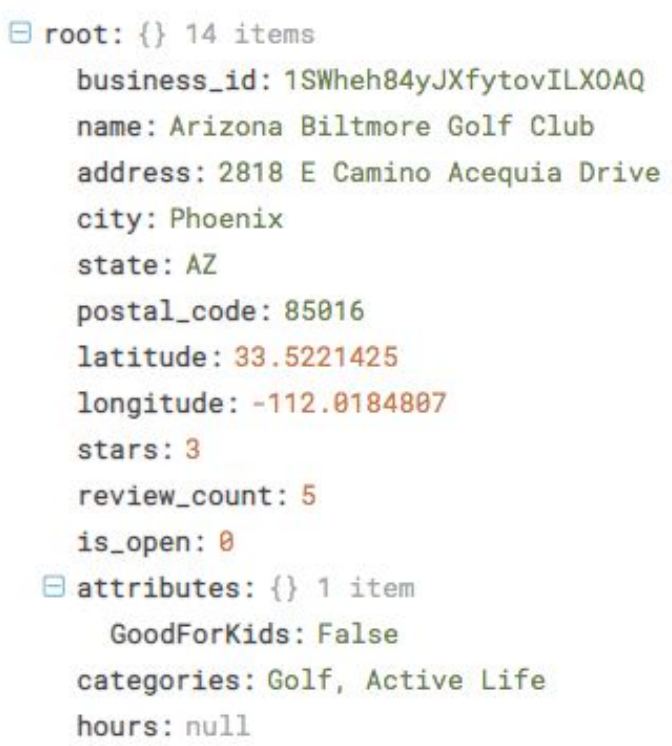
Yelp Dataset Analysis

Big Data Report

Group Members:
Yucheng Qian (yq791)
Hao Qin (hq411)
Zhihao Zhang (zz2347)
Chun Hao (ch2420)

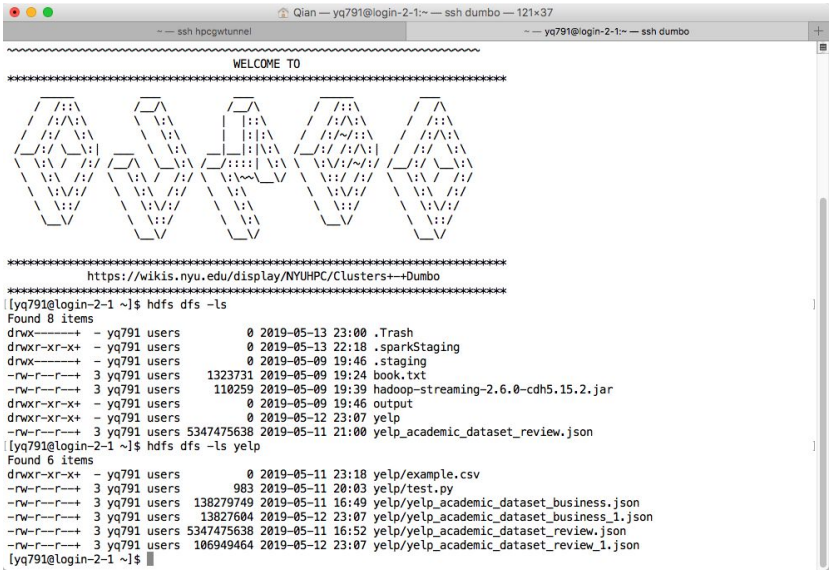
1.Background & Data Source

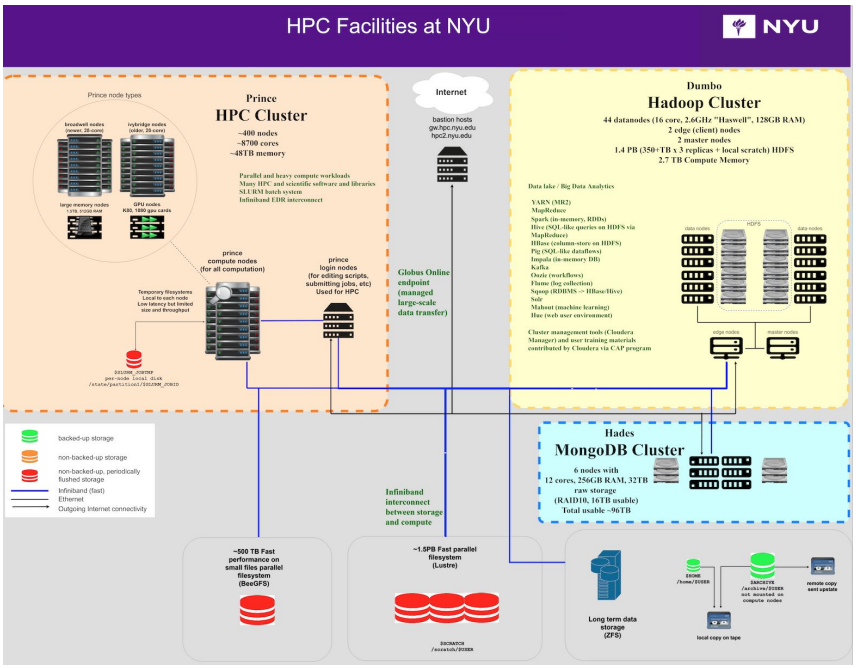
- We explored and downloaded data from Kaggle - Yelp Dataset.
- Dataset link: https://www.kaggle.com/yelp-dataset/yelp-dataset
- Dataset contains 5 json files: business, check in, review, tip, user.
- Total volume of json data: over 8G.
- A snippet of 'Business' data --->



