Predicting the Potential for a New Health Food Venue

Gordon Boulger
June 6th, 2020

Introduction

The Business Problem: One of the biggest challenges for any restaurant is finding the right location. As they say in Real Estate, the three most important things are location, location, location. This would be especially true for healthy food alternatives to fast food that are more of a niche product with a more limited clientele than traditional fast food. The Business Problem I will be solving is to identify high-potential areas in Houston for opening healthy food alternatives to fast food.

Who would be interested: The hypothetical interested party would be an investor looking to open a restaurant that is a healthy food alternative to fast food.

Data Sources and Uses

I will use Starbuck locations in the Houston inner loop area from FourSquare as starting point because Starbucks create traffic, Starbucks corporate picks good locations for their stores, and Starbucks clientele are affluent, health conscious and ecologically aware (from the Motley Fool 6/27/2018).

Houston neighborhood census tract and land use GIS data from Houston MyCity web site combined with Starbucks data to determine high potential locations.

Data on Health food venues is from FourSquare web page of Venue Categories:
Juice Bar
4bf58dd8d48988d112941735
Health Food Store
50aa9e744b90af0d42d5de0e
Poke Place
5bae9231bedf3950379f89d4
Smoothie Shop
52f2ab2ebcbc57f1066b8b41
Vegetarian / Vegan Restaurant
4bf58dd8d48988d1d3941735
I will use this FourSquare web page of Chain Codes:
Starbucks 556f676fbd6a75a99038d8ec

The data will be used to identify high potential Houston areas and then determine the availability of health food related venues near them from FourSquare. The lower the availability of health food related venues in a high potential area, the higher the potential for a new Health food venue in that area.

Data Acquisition and Cleaning

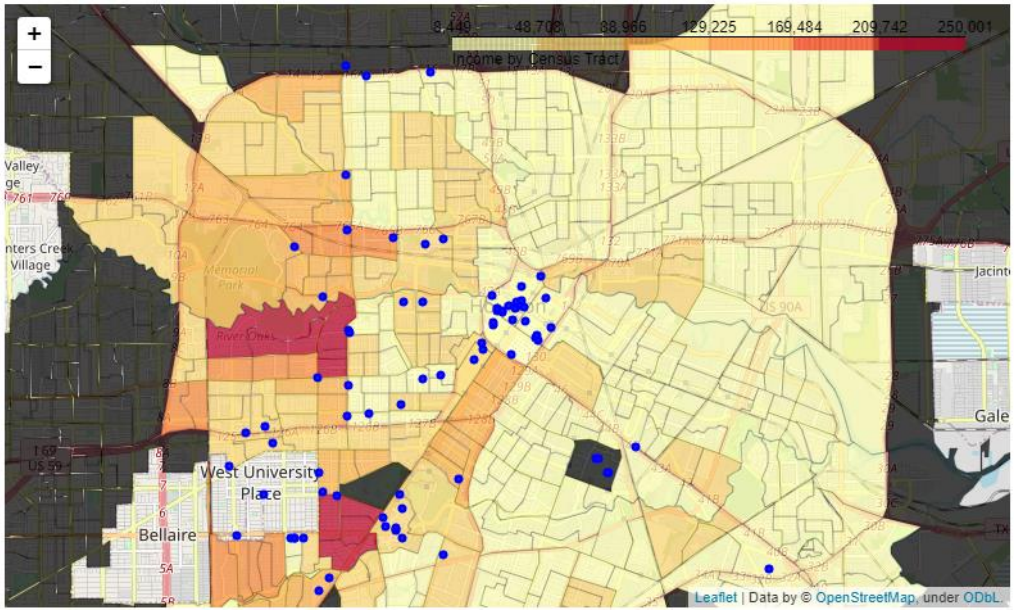
Data was downloaded from the Houston MyCity site for:

- Census Tract demographics and geo data for all of Houston and filtered to limit to just the Houston Inner Loop.
- Land Use data is a 2 gigabyte file that was downloaded with 1.6 million records and filtered to be in just the inner loop. Then the over 180,000 land uses were located within specific census tracts and summarized for total acres by use type by census tract.

FourSquare was used to get data for:

- Starbucks in the inner loop area. Where the queries were limited to 50 venues, it was necessary to break the inner loop into 4 quadrants to get all available Starbucks locations. Then the Starbucks locations were located within specific tracts.
- Health Food venues in the inner loop area. Queries were run for Smoothies and Juice and all other Health Food related queries were found to be redundant with those two. Then the Health Food venues were located within specific census tracts.

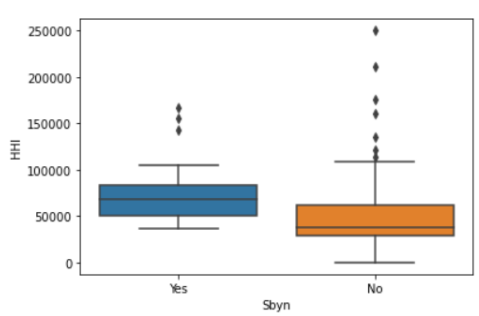
Here is a Folium Choropleth map showing Household Income by Census Tract in the Houston Inner loop with the Starbucks location overlaid.

