Full Report:

Map Risk Clusters of Neighbourhoods in the time of Pandemic

Three Different Lessons from Three Different Clustering Analyses

Introduction/Business Problem

Every year is unique and particular. But, 2020 brought the world the special planetary pandemic challenge of COVID-19. It spread and penetrated rapidly into different parts of the globe. And, the autonomous city of Buenos Aires (CABA: Ciudad Autonoma de Buenos Aires) is not an exception.

In this particular setting, I contemplated for the capstone project a hypothetical corporate client in the food industry from abroad (The Client), that is planning to relocate their representative family to the city of Buenos Aires for their future penetration into Argentina once the pandemic-related restrictions are lifted. Since this would be its very first entry to Buenos Aires, the city is still an unknown territory for the Client. Very concerned with the two risks—the general security risk (crime) and the pandemic risk (COVID-19)—the Client wants to exclude high risk neighbourhoods in the selection of the location for the relocation plan.

At the same time, the Client wants to understand the characteristics of neighbourhoods by popular commercial venue categories such as restaurants, shops, and sports facilities. In this context, the Client hired me as an independent consultant to conduct a preliminary research for its future plan. The Client sets the following three objectives for this assignment.

1. Identify outlier high risk neighbourhoods (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 neighbourhoods into several clusters (the Non-Outlier Neighbourhoods/Clusters) and rank them based on a single quantitative risk metric (a compound risk metric of the general security risk and the pandemic risk).
3. Use Foursquare API to characterize the Non-Outlier Neighbourhoods regarding popular venues. And if possible, segment Non-Outlier Neighbourhoods according to popular venue profiles.

In addition, the Client demonstrated high enthusiasm about Machine Learning and requested us to use machine learning models to achieve all these objectives.

The autonomous city of Buenos Aires is a densely populated city: the total population of approximately 3 million in the area of 203 km^2. And each neighbourhood has its own distinct size of area and population. The city is divided into 48 administrative division, aka ‘barrios’, to which I will refer simply as ‘neighbourhoods’ in this report.

The Clients expressed their concern about the effect of the variability of population density among neighbourhoods. These two risks of the Client’s concern—the general security risk (crime) and the pandemic risk (COVID-19)—are likely affected by the population density profiles. Especially, the fact that as ‘social distancing’ is a key to the prevention of COVID-19 suggests that population density is a significant attribute for the pandemic risk. In other words, the higher the population density, the higher the infection rate. The similar can be said for the general insecurity. Obviously, this preconception needs to be assessed based on the actual data in the course of the project. Nevertheless, the Client made a request to scale risk metrics by 'population density'. This needs to be kept in mind for the analysis.

Data Section

A Data Science project analysis demands a massive effort in data wrangling: collecting and observing the raw data; cleaning it; transforming it; shaping insights from the given data. In this section, I will discuss my findings of the data for this project.

[1] Data Requirements

The three objectives clearly set by the Client 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 profile neighbourhoods from the perspectives of 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.

[2] Data Sources

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 all the relevant neighbourhoods 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.

[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 compress the amount of information into one single metric. 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:

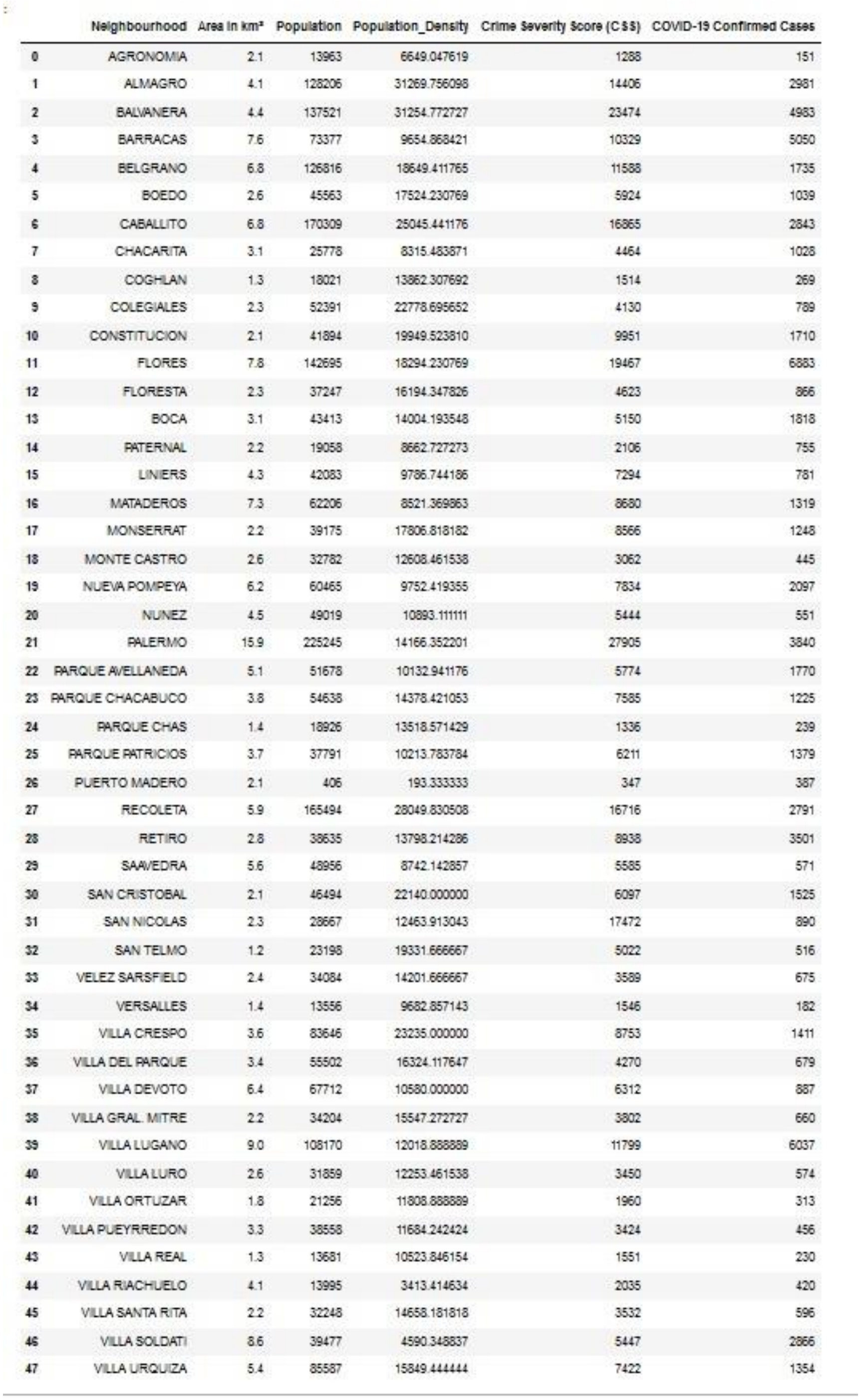
In order to measure the pandemic risk, I simply extracted the cumulative confirmed cases of COVID-19 for each neighbourhood. I did not net out the recovered cases from the data. Thus, the COVID-19 statistics in this analysis is a gross figure. This will proxy the historical cumulative risk of infection.

(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.



	Neighbourhood	Area in km^2	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

(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,

The 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.

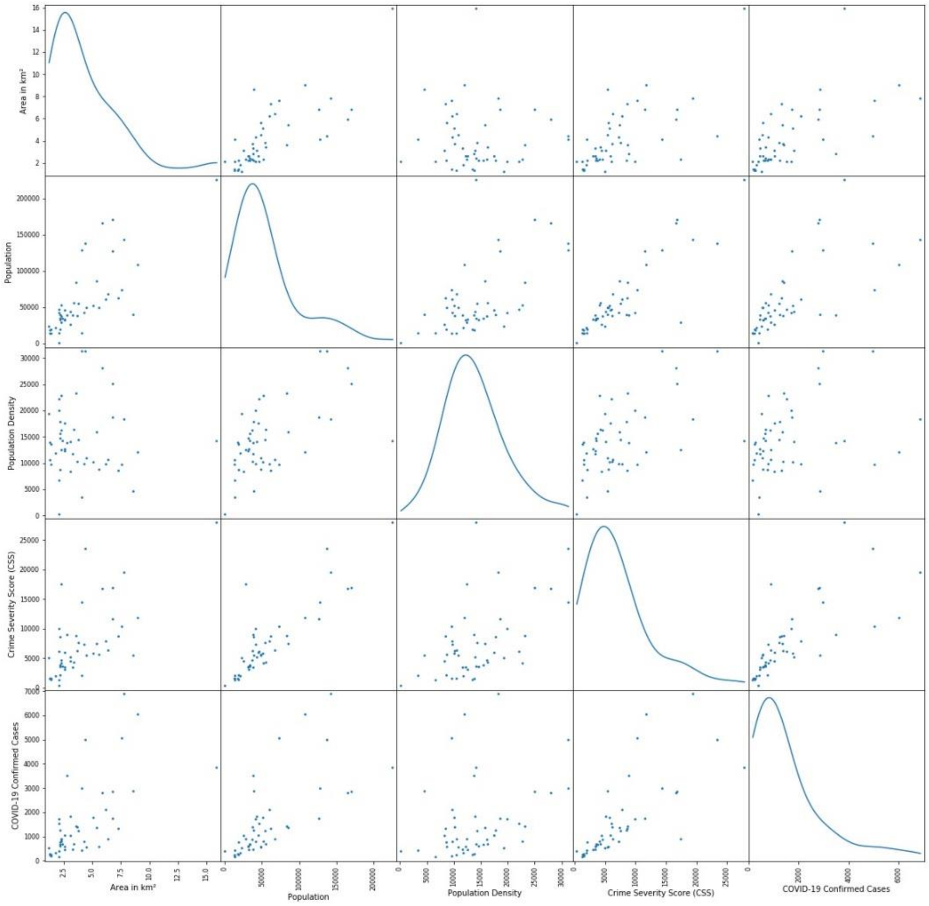
Overall, I take a two-step process of data understanding:

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

Overall, I would divide this section into the following two parts:

- data understanding about the neighbourhoods' basic profile: the size of area, population, population density, crime severity scores, and COVID-19 confirmed cases.
- data understanding about Foursquare data

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



Diagonal cells of the scatter matrix illustrate the distributions of the corresponding variable. All the plots in the diagonal cells are highly skewed, thus, suggest the presence of outliers.

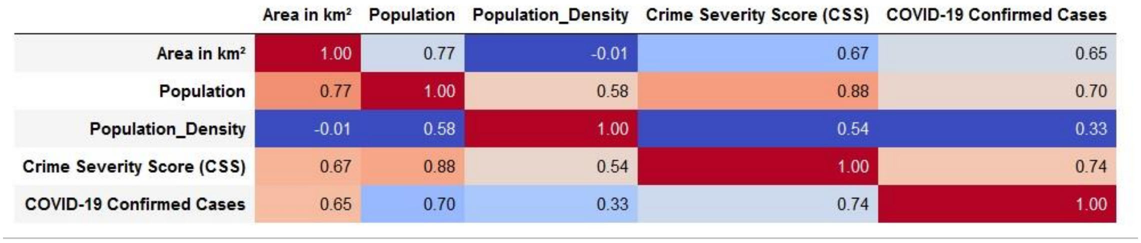
The distributions on the first three diagonal cells are related with the basic profile features of neighbourhoods: area, population, and population density.

- They are all skewed. However, the population density demonstrates the least skewed distribution, if not symmetric one.

