

Clustering of Houston City Neighborhood

for Coursera Capstone Project

Jason Chen

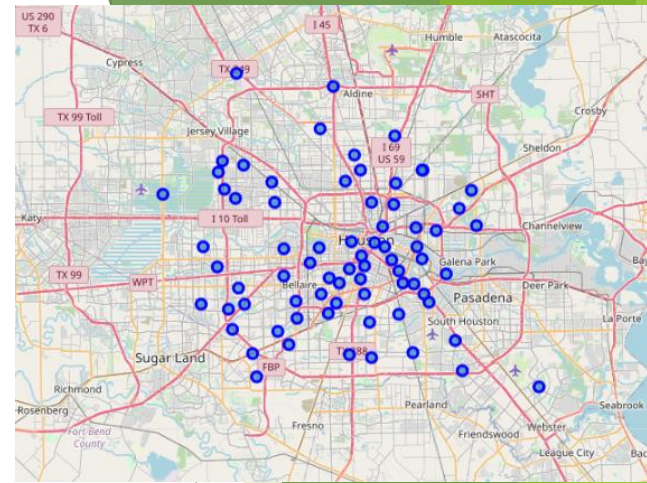
June 2020

Introduction

- ▶ Houston is the most populous city in TX with an estimated population in 2019 of 2.3 million.
- ▶ People move to Houston for its plentiful jobs, education, and favorable housing price.
- ▶ There are approximately 90 so called super neighborhoods in Houston city. 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

