Food Access in Houston, TX, USA

Applied Data Science Capstone Project

Adonis Ichim

June 2020

Problem Statement

- City of Houston, Texas, USA, is the fourth most populous US city, with an estimated 2019 population of 2,320,268 and even higher when considering the entire metropole.
- 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.

Problem Statement

- 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 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.

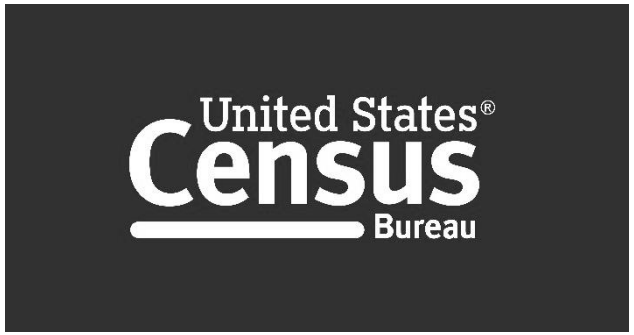
Data

Foursquare API

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

Census Data

- income
- living cost
- internet and vehicle availability



Census Data

- Downloaded as separate files from houstonstateofhealth.com
- Data loaded as streaming body
- Cleaning process:
  - 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.

	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

Census Data – Postal Codes

- Downloaded the Texas Postal Codes Coordinates from the public dataset stored at public.opendatasoft.com into a dataframe.

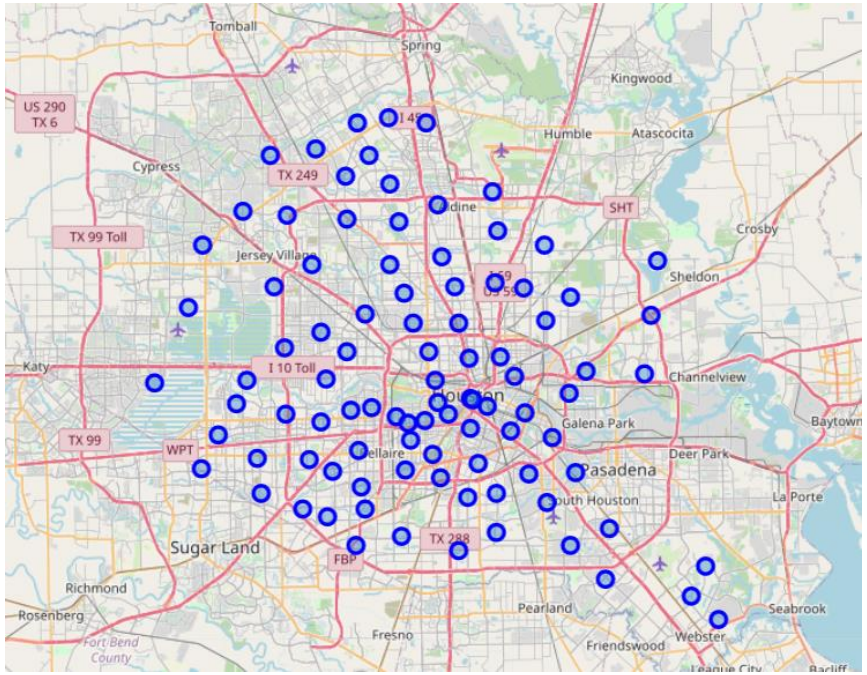
Postal Code	Latitude	Longitude
77046	29.733181	-95.431310
77015	29.778526	-95.181180
77289	29.833990	-95.434241
77072	29.700898	-95.590020
77216	29.833990	-95.434241

- Census data and postal code center coordinates were merged.

Postal Code	Median Rent ($)	Households wihout vehicle (%)	Households with internet (%)	Per Capita Income ($)	Households below poverty level (%)	Homeownership (%)	Latitude	Longitude
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

