Northeastern University, Boston, MA
Group 1

RECOMMENDATION SYSTEM: UPDATE 1

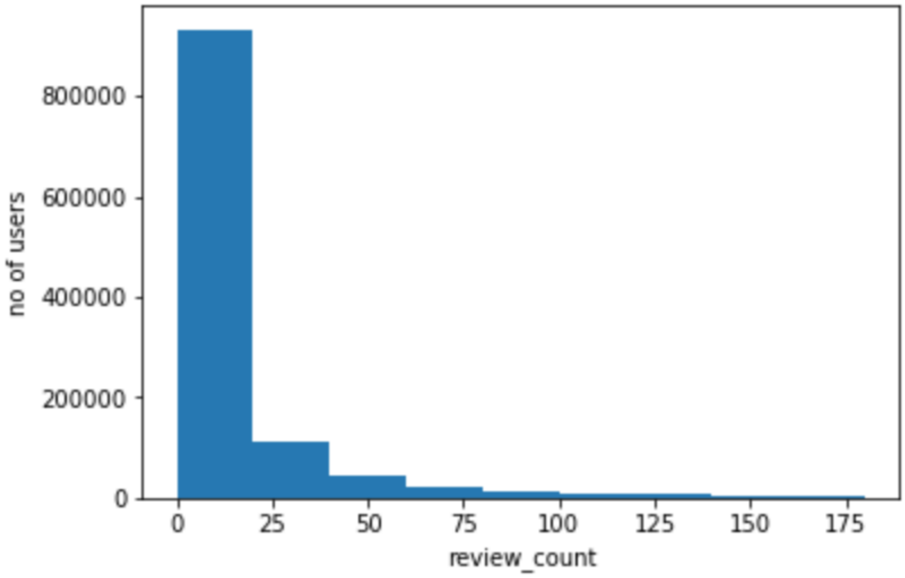
Introduction:

We, the members of group 1 are building a recommendation system on the yelp dataset. We are using three techniques to build this recommendation system namely collaborative filtering, clustering and deep learning. We are not using textual or pictorial data from the dataset for our analysis. We are doing analysis strictly based on star rating which is of type float. Thus we are not using the following files from dataset: review.json, photos.json and tips.json. Currently we have cleaned, preprocessed and performed analysis on the three data files namely: user.json, business.json and checkin.json.

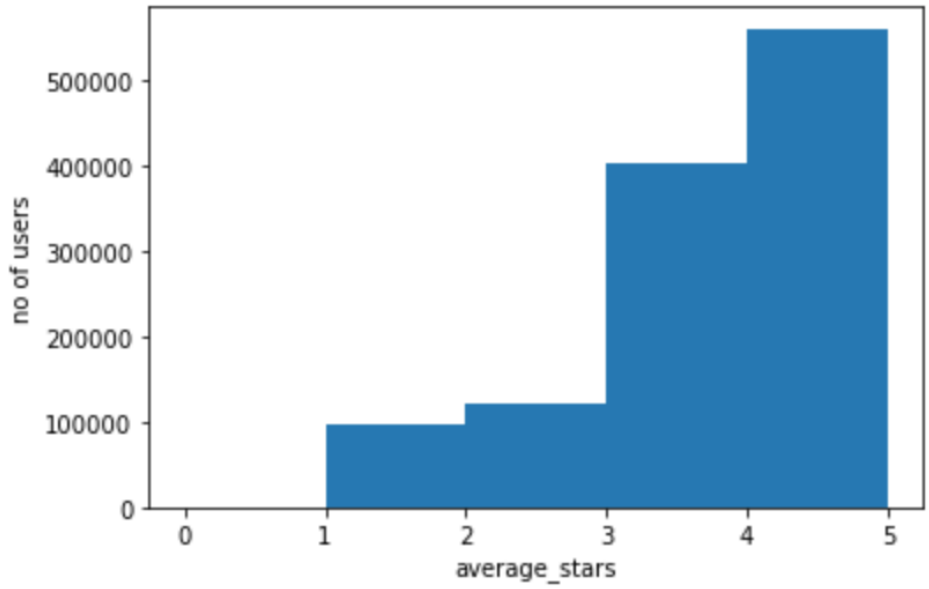
Current Progress:

1) For user.json

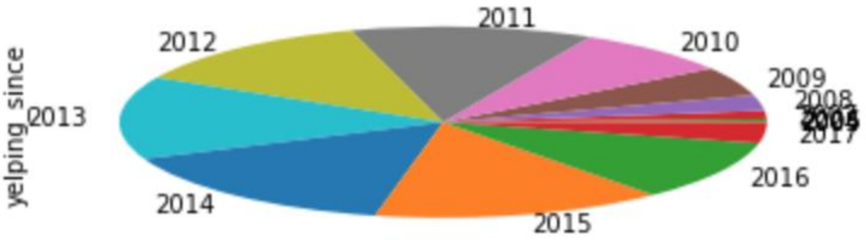
For user.json we first calculated total number of users in our dataset which is 1183362. Then we calculated review count for each user and plotted a histogram shown below:

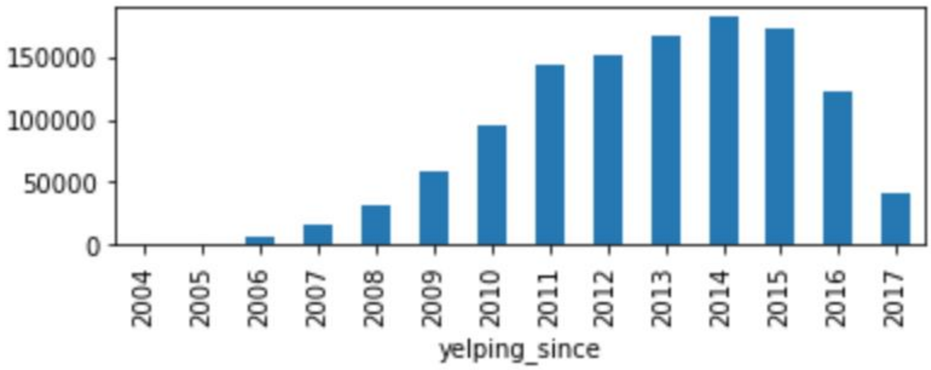


As seen from the histogram majority of the users have given about 25 or less reviews. When we calculate the mean and standard deviation of the given plot we get 23.72 and 80.5 respectively. A higher standard deviation indicates that the data points are far part from each other and spread over a wider range. The maximum number of reviews given by any user is 11656. After that we proceeded to find average rating given by each user. Again, this in float and no textual data is used. We found that about 232987 users have given 5-star rating for each business they have reviewed. Thus, these can be considered as non-trusting users for the purposes of analysis of ratings. Below is a plot to show average rating with number of users:



When we take median of the above plot we get 3.89. Mean and standard deviation for the same are 3.71 and 1.10 respectively. In order to group the reviews as positive, average and negative reviews, we have used the above statistics. We assume that if the rating lies in the range of (mean – standard deviation, mean) which is 2.6 to 3.7 we will categorize the review as average. Review lower than 2.6 will be considered as a negative review and anything greater than 3.7 will be considered as positive reviews with two extremes being 0 and 1. Next we found out data of users joining yelp. Below are the plots to provide a pictorial representation of the data.

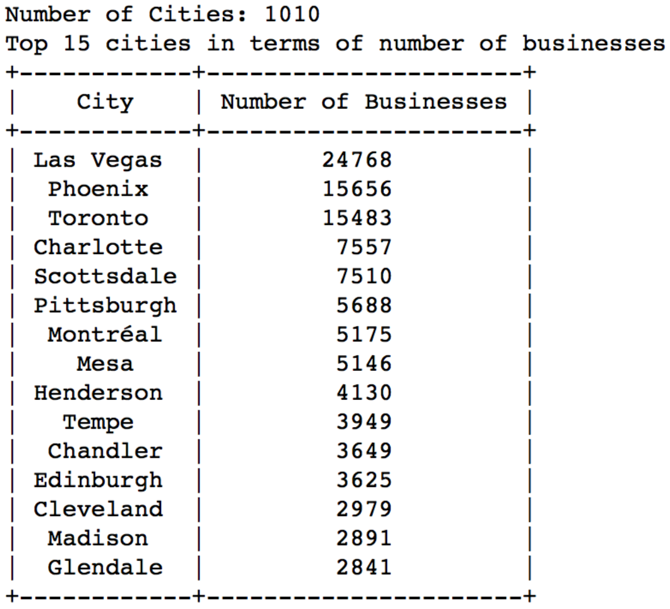




It can be seen from the graph above that there is an increase in users joining yelp from 2005 to 2014 and then there is a continuous dip in number of users joining yelp.