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

- These risk measures 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

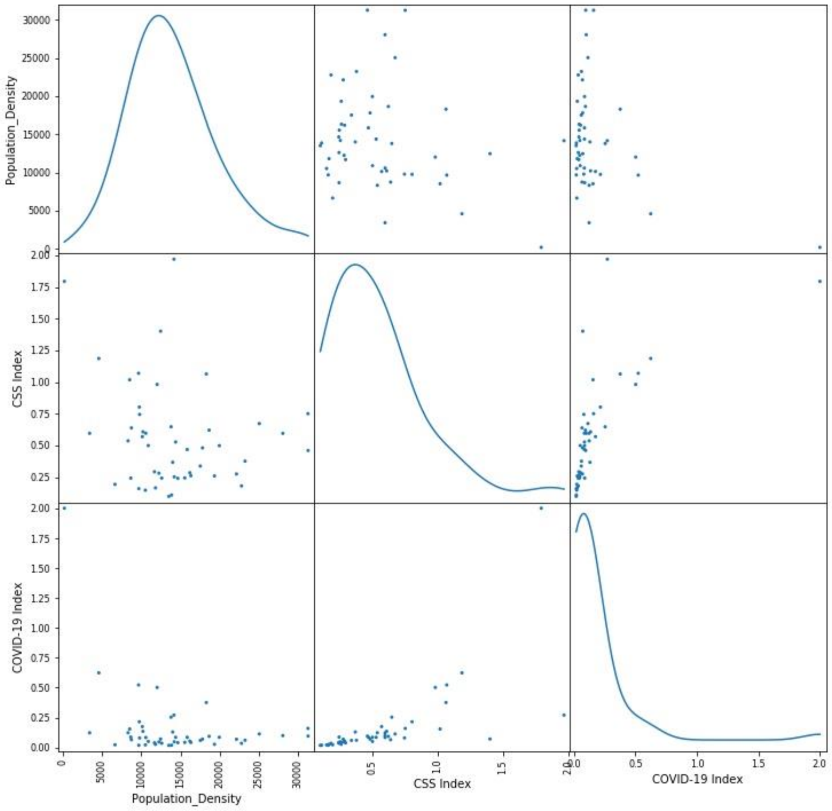
- Population density demonstrates the lowest correlation with other variables.
- Nevertheless, as expected, population density demonstrates positive correlation with the general insecurity risk metric and the COVID-19 risk metric—“Crime Severity Score” (0.54) and “COVID-19 Confirmed Cases” (0.33)—suggesting 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.

Overall, “population density” stands out in the sense that it demonstrates relatively lower correlation with other data. This would raise a question: in order to scale these two risk-metrics—'Crime Severity Score (CSS)’ and ‘COVID-19 Confirmed Cases’—is ‘population density’ the best scaler? For example, population demonstrates the highest correlation with these two risk-metrics. This question needs to be reserved for a suggestion for prospective research.

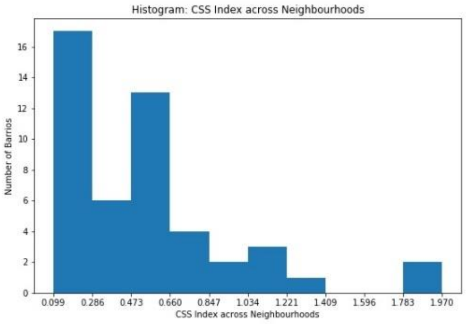
Nevertheless, as per the Client’s general request to scale the risk metrics by population density for this project, I scale these two-risk metrics with population density, by simply dividing the two risk-metrics by population density. As result, we have ‘CSS Index’ and ‘COVID-19 Index’.

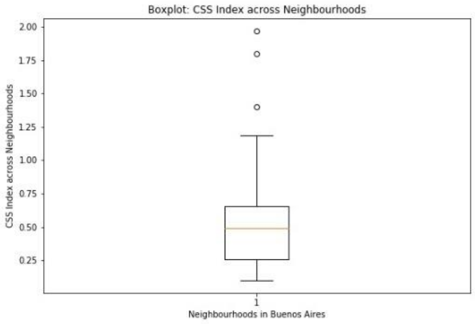
- CSS Index = Crime Severity Scores / Population Density
- COVID-19 Index = COVID-19 Confirmed Case/ Population Density

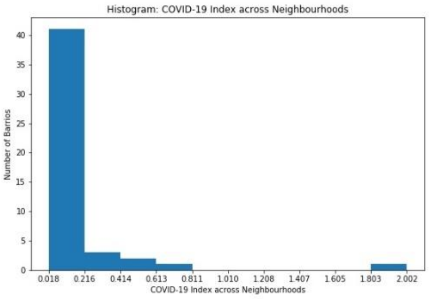
Below, I plot the scatter matrix among these two new indices and the population density. The new scatter matrix also indicates the presence of outliers in these two new indices.

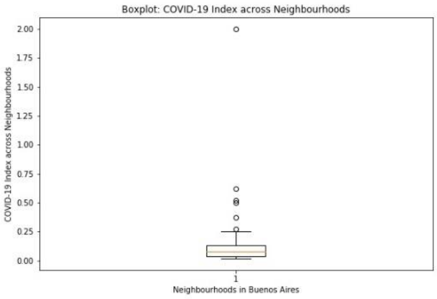


Next, in order to observe the distributions of these two new indices more individually, I plot histograms and boxplots for them. Visualization can be very powerful. Often, it visually reveals some salient characteristics of the data.

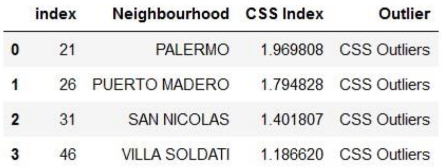




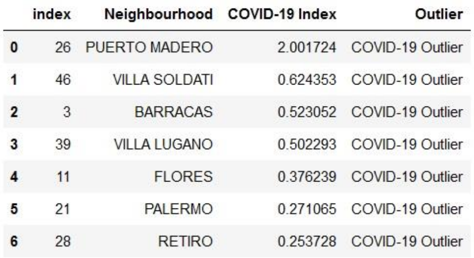




For our data, especially, each boxplot displays outliers beyond its upper whisker. Here is the list of those datapoints beyond the upper whiskers of these two box plots.

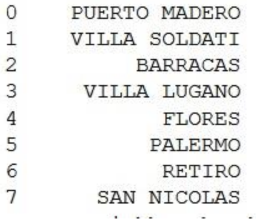


	index	Neighbourhood	CSS Index	Outlier
0	21	PALERMO	1.969808	CSS Outliers
1	26	PUERTO MADERO	1.794828	CSS Outliers
2	31	SAN NICOLAS	1.401807	CSS Outliers
3	46	VILLA SOLDATI	1.186620	CSS Outliers



	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 outlier neighbourhoods between these two lists. Consolidating them, here is the list of overall risk outliers. There are 8 overall risk outliers.

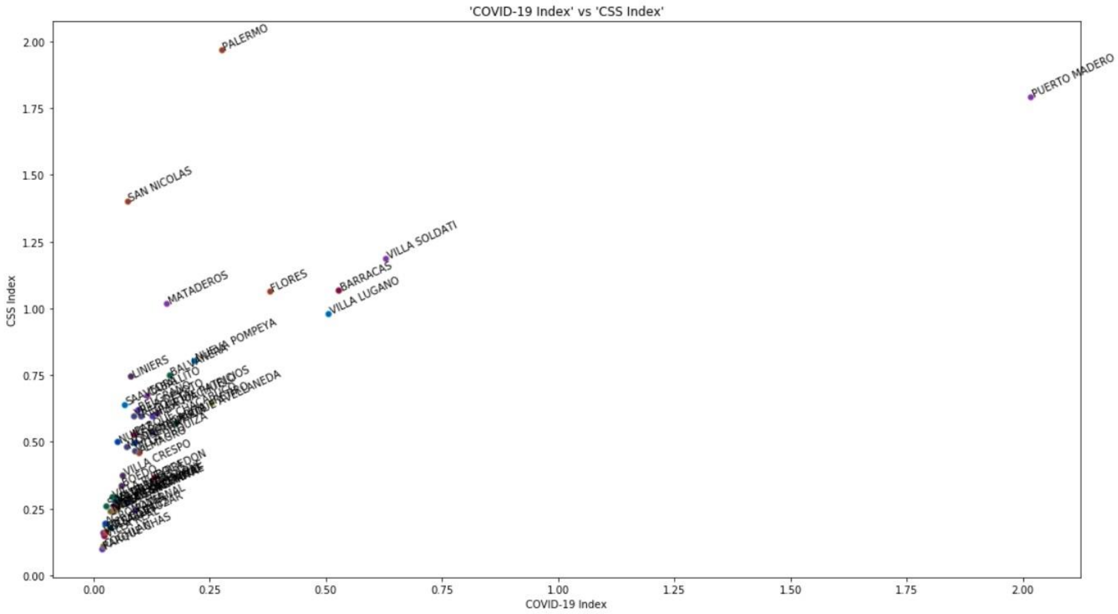


```
0    PUERTO MADERO
1    VILLA SOLDATI
2         BARRACAS
3     VILLA LUGANO
4           FLORES
5          PALERMO
6           RETIRO
7      SAN NICOLAS
```

This is an observation based on the boxplot visualization at the stage of Data Understanding. In the further analysis, as per the Client’s request, I would use a machine learning model, specifically Clustering Machine Learning model, to identify outliers separately and see if the observation above is consistent with the machine learning result.

At last, I would like to present the scatter plot of the neighbourhoods as data points over the two-dimensional risk space: ‘CSS Index’ and ‘COVID-19 Index’.

```
s : 48
```



This visualization is powerful enough to convince us of the presence of outliers.

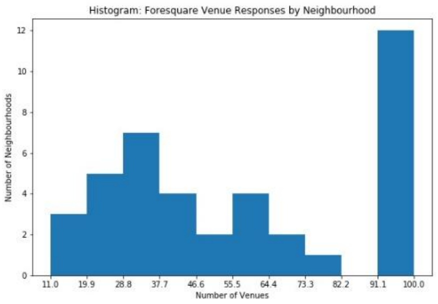
[6] Data Understanding Part 2: Foursquare Data

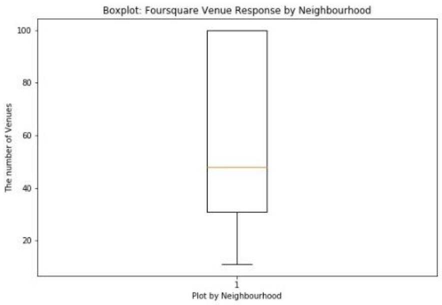
Foursquare API allows the user to explore venues within a user specified radius from one single location point. In other words, the user needs to specify the following parameters:

- 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.

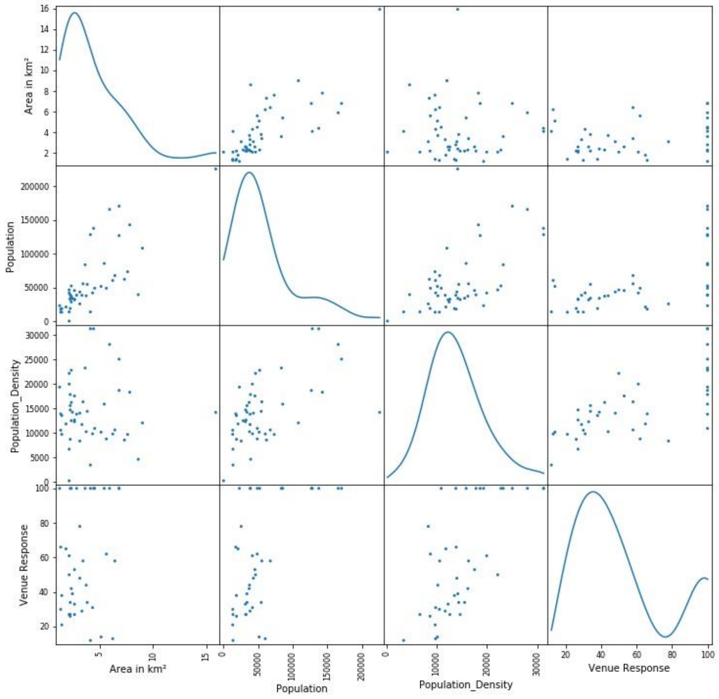
Under this constraint, I set my query with radius=1000 to explore venues within the radius of 1 km from the centre of each neighbourhood. And here is the summary of the Foursquare response to my query: the histogram and the boxplot of the number of venues across different neighbourhoods.

