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

Andrew Ross, Sophie Hilgard, Reiko Nishihara, Nick Hoernle

November 4, 2016

Data Source

We downloaded the data from the 'Yelp Dataset Challenge' (`https://www.yelp.com/dataset_challenge`). The data contains 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.

Data Exploration

Overview Exploration

A simple inspection of the businesses data shows that usful 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.

Businesses dataframe shape (85901, 15)

	attributes	business_id	categories	city	full_address	hours	latitude	longitude	name	neighbor
0	{'Accepts Credit Cards': True, 'Noise Level': ...	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	[]
1	{'Accepts Credit Cards': True, 'Price Range': ...	1	['Nightlife']	Dravosburg	202 McClure St\nDravosburg, PA 15034	{}	40.350553	-79.886814	Clancy's Pub	[]

Figure 1: Head of Businesses Dataframe

Users dataframe shape (686556, 11)

	average_stars	compliments	elite	fans	friends	name	review_count	type	user_id	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	user	0	{'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	user	1	{'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.

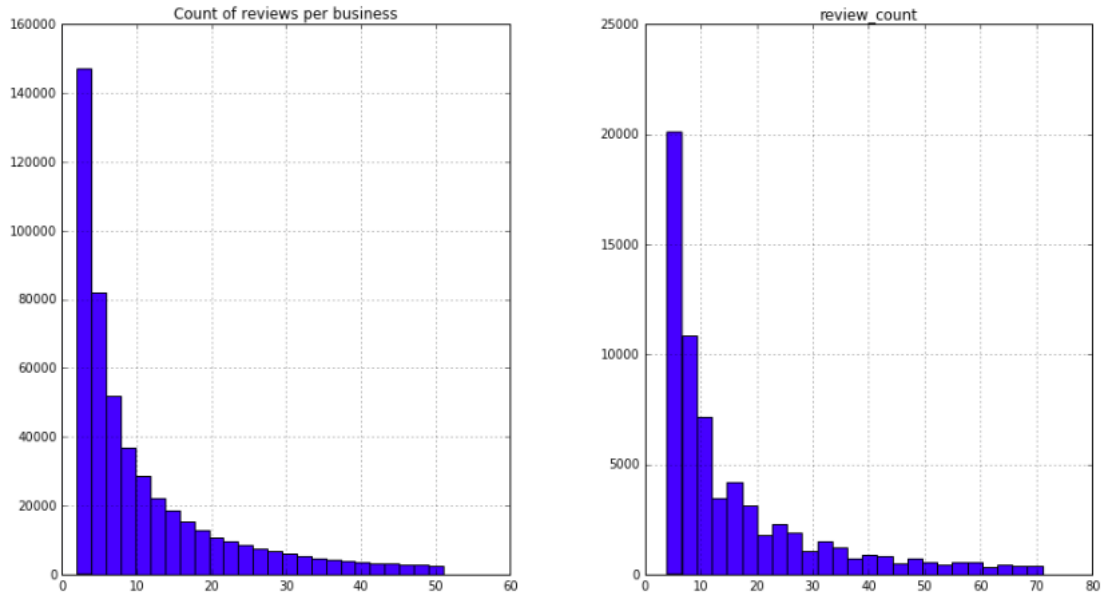


Figure 5: Exponentially distributed count of reviews received per business (left) and count of reviews given per user (right)

