Renting a House in Seoul

Pham Ngoc Son

2021-06-01

1. Introduction

1.1 Background

Renting a house in any place is never an easy task. From renting price to service and utility, everybody want the best apartment with the least pay. Currently, the apartment price in Seoul is readily available in many website such as Naver land or KB land. However, price is only one part of the problem. Customer also cares about many other facility around the place such as transport, park, restaurant, supermarket, cinema to name a few. The problem is this data is not available on any website so customer will have to take a lot of time to investigate each area. This is a quite frustrating task and often they will give up and choose any place with reasonable price.

1.2 Problem

Location data from each apartment can be used together with Foursquare database to get venues around it within a predetermined radius. In this project the venue data from Foursquare will be used to classify each apartment and give user more useful information on their choice of living place.

1.3 Interest

This analysis is not only good for the house finder in Seoul but also for the housing data provider such as Naver & KB. The government also can use this analysis to improve living condition in Seoul.

2. Data acquisition and cleaning

2.1 Data sources

Data for all apartments, buildings and houses in Seoul can be found in the below link:

https://www.juso.go.kr/addrlink/addressBuildDevNew.do?menu=match

The size of this data is very large so take care before downloading it.

The data included all building name, code, detail address included city, district, ward, road number … The X, Y location of each apartment and building is also provided but it is in Korean standard.

2.2 Data cleaning

Due to the size of the data, after downloaded, ~ 1000 rows were randomly extracted to be used

for this analysis.

The X, Y coordinates in Korean standard must be transformed into longitude and latitude to be useful in the analysis. This proves to be a quite difficult task since the transformation from GRS80 UTM-K to latitude and longitude is not readily available. Fortunately, a free software provided by Korean government named NGI pro can be used to do this transformation.

There are some missing values in the building name feature and all of them were removed since this is an important feature.

The feature names were also translated from Korean to English for non-Korean reader to be able to understand the data.

The price was added randomly to each apartment based on the actual renting price in Seoul. It will be better if exact data was used but because the price info was not free so for the scope of this report, we will only use assumed value.

2.3 Feature selection

After data cleaning, some data rows were removed so that exact 1000 rows remained in the database. The data consisted of many not-useful features such as city code, building code, road code …

After carefully checking, the selected features are as below:

City – District – Ward – Road – BuildingName – PostalCode – RentPrice – Longitude – Latitude.

	City	District	Ward	Road	BuildingName	PostalCode	RentPrice	Longitude	Latitude
673	서울특별시	강서구	등촌동	공항대로58나길	힐하우스	7666	970	126.862080	37.549676
378	서울특별시	중랑구	상봉동	신내로7다길	상봉빌라	2079	2295	127.091766	37.603401
118	서울특별시	중구	신당동	동호로8나길	신당힐빌리지	4594	840	127.014236	37.552235
747	서울특별시	강서구	공항동	공항대로	우성빌딩	7623	1335	126.812512	37.558736
8	서울특별시	종로구	누상동	필운대로5가길	창원예가	3037	2865	126.966781	37.580645

The Longitude and Latitude of each building will be used to get venue information using Foursquare API. This venue data will then be used to classify the building into different categories and allow customer to easily select building of their choice.

A new feature, named ‘Neighborhood’ was created as the combination of Building name and postal code to avoid the duplication in ‘Building Name’ feature.

