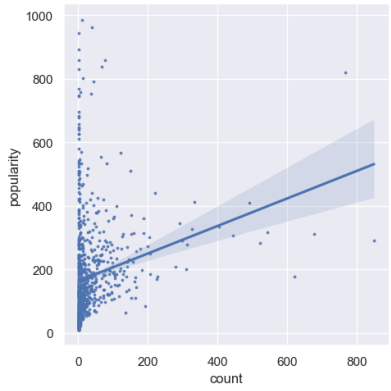

Figure 3: The Relationship Between Business Count and Popularity

Figure 3 shows the the relationship between business count and popularity for each business. (correlation coefficient = 0.21)

3.1.2 The Distribution for Each Restaurant in America

we use the latitude and longitude of each restaurant and plot a scatterplot on the map using **plotly.express.density_mapbox**, from which we could discover that the restaurants almost concentrate in few urban agglomerations and fit the traffic smashes, which is radially distributed.

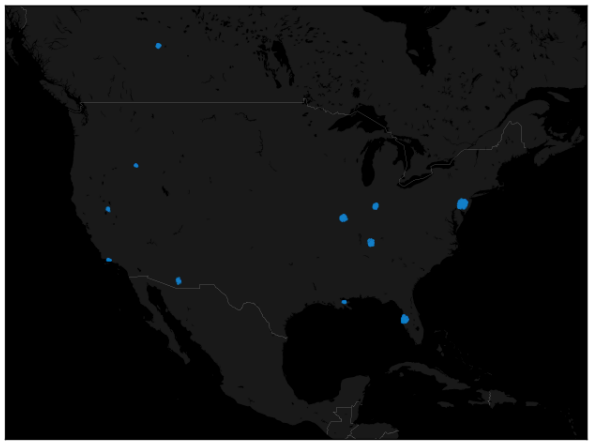

Figure 3: Domestic View

Figure 4: Distribution of Restaurants on Satellite Image

3.1.3 The Adjacent Count vs. Popularity for Each Restaurant In Top5 Cities

We calculate the number of restaurants situated in 500m from each restaurant as column adj_count, and plot a scatterplot to see the relationship between the adj_count and popularity for top5 cities ordered by total popularity(to relieve the calculation).

Analysis shows that the average popularity is increased with the adj_count,which means the more concentratedly where the restaurant situate, the more likely it has a higher popularity.

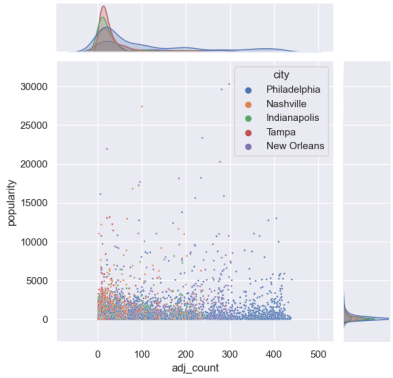

Figure 5: The Relationship between Adjacent Count and Popularity of Top5 Cities

Also,we found that restaurants of the most highest popularities is more likely to situated in the place where restaurants is sparsely distributed.

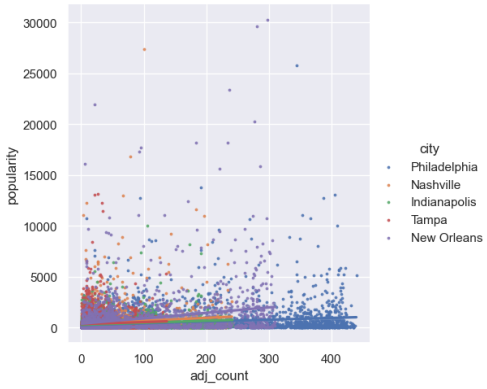

Figure 6: The Regression between Adjacent Count and Popularity of Top5 Cities

3.1.4 The Distribution of Popularity for Top10 Cities

We select out the restaurants in top10 cities, and plot the **kdeplot**, **violinplot** and boxplot of restaurants in top10 cities. the plots shows that the top10 cities are of similar distribution after log transformation(the medians are all between 2 and 3 and the shape are similar).

