Capstone – IBM Data Science

Where to open a mid-range Restaurant in the suburbs of North Dallas

Abstract

A data analysis on how to select the ideal location to open a new restaurant in the North Dallas

Llorenç Aguila de la Morena

Llorenc.aguila@gmail.com

Introduction

There is a saying in Marketing that states that the three most important aspects for a retail business to be successful are: location, location, location. While it is true that customer visits to any retail shop will very much dependent on being in right place: well communicated, nice neighborhood and proximity to other retail business or shopping centers, the reality in North Dallas is a bit different.

Unlike most European or some US cities where there is public transport and neighborhoods are designed to be "walkable", the suburbs in North Dallas are 100% car defendant. Distances from houses and shopping centers or business are very large and people need, most of the times, a car to move around. Commercial centers are designed to cluster a number of business (restaurants, grocery stores, shops) and have large parking lots so customer can park their cars with no issue. With that in mind, the problem of finding a good spot for a restaurant, will not be so dependent on the actual distance from the houses or offices as you will have to get there by car anyway.

The geographical areas covered in this study will be Plano, Frisco and McKinney. These cities are located in the North of Dallas and have been growing economically in the past decades, which made them good candidates to place a restaurant. For this project, the factors that will play a role in determining the ideal location for a midrange Restaurant are the following:

1. Proximity to other successful midrange restaurants will represent competition and will influence negatively the location
2. Proximity to other businesses (groceries, shops, etc) or low range restaurants are regarded as non competition and therefore a positive influence as they will generate customer traffic to the spot.
3. Population density: the more population residing at a reasonable distance from the restaurant, the more chances to convert to customers
4. Population affluency: the higher household income, the more disposable money to spend in a good middle restaurant. Since fine dining can be considered a non essential expenditure, average income in the area plays a role too.

Data Acquisition and Cleaning

Venue and restaurant information can be obtained through Foursquare with any level of granularity and will be used to get information about restaurants and businesses in a given location. For demographics at zip and city level. there is no available source of information that can provide average income or population with any level of granularity. However, there are some sources that provide demographic information at zip/suburb level. These sources will be used to gather data about points:

1. https://www.zip-codes.com/ lists the zip codes that belong to the cities in scope:
    a. Plano https://www.zip-codes.com/city/tx-plano.asp: 75023, 75024, 75025, 75074, 75075,75093
    b. Frisco https://www.zip-codes.com/city/tx-frisco.asp: 75033, 75034, 75035, 75036
    c. McKinney https://www.zip-codes.com/city/tx-frisco.asp: 75069, 75070, 75071, 75072
2. https://www.bestplaces.net/ provides population and income by zip code
3. https://www.census.gov/ provides official data about demographics at city level

These sources are combined into one file with all the relevant information zip code (zip-data.csv) and one file with information at city level (city-data.csv)

The result dataframes for city data looks like this:

Fact		McKinney	Frisco	Plano
0	Population estimates, July 1, 2018, (V2018)	191,645	188,170	288,061
1	Population estimates base, April 1, 2010, (V2...	131,160	117,170	259,857
2	Population, percent change - April 1, 2010 (es...	46.10%	60.60%	10.90%
3	Population, Census, April 1, 2010	131,117	116,989	259,841
4	Persons under 5 years, percent	7.70%	6.60%	5.50%

While the dataframe for zip level data is as the following:

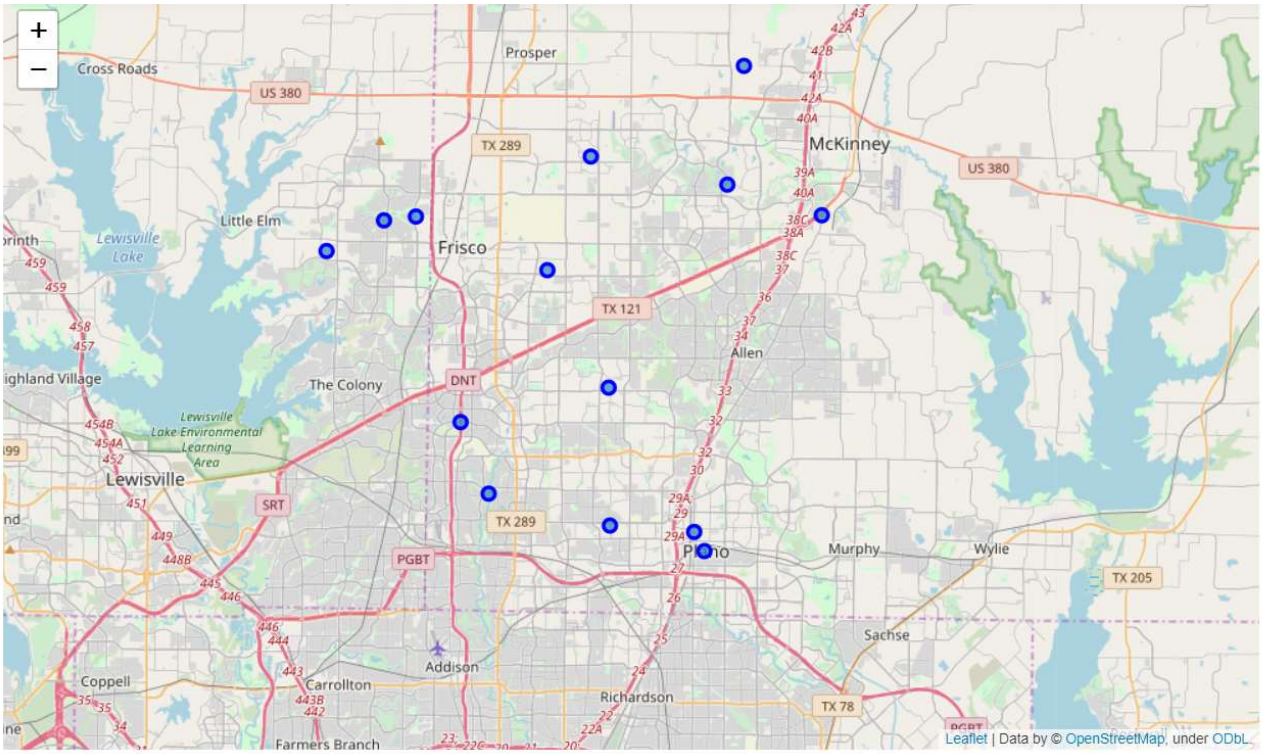
	ZIP	City	Population	Median-Income
9	75036	Frisco	33245	NaN
10	75069	McKinney	36879	47886.0
11	75070	McKinney	93299	100848.0
12	75071	McKinney	49846	81167.0
13	75072	McKinney	52835	NaN

Some of the median income values and median price were not available in the original data source.

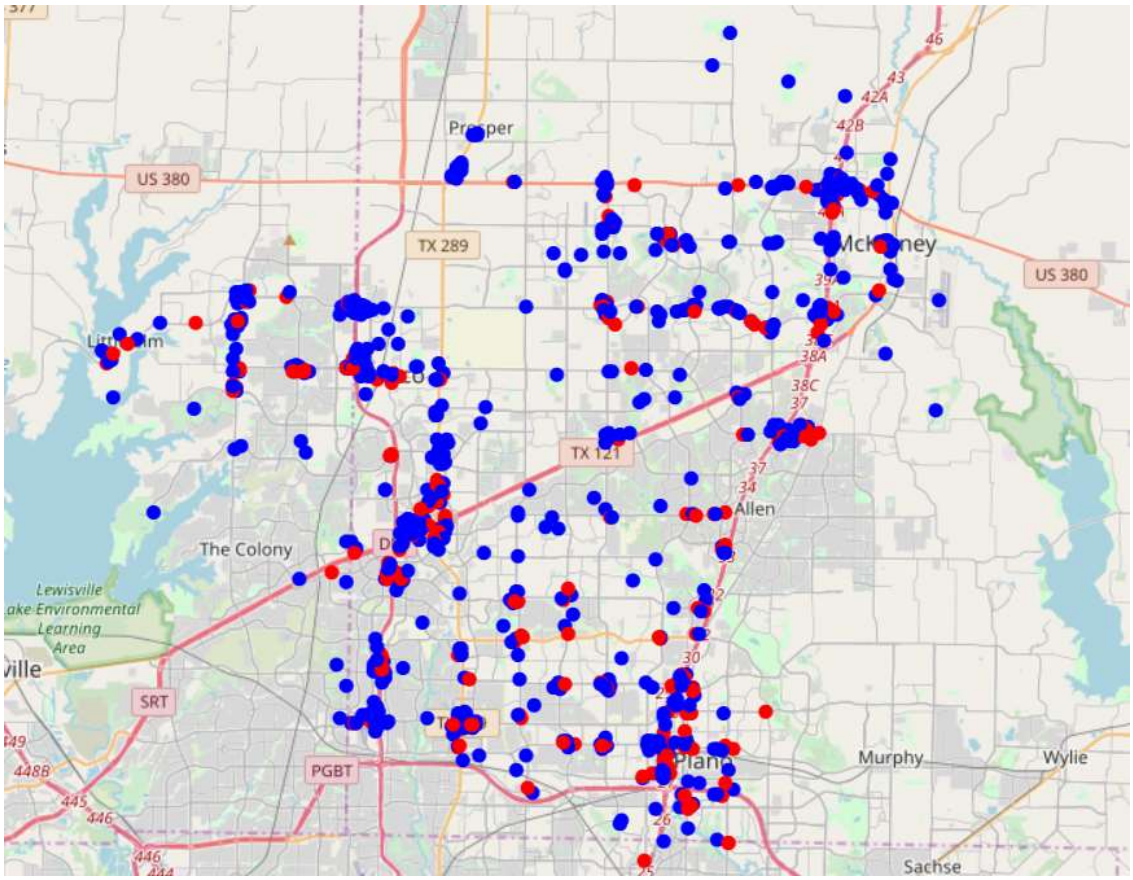
We will use the city average value as an valid approximation for the value

