Introduction

Lately, Bogota has become a great place for doing tourism and business deals. [1] Bogota has slightly over 8 million [2] of people live, and it has population density around 4000 people per square kilometer. Colombian capital does not have enough public transportation supply. Moreover, its public transportation is overcrowding, and it has poor quality service, as well. [4]

You have noticed that Bogota is not a kind place to take a bus or a cab. Furthermore, businessmen and tourists need to move across many boroughs along this city. A comfortable option might be rent a car, so they would rather avoid neighborhoods which have higher traffic accident rate. Another point that we should analyze is to identify similar districts for understanding whether it makes sense go to some communities or not.

Once, we analyze which are our requirements we can propose a map that is going to illustrate traffic accident rate among all boroughs and how they are related.

Data Description

After understanding the problem, I made some decisions about which data I would need and how to gather it. Finally, I decided that my data sources were:

- I found a geo-json file of Bogota and I uploaded it on an Object Storage. [6]

- I used the option 'Search Nearby' on Google Maps to obtain approximately the center coordinates of each Borough. I did manually this task.

- I used Foursquare API to get many venues as possible given center coordinates of each Bogota's district.

- Colombian government has a public web page where anyone can get data about any demographic measures. I chose traffic accidents in Bogota in 2017 (It is the most updated report). [2]

Methodology

I used as notebook container Watson Studio, and I kept my notebook and data sources on an object storage. First, I gathered data al data necessary, so I initialized a dictionary with Bogota's borough coordinates:

	Borough	Latitude	Longitude
1	Usaquen	4.723275	-74.036574
2	Chapinero	4.660363	-74.053383
3	Santa Fe	4.609981	-74.069833
4	San Cristobal	4.562843	-74.089615
5	Usme	4.506400	-74.108000
6	Tunjuelito	4.583216	-74.137272
7	Bosa	4.623055	-74.195987
8	Kennedy	4.627078	-74.151358
9	Fontibon	4.676155	-74.136057
10	Engativa	4.700832	-74.109040
11	Suba	4.741000	-74.084000
12	Barrios Unidos	4.670186	-74.073411
13	Teusaquillo	4.639346	-74.077037
14	Los Martires	4.606241	-74.089889
15	Antonio Nariño	4.587906	-74.099494
16	Puente Aranda	4.613310	-74.114055
17	La Candelaria	4.597014	-74.072150
18	Rafael Uribe Uribe	4.572363	-74.114523
19	Ciudad Bolivar	4.536111	-74.138889
20	Sumapaz	4.260000	-74.178333

The next step was obtaining all venues around each coordinate by using Foursquare API. I decided to set 2.500 meter as radius, and I set maximum 100 of them per borough. I joined that result-set with the above table in order to match each venue with its borough. The joined table was:

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Usaquen	4.723275	-74.036574	Wok	4.719637	-74.037489	Asian Restaurant
1	Usaquen	4.723275	-74.036574	El Kiosco golosinas	4.722658	-74.034035	Snack Place
2	Usaquen	4.723275	-74.036574	Chopinar	4.724496	-74.032496	Buffet
3	Usaquen	4.723275	-74.036574	Parque Cedritos	4.723525	-74.033089	Park
4	Usaquen	4.723275	-74.036574	Harvey	4.726911	-74.035721	Fast Food Restaurant

I wanted to explore that data for understanding what I just gathering. Consequently, I grouped each district's venues and I counted them:

Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Antonio Nariño	100	100	100	100	100	100
Barrios Unidos	100	100	100	100	100	100
Bosa	18	18	18	18	18	18
Chapinero	100	100	100	100	100	100
Ciudad Bolivar	8	8	8	8	8	8
Engativa	100	100	100	100	100	100
Fontibon	100	100	100	100	100	100
Kennedy	87	87	87	87	87	87
La Candelaria	100	100	100	100	100	100
Los Martires	100	100	100	100	100	100
Puente Aranda	100	100	100	100	100	100
Rafael Uribe Uribe	100	100	100	100	100	100
San Cristobal	53	53	53	53	53	53
Santa Fe	100	100	100	100	100	100
Suba	100	100	100	100	100	100
Teusaquillo	100	100	100	100	100	100
Tunjuelito	71	71	71	71	71	71
Usaquen	100	100	100	100	100	100
Usme	7	7	7	7	7	7

I noticed some boroughs did not have enough venues, less than 54 percent. Therefore, I made a decision of dropping them because they might generate noise. The chosen ones were Bosa, Ciudad Bolivar, San Cristobal and Usme. Finally, I needed to become categorical data into numerical, in other words, each venue would become an attribute.

