Capstone Project: Analyzing the risk of reactivation of schools after the Coronavirus lockdown

Avichai Itzhaki

May 8, 2020

1. Introduction

1.1. In the recent months, the majority of the international community is facing with the enormous outbreak of the COVID-19 pandemic when one of the most common tactics among many authorities around the world for decreasing the disease's rate of expansion is closing temporarily public places and institutions in order to ensure the social distancing as much as possible.

1.2. Particularly, many schools and educational institutions are closed to this method and the consequences of this action are serious and problematic – many students are staying at home and having gaps in the study material, many teachers and educators are practically unemployed and the parents of the students are struggling between managing their work and taking care of their children. This problem is becoming more challenging when it comes to exit strategy from the virus. Although the setting of the educational system back to normal is definitely important, reckless and irresponsible process of returning of huge amount of students might enlarge the virus' spread again among the society.

1.3. Therefore, in order to monitor between the needs of reactivating the schools and keeping the people safe from the virus, I created a model that can help the decision makers of cities and even larger authorities to determine which schools in their jurisdiction are safer to be back in action and by methods of clustering set the basics for a more controlled for releasing the entire educational system after the lockdown.

2. Data

2.1. For this model, I chose to focus on the data that's related to the schools and the high schools of New York City. I chose this city since its area has one of the largest amount of known coronavirus cases in the USA and probably in the entire world.

2.2. I arranged the schools' data in the model's database based on the following features:

2.2.1. Name of School, Neighborhood and Borough – In order to get the list of schools in NYC, I used the same database which was introduced in Assignment 3, derived from the link https://geo.nyu.edu/catalog/nyu_2451_34572 , which has the information about the lat-lon locations of each neighborhood in the city. By that data, I used the Foursquare API explore method to scroll over the neighborhoods and scrap the basic information of each school in NYC while filtering the venues for the relevant categories ("School", "High School", "Religious School" and etc.).

2.2.2. Latitude and Longitude of the school – By the process that was mentioned in section 1, I could also retrieve the lat-lon locations of each school in the database.

2.2.3. The Zip Code Area of the school and its amount of Coronavirus Cases Per 1000 people – Based on the map and the CSV data which are introduced in the article: "Coronavirus in New York City, Tracking the spread of the pandemic",
link: https://projects.thecity.nyc/2020_03_covid-19-tracker/.
CSV files for zip code lat-lon locations and the cases for zip code – at my Github
repository: https://github.com/Avichai1125/Coursera_Capstone/blob/master/uszips.csv https://github.com/Avichai1125/Coursera_Capstone/blob/master/cases_per_zip.csv

2.2.4. The composition of the neighboring venues near the school – The model assumes that schools are likely to be opened at the last stages of the exit strategy, and therefore examining the types of the closest venues to the school's area is a good key for examining the risk of the virus' relapse after the school reactivation.
The venues will be retrieved by the Foursquare API explore method and they will be counted and divided into main groups based on their categories, each represented by a column in the database: Restaurants, Shops and Stores, Outdoor Venues, Religious Sites, Other Indoor Venues.

2.2.5. The Risk Factor – A weighted sum of the features mentioned in the sections of 3 and 4 which represents the level of risk at re-opening the school. The weights will be determined by the importance of each factor to the risk itself when a high weight represents a higher threat.

2.3. For Example, for the data in row 20 of the database:

Column Name	Value
School	P.S 207
Borough	Manhattan
Neighborhood	Marble Hill
Latitude	40.87831
Longitude	-73.90597
Zip Code	10463
Cases_Per_1000	20
Restaurants	12
Shops and Stores	11
Outdoor Venues	0
Religious Sites	0
Other Indoor Venues	3
Risk	75

The division of the venues based on their Foursquare's categories:

Group Name	Foursquare Categories	Total
Restaurants	Pizza Place (x2), Caribbean Restaurant, Mexican Restaurant, Spanish Restaurant, Sandwich Place (x3), Seafood Restaurant, American Restaurant, Steakhouse, Café	12
Shops and Stores	Coffee Shop, Supermarket (x2), Dount Shop (x2),	11

	Department Store, Discount Store, Candy Store, Supplement Shop, Ice Cream Shop, Miscellaneous Shop	
Outdoor Venues	-	0
Religious Sites	-	0
Other Indoor Venues	Yoga Studio, Gym (x2)	3

Determining the risk value:

Feature	Weight for feature	Feature*Weight
Cases_Per_1000	1	20
Restaurants	2.5	12*2.5 = 30
Shops and Stores	2	22
Outdoor Venues	0.8	0
Religious Sites	2.5	0
Other Indoor Venues	1	3
	Total:	75

3. Exploratory Data Analysis

3.1. Choosing weights for the risk factor:

As seen in the Data segment, the weights that were chosen for defining the risk factor aren't equally balanced and varied between 0.8 and 2.5.

