The Battle of the Neighborhoods - Week 2

Best Zip Code for Family in Houston

ABC Company Ltd.
19TH Jan 2020

Contents

The Battle of the Neighborhoods - Week 2 1

1. Introduction & Business Problem 2
   a. Problem Background: 3
   b. Problem Description: 3
   c. Target Audience: 3
   d. Success Criteria: 3
2. Data 3
3. Methodology 6
   e. Normalize the Data 7
   f. Cluster the Zip Codes 7
4. Results 8
5. Discussion 11
6. Conclusion 11

1. Introduction & Business Problem

a. Problem Background:

The city of Houston is 4th largest city in the United States and is energy capital of USA. It is multicultural and diverse. For employment opportunities and others, many families move to this city from across the world. When they move their first challenge is to find good neighborhood with good schools and low crime rate.

Most people goes by recommendation from individuals. This is good but having a way to find the best neighborhoods supported by data would enable them to make good choices.

b. Problem Description:

A family who is new to Houston and asks Relocation Consulting Firm- XYZ Company Ltd to find the best neighborhoods for them which have the best public schools for their school going kids and lower crime rate. The choice of neighborhood should also consider the home prices and the number of venues to do activities within that neighborhood. They also want the rational behind the recommendation.

c. Target Audience:

To recommend the correct neighborhood, XYZ Company Ltd has appointed me to lead of the Data Science team. The objective is to locate and recommend to the family which neighborhood of Houston city will be best choice. The Family also expects to understand the rationale of the recommendations made.

This would interest anyone who comes to XYZ Company Ltd with similar request for Houston City.

d. Success Criteria:

The success criteria of the project will be a good recommendation of Neighborhood/Zip Code choice for the family on behalf of XYZ Company Ltd based on data.

2. Data

We need below data for our Analysis -
1. All Zip Codes/Neighborhood in Houston
2. Home Price data for each Zip Code
3. Crime data in every Zip Code
4. Average School Rating in every Zip Code
4. Latitude and Longitude of each Zip Code
5. The Number of venues in each Zip Code from Foursquare APIs based on Latitude and Longitude of Zip Code

We can get All Zip Codes/Neighborhood in Houston with average home prices from website of Houstonia- a popular online magazine in Houston. This is available at - https://www.houstoniamag.com/articles/2017/3/24/neighborhoods-by-the-numbers-real-estate-data-2017

To get this we would need to scrape a page from website of Houstonia magazine to get the neighborhood data of Houston, TX using BeautifulSuop4 Library. This would return neighborhood name, Zip Code and average home price in that Zip Code.

```
page = urlopen('https://www.houstoniamag.com/articles/2017/3/24/neighborhoods-by-the-numbers-real-estate-data-2017').read()
soup = bs(page)
soup.prettify()
table = soup.find('table')
df = pd.read_html(str(table))
nbrs= df[0] # get the first table
nbrs= nbrs[nbrs.columns[0:3]]
nbrs= nbrs.rename(columns={"Unnamed: 0": "Neighborhood", "ZIP Code": "Zip", "2016 Median Home Price":"HomePrice"})
nbrs.head()
```

	Neighborhood	Zip	HomePrice
0	1960/Cypress	77065	$179,000
1	Aldine Area	77039	$133,500
2	Alief	77072	$164,000
3	Alvin North	77511	$227,000
4	Alvin South	77511	$163,900

Next, Crime data in every Zip Code can be read from the published data from Houston Police Department at their website - http://www.houstontx.gov/police/cs/xls/06-2019.NIBRS_Public_Data_Group_A&B.xlsx

```
crimedf = pd.read_excel('http://www.houstontx.gov/police/cs/xls/06-2019.NIBRS_Public_Data_Group_A&B.xlsx', header=11)
crimedf = crimedf[['ZIP', 'Offense Count']].groupby(['ZIP']).sum().reset_index()
crimedf = crimedf.rename(columns={ "ZIP": "Zip", "Offense Count":"Crimes"})
crimedf.head()
```

	Zip	Crimes
0	75248	1
1	77002	547
2	77003	183
3	77004	520
4	77005	91

School Accountability Ratings are available at https://opendata.arcgis.com/datasets/6cf4436417ff43d0a6e741dc83339ae2_0.csv

