Setup a Japanese Restaurant in Gyeonggi, Korea

IBM Applied Data Science Capstone

The Battle of Neighborhoods

By: Pierre Shi

Business Problem

Objective: find a best location in Gyeonggi to setup a Japanese restaurant

- Is there is a clear relationship between the ‘city population’ and ‘number of Japanese restaurant’?
- How to identify the best fit cities for the new Japanese restaurant business, based on the population, existing competitor situation, and other available information?

Data collection

- The Wiki Data for the Korea cities will be used to provide the most recent population and density information.
- Latitude and Longitude data are added to the different cities

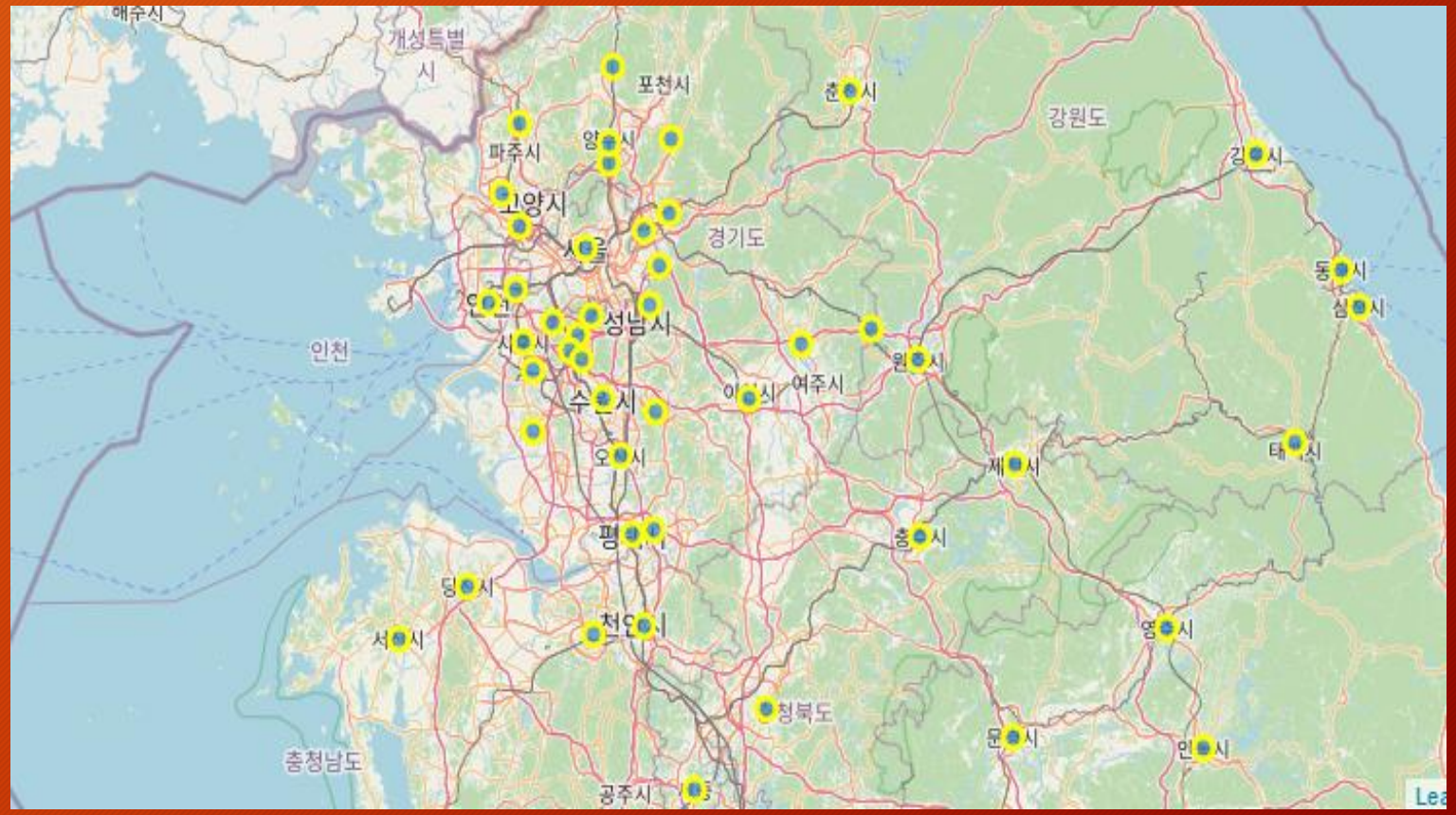
	City	Region	Area	Search Radius	Population	Density	Latitude	Longitude
1	Ansan	Gyeonggi	149.06	6888	689326.0	4624.5	37.321715	126.830860
2	Anseong	Gyeonggi	553.47	13273	182784.0	330.3	37.002048	127.172084
3	Anyang	Gyeonggi	58.46	4314	598392.0	10235.9	37.393853	126.957060

