Adonis Ichim

June 2020

Food Access in Houston, TX, USA

Applied Data Science Capstone

1. Introduction

The City of Houston, Texas, USA is the fourth most populous city in the United States with an estimated 2019 population of 2,320,268. The Houston metropolitan area spans over 1,062 square miles (or 2,750 square kilometers), and is home to more than 6.9 million residents. Houston's economy has a broad base in energy, healthcare, manufacturing, aeronautics, and transportation. Houston is a majority-minority city, with a vibrant, ethnically and culturally diverse community. The city boasts more than 10,000 restaurants, bars, and supporting businesses.

Because of the city's large surface area, limited options of efficient public transportation, zoning laws, and income variability, some zones may not provide adequate food access to the population. According to USDA 2014 data, Houston had 0.17 grocery stores per 1,000 people. Lack of a suitable food supply in the vicinity of one's home can lead to health issues, additional costs and longer times to purchase food. This may also represent lost revenue for the city and the food supply chain.

I will explore what those zones are today and provide solutions to decrease the number of residents living in a food desert. The outcomes of this project should help investors decide where to build food stores and city officials to understand problematic zones where they could intervene with additional support.

2. Data

To determine zones with unsatisfactory food access, I leverage Foursquare's API and access geospatial data. The venue categories I will focus on are:

- grocery store,
- supermarket,
- market,
- big box stores
- food & drink shop.

I combine this with Census data, centralized and published by Houston State of Health. The census data pertains to:

- income,
- living cost,
- internet and vehicle availability,

and enables our understanding of what type of food venue one may consider for a specific location. Data will be analyzed by Postal Code, which will increase the granularity of the analysis.

2.1. Census Data and Postal Code Coordinates Data

The census data has been downloaded as six separate .CSV files from houstonstateofhealth.com and uploaded to the IBM Cloud environment. Figure 1 shows an example of one CSV file loaded as a dataframe. Each file was loaded as a streaming body and went through a similar cleaning process consisting of:

- Dropping columns without values
- Renaming columns to better describe the parameters they contain
- Group the data in each dataframe by Postal Code and use average in case there are multiple reported values for one Postal Code
- Combine the dataframes in a new one.

	Indicator Name	What Is This Indicator	Location Type	Location	Indicator Rate Value	Indicator Rate Value Units	Rate Lower Confidence Interval	Rate Upper Confidence Interval	Indicator Count Value	Indicator Count Value Units	...	Breakout Rate Value	Breakout Rate Value Units	Breakout Rate Lower Confidence Interval	Breakout Rate Upper Confidence Interval	Breakout Count Value	Breakout Count Value Units	Breakout Count Lower Confidence Interval	Breakout Count Upper Confidence Interval	Breakout Unstable	Breakout Footer
0	Homeownership	This indicator shows the percentage of all hou...	Zip Code	77002	15.6	percent	11.8	19.4	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Homeownership	This indicator shows the percentage of all hou...	Zip Code	77002	19.7	percent	16.0	23.4	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 1 - Homeownership Dataframe

The resulting dataframe contains 130 entries and 6 columns, as seen below.

	Postal Code	Median Rent ($)	Households wihout vehicle (%)	Households with internet (%)	Per Capita Income ($)	Households below poverty level (%)	Homeownership (%)
0	77002	1401.666667	17.6750	82.75	34888.500	9.0250	17.8250
1	77003	1008.666667	16.3875	79.35	31134.125	33.3375	33.3500
2	77004	946.666667	20.6000	73.85	30638.000	19.7625	25.2625
3	77005	1713.333333	3.9875	94.45	94654.500	1.5625	66.8625
4	77006	1260.666667	7.2500	87.30	64342.625	3.2625	31.2875

Figure 2 - Census dataframe

I downloaded the Texas Postal Codes Coordinates from the public dataset stored at public.opendatasoft.com.

