Report
Capstone Project - The Battle of Neighborhoods
Cristina Acevedo

1. Introduction

In this project I aim to get some insights into COVID-19 spread in the city of Bogotá, Colombia. As it is known, this is a disease that is very easily transmitted and therefore a los of cities around the world have been taking measures to limit the contagion. Currently, Bogotá is under quarantine and this is likely to continue at least for some weeks. Around some neighborhoods quarantine has been greatly abided; however, in some other regions this has not been the case, mainly because of economic reasons. The main cities in Colombia have an average unemployment rate of 12.2%, but during quarantine this number is expected to rise. Furthermore, a high percentage of the employed population have an informal job, which means they do not have any kind of economical income or insurance to help them through this situation. Since some of the district's help has been delated, this population may be forced to go out and look for ways to earn some money, even though this means putting themselves and others at risk. I seek to understand some of the complex dynamics of the disease spread and its relation to the economic needs around Bogotá. This information might be interesting to government, in order to take decisions regarding how to allocate resources. Particularly, I will be looking at the need of hospital beds.
(As the data is from a spanish speaking country, this report might contain some information in spanish. I tried to translate everything that was necessary in the descriptions)

2. Data

Bogotá's goverment has made public a lot of data related to COVID-19 through the following website. I gathered data from lab confirmed COVID-19 cases up to April 19th. This database contains information about the date, city, localidad age, sex, kind of case, location and state. The variable localidad contains information about where the case was located in the city.

	ID	DATE	CITY	CODE	LOCALIDAD	AGE	SEX	KIND	LOCATION	STATE
0	1	2020-03-06	Bogota	1.0	Usaquen	19.0	F	Importado	Casa	Moderado
1	2	2020-03-10	Bogota	10.0	Engativa	22.0	F	Importado	Casa	Moderado
2	3	2020-03-10	Bogota	10.0	Engativa	28.0	F	Importado	Casa	Moderado
3	4	2020-03-12	Bogota	9.0	Fontibon	36.0	F	Importado	Casa	Moderado
4	5	2020-03-12	Bogota	8.0	Kennedy	42.0	F	Importado	Casa	Moderado

I will also use a database from this website, which contains information about the coordinates of every localidad in the city.

	LOCALIDAD	LONGITUD	LATITUD	CODIGO	gp
0	BARRIOS UNIDOS	-74.084000	4.666400	12	-74.084,4.6664
1	ENGATIVA	-74.107200	4.707100	10	-74.1072,4.7071
2	SUMAPAZ	-74.315224	4.034746	20	-74.315224,4.034746
3	TEUSAQUILLO	-74.093800	4.644800	13	-74.0938,4.6448
4	LA CANDELARIA	-74.073900	4.593900	17	-74.0739,4.5939

Finally, I will gather some data from Foursquare about the venues located in every localidad.

(468, 7)

	Localidad	Localidad Latitude	Localidad Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	BARRIOS UNIDOS	4.6664	-74.084	Centro Canino Cruz Roja	4.665530	-74.086151	Dog Run
1	BARRIOS UNIDOS	4.6664	-74.084	Riquisimo - Postres y Helados Principal	4.668366	-74.083662	Dessert Shop
2	BARRIOS UNIDOS	4.6664	-74.084	Solo Postres	4.667903	-74.083965	Dessert Shop
3	BARRIOS UNIDOS	4.6664	-74.084	Campo de Practica Fedegolf	4.663400	-74.084510	Golf Course
4	BARRIOS UNIDOS	4.6664	-74.084	Postres La Enramada	4.667113	-74.084464	Dessert Shop

3. Methodology

The first step was to download the data from the websites. The database (df1) with information on the coordinates was in a csv file, while the database (df2) related to coronavirus was in an excel file. On df1 I deleted the gp column, because it contained the latitude and longitude already present in other columns and the Bogotá row as it contained the city coordinates. Following this I created a map with a marker for every localidad.

Then, the information of nearby venues was collected for every localidad through a request. The radius was fixed at 1000 and the limit at 100. The obtained database had information about the venue category, which was important to find its economical impact during the quarantine. Venues were found for 17 of the localidades, ranging between 1 and 100. Naturally, there is some bias in this information, since the most touristic localidades might get a higher number of venues.