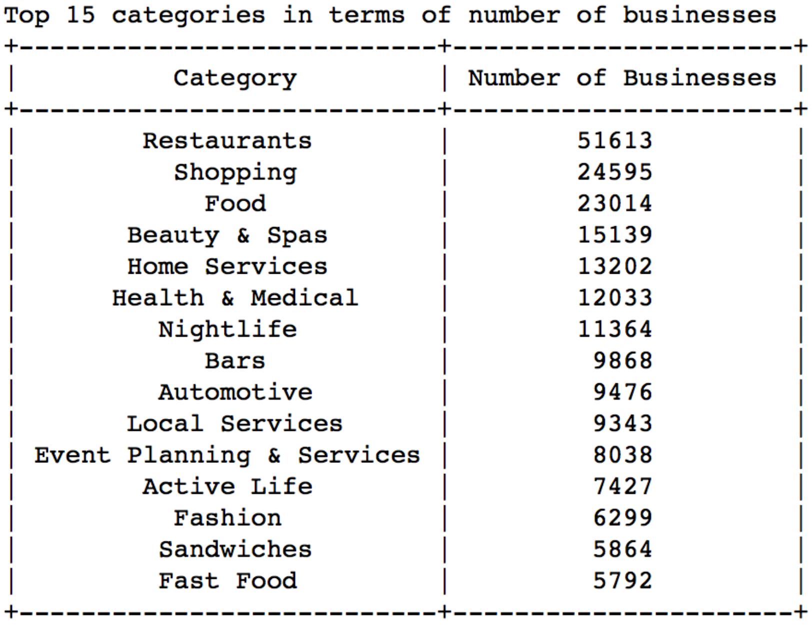
2) For business.json

We first found out total number of businesses which are 156639. Then we calculated number of businesses by city. Below is an example of top 15 cities in terms of number of businesses.



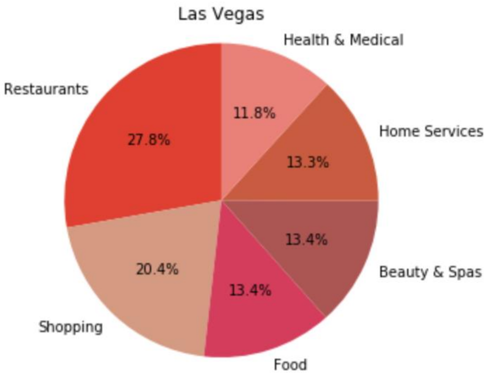
```
Number of Cities: 1010
Top 15 cities in terms of number of businesses
+------------+----------------------+
|    City    | Number of Businesses |
+------------+----------------------+
| Las Vegas  |        24768         |
|  Phoenix   |        15656         |
|  Toronto   |        15483         |
| Charlotte  |         7557         |
| Scottsdale |         7510         |
| Pittsburgh |         5688         |
|  Montréal  |         5175         |
|    Mesa    |         5146         |
| Henderson  |         4130         |
|   Tempe    |         3949         |
|  Chandler  |         3649         |
| Edinburgh  |         3625         |
| Cleveland  |         2979         |
|  Madison   |         2891         |
|  Glendale  |         2841         |
+------------+----------------------+
```

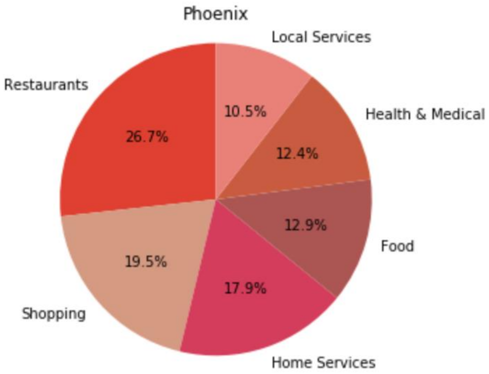
We consider these cities as popular cities. Next, we find out which type of businesses are more in number.