	Zip	City	State	Latitude	Longitude	Timezone	Daylight savings time flag	geopoint
0	75475	Randolph	TX	33.485315	-96.25525	-6	1	33.485315,-96.25525
1	75757	Bullard	TX	32.136787	-95.36710	-6	1	32.136787,-95.3671
2	78650	McDade	TX	30.283941	-97.23563	-6	1	30.283941,-97.23563
3	75010	Carrollton	TX	33.030556	-96.89328	-6	1	33.030556,-96.89328
4	76054	Hurst	TX	32.858398	-97.17681	-6	1	32.858398,-97.17681

Figure 3 - Postal Code Coordinates Dataframe before Processing

I then copied Houston, TX ZIP entries in a separate dataframe, renamed columns, dropped columns of no use (Timezone, Daylight savings time flag, geopoint) and set Postal Code as Index.

	Latitude	Longitude
Postal Code		
77046	29.733181	-95.431310
77015	29.778526	-95.181180
77289	29.833990	-95.434241
77072	29.700898	-95.590020
77216	29.833990	-95.434241

Figure 4 - Postal Code Coordinates Dataframe after Processing

I then merged the postal codes coordinates dataframe with the census dataframe, obtaining 95 entries for each postal code including the data I targeted at the beginning.

	Median Rent ($)	Households wihout vehicle (%)	Households with internet (%)	Per Capita Income ($)	Households below poverty level (%)	Homeownership (%)	Latitude	Longitude
Postal Code								
77002	1401.666667	17.6750	82.75	34888.500	9.0250	17.8250	29.755578	-95.36531
77003	1008.666667	16.3875	79.35	31134.125	33.3375	33.3500	29.749278	-95.34741
77004	946.666667	20.6000	73.85	30638.000	19.7625	25.2625	29.728779	-95.36570
77005	1713.333333	3.9875	94.45	94654.500	1.5625	66.8625	29.717529	-95.42821
77006	1260.666667	7.2500	87.30	64342.625	3.2625	31.2875	29.741878	-95.38944

Figure 5 - Census and Postal Code Dataframe

2.2. Geospatial Data

This step included some data visualization through Folium. City coordinates were obtained via GeoPy Geocoders Nominatim as (29.7589382, -95.3676974). I initialized a map and plotted with blue circles the Postal Code centers.

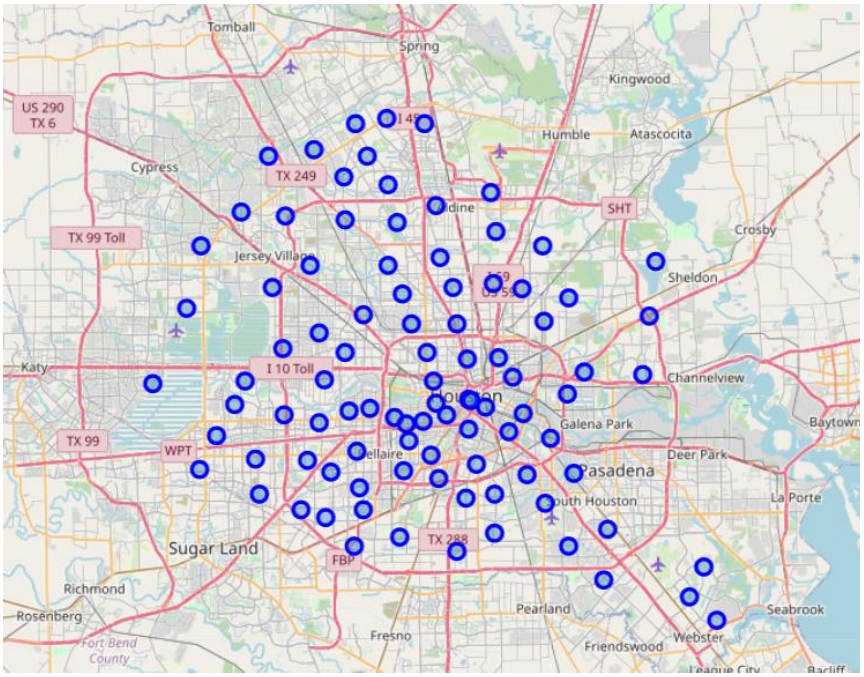


Figure 6 - Houston Postal Code Centers

I used the Foursquare API to retrieve data on stores selling food items in Houston. For that, I wrote a function that uses the Foursquare Category ID as input. These IDs are accessible at [developer.foursquare.com/docs/build-with-foursquare/categories/.](developer.foursquare.com/docs/build-with-foursquare/categories/) The function also takes a radius and venues limit and longitude and latitude as inputs.