After this, the venues were classified as opened or closed during the pandemic. This is due to the fact that opened venues may produce a contagion risk, but offer some economical relief; while closed venues create economical stress and in some cases it might cause people to go out and put themselves at risk. The venues were classified as opened if their venue category contained one of the following words, according to what has been stated by the government.

- Joint
- Restaurant
- Place
- Food
- Bakery
- Breakfast
- Café
- Coffee
- Convenience Store
- Deli
- Dinner
- Grocery
- Supermarket

Next, I imported the data about the COVID-19 confirmed cases. I deleted the rows of cases outside the city and deleted the city column. I decided to create a dataframe of this information summarized. Therefore I created a dataframe grouped by localidad with dummy variables for all of the none quantitative columns and use the describe function to get the descriptive statistics by localidad. Also, a correlation analysis was performed in order to understand the relations between variables.

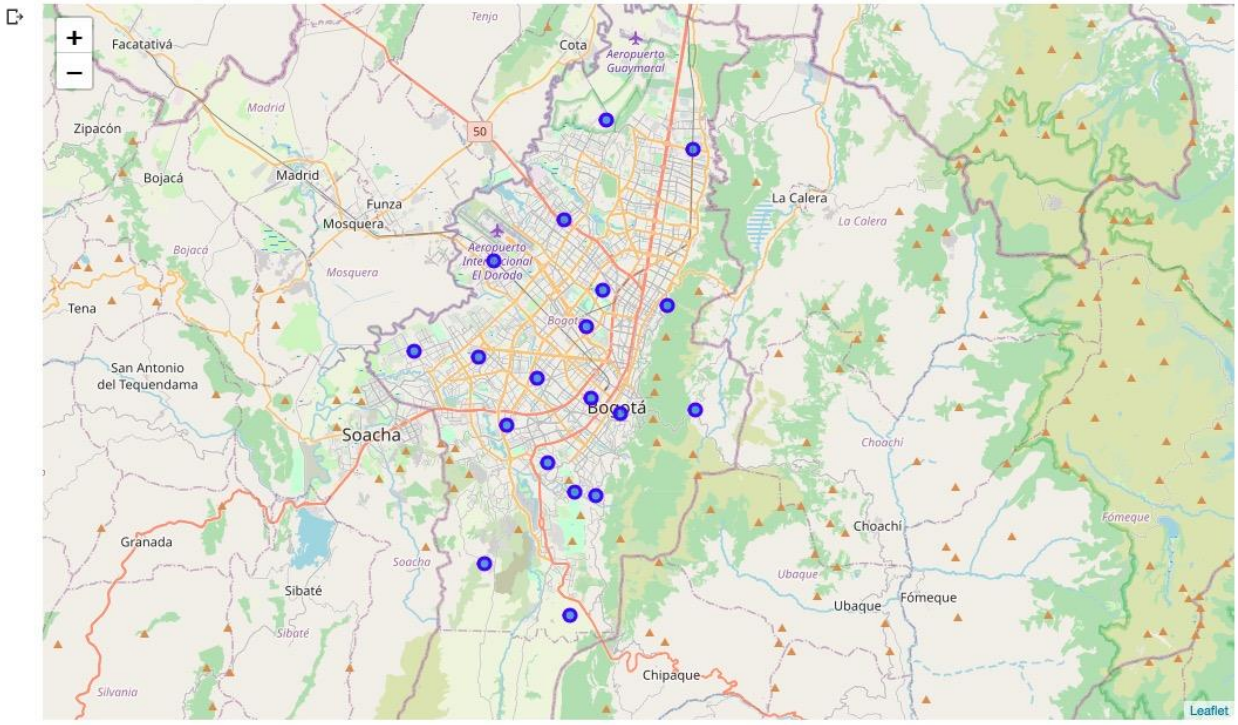
In order to understand better the dynamics in relation to the location in the city, two more maps were created. One with information about the number of cases by localidad and another one with the number of deaths by localidad.

Some more visual information was obtained by graphing variables like the type of case, its location and the state.

Finally, the data was adapted to generate a model. In this part I was interested in predicting wether a case was going to be in a hospital or in a hospital ICU. This might be useful information to plan for the future, as hospital beds are a limited resource. I generated two types of models (SVM and linear regression) and compared them using different metrics.