- Characteristic data for each neighborhood + Four square searching data for vendors



```
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	

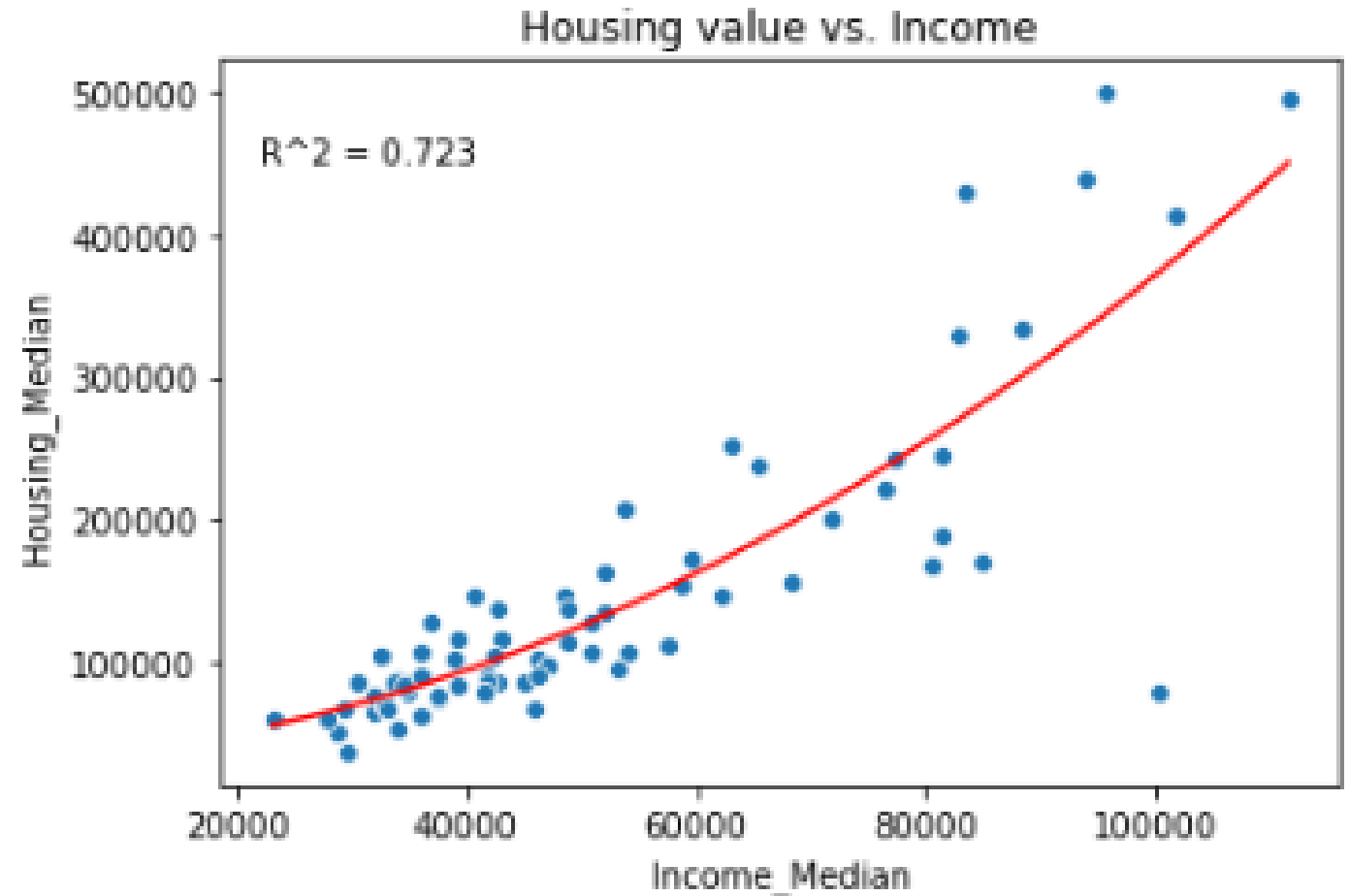
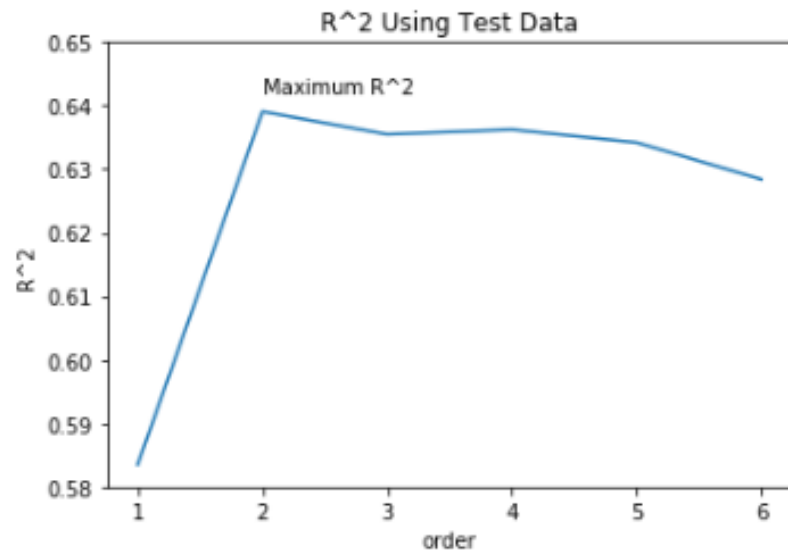
5 rows × 25 columns

Methodology

- ▶ Poly-normial regression
- ▶ K-mean clustering by:
 - ▶ Median household income
 - ▶ Median housing value
 - ▶ Age of population
 - ▶ Ethnicity
 - ▶ Vendor data from Foursquare searching

Results








