STAT-650-FINAL-PROJECT

Raksha Ramaraj and Yogesh Gupta

2022-10-05

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(randomForest)
```

```
## randomForest 4.7-1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(vip)
```

```
##
## Attaching package: 'vip'

## The following object is masked from 'package:utils':
##
##     vi
```

Loading the data set

```
data= read_csv("realtor_data.csv",show_col_types = FALSE)
```

13 variables in data set

```
names(data)
```

```
##  [1] "status"       "price"        "bed"          "bath"         "acre_lot"
##  [6] "full_address" "street"       "city"         "state"        "zip_code"
## [11] "house_size"   "sold_date"    "year"
```

The data set has 106064 rows and we have used only 8 variables for our analysis

```
realestate<-select(data, price,bed,bath,acre_lot,city,state,house_size,year)
dim(realestate)
```

```
## [1] 106064      8
```

```
colSums(is.na(realestate))
```

```
##      price        bed       bath   acre_lot       city      state house_size
##          0       6312       4686      30121          0          0      33873
##       year
##          0
```

Though, there are lot of NA values in the data set, it has not been removed since the variable of interest "Price" has no missing values.

```
typeof(realestate$price)
```

```
## [1] "double"
```

```
summary(realestate)
```

```
##      price               bed              bath           acre_lot
##  Min.   :       0   Min.   : 1.000   Min.   : 1.000   Min.   :    0.00
##  1st Qu.:  299900   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.:    0.09
##  Median :  460000   Median : 3.000   Median : 2.000   Median :    0.19
##  Mean   :  744516   Mean   : 3.276   Mean   : 2.408   Mean   :   17.56
##  3rd Qu.:  775000   3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.:    0.54
##  Max.   :35000000   Max.   :47.000   Max.   :39.000   Max.   :49299.44
##                     NA's   :6312     NA's   :4686     NA's   :30121
##      city              state            house_size         year
##  Length:106064      Length:106064      Min.   :  260   Min.   :2018
```

```
##  Class :character   Class :character   1st Qu.: 1120   1st Qu.:2019
##  Mode  :character   Mode  :character   Median : 1560   Median :2020
##                                        Mean   : 2020   Mean   :2020
##                                        3rd Qu.: 2272   3rd Qu.:2021
##                                        Max.   :99999   Max.   :2022
##                                        NA's   :33873
```

