MA615 Final Project

Kerui Cao

12/13/2019

Data preparation Clean

I use *three dataset*, one contains the information of the number of imigrations of each states of the America, another one is the dataset contains millions of business information, the last one contains reviews about millions of restaurants, there are four sub datasets in total, I use "Business" dataset which contains the infomation of part of businesses listed on Yelp. There are 192609 businesses in this dataset, the main steps of data preparation and clean process are:

- Download the Imigration data;
- Download the Yelp Compititoin Dataset;
- **Using API** from Yelp to download review and match them to selected restaurants;
- Extract restaurant information from all these businesses;
- Extract restaurant with more than 100 reviews;
- Extract cities with more than 150 resturants, which can exactly give us 10 cities after filtering;
- Delete variables with more than 40% missing value;
- Reorganize some variables;
- Delete observations with missing values;

Variables "hours.Monday", "hours.Tuesday", "hours.Wednesday", "hours.Thursday", "hours.Friday", "hours.Saturday", "hours.Sunday" indicate the operational hours for each businesses, we don't need these information, so we delete these variables.

Some variables whose value are list, for example, the value of variables "attributes.GoodForMeal" is ***{'dessert': False, 'latenight': False, 'lunch': True, 'dinner': True, 'brunch': False, 'breakfast': False}***, we simple re-define this variables as a numeric score by counting how many "True" contained, for the example above, the re-defined value is 2. We do the same for variables "attributes.BusinessParking" and "attributes.Ambience".

For variable "categories", we can see that some restaurant contains words like "Chinese", "French" and so on, so I created new binary variables indicating those informations.

Some binary variables contain value "True" and "False" and "None", but only a small part of them are "None", so I simply delete them.

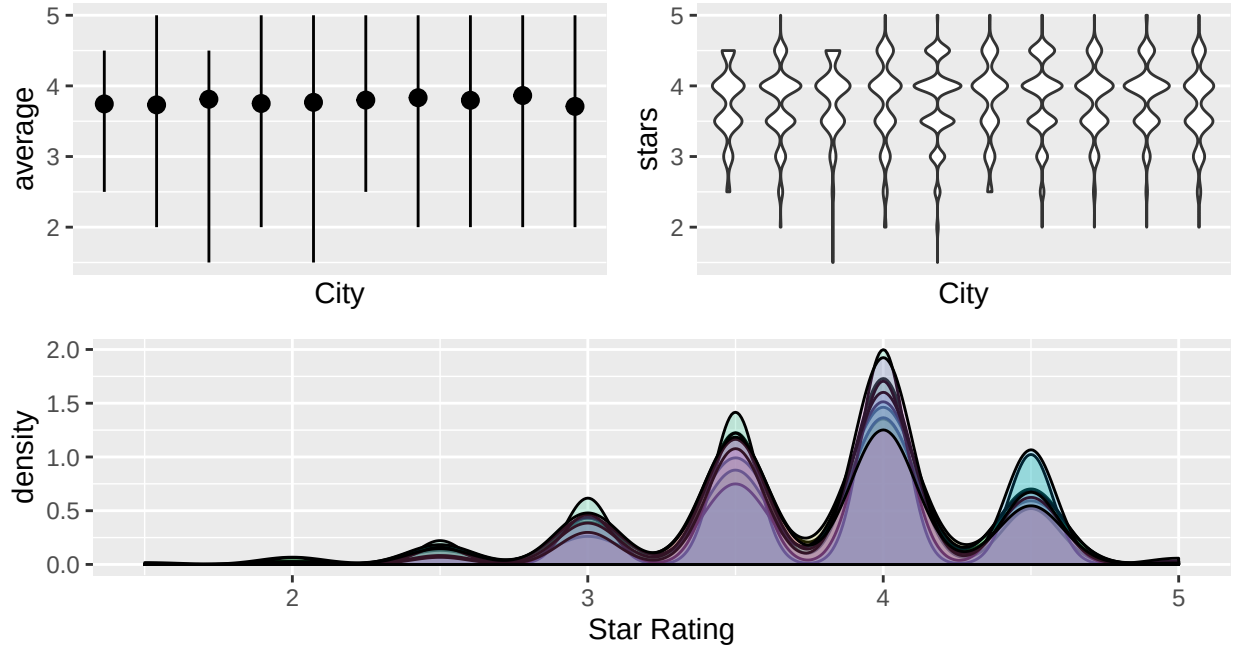
Some variables contain value like "u'average"', we need to transforme it into only "average".

EDA

Distribution of Star ratings

Our research interest lies on the star rating and review for each restaurants, so we try to apply exploratory data analysis around star ratings of restaurants.

First we will see the distributions of star ratings of restaurants, and we are also interested in the difference of distributions across cities.



