Capstone Project- The Battle of the Neighborhoods

High-Grading Optimum Location Zones for a Startup Home Improvement Retailer

Chris Cothran

August 31, 2020

I. Introduction

1.1 Context

The U.S. home sales market continues to be robust. According to the National Association of Realtors, 5.34 million existing homes were sold in 2019, along with 682,000 newly constructed homes. Both contexts support significant investment in home construction, renovation, and maintenance materials. In consideration of the capitalization of a newbuild home improvement retailer, it is vital to be proximal to zones of high newbuild and home renovation and maintenance consumer spending. The patterns of new home construction are often correlated with economic growth and the increase of employment opportunities. As communities grow, the increase in consumption spending grows. This is particularly relevant in areas where new construction and renovation and maintenance spending is robust. Shopping at one retailer over another is often a function of a variety of factors, but understandably convenience is a key consideration for customers. Each year, millions of homes are sold, constructed, or are renovated or repaired. These processes require materials, parts, and expertise that home improvement retailers can provide. The importance of siting a store to garner revenue from all of these activities while being sufficiently distant from competitors so that local activities increasingly rely on your store is one of many factors a company may consider in determining where to locate a new store. It provides a home improvement retailer extensive benefit to be able to predict where optimal zones to build a new store might be. This information can then be incorporated into a wide range of other factors to yield a final business decision.

1.2 Business Problem

In this project we will try to derive a high-graded list of optimal locations for a home improvement retailer. This report will be targeted to stakeholders considering expansion into the home improvement retail market in Houston, Texas, USA. There are multiple well-established home improvement retailers operating in Houston. For this initial exercise, we will consider optimality as a function in part derived from proximity to zones of higher home construction, renovation, and maintenance spending and distance from competing home improvement retail stores. The stakeholders are likely interested in zones that may be underserved in this regard in multiple categories as ideal locations. However, assuming the optimal conditions are met, we would like to locate the store as close to high-population zones within the city to ensure adequate day-to-day foot traffic. We will use multiple data science processes to generate a high-grade list of the most promising neighborhoods for potential sites. Subsequently, we will highlight key features of each area, including the advantages and disadvantages, so that the best possible area can be selected. Foursquare data will be crucial to this process, as it allows the depiction the location of the hypothetical home improvement retailers competitors, so that the potential identification of optimal but underserved locations can be made.

1.3 Stakeholders

Stakeholders in potential home improvement and home construction suppliers would be interested in siting their stores/facilities in optimal locations to garner maximum revenue and market share. An accurate prediction of the location of these sites would bring competitive advantage. Furthermore, the analysis of such siting criteria and findings may be of interest to a variety of related interests, such as home builders and real estate professionals.

II. Data

2.1 Data Sources

Multiple factors influence the decision of a stakeholder on where to site a potential store. This analysis is purposed to provide support to that process by identifying optimum outcomes based on selected criteria. Key factors in this context include:

1) Number of existing home improvement retailers in a city neighborhood
2) Their proximity to zones of high construction, consumer spending, renovation, and maintenance
3) Distance of neighborhood from concentrated population zones
4) Housing listings
5) Economic growth
6) Proximity to existing shopping centers for convenience of access

For this project, I will source data from the following locations: The U.S. Census Bureau, the U.S. Internal Revenue Service, the National Association of Realtors, and Foursquare.

1) Foursquare shall provide the data for the number of existing home improvement retailers in a city neighborhood, distance of neighborhoods from concentrated population zones, and proximity of existing shopping centers for convenience of access.
2) National Association of Realtor's data provided from Realtor.com shall provide the number of housing listings by zip code.
3) The U.S. Internal Revenue Service shall provide population data by zip code used to determine density proximity to potential sites based on number of filed income tax returns.
4) U.S. Internal Revenue Service data for two subsequent years shall provide economic growth indications based on the number of returns by income level in a given zip code year-over-year.

An important consideration is the timeliness of the data for accurate evaluation. Although the Foursquare, Census Zip Code designation, and National Association of Realtors data are dated 2020, the latest available IRS dataset by zip code is for 2017. This reflects information relevant to the 2018 tax filing season. However, the identical methodologies would be utilized for updated datasets, and the findings would not likely differ significantly in the span of two years.