Geospatial Data

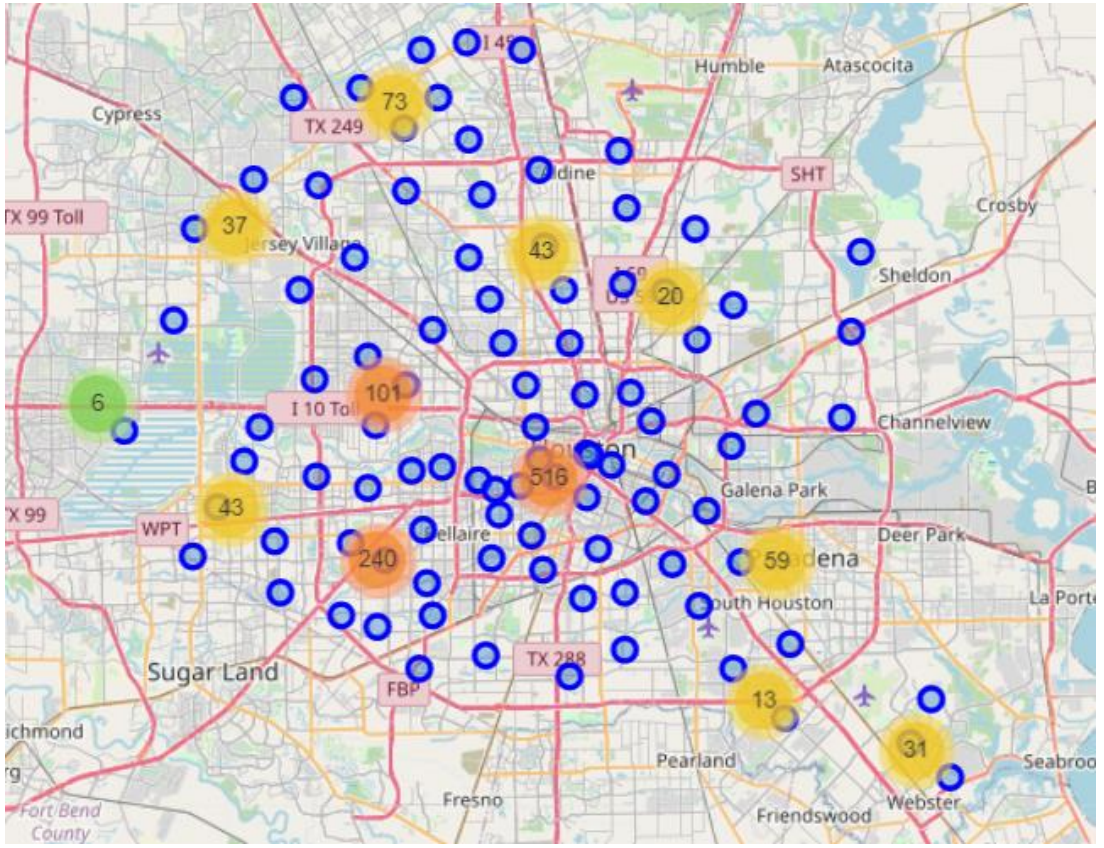
- City coordinates were obtained via GeoPy Geocoders Nominatim as (29.7589382, -95.3676974)
- I initialized a map and plotted with blue circles the Postal Code centers.



Geospatial Data

- I used the Foursquare API to retrieve data on stores selling food items in Houston.
- Wrote a function that uses the Foursquare Category ID as input.
  - IDs are accessible at 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, worked 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
- The GET request returned a dataframe with 1,182 rows and 7 columns. After removing duplicate coordinate pairs, 1,042 entries were left.

Geospatial Data



Exploratory Statistics

- 95 Postal Codes analyzed:
  - the mean of the Median Rent is around 1,035 USD
  - 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,
  - 75.82% have an internet connection.
- Venues dataset:
  - 1042 entries, 794 venues labelled as Grocery Store, 136 as Supermarket, and 30 as Big Box Store

Machine Learning – K Means Clustering

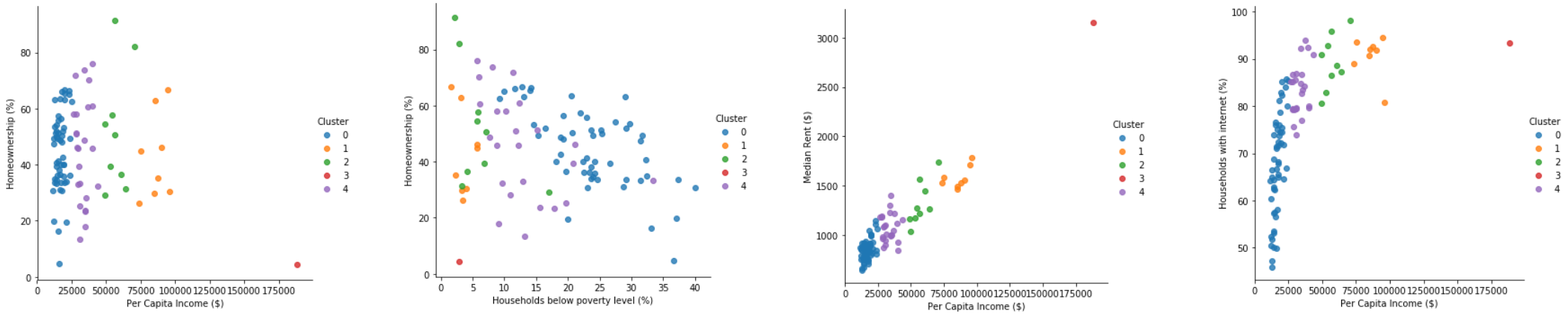
- Used K-Means Clustering to partition census dataset into K pre-defined distinct and non-overlapping subgroups and gain insights about the population of each postal code.
- First, normalized the Census dataset to help algorithms interpret features with different magnitudes and distributions equally.
- Used StandardScaler to normalize the dataset.
- Tried out multiple number of clusters and found best results with five clusters.
- Assigned the cluster label to each data row and added it to the label on the map.