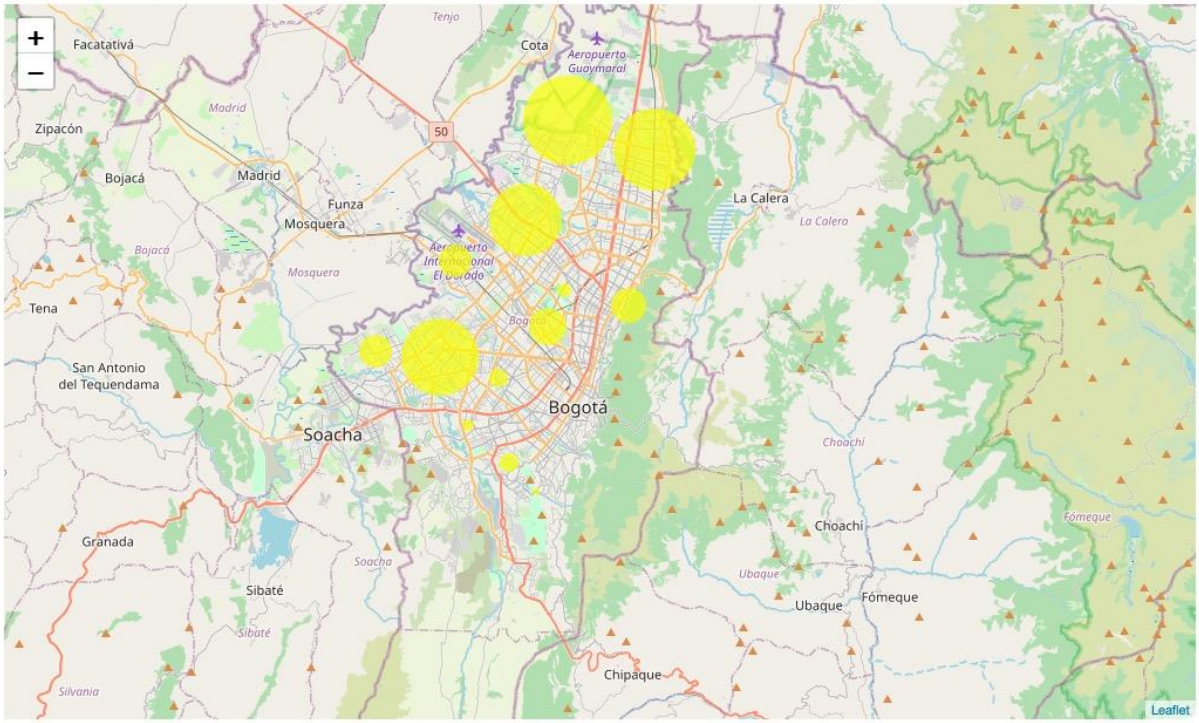
4. Results

First, I graphed some maps of the city with information from the databases. This map represents all of the localidades in Bogotá in the map. This localidades group many neghborhoods and differ in size, population and economic activities.

