Prospective Home Improvement Store Sites in Houston

Applied Data Science Capstone Project

"The Battle of the Neighborhoods"

Christopher Don Cothran

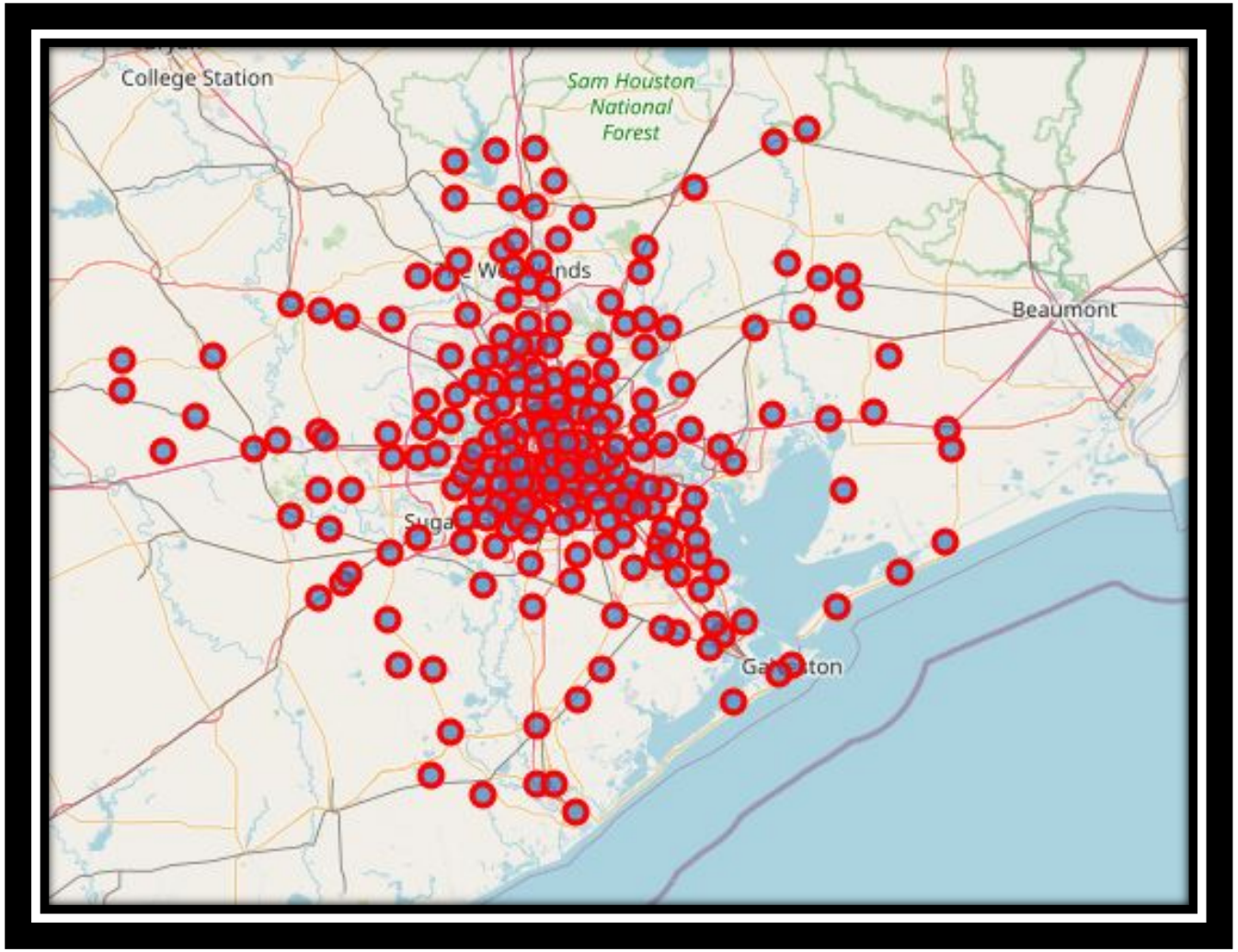
August 2020

High-Grading Prospective Zones in Houston

- Evaluation of potential sites for a new store involves multiple spectrums of consideration.
- Expected performance of a potential store could be highly reflective of the optimality of the location, distance from competitors, and proximity to demand centers.
- Stakeholders would look to analysis such as this for a variety of supportive factors informing decision intelligence processes.
- The high-graded deliverable can be effectively iterated with variations in radius and starting locations, as well as type of business. For this study, we examine Home Improvement retailers.
- Ideally, the stakeholders would like to identify potential underserved niches.

