Clustering of Houston City Neighborhood

For Coursera capstone project

Jason Chen, June 2020

Introduction

According to Wikipedia [1] Houston is the most populous city in TX with an estimated population in 2019 of 2.3 million [2]. **Figure 1** shows that from 1950 to 2010 the population of Houston city has a steady annual growth rate of 20,000 people/year. People move to Houston for its plentiful jobs, education, and favorable housing price [3]. There are approximately 90 so called super neighborhoods (**Figure 2**) in Houston city [4]. The neighborhoods have diverse attributes regarding population characteristics, age of population, ethnicity, income, education status, median household income, median housing value, and so on. For anyone who wants to move to Houston or anyone who wants to open a small business (for example, a day-care center, spider-smart training center, or a restaurant with a typical food type) it would be important to know the clustering and distribution of these diverse neighborhoods, so that an optimal location can be chosen for the small business. This study is about the clustering of Houston city neighborhoods using the various demographic and other data obtained online and the Foursquare location data.

Data

On its main website [4], the super neighborhood's demographics, land use map, and resource assessment are summarized in PDF files accessible through hyperlinks. **Figure 3** shows an example of the first page of the PDF file that contains the feature data regarding each neighborhood, including population characteristics (total population, and persons per sq. mile), age of population (percentage of under 5 years, 5-17 years, 18-64 years, 65 and over), ethnicity (non-Hispanic Whites, non-Hispanic Blacks, Hispanics, Non-Hispanic Asians, and non-Hispanic others), income categories (under 25,000, 25,001 – 50,000, 50,001 – 100,000, over 100,001), median household income, educational status (no diploma, high school diploma, some college, bachelor's or higher), and housing and households (total housing units, occupied, vacant, total households, family households, and median housing value). In total 76 PDF files are downloaded excluding those neighborhoods that do not have such online data.

The other major data type is the Foursquare search results of the various vendors close by each neighborhood. The longitude and latitude of the neighborhood for the Foursquare search is obtained through Google search and summarized in an Excel file. Three neighborhoods are excluded in this process as they are in the airport or lake area.

Data preparation

Data preparation is performed with the following steps:

- Latitude and longitude data for each neighborhood is obtained through Google search
- 73 PDF files are parsed using PDFMINER to convert PDF into text
- The feature data of each neighborhood is obtained through string search of the above text file
- The various data from above is combined into a data frame.

The resulting data frame is shown in **Figure 4**. This data frame along with the Foursquare search results will be used to cluster the neighborhood using different approaches.

Methodology

The methodology applied in the data analysis include poly-normial regression and the k-mean clustering. Poly-normial regression is used to explore the relationship between median household income and median housing value. There are 73 data points in total. The data points are split into training and test data set (15% of data is used for test). The test set is used to determine the degree of the poly-normial fitting. Maximum R^2 is reached when the degree = 2 (**Figure 5**). Therefore, 2^{nd} order poly-normial regression is used to fit the data between median household income and median housing value. R^2 of the regression is 0.723 (**Figure 5**)., which suggests that there is a good correlation between median household income and median housing value. Not surprising as people with higher income tend to buy more expensive houses.

k-mean clustering is used to segment the neighborhoods into different groups. The segmentation is done using 5 types of different data:

1.) median household income,
2.) median housing value,
3.) age of population,
4.) Ethnicity,
5.) Vendor data from Foursquare searching.

The focus of the segmentation or clustering of the neighborhoods is to provide answer to the optimum location for a potential small business, for example, new day-care center for Hispanics or new senior center for Asian group. 7 clusters are chosen as it provides good solutions. 5 clusters are used for the vendor data from Foursquare searching for simplicity purpose.

Results

The clustering results are shown in the following Figures.

Figure 6: clustering by median household income.

Figure 7: clustering by median housing value.

Figure 8: clustering by age of population.

Figure 9: clustering by age of population: potential location for day-care center and senior center.

Figure 10: clustering by Ethnicity.

Figure 11: clustering by Ethnicity: location for Hispanic community and non-Hispanic Asian community.

Figure 12: clustering by vendors from Foursquare searching results: Cluster 1 ●: Asian restaurant, Cluster 2 ●: Mexican restaurant, Cluster 3 ● : Coffee/Hotel, Cluster 4 ● : Mexican / ice-cream, Cluster 5 ● : Discount store.

Figure 13: Potential location for a new day – care center for the Hispanic community.

Figure 14: Potential location for a new senior care center for the non-Hispanic Asian community.

Conclusions

The following conclusion can be drawn from this study:

1.) Clustering method can be used to identify potential business locations. For example, this study identifies potential new day – care center for Hispanic community and new senior care center for non-Hispanic Asian community.
2.) Median housing value has a linear relationship with median household income, the clustering results with respect to median housing value and median household income are identical (shown in **Figure 15**).

Discussion

The foursquare searching results tend to be concentrated with vendors such as restaurants, coffee shops, and stores.

