Determining the lowest priced neighborhood amongst similar Denver neighborhoods

Piyush Mehta

January 16, 2020.

1. Introduction

1.1 Background

Denver is a fast growing city. It was named the best place to live in the United States by U.S. News & World Report.

Denver comprises of 78 neighborhoods. As is to be expected, each neighborhood has its own flavor, with certain neighborhoods exhibiting similar characteristics. Prospective home buyers may benefit with information that identifies neighborhoods with the same set of characteristics and the median prices for houses in neighborhoods with such similar characteristics. The prospective home buyer may then choose the neighborhood with the lowest median price in which to purchase their home.

1.2 Problem

How does a prospective home buyer in Denver determine the least expensive neighborhood to buy a home in amongst neighborhoods with similar characteristics?

1.3 Interest

Clearly, a prospective homebuyer would be very interested in knowing, the neighborhoods that are similar and the median home prices in these neighborhoods. Insight into this information would be one of many inputs that the home buyer will consider when narrowing the neighborhood in which to purchase their home.

2. Data acquisition and cleaning

2.1 Data sources

- Names of the 78 neighborhoods (scraping Wikipedia)
- Latitude and longitude of each neighborhood (using Google Maps API)
- Venue data for the 78 neighborhoods (using FourSquare API)
- Median housing price for each neighborhood (using Zillow API)
- Neighborhood GeoJson data to plot the choropleth map (derived from sources from the Denver Open Data Catalog)

The names of the 78 neighborhoods will be scraped form List of neighborhoods in Denver. From that list of neighborhood names, the Google Maps API will be used to determine the latitude and longitude of each neighborhood. The latitude and longitude information will next be used to get venue data from the FourSquare API. The venue data will be used to cluster the 78 neighborhoods around the venue category types (restaurant, coffee shop, park, etc.). The number of clusters would be determined by evaluating the results of the cluster algorithm. Next, the Zillow APIs will be used to get the median housing prices for the neighborhoods. The clusters will be plotted on a choropleth map to give the home buyer a visual representation of median home prices in clusters of similar neighborhoods. The GeoJson data for the choropleth map will be derived from the shapefile obtained at Statistical Neighborhoods. The shapefile will be loaded in a program such as QGIS and the GeoJson data may then be exported.

Example workflow:
An example neighborhood is 'Central Business District'. We would use this in the Google Maps API 'https://maps.googleapis.com/maps/api/geocode/json?address='. The response would be parsed to determine the latitude and longitude. These in turn would be used as inputs to the FourSquare API 'https://api.foursquare.com/v2/venues/explore' to determine the venues of interest within a certain radius of the location. The venue types will be extracted and the neighborhoods will be clustered around the venue types. The neighborhood name will also be used in the Zillow API 'http://www.zillow.com/webservice/GetDemographics.htm' to retrieve the median home prices in the neighborhood. Finally, the clustered neighborhoods will be visualized against a choropleth map of the neighborhoods based on their median prices.

2.2 Data cleaning
Some of the neighborhood names have special characters in them. The names must be URL encoded prior to invoking the Google Maps API. When retrieving the venue information from the FourSquare API, the radius needs to be tuned to ensure that each neighborhood returns at least one venue result. The Google Maps API and the Zillow API both cue off of the neighborhood name. It is possible however that certain neighborhood names exist in one and not the other. Data from the two APIs must be consolidated. The geographic (GeoJson) data used for the choropleth map is derived from a shapefile representing the neighborhoods. Care must be taken to ensure that this data is consistent with the neighborhood information from other sources.