IBM Data Science specialization

Capstone Project

Final Report

Luxury goods store at Mexico City Borough

Santana Abad Luis Carlos

2020

Table of content

1. Introduction to the business problem

This business problem is part of the first part of the **IBM Data Science specialization**. The purpose of this capstone is to define a problem or an idea of my choice, where I would need to leverage the Foursquare location data to solve or execute. According to these criteria, I've decided to create a case described below.

Find opportunities in the Mexican market for **Luxury Goods** is a task that we can achieve with the right Data Analysis. According to statista.com the luxury market size in Mexico is about **US $3.584 M**, one of the largest in Latin America.

In Mexico City, a client / contractor is interested in knowing where to open a luxury goods store at Coyoacan borough of Mexico City, it is important to know the economic development of the neighborhoods, as well as to know the **10** businesses closest common to know the competition and preferences of the area to study in order to discover the right place to open a store to be successful.

2. Data acquisition and cleaning

As is the case when shopping for ingredients to make a meal, some ingredients might be out of season and more difficult to obtain or cost more than initially thought, collecting data requires that you know the source or, know where to find the data elements that are needed.

Development in neighborhoods reflects social and economic progress and requires economic growth, which means in potential clients to sell, therefore I decided to include the economic development for each neighborhood in Coyoacan. This data was provided by the Mexican government website.

Zip codes I needed to search was on the list at Here API in order to get the coordinates from the neighborhoods with this data, I could be able to graph a map with all the neighborhoods in the city.

Finally, used the Foursquare API to get the data of the nearest venues in different neighborhoods to make a further analysis of the goods that will be selling.

Data URL's:

Zip Code and neighborhoods from Mexico City

Here API to know the neighborhoods location

Social Development Index by neighborhood, Mexico City neighborhood

Foursquare API to make an analysis of the venues nearest to the Neighborhoods

3. Methodology

From Understanding to Preparation.

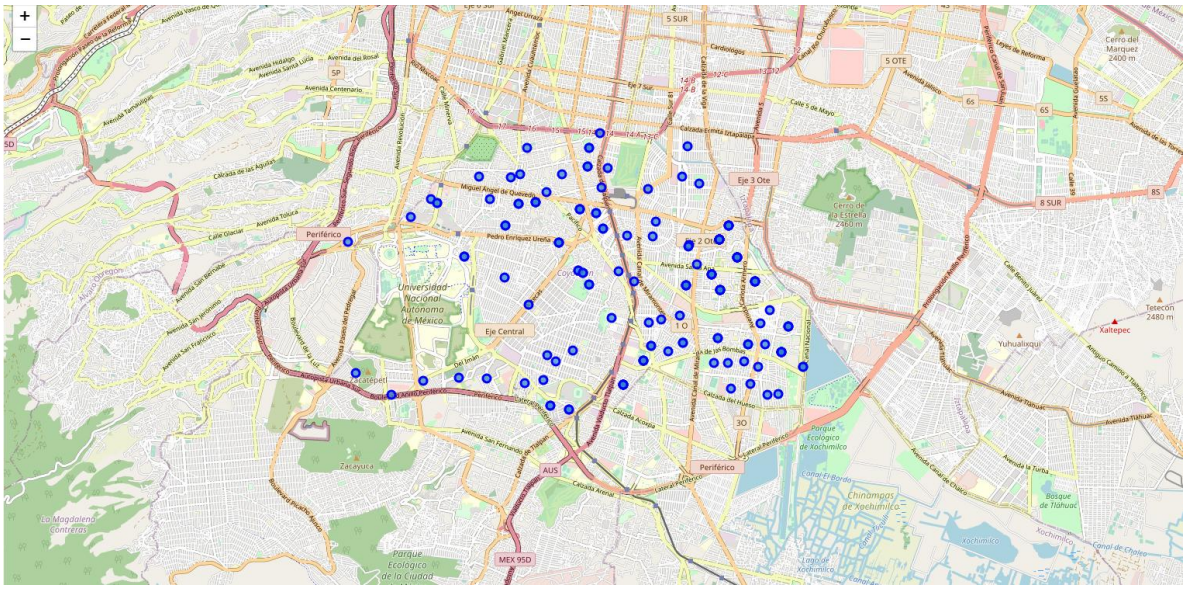
The Zip Code data frame was containing a lot of information, however I drop all the columns that I didn't need and reduce to get just the neighborhood and the zip code.

