Outlier Detection of User behavior on Yelp Data

Find out the customer behavioral outliers to better understand customer behavior in Yelp and define their difference with normal customers in specific aspects.

Jie Dong: jd3191
Lei Sun: ls3816
Enze Wu: ew1238

Project page (on Github): https://github.com/nyu-cs6313-fall2015/Group-9
Video: https://vimeo.com/149665108
Working demo: http://nyu-cs6313-fall2015.github.io/Group-9/index

1. **What is the problem you want to solve and who has this problem?**

Reviews in Yelp play important role in making purchase decisions and help people to narrow down the options and to make decision based on their needs. And from business point of view, positive reviews can result in significant financial benefits. But this also provides opportunities for deceptions, where fake reviews can be generated to garner positive opinions about a product or to disrepute some businesses. Our project focuses on understanding user behaviors on Yelp through explore the datasets given by Yelp. What we will do is to help others define the outliers in our visualization tool as well as their differences. There are some possible applications. For example, help Yelp to better understand about their users; help customers know the reviews written by whom are untrustable, and help business owners to better understand how they are being evaluated and maybe create actions to improve their ratings in Yelp. Thus, our visualization tool can help Yelp company, common customers and business owners to solve their different problems.

2. **What questions do you want to be able to answer with your visualization?**

Question 1: How to recognize outliers from all customers?
We found customers with similar behavior via T-sne algorithm.

Question 2: In each attribute, what is the difference between outliers and normal customers?
We analyzed the outliers in 10 attributes via candlesticks chart to define their maximum, minimum and average. Specifically, we visualized the following attributes between selected outlier and all customers:

- The review total numbers;
- The average star distribution;
- The time of Yelp using;
- The amount of friends;
- The numbers of reviewed business;
- The average amount of reviews that written for different business;
- The average length of reviews;
- The proportion of review (be voted cool, funny and useful respectively)

3. **What is your data about? Where does it come from? What attributes are you going to use? What is their meaning? What are their attribute types (data abstraction)? Do you plan to generate derived attributes? If yes, which and why?**

We get a series of datasets from Yelp. But in our project, we will only use the datasets that are related to user and review. These datasets come from Yelp Challenge website.
In order to find outliers, we chose some features to characterize user behavior and create some vectors to measure distance between users. Based on the attributes in our dataset, we come up with several features described as following table:

attribute names	attribute type	description	range
uname	categories	name of a user	string
user_id	categories	user id of a user	text
review_count	quantitative	the amount of review of a user	1-8433
avg_stars	quantitative	the average rating star of a user	0.00-5.00
how_long_month (derived)	quantitative	the amount of month the user has used yelp	1-124
friends (derived)	quantitative	the amount of friends a user has	0-3830
num_business (derived)	quantitative	the amount of business the user has written reviews on	1-1259
review_per_bus (derived)	quantitative	the average amount of review that user gives to different kinds of business	1.00-7.00
pro_cool (derived)	quantitative	the proportion of review (be voted cool); count of cool votes /over all review count	0.00-151.50
pro_funny	quantitative	the proportion of review (be voted funny)	0.00-

(derived)		count of funny votes/over all review count	265.67
pro_useful (derived)	quantitative	the proportion of review (be voted useful) count of useful votes/over all review count	0.00-459.00
text_length (derived)	quantitative	the average length of text in user's reviews	1-5000

4. **What have others done to solve this or related problems?**

· **Unsupervised sentiment classification of English movie reviews using automatic selection of positive and native**

In this paper, the author examined an unsupervised system of iteratively extracting positive and negative sentiment items which can be used to classify documents. They continued research based on previous paper, but in a different domain which is English movie review. The object they have dealt with is similar to ours, which is review. And their goal is also similar to us. We can refer to their method when we want to derive attribute from the original dataset to get the attitude of the review.