2.Big Data Environment

- We accessed NYU HPC (Dumbo) to build our big data environment.
- Dataset is stored in HDFS.
- Run Python script with PySpark to generate results.



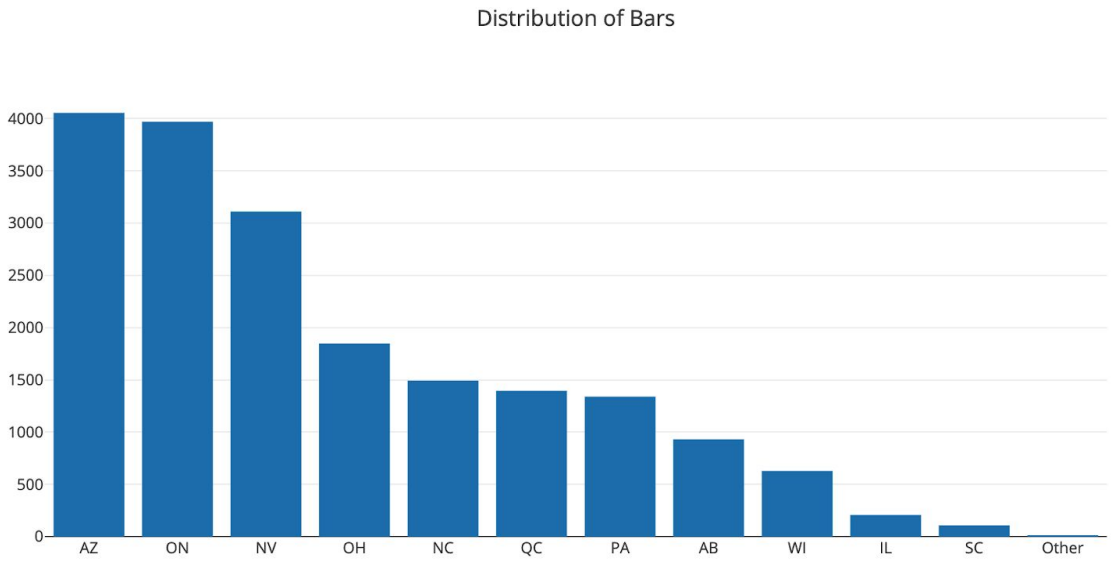


3.Data Analysis and Results

Library & Framework

Spark SQL

Natural Language
Tool Kit (NLTK)
Basic Text Analytics



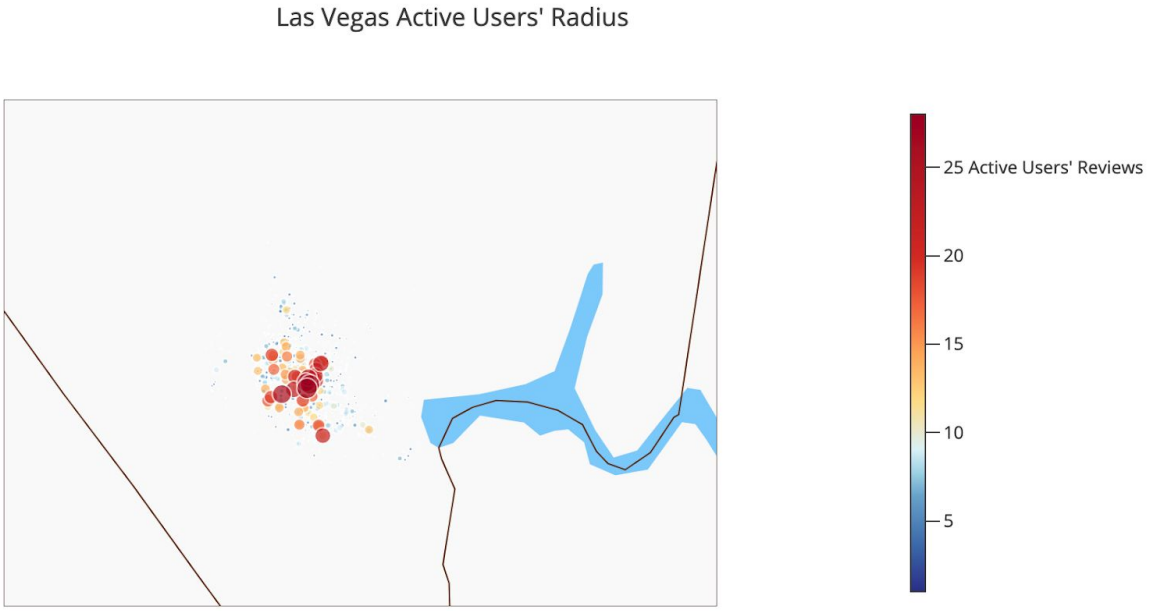
This graph represents the distribution of bars in North America according to our dataset. We use SparkSQL to analyse the data and use the bar chart graph of Plotly to draw this graph. We split the categories attribute and we look for the key-word “bar”.Then we group the data by the states. We combined the states together whose number of bars are less than 100. To draw the graph, we put the result into the Plotly. According to this graph we found that most of bars distribute in Arizona and Ontario of Canada. If we want to make an advertisement about the new beer, we can firstly sell the beer in these areas.

This is word-cloud shows us the most popular words in all the reviews. At first, we use the pyspark sql functions to analyze the data. After that, we use the nltk library in python to filter all the words in review. We filter all of nouns and adjectives and we also set some stop words like I, the, we etc. We can notice that the most popular words are place food good, service etc. Because basically all of yelp data comes from restaurants, it is predictable that we have these words about restaurant.

After we figure out the popular words in the reviews, we diverted to the number of reviews.

This is a review heat map in United State. We group the reviews by area. Darker the

color of the area, more reviews it got.
As we can see in the chart, only few areas are colored. Because the dataset we used is not well distributed, all reviews concentrated on several specific areas. Like vegas, phenix.