Figures

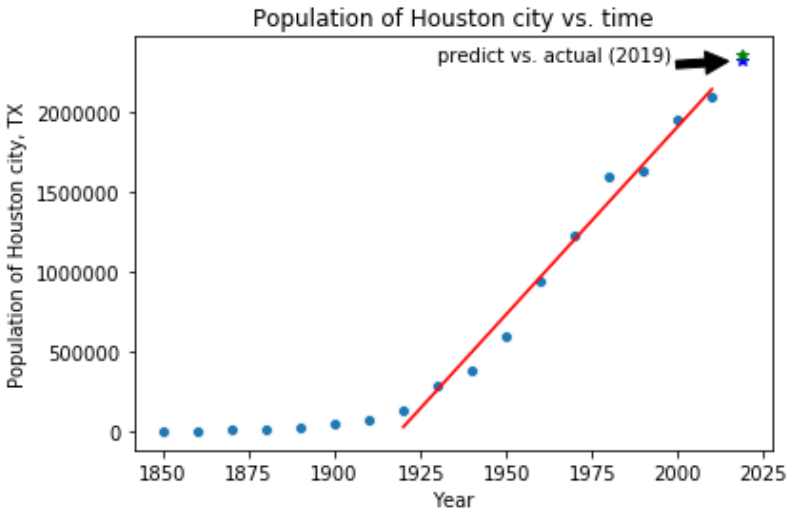


Figure 1. population of Houston from 1850 to 2010. A linear regression was performed using the historical data. The prediction of the population in 2019 matches well with the actual data.

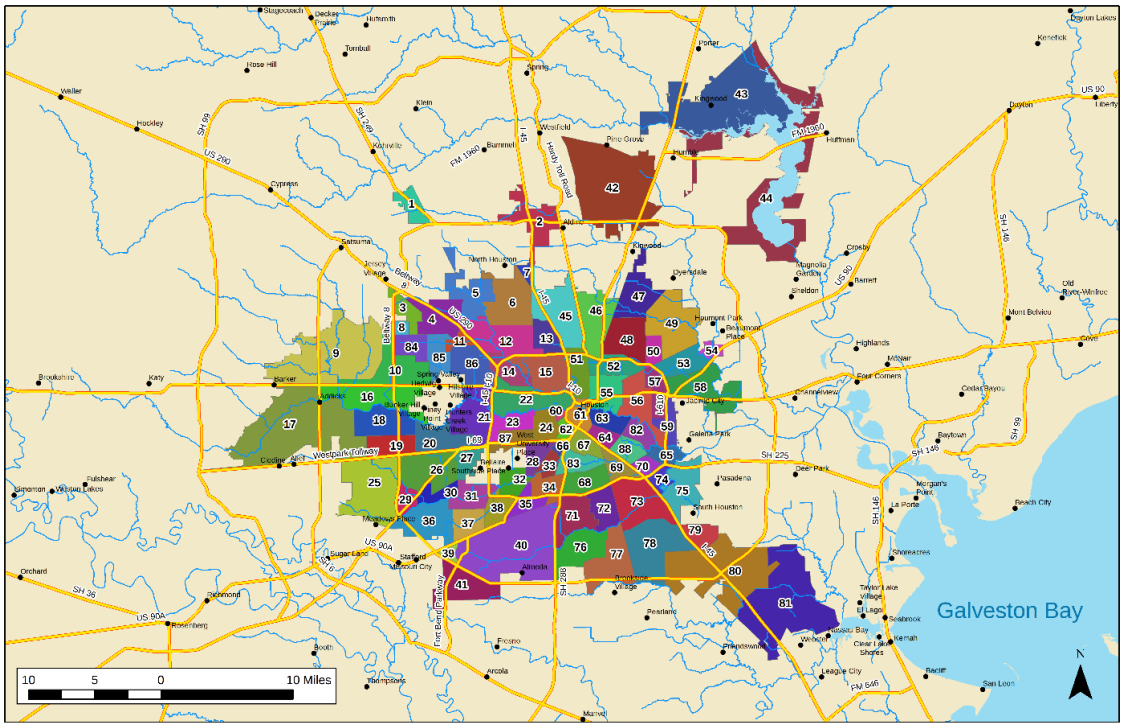
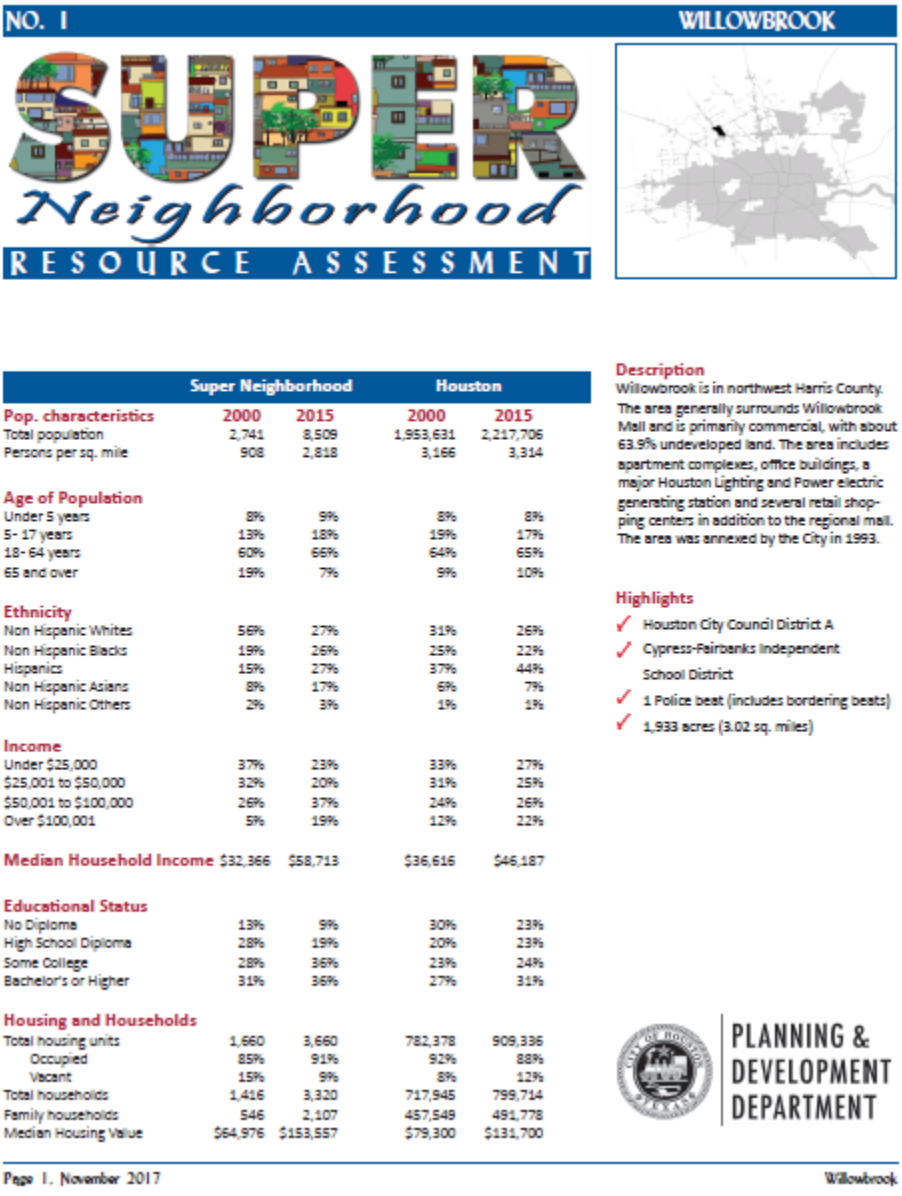


