Lab7.R

rstudio-user

2021-03-12

```
#1.Download the dataset boston.csv
boston <- read.csv("/cloud/project/boston.csv")
summary(boston)
```

```
##      TRACT           LON              LAT             MEDV
##  Min.   :   1   Min.   :-71.29   Min.   :42.03   Min.   : 5.00
##  1st Qu.:1303   1st Qu.:-71.09   1st Qu.:42.18   1st Qu.:17.02
##  Median :3394   Median :-71.05   Median :42.22   Median :21.20
##  Mean   :2700   Mean   :-71.06   Mean   :42.22   Mean   :22.53
##  3rd Qu.:3740   3rd Qu.:-71.02   3rd Qu.:42.25   3rd Qu.:25.00
##  Max.   :5082   Max.   :-70.81   Max.   :42.38   Max.   :50.00
##       CRIM                ZN             INDUS            CHAS
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       NOX               RM              AGE              DIS
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       RAD              TAX           PTRATIO
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40
##  Median : 5.000   Median :330.0   Median :19.05
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20
##  Max.   :24.000   Max.   :711.0   Max.   :22.00
```

```
#Null Values
any(is.na(boston))
```

```
## [1] FALSE
```

```
#2.MEDV is the output /target variable i.e price of the house to be predicted
x <- subset(boston, select=-c(MEDV))
head(x)
```

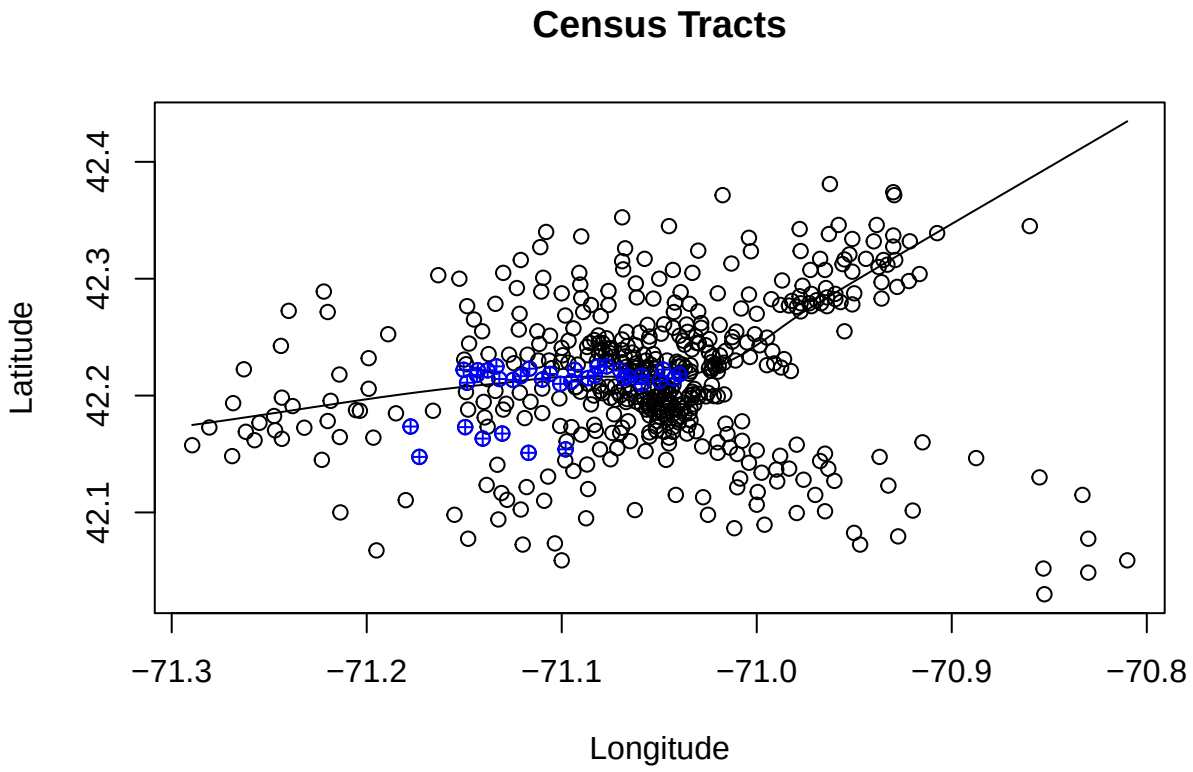
```
##   TRACT      LON     LAT    CRIM ZN INDUS CHAS   NOX    RM  AGE    DIS RAD TAX
## 1  2011 -70.9550 42.2550 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296
```

```
## 2  2021 -70.9500 42.2875 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242
## 3  2022 -70.9360 42.2830 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242
## 4  2031 -70.9280 42.2930 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222
## 5  2032 -70.9220 42.2980 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222
## 6  2033 -70.9165 42.3040 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222
##   PTRATIO
## 1    15.3
## 2    17.8
## 3    17.8
## 4    18.7
## 5    18.7
## 6    18.7
```