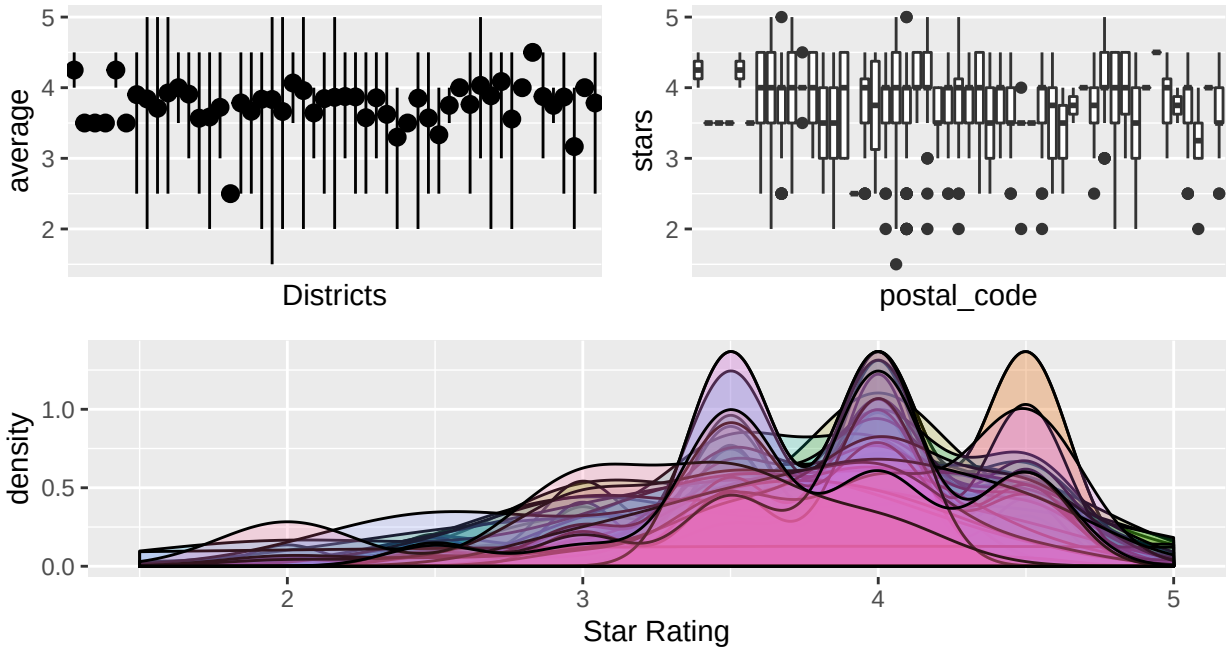
Above plot shows the distribution of star ratings, as for the upper left plot, black points are the average star ratings for selected ten city, vertical lines shows the range of star ratings, upper right plot is the violin plot of star ratings of each cities, lower plot shows the density of star ratings in for each cities. We can tell that restaurants in different cities have similar didtribution, which is centering at 4 stars, and barely seeing restaurants with lower than 2 stars. So we may consider that there is no difference between cities.

We consider that maybe city is not a good standard to separate and group restaurants, so I tried to separate and group restaurants by districts, which can be indicated by variable "postal_code", here I first tried restaurants in Las Vegas, because we have more date from restaurants in Las Vegas, which is 1916 restaurants, and I only pick districts with more than 30 restaurants, below is part of the list of chosen districts:

Table 1: List of districts

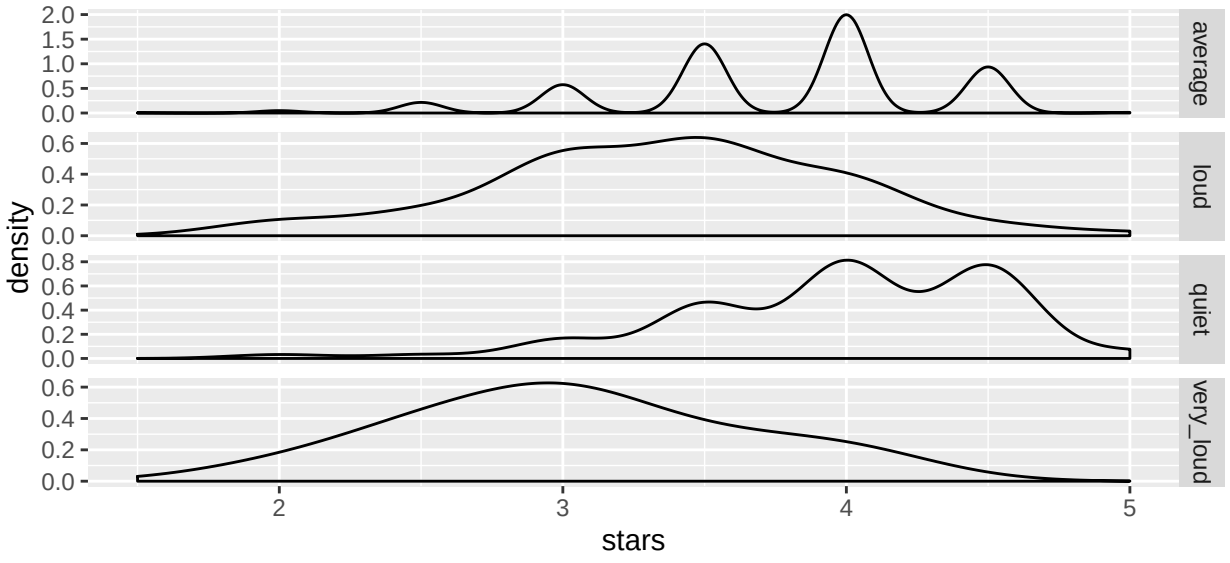
Zip.Code	Num of Restaurant
89109	375
89119	141
89102	113
89103	103
89117	95
89101	92
89123	84
89147	83
89146	77
89118	57

Do the same as grouping restaurants by cities, we drew the same plot shown below, from the plot we can tell that restaurants from different district have quite different distributions, which is more significant than the difference between grouping by cities, so district is a better standard to separate and group restaurants.