Figure 2. Neighborhoods of Houston city, TX [1].



NO. 1

WILLOWBROOK

SUPER Neighborhood

RESOURCE ASSESSMENT

	Super Neighborhood		Houston	
Pop. characteristics	**2000**	**2015**	**2000**	**2015**
Total population	2,741	8,509	1,953,631	2,217,706
Persons per sq. mile	908	2,818	3,166	3,314
Age of Population				
Under 5 years	8%	9%	8%	8%
5- 17 years	13%	18%	19%	17%
18- 64 years	60%	66%	64%	65%
65 and over	19%	7%	9%	10%
Ethnicity				
Non Hispanic Whites	56%	27%	31%	26%
Non Hispanic Blacks	19%	26%	25%	23%
Hispanics	15%	27%	37%	44%
Non Hispanic Asians	8%	17%	6%	7%
Non Hispanic Others	2%	3%	1%	1%
Income				
Under $25,000	37%	23%	33%	27%
$25,001 to $50,000	32%	20%	31%	25%
$50,001 to $100,000	26%	37%	24%	26%
Over $100,001	5%	19%	12%	22%
Median Household Income	$32,366	$58,713	$36,616	$46,187
Educational Status				
No Diploma	13%	9%	30%	23%
High School Diploma	28%	19%	20%	23%
Some College	28%	36%	23%	24%
Bachelor's or Higher	31%	36%	27%	31%
Housing and Households				
Total housing units	1,660	3,660	782,378	909,336
Occupied	85%	91%	92%	88%
Vacant	15%	9%	8%	12%
Total households	1,416	3,320	717,945	799,714
Family households	546	2,107	457,549	491,778
Median Housing Value	$64,976	$153,557	$79,300	$131,700

Description

Willowbrook is in northwest Harris County. The area generally surrounds Willowbrook Mall and is primarily commercial, with about 63.9% undeveloped land. The area includes apartment complexes, office buildings, a major Houston Lighting and Power electric generating station and several retail shopping centers in addition to the regional mall. The area was annexed by the City in 1993.

Highlights

- ✓ Houston City Council District A
- ✓ Cypress-Fairbanks Independent School District
- ✓ 1 Police beat (includes bordering beats)
- ✓ 1,933 acres (3.02 sq. miles)

PLANNING & DEVELOPMENT DEPARTMENT

Page 1. November 2017

Willowbrook

Figure 3. Example of PDF file of super neighborhood data sheet [4].

```
df_data = pd.DataFrame.from_dict(data, orient='index',
                                 columns=['Income_Median', 'Housing_Median', 'Population_Total',
                                          'Psmile_Person', 'Age%_Under5', 'Age%_5_17','Age%_18_64',
                                          'Age%_65up', 'Non_Hispanic%_W', 'Non_Hispanic%_B',
                                          'Hispanic%', 'Non_Hispanic%_A', 'Nin_Hispanic%_O',
                                          'Income%_U25k', 'Income%_U50k', 'Income%_U100k',
                                          'Income%_100kup', 'No_Diploma%', 'High_School%',
                                          'College%', 'Bachelor_Or_Higher%']
                                 )

df_data['Index'] = df_n['Index'].to_numpy()
dft = pd.merge(df_n, df_data, on='Index')
dft.head()
```