```
realestate<-mutate(realestate,price1=price/1000)
realestate
```

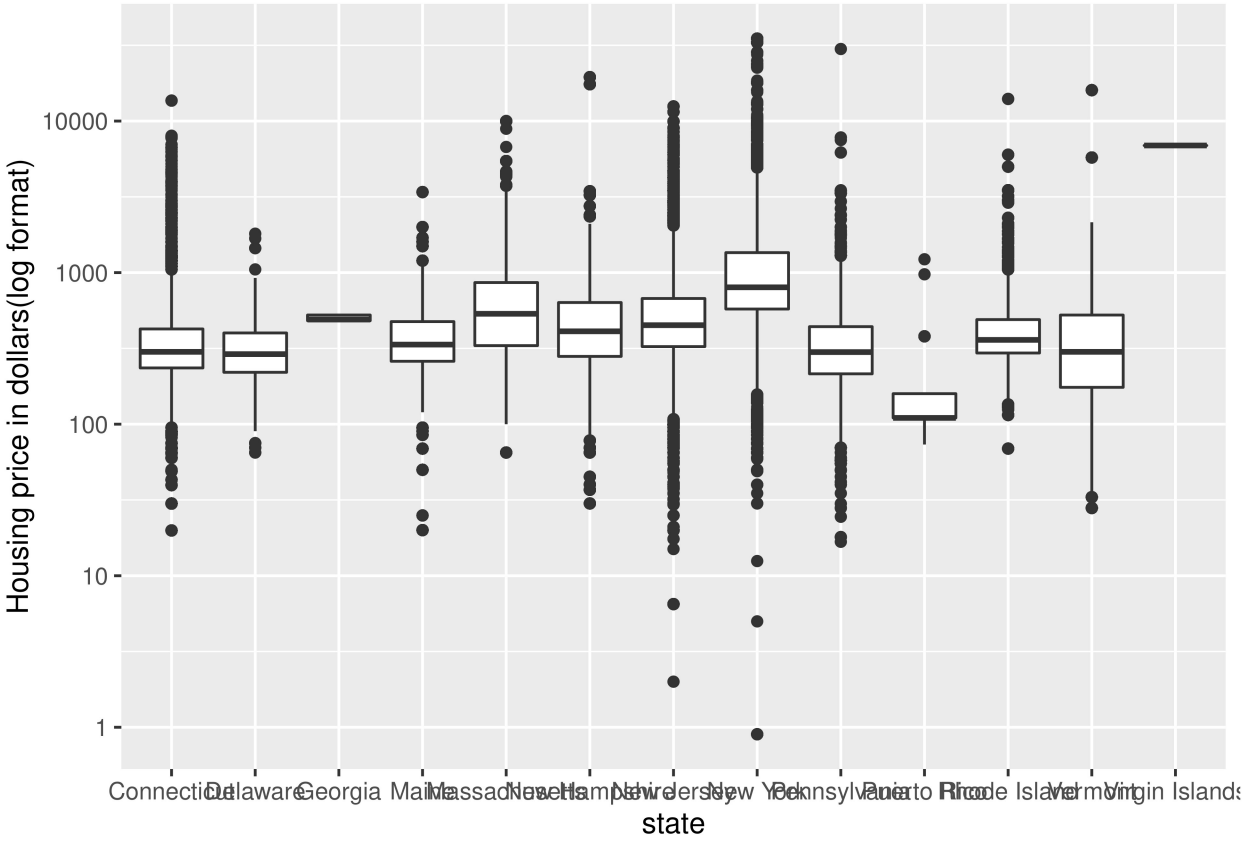
```
## # A tibble: 106,064 x 9
##     price   bed  bath acre_lot city      state       house_size  year price1
##     <dbl> <dbl> <dbl>    <dbl> <chr>     <chr>            <dbl> <dbl>  <dbl>
##  1  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
##  2  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
##  3  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
##  4 110000     7     3     0.09 Dorado    Puerto Rico       1192  2019  110
##  5  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
##  6  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
##  7  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
##  8  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
##  9  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
## 10  76900     3     2    NA    Canovanas Puerto Rico       1200  2020   76.9
## # ... with 106,054 more rows
```

Plot of housing prices over different states

```
ggplot(data = realestate, aes(x = state, y = price1))+
  geom_boxplot() +
  scale_y_log10()+ylab("Housing price in dollars(log format)")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 36 rows containing non-finite values (stat_boxplot).
```



In the above plot we are trying to analyse prices over different states for five years. We can see that most of the house prices are between 100k to 1M. We see that New York state has the highest median compared to other states.The virgin island has only three rows of sales data which is making it hard to interpret.The least median is observed for the Puerto Rico.

```
realestate %>%
  group_by(state) %>%
  summarise(
    count=n(),
    avg_price=mean(price,na.rm=TRUE)
  ) %>% arrange(desc(avg_price))
```

```
## # A tibble: 13 x 3
##    state          count avg_price
##    <chr>          <int>     <dbl>
##  1 Virgin Islands     3  6899000
##  2 New York       22984  1326843.
##  3 Massachusetts   8758   804289.
##  4 New Jersey     41294   631970.
##  5 New Hampshire   4822   620792.
##  6 Vermont         3277   503064.
##  7 Georgia           36   499047.
##  8 Rhode Island    5586   468039.
##  9 Maine           1764   445782.
## 10 Connecticut    13675   421878.
## 11 Pennsylvania    3406   391402.
```