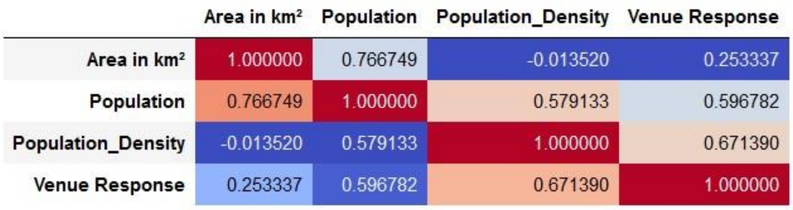




These two plots display the distribution of the Foursquare sample responses. It might suggest that there might be some issues in the coherency of data quality and availability across different neighbourhoods. If that is the case, this might affect the quality of the result of clustering machine learning.

Just in case, I would like to see if there is any relationship between the Foursquare's response and the three basic profiles of neighbourhoods. I generated the correlation matrix and the scatter matrix.





	Area in km^2	Population	Population_Density	Venue Response
Area in km^2	1.000000	0.766749	-0.013520	0.253337
Population	0.766749	1.000000	0.579133	0.596782
Population_Density	-0.013520	0.579133	1.000000	0.671390
Venue Response	0.253337	0.596782	0.671390	1.000000

Here is an interesting outcome. Venue response has the highest correlation with population density and the least correlation with the area size of neighbourhoods. In other words, the scatter matrix and the correlation matrix suggest that the higher the population density, the more venue information Foursquare has for neighbourhoods. It appeals to our common sense in a way: densely populated busy neighbourhoods have more popular venues.

Methodology

Overview of Methodology

The three objectives set by the Client and the data availability determine the scope of methodology.

Once again, let me review the three objectives here.

1. Identify outlier high risk neighbourhoods (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/Clusters) and rank them based on a single quantitative risk metric (a compound risk metric of the general security risk and the pandemic risk).
3. Use Foursquare API to characterize the Non-Outlier Neighbourhoods regarding popular venues. And if possible, segment Non-Outlier Neighbourhoods according to popular venue profiles.

Now, there appears one common salient feature among these three objectives. We have no 'a priori knowledge' about the underlying cluster structure of any of the subjects: outlier neighbourhoods, non-outlier neighbourhoods, and popular venue profiles among non-outlier neighbourhoods. All these three objectives demand us to discover hidden labels, or unknown underlying cluster structures in the data. Simply put, we have no labelled data to train: we have no empirical data about the dependent variable, unlike supervised machine learning models.

This constraint would naturally navigate us to the territory of unsupervised machine learning. Moreover, since all these objectives demand us to segment the given data into clusters, a natural choice would be 'Clustering Machine Learning'.

By its design—in the absence of the labelled data availability (empirical data for the dependent variable)—it would be difficult to automate the validation/evaluation process for an unsupervised machine learning, simply because there is no empirical label to compare the result of the model. According to Dr. Andrew Ng, there seems no widely accepted consensus about clear cut methods to assess the goodness of fit for clustering machine learning models. This creates an ample room for human insight, such as domain/business expertise, to get involved in the validation/evaluation process.

This creates an ample room for human insight, such as domain/business expertise, to get involved in the validation/evaluation process.

In this context, for this project, I will put more emphasis on tuning the model *a priori* rather than pursuing the automation of the *a posteriori* validation/evaluation process.

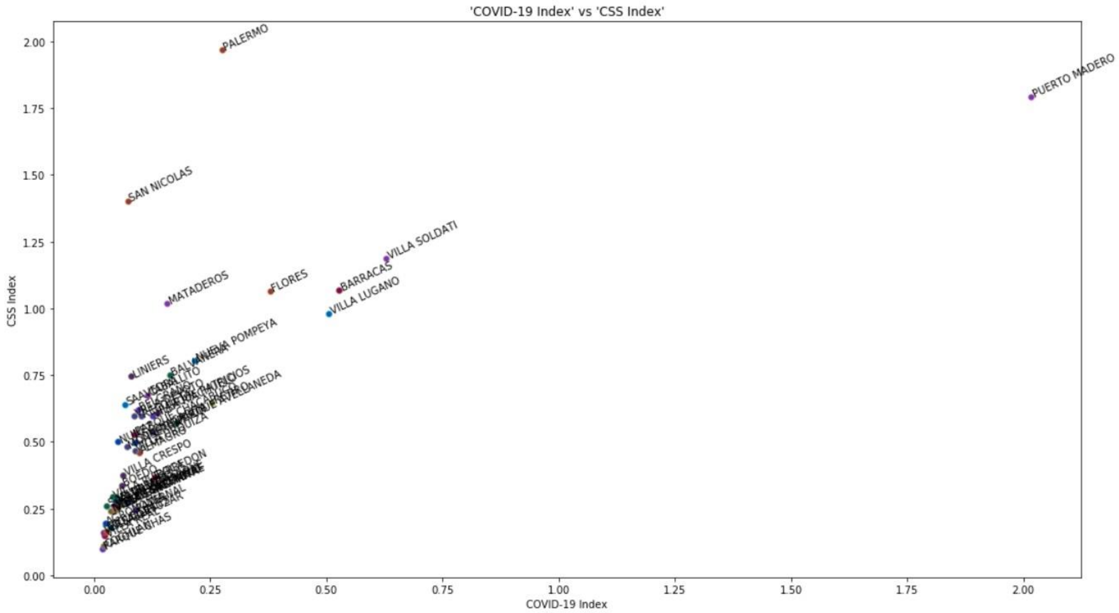
As one more important thing to mention in this overview section, we need to normalize/standardize all the relevant data before passing them to machine learning models.

Now, I will discuss the methodologies for each objective one by one.

[7] Methodology for Objective 1

The first objective is to identify 'Outlier Neighbourhoods'. Now, in the scatter plotting chart below, all the neighbourhoods are plotted in the two-dimensional risk space: 'CSS Index' vs 'COVID-19 Index' space.

```
s : 48
```



In order to identify outliers out of these "two-dimensional spatial data points", I chose ***DBSCAN Clustering model***, or ***Density-based Spatial Clustering of Applications with Noise***. As its name suggests, DBSCAN is a density-based clustering algorithm and deemed appropriate for ***examining spatial data***. Especially, I am very interested in how the density-based clustering algorithm would process outliers which are expected to demonstrate extremely sparse density.

There are several hyperparameters for DBSCAN. And the one considered as the most crucial is 'eps'. According to the Skit-learn.org website, '***eps***' is:

> *"the maximum distance between two samples for one to be considered as in the neighborhood of the other. This is not a maximum bound on the distances of points within a cluster. This is **the most important DBSCAN parameter** to choose appropriately for your data set and distance function."* (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html)

In order to tune 'eps', I will use ***KneeLocator*** of the python library ***kneed*** to identify ***the knee point*** (or elbow point).

What is ***the knee point***?

One way to interpret ***the knee point*** is that it is a point where the tuning result starts converging within a certain acceptable range. Simply put, it is a point where further tuning enhancement would no longer yield a material incremental benefit. In other words, ***the knee point*** determines a cost-benefit boundary for model hyperparameter tuning enhancement.(Source: https://ieeexplore.ieee.org/document/5961514)

[8] Methodology for Objective 2

Now, the objective 2 can be broken down into the following core sub-objectives:

1. Segmentation of 'Non-Outlier Neighbourhoods'.

2. Construction of a single compound risk metric to measure both the general security risk and the pandemic risk.

3. Measuring the cluster level risk.

(1) Given the product of the objective 1, I can remove “Outlier Neighbourhoods” from our dataset and focus only on “Non-Outlier Neighbourhoods” for the next clustering segmentation.

This time, I choose Hierarchical Clustering model. Here are the reasons why I select this particular model for the second objective:

- ➢ I have no advance knowledge how many underlying clusters are expected in the dataset. Although many clustering models, paradoxically, require the number of clusters as a hyperparameter input to tune the models *a priori*, Hierarchical Clustering does not require that.
- ➢ In addition, Hierarchical Clustering algorithm can generate a dendrogram that illustrates a tree-like cluster structure based on the hierarchical structure of the pairwise spatial distance distribution. The ‘dendrogram’ appeals to our human intuition in discovering the underlying cluster structure.

As the action plan, I would do the followings:

- Remove ‘Outlier Neighbourhoods’ to create a newly screened dataset of ‘Non-Outlier Neighbourhoods’.
- Normalize/Standardize the dataset.
- Generate a hierarchical dendrogram.
- Use human insight to discover the most optimal cluster structure in the dendrogram, thus, the corresponding number of the clusters.
- Segment the Non-Outlier Neighbourhoods into the selected number of clusters by hierarchical clustering method.
- Make the scatter plot to visualize the cluster structure on the two risk- metric space: the general insecurity risk and the pandemic risk.

(2) Construction of Compound Risk Metric

Furthermore, in order to achieve one of the Client’s requirement, I need to construct one single compound metric to summarise the two risk profiles of clusters (‘CSS’ and ‘COVID-19’ together).

For this purpose, I designed a compound risk metric as follows.

Compound Risk Metric =

$$[(\text{Standardized CSS Index} – \text{Standardized Origin of CSS Index})^2$$
$$+ (\text{Standardized COVID-19 Index} - \text{Standardized Origin of COVID-19 Index})^2\ ]^{1/2}$$

Although the formula might appear not straightforward, its basic intent is very simple: to measures the risk position of each neighbourhood from the risk-free point in the two-dimensional risk space. For the raw data, the risk-free point is at the origin of the two-risk-metrics space, which is (0,0): 0 represents no risk in the raw data. The formula above is measuring the risk position of data points from the risk-free point after the standardization/normalization transformation. It is because in order to pass the data into the machine learning model, the data needs to be normalized/standardized. In that sense, the formula above measures the distance between the standardized data points and the standardized risk-free position.

That’s all about the architecture of how to construct the compound risk metric.

(3) Risk Profile of Cluster

Now, my ultimate purpose here is to quantify the risk profile at cluster level, not at data point level (or not at a single neighbourhood level).

Each cluster has its own unique centre, aka “centroid”. Thus, in order to represent the risk profile of each cluster, I can use the centroid of each cluster. In this way, I can grade and rank all these clusters according to the compound risk metric of their centroids.

Accordingly, I measure the compound risk metric of the centroids of all these 5 Non-Outlier Clusters and assign each a grade.

[9] Methodology for Objective 3

For the third objective, I will carry out two analyses: Popular Venue Analysis; and Segmentation of Neighbourhoods based on Venue Composition.

a) Popular Venue Analysis:

I will apply One Hot Encoding algorithm to reorganize the data structure of venue category for further data transformation.
With Foursquare data, which is venue-base information, I will use Pandas’ “***groupby***” method to transform it to a neighbourhood-base data and summarise the top 5 popular venue categories for each of 40 ‘Non-Outlier Neighbourhoods’. The result is very long thus, presented in one of the final product at the end of this article.

b) Segmentation of Neighbourhoods based on Venue Composition

Next, I would explore the segmentation of the Foursquare venue data. For this purpose, I contemplate K-Means Clustering Machine Learning.

For a successful K-Means clustering result, I need to determine one of its hyperparameters, n_clusters: the number of clusters to form, thus, the number of centroids to generate. (source: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)

I will run two hyperparameter tuning methods—K-Means Elbow Method and Silhouette Score Analysis—to tune its most important hyperparameter, n_clusters, the number of clusters. These tuning methods would give me an insight about how to cluster the data for a meaningful analysis. Based on the finding from these tuning methods, I would decide how to implement the K-Means Clustering machine learning model.

‘K-Means Elbow Method’

The spirit of ‘K-Means Elbow Method’ is the same as the knee point method that I explained earlier. ***Elbow*** locates a point where further tuning enhancement would no longer yield a material incremental benefit. In other words, ***Elbow*** determines a cost-benefit boundary for model hyperparameter tuning enhancement.

Silhouette Coefficient Analysis

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Cut a long story short, the best value is 1, the worst -1.

- Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighbouring clusters. Which means, the sample is distinguished from the points belonging to other clusters.
- A value of 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters.
- Negative values, (-1,0), indicate that those samples might have been assigned to the wrong cluster.

