Q12

KO/MC/IKM

5/12/2021

First we load the data, and partition into a train and test set. We then fit our model, tuning over a grid of (up to 1000 trees, for computational complexity)

- interaction_depth: [1, 2, 5, 10]
- shrinkage: [.1, .01, .001]
- bag.fraction: [.25, .5]

```
## Using 998 trees...
```

```
## ************************************
## Test Set Results
## ************************************
```

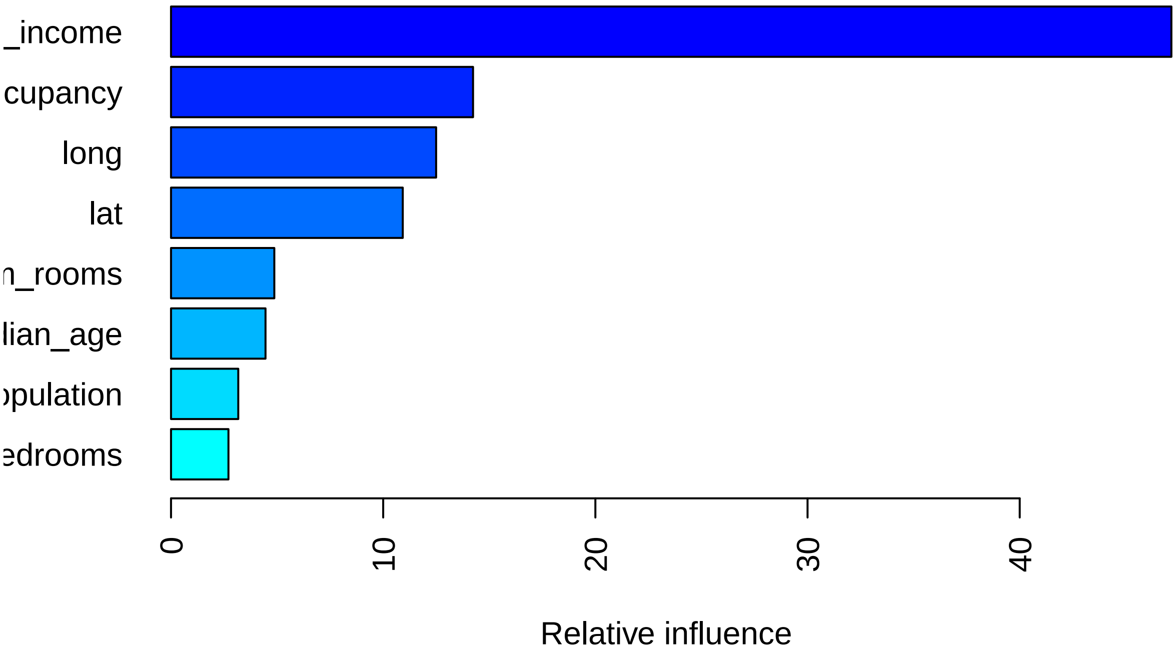
```
## Mean Squared Error:  0.208013
```

```
## Mean Absolute Error:  0.3035144
```

```
## Pearson Correlation:  0.921068
```

