“Capstone Project - The Battle of Neighborhoods”

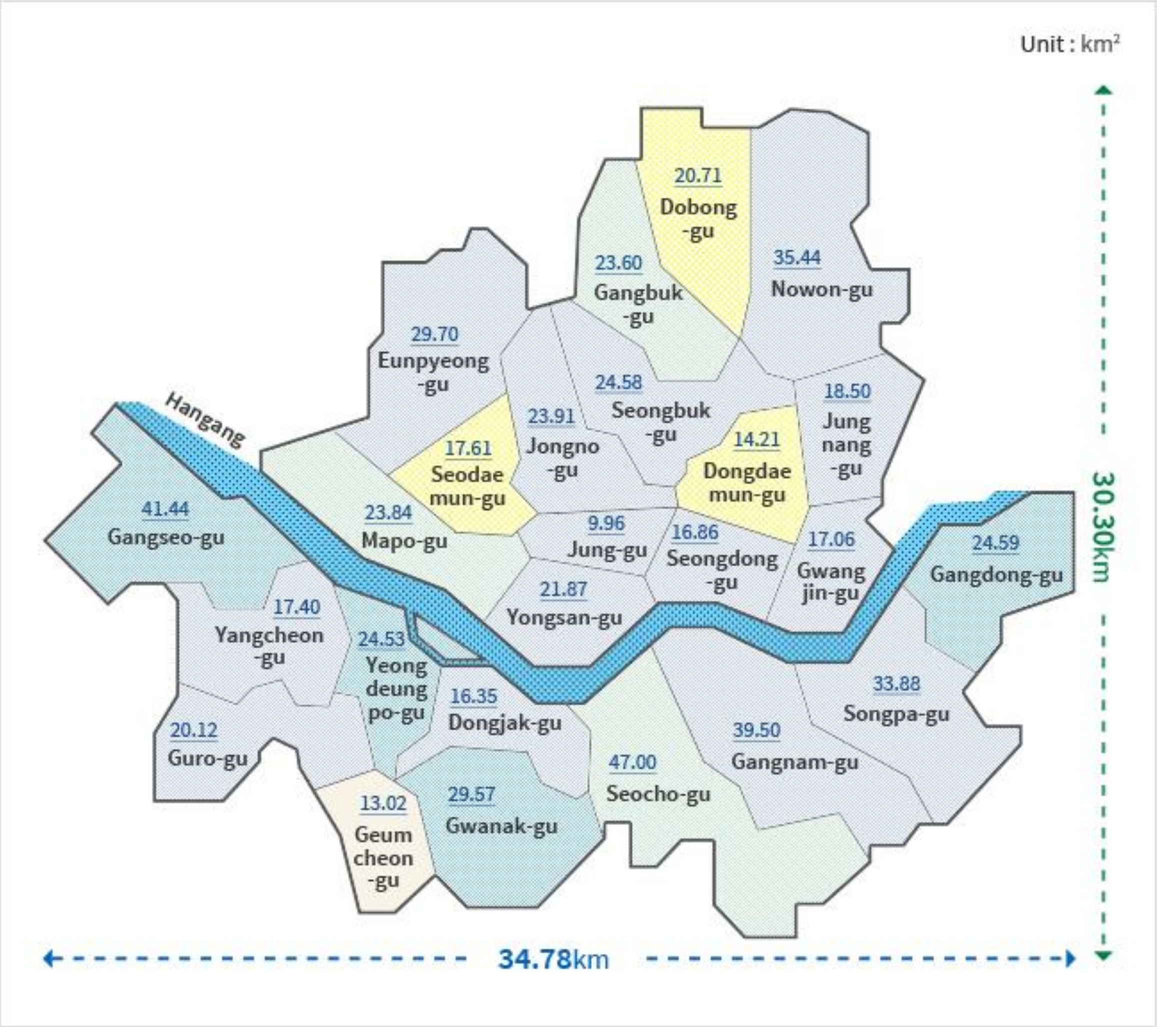
- Seoul, Korea -

Annie

Mar 31, 2019

Introduction/Business Problem Section

The Korean Peninsula lies in the center of Northeast Asia. The coordinates of its capital, Seoul, are 37.34° N and 126.59° E, putting it in close proximity to the Yellow Sea. Seoul is within a three-hour flight from 43 cities with populations of over one million people. Korea's location between China and Japan has been a great geographical advantage for the nation. There are 25 autonomous districts and 423 administrative "dong" units in Seoul. The city covers 0.28% of the entire peninsula (or 0.61% of South Korea), and spans an area 30.30 km north-to-south and 34.78 km west-to-east. If someone moves to Seoul or needs to stay for a certain period of time for business or sightseeing reasons, he or she should decide on the area that meets their requirements. This analysis will help people who want to move to Seoul by classifying the characteristics of each district using machine learning algorithms.



Data Section