- Poly-normal regression between housing value and household income

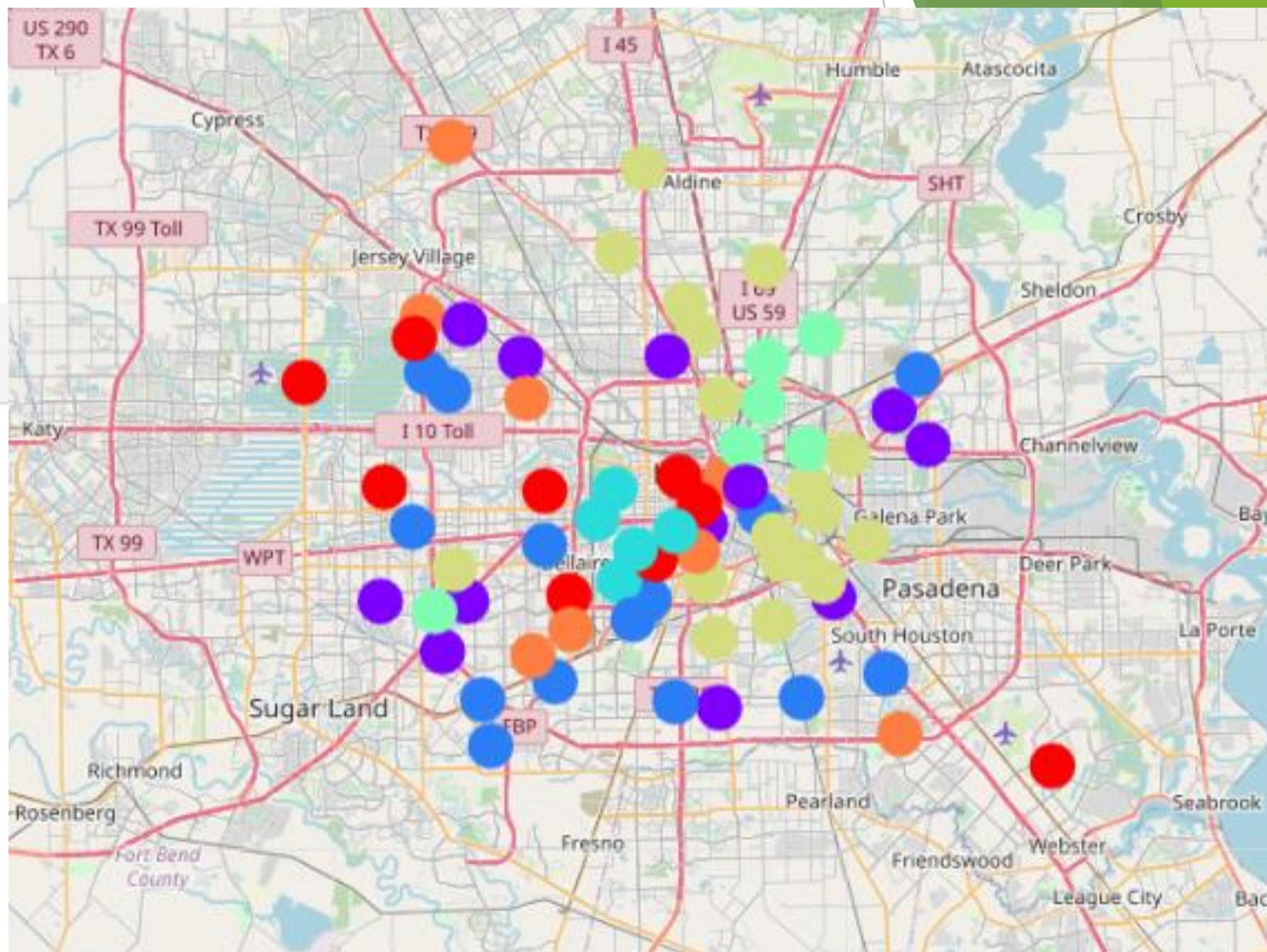


Results

► Clustering by median income

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

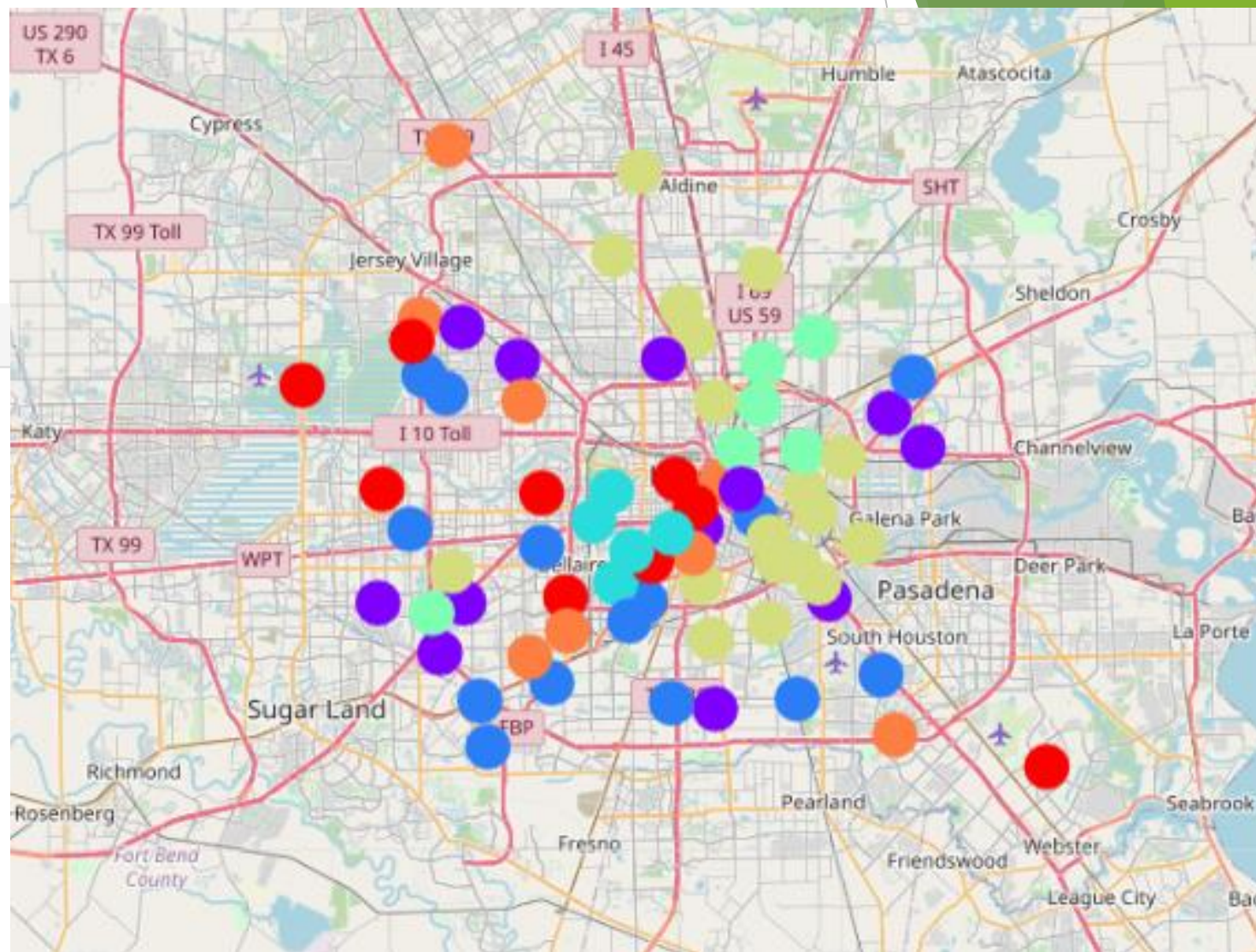


Results

► Clustering by housing value

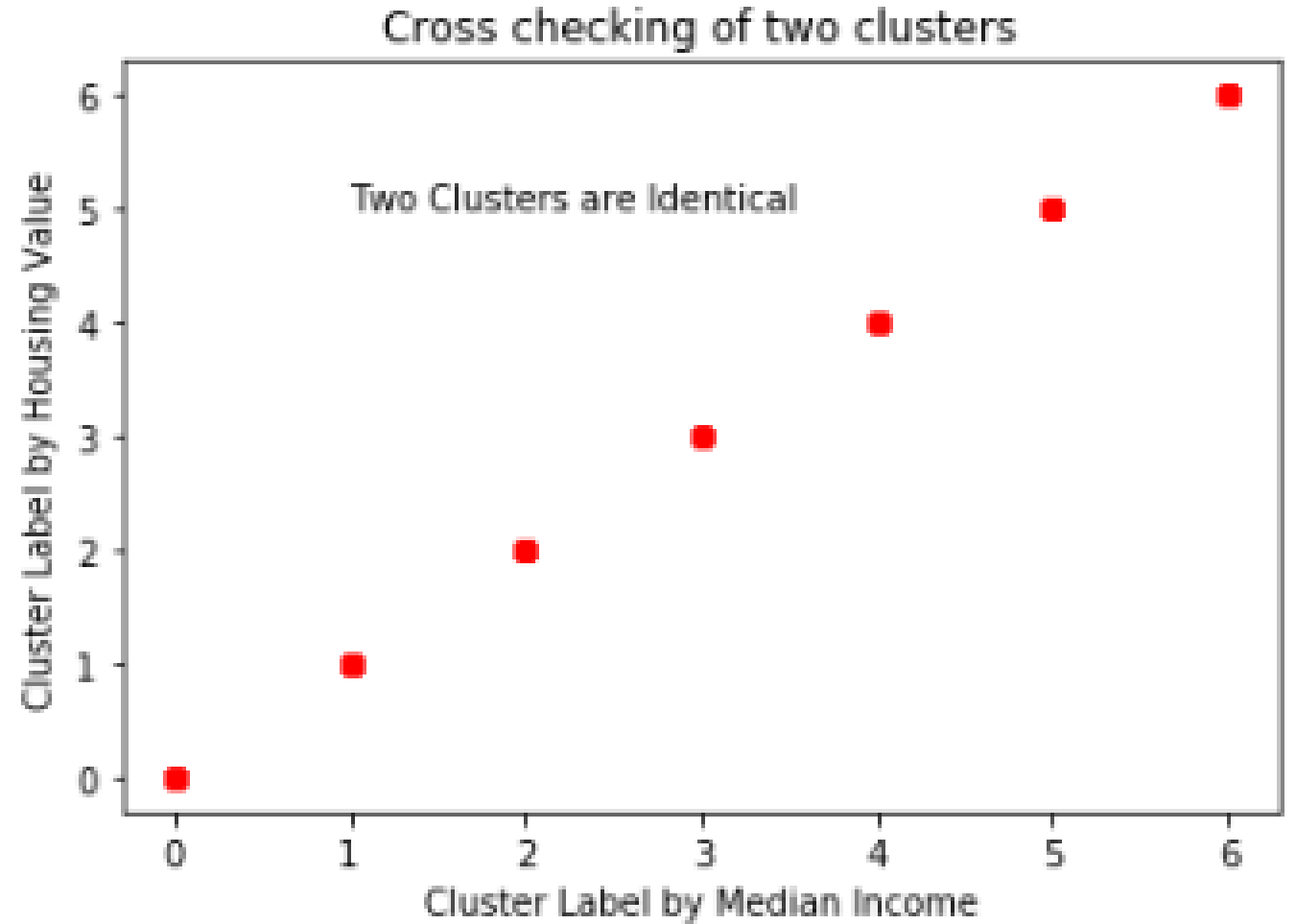
dfc2_clustering_grouped

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



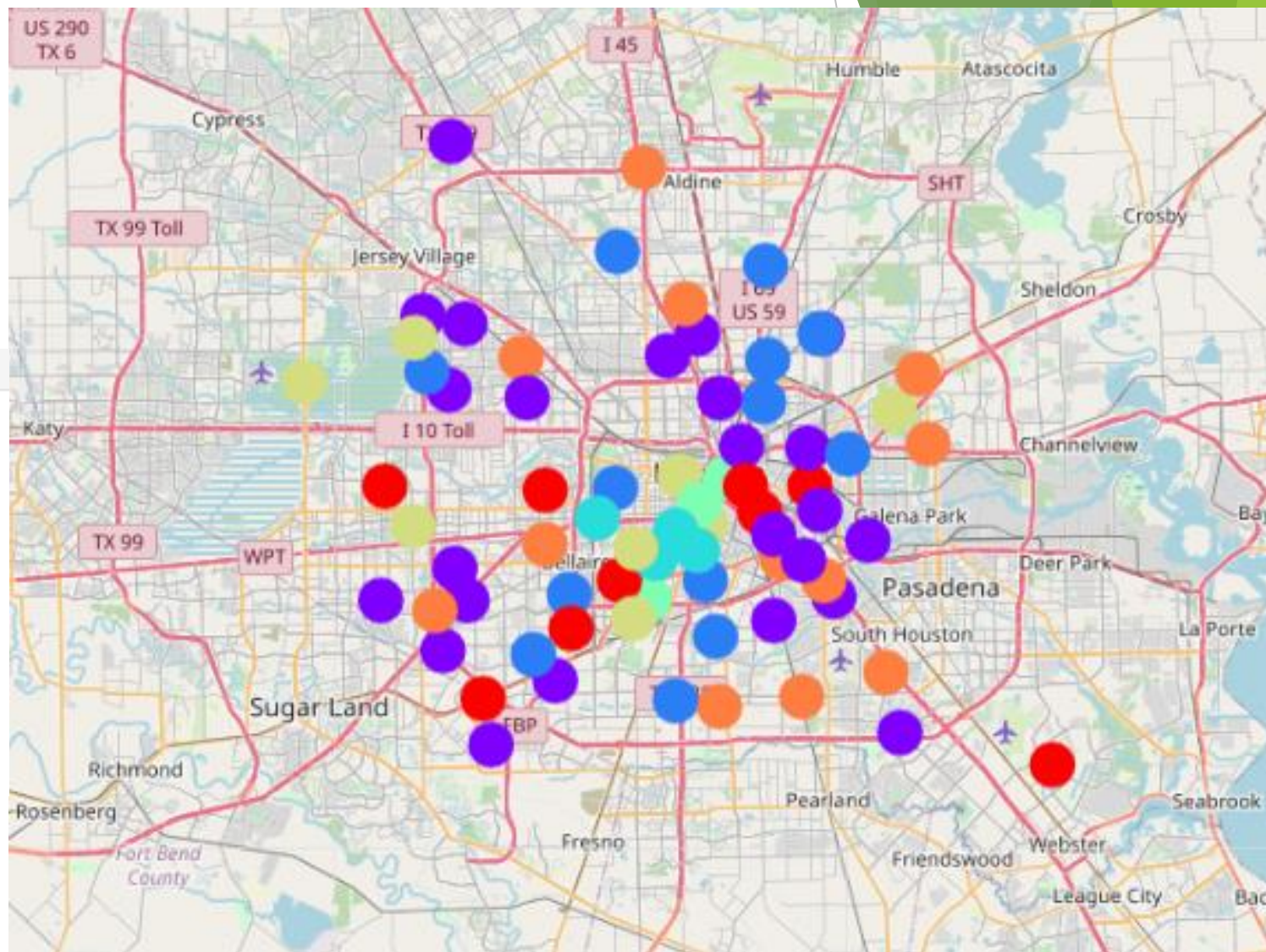
Results

