IBM-COURSERA DATA SCIENCE CAPSTONE

Author: Jesus Leonardo Zorrilla

Title: Santo Domingo neighborhoods analysis for new business opening

Date: 16/11/2020

Introduction

The city is a complex set of systems working in a way that allow human complex interaction and extense range of activities. Those systems intersect on geographical space in a way that can be recognized, that we can identify as a common/similar to, as a known piece/organ, that we can describe and study.

Those patters, the blocks that form a city, are compounded by visible and invisible features that are by themselves products of the evolving of the initial and sequent set of rules, and will contribute to the next stage and apparition of new features.

The business that exists in those areas are a byproduct and a door to the info natively encrypted in the city space, so as this study goes and by using some data processing and modeling tools like pandas and sklearn we are going to try uncover sufficient information to create a recommendation model that understanding those patters Identify places that can be center of new store installations.

As location is one of the key points to be taken in consideration for new Businesses; creating and developing tools that can help us take a better understanding of the city, market and opportunities to take important decisions is a must.

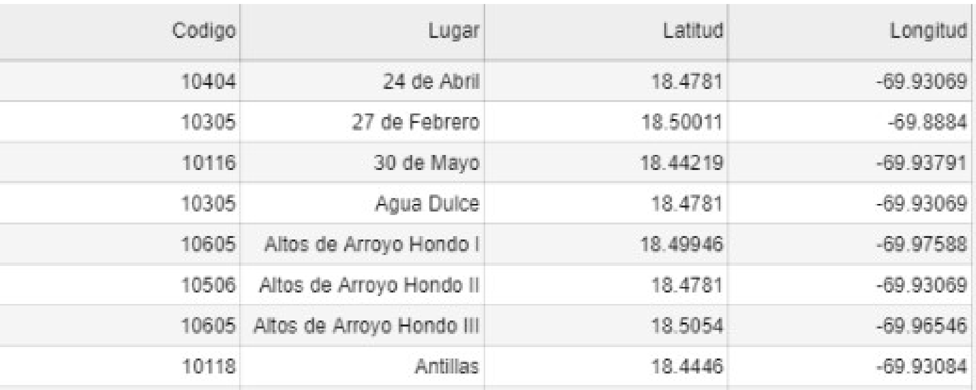
The Problem

The development of this study has aim on **decision making administrative personal, business owners and entrepreneurs** that want to ingress on new Business projects and that need some way to interpret the signals the city present in a way there can be strategical decisions that handle or benefit from existing circumstances.

We are going to concentrate on Pizza Places, on the cities of **Santo Domingo, Distrito Nacional** in the Dominican Republic, as well as the city of **Santiago de los Caballeros** also on the DR. As a requirement for this study, we are going to use Foursquare API to get the venues places, as well as some public data to get the neighborhoods lists and coordinates.

The Data

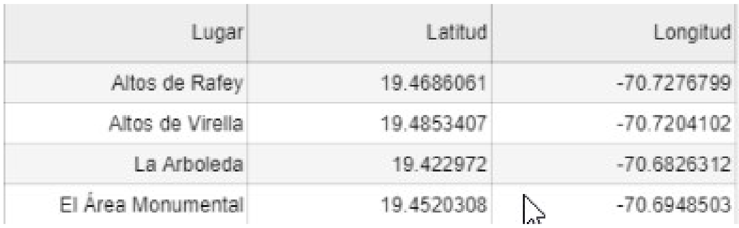
A list of Santo Domingo Neighborhoods was extracted on a csv from xn--cdigos-postales-vrb.cybo.com, where there is the name, coordinates and zip codes, but as I found, the coordinates (Latitude, Longitude) that the source provide didn't represent the center location for some of the areas we were going to use, so there was an initial batch of query's to **Google Places API**, that returned the neighborhood coordinates adjusted. Those are present on a second Santo Domingo neighborhoods csv.



Codigo	Lugar	Latitud	Longitud
10404	24 de Abril	18.4781	-69.93069
10305	27 de Febrero	18.50011	-69.8884
10116	30 de Mayo	18.44219	-69.93791
10305	Agua Dulce	18.4781	-69.93069
10605	Altos de Arroyo Hondo I	18.49946	-69.97588
10506	Altos de Arroyo Hondo II	18.4781	-69.93069
10605	Altos de Arroyo Hondo III	18.5054	-69.96546
10118	Antillas	18.4446	-69.93084

Figure 1 Extract of the Santo Domingo Zip Codes CSV

Also we got from the **Plan de Ordenamiento Territorial (2018)** of the local government of Santiago City a list of their principal neighborhoods, also by using **Google Places API** we got the coordinates of the neighborhood geometric center and saved to a CSV dataframe.