- Use Foursquare data source, to extract the venues information is the next step.

	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

Plot the Korea cities on the map based on Latitude and Longitude data

- Latitude and Longitude data is verified through plotting function
- This confirms the correctness of the downloaded data

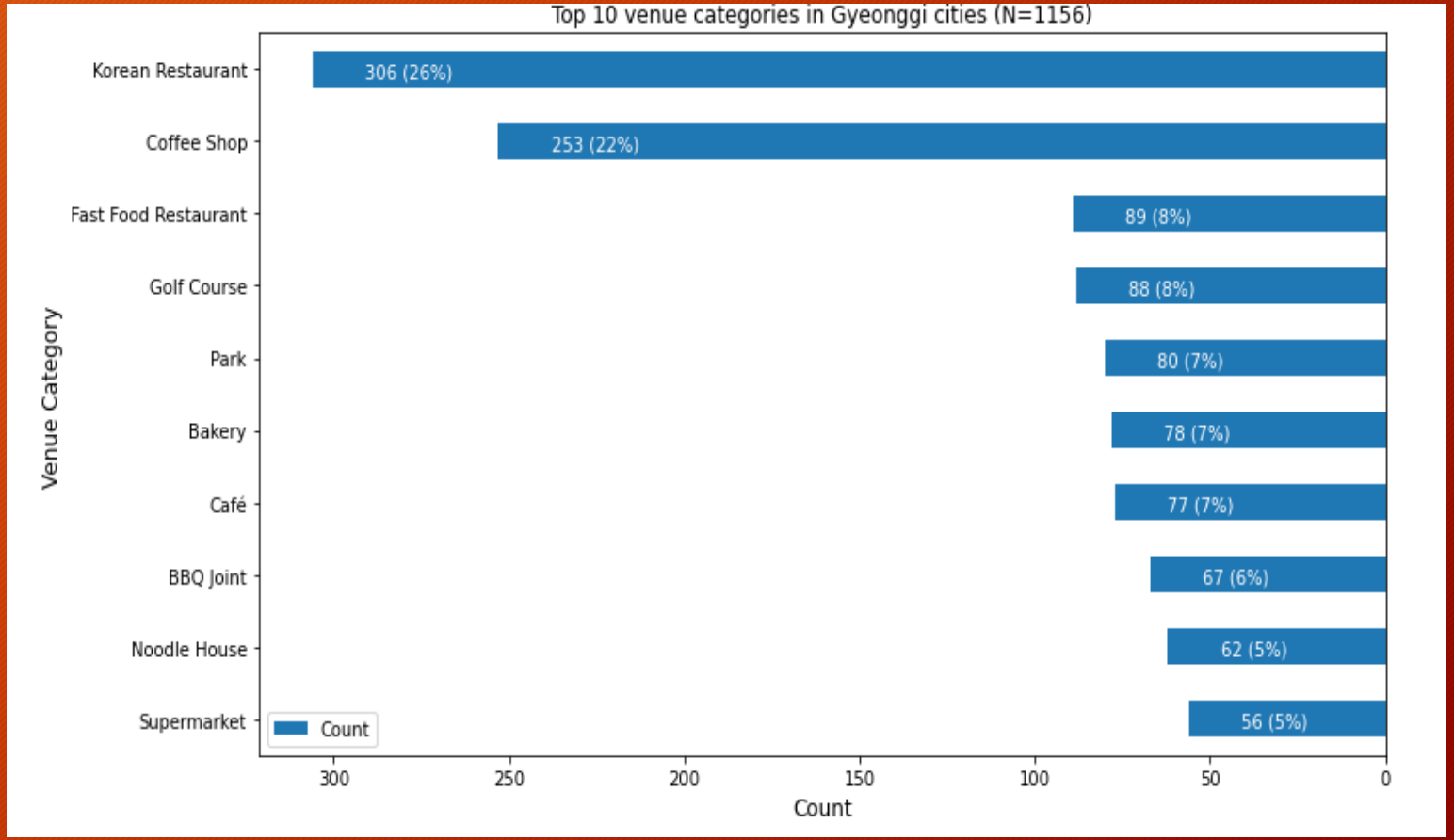


Methodology

- Plot the number of different venues in all the Gyeonggi cities
