(F23) PSTAT 126: Project Step 3

Anthony Cu and William Mahnke

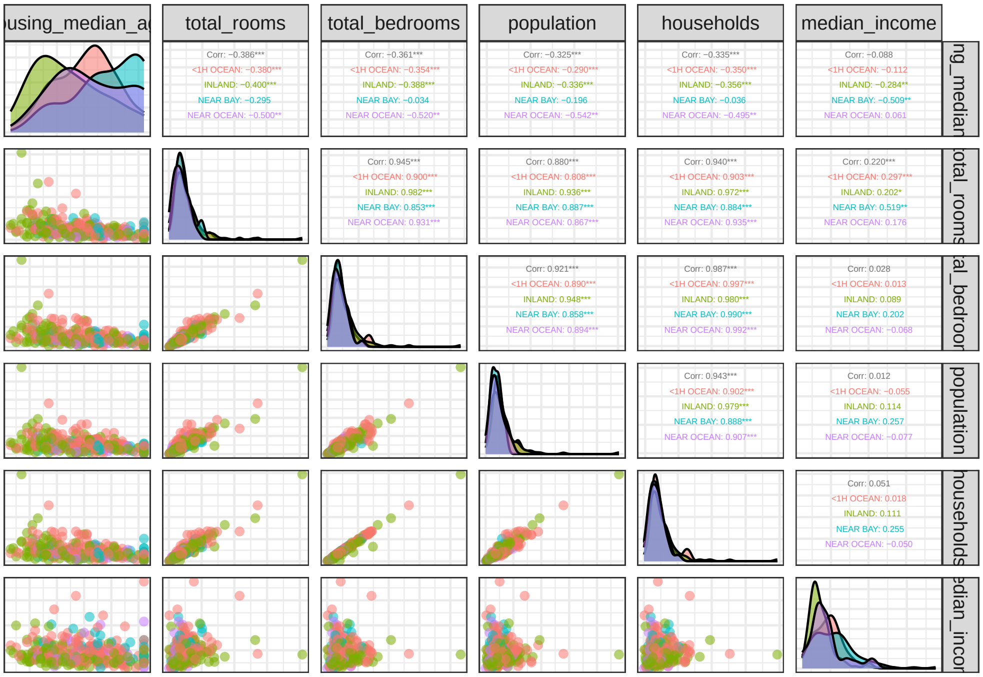
27 December 2023

Introduction

In our project, we explore the 1990 California Housing data set, providing information on a specified district in the state. The survey data describes homes in a district of California in 1990 in order to represent the larger population of this district during the 1990's overall. Each independent observation corresponds to a different block within the district. The data set comes from Kaggle's data repository.

We focus on our response median house value, and its predictors: median house age, median income, population, households, number of bedrooms, number of bathrooms, and ocean proximity.

Pairs Plot on Explanatory and Response Variables



We observe that there is a high correlation between some pairs of explanatory variables, including total bedrooms and total rooms, population and total rooms, population and total bedrooms, and more shown in the correlation plot. Additionally we see that population and households has a positive linear relationship. When finding an optimal model using the training data, we will consider interaction terms for these pairs of highly correlated variables.

