< The Battle of Neighborhoods>

Setup a Japanese Restaurant in Gyeonggi

By: Pierre Shi

Version: 1.0

Version Date: Jan 14, 2021

Table of Contents

1. Introduction

1.1 Background

Gyeonggi Province (Korean) is the most populous province in South Korea. Gyeonggi-do can be translated as "province surrounding Seoul". A group of Japanese investors want to setup some business in this province, but they do not know where to start, so they suggest to open a Japanese restaurant first to evaluate the local business environment.

The question is where to setup such a business is the best choice. A successful start-up could indicate a promising future while a failed investment may frustrate the investors. Comparing the current restaurant setup, using c location data seems to be an idea.

In this project, I am going to leverage the Foursquare location data together with the Gyeonggi cities' population and other restaurant information, to find out where the best place to maximize the success chance of this new Japanese restaurant.

1.2 Business Problem

The main business problems to be solved using advanced data analysis method are:

1. Is there is a clear relationship between the ‘city population’ and ‘number of Japanese restaurant’?
2. How can we identify the best fit cities for the new Japanese restaurant business, based on the population, existing competitor situation?
3. Are there other factors can be used to support this finding, such as using the ‘population density’ information?

Eventually, we want to achieve the objective to find 2-3 city candidates, with relatively higher chance of business success, based on enough population and less competition.

2. Data

This is the description of the data source and method been used in this project:

- Wiki Data for the Korea cities used to provide the most recent population and density information.
- Latitude and Longitude data are added to the different cities
- Then using Foursquare data source, to extract the venues information is the next step.

Combining this information, we can analyze the K-means clusters for the population and current restaurant setup, to find the best city to start the business.

2.1 Download data from Wiki

The list of cities in Korea are collected from a Wikipedia article entitled “List of cities in South Korea” found at https://en.wikipedia.org/wiki/List_of_cities_in_South_Korea.

	City	Hangul	Hanja	Province	Population(2017)	Area	Density	Founded
0	Andong	안동시	安東市	North Gyeongsang	168226	1521.26	110.6	1963-01-01
1	Ansan	안산시	安山市	Gyeonggi	689326	149.06	4624.5	1986-01-01
2	Anseong	안성시	安城市	Gyeonggi	182784	553.47	330.3	1998-04-01
3	Anyang	안양시	安養市	Gyeonggi	598392	58.46	10235.9	1973-07-01
4	Asan	아산시	牙山市	South Chungcheong	303043	542.15	559.0	1986-01-01

Fig. 2.1 Data of Korea cities from Wiki website

This data include the following information needed for further analysis

- City name
- Province
- Population
- Density

2. 2 Data clean up and add latitude and longitude

2.2.1 Data Cleanup

During data cleanup step, we have made the following changes

- Drop some unnecessary columns, such as “Hangul” (Korea name) and “Hanja” (Chinese name)
- Converting the data type – Density and Population from object type to float 64

2.2.2 Add the latitude and longitude to the cities

Sometimes the city itself may not be able to get the latitude and longitude data, especially when some of them share the same English names, for example, “Gwangju”. To get an accurate geocoordinates the

“none” values in column ‘province’ have been replaced by “Korea” and used as combination with “City” to get the precise latitude and longitude
We use the geopy 2.0 as a free service to obtain the latitude and longitude coordinates. This method is simple and it returns appropriate values: we simply need to geocode a pandas dataframe.

The format of the value is verified

	City	Region	Area	Search Radius	Population	Density	Latitude	Longitude
1	Ansan	Gyeonggi	149.06	6888	689326.0	4624.5	37.321715	126.830860
2	Anseong	Gyeonggi	553.47	13273	182784.0	330.3	37.002048	127.172084
3	Anyang	Gyeonggi	58.46	4314	598392.0	10235.9	37.393853	126.957060