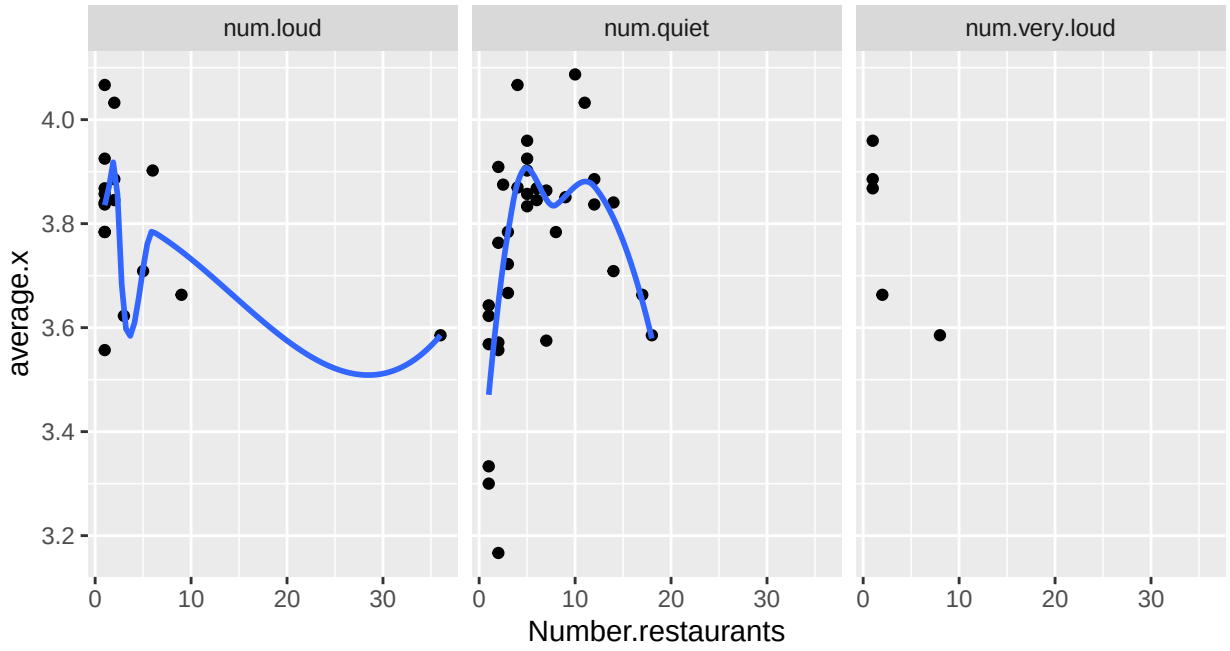
Than we will dive deeper into the data, we try to see the difference between distribution of star ratings between restaurants distinguished by features. I examined all 33 features, below are part of the result.

Noise Level



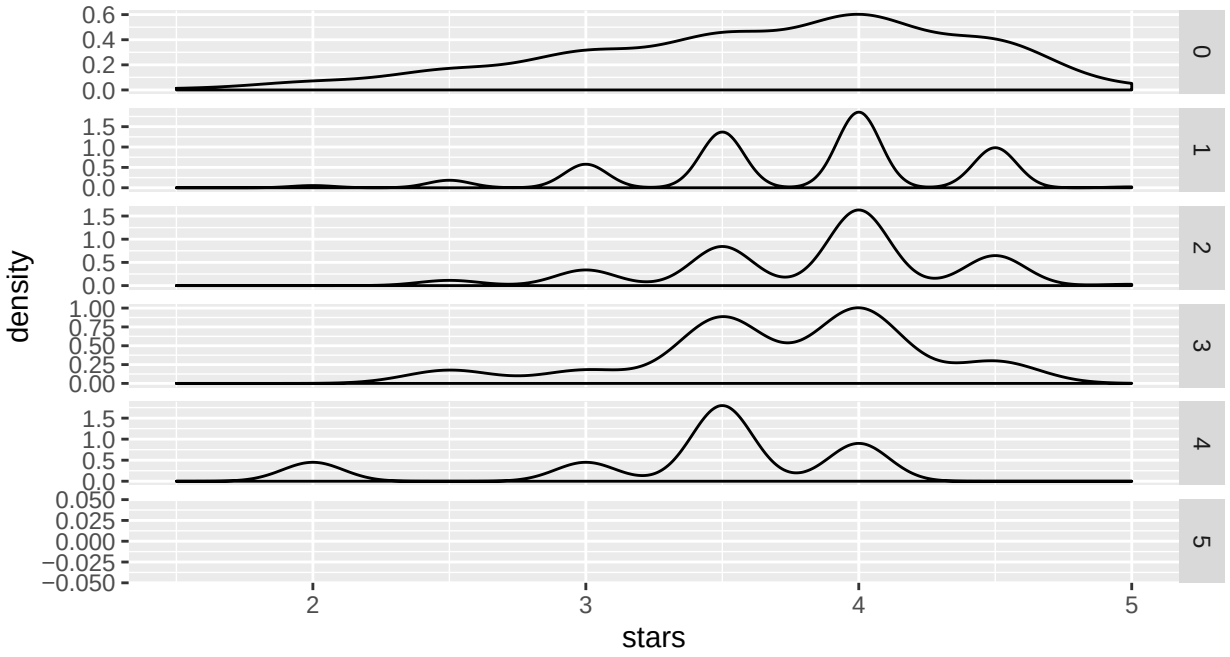
Above plot shows the distribution of stars of restaurants with different noise level, we can see clearly that, ignore difference between districts, as the noise level increase, the stars center at lower score, quiet restaurants center at 4, loud restaurants center at 3.

More detailed we want to see if the proportion of different kinds of noise level restautants in district with highest average star ratings are differnet with district with lowest star ratings.



We can see the if we look at the average star ratings of each district, as the number of loud restaurants increases, the average of star ratings of that districts display a clear trend going down, it is a little confusing that as the number of quiet restaurants increases, the average star ratings will go up and go down, but this may be affect by some other attributes of restaurants, we will continue explore attributes that will affect star ratings of restaurants.