Time based Exploration

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

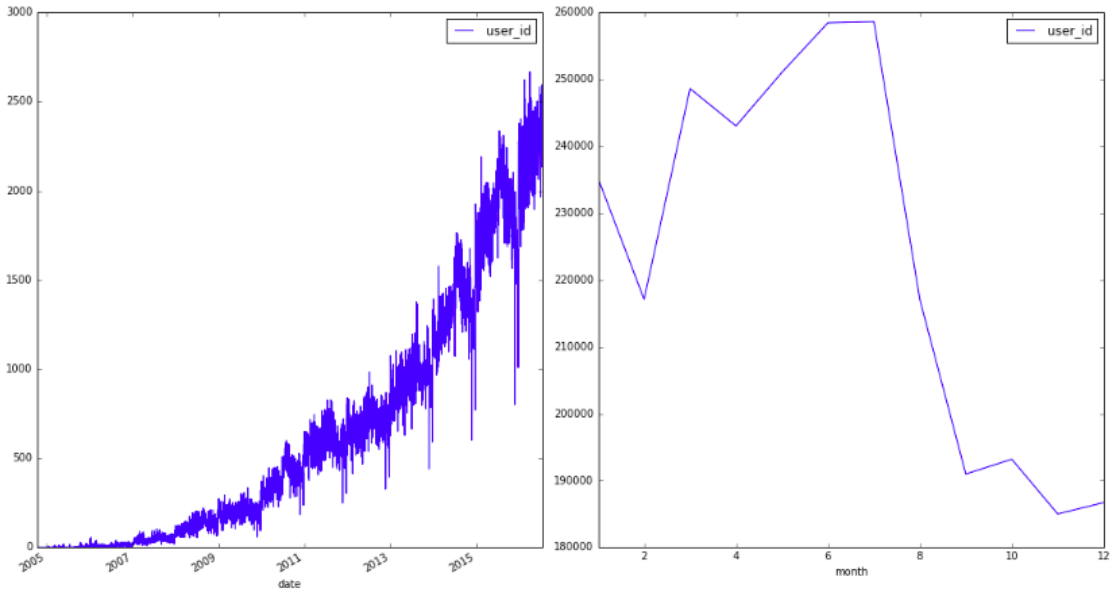


Figure 6: Number of reviews per day on yelp over time

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.

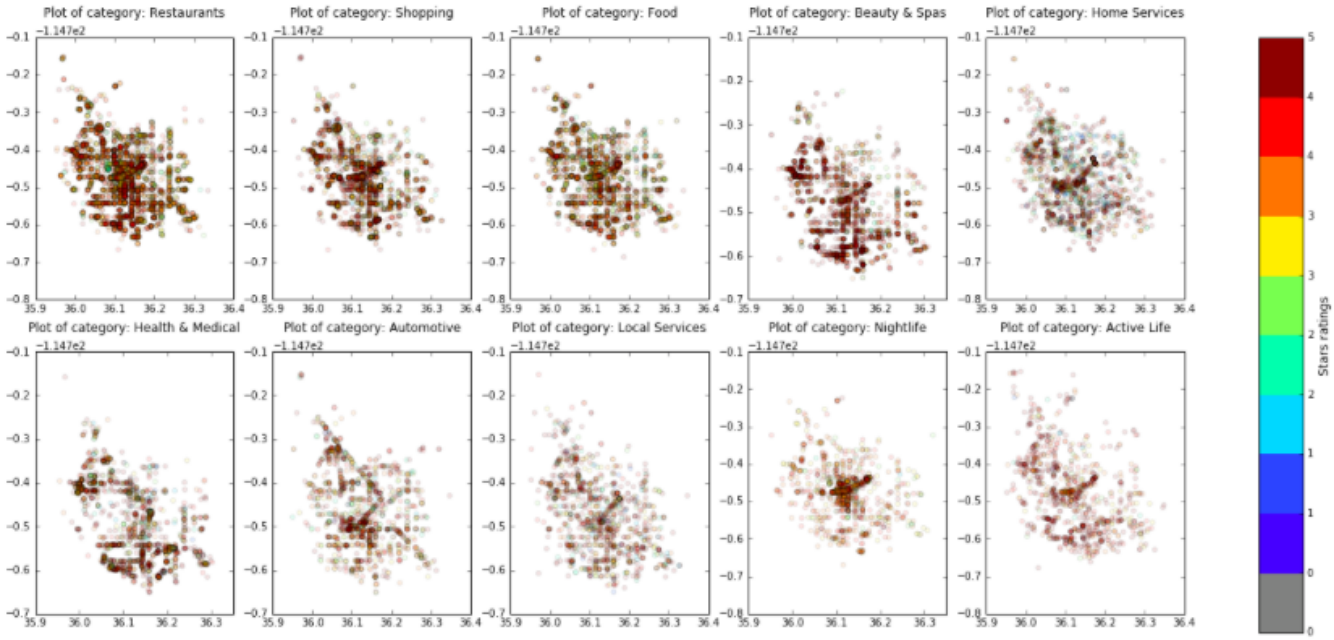


Figure 7: Chart showing the top 10 most popular business category plotted by location and colour coded by average review

