Midterm Report (due 11:59 Wednesday 3/22/17)

Intro

Our team will be analyzing the dataset from the Yelp Challenge (https://www.yelp.com/dataset_challenge). We have decided to work with restaurants and businesses within Nevada (mostly centered around Las Vegas), for this places a feasible limit on the data's scope and is, in general, a familiar area. More specifically, we are interested in the following questions:

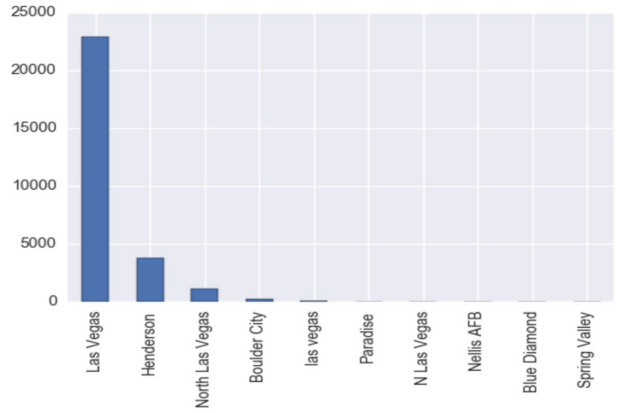
- Clustering users: can we cluster users into various groups, based on their ratings, reviews, and checkins? Can we find interesting clusters (e.g. social trendsetters, clustering based on location, etc?)
- Features that affect checkins/reviews on a restaurant (season, weather, location, users)

Data Exploration

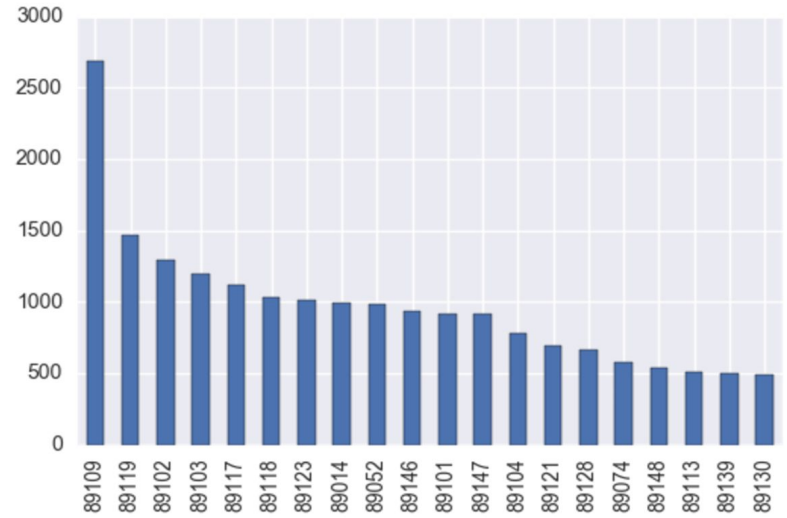
Data provided by Yelp
yelp_academic_dataset_business.json (size: 114.5 MB, records: 144,073)
yelp_academic_dataset_checkin.json (size: 46.2 MB, records: 125,533)
yelp_academic_dataset_review.json (size: 3.46 GB, records: 4,153,151)
yelp_academic_dataset_tip.json (size: 182.2 MB, records: 946,601)
yelp_academic_dataset_user.json (size: 1.18 GB, records: 1,029,433)

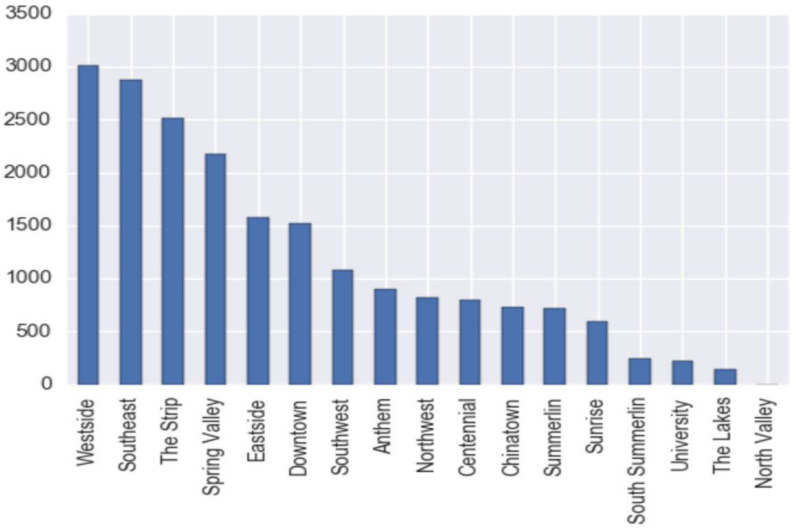
Businesses / Location (region)
In order to get a clear picture of how the restaurants and stores were distributed across Nevada, we ran several basic data explorations in various angles, toggling with zip codes, city and county, and map visualizations