```
## 12 Delaware          397   336458.
## 13 Puerto Rico        62   197727.
```

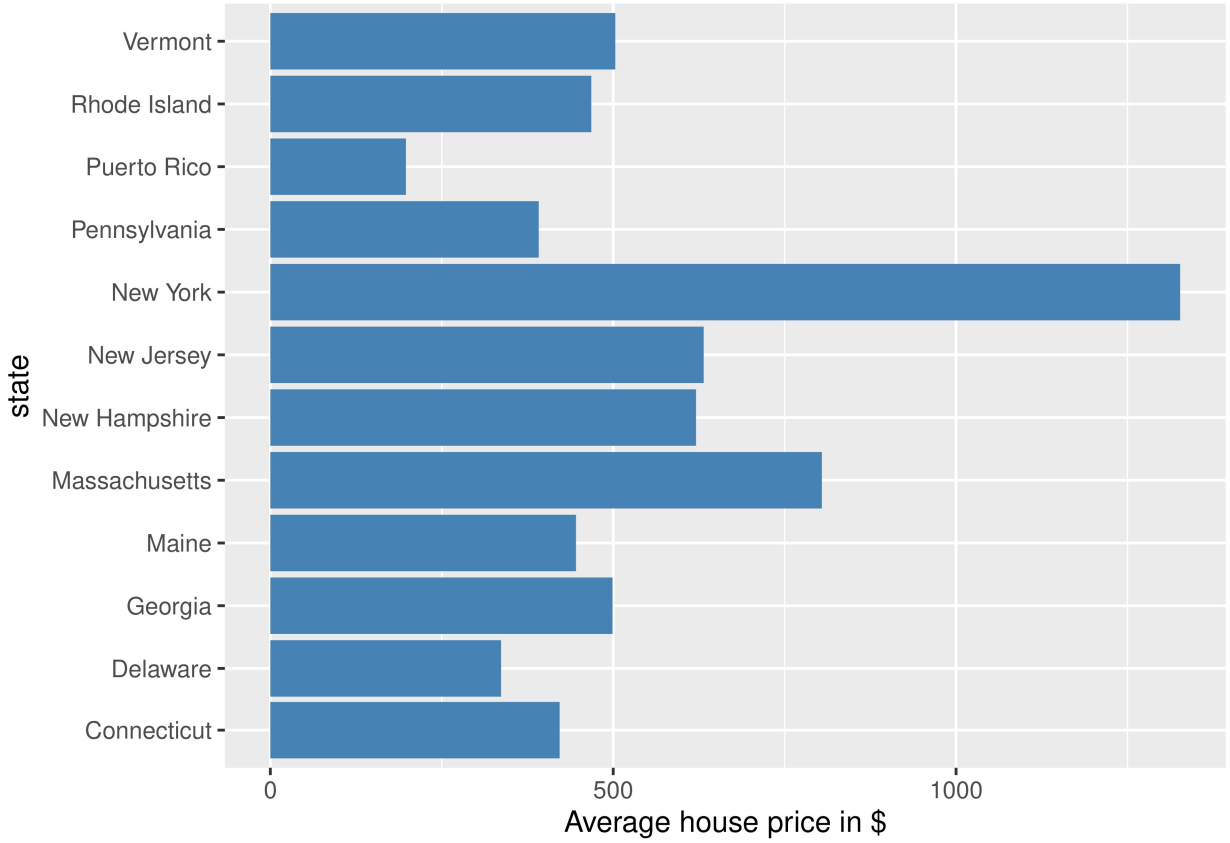
The above table gives an over view of the average prices of the houses sold over different states. Virgin island has an highest average price of 6.8M but only three houses were sold here over a period of 5 years. The least average house price was in Puerto Rico.

```
realestate1<- filter(realestate,state!="Virgin Islands")
dim(realestate1)
```

```
## [1] 106061      9
```

Plot of Average housing price over different states

```
realestate1 %>%
  group_by(state) %>%
summarise(avg=mean(price, na.rm=TRUE)) %>%
  ggplot()+
  geom_col(aes(x = state, y = avg/1000),fill="steelblue")+coord_flip()+ylab("Average house price in $")
```



The virgin Islands has been removed since it has only three data points and it has the maximum average selling price of 6.8M. In the current plot we can see that New York has the highest average selling price.