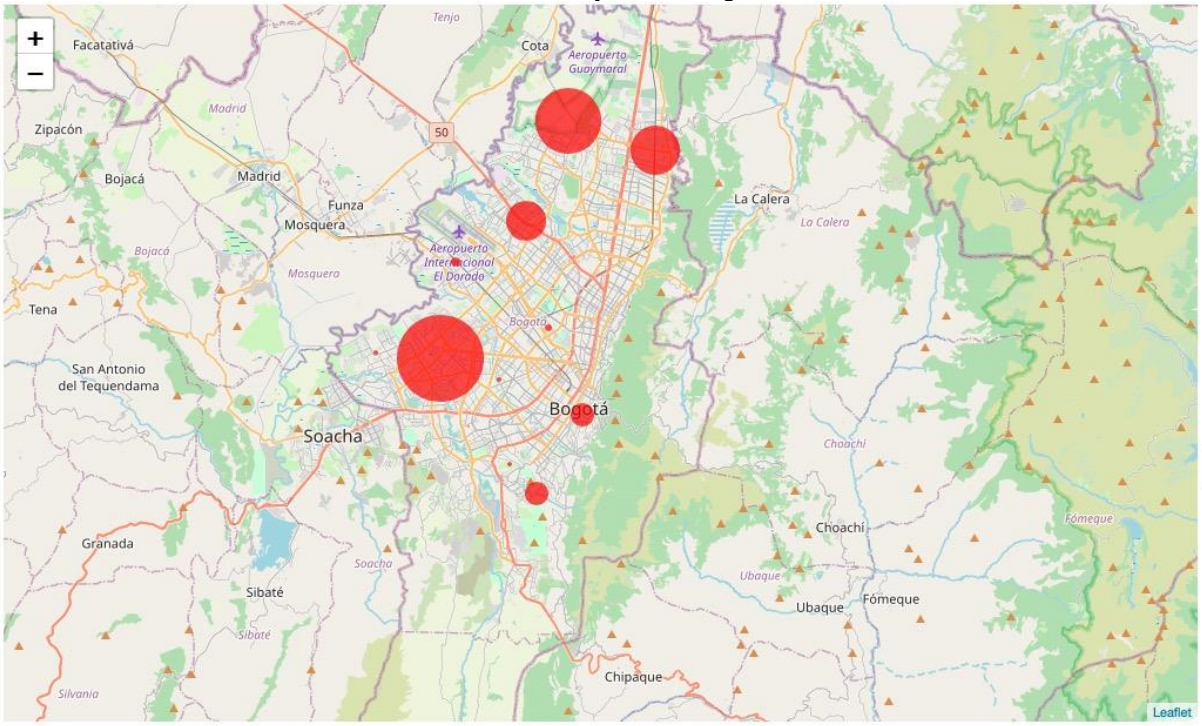


The following map contains information about the number of cases in every localidad. Some of the most affected localidades are Suba, Usaquén and Kennedy. There is an uneven distribution as Kennedy is very far from the other two. Usaquén was the localidad where the first case was reported. It is a localidad where a lot of wealthy people live, which means many of the cases were imported. Kennedy, however, is a working class localidad, so many of the people who live there

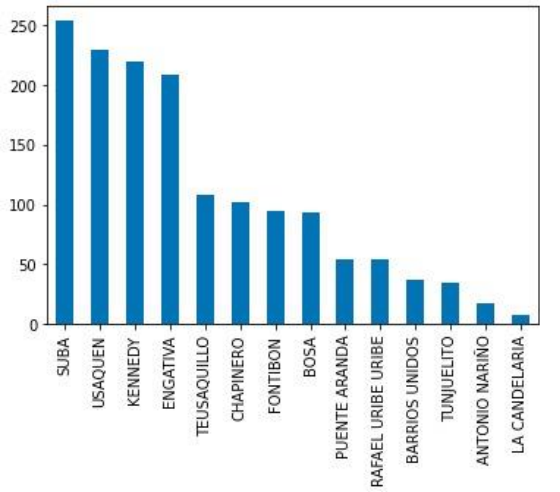
have to travel all the way to places like Usaquén to work, where they might have gotten the dissease.

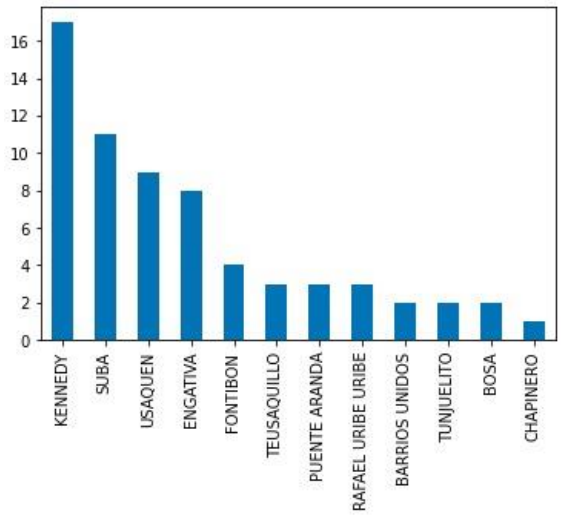


And this one has information on the number of deaths of each localidad. The most affected ones are Kennedy, Suba and Usaquén. This reveals an interesting fact, since Usaquén, a wealthy localidad was a lower death rate than Kennedy, a working class localidad.