- Our data is mostly concentrated in greater Las Vegas area (more than 90%, including Henderson, North Las Vegas):

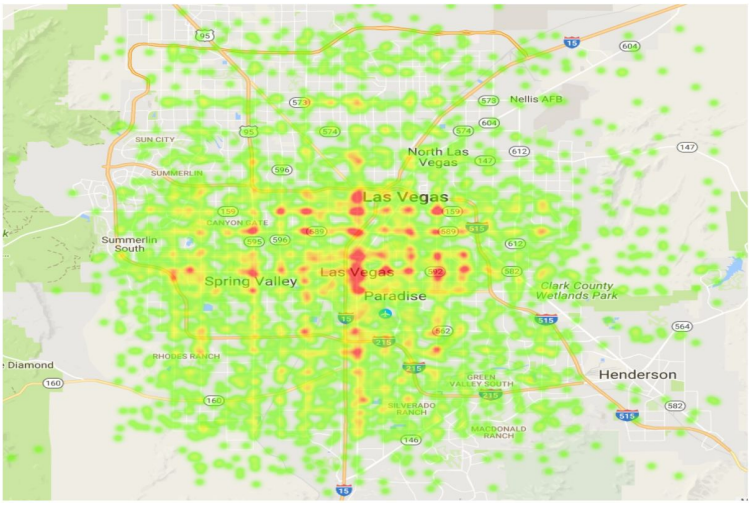


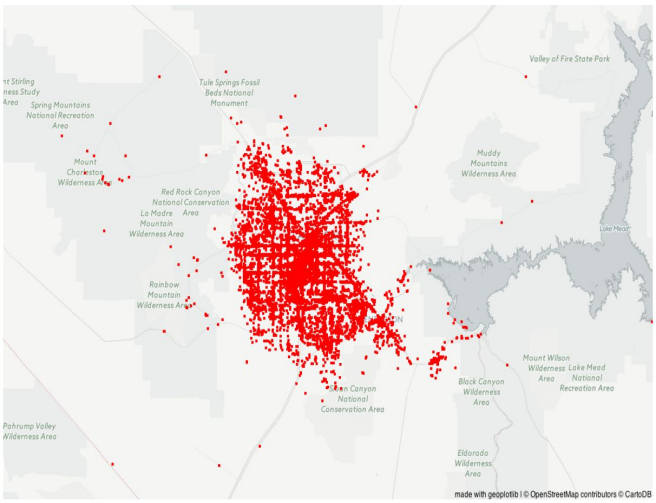
- Allocating stores by their zip codes and by region, we can see most restaurants/businesses are located in Downtown Las Vegas (including the Strip - 89109, 89118, 89119)





Our map visualizations tell us the same story: most of the businesses are located within Las Vegas, and more specifically, around the Strip:

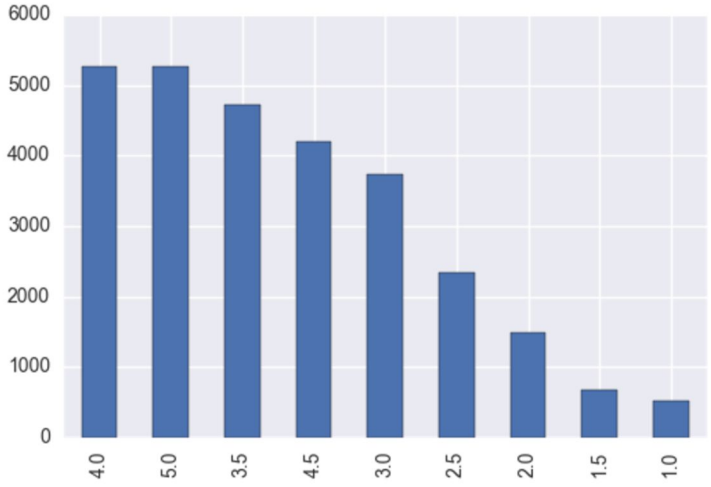




Word Cloud for Business Categories in Nevada

We generated a word cloud for business categories in Nevada. The size of each word demonstrates its frequency. As you can see, Restaurants, Food and Services are the top 3 most popular business categories, and for that reason we will be primarily focusing on the businesses in these categories.

