17 August 2020

Avoid Risk Neighbourhoods: General Insecurity & COVID-19

Data Section

As a recapitulation of the introduction, the Client hired me as an independent consultant to conduct a preliminary research for their future relocation plan. The Client sets the following three objectives for the assignment.

1. Identify outlier high risk neighbours (the Outlier Neighbourhood/Cluster) in terms of these two risks—the general security risk (crime) and the pandemic risk (COVID-19).
2. Segment non-outlier neighbours into several clusters (the Non-Outlier Neighbourhoods) based on a quantitative risk metrics.
3. Use Foursquare API to characterize the Non-Outlier Neighbourhoods regarding popular venues.

[1] Data Requirements

These objectives determine the data requirements as follow:

1. First of all, I would need to gather the following basic information about the neighbourhoods in Buenos Aires.
   - The area and the population for each neighbourhood
   - The geographical coordinates to determine the administrative border of each neighbourhood (for map visualization)
2. Second, for the first and the second objective, I would need to gather the following historical statistics to construct a compound risk measurement to quantify both the general insecurity risk (crime) and the pandemic risk (COVID-19).
   a) general security risk statistics (crime incidences) by neighbourhoods
   b) pandemic risk statistics (COVID-19 confirmed cases) by neighbourhoods
3. Third, for the third objective, since the Client requires me to specifically use Foursquare in order to characterise each Non-Outlier Neighbourhood, I would explore Foursqare data to organize top 10 venue categories.

[2] Data Sources

Michio Suginoo

17 August 2020

I would rely on the following sources to meet the data requirements.

1. Basic info of Neighbourhoods of the Autonomous City of Buenos Aires:
   a) the area and the population of each Barrio from Wikipedia: https://en.wikipedia.org/wiki/Neighbourhoods_of_Buenos_Aires
   b) The city government of Buenos Aires provides a GeoJson file that contains the geographical coordinates which defines the administrative boundary of Barrios (the neighbourhoods) of Buenos Aires. https://data.buenosaires.gob.ar/dataset/barrios/archivo/1c3d185b-fdc9-474b-b41b-9bd960a3806e
2. Historical statistics of the general insecurity risk (crime) and the pandemic risk (COVID-19).
   c) Crime Statistics:
   A csv file which is compiled and uploaded by Rama in his GitHub depository: https://github.com/ramadis/delitos-caba/releases/download/3.0/delitos.csv
   d) COVID-19 Statistics:
   the city government's website provides the COVID-19 statistics by neighbourhood:
   https://cdn.buenosaires.gob.ar/datosabiertos/datasets/salud/casos-covid-19/casos_covid19.xlsx
3. Foursquare Data for Popular Venues by Neighbourhood: as per the Client's requirement to specifically use Foursquare in order to characterise each Non-Outlier Neighbourhood.

[3] Data Cleaning

(1) Consistency of the name of the neighbourhoods:
All these data sources specified above have at least one key variable in common, the name of 'Neighbourhoods'. Nevertheless, there is no consistency in how their names are spelled out. I have to make reconciliation among them in order to merge the relevant data to create one single dataframe for analysis.

(2) The translation from Spanish to English:
All the Spanish descriptions of relevant data are translated into English for the benefit of the English speaking Client.

Michio Suginoo

[4] Limitation of Raw Data & Data Transformation

(1) Basic Info about Neighbourhoods:

The data source a) above contained the following two relevant data for the project.

- The size of area for each neighbourhood.
- The size of population for each neighbourhood.

As mentioned earlier, I would need the population density for the analysis. Thus, I generated the population density by a simple division between these two statistics above.

- Population Density = Population / Area

(2) Crime Statistics:

The compiled crime data covers only the period between Jan 1, 2016 and Dec 31, 2018. It does not provide us with the recent trend. This limitation is noted. There were the following 7 crime types registered in the statistics:

- 'Homicidio Doloso': Homicide (Intentional)
- 'Robo (Con violencia)': Theft (with Violence)
- 'Hurto Automotor': Automotive Theft
- 'Hurto (Sin violencia)': Theft (without violence)
- 'Robo Automotor': Automotive Theft
- 'Homicidio Seg Vial': Homicide (on the street)
- 'Lesiones Seg Vial': Road Injuries

The ultimate purpose to use these data is to quantify the general insecurity risk for each Neighbourhood. In this context, I need to reduce the amount of information to one single metrics. In order to do so, I do the following data transformations:

- Reorganize these 7 crime types into two general types: ‘Homicide’ and ‘Non-Homicide’.
- Assign the score 10 to each ‘Homicide’ incidence and 1 to each ‘Non-Homicide’ incidence.
- Aggregate the scores by Neighbourhood to generate ‘Crime Severity Scores’ per each Neighbourhood.