Although the darkest red represent review numbers above 400 thousand, the hottest area, Las Vegas, actually can reach more than 2 million reviews.



After we find out the most popular area, we are curious about the popular people. What places do they usually go?
So, we chose the hottest area as an example, say Vegas, and try to figure out "where do the most active users in las vegas usually go?"
We select the users who wrote more than one thousand reviews as our active users, and search where they left their reviews, so that we can find out the active users radius.
As we can see in the chart, the active users mainly left their reviews in the centre of the city.

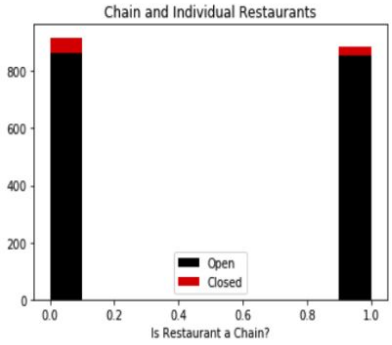
categories	name	review_count	city	state	stars
Comedy Clubs, Hotels & Trave	Jeff Civillico: Comedy In Act	208	Las Vegas	NV	5
Restaurants, Desserts, Food,	Tasty Crepes	260	Las Vegas	NV	4.5
Beer, Wine & Spirits, Bars,	Azuza Hookah Lounge & Cafe	273	Las Vegas	NV	4.5
Playgrounds, Active Life, Pa	Kangamoo Indoor Playground	242	Las Vegas	NV	4.5
Movers, Home Services, Auton	Christopher Moving	212	Las Vegas	NV	4.5
Breakfast & Brunch, Diners,	Lou's Diner	324	Las Vegas	NV	4.5
Argentine, Delis, Specialty	Rincon De Buenos Aires	297	Las Vegas	NV	4.5
Mexican, Restaurants, Breakf	Jefe's Taco Shop	466	Las Vegas	NV	4.5
Asian Fusion, Fast Food, Res	Soho Sushi Burrito	225	Las Vegas	NV	4.5
Burgers, Fast Food, Restaura	In-N-Out Burger	372	Las Vegas	NV	4.5
Parks, Playgrounds, Active L	Exploration Park	205	Las Vegas	NV	4.5
Oil Change Stations, Automot	Superior Tire - Goodyear Auto	259	Las Vegas	NV	4.5
Nightlife, Comedy Clubs, Adu	The Mac King Comedy Magic Sho	492	Las Vegas	NV	4.5
Beer, Wine & Spirits, Food,	Fat Tuesday	428	Las Vegas	NV	4.5
Middle Eastern, Restaurants	Amano Bakery & Deli	317	Las Vegas	NV	4.5