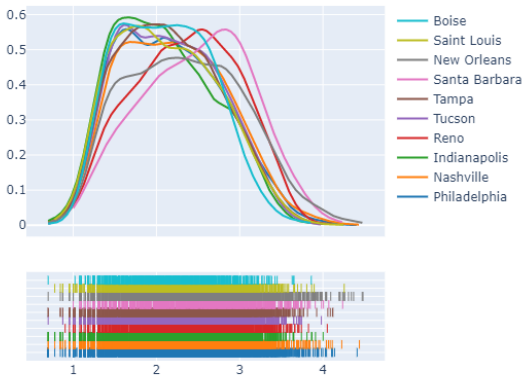

Figure 7: The Distribution of Popularity with Respect to Top10 Cities

3.1.5 Reflect: The Relationship between Review Count and Stars for Restaurant

As mentioned above, we use **review_count*stars** to define the popularity of a store. But wouldn't that be inappropriate? For example, stars are not actually completely separate from review count. Therefore, we studied the relationship between stars and review count, and found that there is a positive correlation between them.

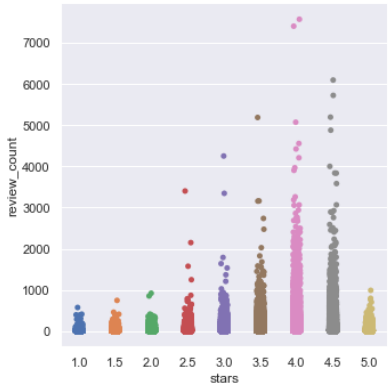

Figure 8: Distribution of Review Count with Respect to Each Stars

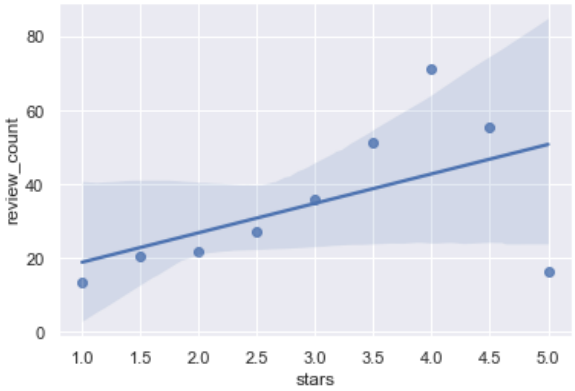

Figure 9: Relationship of Average Review Count and Stars

Although such indicators are widely used, there will be random variable dependence problems in defining popularity in this way, and we can continue to explore how to optimize the evaluation formula in the future.

3.2 Specific Analysis

In previous research, we retrieved the most popular restaurant from the following steps:

1. Find the city with the highest cumulative popularity.
2. Find the most popular restaurant in the city.

The result is **"Reading Terminal Market"** in Philadelphia. Around it, we conducted a specific analysis shown below.

3.2.1 High Frequency Words Of Review

We use Natural Language Processing to find high frequency words of review and make visualizations. First, use regular expression matching to remove the elements that are not letters like symbols and turn the uppercase letters into lowercase. Second, count the quantity of each words and sort them in ascending order. Third, use the **stopwords** which do not make sense for analyzing and delete them. Last, use **WordCloud** to draw a word cloud diagram which the size of words follow the frequency of words and count the most popular food in review.