By using the Geopy libraries and Folium as map visualization tool, we are able to locate these zip codes in the map of North of Dallas.

As obvious, the geographical dispersion of these zip code areas is high. When obtaining venues around these zip codes, we will have to set the radius to at least 6KM in order to be able to capture all the places around the points. Next, we are going to start utilizing the Foursquare API to explore the neighborhoods and segment them.



We use these initial locations to get the top 100 venues that are around the zip code centers within a radius of 6,000 meters by means of the Foursquare API. The results are placed again in the map. In red we are depicting the Restaurants, in blue any other venue that is retrieved by Foursquare. Remember that, we are interested in all venues because we want to place our new restaurant around existing venues. We will use them to create venue activity clusters.



The venues are grouped by zip codes to obtain how many restaurant and other venues we have in the area. The outcome looks like this. This concludes the data gathering phase.

ZIP	Restaurant	Venue
75023	25	25
75024	24	57
75025	22	54

Methodology

As mentioned in the introduction, we want to place our new restaurant ideally in a location where there is already some business activity. The way to achieve this is by using k-means clustering on the venue data obtained from Foursquare to determine what are the points of accumulation of business in the areas in scope.

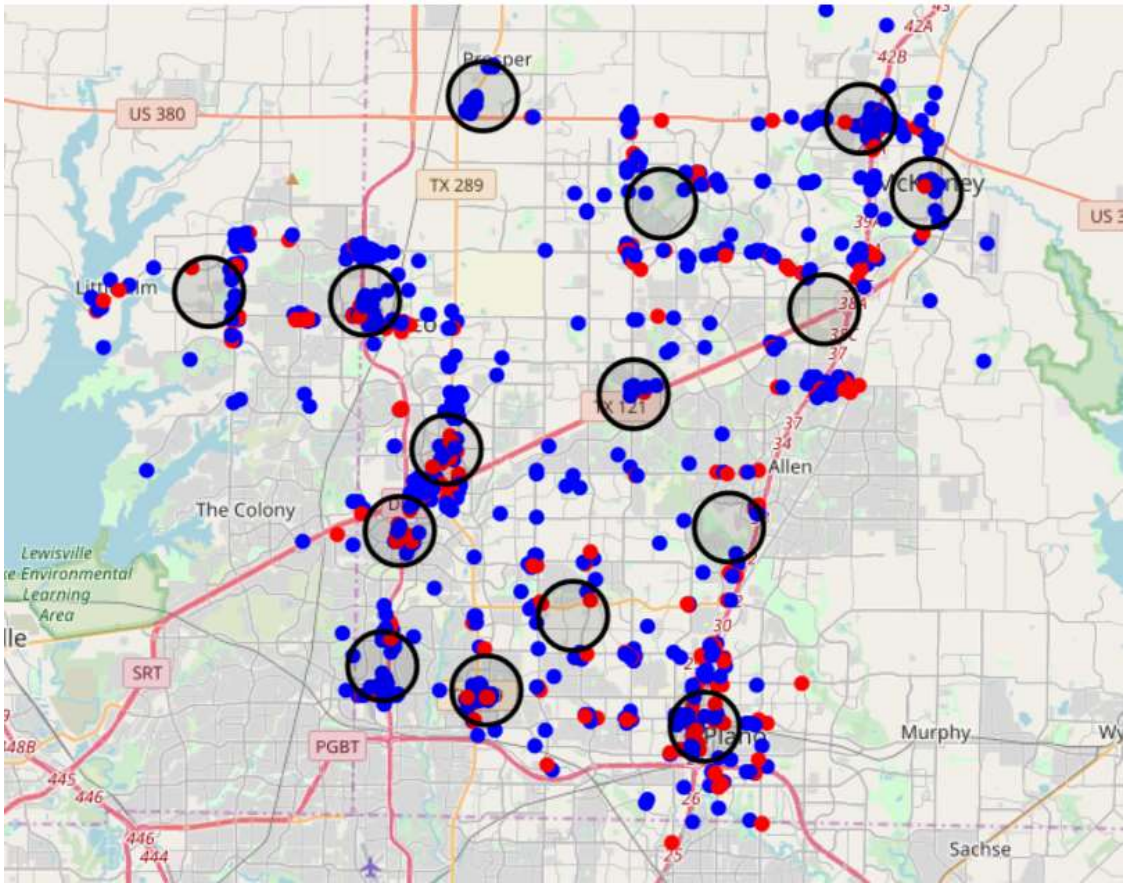
After several simulations with different numbers of clusters, we came to the conclusion that 15 clusters is a reasonable number (close the number of zips in scope:14). Less than 15 will missed some venue accumulations and more than 15 generates way too many clusters in areas with just a handful of venues.

