IBM Applied Data Science

Capstone Project - The Battle of Neighborhoods

Project report

Segmenting and clustering super neighborhoods in Houston, Texas

Samarjit Chakraborty

July 2020

Table of Contents

1. Introduction

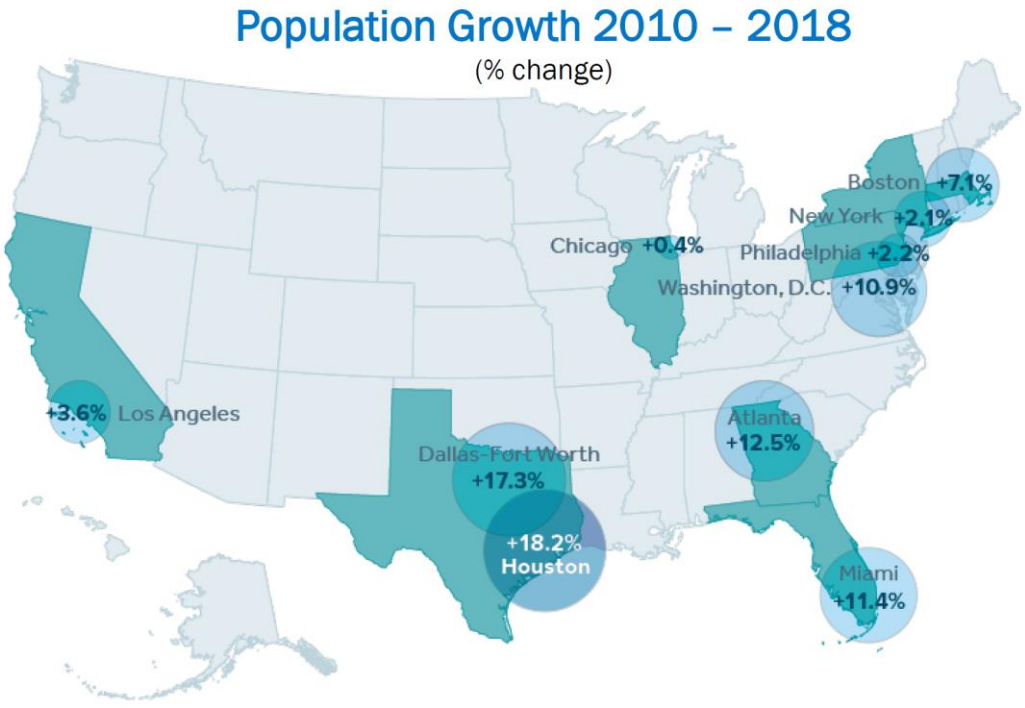
Houston, Texas is the fourth-largest city in the United States. City of Houston has more than 2.3 million residents and covers 634 square miles. Houston's racial or ethnic profile is diverse. Comprehensive, community-based efforts is required to promote healthy living in low-income neighborhoods in Houston. Geospatial analysis of census data with machine learning methods provides decision makers access to tools necessary for resource planning. Most of the open portal data are updated on a regular basis. Therefore, it is important to constantly update the model based on updated data as necessary.

1.1 Background

Houston is one of the youngest, fastest-growing and most diverse populations anywhere in the world. Houston is divided into 88 geographically designated areas, referred to as super neighborhoods. Residents, civic organizations, institutions, and businesses in these super neighborhoods are encouraged to work together to identify, plan, and set priorities to address the needs and concerns of the community. The Houston metro region offers a diverse and extensive labor force of more than three million workers, larger than 35 states. Houston ranks 21st among U.S. metros for venture capital deals.

1.2 Business problem

Every neighborhood in a metropolis should be a neighborhood of promise, hope, and opportunity. However, many of the neighborhoods lack access to quality affordable housing, grocery, schools, and parks. The goal of this project was to analyze data available in public domain to identify zones of opportunity. This in turn would help to attract both practical and innovative investment into underinvested communities while leveraging local and state resources. The super neighborhood elects a council comprised of area residents and stakeholders that serves as a forum to discuss issues and identify and implement priority projects for the area.