- Identical clustering from “median income” and “housing value” when “median income” and “housing value” has a linear relationship ($R^2 = 0.723$)



► Clustering by age of population

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

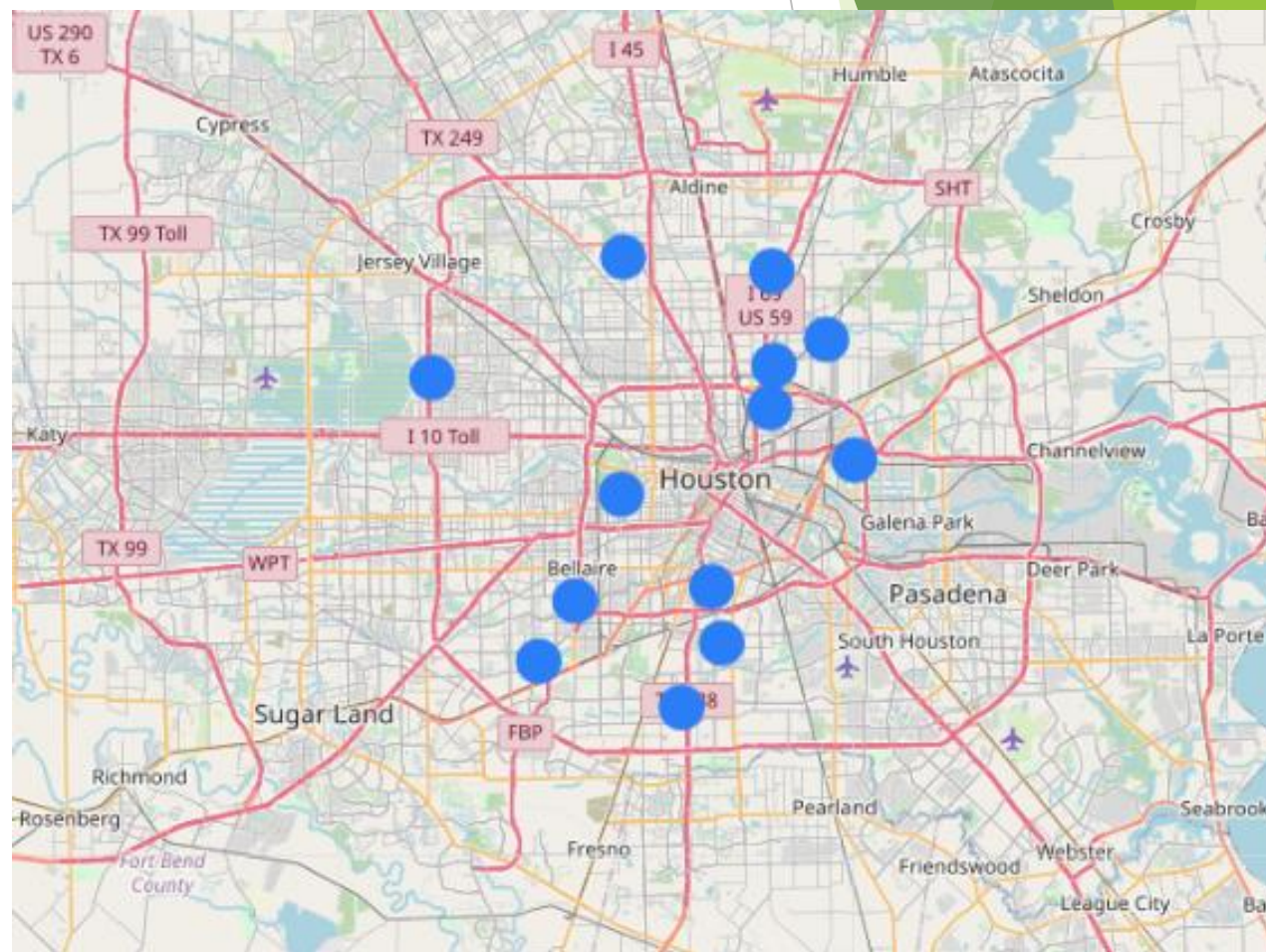


- ▶ Clustering by age of population
 - ▶ Potential location for daycare center








[illegible]

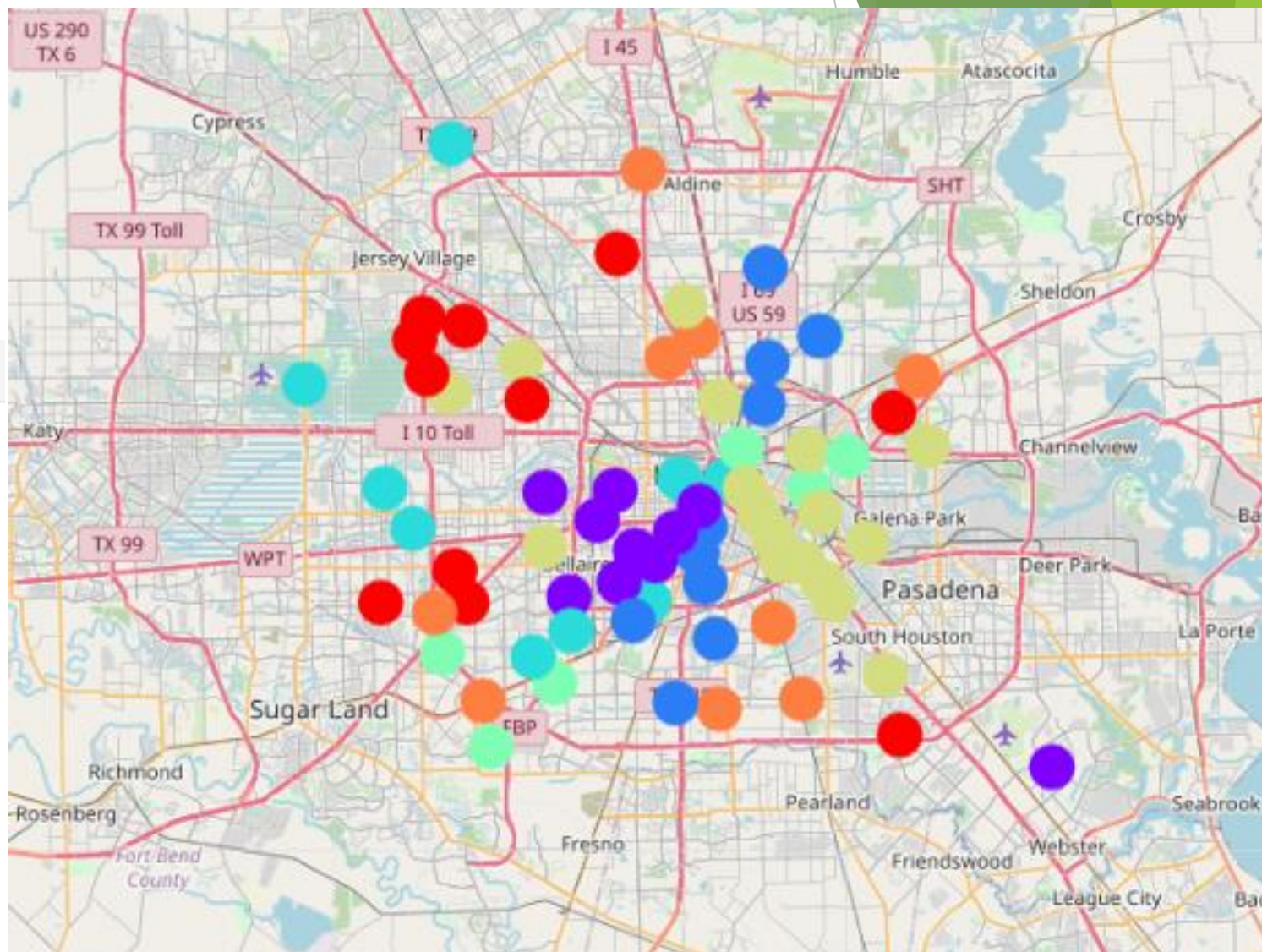
► Clustering by age of population