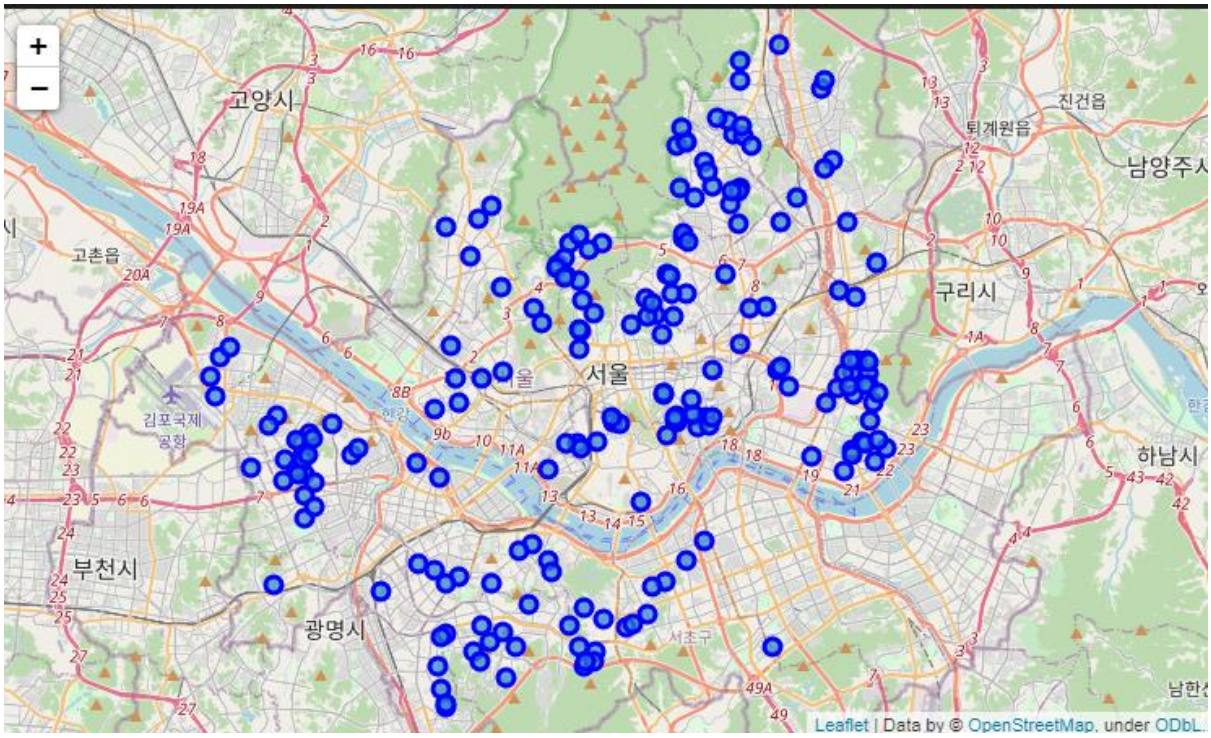
3. Exploratory Data Analysis

3.1 Location of all selected apartment in Seoul

The location of all building and apartment using in this analysis was plotted using Folium library. It

is important to check if the data was properly selected.

The resulting plot was as below:



The distribution of each point was quite good. They are evenly distributed around Seoul, not just clustered in one or two area.

3.2 Getting venue data from Foursquare API

It took too much time to get all data for all 1000 buildings and apartments from Foursquare so the data was further scaled down to 200 rows. Even at this number, it often took 5 to 10 minutes to get the feedback from Foursquare API.

The venue data from Foursquare was acquired using Latitude, Longitude from each building and the result was as below:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	힐하우스_7666	37.549676	126.86208	고양이똥	37.551984	126.864626	Coffee Shop
1	힐하우스_7666	37.549676	126.86208	목동깨비시장	37.548337	126.864731	Market
2	힐하우스_7666	37.549676	126.86208	푸주옥	37.553641	126.853746	Korean Restaurant
3	힐하우스_7666	37.549676	126.86208	한우등촌골	37.547652	126.862748	Korean Restaurant
4	힐하우스_7666	37.549676	126.86208	등촌칼국수 버섯매운탕	37.556144	126.856798	Korean Restaurant

There are a total of 9070 venues data row.

There are a total of 250 unique categories in the 'Venue Category" column.

3.3 Grouping venue into 7 general categories

With over 250 unique categories, the classification may not get high accuracy result and it would be difficult to know the different between each category. Therefore, it is important to group these categories into some general ones to improve the classification effectiveness.

The classification file can be found in the below link:

https://github.com/sonpn82/Coursera_Capstone/blob/cb43ca3b879c0b4deaff8cceafd35df78ca1600d/Seoul_venues_cat.csv

The 1st general category is 'Food and drink' which covers the following venue categories:

- All restaurants
- All bars
- All food place, food course
- All café and tea shop
- All bestiary, pastry shop

The 2nd general category is 'Sports and leisure' which covers the following venue categories:

- All sport centers
- All sport stadium
- All kind of gym, fitness center
- All kind of swimming pools
- Golf course, bowling, basketball, badminton course …
- Water surfing, rock climbing
- Bike trail

The 3rd general category is 'Transport' which covers the following venue categories:

- Airport
- Bus station

- Train station
- Metro station
- Road, tunnel, bridge

The 4th general category is 'Sightseeing and culture' which covers the following venue categories:

- Art gallery
- Museum, Exhibition
- Cultural center
- Flea market
- Historical site
- Mountain, forest, river, park

The 5th general categories is 'Service and store' which covers the following venue categories:

- All kind of shop such as auto workshop, flower shop, mobile phone shop …
- All kind of store such as office supplies store, pet store, electronic store …
- Shopping mall, shopping plaza, outlet mall
- Pharmacy, clinic …

The 6th general categories is 'Entertainment' which covers the following venue categories:

- Arcade, video game store, toy store …
- Bath house, spa, sauna
- Concert hall, theater, rock club, movie theater
- Night club

The final general categories is 'Lodging' which covers the following categories:

- Hotel
- Hostel
- Boarding house
- Resort …

Below is sample of general categories in data frame

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	General Category
0	힐하우스_7666	37.549676	126.86208	고양이똥	37.551984	126.864626	Coffee Shop	Food and drink
1	힐하우스_7666	37.549676	126.86208	목동깨비시장	37.548337	126.864731	Market	Sightseeing and culture
2	힐하우스_7666	37.549676	126.86208	푸주옥	37.553641	126.853746	Korean Restaurant	Food and drink
3	힐하우스_7666	37.549676	126.86208	한우등촌골	37.547652	126.862748	Korean Restaurant	Food and drink
4	힐하우스_7666	37.549676	126.86208	등촌칼국수 버섯매운탕	37.556144	126.856798	Korean Restaurant	Food and drink

3.4 Sort and group the data based on the most common general venue

The general category was broken into each separated column and the data was represented by a value of 1 (has the category) or 0 (not has the category) for each neighborhood. After that, the data was grouped and averaged and the result was as below:

	Neighborhood	Entertainment	Food and drink	Lodging	Service and store	Sightseeing and culture	Sports and leisure	Transport
0	H_3006	0.000000	0.815789	0.000000	0.000000	0.078947	0.078947	0.026316
1	가온누리_7723	0.000000	0.733333	0.000000	0.000000	0.266667	0.000000	0.000000
2	가온빌_7638	0.041667	0.833333	0.000000	0.000000	0.083333	0.041667	0.000000
3	강변아파트_5118	0.035714	0.857143	0.000000	0.035714	0.035714	0.017857	0.017857
4	경신빌라_7679	0.066667	0.666667	0.000000	0.000000	0.266667	0.000000	0.000000
5	골든빌_2716	0.000000	0.850000	0.000000	0.000000	0.100000	0.000000	0.050000

Finally, the result was sorted by the most common categories of each neighborhood, from 1st to 7th. The sorted result was as below:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	H_3006	Food and drink	Sightseeing and culture	Sports and leisure	Transport	Entertainment	Lodging	Service and store
1	가온누리_7723	Food and drink	Sightseeing and culture	Entertainment	Lodging	Service and store	Sports and leisure	Transport
2	가온빌_7638	Food and drink	Sightseeing and culture	Entertainment	Sports and leisure	Lodging	Service and store	Transport
3	강변아파트_5118	Food and drink	Entertainment	Service and store	Sightseeing and culture	Sports and leisure	Transport	Lodging
4	경신빌라_7679	Food and drink	Sightseeing and culture	Entertainment	Lodging	Service and store	Sports and leisure	Transport

Now the data is readily to be analyzed using machine learning algorithm.

4. Non-supervised classification modeling

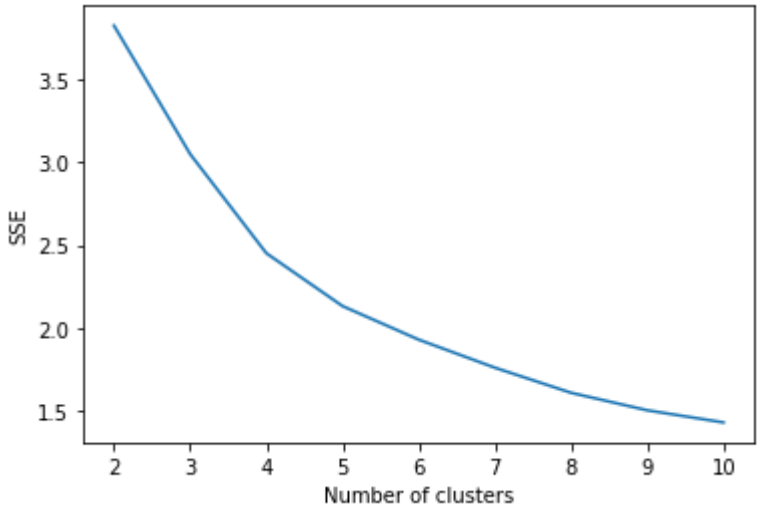
4.1 Model selection

The non-labeled data was put into classification using K-Means clustering algorithm. The sklearn library was used for this kind of analysis.

