Coursera Capstonse Project: The Battle of Neighborhoods

IBM Data Science Professional Certificate

Luis Avila[1]

May, 2019

***Abstract*— The need of knowledge of what kind of neighborhood a person is going to live is very important, to fulfill the expectations and to have a comfortable life in the new place of residence. Sometimes there are opportunities to relocate but not enough time to research more about the new place. In this work is proposed an analysis of different neighborhoods and its nearby venues, in order to compare what is the most suitable neighborhood to live according to a prior neighborhood in another country.**

The comparison between Tecnologico, Monterrey and neighborhoods of Austin takes place, making use of an Exploratory Data Analysis and K-Means clustering. Findings shows that Lower Waller Creek is the most suitable neighborhood to live in Austin, and that is similar to the neighborhood in Mexico.

I. INTRODUCTION

A. Background

Growing cities suppose challenges that many disciplines have to overcome. Construction space, type of land use, mobility, accessibility, streets and highways with enough space for transit, supermarkets and stores to supply basic necessities.

Every time people moves from other localities to their nearest city for many reasons, mainly being for a higher lifestyle, a higher paying job, access to basic services, supermarkets and amenities. When people consider moving to a city they need to know what is the best place to move, according to their profile.

B. Problem

Workers from medium and big companies are often required to relocate in another city and sometimes with no much time in advance in order to make a proper research to know what is the best place to live in the new city. The lack of time to make a research about what is the best place to live can lead to have an unpleasant stay while living in this new city.

An analysis is performed for the case of the company worker Luis, who is moving from Monterrey, Mexico to Austin, Texas due to his company requirements. We need to identify what are the current characteristics of the place were Luis lives, and to compare if there is any similar place in Austin, Texas.

*This work was not supported by any organization

C. Objective

Analyse and identify the neighborhood with the highest percentage of similarity compared with the origin city.

Neighborhood segmentation according to venue information acquired using Foursquare API. Data obtaining and cleaning using web scrapping techniques. Obtaining of latitude and longitude of each one of the cities using the geocoder package. Feature engineering and clustering using k-means algorithm.

The final deliverable is a recommendation based on the percentage of affinity for both areas in Monterrey and Austin.

II. DATA AND METHODS

A. Data Sources

The description of the data to be used in this project is as follows:

- **Foursquare API:** Provides location information of nearby venues given an address. The Foursquare API will be used to extract nearby places for every neighborhood in Monterrey and in Austin. The neighborhood segmentation will take place using this information to extract particular characteristics of each neighborhood.
- **List of Monterrey Postal Codes (1):** Monterrey postal codes are used to extract the main database that are used for this project.
- **List of Austin Postal Codes (2):** Austin postal codes will be used to extract the main database that will be used for this project.

B. Methodology

Web scrapping is used to transform web lists that contains the postal codes for Monterrey and Austin. The creation of a database in form of a pandas dataframe provide the structure required to apply the algorithms proposed. Data cleaning and feature preprocessing is required to avoid duplicates in data, null values and correct data formatting and type.

Figure 1 is an example of the postal codes for Monterrey, Mexico. The list consist of Postal Code, Borough and Neighborhood. In Figure 2 shows a sample of the zip codes available for Austin, Texas. Similar as the Postal codes list from Mexico, the Austin list includes zip code, type and county, also the population for that county is included.

México » Nuevo León » Monterrey

Listado de todos los Códigos Postales de Monterrey, Nuevo León

Buscar en esta tabla (Opcional)

Asentamiento	Tipo de Asentamiento	Código Postal	Municipio	Ciudad	Zona	Mapa
1 de Mayo (F-97)	Colonia	64220	Monterrey	Monterrey	Urbana	Mapa
10 de Junio	Colonia	64268	Monterrey	Monterrey	Urbana	Mapa
10 de Marzo	Colonia	64488	Monterrey	Monterrey	Urbana	Mapa
13 de Junio	Colonia	64780	Monterrey	Monterrey	Urbana	Mapa