Greater Houston Metropolitan Area

- The ZIP code center lines utilized for this analysis are depicted in the Folium map at right.
- This is represented for the purposes of this analysis as the ‘Greater Houston Metropolitan Area’. It includes areas as far north as Conroe, TX and as far south as Freeport, TX, as well as substantially outside the city limits of Houston, TX.
- Many of these Zip codes are not prospective site zones, given one or more of the criteria explained later, but were included from the outset to be as comprehensive as is feasible given the scope of the analysis.



National Association of Realtor's Data

1) National Association of Realtor's data provided from Realtor.com supplied the number of housing listings by zip code.

2) This information was integrated into a pandas dataframe, with an 'investment intensity' derived for a given zip code (YTD 2020 basis) by multiplying the median listing price and active listing count.

3) This data point provides a relative snapshot of the amount of total dollars associated with new and existing home sales.

4) From the standpoint of new (new construction) and existing (upgrades), this information is useful for deriving an area-based indication of home improvement and construction support product demand.

	postal_code	month_date_yyyymm	median_listing_price	active_listing_count	investment intensity
0	77423	202007	307550.0	89	27371950.0
1	77021	202007	292050.0	95	27744750.0
2	77532	202007	241495.0	132	31877340.0
3	77489	202007	189050.0	29	5482450.0
4	77089	202007	298750.0	88	26290000.0

Internal Revenue Service Data

1) The U.S. Internal Revenue Service Census data provides filing data by zip code.
2) This data is used to support the determination of potential prospective zones based on the density of the number of filed income tax returns by income level for each zip code in the analysis coverage area.
3) After converting the Income Levels to a numeric Income Level Score, 'Income Intensity' was calculated based on the relative proportion of the Number of Returns per income level.

	ZIP	Income Level Score	Number of returns	income intensity
0	75001	1	2370	2370
1	75001	2	2440	4880
2	75001	3	1680	5040
3	75001	4	940	3760
4	75001	5	1320	6600

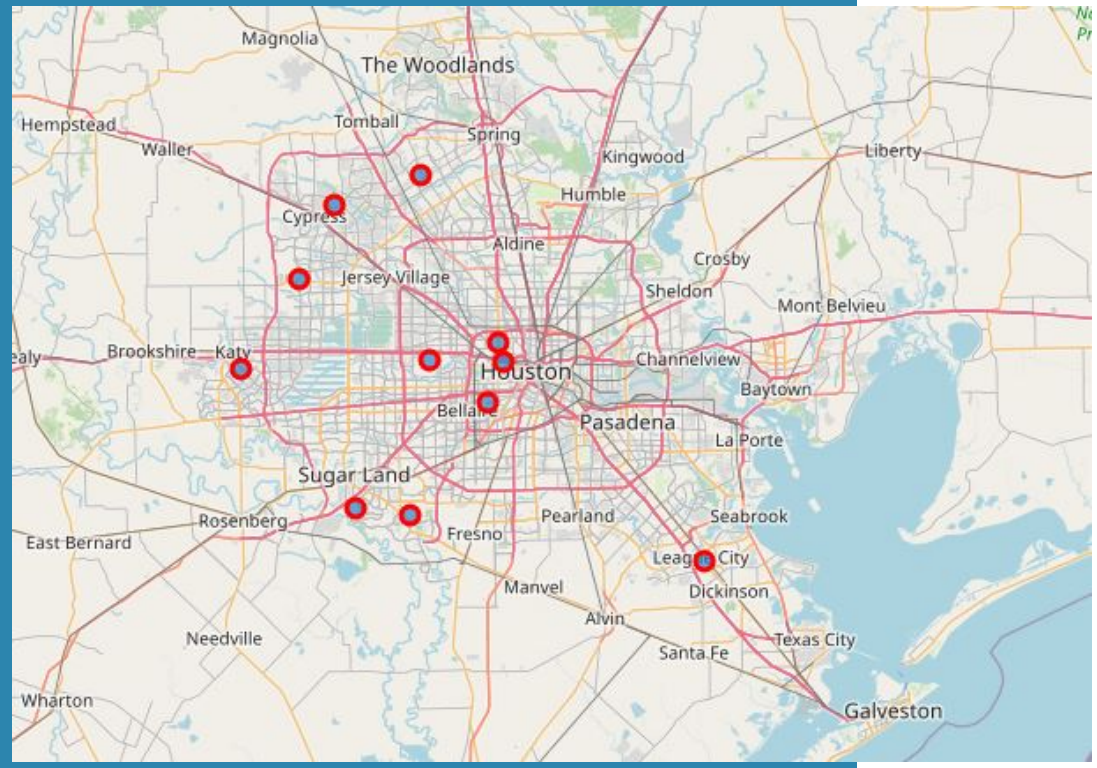
Preliminary Prospectiveness Factors

- Though not the final determination, the derived Prospectiveness Score, a factor of the investment intensity and income intensity values, provides an indication of key zones of investment.
- It is important to note the growth of Katy and Sugar Land in this estimation. Both zones are known for both rapid growth, above average wealth concentration, and maintain sufficient population for adequate foot traffic.
- From the outset, these zones are thus the most prospective.