Fig. 2.2 Latitude and Longitude data added into Korea cities

Then a map is plotted to display the locations

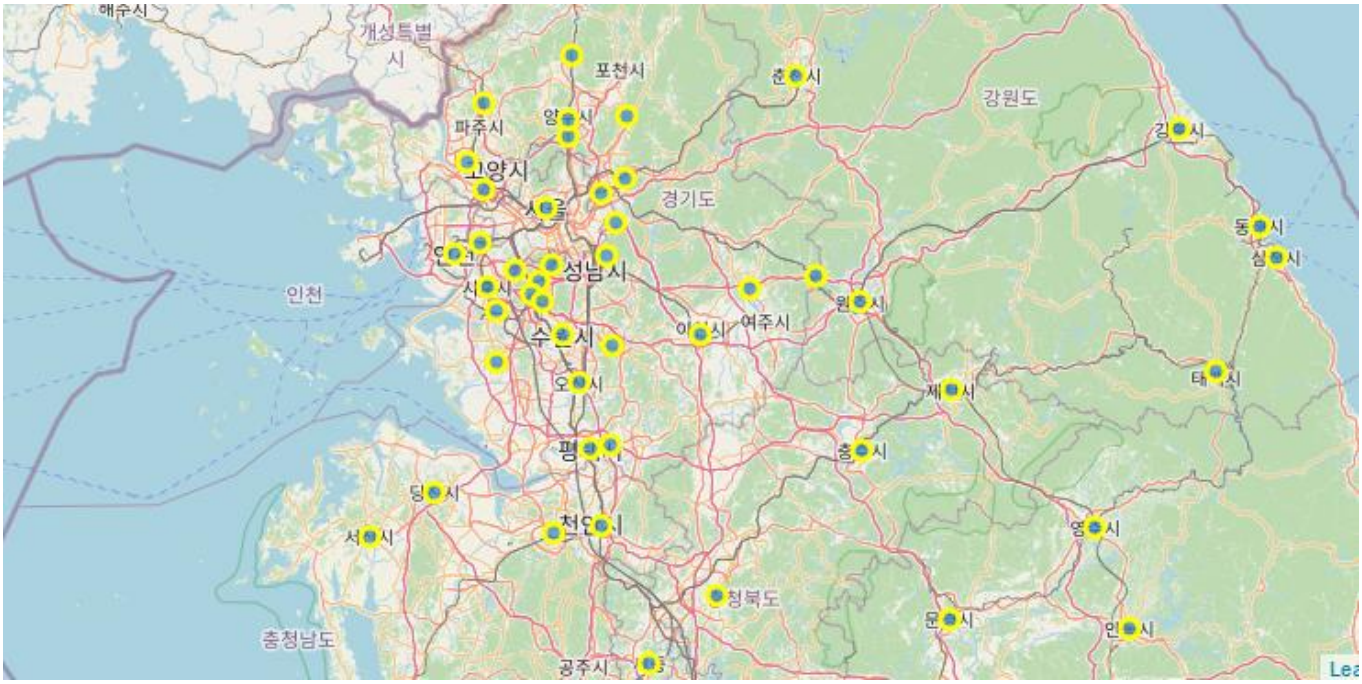

Fig. 2.3 Plotted Korea cities on the map

2.2.3 Remove the rows not belonging to Gyeonggi

Since our target is Gyeonggi, once latitude and longitude quality is verified, we remove those rows not belonging to Gyeonggi. This will make the further data analysis convenient and reduce chance of errors.

	City	Region	Area	Search Radius	Population	Density	Latitude	Longitude
1	Ansan	Gyeonggi	149.06	6888	689326.0	4624.5	37.321715	126.830860
2	Anseong	Gyeonggi	553.47	13273	182784.0	330.3	37.002048	127.172084
3	Anyang	Gyeonggi	58.46	4314	598392.0	10235.9	37.393853	126.957060
6	Bucheon	Gyeonggi	53.40	4123	851245.0	15940.9	37.484110	126.782735
16	Dongducheon	Gyeonggi	95.66	5518	98062.0	1025.1	37.927826	127.054782

Fig. 2.4 Removed data not belonging to Gyeonggi Province

2. 3 Venue data from Foursquare

Setup a personal Foursquare development account allows to have more API calls made in a day. Then use the account to call the database to get the venue information for each city in Gyeonggi.

1. The Foursquare API is used to get the top 100 venues in each city. API calls are made to Foursquare by passing the coordinates and search radius of each city in a Python loop. Here I have attempted to change the limit from 100 to 150, and the result seems to be the same. Fortunately, only some cities have reached 100 that limit.

