Previous Work :
In progress report, we built a recommender system that predicts top N restaurants for a user based on Yelp dataset. We implemented following models for recommending most suitable restaurants to user :

1. **Baseline Model**: A very basic model that would just recommend the restaurant with the best star ratings in the state for user.
2. **Collaborative FIltering Model**: Built a utility matrix of users versus restaurants and exploited users with similar features to predict restaurant ratings. Mean Absolute Error of 0.62.
3. **Model with feature vector**: We built a restaurant feature vector and then used it to build user feature vector. Then similarity was found between the user feature vector and the restaurant feature vector. We also took the distance between the user and the restaurant as one of the tunable parameters. So the user can choose a distance and the recommender system will recommend the restaurants within that distance range and also if two restaurants get the same score according to our model then the restaurant which is closer to the user is recommended to the user. Mean Absolute Error of this model was 0.502. This model performed better than the Collaborative filtering model.

Problem Statement :

Yelp allows its users to form social circles within each other by adding other users as friends. This facilitates the users to share with each other their feedback/experience regarding satisfaction with respect to a particular restaurant. In such a social-driven rating scenario, it would be interesting to explore this user-friends relationship. The ratings or reviews by user's friend definitely becomes important in increasing the importance of a dish or restaurant. As suggested during the office hours by Professor, we are now performing analysis for the below problems.

- Are user and his/her friends are similar when it comes to rating restaurants?
- How does the number of available reviews/ratings for a restaurant impact the user's rating for that restaurant?
- Do users with high number of followers on Yelp tend to rate restaurants higher?

Number of Friends : Distribution
On analysis, we found that for total of **1,183,362** users, there were **39,846,890** friend relationships which gives us that each user is roughly having **33** friends on average. For having a better understanding on how the number of friends per user are really distributed in the system, we made a plot of 'degree 'x' number of friends' vs 'The number of users having 'x' number of friends'. The plot is shown in figure 1.1. The graph output is a **power-law** phenomenon, which is exactly what we expected it to be. The user with the maximum friends

relationship have a value of 14,995. While there are around 507,951 users with no friend and 85,826 users with only 1 friend.
For visualization purposes, we have ignored the users with zero friends and have only taken up users with 300 or less friends.

```
friendCountList = np.array(friendCountList)
friendCountListNonZero = friendCountList[friendCountList != 0]
friendCountListNonZero = friendCountListNonZero[friendCountListNonZero <= 300]
plt.hist(friendCountListNonZero,300, facecolor='blue', alpha=0.75)
plt.xlabel('Friends Count')
plt.ylabel('Number of Yelpers')
plt.rcParams.update({'font.size': 15})
plt.text(25,8000,'8K yelpers have about 25 friends',fontsize=12,bbox=dict(facecolor='red', alpha=0.5))
fig = plt.gcf()
fig.set_size_inches(10.5,5.5)
```