- Plot and compare the number of Japanese restaurant and Total restaurant in each city
- Analyze the relationship between “city population” and “number of Japanese restaurant”
- Cluster the cities based on Population and Number of competitors
- Cluster the cities based on Density information

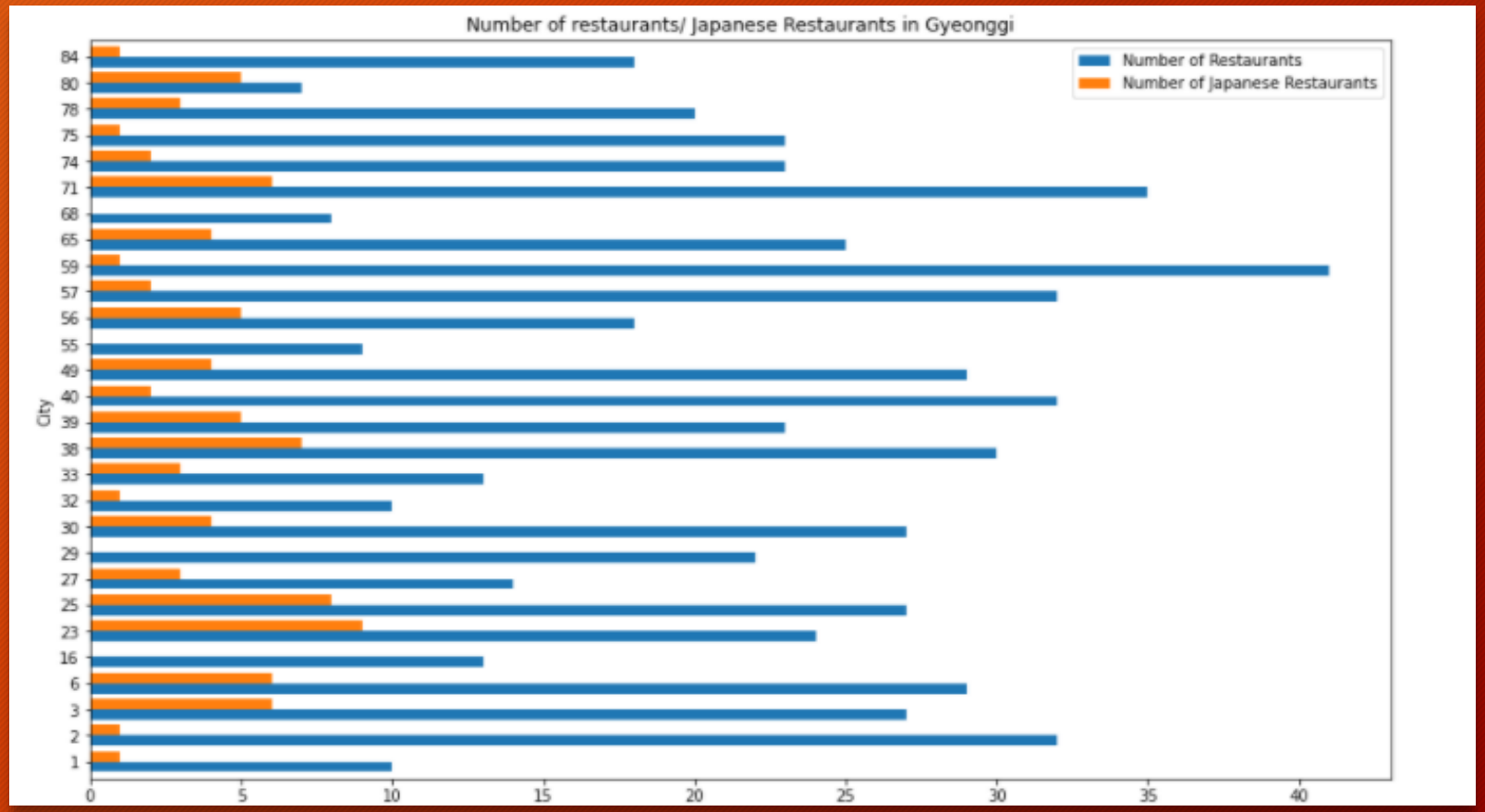
Plot the number of different venues in all the Gyeonggi cities

