Analyzing Houston Neighborhoods for Restaurant Business Opportunity

1. Background

Houston is a well-known city in the US, located in the state of Texas. Houston is the most populous city in Texas and the fourth most populous city in the US and has around 637.4 square miles. The city of Houston contains 88 super neighborhoods. In addition to that, there are many formal and informal small regions. All these neighborhoods comprise of people from different socio-economic and ethnic backgrounds. Hence, the demographics of Houston is very diverse.

Houston is an energy hub in the US, mainly known for oil and natural gas. Renewable energy sources or companies are growing in Houston as well. Also, there are many other big companies in Houston and many attraction centers such as Space Center, Houston Zoo, and Galveston beach, which attract many tourists. All of these make Houston an ideal place for the restaurant business.

2. Problem Statement

Based on the above-stated facts, people are interested in opening a restaurant in Houston. But opening a restaurant business is not easy. It is crucial to have a proper understanding of the area and its surrounding before making any decision. Many factors help in the successful operation of a restaurant, such as the presence of some venues in the surrounding like business, apartments, theatre, attraction centers, etc. Also, knowing the number of nearby restaurants in the targeted area can help make a decision. This study is mainly helpful for people who want to open a restaurant business in Houston. The focus of this study is on:

i) Finding neighborhoods that are suitable for opening a restaurant based on the presence of other venues.

ii) Exploring the type of cuisine (restaurant) that has a high chance of getting success in those areas.

3. Data Description and Source

Next and the most important step is data collection. In order to work on the above stated problem, the following approach should be used.

- Obtain names of neighborhood in Houston using Wikipedia [1].
- Acquire latitude and longitude of neighborhood using GeoPy geocoding. Nominatim is the preferred GeoPy package for this project.
- Manually input latitude and longitude for neighborhoods for which GeoPy could not get the correct location.
- Next find out the most common venues in every neighborhood using Foursquare API.

4. Methodology

The name of the neighborhood along with their locations was extracted from Wikipedia using a web scraping technique with the help of the BeautifulSoup package in Python. Wikipedia data indicates presence of 88 neighborhoods or super neighborhoods in Houston. Next, longitude and latitude of the super neighborhoods were obtained using the Nominatim GeoPy package. Nominatim GeoPy package could only provide us geographical coordinates for 67 neighborhoods and for other 21 the values were null.

After that the map of Houston was created with 67 neighborhoods, for the ones that have longitude and latitude, as shown in figure 1. Based on figure 1, it was observed that the Nominatim provided an incorrect location for 29 neighborhoods. As a result, google search was performed to find the correct geographical coordinates for 50 (21 + 29) neighborhoods.

