AC209a Data Science Project: Data Science with User Ratings and Reviews

Andrew Ross, Sophie Hilgard, Reiko Nishihara, Nick Hoernle

November 5, 2016

Data Source

We downloaded the data from the 'Yelp Dataset Challenge' (`https://www.yelp.com/dataset_challenge`). The data contain in total 2.7M reviews from 687K users for 86K businesses. Business data consist of 15 features including ID, category of business (e.g., fast food, restaurant, nightlife, etc.), city, full address, operation hours, latitude, longitude, review count, and stars earned. User data consist of 11 features including average stars, compliments, elite, number of fans, IDs of his/her friends, name, review count, vote categories, and the month start a yelp review. So in terms of data that would be useful to recommendation systems, we have everything we need to construct a utility matrix (users, businesses, and ratings from 1-5 stars), in addition to a great number of additional interesting features that could be helpful either for recommendation or for general analysis to answer broader questions of interest.

Data Exploration

Overview Exploration

A simple inspection of the businesses data shows that useful information such as the business categories, latitude and longitude and the average ratings given to a business are all easily available. Similarily, for each user we have information about the number of reviews that they have given, the average rating that they give and the date that they have been 'yelping' since.

	categories	city	full_address	hours	latitude	longitude	name	open	review_count	stars	state
0	['Fast Food', 'Restaurants']	Dravosburg	4734 Lebanon Church Rd\nDravosburg, PA 15034	{'Monday': {'close': '21:00', 'open': '11:00'}...	40.354327	-79.900706	Mr Hoagie	True	7	3.5	PA
1	['Nightlife']	Dravosburg	202 McClure St\nDravosburg, PA 15034	{}	40.350553	-79.886814	Clancy's Pub	True	5	3.0	PA

Figure 1: Head of Businesses Dataframe

	average_stars	compliments	elite	fans	friends	name	review_count	votes	yelping_since
0	4.14	{'plain': 25, 'writer': 9, 'cute': 15, 'photos...	[2005, 2006]	69	[1, 2, 3, 5, 93, 12, 99, 464, 1025, 1298, 1388...	Russel	108	{'cool': 246, 'useful': 282, 'funny': 167}	2004-10
1	3.67	{'plain': 970, 'writer': 346, 'cute': 204, 'ph...	[2005, 2006, 2007, 2008, 2009, 2010, 2011, 201...	1345	[0, 2, 3, 4, 5, 6, 8, 9, 12, 95, 97, 187, 465,...	Jeremy	1292	{'cool': 12091, 'useful': 15242, 'funny': 8399}	2004-10

Figure 2: Head of Users Dataframe

A simple summary description of these dataframes is then presented below:

```
businesses.drop(['latitude', 'longitude', 'business_id'], axis=1).describe()
```

	open	review_count	stars
count	85901	85901.000000	85901.000000
mean	0.852272	34.352359	3.694852
std	0.354832	108.677591	0.946045
min	False	3.000000	1.000000
25%	1	5.000000	3.000000
50%	1	10.000000	4.000000
75%	1	26.000000	4.500000
max	True	6200.000000	5.000000

Figure 3: Summary statistics of the Businesses Dataframe

```
users.drop('user_id', axis=1).describe()
```

	average_stars	fans	review_count
count	686556.000000	686556.000000	686556.000000
mean	3.746704	1.290100	25.757102
std	1.086832	11.501621	83.755973
min	0.000000	0.000000	0.000000
25%	3.230000	0.000000	2.000000
50%	3.920000	0.000000	5.000000
75%	4.600000	0.000000	17.000000
max	5.000000	3549.000000	10897.000000

Figure 4: Summary statistics of the User's Dataframe

We see from the summary data that most reviewers do not have fans and wrote a small number of reviews. Concretely, the median number of reviews given by a user is 5 yet the mean is 25. This suggests drastically right skewed data (which is intuitive as there is a lower bound of 0 on the number of reviews that a user can give). The inter-quartile range, IQR, for user rating was 3.2 - 4.6, again suggesting that most users rate businesses higher than the midpoint rating of 2.5. Similarly, we see that the businesses receive a mean rating of 3.69, with a mean count of 34.35. The IQR for review count for businesses is 5 to 26, again suggesting that a majority of businesses have a small number of reviews.