- Korea Restaurant has the largest quantity in all the venues, followed by coffee shops
- Noodle House is one type of Japanese restaurant which has a position in the top 10 list.



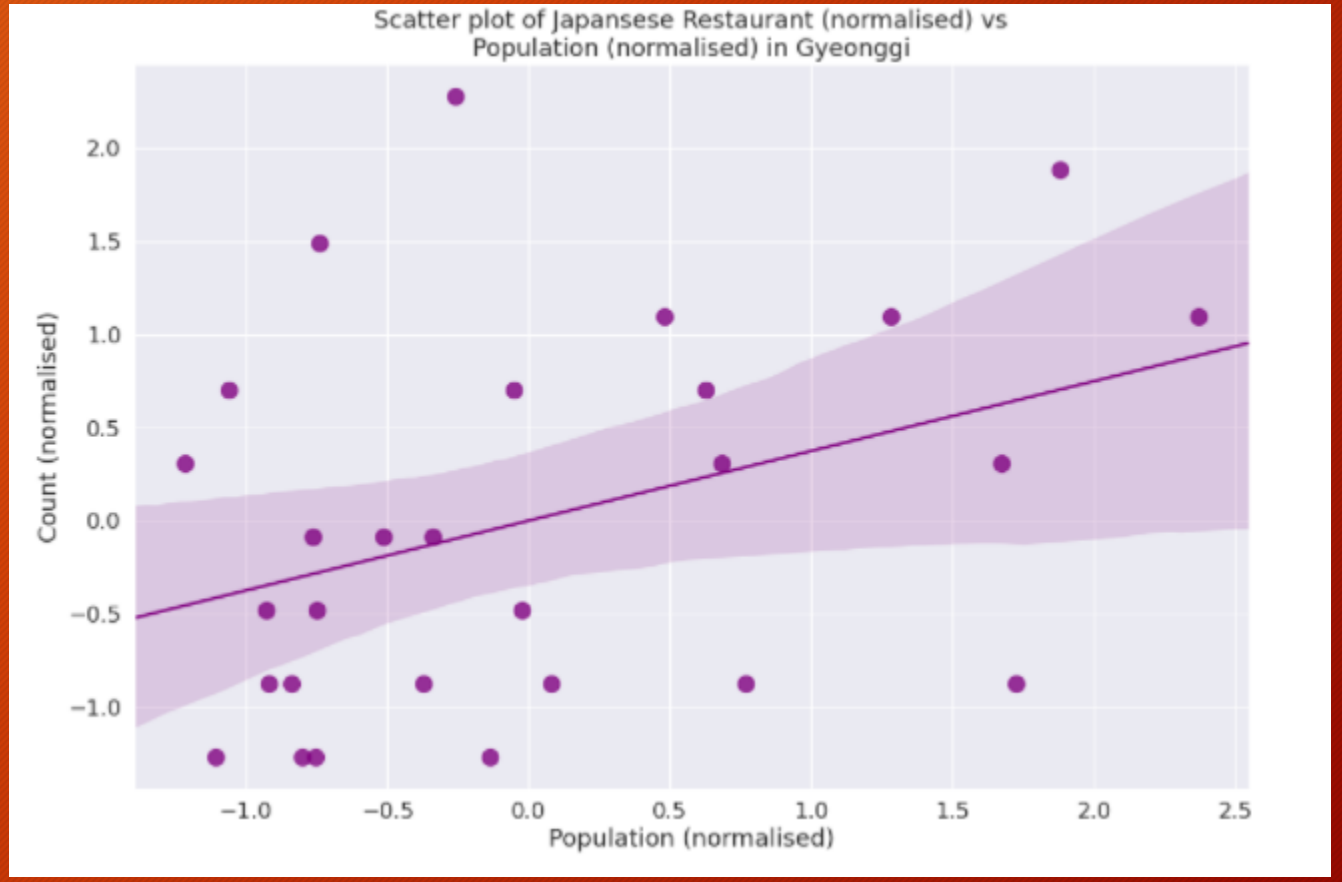
Plot the number of different venues in all the Gyeonggi cities