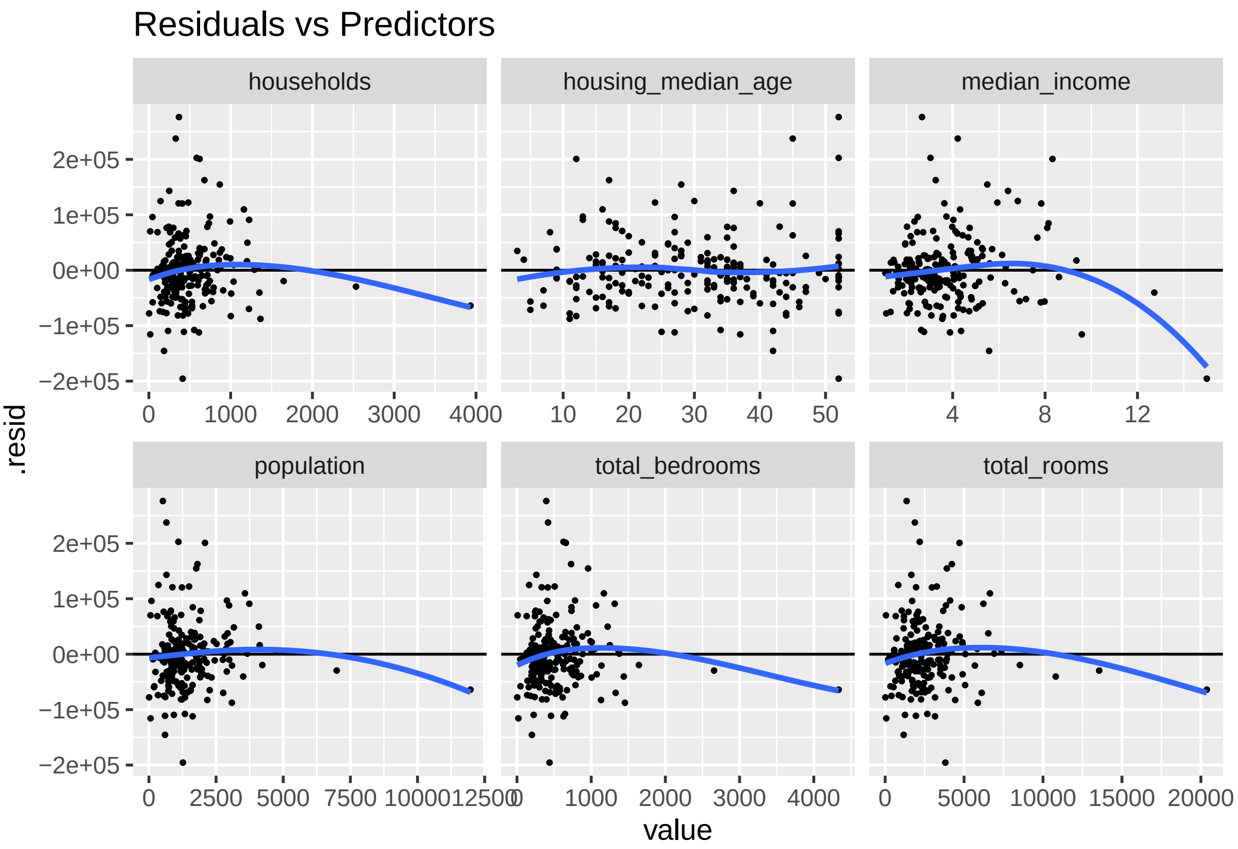
Dividing the data

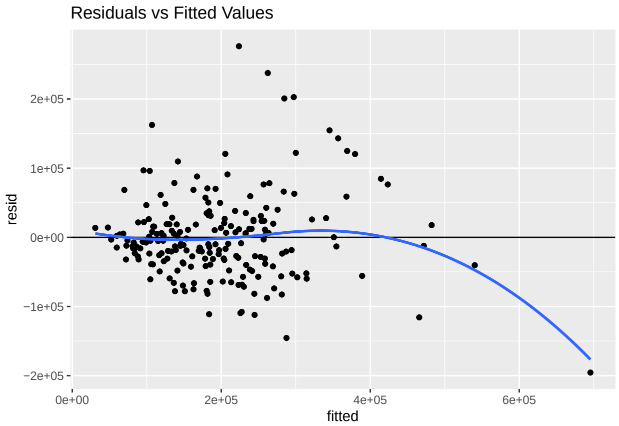
We will split our data into two groups. Seventy percent of the data will be used for finding and training a model while the remaining 30% will be used to perform significance tests, analyze unusual observations, and calculating other important aspects of the model such as R^2_{adj}, R^2, and confidence and prediction intervals.