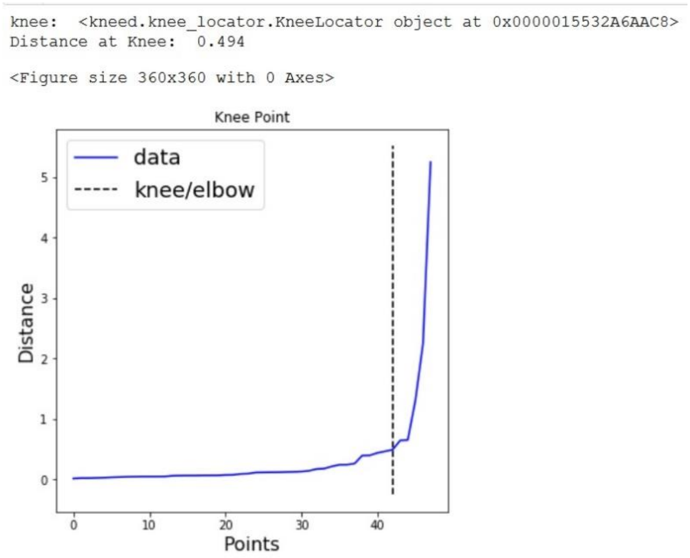
I will run the Silhouette Coefficient Analysis for 4 scenarios: **n_cluster** = [2, 3, 4, 5] to see which value of **n_cluster** yields the result closest to 1.

Result

[10] Result for DBSCAN Clustering for Objective 1:

In order to discover the best value for the model hyperparameter, ‘eps’, for DBSCAN model, I passed the normalized/standardized data for these two risk indices—namely ‘Standardized CSS Index’ and ‘Standardized COVID-19 Index’— into ***the KneeLocator***, which identifies the cost-benefit boundary, the knee point, of the model enhancement.

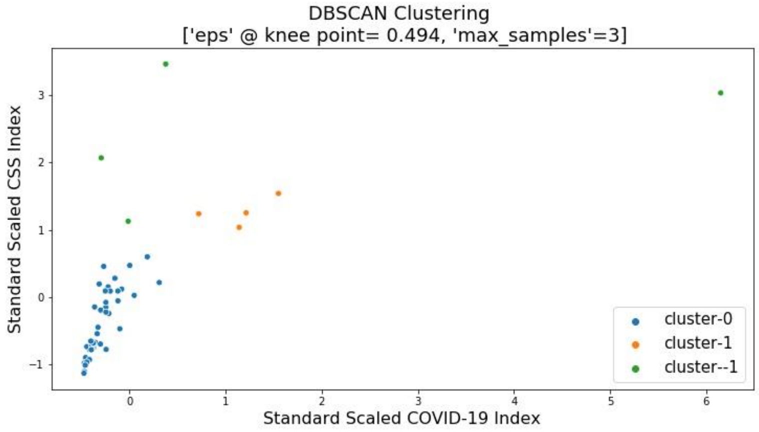
And here is the plot result:



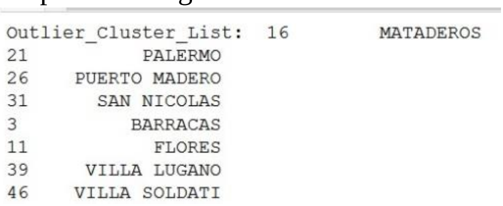
```
knee:  <kneed.knee_locator.KneeLocator object at 0x0000015532A6AAC8>
Distance at Knee:  0.494

<Figure size 360x360 with 0 Axes>
```

The crossing point between the distance curve and the dotted straight vertical line identifies the knee point. Above the chart, ***KneeLocator*** also returned the one single value, 0.494, as ***the knee point***. ***KneeLocator*** is telling me to choose this value as 'eps' to optimize the ***DBSCAN*** model. Accordingly, I plug it into ***DBSCAN***. And here is the result.

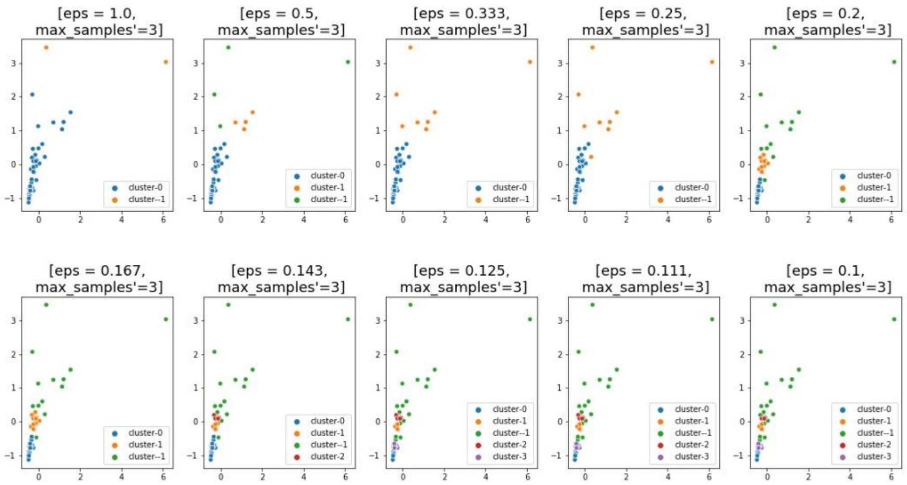


With this plot, I can confirm that DBSCAN distinguished the sparsely distributed outliers from others, yielding two clusters for the outliers: the cluster -1 (light green) and the cluster 1 (orange). Below, I listed up all the neighbourhoods of these two outlier clusters.



```
Outlier_Cluster_List:  16        MATADEROS
21          PALERMO
26    PUERTO MADERO
31      SAN NICOLAS
3          BARRACAS
11           FLORES
39     VILLA LUGANO
46    VILLA SOLDATI
```

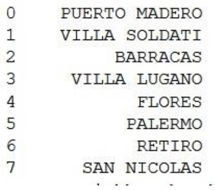
Furthermore, in order to assess if the result at ***the knee point*** is good or not, I run DBSCAN with other different values of 'eps'. Here is the result:



Compared with these various alternative values of 'eps', ***the knee point*** 'eps' gives us a reasonable result. The good thing is that ***KneeLocator*** automatically located the potentially optimal value for 'eps'. Thus, I will not reject ***the knee point,*** the output of ***KneeLocator,*** as the value for the hyperparameter, 'eps'.

When I look at the result of DBSCAN, I realise that this clustering result identifies the same neighbourhoods as outliers as the boxplot visualization performed during the Data Understanding stage.

For your reminder, here is the result of the boxplot once again.



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

The contents of these two results are identical except for the order of the list. What it tells us is that, especially when we are dealing with two dimensional datapoints, a simple boxplot can do the job of finding outliers with low cost.

For the case of our data, we could have avoided to perform a sophisticated, thus, expensive, model such as DBSCAN to identify outliers. In the perspective of cost-benefit management, the simple boxplot did the same job for the less cost. This might not be true when we have different data: especially, in a high-dimensional datapoints. At least, we should take this lesson in modesty so that we should not underestimate the power of simple methods like the boxplot visualisation.

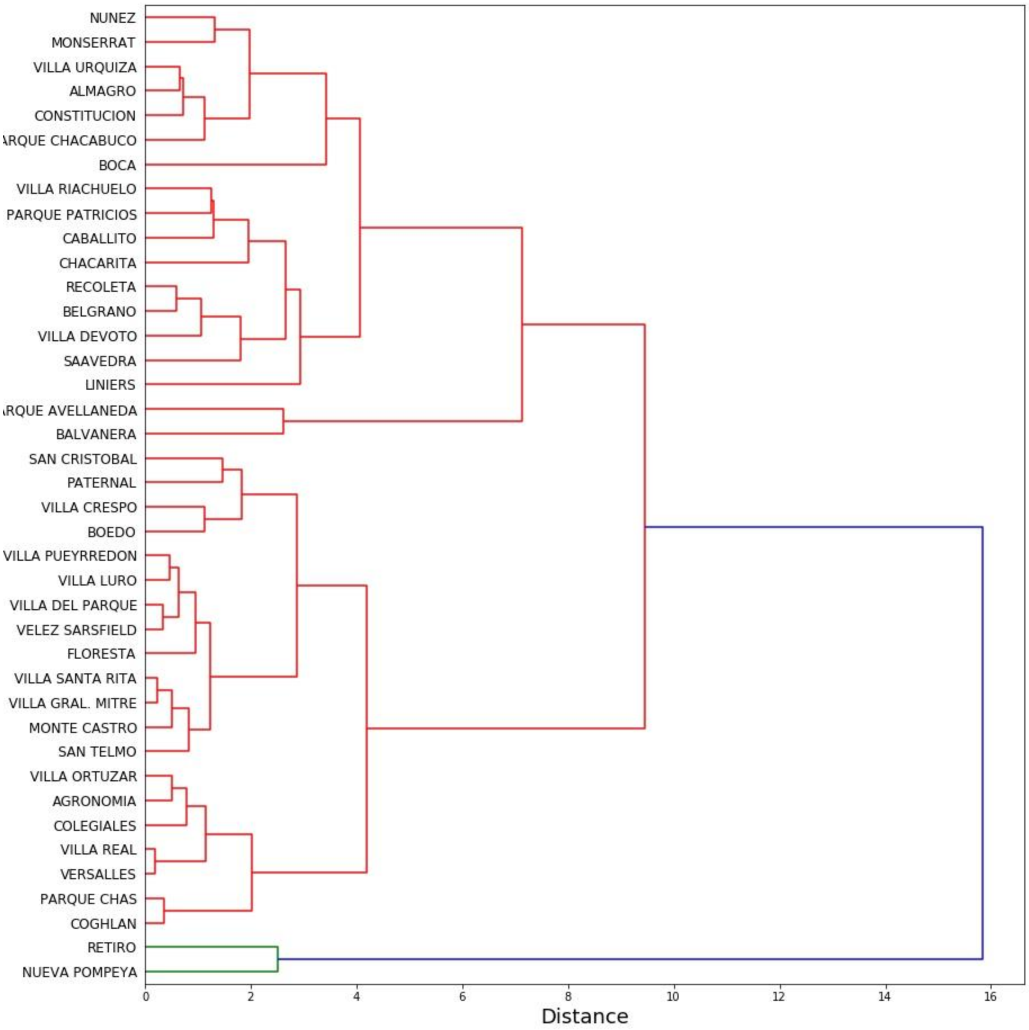
[11] Hierarchical Clustering for Objective 2:

(1) Hierarchical Clustering:

As aforementioned, Hierarchical Clustering does not require the user to set in advance the number of clusters to be discovered in order to generate its dendrogram. It is rather that the dendrogram allows the user to explore the hierarchical distance structure among datapoints and the underlying layers of cluster hierarchy.

The resulting dendrogram illustrates a tree-like cluster structure based on the pairwise distance distribution. In this way, the dendrogram allows the user to design how many clusters to be made for further analysis.

