Best Super Neighborhoods in Houston to open a Restaurant

(Part of IBM Applied Data Science Capstone Course)

- Wahab Nadir Kadiwar

Table of Contents

1.0 Introduction

1.1 Background Information

Houston is the fourth most populous city in the United States located in Southeast Texas near Galveston Bay and the Gulf of Mexico. It is often regarded as one of the most ethnically and culturally diverse metropolitan areas in the country. The city of Houston is divided into 88 Super Neighborhoods. A super neighborhood is a geographically designated area where residents, civic organizations, institutions and businesses work together to identify, plan, and set priorities to address the needs and concerns of their community.

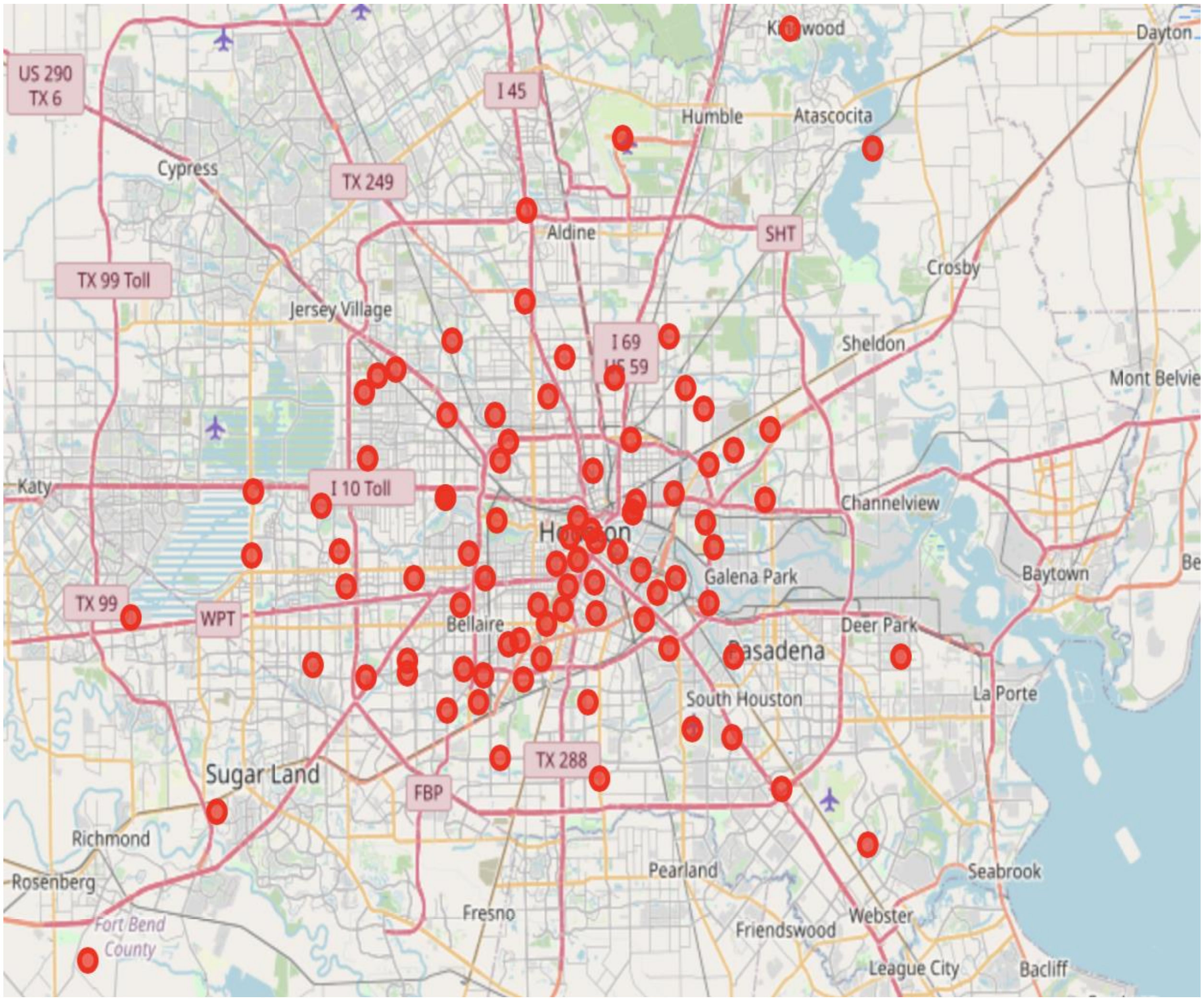


Fig 1. Super Neighborhoods of Houston

1.2 Problem Statement

The goal of this project is to analyze and categorize the Super Neighborhoods of Houston which will determine the optimal location for setting up a restaurant or chain of restaurants. The Super Neighborhoods will be evaluated based on the criteria listed below:

a. Population density
   - Restaurant in densely populated areas have a higher chance of success
b. No. of Housing
   - Greater no. of housing increases the amount of home delivery orders
c. Total Restaurants
   - Restaurant in a lively and happening place has a higher chance of attracting customers
d. Demographics
   - Helps in deciding the cuisines offered in a restaurant

