GEORGETOWN UNIVERSITY
The Graduate School *of* Arts & Sciences

ANLY_501_Project_Report

What Makes a Great Restaurant?

Group 6

Shaoyu Feng, Jiaxuan Sun, Jen Wang, Chelsea Wang

TABLE OF CONTENTS

Data Science Problem

The formula for maintaining a great restaurant has changed nowadays. Customers have new expectations and tend to gravitate towards restaurants that cater to their various needs. The number of articles dissecting and discussing the qualities of a successful restaurant in business have raised to meet the reader's expectation and general trend. In 2011, Stephani Robson, a researcher from Cornell University, published an article discussing the possible strategies to find the ideal restaurant site. The article revealed that the most successful restaurants are in sites that are convenient for the target market. Another article, written by Bryan Keythman, provided a comprehensive viewpoint on the factors that can help determine the quality grade of a restaurant. The article emphasized on quality of food, customers experience, uniqueness and management. Moreover, it also encouraged us to find our own standards on defining the characters and qualities of a restaurant. The purpose of our study is to utilize data-driven methods in analyzing the important factors contributing towards the success of a restaurants and thus provide data-science solutions to restaurant owners to help them to improve their business.

Potential Analysis that Can Be Conducted Using Collected Data

Social medias nowadays allow consumers to share their experiences to the public on designated platforms to help express their opinions, online reviews have become one of the most influential factors in restaurant selection. Because of this, we chose Yelp, a popular diner review site, as our start point to collect data. We primarily focused on analyzing restaurants in large cities like Washington DC, Boston, New York City, Los Angeles, San Francisco, Huston and Seattle.

Since we attempt to analyze factors that impact restaurant rating, we try to collect data relevant to restaurant as much as possible. We categorize interested factors into internal factors and external factors respectively.

For internal factors, we use Yelp Fusion APIs and web scraping to collect data of restaurants in our target cities. We pull the following information of each restaurant from API call: location by latitude and longitude, price level, rating, review count, category, payment method, noise level, open hour, Wi-Fi availability and other relevant information. These variables allowed us to delve into which internal factors would contribute to the success of the restaurant.

For external factors, we collect data based on the location of restaurant. We use Google places API to gather data nearby the restaurant in 500-meter radius: transportation data (like bus station, subway station, train station, taxi stand); the public facilities data (like bank, ATM, museum); the shopping places data (like supermarket and shopping mall); the entertainment venues data (like bar and movie theater). These data can help us analyze the impact of external factors on the restaurant.

We have three potential directions at this stage. Firstly, we will use descriptive and exploratory data analytics methods to find out correlations between different attributes of data. Secondly, we plan to find the key factors that impact the success of a restaurant from all interested factors by using machine learning and multivariate analysis. This could be in the forms of decision trees, hierarchical clustering or ensemble tree approaches. Thirdly, we would love to build advanced visualizations from the data to uncover interesting insights for business users to improve their operation or support their decision.

Data Issues

Data in this project is primarily extracted from Yelp Fusion API. For each of our target city, we make a list of neighbourhoods and pull data accordingly from API. Since Yelp API has limitation on record return for each call, we have to made multiple calls to gather enough information. However, it causes duplicate records in our dataset. Also, there are few restaurants that share the same name but are actually different restaurants. To eliminate confusion in future interpretation, a unique ID is used as identifier for each restaurant.

Moreover, the attributes provided on Yelp were quite not same for every restaurant. Some of the restaurants offer the full list of information such as Wi-Fi availability, parking spaces, credit card acceptance and etc, but other restaurants are not. As a result, there exist some missing values in the dataset. We will properly resolve these missing values in the second part of the project.

Last but not least, over 50% of the dimensions in the data set is binary or categorical, we will have to think about proper ways to interpreted the dimension well in order to facilitate further data analysis.

Collecting New Data

To answer the data science question and support the descriptive and predictive data modelling in next phase, three aspects of data has been collected, namely Yelp API Data, Yelp Web Data and Google Maps Places API Data. Yelp API Data is used as the primary data source, the attributes contain the location, ratings, price ranges and so on. Yelp Web Data refers to the additional information provided online but not available through API calls. Google Maps Places API Data consists of the surrounding information of the restaurants, includes number of bus stops, supermarkets and so on.