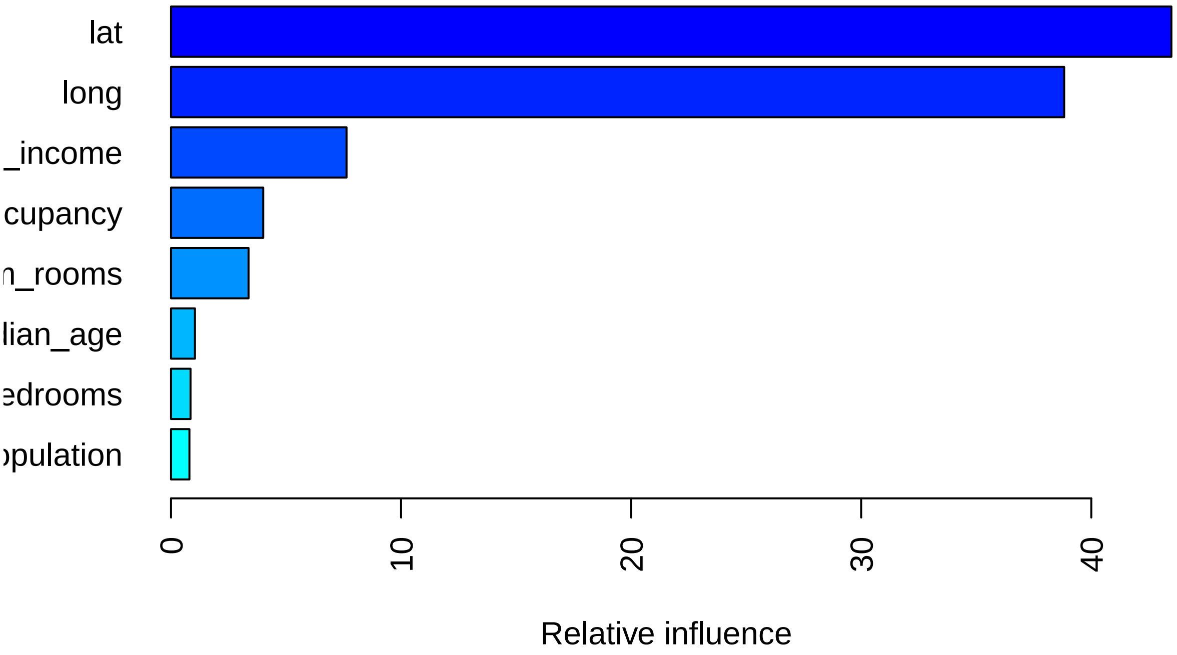
```
## Spearman Correlation:  0.9205144
```

We next plot two measures of variable impotance. First is the relative influence, which follows from average MSE improvement contributed by each split over all the trees/base-learners. According to this measure of importance, median income is far and away the most important predictor. A good distance back and of mild importance are average occupancy and lat/long, while the remaining features are of minor importance according to this importance measure.



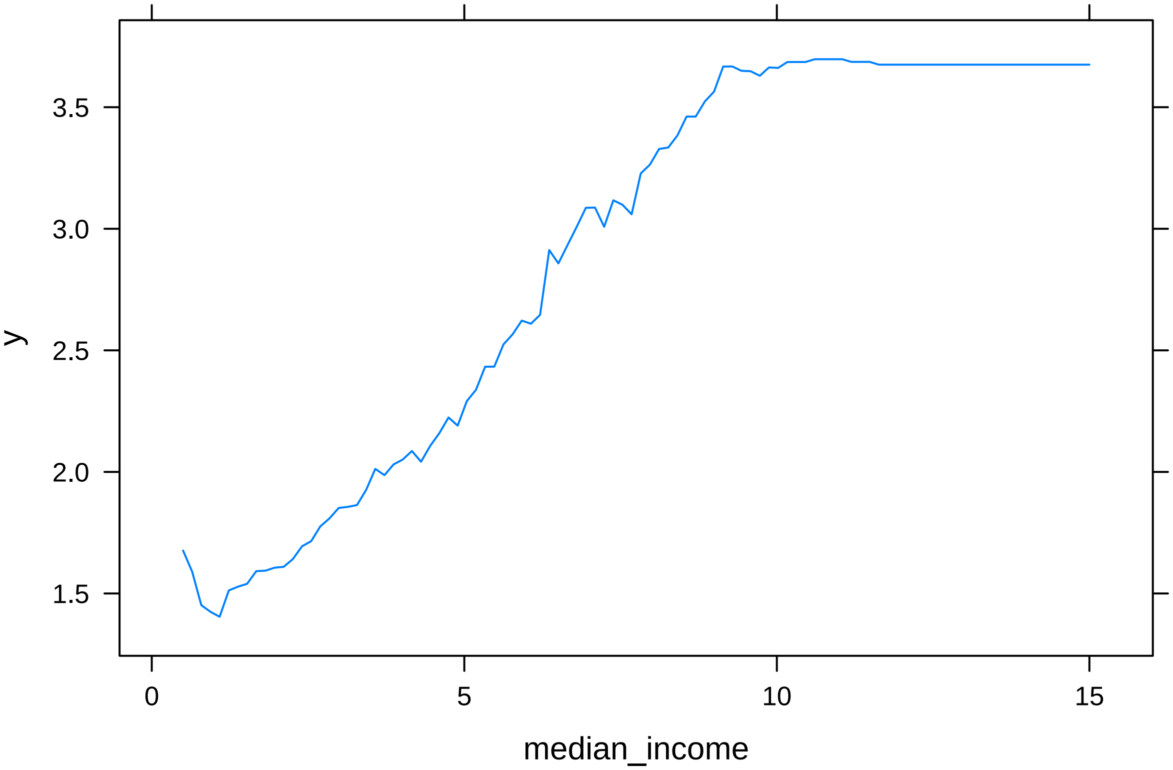
```
##                                     var   rel.inf
## median_income           median_income 47.148222
## avg_occupancy           avg_occupancy 14.234669
## long                             long 12.494728
## lat                               lat 10.922915
## avg_num_rooms           avg_num_rooms  4.869054
## housing_median_age housing_median_age  4.453497
## population                 population  3.169354
## avg_num_bedrooms     avg_num_bedrooms  2.707560
```