This analysis made use of the following data sources:

1. Statistical information by category of each district in Seoul

Data were retrieved from Seoul Open Dataset from https://data.seoul.go.kr website. Various statistical information about welfare, education, traffic, safety, population, etc. can be obtained by district in Seoul. The examples are as follows :

- Number of Public Schools (Elementary/Middle/High)
- Number of Private Academies
- Number of Hospitals
- Number of Crimes
- Traffic Safety Index

Supported Open Dataset types are csv or JSON or XML.

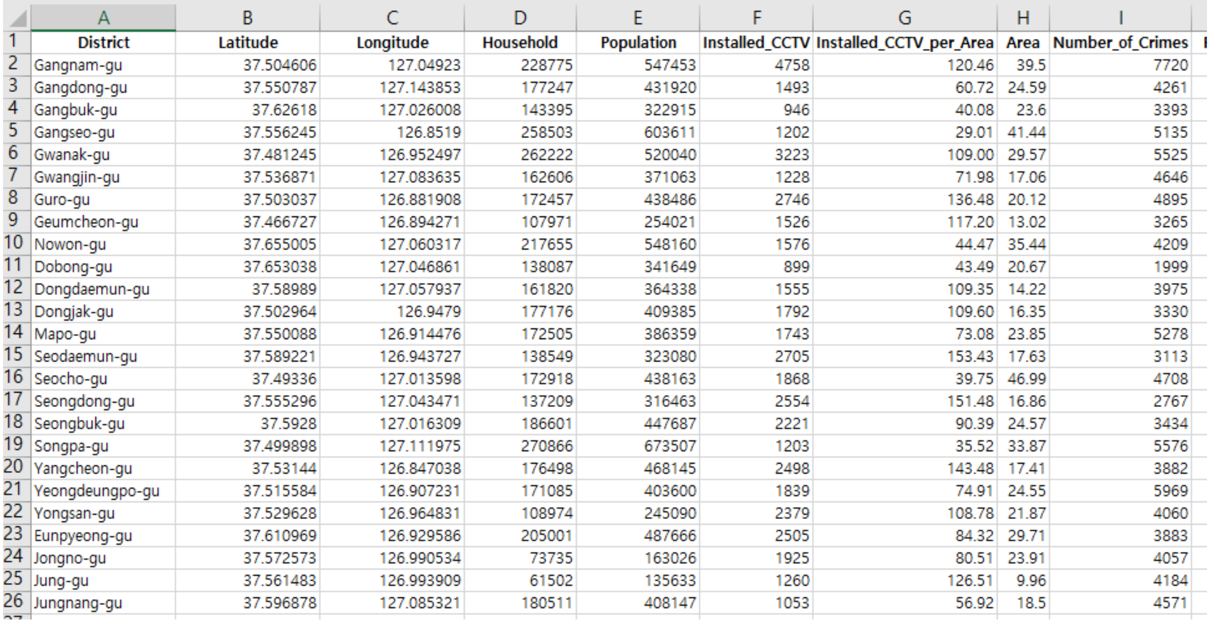
2. Top Venue Recommendations of each district in Seoul

Data were retrieved from FourSquare API (FourSquare website: www.foursquare.com) It includes Venue Name, Venue Category and Score per District.

Methodology section

1. Data Preparation

Seoul Open Dataset was provided in the format of csv file for each district. So the data should be merged, transformed and cleansed. I downloaded all required csv files from Seoul Open Dataset, and merged them into one file using Excel. Also I translated Korean words into English.