	ZIP	City	Latitude	Longitude	investment intensity	income intensity	Prospectivity Score
0	77494	Katy	29.760833	-95.81104	1464.140459	168710	6.634928
1	77024	Houston	29.773994	-95.51771	3104.037550	74780	6.203177
2	77479	Sugar Land	29.573345	-95.63213	1283.715325	143880	4.828698
3	77433	Cypress	29.884175	-95.72219	1089.357898	126820	3.479447
4	77007	Houston	29.772627	-95.40319	1414.952830	88290	3.096075

*Refer to the associated report of this study for the complete list of locations ranked by Prospectiveness score and other factors

Mapping the Preliminary Top Ten

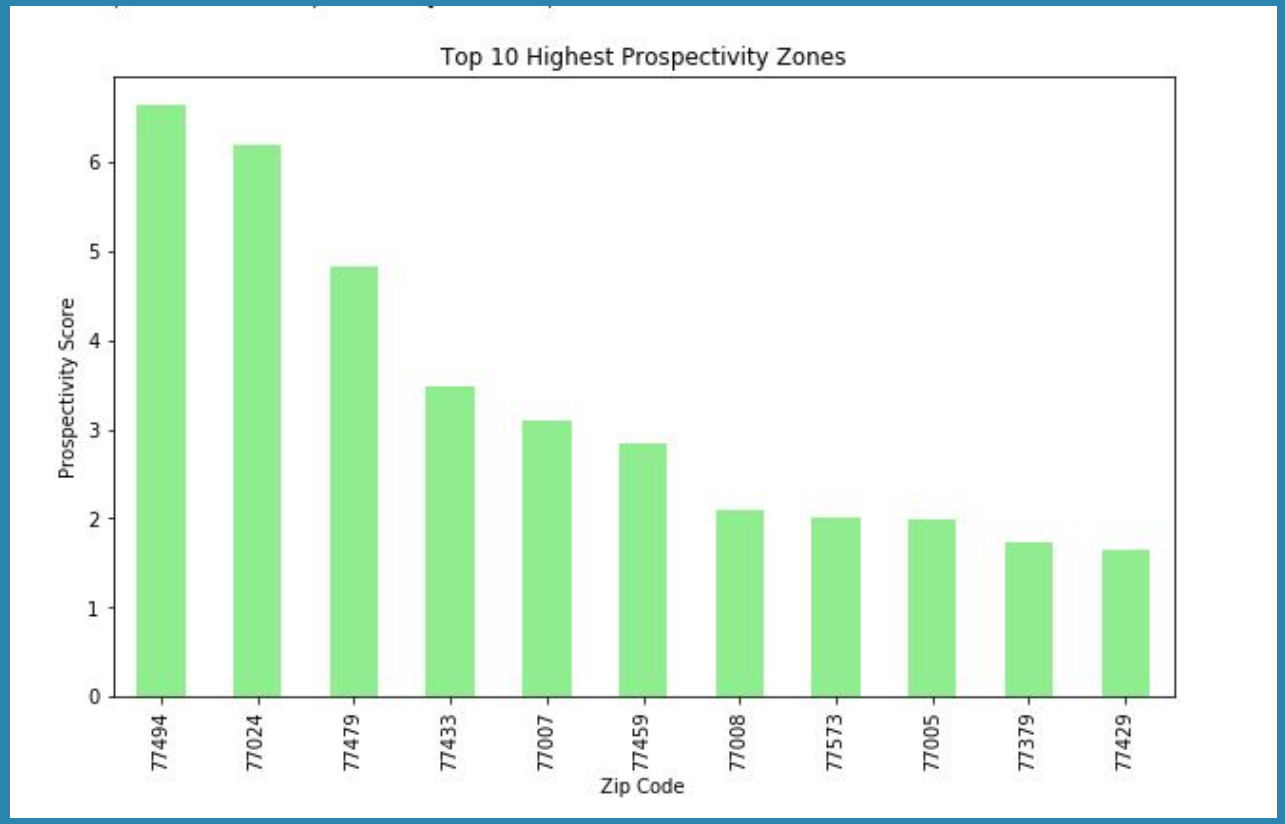


- This map depicts the top ten most prospective zones based on the income intensity and investment intensity (ranked by the composite Prospectiveness Score).
- We note the distribution west of central Houston, with noted clusters in the Uptown and Southwest areas.

