California House Values

Krishna Padayachee

2022-07-19

Housing Values for California

A dataset containing house values for the state of California in 1990 can be obtained from the following site: http://lib.stat.cmu.edu/datasets/ where the zipped file 'houses.zip' may be downloaded and stored in a local folder.

Exploratory Data Analysis

```
   house_value  median_income  ...  latitude  longitude
0     452600.0         8.3252  ...     37.88    -122.23
1     358500.0         8.3014  ...     37.86    -122.22
2     352100.0         7.2574  ...     37.85    -122.24
3     341300.0         5.6431  ...     37.85    -122.25
4     342200.0         3.8462  ...     37.85    -122.25

[5 rows x 9 columns]
```

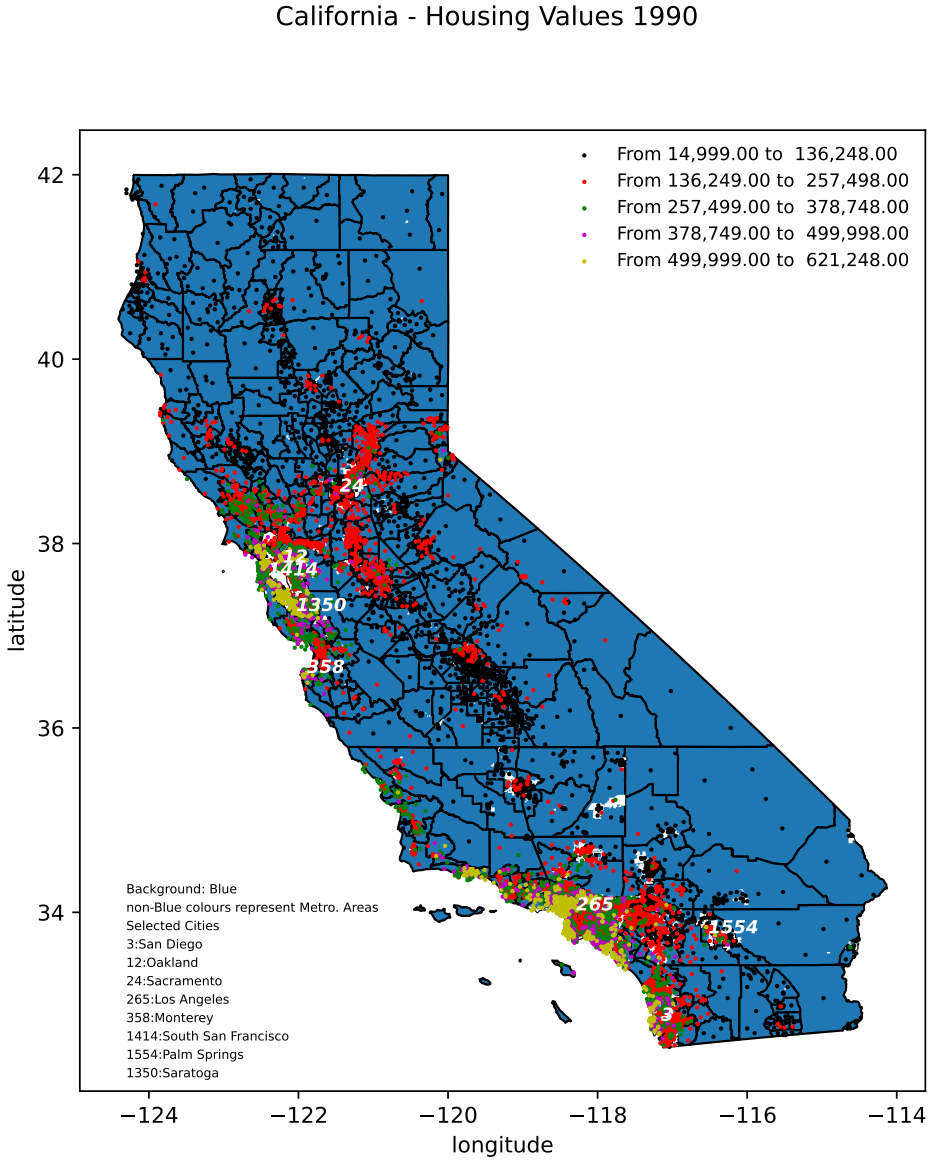
House Values in 1990: the state of California, USA

The python code to generate the map is contained in http://www.github.com/krispad/machine_learning