2. A getNearbyVenues provided from the course learning is used to call the service. From there, all the 28 cities in the Gyeonggi province have returned a list of the venues. There are altogether 2177 rows, with 195 unique venue categories.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ansan	37.321715	126.83086	경기도미술관	37.325320	126.814133	Art Gallery
1	Ansan	37.321715	126.83086	Starbucks (스타벅스)	37.316738	126.837578	Coffee Shop
2	Ansan	37.321715	126.83086	Starbucks (스타벅스)	37.300614	126.838058	Coffee Shop
3	Ansan	37.321715	126.83086	E-Mart (이마트)	37.302689	126.813207	Supermarket
4	Ansan	37.321715	126.83086	일등토종순대감자탕	37.309712	126.869358	Korean Restaurant

Fig. 2.5 Venue data extracted from Foursquare data source

3 Methodology

3.1 Data Exploration

Using basic data exploration method, we can get a first glance of the data we got for analysis:

1. Analyze the venue categories, identify the main “competitors”, such as “Restaurant”, “Japanese Restaurant” and any restaurant types potentially can be competitors

2. A venue quantity per city is displayed, main concern is some of the cities reached the limit of 100, and could be missed key venue values due to the limitation

3. One-hot encoding method on “Venue Category” using pandas.get_dummies() method on the dataframe.

4. Finally, a top 10 venue category analysis is performed

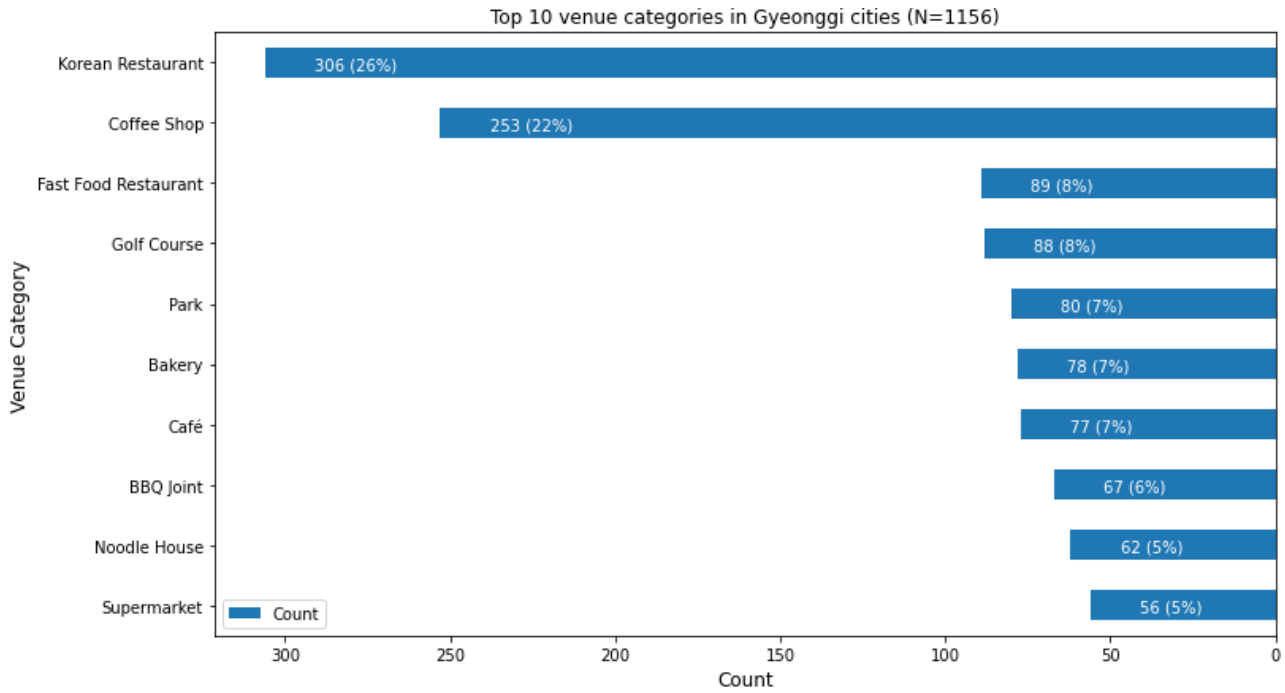

Fig. 3.1 top 10 venue categories in Gyeonggi cities

This graph shows that Korea restaurant and Coffee shop have the biggest quantities in Korea, which makes sense. Noodle House is one type of Japanese restaurant which has a position in the top 10 list.

