House Hunting in San Francisco

Luis Macias-Navarro

Using unsupervised learning to cluster homes for sale in San Francisco based on nearby venues and crime statistics.

Applied Data Science Capstone

Introduction

Problem

- Homebuyers find it difficult to compare houses across different neighborhoods based on nearby venues and crime statistics.

Solution

- Use unsupervised learning to cluster these homes on two dimensions:
  1) Venue-based clustering - using venue data from Foursquare
  2) Crime-based clustering - using crime data from SF Data

Target Audience

- Homebuyers that want to make an informed decision by taking into account nearby venues as well as crime statistics for each one of the homes they are considering.

Data

Source	Type of Data	Features used for modeling	Acquisition criteria	Acquisition Method
Redfin www.redfin.com	Real Estate	For each home: • Street address • Geo coordinates	All homes in SF available for sale mid-Feb 2020, with at least 2 bedrooms and 2 bathrooms, selling price in the $1MM range.	CSV file downloaded from Redfin webpage
Foursquare developer.foursquare.com	Venues	For each venue: • Geo coordinates • Venue category • Venue subcategory	All venues within 500m (0.31miles) of each home, limited to 100 venues	API call to Foursquare via IBM Watson
SF Data datasf.org/opendata/	Crime	For each reported crime: • Geo coordinates • Incident category • Incident subcategory	All crimes committed within 500m (0.31miles) of each home during Feb 2020	CSV file downloaded from SF Data webpage

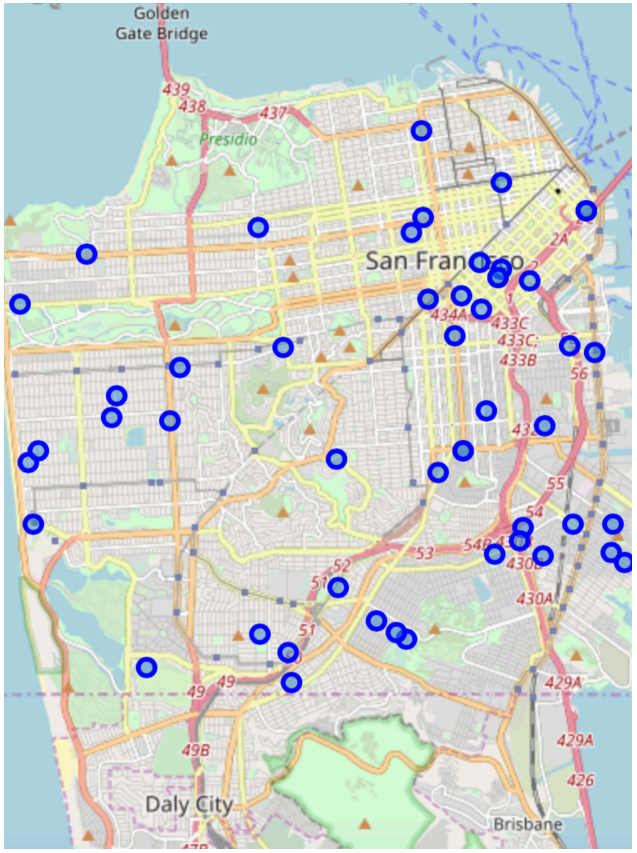
Methodology

Similar methodologies were used for venue-based and crime-based clustering:

1. Import Redfin Data and create map

2. Combine Redfin Data with Foursquare venue data
   - Use *one hot encoding* to convert from category to numerical and establish frequency of venues per home
   - Use K means clustering to obtain clusters of homes that share similar venues

3. Combine Redfin Data with San Francisco crime data
   - Use *one hot encoding* to convert from category to numerical and establish frequency of crimes per home
   - Use K means clustering to obtain clusters of homes that share similar crime incidents



Results: Venue-Based Clustering

Red Cluster – Chinese Restaurants

- easy access to Chinese restaurants and grocery stores. There are other Asian restaurants and pharmacies nearby as well.

Orange Cluster – Art Galleries and Spas

- homes are geographically very close and somewhat isolated
- unique mix of venues available nearby such as art galleries, spas and the marina.

