Analysis of Yelp Dataset Challenge 2017

Yimei Tang

***Abstract*—In my project, I analyze three sub data sets in Yelp Data set Challenge 2017: business data, review data and user data. In particular, I use the business data to perform rating classification, the review data and user data to perform collaborative filtering, and the user data to perform clustering.**

***Keywords*—Yelp Dataset Challenge, Classification, Collaborative Filtering,Matrix Factorization, Clustering**

I. Introduction

A. Research Question

I have three main research questions in this project:

1) Could we classify ratings just by looking at the restaurants' features?
2) Do Yelp users share some common characteristics when their given average ratings are close?
3) Could we use the ratings alone to perform collaborative filtering and matrix factorization to predict ratings?

II. Business Data

A. Purpose of The Analysis

My goal is to use the business data alone to identify useful dependent variables to classify the ratings. Although ratings range from 1-5 as numeric data, since it is discontinuous, it is more useful to treat it as categorical data so we could use classification methods.

In this analysis, I first evaluate each feature in the data set in exploring their relation with the ratings. After that, I use feature selection to choose top features for K Nearest Neighbor, Decision Tree, Naive Bayesian, Linear Discriminant Analysis and Random Forest these classification methods. Last, I compare the results and evaluate these models.

B. Data set

There are 156,639 examples in the business data. Each data is a business unit, it could be restaurant or a salon in different cities. In my project, I will only study the restaurants in Las Vegas. Therefore, my data set has 5,682 examples, each example has 15 features.

These 15 features are:

1) **Address:** Address of the business.
2) **Attributes:** Amenities of the business,such as parking, WiFi and ambience.
3) **Business_id:** A unique ID for each business.
4) **Categories:** The category of the business, for example, it could be restaurants serving Traditional American and Pizza, or it could be Professional Services as Matchmakers.
5) **City:** The city of the business
6) **Hours:** The business hours of the business.
7) **Is_open:** Whether the business is open at the time of fetching the data set from the data base.
8) **Latitude:** The latitude of the business.
9) **Longitude:** The longitude the business.
10) **Name:** The name of the business.
11) **Neighborhood:** The neighborhood of the business.
12) **Postal_code:** The zip code of the business.
13) **Review Count:** Number of reviews given to the business.
14) **Stars:** Ratings of the business.
15) **State:** The state of the business.

C. Restaurant Attributes

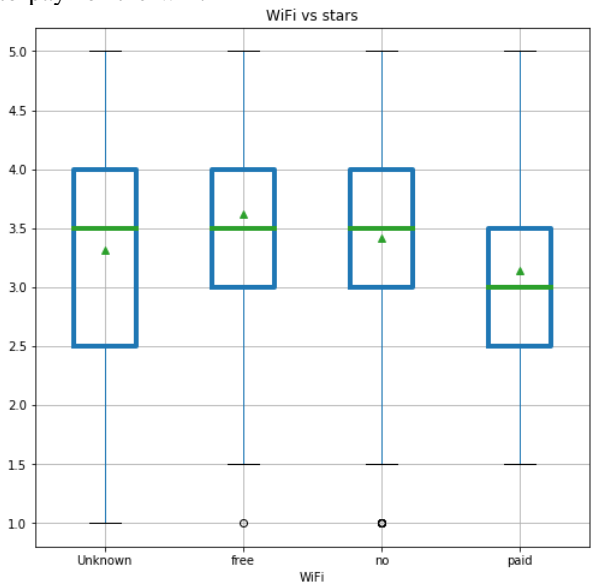
The original "attributes" column in the data set contains many useful information that is related to the rating, such as whether the restaurant accepts credit cards, what is the ambience of the restaurant etc. After parsing the data, there are 37 attributes in this category:

1) Ages Allowed: 21plus,none
2) Alcohol: full_bar,beer_and_wine,open_bar
3) Ambience:classy,divey,hipster,intimate,romantic, touristy,trendy,upscale
4) BYOB: T/F
5) BYOBCorkage: T/F
6) BestNights: Mn,Tu,We,Th,Fr,Sa,Su
7) BikeParking: T/F
8) BusinessAcceptsBitcoin: T/F
9) BusinessAcceptsCreditCards: T/F
10) BusinessParking: garage,lot,street,valet,validated
11) ByAppointmentOnly: T/F
12) Caters: T/F
13) CoatCheck: T/F
14) Corkage: T/F
15) Dietary Restrictions: vegan
16) DogsAllowed: T/F
17) DriveThrough: T/F
18) GoodForDancing: T/F
19) GoodForKids: T/F
20) GoodForMeal:breakfast,brunch,dessert,dinner, latenight,lunch
21) HappyHour: T/F
22) HasTV: T/F
23) Music:background_music,dj,jukebox,karaoke,live, no_music,video
24) NoiseLevel: quiet,average,loud,very_loud
25) Open24Hours: T/F
26) OutdoorSeating: T/F
27) RestaurantsAttire: casual,
28) RestaurantsCounterService: T/F
29) RestaurantsDelivery: T/F
30) RestaurantsGoodForGroups: T/F
31) RestaurantsPriceRange2: 1,2,3,4

