Three Different Lessons from Three Different Clustering Analyses

Map Risk Clusters of Neighbourhoods in the time of Pandemic

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, in order to craft the settings for my capstone project, I contemplated a hypothetical corporate client in the food industry (catering business) from abroad (The Client), that is planning to relocate their representative family to the city of Buenos Aires (CABA) for their future entry 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 plan. In addition, the Client wants to capture the characteristics of neighbourhoods based on popular commercial venue categories such as restaurants, shops, and sports facilities. In this context, the Client hired me as an independent data analyst to conduct a preliminary research for its future plan.

The Client stressed that this is the first-round preliminary analysis for a further extended study for business expansion. And based on the finding from this preliminary analysis, the Client wants to explore the scope of the future analysis. Simply put, the Client wants to conduct the preliminary analysis within a short period of time under a small budget to taste the flavour of the subject.

The Client sets the following three objectives for this preliminary 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 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 Foursquare venue profiles.

The autonomous city of Buenos Aires (CABA) is a densely populated city: the total population of approximately 3 million in the area of 203 km2. 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 ‘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 true for the general security risk. Obviously, this preconception needs to be assessed based on the actual data in the course of the project. This needs to be kept in mind for the analysis. Nevertheless, the Client ask me to scale risk metrics by 'population density' for the first round of the project.

Overall, the Client demonstrated high enthusiasm about Machine Learning and requested me to use machine learning models to achieve all these three objectives aforementioned.

That is the background (business problem) scenario for this capstone project. On one hand, the scenario setting is totally hypothetical. On the other hand, the project handles real data.

Cut a long story short, for these three objectives presented above, I performed three different clustering machine-learning models. And I got three different lessons out of them. All of them are valuable. And in ***Discussion*** section of this article I will stress these different implications from the perspective of Data Science project management.

For now, I will invite you to walk through the process of the analysis.

The code of the project could be viewed in the following link of my GitHub repository:

- Code: https://github.com/Hyper-Phronesis/Capstone-1/blob/master/Capstone%20Three%20Different%20Lessons%20from%20Three%20Different%20Clusterings.ipynb

Now, let's start.

…………………………

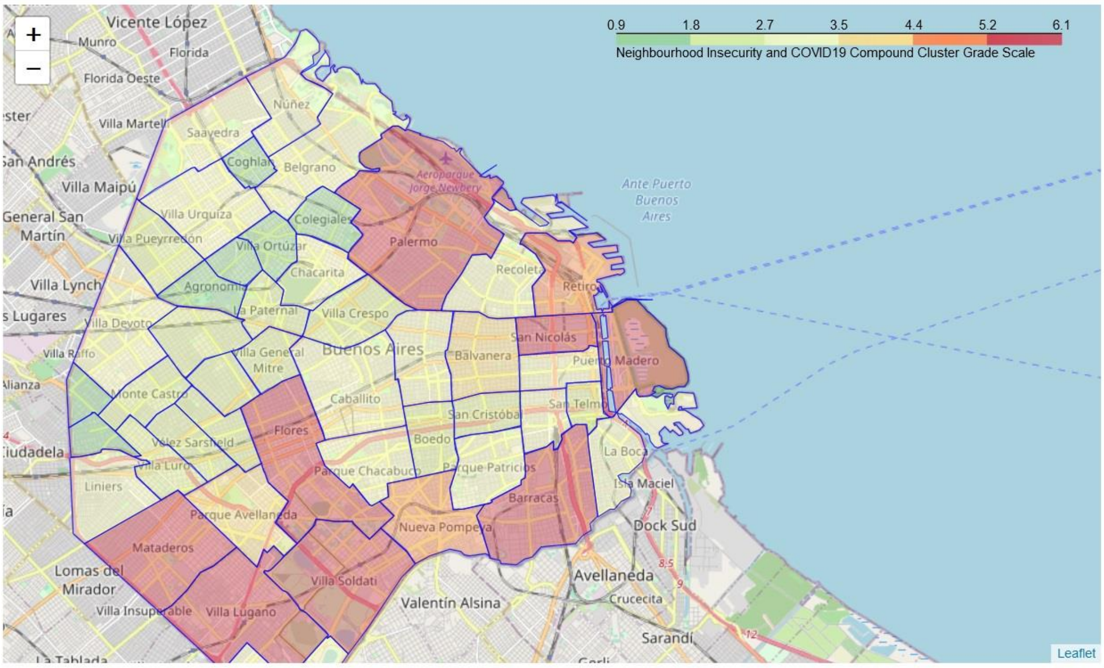
At the beginning of a Data Science project, we need to clarify the following two items:

1) what needs to be solved. (Business Understanding)
2) and what kind of approach we need to make in order to achieve the objective. (Analytical Approach)

By an analogy to cooking, the first part is about deciding what dish you want to cook; then, the second part is about how to cook (grill, boil, steam, or raw-sashimi).

For the case of this project, the Client already has specified both. What the Client wants are risk profiling, venue profiling, and clustering of neighbourhoods. These are all about analysis of the status quo, in other words, descriptive analysis; or potentially, it might involve diagnostic (what happened or what are happening). In other words, the Client is not asking for a forecast (predictive analysis) or how to solve the problem (prescriptive analysis)—at least at this preliminary stage. These navigate the overall direction of our analysis.

Next, we need to talk about data.



A. Data Section

A1. Data Requirements:

By an analogy to cooking, Data Requirements is like a recipe, what ingredients we would need for cooking the dish: thus, what kind of data we would need for the analysis. The three objectives clearly set by the Client determine the data requirements as follow:

(1) 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) Risk statistics:

For the first and the second objective, I would need to gather the following historical statistics to construct a compound risk metric to profile neighbourhoods from the perspectives of both the general insecurity risk (crime) and the pandemic risk (COVID-19).

- general security risk statistics (crime incidences) by neighbourhoods
- pandemic risk statistics (COVID-19 confirmed cases) by neighbourhoods

(3) Foursquare Data:

For the third objective, the Client requires me to specifically use Foursquare in order to characterise each Non-Outlier Neighbourhood.

A2. Data Sources

Based on the data requirements, I explored the publicly available data. Then, I encountered the following relevant souces.

(1) Basic info of the neighbourhoods of CABA:

- the area and the population of all the relevant neighbourhoods from Wikipedia: https://en.wikipedia.org/wiki/Neighbourhoods_of_Buenos_Aires
- 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).

- 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

- 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 API in order to characterise each Non-Outlier Neighbourhood.

A3. Data Collection

What follow now are data collection, data understanding, and data preparation. These parts altogether usually occupy a majority of time for the project, e.g. in a range of 60-70%.

For this article, I would compress the description of these time-consuming parts, only outlining highlights.

After downloading all the relevant data from the data sources above, I have made data reconciliation—cleaning data and transforming it in a coherent format. Thereafter, I consolidated all the data into two datasets: "Risk Profile of Neighbourhoods" dataset and "Foursquare Venue Profile" dataset. The first 4 rows of each dataset are presented below to illustrate their components.