This dataset was made available on ArcGIS website by Texas Education Agency. This Dataset contains the school with its address and its Accountability Rating for whole Texas. We would filter this data set for Houston and get the average rating of all schools in each Zip Code.

```
schooldf  = pd.read_csv('https://opendata.arcgis.com/datasets/6cf4436417ff43d0a6e741dc83339ae2_0.csv', sep=',')
schooldf = schooldf[['School_Nam', 'School_Str','School_Cit', 'School_Sta','School_Zip', 'Acc_Rating']]
schooldf = schooldf[(schooldf['School_Cit'] == 'HOUSTON') & (schooldf['School_Sta'] == 'TX')]
schooldf[['Zip','ZipExtn']]= schooldf['School_Zip'].str.split("-",expand=True)
schooldf['Acc_Rating'] = schooldf[schooldf.columns[5]].replace('[\*,]', '', regex=True)
schooldf.dropna()
invalid_rating = ['Not Rated', 'NULL']
schooldf = schooldf[~schooldf.Acc_Rating.isin(invalid_rating)]
schooldf.head()
```

```
C:\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3057: DtypeWarning: Columns (67,102) have mixed types. Specify
dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

	School_Nam	School_Str	School_Cit	School_Sta	School_Zip	Acc_Rating	Zip	ZipExtn
10	NORTHSIDE H S	1101 QUITMAN	HOUSTON	TX	77009-7815	C	77009	7815
12	ANDERSON ACADEMY	7401 WHEATLEY ST	HOUSTON	TX	77088-7845	F	77088	7845
21	FRAZIER EL	8300 LITTLE RIVER RD	HOUSTON	TX	77064-7904	C	77064	7904
45	YOUNG SCHOLARS ACADEMY FOR EXCELLENCE	1809 LOUISIANA	HOUSTON	TX	77002-8013	D	77002	8013
57	KETELSEN EL	600 QUITMAN	HOUSTON	TX	77009-8113	A	77009	8113

Further. we get the latitude and longitude data of all Zip Codes in US from open data available at https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/download/?format=csv&timezone=America/Chicago&use_labels_for_header=true&csv_separator=%3B.
Again we would filter this dataset for Houston only.

```
zipdf  = pd.read_csv('https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/download/?format=csv&ti
zipdf = zipdf[['Zip', 'Latitude', 'Longitude']]
zipdf["Zip"]= zipdf["Zip"].apply(str)
zipdf.head()
```

	Zip	Latitude	Longitude
0	71937	34.398483	-94.39398
1	72044	35.624351	-92.16056
2	56171	43.660847	-94.74357
3	49430	43.010337	-85.89754
4	52585	41.194129	-91.98027

The Latitude and Longitude of each Zip Code would be used to get all venues in that Zip Code from FourSquare API. We would be interested in only total number of Venues in that zip code.

			Venues
Neighborhood	Latitude	Longitude	
Alief	29.700898	-95.59002	9
Braeswood Place	29.690230	-95.43474	4
Brays Oaks	29.654132	-95.54311	1
Briargrove	29.745129	-95.49131	6
Briargrove Park/Walnut Bend	29.741565	-95.55996	11
Briarmeadow/Tanglewilde	29.734379	-95.52269	21
Champions Area	29.984672	-95.52887	4
Charnwood/Briarbend	29.734379	-95.52269	21
Clear Lake Area	29.574930	-95.13238	3
Cottage Grove	29.772627	-95.40319	48

For analysis, we would join all above data in single dataset.

```
nbrs_merged = pd.merge(nbrs, zipdf, on='Zip', how='left')
nbrs_merged = pd.merge(nbrs_merged, crimedf, on='Zip', how='left')
nbrs_merged = pd.merge(nbrs_merged, schooldf, on='Zip', how='left')
nbrs_merged = pd.merge(nbrs_merged, houston_venues, on='Neighborhood', how='left')
nbrs_merged = nbrs_merged.dropna()
nbrs_merged
```

	Neighborhood	Zip	HomePrice	Latitude	Longitude	Crimes	Rating	Venues
2	Alief	77072	$164,000	29.700898	-95.59002	426.0	4.181818	9.0
13	Braeswood Place	77025	$715,000	29.690230	-95.43474	225.0	5.166667	4.0
14	Brays Oaks	77031	$225,000	29.654132	-95.54311	146.0	5.000000	1.0
15	Briargrove	77057	$824,000	29.745129	-95.49131	432.0	4.666667	6.0
16	Briargrove Park/Walnut Bend	77042	$460,000	29.741565	-95.55996	421.0	4.200000	11.0

The details please look into the Download & Explore Dataset and Explore Neighborhood Of Houston sections of the notebook.

3. Methodology

Our objective is to get the most suitable zip code in Houston area for a family with kids. For thus we would look into the home prices, crime data, school rating and venues of the zip codes and cluster

them using KMeans. Then we would examine each cluster and find the best cluster and then the zip codes.

e. Normalize the Data

In order to do this, first we would need to normalize the data using Min Max