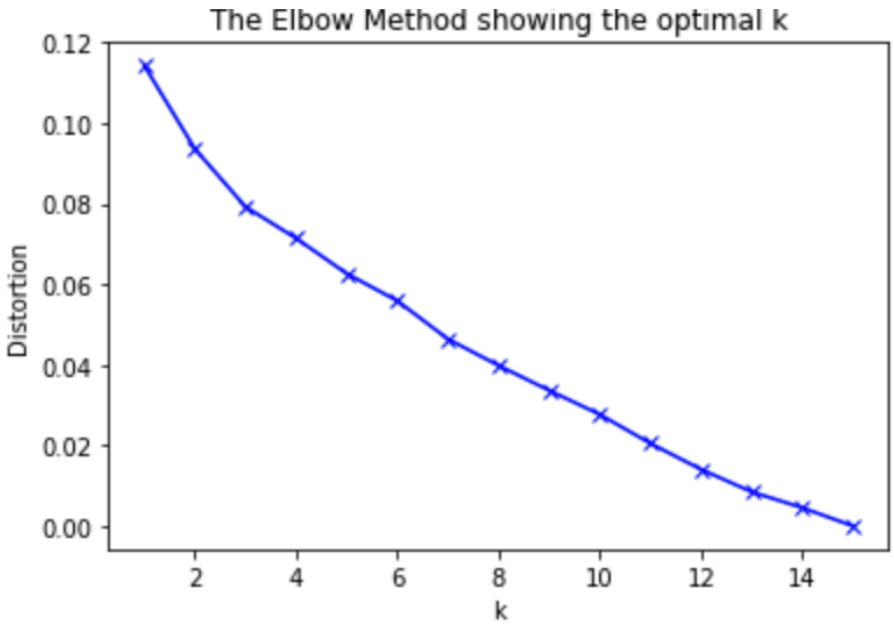
The last step was calculating the frequency of each venue per district:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue
0	Antonio Nariño	Restaurant	Café	Park	Pizza Place	History Museum	Latin American Restaurant	Coffee Shop	Seafood Restaurant	South American Restaurant	French Restaurant	Gym	Hotel	Pub	Plaza	Historic Site
1	Barrios Unidos	Restaurant	Hotel	Coffee Shop	Latin American Restaurant	Bakery	Italian Restaurant	Seafood Restaurant	Pizza Place	Asian Restaurant	Café	Park	Lounge	Bowling Alley	Wings Joint	Gym
2	Chapinero	Restaurant	Hotel	Coffee Shop	Bakery	Italian Restaurant	French Restaurant	Latin American Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Burger Joint	Pizza Place	BBQ Joint	Pub	Wings Joint	Breakfast Spot
3	Engativa	Hotel	Restaurant	Bakery	Pizza Place	Fast Food Restaurant	Coffee Shop	Ice Cream Shop	Bar	Café	Park	Latin American Restaurant	Brewery	Seafood Restaurant	Clothing Store	Cocktail Bar

I explored that data, and I found that some boroughs could be similar. I did not know how to make relationship among them. Therefore, I decided to use K-means algorithm.

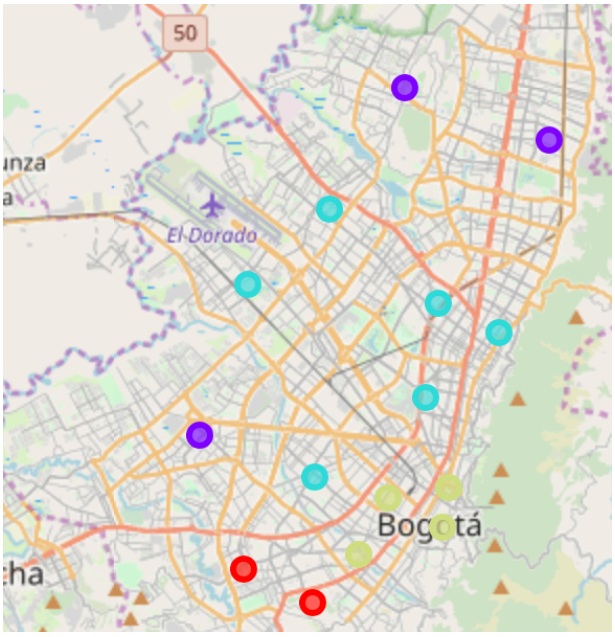
It was necessary to calculate the number of clusters which would be a local optimal point. To achieve this number, I implemented elbow method and I plotted it as well.

The graph was:



Elbow method suggested 4 may be number of clusters, then I merged cluster results with previous data.

I plotted a map with markers for illustrating the clusters



For answering the main question, I also required to import traffic accident data of Bogota.