2.2 Data Cleaning

I downloaded data from multiple web sources using a variety of processes, including use of the WGET Python library, API calls, and web scraping. All these datasets needed to be managed to be combined into a single dataframe. Because of the data gap in filing information after the 2018 tax filing season, I decided to carry this data forward into 2020 without modification. This process would thus lose granularity, but it would maintain functionality as an effective snapshot demonstration once updated data is available. At such time, a historical comparison of 2020 data can be made with the archives of the Foursquare GET request data saved in local files.

There are several problems with the datasets. For the IRS data, a substantial portion of the dataset is unnecessary. For simplicity, the data was managed directly in Excel at first. Several dataset elements were removed. The top four and bottom 12 rows are notation. All but the first three columns are extraneous material not pertinent to the analysis. Blank entries are unhelpful and may skew the results, so this information was removed. Summary columns and rows were also removed. Two columns were renamed to better reflect the data and to maintain consistency with other datasets, such as the ZIP column. Formatting was cleared, and the data was then put into a new spreadsheet for memory management purposes. Once the sheet was saved in a new .xls file, it was uploaded for storage on GitHub.

The other datasets were managed entirely within Python. For the NAR data, we extracted and renamed the required columns, notably postal code, YTD reference, median listing price, and number of active listings. As the NAR datasets contains extensive historical data, data before 2020 was removed.

Zip code data was derived via a table scraping process and isolated into an Excel sheet. This was then imported. Reformatting of this data was necessary as the read_excel process identified the zip code data as an integer rather than an object. The IRS data listed a set of categorical ranges that were not helpful. I assigned to each category an integer. By doing this, this information could be leveraged effectively in calculations.

After resolving these issues, I checked for outliers in the data, notably what information could potentially skew the results. The datasets were well maintained, and all values were within expected bounds, though it was important to remove summary information on a time basis or area basis, as the goal was to render a conclusion based on individual zip codes for a hypothetical time period of July 2020. Invalid values were largely an issue only with the Foursquare query data, as the granularity of the query was sufficient to obtain a variety of results but not tailored specifically to exclude similar concepts. For example, 'home improvement' and 'construction' search terms yield some invalid or non-pertinent results. These were excluded via subsequent dataframes via manual identification and column flagging and filtering into a new dataframe.

2.3 Relevant Data

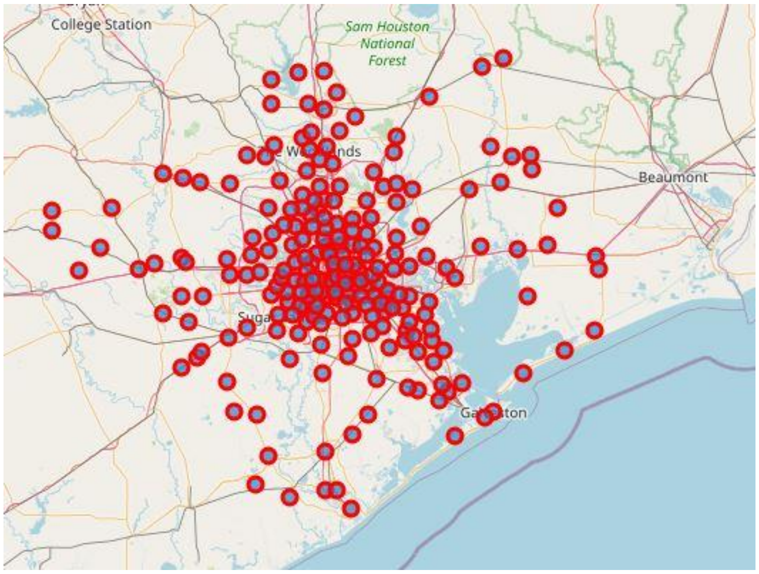
Evaluation of potential sites for a new store involves multiple spectrums of consideration. Stakeholders in new business ventures view location among the paramount consideration of business models predicated on foot traffic. Population is thus an indicator, but it is not the only key or siting considerations.

Expected performance of a potential store could be highly reflective of the optimality of the location, distance from competitors, and proximity to demand centers. For the perspective of this study, we look at potential zones of interest for siting a new home improvement store in the Greater Houston Metropolitan Area.

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. Our potential stakeholders are considering siting a new concept Home Improvement retailer in Houston, but proximity to competitors, such as Home Depot and Lowe's, may impact decisions. Ideally, the stakeholders would like to identify potential underserved niches.