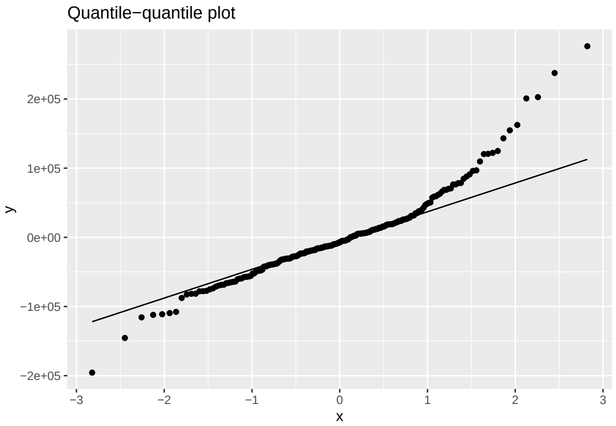
Checking Model Assumptions

We first fit our naive model in which we enter all our predictors linearly, through:

```r
naive_model <- lm(median_house_value ~ ., data = housePartition$train)
```

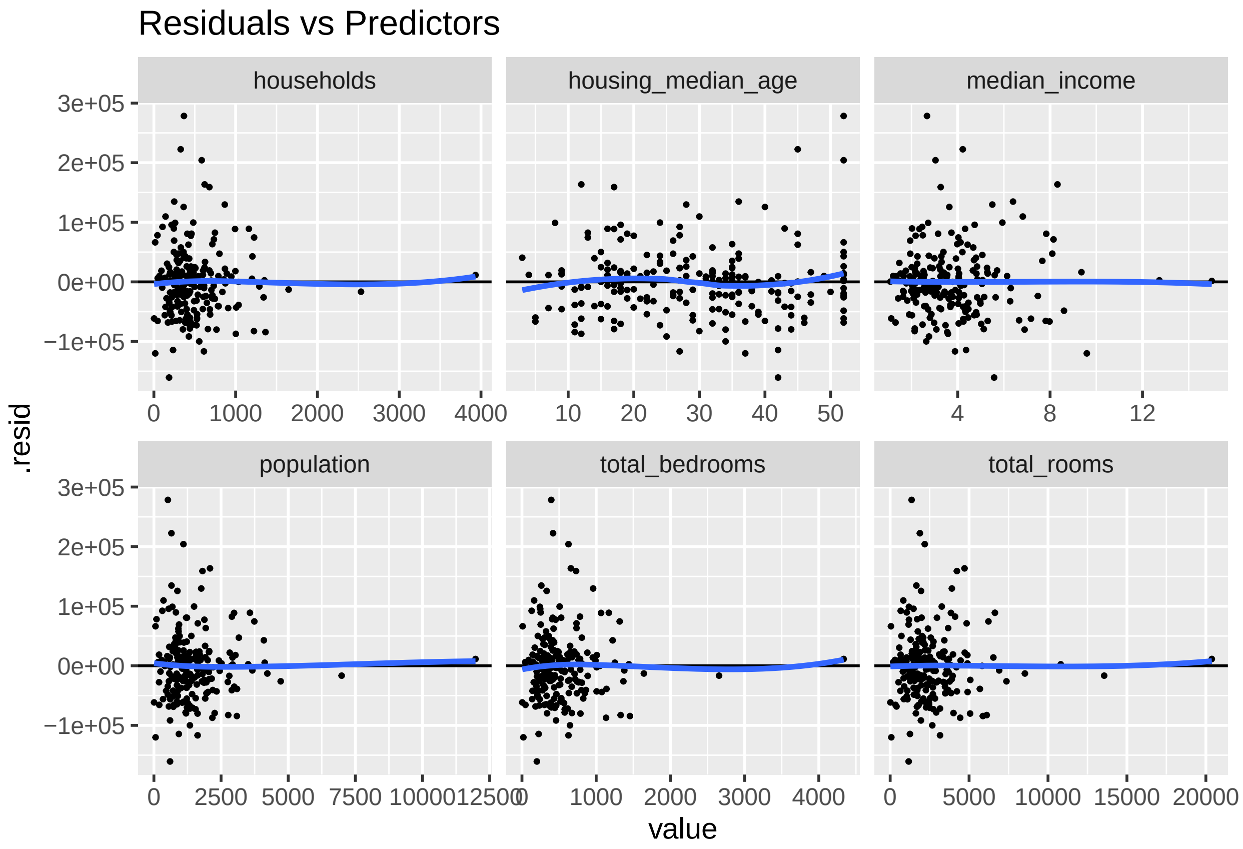


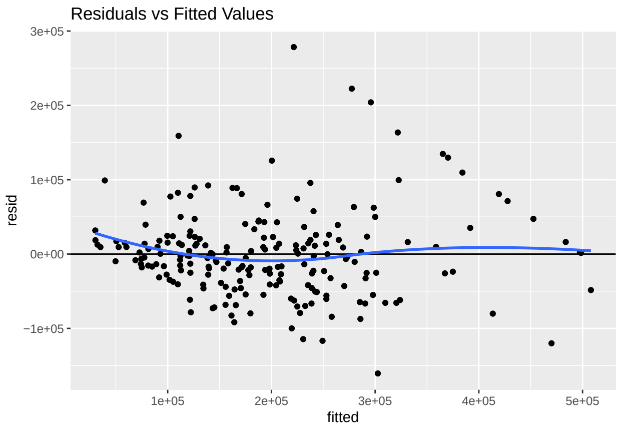


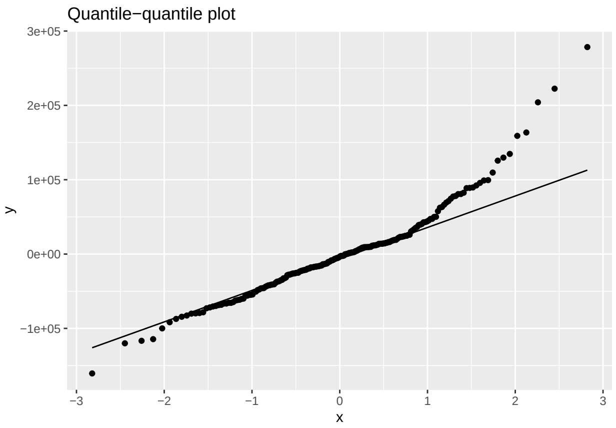


Looking at the diagnostic plots for our naive model, we observed that linear assumption is not met for median income, households, and total rooms. Additionally, the residual vs fitted plot shows the constant variance assumption is also not met. We noticed that households and median income resembled a quadratic shape in its residual vs predictor plot, so we transformed our model by entering households and median income as a quadratic and a cubic respectively.

```r
fitted_model1 <- lm(median_house_value ~ ocean_proximity + housing_median_age +
                      poly(households, 2, raw = T) + total_rooms +
                      total_bedrooms + population + poly(median_income, 3, raw = T),
                    data= housePartition$train)
```