	HoustonNeighborhoods	Index	Latitude	Longitude	Income_Median	Housing_Median	Population_Total	Psmile_Person	Age%_Under5	Age%_5
0	9_Addicks_Park_Ten	9	29.813300	-95.645500	80584.0	168155.0	19683.0	840.0	5.0	
1	23_AftonOaks_RiverOaks	23	29.749994	-95.433234	95682.0	499169.0	14518.0	4021.0	5.0	
2	25_Alief	25	29.682700	-95.593200	41833.0	90655.0	106657.0	7544.0	7.0	
3	34_Astrodome_Area	34	29.685045	-95.409813	46284.0	102268.0	18223.0	4846.0	5.0	
4	30_Braeburn	30	29.682779	-95.534980	42958.0	116547.0	18843.0	4711.0	8.0	

5 rows × 25 columns

Figure 4. The head information for the combined data frame.

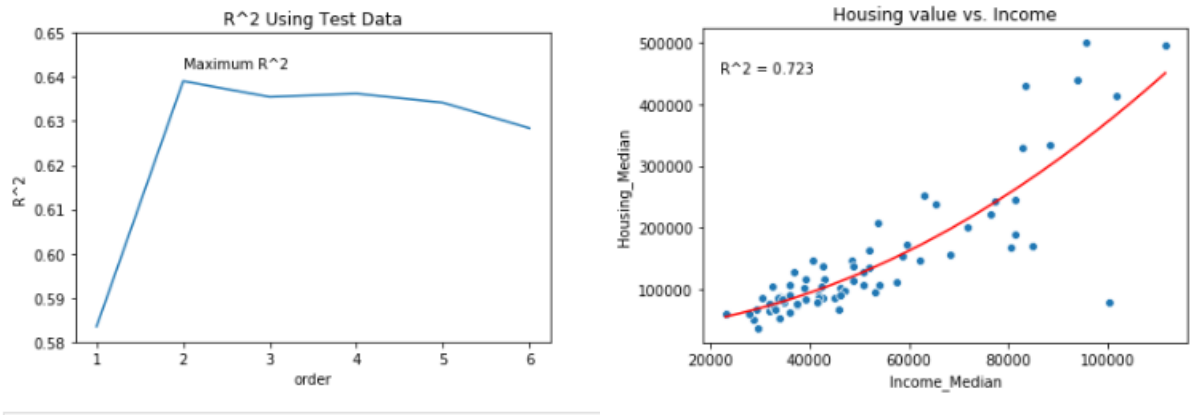
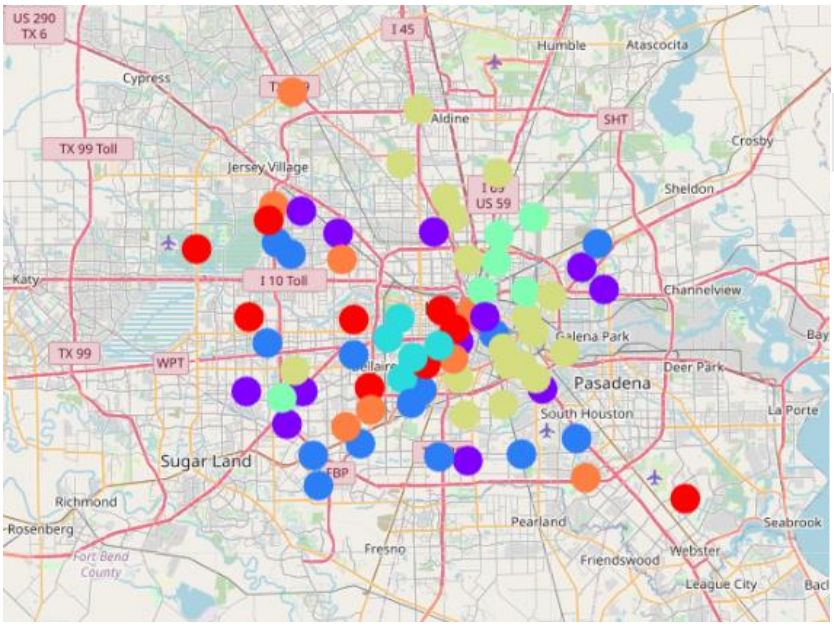


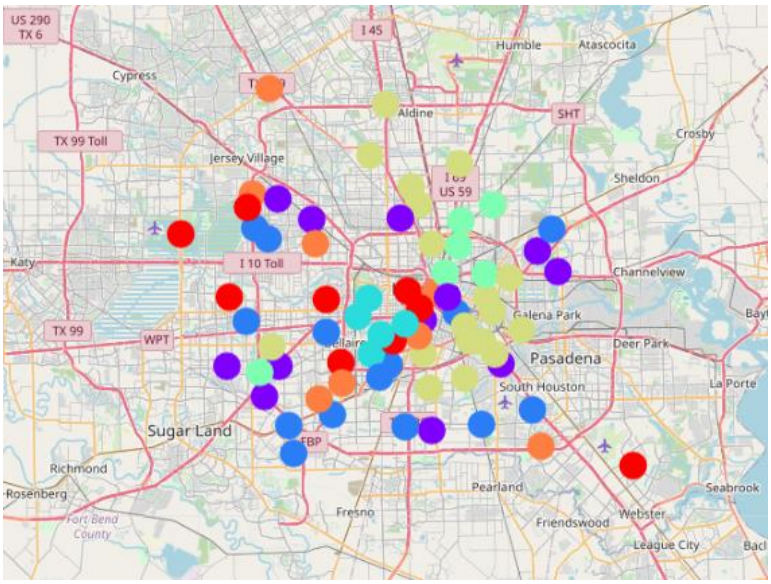
Figure 5. 2nd order poly-normial regression on the relationship between median housing value and median household income (R^2 = 0.723).



