Capstone project - Battle of the nieghborhoods Week 2

Data science course IBM – Coursera

Pedro Cisneros Garza

March 23rd, 2020

Introduction

In Mexico the census of the whole population is taken **every 10 years** (as in the US), and according to the gathered information some classifications are created based on the numbers obtained, one of those classifications is the *ambit of an area, those could be rural or urban* (depending on the population of each area).

Given that the census is taken every 10 years, the conditions of the population change and some areas classified as rural could be on the **verge of becoming an urban areas**.

As an area transits from rural to urban the services needed differ and knowing which areas are on this process of urbanization is helpful to service providers and government institutes in order to provide the necessary services to this areas.

For this project I will be using the information from Mexico. The country is divided into states, each states has municipalities and each municipality has neighborhoods. The information gathered is from the Toluca municipality, located inside the State of Mexico; one of the biggest states in Mexico in terms of population and economic activity.

Data

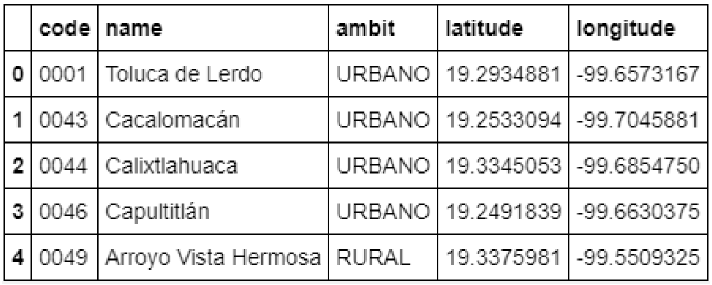
In this project I will use the data available from the National Institute of Information, Stadistics and Georaphy (*INEGI* for its initials in Spanish).

The institute provides a repository with geographic information regarding the neighborhoods of each municipality which contains the ambit (rural or urban) as well as the coordinates (latitude and londigutde) of each neighboorhood.

As requested in the project the information about venues will be consulted using Foursquare API.

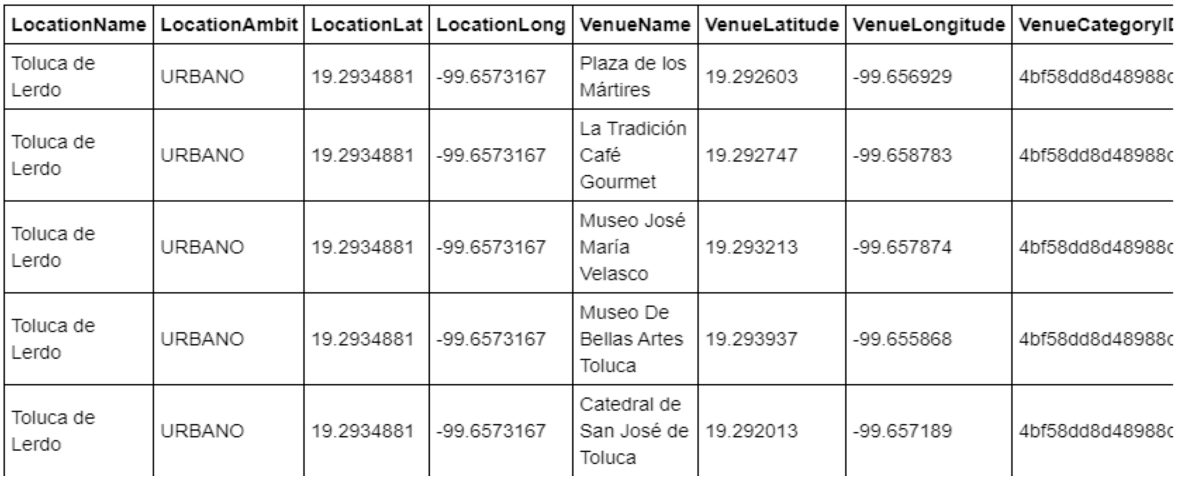
Also to avoid multitude of categories used on Foursquare a process is considered to retrieve the top level category of each venue.

