Zip Code Shopping for High-End Café Establishment

Tae Yoon Kim

May 25, 2020

I. Introduction

The purpose of this project is to determine the best location to open up a high-end cafe in Houston. The definition of a high-end cafe is one that caters to the upper-middle to upper financial class of society. Goods and services will have a premium that is justified by the high quality in comparison to general cafes. Products from the cafe such as baked goods and beverages will be created by distinguished bakers and dessert chefs. The ingredients used will have a greater focus on quality than on affordability. As the economy thrives and a larger number of people realize a personal taste for premium services and goods, intelligent investors will be looking for locales that cater to a higher echelon of society (financially). It is innate human nature that people with greater than average buying power are willing to pay a greater premium for exclusivity.

The ability to purchase and enjoy goods/services in a private, limited community that the general public cannot access has proven to be a strong motivational driver for a lot of consumer habits. The luxury industry is a prime example of the undying thirst of the public for limited edition goods and services. Not everyone can afford a Louis Vuitton product, Mercedes Benz vehicle, or a Thom Browne suit. But people are consistently driven towards these products despite the steep entry price because of the exclusivity and the quality. This thought process carries over to fine-dining restaurants and, in this case, upscale cafes. The Foursquare location data will be vital to determining locations that are reasonably populated, wealthy residential neighborhoods that lack establishments of a similar nature, but contain venues that can create a strong synergy with cafes.

II. Data

As mentioned earlier, the Foursquare data will be essential to determine the number of cafes (especially ones that tailor to the same market as mine will) in all neighborhoods in Houston. This will enable me to make an educated decision on which neighborhoods to filter out and which ones are prime candidates based on market availability. Another source I will use is median household income data per zip code in Houston. This table also contains population per zip code as well as

specific latitude and longitude information (location). Knowing this, I can determine which neighborhoods have the most residents with the ability to afford the products and services my establishment will be offering, and which neighborhoods I will have to filter out due to a lower level of buying power.

By coupling this location/financial data with the Foursquare data, I will be able to cluster specific neighborhoods within Houston that are the most ideal candidates for beginning an upscale cafe business. The specific URL of this dataset is the following: "http://zipatlas.com/us/tx/houston/zip-code-comparison/median-household-income.htm". An example of the data is the following: zip code 77010 of Houston, Texas has a population of 76 people with an average household income of $200,000. The location of this neighborhood is 29.75310, -95.361109. Based on the definition of our problem, the key factors that will influence our decision are: A) the number of existing cafes in the neighborhood (high-end cafes) and B) number of restaurants and their type and location in every neighborhood.

ZipAtlas is an online, structured collection of zip code, area code, city and state demographic, social and economic profiles. Each report is accompanied by a well-constructed, detailed boundary map that allows for visual representation of statistical data. These can range from unemployment rates and property value assessments, to education attainment levels and racial profiles. The data collection spans the entirety of the United States of America.

III. Methodology

The process began with gathering the necessary data. Simply reading the webpage provided by ZipAtlas to a Dataframe through pandas was enough to isolate the desired data. The data had to be edited so that the appearance can be clean and presentable. Location data was divided into individual latitude and longitude columns for easier use in subsequent applications.