	A	B	C	D	E	F	G	H	I
1	District	Latitude	Longitude	Household	Population	Installed_CCTV	Installed_CCTV_per_Area	Area	Number_of_Crimes
2	Gangnam-gu	37.504606	127.04923	228775	547453	4758	120.46	39.5	7720
3	Gangdong-gu	37.550787	127.143853	177247	431920	1493	60.72	24.59	4261
4	Gangbuk-gu	37.62618	127.026008	143395	322915	946	40.08	23.6	3393
5	Gangseo-gu	37.556245	126.8519	258503	603611	1202	29.01	41.44	5135
6	Gwanak-gu	37.481245	126.952497	262222	520040	3223	109.00	29.57	5525
7	Gwangjin-gu	37.536871	127.083635	162606	371063	1228	71.98	17.06	4646
8	Guro-gu	37.503037	126.881908	172457	438486	2746	136.48	20.12	4895
9	Geumcheon-gu	37.466727	126.894271	107971	254021	1526	117.20	13.02	3265
10	Nowon-gu	37.655005	127.060317	217655	548160	1576	44.47	35.44	4209
11	Dobong-gu	37.653038	127.046861	138087	341649	899	43.49	20.67	1999
12	Dongdaemun-gu	37.58989	127.057937	161820	364338	1555	109.35	14.22	3975
13	Dongjak-gu	37.502964	126.9479	177176	409385	1792	109.60	16.35	3330
14	Mapo-gu	37.550088	126.914476	172505	386359	1743	73.08	23.85	5278
15	Seodaemun-gu	37.589221	126.943727	138549	323080	2705	153.43	17.63	3113
16	Seocho-gu	37.49336	127.013598	172918	438163	1868	39.75	46.99	4708
17	Seongdong-gu	37.555296	127.043471	137209	316463	2554	151.48	16.86	2767
18	Seongbuk-gu	37.5928	127.016309	186601	447687	2221	90.39	24.57	3434
19	Songpa-gu	37.499898	127.111975	270866	673507	1203	35.52	33.87	5576
20	Yangcheon-gu	37.53144	126.847038	176498	468145	2498	143.48	17.41	3882
21	Yeongdeungpo-gu	37.515584	126.907231	171085	403600	1839	74.91	24.55	5969
22	Yongsan-gu	37.529628	126.964831	108974	245090	2379	108.78	21.87	4060
23	Eunpyeong-gu	37.610969	126.929586	205001	487666	2505	84.32	29.71	3883
24	Jongno-gu	37.572573	126.990534	73735	163026	1925	80.51	23.91	4057
25	Jung-gu	37.561483	126.993909	61502	135633	1260	126.51	9.96	4184
26	Jungnang-gu	37.596878	127.085321	180511	408147	1053	56.92	18.5	4571

Figure 1 : Integrated Data from Seoul Open Dataset (Excel File)

After uploading the data into jupyter notebook, I standardized data frame using scikit-learn standardscaler.

With location data of each district, I retrieved popular venue information using Foursquare API.

Looking at the venue category data, there was a lot of different duplicated information. I transformed the data into simple categories such as restaurant, theater, and landmark.

2. Data Exploration

Created new data frame and displayed the top 10 most common venues for each district.

Explored public statistical information for each district.

Which districts have the most hospitals ?

Which districts have the most education facilities ?

Which districts have the most crime ratio ?

Which districts have the most population ?

