Capstone Project

California Zoning

Coursera IBM Certified Data Scientist

# Business Problem

A department from the municipality of California wants to propose new classifications to various areas within the state. The team wants this classification to be based on housing and venues within each zone. They are not s ure what venue data is available, but they have access to a dataset containing each zone and housing information.

This problem can be converted into a clustering problem. The clusters will be based on the housing features and number of each venue type. The resulting clusters will then be the new zone classification the department will use in the future.

# Data

There are two data sources that will be used to solve his problem. The first data set come from a paper Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297 (<https://github.com/ageron/handson-ml/tree/master/datasets/housing>). This data set contains around 20k zones and their housing information. The features for each zone are:

* longitude
* latitude
* housing median age
* total rooms
* total bedrooms
* population
* households
* median income
* median house value
* ocean proximity

longitude and latitude will be used as key to match the data with the second data set, which comes from Foursquare. The data will consist of the the different types of venues near a set of coordinates.

Once combined together, the two data sets will contain enough information to cluster the zone into related groups based on the housing information and nearby venues.