· **One submission of 2014 Yelp Dataset Challenge by the students from UCSD**

In their data analysis, they determine the difficulty in predicting user's review stars given the reviews they left as well as providing a classification model, and they also add in a few visualizations to explain their purpose. We can learn a lot from their submission, though we are exploring data from a totally different direction. For example, how they process data, the method they use in their predictive analytics task. Particularly, they mention a kind of machine learning algorithm called Naive Bayes classifier, which uses the positive and negative word rate as conditional probabilities/relative frequencies in its calculation. Naive Bayes classifier seems a alternative way for us to calculate the probability of how positive or negative a review will be. We need to search and discuss more about this later.

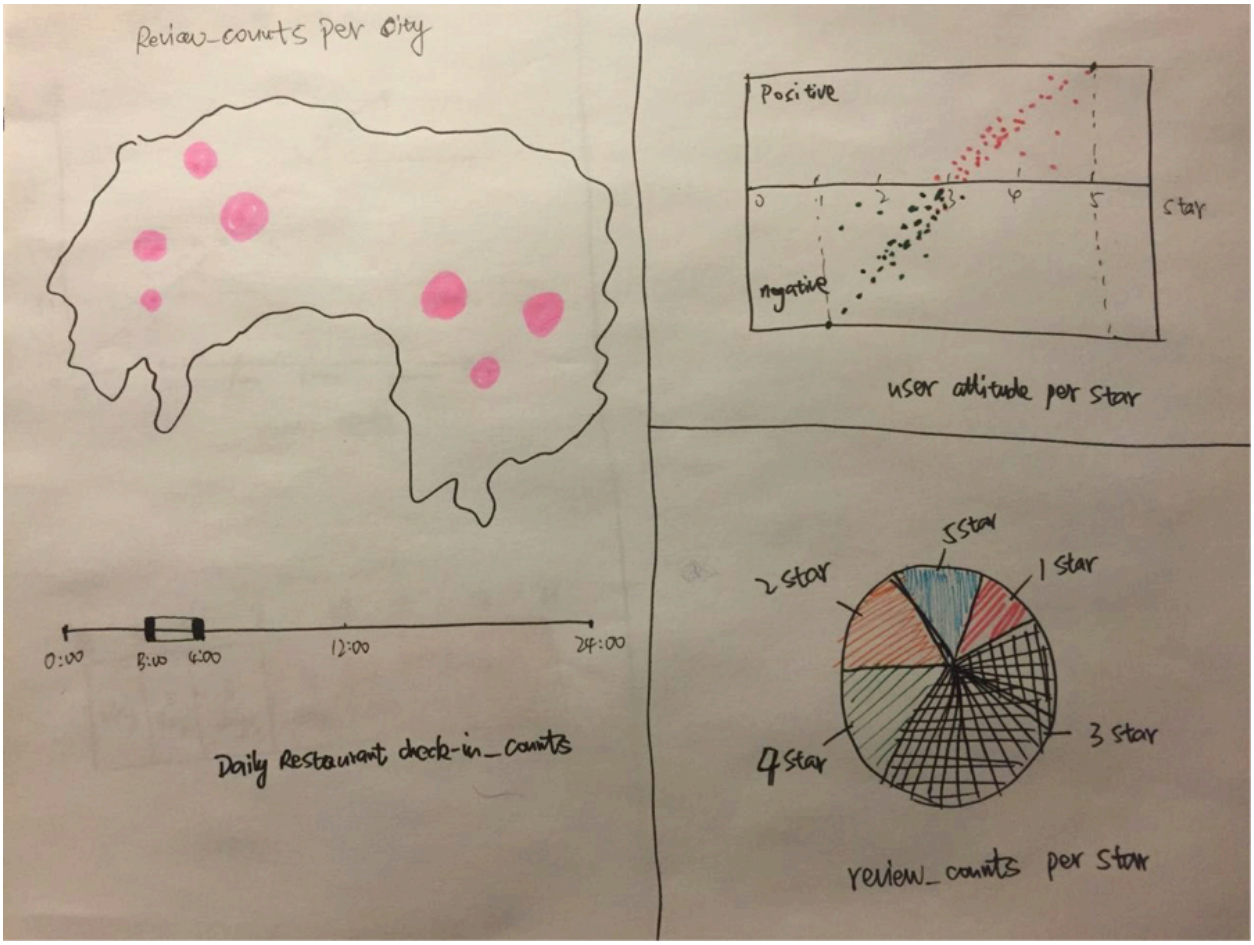
· **Visualizing Yelp Ratings: Interactive Analysis and Comparison of Businesses**

Aid business owners in understanding how their performance compares with other businesses of similar type category or location, to make more informed decisions that could improve their success.

