Predicting AirBnB Rental Prices

Group 11: Trevor Isaacson, Jonathan Olavarria, Jasmine DeMeyer

12/10/2021

Motivation

- You are looking for some additional income and decide renting on AirBnB is the best option
- How much should you rent your extra space for?

Data

- In general, AirBnB data is very open and be easily accessed
- The original dataset is from a past Kaggle competition
  - Contained over 74,000 individual listings
- For sake of time and processing power, we took a random sample of 17,500 from those 74,000 listings
- They also provided a testing file
- Since the competition is over, we will compile our final predictions on that file using our best model

Data

- Consists of 30 variables
- Variables are about the property, property location, the host and host reviews
- After cleaning and eliminating variables, our data consisted of 22 variables

```
str(training_df)
```

```
## 'data.frame':    17440 obs. of  22 variables:
##  $ property_type          : Factor w/ 29 levels "Apartmen
##  $ room_type              : Factor w/ 3 levels "Entire ho
##  $ accommodates           : int  2 2 5 4 3 2 3 8 3 3 ...
##  $ bathrooms              : num  1 2 1 1 1 1 1 2.5 1 1 .
##  $ bed_type               : Factor w/ 5 levels "Airbed","
##  $ cancellation_policy    : Factor w/ 5 levels "flexible"
##  $ cleaning_fee           : Factor w/ 2 levels "FALSE","T
##  $ city                   : Factor w/ 6 levels "Boston","
##  $ host_has_profile_pic   : Factor w/ 3 levels "","f","t"
```

Baseline Regression

```
linear = lm(price ~ ., data = training)
```

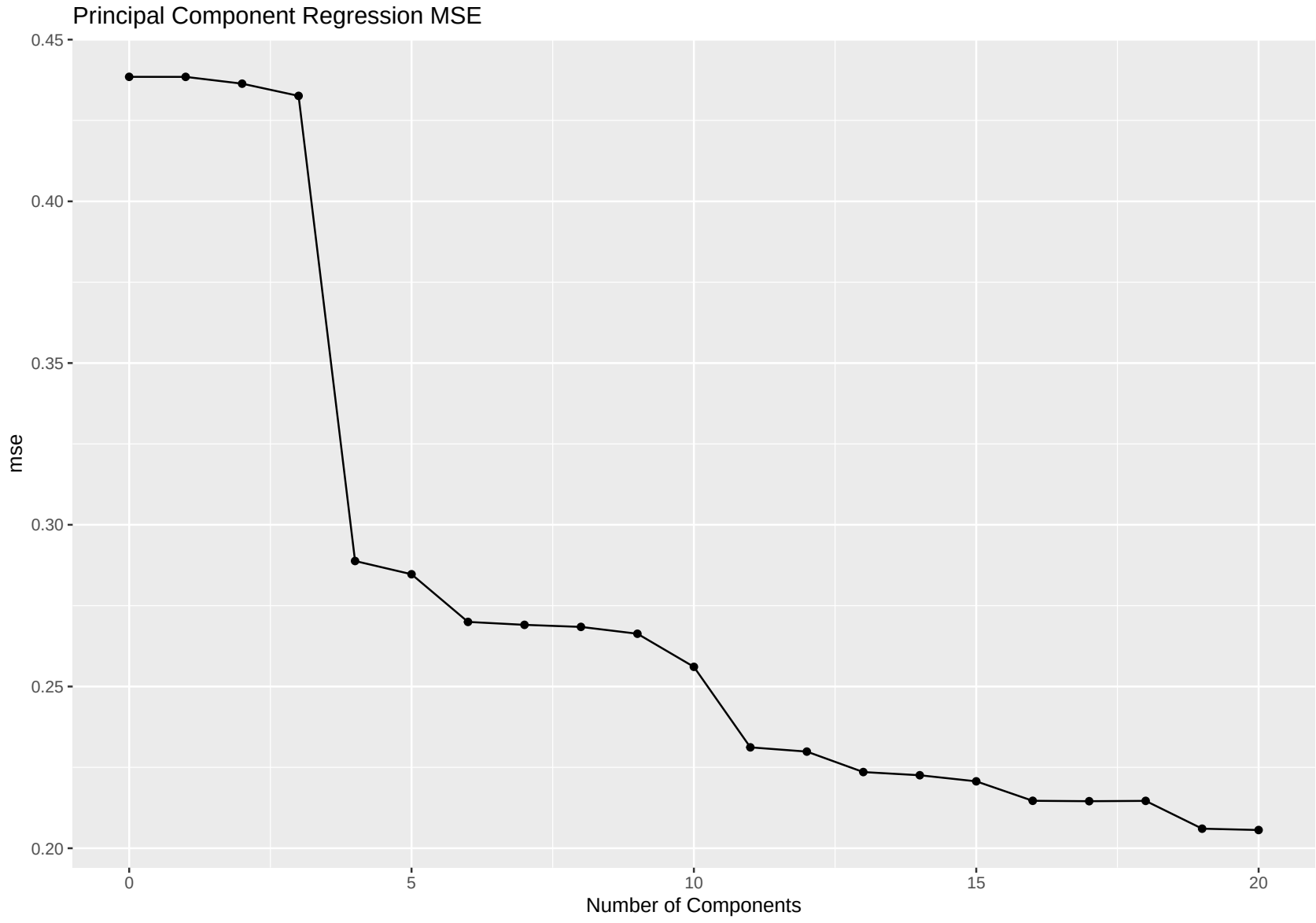
```
## [1] "MSE of Testing Set:  0.165"
```

