IBM capstone Project Course Week 4 Assignment

2020/05/05 Christianus Frederick Hotama

Project Title: Predict the Best District for the House Investment Based on the infrastructure facility

Introduction

Housing is not only the basic commodity for our life but also one of the best investment asset. Therefore, we need to choose wisely before buying it. Unfortunately, besides the house cost, understanding the value of the house is also quite challenging. The value of the house can change from time to time. There are a lot of factors that can change the value of the house. For example, the houses located in a large city usually have a higher value than others, but the cost can be downgraded if the neighborhood lack of public facilities. On the other hand, the house in the rural area can increase dramatically if the government suddenly plans to build a new subway station around that area.

The current twenties generation in Koreans has this issue right now. Right now they work hard so they can buy a new house in the prime area for their future. In this study, the machine learning technique will be used to predict the best location for their housing investment. We will investigate how the public facility in every district in South Korea was developing in the last 10 years. Finally, we will recommend the best district for the housing investment that may keep growing.

Business Problem

The objective of this capstone project is to analyze and select the best housing investment district in South Korea. Using data science methodology and machine learning techniques like regression, this project aims to provide solutions to answer the business question: If I want to buy the house right now, where is the best place so my investment can keep growing from time to time. The main assumption of this study is if the public facilities are growing from year to year, then that area can be recognized as a good place for the investment area. We will simplify our study only in the public infrastructure sector and dismiss the political, social, and even economic factors.

Target Audience of this project

This project is particularly useful to the younger generations in South Korea to decide the location of their future house. Like another advanced country, the property price in South Korea is one of the most expensive in the world. By looking at the economic condition of each person, one family can only afford to buy no more one house. Therefore, the decision about the house location is really important and hopefully, this study can become help them to analyze the value of the property based on the surrounding infrastructure only.

Data:

- List of districts in South Korea (in Korea, it is called -gu) the district. In this study, we will include every district in every city in South Korea
- Latitude and longitude coordinates of those districts. This is required to plot the map and also to get the venue data.
- Infrastructure data for each district, particularly data related to the subway station, bus stop, bus terminal, hospital, restaurant and school. We believe the existence of those facilities may increase the value of the property. We will compare this data with the area, population and the population density of each district. Currently the latest data about the population of each district in south Korea was published in 2015. Later, we will use this data to perform the data prediction about which districts that have the best facility, so this study can help anyone who want to buy the property in Korea.

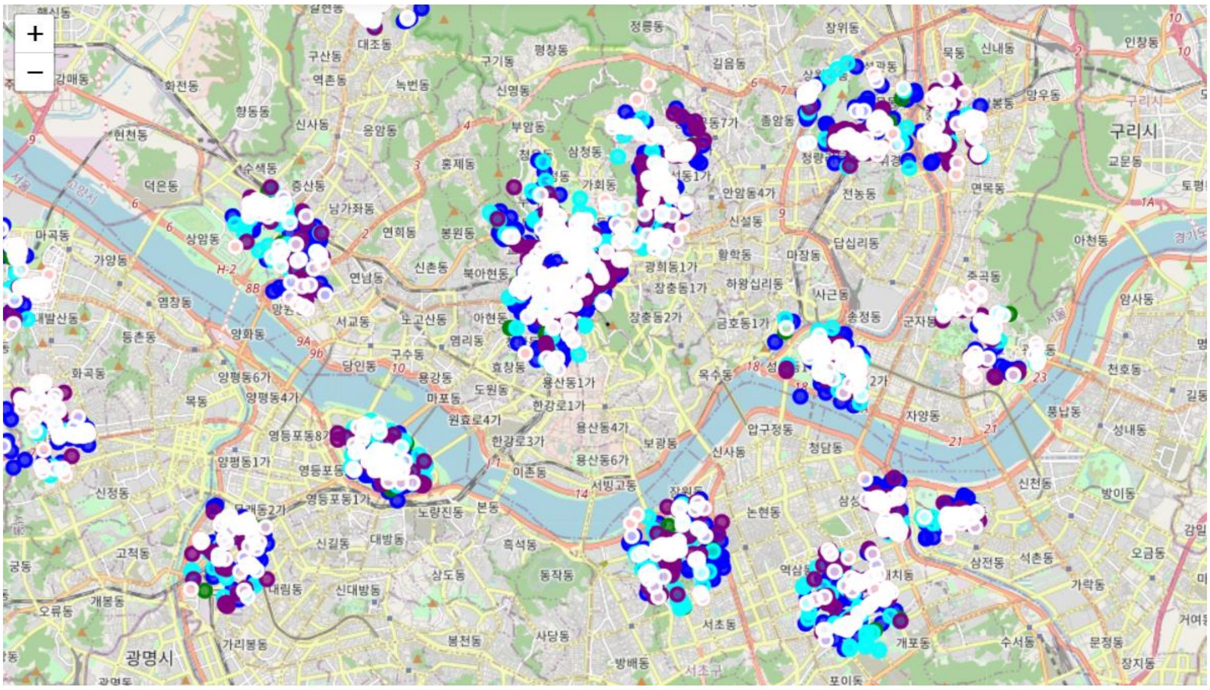
Methodology:

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_districts_in_South_Korea) contains a list of districts in South Korea, with a total of 131 districts from 22 largest cities in Korea. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the districts using Python Geocoder package which will give us the latitude and longitude coordinates of the districts.