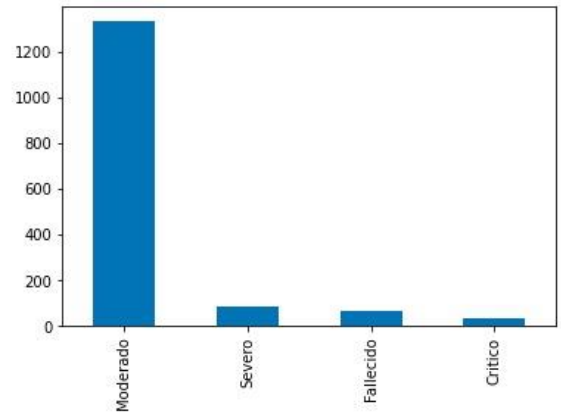


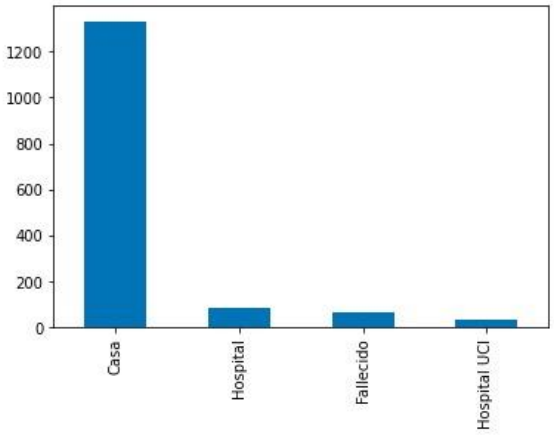
Next, I got some information about the distribution of cases. The following graphs represent the number of cases and deaths by localidad.

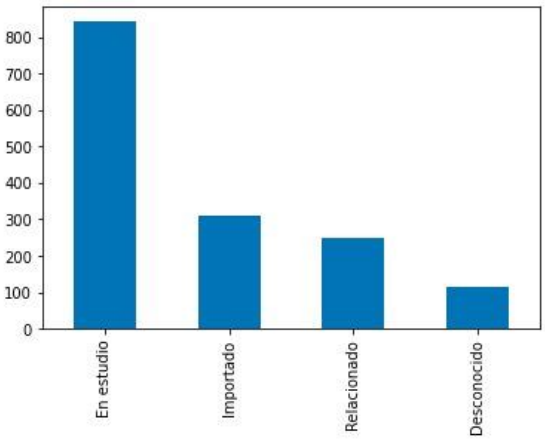




And the following graphs represent the number of cases by kind, location and state, respectively. As might be predicted, the majority of the cases are moderate and most of the people are at home. A great amount of the cases are under study, as It is not known where the contagion might have happened.







The quantitative analysis gave the following results

	Desconocido	En estudio	Importado	Relacionado	Casa	Fallecido	Hospital	Hospital UCI	F	M	Critico	Moderado	Severo	CASES	LONGITUD	LATITUD	CODIGO	NUMOPEN	NUMCLOSE
count	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000	14.000.000
mean	1.485.142.857	10.051.928.571	3.882.142.857	3.018.000.000	16.096.785.714	838.285.714	1.023.571.429	478.571.429	9.489.142.857	8.948.071.429	478.571.429	16.096.785.714	1.023.571.429	108.214.286	-74.107.114	4.645.686	10.357.143	17.428.571	14.071.429
std	2.138.597.163	12.849.706.465	5.500.882.529	3.930.258.594	19.959.713.568	1.221.415.032	1.509.061.691	734.325.930	11.989.810.853	11.031.448.064	734.325.930	19.959.713.568	1.509.061.691	85.113.660	0.043986	0.064497	5.271.216	21.664.271	13.606.923
min	0.000000	28.000.000	0.000000	21.000.000	42.000.000	0.000000	0.000000	0.000000	28.000.000	21.000.000	0.000000	42.000.000	0.000000	7.000.000	-74.194.500	4.548.600	1.000.000	1.000.000	1.000.000
25%	117.000.000	1.150.500.000	165.000.000	175.500.000	1.523.250.000	81.000.000	81.000.000	0.000000	861.000.000	891.750.000	0.000000	1.523.250.000	81.000.000	41.250.000	-74.136.200	4.599.150	7.250.000	2.250.000	4.250.000
50%	315.000.000	3.973.500.000	1.763.000.000	1.324.000.000	7.845.000.000	175.000.000	352.000.000	98.500.000	4.432.500.000	4.490.000.000	98.500.000	7.845.000.000	352.000.000	94.500.000	-74.104.050	4.637.650	10.500.000	8.500.000	9.500.000
75%	2.664.000.000	18.079.500.000	4.169.500.000	4.278.000.000	30.888.000.000	1.349.000.000	1.246.500.000	697.500.000	18.987.000.000	16.689.750.000	697.500.000	30.888.000.000	1.246.500.000	183.750.000	-74.082.800	4.679.000	14.500.000	19.000.000	24.500.000
max	6.604.000.000	36.740.000.000	16.002.000.000	11.500.000.000	56.642.000.000	3.740.000.000	5.060.000.000	2.032.000.000	32.766.000.000	31.750.000.000	2.032.000.000	56.642.000.000	5.060.000.000	254.000.000	-74.031.200	4.765.200	18.000.000	65.000.000	40.000.000