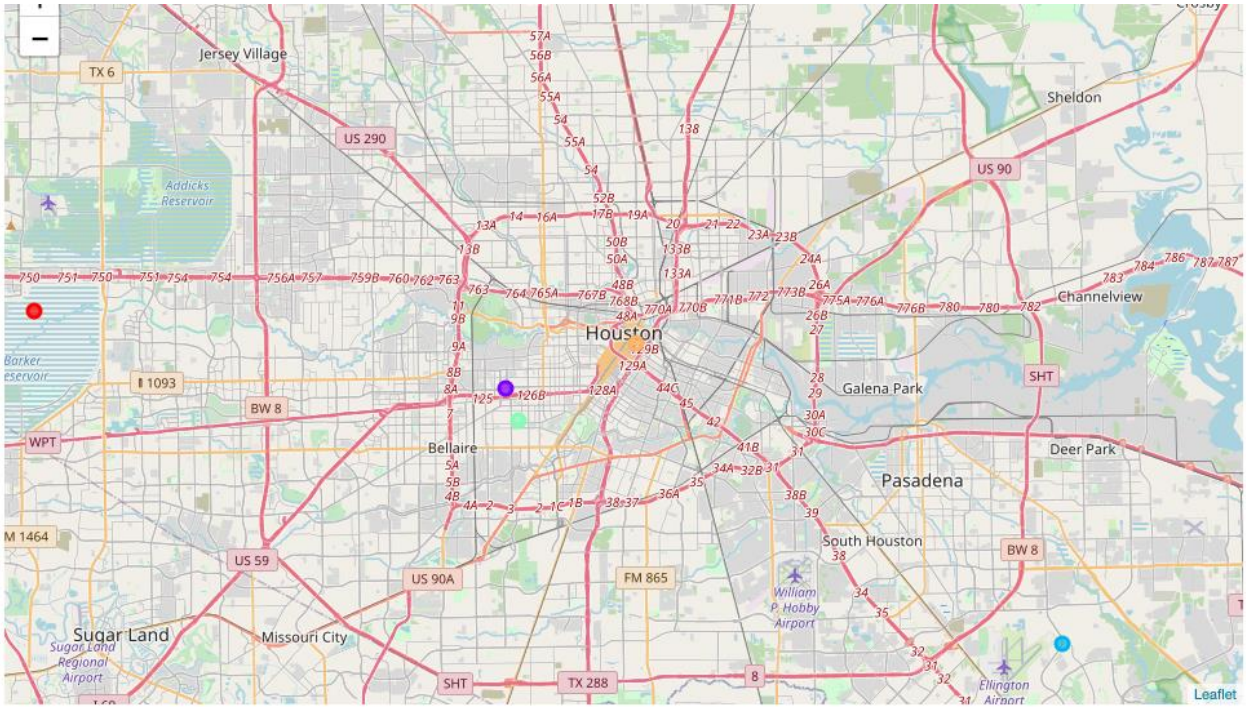
	Zip Code	Population	Avg. Income/H/hold	Latitude	Longitude
1	77010	76	$200,000.00	29.754310	-95.361109
2	77094	7779	$123,244.00	29.769285	-95.681292
3	77046	471	$105,863.00	29.733084	-95.430659
4	77059	16690	$104,844.00	29.615219	-95.134960
5	77005	23338	$104,035.00	29.718435	-95.423555

Keeping track of the datatypes of each column was critical to insure the subsequent use of the data frame's contents in various functions. Basic syntax such as the commas and dollar sign included within the average income per household column had to be removed to avoid critical errors from occurring when the data needed to be manipulated.

The next step was to narrow down the scale of the data to focus on the desired niche of the local, Houston market. To do this, I decided that a solid starting point would be to set the assumption that households with an average income of over six figures (minimum of $100,000) would be best suited to frequent an establishment of this nature. As mentioned in my opening statements, I consider customer buying power one of the most critical factors in determining a suitable neighborhood to setup a high-end café. Admittedly, households with lower levels of average income could still afford the goods and services that I would be offering. However, to be located within a wealthy, well-off neighborhood would naturally give the establishment a positive, expensive vibe that wouldn't be possible in a rundown, isolated one. This narrowed down the potential zip codes to just five from the beginning.

The next set of data to gather was the location and respective venue data. This was made possible by utilizing the Foursquare API. A function to search the vicinity of each of the five zip codes to within a kilometer radius was run to start things off. The information was grouped by zip code and then one hot encoded to condense the surrounding venue information into a single data frame. By calculating the frequency of the different types of venues around each zip code, I was able to output a single data frame with the top 5 venues per zip code. Subsequently, this information was merged with the original data frame containing the location, population, and average income per household data by indexing the zip code column.

To visualize the results of this research, k-means clustering was run with a k value of 5. Data visualization required the importing of geopy, matplotlib, sklearn.cluster, as well as the installation of folium, which took significant time to process. The following map illustrates the clusters and zip codes super-imposed on a map of Houston.



IV. Results & Discussion

The original set of data that was utilized in this project contained an extensive list of zip codes within the Houston area that were ranked by average household income. I targeted the top 5 zip codes based on a personal assumption that yearly incomes of over $100,000 would indicate a certain level of spending and taste for high-end services and products.