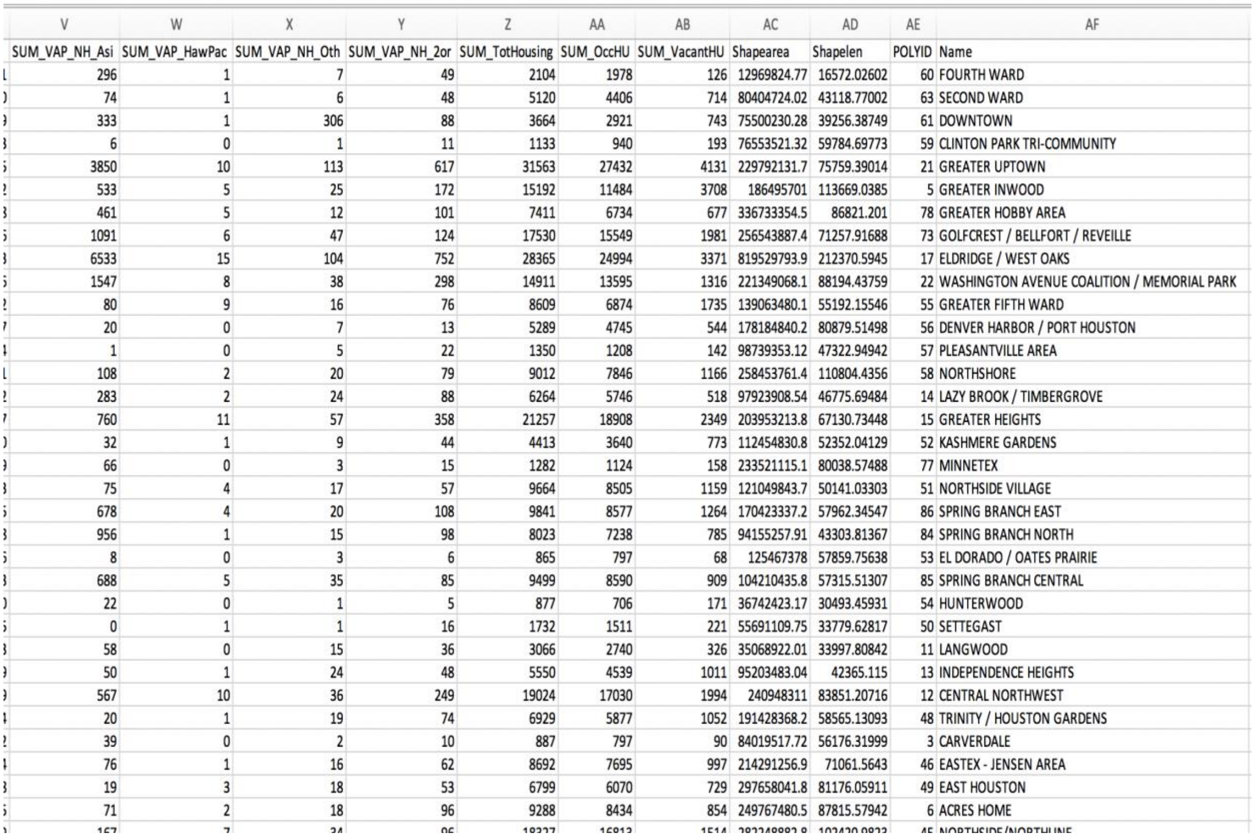
The Super Neighborhoods will be clustered to assist the investors in finalizing the cuisine, menu and location of the restaurant.

2.0 Data

2.1 Houston Census Data 2010

The Data was acquired from 'City of Houston COHGIS Portal' for the 2010 Census of the City of Houston. The data has 88 rows where each row specifies a Super Neighborhood. It comprises of 32 features viz. Name, Total Population, Area, Hispanic Population etc. The following features were fetched form the dataset:

a. Name of the Super Neighborhood
b. Total Population of each Super Neighborhood
c. Hispanic Population
d. Asian Population
e. African American Population
f. White Population
g. Total Area of each Super Neighborhood
h. Total Housing in each Super Neighborhood