- | 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 |



► Clustering by Ethnicity

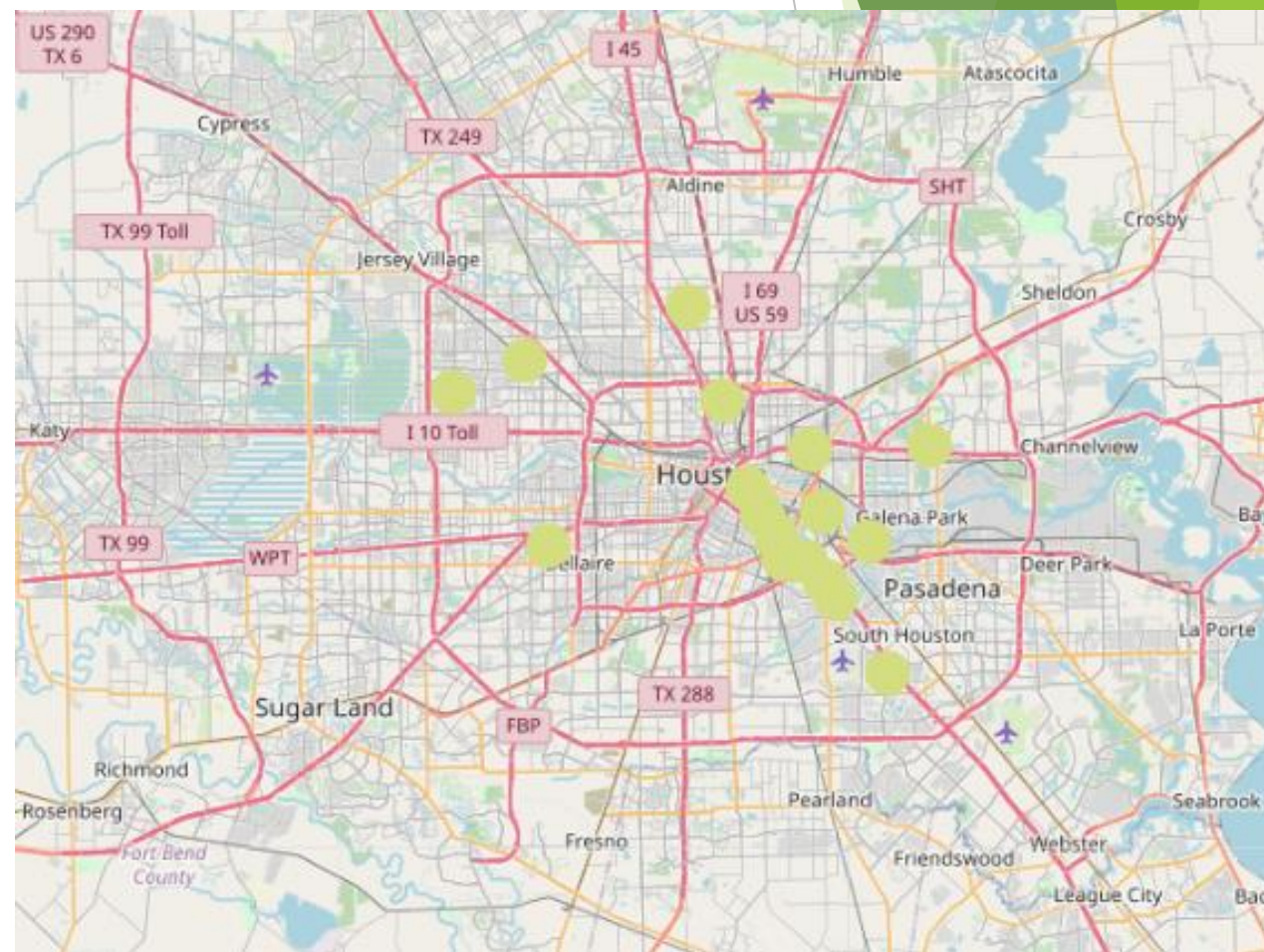
Cluster Labels	Non_Hispanic%_W	Non_Hispanic%_B	Hispanic%	Non_Hispanic%_A	Nin_Hispanic%_O
 0	19.272727	14.636364	56.454545	8.545455	0.727273
 1	62.300000	9.200000	13.800000	12.200000	2.500000
 2	5.400000	73.000000	18.300000	2.200000	1.000000
 3	36.888889	23.666667	25.555556	12.111111	1.888889
 4	5.000000	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