dfc1_clustering_grouped

Cluster Labels	Income_Median
0	81791.666667
1	41576.000000
2	49867.214286
3	100583.400000
4	28174.500000
5	34331.105263
6	63292.250000

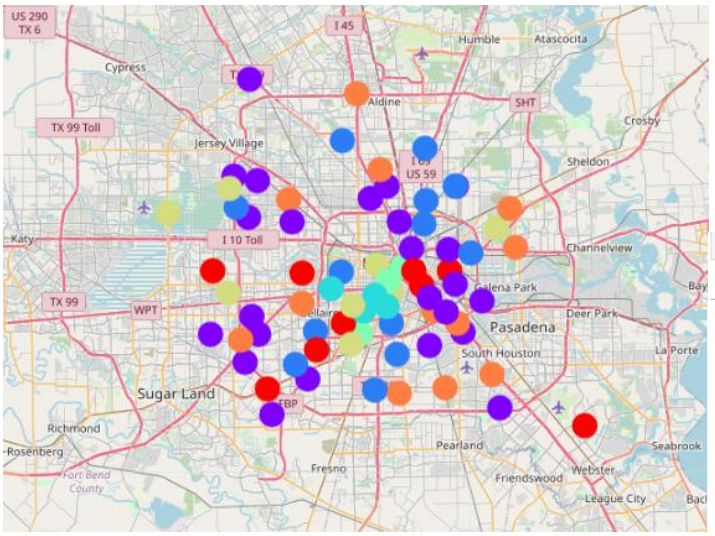
Figure 6: clustering by median household income.



dfc2_clustering_grouped

Cluster Labels	Housing_Median
0	259073.444444
1	103866.916667
2	122021.142857
3	385376.400000
4	61588.500000
5	81776.368421
6	179179.500000

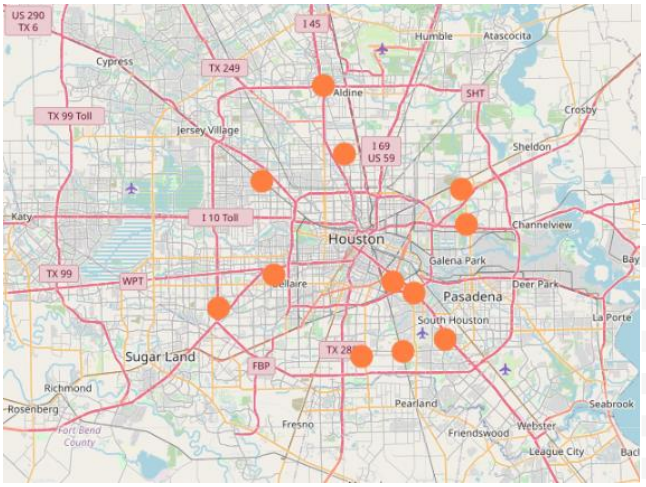
Figure 7: clustering by median housing value.



dfc3_clustering_grouped

Cluster Labels	Age%_Under5	Age%_5_17	Age%_18_64	Age%_65up
0	7.222222	14.777778	65.666667	12.333333
1	8.083333	19.833333	63.291667	8.666667
2	6.692308	18.153846	60.000000	15.384615
3	5.500000	5.500000	76.750000	12.000000
4	3.000000	4.333333	87.666667	4.666667
5	5.500000	13.125000	73.875000	7.500000
6	10.333333	22.916667	61.250000	5.666667

Figure 8: clustering by age of population.



