IBM Applied Data Science Capstone

Capstone Project - The Battle of Neighborhoods

Project report (week 1)

Segmenting and clustering super neighborhoods in Houston, Texas

Samarjit Chakraborty

June 2019

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 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.

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 - http://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, household, income and other social characteristics.