As the product of the Data Wrangling, I generated 'Crime Severity Scores' per each Neighbourhood.

(3) COVID-19 Statistics:

From the downloaded data, I only extracted two info: the name of the neighbourhood and the COVID-19 confirmed incidences. And I aggregated them by neighbourhood to generate COVID-19 Confirmed Cases for each neighbourhood.

(4) Working Data Consolidation

Based on my data wrangling, I consolidate the following relevant data for the analysis into one single Pandas' DataFrame.

1. Neighbourhood: the name of each neighbourhood
2. Area in km^2: the area of each neighbourhood
3. Population: the population of each neighbourhood
4. Population Density: the population density of each neighbourhood
5. Crime Severity Score (CSS): the cumulative CSS of each neighbourhood
6. COVID-19 Confirmed Cases: the cumulative number of COVID-19 confirmed cases.

In the next page, let me present the consolidated dataframe.

(5) Foursquare Data:

Foursquare API allows the user to explore venues from one single point within a user specified radius. This imposes a critical constraint in exploring the venue within a neighbourhood from its corner to corner. I have to set the following variables to make a query:

- The geographical coordinates of one single starting point
- 'radius': he radius to set the geographical scope of the query,

This set of API's query constraints can limit the quality of analysis.

Under this constraint, I set my query with radius=1000 to explore venues within the radius of 1 km from the centre of each neighbourhood.

17 August 2020

	Neighbourhood	Area in km²	Population	Population_Density	Crime Severity Score (CSS)	COVID-19 Confirmed Cases
0	AGRONOMIA	2.1	13963	6649.047619	1288	151
1	ALMAGRO	4.1	128206	31269.756098	14406	2981
2	BALVANERA	4.4	137521	31254.772727	23474	4983
3	BARRACAS	7.6	73377	9654.868421	10329	5050
4	BELGRANO	6.8	126816	18649.411765	11588	1735
5	BOEDO	2.6	45563	17524.230769	5924	1039
6	CABALLITO	6.8	170309	25045.441176	16865	2843
7	CHACARITA	3.1	25778	8315.483871	4464	1028
8	COGHLAN	1.3	18021	13862.307692	1514	269
9	COLEGIALES	2.3	52391	22778.695652	4130	789
10	CONSTITUCION	2.1	41894	19949.523810	9951	1710
11	FLORES	7.8	142695	18294.230769	19467	6883
12	FLORESTA	2.3	37247	16194.347826	4623	866
13	BOCA	3.1	43413	14004.193548	5150	1818
14	PATERNAL	2.2	19058	8662.727273	2106	755
15	LINIERS	4.3	42083	9786.744186	7294	781
16	MATADEROS	7.3	62206	8521.369863	8680	1319
17	MONSERRAT	2.2	39175	17806.818182	8566	1248
18	MONTE CASTRO	2.6	32782	12608.461538	3062	445
19	NUEVA POMPEYA	6.2	60465	9752.419355	7834	2097
20	NUNEZ	4.5	49019	10893.111111	5444	551
21	PALERMO	15.9	225245	14166.352201	27905	3840
22	PARQUE AVELLANEDA	5.1	51678	10132.941176	5774	1770
23	PARQUE CHACABUCO	3.8	54638	14378.421053	7585	1225
24	PARQUE CHAS	1.4	18926	13518.571429	1336	239
25	PARQUE PATRICIOS	3.7	37791	10213.783784	6211	1379
26	PUERTO MADERO	2.1	406	193.333333	347	387
27	RECOLETA	5.9	165494	28049.830508	16716	2791
28	RETIRO	2.8	38635	13798.214286	8938	3501
29	SAAVEDRA	5.6	48956	8742.142857	5585	571
30	SAN CRISTOBAL	2.1	46494	22140.000000	6097	1525
31	SAN NICOLAS	2.3	28667	12463.913043	17472	890
32	SAN TELMO	1.2	23198	19331.666667	5022	516
33	VELEZ SARSFIELD	2.4	34084	14201.666667	3589	675
34	VERSALLES	1.4	13556	9682.857143	1546	182
35	VILLA CRESPO	3.6	83646	23235.000000	8753	1411
36	VILLA DEL PARQUE	3.4	55502	16324.117647	4270	679
37	VILLA DEVOTO	6.4	67712	10580.000000	6312	887
38	VILLA GRAL. MITRE	2.2	34204	15547.272727	3802	660
39	VILLA LUGANO	9.0	108170	12018.888889	11799	6037
40	VILLA LURO	2.6	31859	12253.461538	3450	574
41	VILLA ORTUZAR	1.8	21256	11808.888889	1960	313
42	VILLA PUEYRREDON	3.3	38558	11684.242424	3424	456
43	VILLA REAL	1.3	13681	10523.846154	1551	230
44	VILLA RIACHUELO	4.1	13995	3413.414634	2035	420
45	VILLA SANTA RITA	2.2	32248	14658.181818	3532	596
46	VILLA SOLDATI	8.6	39477	4590.348837	5447	2866
47	VILLA URQUIZA	5.4	85587	15849.444444	7422	1354