The review counts (number of reviews given by a user and number of reviews received by a business) are dramatically right skewed. We thus, omit the outliers in the below plot to understand the distribution of the 10^{th} to 90^{th} percentiles for review counts. We still see a hugely right skewed dataset, again intuitively it is more common for many users to rate a few number of businesses and it is more common for many businesses to receive a low number of reviews.

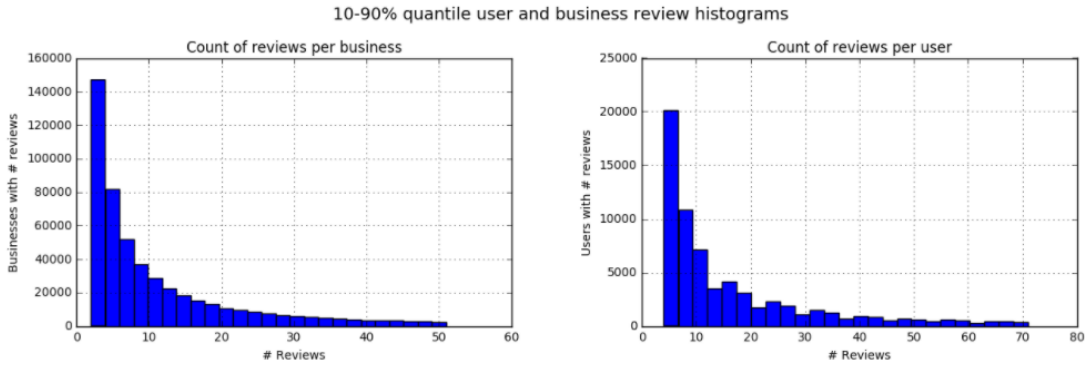

Figure 5: Exponentially distributed count of reviews received per business (left) and count of reviews given per user (right)

Time based Exploration

It was interesting to understand the distribution of the numbers of reviews over time for the yelp data.

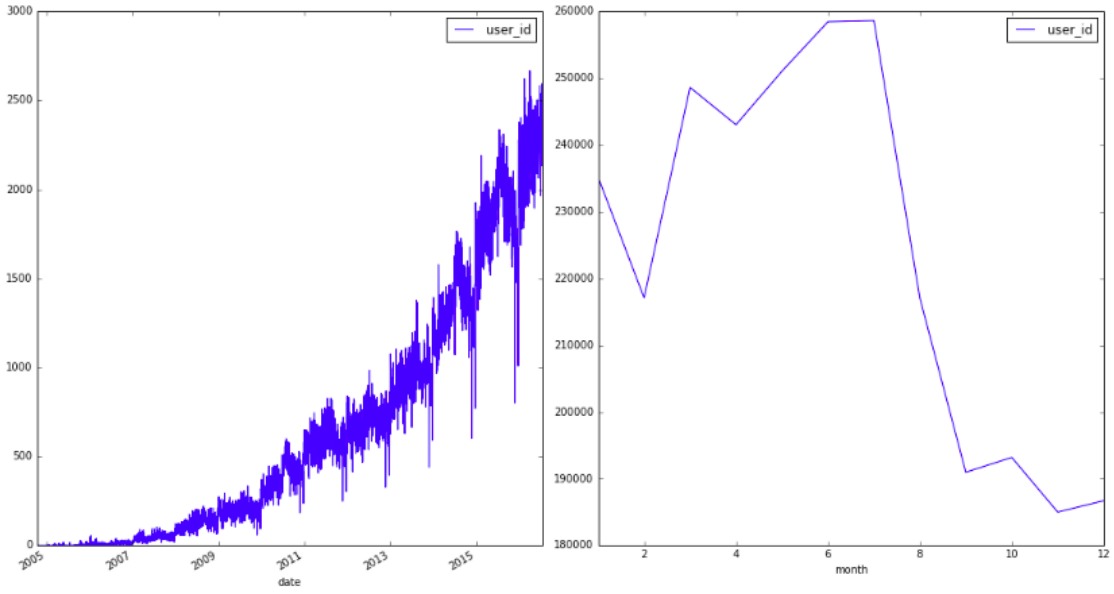

Figure 6: Number of reviews per day on yelp over time

We see that the number of reviews that are submitted on the Yelp platform over time is increasing

dramatically from when Yelp opened in 2004 to present, were approximately 2500 reviews are submitted per day. Analyzing the number of reviews that are submitted over the course of a year shows that June and July are popular months while November and December show a relevantly lower number of reviews. An interesting point for further exploration is if the business type shows different 'high' and 'low' seasons and if the popularity of a business or a category of businesses can be tracked over time.