Fig. 1. Example of Monterrey Postal codes list on internet

Austin, TX Covers 74 ZIP Codes

ZIP Code	Type	County	Population	Area Code(s)
ZIP Code 73301	Unique	Travis	0	512
ZIP Code 73344	Unique	Travis	0	512
ZIP Code 78681	Standard	Williamson	50,606	512
ZIP Code 78701	Standard	Travis	6,841	512 / 737
ZIP Code 78702	Standard	Travis	21,334	512 / 737
ZIP Code 78703	Standard	Travis	19,690	512
ZIP Code 78704	Standard	Travis	42,117	512 / 737
ZIP Code 78705	Standard	Travis	31,340	512 / 737
ZIP Code 78708	P.O. Box	Travis	0	512
ZIP Code 78709	P.O. Box	Travis	0	512
ZIP Code 78710	Standard	Travis	0	512 / 737
ZIP Code 78711	P.O. Box	Travis	0	512
ZIP Code 78712	Standard	Travis	860	512 / 737
ZIP Code 78713	P.O. Box	Travis	0	512
ZIP Code 78714	P.O. Box	Travis	0	512
ZIP Code 78715	P.O. Box	Travis	0	512
ZIP Code 78716	P.O. Box	Travis	0	512
ZIP Code 78717	Standard	Williamson	22,538	512
ZIP Code 78718	P.O. Box	Travis	0	512
ZIP Code 78719	Standard	Travis	1,764	512
ZIP Code 78720	P.O. Box	Travis	0	512
ZIP Code 78721	Standard	Travis	11,425	512
ZIP Code 78722	Standard	Travis	5,901	512 / 737
ZIP Code 78723	Standard	Travis	28,330	512 / 737

Fig. 2. Example of Austin Postal codes list on internet

As described before, the information obtained from the two data sources scraped from the web contains the data required to build a database for both cities. A python package for obtaining the latitude and longitude of each neighborhood is used to complement the information for the analysis.

The Geocoder package is used to transform postal codes and boroughs into Latitude and Longitude coordinates to be used while creating the maps for the two cities. Once the creation of the latitude and longitude fields of the cities is done the exploratory data analysis is performed and the search for venues using the Foursquare API takes place.

Code example for obtaining San Francisco, California coordinates:

```
import geocoder

g = geocoder.arcgis('San Francisco, CA')
g.latlng
```

After obtaining the venues for each cityś neighborhoods they are displayed in a map to visually inspect the elements to proceed to implement the K-Means Clustering algorithm.

The K-Means clustering labels the neighborhoods according to their venues characteristics. The algorithm is applied for each city and the results are presented in a map and in the form of a table, to extract useful patterns. After this step it is selected the Monterrey neighborhood that the test subject lives, extract the venue pattern created from the K-Means labeling and proceed to fin the Austin neighboor with a percentage of similarity of 70% or above.

III. RESULTS

The data from Monterrey in shown in figure 3. Extracted information from Neighborhood and extracted the geographic coordinates, the map was also generated.

	Asentamiento	Tipo de Asentamiento	Código Postal	Municipio	Ciudad	Zona	Mapa	Latitude	Longitude
0	Monterrey Centro	Colonia	64000	Monterrey	Monterrey	Urbana	Mapa	25.664332	-100.317274
1	Nuevo Centro de Monterrey	Equipamiento	64018	Monterrey	Monterrey	Urbana	Mapa	25.671916	-100.300973
2	Condominios Constitución	Unidad habitacional	64019	Monterrey	Monterrey	Urbana	Mapa	25.668419	-100.300588
3	Gonzalitos	Colonia	64020	Monterrey	Monterrey	Urbana	Mapa	25.684598	-100.350584
4	Deportivo Obispado	Colonia	64040	Monterrey	Monterrey	Urbana	Mapa	25.678081	-100.340836
5	María Luisa	Colonia	64040	Monterrey	Monterrey	Urbana	Mapa	25.675132	-100.338444
6	Obispado	Colonia	64060	Monterrey	Monterrey	Urbana	Mapa	25.674287	-100.334677
7	Barrio Antiguo	Colonia	64100	Monterrey	Monterrey	Urbana	Mapa	43.502635	-1.454455
8	Nuevo Repueblo	Colonia	64700	Monterrey	Monterrey	Urbana	Mapa	25.656749	-100.302061
9	Roma	Colonia	64700	Monterrey	Monterrey	Urbana	Mapa	25.658876	-100.296769

Fig. 3. Monterrey Dataframe with Latitude and Longitude

Latitude and Longitude from Austin Neighborhoods were extracted, using the Geocoder package and the ArcGIS platform to make the information extract, due that is open source and better suitable for our application. The extracted information is used to generate the first map in this exercise. The resulting dataframe is shown in figure 4.

	Neighborhood	City	State	Latitude	Longitude
0	Bryker Woods	Austin	Texas	30.305007	-97.754204
1	Caswell Heights	Austin	Texas	30.307865	-97.719418
2	Downtown Austin	Austin	Texas	30.265270	-97.746470
3	Eastwoods	Austin	Texas	30.371418	-97.748008
4	Hancock	Austin	Texas	30.297150	-97.726620

Fig. 4. Austin Dataframe with Latitude and Longitude

The Monterrey map shows all the neighborhoods studied in this work. A total of 31 different neighborhoods covering the central part of the city were analized. Due to the bast amount of data encountered, it was decided to narrow the experiment to the mentioned before neighborhoods, because the final comparison is only between one neighborhood from Monterrey and one neighborhood from Austin. Monterrey map is shown in figure 5.