The ZIP code center lines utilized for this analysis are depicted in the Folium map below.



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 provided from Realtor.com supplied the number of housing listings by zip code. 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. This data point provides a relative snapshot of the amount of total dollars associated with new and existing home sales. 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.

The U.S. Internal Revenue Service Census data provides filing data by zip code. 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.

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.

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.

Once the JSON results were assigned to a python variable, this resulting dictionary data type could be parsed with native pandas dataframe processing. I used a while loop with double iterators to derive the key: value pairs for the Response-Venues-[X]-Location nested chain to derive the longitude, latitude, location name, and city name from the data. Invalid values were dropped from the dataframe, and the helper flag used to identify invalid values was also removed.

III. Methodology

In this project, we ultimately seek to find a top 20 list of prospective Zip code center points for siting a new concept Home Improvement store in the Greater Houston Area. We limit our analysis to a radius of 90 kilometers from the Houston city center. We collect data from the National Association of Realtors (to define areas with housing listings for an intensity score relative to the number of houses listed for a given period (YTD July 2020). We then use Internal Revenue Service Data on IRS filings by adjusted gross income of filers in a zip code to determine aggregate wealth intensity.

We then implement the Haversine formula via a defined function for each of the queried competing Home Improvement store locations to obtain cumulative distances from stores. The areas with the most robust number of listings, highest total income levels, and longest cumulative distances from competitors establish our Top 20 listing results below.

In the third step, we will use K-Means clustering to define cluster labels based on proximity. These will then be mapped, and the Top 20 will be mapped. In the Report, there is a discussion on the data findings specific to these labels.

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, and Machine Learning. We utilize the popular and versatile K-Means Clustering Algorithm for unsupervised learning.

As viability is in many ways associated with proximity to competitors, particularly in new concept startups, we will calculate the distance between each of the Hardware Store Locations Vs the Prospective Zip Code List. This was accomplished using the Haversine mathematical formula.

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.

We implemented the K-Means algorithm as depicted at right with 4 clusters. The cluster results follow. K-Means is a popular clustering algorithm used to analyze difference in areas. One application is partitioning groups into similar characteristics, such as proximity or distance from a mean. The algorithm divides the data into non-overlapping subsets (clusters) without any cluster-internal structure or labels (unsupervised). Objects in clusters are similar (in this case proximal to one another and remote from other clusters). Objects in other clusters are thus dissimilar. Dissimilarity measurement is used to shape the clusters. The algorithm tries to minimize the intra-cluster differences and maximize the inter-cluster distances. The algorithm natively uses the Euclidean distance calculation, but others can be used.

Important notes relevant to this algorithm:

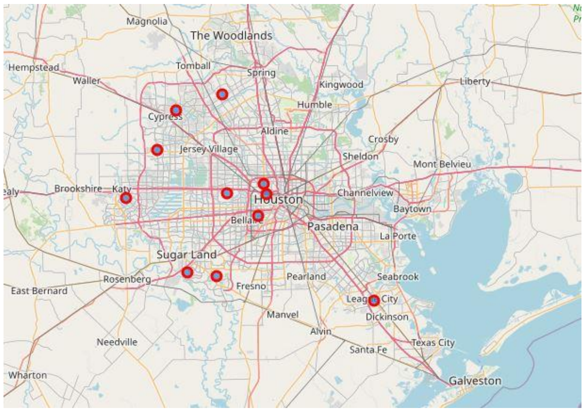
- First, one must normalize the feature set to get the accurate distance measurement
- Must initialize K, which represents the number of clusters
- Centroids selected randomly (number of Centroids is the value of K)
- The algorithm assigns points to the closest centroid, forming a Matrix with each row representing the distance of a customer from each centroid (Distance Matrix)
- Main objective of K-Means clustering is to minimize the distance of a centroid from its cluster and maximize the distance from other cluster centroids
- Use the distance matrix to find the nearest centroid to the data points and assign to a centroid
- All customers fall into a cluster based on their distance from the centroids
- Error is the total distance of each point from its centroid, which is expressed as the sum of the squared difference between each point and its centroid
- Reduce error by moving the centroids according to the mean of the centroid members
- Centroids continue to move until they no longer move to match the mean of the clusters
- K-Means is an iterative algorithm, repeating each step until the Centroids no longer move. This results in the most dense (minimized distance) clusters