```
newyork<- filter(realestate1,state=="New York")
dim(newyork)
```

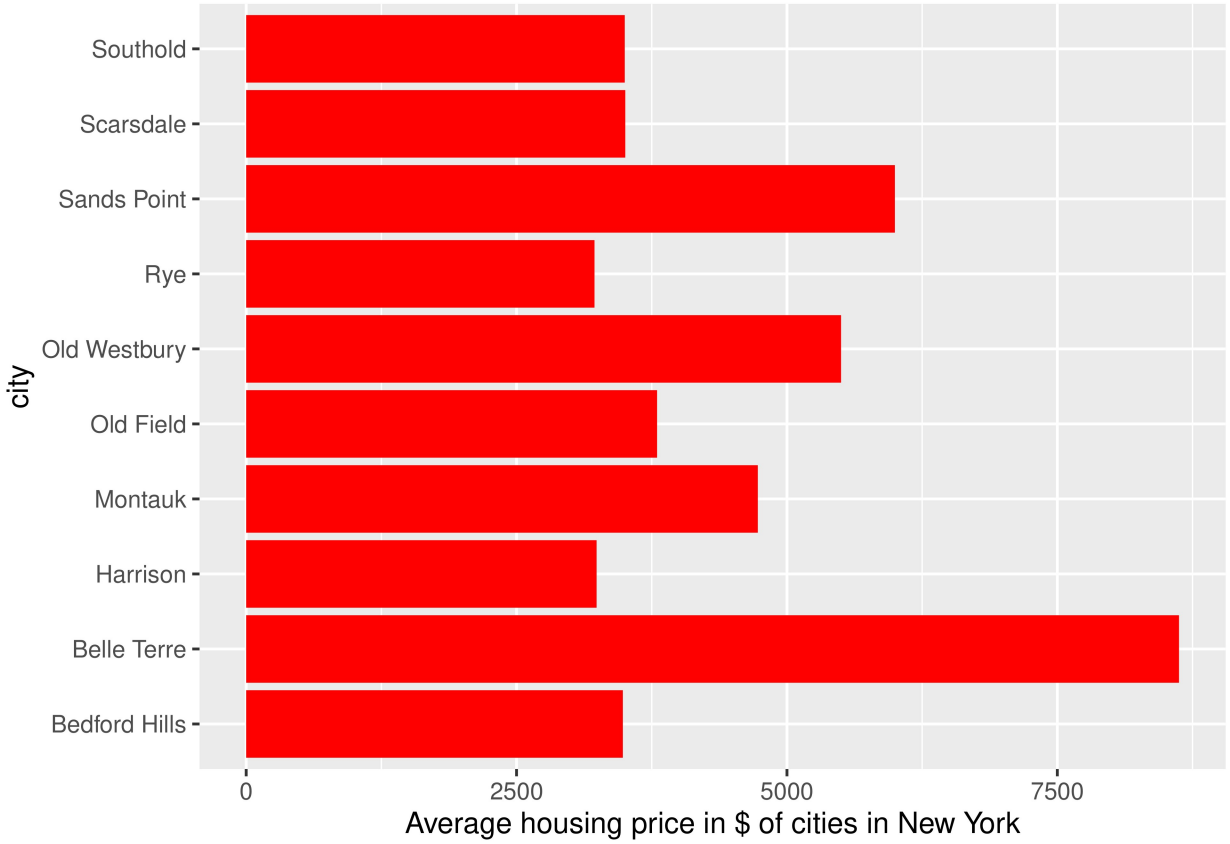
```
## [1] 22984     9
```

```
newyork1<-newyork %>%
  group_by(city) %>%
  summarise(
    count=n(),
    avg_price=mean(price,na.rm=TRUE)
  ) %>%
  arrange(desc(avg_price))
newyork1
```

```
## # A tibble: 296 x 3
##    city          count avg_price
##    <chr>         <int>     <dbl>
##  1 Belle Terre       2  8625000
##  2 Sands Point       5  5998000
##  3 Old Westbury      1  5500000
##  4 Montauk          14  4730714.
##  5 Old Field         1  3799000
##  6 Scarsdale        46  3504958.
##  7 Southold          1  3500000
##  8 Bedford Hills    11  3482727.
##  9 Harrison          4  3240000
## 10 Rye               9  3220555.
## # ... with 286 more rows
```

Among the 296 cities in the New York state, the city of Belle Terre has the maximum average price of $8.63M. Also, New York city recorded the maximun number of houses being sold.