Top Ten List

*Ranked selection based on the prospectiveness of optimal locations as a function of housing sales, gross income, population, and median housing prices in a radius of 150KM from the center of the City of Houston.

ZIP	City	Latitude	Longitude	investment intensity	income intensity	Prospectivity Score
77494	Katy	29.760833	-95.81104	1464.140459	168710	6.634928
77024	Houston	29.773994	-95.51771	3104.037650	74780	6.203177
77479	Sugar Land	29.573345	-95.63213	1283.715325	143880	4.828698
77433	Cypress	29.884175	-95.72219	1089.357898	126820	3.479447
77007	Houston	29.772627	-95.40319	1414.952830	88290	3.096075
77459	Missouri City	29.564347	-95.54762	1053.372022	110360	2.844592
77008	Houston	29.798777	-95.40951	1298.130485	69340	2.084068
77573	League City	29.502759	-95.08906	662.672050	131500	2.000849
77005	Houston	29.717529	-95.42821	1644.197700	52810	1.991826
77379	Spring	30.024749	-95.53215	670.548920	115730	1.724360
77429	Cypress	29.982746	-95.66597	604.343770	123870	1.644865



Prospective Zone Concentration

Prospectiveness independent of distance to competition is an important consideration to disaggregate multiple factors. Examining each independently is useful for noting the difference in ranked lists before and after additional factors are considered.

The Haversine Formula-Computing Distances Across a Sphere

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

- For this project, we developed a function to get the distance from two points using the Haversine mathematical formula depicted above.
- The function was used to derive the distances between each of the zip code center points and the list of existing regional competitors.
- The process was called over 19,200 times to complete the main dataframe used in this analysis.

```
import math
from math import *
#We will need to define a function to calculate distance using the Haversine methematical formula


def get_dist(lata, lona, latb, lonb):
    # approximate radius of earth in km
    R = 6378.0

    lat1 = radians(abs(lata))
    lon1 = radians(abs(lona))
    lat2 = radians(abs(latb))
    lon2 = radians(abs(lonb))

    dlon = lon2 - lon1
    dlat = lat2 - lat1

    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))

    output = R * c
```

Leveraging Foursquare Data

- This analysis leverages the Foursquare Developer login access credentials and API calls to obtain the locations of pertinent venues within a given radius.
- This was accomplished by the following process:
  - Sending a GET request via API call to the Foursquare site with a customized URL referencing the venues subdirectory and the Places API.
  - The JSON results were analyzed via looping into a customized Python dictionary. This dictionary was then converted into a pandas dataframe.

K-Means Clustering- Machine Learning Algorithm

- This analysis ultimately depicts the following key Data Science concepts:
  - Web Scraping for data tables
  - Dataset import and management
  - Dataset wrangling through automated pipelines
  - Descriptive Statistical Analysis
  - Data Visualization
  - RESTful API Calls to the Foursquare API
  - Machine Learning- We utilize the popular and versatile K-Means Clustering Algorithm for unsupervised learning.

- We implemented the K-Means algorithm as depicted at right with 4 clusters. The cluster results follow.

```
from sklearn.cluster import KMeans

number_of_clusters = 4

good_xys = dist_import_df[['Latitude', 'Longitude']].values
kmeans = KMeans(n_clusters=number_of_clusters, random_state=0).fit(good_xys)

cluster_centers = [xy_to_lonlat(cc[0], cc[1]) for cc in kmeans.cluster_centers_]
```

```
labels=kmeans.labels_
```