- In different cities, there is a different proportion of Japanese Styled restaurants



Analyze the relationship between “city population” and “number of Japanese restaurant”

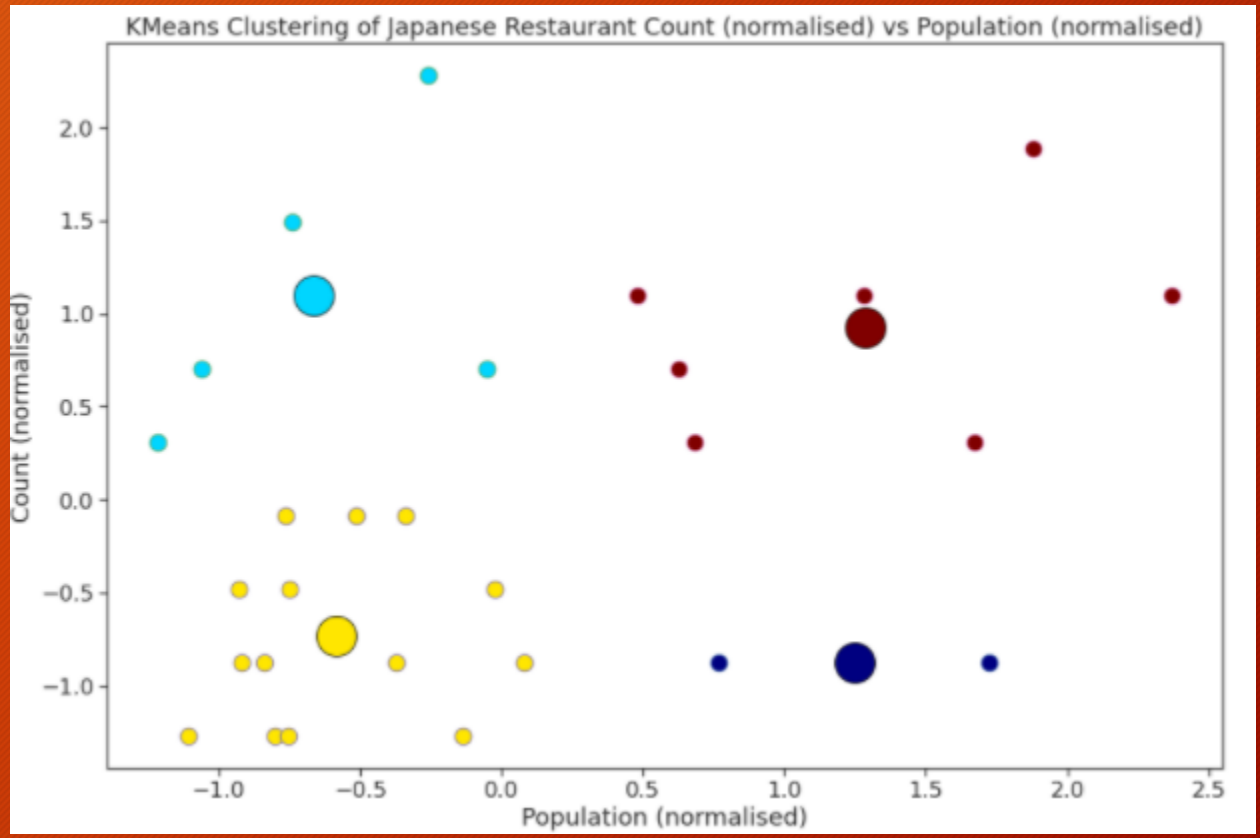
- One-hot encoded dataframe used for the Japanese restaurant venue category
- Visualization is performed to display data distribution
- Data is normalized to analyze
- Regression line is plotted; the relationship exists but not strong
- There are outliers