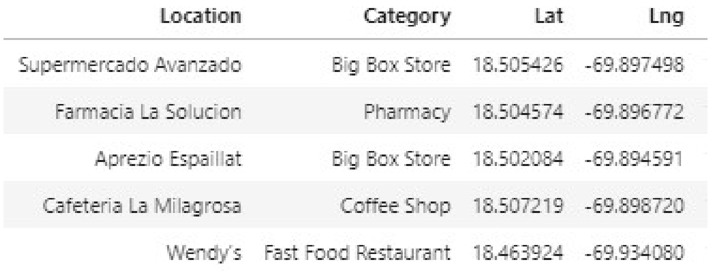


Lugar	Latitud	Longitud
Altos de Rafey	19.4686061	-70.7276799
Altos de Virella	19.4853407	-70.7204102
La Arboleda	19.422972	-70.6826312
El Área Monumental	19.4520308	-70.6948503

2 Extract of Santiago de los Caballeros Neighborhoods with coordinates

Follium maps library is used to present graphically the geographic data using OpenStreetMaps tiles and technology.

Foursquare API was used to get a list of venues for all the neighborhoods for both cities. Those venues list contain data like the name of the venue, the category, latitude and longitude.



Location	Category	Lat	Lng
Supermercado Avanzado	Big Box Store	18.505426	-69.897498
Farmacia La Solucion	Pharmacy	18.504574	-69.896772
Aprezio Espaillat	Big Box Store	18.502084	-69.894591
Cafeteria La Milagrosa	Coffee Shop	18.507219	-69.898720
Wendy's	Fast Food Restaurant	18.463924	-69.934080

Figure 3 Venues ítems

Data Cleaning

For the city of Santiago, we will drop all the venues not appearing in Santo Domingo as a measure for normalizing the present categories in both cities.

Methodology

The data compiled will be used to create a profile for all the neighborhoods, and by examining the venues categories and density composition we will train Kmean classificator to infer similarities between zones and cluster them on similar type super zones.

Those super zones will allow us to detect and analyze characteristics of each zone not as a unique event, but as a common signal of development.

Finally, we will train some Classifications Models, to infer where does a Pizza Place exist, and where it should be installed. Those classificators will be tested and refined to get a relative accurate prediction using correlation matrix, and simple manual maps analytical analysis.

Results

For each neighborhood of the city of Santo Domingo, Distrito Nacional, using its geometric center coordinates, we created a list of venues that fall as far as 650mts radius.