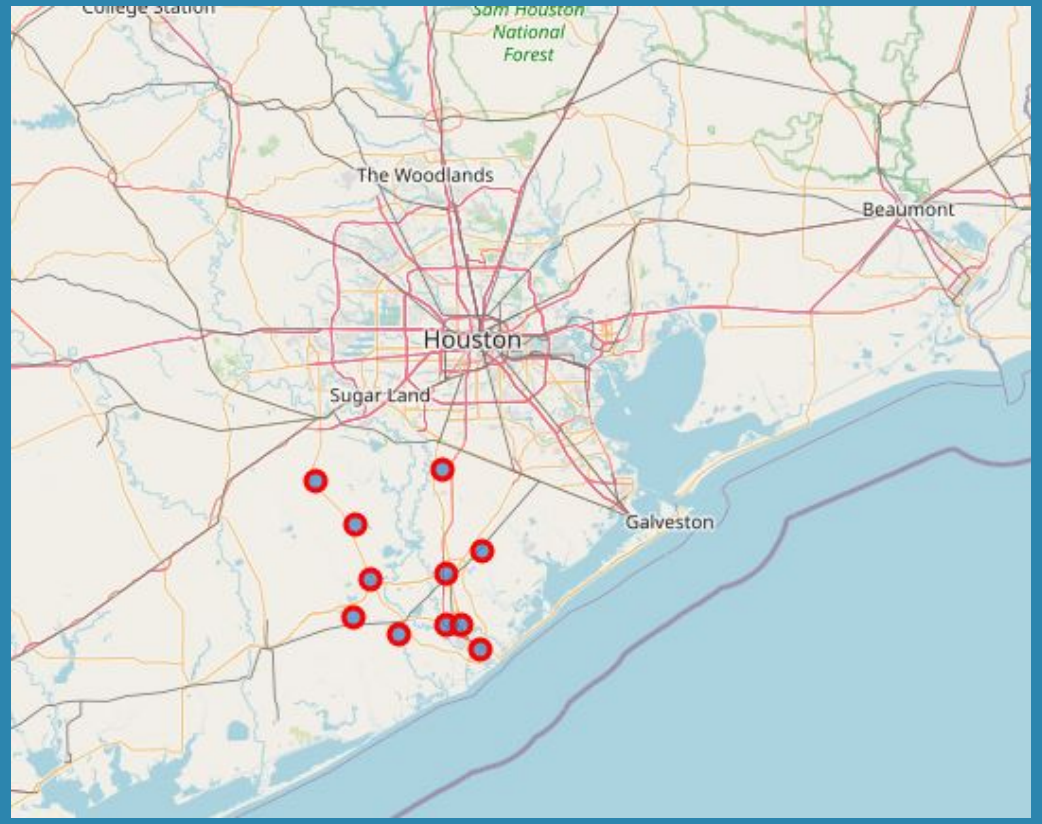
```
labels
```

```
array([2, 2, 2, 2, 2, 2, 3, 2, 1, 2, 1, 3, 2, 2, 2, 2, 1, 1, 2, 2, 2, 2,
       1, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 3, 2, 1, 1, 1, 1, 0, 1, 2, 2, 2,
       2, 1, 3, 3, 2, 2, 2, 2, 2, 1, 3, 3, 2, 2, 1, 2, 1, 1, 1, 3, 1, 3,
       3, 1, 3, 3, 2, 2, 2, 1, 3, 2, 0, 3, 2, 2, 3, 3, 3, 2, 1, 1, 0, 1,
       3, 2, 2, 2, 1, 3, 2, 3, 2, 1, 0, 2, 2, 2, 2, 3, 1, 1, 2, 2, 2, 3,
       1, 3, 2, 1, 3, 2, 2, 3, 3, 3, 3, 2, 1, 3, 2, 1, 2, 2, 2, 1, 3, 0,
       2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 3, 3, 3, 1, 1, 2, 1, 0, 3, 3, 3, 1,
       2, 2, 3, 3, 1, 2, 3, 3, 3, 0, 3, 1, 2, 3, 2, 3, 2, 0, 3, 3, 3, 3,
       3, 0, 2, 1, 3, 1, 2, 1, 3, 3, 3, 3, 1, 2, 3, 2, 3, 0, 2, 1, 3, 2,
       0], dtype=int32)
```

```
dist_import_df['Labels']=labels
```