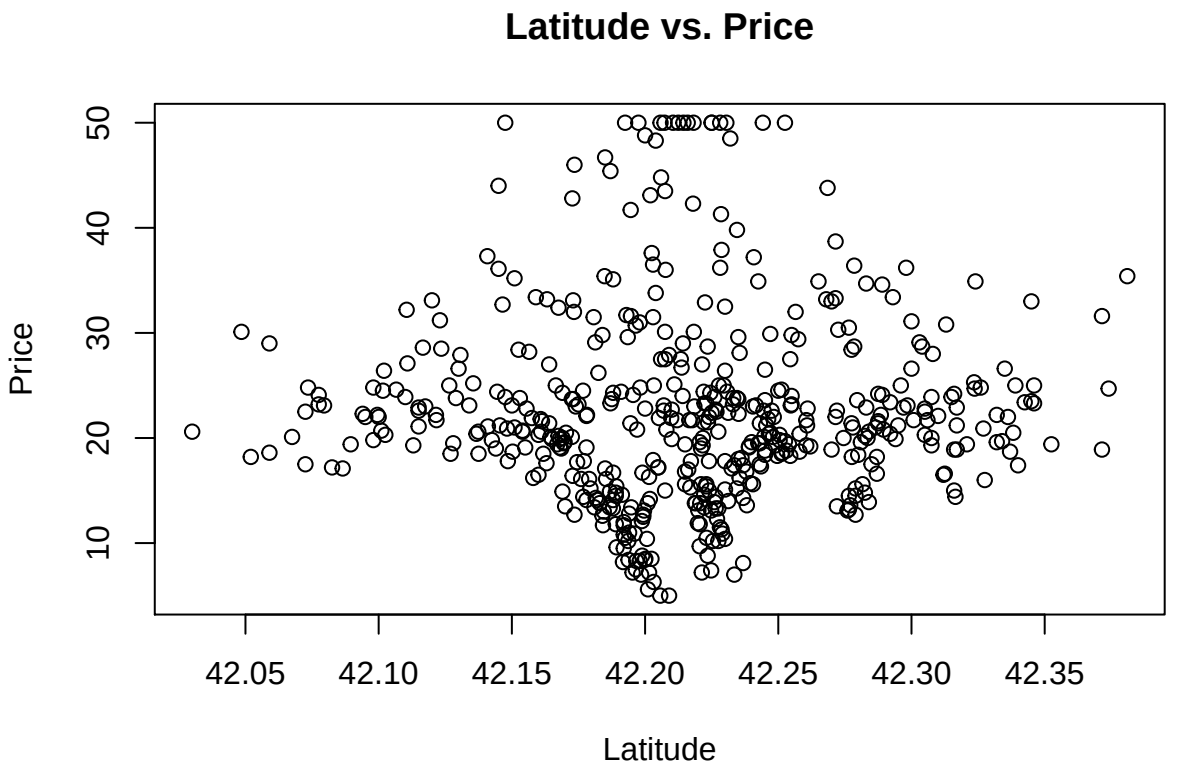
```
y <- subset(boston, select=c(MEDV))
head(y)
```

```
##   MEDV
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

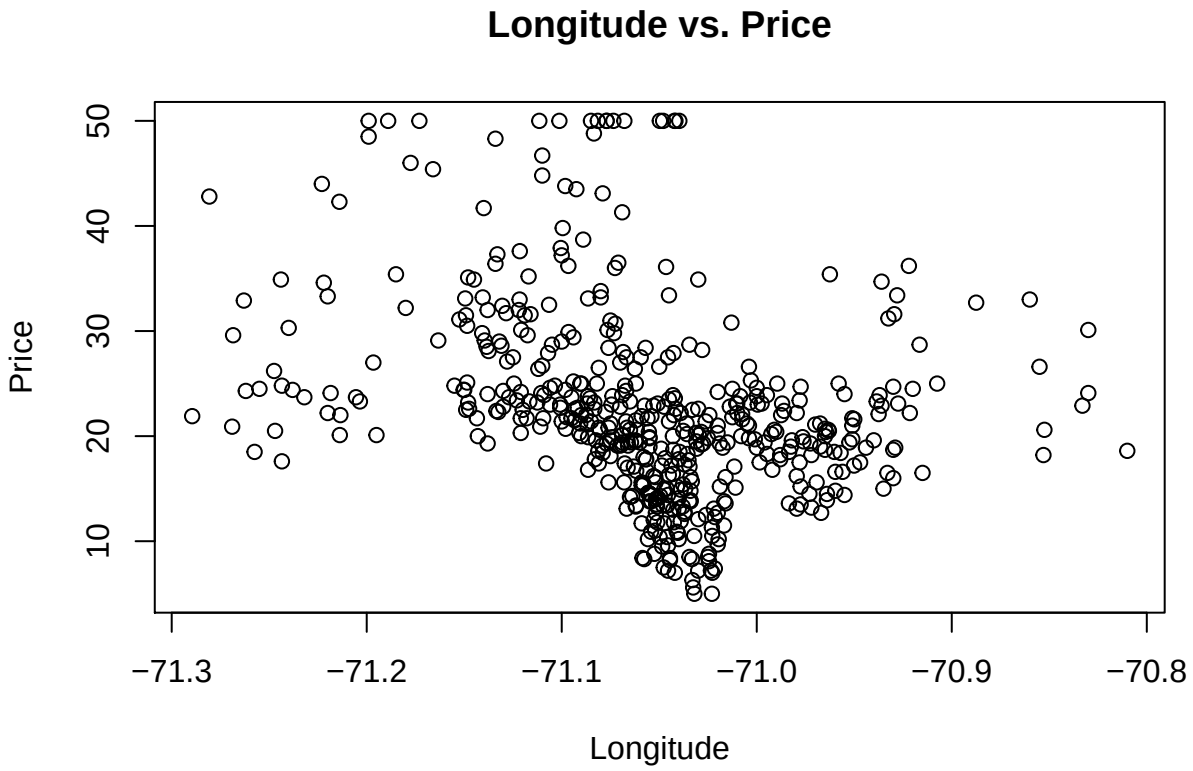
```
#3. Using the plot commands, plot the latitude and longitude of each of our
#census tracts
scatter.smooth(x$LON, x$LAT, main="Census Tracts", xlab="Longitude",
               ylab="Latitude")