One weak point of this kind of model is the user has to manually select the number of cluster. With a multi-dimensions data, it is not simple to suggest a good number of cluster to be used. Therefore, an analysis to check relation between number of cluster and model accuracy was carried out to fulfill this task.

4.2 K number selection

The number of cluster vs model Accuracy was plotted as below, with k from 2 to 10:



The accuracy here is the 'inertial' or the within-cluster sum-of-squares criterion.

It is easy to notice the change of accuracy slope at k = 5, 6. In this study, k= 6 was selected to balance between accuracy and ease of classification work.

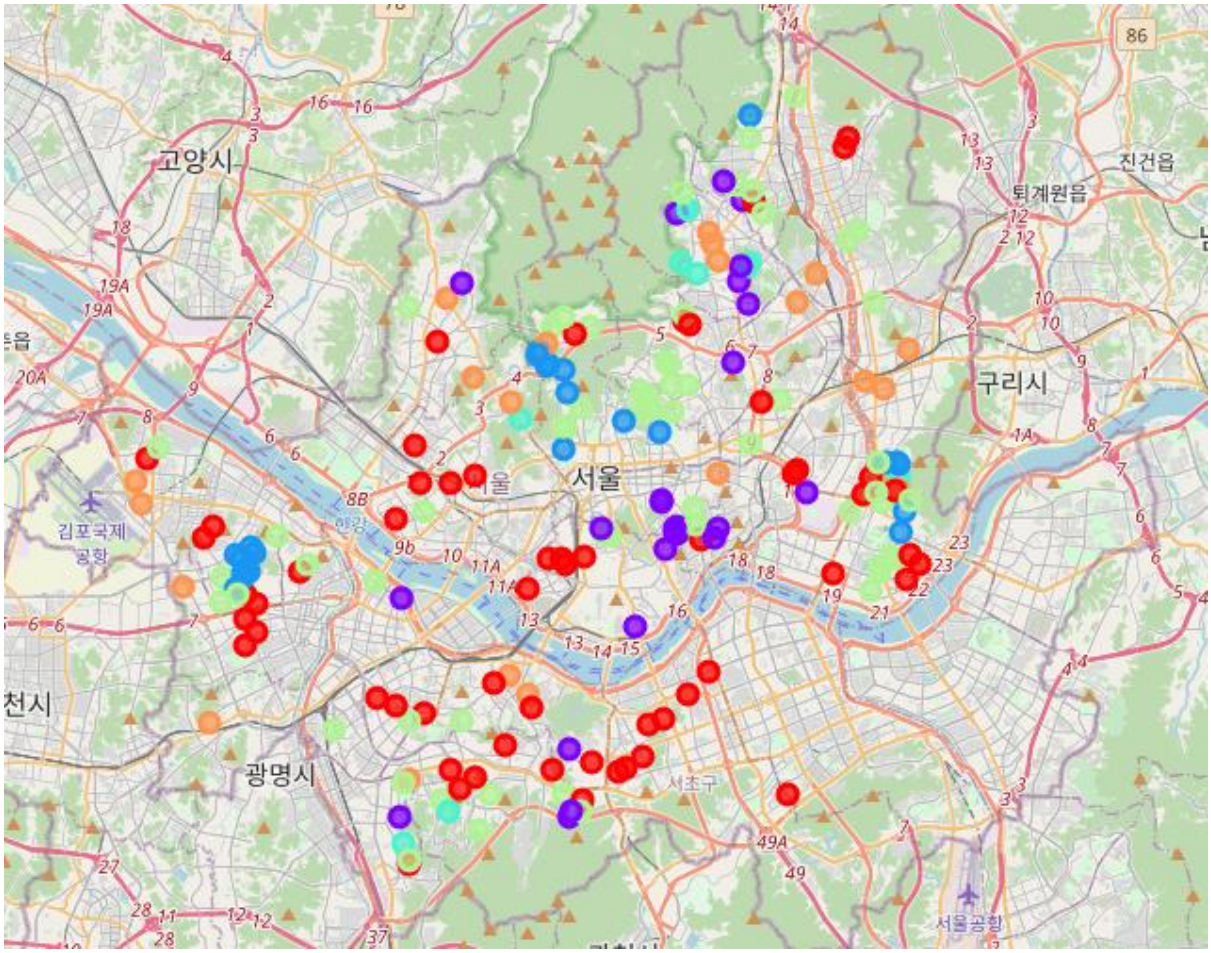
4.3 Add label to data frame

After finishing the model fitting, the resulted labels (0 to 5) were added to the data frame. The result was as below:

	City	District	Ward	Road	BuildingName	PostalCode	RentPrice	Longitude	Latitude	Neighborhood	Cluster Labels
673	서울특별시	강서구	등촌동	공항대로58나길	힐하우스	7666	970	126.862080	37.549676	힐하우스_7666	4
378	서울특별시	중랑구	상봉동	신내로7다길	상봉빌라	2079	2295	127.091766	37.603401	상봉빌라_2079	5
118	서울특별시	중구	신당동	동호로8나길	신당힐빌리지	4594	840	127.014236	37.552235	신당힐빌리지_4594	4
747	서울특별시	강서구	공항동	공항대로	우성빌딩	7623	1335	126.812512	37.558736	우성빌딩_7623	5
8	서울특별시	종로구	누상동	필운대로5가길	창원예가	3037	2865	126.966781	37.580645	창원예가_3037	4

4.4 Visualize the resulting cluster

The labeled data was plotted again using folium. The plot was as below:



Each color represent a cluster, with a total of 6 clusters.

4.5 Create a name for each label

It is important to note that food and drink category ranked number 1 in all 6 clusters. This shows that food and drink is very popular in Seoul and can be found anywhere. Therefore, we will not consider food and drink for naming each cluster

4.5.1 Cluster 1

The data for this cluster is as below:

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Food and drink	Transport	Entertainment	Lodging	Service and store	Sightseeing and culture	Sports and leisure
0	Food and drink	Sports and leisure	Sightseeing and culture	Entertainment	Service and store	Lodging	Transport
0	Food and drink	Service and store	Lodging	Sightseeing and culture	Sports and leisure	Entertainment	Transport
0	Food and drink	Sightseeing and culture	Entertainment	Lodging	Service and store	Sports and leisure	Transport