· **Outlier Detection in Bus Routes of Rio de Janeiro from spring 2015**

This topic is also about outlier detection, though we focus on different domains. What we can learn from their work is that how to proceed our project, for example, the way they tried to deal with their dataset. Firstly, we know their method cannot apply to ours. Their dataset is labeled, while ours is not. However, they find some feature vectors to characterize bus routes, which can guide us to search for some features to characterize customer behavior.

5. Initial Mockup



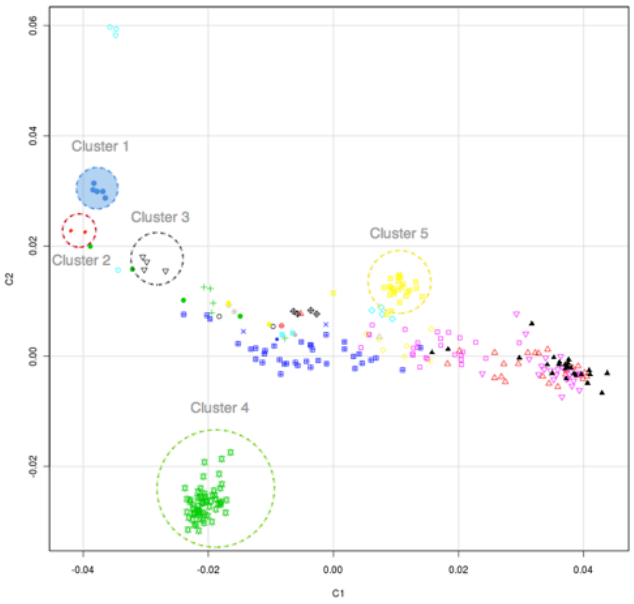
· On the left hand, this visualization shows the change of review amount in an area. You can change the time bar below to get different time period. The red dots on the map show the amount of reviews in each specific area.

· On top of the right hand. The visualization shows the relation between rating stars and positive/negative probability .the probability is derived from data set by machine leaning .Red dots shows review may be positive, while black dots indicate negative. It can help us find some abnormal behavior. For example, review tends to be positive, but the rating stars are low.

6. Project Update

Yelp User Behavior Outlier Detection

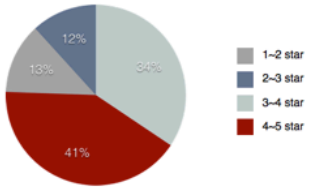
Outlier Groups: Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5



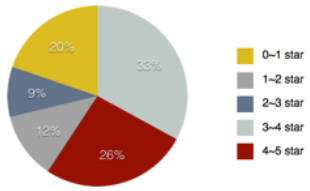
Feature Comparison Filter

Click following features to compare the normal data distribution and the outlier you selected in left graph.

Ave rating | Votes | Num | Blank | Blank



Normal user's Ave rating star distribution



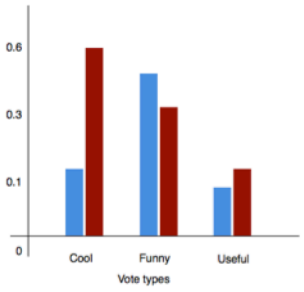
Cluster 1 user's Ave rating star distribution

Feature Comparison Filter

Click following features to compare the normal data distribution and the outlier you selected in left graph.

Ave rating | Votes | Num | Blank | Blank

Votes proportion comparison



Normal user's three vote types proportion

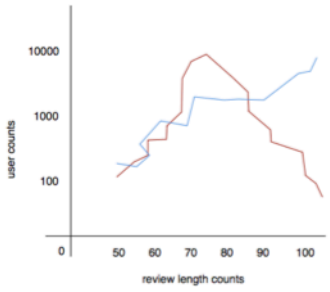
Cluster 1 user's three vote types proportion

Feature Comparison Filter

Click following features to compare the normal data distribution and the outlier you selected in left graph.