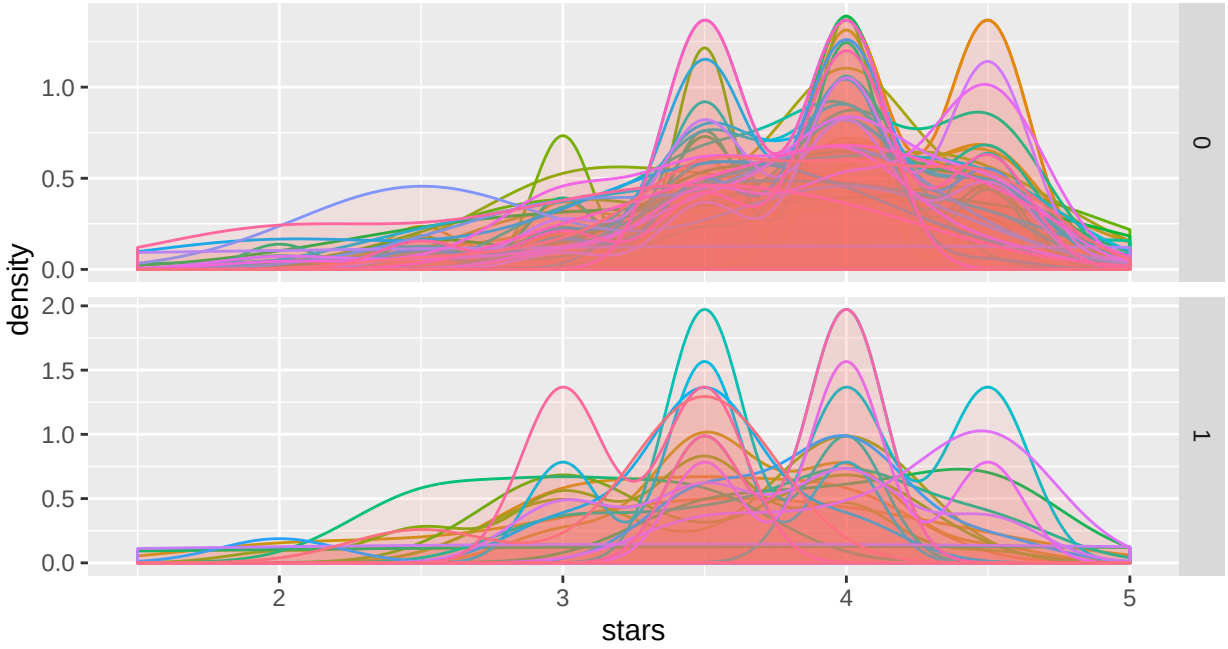
Business Parking



There are five types of business parking, garage, street,validated, lot and valet, if a restaurant has more parking choice, it will have higher score of business parking, above plot shows the distribution of stars of restaurants with different business parking score, here wield thing happend, according to plot, the

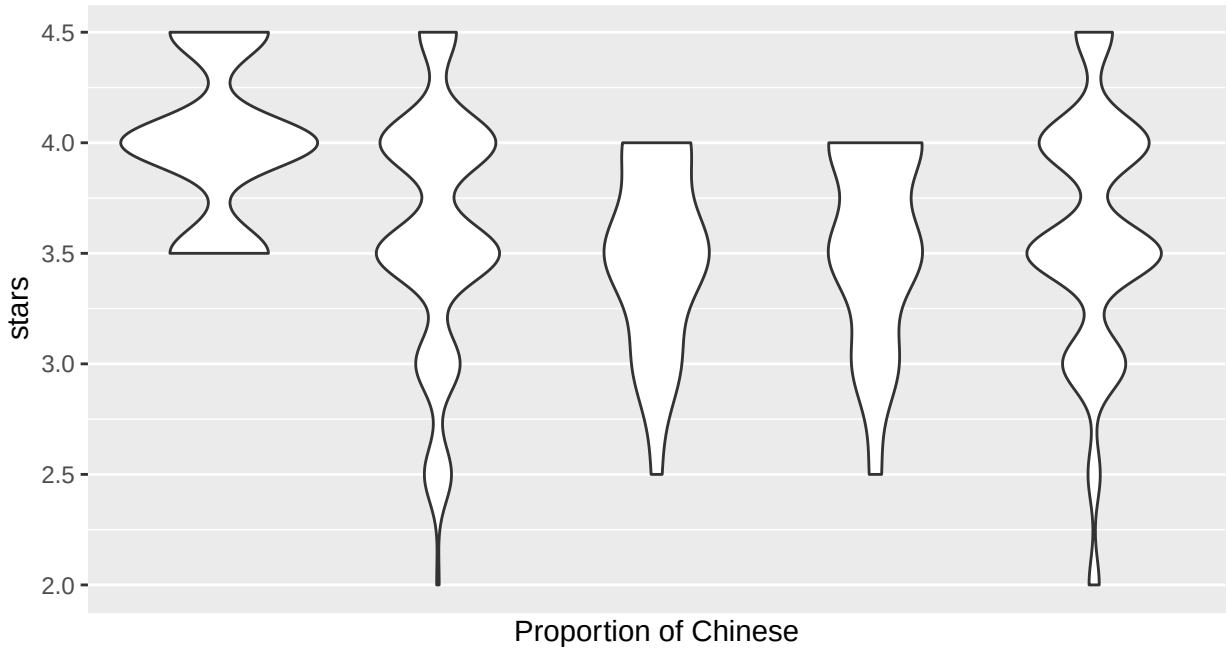
distributions of restaurants with lower business parking scores are centered at higher star ratings, which means the more convenient for parking, the lower star ratings, still this may be biased because we didn't control the influence from other variables, to fix this problem we may need construct models, which is out of the scope of EDA.