Results

- 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.

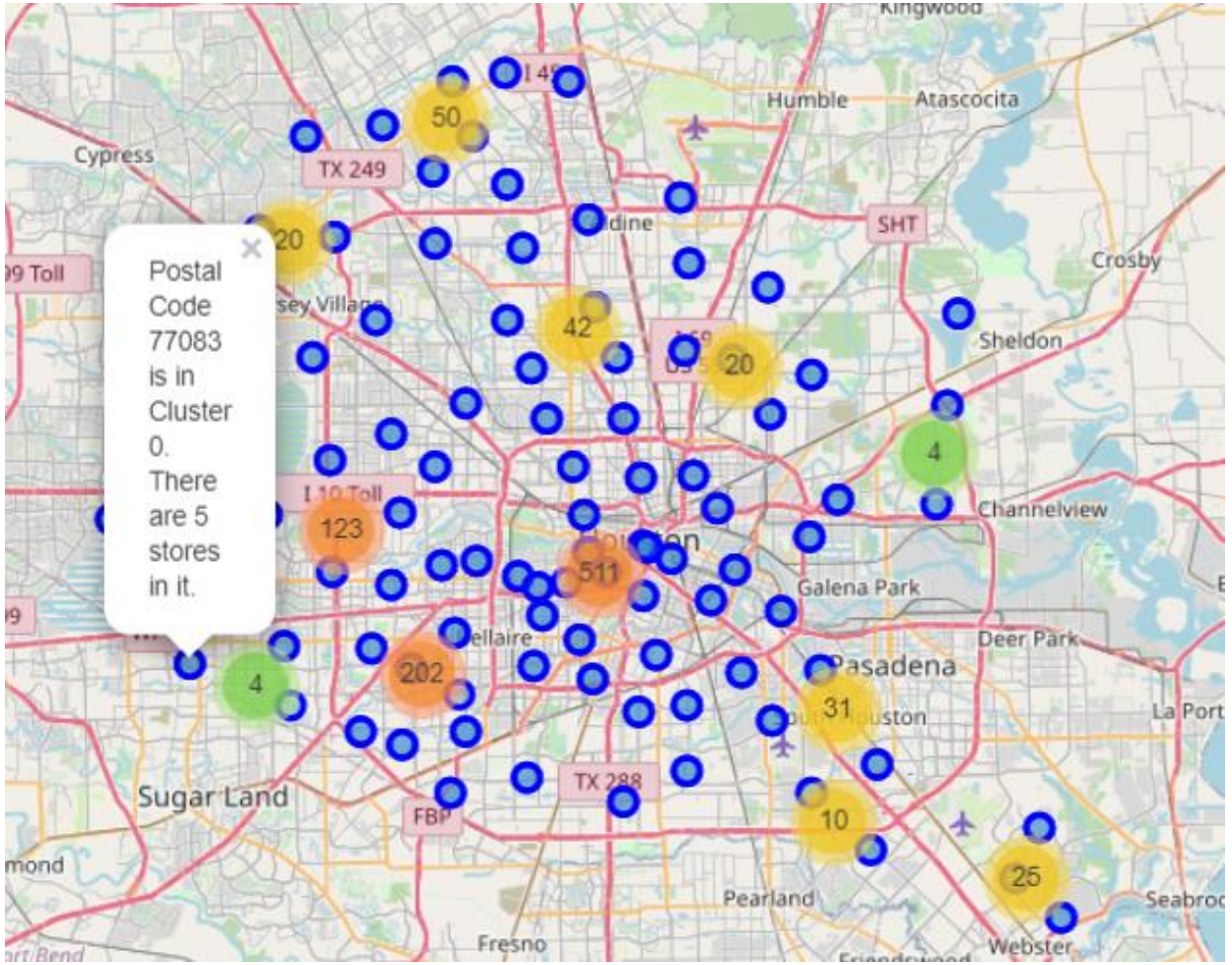
Results

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



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

Main Deliverable



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.
- One may consider investing in **cost-accessible grocery stores** 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.

Discussion

- 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.

Future Work

- This analysis would benefit from the inclusion of additional census data and organizing it by Super Neighborhoods instead of Postal Codes.
- “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.
- Additional types of venues can be added. Data should be crosschecked with an official city source to make sure businesses are active.
- A flooding risk map could be implemented to better select a business location.
- 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.

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,
  - how one may categorize these areas based on census economic indicators.
- I suggested concrete ideas to improve food access such as opening smaller cost-saving stores in some lower income areas or using internet-enabled services in others.