3.2 Correlation analysis – Population, Japanese Restaurant and Restaurant

There are three "objects" we are considering as most related to the analysis:

- The number of Japanese Restaurants in the city
- The city population

- And the number of restaurants in the city

We need to know the relationship between city Population and number of Japanese Restaurant, hopefully it is strong. We also need to consider the other restaurants in the city as for comparison.

Let us check the correlation between the number of Japanese restaurants and the city population.

3.2.1 Pearson Correlation analysis

A first scratch shows us, the Japanese Restaurant is not the most related category to the population. However, the result that “Noodle shop goes” higher than “Korea Restaurant” gives confidence the relationship between “Japanese Restaurant” and “population” exists.

3.2.2 Count the venue quantity – all restaurants and “Japanese Restaurants”

- The Japanese restaurant category will include all those “Sushi shop”, “Ramen shop”, “Noodle shop”.. any of these competitors
- The overall restaurant information is also useful to further analyze the correlation
- The “NaN” values are replaced by “0”

3.2.3 Plot the “number of restaurant” and “Japanese restaurant”

From this plot, we can see two things

- Japanese restaurant has a relative low proportion in the whole restaurant category
- Such proportion shows differently in different cities: sometimes, there is a very large amount of restaurants, but very few “Japanese restaurant”. This becomes a good potential target