	d_codigo	d_asenta	d_tipo_asenta	D_mnpio	d_estado	d_ciudad	d_CP	c_estado	c_oficina	c_CP	c_tipo_asenta	c_mnpio	id_asenta_cpcons	d_zona	c_cve_ciudad
0	1000	San Ángel	Colonia	Álvaro Obregón	Ciudad de México	Ciudad de México	1001	9	1001	NaN	9	10	1	Urbano	1
1	1010	Los Alpes	Colonia	Álvaro Obregón	Ciudad de México	Ciudad de México	1001	9	1001	NaN	9	10	5	Urbano	1
2	1020	Guadalupe Inn	Colonia	Álvaro Obregón	Ciudad de México	Ciudad de México	1001	9	1001	NaN	9	10	6	Urbano	1
3	1030	Axotla	Pueblo	Álvaro Obregón	Ciudad de México	Ciudad de México	1001	9	1001	NaN	28	10	9	Urbano	1
4	1030	Florida	Colonia	Álvaro Obregón	Ciudad de México	Ciudad de México	1001	9	1001	NaN	9	10	10	Urbano	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1534	16840	Santa Cruz Chavarrieta	Colonia	Xochimilco	Ciudad de México	Ciudad de México	16001	9	16001	NaN	9	13	2519	Urbano	16
1535	16850	Chapultepec	Barrio	Xochimilco	Ciudad de México	Ciudad de México	16001	9	16001	NaN	2	13	2520	Urbano	16
1536	16860	Santa Cruz de Guadalupe	Colonia	Xochimilco	Ciudad de México	Ciudad de México	16001	9	16001	NaN	9	13	2521	Urbano	16
1537	16880	Santa Cecilia Tepetlapa	Pueblo	Xochimilco	Ciudad de México	Ciudad de México	16001	9	16001	NaN	28	13	2527	Urbano	16
1538	16900	San Francisco Tlalnepantla	Pueblo	Xochimilco	Ciudad de México	Ciudad de México	16001	9	16001	NaN	28	13	2530	Urbano	16

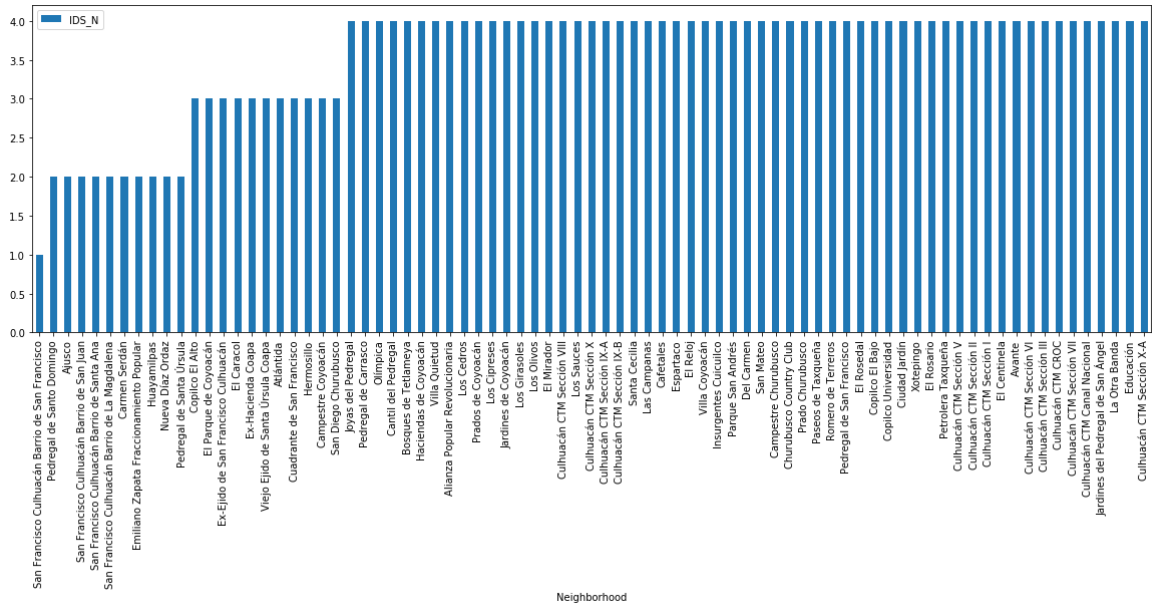
When I merged the data frame generated by Here API it was transformed in this:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	4000	Coyoacán	Villa Coyoacán	19.34850	-99.16331
1	4010	Coyoacán	Santa Catarina	19.34809	-99.17231
2	4020	Coyoacán	La Concepción	19.34480	-99.15750
3	4030	Coyoacán	San Lucas	19.34863	-99.15401
4	4040	Coyoacán	Parque San Andrés	19.34576	-99.14520