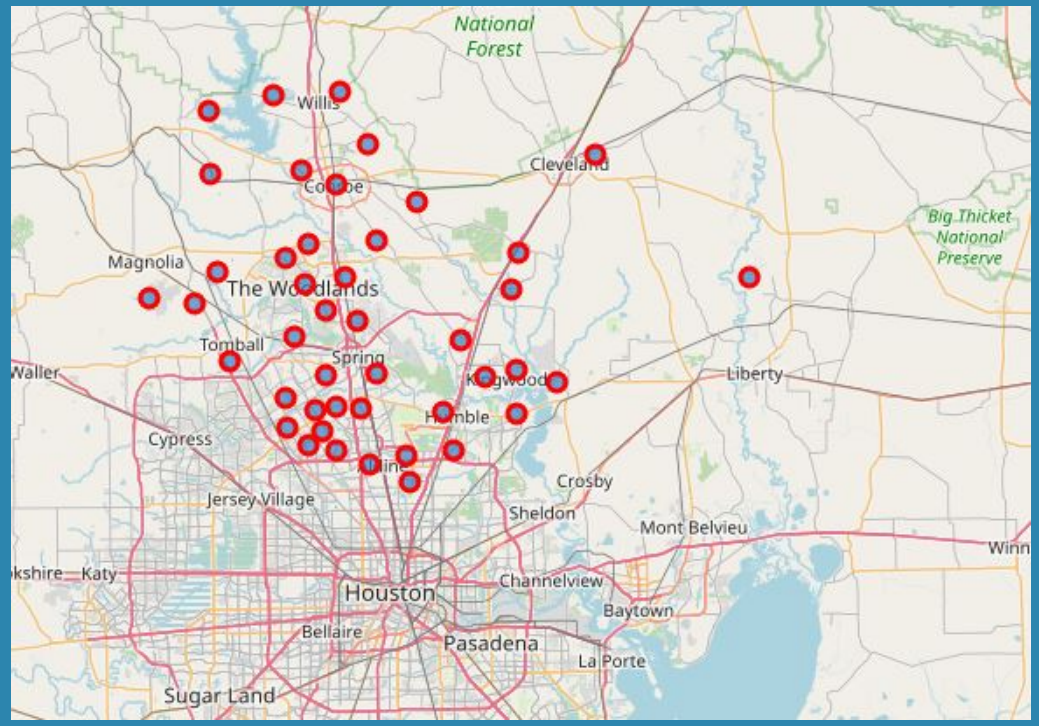
```
dist_import_df.head()
```

1	ZIP	City	Latitude	Longitude	Investment Intensity	Income Intensity	Prospectivity Score	Sum of Distances	Labels
2	77024	Houston	29.774	-95.5177	3104.04	74780	6.20318	2771.53	2
3	77479	Sugar Land	29.5733	-95.6321	1283.72	143880	4.8287	3960.68	2
4	77433	Cypress	29.8842	-95.7222	1089.36	126820	3.47945	3854.89	2
5	77007	Houston	29.7726	-95.4032	1414.95	88290	3.09607	2709.98	2
6	77459	Missouri City	29.5643	-95.5476	1053.37	110360	2.84459	3715.1	2



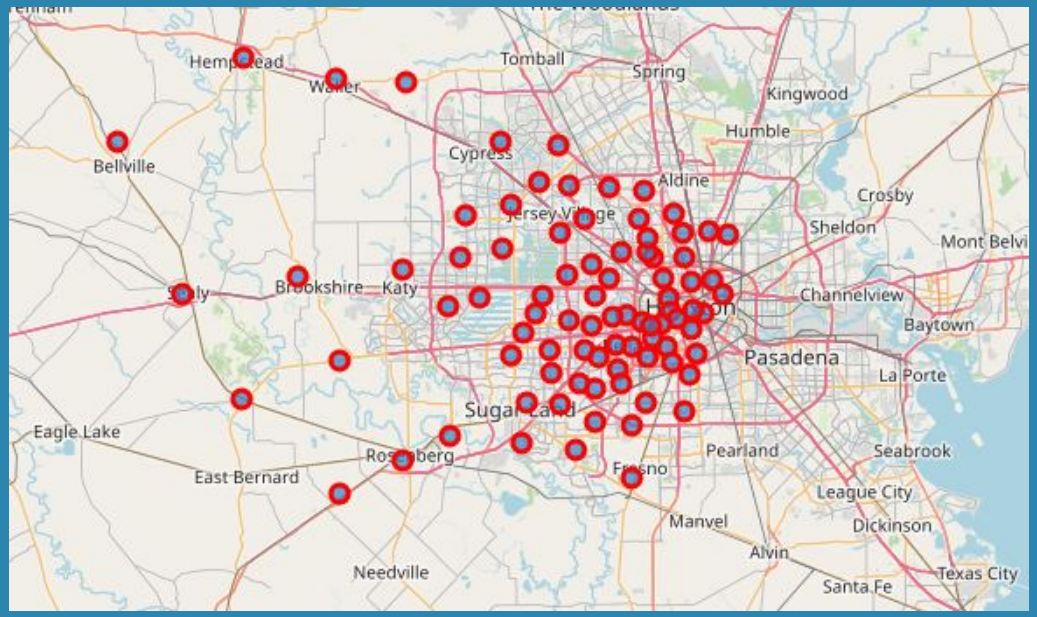
Cluster A

Cluster A- Label 0 is a less-prospective zone overall. The algorithm identified this cluster far south of the City of Houston proper. Of the 20 most prospective sites identified, *none of them are in Cluster A*.