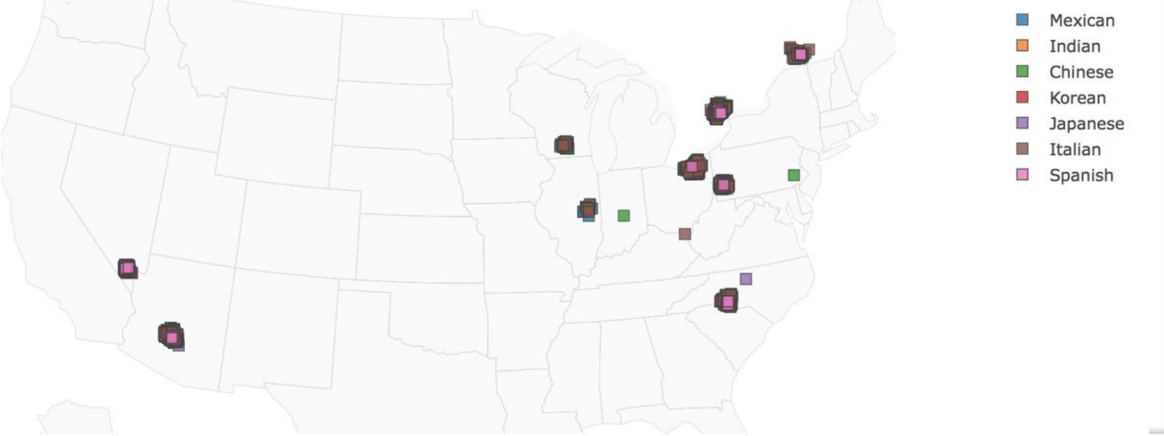
```
Top 15 categories in terms of number of businesses
+---------------------------+----------------------+
|          Category         | Number of Businesses |
+---------------------------+----------------------+
|        Restaurants        |        51613         |
|          Shopping         |        24595         |
|            Food           |        23014         |
|       Beauty & Spas       |        15139         |
|       Home Services       |        13202         |
|      Health & Medical     |        12033         |
|         Nightlife         |        11364         |
|            Bars           |         9868         |
|         Automotive        |         9476         |
|       Local Services      |         9343         |
| Event Planning & Services |         8038         |
|        Active Life        |         7427         |
|          Fashion          |         6299         |
|         Sandwiches        |         5864         |
|         Fast Food         |         5792         |
+---------------------------+----------------------+
```

These are considered as popular businesses. So, for each popular city we can extract which businesses are popular as shown in the pie-chart below:

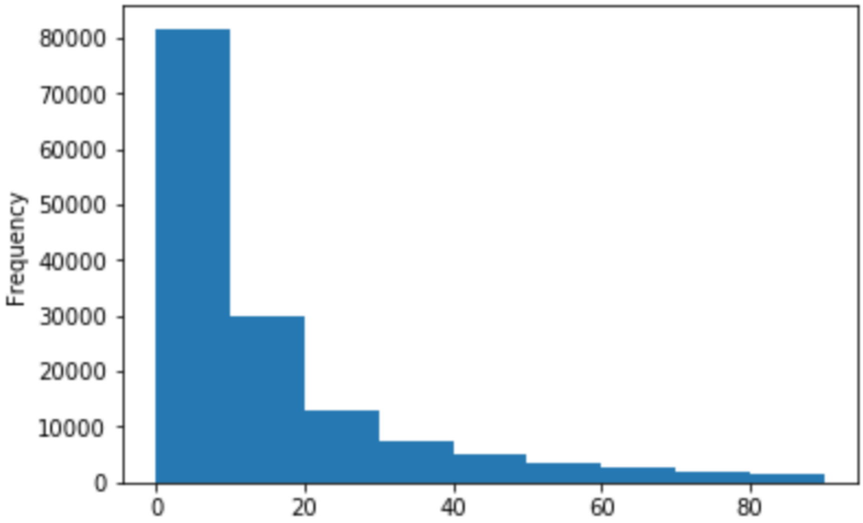


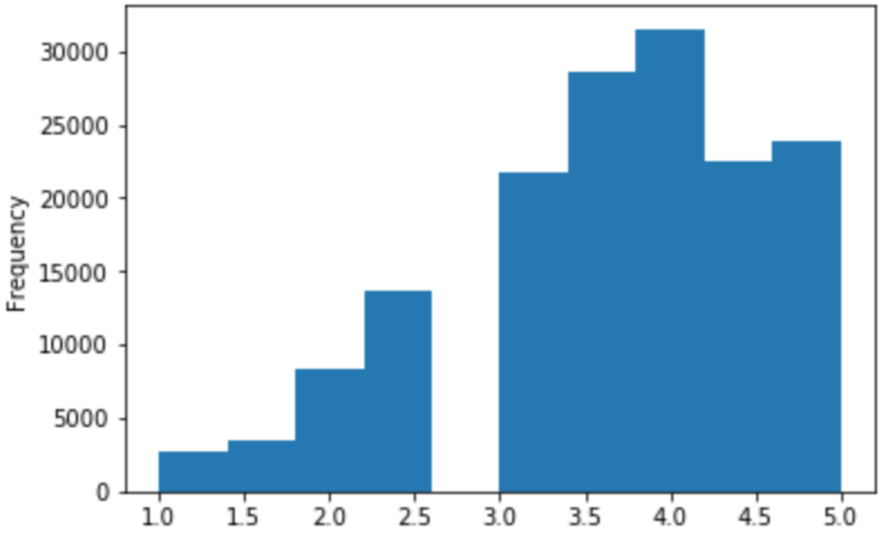


Now, let's take the most popular business say restaurants. We can go into subcategories of restaurants and provide information on type of restaurants in each popular city as shown below:



Thus, our recommendation system can use this data preprocessing to recommend a business to a user. Next, we did some analysis on no of businesses with no of review count and no of businesses with no of average star rating. We got the following two plots below:

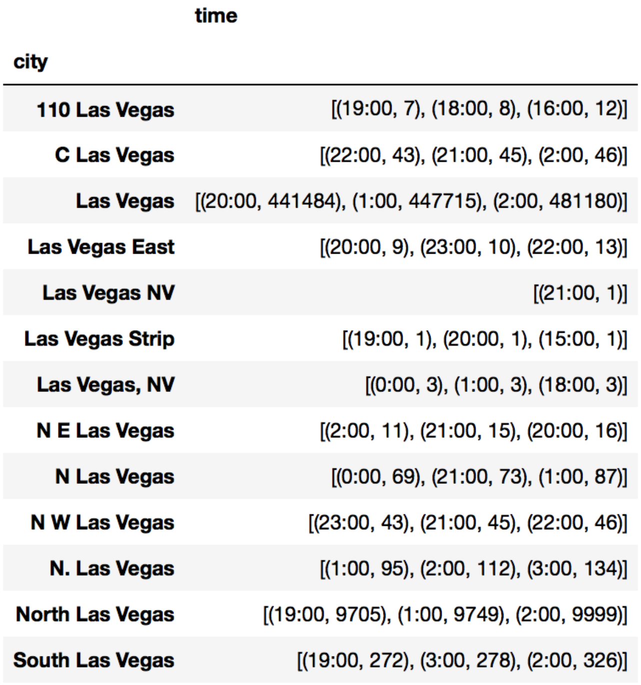




When we compare the no of businesses/no of review count plot with no of users/no of review count plot and no of businesses/average_star_rating plot with no of users/average_star_rating plot we find that they are very similar. Hence the analysis made for the above plots of users can be applied to the above two plots as well.

3) For checkin.json

We can extract popular check in times of users for popular cities in our dataset using the table below:



	time
city	
110 Las Vegas	[(19:00, 7), (18:00, 8), (16:00, 12)]
C Las Vegas	[(22:00, 43), (21:00, 45), (2:00, 46)]
Las Vegas	[(20:00, 441484), (1:00, 447715), (2:00, 481180)]
Las Vegas East	[(20:00, 9), (23:00, 10), (22:00, 13)]
Las Vegas NV	[(21:00, 1)]
Las Vegas Strip	[(19:00, 1), (20:00, 1), (15:00, 1)]
Las Vegas, NV	[(0:00, 3), (1:00, 3), (18:00, 3)]
N E Las Vegas	[(2:00, 11), (21:00, 15), (20:00, 16)]
N Las Vegas	[(0:00, 69), (21:00, 73), (1:00, 87)]
N W Las Vegas	[(23:00, 43), (21:00, 45), (22:00, 46)]
N. Las Vegas	[(1:00, 95), (2:00, 112), (3:00, 134)]
North Las Vegas	[(19:00, 9705), (1:00, 9749), (2:00, 9999)]
South Las Vegas	[(19:00, 272), (3:00, 278), (2:00, 326)]

Thus recommendations can be made based on popular check in times for popular businesses in popular cities.

Next Step:

For update 2, we are planning to finish below tasks –

- Complete pre-processing and cleaning of the data if required.
- Start working on deep learning method as it is going to take longer time to implement.
- In addition, as clustering is already covered in the class, we want to start working on clustering based recommendation system.
- Overall, in update 2, we want to finish pre-processing if needed, complete clustering based approach and make good progress on deep learning based approach.