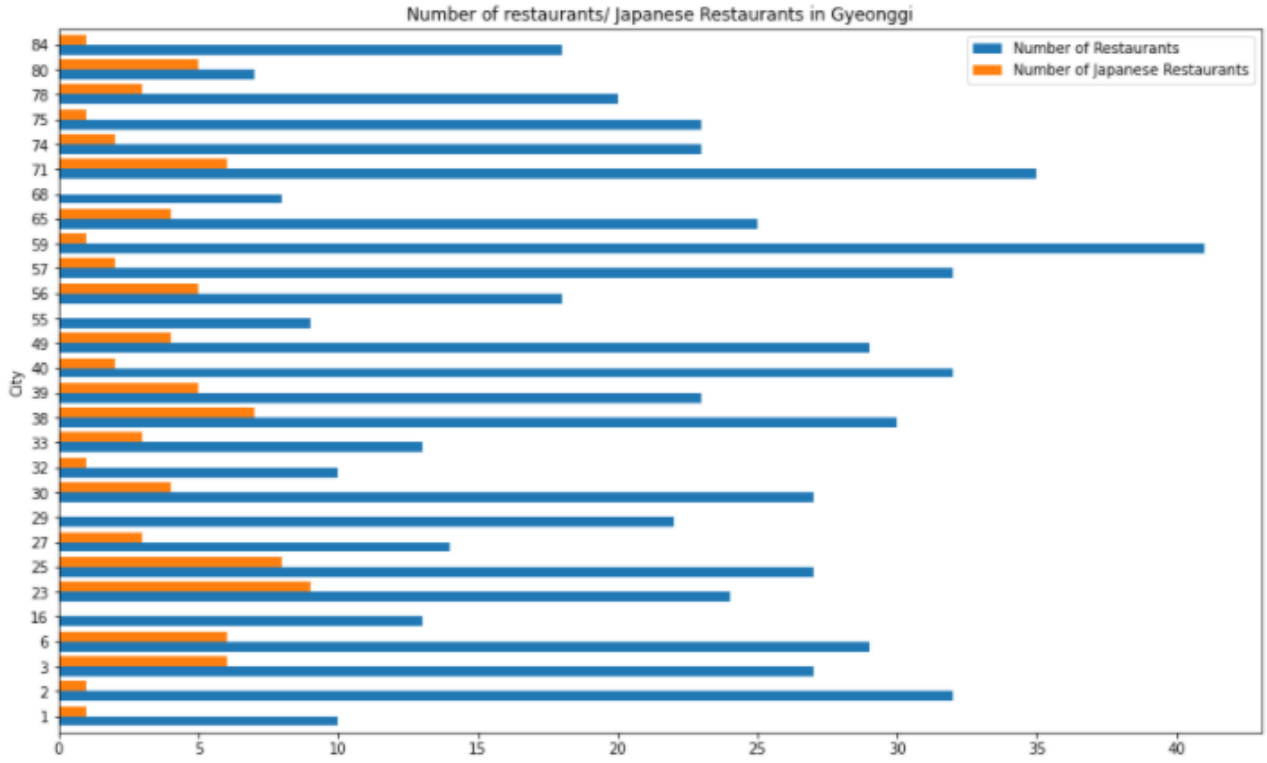

Fig. 3.2 number of restaurants and Japanese restaurants in Gyeonggi province

3.2.4 Plot the “City Population” and “Number of Japanese Restaurant”

The “Venue Category” of interest, Japanese restaurant, is selected from the summed one-hot encoded dataframe. Visualizations are performed to analyze data distribution. Using the sklearn.preprocessing library, these variables are fitted and transformed into new values, with mean = 0 and standard deviation = 1. These standardised values are stored in a new data frame using pandas.DataFrame() method, and a scatter plot with regression line is plotted using seaborn.regplot() function..

Fig. 3.3 Scattered plot of Japanese restaurant and population in Gyeonggi

The plot shows that positive correlative relationship between "Population" and "Japanese Restaurant" exists, but it is not very strong. This means we need to further analyze the data using clustering method to find where the most suitable place should be.

3.3 Cluster analysis – Population vs. Japanese Restaurant

To prepare for the clustering, the data for population and Japanese Restaurant are first normalized.

K-Means clustering is a type of partition clustering that divides data into K non-overlapping subsets or clusters without any cluster internal structure or labels. It is an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.

The standardized values were fitted using the KMeans() function from sklearn.cluster library, with number of clusters = 4. The generated labels and cluster centers are stored in variables. Using a for loop, generated labels, and cluster centers, the K-Means clusters are visualized.