The first 4 rows of "Risk Profile of Neighbourhoods":

	Neighbourhood	Area in km²	Population	Population_Density	Crime Severity Score (CSS)	COVID-19 Confirmed Cases
0	AGRONOMIA	2.1	13963	6649.047619	1288	158
1	ALMAGRO	4.1	128206	31269.756098	14406	3124
2	BALVANERA	4.4	137521	31254.772727	23474	5215
3	BARRACAS	7.6	73377	9654.868421	10329	5199
4	BELGRANO	6.8	126816	18649.411765	11588	1832

The first 4 rows of "Foursquare Venue Profile":

:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	RETIRO	-34.591643	-58.373307	Torre Monumental (ex Torre de los Ingleses)	-34.592166	-58.373752	Monument / Landmark
1	RETIRO	-34.591643	-58.373307	Sheraton Club Lounge - 22nd Floor	-34.593213	-58.372761	Hotel
2	RETIRO	-34.591643	-58.373307	Park Tower Buenos Aires	-34.593560	-58.372863	Hotel
3	RETIRO	-34.591643	-58.373307	Buono Italian Kitchen	-34.593683	-58.373031	Italian Restaurant
4	RETIRO	-34.591643	-58.373307	BASA - Basement Bar & Restaurant	-34.592530	-58.376948	Cocktail Bar

Here is an outline of data limitation below.

(1) Crime Statistics: "Crime Severity Score"

The compiled crime data covers only the period between Jan 1, 2016 and Dec 31, 2018. For the purpose of the project, I would make an assumption that the data during the available period would be good enough to serve a representative proxy for the risk characteristic of each neighbourhood.

The original crime statistics had 7 crime categories. They were weighted according to the severity of crime category and transformed to generate one single metric "Crime Severity Score".

(1) COVID-19 Statistics: "COVID-19 Confirmed Cases"

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. My assumption here is that this will proxy the empirical risk profile of COVID-19 infection.

(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 venues within a neighbourhood from corner to corner. Since there is no uniformity in the area size among neighbourhoods, a compromise would be inevitable, while we want to capture the venue profile of a neighbourhood from corner to corner within its geographical border. Thus, the dataset that I would analyse for Foursquare venue analysis would be a geographically restrained sample set. I will use ***geopy's Nominatim*** to obtain the representative single location point for each Neighbourhood.

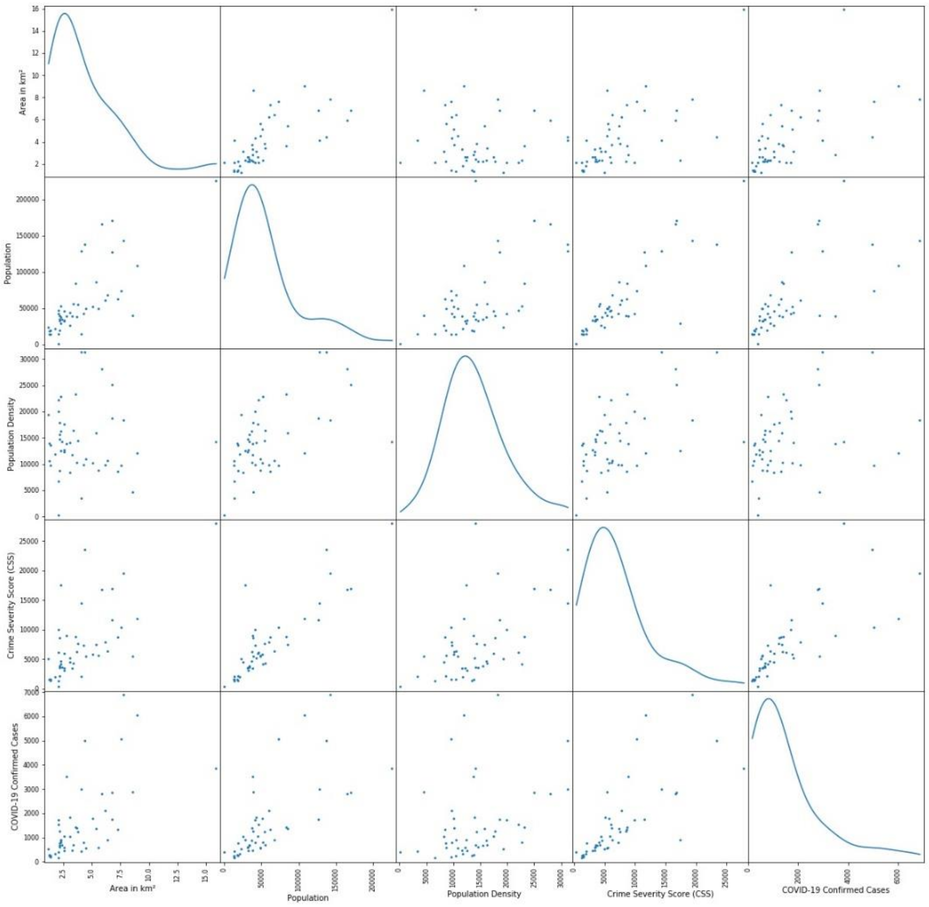
A4. Data Understanding

By now, the required data has been collected and reconciled. By an analogy to cooking, I have already cleaned and chopped the required ingredients according to the cook book. Now, I need to check if the prepared ingredients are representative of what we expected according to the cook book. Analogously, in this step, data understanding, I need to get an insight about the given data.

(1) "Risk Profile" dataset:

For data understanding, there are several basic tools that helps us shape insights about the data distribution. And I performed the following three basic visualizations and generated one basic descriptive statistics:

a) Scatter Matrix:
   The scatter matrix below displays two types of distribution: 1) the individual distribution of each feature variable on the diagonal cells; and ii) the pair-wise distribution of data points for two feature variables.



Here are some insights that I can derived from the scatter plot:

- On the diagonal cells of the scatter matrix, all the data except ‘population density’ demonstrate highly skewed individual distributions, suggesting the presence of outliers.
- In the off-diagonal cells, the most of the pair-wise plots suggest positive correlations in one way or another: except ‘population density’ with the area size and ‘COVID-19 Confirmed Cases’.

b) Correlation Matrix:

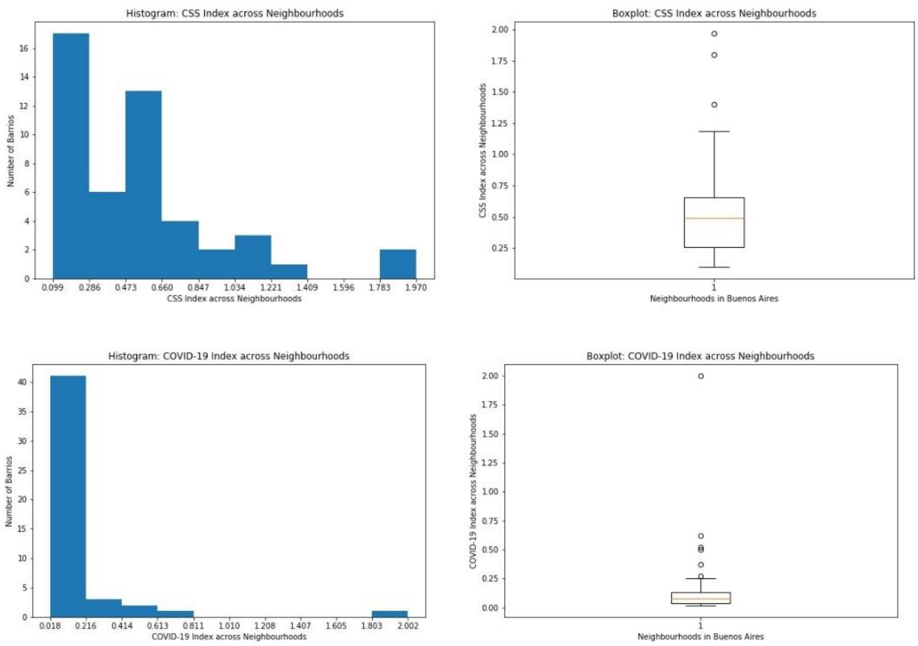
In order to quantitatively capture the second insight above in one single table, I plotted the correlation matrix below.