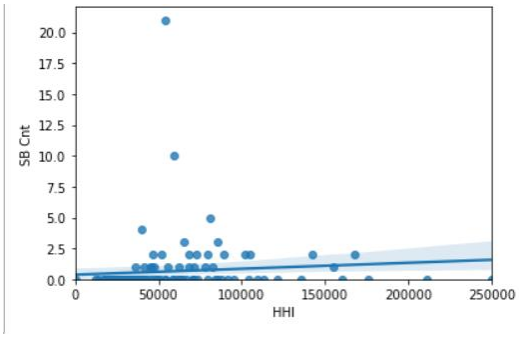


Exploratory Data Analysis

Relationship between Starbucks and Household Income

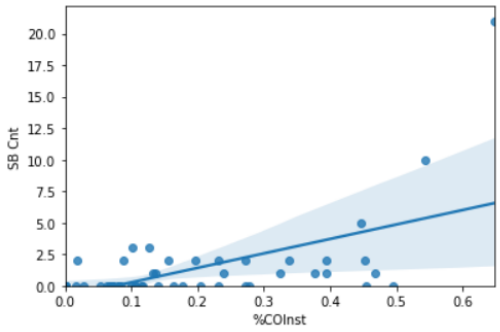
A topline look at the map indicates some correlation between income and a Starbucks location and when we plot whether a tract has a Starbucks, there is a relationship, but the correlation is not very strong:





Relationship between Starbucks and Land Use

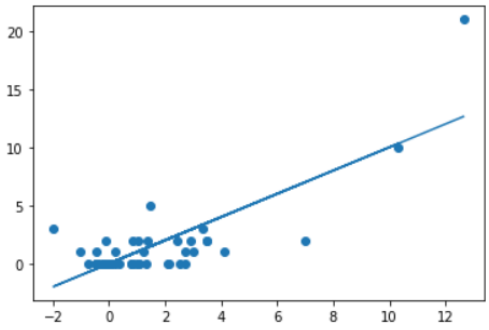
When we limit the tracts to HHI> $50,000 we can see a better correlation with the percent of the land use that is Commercial, Office or Institutional and the correlations by factor:



HCT	-0.578959
SINGLEFAM	-0.230185
HHI	-0.170480
INDUSTRIAL	-0.032001
AGRICULTURE	0.097685
Pop	0.130236
TRANS	0.180907
MULTIFAM	0.215151
PubInst	0.331916
UNDEVELOPED	0.421111
OFFICE	0.562772
%COInst	0.575434
COMMERCIAL	0.619708
COInst	0.736006
SB Cnt	1.000000

Predictive Modeling

Given the above, a Ridge Regression model was developed using tracts with Household Income greater than $50,000 and tract acreage of Office, Commercial, Undeveloped and MultiFamily. The model had a fairly good correlation of .636 and shows that even for Starbucks, there are tracts with opportunities for new venues.

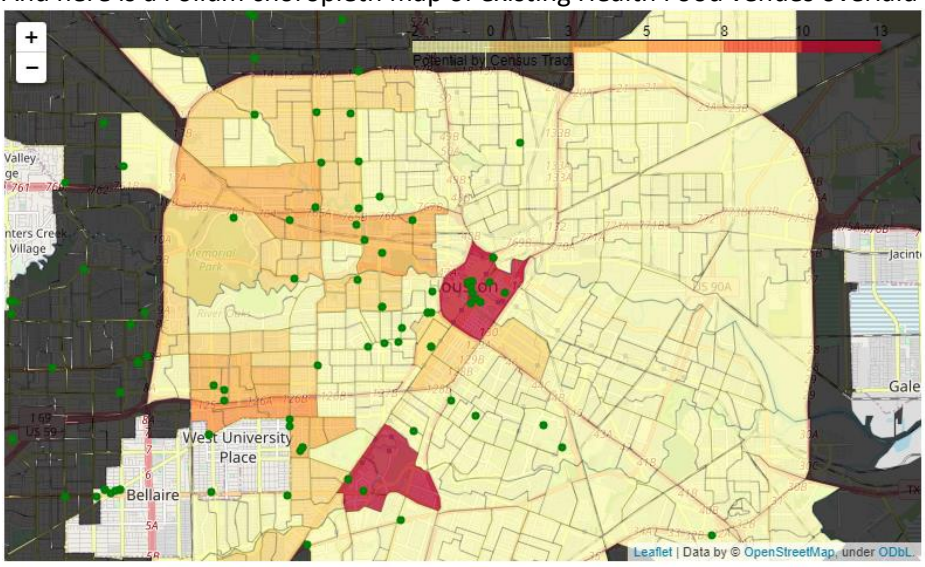


Results

The model was run to give a potential score for each tract in the inner loop and then compared to the number of Health Food venues in those tracts, and sorted by the tracts with the highest differential between potential and Health Food venue counts. Here are the 10 top potential tracts:

	HCT	SBPredicted	HF Cnt	HF Var
53	313100	10.297648	2.0	8.297648
89	411800	6.976731	2.0	4.976731
87	411502	4.107090	0.0	4.107090
47	312500	2.740413	0.0	2.740413
81	411000	2.701852	0.0	2.701852
111	510900	3.338934	1.0	2.338934
84	411300	2.135443	0.0	2.135443
114	511100	2.114070	0.0	2.114070
0	100000	12.640982	11.0	1.640982
108	510600	3.506404	2.0	1.506404

And here is a Folium choropleth map of existing Health Food venues overlaid on tracts by potential:



Conclusion

This study analyzed the locations of Starbucks venues in census tracts in the inner loop of Houston to develop a predictive model that would be useful for selecting locations for a new Health Food restaurant and was able to identify numerous census tracts that have high potential.