Second is the permutation test influence, which shuffles the values of each feature, captures the change in MSE as a result of that shuffle, and then uses that change as a measure of importance (i.e. those variables that when unshuffled, decrease MSE most are most important). According to this measure, lat/long are far and away the most important, with median income, average occupancy, and average number of rooms of lesser but not insignificant importance. Median age, average bedrooms, and population are all of minimal importance.

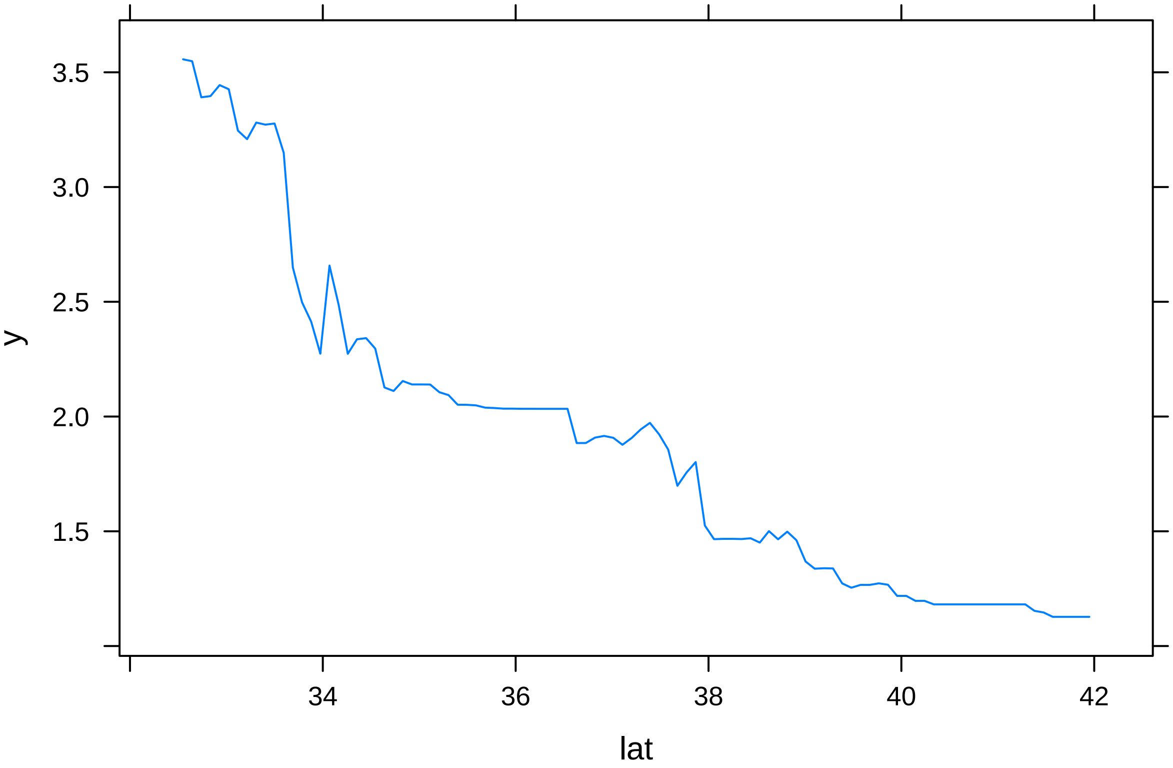


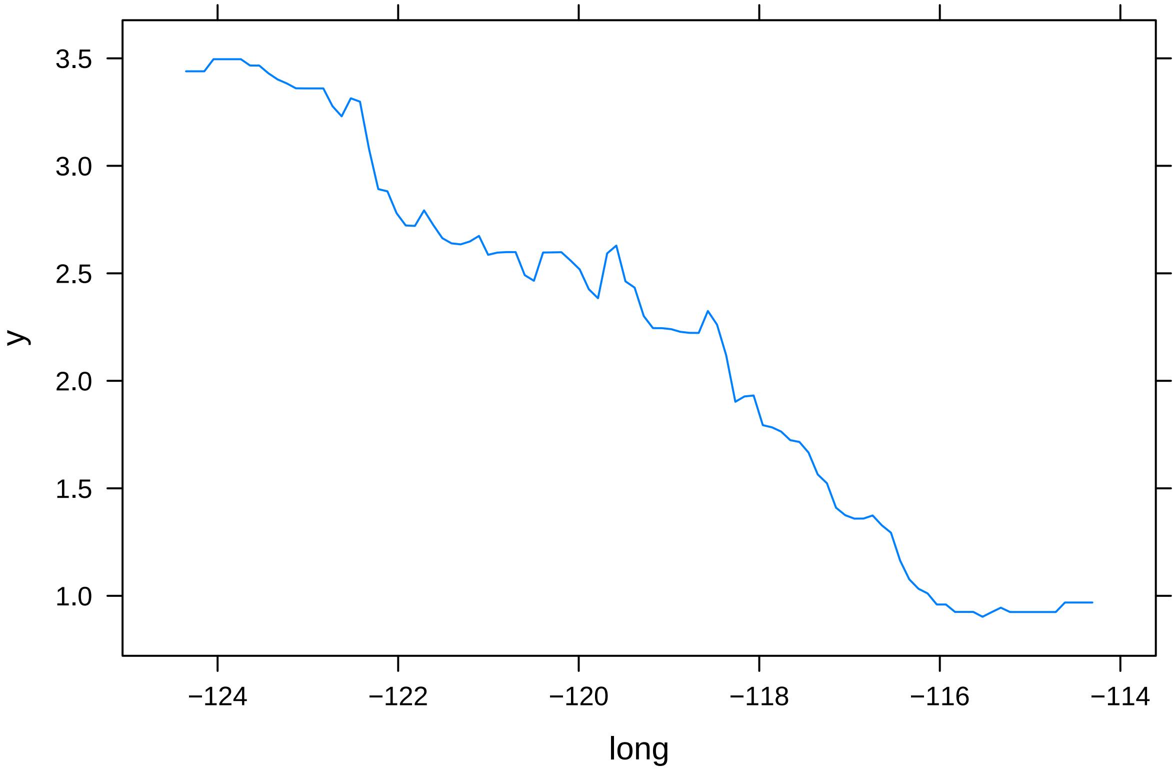
```
##                   var    rel.inf
## 1                 lat 43.4829075
## 2                long 38.8222612
## 3       median_income  7.6290630
## 4       avg_occupancy  4.0085804
## 5       avg_num_rooms  3.3694009
## 6  housing_median_age  1.0403905
## 7    avg_num_bedrooms  0.8472521
## 8          population  0.8001443
```

Next, we plot dependencies. The first batch of partial dependence plots are single variables.

