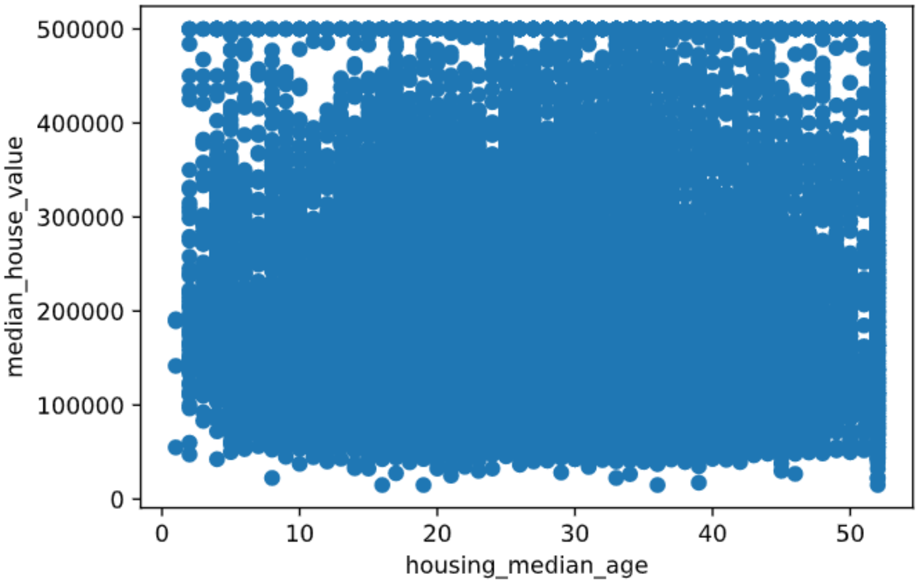
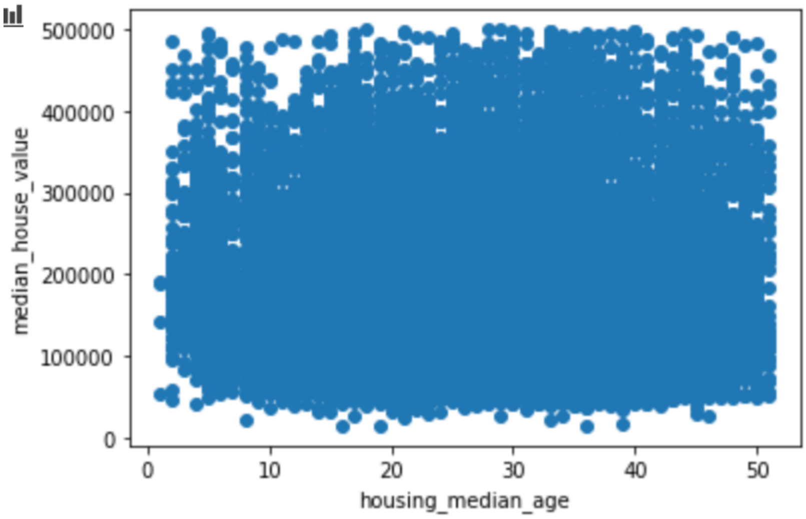
In this experiment, we try to use linear regression and polynomial fitting to process the California housing price data. And we found that the data values have some exceptional distributions at the top or right of plot. So remove them.