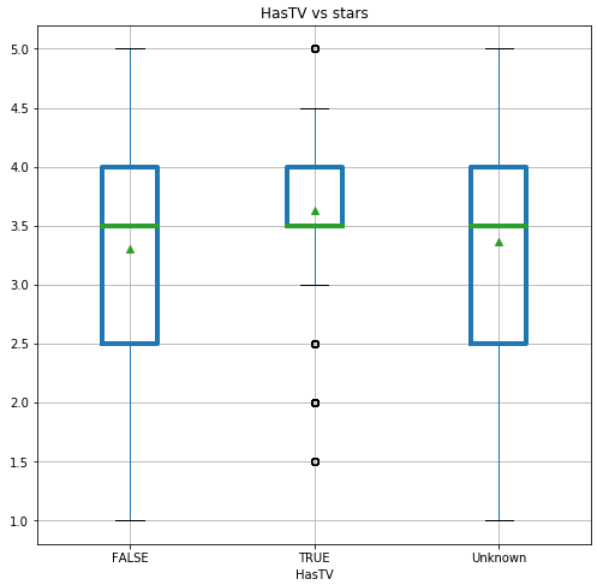
32) RestaurantsReservations: T/F
33) RestaurantsTableService: T/F
34) RestaurantsTakeOut: T/F
35) Smoking: T/F
36) WheelchairAccessible: T/F
37) WiFi: Yes,No,Paid,Free

D. Attributes vs Ratings

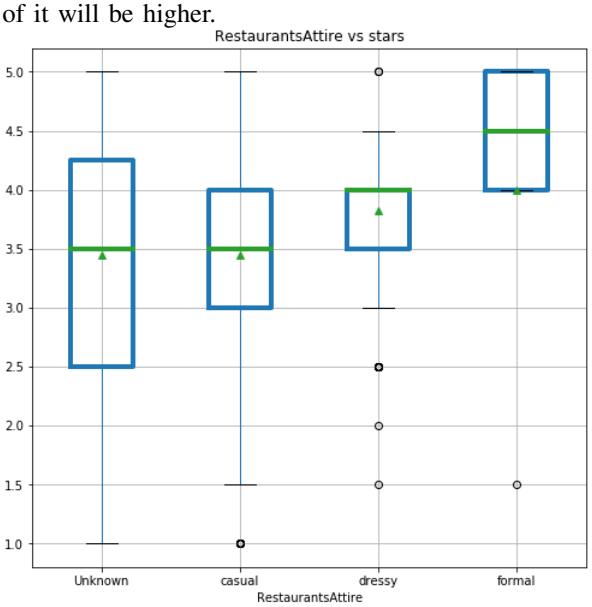
1) **Restaurants TableService:**
If the restaurants have table services, most the ratings tend to be no less than 3.0. Otherwise, a good proportion of the ratings could be lower than 2.5.
2) **Good for meal:**
Usually, restaurants that are good for desserts have the higher ratings. However, restaurants that are good for dinner have very stable ratings in the range of 3 to 4.5.
3) **Alcohol:**
Alcohol didn't play a significant role in influencing the rating.
4) **Caters:**
If the restaurant caters, it tends to get higher rating than non-catering restaurants.
5) **Restaurants Good For Groups:**
If the restaurant is a group-friendly restaurant, it tends to get higher rating.
6) **Noise Level:**
However, if the noise is too loud, the rating will be worse.
7) **Wifi:**
It is better that you have no wifi than the guests need to pay for the wifi.



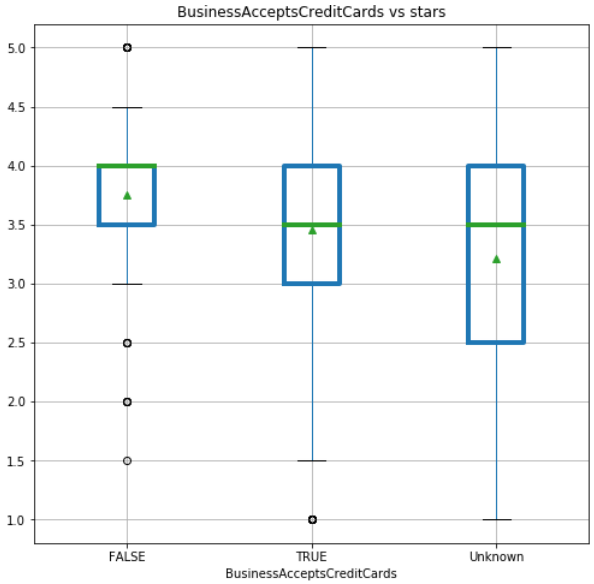
8) **Restaurants Reservations:**
Restaurants require reservations tend to have higher ratings.
9) **Has TV:**
It never hurts to have a TV in the restaurant.