3. Clustering Districts

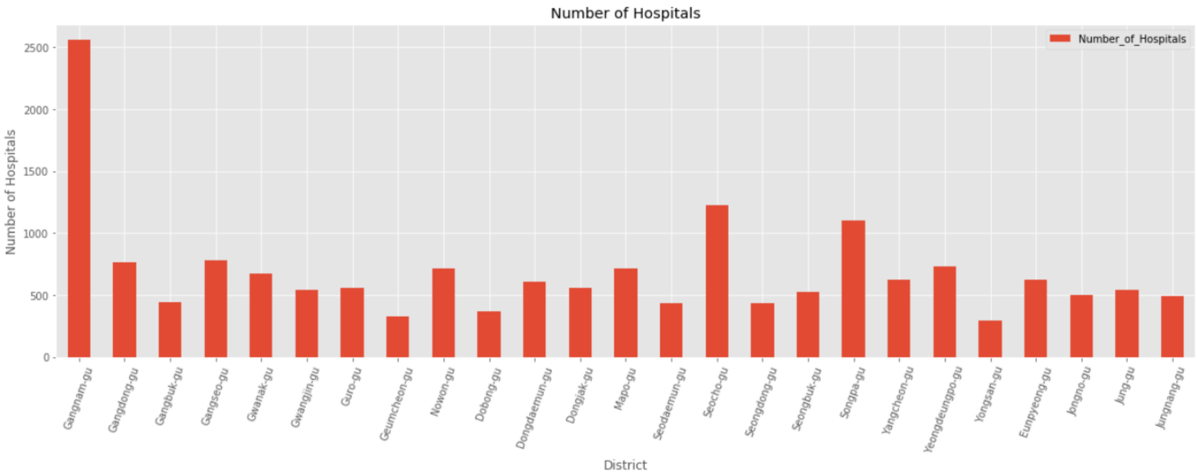
I ran k-means to cluster the districts into 5 clusters.

I used Folium and Choropleth map to visualize them.

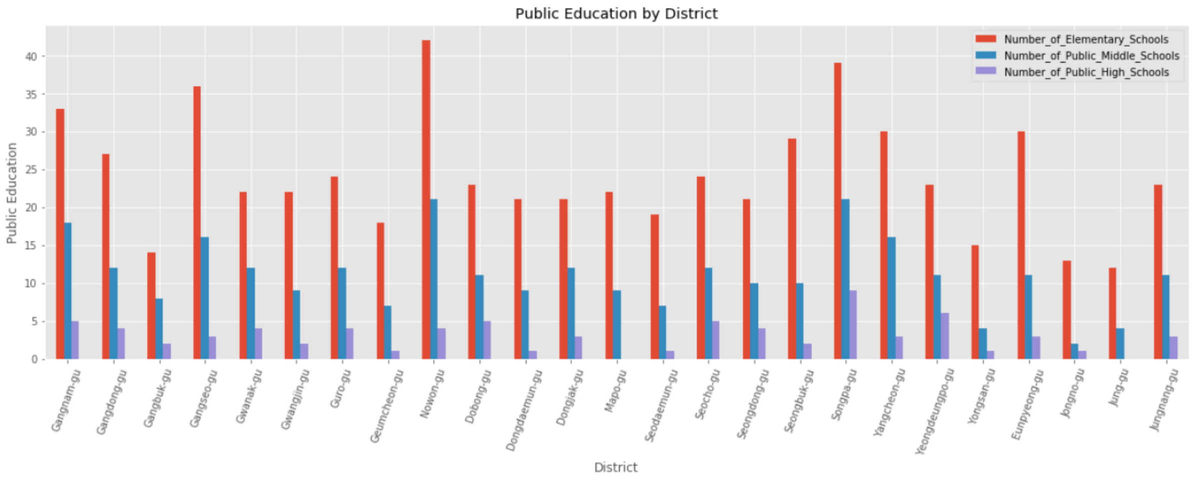
Lastly I examined each cluster and determined the discriminating characteristics that distinguish each cluster.

Results section

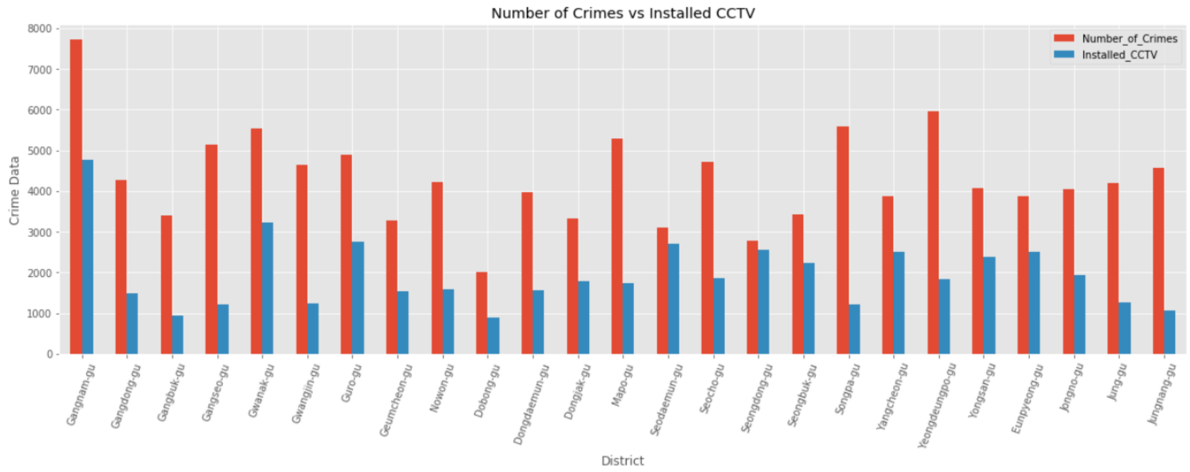
District having the most hospitals : Gangnam-gu