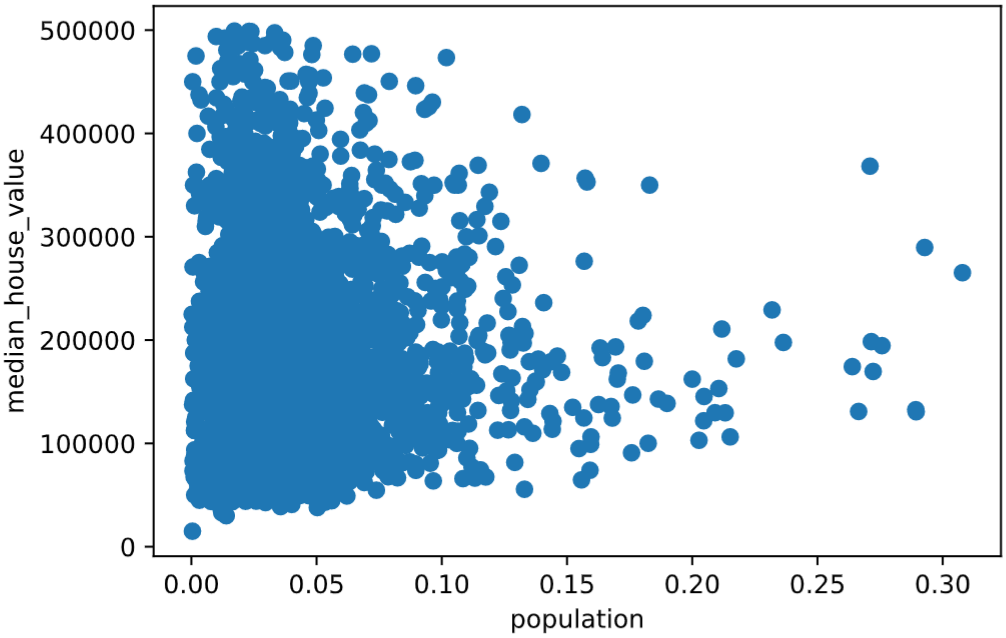


After removing it:



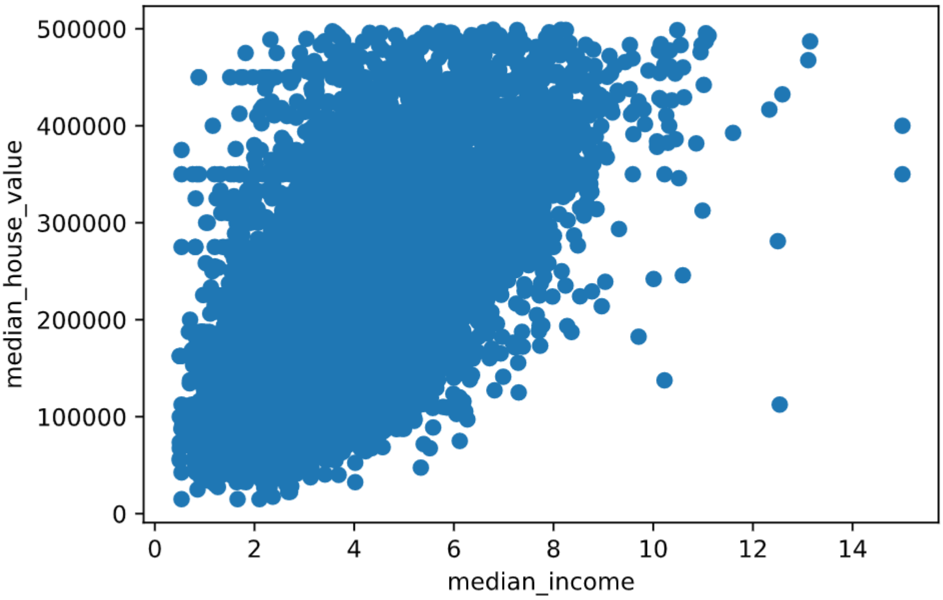
After that, we have the data normalization processing, then use linear regression. Although I think linear regression is more practical, two-dimensional polynomial regression results are better which score is (0.677,0.668) compared to (0.611,0.608) in linear regression. The first data in brackets represents the score of the training set, another respresents the score of the test set.

In fact, we found that linear regression or polynomial regression did not perform well on California data. One possible reason is that most of the data are not closely related to house prices, like as mentioned above median_housing_age, and population and etc.



And because the variance of the data is large and the fluctuation is too large, the score of quadratic polynomial regression is even higher which make no sense.

If we extract the principal component first and the regression, the result may be better, like if we increase the weight of the median_income which is a little more related to median_house_value, the result may be better. The distribution of median_income and median_house_value is shown in the figure below.


median_house_value
median_income