Chinese Restaurants



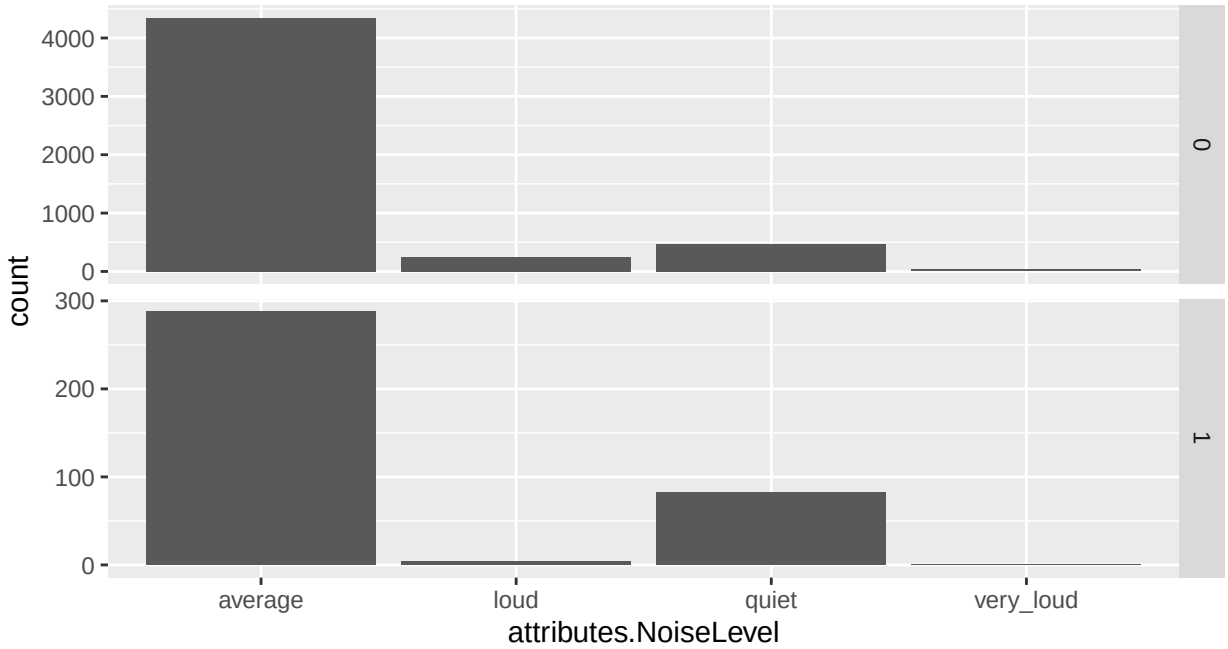
Above plot shows the distribution of star ratings of restaurant providing Chinese dishes and not providing Chinese dishes, we can see clearly that Chinese restaurants tend to have lower stars, as for the difference between districts, we can see that for some districts, Chinese restaurants center at around 4 stars rather than 3.5 stars.

I am a little interested about this phenomenon, so I will explore further. One explanation could be that proportion of Chinese in that city may influence the star ratiings of Chinese Restaurants, so I included a new dataset containing records of migrants numbers of each states, below is the scatter point plot of the proportion of Chinese in each cities and star ratings of restaurants.



Because we have restaurants data from only 5 states, so the number of unique values of proportion of Chinese is only 5, from above plot we can see that there is no clear pattern that indicating that the star ratings of Chinese restuarants are related to the proportion of Chinese imigrants.

To further explore the reason Chinese restaurants having lower stars, and based on previous findings, we suspect that noisy restaurants will have lower stars, below plot shows the distribution of noise level of restaurants providing Chinese dishes and not providing Chinese dishes, we can see that Chinese restaurants have similar distribution and even higher proportion of quiet restaurants, so there may be some other factors. To further analyze those factors affecting star ratings, we need some models to do this.



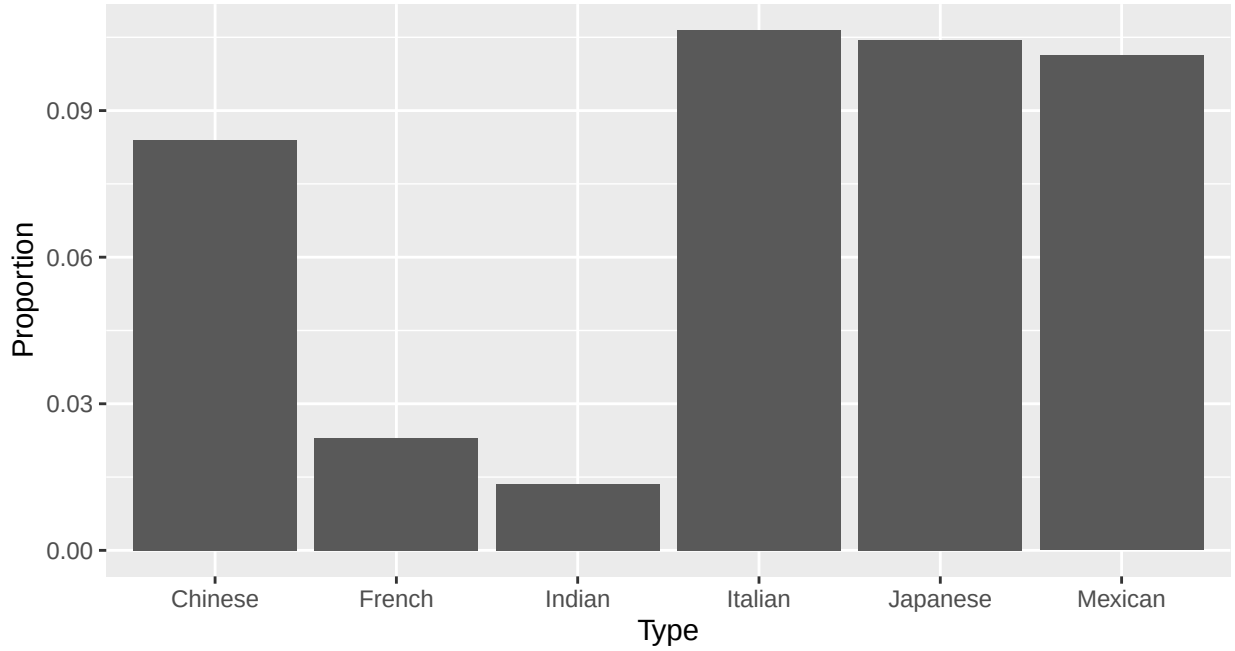
Text Mining on Review

the reviews of restaurants also interests us, so I collected the review of selected restaurants, but due to the limitation of Api, for each restuarant, I can only have 3 reviews, first we will analysis the most comon words appear in reviews.

Table 2: Most common two words appear together

word1	word2	n
las	vegas	253
customer	service	98
happy	hour	78
late	night	39
mexican	food	36
hidden	gem	35
fast	food	28
quick	bite	28

We can see that the most common words appear together is “Las Vegas”, because I choose all the restaurants from Las Vegas, as for the rest, such as “Customer Service”, “Happy Hour”, which means it is likely in most time people like the Las Vegas restaurants, and “Late Night” also tell an interesting story, because Las Vegas is a “Never-sleep-city”, and I also have a question, is there a lot Mexican restaurants in Las Vegas? So I exctract the category of all restaurants, I divided them into “Chinese”, “French”, “Mexican”, “Italian”, “Indian”, “Japanese” and “American”, below is the proportion of each kind of restaurants:



Now we can see that for sure there are a lot Mexican restaurants in Las Vegas.

