[DATA-01P] San Francisco home sales

Miguel-Angel Canela, IESE Business School

September 18, 2016

Introduction

In the San Francisco Bay area, real estate sales are published in the newspapers once a week. To produce an example for his book, J Adler prepared a data set by compiling information from multiple papers. The data were picked as an example to address some questions about the way real estate data are reported in the media. It is a real data set, so it is not clean, and there are many missing values.

We want to know a little more about real estate prices. For instance, a typical question of interest would be: Is there a premium for bedrooms (above square footage)?

The data set

Real estate sales listings were downloaded from San Francisco Bay area newspapers websites (a spider was used to grab and parse the data). Then, longitude and latitude were obtained for each street from web services. The geographic coordinates allowed downloading neighbourhood data from the Zillow API (`http://www.zillow.com/howto/api/APIOverview.htm`).

The data set (file `frisco.csv`) is related to the two-year historical log corresponding to years 2011 and 2012 (17 months of data). It contains 3,281 observations and variables. Among them:

- The street address for the property (`street`).
- The ZIP code for the property (`zip`).
- The approximate date on which the the sale was recorded (`saledate`).
- The sales price for the property (`price`).
- A count of the number of bedrooms (`bedrooms`).
- The interior space in square feet (`squarefeet`).
- The lot size in square feet (`lotsize`)
- The year in which then property was built (`yearbuilt`).
- An attribute derived from the street, indicating if the address was qualified by a unit number (`condolike`). It indicates the presence of a '#' in the variable `street` .
- The geographic coordinates for the property (`latitude` , `longitude`).

Source: J Adler (2012), *R in a Nutshell*, O'Reilly.