Part 1: Yelp API Data:

Yelp Fusion API is a powerful tool providing access to Yelp Open data containing the detailed business information and user reviews for all types of businesses like restaurants,

hotels. In this project, we focus on restaurant data. Utilizing the Business Search options in Yelp Fusion API, we collect about 5500 restaurants data entries in 7 US big cities and metropolitan areas, including New York City Area, Washington DC, Boston, Seattle, Huston, San Francisco and Los Angeles. One thing worth mentioning is that the data is collected in batches of neighborhoods in cities, and we traversed through over 700 neighborhoods in those cities. There are a lot of duplicates in those batches, and the duplicate entries has been removed during data collection phase by selecting unique Yelp Business ID.
Table 1 illustrates the distributions of data in different cities. And Figure 1 shows a snapshot of the data coming from Yelp Data API.

Table 1 Distribution of Restaurant Data in Big Cities

City Name	Restaurant Counts
New York	638
Los Angeles	424
San Francisco	391
Houston	263
Boston	243
Brooklyn	211
Seattle	204
Washington DC	175
Long Beach	105
…	…

Figure 1 Yelp API Data at a Glance

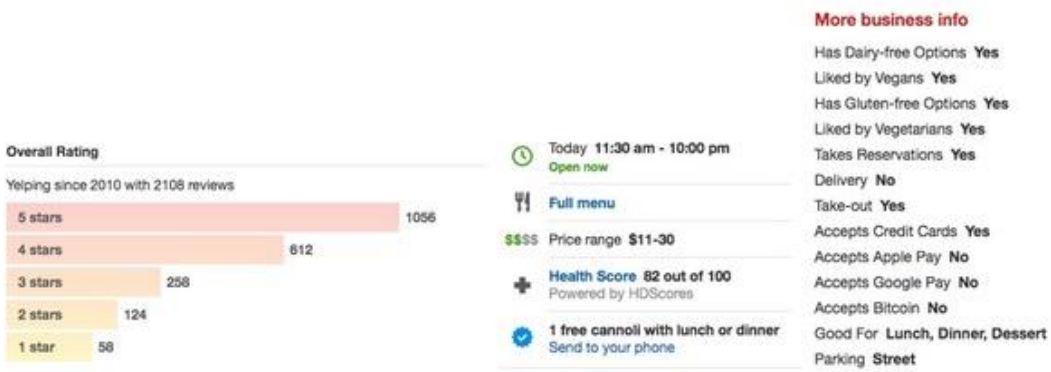
name	latitude	longitude	is_closed	zipcode	city	state	price	rating	url	view_cou	sact	category	id
Middle East Restaurant …	42.3638	-71.1013	False	2139	Cambri…	MA	$$	3.5	https…	531	[…	mideastern…	YFkrZhuA1ph…
Boston Beer Works	42.3472	-71.0991	False	2215	Boston	MA	$$	3.5	https…	691	[]	breweries,…	SNJhevBiD7x…
Craigie On Main	42.3635	-71.0986	False	2139	Cambri…	MA	$$$$	4	https…	1216	[]	newamerica…	58cDJU4cub1…
Gyro City	42.3432	-71.099	False	2215	Boston	MA	$	4	https…	283	[…	greek,medi…	KENZN5yZDjV…
Porcinis Italian Res…	42.3648	-71.1667	False	2472	Watert…	MA	$$	4	https…	167	[]	italian,se…	7200x8GYzQS…
Eastern Standard Ki…	42.3487	-71.096	False	2215	Boston	MA	$$$	4	https…	1749	[]	newamerica…	p8ohzzGvGRC…
Cuchi Cuchi	42.3633	-71.0971	False	2139	Cambri…	MA	$$$	4	https…	1066	[]	tapasmallp…	C7FKNwSUlbU…
Island Creek Oyster Bar	42.3487	-71.0951	False	2215	Boston	MA	$$$	4.5	https…	2486	[]	seafood,ba…	VnuD2cojPTW…
Trattoria Pulcinella	42.3827	-71.1307	False	2138	Cambri…	MA	$$$	3	https…	47	[]	italian	6q5TIpdyLXL…
Bondir Cambridge	42.3683	-71.0978	False	2139	Cambri…	MA	$$$$	4	https…	313	[]	newamerican	T8DpRFty13M…
Giulia	42.3825	-71.1201	False	2138	Cambri…	MA	$$$	4.5	https…	393	[]	italian	9RH1xBIMMAZ…
Temple Bar	42.3827	-71.1201	False	2138	Cambri…	MA	$$	3.5	https…	487	[]	newamerica…	JZPPsLYmLCc…