Cluster the cities based on Population and Number of competitors

- K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.

- 4 clusters identified (based on city population and count of Japanese restaurants)
  - Yellow bubbles
  - Light blue bubbles
  - Red bubbles
  - Dark blue bubbles



Clusters found

- 4 clusters were found during the analysis
- Cluster 0 has only 2 cities, they have high population and very low number of competitors

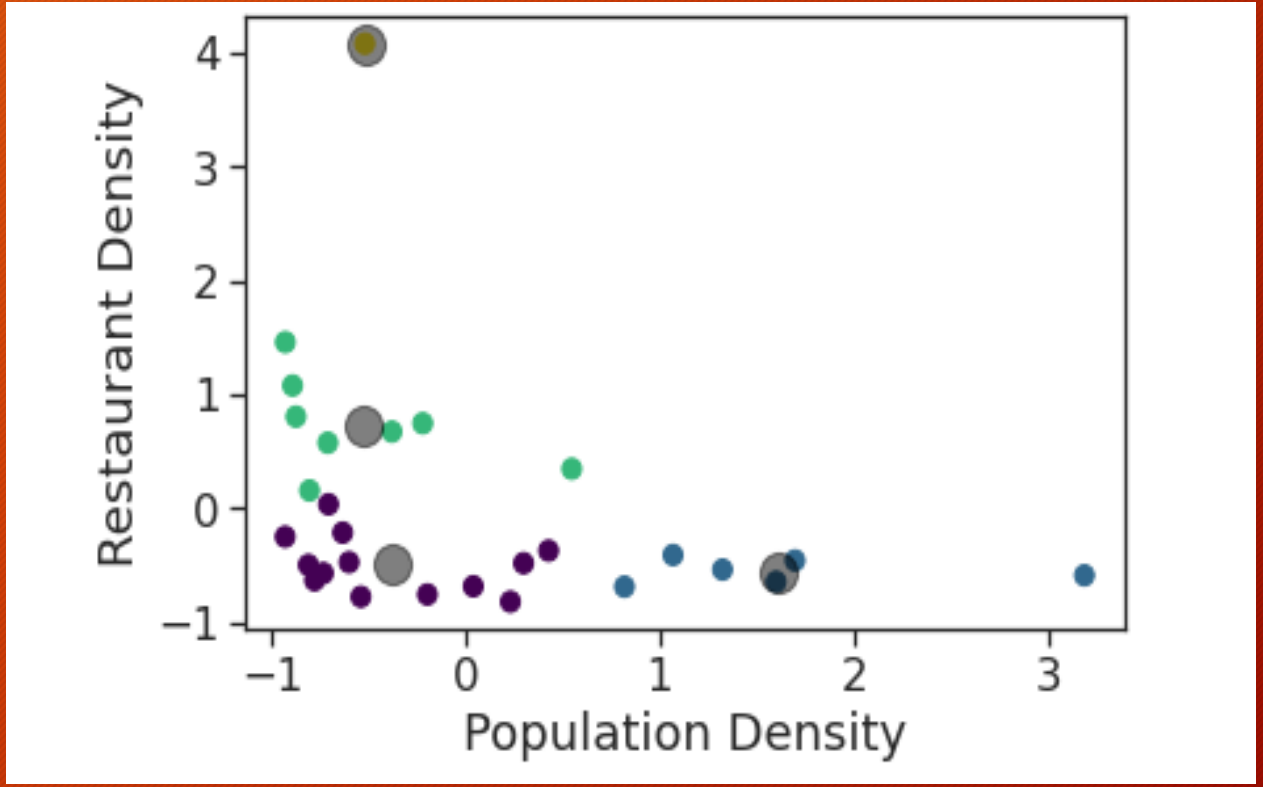
	City	Region	Area	Search Radius	Population	Density	Number of Japanese Restaurants	Number of Restaurants	Cluster Label
1	Ansan	Gyeonggi	149.06	6888	689326.0	4624.5	1.0	10	0
84	Yongin	Gyeonggi	591.36	13720	991622.0	1676.8	1.0	18	0