```
ggplot(newyork1[tail(order(newyork1$avg_price), 10), ], ) +
  geom_col(aes(x = city, y = avg_price/1000),fill="red")+coord_flip()+ylab("Average housing price in $ (
```



Based on previous analysis, we noticed that the state of New York had the highest average housing price over 5 years. The above plot shows us the pricing trend in different cities in New York state.
• The city of Belle Terre, though had only two houses sold over 5 years, had the highest average price of $8.6M across the state. • Among the above considered cities, Scarsdale had higher number of houses sold – 46, the average price being $3.5M.

Random forest analysis for New York state.

Random Forest analysis is used to determine the effect of 7 predictor variables (see VIP plot below for exact variables considered) on the response variable (housing price) across New York state.

```
newyork2<-na.omit(newyork)
newyork2<-select(newyork2, price,bed,bath,acre_lot,city,state,house_size,year)
```

```
names(newyork2)
```

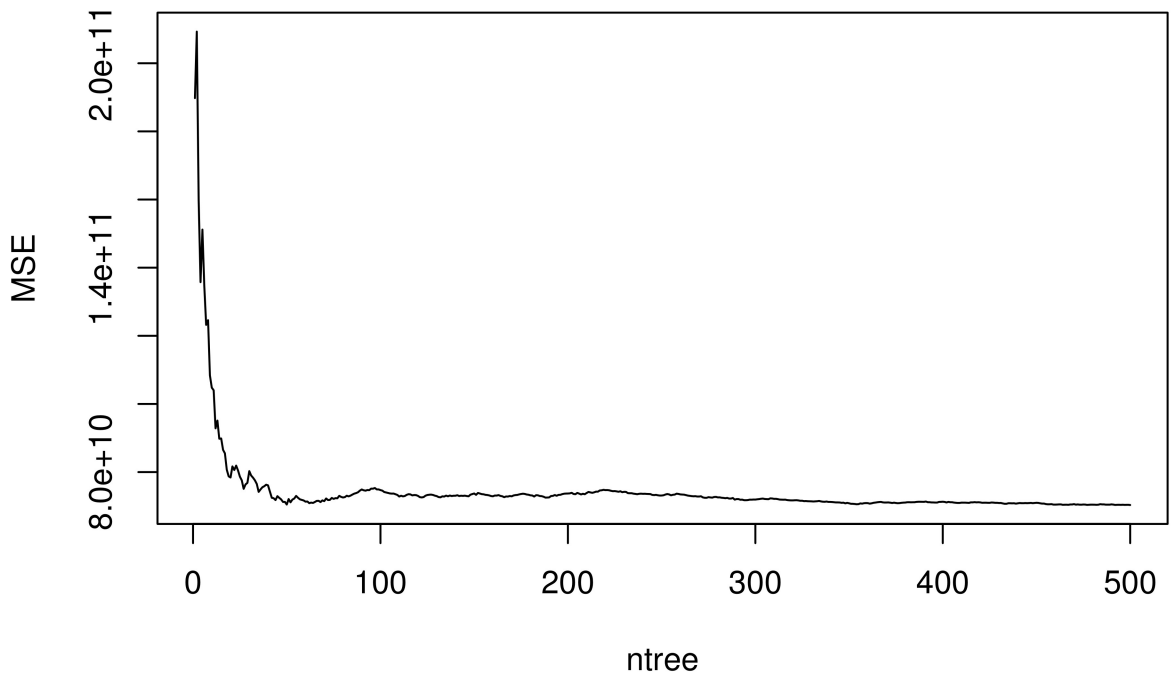
```
## [1] "price"      "bed"        "bath"       "acre_lot"   "city"
## [6] "state"      "house_size" "year"
```