After that, we will use Foursquare API to get the venue data for those districts. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Public Infrastructure to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, machine learning (KNN), and map visualization (Folium).

Result:

Using the foursquare and folium, we can get the location of the public infrastructure (in this case, we only show the Seoul Map:



It is quite complicated to see which facility that each district has the most, therefore, we use the KNN mean analysis, we can get the infrastructure feature of each district. The result can be seen in this table:

	District	Population	Area	Latitude	Longitude	cluster
0	단원구 Ansan	335849	91.23	37.32705	126.7888	0
1	상록구 Ansan	380574	57.83	37.30973	126.8536	0
2	동안구 Anyang	353381	21.92	37.40548	126.947	0
3	만안구 Anyang	265462	36.54	37.41366	126.9284	0
4	북구 Busan	309602	39.44	35.27438	129.0196	0
5	부산진구 Busan	394931	29.69	35.18046	129.0756	0
6	동구 Busan	101251	9.78	35.11573	129.0399	0
7	동래구 Busan	282732	16.63	35.19394	129.0978	0
8	강서구 Busan	62963	180.24	35.11789	128.8467	2
9	금정구 Busan	255979	65.17	35.27896	129.1064	0
10	해운대구 Busan	425872	51.46	35.16375	129.1587	0
11	중구 Busan	49011	2.82	35.10723	129.0358	1
12	남구 Busan	296955	26.77	35.10507	129.0641	0
13	사하구 Busan	357060	40.96	35.08565	128.9783	0
14	사상구 Busan	256347	36.06	35.15985	128.9719	0
15	서구 Busan	124896	13.88	35.11653	129.0145	0
16	수영구 Busan	177575	10.2	35.16541	129.1147	0

17	영도구 Busan	144852	14.13	35.09362	129.0367	0
18	연제구 Busan	214056	12.08	35.18332	129.0976	0
19	진해구 Changwon	179015	120.14	35.15344	128.6601	0
20	마산합포구 Changwon	186757	240.23	35.19434	128.5727	0
21	마산회원구 Changwon	223956	90.58	35.22267	128.5823	0
22	성산구 Changwon	250103	82.09	35.234	128.6647	0
23	의창구 Changwon	250702	211.22	35.25772	128.606	0
24	흥덕구 Cheongju	256681	198.27	36.64711	127.3924	0
25	상당구 Cheongju	179867	404.44	36.61238	127.5058	0
26	상당구 Cheongju	179867	404.44	36.61238	127.5058	0
27	청원구 Cheongju	162422	214.99	36.71574	127.4967	0
28	서원구 Cheongju	228659	114.88	36.61601	127.4844	0
29	동남구 Cheonan	250906	438.52	36.80942	127.1467	0
30	서북구 Cheonan	315577	197.7	36.86179	127.1375	0
31	중구 Daegu	77095	7.06	26.82076	106.795	0
32	동구 Daegu	341616	182.22	35.89665	128.6568	0
33	서구 Daegu	223681	17.48	26.82076	106.795	0
34	남구 Daegu	169765	17.44	35.84404	128.568	0
35	북구 Daegu	450852	94.09	35.87601	128.5961	0
36	수성구 Daegu	461473	76.46	35.82975	128.6901	0
37	달서구 Daegu	606178	62.34	35.84749	128.5299	0
38	대덕구 Daejeon	207312	68.45	36.3682	127.425	0
39	동구 Daejeon	248344	136.61	36.33205	127.4338	0
40	중구 Daejeon	77095	7.06	36.3255	127.4212	3
41	서구 Daejeon	223681	17.48	36.3551	127.3838	0
42	유성구 Daejeon	288618	177.27	36.37657	127.397	0
43	덕양구 Goyang	393479	165.51	37.6369	126.8322	0
44	일산동구 Goyang	275159	59.13	37.69015	126.7804	0
45	일산서구 Goyang	289745	42.77	37.67648	126.7432	0
46	북구 Gwangju	469045	121.74	35.16621	126.9098	0
47	동구 Gwangju	101582	48.86	35.16621	126.9098	0
48	광산구 Gwangju	370527	222.91	35.12431	126.7651	0
49	남구 Gwangju	217934	61.02	35.14908	126.899	0