Once we get the centers of these clusters, we will use the coordinates to get geographical information (address and zip code) and from there we will be able to add the demographics related to the zip code. With this aggregated information, we will be able to decide on the best location based on population per restaurant and income per restaurant ratios.

Analysis

Cluster Neighbors

By utilizing k-means clustering, we can identify what are the business centers in the area in scope. The clustering is performed on all the venues returned by Foursquare and only using the geographical coordinates as parameters. The clusters centers are identified by black circles and are superimposed to the existing distribution of venues:

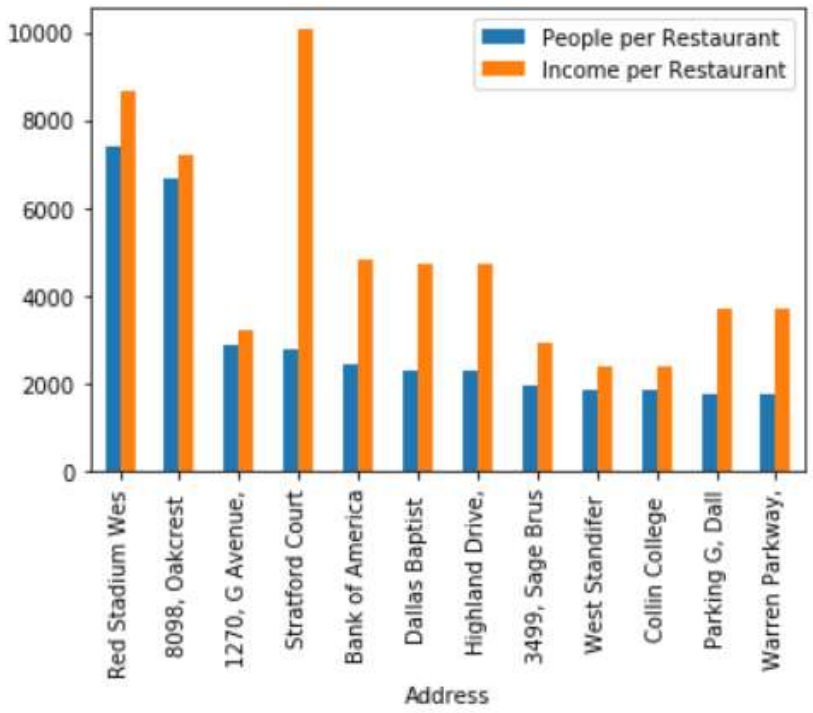


The last step in the analysis is to add the demographic information related to the zip codes. Lastly, we will calculate the ratios People per Restaurant and Income per Restaurant and sort the results by these parameters. These will be decisive to make the final call.