For the data understanding, I would divide it into two section:

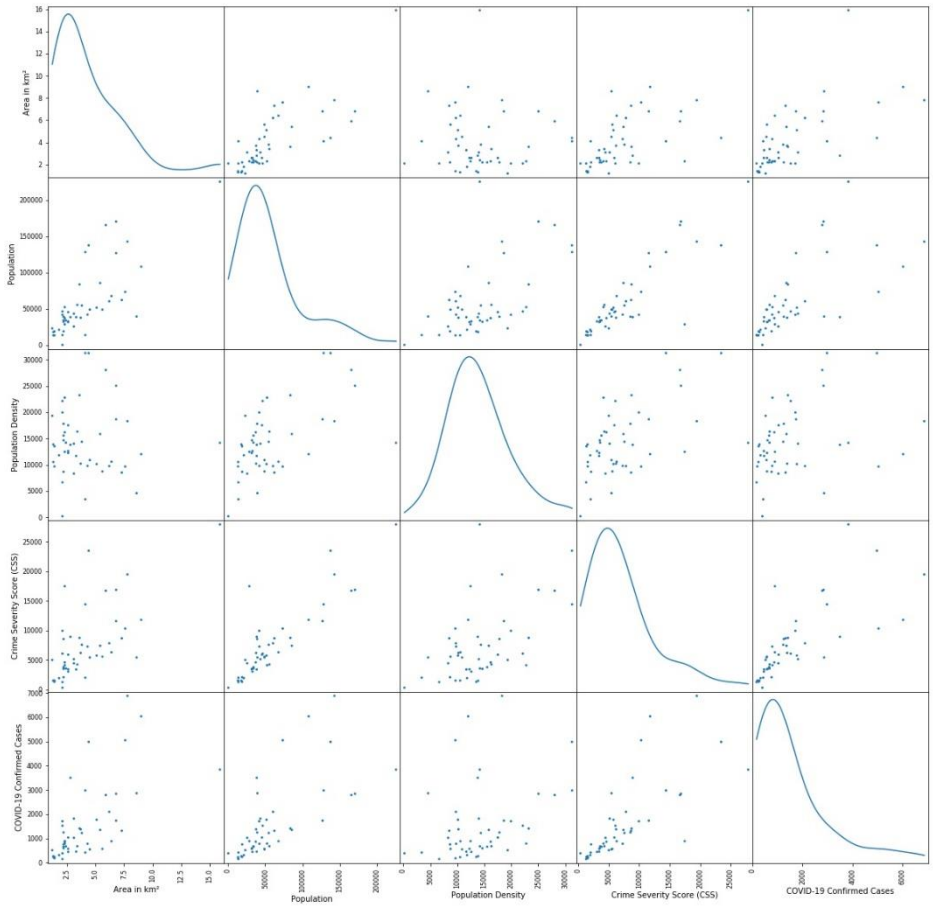
- first, data understanding about the consolidated data: 5 variables above: area, population, population density, crime severity scores, and COVID-19 confirmed cases.
- second, data understanding about Foursquare data

Michio Suginoo

17 August 2020

[5] Data Understanding Part 1: Consolidated Data (5 variables)

In order to understand the underlying structure of the given data, first I plotted the scatter matrix of these 5 variables above: area, population, population density, crime severity scores, and COVID-19 confirmed cases.



A diagonal cell of the scatter plot illustrates the distribution of the corresponding variable. All the plots in the diagonal cells are highly skewed and suggest the presence of outliers.

The distributions on the first three diagonal cells are related with the basic profile features of neighbourhoods: area, population, and population density.