As the same way as the Monterrey map, the Austin map shows the relevant information for the chosen neighborhoods. Due to there was not public information available in form of table about what neighborhoods compose Austin, it was decided to manually extract what neighborhoods are available to analyze and then start to construct the database. In figure 6 is shown the Austin map, composed of the neighborhoods chosen for this analysis.

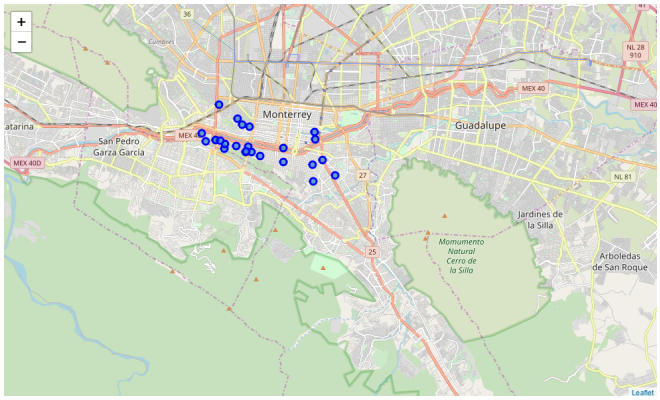

Fig. 5. Monterrey Map

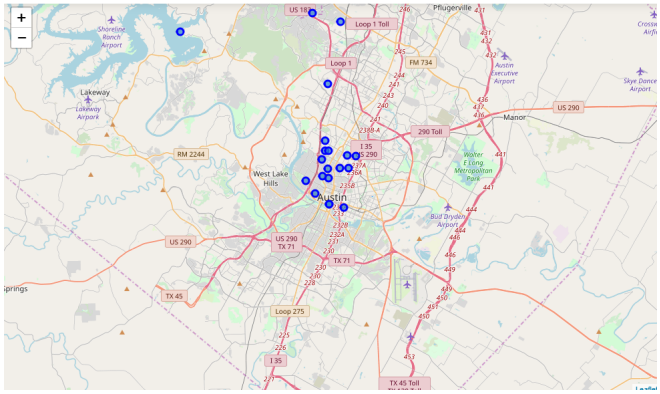

Fig. 6. Austin Map

After obtaining the complete database information of longitude and latitude, necessary for the construction of the maps, the next step is to extract the venue information from the Foursquare API,

The Foursquare API can extract what places are nearby of each one of our neighborhoods. The information can include Coffee shops, bakery, restaurants, fitness centers, parks, schools, between others. The information of the venues were extracted for both Monterrey and Austin neighborhoods, in order to know more about this places. In figure 7 are listed some of the venues available in the Monterrey neighborhoods. While in figure 8 were listed the venues from Austin neighborhoods.

The venues extracted were used to get what are the most common venues by neighborhoods, for each one of our cities studied. In the first place in Monterrey, the figure 9 are listed the most common venues by neighborhood. The most common venues were extracted using the average of venue category listed in each one of the neighborhoods. For Austin the top venues were also calculated by the average occurrence of the venue category in each one of the neighborhoods. Austin top venues for each neighrbood are shown in 10.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Ancira	3	3	3	3	3	3
Balcones del Carmen	3	3	3	3	3	3
Barrio Antiguo	5	5	5	5	5	5
Centro	2	2	2	2	2	2
Comercial Ampliación Doctores	12	12	12	12	12	12
Condominios Constitución	20	20	20	20	20	20
Del Carmen	14	14	14	14	14	14
Deportivo Obispado	11	11	11	11	11	11
El Maguey	7	7	7	7	7	7
Gonzalitos	34	34	34	34	34	34
Hacienda San Francisco	16	16	16	16	16	16
Independencia	9	9	9	9	9	9
Loma Larga	23	23	23	23	23	23
Lomas de San Francisco	20	20	20	20	20	20
Los Doctores	12	12	12	12	12	12
Los Magueyes	6	6	6	6	6	6
María Luisa	15	15	15	15	15	15
Monterrey Centro	58	58	58	58	58	58
Nuevas Colonias	6	6	6	6	6	6
Nuevo Centro de Monterrey	27	27	27	27	27	27
Nuevo Repueblo	13	13	13	13	13	13
Obispado	44	44	44	44	44	44
Pío X	7	7	7	7	7	7
Roma	24	24	24	24	24	24
Roma Sur	12	12	12	12	12	12
Sertoma	24	24	24	24	24	24
Tecnológico	100	100	100	100	100	100