Source: https://www.houstontx.gov/

2. Data

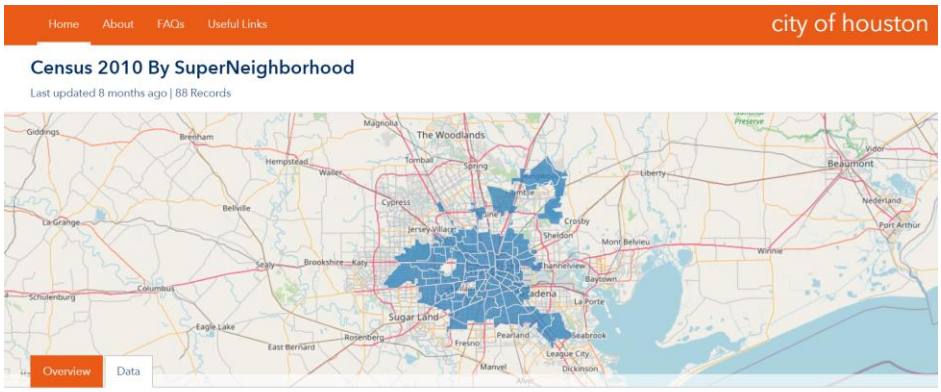
COHGIS stands for City of Houston Geographic Information System. COHGIS dataset is a common GIS dataset published by many departments/business units at the City of Houston. GIS is used by many city departments because it provides decision makers with the tools necessary to answer complex geospatial questions. It integrates spatial and tabular information in a single consistent framework and it provides insight into patterns and spatial relationships within data that might not be obvious outside of a GIS.

2.1 Data sources

2010 census data were obtained, free of charge, from COHGIS GIS Open Data portal - https://cohgis-mycity.opendata.arcgis.com. Both spreadsheet and GeoJSON formats were downloaded from the COHGIS portal.

2.1 Data description

This dataset includes demographic information by super neighborhood. The boundaries of each super neighborhood rely on major physical features (bayous, freeways, etc.) to group together contiguous communities that share common physical characteristics, identity or infrastructure. The Planning and Development Department uses information from the U.S. Census Bureau along with other agencies to develop demographic data and estimates for the City as well as City Council Districts and City Super Neighborhoods. Demographic data includes, but is not limited to: population, housing, and other social characteristics.