Considerations that led for determining the weights include:

3.1.1. Restaurants, shops and stores had the largest weight value since their approachability to the age sections of minor schoolers and high-schoolers is the highest among the other main categories.

3.1.2. Religious sites are also paired with high weight due to the fact that they were proven to be a common crowding spots that unfortunately became prominent centers for the virus to spread.

3.1.3. Outdoor venues however, have a slightly lower than 1 value of weight according to the current perception in the scientific community that the rate of expansion of the COVID-19 virus is decreased in open spaces.

3.1.4. The remaining indoor venues received a neutral weight with a value of 1, due to their span of their types and the fact that a huge part of this group isn't relevant and sometimes prohibited for minor activity (such as gyms, bars and clubs).

3.1.5. Same weight value of 1 was also set for the cases per 1000 feature due to its high mean value compared to the other features and by that to assure its dominancy in the risk factor's essence.

3.2. Overall distribution of the venues' categories:

3.2.1. In order to examine the distribution of the values of the risk factor's components in the school's database, I created a box plot for a visual comparison between them, as seen in Figure 1.

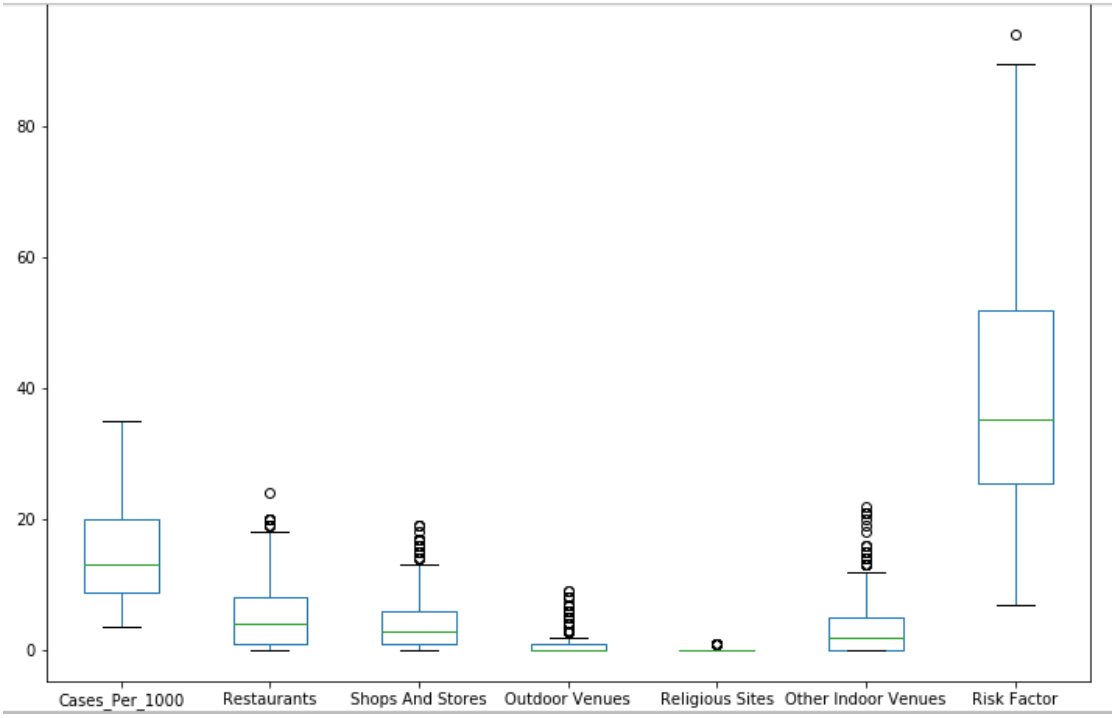

Figure 1: Box plot chart for the neighboring venues' features.

3.2.2. By the box plot's chart, we can see that except from the CP1000 and the Restaurants, all the other features have high number of outliers in their distributions. The reason for this difference is probably due to the fact that the definitions of these features are more general that the other two.

3.2.3. Also, the Restaurants and Shops and Stores features have similar ranges of values, proving that the partition of the entire indoor venues' categories is balanced for the model. However, the Outdoor Venues and Religious Sites features have low mean values and therefore share a small part in the venues' data.

3.2.4. Despite the high amount of the outliers of the risk factor's components, the risk factor's distribution has few of them and thus seems a reliable measure for the model.

3.3. Relationship between the Cases Per 1000 feature and the risk factor:

3.3.1. One assumption that would likely be taken into consideration is there's a strong correlation between the incidence of coronavirus in the school's area (or particularly to our model, its zip code area), demonstrated by the Cases Per 1000, and the risk factor for the school reactivation.

3.3.2. In order to test this assumption, I subtracted the "Cases_Per_1000" feature's part from the risk factor and computed for each of the possible CP1000 values the average for the reduced risk factor values. Then, I created a regression plot for finding the correlation's behavior between the CP1000 values and their corresponding average risk values and the result can be shown at Figure 2.