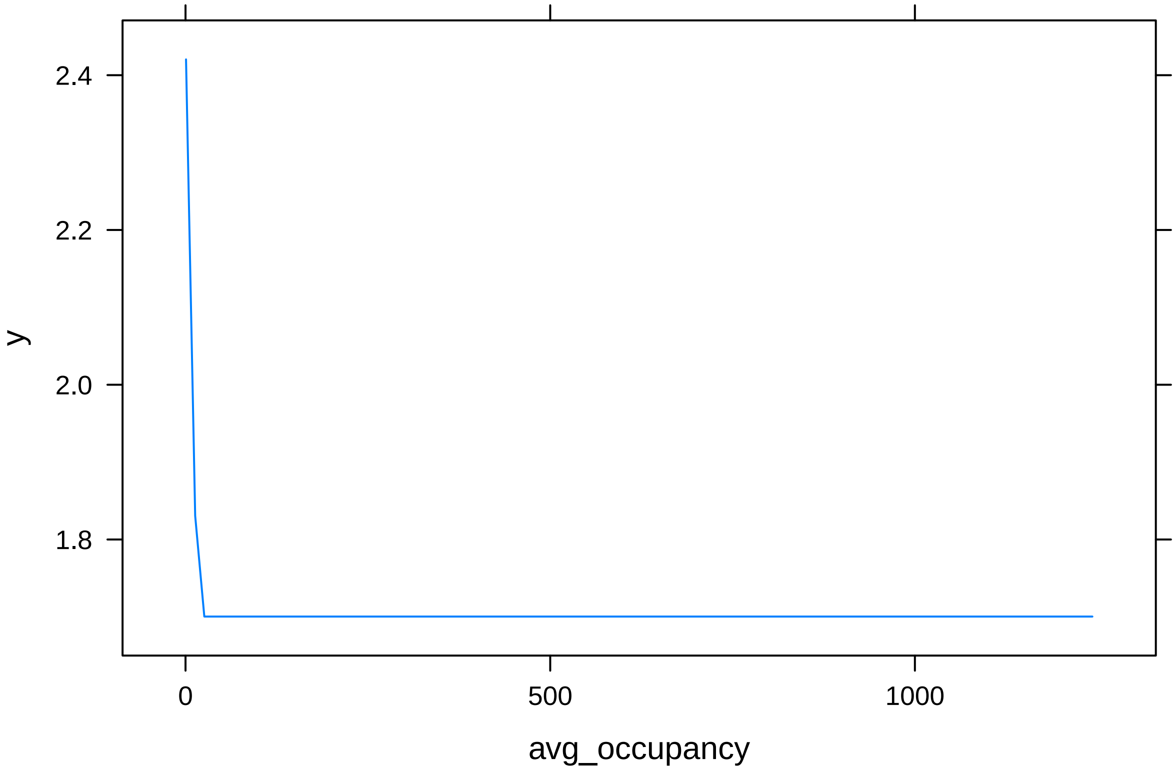


Here, we see the response increase almost linearly with increases in median income, until median income hits approximately 10. At that point, the response levels off, and sits around 3.6. This indicates that median income is helpful to a point, though at some threshold, the respondent is so rich that income itself does not matter and other features are more important.