► Clustering by Ethnicity

- ▶ Clustering by Ethnicity
 - ▶ Hispanic community

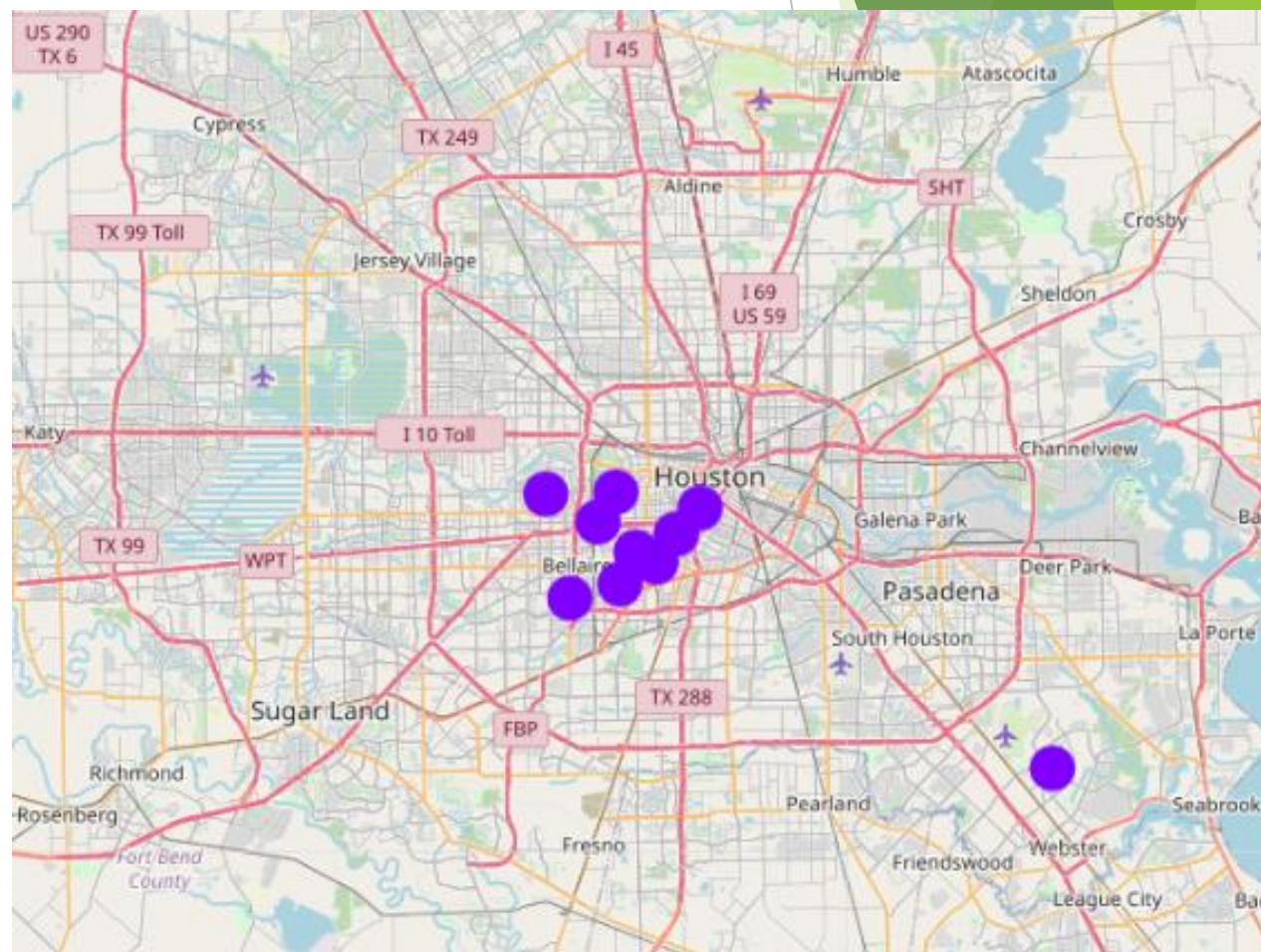
dfc4_c5									
Cluster Labels		HoustonNeighborhoods	Latitude	Longitude	Non_Hispanic%_W	Non_Hispanic%_B	Hispanic%	Non_Hispanic%_A	Nin_Hispanic%_O
12	5	56_DenverHarbor_PortHouston	29.774400	-95.301100	3	7	89	0	0
17	5	79_Edgebrook	29.640117	-95.246973	11	7	80	1	1
24	5	64_Greater_Eastwood	29.735700	-95.334000	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	