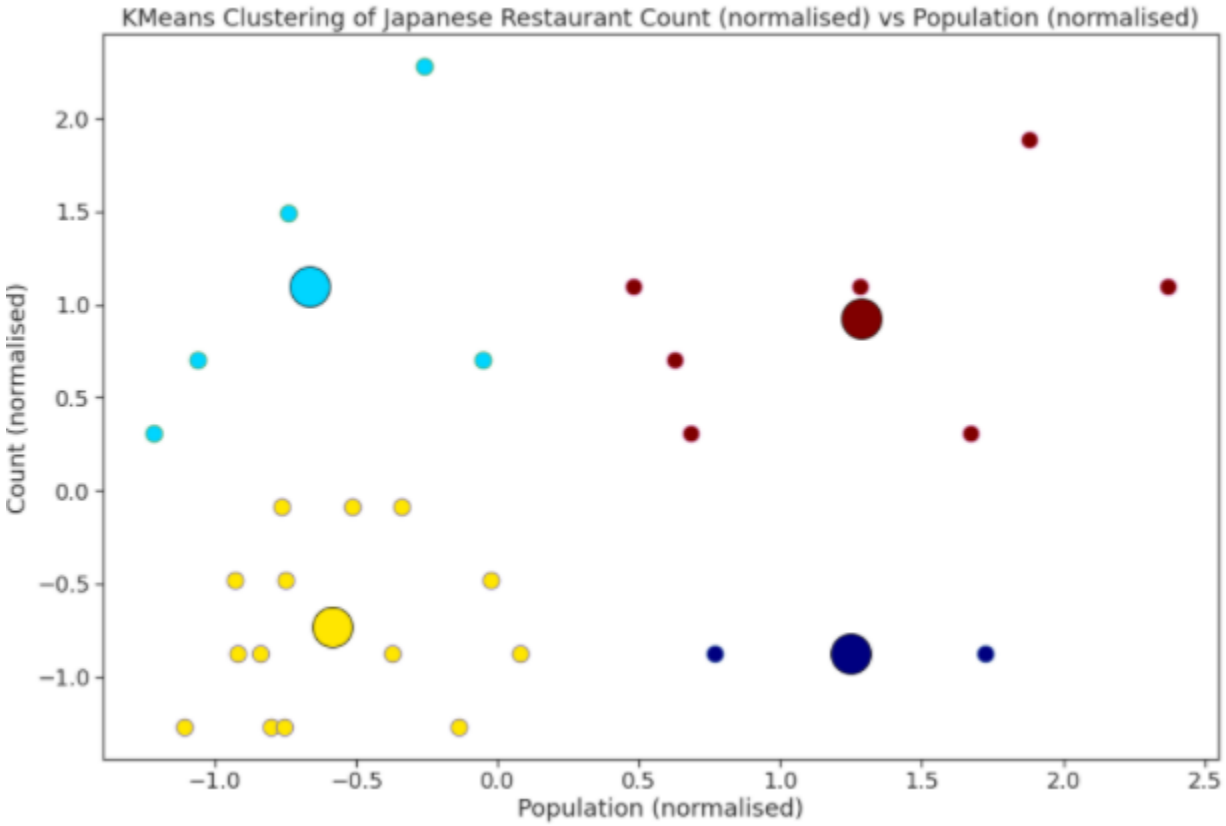

Fig. 3.4 KMeans clustering of Japanese restaurant count vs population

As shown, there are 4 different clusters and confirms the population and quantity of Japanese Restaurants exist, but not very clear. In fact, it separates the cities into 4 groups:
1. Group 1, yellow bubbles, with lower population and low quantity of competitors
2. Group 2, light blue bubbles, with lower population and high quantity of competitors
3. Group 3, red bubbles, with high population and high quantity of competitors
4. Group 4, dark blue bubbles, with high population and low quantity of competitors, which is what we most interested of

The 4 clusters have the following quantities of cities

```
[38]: 2    14
      3     7
      1     5
      0     2
      Name: Cluster Label, dtype: int64
```

Fig. 3.5 clusters found in terms of Japanese restaurant and population

From the analysis, clearly the cluster 0 is of the biggest interest:

	City	Region	Area	Search Radius	Population	Density	Number of Japanese Restaurants	Number of Restaurants	Cluster Label
1	Ansan	Gyeonggi	149.06	6888	689326.0	4624.5	1.0	10	0
84	Yongin	Gyeonggi	591.36	13720	991622.0	1676.8	1.0	18	0

Fig. 3.6 Cluster 0 which contains the target cities

These 2 cities

- Ansan (index 1), has medium high population, only has 1 Japanese restaurant, ratio is 9:1
- Yongin (index 84), has very high population, quite some restaurants but only 1 Japanese restaurant. However, because this city reaches limit of 100 venue, the result is questionable.

Actually, another 1 city in cluster 2 is also maybe one possible candidate: its population is medium and it does not have any Japanese restaurants.

68	Siheung	Gyeonggi	135.02	6556	403398.0	2987.7	0.0	8	2

Fig. 3.7 Potential target city in cluster 2

These 3 cities have a high population but very low quantity of competitors.
Because the relation between Population and Japanese Restaurant, although exists is not enough strong, we are going to further explore other evidence to support these findings.

3.4 Cluster analysis – Population Density vs. Restaurant Density

The density analysis is to investigate the relationship between two densities, the population density (in terms of area space) and the restaurant density (based on population).

It is based on an assumption that a high population density, may indicate more visits to restaurants. We hope this analysis can strengthen the conclusion we got from the last clustering investigation.

To achieve this target, we calculated the restaurant density and put it together with the population density, then we normalized the data

	Density	Number of restaurants per 1000 people
1	4624.5	0.014507
2	330.3	0.175070
3	10235.9	0.045121
6	15940.9	0.034068
16	1025.1	0.132569

Fig. 3.8 Normalized data for population density vs. restaurant density

Use the Elbow method to find the optimal K value to do the clustering analysis, the Elbow analysis suggested the K value should take 4.