- There is a wide range of these three basic features of the neighbourhood profiles: area, population, and population density. They are all skewed. But, the population density has the least skewed distribution, if not symmetric one.

The distributions on the last two diagonal cells are related with these two risk scores: Crime Severity Score and COVID-19 Confirmed Cases.

- These risk metrics are not scaled by the difference in the size of profile features among the neighbourhoods. And both of them demonstrate high skewness, suggesting the presence of outliers.

The other plots off the diagonal cells reflect the pairwise correlations among these 5 variables. In order to be more quantitatively specific, I plotted the correlation matrix among them.

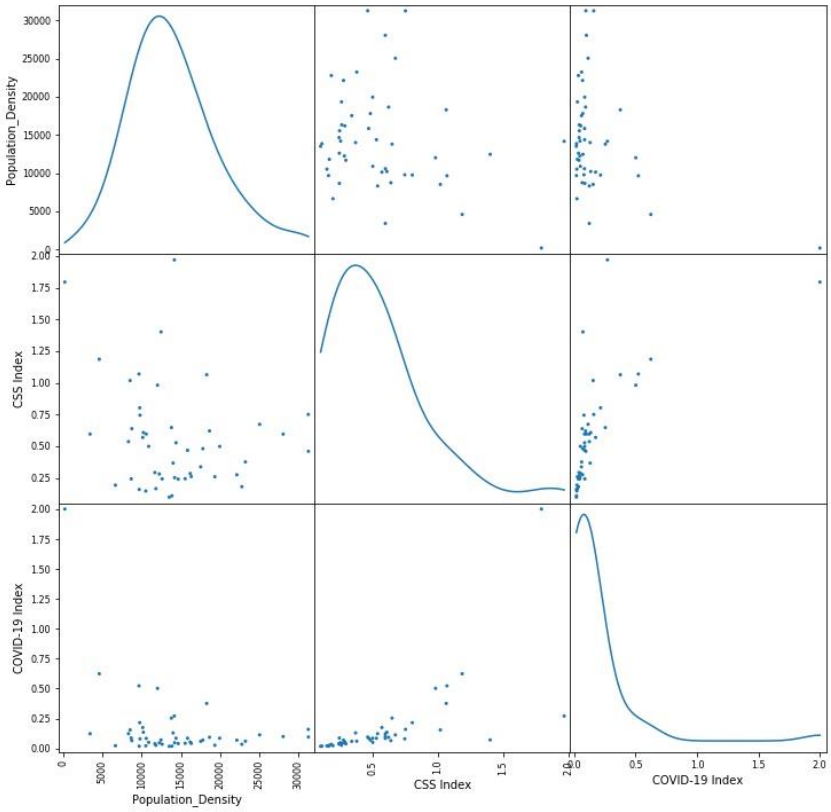
	Area in km²	Population	Population_Density	Crime Severity Score (CSS)	COVID-19 Confirmed Cases
Area in km²	1.00	0.77	-0.01	0.67	0.65
Population	0.77	1.00	0.58	0.88	0.70
Population_Density	-0.01	0.58	1.00	0.54	0.33
Crime Severity Score (CSS)	0.67	0.88	0.54	1.00	0.74
COVID-19 Confirmed Cases	0.65	0.70	0.33	0.74	1.00

- Population density demonstrates the lowest correlation with other variables.
- Nevertheless, as expected, population density demonstrates positive correlation with the general insecurity risk metric and the COVID-19 risk metric—"Crime Severity Score" (0.54) and "COVID-19 Confirmed Cases" (0.33)—suggesting that the crime is more correlated with population density than COVID-19 infection.
- The other four variables—the area, the population, Crime Severity Score, and COVID-19 Confirmed Cases—demonstrate high pairwise correlation among them.