Ave rating | Votes | Num | Review length | Blank

Review length distribution



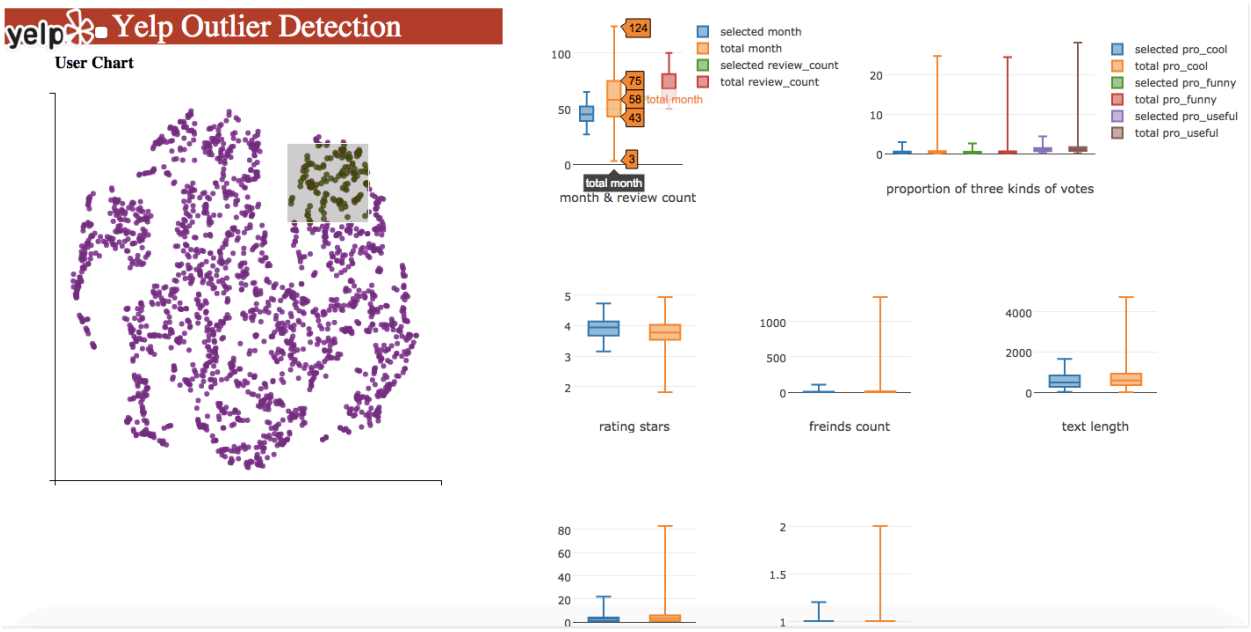
Normal user's review length distribution

Cluster 1 user's review length distribution

The left scatter diagram represents the result of outlier detection which aggregated as 5 groups from cluster-1 to cluster-5. The user can select any of them to see the attribute difference in the right diagram. For example, as the following screen shot, after select cluster-1 in the top left, then if the user want to see the average rating distribution between cluster-1 and all customers, he/she can just click Ave rating button to see the pie chat underneath the button.

As you can see, this data visualization is totally different with our initial mockup. After discussed with our advisor, we found that our initial idea went on the wrong way that was only the statistics work rather than analysis work. What we did was to find some trends which were not problems for any people. Thus, our new solution adopted a machine learning method to detect outliers which would be pretty helpful for Yelp and stakeholders.

7. Final Visualization



1. on the left side select a group of users you are interested in
2. on the right side, ten features of the selected will be compared with the overall user's by using ten pairs of box plot.
3. You can put mouse over the box plot, then it will show the maximum ,75%,median ,25% and the minimum data of a feature. And you can compare they one by one.

