Capstone Project
- The Battle of Neighborhoods:

Comparison of New York Manhattan neighborhoods for the coffee shop Biz investors

Jan. 2019

S.B.Park

1. Introduction

1.1 Business problem

Our customer wish to start a new coffee shop franchise business in New York city. Because there are already many coffee shops and cafés in this big city, in order to catch new market opportunities, our customers want to open coffee shop brand with differentiated characteristics. The customer is preparing a Korean style coffee shop brand considering the popularity of the K-pop music culture which has many fans in this area recently.
The customer wants to determine which of the area(s) in New York city are better suited to open coffee shop businesses.

1.2 Who would be interested in this project

This project report will provide analysis data to the customer who want to open Korean style coffee shop brand in big cities in North America (such as New York, Toronto), the best quality areas for the determine of the coffee shop business.

2. Data sources

Following sources of data are used for the analysis of this Capstone project:

2.1 FourSqure API

- Query to get such categories information.
  . Number of coffee shop & café categories.
  . Number of (Korean) restaurant category.
  * ex) https://api.foursquare.com/v2/venues/explore?client_id=<id>&client_secret=<secret>E&v=X&ll=??.??,-??.?? &radius=??? &limit=999 &categoryId=??,??
- Date type: JSON

※ Used category codes

a) Coffee shop	=>	4bf58dd8d48988d1e0931735
b) cafe	=>	4bf58dd8d48988d16d941735
c) Korean restaurant	=>	4bf58dd8d48988d113941735

2.2 Neighborhood data

- Geospatial postal data for NY, USA (type: CSV)
  http://cocl.us/Geospatial_data

3. Methodology

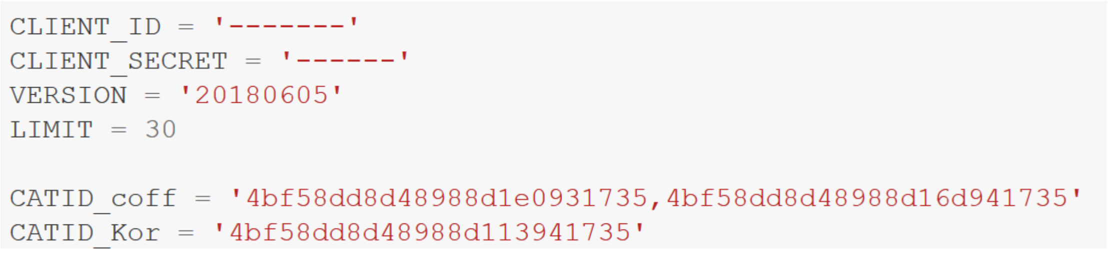
Followed on overall machine learning data analysis methodology.
And it's been used planning, data gathering, cleansing, EDA and machine learning processes.

3.1 Data Download and Explore Dataset

a) NY City has a total of 5 boroughs and 306 neighborhoods. In order to segement the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and logitude coordinates of each neighborhood. I will reuse file that I downloaded before, so I can simply run a wget command and access the data.

b) Use geopy library to get the latitude and longitude values of Manhattan, New York City.
For data analysis purposes, I will simplify segment and cluster only the neighborhoods in Manhattan where is center of Korea culture in NY city. So let's slice the original dataframe and create a new dataframe of the Manhattan data.

c) Define Foursquare Credentials and version



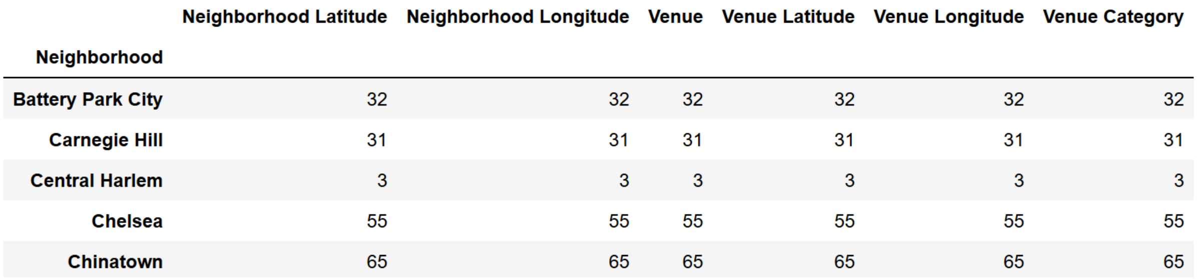
```
CLIENT_ID = '-------'
CLIENT_SECRET = '------'
VERSION = '20180605'
LIMIT = 30

CATID_coff = '4bf58dd8d48988d1e0931735,4bf58dd8d48988d16d941735'
CATID_Kor = '4bf58dd8d48988d113941735'
```