```
nbrs_final = nbrs_merged.copy()
nbrs_final['HomePrice'] = nbrs_final[nbrs_final.columns[2]].replace('[\$,]', '', regex=True).astype(float)
cols_to_norm = ['HomePrice','Crimes','Rating', 'Venues']
nbrs_final[cols_to_norm] = nbrs_final[cols_to_norm].apply(lambda x: (x - x.min()) / (x.max() - x.min()))
nbrs_final.head()
```

	Neighborhood	Zip	HomePrice	Latitude	Longitude	Crimes	Rating	Venues
2	Alief	77072	0.026221	29.700898	-95.59002	0.816764	0.480519	0.106667
13	Braeswood Place	77025	0.224138	29.690230	-95.43474	0.424951	0.761905	0.053333
14	Brays Oaks	77031	0.048132	29.654132	-95.54311	0.270955	0.714286	0.013333
15	Briargrove	77057	0.263290	29.745129	-95.49131	0.828460	0.619048	0.040000
16	Briargrove Park/Walnut Bend	77042	0.132543	29.741565	-95.55996	0.807018	0.485714	0.146667

f. Cluster the Zip Codes

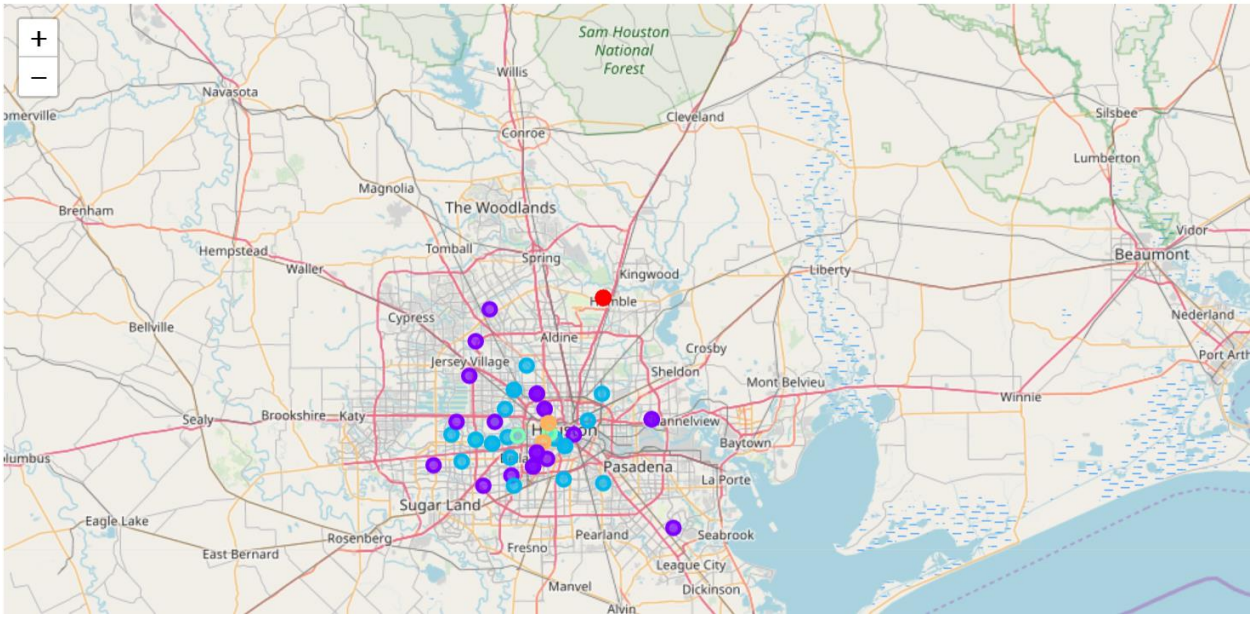
Run *k*-means to cluster the neighborhood into 5 clusters.

