Finding a Suitable Restaurant Location in an Unfamiliar City

Coursera IBM Data Science Capstone Project : Final Report

Ryan Herchig

May 30, 2020

1 Introduction

The business problem I have chosen to address is to determine suitable placement of a new restaurant within a city chosen by the user. Many small businesses fail within the first several years of opening [2] and any advantage that the small business owner can gain could increase their chances of beating these odds. Finding a good location for a new restaurant based on the number of restaurants in various neighborhoods within a city could provide valuable information when making this selection. This product would be beneficial to an entrepreneur who is looking to open a new restaurant in an unfamiliar city and would like to try and find a suitable location.

Additionally, someone who finds themselves in an unfamiliar city and looking to find an area with many restaurant choices could potentially benefit from this as a mobile app. It could point them to neighborhoods with large numbers of restaurants within small areas. Therefore, the target audience could either be someone looking to open a business, or a consumer trying to find areas in an unfamiliar city with many eating options.

2 Data

As previously mentioned, the city will be chosen by the user. Once the city is chosen, the city center will be considered the origin of the grid and the search for a suitable location will be confined to a search area defined by the locations of the zip codes in the city. The application will use the FourSquare API to search the city for different restaurants and subsequently group the data by zip code. This data will then be ordered by the number of restaurants per zip code and the various zip codes will be placed into bins (number of bins being $\approx 10\%$ of the number of zip codes). The application will then generate a map of the city with the different restaurant locations plotted, and the color of the marker representing the latitude and longitude coordinates of the zip code will correspond to the bin that zip code belongs to. From this, the customer will be able to determine at a glance which areas (zip codes) have the highest number of restaurants and which are more sparsely populated. For the entrepreneur looking to open their own restaurant, this will allow them to choose their location based on the density of restaurants in an area. This could mean either introducing a new type of cuisine in an area which is densely populated, or opening in an area which has fewer restaurants in hopes of reducing competition. For the casual user of the mobile app, the benefit would be that they are able to find areas of the city with the largest selections of restaurants.

3 Methodology

The purpose of this software is to provide information about restaurants in any city the user chooses. It should therefore be written as general as possible, accepting any city which is supported by the FourSquare API as input and returning the desired information. As a preliminary data analysis, 5 different cities of various sizes and populations were investigated and compared. The cities used were Los Angeles California, New York New York, Miami Florida, Houston Texas, and Atlanta Georgia. Information about the sizes of the cities in square kilometers as well as population sizes was retrieved using SPARQL queries [1]. The results of these queries for all 5 cities are shown in columns 2 and 3 of table 1.

From table 1 we see that the city with the largest area is Houston Texas while New York City has the most inhabitants. However, a more meaningful metric for comparison is the population density given that it better quantifies the number of people who are likely to live in the vicinity of a given restaurant location. This is calculated by dividing the population of a given city by its total area, giving an answer with units of residents per square kilometer. The data in table 1 shows that New York City has the largest population density, followed by Los Angeles California.

Table 1: City Statistics

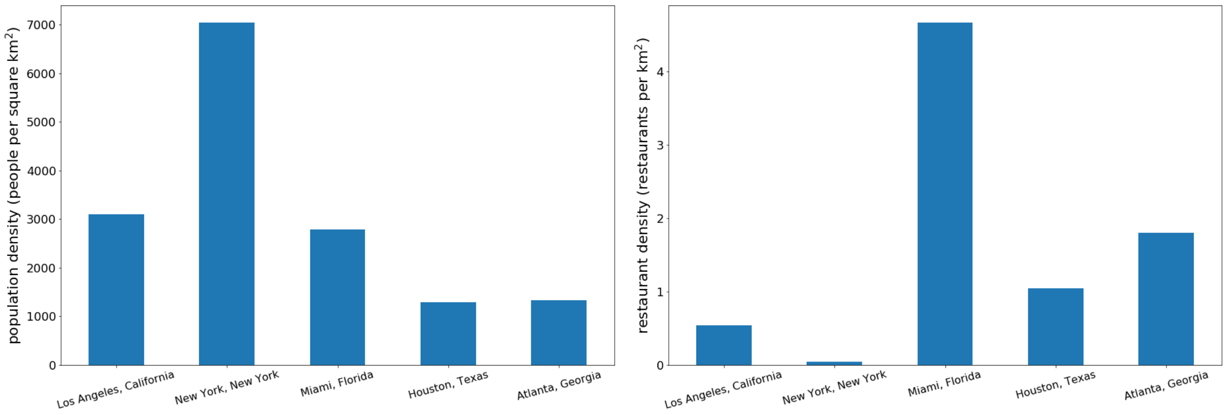
Address	Area (km^2)	Population (millions of residents)	Population Density (residents / km^2)	Number of Venues	Restaurant Density (restaurants / km^2)
Los Angeles, CA	1302.0	4.0	3095.93	701	0.538
New York, NY	1214.0	8.6	7040.56	58	0.048
Miami, FL	143.1	0.4	2790.50	668	4.666
Houston, TX	1625.0	2.1	1291.81	1694	1.042
Atlanta, GA	347.1	0.46	1336.44	625	1.801