3.2 Explore neighborhoods in Manhattan

a) Query Café, coffee shop and Korean restaurant venues through Foursquare API by each neighborhoods in Manhattan.

b) Concat queried data files (Café coffee shop venues file and Korean restaurant venues file.), and grouped by 'neighborhood' for counting venues.



	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Battery Park City	32	32	32	32	32	32
Carnegie Hill	31	31	31	31	31	31
Central Harlem	3	3	3	3	3	3
Chelsea	55	55	55	55	55	55
Chinatown	65	65	65	65	65	65

3.3 Analyze and cleansing each neighborhood's coffee shop & Korean restaurants

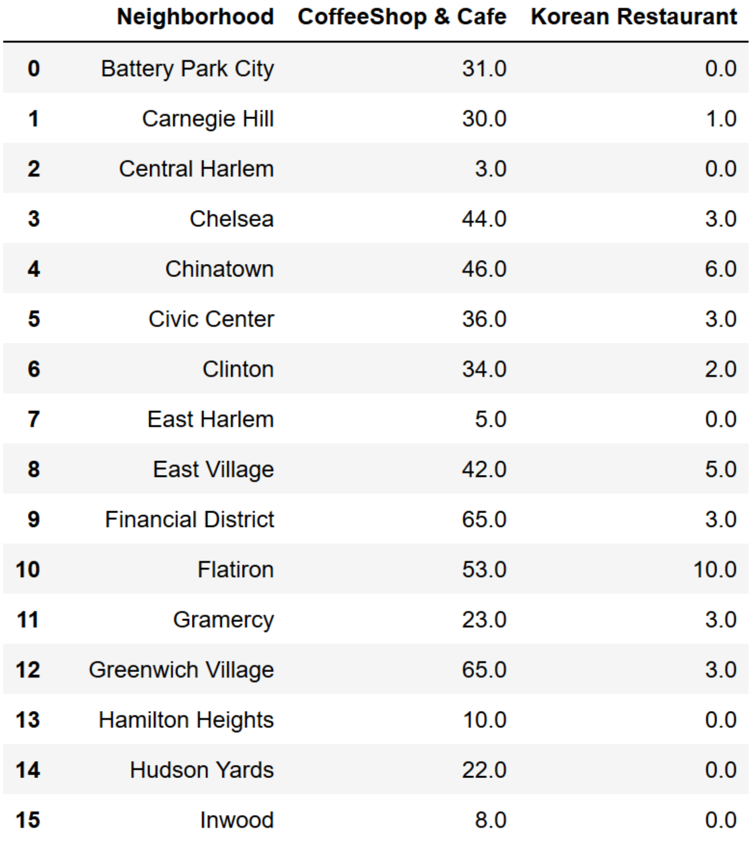
a) Data columns realign by queried category IDs. Then I found there are lots of (about 29) garbage categories.

b) And I reselect correct data columns from prior merged dataframe file.

c) Grouping by each neighborhoods

d) counting total venues by each neighborhoods to calculate by each categories.