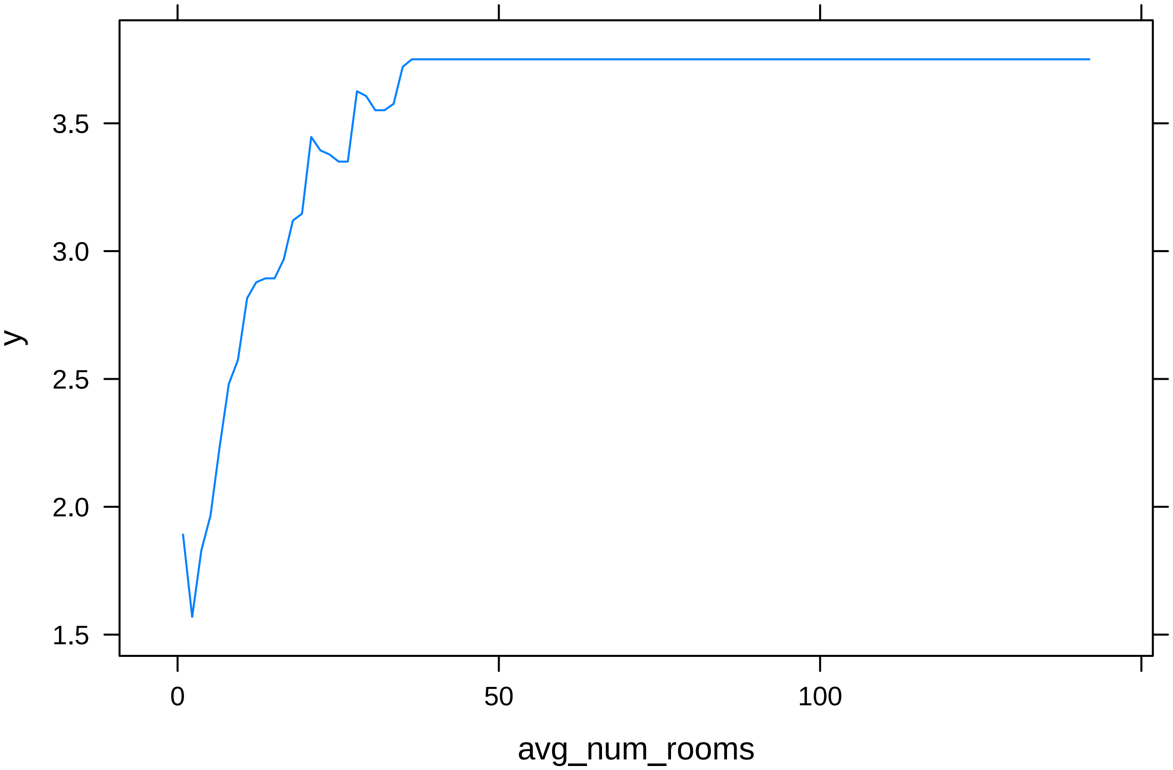




Meanwhile, latitude and longitude show approximately linear decline as both increase. I think longitude has been negated in the dataset (as San Francisco is ~ 37/122, as opposed to 37/-122 here) – this suggests that as you move northeast across the state, median house value declines. This aligns with common sense.