	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

Overall, “population density” stands out in the sense that it demonstrates relatively lower correlation with these two risk-metrics. On the other hand, population demonstrates the highest correlation with these two risk-metrics. 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? This question needs to be reserved for a suggestion for the second round of this project.

Nevertheless, as per the Client’s request to scale the risk metrics by population density for this first round, 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’.

In order to study individual distributions for these newly created indices, I made the following two basic types of visualizations. Here are two pairs of histogram and boxplot, one pair for ‘CSS Index’ and the other pair for ‘COVID-19 Index’.



c) Histogram:

A histogram is useful to capture the shape of the distribution. It displays the distribution of data points across a pre-specified number of segmented ranges of the feature variable called bins. These two histograms above visually warn the presence of outliers.

d) Boxplot:

A boxplot displays the distribution of data according to descriptive statistics of percentiles: e.g. 25%, 50%, 75%. For our data, the boxplots above isolated outliers over their top whiskers. The tables below present more detailed info about these outliers from these two boxplots.

	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

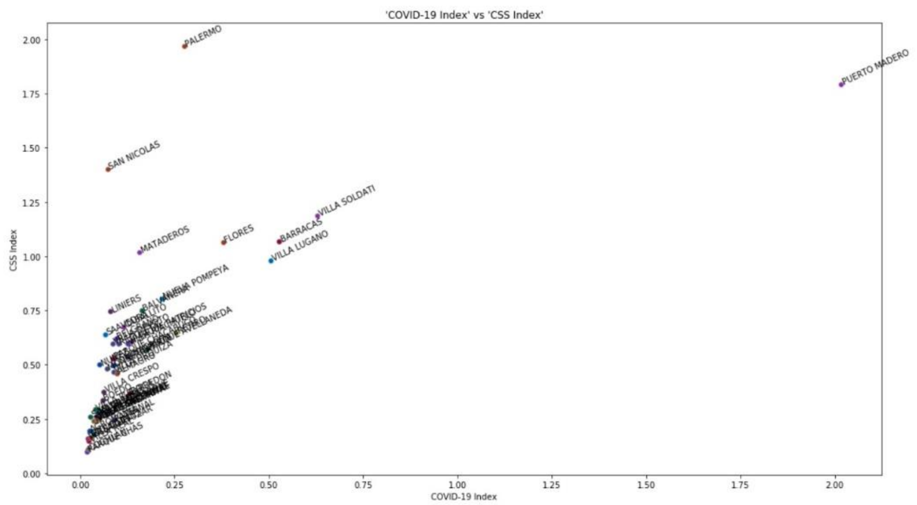
	index	Neighbourhood	COVID-19 Index	Outlier
0	26	PUERTO MADERO	2.001724	COVID-19 Outlier
1	46	VILLA SOLDATI	0.624353	COVID-19 Outlier
2	3	BARRACAS	0.523052	COVID-19 Outlier
3	39	VILLA LUGANO	0.502293	COVID-19 Outlier
4	11	FLORES	0.376239	COVID-19 Outlier
5	21	PALERMO	0.271065	COVID-19 Outlier
6	28	RETIRO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

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

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

Now, let me plot the neighbourhoods on the two-dimensional risk space: ‘CSS Index’ and ‘COVID-19 Index’. The scatter plot below also helps us confirm these outliers visually.

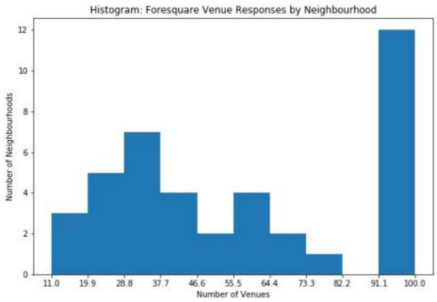
```
s : 48
```

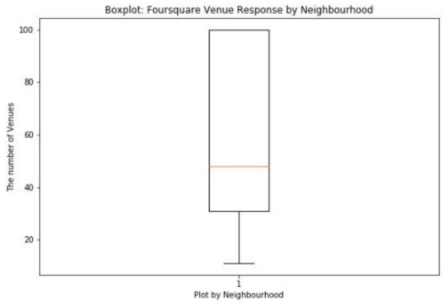


These simple visualizations and descriptive statistics can be a very powerful tool and it helps us shape an insight about the data at the stage of Data Understanding. In a way, before clustering analysis, the boxplot and the scatter plot have already spotted outliers.

(2) "Foursquare Venue Profile" dataset:

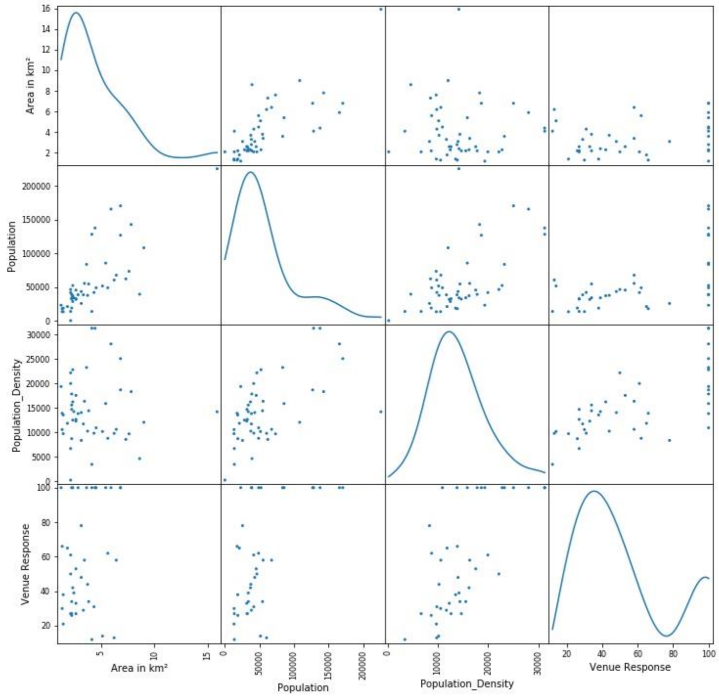
Here is the summary of the Foursquare response to my query. In order to obtain an insight about the distribution of the response across different neighbourhoods, the histogram and the boxplot are presented below.





The histogram 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 km²	Population	Population_Density	Venue Response
Area in km²	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 intuitive 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 venues.