0	Food and drink	Sightseeing and culture	Sports and leisure	Entertainment	Lodging	Service and store	Transport
0	Food and drink	Service and store	Entertainment	Sightseeing and culture	Sports and leisure	Lodging	Transport

Looking at the data, the most common point between each neighborhood is a lot of service, shop and sight-seeing. We can name this cluster as 'Service hub and sight-seeing lover'

4.5.2 Cluster 2

The data for this cluster is as below:

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
1	Food and drink	Sightseeing and culture	Service and store	Sports and leisure	Entertainment	Lodging	Transport
1	Food and drink	Service and store	Entertainment	Sightseeing and culture	Sports and leisure	Lodging	Transport
1	Food and drink	Lodging	Sports and leisure	Transport	Service and store	Entertainment	Sightseeing and culture
1	Food and drink	Service and store	Entertainment	Lodging	Sightseeing and culture	Sports and leisure	Transport
1	Food and drink	Service and store	Sightseeing and culture	Sports and leisure	Entertainment	Transport	Lodging
1	Food and drink	Sports and leisure	Sightseeing and culture	Entertainment	Service and store	Lodging	Transport
1	Food and drink	Service and store	Sports and leisure	Sightseeing and culture	Transport	Entertainment	Lodging

It can be seen that sports dominance this group (total 18 sport rows at the most common venue among 42 data rows). We can call this group 'Sport enthusiasm, pro or amateur!'.

4.5.3 Cluster 3

The data for this group is as below:

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
2	Food and drink	Sightseeing and culture	Transport	Entertainment	Service and store	Lodging	Sports and leisure
2	Food and drink	Sightseeing and culture	Transport	Service and store	Sports and leisure	Entertainment	Lodging
2	Food and drink	Sightseeing and culture	Entertainment	Service and store	Lodging	Sports and leisure	Transport
2	Food and drink	Sightseeing and culture	Entertainment	Lodging	Service and store	Sports and leisure	Transport
2	Food and drink	Sightseeing and culture	Entertainment	Lodging	Service and store	Sports and leisure	Transport
2	Food and drink	Sightseeing and culture	Transport	Entertainment	Service and store	Sports and leisure	Lodging

The data is populated with a lot of sight-seeing and entertainment area. We can call this cluster 'Walk around & play!'