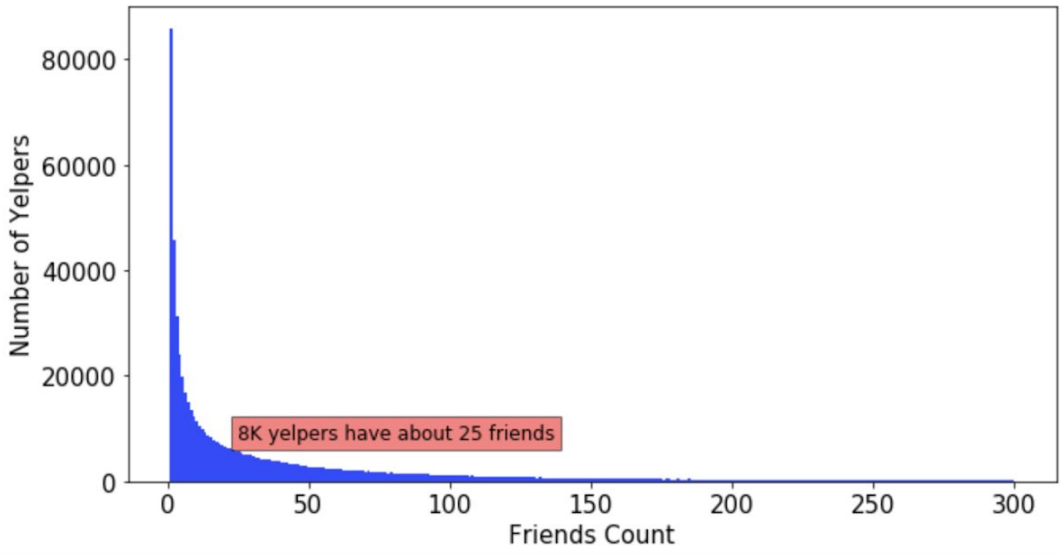


Figure 1.1 : Probability distribution for number of yelp users having a certain friend count

On plotting the same on a cumulative distribution (figure 1.2), we got that around 80% of the Yelp users have 4 or less friends. These users might not turn out to be significant contributors on further exploration of friend analysis.

```
unique, counts = np.unique(friendCountListNonZero, return_counts=True)
counts = np.cumsum(counts)
plt.plot(unique,counts, alpha=0.75)
plt.xlabel('Friends Count')
plt.ylabel('Number of Yelpers')
plt.rcParams.update({'font.size': 15})
plt.text(25,20000,'20K yelpers have about 25 friends',fontsize=12,bbox=dict(facecolor='red', alpha=0.5))
fig = plt.gcf()
fig.set_size_inches(10.5,5.5)
```

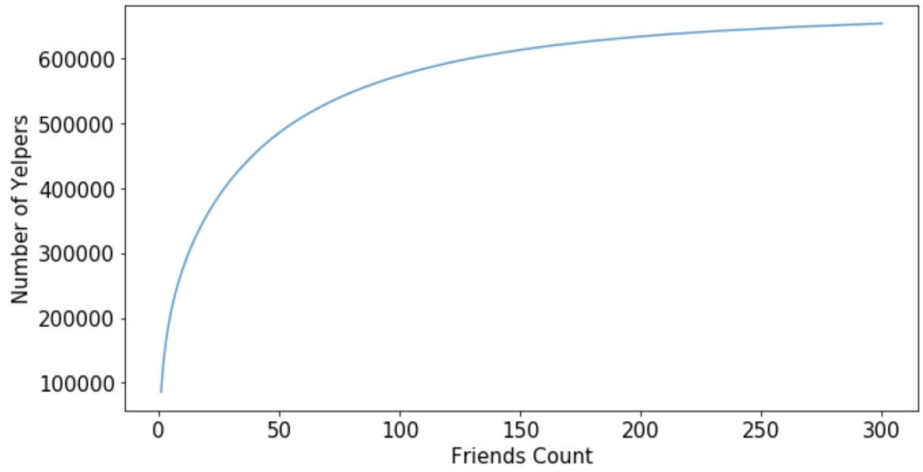


Figure 1.2 : Cumulative distribution showing number of yelp users below a certain friend count

Problem 1 : Impact of friend relations :

Next, we wish to analyze the impact of friend relations on Yelp ratings, i.e. does a friend's rating for a given restaurant influence the user's rating for the same restaurant or in other words: **Are User and his/her Friends Similar?**
Procedure: We analyzed this statewise. We had the list of user's friends in our dataset and rating that they gave to restaurants. So for each user and for each restaurant that the user visited, we find out which of his friends visited the same restaurant and calculated the average rating that they gave to that restaurant. We also have the user rating and the average rating of that restaurant. To be more accurate, we didn't take this average rating of the restaurant, rather we calculated the average for persons who visited that restaurant excluding the user himself and his friends. So this way we calculated the average rating of other people who are not friend of the user on Yelp. Then we used the measure to test results based on below two values, as discussed during the office hours. So, we compared these two values:

| User Rating - Friend's Average Rating| and **| User Rating - Other's Average Rating|**