Comments and where to obtain Shapefiles

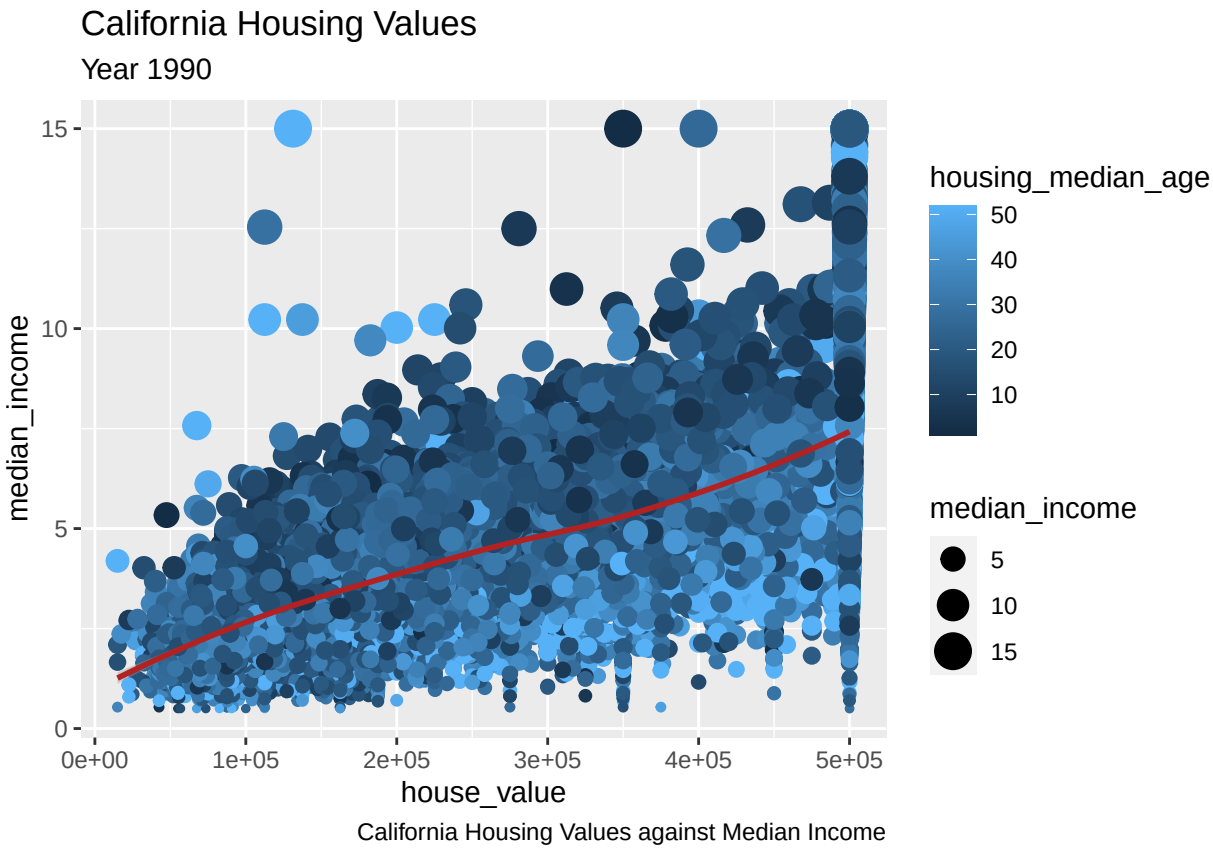
1. Depicting colour-coded areas according to house value ranges
2. *cb_2021_06_cousub_500k.shp*: geodataframe (shapefile) for California Counties - U.S. Geographic Services
   - link: https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html
     - Go to Counties > County Subdivisions and choose the state of California in the (shapefile) rectangular box
   - Unzip the files and place them in a directory and generate the geospatial dataframe
     - e.g. cal _shp2021 = gpd.read_file('~/Documents/'your_path'/cb_2021_06_cousub_500k.shp')
   - Construct a GeoDataFrame from the house_dat dataset
3. The Metropolitan Statistical Area(MSA) for California
   - https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html
   - Go to Metropolitan and Micropolitan Statistical Areas > Places
   - Choose the California shapefile from the rectangular drop down box.
4. Note that **crs** stands for 'coordinate reference system'.