-95.289804	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



► Clustering by Ethnicity

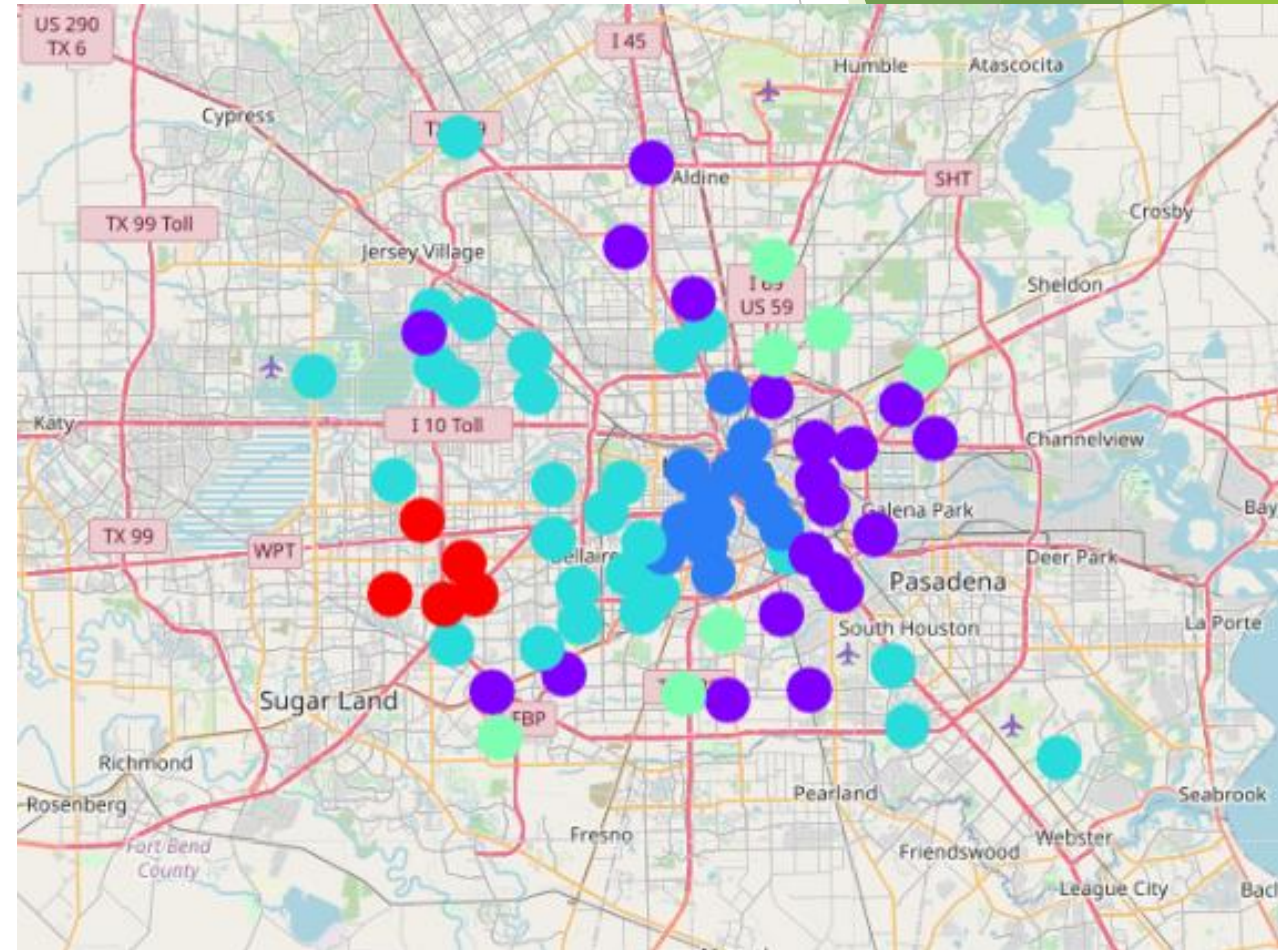
- Non-Hispanic Asian community

	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	29.708573	-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



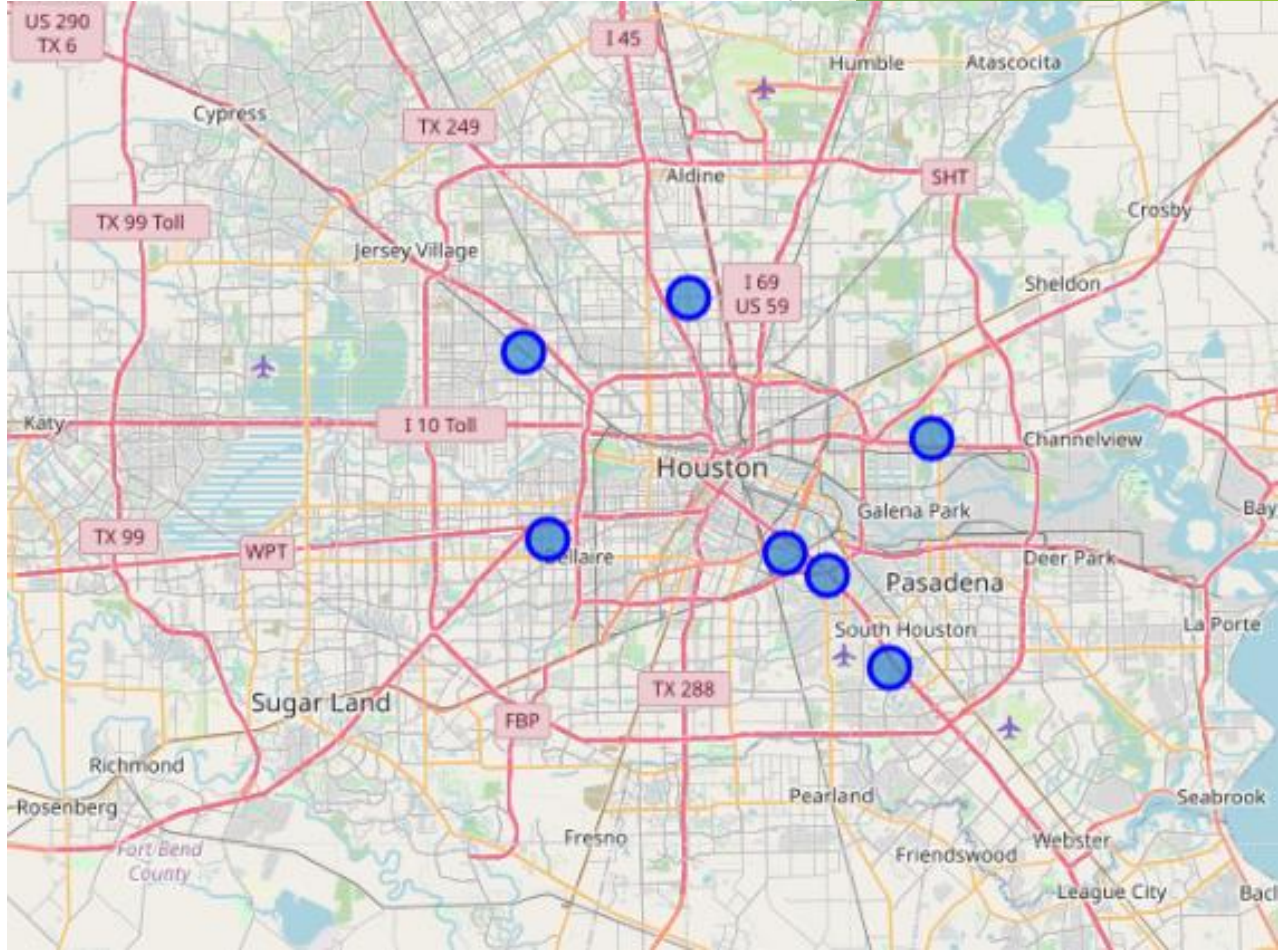
Results

- ▶ 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



► Potential location of a new day-care center for Hispanic community