Into a map of Coyoacan boroughs with all the neighborhoods represented by blue points:



In the next graph, I plotted the economic development for each neighborhood in Coyoacan, it is represented in a range from 1 to 4 max**.**

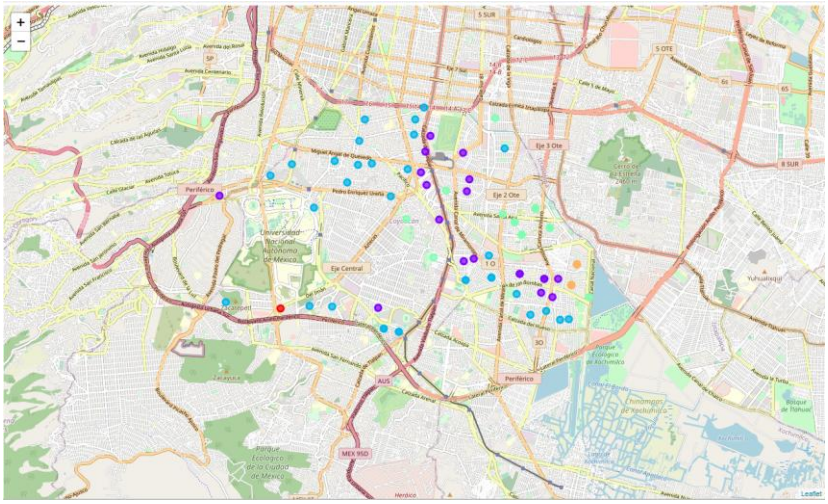


Then I cleaned the data and processed the data filtering the data to only worked with the higher economic development, to train a machine learning algorithm called **"K-means"** which show list of top 10 venue categorized for each neighborhood.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
0	4000	Coyoacán	Villa Coyoacán	19.34850	-99.16331	2	Mexican Restaurant	Coffee Shop	Ice Cream Shop	Café	Plaza	Bookstore	Bar	Art Gallery	Bed & Breakfast	Food Truck
4	4040	Coyoacán	Parque San Andrés	19.34576	-99.14520	1	Mexican Restaurant	Pharmacy	Taco Place	Restaurant	Gym / Fitness Center	Salon / Barbershop	Vegetarian / Vegan Restaurant	Hotel	Office	Smoke Shop
5	4100	Coyoacán	Del Carmen	19.35412	-99.16167	2	Ice Cream Shop	Mexican Restaurant	Café	Plaza	Bakery	Theater	Bar	Liquor Store	Spanish Restaurant	Juice Bar
8	4120	Coyoacán	San Mateo	19.35402	-99.14794	2	Taco Place	Café	Mexican Restaurant	Coffee Shop	History Museum	Convenience Store	Pharmacy	Restaurant	Argentinian Restaurant	BBQ Joint
9	4120	Coyoacán	San Mateo	19.35020	-99.14834	2	Taco Place	Café	Mexican Restaurant	Coffee Shop	History Museum	Convenience Store	Pharmacy	Restaurant	Argentinian Restaurant	BBQ Joint

K-means aims to partition "n" observations into "k" clusters, grouped in base of similar features, this was plotted on a map and showed the neighborhood similarity.

4. Results



The cluster number 1 which is at Insurgentes Cuicuilco neighborhood one the most economically developed in Coyoacan is the best place for a Luxury Goods store.

	Neighborhood	Cluster Labels	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
107	Insurgentes Cuicuilco	0	Cosmetics Shop	Ice Cream Shop	Jewelry Store	Lingerie Store	Snack Place	Candy Store	Costume Shop	Salad Place	Furniture / Home Store	Movie Theater

5. Discussion

The model can always be improved for as long as the solution is needed, regardless of whether the improvements came by examine newly available data sources. The methodology contributed additional value to business units by incorporating data science practices into their daily analysis and reporting functions. Is evident by the fact that the contractor / client is able to make a better decision using a machine learning algorithm. Learned the meaning of a methodology that its purpose is to explain how to look at a problem, work with data in support of solving the problem, and come up with an answer that addresses the root problem. The success of the data model depends on the ability to apply the right tools, at the right time, in the right order, to the address the right problem.

6. Conclusion

The quality of a data after processing and cleaning, along with the tools that we use, help us to generate value information. The methodology can be applied in context, toward successfully achieving the goals that were set out in the business requirements stage to achieve better performance, not only in business, we could apply it to everything, and the reason is because of the greater availability of data and models and analytics to be able to pinpoint where the greatest needs are.

7. References:

https://www.coursera.org/professional-certificates/ibm-data-science