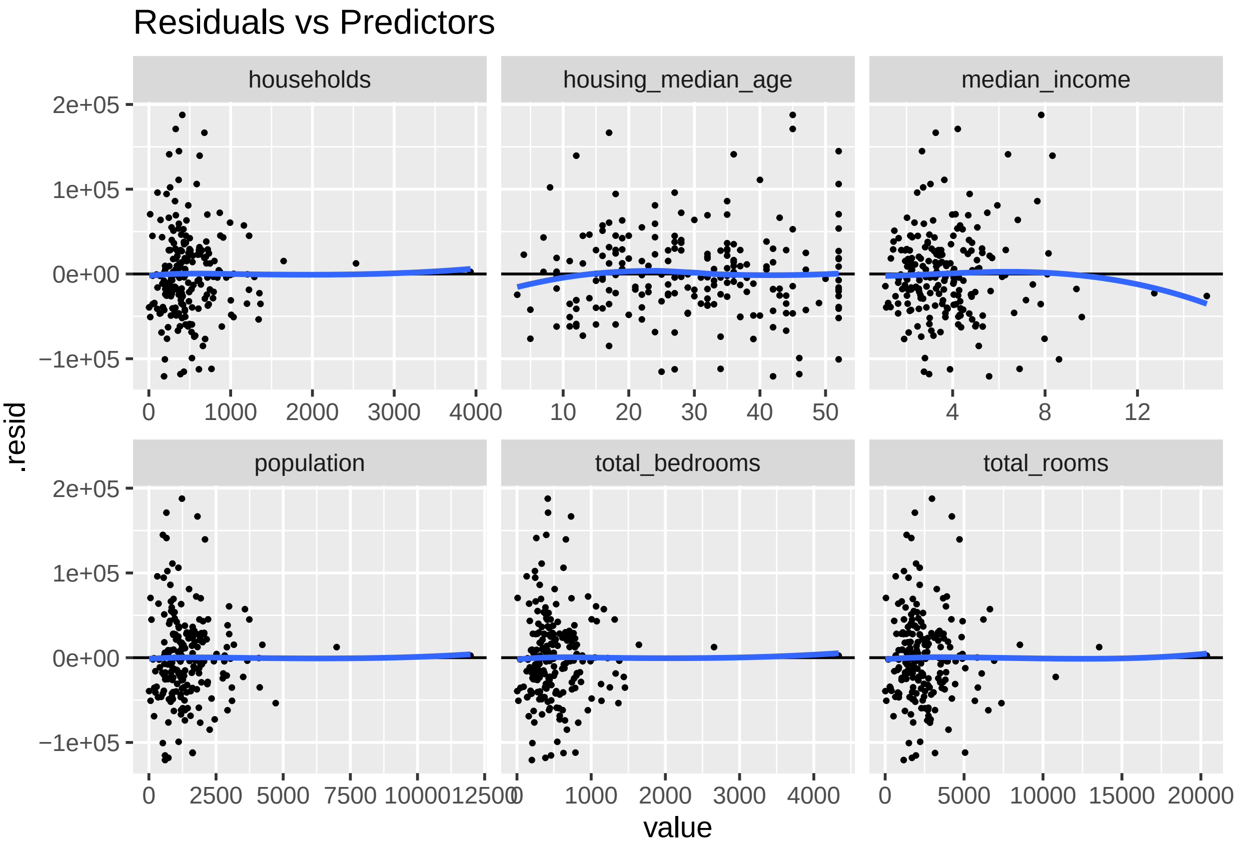
The lack of a pattern in the residual vs predictor plots for our new model suggest that the linearity assumption is now met. The residual vs fitted plot for this new model also doesn't show any recognizable pattern, confirming that the constant variance assumption is met. The quantile-quantile plot shows that normality assumption is also met. Additionally, observations in our data correlate to different blocks within a district. So while there isn't a diagnostic plot to check for the independence assumption, the nature of our data should ensure that assumption is also met.

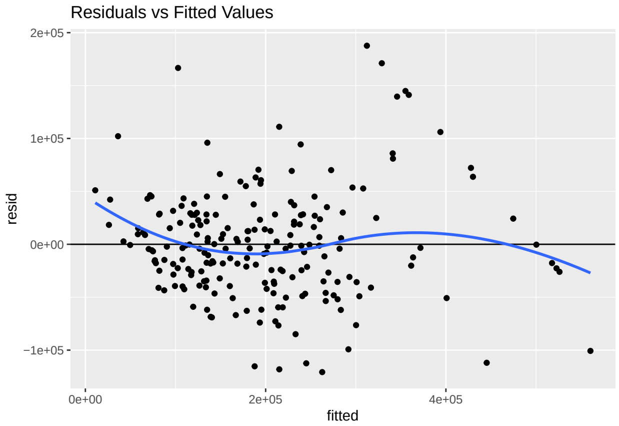
Variable Selection

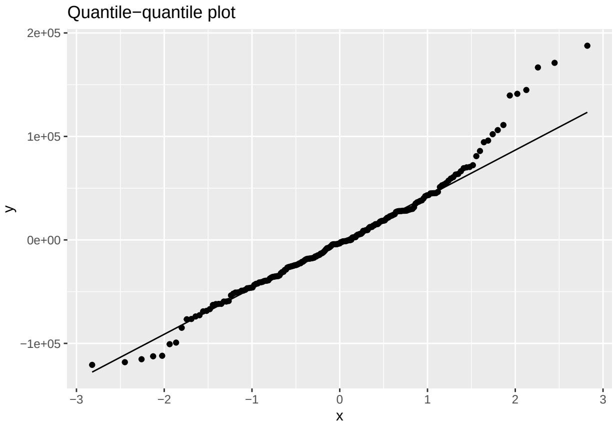
As our pairs plot showed earlier, we thought it was important to include interaction terms when picking an optimal. We begin our variable selection by defining our null and full models using the training data and then performing backwards selection on our full model.