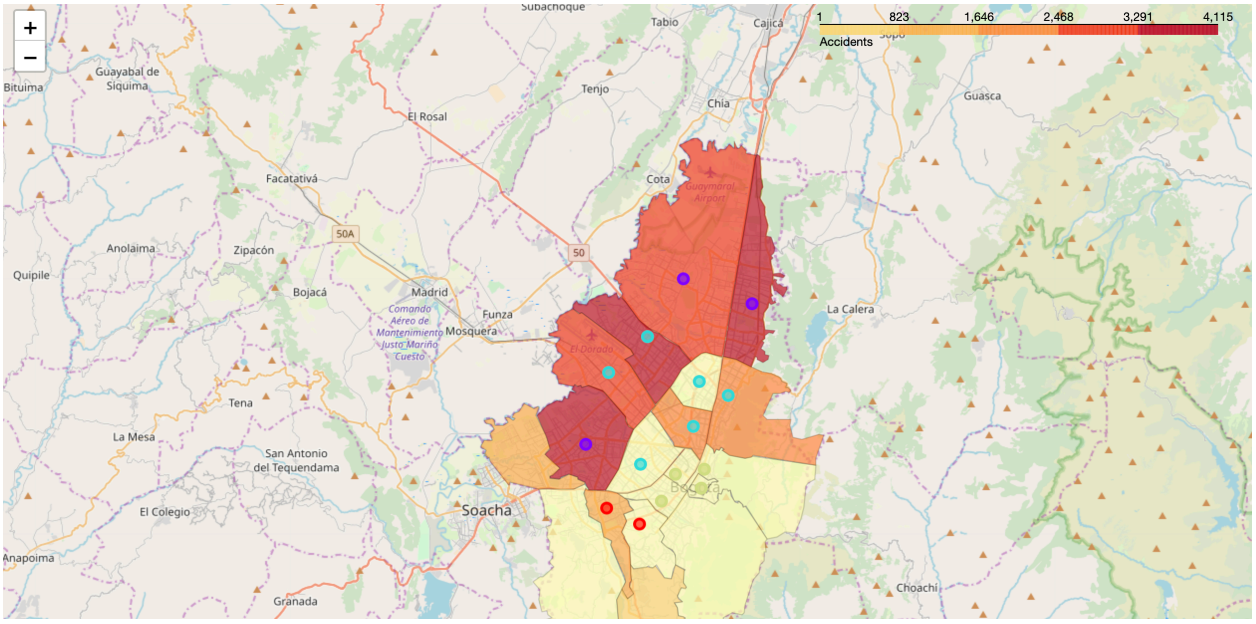
	Día	Fecha	GravedadCod	GravedadNombre	ClaseCodigo	ClaseNombre	Municipio	Localidad	FechaOcurrencia	HoraOcurrencia
0	SÁBADO	4/29/17 0:00	2	Con Heridos	2	Atropello	BOGOTA D.C.	FONTIBON	4/29/17 0:00	12/31/1899 03:40:00 PM
1	SÁBADO	5/6/17 0:00	2	Con Heridos	2	Atropello	BOGOTA D.C.	KENNEDY	5/6/17 0:00	12/31/1899 03:15:00 AM
2	DOMINGO	5/7/17 0:00	1	Con Muertos	2	Atropello	BOGOTA D.C.	KENNEDY	5/7/17 0:00	12/31/1899 04:40:00 AM
3	VIERNES	5/5/17 0:00	3	Solo Daños	3	Volcamiento	BOGOTA D.C.	BARRIOS UNIDOS	5/5/17 0:00	12/31/1899 02:00:00 PM
4	DOMINGO	4/9/17 0:00	1	Con Muertos	2	Atropello	BOGOTA D.C.	TUNJUELITO	4/9/17 0:00	12/31/1899 12:38:00 AM

I transformed and cleaned that data to get useful one. The result was:

	Borough	Quantity of Accidents
0	Kennedy	4114
1	Usaquen	3766
2	Engativa	3592
3	Suba	3073
4	Fontibon	2821

Results

I plotted a choropleth map with all joined variables. I got:



You can see which boroughs have more accidents and how they are clustered.

Discussion

Bogota is a big metropole that have over 8 million people, and it has not had a good city planification because of size of boroughs. Some of them are much bigger than others. For instance, Candelaria district has 1.84 squared kilometer [7] while Ciudad Bolivar has 20.88 squared kilometer [8]. As I said, I dropped some district because of shortage of data. Consequently, I could not analyze all of them.

I use K-means to analyze relationship among Bogota's borough. Additionally, I got a local optimal number of clusters. I think that I would get better result, If I used neighborhoods coordinates instead of district ones. Another recommendation is to use several coordinates from the same district in order to have more accuracy.

I got an overall result that can help businessmen and tourists to make a good decision.

Conclusion

In summarize, people can analyze how a city is and can get insights before they visit it. It is possible to obtain the correct data, but it is mandatory you explore and understand it.

References

[1] https://www.apnews.com/Business%20Wire/bd4954f98b9540c89a1eafd76d4a2b99
[2] Wikipedia
[3] http://worldpopulationreview.com/world-cities/bogota-population/
[4] https://en.wikipedia.org/wiki/Bus_rapid_transit
[5]https://el-tiempo.carto.com/u/datoseltiempo/tables/localidades_bogota/public?redirected=true

[6] https://www.datos.gov.co/Transporte/2017-ACCIDENTES-DE-TR-NSITO-BOGOT-/tcva-ksr4/data
[7] https://es.wikipedia.org/wiki/La_Candelaria
[8] https://es.wikipedia.org/wiki/Ciudad_Bol%C3%ADvar_(Bogot%C3%A1)