For the rest of my work in data collection, data understanding, and data preparation, I would leave it up to the reader to see more detail in my code in the link above.

B. Methodology & Analysis

Now, the data is prepared for analysis. So, I can move on to analysis

The three objectives set by the Client at the outset and the data availability determine the scope of methodology. Cut a long story short, I run three clustering machine learning models for three different objectives and I got three very different lessons from them.

Before proceeding, let me review the three objectives here.

1. Identify outlier high risk neighbourhoods (outlier neighbourhoods/clusters) 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 presents 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. Simply put, unlike supervised machine learning models, we have no labelled data to train: we have no empirical data about the dependent variable. All these three objectives demand us to discover hidden labels, or unknown underlying cluster structures in the data.

This feature would naturally navigate us to the territory of unsupervised machine learning, and more specifically, 'Clustering Machine Learning' in our context.

By its design—in the absence of the labelled data (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 model outputs. 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.

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.

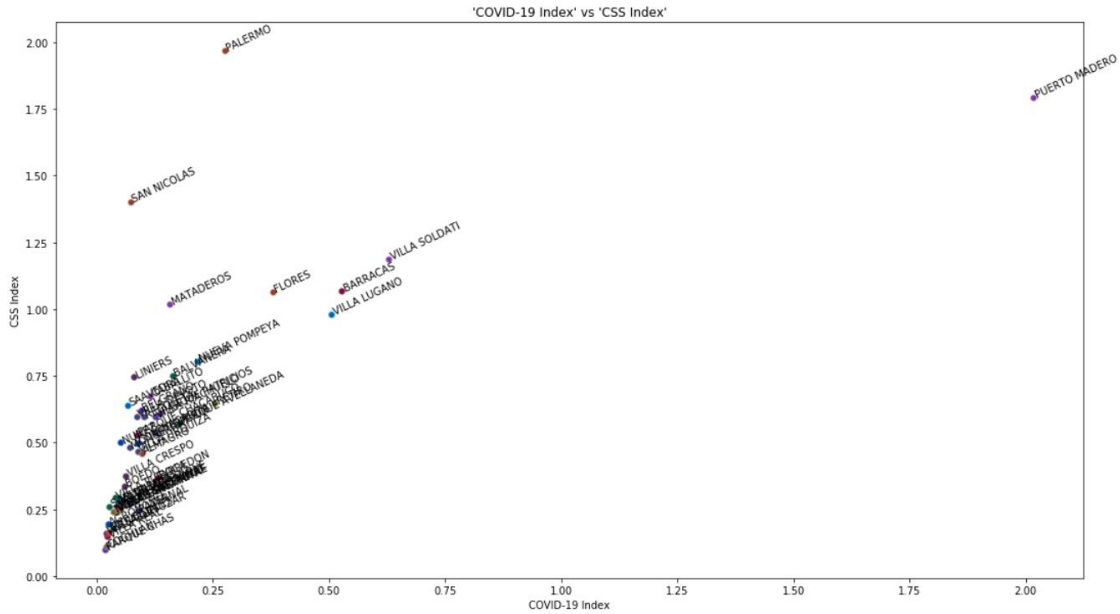
As one more important thing to mention in this overview section, we need to normalize/standardize all the input data before passing them to machine learning models.

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

B.1. DBSCAN Clustering for Objective 1

The first objective is to identify 'Outlier Neighbourhoods'. Now, in the scatter plot 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***?

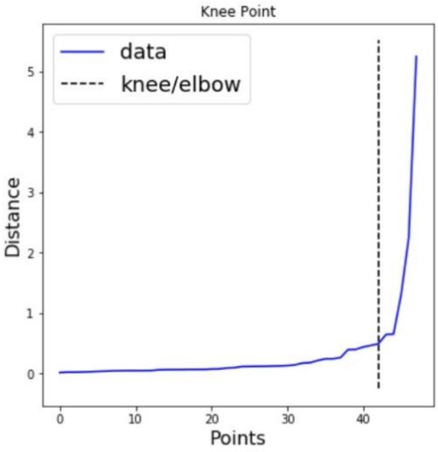
One way to interpret ***the knee point*** is that it is a point where the tuning results start converging within a certain acceptable range. Simply put, it is a point where further tuning enhancement would no longer yield a material incremental benefit. In other words, ***the knee point*** determines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

In order to discover ***the knee point*** of the model hyperparameter, 'eps', for DBSCAN model, I passed the normalized/standardized data of these two risk indices—namely 'Standardized CSS Index' and 'Standardized COVID-19 Index'—into ***the KneeLocator***.

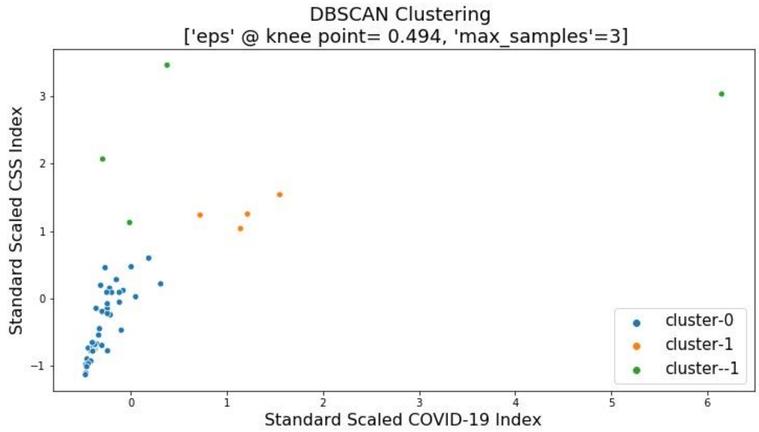
And here is the plot result:

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

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



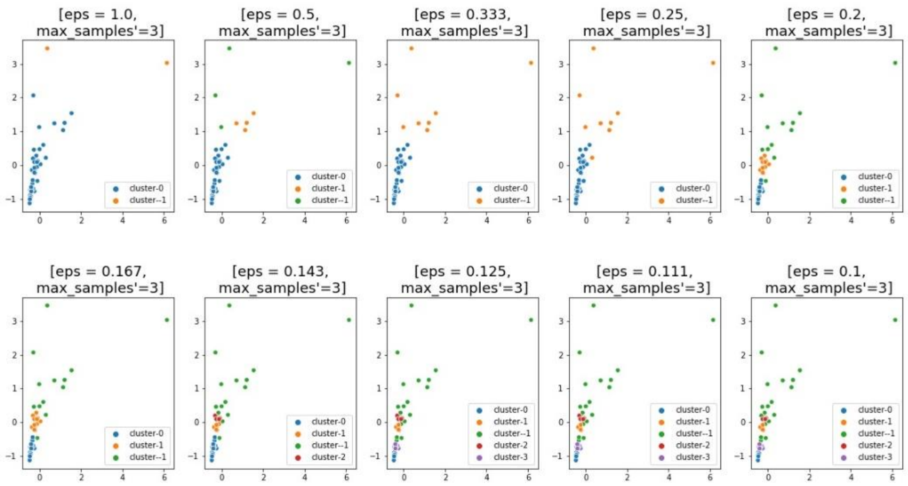
The crossing point between the distance curve and the dotted straight vertical line identifies the knee point. Above the chart, ***KneeLocator*** also returned the one single value, 0.494, as ***the knee point***. ***KneeLocator*** is telling me to choose this value as ‘eps’ to optimize the ***DBSCAN*** 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 them: the cluster -1 (light green) and the cluster 1 (orange). Below, I listed up all the neighbourhoods of these two sparse 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 results from different alternative values of ‘eps’, ***the knee point*** ‘eps’ gave us a reasonable result automatically: 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 isolated into two clusters the same neighbourhoods as the outliers that the boxplot visualization identified 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 does it tell us?