- However, as K-Means is a heuristic algorithm, there is no guarantee that it will converge to the global optimum
- To handle this, repeat the k-Means process and select the best results based on multiple random centroid locations

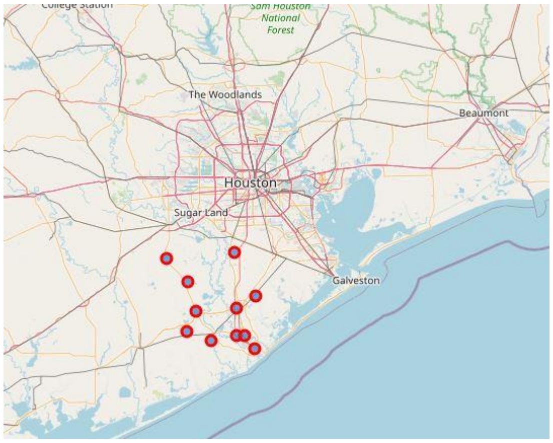
K optimization was not performed in the context of this study, as the algorithm was used to cluster based on data. The centrality of the clusters to each centroid was thus not statistically significant.

IV. Results

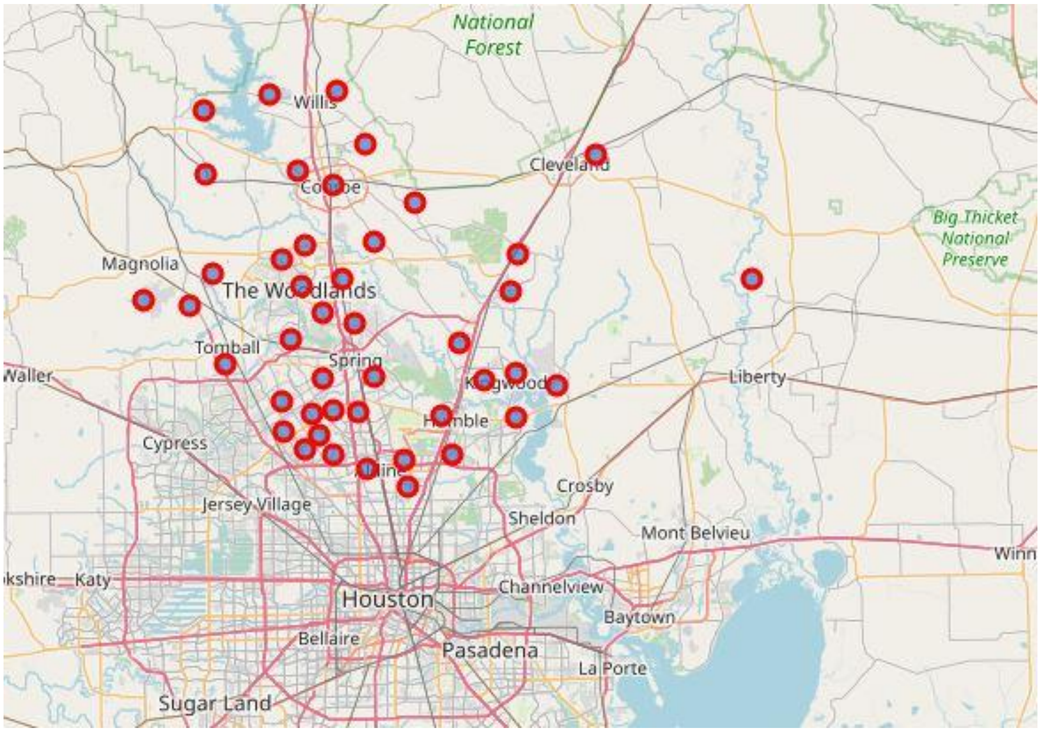
Though not the final determination, the derived Prospectiveness Score, a factor of the investment intensity and income intensity values, provides an indication of key zones of investment. It is important to note the growth of Katy and Sugar Land in this estimation. Both zones are known for both rapid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