Green Cluster – Baseball and Playgrounds

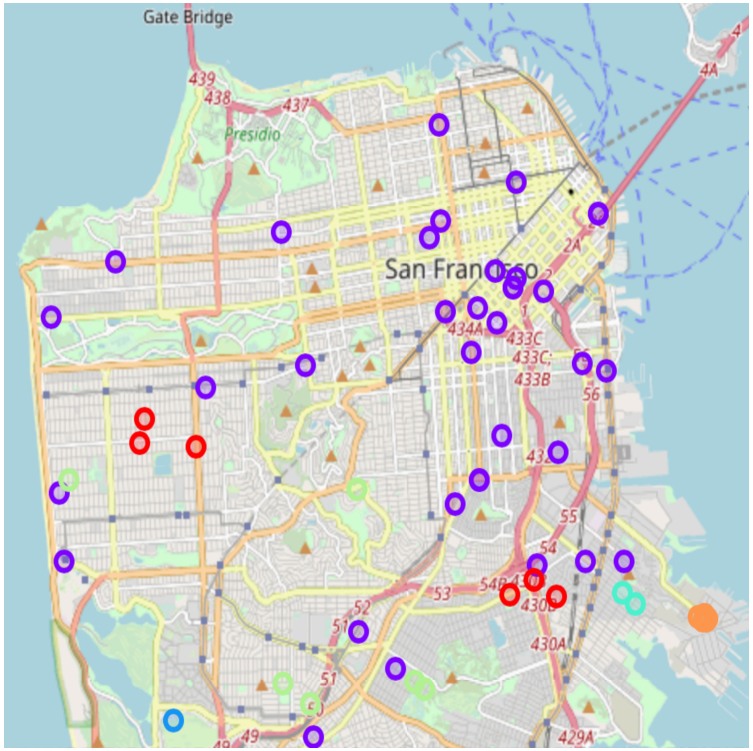
- access to baseball fields and playgrounds as their most common venues.

Purple Cluster – Coffee shops and everything else

- high density of venues nearby
- salient feature: abundance of coffee shops followed by many other types of venues.

Aquamarine and Blue Clusters

- these clusters are very small and boast unique venues such as motorcycle and electronic shops.



Results: Crime-Based Clustering

Red Cluster – Larceny theft from vehicle

- larceny theft from vehicle occurs overwhelmingly
- second most common is Vandalism

Purple cluster - Larceny theft from vehicle mixed with a variety of others

- large amount of Larceny theft from vehicle; it's the most and second most common crime
- number of other types of crimes committed frequently such as Vandalism, Other Larceny, Drug Violations.

Blue cluster - Larceny theft from vehicle and Vandalism

- cluster contains only two homes (at opposite ends of the city) and they have the exact same crime pattern for most common crimes

Green Cluster – Simple Assault

- cluster contains only one home but the crime pattern is different enough that the algorithm gave it its own cluster
- More data needed to know if violent crimes occur often

Orange Cluster - Motor Vehicle Theft

- The most common crime is Motor Vehicle theft followed by Larceny from vehicle

Discussion

- Two large clusters emerged from both venue and crime based clustering (red and purple for both)
  - homes in the purple cluster have a variety of venues to enjoy, however this high level of activity brings with it a high level and variety of criminal activity
  - assumption is that the purple cluster's proximity to notorious neighborhoods like the Tenderloin accounts for this criminal activity
  - homes in red cluster for crime ("Larceny theft from vehicle") appear to be farther away from the central axis of criminal activity

- Orange crime cluster ("Motor vehicle theft") makes sense because these homes are far from epicenter of criminal activity
  - homes in the orange crime cluster overlap somewhat with the green venue cluster ("baseball and playgrounds")
  - These homes would be good for families that enjoy playing outside but do not have a car, or do not mind having it stolen!

- We can see relatively well-delineated patterns across clusters despite crime data being for Feb 2020

- Further study would be required with larger data sets to see if these crime patterns hold. My hunch is that the patterns will largely hold.

Conclusion

- Fifty seven homes that were for sale in San Francisco were clustered based on their proximity to venues and crime incidents reported.
- This project showed that unsupervised learning can be used as a complementary tool for homebuyers to be able to make an informed, data driven decision.
- This tool can be refined and developed to see what type of cluster each home belongs to based on the venues that are nearby as well as crime statistics. This would make comparing homes very easy and convenient.
- Such a tool would be particularly useful in a city like San Francisco where nearby venues are important since it's such a walkable city.
- Unfortunately, any homebuyer must understand what types of crimes he or she would have to potentially deal with on the way to those venues or at home.