However, we didn't find any significant difference between the two values.
For IL state :
For All Users (User Rating - Friend's Average Rating) = **0.85067**
For All Users (User Rating - Other's Average Rating) = **0.85162**

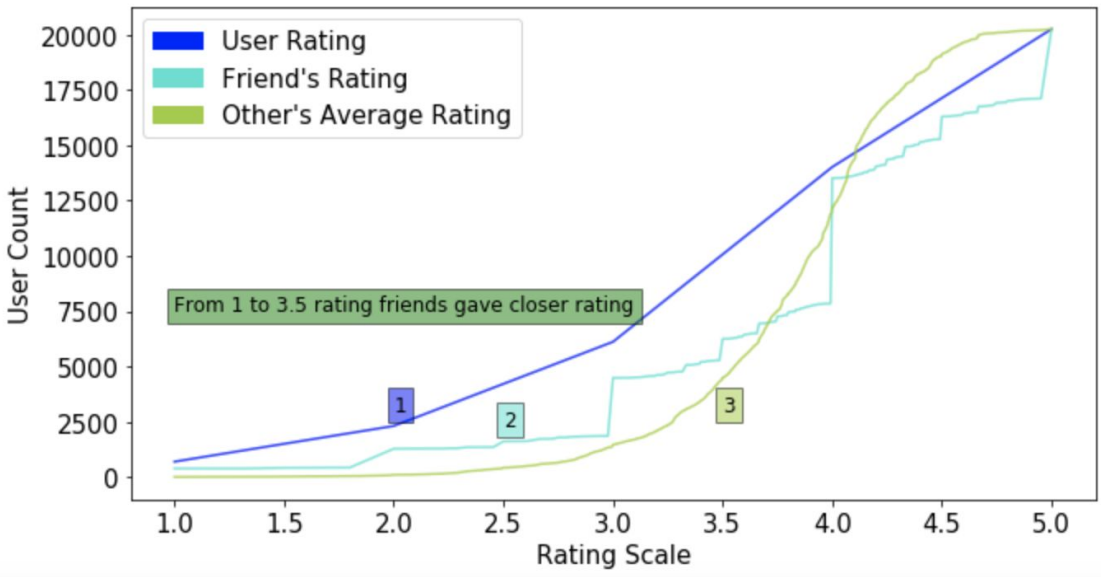
Code Snippet for IL State:

```
In [196]: a = 0
          b = 0
          for i in range(len(userPlotList)):
              a += abs(userPlotList[i] - friendPlotList[i])
              b += abs(userPlotList[i] - averagePlotList[i])
          print(a/len(userPlotList), "===============", b/len(userPlotList))

          0.8506702046587706 =============== 0.8516298985320394
```

As we can see the difference is extremely minimal. Same was the result for other states also. So we can make an inference based on given data that a user and user's friends rating are not more similar to user's rating and other people rating. The user rates restaurants as per his own experience at the restaurant for that day.

Below is a graph showing values of user rating, friends' rating and other's average rating for state WI. This helps to visualize the difference between the 2 absolute values :
| User Rating - Friends' Average Rating| and | User Rating - Other's Average Rating|



As we can see that from rating 1.0 to 3.5, friends of the user gave ratings which are more closer to user's rating than the other people average rating to the user. After that both are equally closer to the user's rating for a given restaurant because rating from 3.5 to 4.5 are general in the sense that most people give these ratings. So the closeness of friends is not greater than the other user to a given user after 3.5 rating.

On finding correlation between ‘avg friend rating’ and the ‘user rating’ we got a value (for all states) of 0.37, which though positive but is very low to state any kind of positive relationship between user’s rating and the friends’ average rating.