4.5.4 Cluster 4

The data for this group is as below:

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
3	Food and drink	Transport	Sightseeing and culture	Sports and leisure	Entertainment	Lodging	Service and store
3	Food and drink	Sports and leisure	Sightseeing and culture	Service and store	Entertainment	Lodging	Transport
3	Food and drink	Sports and leisure	Service and store	Transport	Entertainment	Lodging	Sightseeing and culture
3	Food and drink	Transport	Sightseeing and culture	Service and store	Sports and leisure	Entertainment	Lodging
3	Food and drink	Sightseeing and culture	Sports and leisure	Service and store	Entertainment	Lodging	Transport

The data is balanced between all categories. We have sight-seeing, transportation, sport center, service center all around. This cluster can be called 'All around'.

4.5.5 Cluster 5

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
4	Food and drink	Sightseeing and culture	Service and store	Transport	Entertainment	Lodging	Sports and leisure
4	Food and drink	Sports and leisure	Sightseeing and culture	Service and store	Entertainment	Lodging	Transport
4	Food and drink	Sightseeing and culture	Service and store	Sports and leisure	Entertainment	Lodging	Transport
4	Food and drink	Sightseeing and culture	Entertainment	Service and store	Lodging	Sports and leisure	Transport
4	Food and drink	Sightseeing and culture	Entertainment	Sports and leisure	Lodging	Service and store	Transport

It is easy to see many sports and sight-seeing here. So we can call this one 'Sport lovers and sight-seeing goers'

4.5.6 Group 6

The data is as below

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
5	Food and drink	Service and store	Transport	Entertainment	Sports and leisure	Lodging	Sightseeing and culture
5	Food and drink	Service and store	Transport	Entertainment	Sightseeing and culture	Sports and leisure	Lodging
5	Food and drink	Service and store	Sports and leisure	Transport	Lodging	Sightseeing and culture	Entertainment
5	Food and drink	Service and store	Transport	Entertainment	Sightseeing and culture	Lodging	Sports and leisure
5	Food and drink	Transport	Service and store	Sightseeing and culture	Entertainment	Lodging	Sports and leisure
5	Food and drink	Transport	Service and store	Sports and leisure	Entertainment	Lodging	Sightseeing and culture

The 'Transport' venue dominates this group. Besides, we can also find a lot of service and store close by. This group can be called 'Transportation hub and a convenience life'.