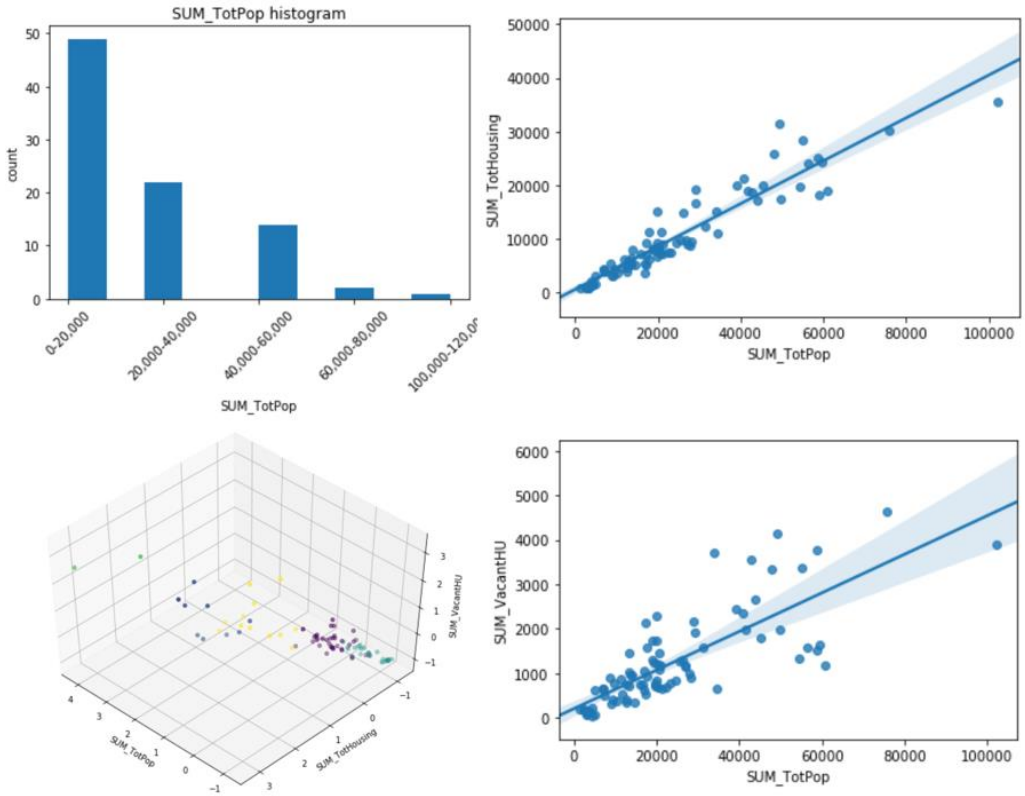
Showing 1 to 10 of 88

Hint: Filter columns using

OBJECTID	SUM_POP100	SUM_HU100	SUM_TotPop	SUM_HispPop	SUM_NonHispPop	SUM_NHOneRace	SUM_NH_White	SUM
1	3881	2104	3881	711	3170	3091	1353	1396
2	13471	5120	13471	10603	2868	2798	1063	1623
3	16716	3664	16716	3266	13450	13347	6092	6562
4	2497	1133	2497	603	1894	1880	15	1853
5	49277	31563	49277	7311	41966	41129	34063	2291
6	33973	15192	33973	15914	18059	17769	4920	12080
7	22317	7411	22317	15338	6979	6806	1824	4364
8	49757	17530	49757	34780	14977	14786	3805	9403
9	54978	28365	54978	12827	42151	40969	19487	12911

3. Methodology

In the Jupyter notebook, geographical coordinates from polygons were converted into their equivalent centroid latitude and longitude values. Also, Foursquare API was used to explore super neighborhoods in Houston. The explore function was then used to get the most common venue categories in each neighborhood, and this feature was then used to group the neighborhoods into clusters. The k-means clustering algorithm was used to complete this task. Census data were also used separately to group the neighborhoods into clusters for initial analysis. Finally, the Folium library was used to visualize the neighborhoods in Houston and their emerging clusters.

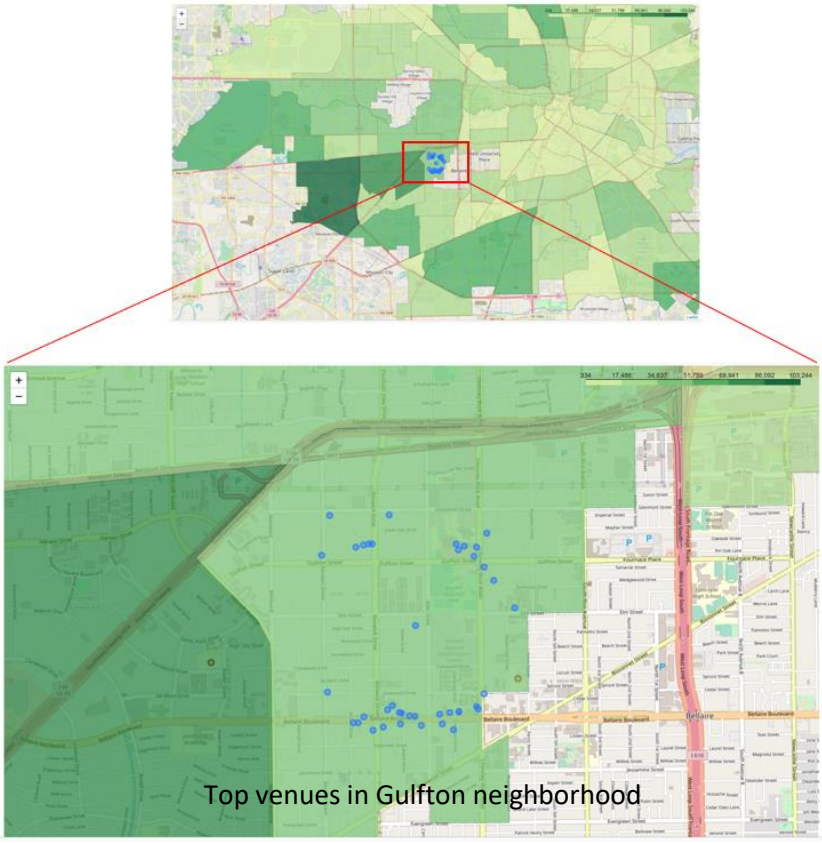


3.1 Download and explore dataset

Data were visualized to find correlation between variables. Histograms were plotted to see the distribution of total population. Scatterplots (with fitted regression lines) were generated using 'regplot' to understand relationship between variables. Basic statistics were computed for all continuous variables.