```
set.seed(652) # make results reproducible
rf1 <- randomForest(price ~ ., data = newyork2, importance = TRUE)
rf1
```

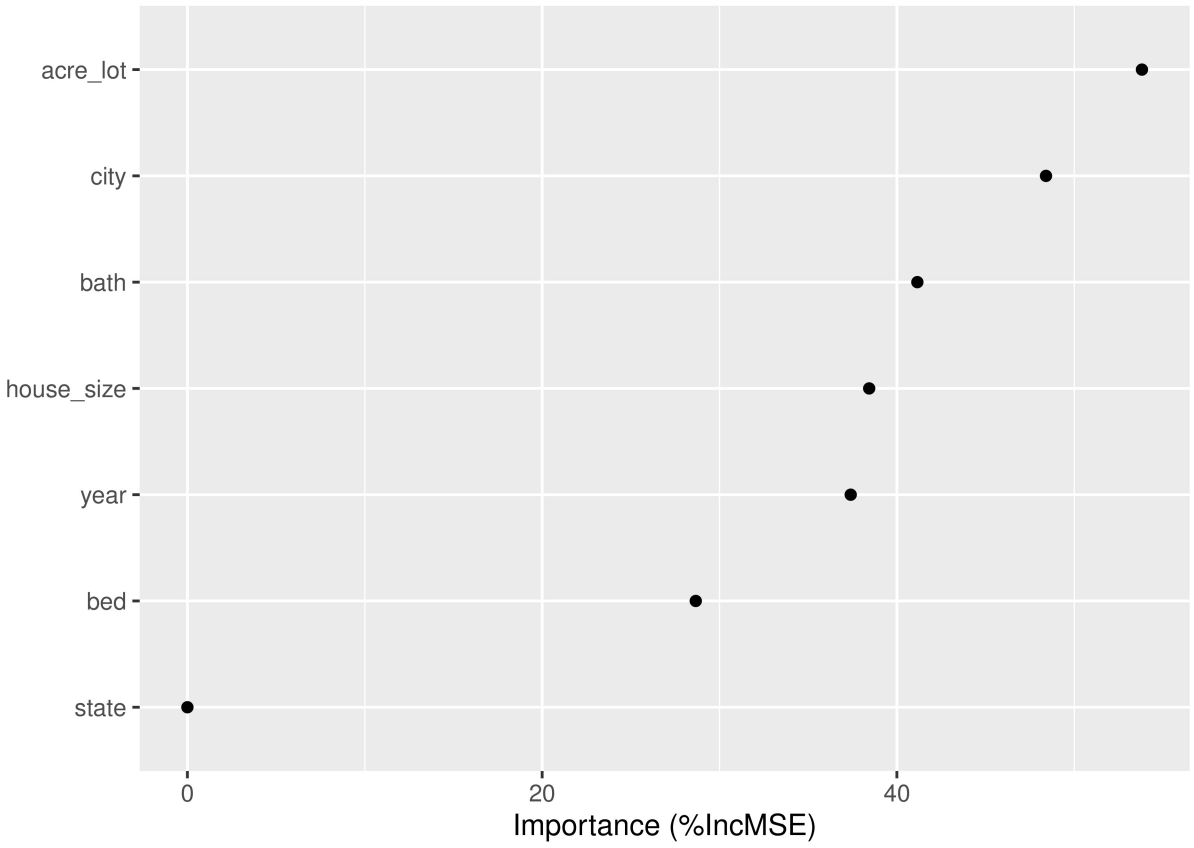
```
##
```

```
## Call:
##  randomForest(formula = price ~ ., data = newyork2, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 70342497584
##                     % Var explained: 95.1
```

```
plot(c(1: 500), rf1$mse, xlab="ntree", ylab="MSE", type="l")
```




```
vip(rf1, num_features = 8, geom = "point", include_type = TRUE)
```



Our Random Forest model shows us that 95% of the variance is explained by our out-of-bag predictions. Looking at the variable importance plot it is clear that the plot size(acre_plot) has the maximum importance compared to the rest of the variables.