Example of the location data obtained from INEGI web site.



	code	name	ambit	latitude	longitude
0	0001	Toluca de Lerdo	URBANO	19.2934881	-99.6573167
1	0043	Cacalomacán	URBANO	19.2533094	-99.7045881
2	0044	Calixtlahuaca	URBANO	19.3345053	-99.6854750
3	0046	Capultitlán	URBANO	19.2491839	-99.6630375
4	0049	Arroyo Vista Hermosa	RURAL	19.3375981	-99.5509325

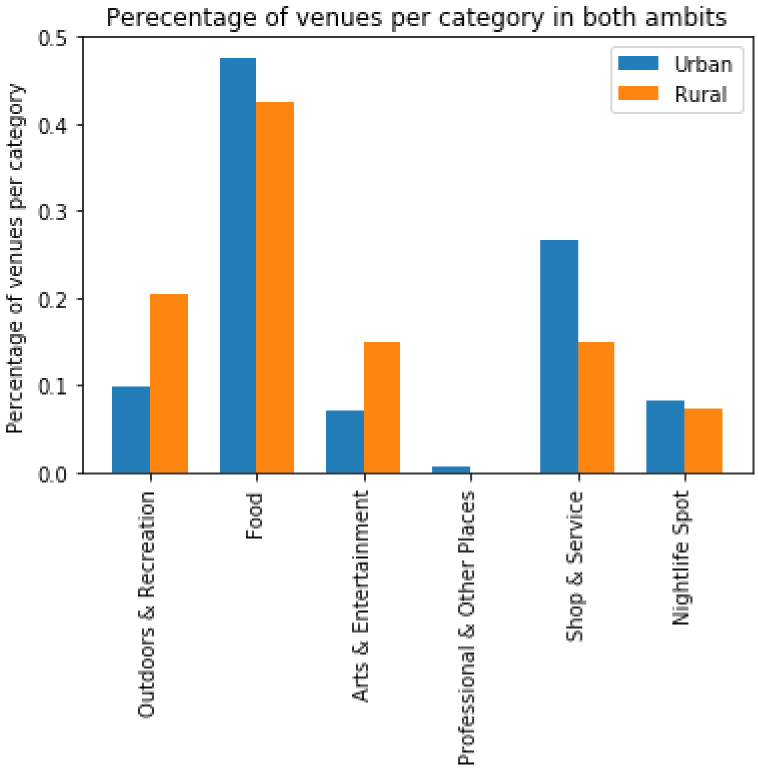
Example of the venue data gathered



LocationName	LocationAmbit	LocationLat	LocationLong	VenueName	VenueLatitude	VenueLongitude	VenueCategoryI
Toluca de Lerdo	URBANO	19.2934881	-99.6573167	Plaza de los Mártires	19.292603	-99.656929	4bf58dd8d48988c
Toluca de Lerdo	URBANO	19.2934881	-99.6573167	La Tradición Café Gourmet	19.292747	-99.658783	4bf58dd8d48988c
Toluca de Lerdo	URBANO	19.2934881	-99.6573167	Museo José María Velasco	19.293213	-99.657874	4bf58dd8d48988c
Toluca de Lerdo	URBANO	19.2934881	-99.6573167	Museo De Bellas Artes Toluca	19.293937	-99.655868	4bf58dd8d48988c
Toluca de Lerdo	URBANO	19.2934881	-99.6573167	Catedral de San José de Toluca	19.292013	-99.657189	4bf58dd8d48988c

Exploratory analysis of data

As initial exploration of the data adquired, let's see a bar chart of the percentage of venues per category from both ambits



As we can appreciate on the bar chart there is not big difference on the percentages of venues per category among both ambits across all categories.

Methodology

In the project we will review two methods to predict which neighborhoods are in the process of urbanization:

>Logistic regresion (supervised method)
This method is used to predict a categorical value (rural or urban) with a given set of independent data

>Clustering (unsupervised method)
For this we will use K-Means for clustering rural neighborhoods given the quantity of venues around them

Logistic regresion

For this model we took all the data gathered and splitted on 70% training and 30% for testing purposes. Given that this data set is reduce the method to train this model is liblinear solver.