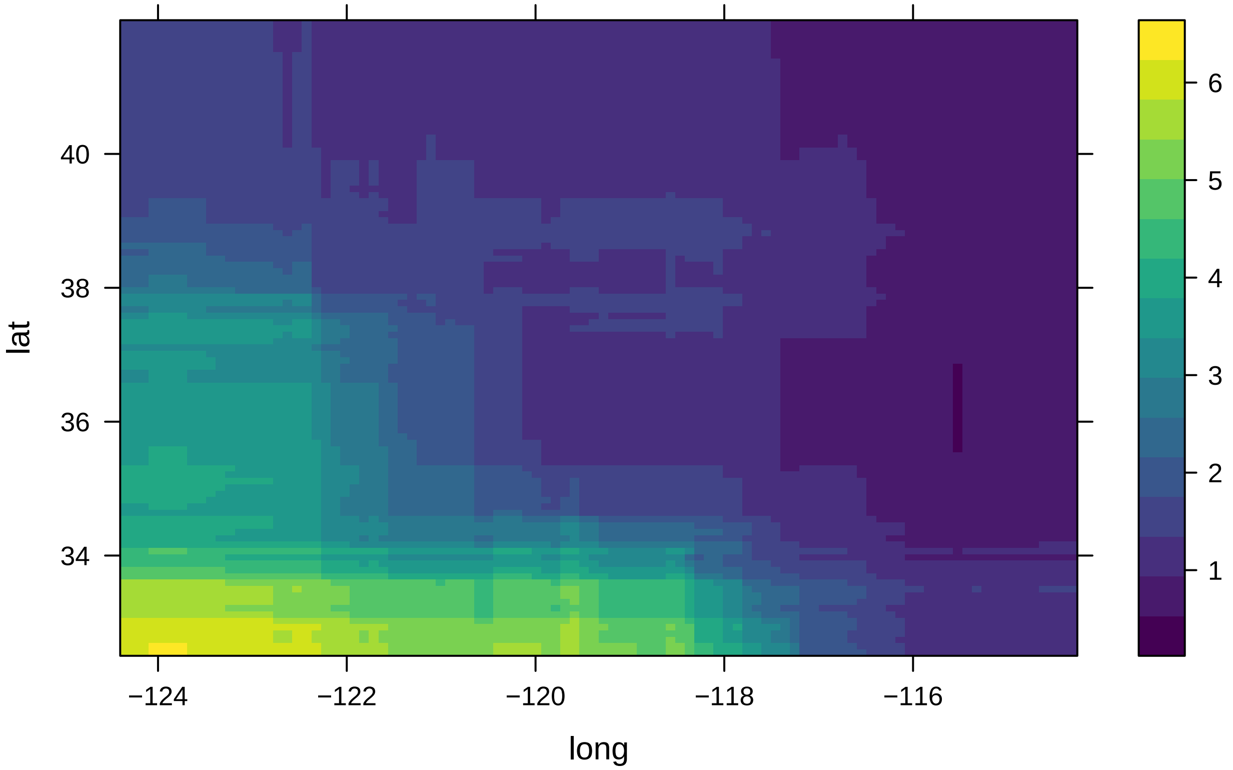


Next, we see a sharp cut on average occupancy. In this way, it almost serves as a thresholding/indicator function, wherein zero occupancy corresponds to greater home value (perhaps more expensive apartments in “desirable”" cities like SF, LA, or SD), followed by a sharp drop as