From performing the backward selection, we fit our second linear model accounting for certain interaction terms between predictors. Some of the interaction terms are reflected in the variables with high correlation from the pairs plot shown earlier.

```r
fitted_model2 <- lm(median_house_value ~ housing_median_age + total_rooms +
    total_bedrooms + population + households + median_income +
    ocean_proximity + housing_median_age:total_rooms + housing_median_age:population +
    housing_median_age:households + total_rooms:total_bedrooms +
    total_rooms:population + total_bedrooms:population + total_bedrooms:households +
    total_bedrooms:median_income + population:median_income +
    median_income:ocean_proximity, data = housePartition$train)
```







Checking the model assumptions for this new model, we see that linearity, constant variance, and normality are satisfied again. Independence follows too since the nature of the data hasn't changed. Now that we have two valid models, one produced from entering in terms to satisfy the model assumptions and the other from backwards selection, we will compared their AIC, BIC, and R^2_{adj} to select a single 'best' model to use on the testing data we set aside earlier.

We now check criteria to determine which fitted model is better.

- The AIC (Akaike Information Criterion) of fitted_model1 is 5210.7327151 $>$ 5171.7825392, the AIC of fitted_model2. So, fitted_model2 minimizes AIC and thus prioritizes predictive accuracy.
- To add on, the BIC (Bayesia Information Criterion) of fitted_model1 is 5257.5253946 $>$ 5248.656227, the BIC of fitted_model2. So, fitted_model2 minimizes BIC and thus priorities selection consistency.
- Finally, the adjusted R^2 of fitted_model1 is 0.7163487 $<$ 0.7736092, the adjusted R^2 of fitted_model2. So, fitted_model2 maximizes adjusted R^2 and thus prioritzies model fit.

Statistical Method

```
Call:
lm(formula = median_house_value ~ housing_median_age + total_rooms +
    total_bedrooms + population + households + median_income +
    ocean_proximity + housing_median_age:total_rooms + housing_median_age:population +
    housing_median_age:households + total_rooms:total_bedrooms +
    total_rooms:population + total_bedrooms:population + total_bedrooms:households +
    total_bedrooms:median_income + population:median_income +
    median_income:ocean_proximity, data = housePartition$test)

Residuals:
   Min     1Q Median     3Q    Max
-89228 -31481  -6241  27092 175684

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -2.580e+04  6.444e+04  -0.400  0.69018
housing_median_age                       4.506e+02  1.184e+03   0.381  0.70467
total_rooms                             -8.889e+01  5.835e+01  -1.523  0.13227
total_bedrooms                          -2.109e+01  2.897e+02  -0.073  0.94218
population                              -2.738e+00  1.018e+02  -0.027  0.97862
households                               6.174e+02  3.258e+02   1.895  0.06226
median_income                            5.204e+04  9.214e+03   5.648 3.35e-07
```

```
ocean_proximityINLAND                    1.782e+04  4.097e+04   0.435  0.66484
ocean_proximityNEAR BAY                 -1.544e+04  5.312e+04  -0.291  0.77223
ocean_proximityNEAR OCEAN               -1.665e+05  8.899e+04  -1.871  0.06557
housing_median_age:total_rooms           4.684e-01  1.496e+00   0.313  0.75511
housing_median_age:population           -4.745e-01  1.646e+00  -0.288  0.77393
housing_median_age:households           -5.012e-02  7.694e+00  -0.007  0.99482
total_rooms:total_bedrooms               3.212e-02  3.601e-02   0.892  0.37556
total_rooms:population                   1.848e-02  1.955e-02   0.945  0.34790
total_bedrooms:population               -1.355e-02  1.375e-01  -0.099  0.92180
total_bedrooms:households               -4.762e-01  2.132e-01  -2.233  0.02877
total_bedrooms:median_income             1.385e+02  4.128e+01   3.356  0.00129
population:median_income                -5.224e+01  2.014e+01  -2.594  0.01158
median_income:ocean_proximityINLAND     -2.732e+04  1.286e+04  -2.125  0.03718
median_income:ocean_proximityNEAR BAY   -7.118e+03  1.504e+04  -0.473  0.63755
median_income:ocean_proximityNEAR OCEAN  4.767e+04  2.360e+04   2.020  0.04726

(Intercept)
housing_median_age
total_rooms
total_bedrooms
population
households                              .
median_income                           ***
ocean_proximityINLAND
ocean_proximityNEAR BAY
ocean_proximityNEAR OCEAN               .
housing_median_age:total_rooms
housing_median_age:population
housing_median_age:households
total_rooms:total_bedrooms
total_rooms:population
total_bedrooms:population
total_bedrooms:households               *
total_bedrooms:median_income            **
population:median_income                *
median_income:ocean_proximityINLAND     *
median_income:ocean_proximityNEAR BAY
median_income:ocean_proximityNEAR OCEAN *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 51690 on 69 degrees of freedom
Multiple R-squared:  0.8517,    Adjusted R-squared:  0.8066
F-statistic: 18.87 on 21 and 69 DF,  p-value: < 2.2e-16
```

