MANOJ KUMAR - 2048015

R-Laboratory 7

12/03/2021

1.Download the dataset BOSTON.csv

```
# Loading BOSTON dataset.
dataset <- read.csv("boston.csv")
head(dataset)
```

	TOWN <chr>	TRACT <int>	LON <dbl>	LAT <dbl>	MEDV <dbl>	CRIM <dbl>	ZN <dbl>	INDUS <dbl>	CHAS <int>
1	Nahant	2011	-70.9550	42.2550	24.0	0.00632	18	2.31	0
2	Swampscott	2021	-70.9500	42.2875	21.6	0.02731	0	7.07	0
3	Swampscott	2022	-70.9360	42.2830	34.7	0.02729	0	7.07	0
4	Marblehead	2031	-70.9280	42.2930	33.4	0.03237	0	2.18	0
5	Marblehead	2032	-70.9220	42.2980	36.2	0.06905	0	2.18	0
6	Marblehead	2033	-70.9165	42.3040	28.7	0.02985	0	2.18	0

6 rows | 1-10 of 17 columns

```
- LON and LAT are the longitude and latitude of the center of the census tract.
- MEDV is the median value of owner-occupied homes, measured in thousands of dollars.
- CRIM is the per capita crime rate.
- ZN is related to how much of the land is zoned for large residential properties.
- INDUS is the proportion of the area used for industry.
- CHAS is 1 if a census tract is next to the Charles River else 0
- NOX is the concentration of nitrous oxides in the air, a measure of air pollution.
- RM is the average number of rooms per dwelling.
- AGE is the proportion of owner-occupied units built before 1940.
- DIS is a measure of how far the tract is from centres of employment in Boston.
- RAD is a measure of closeness to important highways.
- TAX is the property tax per $10,000 of value.
- PTRATIO is the pupil to teacher ratio by town
```

```
str(dataset)
```

```
## 'data.frame':    506 obs. of  16 variables:
##  $ TOWN   : chr  "Nahant" "Swampscott" "Swampscott" "Marblehead" ...
##  $ TRACT  : int  2011 2021 2022 2031 2032 2033 2041 2042 2043 2044 ...
##  $ LON    : num  -71 -71 -70.9 -70.9 -70.9 ...
##  $ LAT    : num  42.3 42.3 42.3 42.3 42.3 ...
##  $ MEDV   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 22.1 16.5 18.9 ...
##  $ CRIM   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ ZN     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ INDUS  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ CHAS   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ NOX    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ RM     : num  6.58 6.42 7.18 7 7.15 ...
##  $ AGE    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ DIS    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ RAD    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ TAX    : int  296 242 242 222 222 222 311 311 311 311 ...
##  $ PTRATIO: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
```

```
summary(dataset)
```

```
##      TOWN               TRACT           LON              LAT       
##  Length:506         Min.   :   1   Min.   :-71.29   Min.   :42.03  
##  Class :character   1st Qu.:1303   1st Qu.:-71.09   1st Qu.:42.18  
##  Mode  :character   Median :3394   Median :-71.05   Median :42.22  
##                     Mean   :2700   Mean   :-71.06   Mean   :42.22  
##                     3rd Qu.:3740   3rd Qu.:-71.02   3rd Qu.:42.25  
##                     Max.   :5082   Max.   :-70.81   Max.   :42.38  
##       MEDV            CRIM                ZN             INDUS      
##  Min.   : 5.00   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46  
##  1st Qu.:17.02   1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19  
##  Median :21.20   Median : 0.25651   Median :  0.00   Median : 9.69  
##  Mean   :22.53   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14  
##  3rd Qu.:25.00   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10  
##  Max.   :50.00   Max.   :88.97620   Max.   :100.00   Max.   :27.74  
##       CHAS              NOX               RM             AGE        
##  Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
##  1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
##  Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
##  Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
##  3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
##  Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
##       DIS              RAD              TAX           PTRATIO      
##  Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60  
##  1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40  
##  Median : 3.207   Median : 5.000   Median :330.0   Median :19.05  
##  Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46  
##  3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20  
##  Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00
```

```
# Checking for missing values.
colSums(is.na(dataset))
```

```
##    TOWN   TRACT     LON     LAT    MEDV    CRIM      ZN   INDUS    CHAS     NOX 
##       0       0       0       0       0       0       0       0       0       0 
##      RM     AGE     DIS     RAD     TAX PTRATIO 
##       0       0       0       0       0       0
```

```
colSums(dataset=='')
```

```
##    TOWN   TRACT     LON     LAT    MEDV    CRIM      ZN   INDUS    CHAS     NOX 
##       0       0       0       0       0       0       0       0       0       0 
##      RM     AGE     DIS     RAD     TAX PTRATIO 
##       0       0       0       0       0       0
```