Part2: Yelp Web Data

Yelp API data provides a general overview of a restaurant business; however, no further details are given to describe the restaurant in terms of its service, environment and so on. To make the data set more comprehensive, a web scrape script has been developed to grab

additional information about the restaurants. Figure2 shows a snapshot of the Yelp website and the list of information available.

Figure 2 Yelp Business Details From: https://www.yelp.com/biz/il-canale-washington-2?osq=Restaurants



The Scrape program makes uses of html parser to extract information from the yelp page for all the restaurants we retrieved from Yelp Data API. The summary of all the available features from the scrape program are shown in table2.

Table 2 Additional Information Gathered from Web

Name	*Type*	*Data Description*	*Name*	*Type*	*Data Description*
Id	object	Yelp ID	**Caters**	object	Holds Catering or Not
Name	object	Name	**Delivers**	float64	Delivery Option
category	object	Category of Resturant	**Dogs_Allowed**	object	Whether Pets Allowed
lowprice	object	Price Range	**Outdoor_Seating**	object	Is Outdoor Seating Available
highprice	object	Price Range	**Parking**	object	Parking Option
health_index	object	Health Score in 100 Scale	**Smoking_allowed**	float64	Smoker Free
star1	int64	Number of 1 Star Reviews	**Music**	object	Background Music
star2	int64	Number of 2 Start Reviews	**Takes_Reservations**	object	Whether take Reservation
star3	int64	Number of 3 Start Reviews	**Wheelchair_Accessible**	object	Diabled Person Friendly
star4	int64	Number of 4 Start Reviews	**WIFI**	object	Free WiFi Option
star5	int64	Number of 5 Start Reviews	**Opened_24hrs**	float64	Whether Open 24 hrs
Accept_Credit_Card	object	Payment Option	**Ambience**	object	
Alcohol	object	Types of Alcohol	**Attire**	object	Dress Code
Appointment_Only	float64	By Reservation Only	**Noise_Level**	object	

Part 3 Google Maps Places API:

In additional to the internal factors of the restaurant, we also consider the external locational factors of the restaurants, like whether the transportation is convenient or whether there are places to go nearby. Google Maps Places API provides a comprehensive function in retrieving nearby places data of the restaurants. Making use of the latitude and longitude pulled from Yelp API, we draw a cycle with 500 meters in radius, list out all the places of interest within that cycle. The places of interest include bus stops, train stations, supermarkets and so on. The retrieved raw data is originally in the form of a tall table, we then pivot the

table in to a wide table for data join. The transformed data contains the number of different places near the restaurants. Figure 3 and 4 show the raw data and pivoted data retrieved from Google Maps Places API respectively.