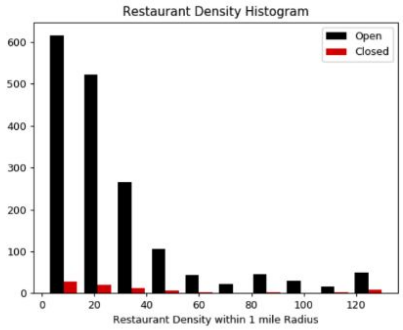
We already learned much about the hottest word, the hottest place, the hottest user.
Now, what if somebody wanna find a quality place, but not so hot, so that they can enjoy the place in a quiet, less busy environment.
Here are some small but special places we found in dataset by using a filter “stars above 4.5’ and ‘review number lower than 500”.

4.Machine Learning - Predict if a restaurant is likely to close in the next 6 years (2013-2019)

We try to make the dataset much more meaningful. So we came up with an idea with using Machine Learning Model to predict if a restaurant is likely to close in the next 6 years.
In reality, to those entrepreneur, before start a restaurant, it is important to know that what factors might lead to a success or failure, factors like chain, density, price, and so on.
To those investors, it is more likely to make a good decision if they know if invest to those restaurants might be a failure.
The dataset we use is an yelp dataset released in 2013. It contains some originals field but it provided a very poor ability of prediction.
We need a new generated datasets because we decided to only work with the restaurant that is currently foundable.
Given an dataset that was released in 2013, we came up with an idea which is to find if a restaurant is still open in 2019. We use yelp search api to solve this problem but only 68% can still be found. Because a restaurant might be closed so it is not searchable or because it it has changed its name. To those not searchable

ones, we use google search api to extract its business_id and use this business_id to query yelp business api to extract its information. In total we found 23% of the restaurants were actually closed.
The original dataset fields have very poor predictive ability, so we consider generating more meaningful features like chain, density, z-scores, age, and so forth.
We consider restaurant as a chain if its name appears more than once in the business datasets. After analysis we generate graphs shows that if a restaurant is a chain it's much likely to be existed.
We also defines density as the count of restaurant within 1 mile radius. From the picture we see density means traffic, which means it can bring more people here. It also means competition.

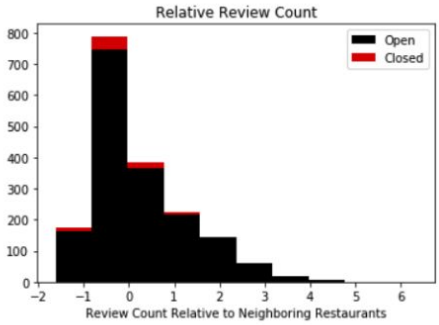




- Chain is much likely to win

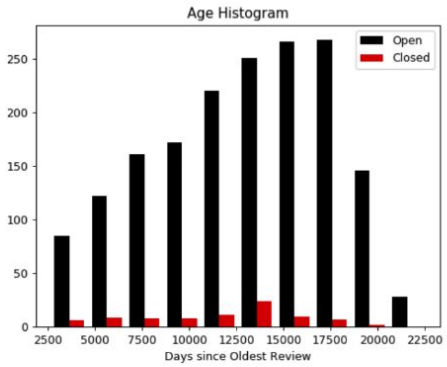
- Density means traffic / competition

Simply z-score is the number of standard deviations from the mean a data point is.
We calculated Z-scores to represent Relative values with surrounding restaurants used in review count, star rating and price and so on by subtracting the mean of this group of restaurants from each individual restaurant and dividing with the standard deviation of the value for this group of restaurants.
We also mark the oldest review from now of each individual restaurant as the age of the restaurant.