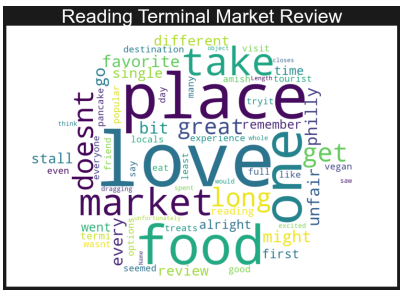

Figure 10: Word Cloud Diagram

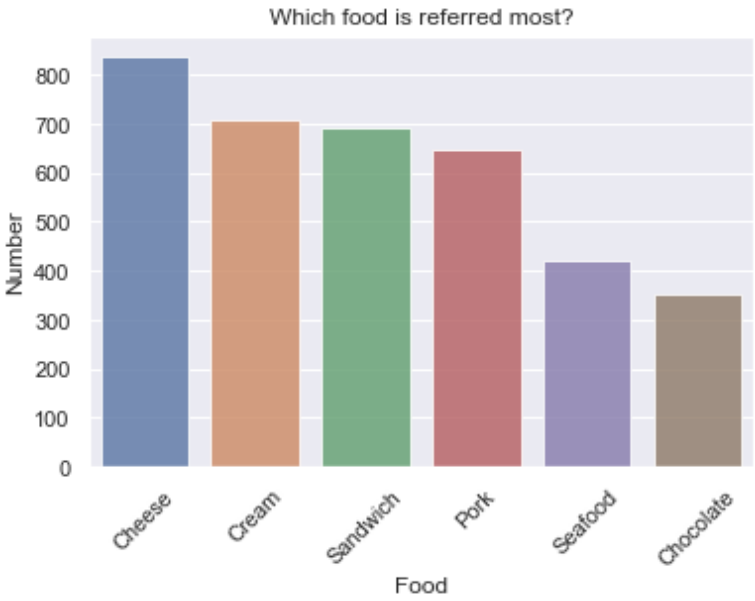

Figure 11: Popular Food

The function of word cloud diagram is to show the high frequency words and let readers catch key words quickly. From Figure 10, the first word coming to our visual field is "love", and we can also see "favorite", "great" which mean the positive attitudes of users. That is to say, most of the users love the food in this restaurant. And it also shows the attribute of the business, using "food", "market", "eat". The words "remember", "many", "long" show the impressions of users. And words "unfair", "doesn‘t", "isn’t" show there might be some problems of the place. From Figure 11, it is clear that people like to eat cheese, cream, sandwich and so on in the market.

3.2.2 Stars And Numbers Of Comments

We count the number of stars in all years and make a bar chart.

Figure 12: Average Stars And Counts Of Review

We calculate the average stars and combine a line chart of average stars as years' change with a bar chart of counts of review.

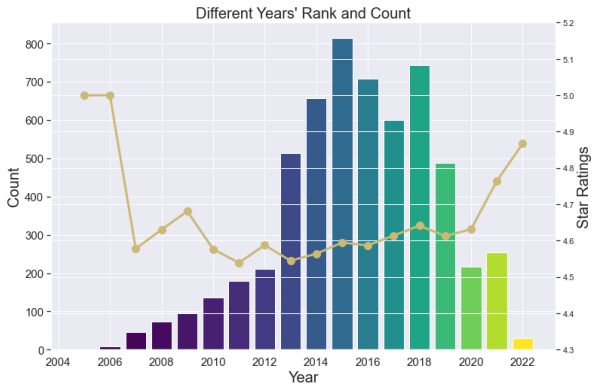

Figure 13: Average Stars And Counts Of Review

We paint a **catplot** and a **jointplot** of stars in this business and average stars by elite.