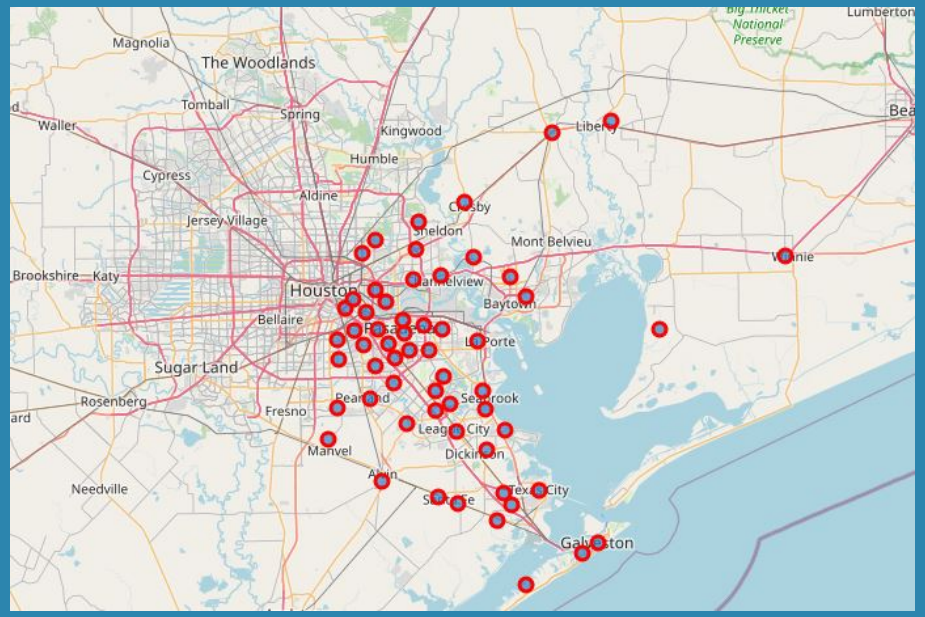
Cluster B

Cluster B- Label 1 is a less-prospective zone overall. The algorithm identified this cluster north of the City of Houston. Of the 20 most prospective sites identified, *four of them are in Cluster B. (20%).*



Cluster C

Cluster C- Label 2 is the most prospective zone overall, and consists of the greatest concentration of population, gross income, likely potential foot traffic, and demand for home improvement and construction products. 14 of the prospective sites are located in Cluster C (70%), including the top eight prospects.



Cluster D

Cluster B- Label 1 is a less-prospective zone overall. The algorithm identified this cluster north of the City of Houston. Of the 20 most prospective sites identified, *only two of them are in Cluster B. (10%).*

Top 20 List- Final Score

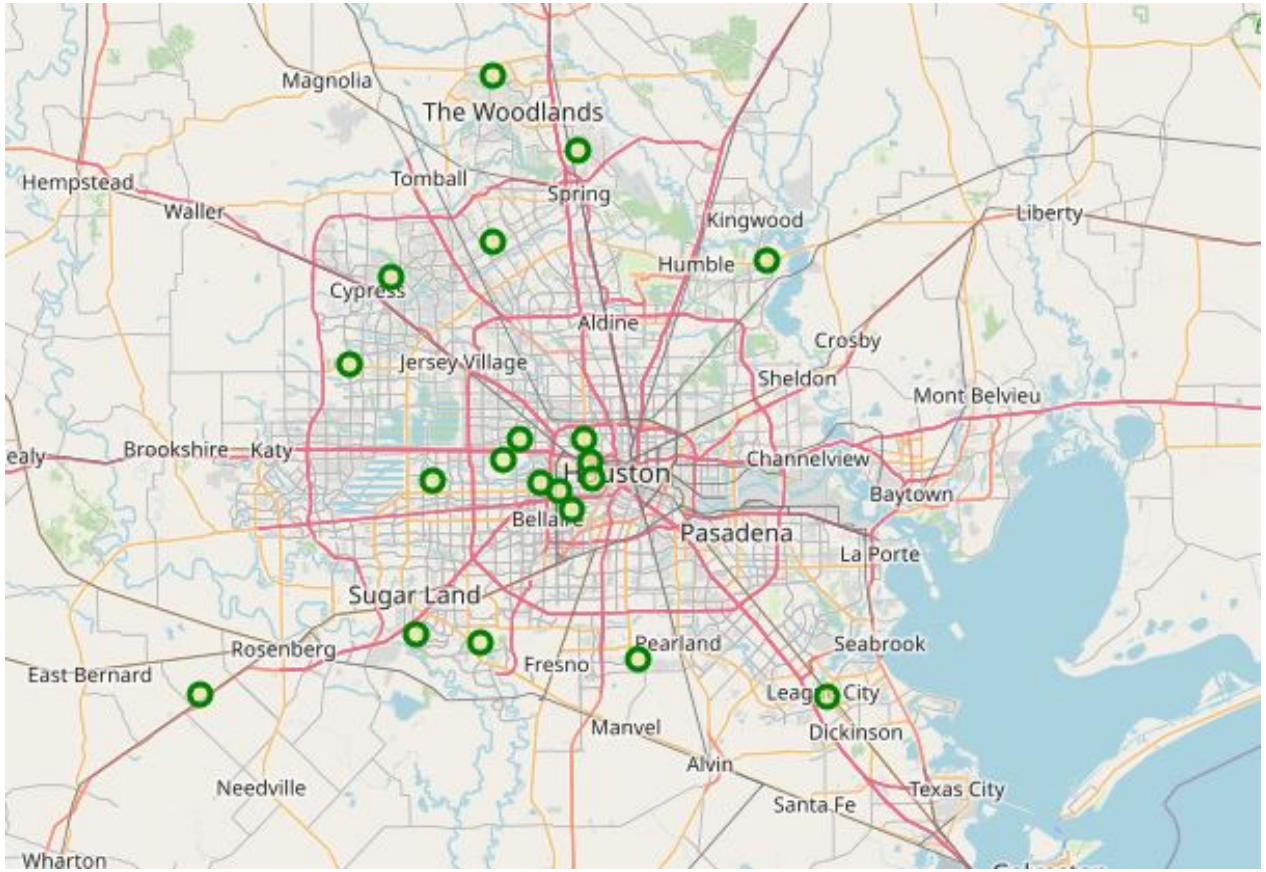
*Ranked selection based on the prospectiveness of optimal locations, including the sum of distances to competitor locations as a function of housing sales, gross income, population, and median housing prices in a radius of 150KM from the center of the City of Houston.