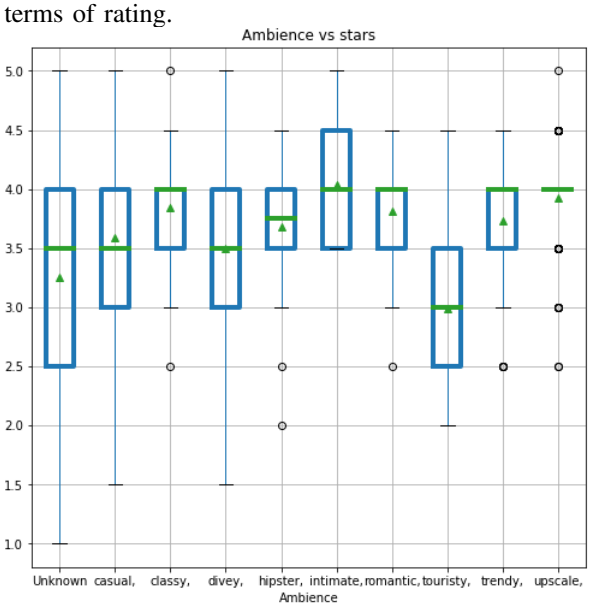
10) **Restaurants Attire:**
If you need to dress up for the restaurant, the rating of it will be higher.



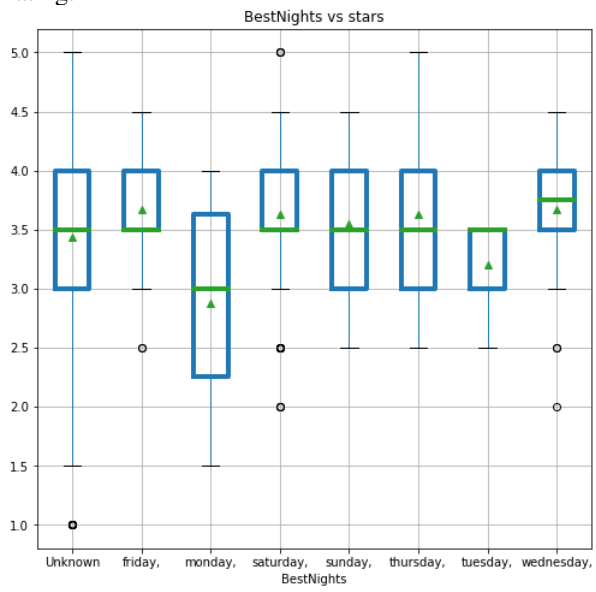
11) **Outdoor Seating:**
Outdoor Seating doesn't make much difference.
12) **Business Accepts Credit Cards:**
The restaurant must be very good that guests are willing to pay with cash, therefore, the ratings tend to be higher.



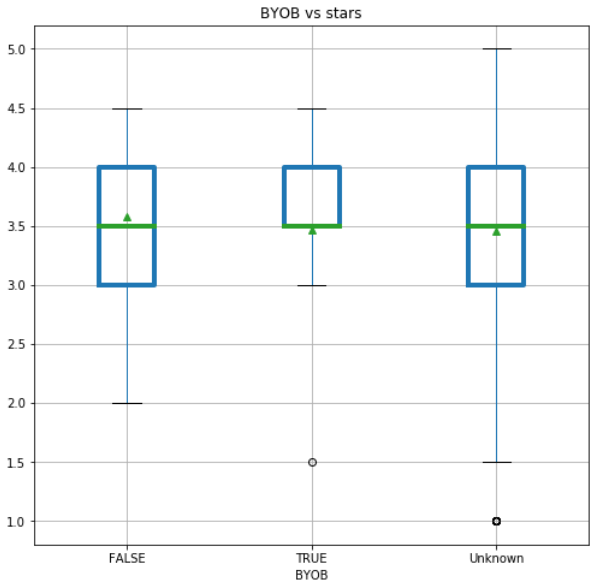
13) **Restaurants Price Range2:**
The higher the price range is, the higher the rating is.
14) **Bike Parking:**
It doesn't matter much if you have bike parking or not.
15) **Restaurants Delivery:**
If the restaurant offers delivery, the rating will be higher.
16) **Ambience:**
The ambience of a restaurant is very important in terms of rating.



17) **Restaurants Take Out:**
Restaurants Take Out is of little importance of rating.
18) **Good For Kids:**
Whether restaurants are good for kids or not is not influential in ratings.
19) **Drive Through:**
If the restaurant offers drive through, it will tend to have lower ratings.
20) **Business Parking:**
If the restaurant has validated parking, people tend to give higher ratings.
21) **Dogs Allowed:**
Whether dogs are allowed or not is not important.
22) **Wheelchair Accessible:**
Wheelchair Accessible is not an important feature.
23) **Music:**
Music is not important in terms of ratings.
24) **Happy Hour:**
The restaurant tend to have a higher rating if it doesn't have happy hour.
25) **Good For Dancing:**
The restaurant tend to have a higher rating if it is not good for dancing.
26) **Best Nights:**
If the best night is Monday, the restaurant has lower rating.