Figure 3 Raw Data From Google Map Places API

yelp_id	estaurant_locatio	id	place_id	types	name	rating ▲	geometry	plus_code	vicinity
wQ9e-NbQH2bX7OQZ…	34.06621,-1…	078ee2bd904…	ChIJm8T8Plr…	cafe	Gio's Donuts	1	34.06282349…	3Q76+4G Los Angeles, Ca…	Alpine Street, Los…
UZlLBVnqs9I…	34.052601,-…	0a49514a3fe…	ChIJNwly-7D…	bus_station	7th / Francisco	1	34.050087,-…	3P2Q+27 Los Angeles, Ca…	United States
UZlLBVnqs9I…	34.052601,-…	ed426931574…	ChIJl4EnQa7…	school	Film Connection	1	34.05482289…	3P3R+W2 Los Angeles, Ca…	1201 West 5th Street …
UZlLBVnqs9I…	34.052601,-…	349836bfd33…	ChIJgz7VY6X…	atm	Los Angeles Health Care…	1	34.05376039…	3P3M+GX Los Angeles, Ca…	637 Lucas Avenue, Los…
UZlLBVnqs9I…	34.052601,-…	0e8d630622a…	ChIJNSo-YrHHwoAR2zH…	atm	Chase ATM	1	34.0508852,…	3P2R+9F Los Angeles, Ca…	Los Angeles
VdpxYO4hCKY…	34.06518460…	9f3875093b4…	ChIJd4K_Y1z…	bus_station	Main / College	1	34.062889,-…	3Q78+59 Los Angeles, Ca…	United States
VdpxYO4hCKY…	34.06518460…	078ee2bd904…	ChIJm8T8Plr…	cafe	Gio's Donuts	1	34.06282349…	3Q76+4G Los Angeles, Ca…	Alpine Street, Los…
XKcHgQ75rl0…	34.05964799…	3974950cc40…	ChIJT5Ebo4L…	school	Joyful Presbyteria…	1	34.0560956,…	3P46+CC Los Angeles, Ca…	866 South Westmorelan…
XKcHgQ75rl0…	34.05964799…	be363cd883e…	ChIJJXlr_nn…	atm	Access to Money	1	34.0616458,…	3P66+MQ Los Angeles, Ca…	3050 Wilshire Bo…
XKcHgQ75rl0…	34.05964799…	f626093c1b3…	ChIJxWiHknj…	atm	Coin Cloud Bitcoin ATM	1	34.0571716,…	3P45+VC Los Angeles, Ca…	824 Vermont Avenue, Los…
XKcHgQ75rl0…	34.05964799…	fd1a8212b93…	ChIJxc4373j…	bank	Bank Of Hope	1	34.0577983,…	3P55+48 Los Angeles, Ca…	3003 West 8th Street,…
4lfzJCUnLiH…	34.06100700…	b36948cfe30…	ChIJVcj-7nz…	bus_station	6th / Catalina	1	34.063503,-…	3P73+CQ Los Angeles, Ca…	United States
4lfzJCUnLiH…	34.06100700…	465eb985240…	ChIJtWENUnz…	school	California School of H…	1	34.0620708,…	3P63+RC Los Angeles, Ca…	3345 Wilshire Bo…
4lfzJCUnLiH…	34.06100700…	be363cd883e…	ChIJJXlr_nn…	atm	Access to Money	1	34.0616458,…	3P66+MQ Los Angeles, Ca…	3050 Wilshire Bo…
4lfzJCUnLiH…	34.06100700…	f626093c1b3…	ChIJxWiHknj…	atm	Coin Cloud Bitcoin ATM	1	34.0571716,…	3P45+VC Los Angeles, Ca…	824 Vermont Avenue, Los…

Figure 4 Pivoted Data From Google Maps API

atm	bank	bar	eauty_salo	ook_stor	bus_station	cafe	gas_station	gym	movie_theater	museum	school	shopping_mall	subway_statior	supermarket	taxi_stand	train_station	yelp_id
6	2	1	1	nan	10	2	nan	1	nan	nan	14	nan	nan	nan	nan	nan	-iOW4FIS61Tb…
31	14	37	15	1	29	23	nan	12	1	6	15	6	1	1	nan	1	-pK1SR3blPDE…
14	12	11	34	6	4	13	1	14	1	nan	18	1	nan	2	nan	nan	0f7KLXeVKJ1…
27	11	42	37	4	22	35	nan	7	3	4	43	4	2	nan	nan	nan	0s8iv660dB3…
13	9	3	18	1	16	4	1	6	nan	5	14	nan	nan	nan	nan	nan	0ulC3Jz1eXn…
7	2	11	8	1	34	9	nan	6	nan	2	15	1	nan	nan	nan	nan	1A9alCsYB5Z…
5	nan	6	4	nan	15	4	1	6	nan	nan	2	nan	1	nan	nan	nan	1C32E52HmhR…
11	2	10	10	nan	19	8	nan	10	nan	nan	11	nan	nan	2	nan	1	1exJzD7WPZz…
52	31	41	60	5	19	35	nan	26	1	8	27	16	1	nan	nan	1	1hWZ0K2DOKZ…
19	10	18	9	6	13	22	nan	9	1	nan	31	3	1	nan	nan	nan	1iY5q6spBS_…
15	4	16	1	nan	14	6	nan	2	nan	nan	3	nan	nan	nan	nan	nan	25sny_QPRSF…
1	nan	2	4	nan	nan	2	nan	8	nan	nan	4	1	nan	nan	nan	nan	2K1A9fEqmo-N2J6VEBXcDA
46	25	42	47	6	23	60	1	16	3	4	39	6	4	3	nan	1	2bncbx08BFs…
21	9	47	23	3	12	20	1	9	nan	6	11	6	1	1	nan	nan	2u_w3rthRzR…