1	ZIP	City	Latitude	Longitude	Investment Intensity	Income Intensity	Prospectivity Score	Sum of Distances	Labels	Final Score
0	77024	Houston	29.774	-95.5177	3104.04	74780	6.20318	2771.53	2	17192.3
1	77479	Sugar Land	29.5733	-95.6321	1283.72	143880	4.8287	3960.68	2	19124.9
2	77433	Cypress	29.8842	-95.7222	1089.36	126820	3.47945	3854.89	2	13412.9
3	77007	Houston	29.7726	-95.4032	1414.95	88290	3.09607	2709.98	2	8390.3
4	77459	Missouri City	29.5643	-95.5476	1053.37	110360	2.84459	3715.1	2	10568
5	77008	Houston	29.7988	-95.4095	1298.13	69340	2.08407	2695.13	2	5616.82
6	77573	League City	29.5028	-95.0891	662.672	131500	2.00085	4759.28	3	9522.59
7	77005	Houston	29.7175	-95.4282	1644.2	52810	1.99183	2784.67	2	5546.58
8	77379	Spring	30.0247	-95.5322	670.549	115730	1.72436	3551.64	1	6124.31
9	77429	Cypress	29.9827	-95.666	604.344	123870	1.64487	3854.03	2	6339.35
10	77346	Humble	30.0019	-95.1696	805.103	88090	1.5307	4072.27	1	6233.43
11	77584	Pearland	29.5437	-95.3404	556.37	124010	1.47487	3704.33	3	5463.43
12	77056	Houston	29.7473	-95.4693	1411.95	47820	1.4321	2734.48	2	3916.05
13	77019	Houston	29.7525	-95.3992	1338.49	48860	1.37061	2734.24	2	3747.59
14	77055	Houston	29.7989	-95.4963	1266.87	50170	1.31728	2727.51	2	3592.91
15	77406	Richmond	29.504	-95.9191	800.16	75730	1.23141	6139.18	2	7559.86
16	77382	Spring	30.2147	-95.5321	949.881	59440	1.11155	5040.08	1	5602.31
17	77386	Spring	30.1289	-95.419	659.7	80910	1.02214	4210.67	1	4303.89
18	77077	Houston	29.7509	-95.6125	609.744	81250	0.910991	3145.19	2	2865.24
19	77027	Houston	29.739	-95.4436	1217.47	37130	0.785273	2736.09	2	2148.58

Top 20 List- Prospective Site Zones

Clear Indications:

- High population centers rank high.
- Areas of rapid growth and above-average gross income by tax filings trend in the most prospective clusters.
- Outlier areas, notably east and far South Houston are less prospective.
- Prospectiveness factors such as distance from competitors are in some cases entirely offset by the concentration of wealth and construction levels.
- Remote areas with robust home sales stats trend higher in the list on greater transaction volume remote from clusters of competing store locations.



Results and Conclusions

- Separation of distance, home sales, and general financial conditions are good but preliminary measures of prospectiveness.
- Nuance requires parameter weighting, as outliers trend extremely high or low in prospectiveness due to extreme factors (such as the benefits of remoteness that may not be correlated with highly prospective population and income levels).
- K-Means was an optimal Machine Learning Algorithm for this use case but alternatives such as DBSCAN would offer clustering based on parametric factors that cannot be linearly separated, such as identifying prospective pockets within Clusters that generally trend low in prospectvity.