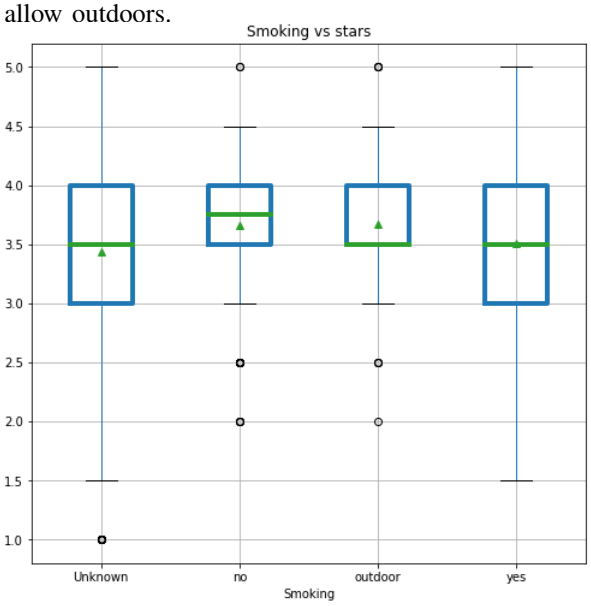
27) **Coat Check:**
Having the service of checking coats could help the restaurant get higher rating.
28) **By Appointment Only:**
Restricting the guest to make appointment beforehand is not hurting the rating of a restaurant.
29) **Business Accepts Bitcoin:**
It is not showing a big difference in rating whether the restaurant accepts bitcoin or not.
30) **BYOBCorkage:**
Compared to free corkage, restaurants which charge fee for corkage have higher ratings.
31) **BYOB:**
People will give higher ratings for restaurant that allow them to bring their favorite bottles.



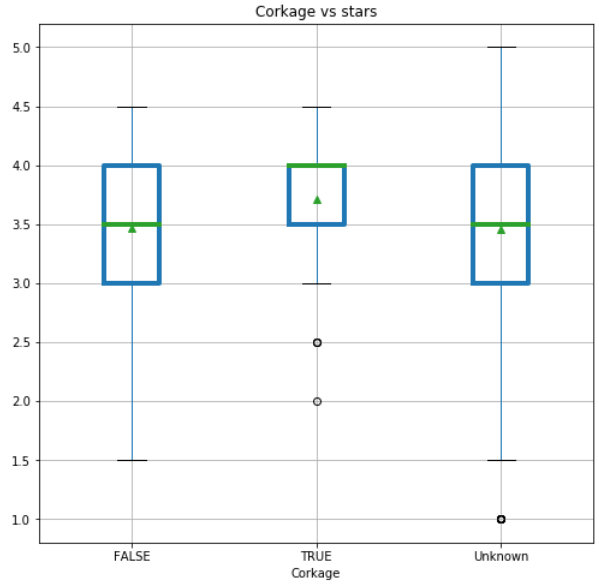
32) **Open 24 Hours:**
Restaurant that are open 24 hours tend to receive lower ratings.

33) **Restaurants Counter Service:**
If Restaurants offer counter service, the ratings will tend to be lower.

34) **Smoking :**
Restaurants allow smoking indoors will have lower ratings that those that don't allow smoking or only allow outdoors.



35) **Corkage :**
Restaurant charge for opening bottles have higher ratings.



36) **Ages Allowed :**
Ages allowed has no significant relationship with ratings.

37) **Dietary Restrictions :**
We only have 10 data in this feature, so we can't make any feasible inference.

Based on above analysis, I will include the following 14 features as dependent variables: Restaurants Table Service,Good For Meal,Caters, Has TV, Restaurants Good For Groups,Noise Level,WiFi,Restaurants Attire,Restaurants Reservations,Business Accepts Credit Cards,Restaurants Delivery,Ambience,Drive through and Business Parking.

They are all categorical variables so I will convert them into dummy variables when building the model.

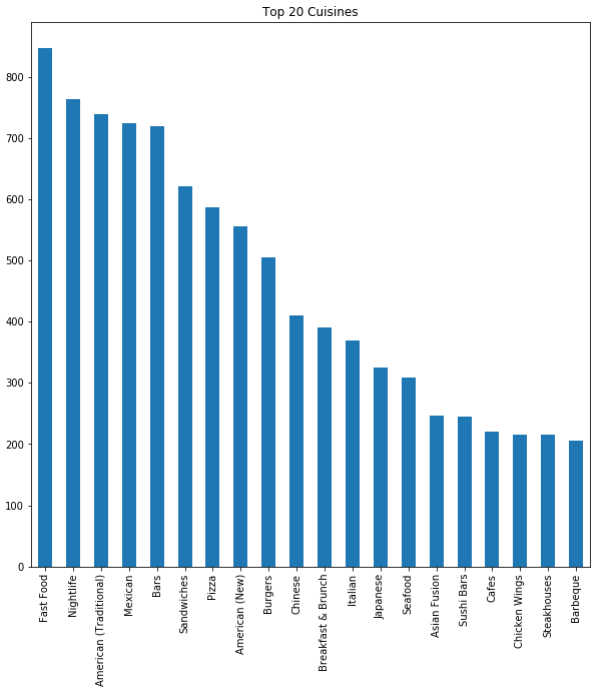
E. Category - Cuisine Types

After analyzing the categories column, I found 365 cuisine types.

```
cuisines.sum().sort_values(ascending=False)
```

```
Fast Food                847
Nightlife                764
American (Traditional)   739
Mexican                  724
Bars                     720
Sandwiches               621
Pizza                    587
American (New)           555
Burgers                  505
Chinese                  410
Breakfast & Brunch       390
Italian                  370
Japanese                 325
Seafood                  308
Asian Fusion             247
                         ...
```

