Homework 4: OLS vs Random Forest

A battle for the ages

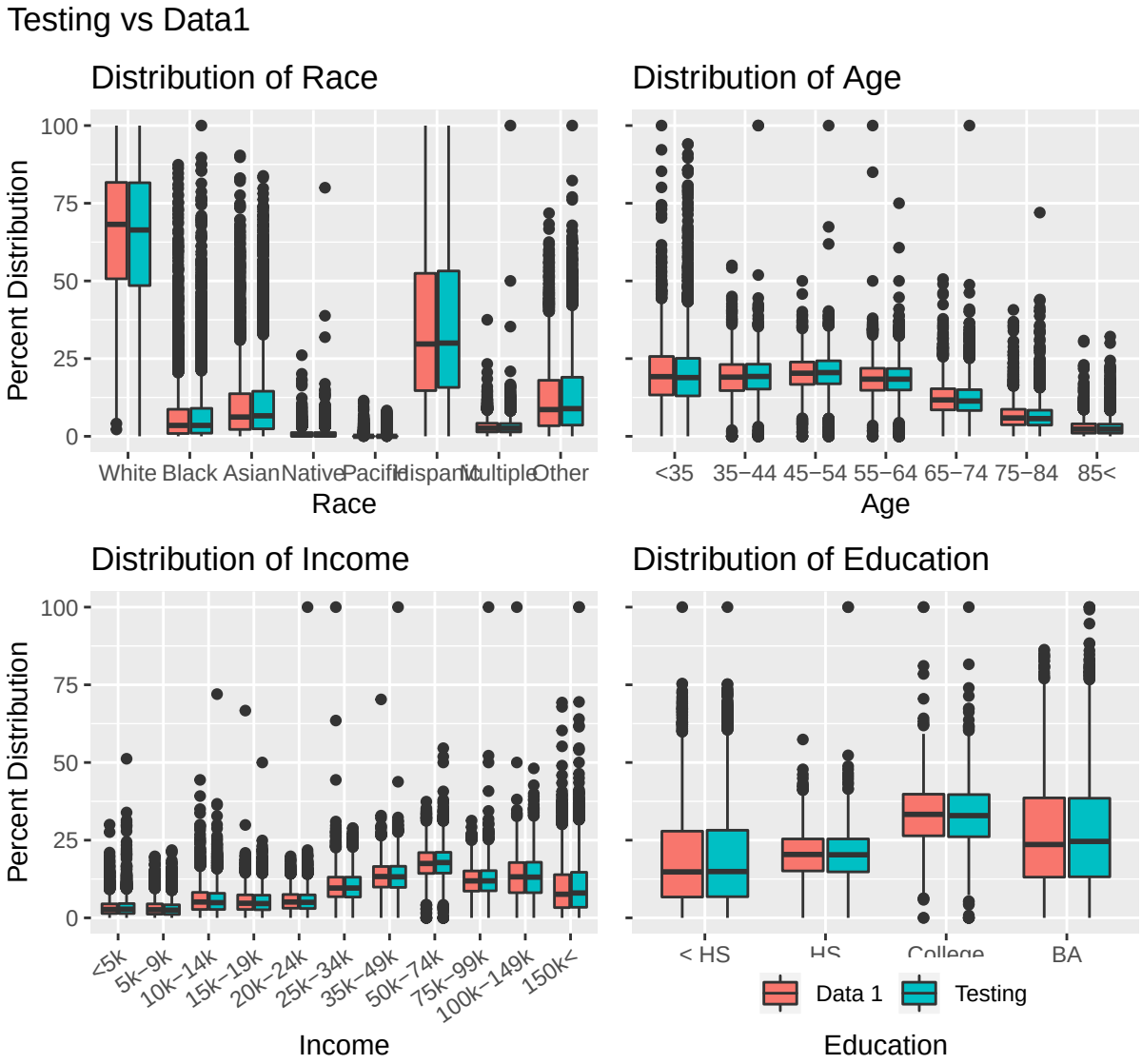
Neeraj Sharma

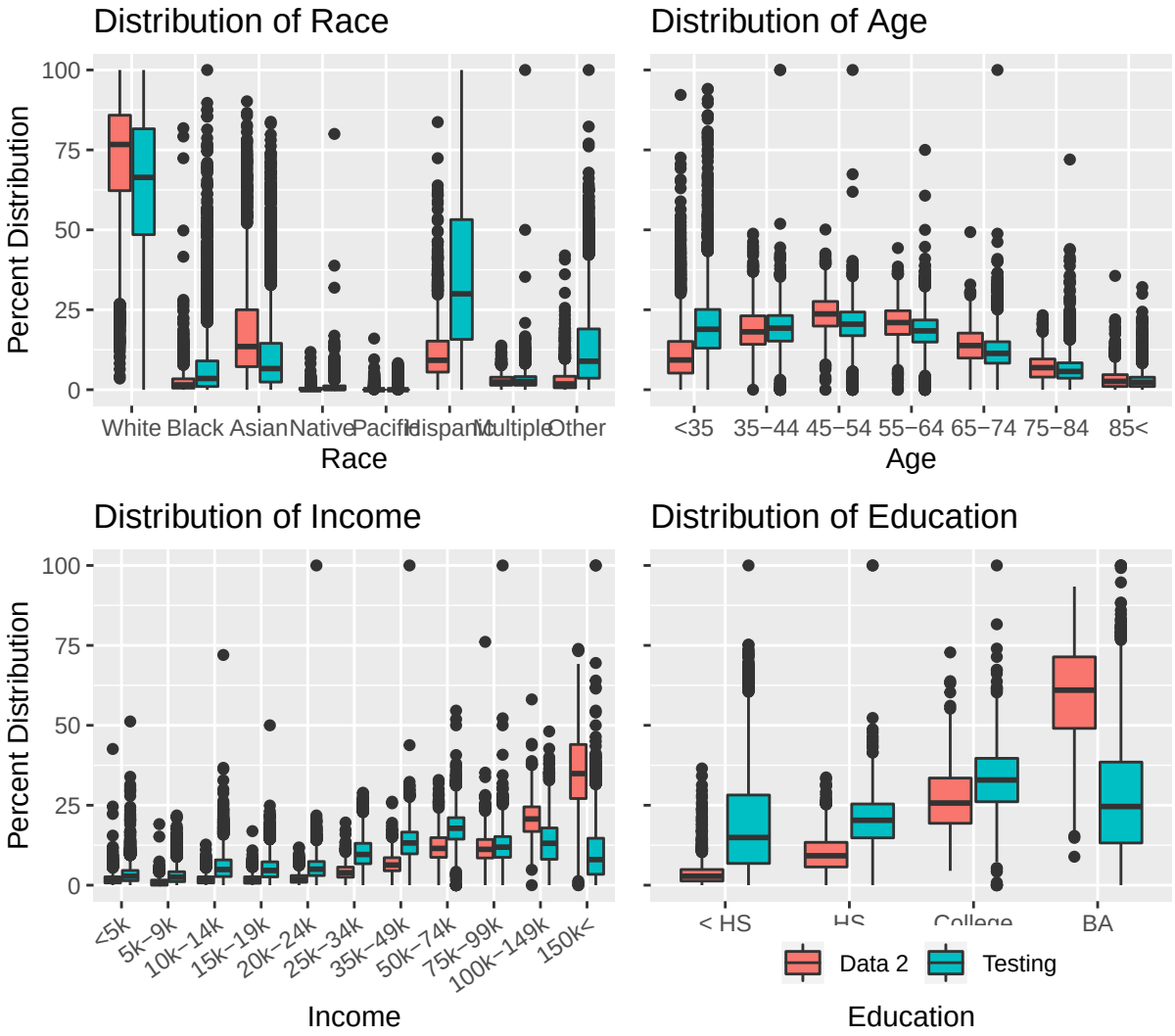
06/01/2020

1) OLS - prediction for median housing costs in data set 2

I approach creating my OLS model two ways. The first way is through a modeling and prediction-oriented approach using a LASSO regression to select variables, and the second approach is to intuitively reason which variables will contribute to housing prices.