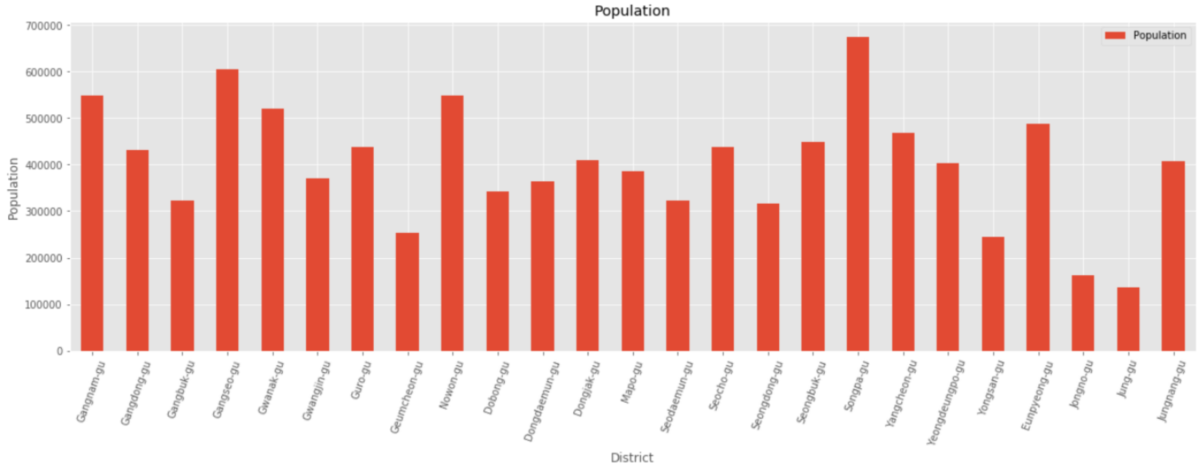
District having the most public education : Nowon-gu