States	Corr Coeff.
EDH	0.43
PA	0.34
WI	0.32
NV	0.35
QC	0.35
OH	0.36
ON	0.34
BW	0.45
IL	0.4
Average	0.371111111

Figure 1.3 : Correlation Coefficient between User and his Friend’s rating for different states.

The highest correlation was for state BW (0.45) and least was for WI (0.32) and the average is 0.37.

Next we plotted some plots to analyze more. Below are the plots for 4 different states which shows the relationship between rating given by a user to a restaurant and the average rating that his friends gave to the restaurant that they visited in common with the said user.

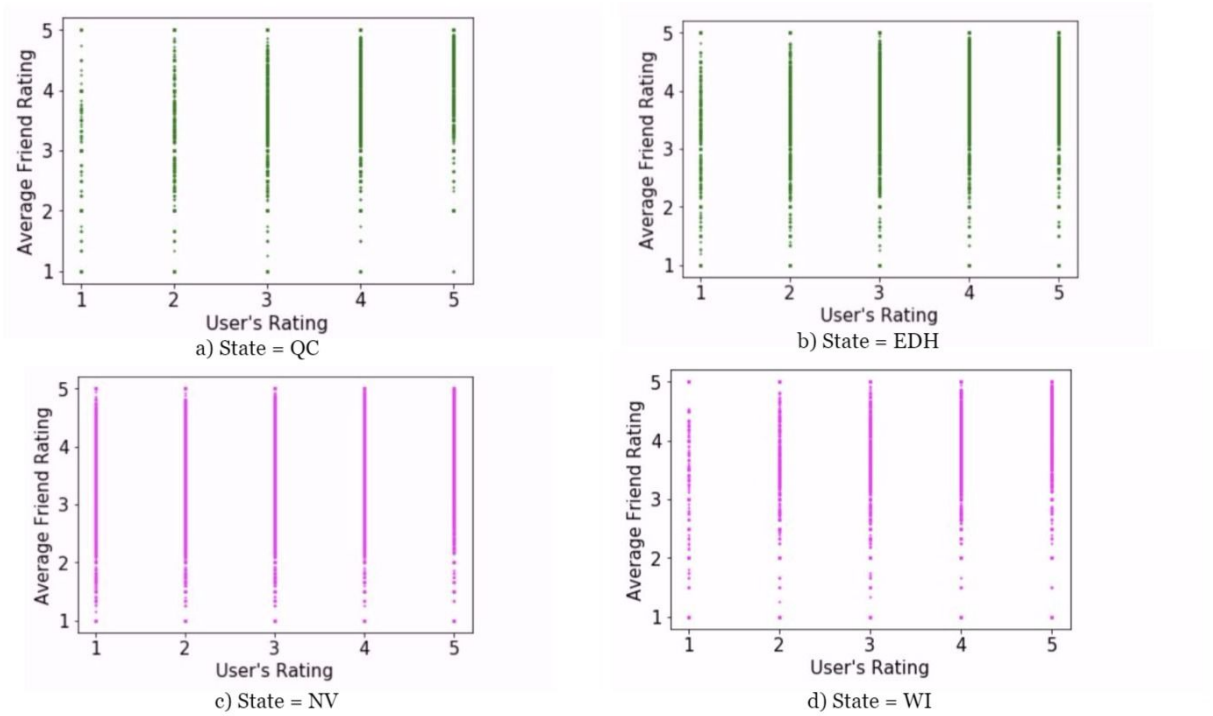


Figure 1.4 : Plot showing Average friend rating to the user's given rating for 4 states (QC, EDH, NV and WI)

All the 4 plots are inline with each other. Firstly, a very basic observation is that a rating of 1 is given very infrequently both by the user and his friends which is also true for general public. As we move from left to right, as the user rating increases and so is the user's friend average rating. So the dots start concentrating towards more and more top when we move from left to right. So this little relationship do explains the correlation coefficient that we got from our research findings but this is not substantial to infer anything.

