

Houston Area Zip Code Commercial Real Estate Analysis

Introduction

This Jupyter notebook will provide Houston zip code analysis to help commercial real estate agents & developers make decisions for their business center projects.

Using Foursquare's crowd-sourced existing venue data grouped by Houston super-neighborhood locations, clustering neighborhoods into quintiles will help provide data-driven guidance for:

- customer base (who is already shopping in the neighborhoods)

- business-types (what businesses are in demand in the neighborhoods)

- locations (where are neighborhoods for associated-type businesses)

Additionally, valuable summary data for the area zip codes helps further explain focused areas within a cluster.

Commercial leases and development are very large capital investments over a long period of time. The combination of these two factors alone greatly affect the success and survival of a business and individual projects. Available business data map analysis helps steer important decisions towards a successful destination.

Data

Houston-region data is available from the City of Houston's Zip Code map json data. The map data includes latitude-longitude perimeter information for all 88 neighborhoods around the city, but zip code GIS point data will be pulled from the "geocoder" ArcGIS module.

Foursquare's API provides crowd-sourced existing business data that can be pulled relative to a given geo-location, radius, and then "exploring" the top 100 venues in that radius. The Foursquare data has businesses segmented among 332 categories such as "coffee shop" and "farm".

Methodology

After all the zip codes and latitude-longitude data are acquired, the data will be organized and wrangled into a clean Pandas dataframe for further analysis. From the dataframe, we first visualize and confirm the zip codes location markers on a Houston region "Folium" map.

Then, using the Foursquare API, we explore the Houston area zip codes and segment them. Initial exploration is of the 1st zip code in the dataframe. The Foursquare API json data returned

for the zip code is for, at most, 100 popular venues in a 500 meter radius around the zip code latitude-longitude point. Following the model from the Foursquare & NY project labs, we use the functions to pull the desired venue detail into a clean venue dataframe.

After sampling the process for one zip code, we can modify additional borrowed functions from the NY lab to pull popular venue data from the Foursquare API for all 213 zip codes in the Houston region, using the same parameters as above:

0 to 100 popular venues in a 500 meter radius around the zip code geocoded point.

This pulls a very large amount of data extracted into a new dataframe for “Houston venues”. “Houston venues” data can then be organized, reviewed, and analyzed. The data is grouped to show how many venues are pulled per zip code, and how many different categories available across the region.

Continuing to follow the processes from the NY lab, we can modify the borrowed one-hot encoding functions to analyze each zip code’s trends and details. Then grouping the zip code venue data by rows, we can see the frequency of the area categories within each respective zip code. To close out categorical review, we get a top 10 list per zip code that provides a general summary picture of each zip code.

Finally, we combine our categorical data, machine learning capability, and mapping to analyze the data. By running the zip code venue data thru *k*-means machine learning to cluster the neighborhood into 5 similar clusters, the clusters can be mapped & examined to determine the venue category factors that distinguish each cluster.

Based on the defining categories, a rank can be assigned for each cluster to help guide different type business projects.

Results

The results of the Houston zip code analysis show a wide variety of venue categories since the zip codes range from downtown urban areas, coastal beach areas, industrial / chemical plant areas, suburban areas, and rural areas.

From the Foursquare API, explore mode, at a 500 meter radius:

1993 venues were pulled across the region

Each zip code pulled between 1 and 89 venues (none maxed out at the 100 venue limit)

226 total categories were pulled across the region

The cluster data shows a majority of active venue zip codes all around the city, which makes sense for the 4th largest city in the US. Other clusters are focused on parks, trails, construction and landscaping, or Mexican restaurants.

The zip code frequency and Top 10 listings are interesting to get more focused zip code data within the 5 clusters.

Discussion

Based on the cluster breakdowns, there seems to be bias based on parks, trails, construction and landscaping, and Mexican restaurants. Regarding parks and trails, Houston is not known for having comparably large number of parks compared to other cities so the geo-code point location and limited 500m radius may have overly influenced these zip code clusters. Regarding Mexican restaurants, Houston has a majority Hispanic population and a very large restaurant base, so I think this should not steer a small cluster as much as it is here.

I think the bias for the 4 smaller clusters is resulting primarily from a limited area used to pull popular venues per zip code. Future analysis can include larger radii as able, since the Foursquare API used is limited transaction-wise. Additional analysis could be done for different zip codes that focus on proximity from the business downtown and shopping district downtown areas.

Conclusion

In conclusion, this notebook is currently a crude general level analysis of zip codes that can be used to analyze zip codes vs commercial development and leasing activity. Further work can be applied to try different factors and data sampling to provide a better general analysis tool. For a more focused project analysis tool, the framework is in place to run deep analysis to a smaller sample of comparable areas of interest.