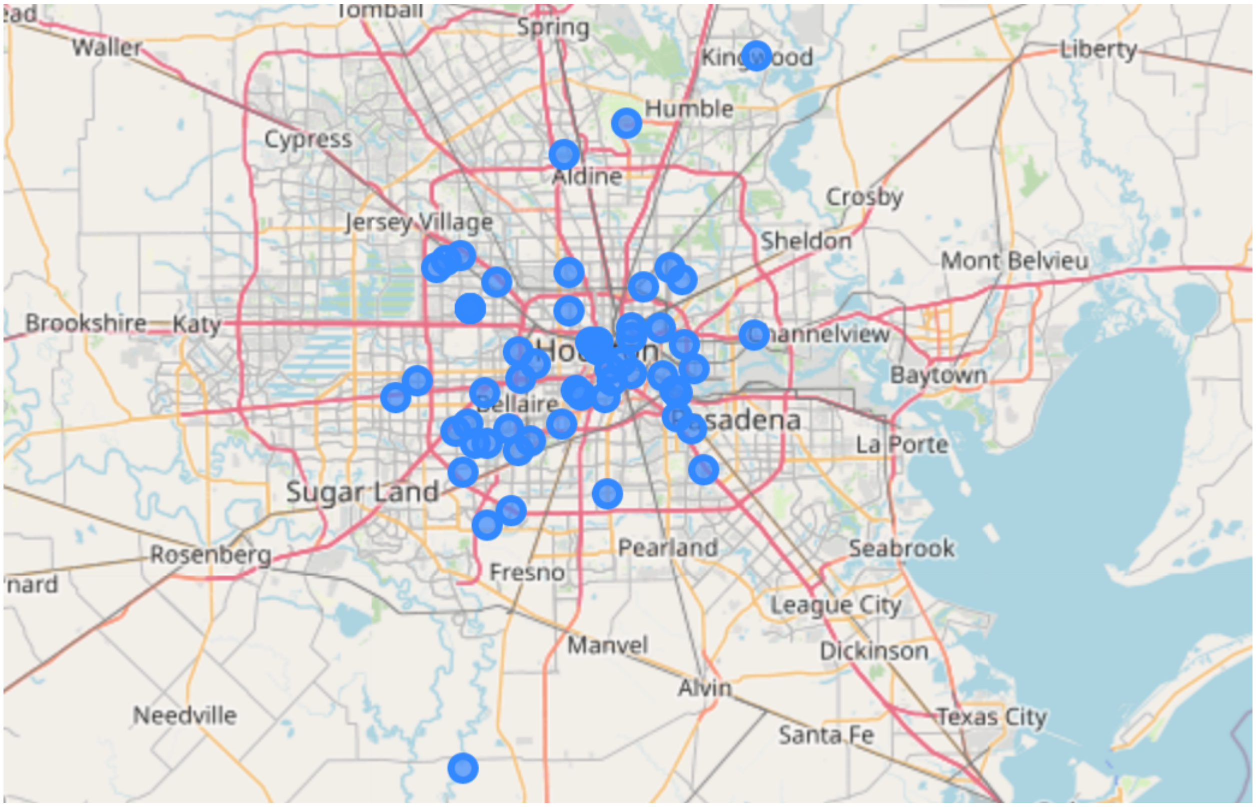

Figure 1

Next, the FourSquare API was used to get information about venues in the neighborhoods. FourSquare is the most trusted, independent location data platform for understanding how people move through the real world [2]. FourSquare developer account allows a developer to make API calls and get various information about locations and venues. Hence, foursquare API was utilized to get the top 100 venues that are within 1km of a neighborhood. Although some of the venues are similar in nature, venue categories were named differently as per data pull after calling FourSquare API.

To make data analysis more efficient, the category name for some venues were renamed. Next One-hot encoding was used in python. One-hot encoding is essentially the representation of categorical variables as binary vectors [3]. It helps the Machine Learning algorithm to do a better job as most of the existing machine learning algorithms do not work well with categorical variables. After performing One-hot encoding on venue categories, a pandas DataFrame was created by grouping neighborhoods and taking the sum of occurrence of each venue category. Again, the DataFrame was divided into two separate DataFrames based on popular venue categories selected for further clustering the neighborhood and venue categories for restaurants only. Lists of selected popular venues categories are Arts & Entertainment, Aquarium & Zoo, Auto Shop, Bank, Building / Office, Community Activities, College Place, Health & Beauty Service, Market, Museum, Shopping Mall, Stadium / Ground, Station, and Store.

Next, the K-Means clustering method was performed using the DataFrame created based on popular venues. K-Means clustering is one of the simplest unsupervised machine learning algorithms that help in clustering objects based on similar attributes. K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid [4]. In this study, the elbow method using a within-cluster sum of square (WCSS) was used to find the optimum number of clusters. As shown in figure 2, the WCSS is showing a decreasing trend even at n_cluster = 15. A cluster of 15 or more is very high for just 88 samples. Therefore, to apply K-Means, n = 6 was used. Also, at n = 6 WCSS almost got down by 1/3rd and thereafter WCSS did not drop by much. Then, further analysis was done using the DataFrame that was created using a venue category that includes only restaurants.