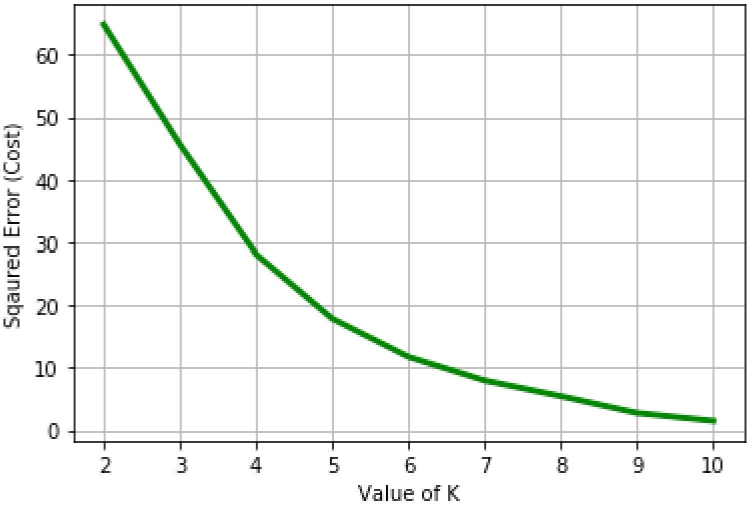
This are the results on the model

Jaccard score 0.5

	precision	recall	f1-score	support
0	0.38	0.75	0.50	4
1	0.86	0.55	0.67	11
accuracy			0.60	15
macro avg	0.62	0.65	0.58	15
weighted avg	0.73	0.60	0.62	15

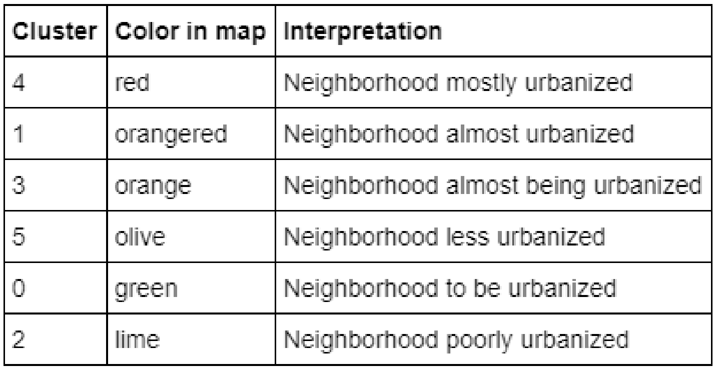
K-Means

To apply K-Means clustering I excluded the urban areas, as we are interested in the characteristics of rural areas and how those can indicate a lever or urbanization of the area.
The first thing to to do is find the best K for the method, for this purpose we iterate from a K value of 2 up to 11 and created the correspondent graphic.