	Desconocido	En estudio	Importado	Relacionado	Casa	Fallecido	Hospital	Hospital UCI	F	M	Critico	Moderado	Severo	CASES	LONGITUD	LATITUD	CODIGO	NUMOPEN	NUMCLOSE
Desconocido	1.000.000	0.917055	0.874892	0.927016	0.975226	0.864588	0.788230	0.909438	0.971463	0.972757	0.909438	0.975226	0.788230	0.927827	0.204603	0.759177	-0.283637	-0.194209	-0.176270
En estudio	0.917055	1.000.000	0.718332	0.831578	0.935164	0.983283	0.959071	0.976134	0.967033	0.946036	0.976134	0.935164	0.959071	0.939073	0.008553	0.619601	-0.338834	-0.217746	-0.210877
Importado	0.874892	0.718332	1.000.000	0.958479	0.913843	0.680910	0.498831	0.741994	0.866932	0.904233	0.741994	0.913843	0.498831	0.857858	0.495760	0.870082	-0.464639	-0.074837	-0.071273
Relacionado	0.927016	0.831578	0.958479	1.000.000	0.969135	0.785698	0.644270	0.810571	0.942937	0.957730	0.810571	0.969135	0.644270	0.931167	0.388200	0.845139	-0.477298	-0.092491	-0.108602
Casa	0.975226	0.935164	0.913843	0.969135	1.000.000	0.898210	0.797591	0.922856	0.994242	0.998719	0.922856	1.000.000	0.797591	0.973230	0.251514	0.797974	-0.422446	-0.160401	-0.160264
Fallecido	0.864588	0.983283	0.680910	0.785698	0.898210	1.000.000	0.960365	0.978900	0.936611	0.914451	0.978900	0.898210	0.960365	0.895127	0.017615	0.557187	-0.312448	-0.229120	-0.204578
Hospital	0.788230	0.959071	0.498831	0.644270	0.797591	0.960365	1.000.000	0.931321	0.856137	0.817727	0.931321	0.797591	1.000.000	0.826064	-0.181758	0.422829	-0.231428	-0.228692	-0.216086
Hospital UCI	0.909438	0.976134	0.741994	0.810571	0.922856	0.978900	0.931321	1.000.000	0.950265	0.939300	1.000.000	0.922856	0.931321	0.904882	0.044855	0.608908	-0.312615	-0.220579	-0.178118
F	0.971463	0.967033	0.866932	0.942937	0.994242	0.936611	0.856137	0.950265	1.000.000	0.996128	0.950265	0.994242	0.856137	0.973753	0.185544	0.752299	-0.398385	-0.179569	-0.177138
M	0.972757	0.946036	0.904233	0.957730	0.998719	0.914451	0.817727	0.939300	0.996128	1.000.000	0.939300	0.998719	0.817727	0.974908	0.233484	0.786223	-0.418419	-0.166388	-0.161514
Critico	0.909438	0.976134	0.741994	0.810571	0.922856	0.978900	0.931321	1.000.000	0.950265	0.939300	1.000.000	0.922856	0.931321	0.904882	0.044855	0.608908	-0.312615	-0.220579	-0.178118
Moderado	0.975226	0.935164	0.913843	0.969135	1.000.000	0.898210	0.797591	0.922856	0.994242	0.998719	0.922856	1.000.000	0.797591	0.973230	0.251514	0.797974	-0.422446	-0.160401	-0.160264
Severo	0.788230	0.959071	0.498831	0.644270	0.797591	0.960365	1.000.000	0.931321	0.856137	0.817727	0.931321	0.797591	1.000.000	0.826064	-0.181758	0.422829	-0.231428	-0.228692	-0.216086
CASES	0.927827	0.939073	0.857858	0.931167	0.973230	0.895127	0.826064	0.904882	0.973753	0.974908	0.904882	0.973230	0.826064	1.000.000	0.141743	0.793683	-0.486598	-0.146647	-0.180211
LONGITUD	0.204603	0.008553	0.495760	0.388200	0.251514	0.017615	-0.181758	0.044855	0.185544	0.233484	0.044855	0.251514	-0.181758	0.141743	1.000.000	0.366038	-0.140080	0.467962	0.524950
LATITUD	0.759177	0.619601	0.870082	0.845139	0.797974	0.557187	0.422829	0.608908	0.752299	0.786223	0.608908	0.797974	0.422829	0.793683	0.366038	1.000.000	-0.509969	-0.013758	0.086284
CODIGO	-0.283637	-0.338834	-0.464639	-0.477298	-0.422446	-0.312448	-0.231428	-0.312615	-0.398385	-0.418419	-0.312615	-0.422446	-0.231428	-0.486598	-0.140080	-0.509969	1.000.000	0.087472	0.065038
NUMOPEN	-0.194209	-0.217746	-0.074837	-0.092491	-0.160401	-0.229120	-0.228692	-0.220579	-0.179569	-0.166388	-0.220579	-0.160401	-0.228692	-0.146647	0.467962	-0.013758	0.087472	1.000.000	0.892067
NUMCLOSE	-0.176270	-0.210877	-0.071273	-0.108602	-0.160264	-0.204578	-0.216086	-0.178118	-0.177138	-0.161514	-0.178118	-0.160264	-0.216086	-0.180211	0.524950	0.086284	0.065038	0.892067	1.000.000