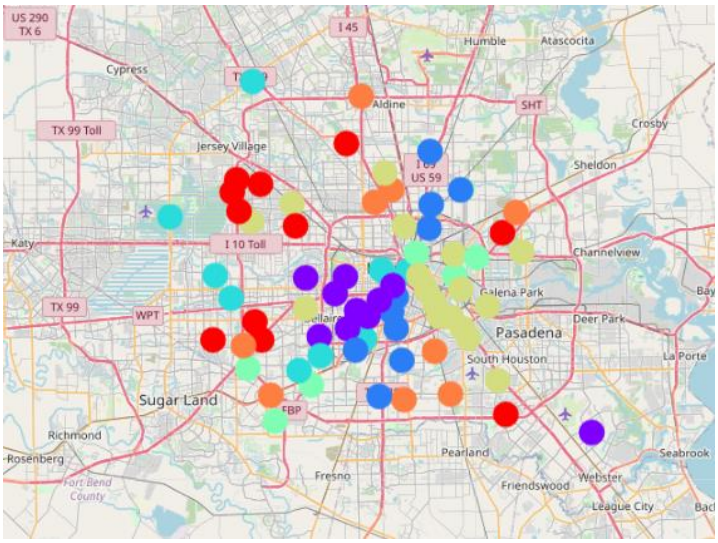
dfc3_c6

	Cluster Labels	HoustonNeighborhoods	Latitude	Longitude	Age%_Under5	Age%_5_17	Age%_18_64	Age%_65up
17	6	79_Edgebrook	29.640117	-95.246973	9	23	60	8
26	6	2_Greater_Greenspoint	29.940700	-95.413900	14	23	59	4
27	6	78_Greater_HobbyArea	29.626363	-95.305170	10	21	63	6
32	6	69_GulfgateRiverview_PineValley	29.708500	-95.318100	10	22	61	8
33	6	27_Gulfton	29.716500	-95.480600	9	23	64	4
36	6	54_Hunterwood	29.817604	-95.225253	13	24	61	3
39	6	11_Langwood	29.827531	-95.497872	7	24	64	4
47	6	77_Minnetex	29.620402	-95.361335	10	25	58	7
51	6	58_Northshore	29.775800	-95.218300	11	21	61	7
52	6	45_Northside_Northline	29.859500	-95.384700	9	23	60	8
53	6	74_Park_Place	29.694865	-95.289804	9	22	63	6
71	6	29_Westwood	29.677500	-95.557000	13	24	61	3

dfc3_c2

	Cluster Labels	HoustonNeighborhoods	Latitude	Longitude	Age%_Under5	Age%_5_17	Age%_18_64	Age%_65up
1	2	23_AftonOaks_RiverOaks	29.749994	-95.433234	5	14	61	20
15	2	47_EastLittleYork_Homestead	29.881945	-95.329744	6	18	59	18
28	2	68_OST_SouthUnion	29.695266	-95.371262	7	18	61	14
35	2	7_Hidden_Valley	29.890077	-95.431909	5	21	60	14
38	2	52_Kashmere_Gardens	29.800757	-95.331510	7	17	61	15
45	2	31_Meyerland	29.686600	-95.464700	6	17	61	16
55	2	57_Pleasantville	29.770800	-95.272900	8	20	54	18
57	2	50_Settegast	29.841600	-95.293100	7	20	61	12
59	2	76_Southacres_CrestmontPark	29.623530	-95.392020	6	18	62	15
64	2	84_SpringBranch_North	29.819482	-95.561616	6	18	62	14
65	2	71_Sunnyside	29.662000	-95.364000	9	17	59	15
66	2	48_Trinity_HoustonGardens	29.826362	-95.328715	7	19	59	15
69	2	37_Westbury	29.650600	-95.489700	8	19	60	14

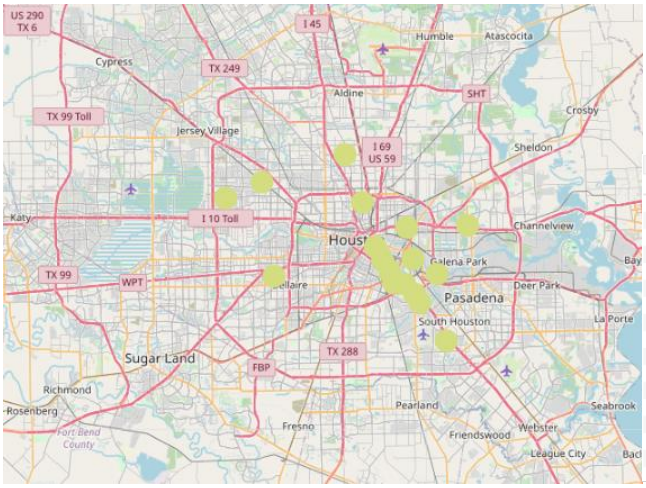
Figure 9: clustering by age of population: potential location for day-care center (top figure) and senior center (bottom figure).



dfc4_clustering_grouped

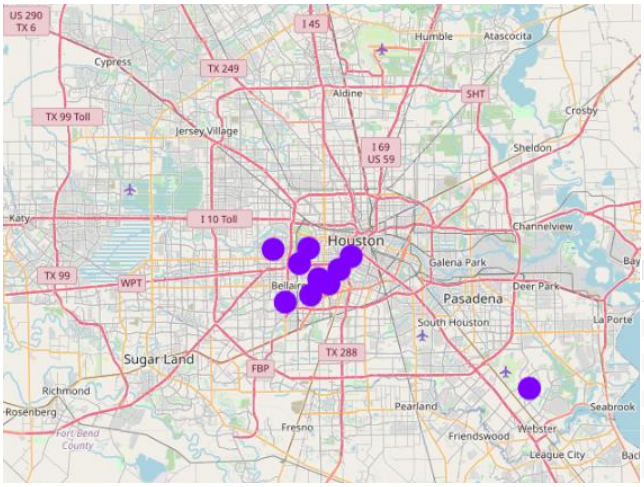
Cluster Labels	Non_Hispanic%_W	Non_Hispanic%_B	Hispanic%	Non_Hispanic%_A	Nin_Hispanic%_O
0	19.272727	14.636364	56.454545	8.545455	0.727273
1	62.300000	9.200000	13.800000	12.200000	2.500000
2	5.400000	73.000000	18.300000	2.200000	1.000000
3	36.888889	23.666667	25.555556	12.111111	1.888889
4	5.000000	50.571429	41.428571	2.000000	1.142857
5	8.235294	6.941176	82.529412	1.470588	0.588235
6	6.666667	27.444444	63.444444	2.000000	0.333333