Insight

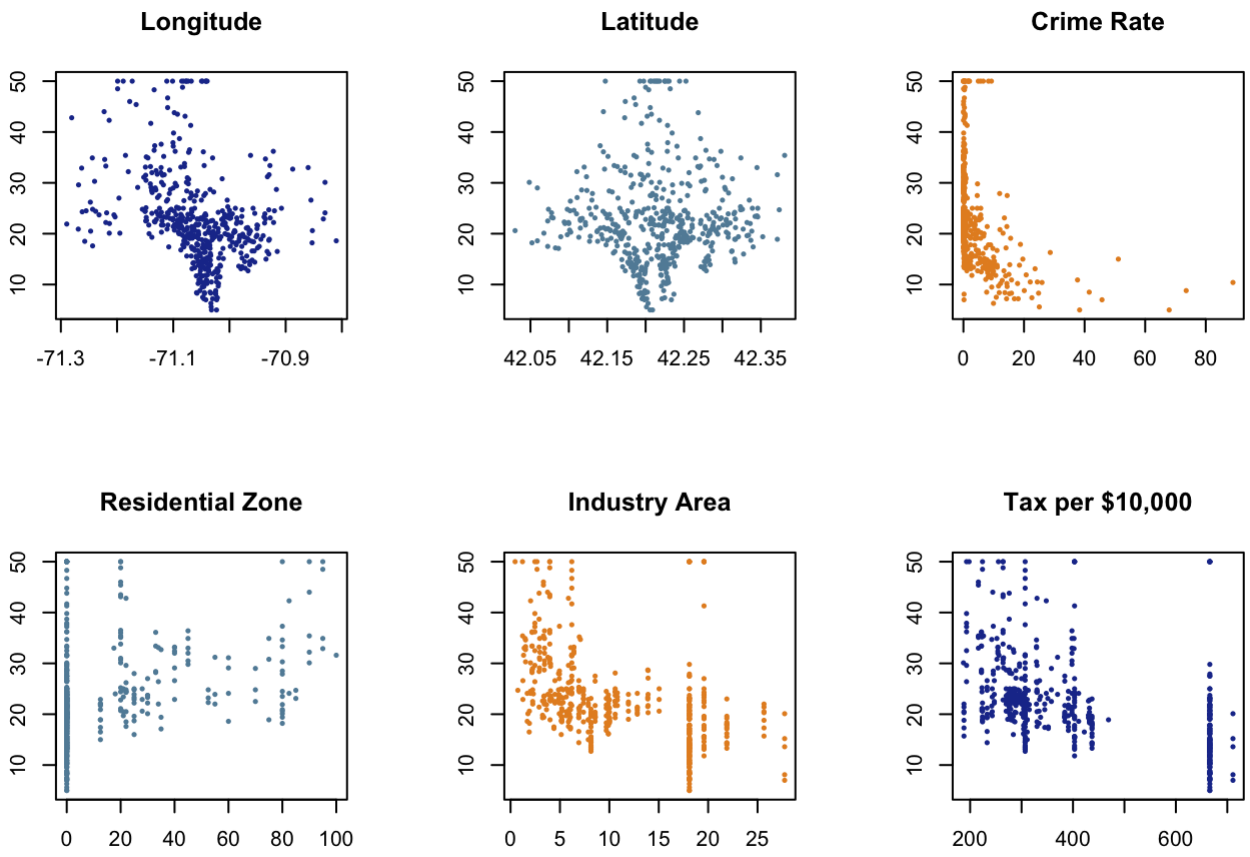
```
-Character  - TOWN
-Integer     - TRACT,CHAS,RAD,TAX
-Numeric     - LON,LAT,MEDV,CRIM,ZN,INDUS,NOX,RM,AGE,DIS,PTRATIO
-Number of columns:16 5.Number of observation:506

-The house price ranges between 5000 and 50000 dollars.
-The Longitude and Latitude lies on -71 and 42 coordinate points respectively in Boston.
```