Looking at the summary of our model using the testing data and assuming a p-value of 0.1, we see that the statistically significant predictors in our model are median income, households, the indicator variable when ocean proximity is near the ocean, total bedrooms interacting with households, total bedrooms interacting with median income, population interacting with median_income, median income when the ocean proximity is inland, and median income when the ocean proximity is near ocean.

Our model shows that fixing all other variables, a $10000 increase in median income increases the average median house value by about $52000. The significance in the relationship makes sense. Looking at the interaction terms involving median income and ocean proximity, we see that the association between mean median house value and median income decreases by about $7118 per $10000 of median income when the neighborhood is considered inland. Additionally, the association between mean median house value and median income increases by about $47670 per $10000 of median income when the neighborhood is considered near the ocean. In real life, these two relationships make sense as well.

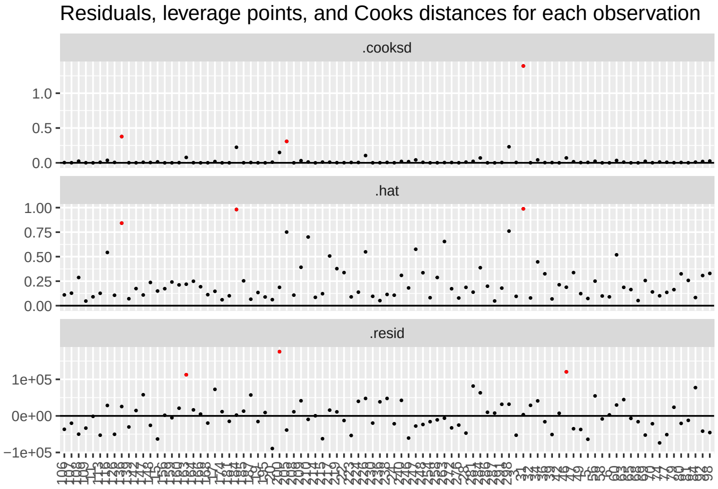
The significance of certain predictors over others indicates that median income, households, and ocean proximity are the most important factors when estimating the median house price of the sample. While other factors are just as important in determining the value of a household, it makes sense that location and the quality of families (determined by median income and households) are the most important indicators of house value, especially in California. Families with a higher income are going to be able to afford more expensive houses and houses by the ocean are generally more expensive because of they're in a more preferable location.

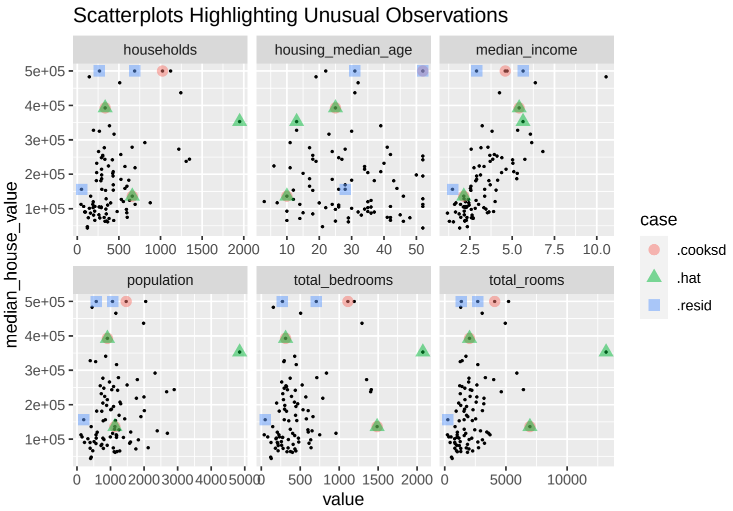
On the test data, the $R^2 = 0.851702$ and the $R^2_{adj} = 0.8065678$. This means that about 80% of the variance is explained by our model.