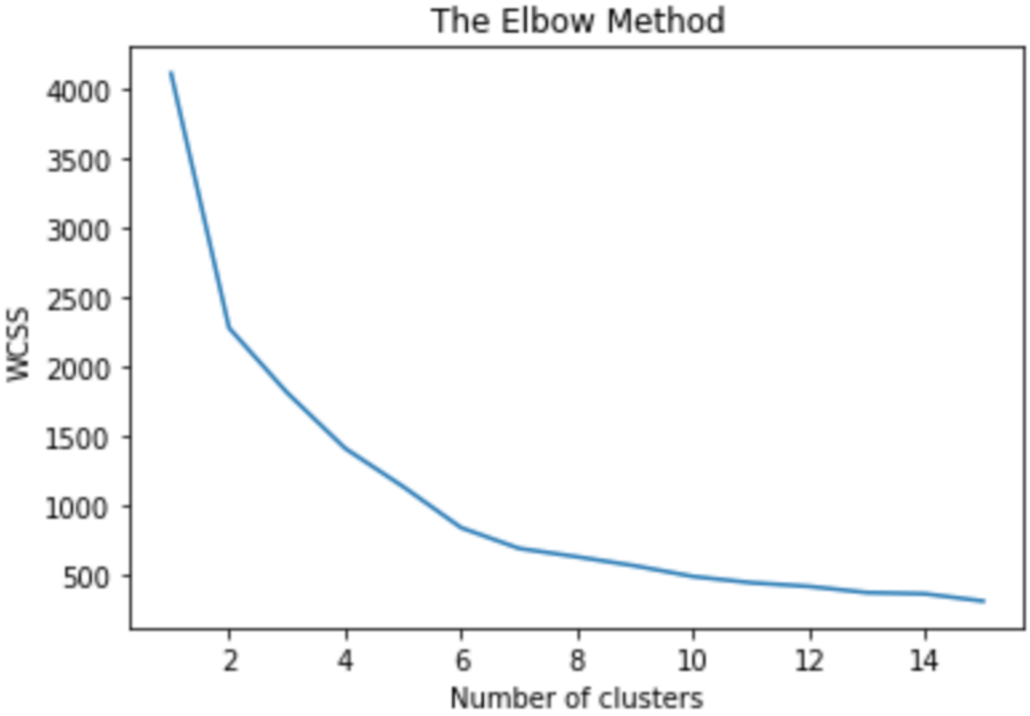

Figure 2

5. **Results**

K-Means algorithm indicated that the neighborhoods can be divided into six clusters based on popular venue categories. All the neighborhoods that belong to cluster 0 (light green) as shown in figure 3 have many popular sites (see figure 4) and hence probably would be favorable for any new owner to open a restaurant. Some of the neighborhoods in cluster 0 already have a lot of restaurants (please see figure 5) so a more granular analysis was performed.