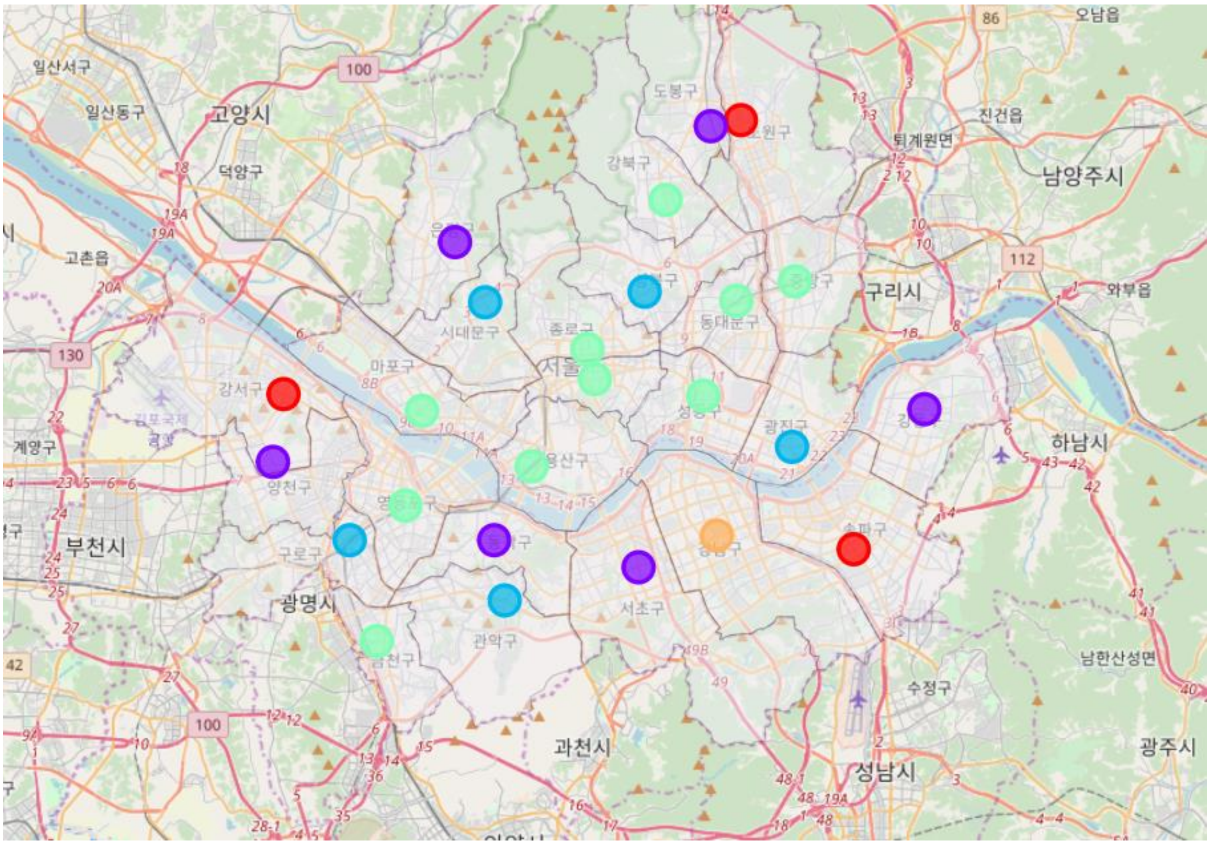
District having the most crimes and CCTV : Gangnam-gu



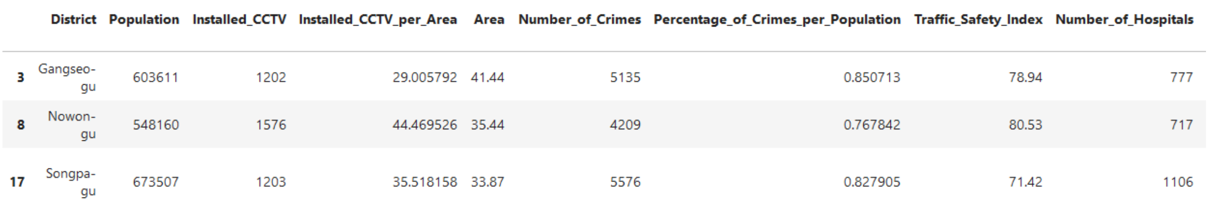
District having the most population : Songpa-gu



K-means clustering result is as follows.