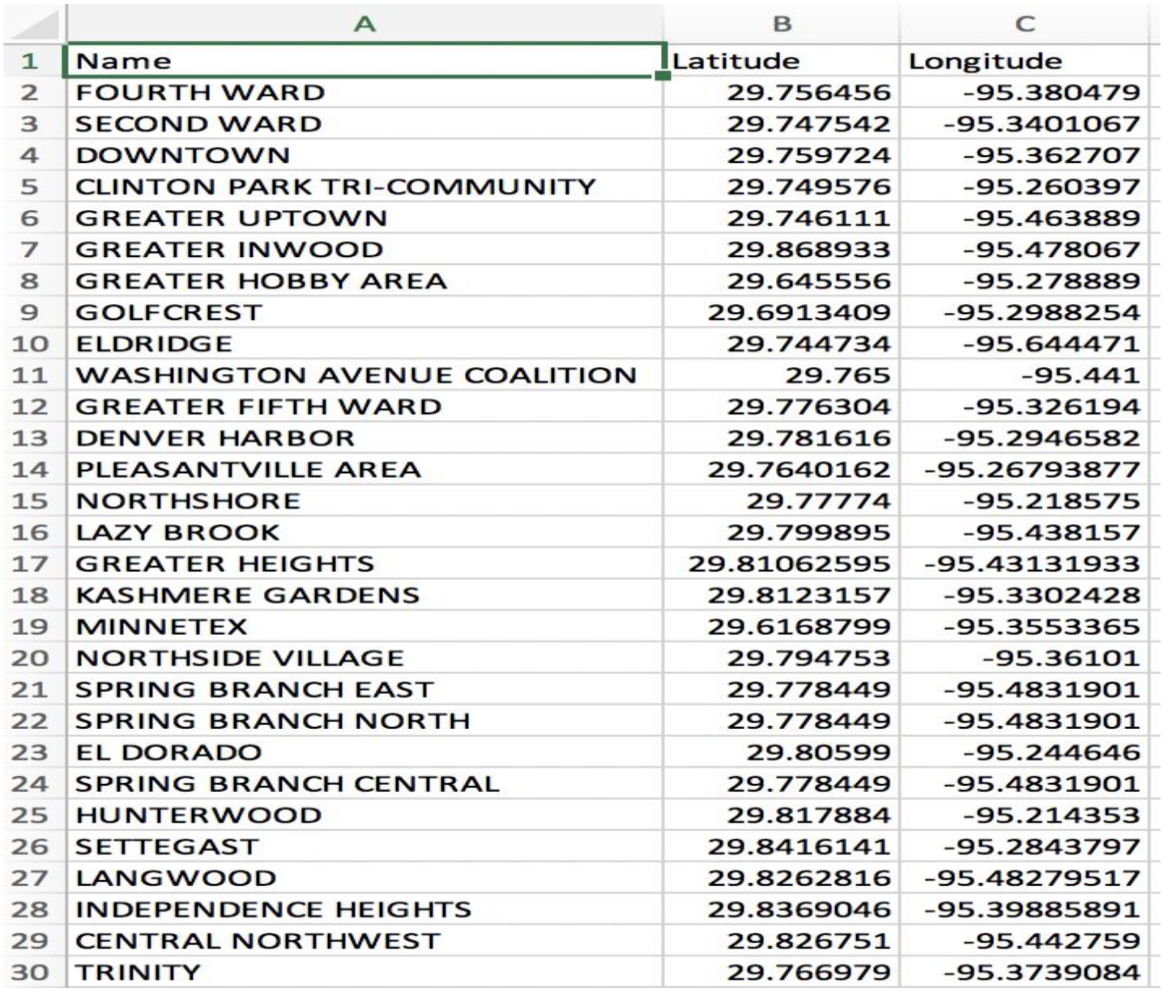


V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
SUM_VAP_NH_Asi	SUM_VAP_HawPac	SUM_VAP_NH_Oth	SUM_VAP_NH_2or	SUM_TotHousing	SUM_OccHU	SUM_VacantHU	Shapearea	Shapelen	POLYID	Name
296	1	7	49	2104	1978	126	12969824.77	16572.02602	60	FOURTH WARD
74	1	6	48	5120	4406	714	80404724.02	43118.77002	63	SECOND WARD
333	1	306	88	3664	2921	743	75500230.28	39256.38749	61	DOWNTOWN
6	0	1	11	1133	940	193	76553521.32	59784.69773	59	CLINTON PARK TRI-COMMUNITY
3850	10	113	617	31563	27432	4131	229792131.7	75759.39014	21	GREATER UPTOWN
533	5	25	172	15192	11484	3708	186495701	113669.0385	5	GREATER INWOOD
461	5	12	101	7411	6734	677	336733354.5	86821.201	78	GREATER HOBBY AREA
1091	6	47	124	17530	15549	1981	256543887.4	71257.91688	73	GOLFCREST / BELLFORT / REVEILLE
6533	15	104	752	28365	24994	3371	819529793.9	212370.5945	17	ELDRIDGE / WEST OAKS
1547	8	38	298	14911	13595	1316	221349068.1	88194.43759	22	WASHINGTON AVENUE COALITION / MEMORIAL PARK
80	9	16	76	8609	6874	1735	139063480.1	55192.15546	55	GREATER FIFTH WARD
20	0	7	13	5289	4745	544	178184840.2	80879.51498	56	DENVER HARBOR / PORT HOUSTON
1	0	5	22	1350	1208	142	98739353.12	47322.94942	57	PLEASANTVILLE AREA
108	2	20	79	9012	7846	1166	258453761.4	110804.4356	58	NORTHSHORE
283	2	24	88	6264	5746	518	97923908.54	46775.69484	14	LAZY BROOK / TIMBERGROVE
760	11	57	358	21257	18908	2349	203953213.8	67130.73448	15	GREATER HEIGHTS
32	1	9	44	4413	3640	773	112454830.8	52352.04129	52	KASHMERE GARDENS
66	0	3	15	1282	1124	158	233521115.1	80038.57488	77	MINNETEX
75	4	17	57	9664	8505	1159	121049843.7	50141.03303	51	NORTHSIDE VILLAGE
678	4	20	108	9841	8577	1264	170423337.2	57962.34547	86	SPRING BRANCH EAST
956	1	15	98	8023	7238	785	94155257.91	43303.81367	84	SPRING BRANCH NORTH
8	0	3	6	865	797	68	125467378	57859.75638	53	EL DORADO / OATES PRAIRIE
688	5	35	85	9499	8590	909	104210435.8	57315.51307	85	SPRING BRANCH CENTRAL
22	0	1	5	877	706	171	36742423.17	30493.45931	54	HUNTERWOOD
0	1	1	16	1732	1511	221	55691109.75	33779.62817	50	SETTEGAST
58	0	15	36	3066	2740	326	35068922.01	33997.80842	11	LANGWOOD
50	1	24	48	5550	4539	1011	95203483.04	42365.115	13	INDEPENDENCE HEIGHTS
567	10	36	249	19024	17030	1994	240948311	83851.20716	12	CENTRAL NORTHWEST
20	1	19	74	6929	5877	1052	191428368.2	58565.13093	48	TRINITY / HOUSTON GARDENS
39	0	2	10	887	797	90	84019517.72	56176.31999	3	CARVERDALE
76	1	16	62	8692	7695	997	214291256.9	71061.5643	46	EASTEX - JENSEN AREA
19	3	18	53	6799	6070	729	297658041.8	81176.05911	49	EAST HOUSTON
71	2	18	96	9288	8434	854	249767480.5	87815.57942	6	ACRES HOME
167	7	34	96	18327	16813	1514	[illegible]	[illegible]	45	NORTHSIDE/NORTHLINE

Fig 2. City of Houston 2010 Census Data

2.2 Super Neighborhood Coordinates

The latitude and longitude of each Super Neighborhood was retrieved using GeoPy. This information is useful to retrieve venue data for each Super Neighborhood from Four Square.