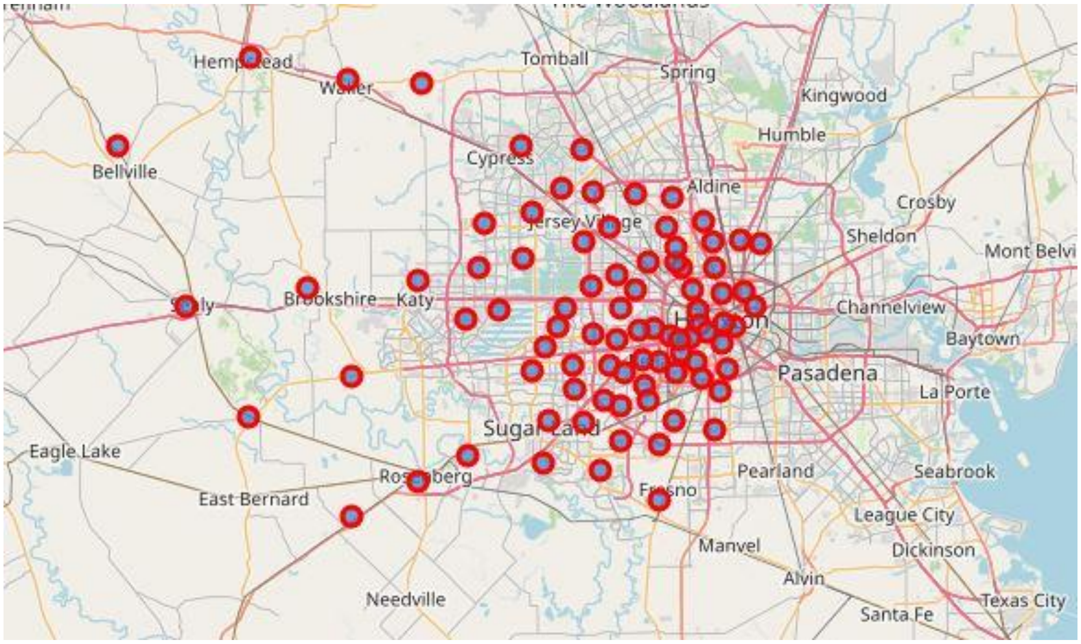
Cluster A- Label 0 is a less-prospective zone overall. The algorithm identified this cluster far south of the City of Houston proper. Of the 20 most prospective sites identified, *none of them are in Cluster A*.



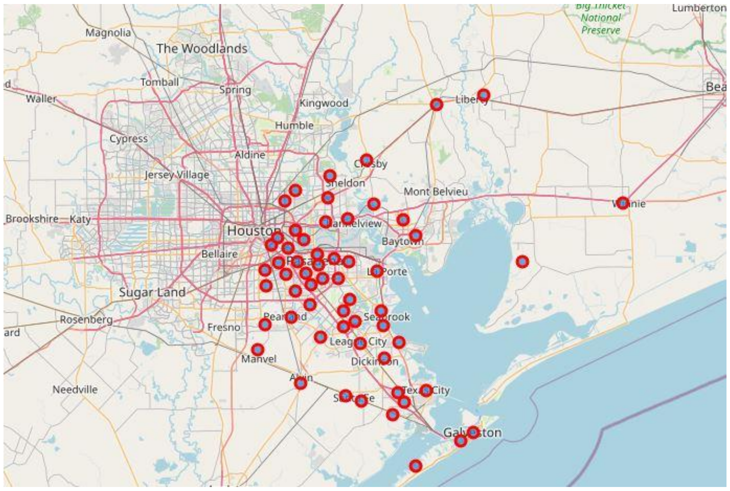
Cluster B- Label 1 is a less-prospective zone overall. The algorithm identified this cluster north of the City of Houston. Of the 20 most prospective sites identified, *four of them are in Cluster B. (20%).*



Cluster C- 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 in Cluster C (70%), including the top eight prospects.