At this point, calls were made to the FourSquare API for each of these cities using the latitude and longitude of the physical zip code locations as the coordinates for the search query. The term physical zip codes refers to the fact that the list of all zip codes in each city was first purged of zip codes corresponding to P.O. boxes before being used to search for restaurant venues with the FourSquare API. Column 5 of table 1 shows the number of restaurant venues which were returned for each respective city. For all the cities, the search queries were run iteratively with the search radius parameter being varied for each iteration. At the end of each iteration, the full list of returned venues was purged of duplicates by referencing the venue ID which was taken to be a unique identifier for each restaurant. The search radii used are 5000, 8000, and 12000 meters. Results from the different searches were compared and it was found that an 8000 meter search radius is sufficient to capture the same number of venues as yielded by a 12000 meter radius. The 5000 meter radius produced slightly fewer returns the the 8000 meter search radius.

The "Number of Venues" column of table 1 shows that FourSquare returned the largest number of restaurant venues for Houston Texas which also has the largest area in square kilometers. Surprisingly, only 58 venues were returned for New York. The reason for this was not explored, but might be due to excessive overlap between the search radii centered on the different zip codes.

The final column of table 1 displays the restaurant density or the number of restaurants per square kilometer for each city investigated. This figure is proportional not to the actual number of restaurants in each city, but rather to the number which were returned from the sequence of queries for each city. Consequently, New York has a very low restaurant density while Miami's is extremely high.

Figures 1a and 1b show a graphical representation, in the form of a bar chart, of columns 4 and 6 of table 1, the population densities and restaurant densities respectively. This is shown for each city included in the preliminary analysis. The figure shows how extreme the population density of New York is compared to the other cities considered during this study, more than double Los Angeles which has the next highest. The figure also shows the small number of venues returned for New York City as compared to the other cities, despite its small size. Miami Florida on the other hand has a restaurant density 2 to 3 times that of Atlanta Georgia with the next highest restaurant density.



(a) population densities for each city

(b) restaurant densities for each city

Figure 1

As a case study, Los Angeles was chosen due to the fact that it has a moderate size and number of inhabitants compared to the other cities which were investigated. Only physical zip codes which fall inside the city limits were considered. Figure 2 shows the latitude and longitude coordinates plotted on a Folium Leaflet corresponding to the zip code locations (blue circles) and the restaurant venues (red circles). This alone could serve as a visual aid to a customer searching for a suitable location for a new restaurant. The figure shows streets with large concentrations of restaurants and other areas which are devoid of restaurants. The areas with relatively low concentrations of restaurants could be residential areas in which businesses are not allowed to be open.

The primary purpose of this software is to aid the customer in finding a suitable location to open a new restaurant. As mentioned in section 2, the application uses the FourSquare API to search the city for different restaurants and subsequently group the data by zip code. This was done for the city of Los Angeles. This data was then ordered by the number of restaurants per zip code and the various zip codes were be placed into bins with the number of bins being $\approx 10\%$ of the number of zip codes. For Los Angeles, the minimum number of restaurant venues returned for any zip code was 1 and the maximum number returned was 30. When calculating the total range that the bins will span, the minimum of 0 and the smallest number of venues