Part 4 Data Joining and Additional Information:

In this step we join the three sets of data together for data cleanness assessment. Among the three sets of data, the join key is Yelp Business ID, a unique identifier of the restaurants. There are altogether 5133 rows data with 59 columns after the data joining. And 50 out of the 59 columns are considered as feature columns, and out of the 50 dimensions, 24 of them are considered as numeric columns before data cleaning and feature engineering.

One thing worth mentioning is that we will add another dimension into the data sets, which is the demography of the area. The data scrapping is still in progress due to the data volume.

Data Cleanness Assessment

Prior to data cleaning, one important step is to evaluate the cleanness of the data and understand any potential issues within your dataset. In this case, we need to evaluate the cleanness of the joined data, since it is the primary data source used for future data analytics. A data quality score will be given to indicate the cleanness of the data set.

Data Quality Score and the method

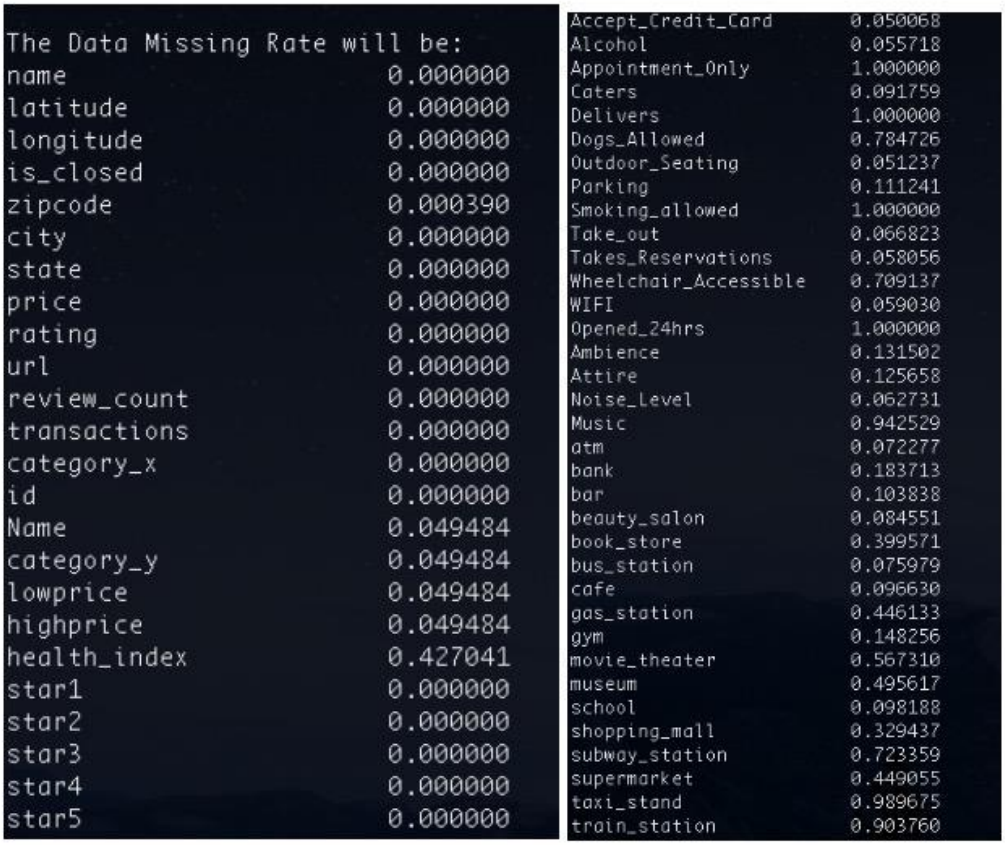
The first assumption we have made is that all dimensions carry the same feature importance for now, since we do not have any understandings about the contributions of each feature towards our targets. In this metric, each column has a base score of 10. We primarily assess the cleanness and validity of the data entry column by column. The base score for each column is deducted proportional to data missing rate for each column. And each invalid/out-of-range-data entry contributes towards a 0.5-point reduction until the base score for a particular column becomes 0. Each column score is added up together and the sum is divided by the total score to obtain an overall cleanness percentage.