- Z-View-Count

More views mean more traffic

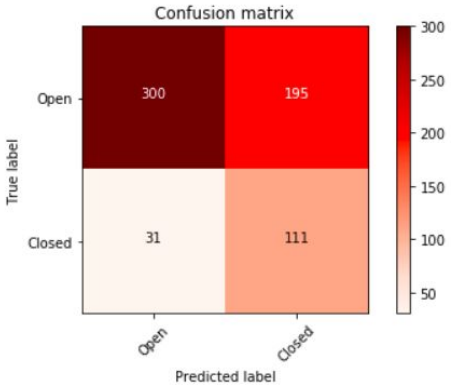


Age

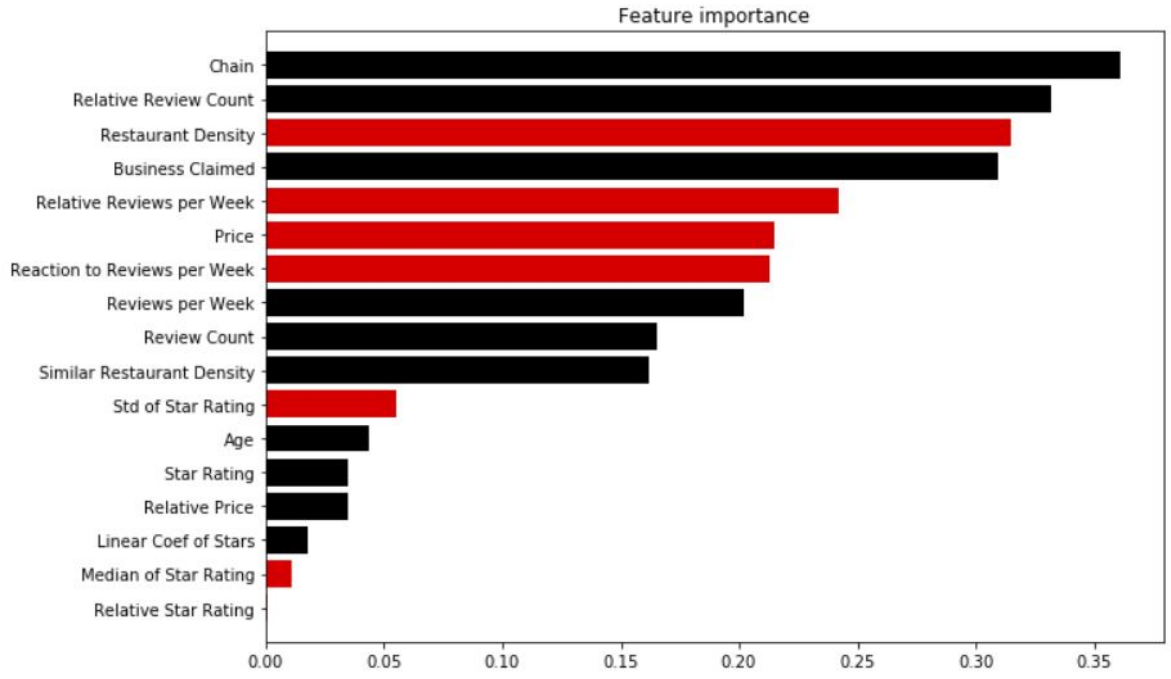
After feature engineering we built machine learning models using logistic regression which is simple and has good interpretability. We split data with 8 and 2. And choose grid search with cross-validation for optimization the parameters of Logistic regression.
One thing we need to mention here is that our goal is to determine which restaurant are much more worthy to be invest so we need to minimize the number of False Positive and focus on true positive.

We use Python (scikit) sai kit learn library and after optimization we get a precision of 90.6%.
Precision 90.6% means among the prediction of open restaurants, 90.6% of them are actually open, this is the result after we minimize the FP. In reality a bank would give loans based on this model would have a 6-year rates of 9%. However if it give loans Indiscriminately would have a 6-year rates of 23%.

```
Accuracy:  0.645211930926
Precision:  0.906344410876
Recall:  0.606060606061
F1 Score:  0.726392251816
Confusion Matrix:
 [[111  31]
 [195 300]]
```



We generated feature importance rankings afterwards. Black color here means open and red means close. We found that chain restaurant is the most important factors for a restaurant open. This is not surprising because always chaining has higher profit.
Review counts ranks the second because higher review counts means higher traffic. Restaurant density is pretty interesting. We found High restaurant density is negative for restaurant success, while high similar restaurant density is positive. For example if a chinese restaurants located in chinatown reduces the chance of failure than it locates in somewhere else.



5.Summary & Future Scope

- About Dataset: Not well-distributed.
- HPC Performance: not stable.
- Plotly SVG drawing functions have a hard time graphing more than 500k data

points for line charts, or 40k points for other types of charts.

- There are other reasons like surrounding venues can also increase the precision. The model can be improved with incorporating more datasets.

6.Code & Other Materials

https://github.com/Chandler-Qian/Yelp-Big-Data-Analysis