Location based Exploration

We see that we are given 10 cities from the yelp data. From the metatdata about this database we know that these correspond to (the corresponding number of businesses was calculated for each city):

- U.K.: Edinburgh (3480)
- Germany: Karlsruhe (1074)
- Canada:
  - Montreal (5592)
  - Waterloo (530)
- U.S.:
  - Pittsburgh (4088)
  - Charlotte (7160)
  - Urbana-Champaign (807)
  - Phoenix (36505)
  - Las Vegas (23598)
  - Madison (3067)

For the locations part of this study we will therefore narrow the focus to Las Vegas and Phoenix due to the large number of businesses in these cities and the expected restaurant culture that is a perception that is given of these cities.

For Phoenix the various categories were extracted from the data and the top business categories (by count) are shown below:

- 'Restaurants', 9428
- 'Shopping', 5424
- 'Food', 3637
- 'Beauty & Spas', 3603
- 'Home Services', 3466
- 'Health & Medical', 3420

- ‘Automotive’, 2629
- ‘Local Services’, 2119
- ‘Nightlife’, 1599
- ‘Active Life’, 1551

We explored any trends in these categories and plotted the data by location. We also color coded the businesses by their average rating (with brown being the highest rating of 5 and gray being the lowest rating of 1). Restaurants are evenly distributed throughout Phoenix with an overwhelming number of highly rated businesses. The other business types are more evenly spread out but the Nightlife category shows a clear ‘hub’ in the city center. Unfortunately, we see no clear areas that are high rated vs low rated areas. This is a specific point of interest (i.e. do certain areas come into and out of popularity) and thus will be investigated further throughout the course of this project.

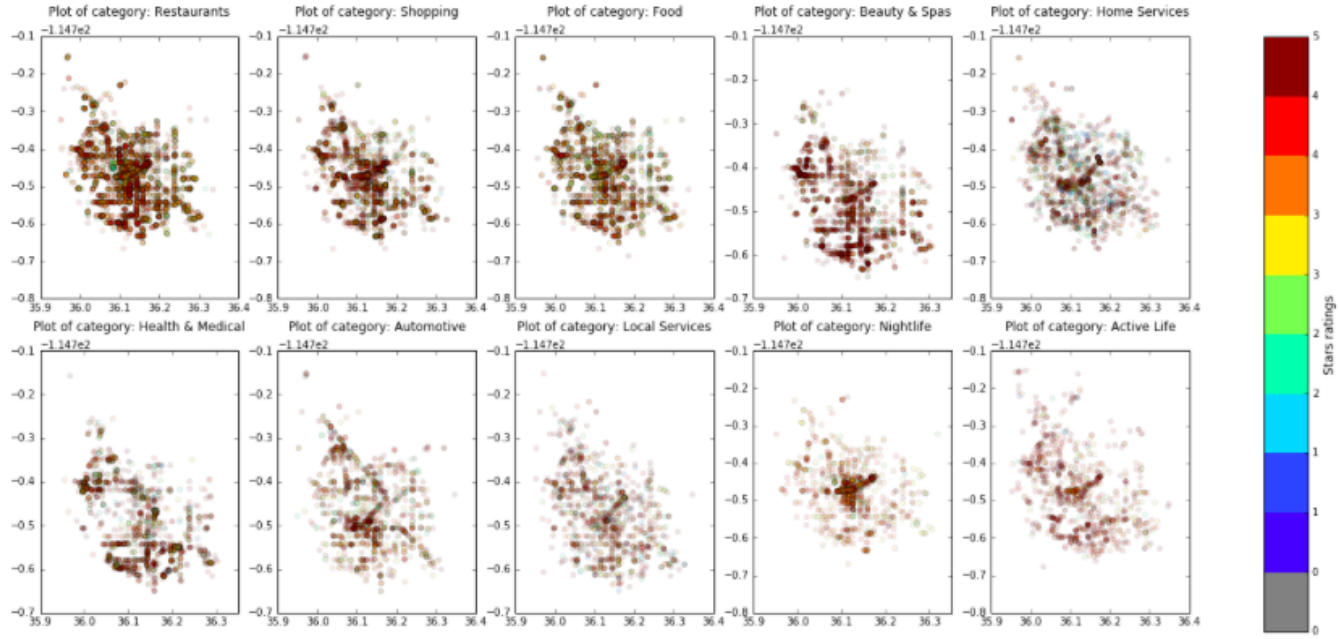

Figure 7: Chart showing the top 10 most popular business category plotted by location and colour coded by average review