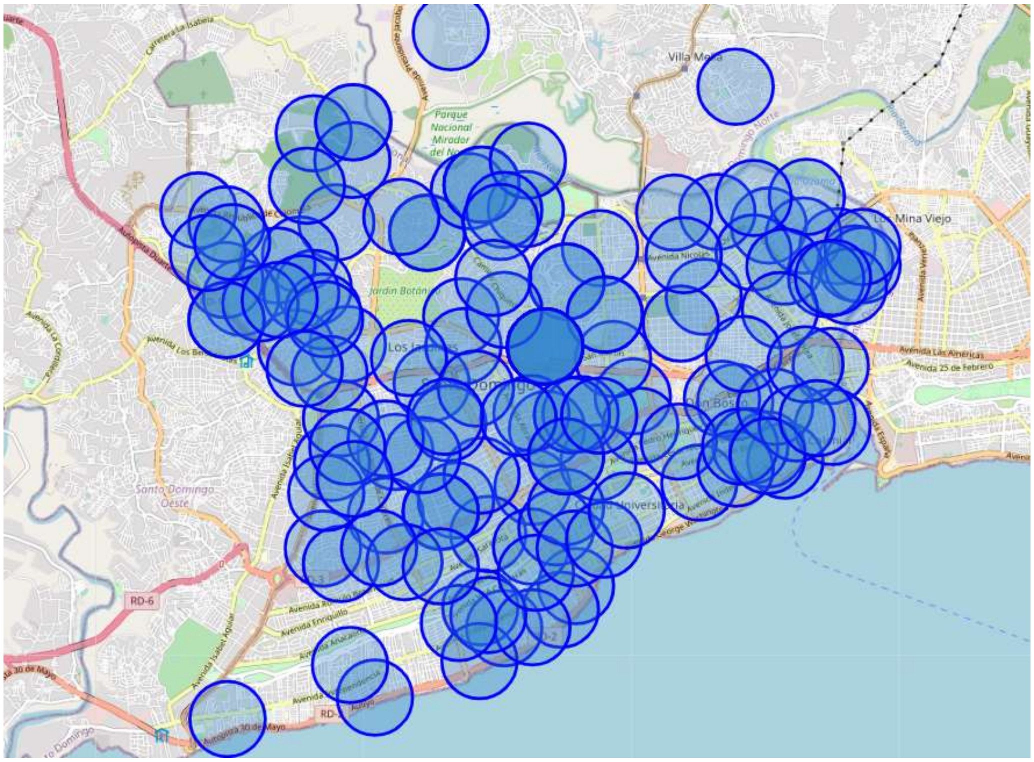


Figure 4 Neighborhoods radius

Using KMeans we were able by using the venues categories compositions for each neighborhood to make a Cluster classification witch can be seen in the next graphic.

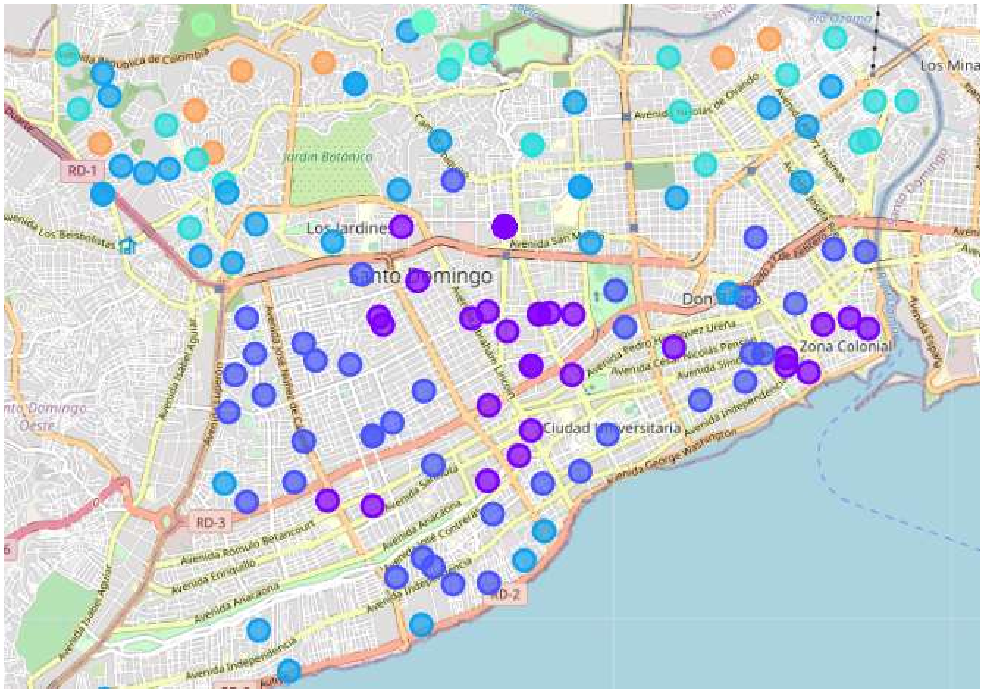
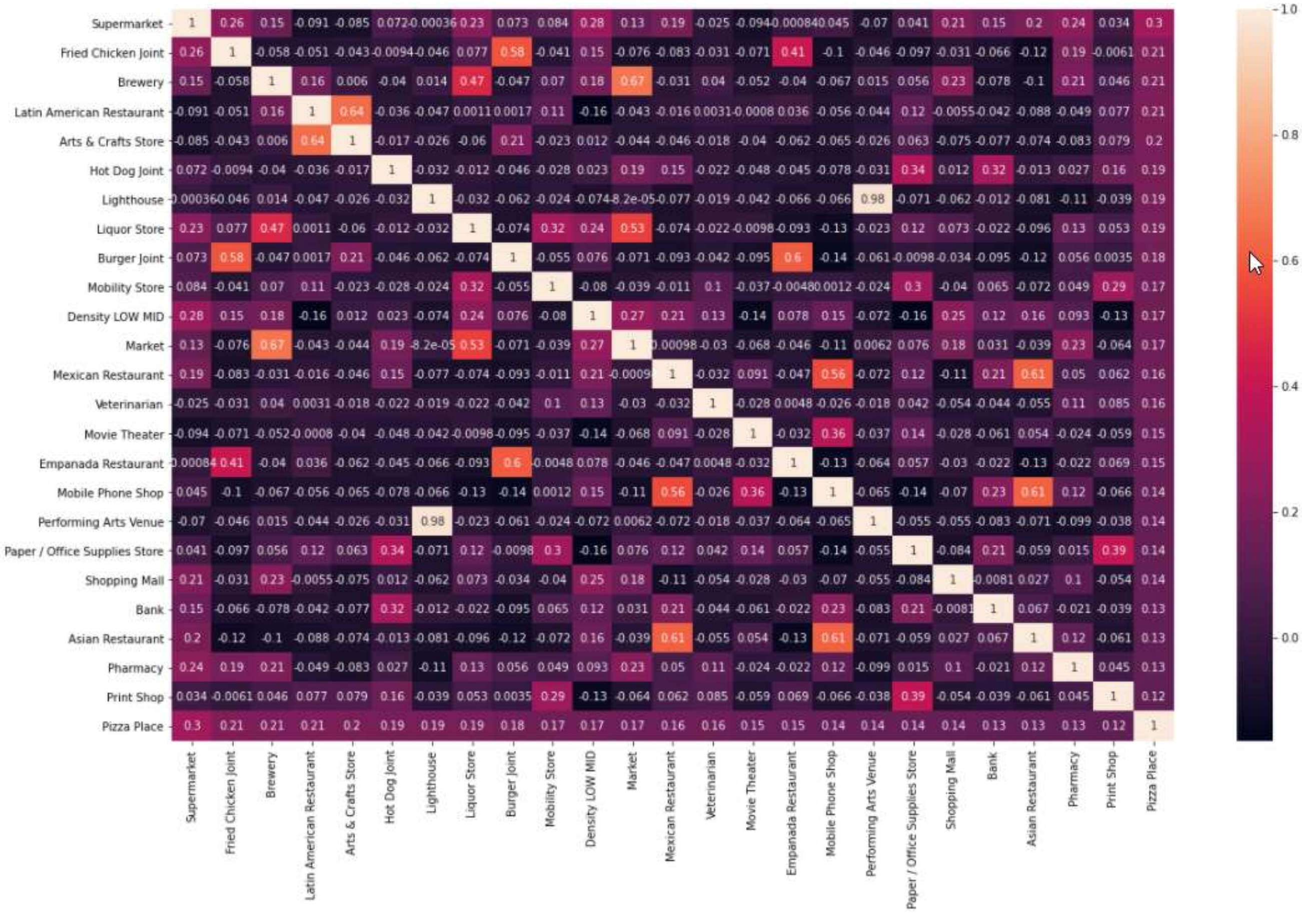