Ratings based Exploration

We next explored the top rated businesses by introducing a prior model to the rating. We introduced a prior that most businesses will by rated at 2.5 and this has the effect of normalizing the distribution such that the higher rated businesses are required to have a large number of high ratings and similarly, the lower rated businesses are required to have a low number of ratings. The resulting distribution is shown below:

```
businesses_percentage_rating = businesses_sum_reviews.apply(sample_posterior, axis=1, args=(a, b,
n_samples))
businesses_percentage_rating.hist(bins=10)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x115df7c50>
```

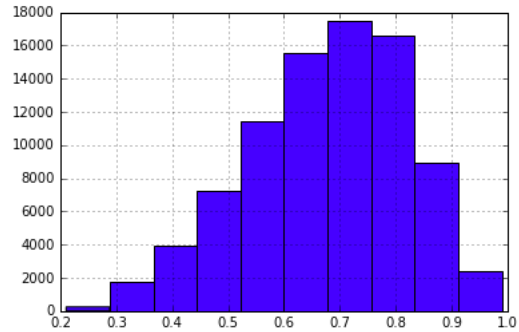


Figure 8: Chart showing the top biased businesses

Examples from the highest and lowest ranked businesses:

Table 1: Examples of low and highly ranked businesses

Lowest Ranked Businesses	**Lowest Ranked Businesses**
A Victory Inn	Blue Chip Auto Glass
OnTrac	Lockaid USA
Monitronics Security	Stell Roofing
Anjile Cleaning Service LLC	Simply Skin Las Vegas
Website Backup	Khina Eyebrow Threading & Henna Art

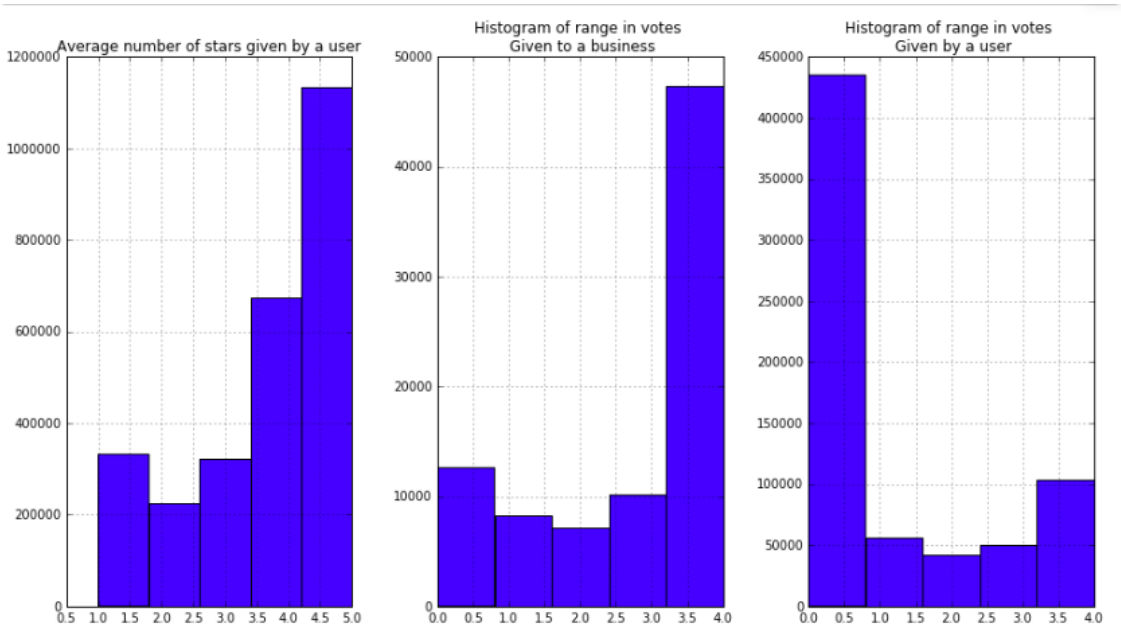


Figure 9: Plot of the average and range of ratings that a user gives and the range of ratings that a business receives.