5. Results

5.1 Finding a house using clustering data

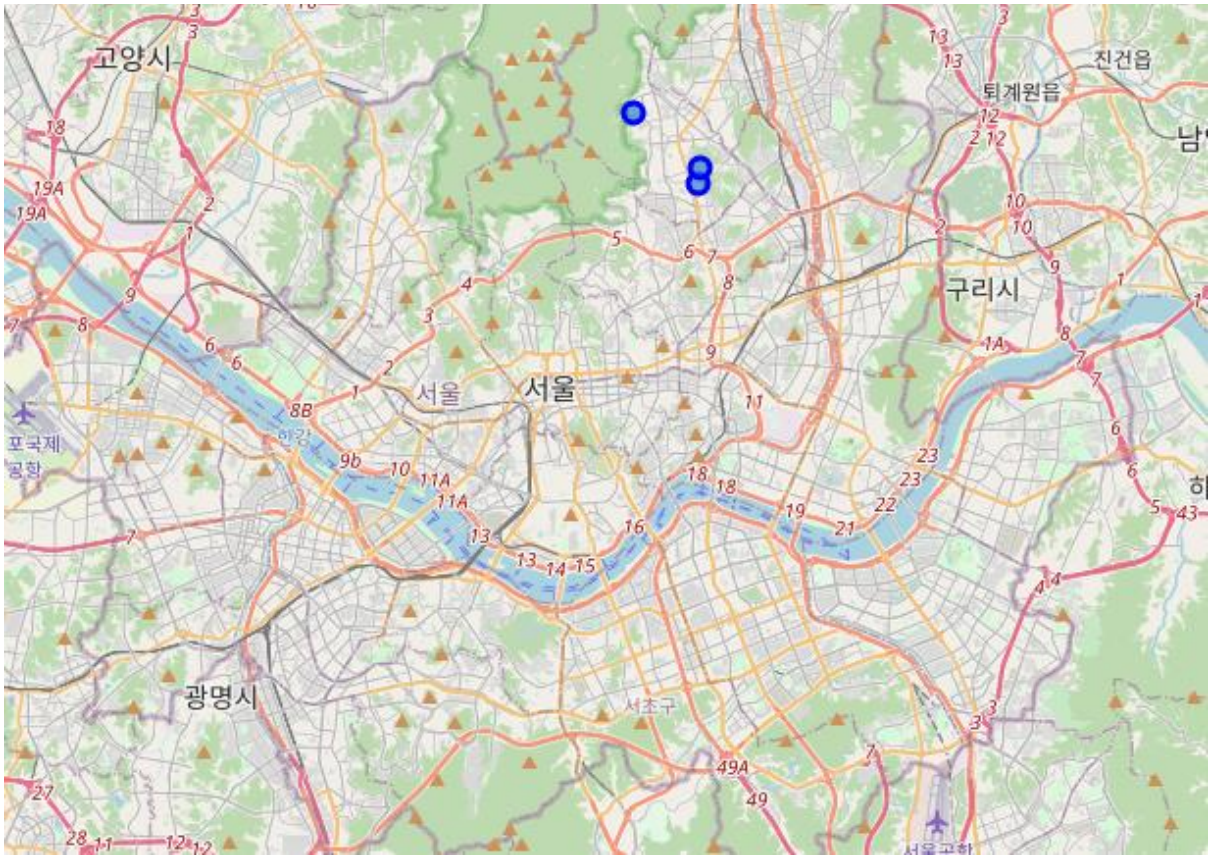
After finishing all of the modeling & model building, we can put our model to real use case. In this example, a customer wants to choose a house with below spec:

- A rent price less than 2000$ per month
- A location in '강북구' or 'Kangbukgu' district.
- Customer is a sport lover so he chose group 2.

Our program gives out the 3 choices as below:

	City	District	Ward	Road	BuildingName	PostalCode	RentPrice	Longitude	Latitude	Neighborhood	Cluster Labels
439	서울특별시	강북구	미아동	도봉로30길	타임아트빌	1162	1665	127.029778	37.622958	타임아트빌_1162	1
451	서울특별시	강북구	미아동	오패산로52바길	홍익빌라	1142	1855	127.030940	37.627415	홍익빌라_1142	1
488	서울특별시	강북구	수유동	인수봉로73나길	홍익빌라	1020	1325	127.007724	37.642437	홍익빌라_1020	1

Their location on the map is as below:



With the above data, customer can furtherly check the detail of each location to find their best place. It will not be a hard time for him to look for a renting place anymore.

5.2 Finding a house using customized data

Sometimes, customer will not fall into any of the above 6 groups. In this case, we will let them customize their selection based on the 7 parameters as below:

- Food and store
- Sight-seeing

- Service
- Sport and leisure
- Entertainment
- Lodging option
- Transportation

All parameters will be rated from 0 to 100 based on customer selection. Below is one example of customer customized selection:

```
lv_food = 100                # food and drink
lv_sightSeeing = 20         # sight-seeing and culture
lv_service = 20              # service and shop
lv_sport = 100                # sport and leisure
lv_entertain = 100           # entertainment
lv_lodging = 100             # lodging
lv_transport = 20           # trasportation
```

The rental fee is limited at 2000$ per month.