2.MEDV is the output /target variable i.e price of the house to be predicted.

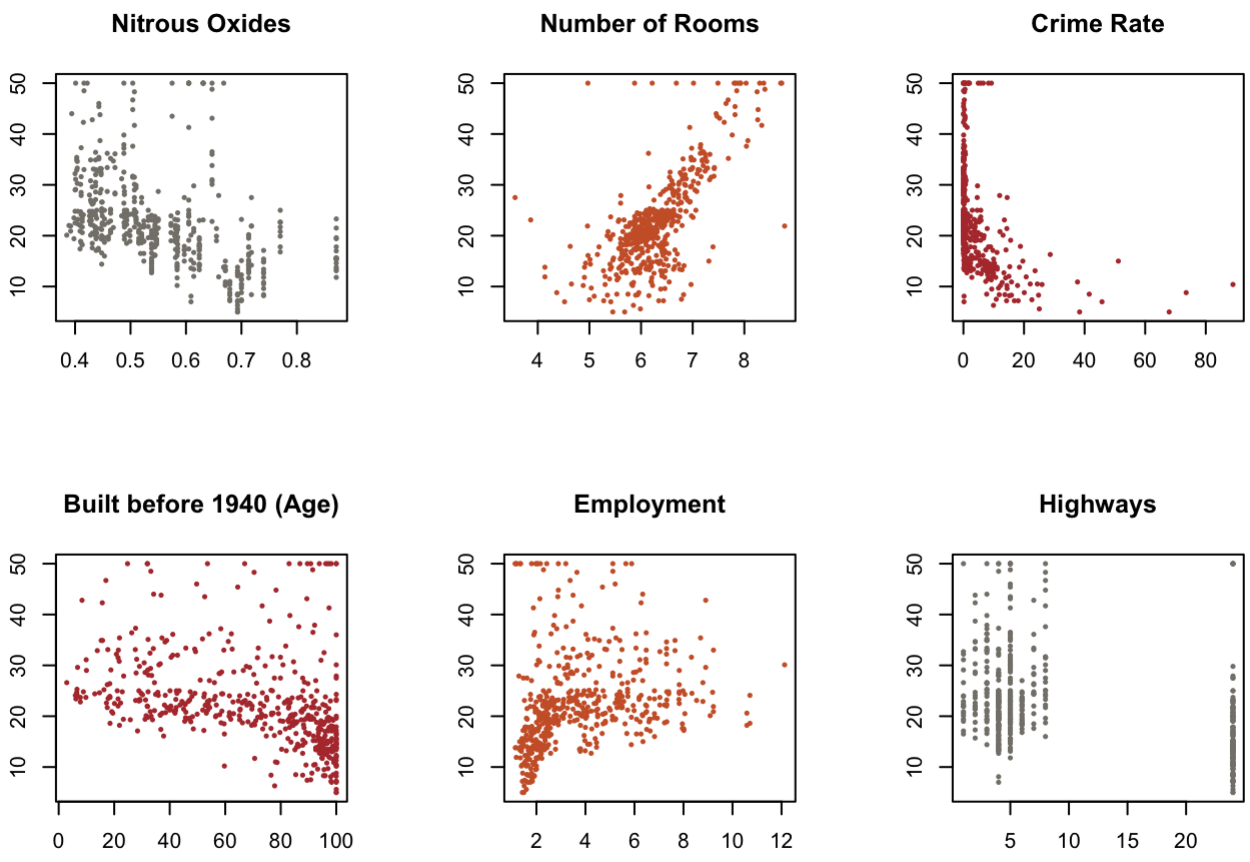
```
par(mfrow=c(2,3))

plot(dataset$LON, dataset$MEDV, main = "Longitude", xlab="", ylab="", cex.lab="1.5", col="#1e3799", pch=20, cex=
0.5 )
plot(dataset$LAT, dataset$MEDV, main = "Latitude", xlab="", ylab="", cex.lab="1.5", col="#628ca6", pch=20, cex=
0.5)
plot(dataset$CRIM, dataset$MEDV, main = "Crime Rate", xlab="", ylab="", cex.lab="1.5", col="#e58e26", pch=20, cex
= 0.5)
plot(dataset$ZN, dataset$MEDV, main = "Residential Zone", xlab="", ylab="", cex.lab="1.5", col="#628ca6", pch=20,
cex= 0.5)
plot(dataset$INDUS, dataset$MEDV, main = "Industry Area", xlab="", ylab="", cex.lab="1.5", col="#e58e26", pch=20,
cex= 0.5)
plot(dataset$TAX, dataset$MEDV, main = "Tax per $10,000", xlab="", ylab="", cex.lab="1.5", col="#1e3799", pch=20,
cex= 0.5)
```




```
par(mfrow=c(2,3))

plot(dataset$NOX, dataset$MEDV, main = "Nitrous Oxides", xlab="", ylab="", cex.lab="1.5", col="#84817a", pch=20,
 cex= 0.5 )
plot(dataset$RM, dataset$MEDV, main = "Number of Rooms", xlab="", ylab="", cex.lab="1.5", col="#cd6133", pch=20,
 cex= 0.5 )
plot(dataset$CRIM, dataset$MEDV, main = "Crime Rate", xlab="", ylab="", cex.lab="1.5", col="#b33939", pch=20, cex
= 0.5 )
plot(dataset$AGE, dataset$MEDV, main = "Built before 1940 (Age)", xlab="", ylab="", cex.lab="1.5", col="#b33939",
pch=20, cex= 0.5 )
plot(dataset$DIS, dataset$MEDV, main = "Employment", xlab="", ylab="", cex.lab="1.5", col="#cd6133", pch=20, cex=
0.5 )
plot(dataset$RAD, dataset$MEDV, main = "Highways", xlab="", ylab="", cex.lab="1.5", col="#84817a", pch=20, cex=
0.5 )
```

