



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.



1. Introduction

1.1 Background

Purchasing a home is the largest financial transaction most people will conduct in their lifetimes. What's more, this purchase usually represents the lion's share of an individual or couple's net worth. As a result, such a large and important transaction requires a fair amount of time and due diligence to complete.

Luckily there are tools available online such as Redfin and Zillow that make searching for a home very easy and convenient. These real estate platforms provide data on a home's listing price, days on market, estimated sales price, property history, etc. and can even connect homebuyers with a real estate agent.

However, these are not the best tools when it comes to comparing many homes against one another across any number of features.

1.2 Problem

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

Online real estate platforms such as Redfin provide generic scores based on walkability, transit options and cycling infrastructure. While the walkability score is useful it does not provide any granularity regarding what type of venues are in the vicinity (Figure 1.). Furthermore Redfin does not explicitly provide any information related to crime statistics for a given neighborhood.

Transportation in Ashbury Heights / Parnassus



This area is a **walker's paradise** — daily errands do not require a car. **Transit is excellent** and convenient for most trips. There is **some amount of infrastructure for biking**.

Figure 1. Redfin neighborhood information

The solution is to 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

Having homes clustered based on venues allows buyers to understand whether homes in different parts of the city have access to similar type venues. Crime-based clustering tells them which homes have similar crime rates even if they're in different parts of the city.

1.3 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.

2. Data




2.1 Data Gathering

The basis for this project was data generated from Redfin on 57 homes for sale in San Francisco that met the *Acquisition criteria* displayed in Table 1 below. Each one of those homes had a set of geo coordinates (latitude and longitude) that were then used as centers to which venue and crime data was assigned based on distance. The data was created by exporting from the Redfin website as a CSV file.

Venue data for each of the 57 homes was generated using the Foursquare API. Foursquare returned all the venues (limited to 100) within a 500-meter (0.31 miles) radius of each home. The distance of 500 meters was chosen because it is a reasonable, maximum distance to walk without needing to take public transportation or drive.

Crime data was acquired through the SF Data portal and downloaded into a CSV file; crime incidents reported had a unique set of geo coordinates (latitude and longitude). Similar to the venue-based data, crime data was generated for each home within a 500 meter radius. This was chosen to reflect how safe the immediate vicinity of each home is and what types of crime are perpetrated within walking distance. In addition, and due to the huge volume of incidents reported (yikes), only incidents reported during February of 2020 were included in this project.

Table 1 Summary of data sources

Source	Type of Data	Features used for modeling	Acquisition criteria	Acquisition Method
 Redfin www.redfin.com	Real Estate	For each home: <ul style="list-style-type: none"> • Street address • Geographic 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: <ul style="list-style-type: none"> • Geographic 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 <ul style="list-style-type: none"> • Geographic 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

2.2 Data cleaning

Data downloaded from Redfin required minimal cleaning because the export feature on the website allows for a CSV download with the selected criteria. In the CSV, each home was listed in a separate row and the geo coordinates were right alongside each address in a separate column for latitude and longitude. That was all the data needed from Redfin. Foursquare data required minimal wrangling, once the API was called, to get the data into a dataframe and matched up with the Redfin data.

Crime data from SF Data was generated into a CSV. This data required the most wrangling because the distance between each home and each location where a crime had been reported had to be calculated. This was done in Excel using the Haversine formula:

$$d = 2r \arcsin\left(\sqrt{\text{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1) \cos(\varphi_2) \text{hav}(\lambda_2 - \lambda_1)}\right)$$

Where:

- ϕ_1, ϕ_2 : latitude of point 1 and latitude of point 2 (in radians),
- λ_1, λ_2 : longitude of point 1 and longitude of point 2 (in radians).

After each distance was calculated, only distances ≤ 500 meters between a crime location and a home were taken into account. This resulted in a CSV file that matched up crime locations with each one of the 57 home addresses.