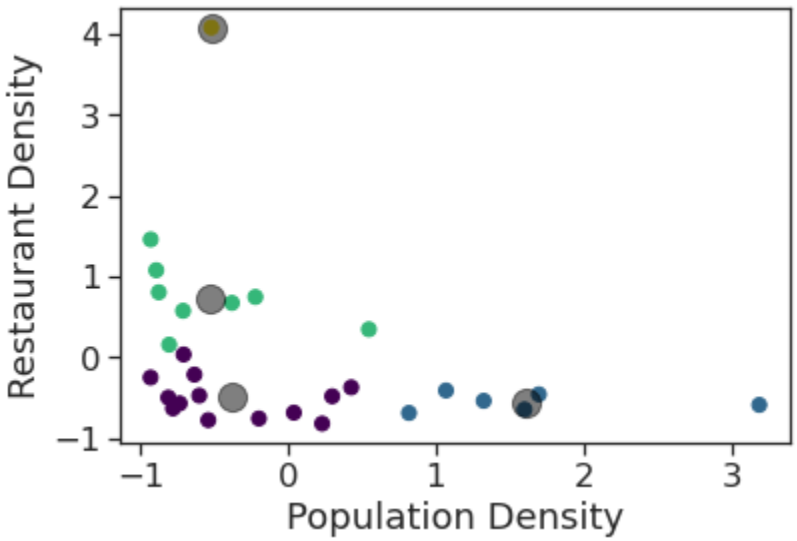

Fig. 3.9 Plotted population density vs. restaurant density

As shown this relation is not clearly strong. In fact there is two outier then suggesting “there is no absolute rules”. But overall, we are able to get 4 clusters
- Group 1 - the purple dots, has low population density, and low restaurant density
- Group 2 - the green dots, has low population density, and relatively high restaurant density
- Group 3 - the blue dots, has high population density, and low restaurant density
- Group 4 - the single outlier, has very high restaurant density as compared to its population density

Group 3 data (cluster 1 below)– high population density but low restaurant density:

	Kmeans-Cluster Labels	City	Population	Density	Number of Japanese Restaurants	Number of Restaurants	Cluster Label	Number of restaurants per 1000 people	Number of Japanese restaurants per 1000 people
6	1	Bucheon	851245.0	15940.9	6.0	29	3	0.034068	0.007048
3	1	Anyang	598392.0	10235.9	6.0	27	3	0.045121	0.010027
71	1	Suwon	1194276.0	9862.7	6.0	35	3	0.029306	0.005024
33	1	Gwangmyeong	339071.0	8807.0	3.0	13	2	0.038340	0.008848
27	1	Gunpo	284735.0	7831.0	3.0	14	2	0.049169	0.010536
65	1	Seongnam	974755.0	6873.2	4.0	25	3	0.025647	0.004104

Fig. 3.10 High population density cities with low restaurant density

This should be the main objective in our analysis, however, the result somehow suggested that, in many cases, there could be naturally a very high population density with lower restaurant density (by population) density. This is probably because in these cases, there is simply a lack of land to setup more restaurants. So we should not rely on this group to setup the target business, because first you may not be easy to find land, and second the current business competitors there could be pretty strong.

On the other hand, the Medium population (by area) density and medium low restaurant (by population) density (in the cluster 0 below) suggests a more suitable target area.

In fact the 3 objectives we identified through the first clustering analysis, index 1, 84 and 68 all sit in this cluster.

	Kmeans-Cluster Labels	City	Population	Density	Number of Japanese Restaurants	Number of Restaurants	Cluster Label	Number of restaurants per 1000 people	Number of Japanese restaurants per 1000 people
74	0	Uijeongbu	438753.0	5377.5	2.0	23	2	0.052421	0.004558
55	0	Osan	208873.0	4884.8	0.0	9	2	0.043088	0.000000
1	0	Ansan	689326.0	4624.5	1.0	10	0	0.014507	0.001451
25	0	Goyang	1040648.0	3893.0	8.0	27	3	0.025945	0.007688
68	0	Siheung	403398.0	2987.7	0.0	8	2	0.019832	0.000000
84	0	Yongin	991622.0	1676.8	1.0	18	0	0.018152	0.001008
49	0	Namyangju	662183.0	1444.1	4.0	29	3	0.043795	0.006041
23	0	Gimpo	364808.0	1318.7	9.0	24	1	0.065788	0.024671
59	0	Pyeongtaek	472141.0	1038.5	1.0	41	2	0.086838	0.002118
39	0	Hwaseong	644498.0	937.4	5.0	23	3	0.035687	0.007758

Fig. 3.11 Cities from first cluster shows in medium population density and low restaurant density cluster

This data has provided supporting evidence to the clustering analysis in the previous part. The 3 cities we found before, 1, 84 and 68 are all in the same cluster, which has relatively high population density, and relatively low restaurant density.