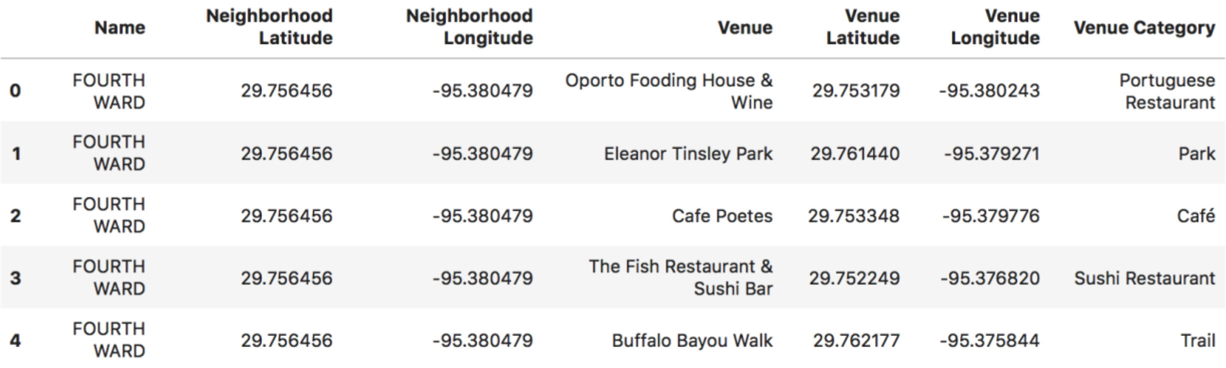


	A	B	C
1	Name	Latitude	Longitude
2	FOURTH WARD	29.756456	-95.380479
3	SECOND WARD	29.747542	-95.3401067
4	DOWNTOWN	29.759724	-95.362707
5	CLINTON PARK TRI-COMMUNITY	29.749576	-95.260397
6	GREATER UPTOWN	29.746111	-95.463889
7	GREATER INWOOD	29.868933	-95.478067
8	GREATER HOBBY AREA	29.645556	-95.278889
9	GOLFCREST	29.6913409	-95.2988254
10	ELDRIDGE	29.744734	-95.644471
11	WASHINGTON AVENUE COALITION	29.765	-95.441
12	GREATER FIFTH WARD	29.776304	-95.326194
13	DENVER HARBOR	29.781616	-95.2946582
14	PLEASANTVILLE AREA	29.7640162	-95.26793877
15	NORTHSHORE	29.77774	-95.218575
16	LAZY BROOK	29.799895	-95.438157
17	GREATER HEIGHTS	29.81062595	-95.43131933
18	KASHMERE GARDENS	29.8123157	-95.3302428
19	MINNETEX	29.6168799	-95.3553365
20	NORTHSIDE VILLAGE	29.794753	-95.36101
21	SPRING BRANCH EAST	29.778449	-95.4831901
22	SPRING BRANCH NORTH	29.778449	-95.4831901
23	EL DORADO	29.80599	-95.244646
24	SPRING BRANCH CENTRAL	29.778449	-95.4831901
25	HUNTERWOOD	29.817884	-95.214353
26	SETTEGAST	29.8416141	-95.2843797
27	LANGWOOD	29.8262816	-95.48279517
28	INDEPENDENCE HEIGHTS	29.8369046	-95.39885891
29	CENTRAL NORTHWEST	29.826751	-95.442759
30	TRINITY	29.766979	-95.3739084

Fig 3. Latitude and Longitude data retrieved from GeoPy for all Super Neighborhoods

2.3 Venue Data from Four Square

For each Super Neighborhood, 100 venues were fetched from Four Square with radius being the actual Radius evaluated for each Super Neighborhood from its total area.



	Name	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	FOURTH WARD	29.756456	-95.380479	Oporto Fooding House & Wine	29.753179	-95.380243	Portuguese Restaurant
1	FOURTH WARD	29.756456	-95.380479	Eleanor Tinsley Park	29.761440	-95.379271	Park
2	FOURTH WARD	29.756456	-95.380479	Cafe Poetes	29.753348	-95.379776	Café
3	FOURTH WARD	29.756456	-95.380479	The Fish Restaurant & Sushi Bar	29.752249	-95.376820	Sushi Restaurant
4	FOURTH WARD	29.756456	-95.380479	Buffalo Bayou Walk	29.762177	-95.375844	Trail

Fig 4. Venue Data Fetched from Four Square

3.0 Methodology

3.1 Data Cleaning

Names of the Super Neighborhoods contained alternative names and special symbols which could have caused a problem while fetching the latitude and longitude data from GeoPy. So, all the special symbols and alternate names were removed.

3.2 Data Preparation

The Names of Super Neighborhoods were used to fetch the latitude and longitude using GeoPy. Further, the below columns were created using the data from 2010 Houston City Census Dataset:

a. Population Density
   - Total Population / Total Area
b. Housing Ratio
   - Total Housing / Total Area
c. Percentage of Hispanic Population
   - Total Hispanic Population / Total Population
d. Percentage of Asian Population
   - Total Asian Population / Total Population
e. Percentage of African American Population
   - Total African American Population / Total Population
f. Percentage of White Population
   - Total White Population / Total Population
g. Total Area

Four Square API was used fetch 100 venues per Super Neighborhood. The radius in which the venues will be searched was evaluated using the Total Area of each Super Neighborhood. Further, for finding out the total restaurants from the 100 venues fetched, most common food or restaurant keywords were checked in the venue category. The list of keywords used to filter the restaurants is as follows:

["Restaurant", "Beer", "Pub", "Lounge", "Tea", "Breakfast", "Buffet", "Burrito", "Bar", "Snack", "Taco", "Food", "Hot Dog", "Chicken", "Cafe", "Steakhouse", "Sandwich", "Wings", "Deli", "Donut", "Bakery", "Salad", "Juice", "Pizza", "BBQ", "Nightclub", "Café", "Empanada", "Dessert", "Mac & Cheese", "Noodle", "Cupcake"]

After finding out if a venue was a restaurant or café, the count of restaurants per 100 venues for each Super Neighborhood was calculated and stored in the dataset.

3.3 Exploratory Data Analysis / Data Understanding

Features of the dataset were explored by plotting bar graphs of relevant columns.

a. The bar graph depicts top 10 Super Neighborhoods by highest number of restaurants per 100 venues (fetched from Four Square).