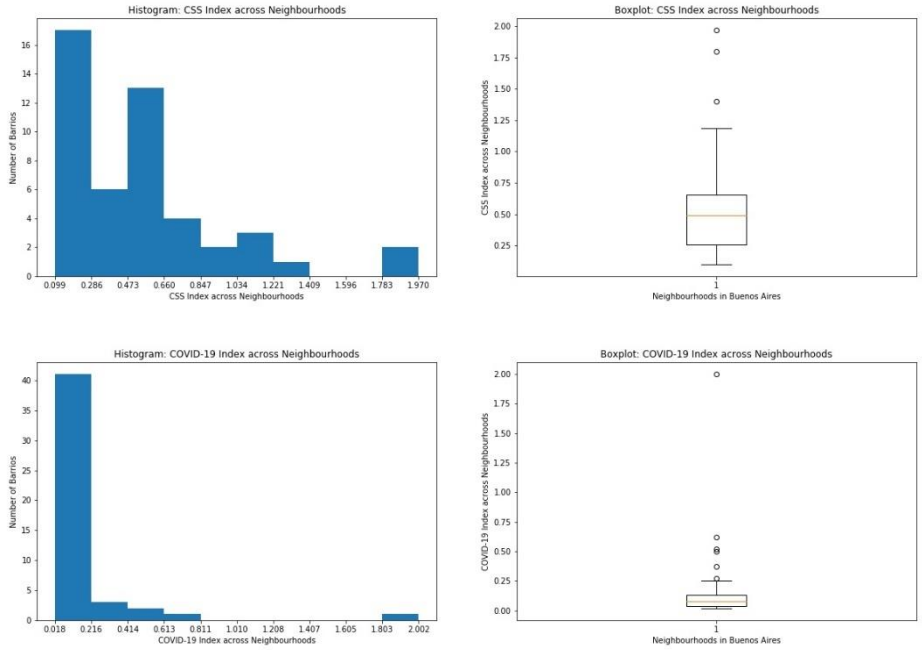
One option to scale the indices is to divide them by population, which demonstrates the highest correlation with these indices. Nevertheless, for the first submission of the report, in order to reflect the concern of social distancing during the pandemic, I choose population density to generate the following scaled indices:

- CSS Index = Crime Severity Scores / Population Density
- COVID-19 Index = COVID-19 Confirmed Case/ Population Density

Below, I plot the scatter matrix among these two new indices and the population density. The new scatter matrix also indicates the presence of outliers in these two new indices.



Next, in order to observe the distributions of these two new indices more individually, I plot histograms and boxplots for them. Visualization can be very powerful. Just looking at them, we can identify some salient characteristics of the variables. Especially, each boxplot displays outliers beyond its upper whisker.



Here is the list of those datapoints beyond the upper whiskers of these two box plots.

	index	Neighbourhood	CSS Index	Outlier
0	21	PALERMO	1.969808	CSS Outliers
1	26	PUERTO MADERO	1.794828	CSS Outliers
2	31	SAN NICOLAS	1.401807	CSS Outliers
3	46	VILLA SOLDATI	1.186620	CSS Outliers

	index	Neighbourhood	COVID-19 Index	Outlier
0	26	PUERTO MADERO	2.001724	COVID-19 Outlier
1	46	VILLA SOLDATI	0.624353	COVID-19 Outlier
2	3	BARRACAS	0.523052	COVID-19 Outlier
3	39	VILLA LUGANO	0.502293	COVID-19 Outlier
4	11	FLORES	0.376239	COVID-19 Outlier
5	21	PALERMO	0.271065	COVID-19 Outlier
6	28	RETIRO	0.253728	COVID-19 Outlier

There are some overlapping cases. Consolidating them, here is the list of overall risk outliers. There are 8 overall risk outliers.

```
0    PUERTO MADERO
1    VILLA SOLDATI
2        BARRACAS
3     VILLA LUGANO
4           FLORES
5          PALERMO
6           RETIRO
7     SAN NICOLAS
```

This is an observation based on the boxplot visualization at the stage of Data Understanding. In the further analysis, I would use a Clustering Machine Learning method to identify outliers separately to see if this observation is consistent with the machine learning result.

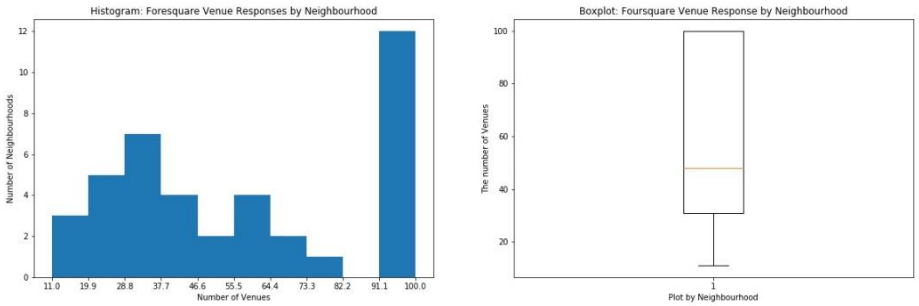
[6] Data Understanding Part 2: Foursquare Data

Foursquare API allows the user to explore venues within a user specified radius from one single location point. In other words, the user needs to specify the following parameters:

- The geographical coordinates of one single starting point 'radius'.
- The radius to set the geographical scope of the query,

This imposes a critical constraint in exploring the venue within a neighbourhood from its corner to corner. Since there is no uniformity in the area size among neighbourhoods, a compromise would be inevitable when we want to capture the venue profile of a neighbourhood from corner to corner within its geographical border. I will use ***geopy's Nominatim*** to obtain the representative single location point for each Neighbourhood.