For grocery stores, I chose to work with "Supermarkets", there are other IDs that may fall within this classification such as:

- #4bf58dd8d48988d118951735 - grocery store
- #52f2ab2ebcbc57f1066b8b45 - organic grocery
- #4bf58dd8d48988d1f9941735 - food and drink shop
- #52f2ab2ebcbc57f1066b8b46 - supermarket

Executing the function returns a dataframe with 1,182 entries and 7 columns which includes the postal code and venue coordinates, venue name, and venue category.

	Postal Code	Postal Code Latitude	Postal Code Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	77002	29.755578	-95.36531	A&K Sammis Convenient & Gift Store	29.755707	-95.364785	Grocery Store
1	77002	29.755578	-95.36531	Whole Foods Market	29.744379	-95.381412	Grocery Store
2	77002	29.755578	-95.36531	D J Super Market	29.744470	-95.332044	Grocery Store
3	77002	29.755578	-95.36531	Grit Grocery	29.761934	-95.362135	Food Truck
4	77002	29.755578	-95.36531	Midtown Food Store	29.749687	-95.377785	Grocery Store
5	77002	29.755578	-95.36531	Hollywood Food & Cigars	29.747169	-95.391777	Smoke Shop
6	77002	29.755578	-95.36531	Sunrise Grocery	29.781233	-95.372794	Grocery Store
7	77002	29.755578	-95.36531	Sunny's Food Store	29.747130	-95.381287	Grocery Store
8	77002	29.755578	-95.36531	1st Stop Convenience Store	29.742616	-95.388596	Grocery Store
9	77002	29.755578	-95.36531	Kim Hung Market	29.750712	-95.355355	Grocery Store

Figure 7 - Venues Dataframe

Then I plotted the venue datapoints on the previously shown map.