Now, the question worthwhile to ask would be: if we needed to perform a sophisticated and expensive model such as DBSCAN to identify outliers, when the simple boxplot can do that job.

In the perspective of cost-benefit management, the simple boxplot did the same job for the less cost—almost no 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.

B.2. Hierarchical Clustering for the second objective

Now, the second objective 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 risk profile at cluster level (not datapoints/neighbourhoods level).

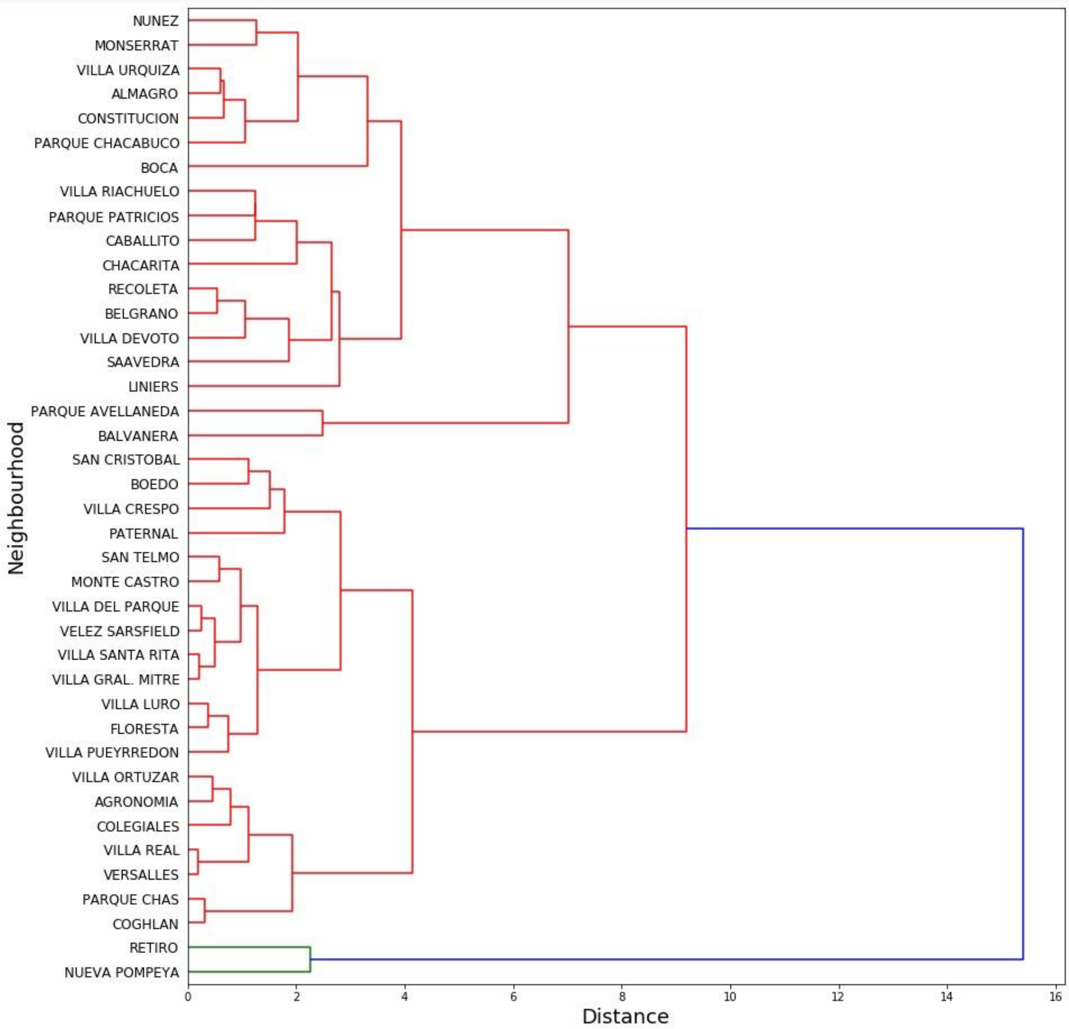
a) Segmentation of 'Non-Outlier Neighbourhoods'.

Given the result of the first objective, now I can remove "Outlier Neighbourhoods" from our dataset and focus only on "Non-Outlier Neighbourhoods" for further clustering segmentations.

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

- I have no advance knowledge how many underlying clusters are expected in the dataset. Many clustering models, paradoxically, require the number of clusters as a hyperparameter input to tune the models *a priori*. But, Hierarchical Clustering doesn't.
- 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.

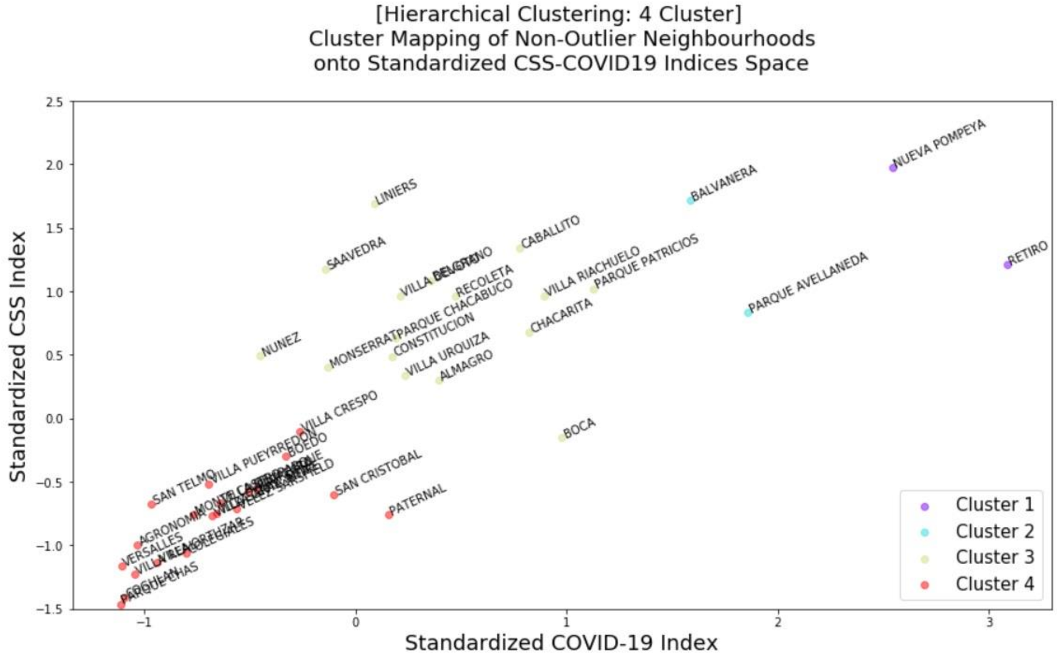
What is a dendrogram? Seeing is understanding! Maybe. Here you go:

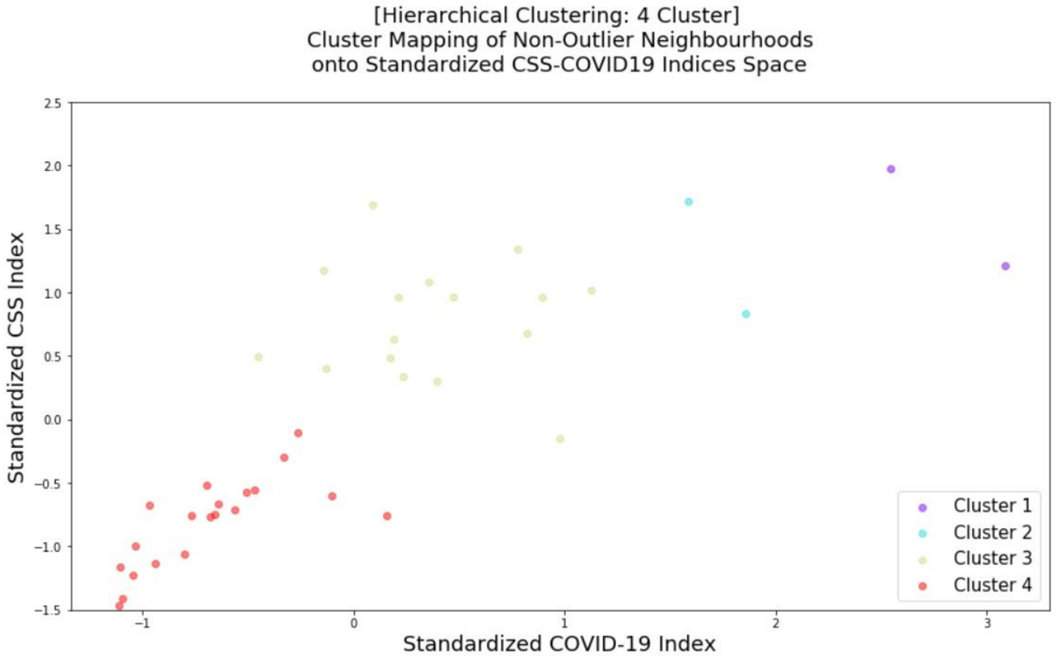


The dendrogram allows the user to study the hierarchical structure of distances among datapoints and the underlying layers of cluster hierarchy. The resulting dendrogram illustrates a tree-like cluster structure based on the pairwise distance distribution. In this way, the dendrogram allows the user to design how many clusters to be made for further analysis. 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 4 (at the distance of 5 or 6 on the x-axis in the dendrogram) as the number of clusters to be shaped.
- Then, I run Hierarchical Cluster Model for the second time, this time with the specification of the number of the clusters, 4.

Accordingly, I got the 4 clusters of the neighbourhoods. The following two charts present the clustered neighbourhoods on the two risk-metrics space: one with neighbourhoods' names and the other without.





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

In order to assign these clusters risk values. I will construct one single compound risk metric.

b) Construction of Compound Risk Metric

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

For this purpose, I formulated 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 measure 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 a data point 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. Nothing else. That's all and simple.

c) Risk Profile of Cluster

Now, my ultimate purpose here is to quantify the risk profile at cluster level, not at data point/neighbourhood level.

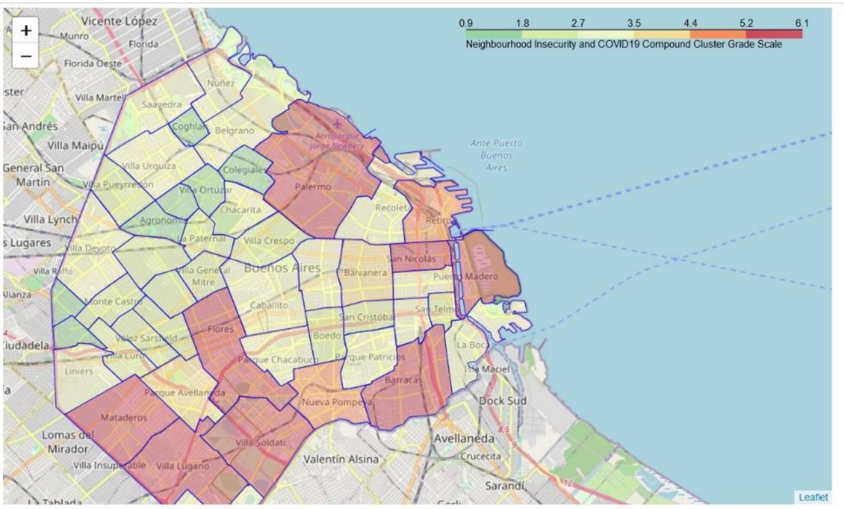
Each cluster has its own unique centre, aka "***centroid***". Thus, in order to measure the risk profile of each cluster, I can refer to the centroid for each cluster. In this way, I can grade and rank all these clusters according to the compound risk metric of their centroids.

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

Here is the result.

	Cluster	Centroid_Risk_Metric	Centroid_Grade
0	1	5.559143	4
1	2	4.536557	3
2	3	3.286430	2
3	4	1.384245	1

The higher the grade, the riskier the cluster. I merged this result with the master dataset and assigned the cluster grade 5 to the 2 outlier clusters. Then, I mapped these cluster grades of all the neighbourhoods across CABA in the following Choropleth Map.



This map visually summarises the findings for 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.

B.3. Foursquare Analysis for the third objective

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

a) Popular Venue Analysis:

I apply One Hot Encoding algorithm to transform the data structure of venue category for further data transformation.

With Foursquare data, which has 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 a very long list thus, I only display the first 7 lines.

Out[96]:

	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	Bakery	Café	Coffee Shop
1	SAN TELMO	1	100	Argentinian Restaurant	Bar	BBQ Joint	Hostel	Coffee Shop
2	VILLA CRESPO	1	100	Pizza Place	Argentinian Restaurant	Café	BBQ Joint	Ice Cream Shop
3	VILLA ORTUZAR	1	71	BBQ Joint	Café	Bakery	Restaurant	Pharmacy
4	COGHLAN	1	68	Pizza Place	Argentinian Restaurant	Bakery	BBQ Joint	Plaza
5	VILLA DEL PARQUE	1	56	Café	Pharmacy	Bar	Argentinian Restaurant	Pizza Place
6	BOEDO	1	54	Argentinian Restaurant	Pizza Place	Bakery	Burger Joint	Café

b) Segmentation of Neighbourhoods based on Venue Profile

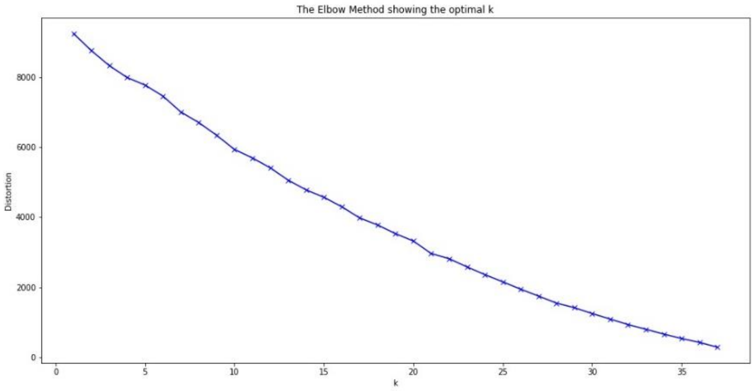
Next, I need to segment the Foursquare venue profile of each neighbourhood. For this purpose, I contemplate K-Means Clustering Machine Learning.