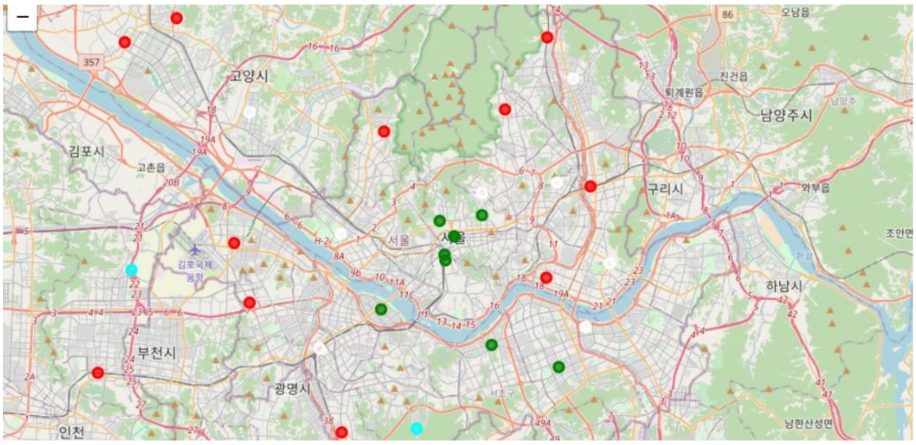
50	서구 Gwangju	302280	46.71	35.1337	126.875	0
51	부평구 Incheon	562110	31.99	37.48957	126.7238	0
52	동구 Incheon	79624	7.19	37.46611	126.643	5
53	계양구 Incheon	345671	45.58	37.5479	126.7477	0
54	중구 Incheon	93520	123.09	37.47589	126.617	4
55	남동구 Incheon	491038	56.99	37.43693	126.6934	0
56	미추홀구 Incheon	419683	24.85	37.43693	126.6934	0
57	서구 Incheon	420939	113.91	37.47589	126.617	0
58	연수구 Incheon	283840	42.74	37.39017	126.6455	0
59	덕진구 Jeonju	283813	110.79	35.85005	127.1624	0
60	완산구 Jeonju	361038	95.22	35.81219	127.1474	0
61	북구 Pohang	262581	393.33	36.08186	129.3369	0
62	남구 Pohang	253278	735.48	35.99765	129.3845	0
63	분당구 Seongnam	485767	69.35	37.3947	127.1205	0
64	중원구 Seongnam	256298	26.38	37.43068	127.1336	0
65	수정구 Seongnam	237986	45.99	37.44256	127.1264	0
66	도봉구 Seoul	366879	20.7	37.67921	127.0455	0
67	동대문구 Seoul	366633	14.2	37.59712	127.052	0
68	동작구 Seoul	402567	16.35	37.56668	126.9783	0
69	은평구 Seoul	491741	29.71	37.62637	126.9282	0
70	강북구 Seoul	345502	23.61	37.63862	127.0149	0
71	강동구 Seoul	496364	24.58	37.56668	126.9783	0
72	강남구 Seoul	570392	39.54	37.4928	127.0535	0
73	강서구 Seoul	571526	41.42	37.56297	126.8216	0
74	금천구 Seoul	243280	13.01	37.45602	126.8981	0
75	구로구 Seoul	422322	20.12	37.50345	126.8825	0
76	관악구 Seoul	529195	29.57	37.45813	126.9521	0
77	광진구 Seoul	373608	17.05	37.5517	127.0898	0
78	종로구 Seoul	169217	23.91	37.57906	126.9987	0
79	중구 Seoul	132224	9.96	37.55635	126.9715	0
80	중랑구 Seoul	425668	18.51	37.59495	127.0763	0
81	마포구 Seoul	392635	23.87	37.5683	126.8972	0
82	노원구 Seoul	605756	35.44	37.65636	127.0635	0
83	서초구 Seoul	432934	47	37.50549	127.0055	0
84	서대문구 Seoul	318467	17.6	37.57556	126.9686	0

85	성북구 Seoul	488036	24.57	37.59149	126.9984	0
86	성동구 Seoul	127748	16.85	37.54359	127.0447	0
87	송파구 Seoul	684028	33.88	37.51582	127.0727	0
88	양천구 Seoul	498819	17.4	37.52916	126.8326	0
89	영등포구 Seoul	403062	24.56	37.52539	126.9266	0
90	용산구 Seoul	247206	21.87	37.5531	126.9726	0
91	권선구 Suwon	307410	47.3	37.26165	127.0318	0
92	팔달구 Suwon	214653	13.08	37.26618	127.0002	0
93	영통구 Suwon	261008	27.46	37.25136	127.0713	0
94	북구 Ulsan	181611	157.35	35.5826	129.3604	0
95	동구 Ulsan	170639	36.01	35.50467	129.417	0
96	중구 Ulsan	232421	36.99	35.5691	129.333	0
97	남구 Ulsan	343487	72.55	35.49879	129.3459	0
98	처인구 Yongin	209893	467.57	37.23567	127.1925	0
99	기흥구 Yongin	365632	81.68	37.25214	127.1681	0
100	수지구 Yongin	314757	42.1	37.3348	127.1002	0