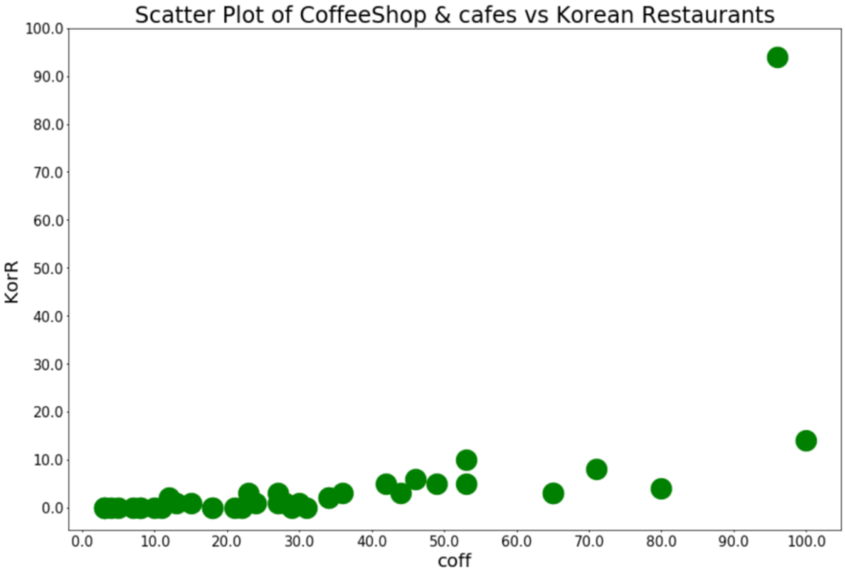
e) concat two dataframe files, and replace each categories data using calculated venues data.



	Neighborhood	CoffeeShop & Cafe	Korean Restaurant
0	Battery Park City	31.0	0.0
1	Carnegie Hill	30.0	1.0
2	Central Harlem	3.0	0.0
3	Chelsea	44.0	3.0
4	Chinatown	46.0	6.0
5	Civic Center	36.0	3.0
6	Clinton	34.0	2.0
7	East Harlem	5.0	0.0
8	East Village	42.0	5.0
9	Financial District	65.0	3.0
10	Flatiron	53.0	10.0
11	Gramercy	23.0	3.0
12	Greenwich Village	65.0	3.0
13	Hamilton Heights	10.0	0.0
14	Hudson Yards	22.0	0.0
15	Inwood	8.0	0.0

3.4 Cluster venues (=coffee shops and Korean restaurants) by neighborhoods.

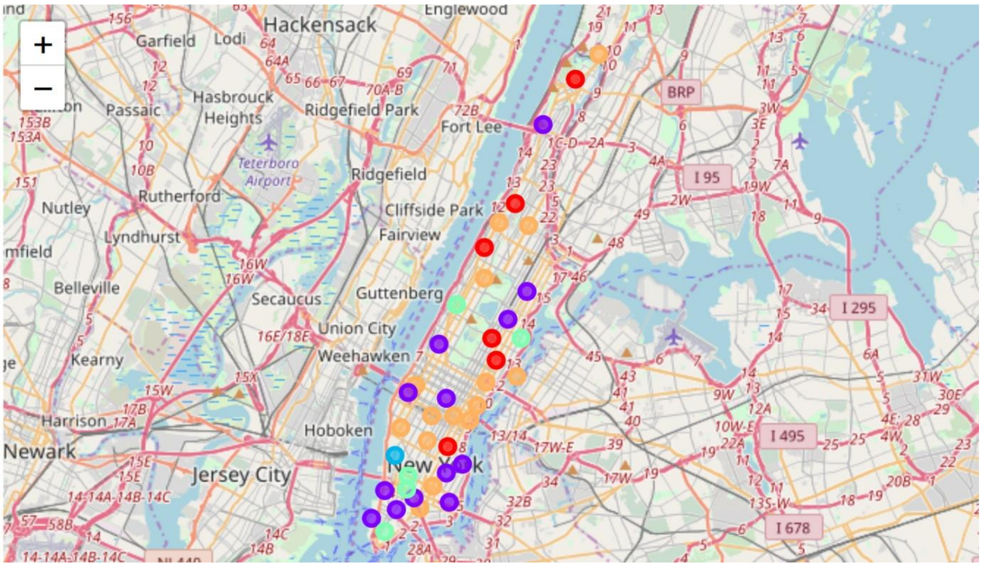
a) Initialize k-means – plot data point and displaying it using Scatter Plot function.