Exploring the ”Stars”

Since a major focus of this project is on recommendations, we want to dive deeper into the star ratings data and see how we might use it to compute both baseline rankings for businesses and users (which was a major component of the model that won the Netflix challenge), as well as user-specific modifications. First, let’s just look at the overall distribution of stars that users give to businesses, as well as the range of stars each user gives / each business receives:

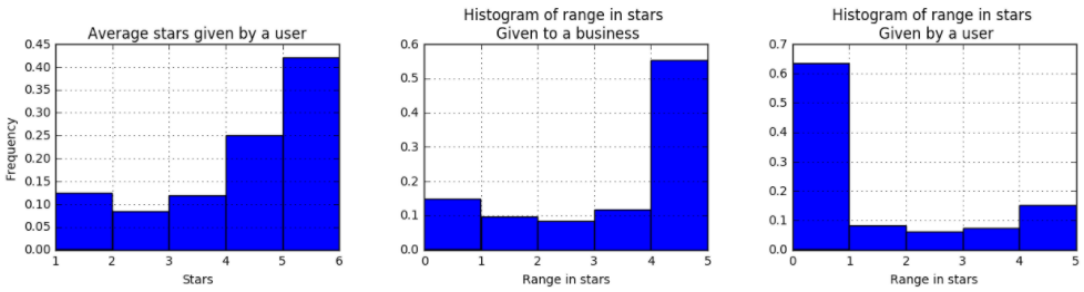

Figure 8: Plot of the average and range of ratings that a user gives and the range of ratings that a business receives.

Here we see that users typically rate businesses highly (with a clear mode being 5 stars). However, we also note that businesses have a high range of votes (a mode of 4 indicates that users differ in their opinions). The figure also shows that users did not change their rating schema based on the business, and the majority of users rate all businesses within 1 point of each other. When we limit our selection to users who have given more than 15 reviews, this distribution changes a little bit:

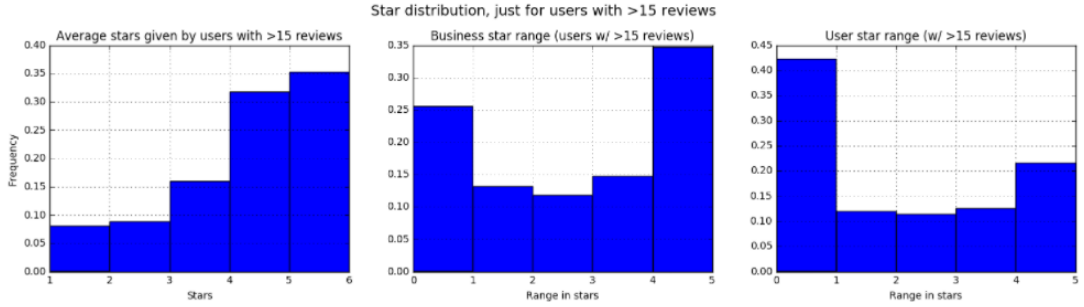

Figure 9

In particular, users are more likely to give 4s and less likely to give 5s and 1s. This is remniscent of the effect in the Netflix paper, when users who were giving more than one review per day were likely to give less extreme reviews (which was a major way they reduced their MSE that final amount to win). We can actually investigate the exact same phenomenon:

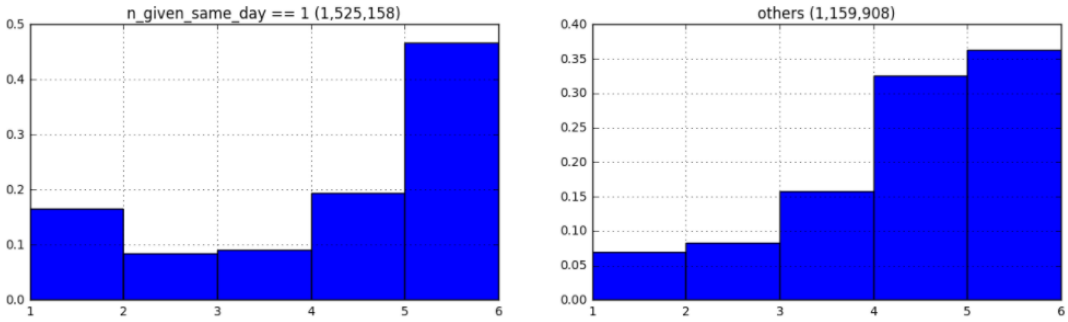

Figure 10: Difference in stars when giving one or many reviews per day

When users give multiple reviews on a day, they tend to be more moderate in their ratings. If users are only giving one review per day, it might be because they just had a very positive or very negative experience on that day, and immediately log on to rate the business. Note that the fraction of users who only give one review per day has changed over the course of Yelp's history:

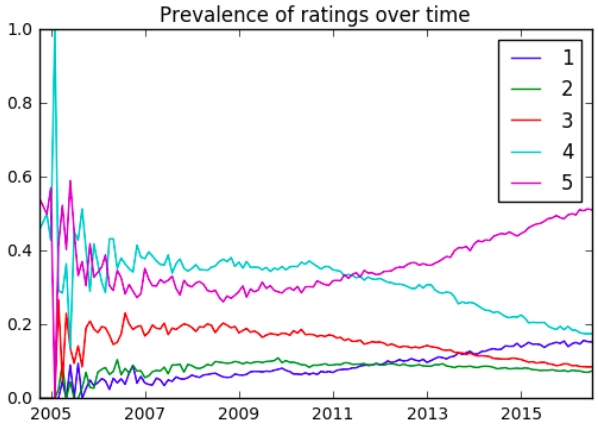

Figure 11: Prevalence of ratings over time

A natural question to ask is whether users who give one review per day actually give more extreme reviews, or whether users in general have started giving more extreme reviews, and there are just more users who give one review per day. However, even when we conditioned on short time slices, we saw the same effect. A more extreme example is to condition on number of fans:

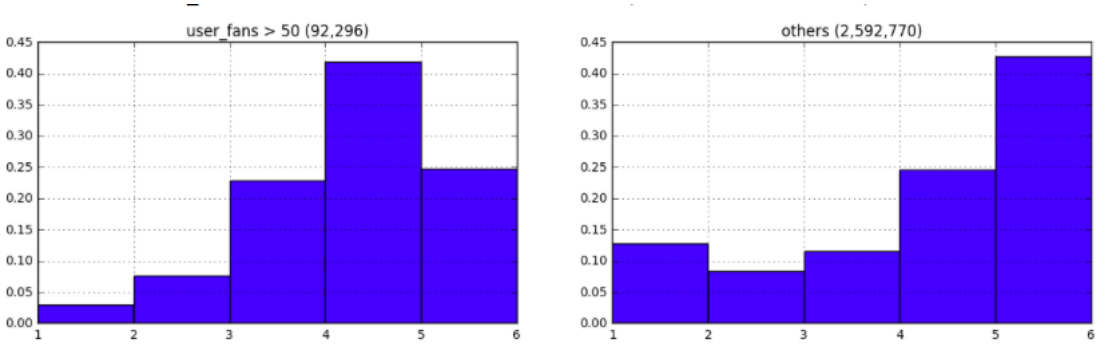

Figure 12: Difference in stars between users with ¿50 and ¡=50 fans

If we condition on how many fans the user has, we find that that also plays a significant role in how extremely they rate users. However, we also discovered that even for users with more than 50 fans, when we condition on whether *they* were giving one review per day or many, they were also more extreme when giving just one review (although the effect was less strong). This will be important to consider when computing a baseline model for business likeability.

Another issue in computing baselines is that we do have sparse data. To address this, we are considering adopting an approach similar to the movie recommendation homework, where we begin with a Beta prior assumption that every business has an average rating of 2.5 stars. More concretely, we introduced a 0 to 1 scale, assumed a Beta(8,8) prior, and added the ratings to rank all business, which gave us the following distribution of posterior "business ratings:"

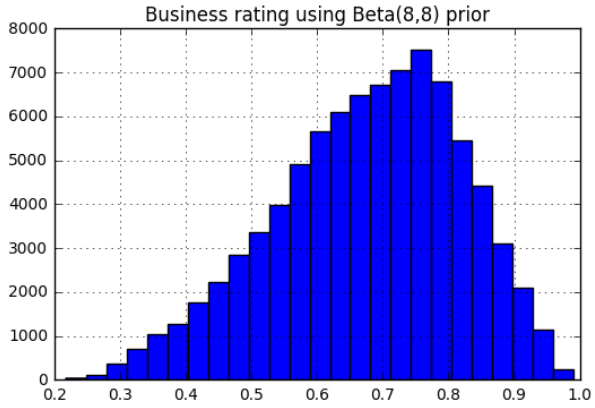

Figure 13: Chart showing the top biased businesses

To give an intuition for what the highest and lowest ranked businesses are, we can just examine the top 5 of each:

Table 1: Examples of low and highly ranked businesses (w/ rounded average stars, number of reviews, and whether they have since closed)

Lowest Ranked	*s	#	Closed	**Highest Ranked**	*s	#	Closed
A Victory Inn	1	348		Blue Chip Auto Glass	5	121	
OnTrac	1	28	x	Lockaid USA	5	151	
Monitronics Security	1	22		Stell Roofing	5	124	
Anjile Cleaning Service	1	26	x	Simply Skin Las Vegas	5	133	
Website Backup	1	63		Khina Eyebrow Threading	5	105	

Although the average stars here are rounded (a limitation in the summary statistics provided by the dataset), we can see that businesses with lots of high ratings rise to the top and businesses with lots of low ratings sink to the bottom. Also, some of those poorly-rated businesses are now closed, which might be interesting to explore further. How well can we use rating data to predict whether a business closes?

We are also interested in determining the effect that location plays in determining ranking. When we locate businesses on top of the actual map, we see that highly-rated businesses tend to locate on highways. This suggests that the accessibility is important to earn positive reviews as well as reviews itself, especially when the businesses are not located in the center of the city.

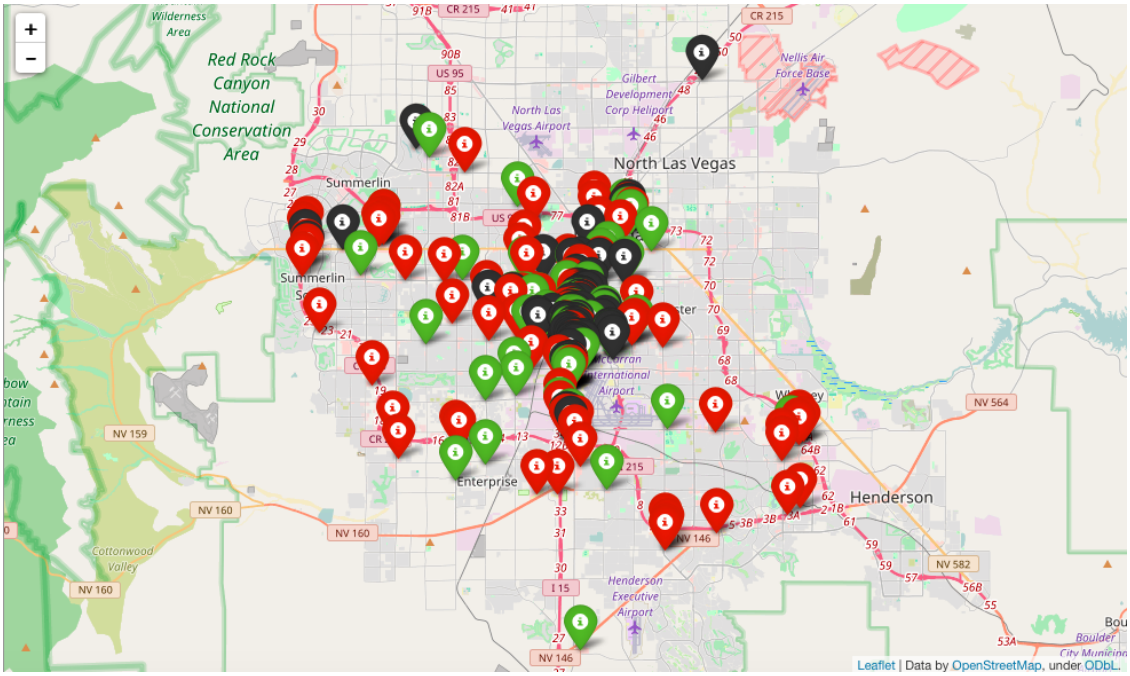

Figure 14: Geographic location of top-rated businesses

Network Analysis

We conducted a network analysis to see the topology and relationship among businesses based on user-rated 'stars'. When the star-ranking is positively correlated between two businesses, they

are connected with a red edge in the business network. When the star-ranking is negatively correlated between two businesses, they are connected with a blue edge. We used Spearman's ranked correlation analysis to calculate correlation coefficients.