At first glance, this thought process would lead one to believe the area pertaining to the zip code 77010 is the most ideal location for an upscale cafe due to the fact that it has the highest average household income out of the entire list. However, it is important to take into consideration that for a retail operation, buying power is secondary to customer attendance. What this means is that while the ability for the surrounding population to buy the services and goods that I am offering is important, the frequency of customers is by far the most important factor. As such, to best garner a larger consumer base, 77010 is not an option.

Next, we have to take a careful look at the surrounding venues. Foursquare data illustrates that zip codes 77094 and 77005 are not ideal locations for a coffee shop. 77005 simply because there are already a high density of cafes within this region. The greater the competition, the less appealing a certain location will be. 77094 lacks the appeal due to the fact that the surrounding

area is dominated by service providers that are uncharacteristic of a typical, high-income residential area. The atmosphere and vibe of a cafe is its lifeforce and being surrounded by gas stations, construction and landscaping, auto dealerships, and other repair shops would be difficult to justify for high-end consumers. That leaves two candidates - 77046 and 77059.

Taking a look at the surrounding venues in zip code 77059, one can realize that there is a similar issue to 77094. While, admittedly, 77059 does have a significantly larger population of food providers in the vicinity, the price level and target population are geared more toward the general public. The area is dominated by cheap and quick eats that cater to the middle to lower financial classes of society. The people that I am targeting with this upscale cafe do not frequently visit these areas. As such, an upscale café would attract a completely different set of customers than the surrounding venues. This indicates that the synergy with the venues in the vicinity would be nonexistent for my establishment.

The reason 77046 is the perfect location for an establishment of this nature is due to two factors. The first factor is the surrounding venues. According to the data drawn from Foursquare, the top 5 most common venues in this location are all food providers (restaurants) of varying nature - Italian, New American, Mexican and Seafood. This variety facilitates the perfect location for a cafe to come in. The synergy between an upscale dessert cafe in a neighborhood dominated by middle to upscale restaurants would create the perfect location for families to stop by and spend their time and money for leisurely or formal enjoyment. A mutually beneficial relationship could easily form where the customers looking for a restaurant could chance upon my café and vice versa. Since the local establishments would all be targeting the same niche of the consumer market, there would be minimal clash of interest. Consumers wouldn't have to travel farther than a kilometer to be faced with a variety of delectable options for any day and any occasion. The second factor is the location relative to local landmarks. This zip code is located within the trifecta of Houston's most well-off areas - the Galleria, River Oaks, and Rice University. This is a strategically superior location to any other zip code due to this placement as it could draw on the collective population of all three areas, which have plenty of consumers with plenty of buying power.

V. Conclusion

The purpose of this project was to identify areas around Houston with a low number of cafes (particularly upscale establishments) in order to aid stakeholders in narrowing down the search for

the optimal location to build a new high-end bakery/cafe. By sorting zip codes within Houston based on average household income, we have first narrowed down an extensive list of potential locations throughout Houston that justify further analysis. Then, we proceeded to access Foursquare data to identify what the characteristics of each of these locations were like. The information was primarily concerning what were the five most common venues in each of the zip codes. Clustering of each of these locations was then performed in order to visualize these points of interest on a map of Houston pending final decision making by stakeholders.

The final decision on an optimal cafe location will be made by stakeholders based on the unique characteristics of each zip code location. The process will take into consideration additional factors such as the appeal of each region (proximity to high-end neighborhoods), levels of human traffic, as well as social and economic dynamics based on most common venues. The recommendation of this project is that the establishment be built in the location of zip code 77046.

VI. Possible (Future) Directions

While simple cluster points on a map is one way to visualize data, a more efficient method of data visualization for this particular application would've been using choropleth maps. This could be used in conjunction with geojson datafiles containing zip code boundary coordinates in latitude and longitude in order to draw accurate zip code boundaries throughout Houston. The choropleth maps would be useful to vividly depict the population densities throughout each wealthy neighborhood or even the real estate prices throughout each zip code region. Color-coding could be set so that ideal candidate neighborhoods that fall under specific zip codes could be colored in a brighter shade whereas unsuitable zip codes could be filtered out through darker color fills. This would enable a far more attractive and comprehensive visualization of the original dataset to present to stakeholders.