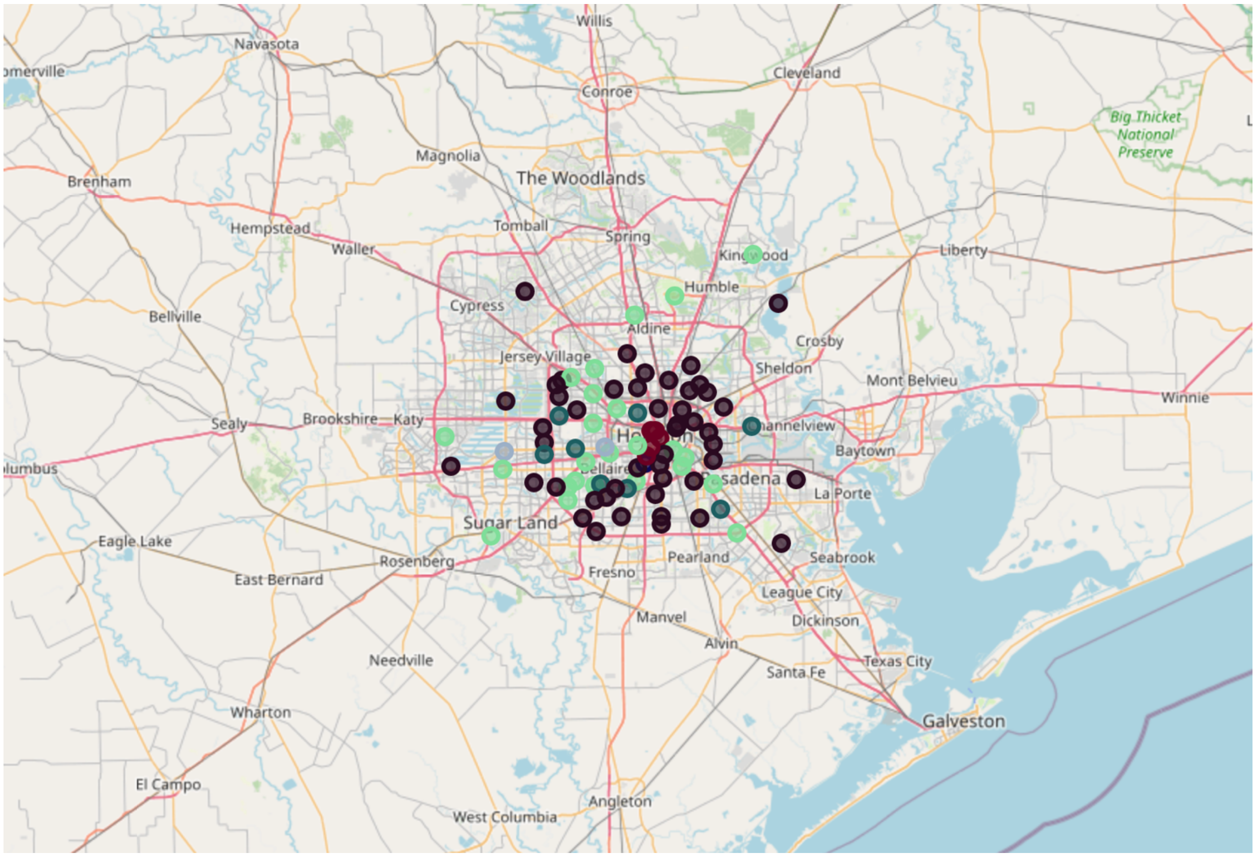
Willis
Navasota
omerville
Cleveland
Conroe
Big Thicket National Preserve
Magnolia
Brenham
The Woodlands
Tomball
Spring
Hempstead
Waller
Kingwood
Liberty
Humble
Cypress
Bellville
Aldine
Crosby
Jersey Village
Sheldon
Mont Belvieu
Winnie
Sealy
Brookshire
Katy
Houston
hannelview
Baytown
olumbus
Bellaire
sadena
La Porte
Eagle Lake
Sugar Land
Seabrook
Rosenberg
Pearland
East Bernard
Fresno
League City
Manvel
Dickinson
Alvin
Needville
Texas City
Santa Fe
Galveston
Wharton
El Campo
Angleton
West Columbia