Fig. 7. Monterrey venues extracted from Foursquare

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bryker Woods	23	23	23	23	23	23
Caswell Heights	31	31	31	31	31	31
Downtown Austin	100	100	100	100	100	100
Hancock	3	3	3	3	3	3
Heritage	5	5	5	5	5	5
Hyde Park	15	15	15	15	15	15
Lower Waller Creek	79	79	79	79	79	79
North University	4	4	4	4	4	4
Oakmont Heights	3	3	3	3	3	3
Old Enfield	4	4	4	4	4	4
Old Pecan Street	24	24	24	24	24	24
Old West Austin	43	43	43	43	43	43
Original Austin	2	2	2	2	2	2
Original West University	28	28	28	28	28	28
Pemberton Heights	6	6	6	6	6	6
Ridgelea	10	10	10	10	10	10
Ridgetop	2	2	2	2	2	2
Rosedale	4	4	4	4	4	4
Shoal Crest	10	10	10	10	10	10
West Downtown	100	100	100	100	100	100

Fig. 8. Austin Venues extracted from Foursquare

	Neighborhood	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	Ancira	Mexican Restaurant	Bed & Breakfast	Market	Yoga Studio	Drugstore	Construction & Landscaping	Convenience Store	Cosmetics Shop	Cuban Restaurant	Cupcake Shop
1	Balcones del Carmen	Pharmacy	Casino	Mexican Restaurant	Yoga Studio	College Quad	Concert Hall	Construction & Landscaping	Convenience Store	Cosmetics Shop	Cuban Restaurant
2	Barrio Antiguo	Yoga Studio	Supermarket	Water Park	Pharmacy	Hotel	Dessert Shop	Department Store	Deli / Bodega	Dance Studio	Cupcake Shop
3	Centro	Construction & Landscaping	Automotive Shop	Yoga Studio	Electronics Store	Concert Hall	Convenience Store	Cosmetics Shop	Cuban Restaurant	Cupcake Shop	Dance Studio
4	Comercial Ampliación Doctores	Mexican Restaurant	Pharmacy	Convenience Store	College Cafeteria	Café	Salad Place	Casino	Pizza Place	Diner	Cosmetics Shop

Fig. 9. Monterrey Top Venues by Neighborhood

	Neighborhood	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	Bryker Woods	Park	Italian Restaurant	Yoga Studio	Gift Shop	Pizza Place	Pharmacy	Music Store	Massage Studio	Garden Center	Rental Car Location
1	Caswell Heights	Bus Stop	Gym	Ice Cream Shop	Pet Store	Beer Store	Breakfast Spot	Pool Hall	Pool	Pizza Place	Comedy Club
2	Downtown Austin	Coffee Shop	Hotel	Restaurant	New American Restaurant	Italian Restaurant	Bar	Gay Bar	Music Venue	Clothing Store	Lounge
3	Hancock	Golf Course	Intersection	Park	Yoga Studio	Fast Food Restaurant	Farmers Market	Fabric Shop	Event Service	Donut Shop	Dog Run
4	Heritage	Construction & Landscaping	Bus Stop	Trail	Thrift / Vintage Store	Dog Run	Dive Bar	Fast Food Restaurant	Farmers Market	Fabric Shop	Event Service

Fig. 10. Austin Top Venues by Neighborhood

The clustering method applied is the K-Means. The algorithm takes the data and place a n-quantity of points randomly put on the data. The algorithm tries to find the k-nearest neighbor to this centroid, and calculates the average distance. Every time the algorithm loops through this cycle, the average distance is minimized until the process get the minimum distance possible.

For this exercise this means that the algorithm finds as is centroid for each cluster, the average distance for the points. The points in this dataset are the average venues available in each one of the neighborhoods, meaning that the process classifies the neighborhoods according to the types of venues near them. The number of clusters given to the algorithm was 5.

The Monterrey map in figure 11 shows the result of the clustering. Most of the neighborhoods correspond to the first cluster, where are located a variety of venues, including coffee shops, restaurants and taco places. The other clusters have parks or construction and landscape venues.

Similar to the Monterrey clustering, the map shown in figure 12 correspond to the map of Austin after the clustering technique were applied. And also as the Monterrey clustering, in Austin most of the venues were labeled in the first cluster, where this cluster have more variety of venues as other neighborhoods.