We can also see the most seen three words that appears together.

Table 3: Te two words appear together

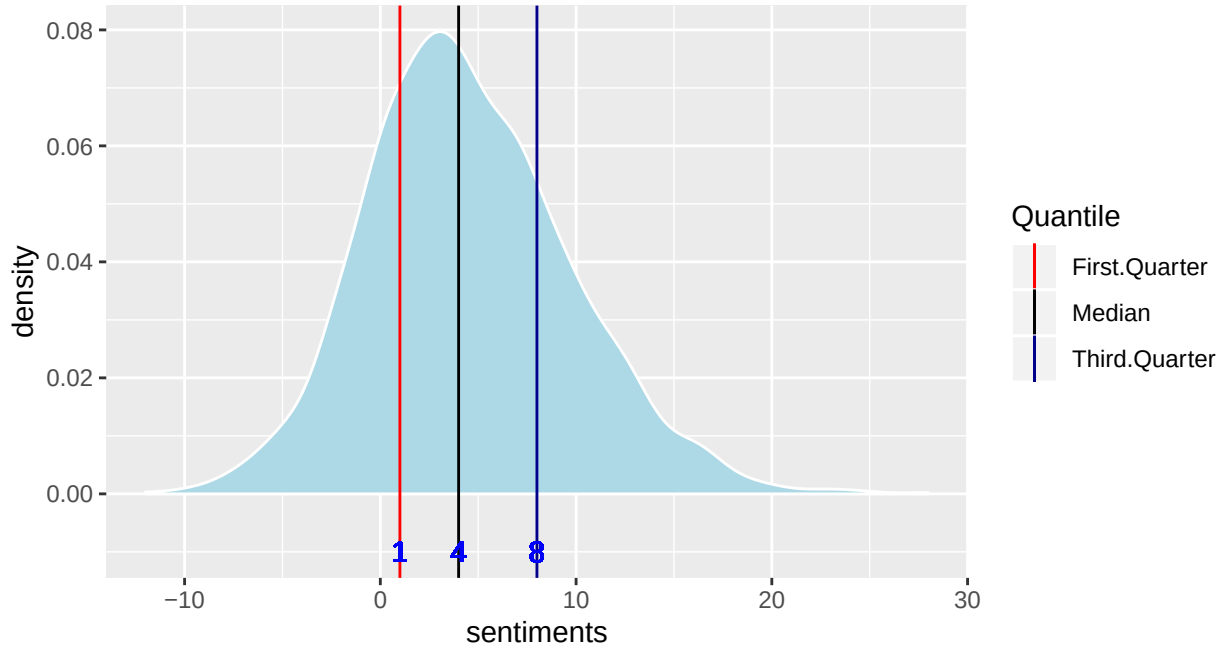
word1	word2	word3	n
love	love	love	12
excellent	customer	service	7
buffalo	wild	wings	5
downtown	las	vegas	5
las	vegas	strip	5
late	night	dinner	5
miracle	mile	shops	5
ayce	sushi	spot	4
birthday	celebration	week	4
carne	asada	fries	4

According to above table, we can tell that:

- Customer really love the restaurants;
- If we go to Las Vegas, we need try the Buffalo Wild Wings;
- A big proportion of customers of Las Vegas restaurants may be tourists;

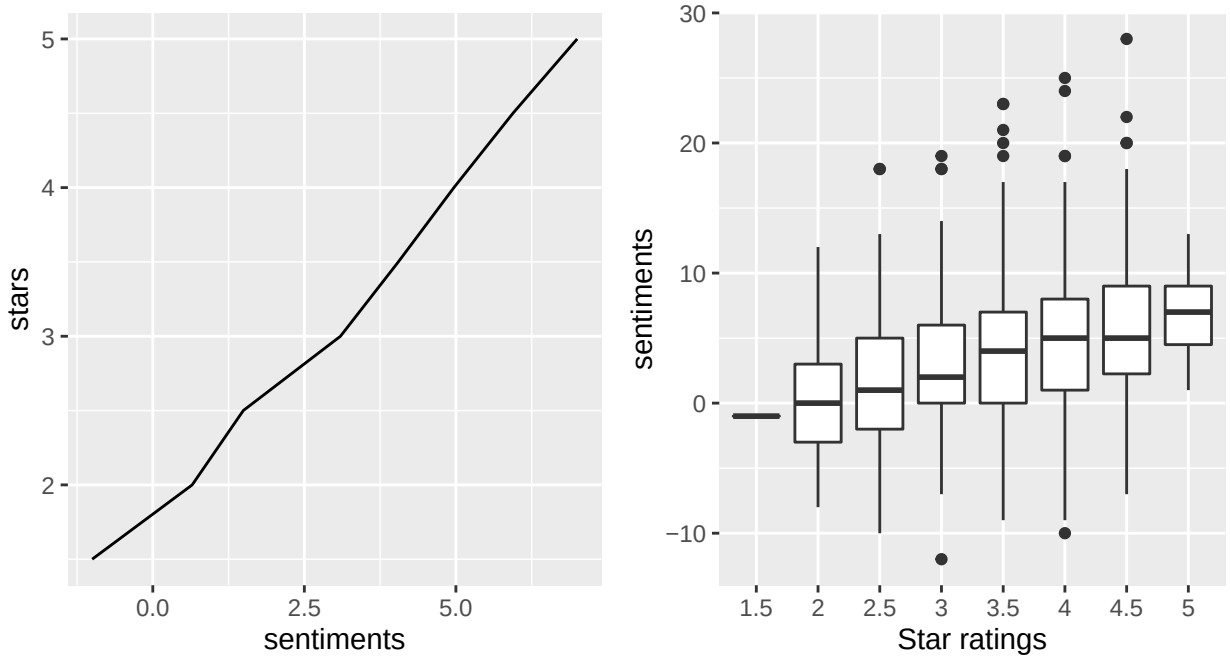
Sentiment Analysis

From above text analysis we see that customers love the restaurants in Las Vegas, so now I want to know more about do they really enjoy the local restaurants and how happy they are about the restaurants, so I applied sentiments analysis:



We can see that the 25% quantile of sentiments is 1, which means less than 25% feel bad about the restaurants, and also we can see that 75% quantile is 8, which means over 25% customers are very happy about the restaurants.