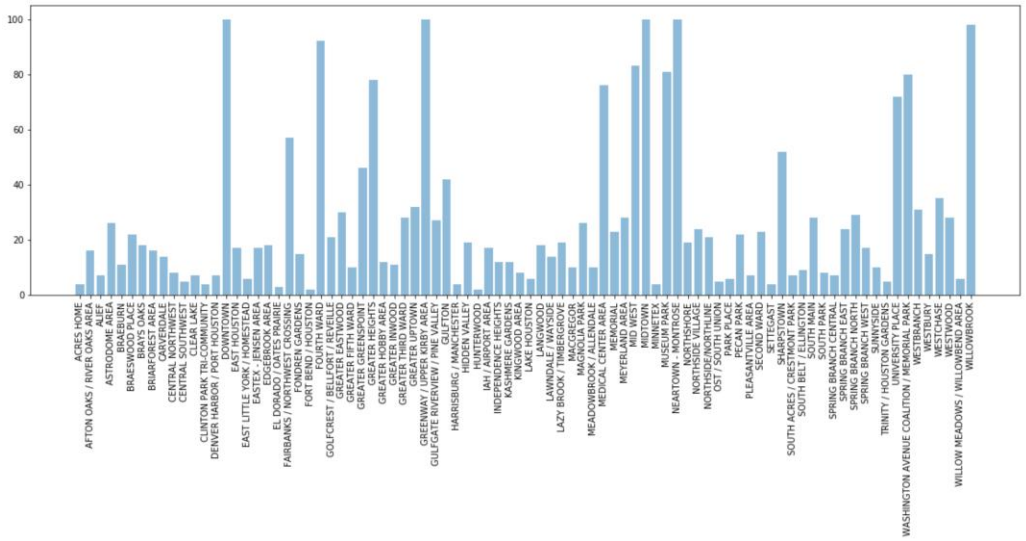
3.2 Explore single neighborhood

For illustration purposes, a single neighborhood was selected from 88 neighborhoods in Houston. Foursquare API was used to get top 100 venues within a radius of 1 km from the neighborhood centroid latitude and longitude values. Python Folium was then used to visualize the venues in the Gulfton neighborhood.