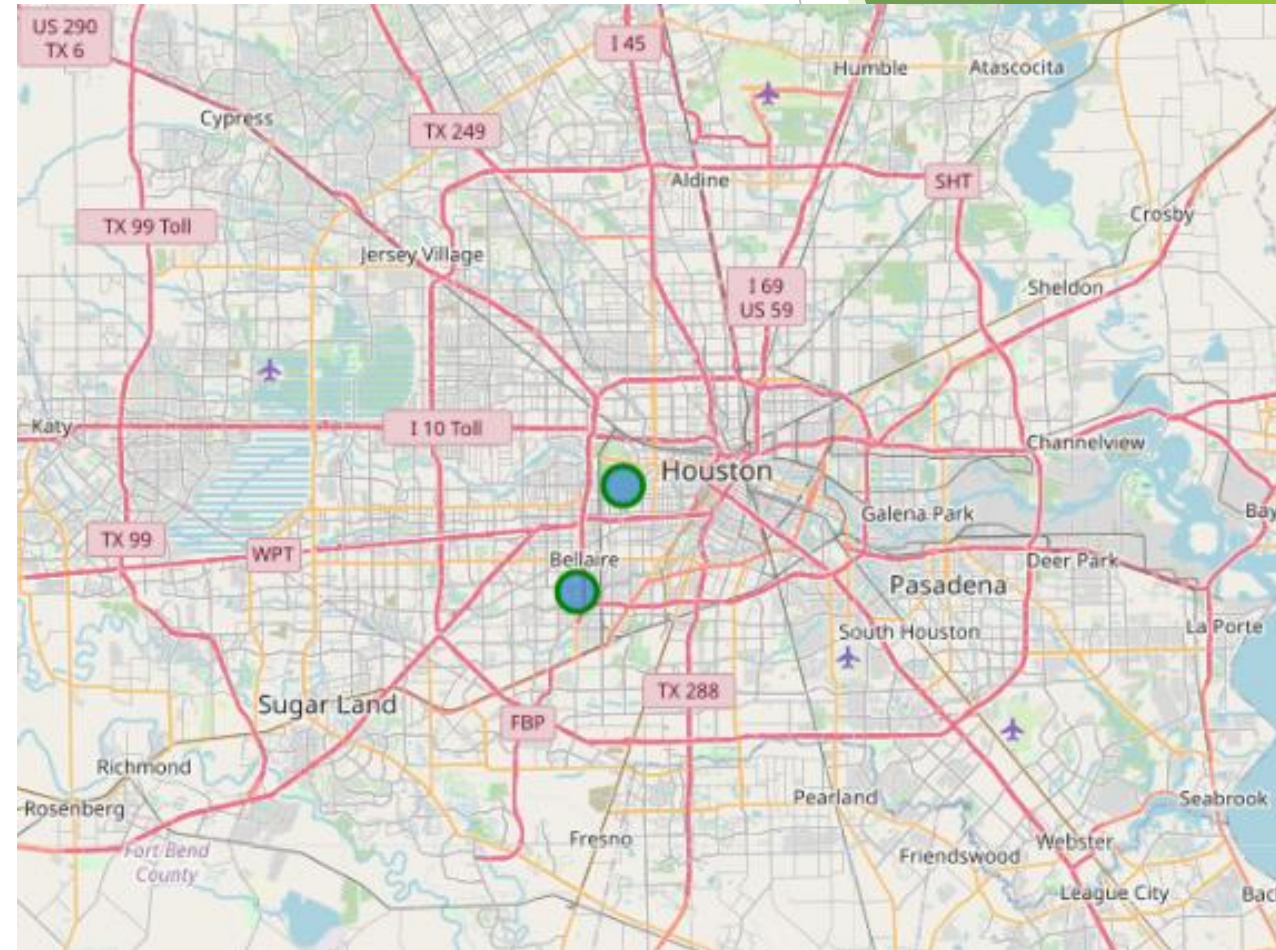
9	23	80
10	22	88
9	23	73
7	24	76
11	21	76
9	23	84
9	22	87



Application

- Potential location of a new senior center for Non-Hispanic Asian 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



Conclusion

- ▶ 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.
- ▶ Clustering method can be used to identify potential business locations.

Discussion

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