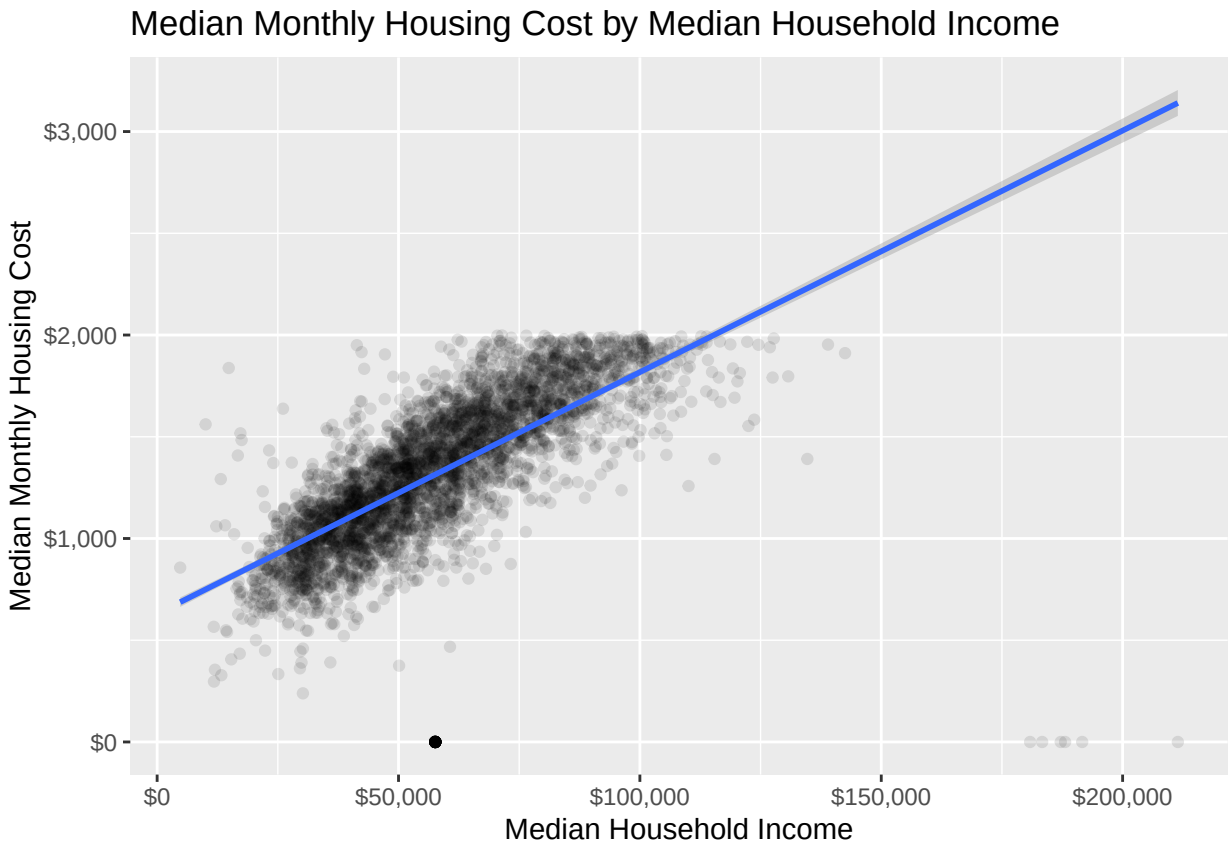
In order to effectively apply the model I train on the training dataset to experimental dataset 1, I first need to substantiate my belief that the two datasets look similar.