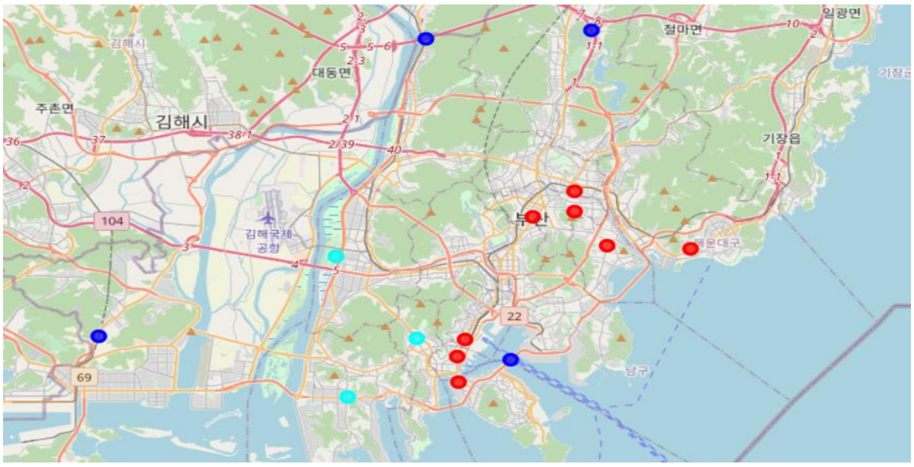
Note the number is represent the public infrastructure categories in this table

Cluster NO.	**Category**
0	Restaurant
1	Apartment
2	Subway Station
3	Bus Stop
4	Hospital
5	School

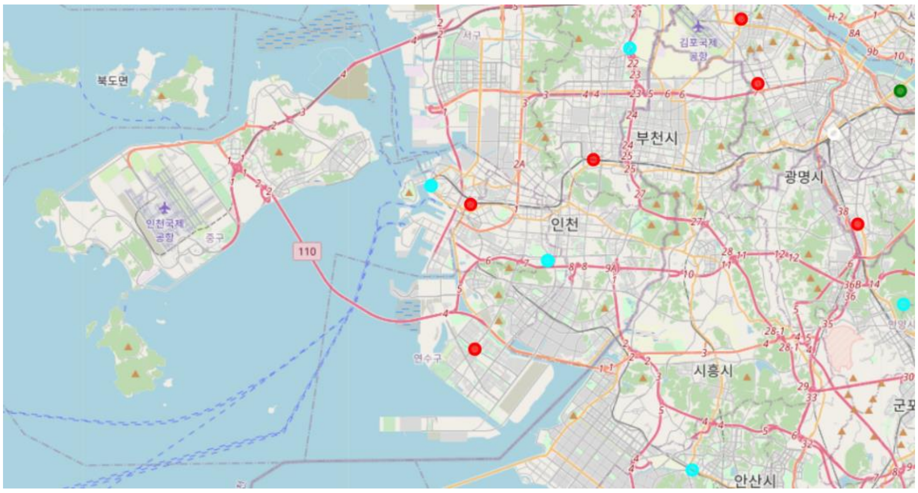
With the simplified data from KNN mean analysis, we can visualize the data to become more simplify, in this report we will show the analysis results in three biggest city in Korea, (Seoul, Busan, Incheon)



Result analysis in Seoul



Result analysis in Busan



Result analysis in Incheon

Discussion:

Based on the K means analysis we can assume that:

1. Most of the District has the number of the Restaurant that quite have a good proportion with the number of the population, therefore almost every area in the big city of Korea is recommended.
2. The number of the subway station in central Seoul is higher than others, therefore if the buyers must to commute by train, that area is the best place to stay
3. Overall, the number of the bus in Incheon is higher than others, therefore if the buyers must to commute by bus, that area is the best place to stay

Our recommendation for the future study:

1. The K means can simplify the result of the analysis
2. This result is quite limited because we only analyze the number of the infrastructure compare with the area and population of each district, therefore we need to add more data to make it more realistic
3. The data in the foursquare is quite limited compared to other provider such as google maps. Therefore it is highly recommended to compare the data from foursquare with another API location provider