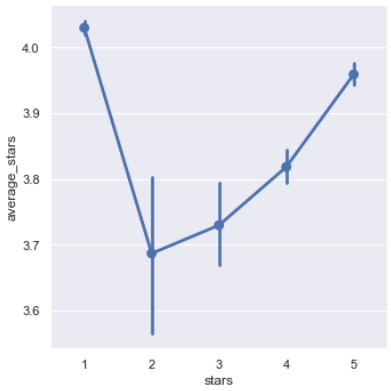

Figure 14: Catplot Of Stars And Average Stars

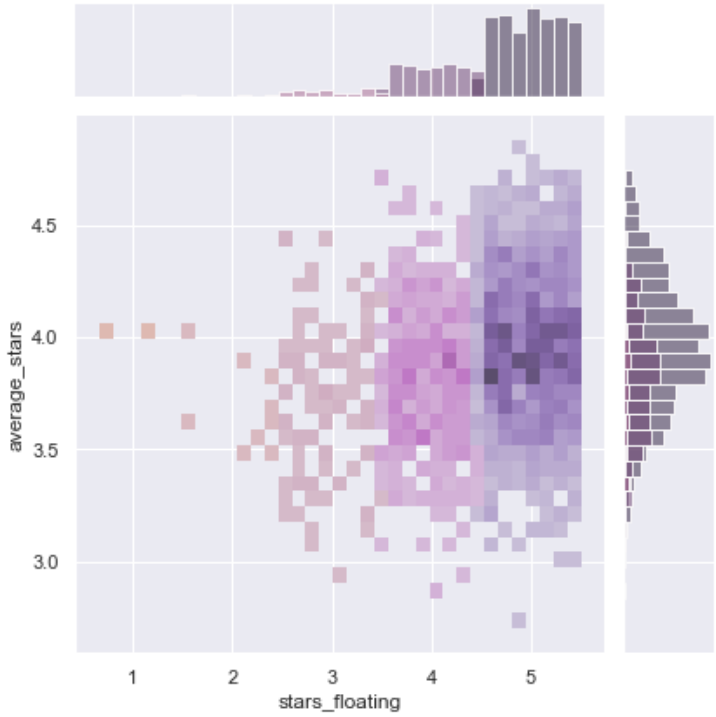

Figure 15: Jointplot Of Stars And Average Stars

From Figure 12, users tend to give 4 or 5 stars to the store. From Figure 13, the average stars given by users is range from 4.5 to 5 and steady. The counts of review are increasing firstly and then decreasing. Notice that in 2005 and 2005, the average star is 5, but these two points do not have reference value because the number of review is too few. From Figure 15, we can see the average stars by elites follows a normal distribution but they tend to give higher stars to the store. From Figure 14 and Figure 15, we can find that a user's usual scoring habits will affect its scoring habits for this restaurant, if he usually scores high, then it is also more inclined to give this restaurant a high score, which indicates that the user's scoring is habitual, not arbitrary. So the rating of a store by individual users is affected by their own scoring habits. To sum up, the store is of great popularity and people tend to give higher rating.

3.2.3 Correlation

We make a **heapmap** of stars, useful, funny, cool, text_count and weekday. And also, make a **pairplot** of the varibles using stars as hue. Among the varibles, **useful**, **funny**, **cool** refer to the evaluation of other people who see the review and make their attitudes, as you can see in Figure 16.