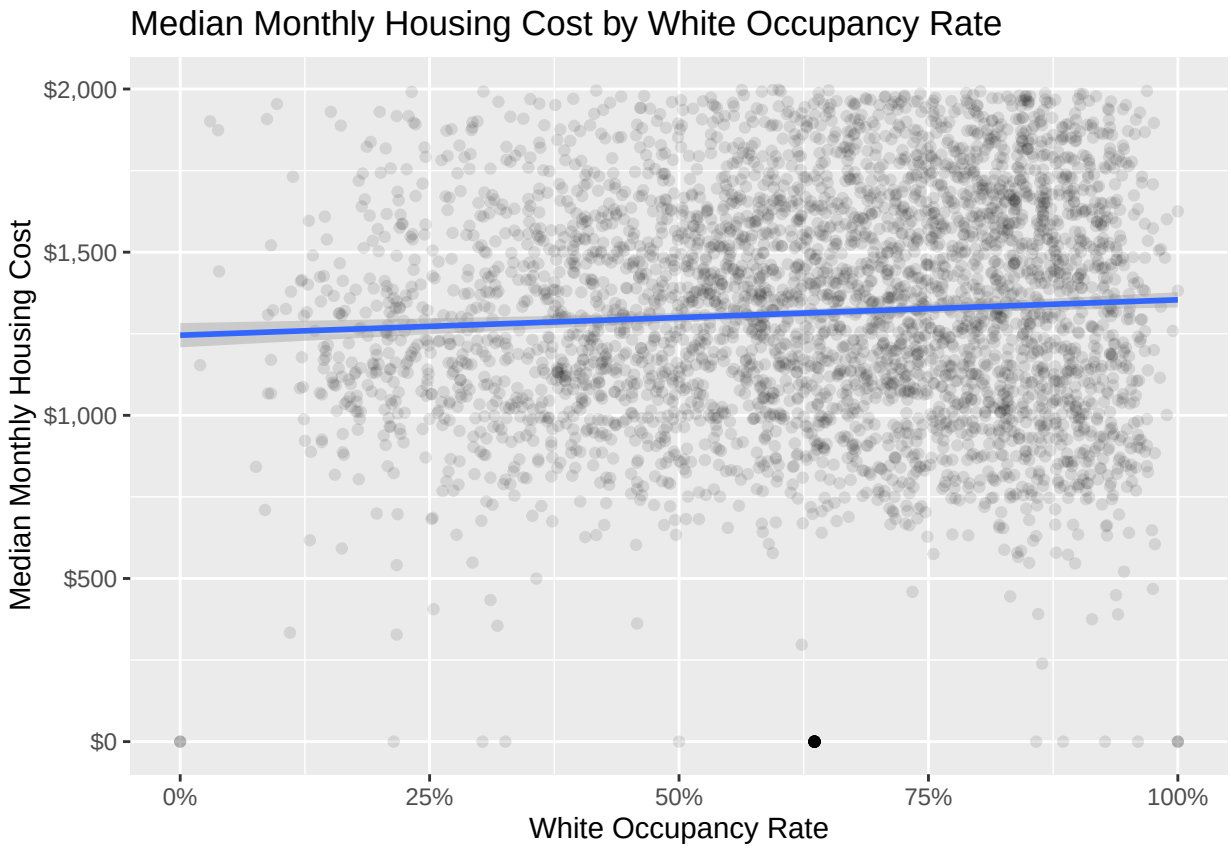




a. Describe both your model (as in, the regression you ran) and the thinking that underlay the choices of what to put in your model. (5 points)

```
## # A tibble: 3,894 x 6
##    MedianHousehold~ MeanHouseholdIn~ WhiteOcc CollegeOcc BAOcc
##               <dbl>            <dbl>    <dbl>      <dbl> <dbl>
##  1            61968            80794    0.563      0.332 0.135
##  2            56049            69944    0.528      0.309 0.192
##  3            49485            85135    0.866      0.377 0.282
##  4           101042           125475    0.856      0.215 0.649
##  5            58808            77753    0.859      0.301 0.195
##  6            53224            71563    0.911      0.319 0.25
##  7            31682            41053    0.207      0.288 0.08
##  8            41399            52785    0.39       0.229 0.057
##  9            59643            73596    0.639      0.371 0.103
## 10            51206            56815    0.832      0.39  0.161
## # ... with 3,884 more rows, and 1 more variable:
## #   MedianMonthlyHousingCosts <dbl>
```


Median Monthly Housing Cost by Median Household Income
Median Monthly Housing Cost
$3,000
$2,000
$1,000
$0
$0
$50,000
$100,000
$150,000
$200,000
Median Household Income


Median Monthly Housing Cost by White Occupancy Rate
Median Monthly Housing Cost
$2,000
$1,500
$1,000
$500
$0
0%
25%
50%
75%
100%
White Occupancy Rate

b. Guess what your performance will be in terms of R-squared and beta, when, using data set 2 we run a regression of the form:

y = a + beta*y-hat

where y-hat is your predicted housing costs and y is the true housing costs. We'd like numeric answers for both the r-squared and beta. Emphasize the logic of why you guessed your guesses. (5 points)

2. Random Forest - 15 points

Using random forest techniques, come up with your best prediction for median housing costs In data set 2..

a.Describe both your model (as in, the regression you ran) and the thinking that underlay the choices of what to put in your model. (5 points)

b. Guess what your performance will be in terms of R-squared and beta, when, using data set 2 we run a regression of the form:

y = a + beta*y-hat

where y-hat is your predicted housing costs and y is the true housing costs. We'd like numeric answers for both the r-squared and beta. Emphasize the logic of why you guessed your guesses. (5 points)

c. Which do you think will do better in out-of-sample predictions, random forest or OLS? (5 points)