Distribution of Ratings



Here, we see a basic distribution of ratings over the stores in Las Vegas region. There seem to be the most 4 and 5 star ratings and the least 1 and 1.5 star ratings. We plan to investigate this further in the following blog posts / analysis.

Most frequent words in 5-star and 1-star reviews (Naive Bayes Classifier)

	five_star_ratio	five_star_tokens	one_star_ratio	one_star_tokens
token				
ontrac	4.693521e-06	0.003735	0.001257	267.722368
frechheit	7.822535e-07	0.004527	0.000173	220.910402
rudest	1.877408e-05	0.004646	0.004041	215.256147
discriminates	7.822535e-07	0.004951	0.000158	201.975224
unprofessionally	2.346760e-06	0.005432	0.000432	184.092001
telemarketing	7.822535e-07	0.005559	0.000141	179.884184
unprofessional	1.853941e-04	0.005988	0.030960	166.994415
insinuating	1.564507e-06	0.006213	0.000252	160.949007
discusting	1.564507e-06	0.006888	0.000227	145.169692
transcripts	7.822535e-07	0.007042	0.000111	142.013830

	five_star_ratio	five_star_tokens	one_star_ratio	one_star_tokens
token				
delicioso	0.000250	101.398575	0.000002	0.009862
smoothest	0.000188	76.048932	0.000002	0.013149
foodgasm	0.000139	56.402958	0.000002	0.017730
deelish	0.000138	55.769216	0.000002	0.017931
eloff	0.000136	55.135475	0.000002	0.018137
yummmmmm	0.000110	44.361877	0.000002	0.022542
wac	0.000108	43.728136	0.000002	0.022869
addicting	0.001177	43.324846	0.000027	0.023081
uuu	0.000203	41.034736	0.000005	0.024370
gluch	0.000098	39.608818	0.000002	0.025247

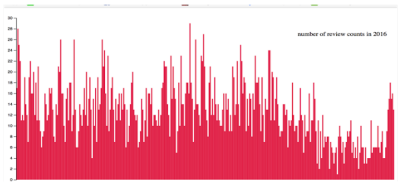
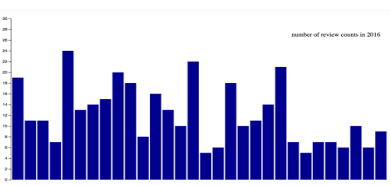
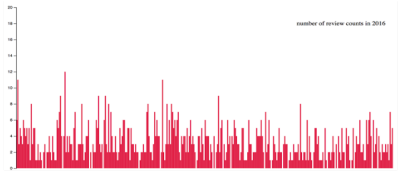
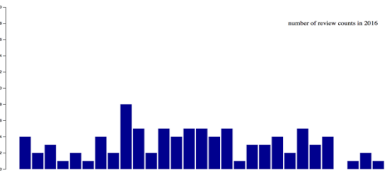
We wanted to find out which words are accounted the most for reviews with two extreme star values, 1 and 5. For the text analysis, we used sklearn.feature_extraction.text.CountVectorizer to fit and transform text data into a matrix of token counts. By fitting the model using Multinomial Naive Bayes classifier, we were able to get two interesting results for each case.

Top 10 one-star tokens: 1) ontrac, 2) frechheit, 3) rudent, 4) discriminates, 5) unprofessionally, 6) telemarketing, 7) unprofessional, 8) insinuating, 9) discusting, 10) transcripts

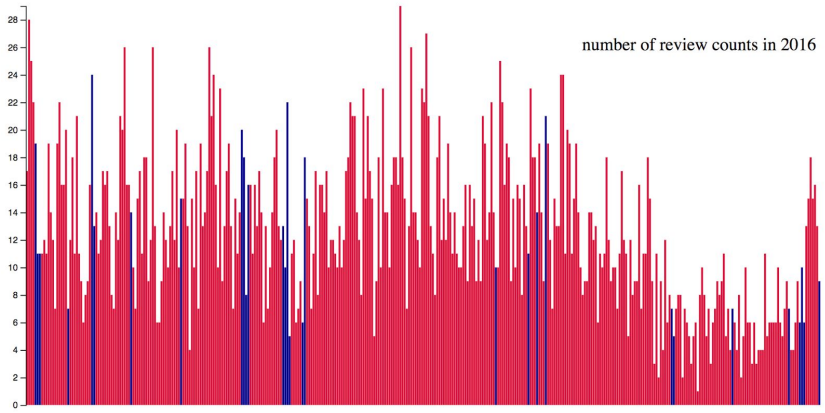
Top 10 five-star tokens: 1) delivioso, 2) smoothest, 3) foodgasm, 4) deelish, 5) eloff, 6) yummmmmm, 7) wac, 8) addicting, 9) uuu, 10) gluch