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.

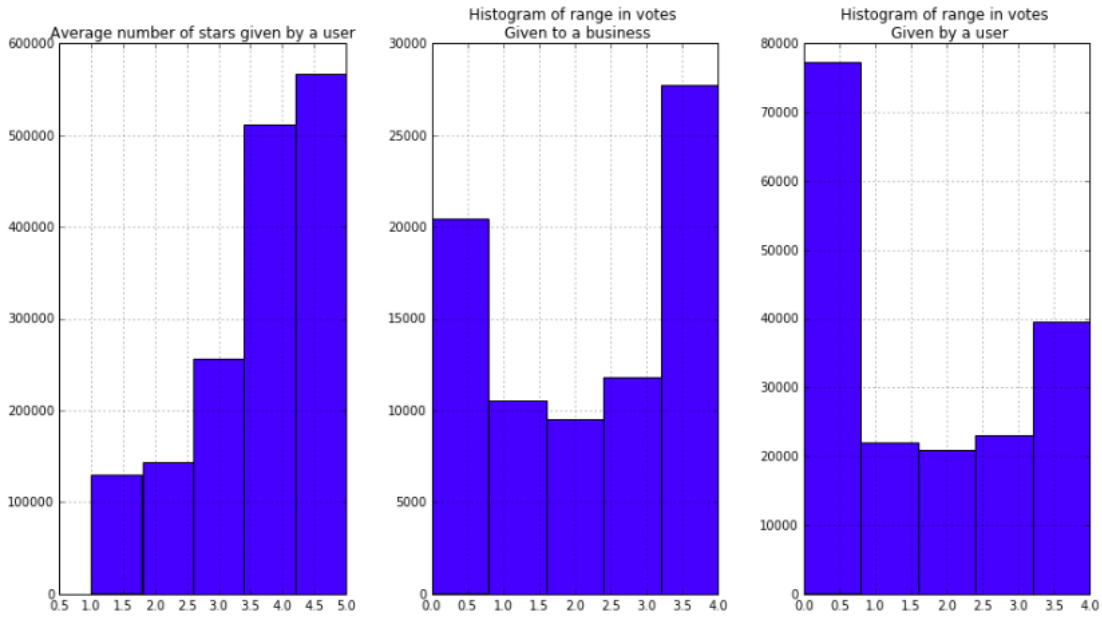
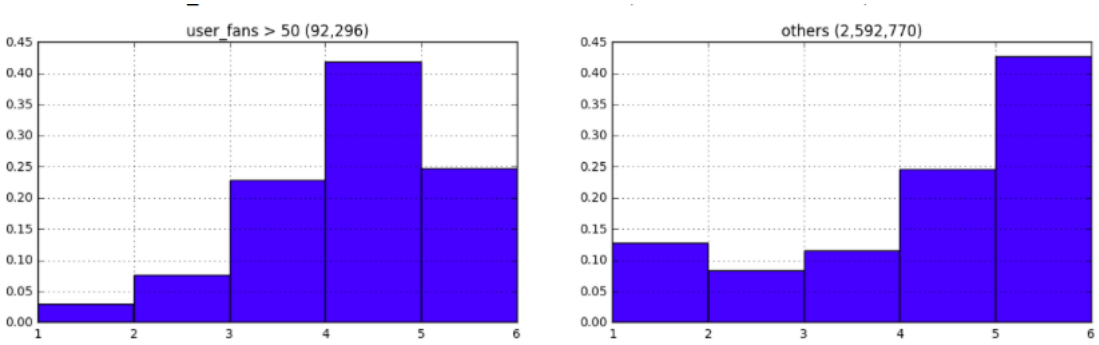
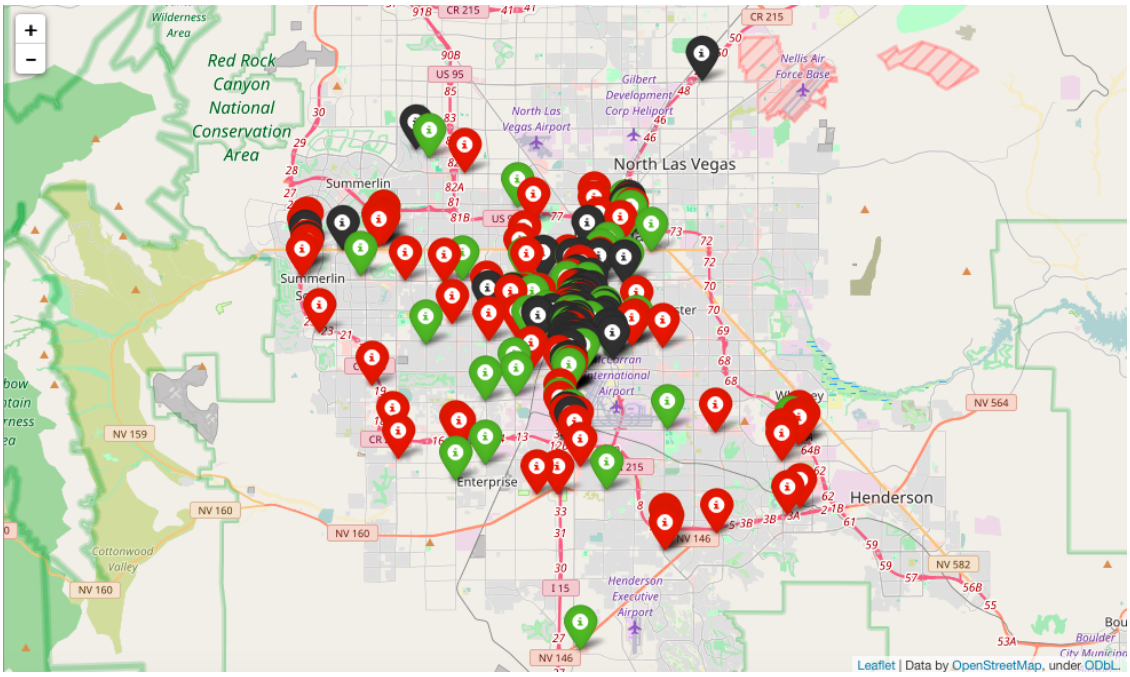
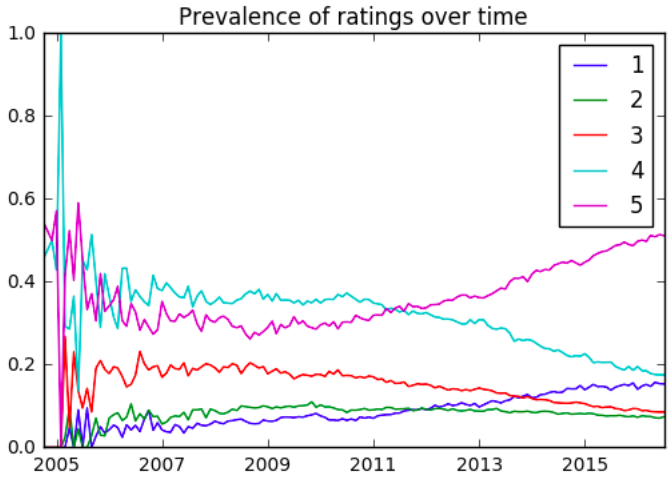

Average number of stars given by a user
Histogram of range in votes
Given to a business
Histogram of range in votes
Given by a user