Regression Splines/Generalized Additive Models

PCR and PLS

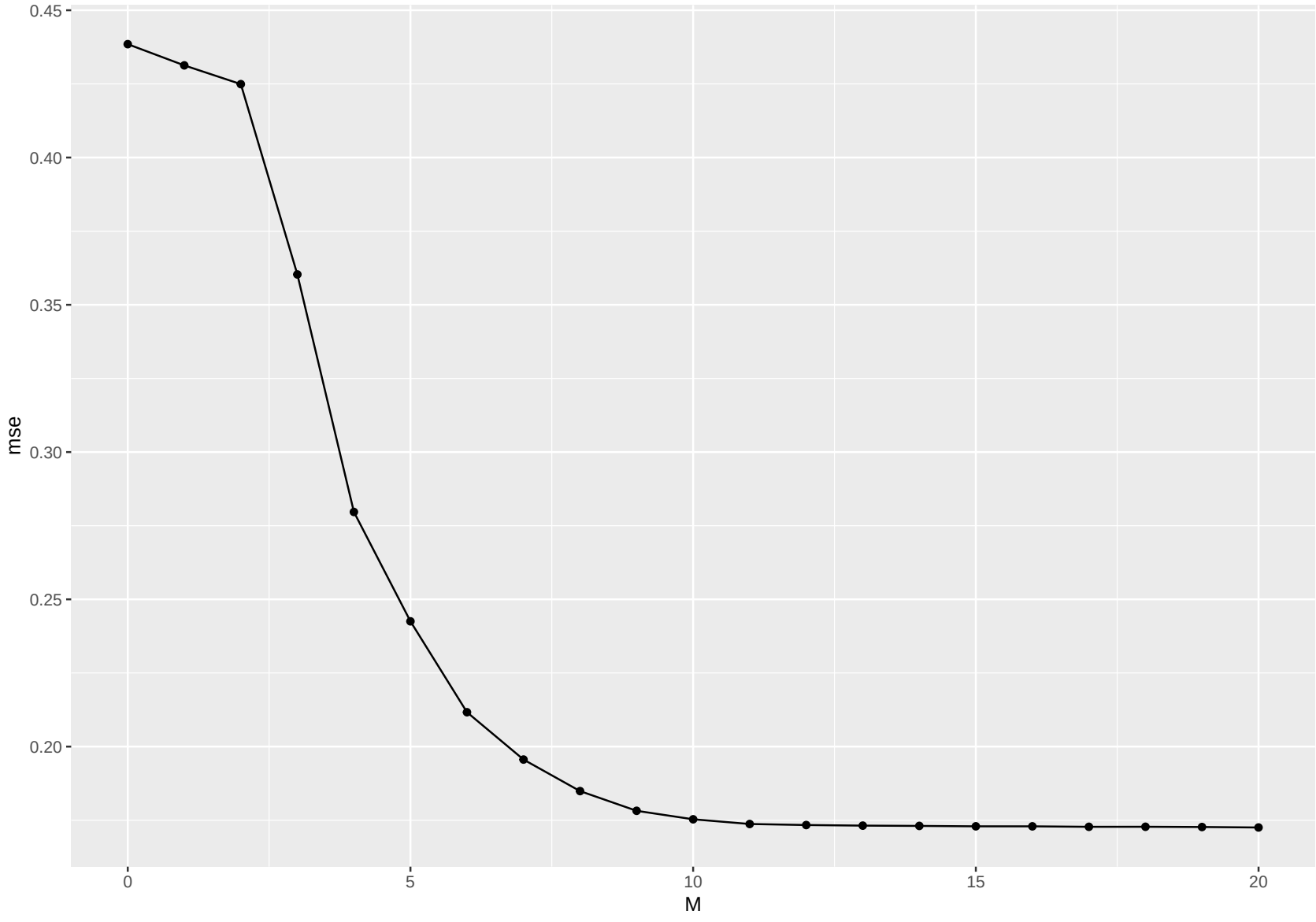
- 10 Fold Cross-Validation was performed for number of components ranging from 1 to 20.
- The Cross-Validation MSE was used to pick optimal number of components for both models.