Visualization

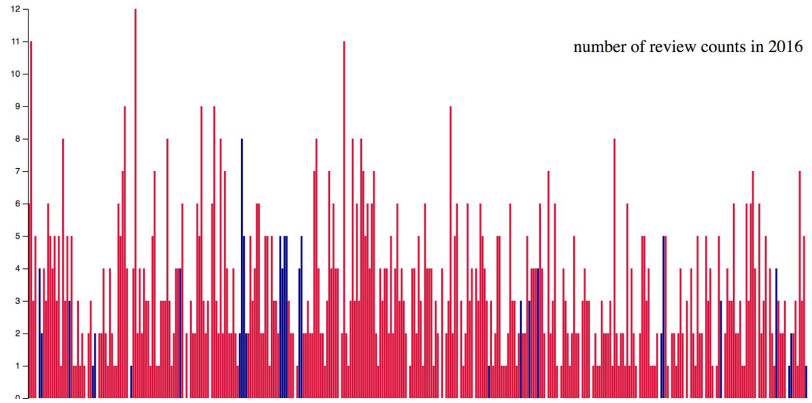
We wanted to find out which attributes affect the success of the business. Among various attributes, we first hypothesized that restaurants with delivery services will get more customers thus receive more reviews when weather is not clear. To make a sample group to test, we categorized restaurants by delivery option, and chose 5 business with most reviews from each group. With the open weather dataset provided by NOAA, we sorted review counts according to weather in 2016. Following bar-chart is the result of that.

	Number of Reviews on Sunny Days in 2016	Number of Reviews on Rainy/Snowy Days in 2016
Top 5 Restaurants without delivery service		
Top 5 Restaurants with delivery service		

Even though there was limited data on rainy days in 2016, we wanted to see the overall flow and changes of the review counts, so we also tried to combine data on rainy days and sunny days and identify pattern. For the hypothesis to be true, restaurants without delivery service should have higher drop rates on rainy days compared to restaurants with delivery service.



<Restaurants without Delivery>



<Restaurants with Delivery>

With the tested sample, it is hard to tell whether delivery service attracts more customers on rainy days. To prove our hypothesis, we have to expand the range of our exploration, so that we can try more direct comparison.

Other

- What is hardest part of the project that you've encountered so far?

The hardest part of the project so far was the preparation and cleaning. For example, the yelp dataset values contained objects (that contain many other objects), but the most challenging part was that it was inconsistent. There were double quotes, single quotes, no quotes, etc and it took us a while to clean the data and retrieve the attributes / categories. Also, it was difficult to make sense of the data: for example, the check-in data provided by Yelp are in the format Monday-1:3, which amazingly does not give us the date of the checkin (or the range of dates the check-ins were recorded).

- What are your initial insights?

As discussed above, we have done a lot of basic exploratory analysis. Most of the data we are working with are focused on a specific region (Las Vegas) and the reviews seem to indicate a near bipolarity (either a 5 or a 1).

- Are there any concrete results you can show at this point? If not, why not?

Not yet. Concrete results will be available after we apply some machine learning algorithms to build and test some models. Up until this point, we primarily focused on data preprocessing and exploratory data analysis really trying to grasp the overall picture of our datasets.

- Going forward, what are the current biggest problems you're facing?

The current biggest problem we are facing is the lack of feasible check-in data (we had originally planned on making extensive use of it) and our need to pivot our hypothesis to fit the available data.

- Do you think you are on track with your project? If not, what parts do you need to dedicate more time to?

Yes, I think the first few stages of the project were focused on data exploration and creating visualizations. I think the only step we have left is aggregating our exploration and analysis to come up with an exciting, yet viable hypothesis to test.

- Given your initial exploration of the data, is it worth proceeding with your project, why? If not, how are you going to change your project and why do you think it's better than your current results?

Yes, it is worth proceeding with the project. There are many interesting aspects about the Yelp dataset we can investigate (what links to higher # of reviews, higher ratings) and we believe the more exciting parts of the project lie ahead.