Figure 10

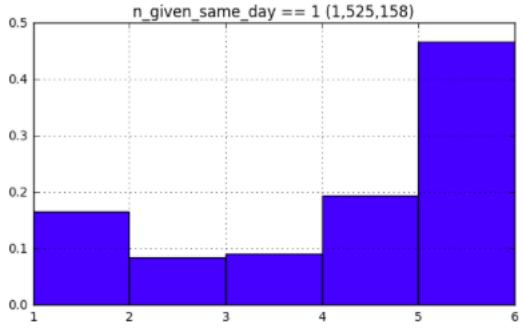

user_fans > 50 (92,296)
others (2,592,770)

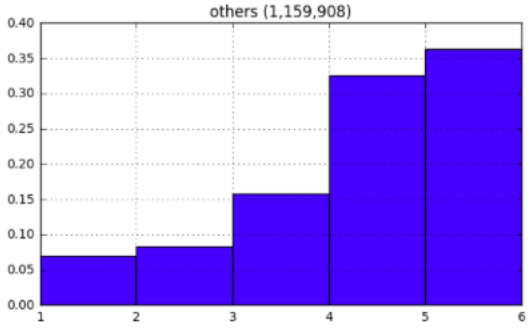

Red Rock Canyon National Conservation Area
Summerlin
North Las Vegas
Henderson
Enterprise
Nellis Air Force Base
North Las Vegas Airport
Henderson Executive Airport
Cottonwood Valley
NV 159
NV 160
NV 564
NV 582
NV 146
US 95
I 15
CR 215
Leaflet | Data by OpenStreetMap, under ODbL.