Next we want to explore the relationship between sentiments and star ratings:



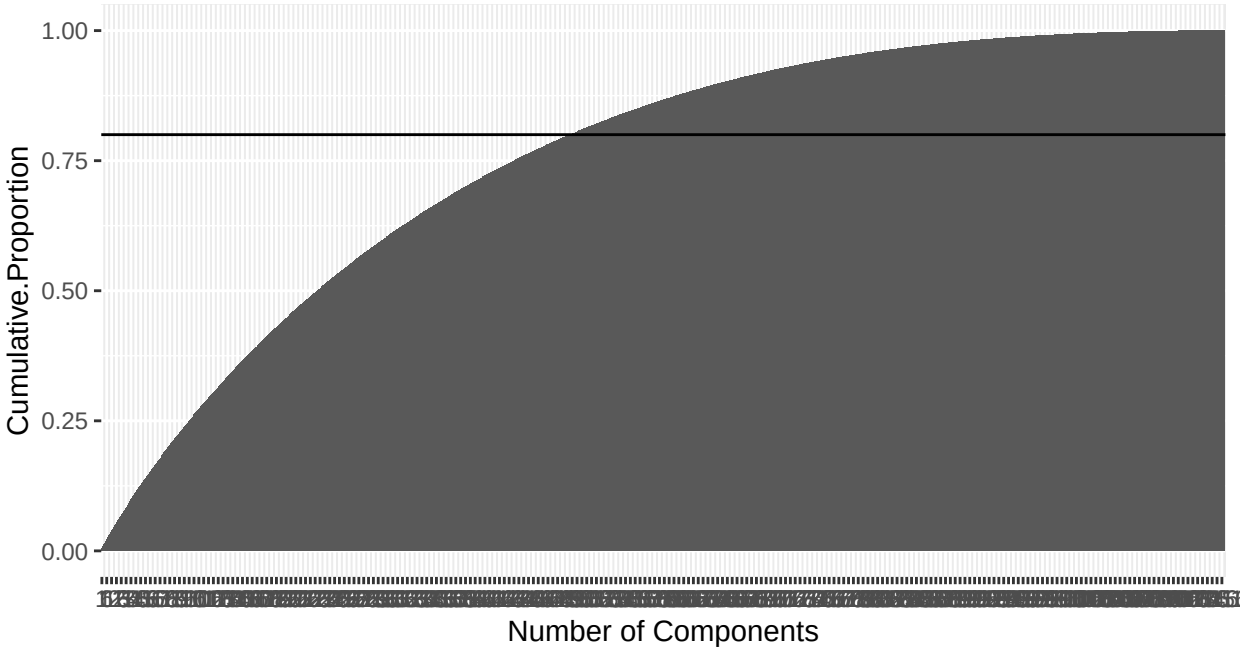
Above plot shows the higher sentiment, the higher star ratings, which perfectly make sence.

CPA and Cluster Analysis based on Text mining.

Now I want to go further, I want to see how can we cluster these restaurants with similar reviews, how does the cluster looks like and what is the common review within an cluster?

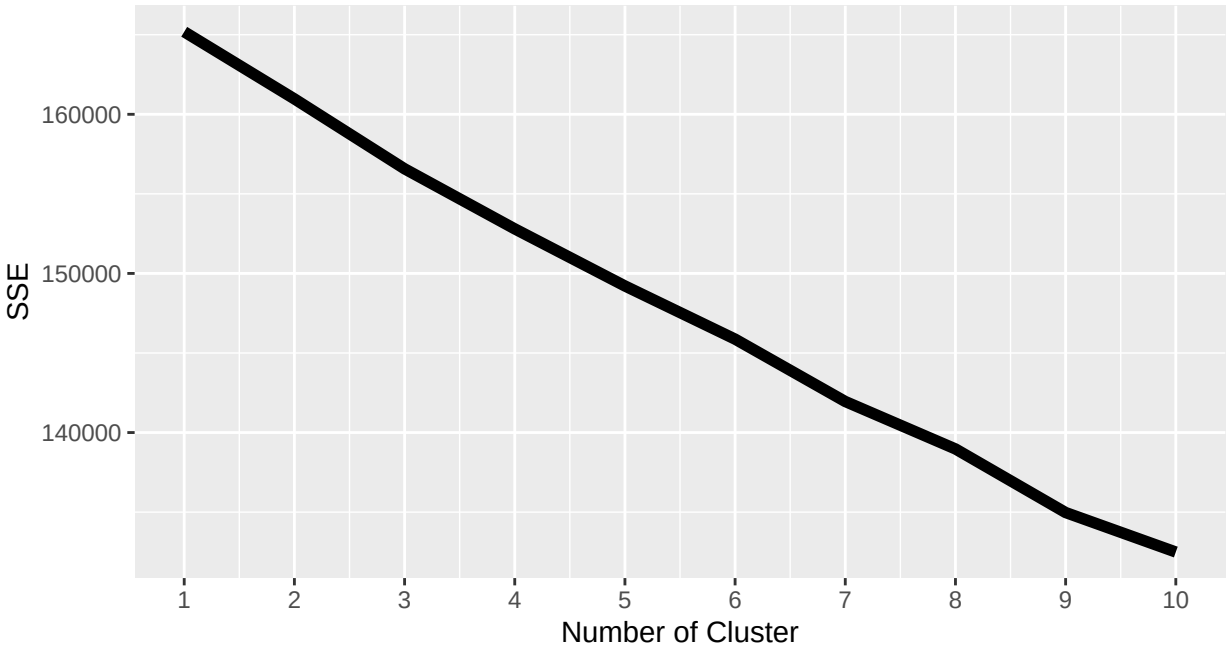
First we just count the words across all the restautants, I get a table of 100 rows and over 1300 colomns, so I decided to drop some words that are really rare across restaurants, after that I get a table of 100 rows and 75 colomns, but 75 colomns is still too much to be processed, so I used PCA to condense the data to certain amount.