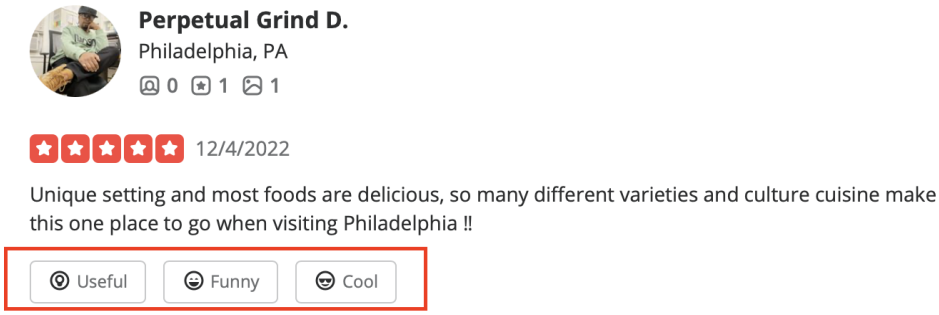

Figure 16: Explanation

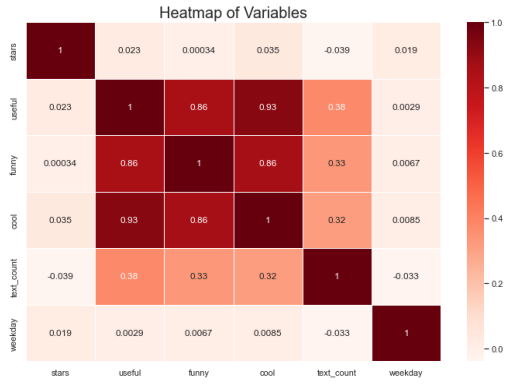

Figure 17: Heapmap Of Correlation

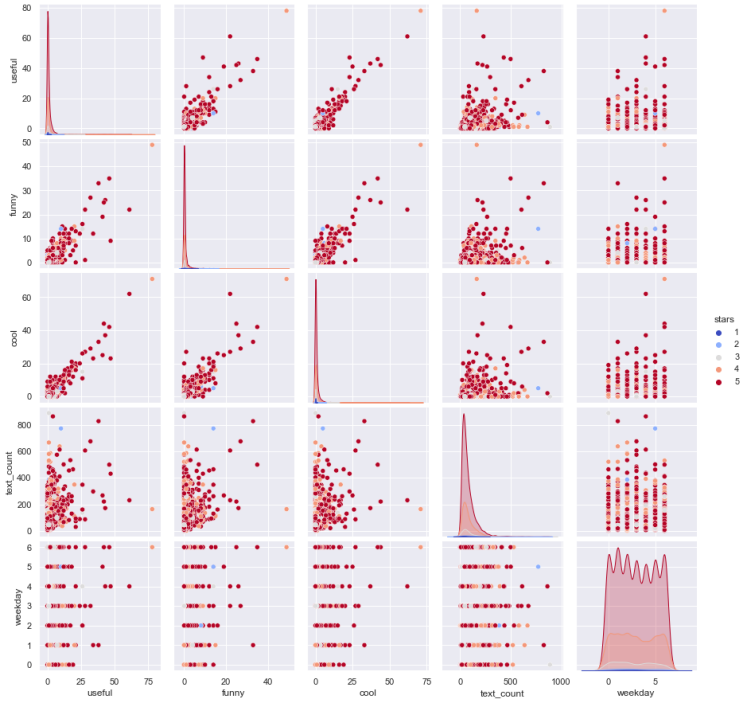

Figure 18: Pairplot Of Correlation

We make a **boxplot** of correalation between stars and text_count.

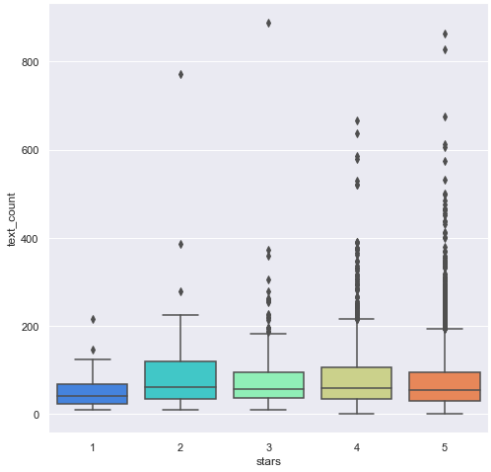

Figure 19: Boxplot Of Correlation

From Figure 17, Figure 18 and Figure 19, **useful**, **funny**, **cool** have strong linear relationship because they are all positive compliment of review. But other varibles have weak relationship. That is to say, the length of review do not influence the stars. No matter which day to go to the store, the stars are given stably.

3.2.4 Elites' Impact On Friends

We import the data of elites who have gone to the store and their friend into Database, and then extract data of the nodes which contains the elites and their friends and whether they have gone to the store, the edges which contains the elites and their friends. Then we import the nodes and edges into **Gephi**. Next, use the layout of Yifan Hu to disperse nodes. In Figure 20, we modularize the nodes. In Figure 21, we paint green to represent the person have gone to the store and pink to represent the person have not.