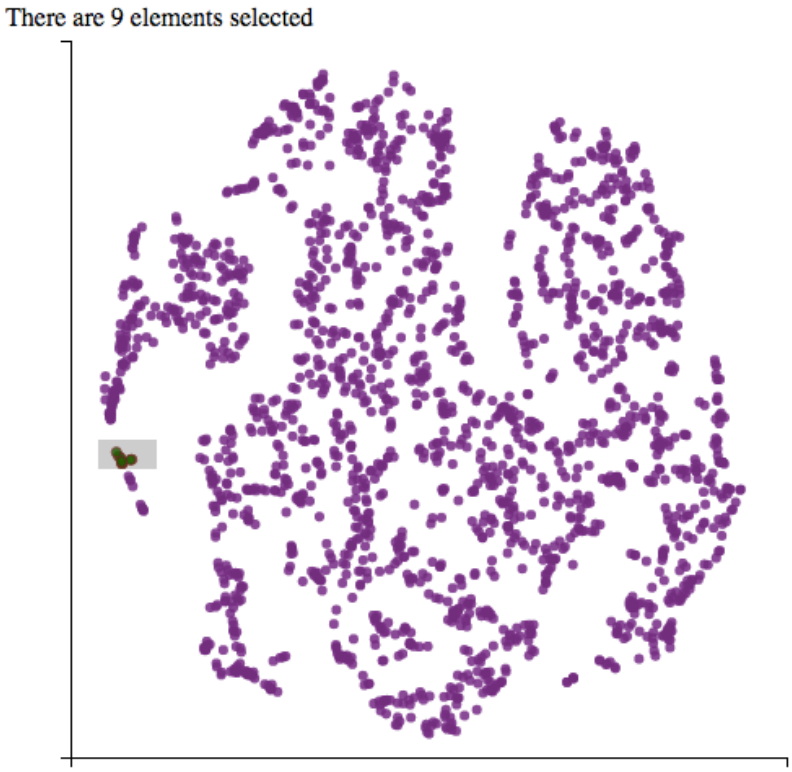
How and why it changed from your project update?

We show all ten features' box plot on the right side of our screen instead of displaying them by using button select them one by one. Because it can improve the space use and it will be easy to get more information at the same time and may discover some relations between these ten features

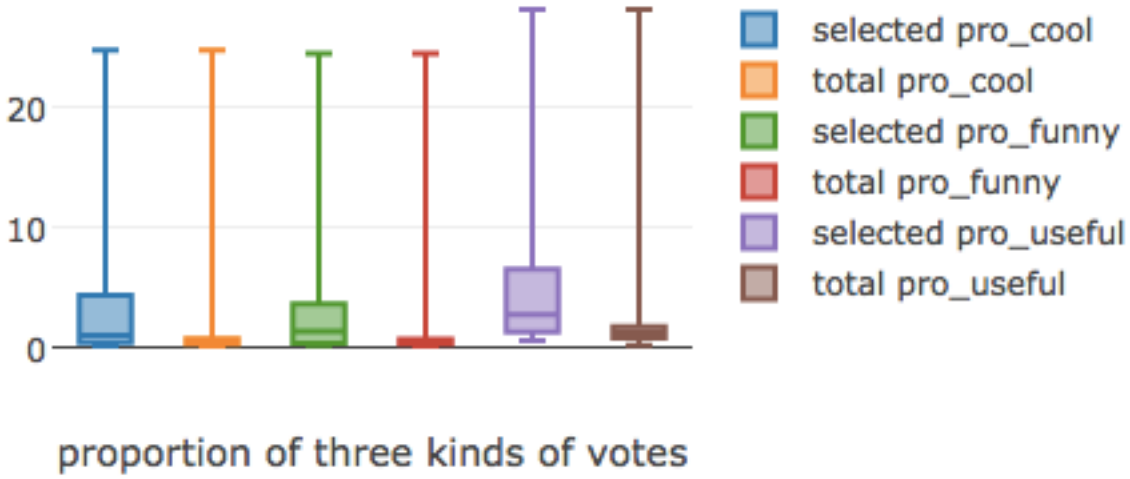
8. Data Analysis

interesting group1:

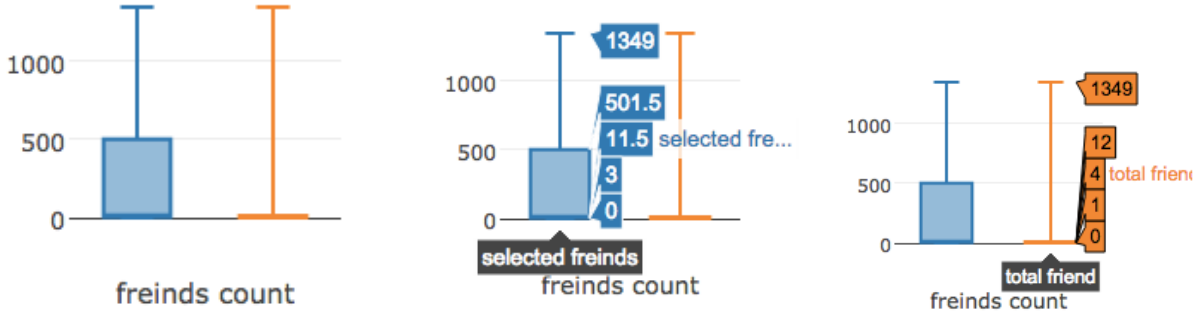
1.First, we'll show a small group of people may be "elite" users(elite means they are active in yelp and their review may be very helpful)



we select a small group in the location on the above picture.
Then we examine 4 features of them (proportion of three kinds of votes, friends count)

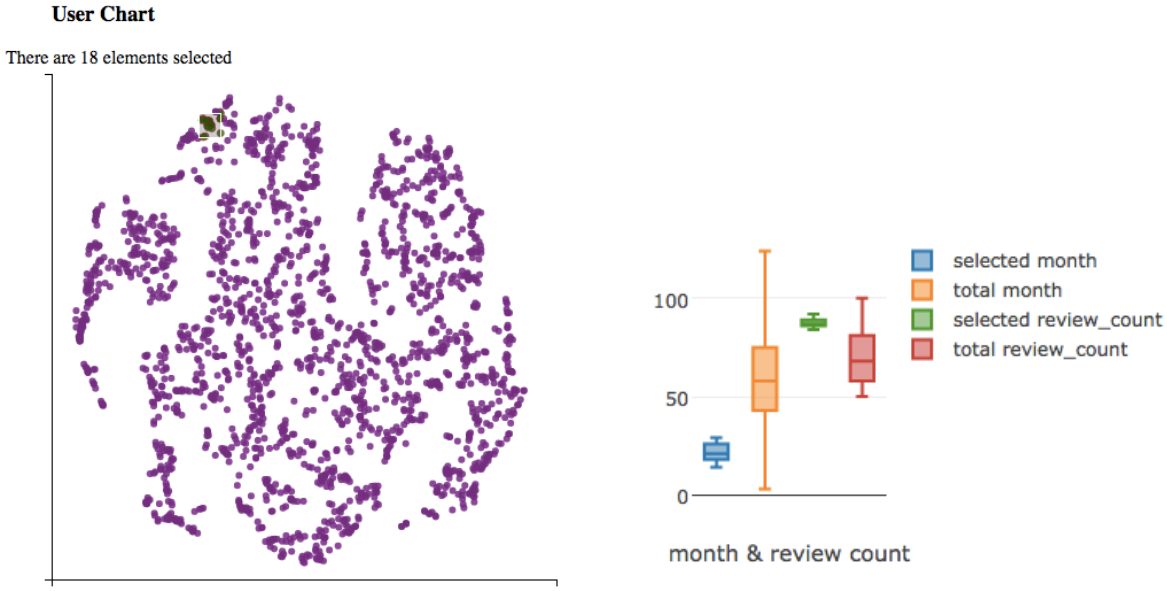


In this picture ,we can see this group has a much higher proportion than overall users. This means every review they write will get more votes(cool,funny,useful) from others. And what is interesting is that, the maximum proportion of the overall user is in this group.