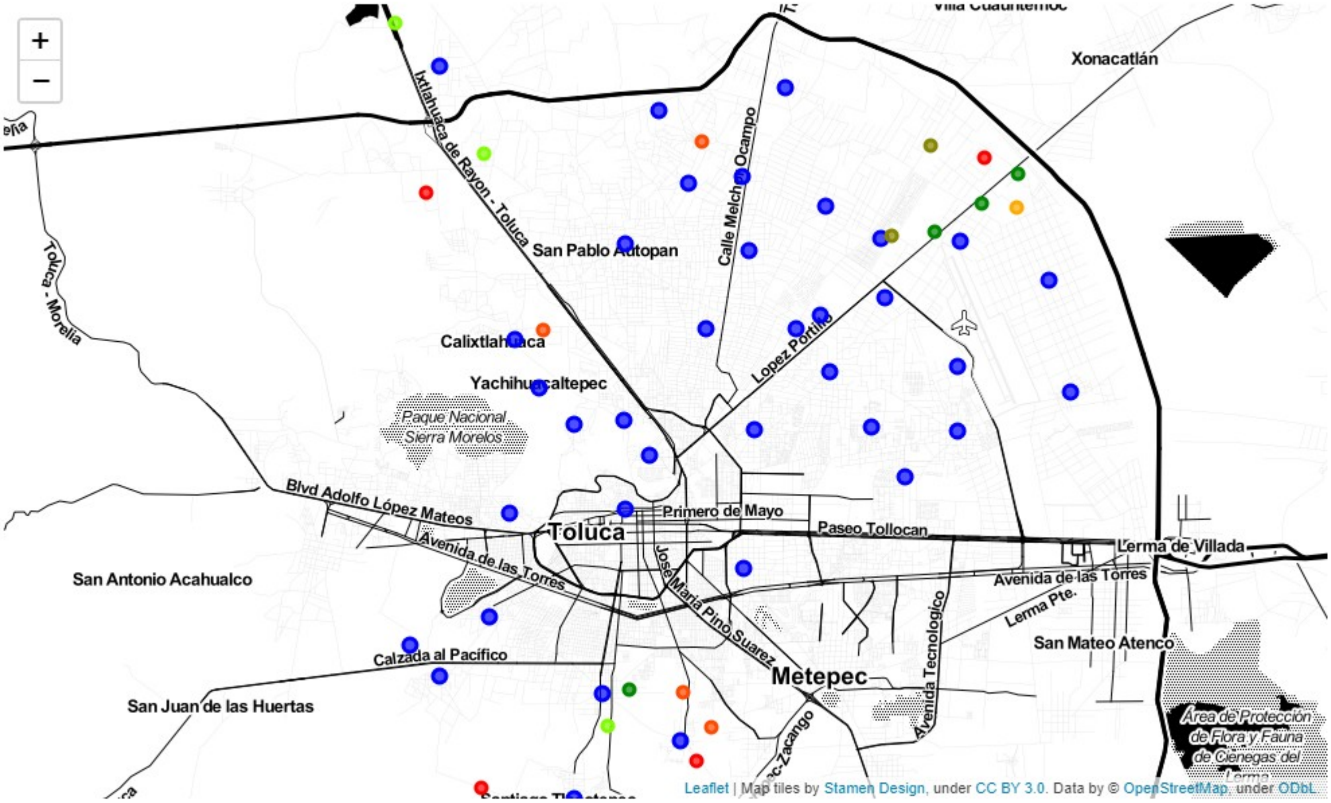
Using the elbow method we find that the best K for this data set is 6.

After runing the method we obtained the following clusters and a possible interpretation of each cluster.



Cluster	Color in map	Interpretation
4	red	Neighborhood mostly urbanized
1	orangered	Neighborhood almost urbanized
3	orange	Neighborhood almost being urbanized
5	olive	Neighborhood less urbanized
0	green	Neighborhood to be urbanized
2	lime	Neighborhood poorly urbanized

Below we see a map with the urban areas in blue and clustered area with the color indicated in the chart above.



Results

Logistic regresion model

As seen by the Jaccard score and accuracy in classification report the data set is not the most appropriate to predict if a given neighboorhood is rural or urban according to the summary of venues, this could be originated in the fact that Toluca municipality is almost urban and the difference between these two ambits is no longer that evident.

K-Means

Using the clustering of K-Means we can see that despite the fact that almost all the municipality is becoming urbanized, there are neighboorhoods which are ahead of this process.

As we can see the first three clusters (from red to orange) are located near the airport and also have been an increase of industrial parks created in that zone.

In the other hand we can see in clusters 3-6 (from olive to lime) the venues are more related to recreation and shops and this concurs the real state developments around that zone of the municipality of Metepec.

Discusion

Finding which neighborhoods are going trough a process of urbanization is an important task due to the different services a community needs.

Unfortunately the first approach was not highly accurate given the situation of the municipality studied; however I believe this could be achieved with a different set of data in which the difference of these two ambits is more broad.

This step back on a prediction algorithm allow me to explore other methods; and using an automatic clustering I was able to find some insights of this neighborhoods and how are they changing. Luckily this work is being done in year 2020 and the census result can be expected soon in order to validate the methods and findings of this work.

Conclusion

The results of this excercise could be useful to service providers such as mobile conections, broadband and electricity just to name a few particular providers.

These results could be useful to government agencies on the three levels (municipality, state and federal) as the changes in the population lead to the need of new roads, increase of police force and water supply as some examples.