From the evaluation of the models I got the following scores using Jaccard's index and F1-score, which indicate good results, especially for the logistic regression.

Algorithm	Jaccard	F1-score
SVM	0.837545	0.827067
Logistic regression	0.898917	0.851066

5. Discussion

From the results a lot of information may be deducted. First, it was evidenced that there is contagion between localidades that are far away and this is due to the movement of the people across the city for daily activities. Therefore, it may be inferred that the restrictions on mobility may limit the spread of the disease to localidades where a lower number of cases has been reported. This may help avoid what happened in Kennedy. Also, it was found that the death rate varies between localidades. This is an important fact because it points at the fact that some localidades may have greater risks for mortality such as heath risk factors (obesity, smoking, breathing problems, etc.) or may have more limited resources at hospitals. As a consequence, the government might be interested in implementing higher restrictions for mobility in this localidades.

Also, it was confirmed that most of the cases are moderate and the infected are at home. This means that fatalities are low and that the hospital system may have time to prepare for the peak of the curve. The graph on the kind of cases indicates that most are under study. There is a high amount of imported cases; however, this number is not expected to grow much as airports and frontiers have been closed. It will be interesting to study the dynamic of this graph, as it may be an indicator of the dynamics of the disease among the population. A good result will be indicated by a low number of unknown cases, as it is the case right now.

In the descriptive statistics, there seem to be more open businesses than closed ones. This may actually not be true, since most of the economy has stopped. Actually, this may be a result of the bias in the venues data, since a lot of the information comes from places that may be interested to tourist. This includes an overrepresented number of restaurants and venues related to food.

The correlation analysis shows that there is a high correlation between many of the variables. Part of this is due to the fact that many are generated as dummies. It is also interesting the high correlation between opened and closed venues. This could be due to the fact that places that are more touristic have a higher number of places in foursquare and the proportion between opened and closed is similar between localidades.

Finally, the models generated have a very good prediction power and may be interesting in predicting the disease dynamics.

6. Conclusion

The information available from COVID-19 is a very interesting source to get insights into how the disease is developing in different locations. I think it is important that governments open this data so that the general population may gain a better understanding of how decisions affect the spread. Also, although it is very interesting, is important to note that information coming from websites such as foursquare may be biased. This may impact the results and give a distorted view of the actual data. Finally, the generation of models such as the ones presented may be a fundamental tool in understanding data and planning for the future.