occupancy increases (perhaps more suburban/inland homes have lower values on account of location, but more space?).

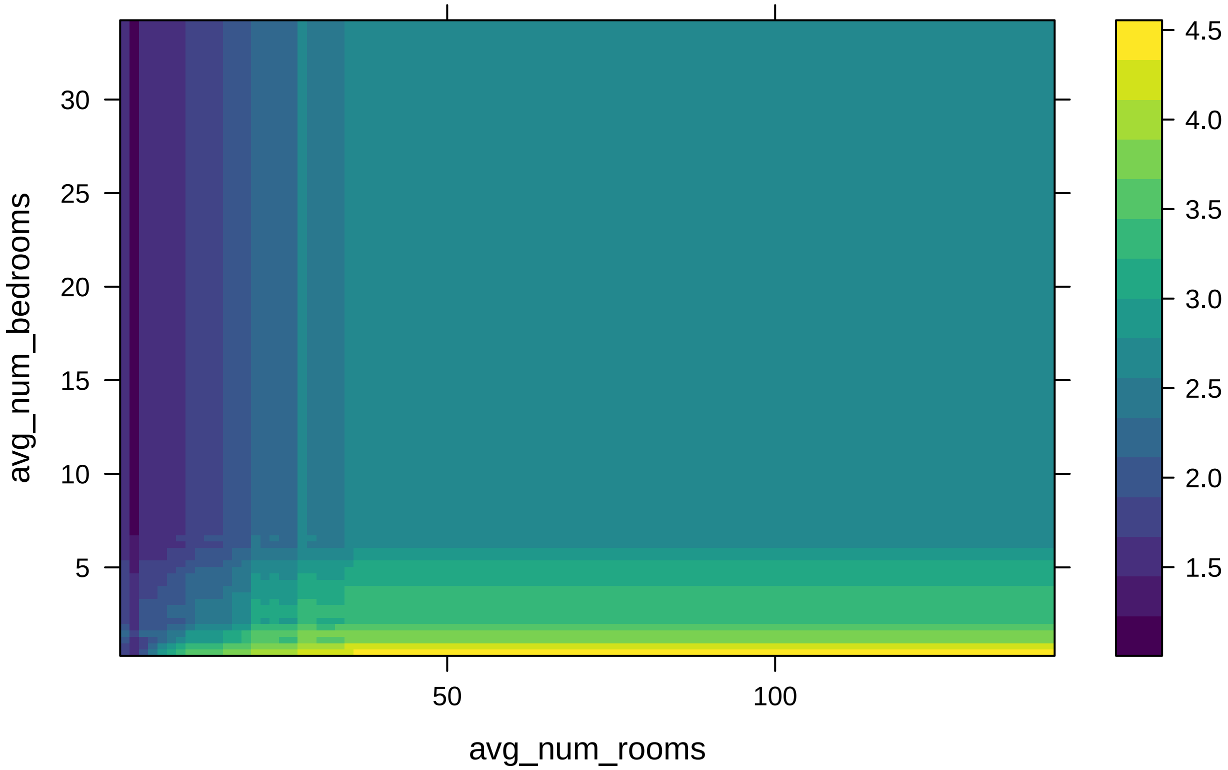


And for rooms, the more rooms you have, the more mansion-like the house probably is, so as expected median value increases (up to a certain point, where it levels off).