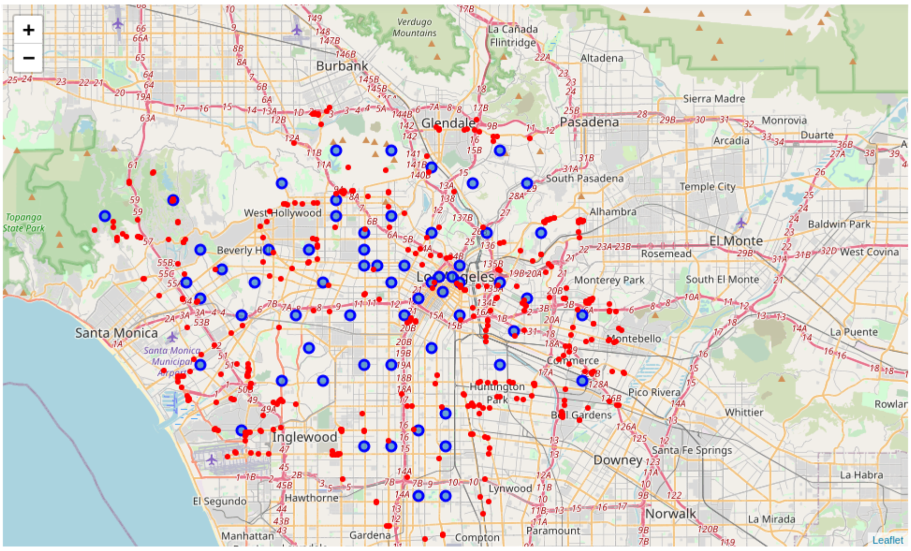


Figure 2: Folium Leaflet map showing locations of all zip codes in Los Angeles proper and the locations of all restaurants returned by the FourSquare API.

returned minus 1 was taken to be the lower bound and the maximum number of venues returned was taken as the maximum value. Consequently, the bins were calculated using 0 as the absolute minimum and 31 as the absolute maximum. The total number of zip codes was 52 with 10% of this being ≈ 5, rounded to the nearest integer. The dividing values for the histogram bins were then calculated to be [0.0, 6.0, 12.0, 18.0, 24.0, 30.0]. Therefore, cluster 1 includes all zip codes where the number of restaurants falls between 0.0 and 6.0, cluster 2 includes all zip codes for which the number is greater than 6.0 and less than or equal to 12.0, etc. Once all the restaurants were placed into the correct bins, the data was ready to be used in a figure where the color of the marker representing the latitude and longitude coordinates of the zip code corresponds to the bin that zip code belongs to.

4 Results

Figure 3 shows the results of running the previously described binning function on the data returned from the FourSquare API for Los Angeles. The restaurant locations are represented by small black circles while the different zip code centers are depicted by the larger circles of various colors. The colors correspond to the 5 different bins as follows: bin 1 = red, bin 2 = purple, bin 3 = blue, bin 4 = light green, bin 5 = orange. The astute reader may notice that the number of zip code markers differs from figure 2 and figure 3. The reason for difference in number of zip code markers between these figures is due to the fact that the first figure plots all zip codes, regardless of if they have any venues assigned to them while the second figure only plots zip codes which have venues associated with them. The reason that some of the zip codes have no venues associated with them is that the algorithm assigns each venue to only 1 zip code (the first one it encounters while looping through list of zip codes). After the query is ran for each zip code, the list is purged of duplicates. If a venue happens to be returned in multiple queries, i.e. belongs to more than one zip code, the 2nd, 3rd, etc. instances are removed from the list since these would represent duplicates. Therefore, a venue will only “belong” to 1 zip code, specifically the first zip code which returns the venue in the result of the FourSquare query. This effect is an unfortunate consequence of the fact that the search area is specified as a radius in the FourSquare API, coupled with the fact that the zip code centers are not spaced regularly. If the search radius is made too small, venues will be missed in the intermediate space. To ensure all discoverable venues are found, the search radius is made excessively large and duplicate returns are subsequently removed.