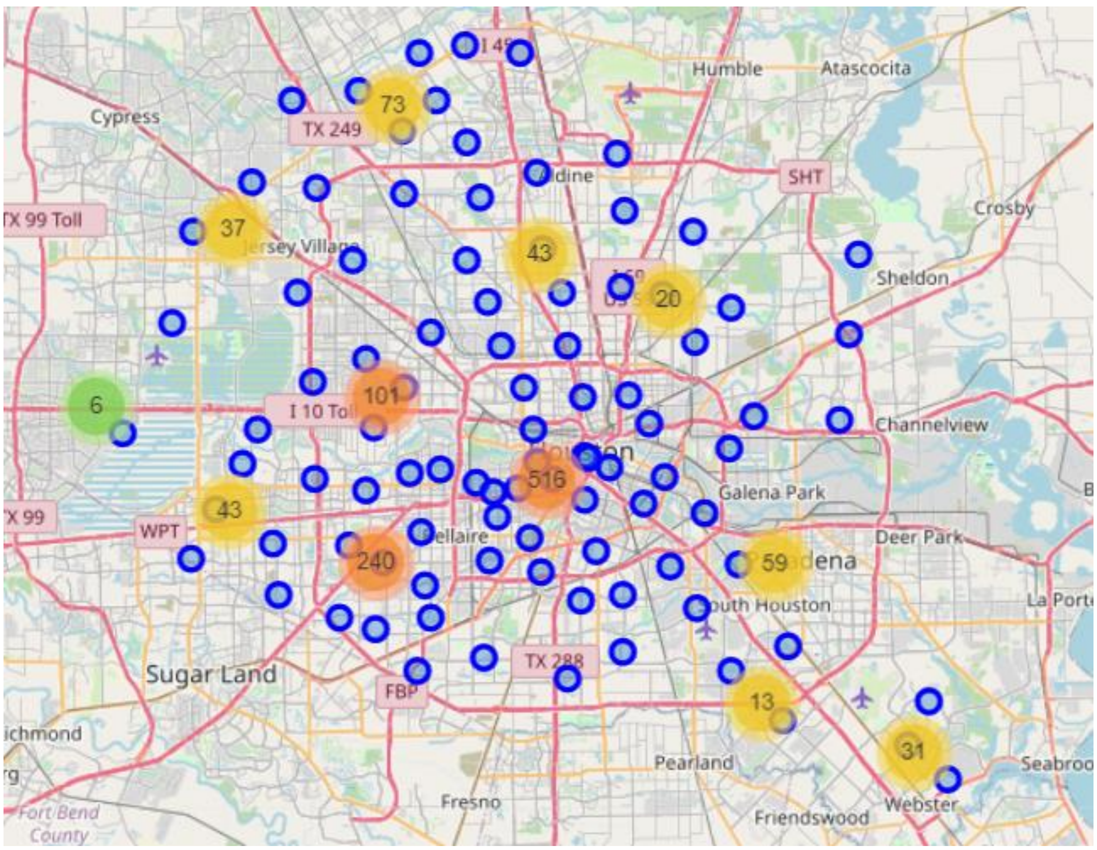


Figure 8 - Houston Postal Code Centers and Clustered Stores Map

When zooming in certain areas, I found some duplicate pairs of longitude and latitude coordinates. I removed duplicates and ended up with 1,042 entries.

3. Methodology

3.1. Exploratory Statistics

I first explore the Census dataset described earlier. For the 95 Postal Codes analyzed, the mean of the Median Rent is around 1,035 USD and the average per capita income is 33,400 USD. About 8.5% of the households do not have an internet connection, 17.33% find themselves below poverty level, and 75.82% have an internet connection.

In the venues dataset, there are 1042 entries, with 794 venues labelled as Grocery Store, 136 as Supermarket, and 30 as Big Box Store. There are some other ones in there like Convenience Store, Fish Market, etc., which may sell grocery items, so I will include these in the analysis.

I also observed that Postal Code 77081 has 32 grocery stores, while some other postal codes have less than 5.

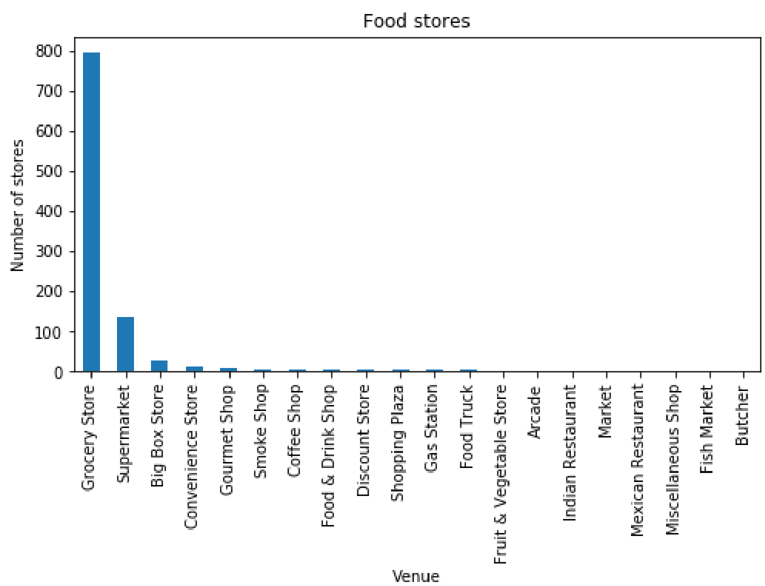


Figure 9 - Stores Selling Grocery Items in Houston

3.2. Machine Learning – Census Data Clustering

I use K-Means Clustering to partition my census dataset into K pre-defined distinct and non-overlapping subgroups and gain insights about the population of each postal code. First, I normalize the Census dataset to help algorithms interpret features with different magnitudes and distributions equally. I use StandardScaler to normalize the dataset. Both are available in the SciKit Learn library.