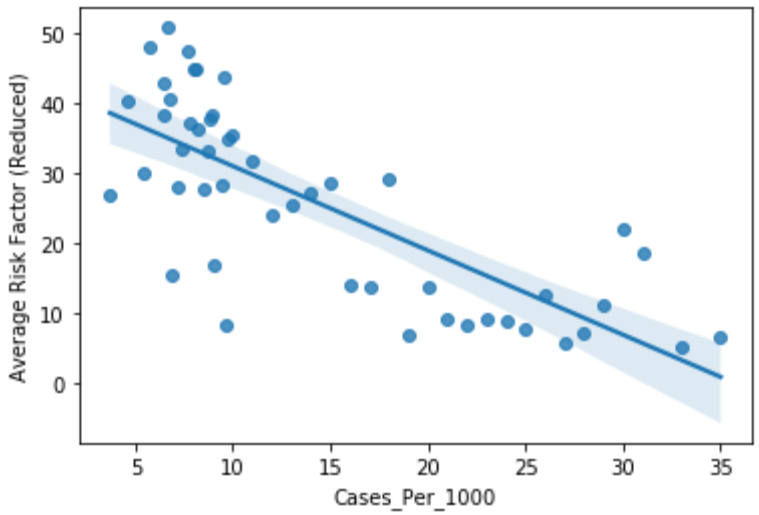

Figure 2: Regression plot for the correlation between the Cases_Per_1000 feature and the corresponding average risk factor (excluding the CP1000 part).

3.3.3. The graph proves the assumption and it demonstrated a strong correlation between those variables, however, the correlation is surprisingly negative. This is probably caused by the fact that the remaining reduced risk factor is based on the venues' data, and since the places with the highest CP1000 values are mostly located in the outskirts and in the poor borough of Bronx, places with lack of tourist spots and places of entertainment, the risk in those areas from the venues is getting lower.

3.3.4. Therefore, it seems that the full risk factor balances pretty well between these sub-factors.

4. Analyzing the Data by the Model

4.1. During the process of the data preparation it became more clearly to see that the chosen features for analyzing the schools in New York City for the model, can be divided into main groups: geographical features – latitude, longitude, borough, neighborhood and zip code and environmental features – cases per 1000, restaurants, shops and stores, outdoor venues, religious sites, and other indoor venues.

4.2. Therefore, in order to have proper insights for setting the key points of the schools' reactivation process, we need to examine the interactions between the geographical and the environmental concepts of the data's features as well as analyzing each of their influences on the risk factor.

4.3. For that matter, the analysis, as it performed by the model, <u>will have three main parts:</u>

4.3.1. <u>Analyzing the relationship between the geographical features and the risk factor:</u>

By applying the K-means algorithm over the data of the lat-lon coordinates, we will test the dependency of the risk factor based on the geographical location of the school.

Since the population of New York City is divided in 5 main boroughs, we set the number of clusters as K=5 in order to have a more realistic and easy understanding results for the decision makers.

4.3.2. Analyzing the relationship between the environmental features and the risk factor:

By applying the K-means algorithm over the environmental features we'll be able to find complex characteristics of the schools' area based on the composition of their surrounding venues. In that case, the objective is to detect special cases and phenomena that might be taken into considering in the reactivation process as well as their effects on the risk factor.

Since there are 5 features for describing the venues' distribution, we set the number of clusters as K=5 for assuring that each feature for the most will have a dominant influence in one of the resulting types.

4.3.3. Analyzing the connection between the environmental and the geographical features

By the knowledge that was earned in the two previous parts of the model and by using statistical methods, we will examine the correlation between the environmental and the geographical according to the two different types of labeling we offered for each school.

4.4. Model's Output:

4.4.1. For each set of labels (geographical and environmental) the model will return a dataframe containing summarized and significant statistical values for examining the clusters.

4.4.2. For each cluster the dataframe will return the following informatin:

- The label of the cluster
- The mean values of the features and the risk factor of each cluster
- Standard deviation of the lat-lon information (for reviewing the span of the cluster's components).
- The partition of the cluster's components among the New York City's boroughs
- The neighborhood with the highest number of components of the cluster

5. Results

5.1. Analysis of the risk factor based on geographical features (location labels):

5.1.1. Based on the results after applying the K-means algorithm over the geographical features of the NYC schools' data, we received 5 clusters whose geographical locations are shown below:

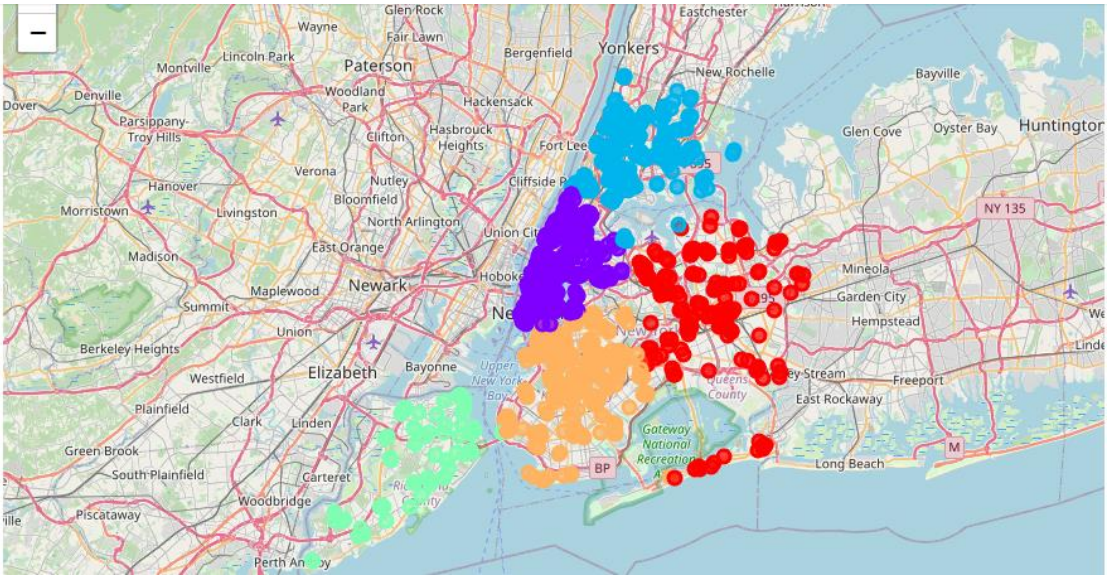

Figure 3: Geographical location of the location-labeled clusters of NYC schools' data

Legend

Label 1, **Label 2**, **Label 3**, **Label 4**, **Label 5**

5.1.2. Full exploratory data of the location-labeled clusters:

Label	1	2	3	4	5
Cluster Size	649	301	72	358	176
Bronx	0	239	0	0	0
Brooklyn	34	0	1	353	8
Manhattan	579	58	0	0	0
Staten Island	0	0	71	0	0
Queens	36	4	0	5	168
Top Neighborhood	East Harlem	Morrisania	Tompkinsville	Boerum Hill	Rego Park
Latitude Mean	40.7465	40.8449	40.5959	40.6682	40.7132
Latitude STD	0.0270676	0.0245274	0.0365428	0.0300619	0.0477948
Longitude Mean	-73.9782	-73.8942	-74.1251	-73.9599	-73.8144
Longitude STD	0.0217996	0.0347936	0.0476901	0.0332348	0.045128
Cases_Per_1000	9.647	21.3987	24.2917	13.4374	20.6818
Restaurants	7.94145	3.02658	1.59722	4.18436	3.15341
Shops And Stores	5.49615	1.88704	0.847222	3.4581	2.32386
Outdoor Venues	1.0077	0.598007	0.486111	0.564246	0.590909
Religious Sites	0.00924499	0	0	0.00837989	0.0113636
Other Indoor Venues	5.55932	1.08306	0.513889	2.38547	0.909091
Risk Factor	46.8815	34.3007	30.8819	33.6723	34.6233

Figure 4: Table of statistical values of the location-labeled clusters

5.1.3. In order to examine the geographical aspects of the risk factor for the schools' reactivation, we'll analyze the presence of the clusters in the boroughs of New York City along with their connection the risk factor itself:

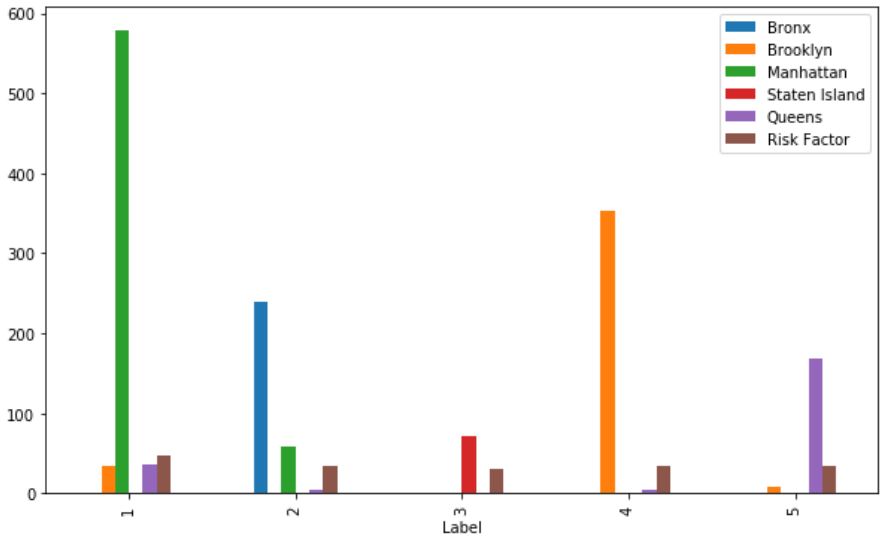

Figure 5: The partition of the location labeled clusters over the boroughs of New York City

5.1.4. Same with the motive of choosing the number of the clusters, the map and the chart demonstrate the strong connection between the labels and the boroughs of NYC. Label 1 is focused on Manhattan, label 2 on Bronx, label 3 on Staten Island, label 4 on Brooklyn and label 5 on Queens.