This is the outcome:

	Address	latitude	longitude	ZIP	Population	Median-Income	Restaurant	Venue	People per Restaurant	Income per Restaurant
1	Red Stadium West, World Cup Way, Frisco, Colli...	33.156902	-96.838496	75034	95996	112626.0	13	30	7384.307692	8663.538462
7	8098, Oakcrest Drive, McKinney, Collin County,...	33.190316	-96.722366	75070	93299	100848.0	14	47	6664.214286	7203.428571
4	1270, G Avenue, Plano, Collin County, Texas, 7...	33.016596	-96.704431	75074	48977	55152.0	17	29	2881.000000	3244.235294
10	Stratford Court, Little Elm, Denton County, Te...	33.160606	-96.900746	75036	33245	120701.0	12	44	2770.416667	10058.416667
0	Bank of America, Custer Road, Frisco, Collin C...	33.127015	-96.732617	75025	53559	106301.0	22	54	2434.500000	4831.863636
2	Dallas Baptist University (DBU North), Dallas ...	33.038672	-96.831319	75093	48021	99378.0	21	63	2286.714286	4732.285714
3	Highland Drive, Plano, Collin County, Texas, 7...	33.029622	-96.790684	75093	48021	99378.0	21	63	2286.714286	4732.285714
11	3499, Sage Brush Trail, Plano, Collin County, ...	33.053191	-96.756657	75023	49563	73602.0	25	25	1982.520000	2944.080000
5	West Standifer Street, McKinney, Collin County...	33.193276	-96.616711	75069	36879	47886.0	20	62	1843.950000	2394.300000
6	Collin College - Central Park Campus, West Uni...	33.218641	-96.642309	75069	36879	47886.0	20	62	1843.950000	2394.300000
8	Parking G, Dallas Parkway, Plano, Collin Count...	33.081428	-96.825272	75024	42405	89119.0	24	57	1766.875000	3713.291667
9	Warren Parkway, The Centre at Preston Ridge, F...	33.109010	-96.806574	75024	42405	89119.0	24	57	1766.875000	3713.291667

In order to facilitate the decision, we create the following chart depicting the two decision parameters: People per Restaurant and Income per Restaurant:

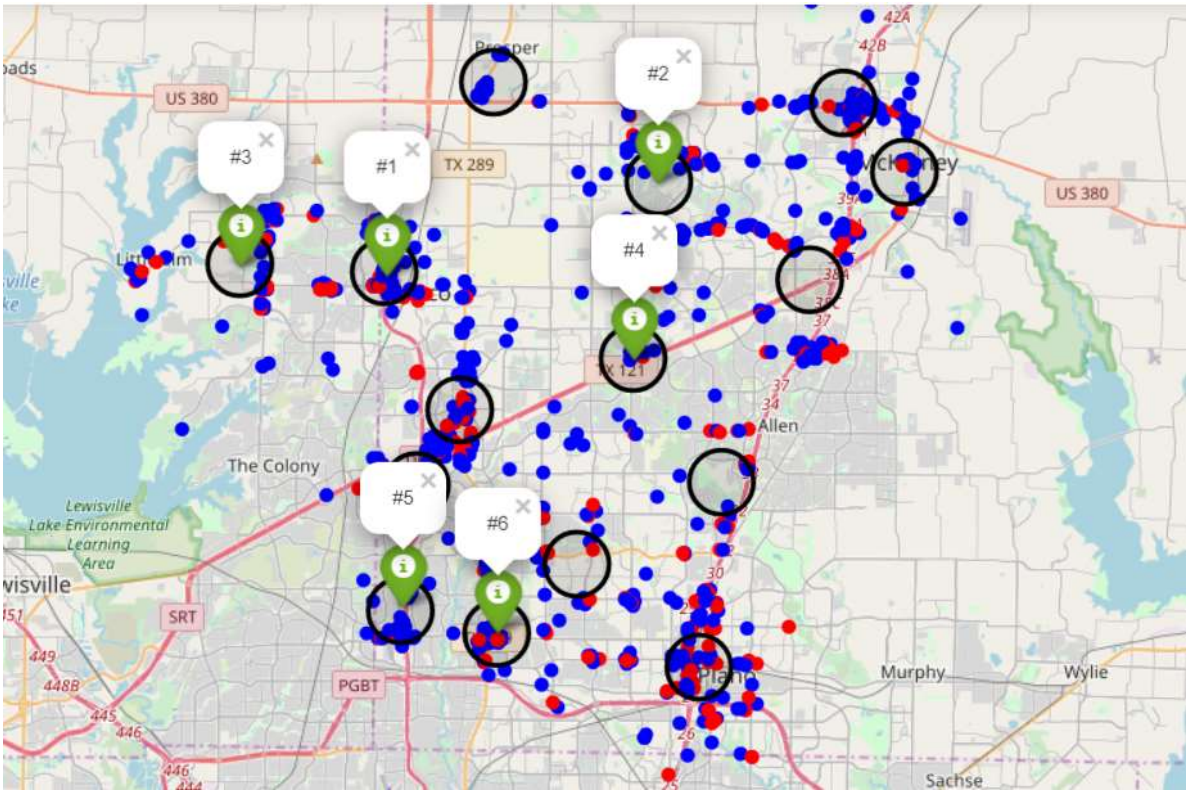


The top two in Frisco and McKinney clearly stand out as being the most suitable with very high number of people per restaurant(low density) and similar income per restaurant. No surprise on the areas as these are known to be developing and therefore not a lot of restaurants are in operations. These two are clear candidates.

In contrast, more developed areas such as Plano have lower number of people per restaurant (high density) as they have established for a longer period in time. We will not considered the 1270G Avenue as a good candidate as, while it show high people per restaurant ratio, it related income per Restaurant is much lower than the following location candidates.

Last, the next three cluster centers will also make the final cutover list of recommended places because their combination of Income and People per Restaurant ratios. We leave the remainder out as they show less favorable combinations

The top 6 picks for ideal Restaurant locations in the map along with the other clusters and venues.



Confirming expectations. The northern suburbs offer more opportunities to open a new Restaurant as they are relatively new neighborhoods that are still being developed. Two of the more established areas in Plano (choice#5&6) also come up but are clearly less favorable as the ones on top of the list.

Results and Discussion

The analysis performed demonstrates that not all places are ideal to locate a new restaurant in the growing Northern suburbs of Dallas. The venue data obtained from Foursquare shows that the distribution of business across the area is uneven and there are venue concentrations that need to be taken into account when considering opening a new retail business.

In that respect, the geographical clustering of venues obtained through K-means is enlightening. Several clusters emerge from the analysis and they are not necessarily aligned with purely geographical locations. With that information at hand and the map visualizations, it has become apparent where are the best candidate locations are, at least from a pure business density point of view. These clusters are scattered all across Frisco, Plano and McKinney.

By adding the demographic information to the data mix, we have been able to narrow down the list of locations most suitable to open a new restaurant. These data elements have made clear that the best locations are in the north. The locations are less dense in terms of existing restaurant per person (high Person per Restaurant ratio) and more income per restaurant. In simple terms that means that people have less restaurant choices (less competition) and more money to spend! which made those locations ideal to setup new restaurants. In that respect, two locations in Frisco and McKinney clearly stand out from the rest and are the top picks by far. This comes as no surprise as these norther areas are known to be under development and are some of the most fast-growing places in the whole US.

This analysis is conducted without taking into account other economic viability factors such as availably of commercial spaces, lease prices, etc. However, it proves very good starting point to continue the analysis focused in very specific geographical places determined by the top picks.

Conclusion

The purpose of this project was to identify ideal locations to open a new mid-range restaurant in the Northern suburbs of Dallas (Frisco, McKinney, Plano). Since the transportation context in this part of the country makes it mandatory to access the restaurant by car, we have discarded the restaurant density by physical area in favor of looking for concentration of existing venues as attractors of clients. Our criteria to identify ideal locations have been established as proximity to existing business, restaurant density per person and average income for the residents in the area.

In order to perform the geographical analysis, we have divided the area in zip codes and obtained demographic information.

By means of the Foursquare API, we have been able to gather abundant information of venues in the area in scope. Clustering venues by using K-means clustering has given us the necessary insights to visualize and select the densest venues locations. Last, by using the corresponding demographic information we have been able to single out the most attractive areas in terms of lower restaurant saturation per population and highest income.

The final decision will necessarily involve a physical inspection of the recommended areas as well as other economic factors not considered in this study such as availability of commercial space, lease prices and other viability factors. The availability of more granular data and other sources of information could certainly enrich this analysis and make it much more precise. Nevertheless, the purpose of the project has been accomplished with the existing tools and resources.

Future Directions

The accuracy of the analysis could be further improved if we had access to demographic data at household level, which as of today is not publicly available. We could also add some additional attributes to the mix such as lease prices in the different areas to help take the optimal decision.

Furthermore, if we could have commercial data about existing and past restaurants, we could envision a predictive model that can anticipate the success probability of a new restaurant based on past performance.