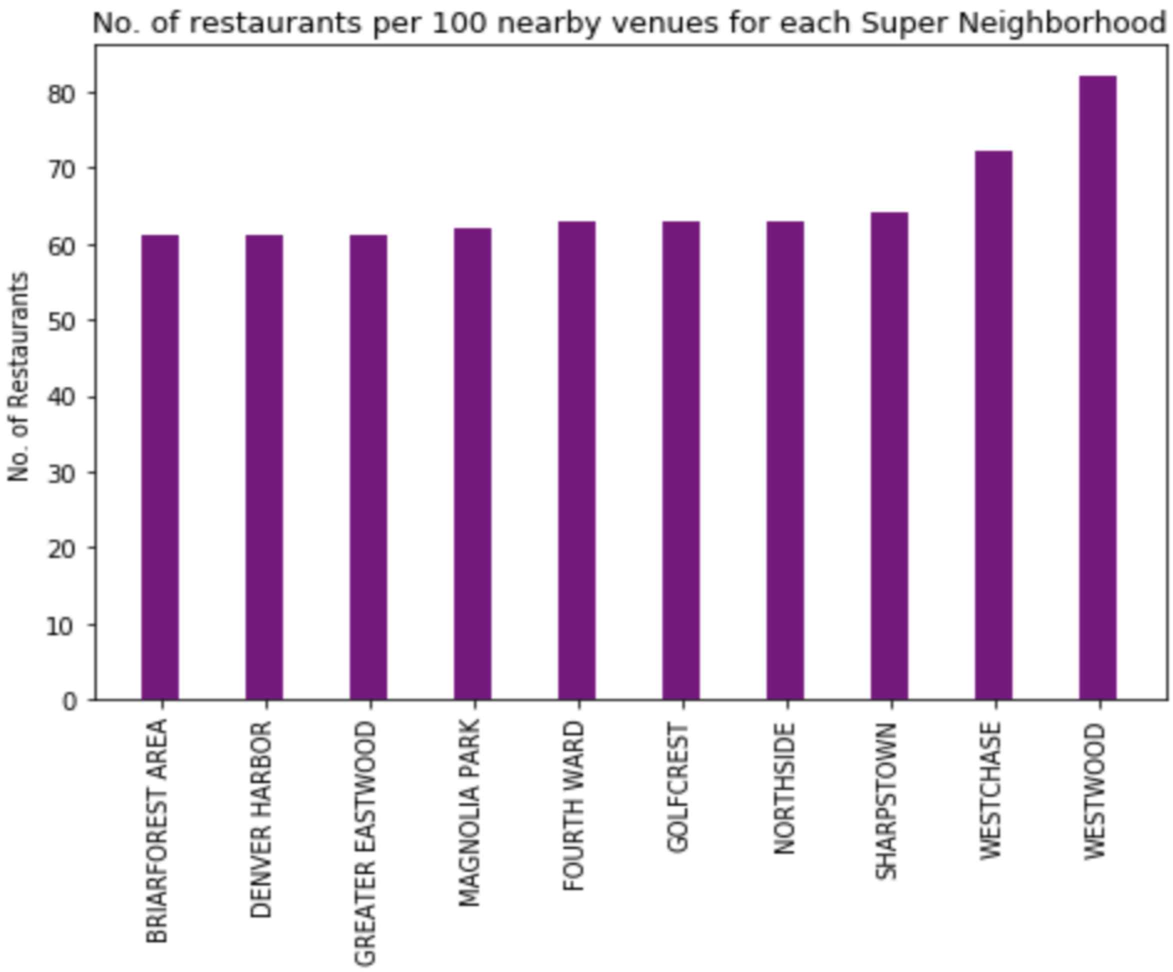


Fig 5. Bar Graph of top 10 Super Neighborhoods by most no. of restaurants

b. The below graph was plotted to visualize the top 10 Super Neighborhoods by Housing per Area.