5.1.5. The highest mean value of the risk factor among the labels belongs to label 1 (46.8815) while in other clusters the value is around 35, fact that implies about Manhattan being a very challenging place for the schools' reactivation procedure. That claim can also be confirmed by looking the venues' data of label 1 where the mean values of environmental features are significantly higher than the other labels.

5.1.6. On the other hand, Staten Island is shown to be the least dangerous place for the reactivation procedure due its low risk factor value, in contrast to its high CP1000 value, proving the negative correlation we examined between this value and the risk factor.

5.2. Analysis of the risk factor based on environmental features (composition labels):

5.2.1. Based on the results after applying the K-means algorithm over the environmental features of the NYC schools' data, we received 5 clusters whose geographical locations are shown below:

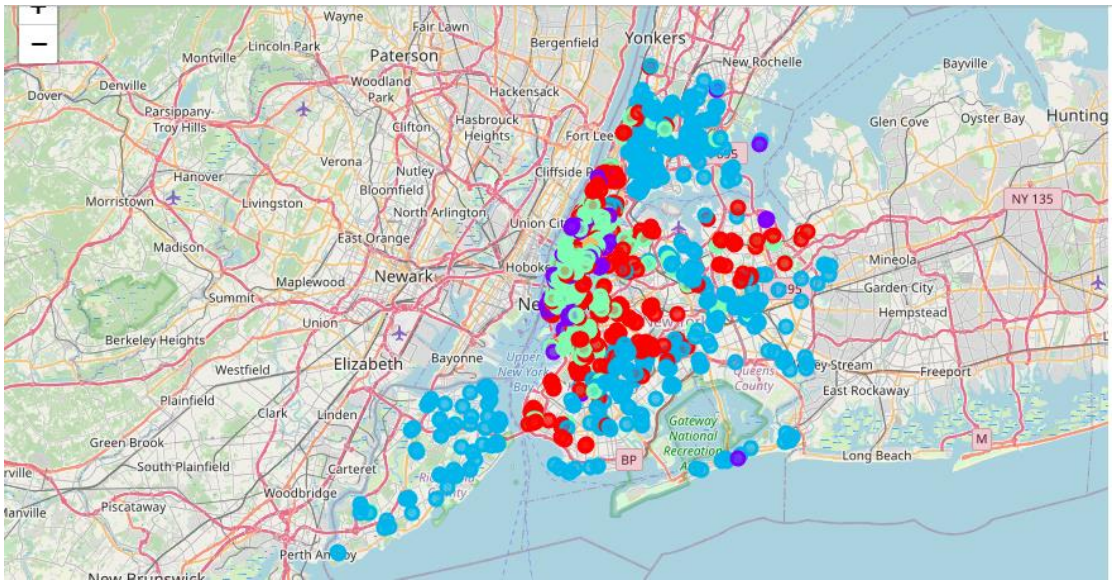

Figure 6: Geographical location of the composition-labeled clusters of NYC schools' data

Legend

Label 1, **Label 2**, **Label 3**, **Label 4**, **Label 5**

5.2.2. Full exploratory data of the composition-labeled clusters:

Label	1	2	3	4	5
Cluster Size	95	526	499	11	425
Bronx	2	217	12	0	8
Brooklyn	14	96	112	3	171
Manhattan	71	25	348	6	187
Staten Island	0	70	1	0	0
Queens	8	118	26	2	59
Top Neighborhood	Financial District	Morrisania	Midtown South	Carroll Gardens	Lower East Side
Latitude Mean	40.7393	40.7416	40.7393	40.7339	40.7274
Latitude STD	0.0422661	0.104357	0.0444934	0.0353402	0.0558987
Longitude Mean	-73.973	-73.9122	-73.9716	-73.9617	-73.9514
Longitude STD	0.0515442	0.100392	0.0385996	0.0658171	0.0521964
Cases_Per_1000	8.90105	22.4049	10.3573	11.0818	11.7193
Restaurants	5.50526	1.61787	10.4168	6.09091	3.75059
Shops And Stores	4.25263	1.03232	7.6012	5.81818	2.44471
Outdoor Venues	4.22105	0.545627	0.537074	0.545455	0.501176
Religious Sites	0	0	0	1	0
Other Indoor Venues	5.83158	0.572243	6.34269	4.90909	2.14353
Risk Factor	40.3779	29.523	58.3741	45.7909	28.5296

Figure 7: Table of statistical values of the composition-labeled clusters

5.2.3. For having a clear definition for each group in the data which was presented by the clusters, we'll analyze the mean values for the venues' data for each one of them, along with the corresponding risk factor values:

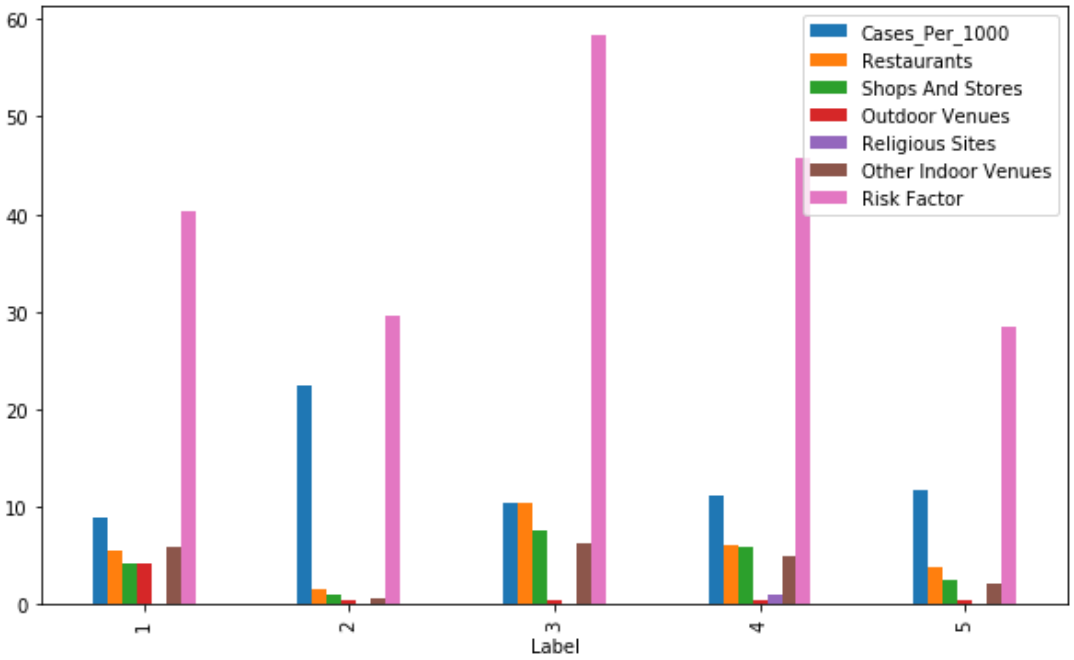

Figure 8: The mean values of the venues' data and the risk factor for each of the clusters.

5.2.4. By analyzing the mean values of the venues' data for each of the labels, compared to the corresponding mean value for the risk factor, we can properly define the sub groups of the NYC schools which the composition labels are referring to:

- Comp label 3 - A large cluster of schools with low CP1000 values, but with high amount of restaurants, shops and stores around them, with average risk factor of 58.3741 (the highest).
- Comp label 4 - The smallest cluster among the other of schools with low CP1000 values, but with high presence of restaurants and religious sites, with average risk factor of 45.7909.
- Comp label 1 - A small-medium sized cluster of schools with the lowest CP1000 values among the other labels, but with a significant

presence of indoor and outdoor venues around them, with average risk factor of 40.3779.

- Comp label 2 - The largest cluster of among the others of schools with very high CP1000 values and a very low presence of venues' around them, with average risk factor of 29.523.
- Comp label 5 - A very large group of schools with low CP1000 values and low presence of venues around them, with average risk factor of 28.5296 (the lowest).

5.2.5. The huge amount of schools with low risk values (labels 2 and 5) can enlarge the operating space and the requested time limit for the reactivation process, but the high standard deviation values of the more problematic clusters (label 3 and label 1) might be a huge obstacle for initiating the plan.

5.3. Relationship between the location labels and the composition labels

5.3.1. By computing the Pearson correlation coefficient between the vectors of the location labels and the composition labels (derived from the columns of the model's dataframe), we get a value of **-0.0117** , which is pretty close to 0 and therefore indicates that there is no sign of dependency between these two labeling types.

5.3.2. However, we can still takes notes for the relationship between them by analyzing the partition of the composition labels over the location labels (How many schools with location label of 1 have a composition label of 1 and so on…):

