Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : “Improving Restaurant Reputation Using Yelp User Reviews”
Part 1: Data Description & Data Wrangling

How to improve restaurant reputation even just one star using Yelp user reviews !
“ The path from being normal to good, and good to perfect !"

Data Description:

The dataset used in this project is part of the dataset which is provided by Yelp for its round 10th challenge[1]. The whole dataset includes several datasets (business, checkins, photos, review, tip, user). Based on my capstone proposal and the attributes[2] of the mentioned sub datasets, for this project I just need two specific sub datasets, “review” and “business”. Both of these sub datasets have .json format.
The datasets have been downloaded from Yelp server on January 9, 2018 and kept locally during this project. I opened the review and business datasets separately as CSV format.

Data Attributes:

[1] https://www.yelp.com/dataset/challenge
[2] https://www.yelp.com/dataset/documentation/json

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : “Improving Restaurant Reputation Using Yelp User Reviews”
Part 1: Data Description & Data Wrangling

Business

Dimension :

174567 * 15

Memory Usage:

20 MB

Attributes:

Name	Description	Type
business_id	22 character unique string business id	object
name	Business’s name	object
neighborhood	Business’s neighborhood	object
address	The full address of business	object
city	The city	object
state	2 character state code	object
postal_code	The postal code	object
latitude	Latitude	float64
longitude	Longitude	float64
stars	The star rating (rounded to half-stars)	float64
review_count	Number of reviews	int64
is_open	0 or 1 for closed or open, respectively	int64
attributes	Business attributes to values	object
categories	An array of strings of business categories	object
hours	An object of key day to value hours	object

Review

Dimension:

5261669 * 9

Memory Usage:

361 MB

Attributes:

Name	Description	Type
review_id	22 character unique review id	object
user_id	22 character unique user id	object
business_id	22 character unique business id	object
stars	Star rating	int64
date	Date format YYYY-MM-DD	object
text	The review itself	object
useful	Number of useful votes received	int64
funny	Number of funny votes received	int64
cool	Number of cool votes received	int64

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 1: Data Description & Data Wrangling

By looking at the attributes of these two datasets and my goal in this project, initially I focused on some of these attributes:

Attributes	Example	Attributes	Example
business_id	tnhfDv5Il8EaGSXZGiuQGg	name	Garaje
user_id	Ha3iJu77CxlrFm-vQRs_8g	state	CA
postal_code	94107	review_count	1198
stars	4.5	date	2016-03-09
categories	["Mexican","Burgers","Gastropubs"]	attributes	{"RestaurantsTakeOut":true, "BusinessParking": {"garage": false,"street": true,"validated": false,"lot": false,"valet": false},}
text	Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks.		

- **Filtering "Restaurants" from our original dataset**

Since my project is just about the "Restaurants", I have to separate the "Restaurant" category. I looked into the "categories" column to figure out what are the **words** that point out to restaurants and foods. For now, there is no "Null" entry for categories column, but I have to look more in depth to see if we have some untidy entries. Initially I divided the original business dataset into "restaurant" and "non_restaurant" subsets.
Glancing to the "categories" column for "non_restaurant" subset, I noticed another similar word to "Restaurant" which is **Food** . There are around 14500 observations that ***do not have RESTAURANT but have FOOD as one of their categories tags.*** So I need to look at this tag (Food) more carefully. However I need to focus more on the entries that have "Food" as one of their categories, since some of them may provide some products or services that are related to the food (such as groceries, etc.) and **Not Serving Food** .
Looking at the part of the data that have **Food** and not **Restaurant** I figured out that majority of them have "Coffee and Tea" , "Grocery", "Ice Cream and Frozen Yogurt", "Bakeries", "Convenience Store" , and so on. So let's just keep the ones that have **Restaurant**, since there are enough observations (more than 54 K) and keeping the ones with **Food** may make it very complex and may bring some biases to our final analysis on reviews (for example the businesses that serve Italian ice cream or american coffee may affect our analysis of whole Italian or

American restaurants group). Finally, Finally, let's exclude the part of the data that has ***Grocery , Convenience Store, Coffee and Tea, Ice Cream and Frozen Yogurt, and Bakeries*** , Because some users might write review about the other parts and not about restaurant. (For example a restaurant that serves Ice Cream or Coffee and Tea may have reviews about these items and not food)

- **Filtering "Non_English" reviews**

My project is based on analysing user reviews (texts). Yelp is a globally used platform and the downloaded dataset includes non_english texts that are needed to be filtered. I did the filtering based on the "state" column of dataset.

- **Adding "Country" column**

Using "state" column of original dataset, I added another column "Country" for future analysis. Separating original dataset into 3 English spoken countries (USA, Canada, Britain - including UK, Scotland -) may expand our insight of this dataset.