```
# set number of clusters
kclusters = 5

clustering_input = nbrs_final.drop(['Neighborhood','Zip', 'Latitude','Longitude'],  1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(clustering_input)
```

And plot the clusters using Folium



4. Results

Now we have all 5 clusters we will analyze them one by one to come up with few of the best Zip Codes.

i. *Cluster 1 – Marked by Red cluster*

	Neighborhood	Zip	HomePrice	Latitude	Longitude	Crimes	Rating	Venues
60	Humble Area East	77338	0.025144	30.005691	-95.28488	0.033138	0.0	0.426667
61	Humble Area South	77338	0.006214	30.005691	-95.28488	0.033138	0.0	0.426667
62	Humble Area West	77338	0.021552	30.005691	-95.28488	0.033138	0.0	0.426667
70	Kingwood South	77338	0.053520	30.005691	-95.28488	0.033138	0.0	0.426667

Neighborhoods in this cluster have low home prices and low crime rates. Public school ratings are also not available or very low. So, we are unable to recommend them

ii. *Cluster 2 – Marked by Sky Blue*

	Neighborhood	Zip	HomePrice	Latitude	Longitude	Crimes	Rating	Venues
13	Braeswood Place	77025	0.224138	29.690230	-95.43474	0.424951	0.761905	0.053333
14	Brays Oaks	77031	0.048132	29.654132	-95.54311	0.270955	0.714286	0.013333
21	Champions Area	77069	0.051435	29.984672	-95.52887	0.000000	0.714286	0.053333
23	Clear Lake Area	77062	0.051006	29.574930	-95.13238	0.097466	0.771429	0.013333
40	East End Revitalized	77003	0.058190	29.749278	-95.34741	0.343080	0.857143	0.080000
41	Eldridge North	77041	0.043642	29.858730	-95.57243	0.120858	0.785714	0.053333
49	Galleria	77056	0.209770	29.747328	-95.46931	0.637427	0.809524	0.186667
50	Garden Oaks	77015	0.127155	29.778526	-95.18118	0.294347	0.688312	0.013333
53	Heights/Greater Heights	77008	0.132507	29.798777	-95.40951	0.481481	0.714286	0.226667
72	Knollwood/Woodside Area	77025	0.121767	29.690230	-95.43474	0.424951	0.761905	0.053333
81	Medical Center Area	77030	0.084052	29.704584	-95.40466	0.249513	1.000000	0.146667
85	Memorial Villages	77024	0.458724	29.773994	-95.51771	0.368421	0.959184	0.000000
86	Memorial West	77079	0.209770	29.773018	-95.60125	0.360624	0.714286	0.293333
87	Meyerland Area	77096	0.106501	29.674336	-95.48123	0.539961	0.771429	0.160000
90	Mission Bend Area	77083	0.028736	29.691714	-95.64978	0.017544	0.587302	0.106667
94	North Channel	77015	0.016613	29.778526	-95.18118	0.294347	0.688312	0.013333
98	Oak Forest East Area	77018	0.106681	29.825476	-95.42619	0.346979	0.510204	0.173333
107	Rice/Museum District	77005	0.261566	29.717529	-95.42821	0.163743	1.000000	0.026667
114	Shepherd Park Plaza Area	77018	0.120690	29.825476	-95.42619	0.346979	0.510204	0.173333
133	Timbergrove/Lazybrook	77008	0.119971	29.798777	-95.40951	0.481481	0.714286	0.226667
142	West University/Southside Area	77005	0.395474	29.717529	-95.42821	0.163743	1.000000	0.026667
146	Willowbrook	77064	0.022450	29.923638	-95.55919	0.116959	0.673469	0.026667

When we examine cluster 2, we find that in this cluster, the home price is on lower side except Memorial Villages, Crime rate is lower except Sharpstown area and Number of Venues are also lower. Some of the neighborhoods seems interesting -

1. Medical Center Area
2. Memorial Villages
3. West University
4. Rice/Museum District

These neighborhoods have higher school ratings, lower crimes but have lower venues for other activities. Further Medical center area has lower home prices while others have higher home prices.

We would recommend Zip Code - 77030 and 77005 for the families who are okay with little less venues for activities in their own neighborhood.

iii. *Cluster 3 – Marked by Purple*

	Neighborhood	Zip	HomePrice	Latitude	Longitude	Crimes	Rating	Venues
2	Alief	77072	0.026221	29.700898	-95.59002	0.816764	0.480519	0.106667
15	Briargrove	77057	0.263290	29.745129	-95.49131	0.828460	0.619048	0.040000
16	Briargrove Park/Walnut Bend	77042	0.132543	29.741565	-95.55996	0.807018	0.485714	0.146667
17	Briarmeadow/Tanglewilde	77063	0.072198	29.734379	-95.52269	0.771930	0.673469	0.293333
22	Charnwood/Briarbend	77063	0.156789	29.734379	-95.52269	0.771930	0.673469	0.293333
37	Denver Harbor	77020	0.005029	29.775927	-95.31836	0.526316	0.532468	0.146667
42	Energy Corridor	77077	0.087644	29.750897	-95.61255	0.851852	0.510204	0.040000
52	Gulfton	77081	0.063398	29.708280	-95.48361	0.697856	0.693878	0.226667