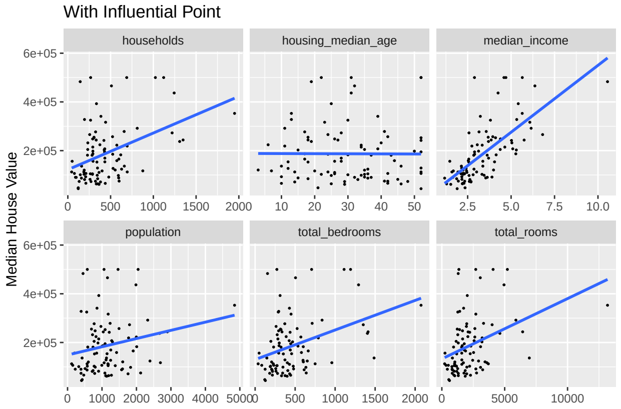
A high R^2 is not necessarily a guarantee that the model will accurately describe the population because the calculated R^2 value only accounts for values within the sample and doesn't include the entire population.

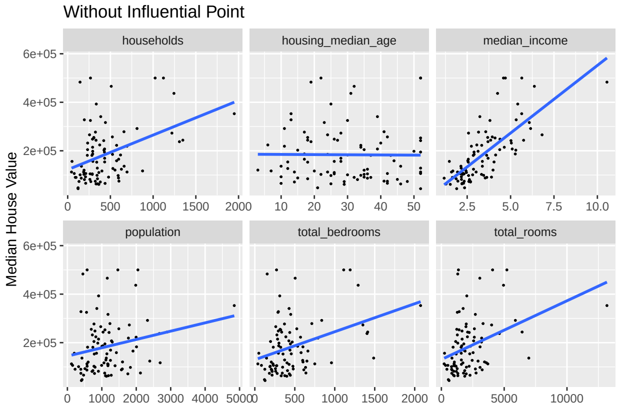
Analysis

```
# A tibble: 9 x 3
# Groups:   name [3]
  .rownames name       value
  <chr>     <chr>      <dbl>
1 31        .cooksd    1.39
2 136       .cooksd    0.377
3 205       .cooksd    0.308
4 31        .hat       0.988
5 184       .hat       0.981
6 136       .hat       0.842
7 200       .resid  175684.
8 46        .resid  120499.
9 163       .resid  112538.
```









Analyzing the residuals, leverage values, and cook's distances for all of the testing data, we noticed there was one row with a significantly higher cook's distance than other points (more than the other two points that unusual cook's distances) When plotting the model fit with and without that particular row, we observed that the fit was pulled a little towards the influential point. However, we agreed that the fit didn't change a considerable amount, so we continued with our confidence and prediction interval calculations including the row in our data.

Confidence and Prediction Intervals

```
       fit      lwr      upr
1 202165.7 184570.2 219761.1
```

With 95% confidence, the mean median house value for a block with measurements equal to the mean of average in the data is estimated to be between $184570.20 and $219761.1

```
       fit      lwr      upr
1 193561.4 73722.94 313399.9
```

With 95% confidence, the median house value for the particular sampled block is estimated to be between $73722.94 and $313399.90 (the particular block has the measurements: housing_median_age = 44, total_rooms = 2526, total_bedrooms = 579, population = 1423, households = 573, median_income = 2.5363, and ocean_proximity of <1H OCEAN).

Conclusion

We first created two models, one designed with the intention of satisfying the model assumptions and one designed using backwards selection (that satisfied the model assumptions too). Using AIC, BIC, and R^2_{adj}, we selected a model with a lot of interaction terms that reflected the high correlation we saw between some of the explanatory variables. We then analyzed the significance of each predictor in our model and the R^2_{adj} to explain what variables were the best predictors for median house value. With the model, we also analyzed the unusual observations in our data and how the points with high influence affected the fit of the model. Using the same model, we also calculated a 95% confidence interval for the mean median house price and a 95% prediction interval for a particular neighborhood.