Figure 10: clustering by Ethnicity.



dfc4_c5

Cluster Labels	HoustonNeighborhoods	Latitude	Longitude	Non_Hispanic%_W	Non_Hispanic%_B	Hispanic%	Non_Hispanic%_A	Nin_Hispanic%_O
12	5 56_DenverHarbor_PortHouston	29.774400	-95.301100	3	7	89	0	0
17	5 79_Edgebrook	29.640117	-95.246973	11	7	80	1	1
24	5 64_Greater_Eastwood	29.735700	-95.334000	15	5	78	2	0
32	5 69_GulfgateRiverview_PineValley	29.708500	-95.318100	5	6	88	1	0
33	5 27_Gulfton	29.716500	-95.480600	7	12	73	6	2
34	5 65_Harrisburg_Manchester	29.718500	-95.260200	3	14	82	0	0
39	5 11_Langwood	29.827531	-95.497872	13	9	76	0	2
40	5 88_Lawndale_Wayside	29.722815	-95.324774	9	3	87	1	1
42	5 82_Magnolia_Park	29.737200	-95.292500	3	1	95	0	0
43	5 75_Meadowbrook_Allendale	29.685181	-95.283277	10	3	85	2	0
49	5 51_Northside_Village	29.803300	-95.361000	8	10	81	0	1
51	5 58_Northshore	29.775800	-95.218300	12	11	76	1	0
52	5 45_Northside_Northline	29.859500	-95.384700	7	9	84	0	0
53	5 74_Park_Place	29.694865	-95.289604	4	4	87	4	1
54	5 70_Pecan_Park	29.707132	-95.303723	3	2	92	2	0
56	5 63_Second_Ward	29.751818	-95.343899	11	11	76	1	1
62	5 85_SpringBranch_Central	29.808792	-95.546308	16	4	74	4	1



dfc4_c1

Cluster Labels	HoustonNeighborhoods	Latitude	Longitude	Non_Hispanic%_W	Non_Hispanic%_B	Hispanic%	Non_Hispanic%_A	Nin_Hispanic%_O
1	1 23_AftonOaks_RiverOaks	29.749994	-95.433234	77	4	10	7	2
5	1 32_Braeswood_Place	29.695000	-95.430000	61	10	12	15	2
10	1 81_Clear_Lake	29.585700	-95.132800	54	8	20	14	4
30	1 21_Greater_Uptown	29.748729	-95.480928	67	5	16	10	2
31	1 87_Greenway_UpperKirby	29.731922	-95.444807	69	6	14	9	2
44	1 33_Medical_Center	29.706573	-95.405690	52	16	12	16	4
45	1 31_Meyerland	29.686600	-95.464700	58	11	16	13	2
46	1 62_Midtown	29.740800	-95.375600	63	16	13	6	1
48	1 66_MuseumPark	29.724721	-95.391874	54	10	16	17	3
67	1 28_University_Place	29.714344	-95.419076	68	6	9	15	3

Figure 11: clustering by Ethnicity: location for Hispanic community (top figure) and non-Hispanic Asian community (bottom figure).