Under this constraint, I set my query with radius=1000 to explore venues within the radius of 1 km from the centre of each neighbourhood. And here is the summary of the Foursquare response to my query: the histogram and the boxplot of the number of venues across different neighbourhoods.



These two plots display the distribution of the Foursquare sample responses. It might suggest that there might be some issues in the coherency of data availability across different neighbourhoods. This might affect the quality of the result of clustering machine learning.

[7] How to use Data for Analysis

This section will brief how to use the data to meet the three objectives of the Client.

Over the course of analysis, I will ***normalize/standardize*** 'CSS Index' and 'COVID-19 Index' and plot the neighbourhoods as the two-dimensional spatial data points over the space of these two standardized risk indices, 'Standardized CSS Index' vs 'Standardized COVID-19 Index' space. Then, for these standardized indices, I will

explore three types of Clustering Machine Learning Models in one way or another to meet the three objectives that the Client demanded:

- DBSCAN Clustering to identify 'Outlier Neighbourhoods' with respect to their risk profiles.
- Hierarchical Clustering to discover hidden underlying clusters among 'Non-Outlier Neighbourhoods' with respect to their risk profiles.
- K-Means Clustering to discover hidden underlying clusters among 'Non-Outlier Neighbourhoods' with respect to Venue Composition (Foursquare Data).

For DBSCAN and K-Means models, I would use hyperparameter tuning models to identify appropriate hyperparameters for machine learning model setting: such as K, the number of clusters, for K-Mean Clustering.

I use these unsupervised machine learning models since there is no label (cluster) in the data. In other words, we have no empirical idea about the cluster structure among the neighbourhoods for these two risks. By design, no supervised machine learning model can handle this case.

(1) Objective 1:

In order to identify 'Outlier Neighbourhoods', I will run DBSCAN Clustering.

- Normalize the data over the entire neighbourhoods.
- Tune the hyperparameter, 'eps' I would use ***KneeLocator*** of the python library ***kneed***. For more details: https://towardsdatascience.com/how-to-use-dbscan-effectively-ed212c02e62
- Run DBSCAN model over the entire neighbourhoods with the tuned 'eps' value.

(2) Objective 2: Segment non-outlier neighbours into several clusters based on a quantitative risk metrics.

Since I would not have an 'a priori knowledge' what number of clusters would be appropriate, Hierarchical clustering would give me a flexible solution. A

hierarchical clustering method does not require a predetermined value for the number of clusters. It generates a dendrogram to illustrate the tree-like cluster structure based on the pairwise distance distribution. It appeals to our intuition to pick the most desirable cluster structure from the 'dendrogram'.

- Normalize only data of the Non-Outlier Neighbourhoods
- Generate a hierarchical dendrogram and discover the most appropriate structure of the cluster and choose the corresponding K, the number of the clusters.
- Segment the Non-Outlier Neighbourhoods into the selected number of clusters by hierarchical clustering method. Make the scatter plot of the K clusters.

I will present Choropleth Map to visualize the findings for these first two objectives.

(3) Objective 3: Characterize the Non-Outlier Neighbourhoods based on popular venues, using Foursquare API.

I will carry out two analysis: 1) Popular Venue Analysis and 2) Segmentation of Neighbourhoods based on Venue Composition.

a) Popular Venue Analysis:

With Foursquare data, which is venue base information, I will apply One Hot Encoding to transform the venue base data to a neighbourhood base data: to summarise top 10 popular venue categories for each neighbourhood for 'Non-Outlier Neighbourhoods'.

b) Segmentation of Neighbourhoods based on Venue Composition

For this purpose, I contemplate to run K-Means Clustering Machine Learning. Before running the model, I will run two hyperparameter tuning models—K-Means Elbow Method and Silhouette Score Analysis—to identify its most important hyperparameter, K, the number of clusters.

For the detail of the hyperparameter tuning methods, here are references:

- Silhouette Score Analysis: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
- K-Mean Elbow Method: https://predictivehacks.com/k-means-elbow-method-code-for-python/

These models would give me an insight about how to cluster the data for a meaningful analysis. Based on the finding from these models, I would decide how to implement the clustering machine learning model.