I tried out multiple number of clusters and found best results with five clusters. I assign the cluster label to each data row and add it to the label on the map.

```
[4 4 4 1 2 1 2 4 3 0 0 0 0 0 0 0 4 1 0 0 0 0 1 2 0 1 0 0 2 0 0 0 0 0 0 0 0
 0 4 4 4 4 4 0 1 0 0 0 0 0 4 4 1 2 4 2 0 0 4 4 4 4 0 0 4 2 4 0 0 0 0 0 0 4
 0 2 0 0 4 0 4 0 0 0 0 0 0 0 0 0 2 4 4 1 0]
```

Figure 10 - Labels Output from K Means Clustering

4. Results

We can observe a few trends from the K Means clustering:

- Cluster 0 has the lowest Per Capita Income, Median Rent Value and Internet Subscriptions. 23.6% of the households live below the poverty level, and ~45% own their homes.
- Cluster 4 has a higher Per Capita Income and Median Rent than Cluster 0, Internet usage above 80%. 12.86% of the households in these postal codes live below the poverty level, and 43.5% own their homes.
- Cluster 2 has a higher Per Capita Income and Median Rent than Cluster 4, Internet usage close to 90%. 6% of the households in these postal codes live below the poverty level, and homeownership is at 52.5%.
- Cluster 1 has a higher Per Capita Income and Median Rent than Cluster 2, and Internet usage close to 90%. 3.65% of the households in these postal codes live below the poverty level, and homeownership is at 42.8%.
- Cluster 3 seems to be an outlier. Here, Per capita income is 120% higher than in Cluster 1 Postal Codes. Rent is also much higher. Internet usage is above 90% and homeownership is low, as well as the percentage of households below poverty level.

Cluster	Median Rent ($)	Households wihout vehicle (%)	Households with internet (%)	Per Capita Income ($)	Households below poverty level (%)	Homeownership (%)
0	849.528302	10.442925	67.488679	16815.528302	23.612500	44.721226
4	1071.902778	7.006771	83.495833	33270.895833	12.864583	45.350521
2	1320.407407	4.701389	89.277778	57155.888889	6.094444	52.537500
1	1580.666667	4.904688	90.600000	86004.625000	3.651562	42.828125
3	3154.000000	4.925000	93.400000	188382.250000	2.862500	4.362500

Figure 11 - Census Data Labels

The figures below show clustering results on different parameters. We see that:

- Strong Homeownership is not really influenced by Per Capita Income (upper left).
- There are households below poverty level in all clusters (upper right).
- Median rent and Per Capita Income are almost linearly correlated (bottom left).
- There is a power law increase of internet usage with per capita income.