Cluster Label : 0

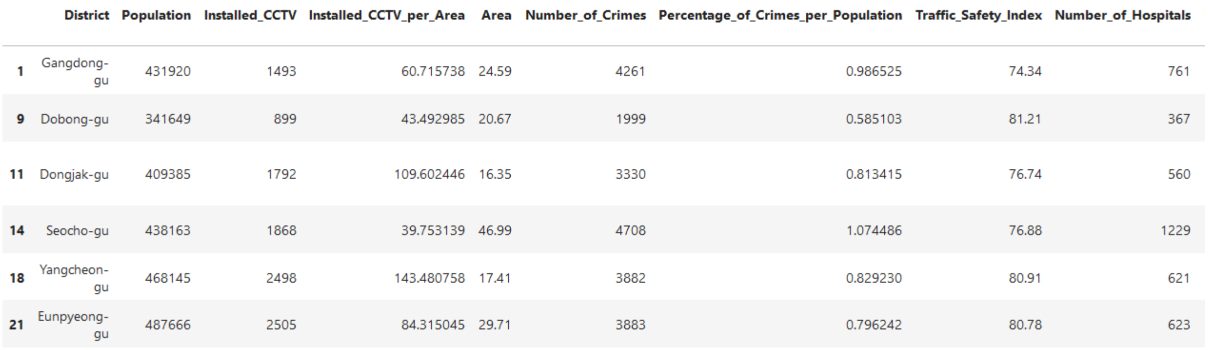


	District	Population	Installed_CCTV	Installed_CCTV_per_Area	Area	Number_of_Crimes	Percentage_of_Crimes_per_Population	Traffic_Safety_Index	Number_of_Hospitals
3	Gangseo-gu	603611	1202	29.005792	41.44	5135	0.850713	78.94	777
8	Nowon-gu	548160	1576	44.469526	35.44	4209	0.767842	80.53	717
17	Songpa-gu	673507	1203	35.518158	33.87	5576	0.827905	71.42	1106

- Characteristics : High Population, High Crimes but low CCTV, High Houses, Medium Private Institutes

Cluster Label : 1

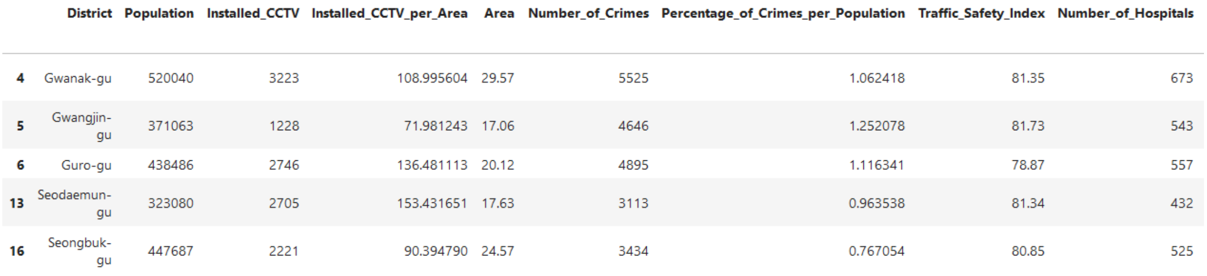
- Characteristics : Medium Population, Medium Crimes, Medium Houses



	District	Population	Installed_CCTV	Installed_CCTV_per_Area	Area	Number_of_Crimes	Percentage_of_Crimes_per_Population	Traffic_Safety_Index	Number_of_Hospitals
1	Gangdong-gu	431920	1493	60.715738	24.59	4261	0.986525	74.34	761
9	Dobong-gu	341649	899	43.492985	20.67	1999	0.585103	81.21	367
11	Dongjak-gu	409385	1792	109.602446	16.35	3330	0.813415	76.74	560
14	Seocho-gu	438163	1868	39.753139	46.99	4708	1.074486	76.88	1229
18	Yangcheon-gu	468145	2498	143.480758	17.41	3882	0.829230	80.91	621
21	Eunpyeong-gu	487666	2505	84.315045	29.71	3883	0.796242	80.78	623

Cluster Label : 2

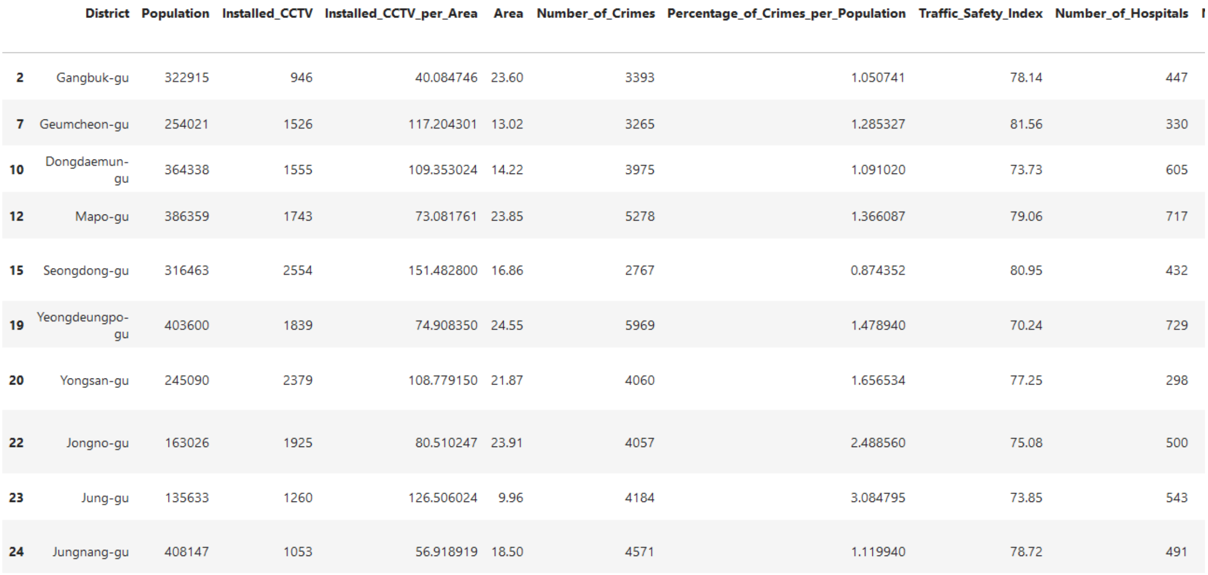
- Characteristics : Medium Population, Medium Crimes, Medium Houses, Low Private Institutes