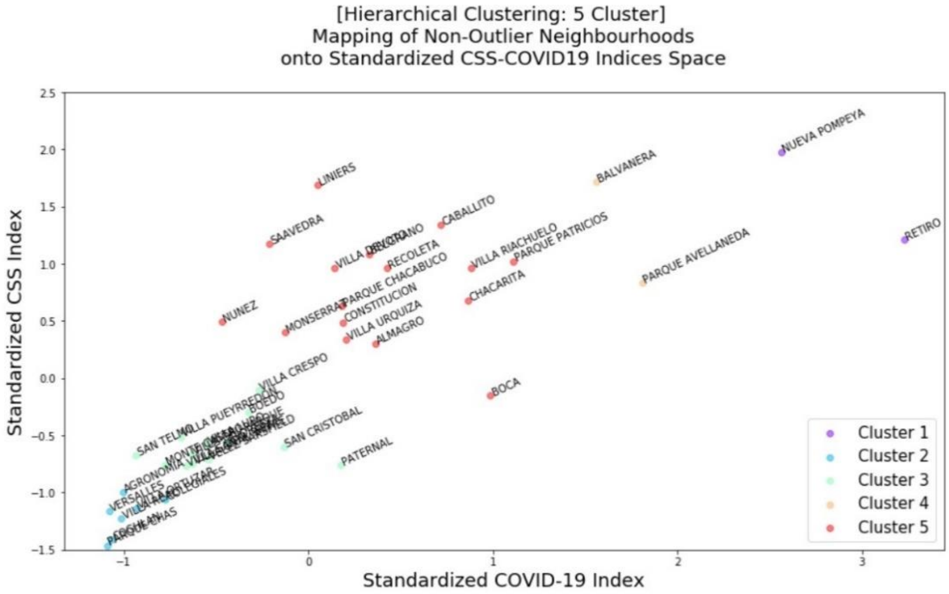
Here is the result of the hierarchical dendrogram:

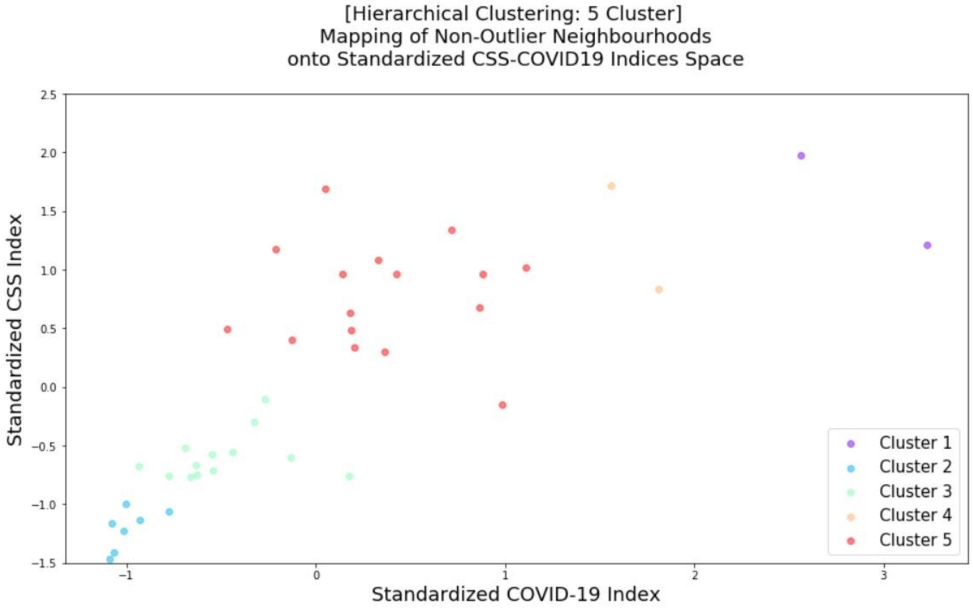


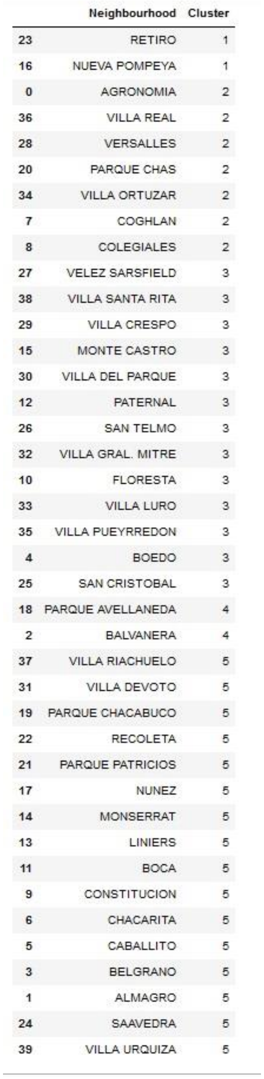
We can visually confirm the hierarchy of the distances among data points and the layers of cluster structure in the dendrogram. From this dendrogram, I choose 5 as the number of clusters: we can identify that at distance of 4 in the dendrogram on the x-axis. Then, I run Hierarchical Cluster Model, this time with the specification of the number of the clusters as 5, and got the 5 clusters of the neighbourhoods. Here are:

1. the visualization of the clustered neighbourhoods on the two risk-metrics space: I will plot one with neighbourhoods' names and the other without.

2. And the list of the clustering result:





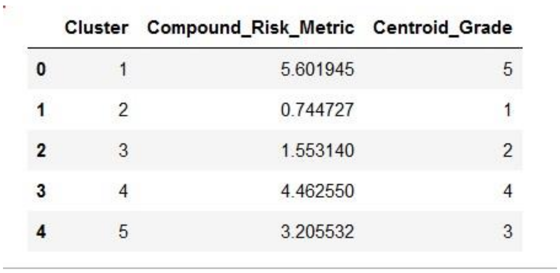


	Neighbourhood	Cluster
23	RETIRO	1
16	NUEVA POMPEYA	1
0	AGRONOMIA	2
36	VILLA REAL	2
28	VERSALLES	2
20	PARQUE CHAS	2
34	VILLA ORTUZAR	2
7	COGHLAN	2
8	COLEGIALES	2
27	VELEZ SARSFIELD	3
38	VILLA SANTA RITA	3
29	VILLA CRESPO	3
15	MONTE CASTRO	3
30	VILLA DEL PARQUE	3
12	PATERNAL	3
26	SAN TELMO	3
32	VILLA GRAL. MITRE	3
10	FLORESTA	3
33	VILLA LURO	3
35	VILLA PUEYRREDON	3
4	BOEDO	3
25	SAN CRISTOBAL	3
18	PARQUE AVELLANEDA	4
2	BALVANERA	4
37	VILLA RIACHUELO	5
31	VILLA DEVOTO	5
19	PARQUE CHACABUCO	5
22	RECOLETA	5
21	PARQUE PATRICIOS	5
17	NUNEZ	5
14	MONSERRAT	5
13	LINIERS	5
11	BOCA	5
9	CONSTITUCION	5
6	CHACARITA	5
5	CABALLITO	5
3	BELGRANO	5
1	ALMAGRO	5
24	SAAVEDRA	5
39	VILLA URQUIZA	5

As you see in the scatterplot, the clusters' ID in the legend does not really represent the risk scale of each cluster. Next, in order to compare these clusters along a risk scale, I will construct one single compound risk metric. For this purpose, I will use the risk profile of the centroid of each cluster.

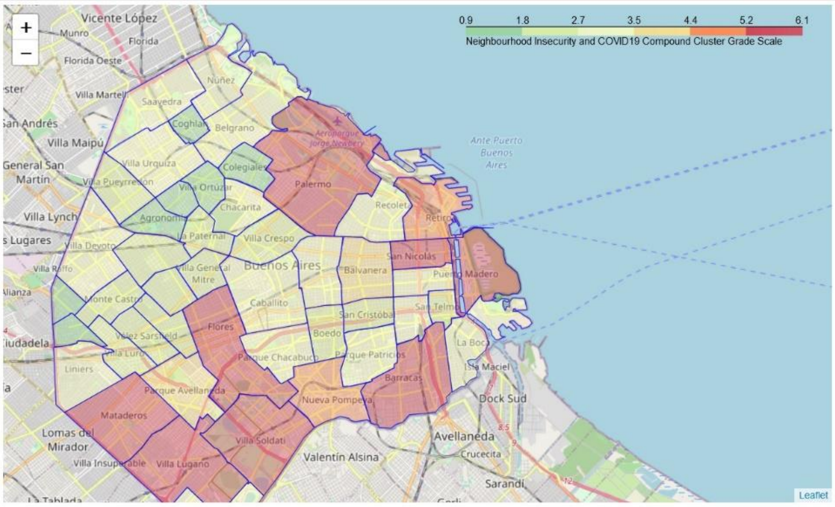
(2) Compound Risk Metric

As outlined in the Methodology section, in order to summarise two risk metrics—the general security risk (crime) and the pandemic risk (COVID-19)—for each cluster with one single risk metric, I constructed a compound risk metric. Each cluster has its own unique centre, aka "***centroid***". Thus, in order to represent the risk profile of each cluster, I will use the centroid of each cluster. In my code, I will calculate the compound risk metric of the centroid, '**Compound_Risk_Metric**', as the risk profile of each cluster. Based on '**Compound_Risk_Metric**', I assign a unique risk grading, '**Centroid_Grade**', to each cluster. Here is the result.



	Cluster	Compound_Risk_Metric	Centroid_Grade
0	1	5.601945	5
1	2	0.744727	1
2	3	1.553140	2
3	4	4.462550	4
4	5	3.205532	3

The higher the grade, the riskier the cluster. I merged this result with the original dataframe and assigned the cluster grade 6 to the 2 outlier clusters. Then, I mapped these cluster grades of all the neighbourhoods across the autonomous city of Buenos Aires in the following Choropleth Map.



This map visually summarises the findings for both of these first two objectives. It allows the user to visually distinguish neighbourhood clusters across the autonomous city of Buenos Aires based on their cluster risk grade.

[12] Foursquare Analysis for Objective 3:

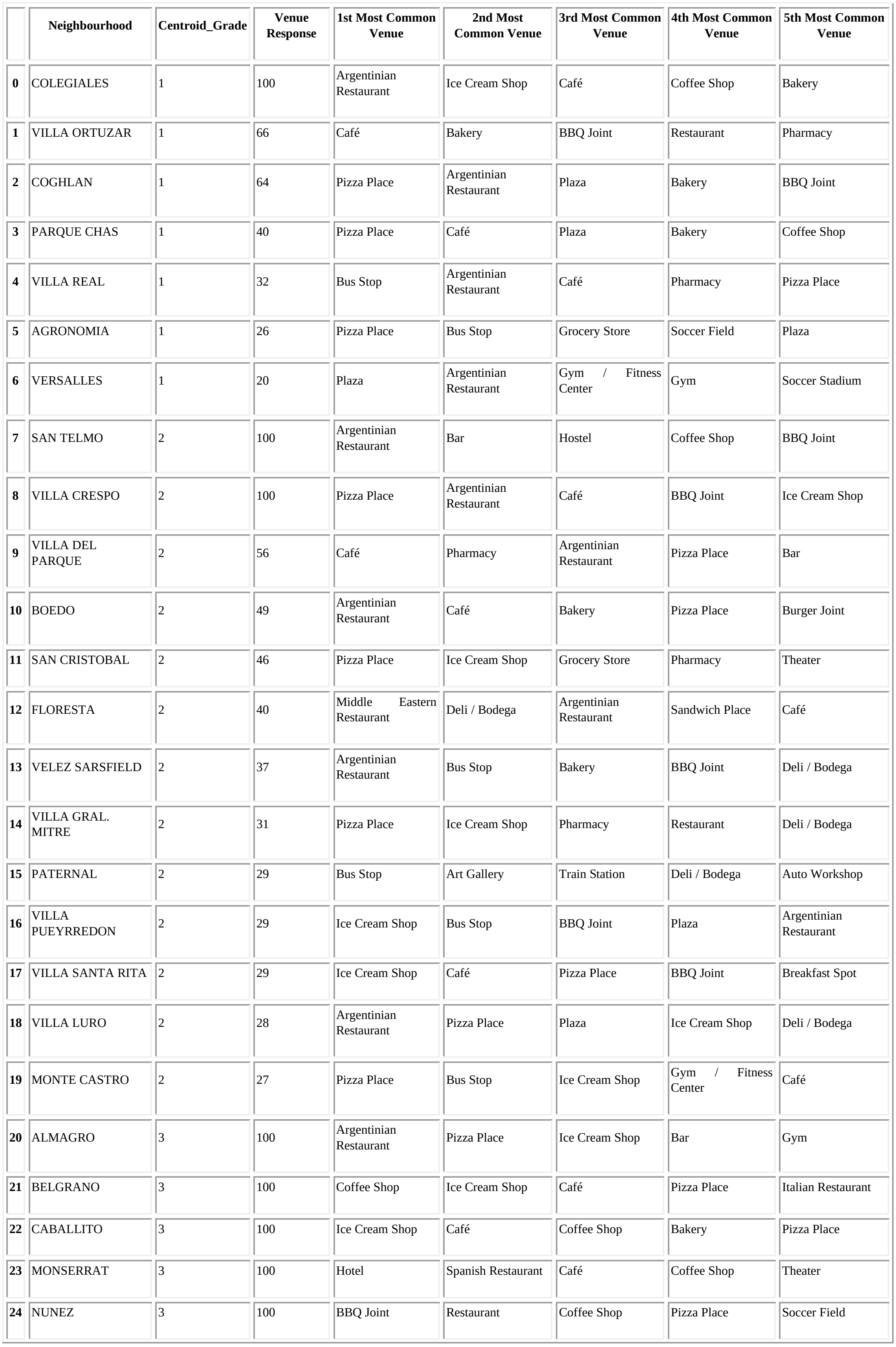
For the third objective, ***Foursquare analysis***, I carried out two analyses:

- Popular Venue Analysis, and
- Segmentation of Neighbourhoods based on Venue Composition.

c) Popular Venue Analysis:

This part of analysis requires a bit of data wrangling. I used the technique called ‘One Hot Encoding’ to organize venue information according to the venue categories. Further, I used Pandas’ ‘groupby’ method to summarise the resulting venue category base information by each neighbourhood. And as a presentation, I will list up top 5 popular venue categories for each neighbourhood.