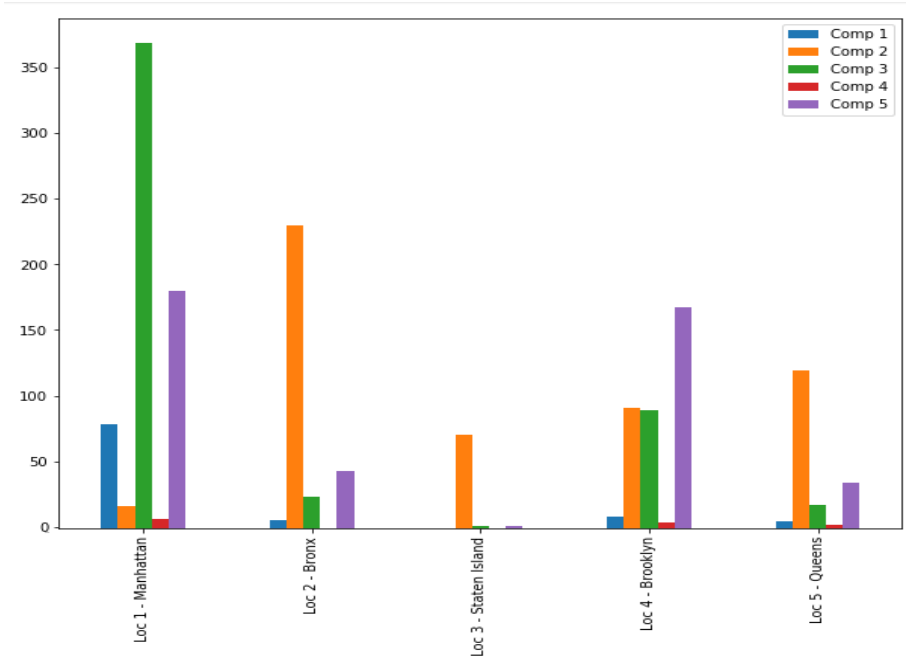

Figure 9: The partition of the composition labels over the location labels

5.3.3. By looking at this chart along with the geographical span of the composition labels' clusters, it's pretty clear to see that the largest number of high risk labeled schools (labels 1, 3 and 4) are mostly located in Manhattan area and as long as the school is far away from the center of the city, its affiliation with the "safe" clusters (labels 2 and 5) is more common.

5.3.4. This tendency can serve the decision makers as a keypoint for determining the order of the schools for reactivation.

5.4. Ordering the schools by the risk factor and the distribution of the labels

5.4.1. The final model's dataframe can be ordered by the risk factor column and therefore can be a great source of information for the decision makers to scroll over the schools' data meticulously and define the order of the schools to be reopened.

5.4.2. Also, by reviewing the ordered dataframe we can analyze the distribution of the labels as long as the risk factor increases. The following graphs will demonstrate the cumulative incidence for each of the labels according to the risk factor ordered list of NYC schools' data.

5.4.3. Distribution of the location labels along the increasing risk factor:

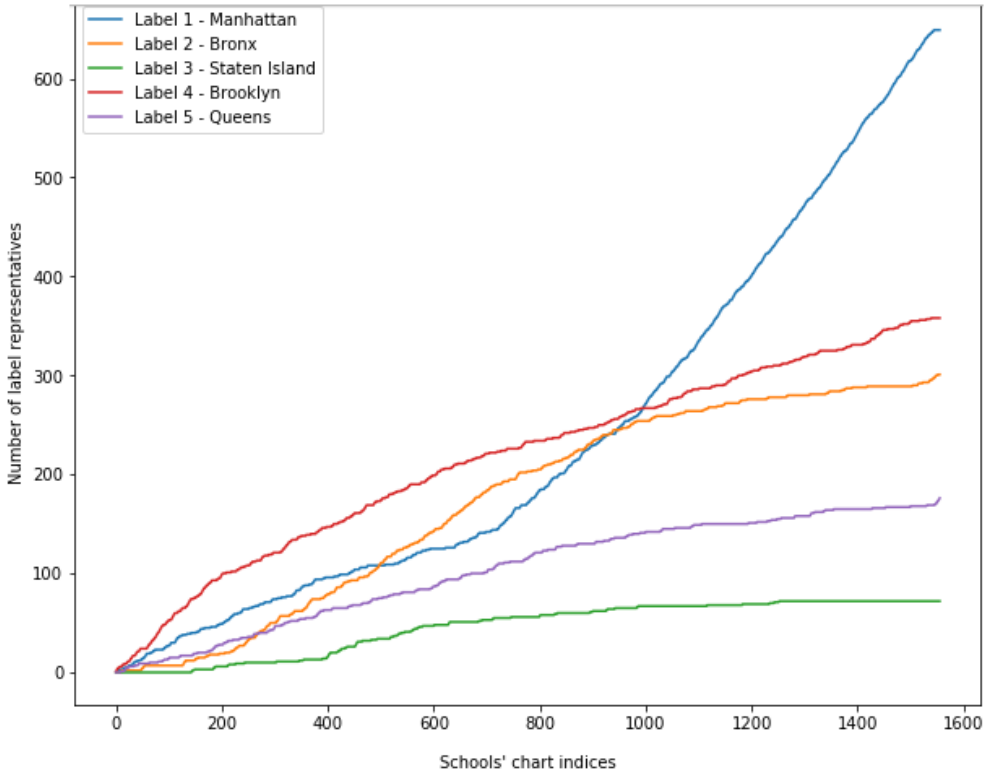

Figure 10: Distribution of the location labels along the increasing risk factor

- The chart shows that the Manhattan related cluster has a decent number of low risk schools while it also has a dominant amount of representatives at the top 600 highest risk schools in NYC. The increase in its growth on list occurs at the same place where the other labels' schools became less common.