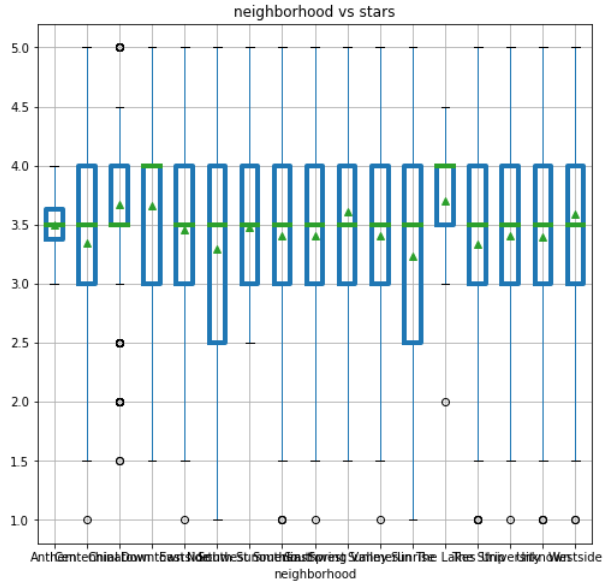
A restaurant could have more than 5 cuisine types: Fastfood, Nightlife ,American (Traditional),Bars, Pizza. It is hard to determine the overall style of the restaurant. Therefore, I will not include this "category" feature in my analysis.



F. Neighborhoods

Even though all these 5,826 restaurants are scattered in different areas in Las Vegas. However, the neighbourhoods itself have little influence on the ratings.

Therefore, I will not include ”neighborhood” feature in my analysis.



G. Postal Code

Since postal code is correlated with the neighborhood, I will not include ”postal code” feature in the analysis.

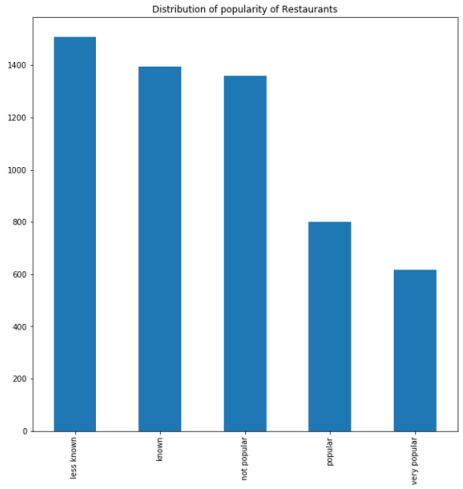
H. Review Counts

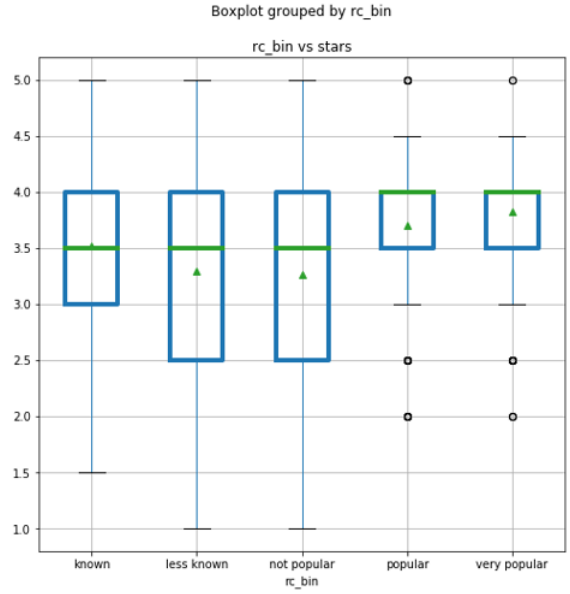
```
count    5682.000000
mean      149.582541
std       341.862067
min         3.000000
25%        14.000000
50%        44.000000
75%       148.000000
max      6979.000000
Name: review_count, dtype: float64
```

More than 50% of the restaurants have 44 reviews or more. Only 25% of the restaurant has reviews of more than 148. Therefore, I will use these critical values as my binning threshold to create five bins.

```
bins=[0,14,44,148,350,6979]
labels=["less known","not popular","known","popular","very popular"]
```

After binning, we could tell the number of reviews do correlate with ratings.





Therefore, I will include the binned review_count feature in my analysis.