Conclusion: We conclude that based on given data, the user's rating for a restaurant and user's friends' average rating for the restaurant are not dependent on each other.

Inference: As, we see that there is no direct relationship between how a user's friends rate a restaurant and the user' rating. One of the foremost reason we can think of is as friends are actually a group of people whom the user have simply added as friend on Yelp. These persons might not be user's true friends and the user might not have ever met these people. It is highly possible that these people's restaurant preferences, price range and taste do not match the user. Moreover, it might be that Yelp friends do not dine-in together.

Problem 2 : Impact of number of reviews for a restaurant :

Another interesting problem is to how the number of reviews available for a restaurant effect a user's mindset while rating that restaurant. What seems plausible is that more the number of

reviews (with good scores), more likely is a user expected to give a good rating to this restaurant.
From our dataset formed by reviews.json file (after extracting restaurants from all businesses), we formed a new column for every restaurant which gives us the number of reviews available for that restaurant. Now as we wish to check the correlation, we make the ‘number of reviews for a restaurant as our independent variable and user’s rating as predictor or dependent variable. This gives us a correlation coefficient of 0.07 which is very less and signifies that our correlation is not able to explain our data.

	State	Corr Coeff.
0	QC	0.063095
1	BW	-0.004547
2	PA	0.098605
3	WI	0.125087
4	EDH	0.087021
5	IL	0.204120
6	OH	0.102344

Fig 1.5 : For each state,correlation b/w user rating and number of times restaurant has been reviewed.

To visualize this, we plotted a scatter graph (figure 1.6) and as in accord with our low value of correlation coefficient, the plot does not provide any sharp outcome. This again shows that there is not much correlation among the number of reviews/ratings a restaurant has received versus the rating a user has given. For each state we can observe that irrespective of number of reviews a restaurant has, the restaurant is likely to get all kinds of rating i.e. 1,2,3,4,5.