And here is the result of the analysis:



	Neighbourhood	Centroid_Grade	Venue Response	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	COLEGIALES	1	100	Argentinian Restaurant	Ice Cream Shop	Café	Coffee Shop	Bakery
1	VILLA ORTUZAR	1	66	Café	Bakery	BBQ Joint	Restaurant	Pharmacy
2	COGHLAN	1	64	Pizza Place	Argentinian Restaurant	Plaza	Bakery	BBQ Joint
3	PARQUE CHAS	1	40	Pizza Place	Café	Plaza	Bakery	Coffee Shop
4	VILLA REAL	1	32	Bus Stop	Argentinian Restaurant	Café	Pharmacy	Pizza Place
5	AGRONOMIA	1	26	Pizza Place	Bus Stop	Grocery Store	Soccer Field	Plaza
6	VERSALLES	1	20	Plaza	Argentinian Restaurant	Gym / Fitness Center	Gym	Soccer Stadium
7	SAN TELMO	2	100	Argentinian Restaurant	Bar	Hostel	Coffee Shop	BBQ Joint
8	VILLA CRESPO	2	100	Pizza Place	Argentinian Restaurant	Café	BBQ Joint	Ice Cream Shop
9	VILLA DEL PARQUE	2	56	Café	Pharmacy	Argentinian Restaurant	Pizza Place	Bar
10	BOEDO	2	49	Argentinian Restaurant	Café	Bakery	Pizza Place	Burger Joint
11	SAN CRISTOBAL	2	46	Pizza Place	Ice Cream Shop	Grocery Store	Pharmacy	Theater
12	FLORESTA	2	40	Middle Eastern Restaurant	Deli / Bodega	Argentinian Restaurant	Sandwich Place	Café
13	VELEZ SARSFIELD	2	37	Argentinian Restaurant	Bus Stop	Bakery	BBQ Joint	Deli / Bodega
14	VILLA GRAL. MITRE	2	31	Pizza Place	Ice Cream Shop	Pharmacy	Restaurant	Deli / Bodega
15	PATERNAL	2	29	Bus Stop	Art Gallery	Train Station	Deli / Bodega	Auto Workshop
16	VILLA PUEYRREDON	2	29	Ice Cream Shop	Bus Stop	BBQ Joint	Plaza	Argentinian Restaurant
17	VILLA SANTA RITA	2	29	Ice Cream Shop	Café	Pizza Place	BBQ Joint	Breakfast Spot
18	VILLA LURO	2	28	Argentinian Restaurant	Pizza Place	Plaza	Ice Cream Shop	Deli / Bodega
19	MONTE CASTRO	2	27	Pizza Place	Bus Stop	Ice Cream Shop	Gym / Fitness Center	Café
20	ALMAGRO	3	100	Argentinian Restaurant	Pizza Place	Ice Cream Shop	Bar	Gym
21	BELGRANO	3	100	Coffee Shop	Ice Cream Shop	Café	Pizza Place	Italian Restaurant
22	CABALLITO	3	100	Ice Cream Shop	Café	Coffee Shop	Bakery	Pizza Place
23	MONSERRAT	3	100	Hotel	Spanish Restaurant	Café	Coffee Shop	Theater
24	NUNEZ	3	100	BBQ Joint	Restaurant	Coffee Shop	Pizza Place	Soccer Field

	Neighbourhood	Centroid_Grade	Venue Response	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
25	RECOLETA	3	100	Hotel	Plaza	Ice Cream Shop	Italian Restaurant	Argentinian Restaurant
26	VILLA URQUIZA	3	99	Pizza Place	Bakery	Ice Cream Shop	Coffee Shop	Grocery Store
27	CHACARITA	3	77	Restaurant	Café	Pizza Place	Bakery	Theater
28	CONSTITUCION	3	65	Argentinian Restaurant	Hotel	Pizza Place	Café	Restaurant
29	VILLA DEVOTO	3	64	Coffee Shop	Ice Cream Shop	Café	Beer Bar	Train Station
30	SAAVEDRA	3	59	Ice Cream Shop	Restaurant	Bus Stop	Bookstore	Grocery Store
31	BOCA	3	49	Argentinian Restaurant	Café	Art Gallery	BBQ Joint	Seafood Restaurant
32	PARQUE PATRICIOS	3	42	Ice Cream Shop	Bakery	Pizza Place	Gym	Argentinian Restaurant
33	PARQUE CHACABUCO	3	31	Ice Cream Shop	Pharmacy	Korean Restaurant	Pizza Place	Argentinian Restaurant
34	LINIERS	3	29	Fast Food Restaurant	Bus Stop	Pizza Place	Argentinian Restaurant	Sandwich Place
35	VILLA RIACHUELO	3	12	Racetrack	Bakery	Plaza	Pizza Place	Recreation Center
36	BALVANERA	4	100	Café	Gym	Theater	Japanese Restaurant	Ice Cream Shop
37	PARQUE AVELLANEDA	4	15	Pizza Place	Park	Ice Cream Shop	Athletics & Sports	Convenience Store
38	RETIRO	5	100	Hotel	Coffee Shop	Argentinian Restaurant	Café	Restaurant
39	NUEVA POMPEYA	5	13	Plaza	Pizza Place	Thrift / Vintage Store	Train Station	Soccer Field

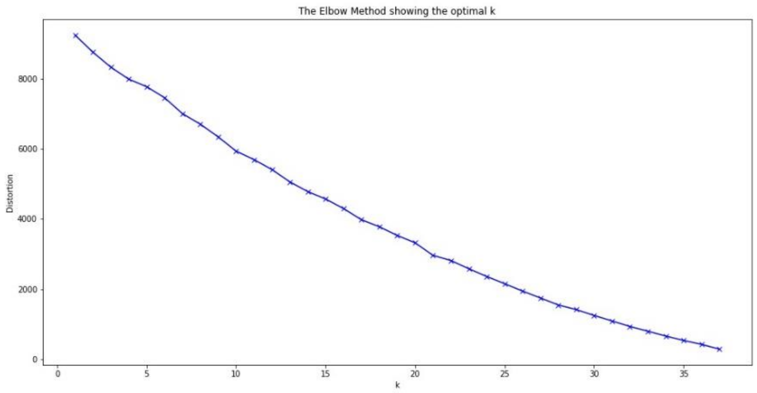
d) Segmentation of Neighbourhoods based on Venue Composition

In order to discover the optimal value for the model hyperparameter, ‘eps’, I run two model hyperparameter methods: ‘K-Means Elbow Method’ and ‘Silhouette Coefficient Analysis’.

a) ‘K-Means Elbow Method’

The spirit to use this method is as same as for the case of the knee point, which was explained earlier in the section of the first objective. This method also intends to discover the cost-benefit boundary of the model enhancement.

Here is the result of K-Means Elbow Method:



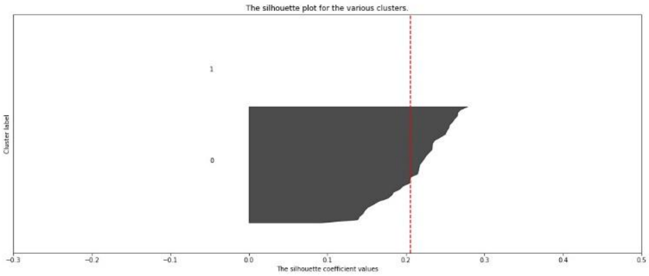
As the number of clusters increases, the response does not converge into any range; instead, it keeps dropping. There is no knee/elbow, the cost-benefit boundary, in the entire space. This suggests that there might be no meaningful cluster structure in the dataset. This is a disappointing result.

b) Silhouette Coefficient
Cut a long story short, Silhouette Coefficient has a range of [-1, 1] and its best value is 1, the worst -1.

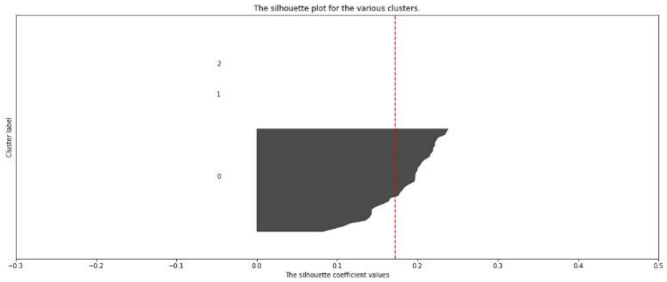
- Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighbouring clusters. Which means, the sample is distinguished from the points belonging to other clusters.
- A value of 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters.
- Negative values, (-1,0), indicate that those samples might have been assigned to the wrong cluster.

I ran the Silhouette Coefficient Analysis for 4 scenarios: **n_cluster** = [2, 3, 4, 5]. Unfortunately, all the results from these scenario settings turned out to be close to 0, indicating that the sample is on or very close to the decision boundary between two neighbouring clusters. And here are the results:

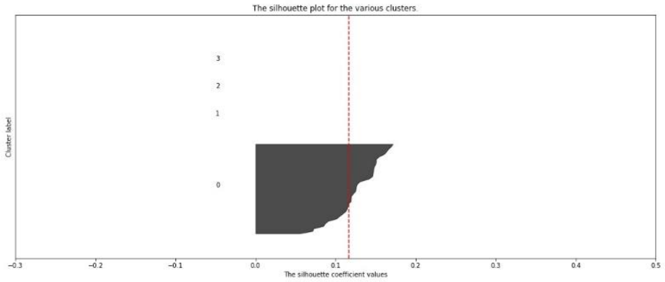
For **n_cluster** = 2:



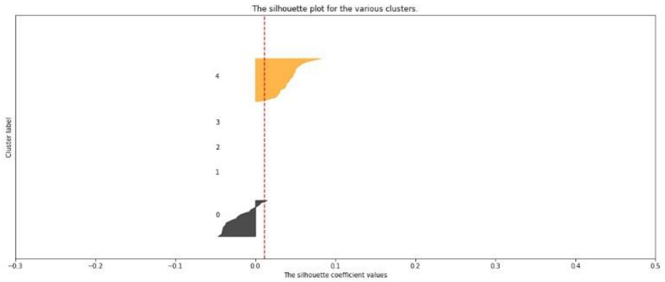
For **n_cluster** = 3:



For **n_cluster** = 4:



For **n_cluster** = 5:



This suggests that we cannot derive an inference about the underlying cluster structure in the data set. It might be due to the characteristics of the city. Or it could be due to the quality of available data.

Discussion

I explored three clustering machine learning models for three different objectives that the Client set. And, I came across with quite different results. I would like to present the implication of each case below.

Lesson from the first objective:

In the course of the data understanding—at an early stage of the project—basic visualizations by the scatter matrix and the correlation matrix illustrated highly skewed and positively correlated distributions among relevant data, suggesting the presence of outliers in the data. As a matter of fact, simple boxplots identified 8 outliers for the general security risk metric (Crime Severity Index) and the pandemic risk metric (COVID-19 Index), by plotting them above their upper whiskers. This validated the Client's concern about the potential presence of high-risk outlier neighbourhoods.