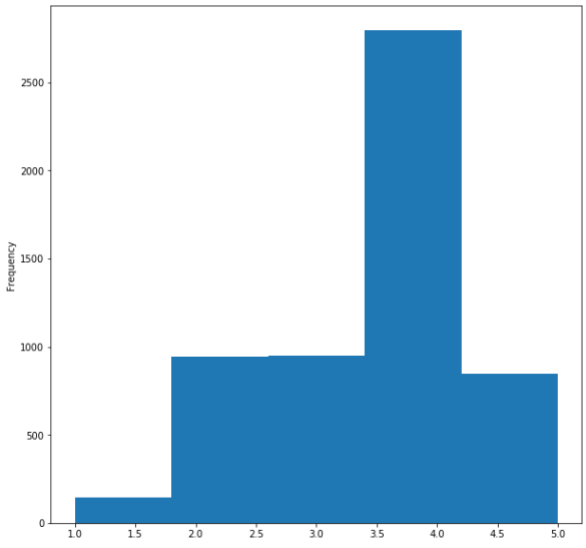
I. Ratings

The rating is skewed to the left since the mean is smaller than the median.

We have 9 classes of ratings: 1,1.5,2,2.5,3,3.5,4,4.5,5

```
count    5682.000000
mean        3.455913
std         0.798719
min         1.000000
25%         3.000000
50%         3.500000
75%         4.000000
max         5.000000
Name: stars, dtype: float64
```

As we could tell from the histogram, most of the ratings are between 3.5 to 4.0. However, the number of restaurants receive rating of 5 is larger than that of rating less than 1. It shows that not so many restaurants in Las Vegas are poorly rated and there is a handful of excellent restaurants.



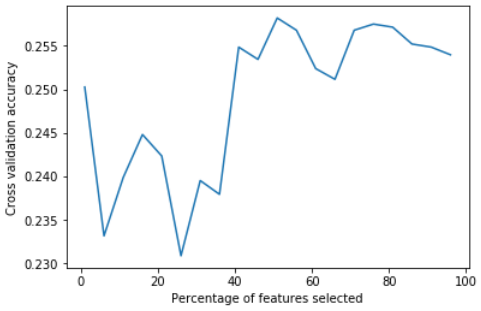
J. Models and Comparison of Performance

1) Variables:

Target: 1 categorical Variable: Ratings (9 classes) Independent Variables: 15 Categorical Variables After converting the categorical variables to dummy variables, I now have 65 dummy variables.

2) K Nearest Neighbours:

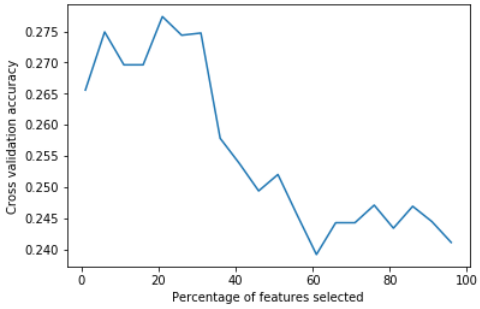
The best model of KNN found by using GridSearchCV is : 10 nearest neighbours and use uniform weights. By using feature selection, out of the 65 dummy variables, the top 51% features(33 features) will have highest accuracy score on the test set.




```
The accuracy score of KNN Method in train set is 0.400440
The accuracy score of KNN Method in test set is 0.276165
```

3) Decision Tree: The best model of Decision Tree found by using GridSearchCV is : using entropy as criterion, the max_depth is 7, the min_samples_split is 12, and the min_samples_leaf is 1.

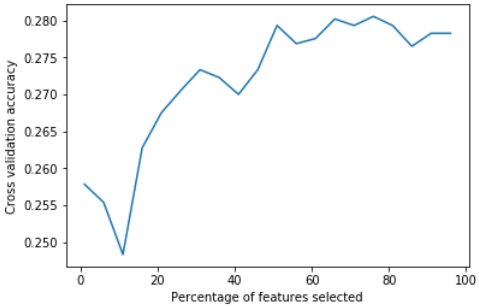
By using feature selection, out of the 65 dummy variables, the top 21% features(13 features) will have highest accuracy score on the test set.




```
The accuracy score of Decision Tree in train set is 0.318152
The accuracy score of Decision Tree in test set is 0.270888
```

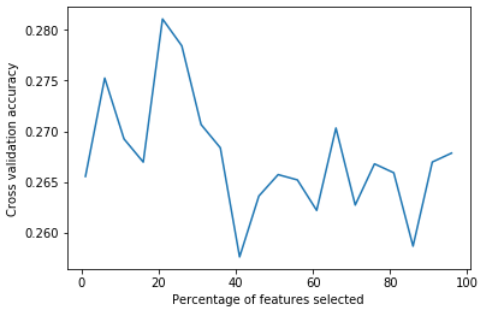
4) Naive Bayesian Classification: The Naive Bayesian model is fitting poorly on the data set with accuracy score 0.056 on training set. The reason might be that many of the features in the dummy variables are correlated.