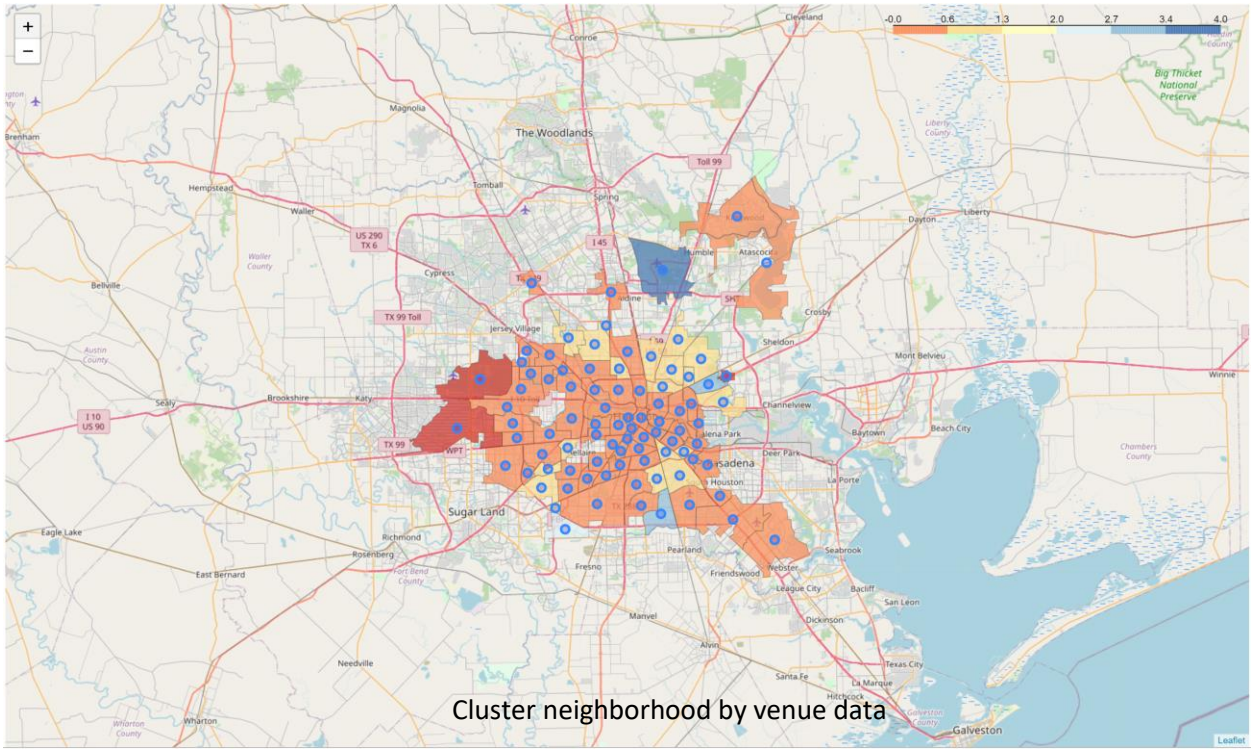
3.3 Analyze each neighborhood

A function was created to extract venue information for all the neighborhoods in Houston. Foursquare API was used to get top 100 venues within a radius of 1 km from the neighborhood centroid latitude and longitude values. A bar plot was generated to check venues returned for each neighborhood. Some neighborhoods (Downtown, Midtown, etc.) reached the venue limit of 100. 301 unique venue categories were found from all the returned venues. A new dataframe was created to display the top 10 venues for each neighborhood.