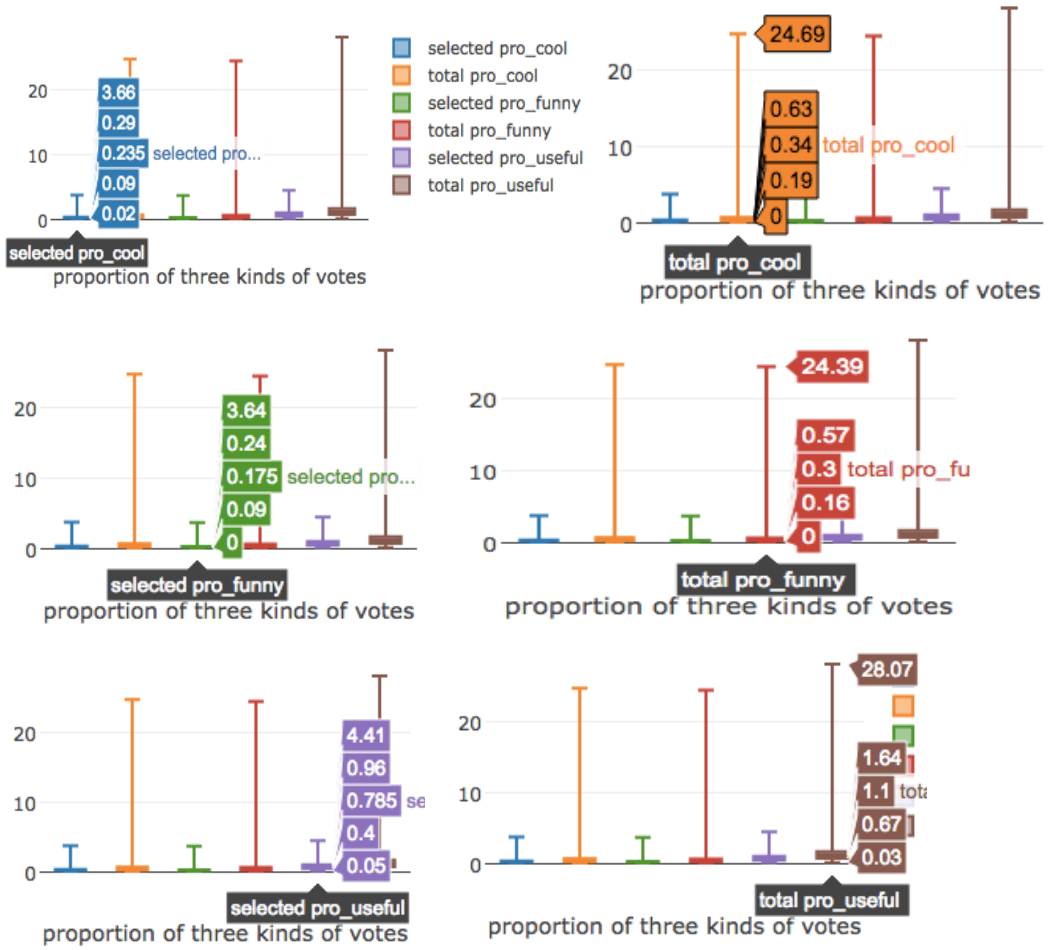


As we can see, member in this group has more friends ,and the user who has maximum friends count is in this group!
In conclusion, we suspect this group is kind of “elite” . Their review is highly agreed by others and they have much more friends, which means they are trustworthy.

Interesting group 2:
Another interesting group ,we may suspect it “spam” users since they have large amount of review in short time and their review may be useless.



In month & review count comparison box plot. We discover that selected group has shorter time use yelp than overall group. But their review amount is much higher than overall group



The selected group gets a lower proportion of votes than overall users.
In conclusion, we deduce this group may be “spam” users. They give more review in short time and their review gets less votes from others.

9. Limitations and Future Works

Limatation1: In some features plot like proportion of votes , friend amount, it is difficult for user to see the box plot clearly since they has very large maximum numbers.

Limatation2: It is hard for a user to examine some suspicious individuals in a group.

Further steps:

We want to add a list below the user chart to show detail information of every user in selected group so we can examine individual we are interested in.