5) Linear Discriminant Analysis: The best model of LDA found by using GridSearchCV is : using SVD as solver. By using feature selection, out of the 65 dummy variables, the top 76% features(49 features) will have highest accuracy score on the test set.




```
The accuracy score of LDA in train set is 0.311551
The accuracy score of LDA in test set is 0.286719
```

6) Random Forest: The best model of random forest found by using GridSearchCV is : using entropy as criterion, max_depth is 9, min_samples_leaf is 1 , min_samples_split is 14 and n_estimators is 7. By using feature selection, out of the 65 dummy variables, the top 21% features(13 features) will have highest accuracy score on the test set.




```
The accuracy score of Random Forest in train set is 0.405941
The accuracy score of Random Forest in test set is 0.264732
```

7) Comparison:

	Accuracy Score (5 fold CV)		# Of Features
	Train	Test	
K Nearest Neighbour	0.40	0.27	33
Decision Tree	0.31	0.27	13
Linear Discriminant Analysis	0.31	0.28	49
Random Forest	0.40	0.26	13

The best model is Decision Tree model(entropy as criterion, the max_depth is 7, the min_samples_split is 12, and the min_samples_leaf is 1) with 13 features. The reason is that it has the highest accuracy score of the four(since the accuracy score of Naive Bayesian classifier is extremely low, we don't include it here in comparison), and the required number of features is smallest.

The most important features are:

```
RestaurantsTableService_TRUE          161.7231128687851
RestaurantsTableService_Unknown       232.0059301600887
GoodForMeal_Unknown                    331.059030265652
GoodForMeal_dinner,                    227.406020389406
Caters_TRUE                           161.90251059581226
HasTV_Unknown                         169.15306593270094
RestaurantsGoodForGroups_Unknown      184.17903708845438
NoiseLevel_Unknown                     239.9966829824095
WiFi_Unknown                          169.41467630548863
RestaurantsAttire_Unknown             211.33626339556739
Ambience_Unknown                      297.03316500009373
DriveThru_TRUE                         405.8077269705346
BusinessParking_Unknown                311.4898304873481
BusinessParking_lot,                  167.05291946426232
```

The most important feature is DRIVEThru(TRUE). Usually only fast food and coffee shops have drive through service, therefore, it explains why this feature is important. It is a determining factor whether this is a fast food restaurant or other types of restaurant.

Filtering out the features that are unknown, we see that Restaurants Table Service, Good For Meal(dinner) , Caters and Business Parking_lot are important features.

III. Review Data and User Data

A. Purpose of The Analysis

In this part of analysis, I aim to use the review data and user data together to perform collaborative filtering. In order to do this, I collected the ratings of users for restaurants in Las Vegas. The original data set has 849,883 ratings for the 5,682 restaurants in Las Vegas from 307,484 users, which is too large to train. Therefore, I randomly select 8,000 ratings from the orginal data set to perform matrix factorization in collaborative filtering.

In another part of the analysis, I will use the features in user data set to perform K-Means analysis. Since it is unsupervised, I will not have any target variables.

B. Review Data set

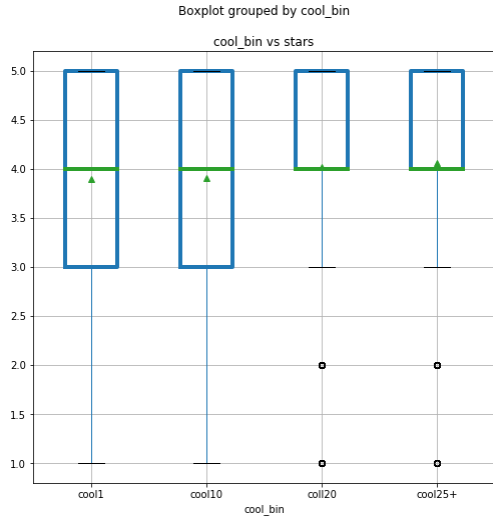
There are 4,736,897 reviews in the original data set. Each review has 9 features:

1) **Business_id:**
The unique business id of the business that the review is given to.
2) **Cool:**
The number of people who think this review is cool.
3) **Date:**
Date of the review is written.
4) **Funny:**
The number of people who think this review is funny.
5) **Reviewid:**
The unique id of the review.
6) **Stars :**
The rating/stars of the review, from 1 to 5 as 1 is the lowest.
7) **Text:**
The text of the review.
8) **Useful:**
The number of people who think this review is useful.
9) **User_id:**
The unique id of the reviewer.

C. Review Features

I am only analyzing the reviews for restaurants in Las Vegas. There are 849,883 reviews in total.

1) **Cool :**
Not many reviews are considered "cool" by more than 1 person. However, when the number of "cool" increase to surpass 20, the rating tend to go up as well. The ratings of reviews that receive fewer than 20 "cools" are not that different.



2) **Useful:**
However, the number of useful reviews might indicate the red flags of a restaurant so the number of "useful" is not always positively correlated with the ratings.
3) **Funny :**
Funny feature has the similar pattern. It shows that when people think a review is cool, usually they will think the review is funny as well.

D. User Data set

In total, there are 1,183,362 users in the data set, each user has 22 features.