The additional clustering analysis with DBSCAN algorithm also identified the same 8 datapoints in two clusters remote from the concentration of the other data points. Simply put, the machine learning analysis only confirmed the earlier automatic identification of outliers made by the boxplot. As result, it was imperative to remove these outliers from the analysis.

Simply put, the boxplot has the whisker, its built-in criteria to automatically distinguish outliers from the rest. In contrast, the clustering model, although identified these 8 outliers as two remote clusters, did not provide a means to automatically determine them as outliers: although potentially some conditions might be added to enable it automatically identify the outliers.

So, the lesson from the first objective was that a sophisticated method is not necessarily superior to a simple method. We should take this lesson in modesty.

Lesson from the second objective

After removing the outlier neighbourhoods, I performed Hierarchical Clustering Model and generated its dendrogram on the dataset of the non-outlier neighbourhoods. For the size of 40 non-outlier neighbourhoods, Hierarchical Clustering model worked very efficiently. And the dendrogram arranged neighbourhoods accordingly to the pairwise distance hierarchy and made it easy for me to shape human insight to decide how to cluster the subject.

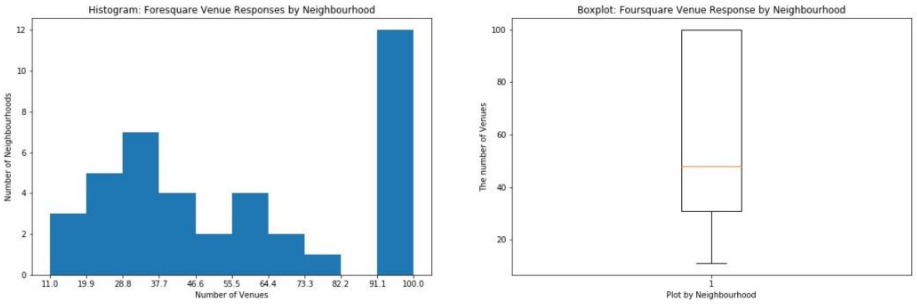
This presents a case that a machine learning model can play a great role in supporting human decision-making process. A user can leverage his/her profound domain expertise in the use of Hierarchical Clustering Dendrogram to effectively achieve the given objective. By interacting with a machine learning model in an organic way, the user can optimise the use of machine learning and make a better decision.

The Client can use their insight to discover their desirable clustering result based on the dendrogram and their human insight.

Lesson from the third objective

I performed two hyperparameter tuning methods to discover the best ***n_clusters***, the number of clusters, for K-Mean Clustering model. Unfortunately, neither of them yielded a convincing implication about the underlying cluster structure. This suggests that a clustering model would unlikely yield a reliable result for the given dataset.

The output of the machine learning is as good as the data input. The disappointing hyperparameter tuning result might have something to do with earlier concern about the data quality. Here, for your reminder, I present the distribution of the Foursquare responses to my ‘explore’ query within the radius of 1,000 m to the single representative location point for each neighbourhood.



Especially, the histogram on the left side demonstrated an awkward shape of the distribution of the Foursquare responses across the neighbourhoods in the city. This might suggest:

- The Foursquare might not have a coherent quality of data across different neighbourhoods sufficient to be a representative of their general profiles for this particular city.
- Or, otherwise, venues might be actually highly concentrated in those 12 neighbourhoods. And the Foursquare sample responses could be a good representation of the overall profile of the city (only about the Non-Outlier Neighbourhoods). Moreover, there could be actually no particular underlying cluster structure among the neighbourhoods in the autonomous city of Buenos Aires.

Which is correct? This question requires a comparative study with data from other sources and would be a good topic for the prospective research, but beyond the scope of this project.

Whatever the real reason might be, all I know from these tuning results is that there is no convincing implication regarding the underlying cluster structure in the given data. In order to avoid an unreliable and misleading recommendation, I would rather refrain from performing K-Means Clustering Model for the given dataset. And I would rather prepare summary materials so that the Client can use their human insight/domain expertise to explore them and identify the popular venue profiles of neighbourhoods.

Overall

Overall, the Client now has all the answers to the first two objective and the first part of the third objective. For the second part of the third objective, I would make the following recommendations:

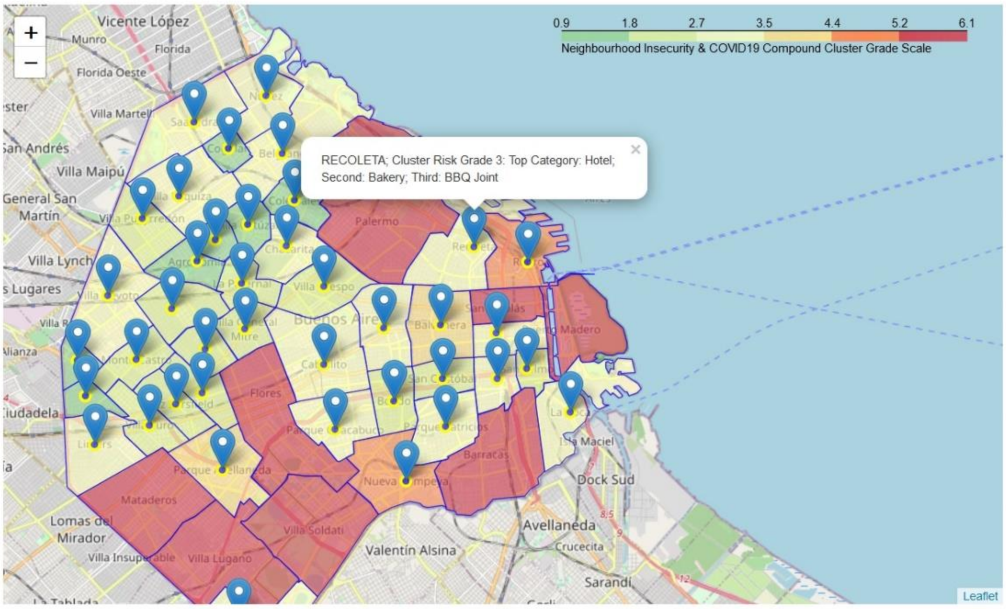
- For the prospective research, the Client can explore other data source, especially local ones, to
- For the immediate use, the Client can refer to the top popular venue profiles and use their business insight/domain expertise to characterise each neighbourhood’s popular venue characteristics.

Now, I would like to present the following summary materials that the Client can utilize for their ultimate decision-making: a map, two scatter plots, and a summary table.

Choropleth Map

In order to summarize the results for all these objectives, I added a pop-up feature to the choropleth map that I had created for the presentation of the objective 1 and 2 in order to display the following information of ‘Non-Outlier Neighbourhoods’.

- Name of the Neighbourhood
- Cluster Risk Grade, ‘Centroid_Grade’, to show the cluster risk profile of the neighbourhood.
- Top 3 Venue Categories

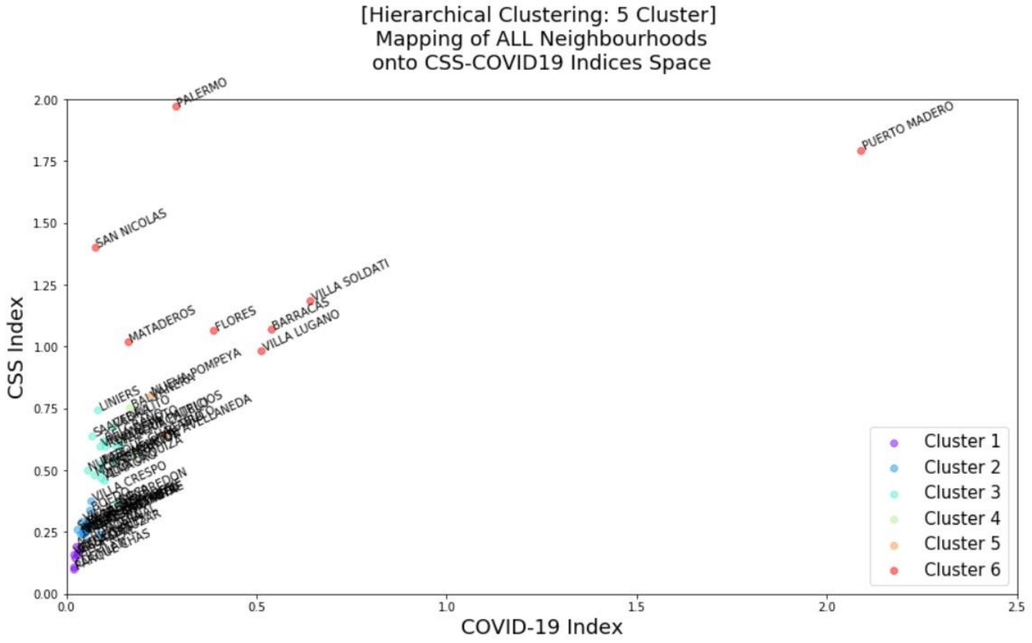

RECOLETA; Cluster Risk Grade 3: Top Category: Hotel; Second: Bakery; Third: BBQ Joint

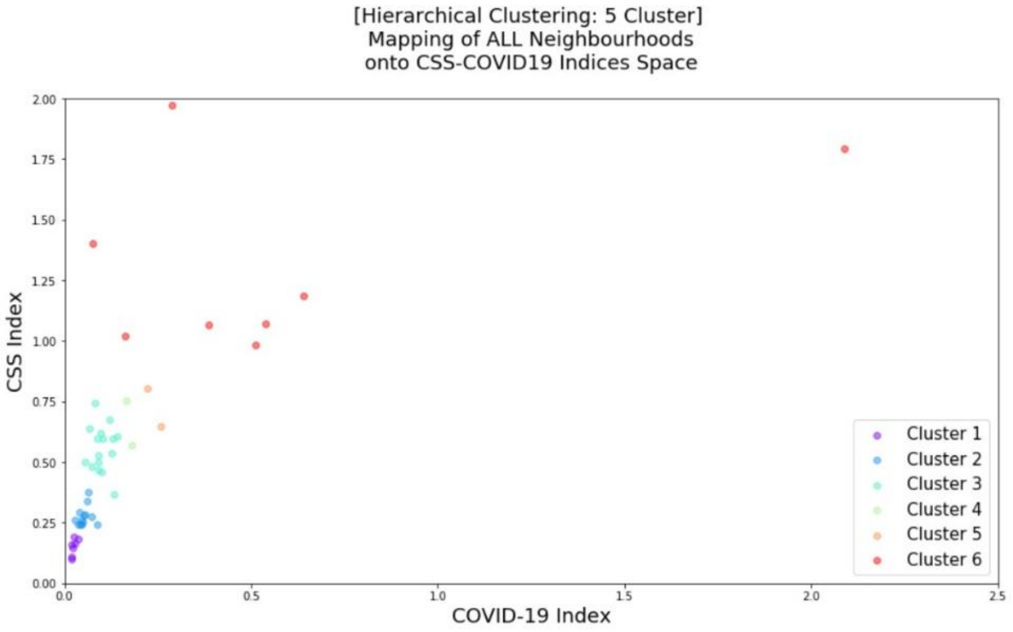
The map above displays an example of the pop-up feature.

For ‘Outlier Neighbourhoods’, I controlled the pop-up feature, since the Client wants to exclude them from consideration.

This map would compensate for the absence of the “venue category” base segmentation, allowing the Client to explore the popular venue profile for each neighbourhood individually.