Figure 3

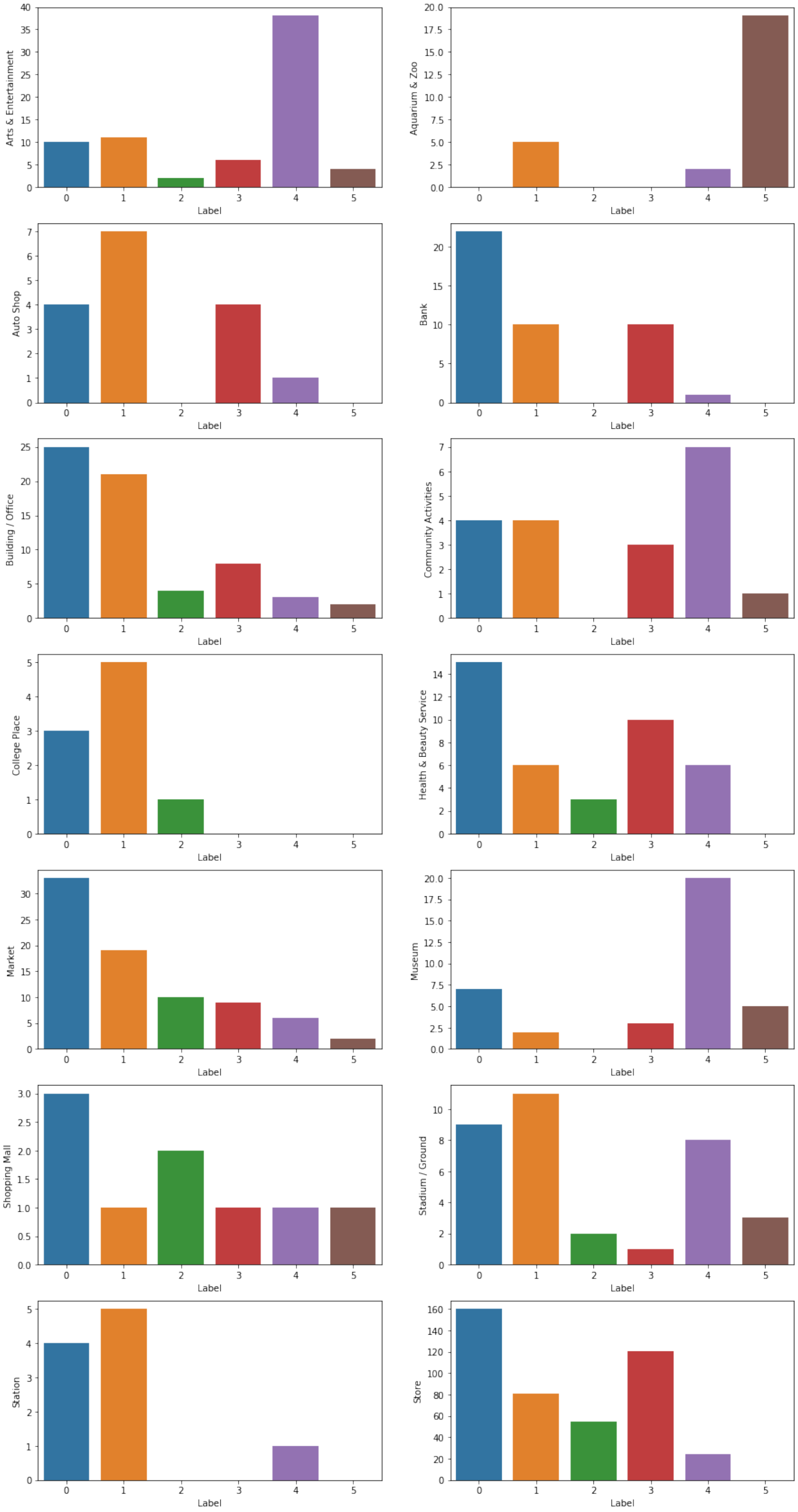
Arts & Entertainment
0
5
10
15
20
25
30
35
40
Aquarium & Zoo
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Auto Shop
0
1
2
3
4
5
6
7
Bank
0
5
10
15
20
Building / Office
0
5
10
15
20
25
Community Activities
0
1
2
3
4
5
6
7
College Place
0
1
2
3
4
5
Health & Beauty Service
0
2
4
6
8
10
12
14
Market
0
5
10
15
20
25
30
Museum
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Shopping Mall
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Stadium / Ground
0
2
4
6
8
10
Station
0
1
2
3
4
5
Store
0
20
40
60
80
100
120
140
160
0
1
2
3
4
5
Label