PCR


Principal Component Regression MSE
mse
0.45
0.40
0.35
0.30
0.25
0.20
0
5
10
15
20
Number of Components

PCR Predictions

PLS



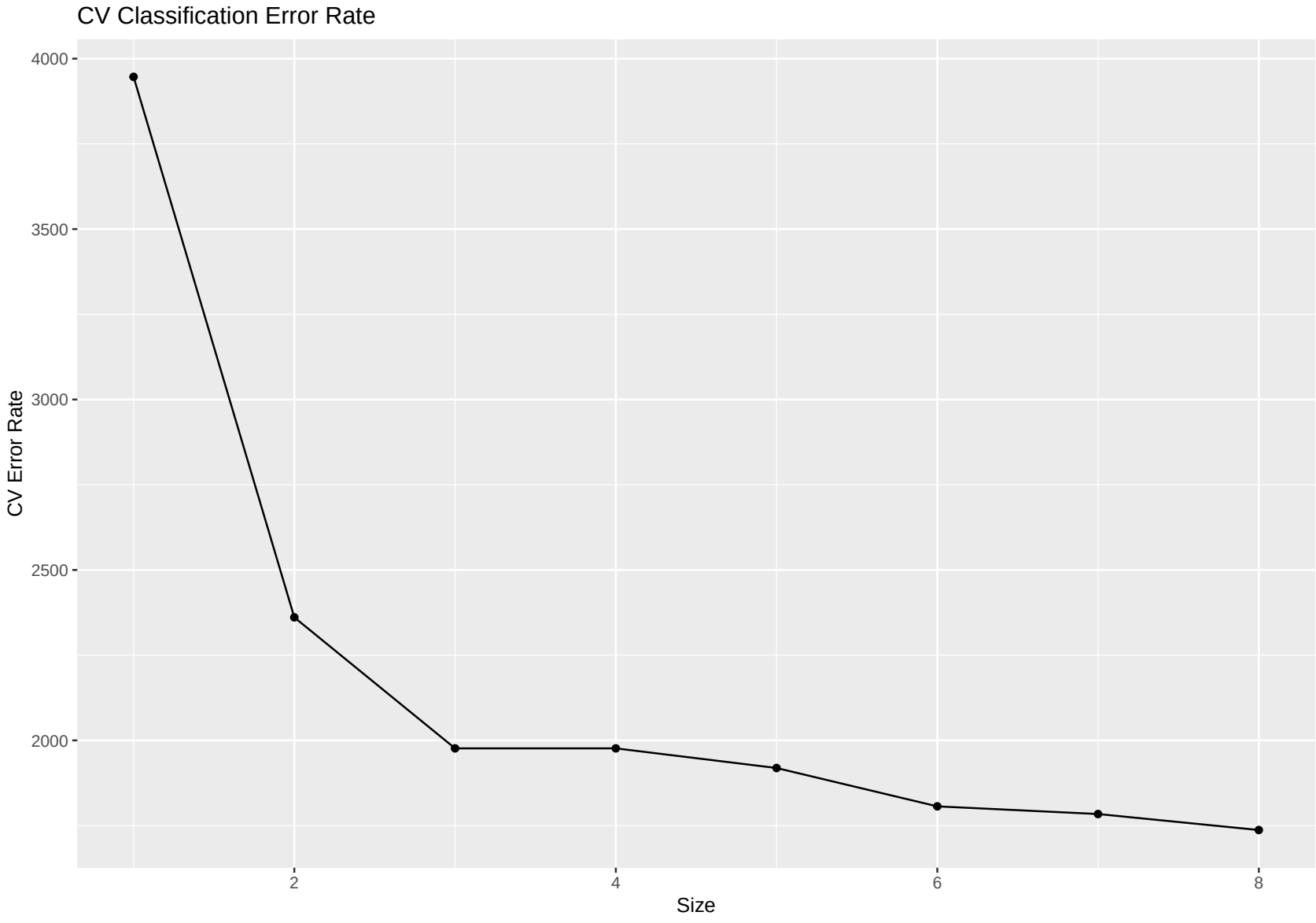
PLS Predictions

Regression Trees

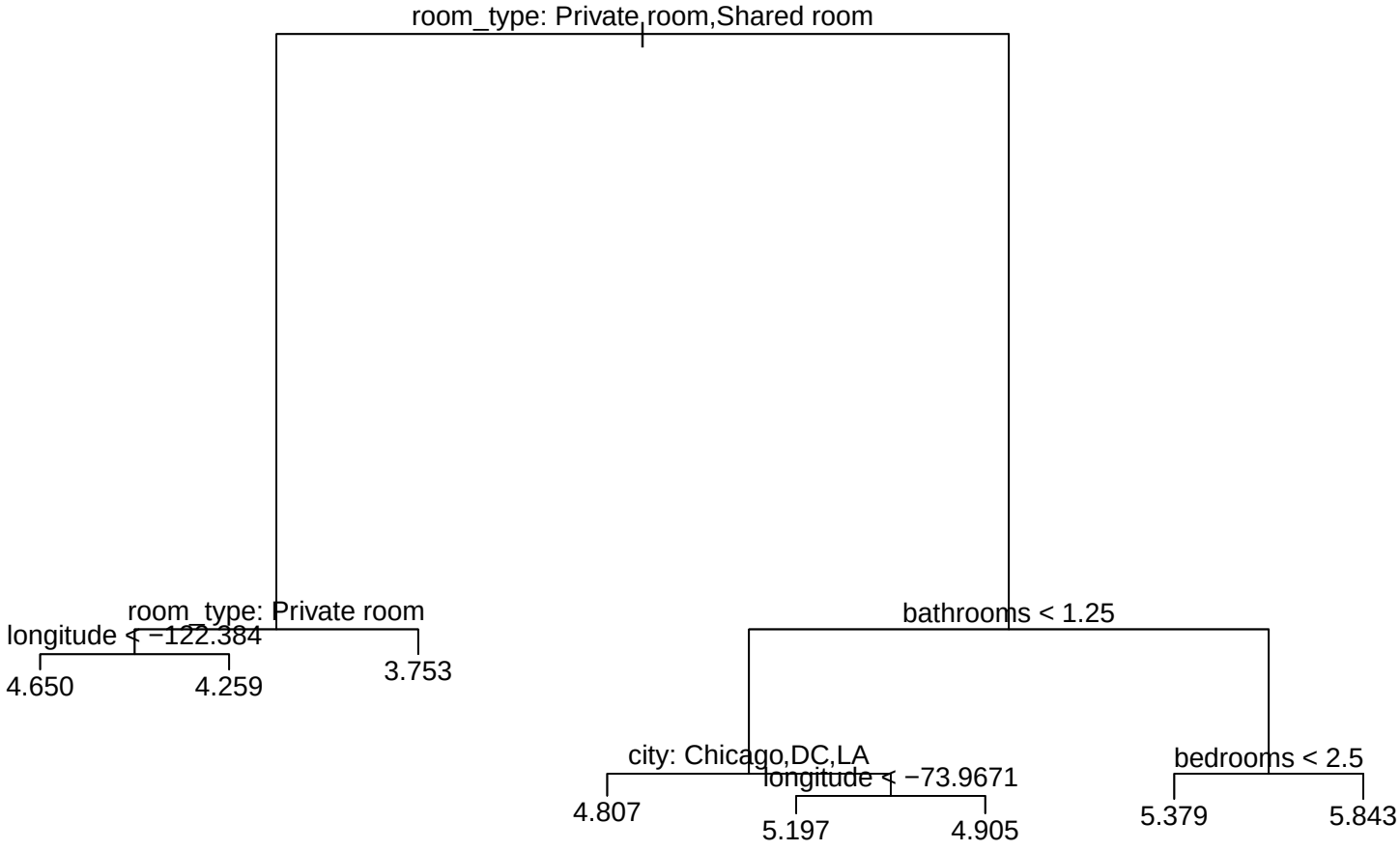
```
##
## Regression tree:
## tree(formula = price ~ ., data = training)
## Variables actually used in tree construction:
## [1] "room_type" "longitude" "bathrooms" "city"        "bed
## Number of terminal nodes:  8
## Residual mean deviance:  0.1885 = 1695 / 8992
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.5050 -0.2999 -0.0196  0.0000  0.2558  2.8310

## [1] "Test MSE of Initial Tree:  0.1926"
```

Regression Trees



Regression Trees



Questions?

References