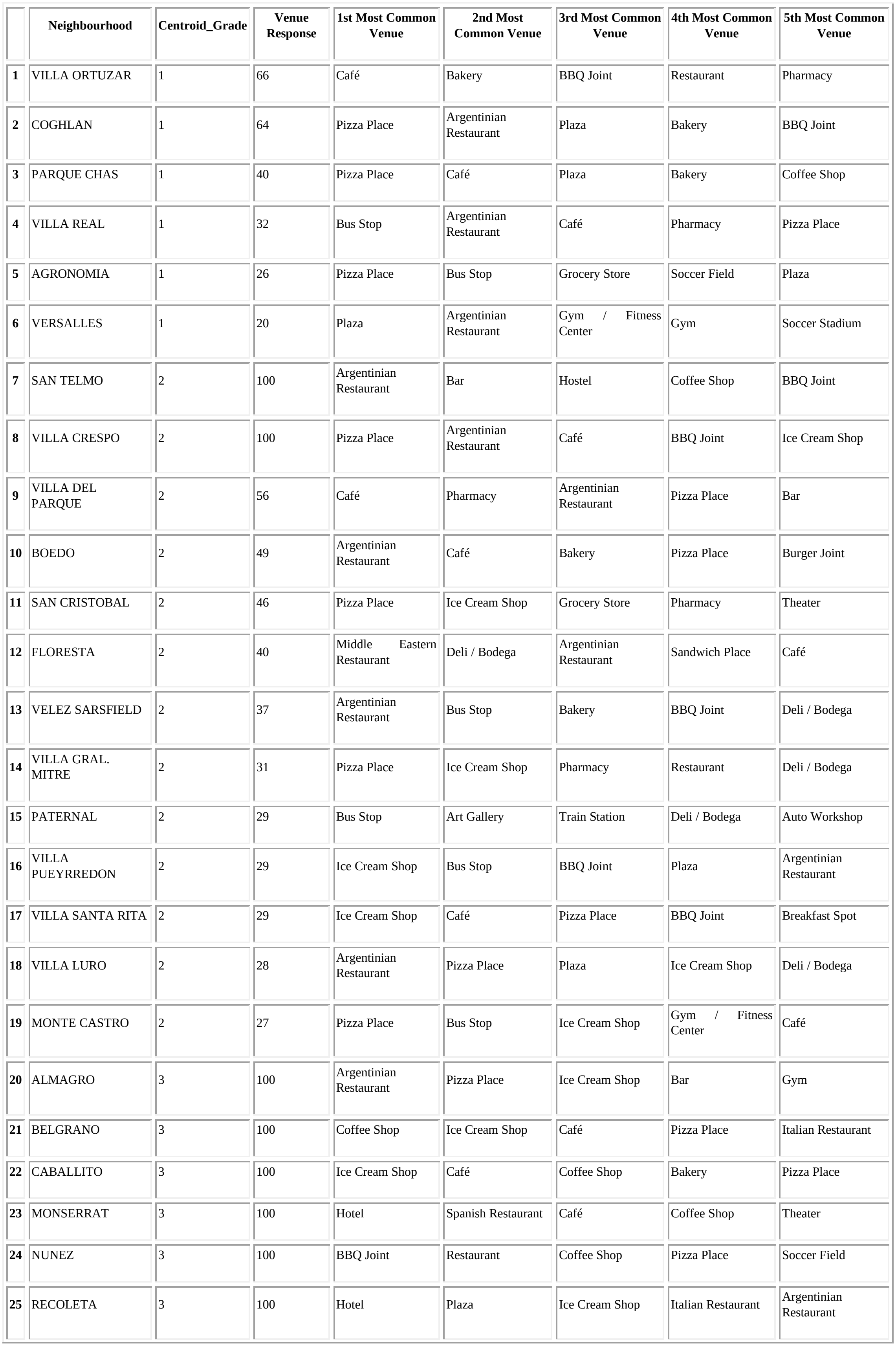
The following two scatter plots display the underlying cluster risk structure in the two-risk-metrics space. Although the first one with the names of neighbourhoods can be very helpful in identifying individual outliers, the densely plotted names at the left bottom are hiding the individual pots for ‘Non-Outlier Neighbourhoods’. The second scatter plot without the name is useful to make an overview of the entire cluster structure.





In addition, as a summary table, I also present a table of the top 5 most popular venue categories below, sorted by the cluster’s risk profile (in ascending order of Centroid_Grade) and the number of Foursquare venue response (in descending order of Venue Response. In this sorted order, the Client can view the list of neighbourhoods in an organized order: from the safest cluster to more dangerous clusters; from presumably the most popular venue neighbourhood to the less)

	Neighbourhood	Centroid_Grade	Venue Response	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	COLEGIALES	1	100	Argentinian Restaurant	Ice Cream Shop	Café	Coffee Shop	Bakery



	Neighbourhood	Centroid_Grade	Venue Response	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	VILLA ORTUZAR	1	66	Café	Bakery	BBQ Joint	Restaurant	Pharmacy
2	COGHLAN	1	64	Pizza Place	Argentinian Restaurant	Plaza	Bakery	BBQ Joint
3	PARQUE CHAS	1	40	Pizza Place	Café	Plaza	Bakery	Coffee Shop
4	VILLA REAL	1	32	Bus Stop	Argentinian Restaurant	Café	Pharmacy	Pizza Place
5	AGRONOMIA	1	26	Pizza Place	Bus Stop	Grocery Store	Soccer Field	Plaza
6	VERSALLES	1	20	Plaza	Argentinian Restaurant	Gym / Fitness Center	Gym	Soccer Stadium
7	SAN TELMO	2	100	Argentinian Restaurant	Bar	Hostel	Coffee Shop	BBQ Joint
8	VILLA CRESPO	2	100	Pizza Place	Argentinian Restaurant	Café	BBQ Joint	Ice Cream Shop
9	VILLA DEL PARQUE	2	56	Café	Pharmacy	Argentinian Restaurant	Pizza Place	Bar
10	BOEDO	2	49	Argentinian Restaurant	Café	Bakery	Pizza Place	Burger Joint
11	SAN CRISTOBAL	2	46	Pizza Place	Ice Cream Shop	Grocery Store	Pharmacy	Theater
12	FLORESTA	2	40	Middle Eastern Restaurant	Deli / Bodega	Argentinian Restaurant	Sandwich Place	Café
13	VELEZ SARSFIELD	2	37	Argentinian Restaurant	Bus Stop	Bakery	BBQ Joint	Deli / Bodega
14	VILLA GRAL. MITRE	2	31	Pizza Place	Ice Cream Shop	Pharmacy	Restaurant	Deli / Bodega
15	PATERNAL	2	29	Bus Stop	Art Gallery	Train Station	Deli / Bodega	Auto Workshop
16	VILLA PUEYRREDON	2	29	Ice Cream Shop	Bus Stop	BBQ Joint	Plaza	Argentinian Restaurant
17	VILLA SANTA RITA	2	29	Ice Cream Shop	Café	Pizza Place	BBQ Joint	Breakfast Spot
18	VILLA LURO	2	28	Argentinian Restaurant	Pizza Place	Plaza	Ice Cream Shop	Deli / Bodega
19	MONTE CASTRO	2	27	Pizza Place	Bus Stop	Ice Cream Shop	Gym / Fitness Center	Café
20	ALMAGRO	3	100	Argentinian Restaurant	Pizza Place	Ice Cream Shop	Bar	Gym
21	BELGRANO	3	100	Coffee Shop	Ice Cream Shop	Café	Pizza Place	Italian Restaurant
22	CABALLITO	3	100	Ice Cream Shop	Café	Coffee Shop	Bakery	Pizza Place
23	MONSERRAT	3	100	Hotel	Spanish Restaurant	Café	Coffee Shop	Theater
24	NUNEZ	3	100	BBQ Joint	Restaurant	Coffee Shop	Pizza Place	Soccer Field
25	RECOLETA	3	100	Hotel	Plaza	Ice Cream Shop	Italian Restaurant	Argentinian Restaurant

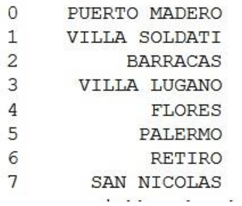
	Neighbourhood	Centroid_Grade	Venue Response	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
26	VILLA URQUIZA	3	99	Pizza Place	Bakery	Ice Cream Shop	Coffee Shop	Grocery Store
27	CHACARITA	3	77	Restaurant	Café	Pizza Place	Bakery	Theater
28	CONSTITUCION	3	65	Argentinian Restaurant	Hotel	Pizza Place	Café	Restaurant
29	VILLA DEVOTO	3	64	Coffee Shop	Ice Cream Shop	Café	Beer Bar	Train Station
30	SAAVEDRA	3	59	Ice Cream Shop	Restaurant	Bus Stop	Bookstore	Grocery Store
31	BOCA	3	49	Argentinian Restaurant	Café	Art Gallery	BBQ Joint	Seafood Restaurant
32	PARQUE PATRICIOS	3	42	Ice Cream Shop	Bakery	Pizza Place	Gym	Argentinian Restaurant
33	PARQUE CHACABUCO	3	31	Ice Cream Shop	Pharmacy	Korean Restaurant	Pizza Place	Argentinian Restaurant
34	LINIERS	3	29	Fast Food Restaurant	Bus Stop	Pizza Place	Argentinian Restaurant	Sandwich Place
35	VILLA RIACHUELO	3	12	Racetrack	Bakery	Plaza	Pizza Place	Recreation Center
36	BALVANERA	4	100	Café	Gym	Theater	Japanese Restaurant	Ice Cream Shop
37	PARQUE AVELLANEDA	4	15	Pizza Place	Park	Ice Cream Shop	Athletics & Sports	Convenience Store
38	RETIRO	5	100	Hotel	Coffee Shop	Argentinian Restaurant	Café	Restaurant
39	NUEVA POMPEYA	5	13	Plaza	Pizza Place	Thrift / Vintage Store	Train Station	Soccer Field

Conclusion

The scope of this project was well defined by the three objectives set by the Client: presented in the ***Introduction*** section.

During the data understanding—at an early stage of the project—basic visualizations by the scatter matrix and the correlation matrix suggested highly skewed and positively correlated distributions among relevant data. That suggested the presence of outliers in the data. As a matter of fact, simple boxplots identified outliers for the general security risk metric (Crime Severity Index) and the pandemic risk metric (COVID-19 Index).

This validated the Client’s concern about the potential presence of high-risk outlier neighbourhoods. The following outliers were identified.



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

The Client demonstrated their strong enthusiasm to run machine learning model to achieve their objectives. In response to the Client's enthusiasm toward Machine Learning, I run three different clustering machine learning models for three different objectives.

In hindsight, all of these three attempts yielded totally different implications about machine learning analysis. All implications from these attempts are valuable. And these implications are discussed earlier in ***Discussion*** section of this report.

Overall, the result of the analysis yielded the following mixed picture.

First, DBSCAN Clustering and Hierarchical Clustering produced successful results to achieve the first two objectives and the first part of the third objective.

Unfortunately, for the second part of the third objective—to segment non-outlier neighbourhoods into clusters based on the popular venue profile—I could not derive any convincing inference regarding the underlying cluster structure among non-outlier neighbourhoods. Rather than deriving an unreliable, and potentially misleading, inference, I judged that it would be wise to refrain from running the machine learning model for this particular case. I would recommend the Client to utilize these summary materials provided in ***Discussion*** section to apply their business insights/domain expertise to analyse the popular venue profile of neighbourhoods.

Moreover, for the future potential improvement, I would suggest that the Client might want to explore other sources than Foursquare to examine the popular venue profiles of these neighbourhoods. That would allow the Client to assess by comparison if the Foursquare data is representative of the actual state of popular venues in the particular city.

In addition, as a suggestion, the Client might benefit from conducting a similar analysis based on per-capita data. In this project the risk metrics were scaled by population-density. A per-capita base scaling might yield a different picture about the risk profile of the neighbourhoods.

At last, for the prospective development, I would recommend that the Client should incorporate the following two lessons from this project into their analytical project management policy building.

1. When a basic tool can achieve the intended objective, it would be cost-effective to embrace the basic tool to derive a conclusion/inference, rather than blindly implement an advanced tool, such as machine learning. Here, by basic tools, I am contemplating basic visualizations and basic exploratory statistics, such as correlation and regressions.

2. When the given data is not suitable for the design of a machine learning model, it might be unproductive to run it. If that is the case, there would be no point in wasting the precious resource to end up yielding a potentially misleading result.

Due to the hype for Machine Learning among the public, some clients demonstrate some blind craving for it, assuming that such an advanced method would yield a superior result than basic tools. Nevertheless, this project yielded a mixed set of answers of both 'yes' and 'no'. Especially since the Client demonstrated an exceptional enthusiasm towards Machine Learning for their future business decision making, I believe that it would be worthwhile reflecting these lessons for their future productive conduct of data analysis.

Thank you very much