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. These tuning methods would give me an insight about how to cluster the data for a meaningful analysis. Based on the findings 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. 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.

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

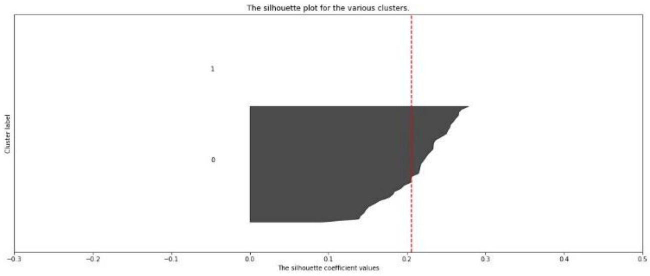
Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This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

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

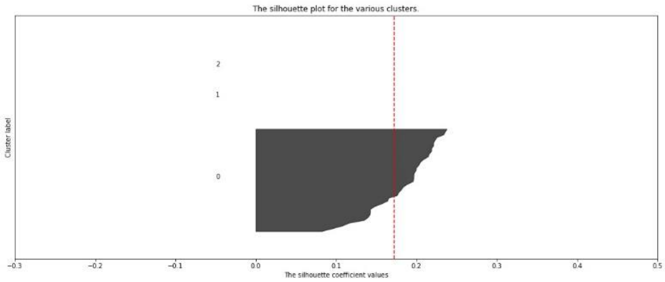
- Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighbouring clusters. Which means, the sample is distinguished from the points belonging to other clusters.
- A value of 0 indicates that the sample is on or very close to the decision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

I 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. And here are the results:

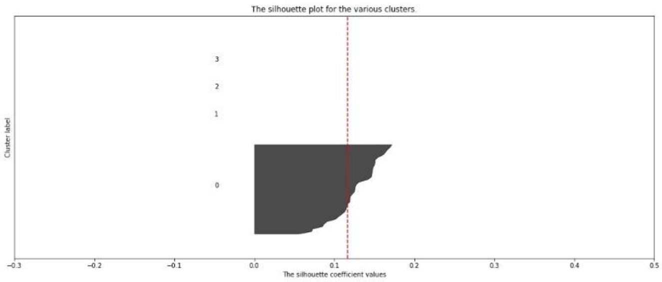
For **n_cluster** = 2:



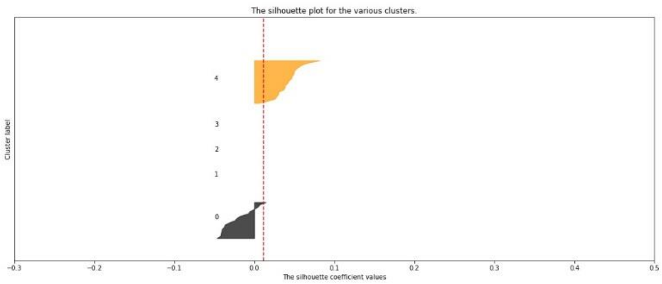
For **n_cluster** = 3:



For **n_cluster** = 4:



For **n_cluster** = 5:



All results are close to 0, suggesting that the sample is on or very close to the decision boundary between two neighbouring clusters. In other words, there is no apparent indication of an underlying cluster structure in the dataset.

Both K-Means Elbow Method and Silhouette Analysis suggest that we cannot confirm an indication about the presence of 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.

Whatever real reason it might be, all we 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 potentially misleading, recommendation, I would rather refrain from performing K-Means Clustering Model for the given dataset.

Discussion

Three Lessons from the Project

Overall, I explored three clustering machine learning models for three different objectives. And, I came across with quite different results among them. I would like to walk you through my observations one by one.

Lesson from the first objective:

At the stage of data understanding, before conducting clustering analysis, the simple boxplots that I made automatically isolated outliers above their top whiskers from the rest: 8 in total for both of these two risk indices—the general security risk metric (Crime Severity Index) and the pandemic risk metric (COVID-19 Index).

The clustering analysis with DBSCAN algorithm identified the exactly same 8 datapoints—as the 8 outliers that the box plots identified—in two clusters remote from the concentration of the rest of the data points. Simply put, the machine learning model only confirmed the validity of the boxplots' earlier automatic identification of those outliers.

This case tells us a lesson that a sophisticated method is not necessarily superior to a simple method. We should take this lesson in modesty from the cost-benefit management perspective.

Lesson from the second objective

For the second objective, I performed Hierarchical Clustering Model and generated its dendrogram on the non-outlier neighbourhoods' dataset. The dendrogram arranged 40 non-outlier neighbourhoods accordingly to their pairwise distance hierarchy. The dendrogram visually made it easy for me to shape human insight about the underlying cluster structural hierarchy. That helped me decide how many clusters to generate with Hierarchical Clustering algorithm for the second run.

This presents a successful case that a machine learning model can play a productive role in supporting human decision-making process. A user can leverage one's own profound domain expertise or human insight in the use of the dendrogram and effectively achieve the given objective.

The lesson here is that the user can proactively interact with machine learning algorithm to optimise the performance of machine learning and make a better decision.

Lesson from the third objective

For the third objective, I performed two hyperparameter tuning methods (K-Means Elbow Method and Silhouette Score Analysis) to discover the best ***n_clusters***, one of the hyperparameters for K-Mean Clustering algorithm. Unfortunately, neither of them yielded a convincing implication about the underlying cluster structure in the Foursquare venue dataset. 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 quality of the Foursquare data.

Or, possibly there could be actually no particular underlying venue-based cluster structure among the neighbourhoods in CABA. That case, there would be nothing wrong with the quality of the Foursquare sample responses. They rather represent the overall actual profile of the city (only regarding the Non-Outlier Neighbourhoods).

Which is correct? This question, requiring a comparative study with data from other sources, might be a good topic for the prospective study.

Nonetheless, whatever real reason it 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. The lesson here would be: in the absence of supporting indication for the use of machine learning, I would be better off refraining from performing it in order to avoid a potentially misleading inference. Instead, I could rather provide more basic materials that can assist the Client use their human insight/domain expertise to analyse the subject.

With these different implications given, it would be naïve to believe that we can simply automate machine learning process from the beginning to the end. Overall, all these cases support that human involvement would make the use of machine learning more productive.

Suggestions for Future Development

As the Client stressed at the outset of the project, this analysis was conducted to yield some flavour of the subject as the preliminary analysis for a further extended study for their business expansion. Now, based on the findings from this analysis I would like to make some suggestions for the next round. Let me start.