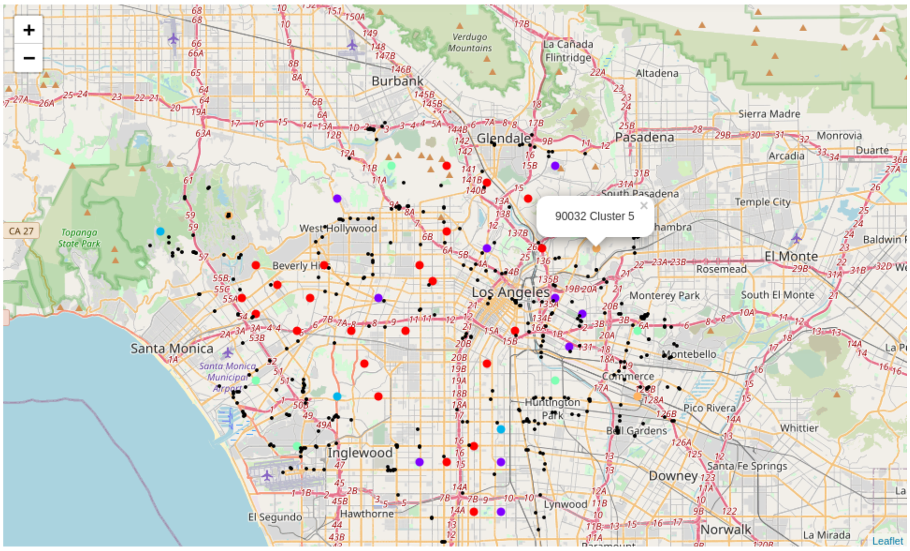


Figure 3: Folium Leaflet map of Los Angeles proper showing which areas have a greater number of restaurants and which have fewer. The figure also shows the locations of all restaurants returned by the FourSquare API.

Essentially, this means the circles associated with the different zip codes in figure 3 serve as markers for distinct areas which have venues associated with them. It is assumed that if 2 zip codes are close enough together that they return the same venues to within the search radius, either one of the 2 zip code markers will work well to define that given area on the map.

5 Discussion

From the perspective of someone wanting to open a restaurant in a new city, several promising areas can be identified from figure 3. Setting the criteria for a good restaurant location as an area with numerous other restaurants in the immediate area (the logic being the user could open a restaurant which serves a different cuisine not available in the area), the light green and orange markers fit this description. These represent areas where $18 > N_{rest} \leq 24$ and $24 > N_{rest} \leq 30$ respectively. Consequently, the area on the map in the vicinity of Commerce and Montebello would make excellent locations. Another orange marker (labeled by the pop-up which reads “90032 Cluster 5”) exists in the northeastern area of the city and would make a good location given the above criteria. Lastly, another orange marker in the northwestern corner of the city which seems to be in an outdoor/forested area as compared to intercity Los Angeles.

If on the other hand, the user decides to take the opposite approach and search for areas with few restaurants (or none at all), the red markers denote areas with few restaurants in their immediate vicinity. Adopting this approach, the user would need to be conscious of the fact that some areas may be devoid of restaurants simply because the zoning restrictions do not allow businesses/commercial establishments. However, given that the city regulation allow restaurants in such an area and the user decides it would be a suitable location, a well placed restaurant of a specific type could be very profitable given the lack of competition. Perhaps the user could decide on the restaurant type by taking into account the ethnicity of the surrounding residents, or establish a type of restaurant which is not prevalent throughout the city in general.

Analyzing the situation form the perspective of a casual user trying to find an area of an unfamiliar city with a high concentration of restaurants, the same areas mentioned in the first paragraph of this section would be of interest. In particular, the neighborhoods of Huntington Park, Montebello, and the area east of Topango State Park would all be good choices for someone looking for a variety of restaurant choices in a small area.

In general, the areas which seem to have the highest concentration of restaurant venues, at least judging from the Los Angeles case study, are those near highways and main roads and which lie on the outskirts of town in the suburbs. The lowest concentrations seem to be found in what appear to be highly residential areas. This is not surprising given that building regulations likely restrict the opening of commercial businesses in some of these areas.

6 Conclusion

In conclusion, the software which has been developed as a part of this project has been demonstrated to effectively search a city (specified by the user) for restaurant venues and subsequently plot these venues over a map of the city. Using Los Angeles as a specific case study, several areas were identified as promising due to either their high concentration of restaurant venues or low concentration. The areas with high concentrations of restaurants are also attractive areas for a person using the software to find areas in an unfamiliar city which have many restaurant choices in a small geographic region. Though Los Angeles was used as an example in this report, the application has been written in a general way such that similar results could be found for any city supported by the FourSquare database.

References

[1] sparql-client 3.6. https://pypi.org/project/sparql-client/.

[2] Tom Sumrak. lendingtree : What is the bureau of labor stats small business failure rate in 2020? https://www.lendingtree.com/business/small/failure-rate/.