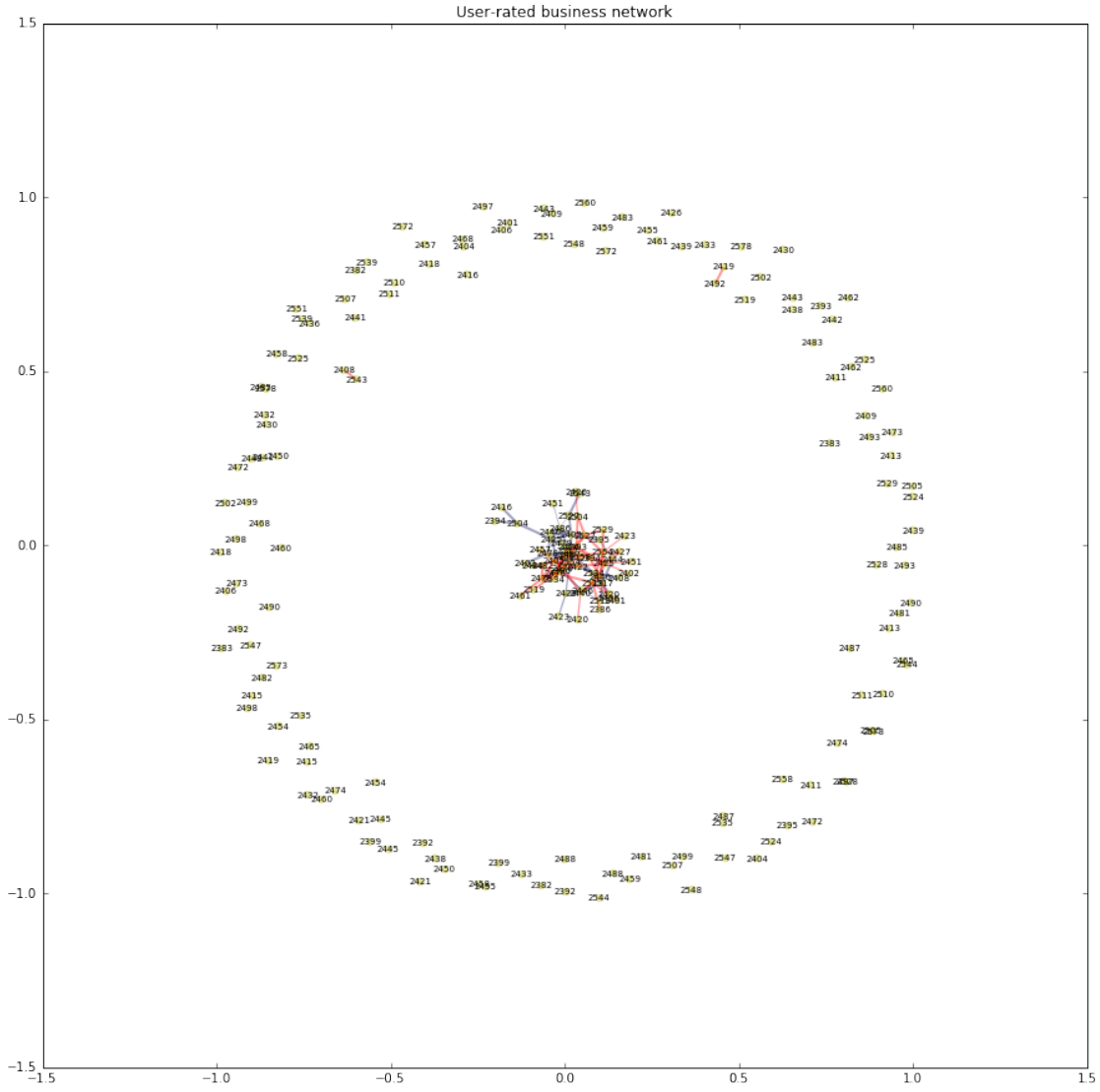

Prevalence of ratings over time
1
2
3
4
5
1.0
0.8
0.6
0.4
0.2
0.0
2005
2007
2009
2011
2013
2015

```
compare_star_distribution(reviews, 'n_given_same_day == 1')
```

```
1 star % for n_given_same_day == 1 is 0.16 vs. 0.07 for others (difference of 81.9%)
2 star % for n_given_same_day == 1 is 0.08 vs. 0.08 for others (difference of 1.0%)
3 star % for n_given_same_day == 1 is 0.09 vs. 0.16 for others (difference of -54.9%)
4 star % for n_given_same_day == 1 is 0.19 vs. 0.33 for others (difference of -50.5%)
5 star % for n_given_same_day == 1 is 0.47 vs. 0.36 for others (difference of 24.8%)
```







Now, we plotted average numbers of stars by the number of reviews of users according to the number of reviews in a business. Figures show a modest trend of having higher stars when a business was reviewed by many people. However, we don't see clear relationship among the number of reviews given by a user (or to a business) and the average rating that the business has (or user gives). We can further corroborate this result by analysing the standard deviation in the reviews that a business receives (and a user gives).

Using a prior that assumes businesses will not be enjoyed (we wish to overcompensate for the average high ratings of users), we are able to build a more comprehensive likelyhood and thus filter the review somewhat into a more confident top and bottom grouping. The ratings alone are unreliable and we rather wish to include the number of times a business has been rated in a particular manner to calculate the likelihood that this is a favorable (or not) business). For example, if a business is rated poorly twice, this is an unreliable statistic and we do not wish to penalise the business from a median likelihood too dramatically. On the contrary, if a business is rated poorly 100 times, we are confident that users rate the business poorly and we wish to have a low likelihood

for enjoyment. In the figure, we show a distribution that corresponds to a probability that this is a good business (i.e., users will like it). On the low end of the scale we have a small number of very poor businesses. We note that these businesses will have a large number of poor ratings. Similarly, on the high end, we see that there is a relatively small number of highly rated businesses. These businesses require a high number of very positive ratings to be rated into this category. We note that we used a fairly aggressive prior assumption that all businesses have a median probability for being liked that is less than 2.5. This is to overcompensate for the generally high ratings. We now have a more normally distributed dataset.

In earlier years (i.e., before 2008), the average stars varied very widely from 1 to 5 stars. In contrast, after 2009 onward, the average stars converged to around 4. This is mainly because we have more reviews over time and the mean value became less variable. The rating did not differ according to the month reviewed.

When we analyzed the number of repeated reviews by a user, we found that majority of reviews reviewed once.