California Housing Data – ADS1001

This dataset serves as an excellent introduction to implementing data science procedures because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size for experimentation.

The data contains information from the 1990 California census and will be provided to you by the project mentor.

Content

The data pertains to the houses found in a given California district (block) and some summary stats about them based on the 1990 census data. Be warned the data is not cleaned so there are some preprocessing steps required! The columns are as follows, their names are pretty self-explanatory:

- longitude: A measure of how far west a house is; a higher value is farther west
- latitude: A measure of how far north a house is; a higher value is farther north
- housingMedianAge: Median age of a house within a block; a lower number is a newer building
- totalRooms: Total number of rooms within a block
- totalBedrooms: Total number of bedrooms within a block
- population: Total number of people residing within a block
- households: Total number of households, a group of people residing within a home unit, for a block
- medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- medianHouseValue: Median house value for households within a block (measured in US Dollars)
- oceanProximity: Location of the house w.r.t ocean/sea

Project Objectives

The objectives of this project are primarily to

- clean the data and remove outliers,
- undertake some exploratory data analysis,
- investigate correlations between variables,
- investigate the effect of proximity to the ocean,
- create some graphics which illustrate the geographical distribution of house prices and the relation to other variables.