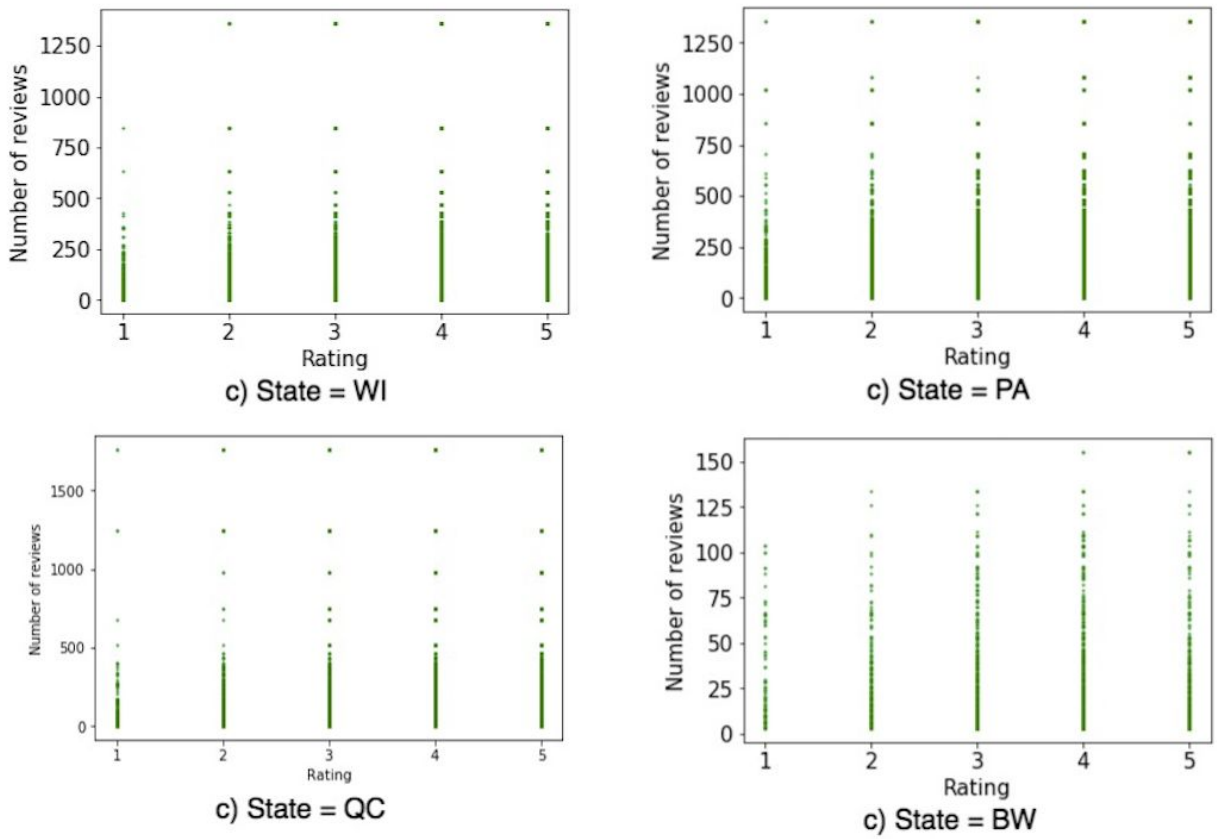


Fig 1.6 : Scatter plot b/w Number of reviews and Rating given by user for 4 states (WI, PA, QC and BW)

Though the number of ratings a restaurant has received show the authenticity of the ratings. Combined with other Yelp features like reviews and photographs help to provide an overall picture of the restaurant which can help in deciding restaurants. A given rating/review (either good or bad) might say about a person's preference, taste or mood at that point of time. But high number of such ratings need not necessarily show how a given user would experience on visiting and eating at the restaurant. It can be that the user gave negative ratings as he/she was in a grumpy mood all day or maybe the food at a very fine restaurant was not good just because the chef had a bad day. While on the other hand, a user can like a restaurant poorly rated by many users because restaurant provided certain amenities that are a priority to the user. Maybe the user was in a hurry and the restaurant manager agreed on a really quick service that day and so, the user might rate even the average restaurant in the city very highly.

Problem 3 : Impact of number of fans on the user's rating :

In addition to making friends on Yelp, one can also follow another user and thus become their fan. While the friend relations requires an acceptance of friend request from another party, but for a fan relationship, there is no such acceptance of request. This distribution, like the friend

distribution, also follows power law. While there are **915,264 users** with zero fans, there are **607** users with 200 or more fans. The most number of fans for a user is **6087**. Most of the users with high number of fans are the ones whose reviews have been upvoted heavily and hence people find their ratings informative and authentic. These people mostly do comprise of celebrity chefs or foodie celebs. Here, we want to analyze users who have high fan following and how (if any) their ratings are affected.
For this, we parsed our user.json file again and added a new column for the 'number of fans' for every user.

Code Snippet:

```
f = open('user.json','r')
num_fans = list()
avg_rating = list()
for line in f:
    jsonObject = json.loads(line)
    fans = jsonObject.get("fans")
    avg_star = jsonObject.get('average_stars')
    if fans<1000:
        num_fans.append(fans)
        avg_rating.append(avg_star)
f.close()
plot(avg_rating,num_fans,"Avg Rating","Number of fans")
```