- **Filtering 7 most popular types of Food (Adding "food_type" column)**

In this project, I also wondered whether type of food (Italian, American, Mexican, etc.) may affect factors or their weights that influence customers' text reviews and stars (ratings). Therefore, I checked "Categories" column and add another column ("food_type"). Since there are some restaurants that serve different types of food at the same time, I just filtered the ones that are specifically serve one of these 7 most popular types of food to avoid any complexity and bias that may combination of foods may bring into our analysis.

Eventually I came up with the final version of dataset , "**df_rest_eng_top7**" , which consists of ***English spoken restaurants which serve one of the top 7 popular types of food*** (American, Italian, Mexican, Chinese, Japanese, Indian, and Thai) based on the original dataset.

American	8401
Italian	3788
Mexican	3604
Chinese	3318
Japanese	1972
Indian	1209
Thai	989

Number of restaurants (in English Spoken Countries) based on each type of food

Our final dataset has 23281 entries.
Since the ultimate goal of this project is coming up with the model that can predict the rating of a restaurant based on its user reviews and type of food, I had to check whether there are enough observations (number of restaurants and review counts) for any type of food and

stars, as you can see in the below tables there are almost enough observations for each food type and star.

food_type	stars	count	mean	sum
American	1.0	21	4.952381	104
	1.5	99	15.292929	1514
	2.0	398	26.060302	10372
	2.5	965	48.964767	47251
	3.0	1821	61.663921	112290
	3.5	2273	92.252970	209691
	4.0	1970	160.464975	316116
	4.5	717	137.789400	98795
	5.0	137	20.087591	2752
Chinese	1.0	19	6.421053	122
	1.5	41	8.634146	354
	2.0	188	20.265957	3810
	2.5	410	29.087805	11926
	3.0	852	35.836854	30533
	3.5	962	59.067568	56823
	4.0	644	64.026398	41233
	4.5	172	46.046512	7920
	5.0	30	6.533333	196
Indian	1.0	2	3.000000	6
	1.5	14	10.642857	149
	2.0	27	11.407407	308
	2.5	99	12.717172	1259
	3.0	204	23.058824	4704
	3.5	338	49.559172	16751
	4.0	332	67.650602	22460
	4.5	170	52.200000	8874
	5.0	23	6.391304	147
Italian	1.0	28	7.321429	205
	1.5	88	8.863636	780
	2.0	180	15.450000	2781
...		...	...	...
	4.0	1042	95.005758	98996
	4.5	431	78.969838	34036
	5.0	75	11.946667	896
Japanese	1.0	6	6.000000	36
	1.5	10	20.900000	209
	2.0	48	28.416667	1364
	2.5	135	37.155556	5016
	3.0	357	54.577031	19484
	3.5	564	92.382979	52104
	4.0	574	136.843206	78548
	4.5	241	142.236515	34279
	5.0	37	16.378378	606
Mexican	1.0	15	6.266667	94
	1.5	65	15.630769	1016
	2.0	192	24.322917	4670
	2.5	475	37.947368	18025
	3.0	741	49.659919	36798
	3.5	887	80.350620	71271
	4.0	792	108.862374	86219
	4.5	360	82.805556	29810
	5.0	77	16.350649	1259
Thai	1.0	2	4.500000	9
	1.5	13	6.307692	82
	2.0	46	11.217391	516
	2.5	60	24.550000	1473
	3.0	134	27.552239	3692
	3.5	269	77.334572	20803
	4.0	331	126.395770	41837
	4.5	116	112.646552	13067
	5.0	18	8.055556	145

Number of restaurant (count) and number of reviews (sum) for each type of food and star

This is the general info of our final dataset (Restaurants in English Spoken Countries and serve one of the seven most popular foods):

```
business_id     23281 non-null object
name            23281 non-null object
neighborhood    9786 non-null object
address         23196 non-null object
city            23281 non-null object
state           23281 non-null object
postal_code     23256 non-null object
latitude        23281 non-null float64
longitude       23281 non-null float64
stars           23281 non-null float64
review_count    23281 non-null int64
is_open         23281 non-null int64
attributes      23281 non-null object
categories      23281 non-null object
hours           23281 non-null object
country         23281 non-null object
food_type       23281 non-null object
```

I also took a deep dive into the dataset and found out that there is no missing value for the columns that I need in my analysis (I do not need "neighborhood" column in my analysis).
Review dataset also consists of texts, so there is no missing value in the datasets.

The main columns which I used in this project are review texts, star ratings and food type, so there is no outliers or at least there is no obvious outlier at this stage of project. However some text pre-processing techniques have been done on "review" datasets which are mentioned in the next section.