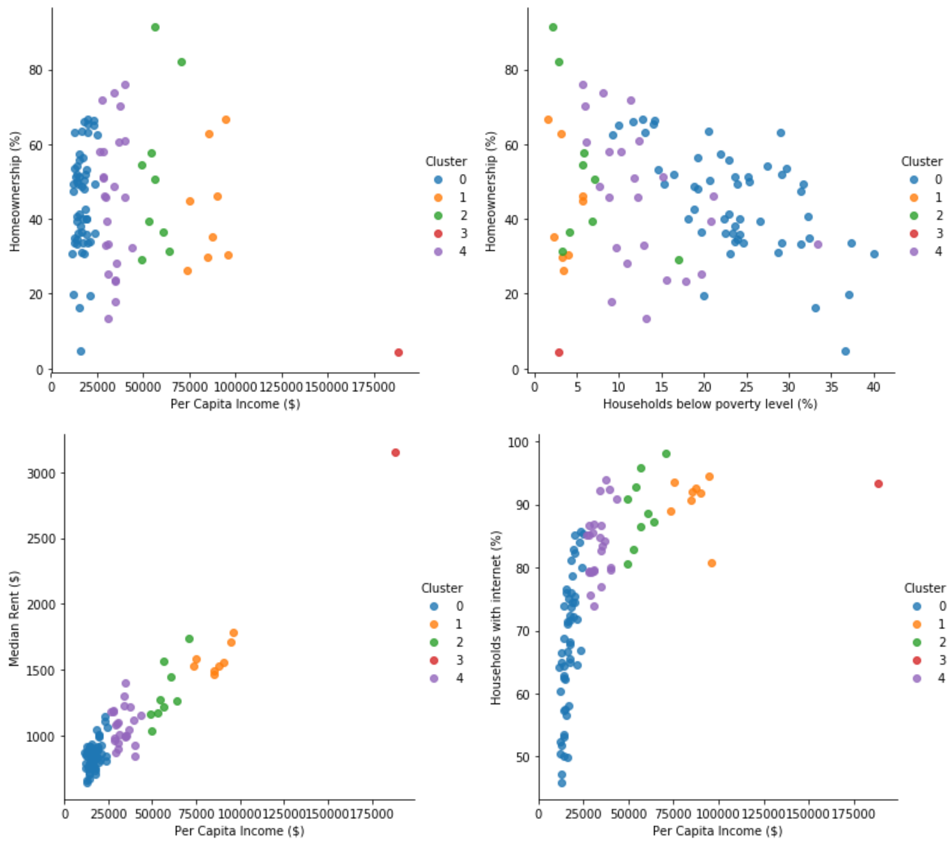
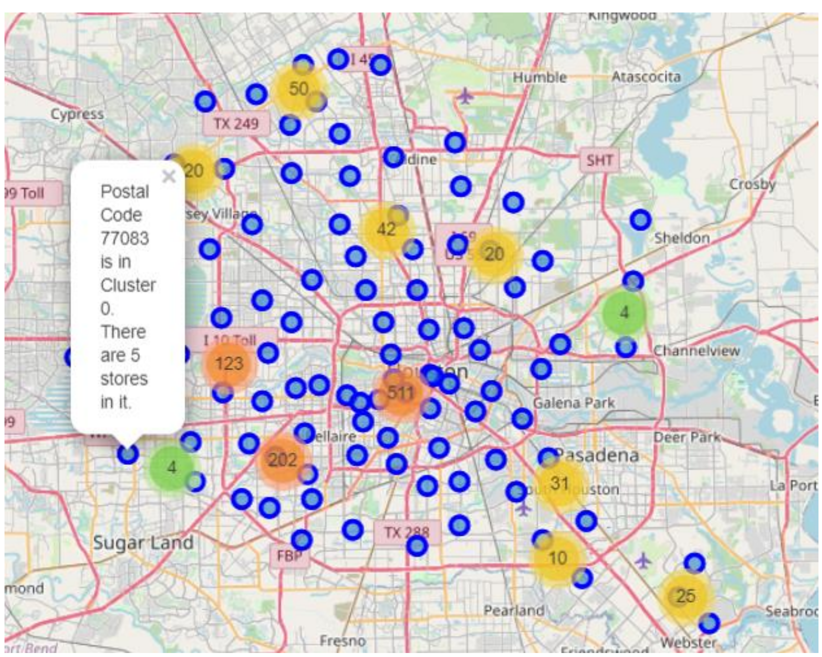


Figure 12 - Clustering and Census Data

Putting all results together on a map allows an enhanced visualization of the analysis with access to the postal code, the cluster identified for the postal code, and the number of tores. This map is the main deliverable of the project.



5. Discussion

5.1. General Discussion

Based on the analyzed data, I observe that most of the food stores in Houston, Texas are in the center and West part of the city (the so-called Inner Loop, Westside, and Bellaire areas). Developers or city officials can access this data and understand which areas might need additional stores and what type of stores one may develop.

For example, postal code 77085 only has 4 stores, all in the same spot. One may consider investing in a cost-accessible grocery store that would be valuable for the identified Cluster 0, which appears to be lower income. Since vehicle ownership is lower, smaller, denser stores within walking distance to living communities could be profitable. Similarly, city officials might consider opening weekend markets to increase food access.