57	Hobby Area	77061	0.012033	29.660280	-95.28446	0.703704	0.542857	0.120000
82	Medical Center South	77051	0.011135	29.665430	-95.36871	0.530214	0.142857	0.040000
89	Midtown-Houston	77004	0.109914	29.728779	-95.36570	1.000000	0.480519	0.066667
92	Montrose	77006	0.188218	29.741878	-95.38944	0.686160	0.428571	0.440000
95	Northeast Houston	77028	0.000000	29.827315	-95.28631	0.426901	0.357143	0.026667
96	Northside	77092	0.005029	29.833326	-95.47644	0.912281	0.357143	0.000000
97	Northwest Houston	77088	0.017241	29.879213	-95.45028	0.789474	0.357143	0.000000
99	Oak Forest West Area	77092	0.057112	29.833326	-95.47644	0.912281	0.357143	0.000000
110	Riverside	77004	0.078305	29.728779	-95.36570	1.000000	0.480519	0.066667
118	Spring Branch	77055	0.084052	29.798877	-95.49629	0.643275	0.514286	0.053333
143	Westchase Area	77042	0.184896	29.741565	-95.55996	0.807018	0.485714	0.146667
145	Willow Meadows Area	77035	0.078664	29.654108	-95.47692	0.508772	0.200000	0.146667

When we look into the cluster 5, we find that in this cluster the Home Price is on lower side, Crime rate is average to high. School Rating are average. Number of Venues in neighborhoods are also lower.

iv. *Cluster 4 – Marked by Green*

	Neighborhood	Zip	HomePrice	Latitude	Longitude	Crimes	Rating	Venues
83	Memorial Close In	77024	0.810776	29.773994	-95.51771	0.368421	0.959184	0.000000
108	River Oaks Area	77019	0.755388	29.752528	-95.39923	0.401559	0.714286	0.346667
109	Rivercrest	77042	1.000000	29.741565	-95.55996	0.807018	0.485714	0.146667
129	Tanglewood Area	77056	0.557651	29.747328	-95.46931	0.637427	0.809524	0.186667

Neighborhoods in this cluster have highest home prices. Public school ratings are not the best except 77024. Also, number of venues in the neighborhoods are not that many.

We would recommend Zip Code - 77024 for affluent families who are okay with little less venues for activities in their own neighborhood.

v. *Cluster 5 – Marked by Orange*

	Neighborhood	Zip	HomePrice	Latitude	Longitude	Crimes	Rating	Venues
30	Cottage Grove	77007	0.108297	29.772627	-95.40319	0.803119	1.000000	0.613333
84	Memorial Park	77007	0.373204	29.772627	-95.40319	0.803119	1.000000	0.613333
106	Rice Military/Washington Corridor	77007	0.128915	29.772627	-95.40319	0.803119	1.000000	0.613333
137	Upper Kirby	77098	0.258064	29.735529	-95.41405	0.360624	0.657143	1.000000
139	Washington East/Sabine	77007	0.114224	29.772627	-95.40319	0.803119	1.000000	0.613333

When we examine cluster 5, we find that in this cluster, the home prices are on relatively reasonable side, Crime rate is average. School ratings are high except Upper Kirby. These neighborhoods have many venues for different activities.

We would like to recommend all neighborhoods in Zip Code -77007

5. Discussion

Going by the results we see it is very difficult to recommend one single zip code where we get all 4 things together – Low Home Price, School Ratings, Lot of venues and Lower Crimes. It is generally observed that home prices would be higher in those area where schools are good, crimes are low and lot of venues to explore. But again, good schools and low crimes are very important so we will not compromise with these two attributes. Home prices as optional choices for family depending on the affluency.

Considering all of this, we will finally recommend 4 zip codes. Families can choose based on their priorities –

1. **77007** - We would like to recommend all neighborhoods in Zip Code. This zip code has all favorable attributes except Memorial Park are in this zip code may have costly homes.
2. **77024** - We would recommend this zip code for affluent families who are okay with little less venues for activities in their own neighborhood.
3. **77030 and 77005** - We would recommend zip code for the families who are okay with little less venues for activities in their own neighborhood.

6. Conclusion

This analysis has been done with limited set of data and conserving the data which is available is correct. There are many other things which impact suitability of a neighborhood for a family like distance from office etc. Also, affluent facilities put their kids in private schools also. So before recommending a

neighborhood for a family this individual, preferences should also be considered. But certainly, the above analysis would form a base for recommendation.