Below is the polt of cumulative variance plot against number of components:



As shown above, we use the accumulative variance as standard to select how many components we want, here I set the threshhold of 0.8, so we will have 26 Principal components.

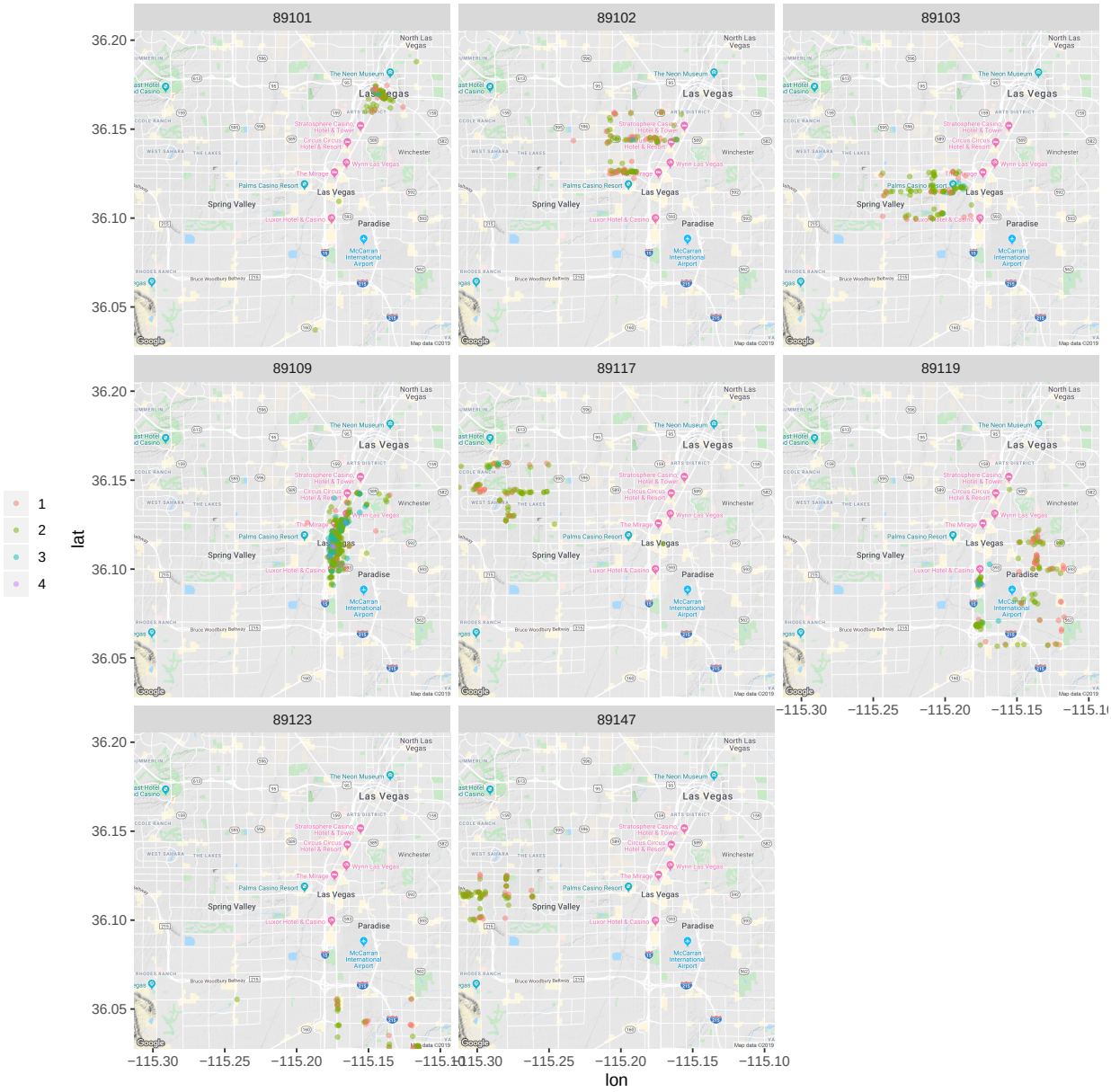
After condensing the data, we continue doing cluster analysis, as for the number of clusters, I draw below plot to help me to decided:



According to above plot, y axis shows the total Sum of square residuals within clusters, the idea is if there exists a good cluster number, we just increase the number of cluster, as it is close to the real number, the SSE will drop significantly, but now I didn't any pattern like that, so I cannot decide the best number of cluster, which may suggests there is clusters, which also means we can not do cluster on this, so I give up on cluster analysis.

Mapping

Besides the plot and data, I want see the geographical distribution of restaurants that comply with certain features, for example I want too see the distribution of Chinese restaurants or high rated restaurants and so on.



We can see that each district has restaurants with all price range, and the proportion of each price range looks quite same, so we may conclude that there is no rich districts or poor districts.

Summary of EDA

For the entire EDA I did looked at the distribution of star ratings of restaurants, applied text mining on reviews of restaurants, tried to apply cluster analysis over the result of text analysis, finally draw maps to check the geographical distribution of restaurants, below are main findings:

1. Star ratings do not vary across cities but vary across district;
2. Noise level of a restaurants will centainly affect the star ratings of a restaurant;
3. Chinese restaurants potentially have lower star ratings which are affected by the proportion of Chinese imigrations;
4. From a geography perspective, restaurants with different price range are evenly distributed;