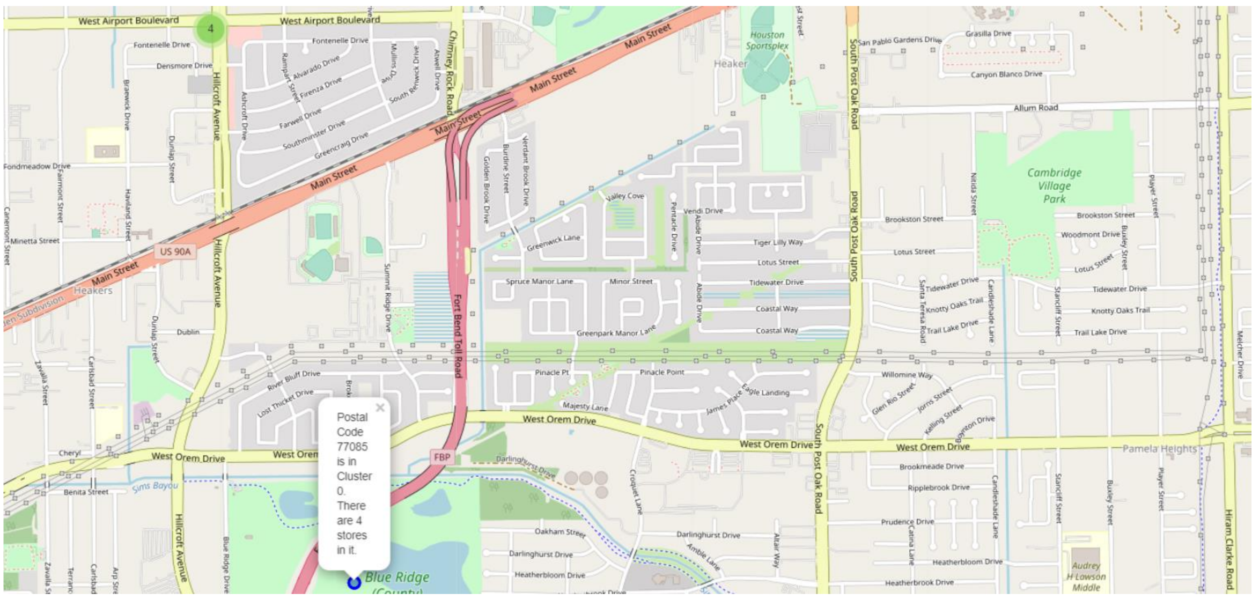


Figure 13 - Postal Code 77085

Furthermore, a fresh food delivery company may consider opening a distribution point and advertise in an area with less stores and less vehicle ownership. Cluster 4 Postal Codes would be a good target, as per capita income and internet usage are higher. The latter would allow the implementation of an online order system.

The data can also be used to decide where not to open a store, which is equally important in the decision-making process of selecting a location.

5.2. Future Work

This analysis would benefit from the inclusion of additional census data and organizing it by Super Neighborhoods instead of Postal Codes.

Throughout this project, I have realized that working with Postal Codes, albeit its granularity, might not be the best method, as Postal Codes' boundaries are hard to identify. “Super Neighborhoods” (defined in Houston as just larger areas of the city) or Census Tracts may be easier to work with and understand.

An additional important point would be the addition of population data, as this may enable decision-makers to understand potential exposure to customers.

Regarding the geospatial data, additional types of venues can be added. Data should be crosschecked with an official city source to make sure businesses are active. Another risk in Houston for businesses is flooding. A flooding risk map could be implemented to better select a business location.

Finally, when visualizing the results, a chloropleth map may prove itself useful. This ties into working with Postal Codes, as there was no reliable dataset allowing plotting of the Postal Codes boundaries and overlaying results of the clustering algorithm, reason why I went with map labels instead.

6. Conclusions

“Food deserts” are a real issue even in large cities across the United States. Through accessing data and leveraging machine learning capabilities, I was able to draw conclusions on which areas seem to be impacted mostly by lack of food access, and how one may categorize these areas based on census economic indicators. I suggested a few ideas to improve food access such as opening smaller cost-saving stores in some lower income areas or using internet-enabled services in others.

7. References

1. www.stateofhoustonhealth.com
2. www.public.opendatasoft.com
3. www.developer.foursquare.com/docs/build-with-foursquare/categories/
4. www.communityimpact.com (Picture Credit)