	District	Population	Installed_CCTV	Installed_CCTV_per_Area	Area	Number_of_Crimes	Percentage_of_Crimes_per_Population	Traffic_Safety_Index	Number_of_Hospitals
4	Gwanak-gu	520040	3223	108.995604	29.57	5525	1.062418	81.35	673
5	Gwangjin-gu	371063	1228	71.981243	17.06	4646	1.252078	81.73	543
6	Guro-gu	438486	2746	136.481113	20.12	4895	1.116341	78.87	557
13	Seodaemun-gu	323080	2705	153.431651	17.63	3113	0.963538	81.34	432
16	Seongbuk-gu	447687	2221	90.394790	24.57	3434	0.767054	80.85	525

Cluster Label : 3

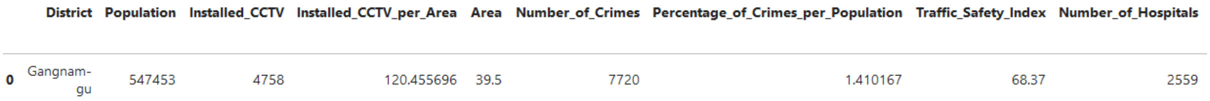
- Characteristics : Low Population, Medium Crimes, Low Houses, Less Students



	District	Population	Installed_CCTV	Installed_CCTV_per_Area	Area	Number_of_Crimes	Percentage_of_Crimes_per_Population	Traffic_Safety_Index	Number_of_Hospitals
2	Gangbuk-gu	322915	946	40.084746	23.60	3393	1.050741	78.14	447
7	Geumcheon-gu	254021	1526	117.204301	13.02	3265	1.285327	81.56	330
10	Dongdaemun-gu	364338	1555	109.353024	14.22	3975	1.091020	73.73	605
12	Mapo-gu	386359	1743	73.081761	23.85	5278	1.366087	79.06	717
15	Seongdong-gu	316463	2554	151.482800	16.86	2767	0.874352	80.95	432
19	Yeongdeungpo-gu	403600	1839	74.908350	24.55	5969	1.478940	70.24	729
20	Yongsan-gu	245090	2379	108.779150	21.87	4060	1.656534	77.25	298
22	Jongno-gu	163026	1925	80.510247	23.91	4057	2.488560	75.08	500
23	Jung-gu	135633	1260	126.506024	9.96	4184	3.084795	73.85	543
24	Jungnang-gu	408147	1053	56.918919	18.50	4571	1.119940	78.72	491

Cluster Label : 4

- Characteristics : High Population, High Crimes, High Houses, High Hospitals, High Students



	District	Population	Installed_CCTV	Installed_CCTV_per_Area	Area	Number_of_Crimes	Percentage_of_Crimes_per_Population	Traffic_Safety_Index	Number_of_Hospitals
0	Gangnam-gu	547453	4758	120.455696	39.5	7720	1.410167	68.37	2559

Discussion section

Conclusion section

Through this analysis, I classified the characteristics of each district using machine learning algorithms. This report will help those who are planning to move to Seoul to decide which district to go.