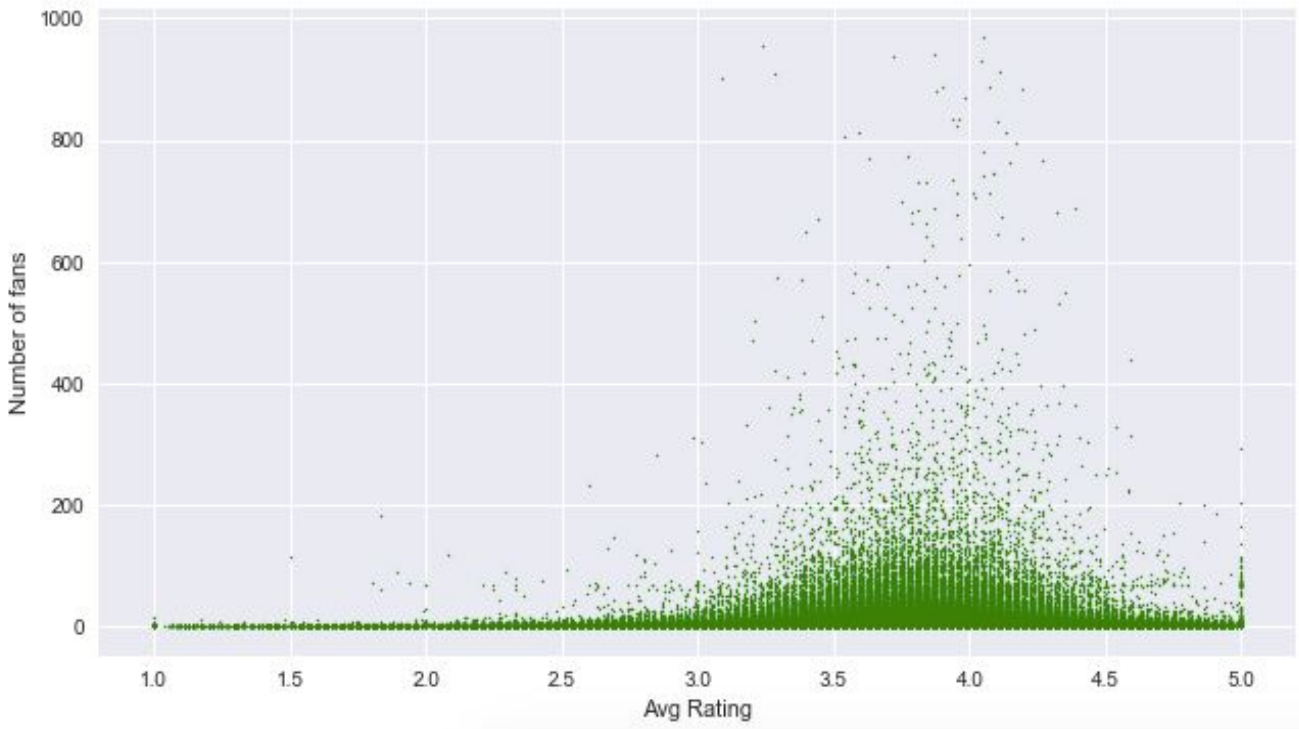


Fig 1.7 : Relationship b/w Avg Rating by user and Number of fans he/she has

We built a scatter plot (figure 1.7) for the ‘number of fans’ against the ‘average rating’ given by that user. We observe that for users with 300 or more fans have their ratings consolidated around 3.5 to 4.5. This shows that such elite users don’t tend to give bad ratings to restaurants on average. These elite users rate more restaurants highly and rate less restaurants poorly. This might be due to the fact that foodies with their experience often find some shortcomings in highly rated restaurants while they’ll find out some good plus points about averagely rated restaurants.

Conclusion :

So, we conclude our 3 problems by saying that :

1. Average friend rating for a given restaurant do not impact on the user’s rating, although there is a weak correlation of 0.3.
2. The number of ratings for a restaurant and the user’s rating have a very low correlation of 0.07 which is very weak to provide any kind of direct influence. We can better say that both these events are uncorrelated.
3. And for the last problem, we observed that users having more fans tend to give higher ratings to restaurants.

Future Work :

In the future, we would like to expand the current analysis to include review text and user rating evaluations (whether other users thought a particular user’s review was funny, useful, or helpful) as features in the prediction model. We would also, explore further hybrid approaches and evaluate their performances.