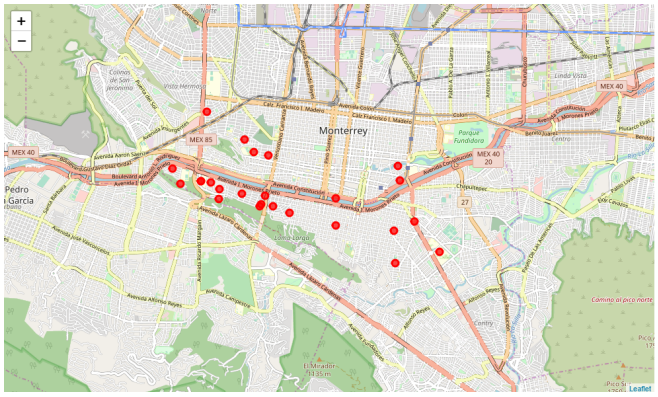

Fig. 11. Monterrey Map after Clustering applied

After analyzing the data obtained, the neighborhood Tecnologico, which at the start of this work was the point of interest correspond to the first cluster, with Taco Place, Mexican Restaurant, Italian Restaurant, COffee Shop and Restaurant as the most common venues for this cluster.

Analogously, the neighborhood Lower Waller Creek also corresponding to cluster 1, share most of the venues observed in Tecnologico, Monterrey, such as Coffee Shop, Mexican Restaurant and Taco Place, having some differences as Dive Bar, Cocktail Bar and Ice Cream Shop, but we can assume this neighborhood in Austin is pretty similar in terms of

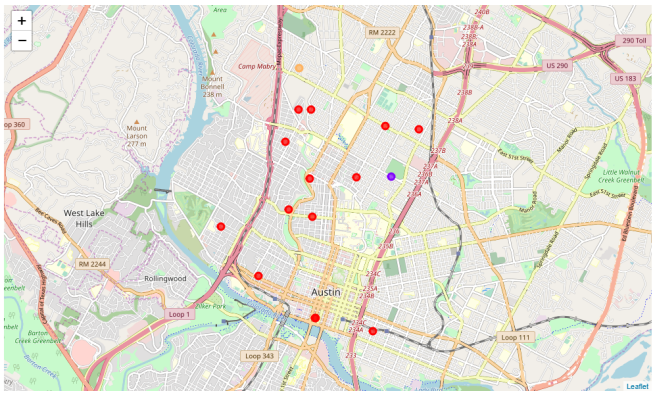

Fig. 12. Austin Map after Clustering applied

venues as the neighborhood in Monterrey.

IV. DISCUSSION

A high percentage of real world data sets are not available to be used directly to extract some insights. The real world have messy and diffuse data, sometimes not reachable very easily and some times even not available to the public. The data available for this work was not too much detailed, so we proceeded to extract manually some information in order to complete the database that serves for the construction of the dataset, the map and the clustering. Foursquare was a valuable tool that let us to extract venue information for the neighborhoods analyzed. The information included in Foursquare is very valuable and can be take advantage of with GIS projects on data science.

Regarding the final steps on this work, it was found that there is one similar neighborhood in Austin, and that was the point of this project, to make a recommendation about what neighborhood is similar enough for a person who is moving to another country and another city that he cannot had the opportunity to explore in advance. The recommendation is the Lower Waller Creek neighborhood, because it shares most of the common venues from cluster 1.

V. CONCLUSIONS

In this work we analized the information of neighborhoods corresponding to Monterrey and Austin, to find the best suitable neighborhoood that shares most of the venues from Tecnologico, Monterrey, for a person who is moving from Mexico to Texas and have no prior information about the new city.

The information of the neighborhoods were obtained from web pages, some manual handling took place because all the information was not available. The venues nearby were obtained with the Foursquare API and a exploratory data analysis took place to know more about this venues.

The K-Means clustering technique was used to classify the neighborhoods into the most suitable categories. After the completion of the algorithm it was found that Lower Waller

Creek corresponds in most of the venues to Tecnologico, Monyterrey. The recommendation obtained is that Lower Waller Creek is the most suitable neighborhood to live according to the similarities with the mexican neighborhood.

ACKNOWLEDGMENT

Thanks to the IBM Team for offering the Data Science courses available through Coursera. The courses syllabus, videos and practice labs are high quality resources that help introduce us to the Data Science world.

REFERENCES

[1] Monterrey List of Postal Codes. (n.d.). Retrieved April 11, 2019, from https://micodigopostal.org/nuevo-leon/monterrey/

[2] List of Austin, Texas Postal Codes. (n.d.). Retrieved April 11, 2019, from https://www.zip-codes.com/city/tx-austin.asp