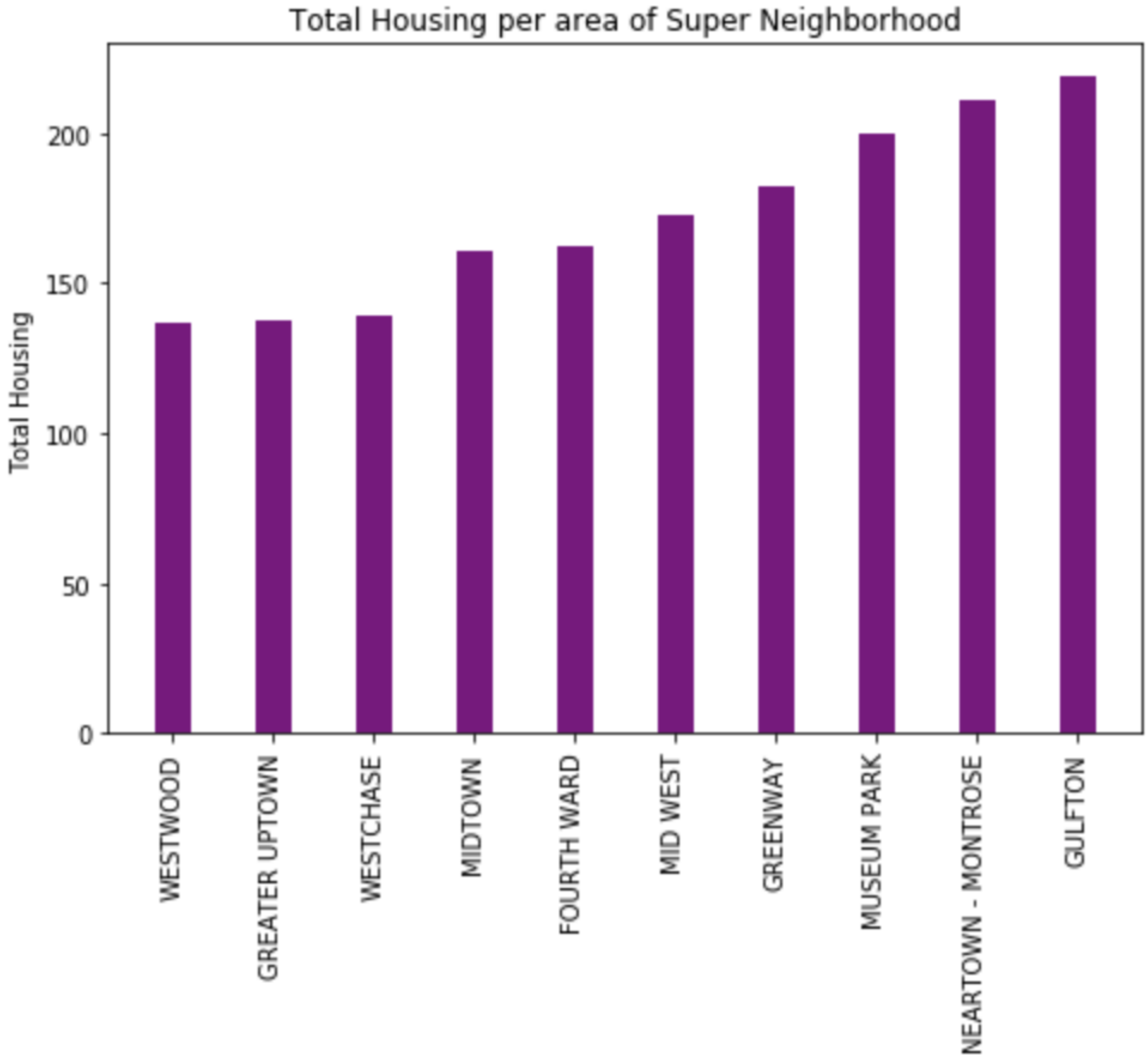


Fig 6. Bar Graph of top 10 Super Neighborhoods of Houston City by Housing per Area

c. Stacked bar graph was plotted for the percentage of population for the below Race and Ethnicity:
- Hispanic
- Asian
- African American
- White
- Others

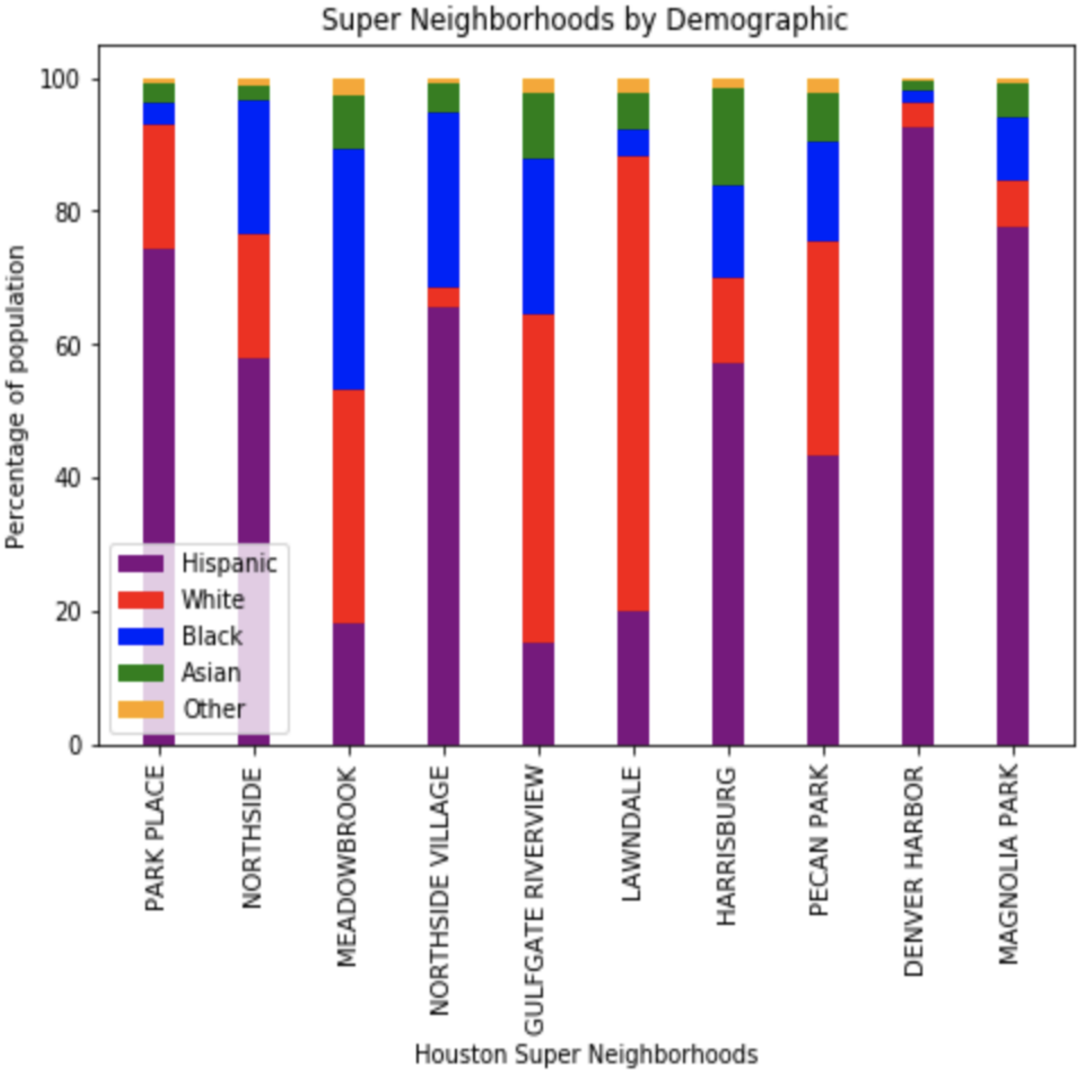


Fig 7. Stacked Bar Graph depicting the Race and Ethnicity of Super Neighborhoods

3.4 Data Scaling

The dataset contained values in different units which can affect clustering results. For example, race and ethnicity columns are in percentage but ‘total restaurants’ is a count of restaurants. To avoid any unexpected results, the data was scaled using the Min Max Scaler, which converts all the data into values ranging from 0 to 1.