- In cluster 2: there is 1 city may also be suitable: medium population and no competitor

68	Siheung	Gyeonggi	135.02	6556	403398.0	2987.7	0.0	8	2

Cluster the cities based on Density information

- Population density = population/ area
- Restaurant density = restaurant / population
- 4 clusters identified (based on population density and restaurant density)
  - Purple dots
  - Green dots
  - Blue dots
  - Single outlier



Clusters found

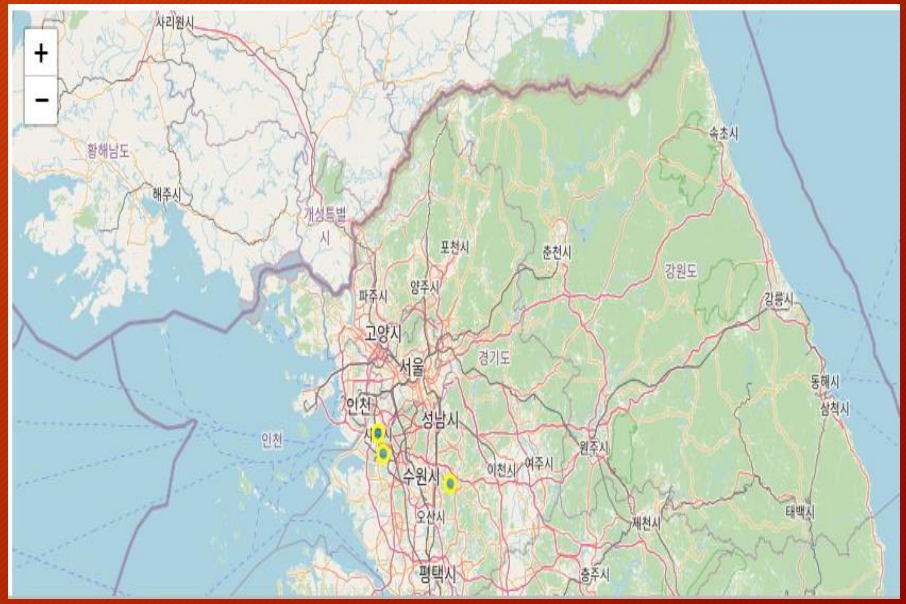
- 4 clusters found, all the interested cities are in same cluster 0, these cities have
- Medium population density
- Low restaurant density
- Result supports these 3 cities as good potential targets

	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

Results

- 3 cities found most suitable for the setup of new Japanese restaurant:
  Ansan, Yongin, and Siheung

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



Discussion

- Limitation of approach:
  - Relationship between “City population” and “number of Japanese restaurant” is not strong enough, suggesting other factors should be considered
  - How to overcome: collect and consider other data into further analysis, such as : local population structure, city growth phase..
- Limitation of data
  - Foursquare has a limit of 100 venues per city, and there are some venues not included in the database
  - How to overcome: consider collect data from other sources to further analyze and confirm the finding

Conclusion

- Information comes from internet public source is downloaded, extracted, cleaned up
- Important statistical features such as relationship between variants are explored, investigated, and cross verified.
- Clustering methodology is applied to analyze the data
- 3 potential candidates cities, namely, Ansan, Yongin and Siheung
- Further verification and confirmation is required