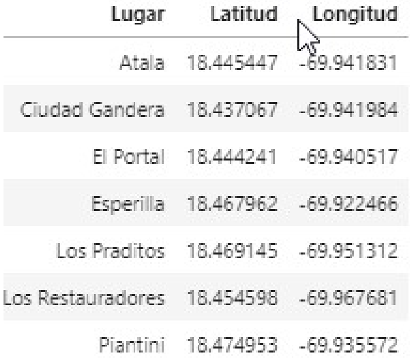
Figure 5 KMean Clusters for neighborhoods

A Correlation matrix will allow us to analyze the relationship between Pizza and other venues categories, as we can see supermarkets and fried chicken joints are between the most correlated venues categories.



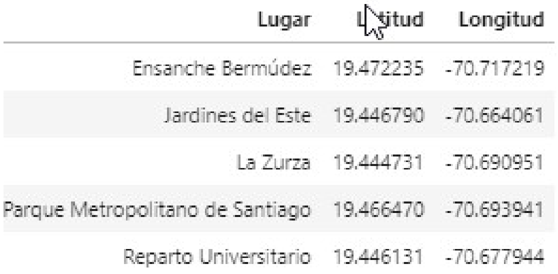
For both cities where identified a set of places that can be recommended for business market study for a new Pizza place. Both list also present great correlation to what I, as a local citizen, can predict for the suggested neighborhoods.

Some of the places where on neighborhoods well in a near radius to a Pizza Place, so increasing neighborhood radius will eliminate some of the recommendations.