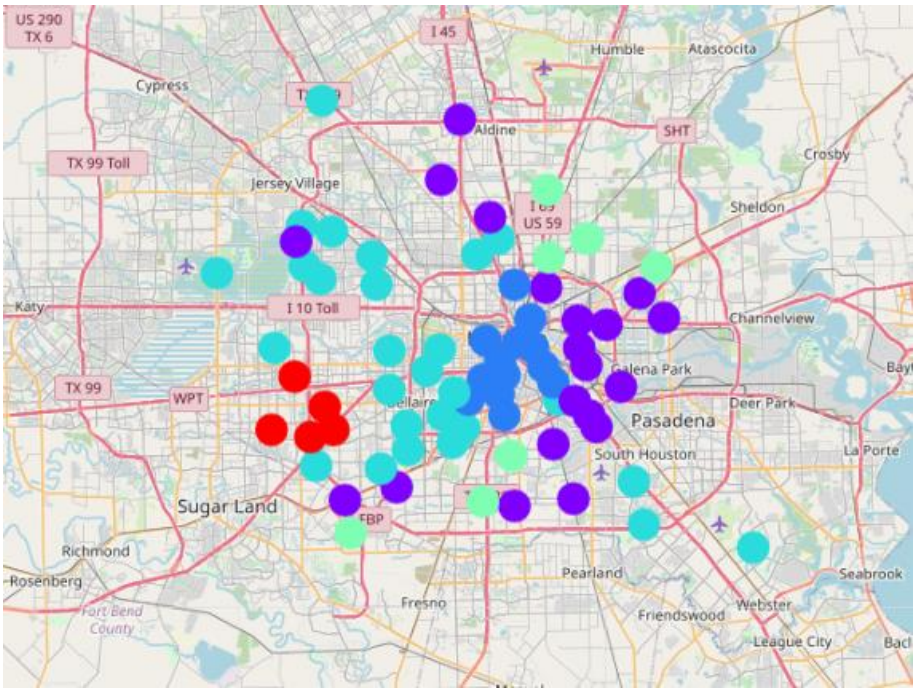
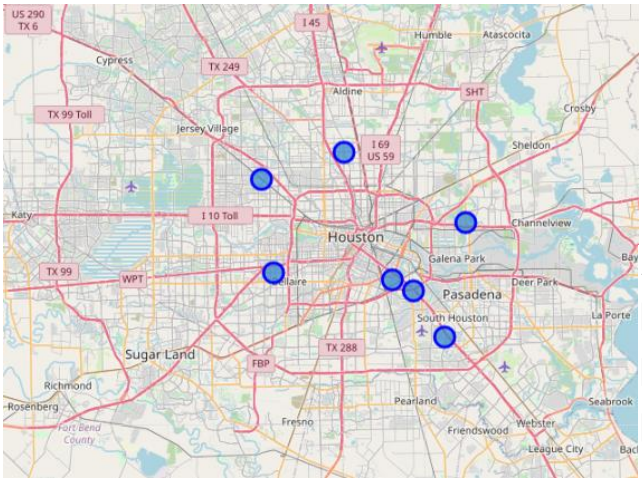
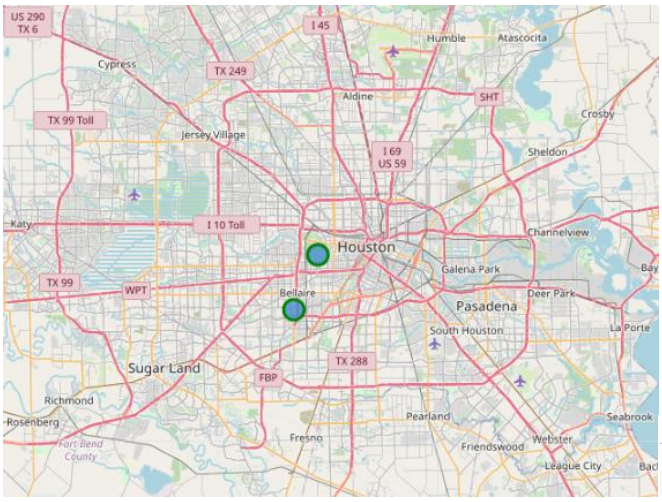


Figure 12: clustering by vendors from Foursquare searching results: Cluster 1 ●: Asian restaurant, Cluster 2 ●: Mexican restaurant, Cluster 3 ● : Coffee/Hotel, Cluster 4 ● : Mexican / ice-cream, Cluster 5 ● : Discount store.



	HoustonNeighborhoods	Latitude	Longitude	Age%_Under5	Age%_5_17	Hispanic%
0	79_Edgebrook	29.640117	-95.246973	9	23	80
1	69_GulfgateRiverview_PineValley	29.708500	-95.318100	10	22	88
2	27_Gulfton	29.716500	-95.480600	9	23	73
3	11_Langwood	29.827531	-95.497872	7	24	76
4	58_Northshore	29.775800	-95.218300	11	21	76
5	45_Northside_Northline	29.859500	-95.384700	9	23	84
6	74_Park_Place	29.694865	-95.289804	9	22	87

Figure 13: Potential location for a new day – care center for the Hispanic community.



	HoustonNeighborhoods	Latitude	Longitude	Age%_18_64	Age%_65up	Non_Hispanic%_A
0	23_AftonOaks_RiverOaks	29.749994	-95.433234	61	20	7
1	31_Meyerland	29.686600	-95.464700	61	16	13

Figure 14: Potential location for a new senior care center for the non-Hispanic Asian community.

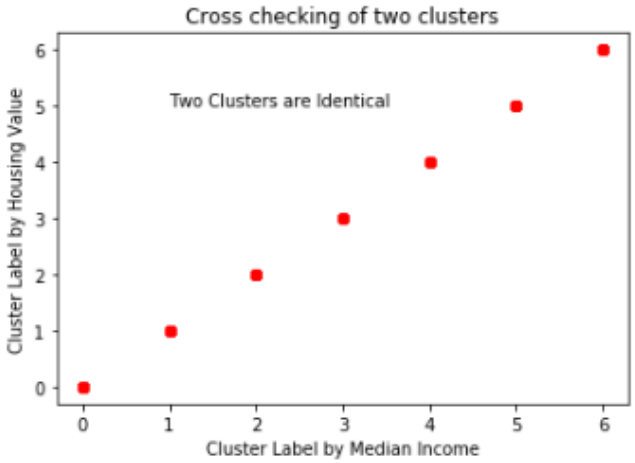


Figure 15: Identical clustering from “median income” and “housing value” when “median income” and “housing value” has a linear relationship (R^2 = 0.723).

References

[1]. https://en.wikipedia.org/wiki/Houston, accessed on June 2020.

[2]. Annual Estimates of the Resident Population for Incorporated Places of 50,000 or More, Ranked by July 1, 2019 Population: April 1, 2010 to July 1, 2019". United States Census Bureau, Population Division. Retrieved May 21, 2020.

[3]. https://www.houstonchronicle.com/news/houston-texas/houston/article/Houston-growth-shows-no-sign-of-waning-4540887.php, accessed on June 14, 2020.

[4]. Super neighborhood, https://www.houstontx.gov/superneighborhoods/recognized.html, accessed on June 14, 2020.