2 Data

In dataset for the machine learning algorithm, each row is information of the restaurant. I select distance from the office, price level, ratings by Foursquare users, the number of shared tips and photos by Foursquare users as features of my data. I used Foursquare API to collect data

Foursquare API will provide information about places including non-restaurant, so I need to do feature engineering to remove rows which is useless for this project. First, I drop the duplicated rows by place id, 285 rows remained. Second, I remove rows whose 'categories' is non-restaurant and like café or Bar where we will not go to have lunch. After these two steps, 173 restaurants' data retained

The feature 'categories' in the dataset is important information about restaurants, but it refines restaurants into too many types. For example, Janguh Restaurant category only has one place in the dataset, which actually can be in Seafood Restaurant category. For another, someplace is an American Restaurant, some are Sandwich Place, some are Taco Place, all of these only have less than 2 places in the same category. So I create 'New_categories' to turn 36 'categories' into 9 'New_categories'

	name	categories	lat	lng	id	Station	New_categories	distance
3	불이아 (弗二我)	Chinese Restaurant	37.502396	127.036219	52aa9d0111d25f79bea32b9d	Yeoksam	Chinese Restaurant	0.010907
4	Yang Good (양국)	BBQ Joint	37.502146	127.034727	5703972bcd107356d0a6cee9	Yeoksam	BBQ Joint	0.010252
5	고갯마루집	Korean Restaurant	37.499361	127.039308	4d12cb2a12916dcb3d1fd98a	Yeoksam	Korean Restaurant	0.009911
6	BAS BURGER (바스버거)	Burger Joint	37.499723	127.035882	57c65f00498ed5b609eaf4ef	Yeoksam	Burger Joint	0.008299
10	대우식당	Korean Restaurant	37.502750	127.035137	4b8a4a0ef964a520246732e3	Yeoksam	Korean Restaurant	0.010933

About information shared by Foursquare users, I find that many restaurants have more than 50 tips or photos shared. I think the number of shared tips and photos is very important for the decision, but too many tips are useless. Because when I'm searching for the lunch restaurant in office before lunchtime, I see many blog posts or youtube videos on the internet but no more than 30~40 contents. So I refine the number of shared tips and photos to 5 levels. 1st level means there are less than 5 tips that not enough to use for make decision, 5th level means there are more than 30 tips which can fully help to make a decision. Like the table below.

	id	price	rating	photos_count	tips_count	Photos_count_level	tips_count_level
0	52aa9d0111d25f79bea32b9d	1.0	8.0	84	13	5.0	3.0
1	5703972bcd107356d0a6cee9	2.0	8.3	134	20	5.0	3.0
2	4d12cb2a12916dcb3d1fd98a	2.0	8.4	69	16	5.0	3.0
3	57c65f00498ed5b609eaf4ef	1.0	8.0	30	7	4.0	2.0
4	4b8a4a0ef964a520246732e3	2.0	8.2	169	26	5.0	4.0

As I mentioned before, this project's purpose is to predict ratings of restaurants near Yeoksam

station, so I need my ratings of restaurants near the office, Gangnam station, Sinnonhyeon station ,and Yangjea station. I do survey myself based on memos, Instagram, blog and more, add My rating to the dataset. What if there are restaurants which I never visited even near the office? I will keep it blank to predict, but fortunately, except for 44 restaurants near Yeoksam station, I had been all of the restaurants in dataset! Thanks to my curiosity about the new taste, 3 years was enough time to visit these 129 restaurants.

Now, the dataset is ready, including information of restaurants ID, distance, categories, price, four square user rating, counts of shared tips and photos, and my rating.