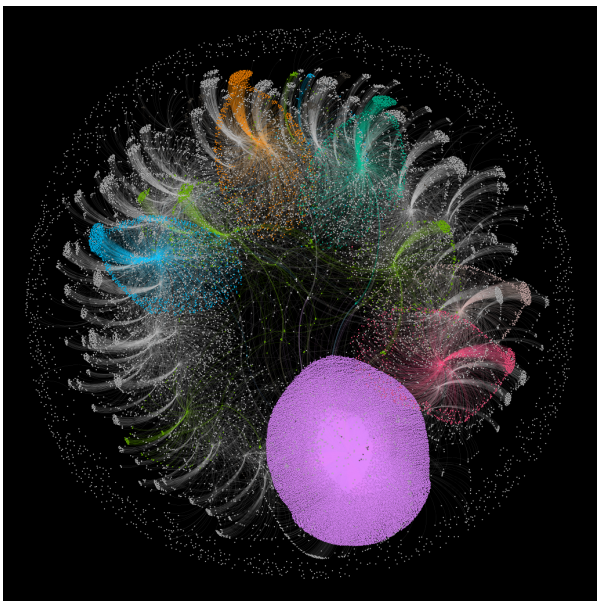

Figure 20: Modularization Of Elites And Friends

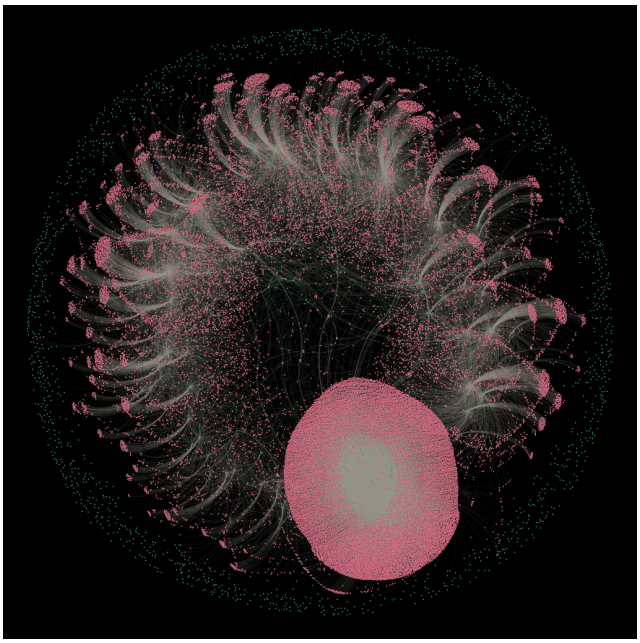

Figure 21: Whether go or not

From Figure 20, it is easy to see that some elites surrounded by many nodes have a lot of friends but some elites are isolated points. From Figure 21, most of nodes are pink and most of the green nodes are surrounded by pink nodes. Hence, if one has gone to the store, his friends are not influenced by the person and they do not go to the store.

4 Conclusion

Final conclusions:

- Specific Analysis:

  1. For distribution in each state, costal more than continental(vacant in many state), eastern more than western.
  2. For distribution of each restaurants, they all highly concentrate in limited area(city group). In city group, they almost fit traffic nets.
  3. When restaurants concentrate, the average popularity increase, but the highest popularity decrease.
  4. The distribution of popularity in top 10 cities are similar, almost situate between 100 and 1000, and Santa Barbara has the highest average popularity, Philadelphia has the highest total popularity.

- Broad Analysis:

  1. Word Cloud Diagram reflects the restaurant's attributes and problems, customers' love and impressions. The restaurant sells cheese, cream, sandwich, etc.
  2. No matter seeing the stars from total, average stars of different years or stars given by elites, all of them think highly of the restaurant and stars are between 4 to 5.
  3. Stars have little correlation with useful, funny, cool, text_count and weekday. Useful, funny, cool have strong positive correlation.
  4. Elites who have gone to the restaurant have little impact on their friends.

Reference

1. Wikipedia contributors. (2022, December 13). *Yelp*. Wikipedia. https://en.wikipedia.org/wiki/Yelp ↩