5.4.4. Distribution of the composition labels along the increasing risk factor:

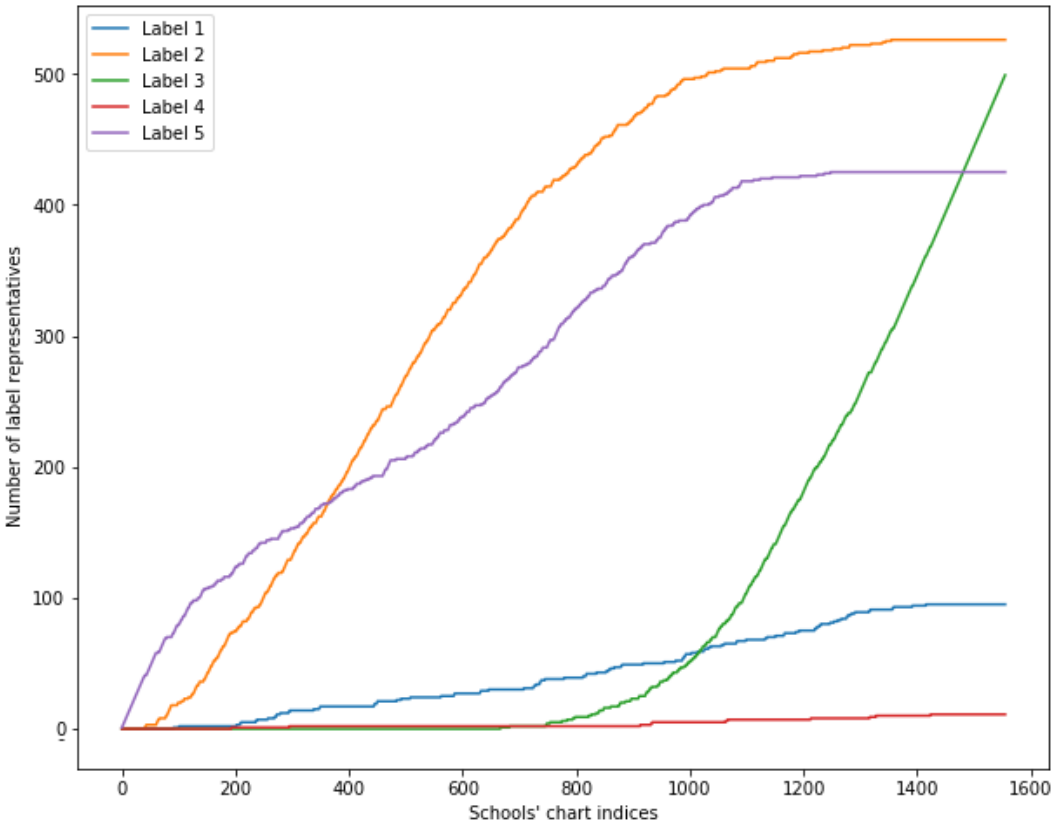

Figure 10: Distribution of the composition labels along the increasing risk factor

- The chart shows that the Manhattan related cluster has a decent number of low risk schools while it also has a dominant amount of representatives at the top 600 highest risk schools in NYC. The increase in its growth on list occurs at the same place where the other labels' schools became less common.

6. Conclusions

6.1. The model's results reveal that there is a contrast between two main factors which probably had the largest effect of the risk factor's determining: The cases per 1000 value and the amount of venues around the schools. Although at first glance these two factors affect each other in a linearly positive relation, the relation itself, as it was depicted by the model, was actually inverse - schools from areas with high CP1000 values were unlikely to have exposure to venues around them (such as Bronx and Staten Island) when on

the other hand, schools from areas with low CP1000 values were mostly had a significant amount of venues in a close distance.

6.2. As I mentioned before at the Exploratory Data Analysis segment, this phenomenon is probably caused by external economic factors that increases the CP1000 value and decrease the risk from the venues at the same time (poor conditions, lack of entertainment spots and etc) and even though the risk factor seems to target schools with higher presence of neighboring venues and small CP1000 values as more problematic to be reopened, the decision makers need to focus equally on both of these factors and find the compromise between them at each stage of the reactivation procedure.

6.3. Also, the results indicate the general distribution of the schools based on their risk factor, when the risk is more present at the center of the city (especially at the Manhattan area) and it's getting weaker as long as we move to the outskirts of city. Therefore, this insight gives the decision makers a reliable key point for determining the order of the schools to be reopened at the reactivation procedure along with the ordered list of schools that have already derived from the model's results.

6.4. However, following the general direction of the risk levels shouldn't come on the expense of detecting special cases. For example, schools with religious sites around them made up a very small part of the entire data, but in the final results they are considered as a widely spread group with very high risk factor, showing that these kinds of anomalies must be paid attention at the procedure.