We firstly calculate the missing rate for each column using Pandas operation. And based on heuristics and prior understanding about the dataset, we pre-define a list consisting whether the column is a context column or feature column, and we list down the maximum and minimum value for each numeric column. We then traverse through the columns to find out out-of-range values.

The Results and Interpretation

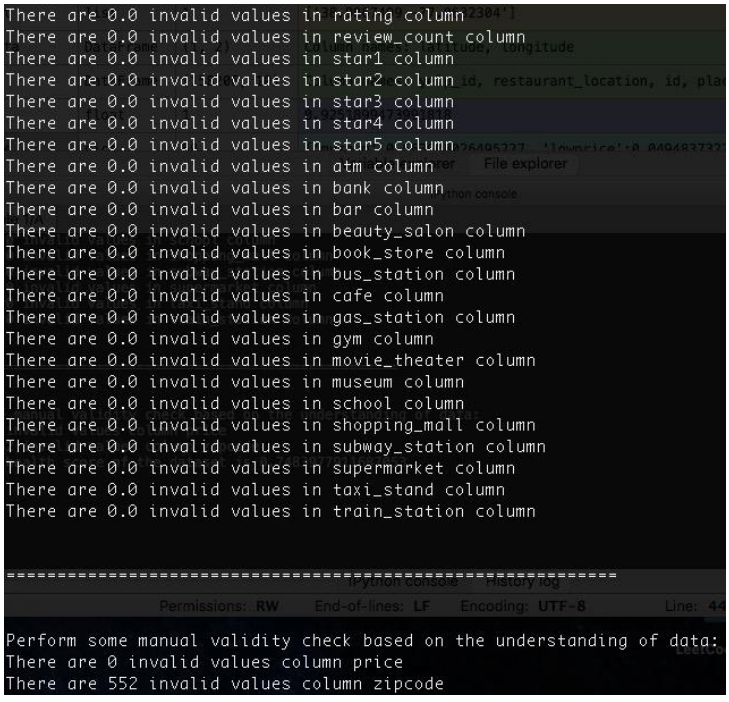
Figure 5 illustrates a screenshot for the heath report generated by our data cleanness check script. We can see that there are large variations in missing rates for different columns, and we can even observe total missing column in the dataset. However, some of the missing values do not indicate missing but rather ‘0’. Special considerations will be taken care during actual data cleaning stage.

Figure 6 illustrates the invalid data entries in the dataset. Upon scrutinization, there are few invalid data entries errors. The data entry error on ‘zipcode’ column is due to data type conversion, can be easily corrected by adding leading ‘0’ before the data entry.



```
The Data Missing Rate will be:
name                    0.000000
latitude                0.000000
longitude               0.000000
is_closed               0.000000
zipcode                 0.000390
city                    0.000000
state                   0.000000
price                   0.000000
rating                  0.000000
url                     0.000000
review_count            0.000000
transactions            0.000000
category_x              0.000000
id                      0.000000
Name                    0.049484
category_y              0.049484
lowprice                0.049484
highprice               0.049484
health_index            0.427041
star1                   0.000000
star2                   0.000000
star3                   0.000000
star4                   0.000000
star5                   0.000000
Accept_Credit_Card      0.050068
Alcohol                 0.055718
Appointment_Only        1.000000
Caters                  0.091759
Delivers                1.000000
Dogs_Allowed            0.784726
Outdoor_Seating         0.051237
Parking                 0.111241
Smoking_allowed         1.000000
Take_out                0.066823
Takes_Reservations      0.058056
Wheelchair_Accessible   0.709137
WIFI                    0.059030
Opened_24hrs            1.000000
Ambience                0.131502
Attire                  0.125658
Noise_Level             0.062731
Music                   0.942529
atm                     0.072277
bank                    0.183713
bar                     0.103838
beauty_salon            0.084551
book_store              0.399571
bus_station             0.075979
cafe                    0.096630
gas_station             0.446133
gym                     0.148256
movie_theater           0.567310
museum                  0.495617
school                  0.098188
shopping_mall           0.329437
subway_station          0.723359
supermarket             0.449055
taxi_stand              0.989675
train_station           0.903760
```