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%).*

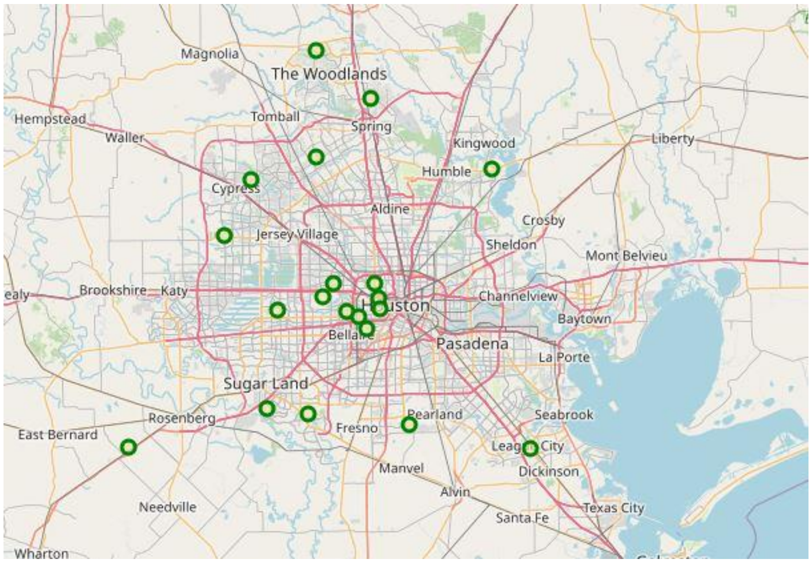


Ranked selection based on the prospectiveness of optimal locations, including the sum of distances to competitor locations as a function of housing sales, gross income, population, and median housing prices in a radius of 150KM from the center of the City of Houston.

Top 20 List- Prospective Site Zones

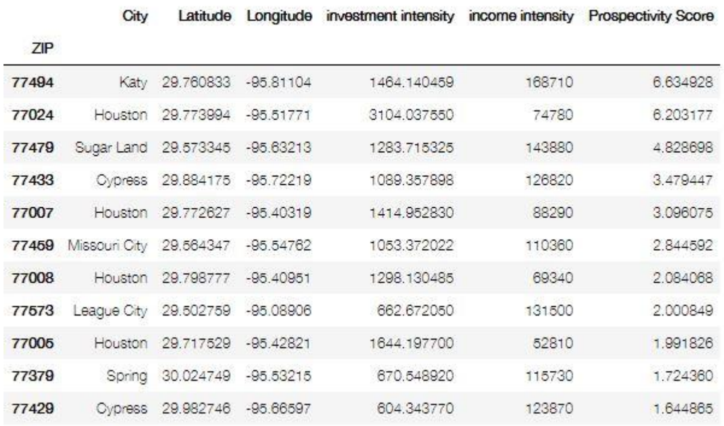
Clear Indications:

- 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.



V. Discussion

- 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 prospectiveness.
- 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 following is a 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.



	City	Latitude	Longitude	investment intensity	income intensity	Prospectivity Score
ZIP						
77494	Katy	29.760833	-95.81104	1464.140459	168710	6.634928
77024	Houston	29.773994	-95.51771	3104.037550	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

VI. Conclusion and Future Directions

In this study, I analyzed potential prospective zones for siting a home improvement store, leveraging location data via API calls, a machine learning algorithm to cluster this data, and mapping and data visualization functionality to depict the most prospective locations within the Greater Houston Area. I identified population as the most evident indication of location clustering, largely due to foot traffic, but there are other factors. Expected performance considerations would be one, as well as long-term viability. I built a machine learning model to utilize location data to cluster locations based on distances. Interestingly, this same factor was utilized in the viability scoring, as the Haversine formula derived a sum of distances for each location relevant to competing stores. The findings show that remote stores have higher ratings, but only when corroborated with robust housing sales and gross income data. The top 10 reflects areas of significant and ongoing suburban development. The most prospective zone is both comparatively remote from most competing locations and both active in real estate sales, including newbuilds, maintenance, and restoration investment, as well as relatively intensive in per-capita income levels.

Future analyses could examine the parameters with a weighting consideration. To do this was out of scope for this project. However, by examining modification of weighting scoring parameters as a function of regression analysis of the relationship between scoring outcomes and actual newbuilds a more nuanced view of the relative viability of prospective zones could be established. Financial and marketing considerations are also key. For example, the importance of foot traffic may completely outweigh proximity and convenience. Furthermore, at times competing stores opt to co-locate at prime locations. This preference for proximity to economic and construction zones far outweighs the negative impact of lower combined sum of distances from competitors. In some contexts, that factor may less heavily than the others. In conclusion, the top 20 prospective zones identified in this analysis correlates highly with known opportunity areas, and thus we anticipate that with further testing the model highlighted in this analysis could be reliably used in commercial application as a preliminary decision intelligence and analytics process.