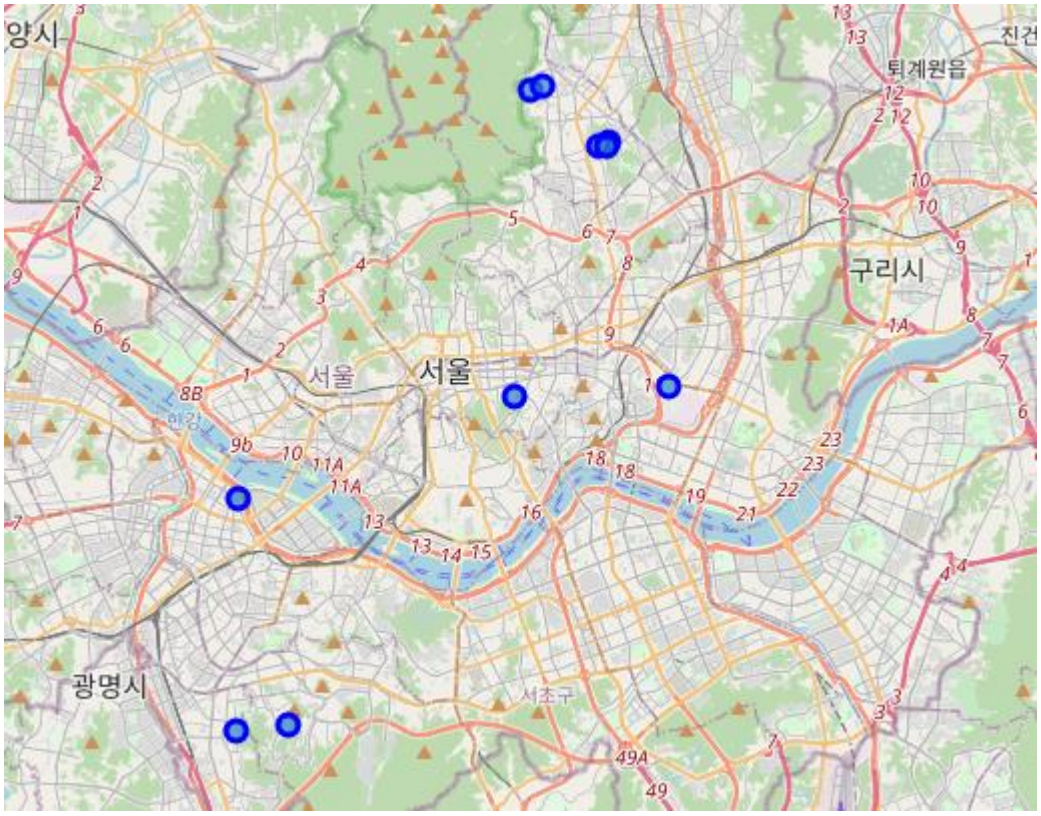
At this time, there is no limited to any particular district.

The customer selected data was normalized and then using the Minkowski distance from scipy library, we can calculate the similarity between customer profile and every neighborhood row in our database.

The resulted dataset was then sorted by the calculated distance. The result was as below:

Neighborhood	Entertainment	Food and drink	Lodging	Service and store	Sightseeing and culture	Sports and leisure	Transport	Mindis
유진빌라_1147	0.055556	0.444444	0.000000	0.111111	0.166667	0.222222	0.000000	0.845411
홍익빌라_1142	0.055556	0.500000	0.000000	0.166667	0.055556	0.222222	0.000000	0.845411
아이나라놀이방_4805	0.000000	0.590909	0.136364	0.045455	0.000000	0.136364	0.090909	0.845850
현대하이츠빌라_8557	0.050000	0.500000	0.100000	0.150000	0.000000	0.100000	0.100000	0.891304
홍익빌라_1020	0.000000	0.600000	0.040000	0.040000	0.080000	0.160000	0.080000	0.911304

The top 10 locations were shown to customer for further selection



The customer now can easily find the location they want to live among our suggested locations.

6. Discussion

The classification result show several quite distinctive group such as sport dominance or transport dominance. This shows the effectiveness of clustering algorithm in grouping data.

The clustered result can also show the strength and weakness in each group based on what they have and what they lack. The user can therefore select area that best matched their need. The government can also know the weakness of each area to improve its living quality.

7. Conclusion

In this study, the problem of renting a house in Seoul was solved using K-Means clustering algorithm and venue data from Foursquare API. It can be seen that the neighborhood can be clustered in 6 different groups with acceptable accuracy. The user now can easily select a rent house based on their spec such as renting price, renting location and renting group. If a user want a customized result, the program can provide him with customized profile selection. The profile then will be matched with all neighborhood data using Minkowski distance to find the best location.

8. Future direction

Other model of non-supervised clustering algorithm should also be used to check their effectiveness in classification compared with K-Means clustering. A best model will then be selected to be our main engine in the selection process.

On the other hand, a connection with actual house renting price will become a real useful application. This will require the cooperation between real estate data provider and Foursquare data.

9. Acknowledgement

Deeply thanks the Coursera team & IBM for helping me finish this project.