Lugar	Latitud	Longitud
Atala	18.445447	-69.941831
Ciudad Gandera	18.437067	-69.941984
El Portal	18.444241	-69.940517
Esperilla	18.467962	-69.922466
Los Praditos	18.469145	-69.951312
Los Restauradores	18.454598	-69.967681
Piantini	18.474953	-69.935572

Figure 7 Prediction for Distrito Nacional

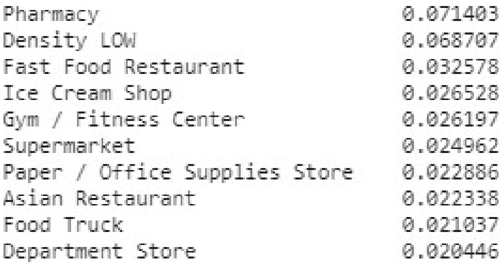


Lugar	Latitud	Longitud
Ensanche Bermúdez	19.472235	-70.717219
Jardines del Este	19.446790	-70.664061
La Zurza	19.444731	-70.690951
Parque Metropolitano de Santiago	19.466470	-70.693941
Reparto Universitario	19.446131	-70.677944

Figure 6 Prediction for Santiago

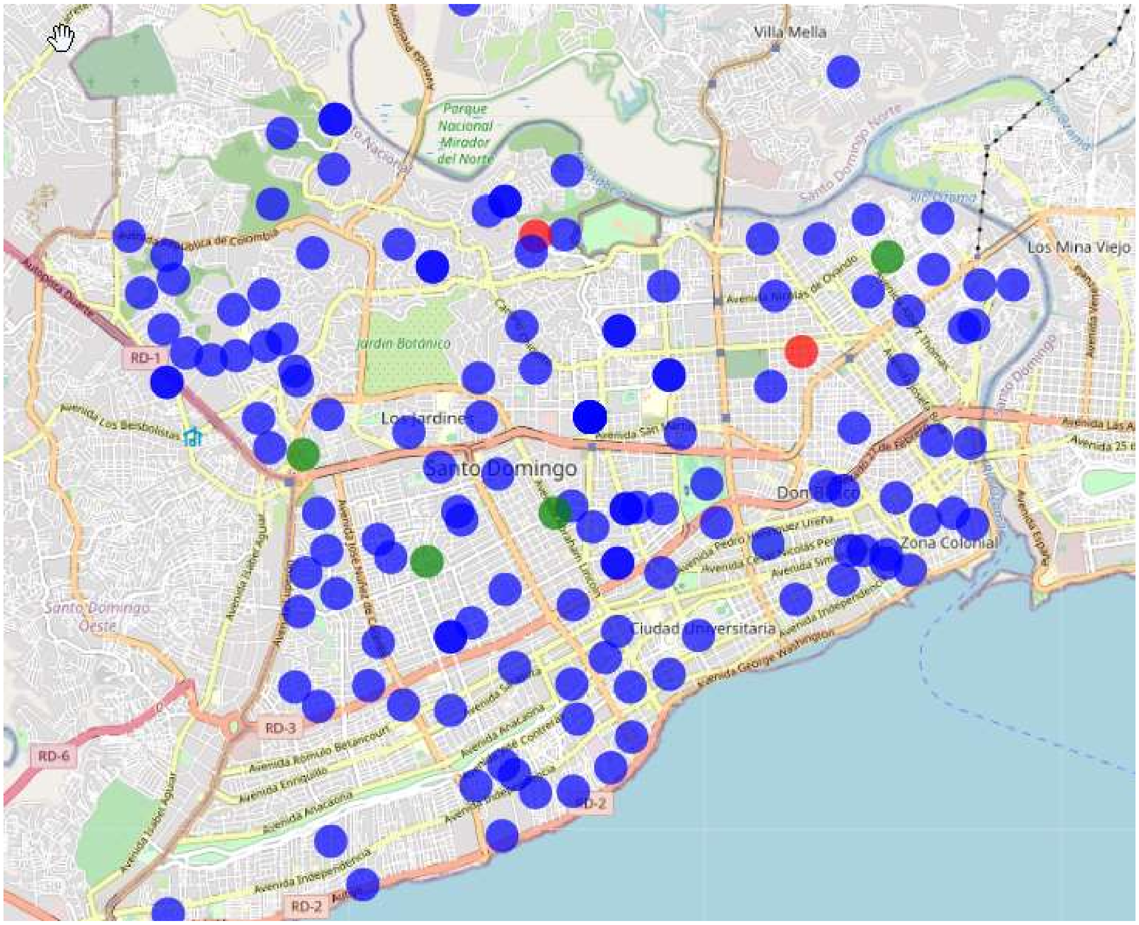
We created some features based on the density (quantity of business in the reach radio), the reason is that some business are more common where there are a great quantity of stores, like in Malls and near super markets.