Figure 4

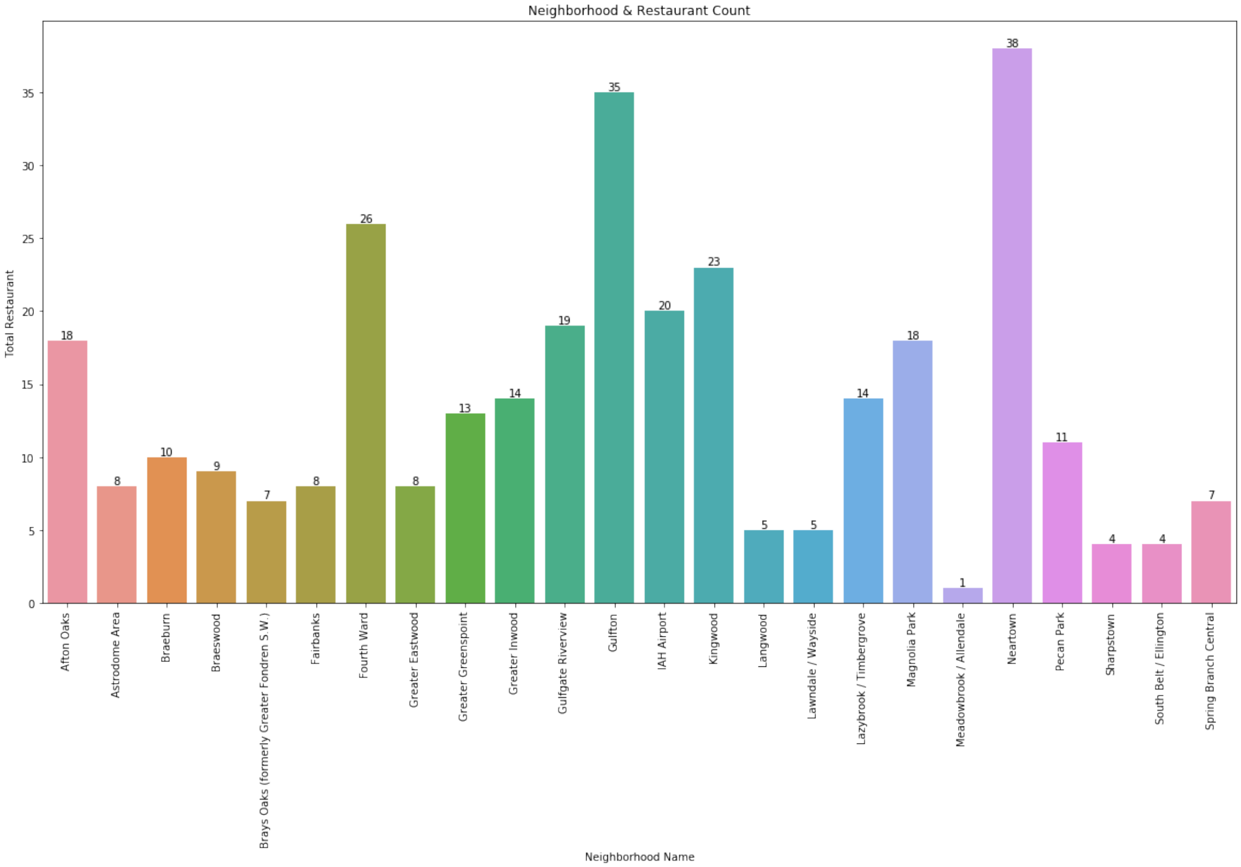

Figure 5

The analysis involved the discovery of neighborhoods with less than 10 restaurants in cluster 0. There are 11 neighborhoods in cluster 0 that have less than 10 restaurants. Hence, all of them are probably a perfect location to open a new restaurant. The names of neighborhoods are Astrodome Area, Braeswood, Brays Oaks (formerly Greater Fondren S.W.), Fairbanks, Greater Eastwood, Langwood, Lawndale / Wayside, Meadowbrook / Allendale, Sharpstown, South Belt / Ellington, and Spring Branch Central. Further analysis on these neighborhoods was done and the result is shown in figure 6.