#4. Show all the points that lie along the Charles River in a blue colour.
points(boston$LON[boston$CHAS==1], boston$LAT[boston$CHAS==1], col="blue",
       pch=10)
```




```
#5.Apply Linear Regression by plotting the relationship between latitude and
#house prices and the longitude and the house prices.
plot(boston$LAT, boston$MEDV, main="Latitude vs. Price", xlab="Latitude", ylab="Price")
```



```r
plot(boston$LON, boston$MEDV, main="Longitude vs. Price", xlab="Longitude", ylab="Price")
```




```r
lmmodel <- lm(MEDV ~ LAT+LON, data=boston)
summary(lmmodel)
```

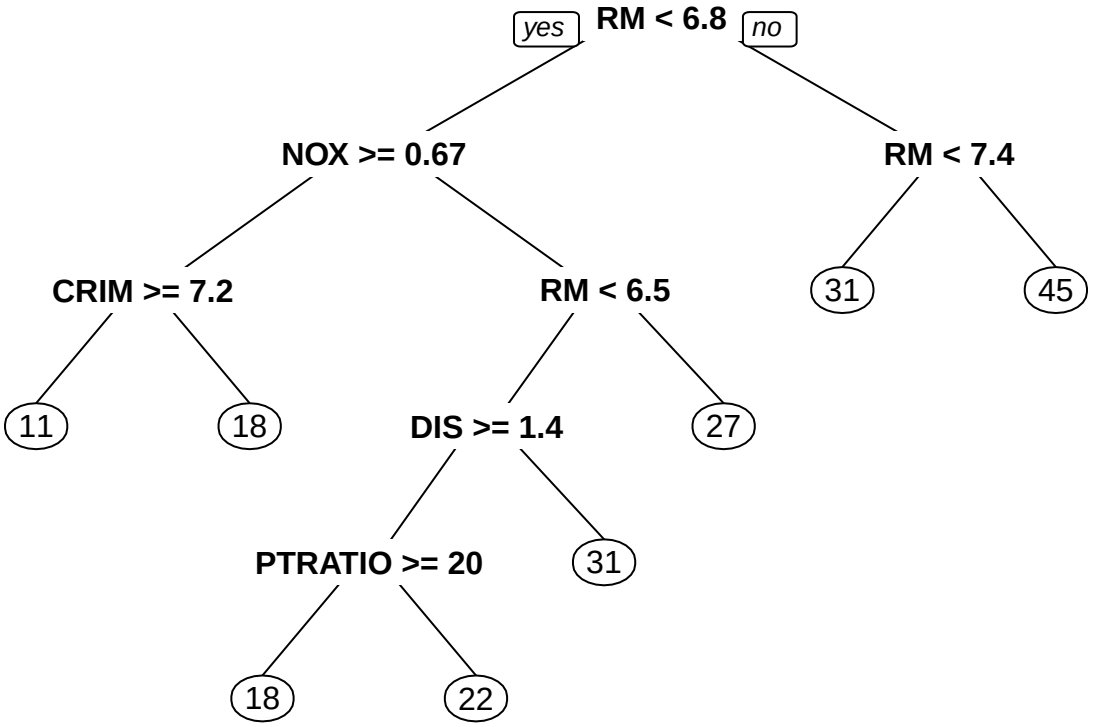
```
##
## Call:
## lm(formula = MEDV ~ LAT + LON, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.460  -5.590  -1.299   3.695  28.129
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3178.472    484.937  -6.554 1.39e-10 ***
## LAT             8.046      6.327   1.272    0.204
## LON           -40.268      5.184  -7.768 4.50e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.693 on 503 degrees of freedom
## Multiple R-squared:  0.1072, Adjusted R-squared:  0.1036
## F-statistic: 30.19 on 2 and 503 DF,  p-value: 4.159e-13
```

```r
#R squared is 0.1, which is bad
#The latitude is not significant, which means the north-south location
#differences aren't going to be really used at all. This also seems unlikely.

#Longitude is significant, but negative which means that as we go towards the
```

```
#east house prices decrease linearly, which is also unlikely.

#6.Apply Regression Tree to the problem and draw conclusions from it.
library(rpart)
install.packages("rpart.plot")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(rpart.plot)
tree = rpart(MEDV ~ LAT + LON + CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO,
prp(tree)
```




```
#We can say that the latitude and longitude arent really important. Rooms are
#the most important. Pollution appears in there twice, so it's, in some sense,
#nonlinear on the amount of pollution i.e if it's greater than a certain amount
#or less than a certain amount, it does different things. Very nonlinear on the
#number of rooms.
```