Figure 5 Missing Rate by Columns



```
There are 0.0 invalid values in rating column
There are 0.0 invalid values in review_count column
There are 0.0 invalid values in star1 column
There are 0.0 invalid values in star2 column
There are 0.0 invalid values in star3 column
There are 0.0 invalid values in star4 column
There are 0.0 invalid values in star5 column
There are 0.0 invalid values in atm column
There are 0.0 invalid values in bank column
There are 0.0 invalid values in bar column
There are 0.0 invalid values in beauty_salon column
There are 0.0 invalid values in book_store column
There are 0.0 invalid values in bus_station column
There are 0.0 invalid values in cafe column
There are 0.0 invalid values in gas_station column
There are 0.0 invalid values in gym column
There are 0.0 invalid values in movie_theater column
There are 0.0 invalid values in museum column
There are 0.0 invalid values in school column
There are 0.0 invalid values in shopping_mall column
There are 0.0 invalid values in subway_station column
There are 0.0 invalid values in supermarket column
There are 0.0 invalid values in taxi_stand column
There are 0.0 invalid values in train_station column


============================================================
Perform some manual validity check based on the understanding of data:
There are 0 invalid values column price
There are 552 invalid values column zipcode
```

Figure 6 Invalid Data Entries by Columns

The overall data quality score in this dataset is 0.7032. The score implies that data cleaning will be a must step before any further feature engineering and data cleaning. The biggest contributors of the score loss is missing data, and we will have to deal the miss case by case in later stage of the project.

Feature Generation

In this section, we generate three features from the data set. First, we assign letter grade to represent popularity of a restaurant by categorizing review count. We believe that if the restaurant has too few reviews on Yelp, the rating could be biased; thus, we assign 'biased' if the restaurant has less than 10 reviews. The letter grade goes up from 'F' to 'A' when the restaurant has 200 more reviews. The top-grade A means more than 1000 reviews are written below the restaurant on the Yelp website. This feature helps us interpret popularity of a restaurant.

Second feature is to represent price level in a numeric form. On the Yelp website, the average price of a restaurant is represented as '$$$$', '$$$', '$$', '$' from the most expensive to least expensive. We assigned 4 to 1 accordingly to help with analysis.

Third feature is to sum the transport spots including bus stations, subway stations, train station and taxi stand and then categorize them into 4 levels of convenience. The new feature evaluates the convenience of transportation in a comprehensive manner.

Next Step

Our next step is to reorganize data set in an organized format. We are also interested in adding more demographic information of each by zip code. At the same time, we will attempt to use EDA (exploratory data analysis) to summarize and visualize main characteristics of data. Finally, we will apply the techniques that we learn in this class to help diagnose the dataset.

Reference:

Robson, S. (2011). That's the spot! Strategies for finding the ideal restaurant site [Electronic version]. Restaurant Startup and Growth, 8(5), 24-29. Retrieved October 6th, 2018, from Cornell University, School of Hospitality Administration site: http://scholarship.sha.cornell.edu/articles/146/

Keythman, Bryan. (n.d.). What Are Some Qualities or Characteristics That Make a Good Restaurant? Small Business - Chron.com. Retrieved from http://smallbusiness.chron.com/qualities-characteristics-make-good-restaurant-38863.html