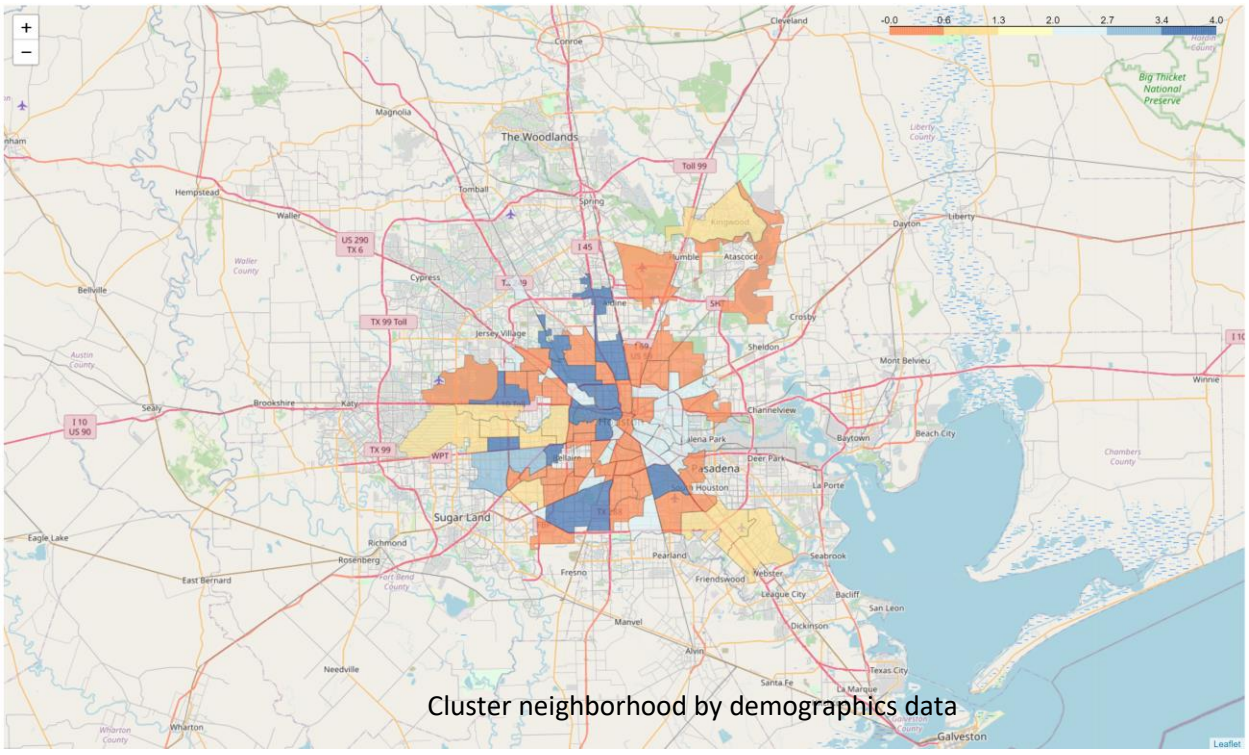


3.4 Cluster neighborhoods

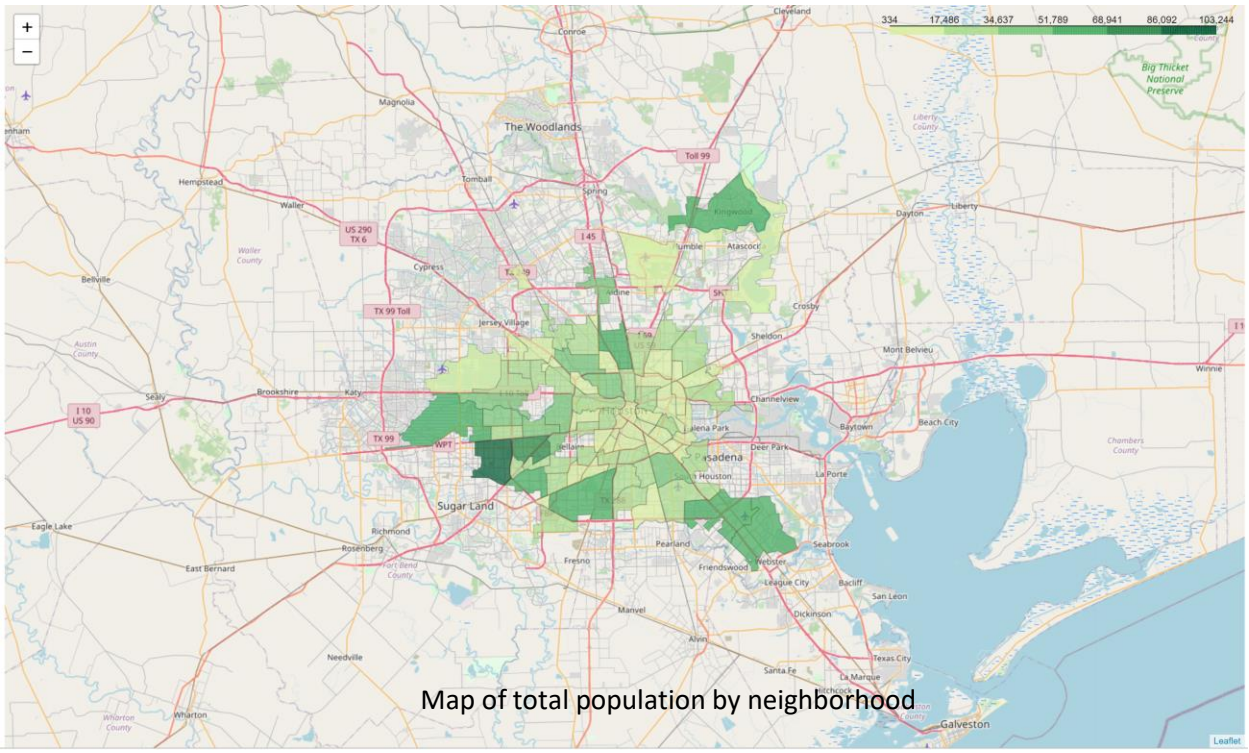
There are many models for clustering. In this project, one of the simplest model K-means was used to cluster demographics and top 10 venues in each neighborhood. Despite its simplicity, the K-means method was useful to quickly discover insights from unlabeled data. In this project K-means was used for neighborhood segmentation. Clusters were then visualized with folium.



Cluster neighborhood by venue data



Cluster neighborhood by demographics data



Map of total population by neighborhood

3.5 Examine clusters

Each cluster was examined to determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, each cluster can be assigned a name as necessary.

4. Results

Descriptive statistical analysis was done to gain insights. Census and venue data were used to cluster neighborhood. In future other data (household, income, etc.) can be visualized along with cluster data for further analysis.

5. Discussions

Datasets were read from the Houston GIS open data portal. Data wrangling was done to convert spreadsheet and GeoJSON data to Pandas dataframes for analysis. Histograms were plotted to see the distribution of total population. Scatterplots (with fitted regression lines) were generated using 'regplot' to understand relationship between variables. Polygon coordinates were converted into their centroid latitude and longitude values. Foursquare API was used to explore neighborhoods in Houston. The explore function was then used to get the most common venue categories in each neighborhood, and this feature was then used to group the neighborhoods into clusters. The k-means clustering algorithm was used to complete this task. Data were normalized so the variable average is 0 and variance is 1 before clustering. Finally, the Folium library was used to visualize the neighborhoods in Houston and their emerging clusters. Clusters obtained from two datasets (demographics and venues) were then compared.

6. Conclusions

The main goal of this project was to analyze data available in public domain to identify zones of opportunity. This in turn would help to attract both practical and innovative investment into underinvested communities while leveraging local and state resources. This information would be useful for area residents and stakeholders that serves as a forum to discuss issues and identify and implement priority projects for the area. Most of the data (census, venue, etc.) are updated on a regular basis. Therefore, it is important to constantly update the underlying model based on newly available data as necessary. In future other data (household, income, etc.) can be visualized along with cluster data for further analysis.

7. References

- Census 2010 By SuperNeighborhood
- SuperNeighborhood
- Foursquare
- Github