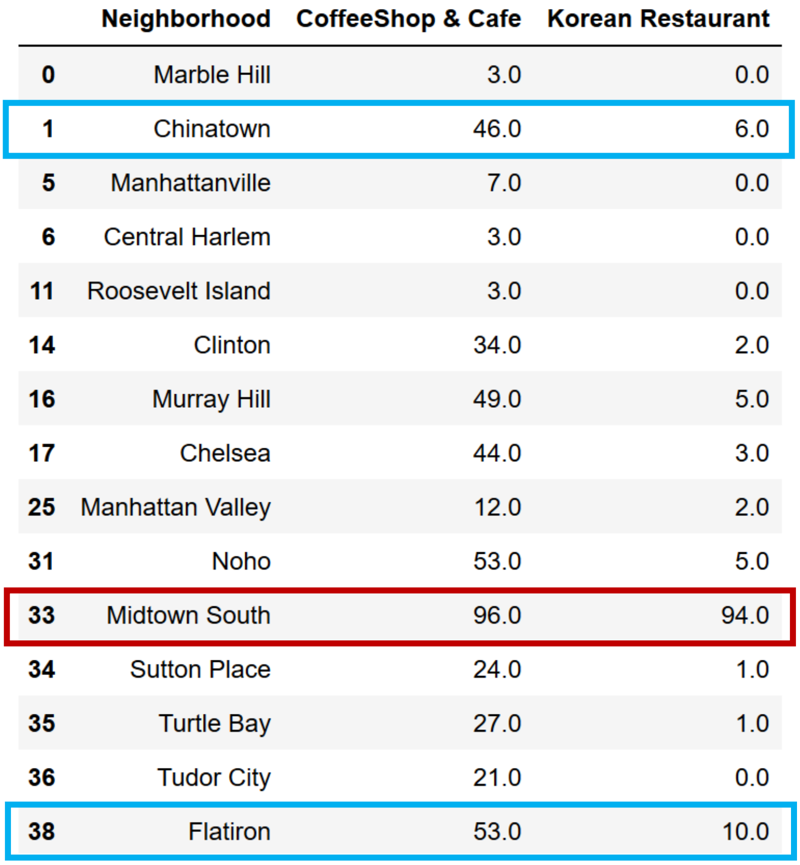


b) Run k-means to cluster the neighborhood into 5 clusters.

4. Results

By running the exploratory data analysis for the Manhattan of New York city to find recommending place for the new Korean style coffee shop, the machine learning algorithm recommends 'Midtown South' and some neighborhoods.





	Neighborhood	CoffeeShop & Cafe	Korean Restaurant
0	Marble Hill	3.0	0.0
1	Chinatown	46.0	6.0
5	Manhattanville	7.0	0.0
6	Central Harlem	3.0	0.0
11	Roosevelt Island	3.0	0.0
14	Clinton	34.0	2.0
16	Murray Hill	49.0	5.0
17	Chelsea	44.0	3.0
25	Manhattan Valley	12.0	2.0
31	Noho	53.0	5.0
33	Midtown South	96.0	94.0
34	Sutton Place	24.0	1.0
35	Turtle Bay	27.0	1.0
36	Tudor City	21.0	0.0
38	Flatiron	53.0	10.0

5. Discussion

Based on By running the exploratory data analysis for the Manhattan of New York city to find recommending place for the new Korean style coffee shop, the machine learning algorithms recommends 'Midtown South' and some neighborhoods.

Based on the data, two types of machine learning analysis techniques were used. We have looked at two criteria for selecting a recommended area for Korean style coffee shop.
First, the area where many coffee shops are currently operating.
Second, there are many Korean restaurants nearby.

The area that meets both of these two conditions is assumed to be the region with the highest recommend for the Korean style coffee shop. Based on these assumptions, the results of data analysis and machine learning analysis are shown.

6. Conclusion

Based on by results and discussion, our best recommending neighbors are as below

1) Best recommending neighbor is : Midtown South
2) Second recommending neighbor is : Flatiron
3) Third recommending neighbor is : Chinatown

Among the adopted two machine learning algorithms, the scatter plot graph was easier to select a region that meets the pre-conditions. From the K-means clustering, we have been recommended for more areas, but further analysis will be needed to determine which criteria are clearly recommended. If these are near the top of the region, or for any other reasons. Supplementing the referral system in the future will allow more sophisticated recommendations.

Nevertheless, based on the data analysis, it is considered that sufficient analysis results are recommended to recommend the best place to carry out the business in Korea style coffee shop.