For Pizza places the next graph show a correlation found with a Forest Tree Classificator of some venues types with the target one.



```
Pharmacy                          0.071403
Density LOW                       0.068707
Fast Food Restaurant              0.032578
Ice Cream Shop                    0.026528
Gym / Fitness Center              0.026197
Supermarket                       0.024962
Paper / Office Supplies Store     0.022886
Asian Restaurant                  0.022338
Food Truck                        0.021037
Department Store                  0.020446
```

The blue Dots are some Pizza places correctly predicted, the Green are suggestions for new ventures in Pizza Places, the red ones are false predictions of business that already exist.



The Confusion Matrix analysis got us a near 97% true positive predictions, and an 80% true negative predictions using 14 as max deep tree, and 99% and 75% for a max deep 9. While the true positive where higher in Santiago, is was good enough to get a decent prediction on Santo Domingo Dataset.

In Both cases the false positive was under 3%, and in the lesser deeper (9) we had a 25% chance on false negative.

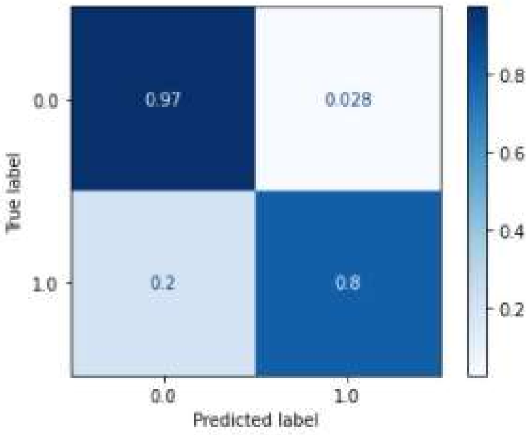


Figure 9 Random Forest, max DEEP 14

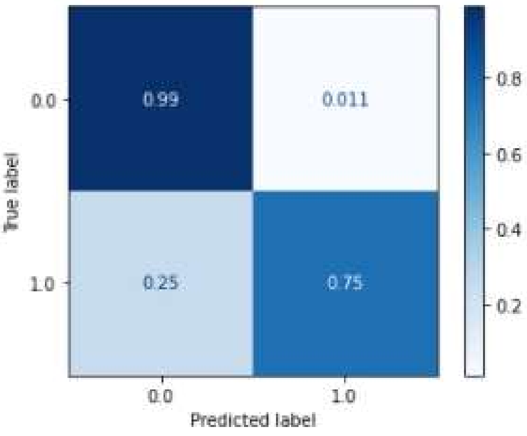


Figure 8Random Forest, max DEEP 9

Discussions

As we progressed we were tempted to extend the analysis to more cities of Latino America, that way the models can be refined to a greater by been exposed to variety of patters, and tested against distinct compositional cities. Also the real relationship between some venues can be better described by higher feature understanding when a more diverse and ample dataset is used.

Venues categories should be first reduced, and clustered into similar super Categories; If done well, the resulting model can be more powerful and general.

That is a limitation of the dataset used that diminished the Supervised and Unsupervised training models, as a lot of features that we can consider similar to some extends (like Restaurants and Other kind of Restaurants) were managed as individual features, without relationship between them.

Also the Foursquare API had a limited set of places available as it's not so used in the Dominican Republic. We think using Google Places will allow a greater set of Venue types/categories that can be used to get a better clustering of the different zones.

Conclusion

The potential for city zones clustering is only limited by the quality of the data we can get to feed our models, helpfully that as things are going, it continuous bettering.

Known this, real estate and construction industry will benefit by gaining knowledge before it become crucial about places to build, develop or invest. Big data sources with historical data are needed for this to happen, so there is an opportunity for any data source historical collector.

Also big and small business can benefit from knowing the needs present on cities commercial systems and fulfill those needs before its even know. If that info is correlated with the historical development of the city, there will be big advantages to business managing it.

This study also touch some techniques not present on the notebook provided, like feature evaluation for Random Forest Classifiers, and the comparison to others classifiers like GradientBoostingClassifier and logistical regression, both with results not as good as the Forest implementation.