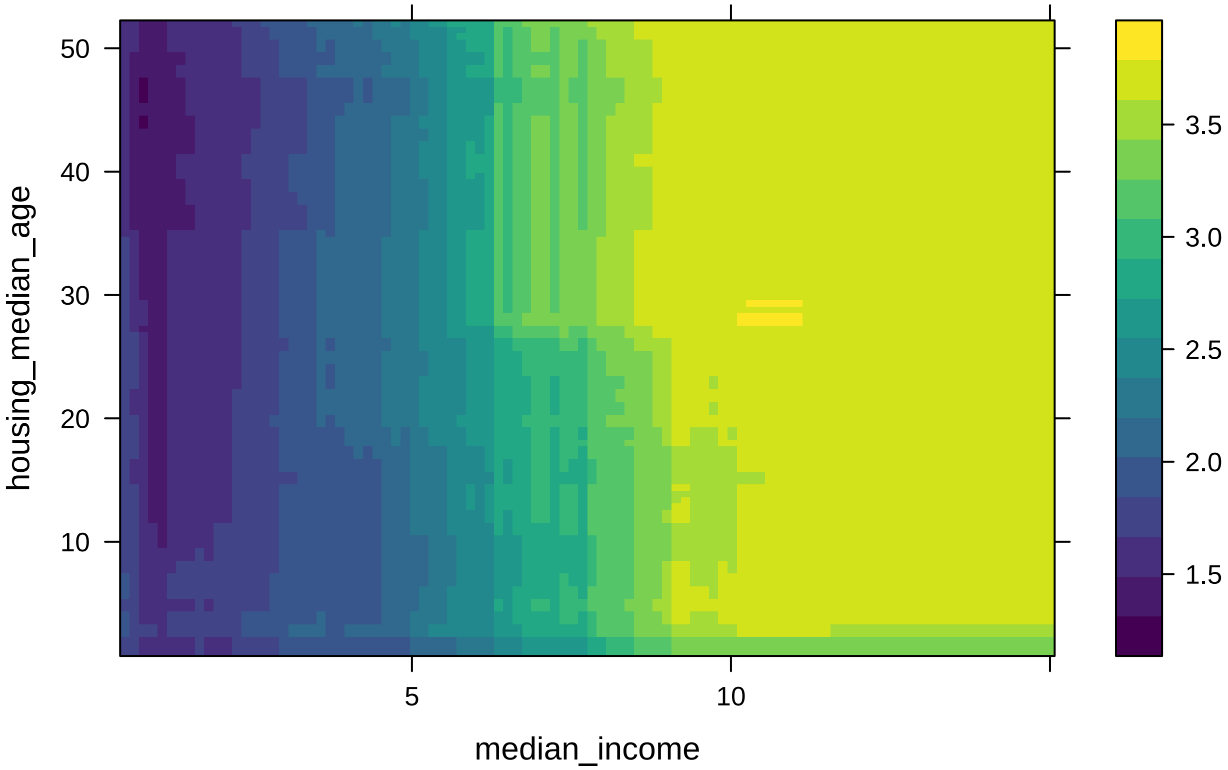
Next, we proceed to pairwise partial dependence plots. A natural pairing is the lat/long dependence, which gives:



Again, the general southwest direction across the state corresponds to higher median home values – this we expected from above. Furthermore, if we pair average number of rooms and average number of bedrooms, we see



The pairwise plot really just reflects the two marginal plots above – nothing too interesting here. Finally, for fun, we look at median income vs. population (admittedly, housing median age is not a wildly important feature, by the two importance measures above), and see:



Again, there's not a whole lot interesting about this plot: as one would expect, home value increases with median income, and to a much lesser extent with age. The noteable find here is that areas with high median incomes but low age (green in the bottom right) show high, but not extremely high, median home values. One explanation here might be that many of these high-earning young people have simply yet to buy an expensive home. Alternatively, perhaps this might result from areas where parents are wealthy, but there are many kids in each household. Here, parents might opt for less expensive/top-end housing, in anticipation of the cost of raising all of these children.