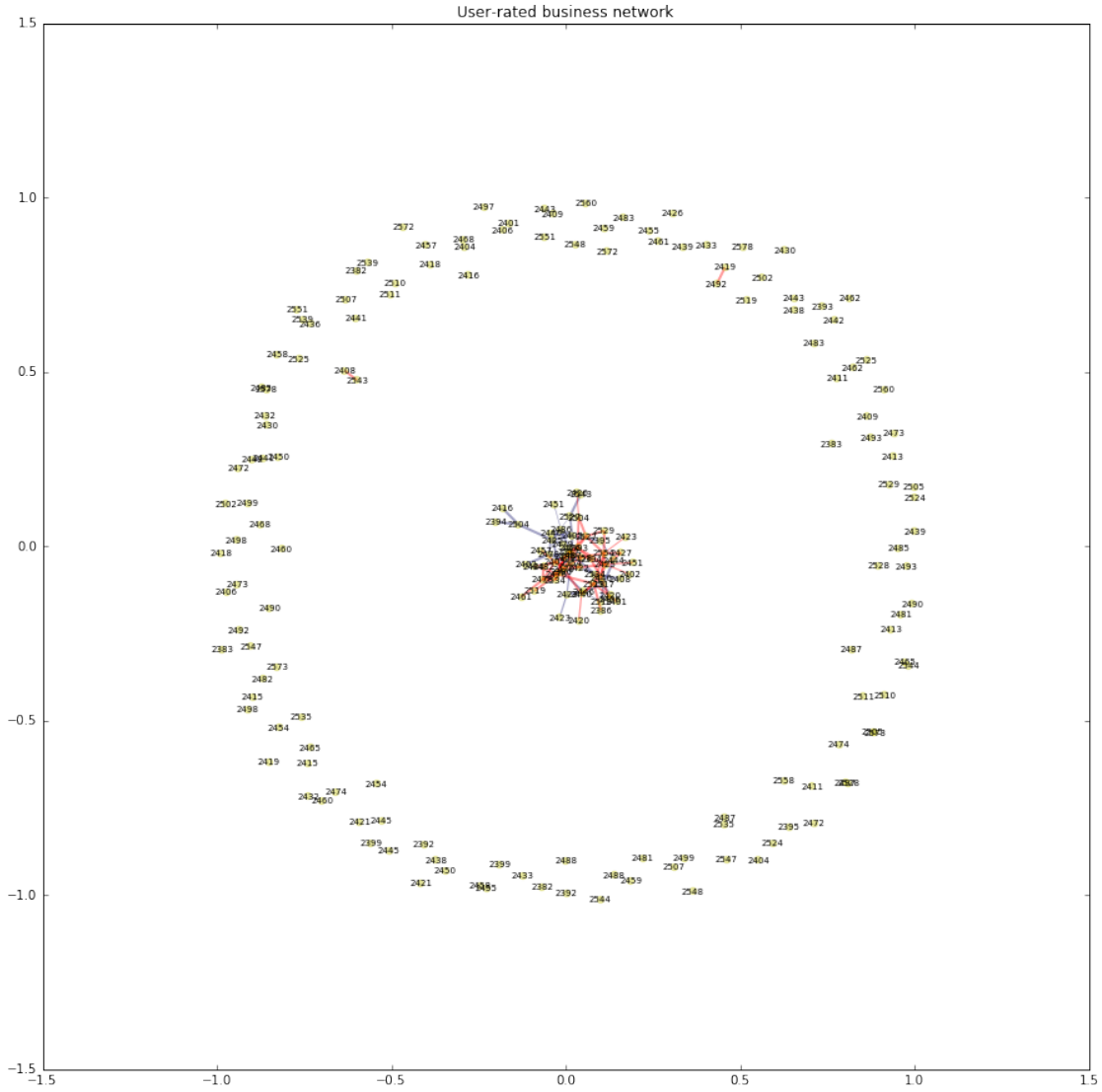

Figure 15: Business correlation network based on user-rated stars in Phoenix.

This preliminary analysis only included 100 businesses (to clearly see the network structure) located in Phoenix. The node in the network figure indicates a business ID, and an edge represents a connection between businesses. The above figure shows that some businesses are highly connected with others, and there are businesses not connected (i.e., businesses were not rated by users). In the next step, we will further focus on a highly-rated businesses and specific type of businesses (e.g., restaurant).