4 Results

When combining the cluster analysis based on ‘City Population’ and ‘number of Japanese restaurant’ with the analysis of ‘population’ and ‘restaurant’ densities, we can identify 3 target cities of the biggest interest to setup the new Japanese restaurant:

1. Ansan, original index 1
2. Yongin, original index 84
3. Siheung, original index 68

Their features can be summarized as below

City	Population in city	Population Density	Restaurant Density	Qty of Japanese Restaurant
Ansan	High-medium	High-Medium	Low	1
Yongin	Very High	Medium	Low	1
Siheung	Medium	Medium	Low	0

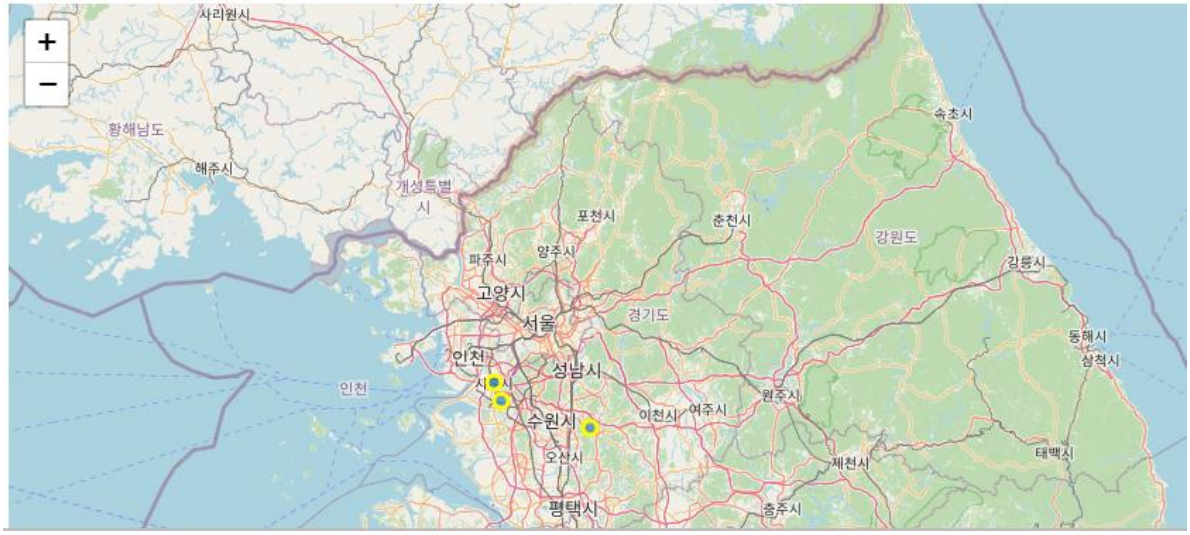

Fig. 3.12 The 3 selected cities and their location

There 3 cities meet the objectives such as

- High Population
- Low number of Japanese Restaurants
- Also a higher population density vs. a lower restaurant density.

These factors make it more possible to achieve success when setting up a Japanese restaurant inside the city.

Among these 3 objects, Yongin needs to be further confirmed of its Japanese restaurant number because it reached FourSquare limit of 100 venues. Ansan and Siheung are safer choices, because they did not reach FourSquare limit. So the data from Foursquare is more trustable.

5 Discussion

5.1 Approach

The strength of linear relationship between ‘city population’ and ‘number of Japanese restaurant’ is not very high. This means there are other factors need to be considered and included into the analysis. This could include for instance, the local population structure, as important factors to be considered.

On the other side, there is a couple of ‘Outliers’ in the analysis, especially in the density analysis. This type of object always worth individual analysis and find out what the reason causing such ‘outstanding’ status. The result may impact our selection consideration.

5.2 Source of Data

Even though the data source from Wiki and Foursquare are reliable and updated, there is clearly a limitation of the ‘venue information’: first a limitation to 100 per query and secondly, not all the venues are included on their list.

Considering the target objective ‘Japanese restaurant’ has a lower proportion and representation in some of the target cities, even a small miss of the venue cause may cause great deviation. This maybe a problem for one of the city we have chosen: Yongin.

Therefore before selecting the final target, these 3 suggested cities’ actual venue information needs to be further verified, probably through different sources of data or websites.

6 Conclusion

In our IBM Data Science Capstone final project, we analyzed location and current restaurant information of Gyeonggi, Korea, to determine where is the best spot city to setup a new Japanese restaurant.

The information comes from internet public source is downloaded, extracted, cleaned up for clustering analysis. Normalization is used before data modeling.

During analysis, several important statistical features such as relationship between variants are explored, investigated, and cross verified.

3 potential candidates cities, namely, Ansan, Yongin and Siheung, are identified as most attractive cities to setup the Japanese restaurant business. It is also recommended to have follow up further analysis of these cities, before a business decision been finalized.