Extracting the Metropolitan Statistical Areas (MSAs) and selecting a subset to be identified on a map of California



Plot of house_value as a function of the median income.

```
`geom_smooth()` using formula 'y ~ x'
```



Transforming the initial Explanatory Variables

The current variables are transformed by applying the function 'houses_cal_mod'. The R code for this function is contained in http://www.github.com/krispad/machine_learning

Arguments for 'houses_cal_mod':

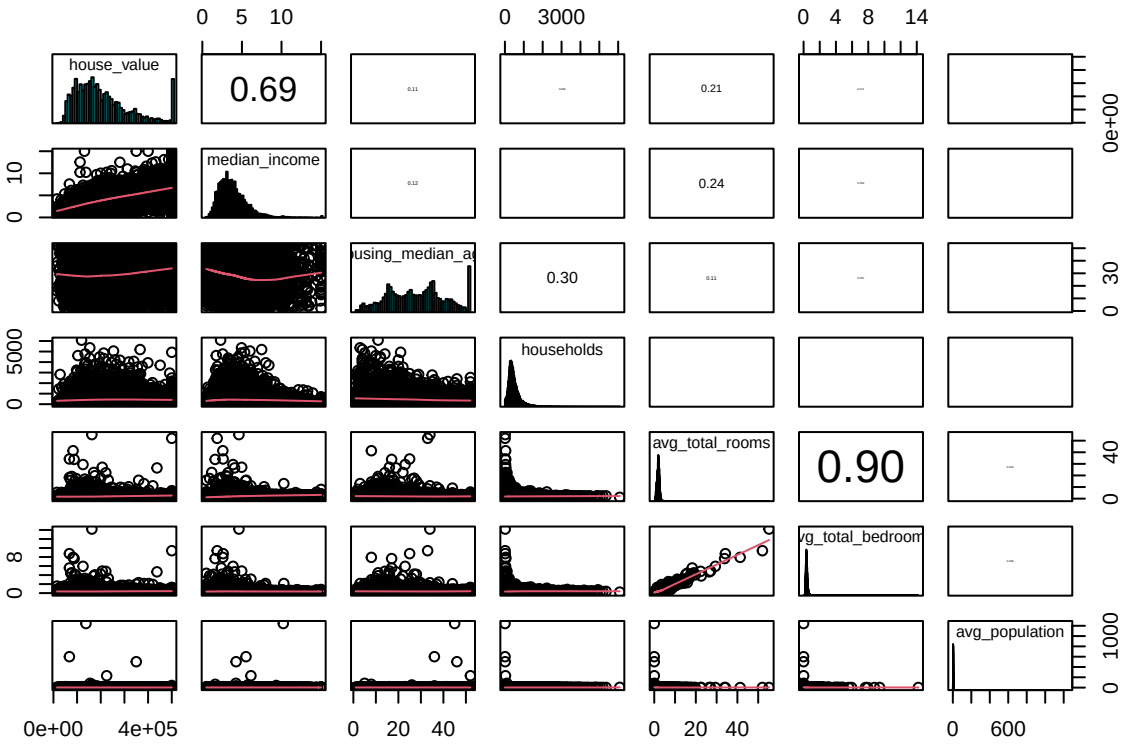
1. *dat* – name of the data set
2. *varnames* - variable names whose values need to be modified, e.g. *total_rooms*, *total_bedrooms*, *population*.
3. *modnames* -names of variables that modify variables in varnames (for the housing data 'modnames' has two entries, *population*, *households*)
4. *first* - a number indicating the initial r elements of varnames that are modified by modnames[1]; in the case of the housing data use 'population'

```
  house_value median_income housing_median_age households latitude longitude
1      452600        8.3252                 41        126    37.88   -122.23
2      358500        8.3014                 21       1138    37.86   -122.22
3      352100        7.2574                 52        177    37.85   -122.24
4      341300        5.6431                 52        219    37.85   -122.25
5      342200        3.8462                 52        259    37.85   -122.25
6      269700        4.0368                 52        193    37.85   -122.25
  avg_total_rooms avg_total_bedrooms avg_population
1        2.732919          0.4006211       2.555556
2        2.956685          0.4606414       2.109842
3        2.957661          0.3830645       2.802260
4        2.283154          0.4211470       2.547945
5        2.879646          0.4955752       2.181467
6        2.225182          0.5157385       2.139896
```

Panel Plots of the variables in the transformed dataset house_dat2.

Note: The panel functions are taken from examples on the R help pages for 'pairs' and developed by the R-Core Team. They are included here for convenience.

1. Histograms on the diagonals
2. Absolute correlations on the upper right diagonals with sizes proportional to the correlations.
3. Scatter plots and 'smoother' fit lines on the the lower triangle of the grid.



Conclusions drawn from the Exploratory Data Analyses:

1. 'house_value' appears to be dependent on 'median_income'
2. 'avg_total_rooms' is highly correlated to 'avg_total_bedrooms'
3. Values of 'housing_median_age' is 'distributed across the range of values for 'house_value'
4. High 'house_value' homes are located on the coastal areas around cities such as Los Angeles, San Francisco, San Diego, Monterey, ...