3.5 K-Means Clustering

To categorize the Super Neighborhoods using K-Means clustering algorithm, number of clusters has to be passed as an input. So, to evaluate the optimal value of number of clusters, Elbow method was used, which utilizes the sum of squared distance. The optimal value is when the line graph makes around 90 degrees rotation. As seen from the graph, the optimal value for the number of clusters should be 4 or 5.

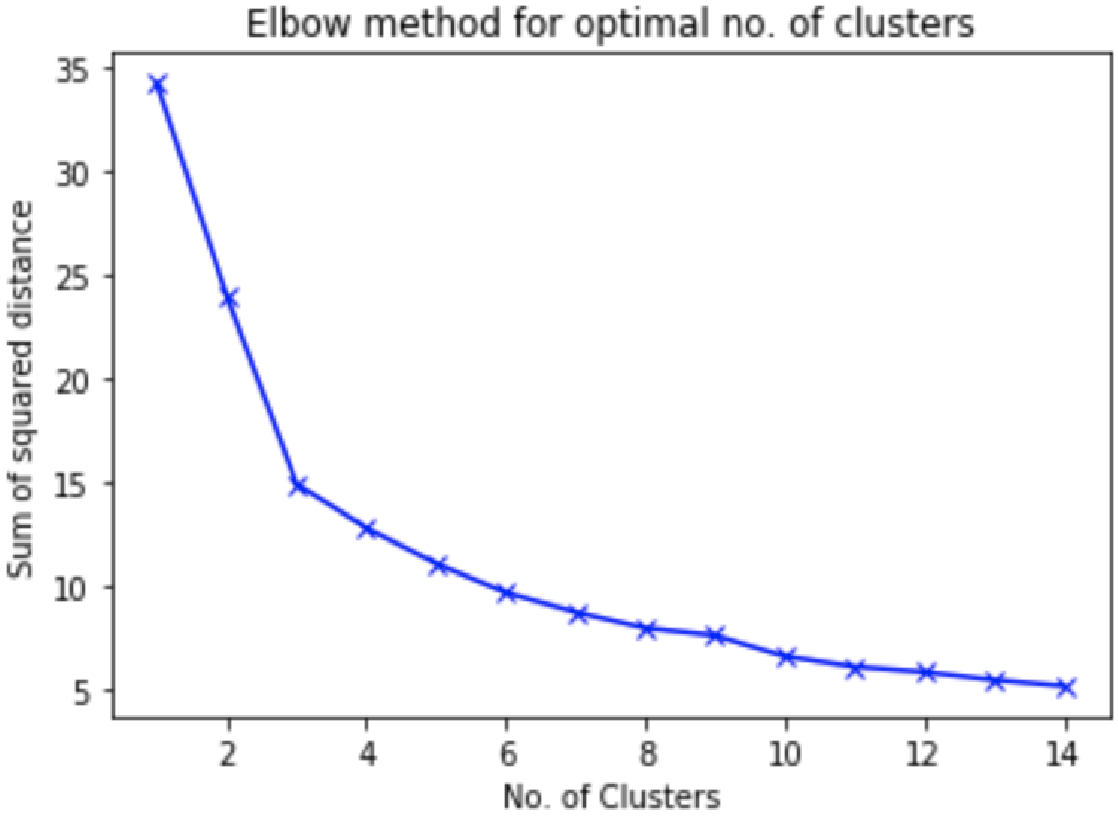


Fig 8. Elbow method to find out the optimal value for cluster centers

After finding out the optimal value of number of clusters as 5, K-Means algorithm was run on the dataset to categorize the Super Neighborhoods. After obtaining the cluster labels for each record in the dataset, the columns such as ‘Latitude’, ‘Longitude’ and Cluster Labels were added to the dataset for visualizing on a map. Using Folium, the map of Houston was marked with all the Super Neighborhoods in different colors according to the clusters they belong.

4.0 Results

K-Means clustering categorized the Super Neighborhoods of Houston City into 5 clusters which can be seen in the map below:

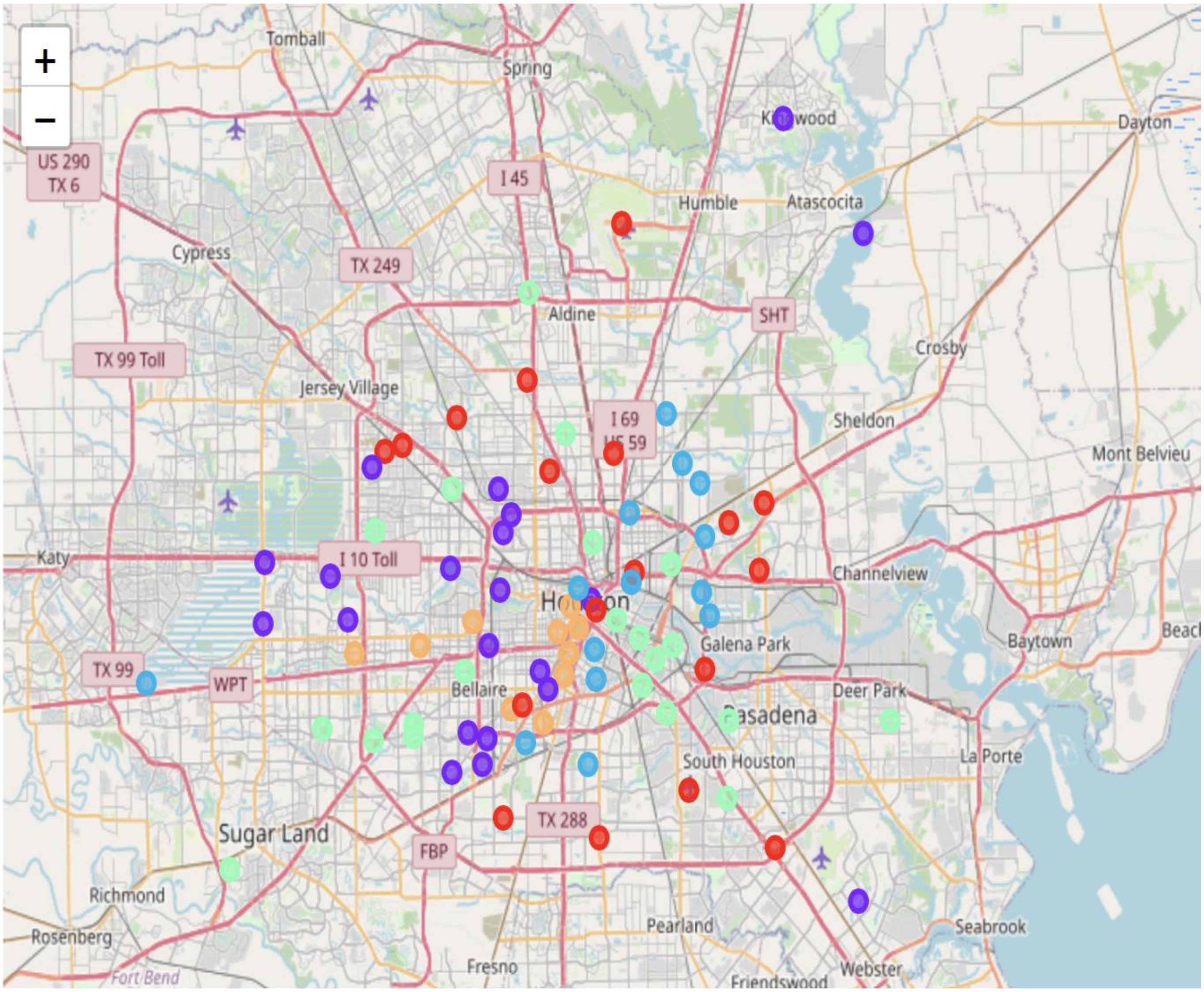


Fig 9. Map of Houston with Clustered Super Neighborhoods

The Clusters can be categorized as follows:
Cluster 1: Orange
Characteristics: Happening places of Houston. Some of the best residential areas.
Some Super Neighborhoods: Midtown, Montrose, Greenway, Greater Uptown

Cluster 2: Blue
Characteristics: Not very developed areas.
Some Super Neighborhoods: Mac Gregor, South Main, Sunny Side

Cluster 3: Green
Characteristics: Areas with high population density but low housing ratio
Some Super Neighborhoods: Westwood, Sharpstown, Golfcrest

Cluster 4: Purple
Characteristics: Areas with very less housing and population. Generally, office and work areas.
Some Super Neighborhoods: Downtown, Medical Center, University Place

Cluster 5: Red
Characteristics: Most of the super neighborhoods in this cluster are on the outskirts of the city.
Some Super Neighborhoods: South Belt, IAH, Hidden Valley

5.0 Discussion

Investors who are planning to invest in a restaurant or restaurant chain in Houston City can get insights from the clustering results as below:

a. Best Place to Open a Fancy and lavish Restaurant -> Cluster 1
   Cluster 1 had all the lively and happening super neighborhoods, so it gives the restaurant a higher chance of success. These places are already filled with crowds; therefore, the investors don't have to worry about drawing people. Also, as these super neighborhoods have some of the highest housing ratios, there will be more home delivery orders compared to other super neighborhoods.

b. Best Place to open Fast food restaurant -> Cluster 2
   The Super Neighborhoods belonging to cluster 2 are at a developing stage. So, investing in a Fancy Restaurant could be risky. Therefore, the best option for these areas is a fast food restaurant like Subway or Dominos. Food at these places is cheap when compared to other lavish

restaurants and also people have already heard about these places, so this gives investor one less task.

c. Best Place for a Deli / Lunch /Breakfast Places -> Cluster 4
Most of the Super Neighborhoods in this cluster are in the office and work areas, so there are high chances that there will be more people interested in Lunch or to-go boxes. For investor, the only concern in these areas should be that the orders need to process at a high rate because there will be a lot of crowd during the lunch hours for a relatively shorter period of time.

Some of the Super Neighborhoods of Green Cluster could also be good spots for a restaurant but some additional research needs to be done as these places are very crowded, but they do not have high housing ratio, so there may be high chances of crime. Similarly, additional research needs to be done for super neighborhoods belonging to red cluster as these places have less population and also, they are located on the outskirts of the city.

6.0 Conclusion

A research has been conducted in identifying the best super neighborhoods of Houston to open different kinds of restaurant. Features such No. of Housing, population density, demographics and how happening the place is were taken into consideration. Cluster 1 was identified as the best one for opening a Fancy and lavish restaurant. Cluster 2 was best for a fast food restaurant and cluster 3 was best for a Deli/Breakfast/Lunch places.

8.0 Future Work

The research conducted can be made useful to audiences that are looking for a good super neighborhood to live by adding the below features:

a. Crime Rate in each area
b. Housing Cost
c. No. of schools in the neighborhood
d. No. of Leisure places in the neighborhood (Example Trail, parks, etc.)

References

1. https://en.wikipedia.org/wiki/Houston
2. https://cohgis-mycity.opendata.arcgis.com
3. https://en.wikipedia.org/wiki/List_of_Houston_neighborhoods
4. https://developer.foursquare.com/