1) **User_id:**
The unique id of the user.
2) **Name:**
The name of user.
3) **Review_count:**
The number of reviews they've written
4) **Yelping_since:**
When the user joined Yelp.
5) **Friends:**
User ids of their friends.
6) **Useful:**
Number of useful votes sent by the user.
7) **Funny:**
Number of funny votes sent by the user.
8) **Cool:**
Number of cool votes sent by the user.
9) **Fans:**
Number of fans the user has.
10) **Elite:**
The years the user was elite
11) **Average_stars:**
Average rating of all reviews.
12) **Compliment_hot:**
Number of hot compliments received by the user.
13) **Compliment_more:**
Number of more compliments received by the user.
14) **Compliment_profile:**
Number of profile compliments received by the user
15) **Compliment_cute:**
Number of cute compliments received by the user.
16) **Compliment_list:**
Number of list compliments received by the user.
17) **Compliment_note:**
Number of note compliments received by the user
18) **Compliment_plain**
Number of plain compliments received by the user
19) **Compliment_cool**
Number of cool compliments received by the user
20) **Compliment_funny**
Number of funny compliments received by the user
21) **Compliment_writer**
Number of writer compliments received by the user
22) **Compliment_photos**
Number of photo compliments received by the user

E. K-Means Clustering and Principal Component Analysis

We will only use numerical data in this analysis since K-Means is mainly for numeric data.

1) K-Means with Normalized Data: The first time, I just use normalized data with cluster = 3.

Centroids of Clusters

	0	1	2
average_stars	0.641657	0.896427	2.271727e-01
compliment_cool	0.000807	0.000271	6.294725e-06
compliment_cute	0.000476	0.000137	5.608098e-06
compliment_funny	0.000807	0.000271	6.294725e-06
compliment_hot	0.000569	0.000192	3.136609e-06
compliment_list	0.000185	0.000044	1.098372e-06
compliment_more	0.000419	0.000121	1.513781e-05
compliment_note	0.001027	0.000304	4.004584e-05
ompliment_photos	0.000142	0.000058	7.771922e-07
compliment_plain	0.000754	0.000279	1.498663e-05
ompliment_profile	0.000186	0.000054	1.395265e-06
compliment_writer	0.000715	0.000217	9.649058e-06
cool	0.000388	0.000145	6.498652e-06
fans	0.000850	0.000337	4.238234e-05
funny	0.000439	0.000151	1.445935e-05
review_count	0.006276	0.002354	8.190951e-04
useful	0.000553	0.000202	3.369112e-05

As we could tell here, the cluster 0 seems to be of users who is relatively "conservative" in giving high ratings. Users from cluster 1 are the most "generous" type of users, they give the highest ratings among these three groups. On the contrary, users from cluster 2 are very "picky" that they give the lowest ratings of all.

However, it is the users from cluster 0 receives the highest compliments across all categories in these three clusters. Users from cluster 1 receives the second highest compliments among these three clusters. What is surprising is that , event though

users from cluster 2 are critical, their reviews are far less helpful or cool than other users.

2) K-Means with PCA Transformed Data: The second time, I use PCA and the 98% of the total variance of original 17 features could be represented by 1 principal component. And this PC has the largest value on the axis of the feature average_rating. These two K-Means clustering findings let us know that we could tell a lot just by looking at the average rating of a user alone.

Centroids of Clusters

	0	1	2
PC1	-0.210469	0.044842	0.460436

F. Collaborative Filtering with Matrix Factorization

In this analysis, I only use a sample size of 8,000 ratings to perform collaborative filtering with matrix factorization. Although I only train the model with 100 steps, the outcome is very good with an overal MAE of 1.69 for 7,389 users and 5,682 restaurants in total.

```
MAE for User 7381 = 1.699749282095247
MAE for User 7382 = 2.9479590446935586
MAE for User 7383 = 0.8668159173845851
MAE for User 7384 = 3.665557587731295
MAE for User 7385 = 0.5853615133036634
MAE for User 7386 = 0.6599671020832862
MAE for User 7387 = 1.296334562858389
MAE for User 7388 = 2.5863344566861466
MAE for User 7389 = 1.0436817209027982
Overal MAE = 1.6894561998158202
```

IV. Conclusion

A. Could we classify ratings just by looking at the restaurants' features?

The answer is no. Given the low accuracy score of four classification models, we can't just rely on the restaurant features alone to classify the ratings. However, these features will be helpful in future analysis.

B. Do Yelp users share some common characteristics when their given average ratings are close?

The answer is yes. If your average rating is in the middle level yelp users(in another way, you are not too critical or too easy to satisfied), you will receive more compliments from others. People are more likely to think your reviews are helpful or cool. However, if your average ratings are extremely low, then people are less likely to compliment your reviews.

C. Could we use the ratings alone to perform collaborative filtering and matrix factorization to predict ratings

The answer is yes. Although I only use a sample size of 8,000 out of 850,000 records, the performance of the collaborative filtering model is very good with only 1.69 in MAE for 7,389 users and 5,682 restaurants.

V. Future Work

I enjoy this fun exploration with Yelp Dataset. I hope to learn how to use Hadoop to analyze this gigantic Yelp Dataset and use other machine learning methods, especially in natural language processing.