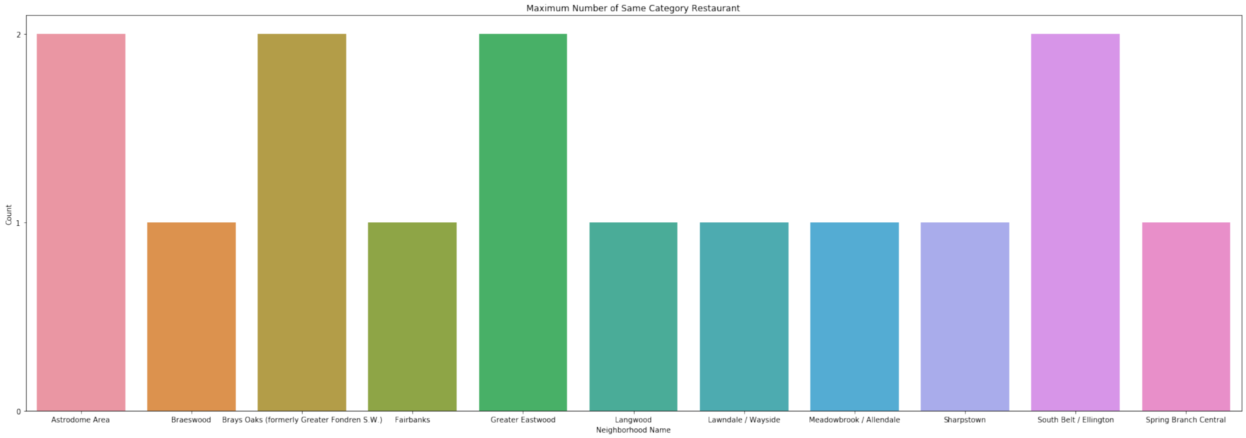

Figure 6

6. Discussion

Figure 5 indicated that the 11 neighborhoods in cluster 0 with the number of restaurants lower than 10 are ideal locations to start a new restaurant. On top of that figure 6 showed that none of those 11 neighborhoods have a high number of restaurants that fall in same category; only Astrodome Area, Brays Oaks, Greater Eastwood, and South Belt / Ellington have two of the same kind. Two is not a high number and therefore all kinds of cuisine should be able to do a good business in those areas based on this study.

The study is performed by manually selecting some of the popular venue categories with the potential to attract people. It should be noted that if different categories were chosen, the result could have been slightly different. Also, during the study, many assumptions and adjustments were made such as manual input of geographical co-ordinates for 50 neighborhoods by using google and renaming some of the venue's categories to match up with similar categories. All of these influence results. Also, this project does not include the study of population density in the neighborhood. Population density can definitely have a big impact on selecting a proper location for opening a restaurant. Higher population density means more business opportunities. For the future, if anyone is interested, a project on selecting a restaurant location in Houston neighborhood by including population density will be a good one.

7. Conclusion

Houston has a large area with 88 neighborhood and is a very popular site for any entrepreneurs who want to open a new restaurant. This project was primarily focused on finding the optimal locations to start a new restaurant in Houston along with cuisine type as described in section 2. The study commenced with getting the name of neighborhoods of Houston from Wikipedia and ended by naming some neighborhoods that have a high prospect for new restaurant and restaurant type. After extracting neighborhoods names from Wikipedia, this study went through various stages such as data pre-processing, exploratory data analysis, and implementing a machine learning algorithm before finding an answer to the problem statement in section 2. Before completing this report, it is necessary to mention that this study does have some limitations as it does not include population density into consideration while analyzing and exploring the data. Hence, a future study incorporating population density and possible other attributes could probably provide a better recommendation.

References:

1. List of Houston neighborhoods. (2020, August 07). Retrieved August 23, 2020, from https://en.wikipedia.org/wiki/List_of_Houston_neighborhoods

2. FourSquare. (n.d.). Retrieved August 23, 2020, from https://foursquare.com/about

3. One-hot encoding in Python. (n.d.). Retrieved August 23, 2020, from https://www.educative.io/edpresso/one-hot-encoding-in-python

4. Pulkit Sharma. (2020, April 23). The Most Comprehensive Guide to K-Means Clustering You’ll Ever Need. Retrieved August 23, 2020, from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/