Different Local Venue Data Source

Unfortunately, for the second part of the third objective—to segment non-outlier neighbourhoods into clusters based on their venue profile—I could not derive any convincing inference regarding the underlying cluster structure among non-outlier neighbourhoods. In this context, I would suggest that the Client might want to explore other sources than Foursquare to examine the 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 this particular city. Furthermore, the Client might benefit from exploring other analysis than clustering in order to better understand the subject.

Different Scaling

In addition, the Client might benefit from conducting a similar analysis based on per-capita data. As per the Client's request, the risk metrics were scaled by population-density for the first round of the project. A per-capita base scaling might yield a different picture about the risk profile of the neighbourhoods. As a note, population demonstrated the highest correlation with the risk metrics among other feature data.

Effective Data Science Project Management Policy Making

At last, from the perspective of an effective Data Science project management, I would recommend that the Client should incorporate into their data analysis management policy the following two lessons from this project.

1. When a basic tool can achieve the intended objective, it would be cost-effective to embrace it in deriving a conclusion/inference, rather than blindly implementing an advanced tool, such as machine learning.
2. Unless we are certain that the given data is suitable for the design of a machine learning model, it might be unproductive to run it. In such a 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 tool would yield a superior result. 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 in this report for their future productive conduct of data analysis.

Final Products

Now, as the final products for this preliminary project, I decided to present the following summary materials—a pop-up Choropleth map, two scatter plots, and a summary table—that can help the Client use their domain expertise for their own analysis.

Pop Up Choropleth Map

In order to summarize the results for all these objectives, I incorporated a pop-up feature into the choropleth map that I had created for the presentation of the objective 1 and 2. Each pop-up would display the following additional information of the corresponding 'non-outlier neighbourhood'.

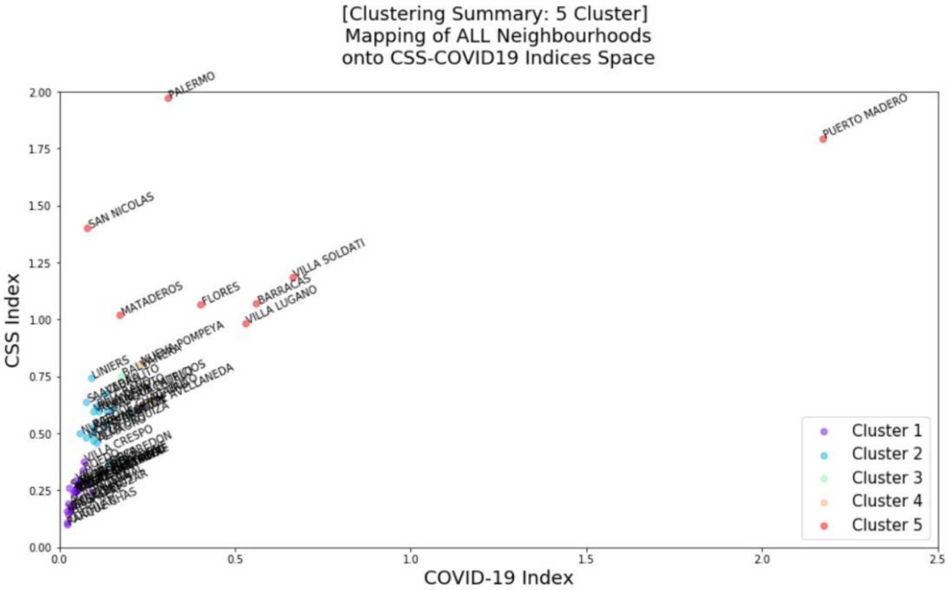
- Name of the Neighbourhood
- Cluster Risk Grade: to show 'Centroid_Grade', the cluster risk profile of the neighbourhood.
- Top 3 Venue Categories

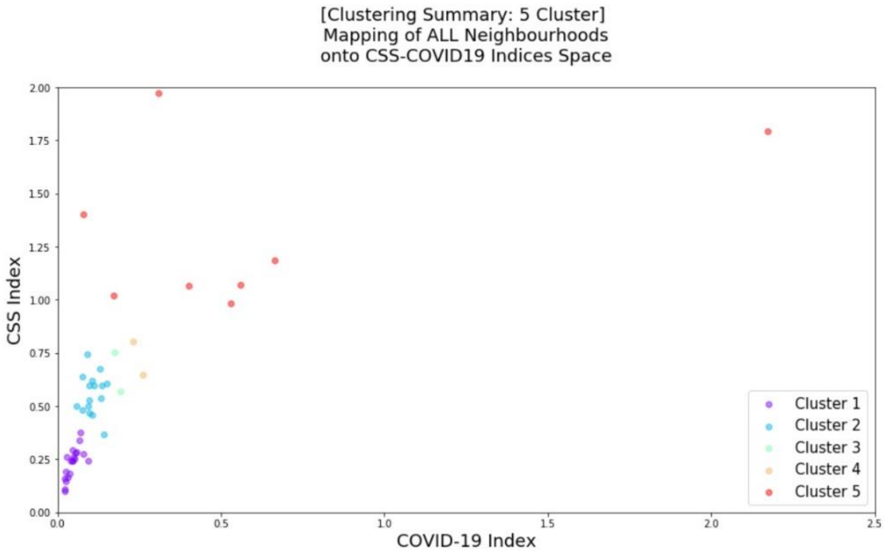
The map above illustrates an example of the pop-up feature.

For high risk ‘outlier neighbourhoods’, I controlled the pop-up feature, since the Client wants to exclude them from consideration.

As a precaution, the colour on the map represents the Risk Cluster, not the Venue Cluster. Since I refrained from generating Venue Cluster, there is no Venue Cluster information. This map would compensate for the absence of the “venue category” base segmentation, allowing the Client to individually explore the popular venue profile for each neighbourhood individually.

The following two scatter plots display the same underlying cluster structure in the two-dimensional risk space: the first one with the names of neighbourhoods; the second without the names. Since the densely plotted names at the left bottom are hiding the view of individual pots for ‘non-outlier neighbourhoods’ in the first plot. The second one 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 for each neighbourhood, 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 the following manner: from the safest cluster to more risker ones; from presumably the commercially busiest neighbourhood to less busier ones.

Since the table is very long, here I would present only the top 7 rows of the table.

Out[96]:

	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	Bakery	Café	Coffee Shop
1	SAN TELMO	1	100	Argentinian Restaurant	Bar	BBQ Joint	Hostel	Coffee Shop
2	VILLA CRESPO	1	100	Pizza Place	Argentinian Restaurant	Café	BBQ Joint	Ice Cream Shop
3	VILLA ORTUZAR	1	71	BBQ Joint	Café	Bakery	Restaurant	Pharmacy
4	COGHLAN	1	68	Pizza Place	Argentinian Restaurant	Bakery	BBQ Joint	Plaza
5	VILLA DEL PARQUE	1	56	Café	Pharmacy	Bar	Argentinian Restaurant	Pizza Place
6	BOEDO	1	54	Argentinian Restaurant	Pizza Place	Bakery	Burger Joint	Café

That's all about my capstone project.

Thank you very much for reading through this article.