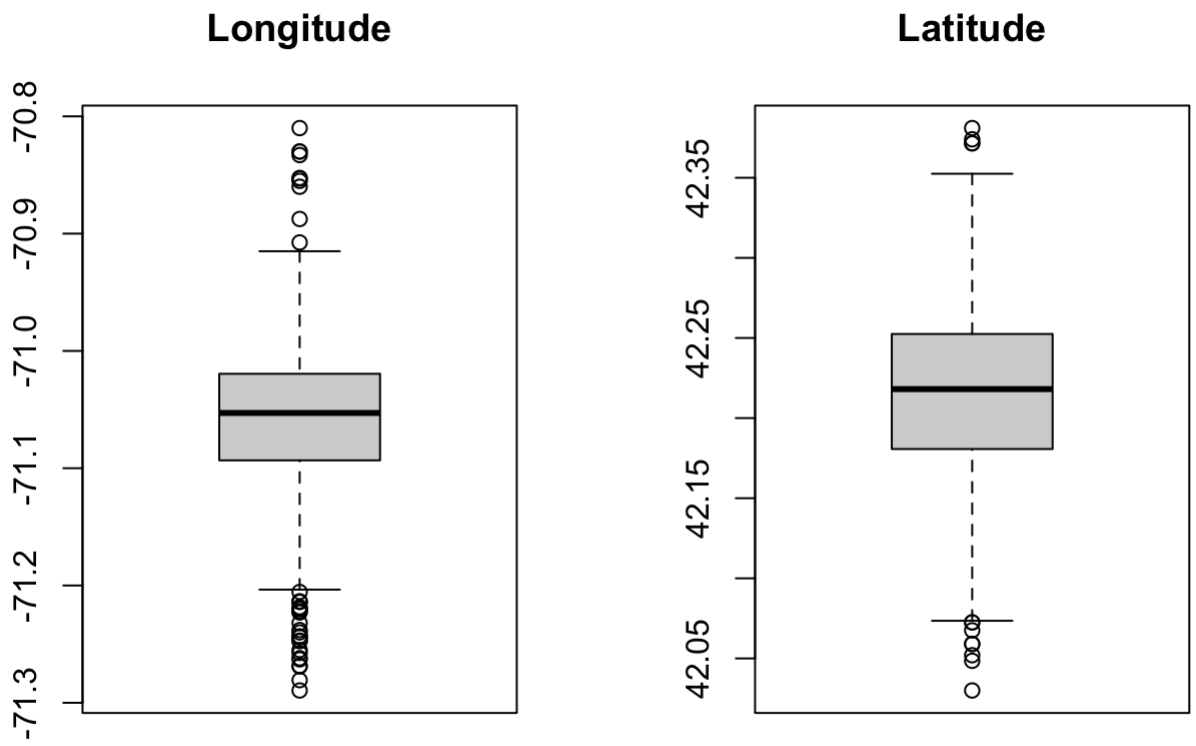


Insight

```
  -There is an observation that when house prices go down there is a significant increase in crime rate.
  -There is presence of nitous oxide in air which cause air pollution.And its normally distributed in all range o
f house price regions.
  -There is higher probability of employment opportunity when the housing price significantly increases.
```

```
par(mfrow=c(1,2))

boxplot(dataset$LON, main = "Longitude")
boxplot(dataset$LAT, main = "Latitude")
```




```
outliers <- boxplot(dataset$LON, plot=FALSE)$out
dataset[which(dataset$LON %in% outliers),]
```

	TOWN <chr>	TRACT <int>	LON <dbl>	LAT <dbl>	MEDV <dbl>	CRIM <dbl>	ZN <dbl>	INDUS <dbl>	CHAS <int>
64	Beverly	2176	-70.9075	42.3390	25.0	0.12650	25.0	5.13	0
65	Manchester	2181	-70.8600	42.3450	33.0	0.01951	17.5	1.38	0
197	Concord	3611	-71.2200	42.2715	33.3	0.04011	80.0	1.52	0
198	Concord	3612	-71.2400	42.2725	30.3	0.04666	80.0	1.52	0
199	Concord	3613	-71.2220	42.2890	34.6	0.03768	80.0	1.52	0
200	Sudbury	3651	-71.2440	42.2425	34.9	0.03150	95.0	1.47	0
201	Sudbury	3652	-71.2630	42.2225	32.9	0.01778	95.0	1.47	0
202	Wayland	3661	-71.2185	42.1955	24.1	0.03445	82.5	2.03	0
203	Wayland	3662	-71.2140	42.2180	42.3	0.02177	82.5	2.03	0
239	Natick	3821	-71.2055	42.1875	23.7	0.08244	30.0	4.93	0

1-10 of 35 rows | 1-10 of 17 columns Previous **1** 2 3 4 Next

```
dataset <- dataset[-which(dataset$LON %in% outliers),]
```

```
outliers <- boxplot(dataset$LAT, plot=FALSE)$out
dataset[which(dataset$LAT %in% outliers),]
```

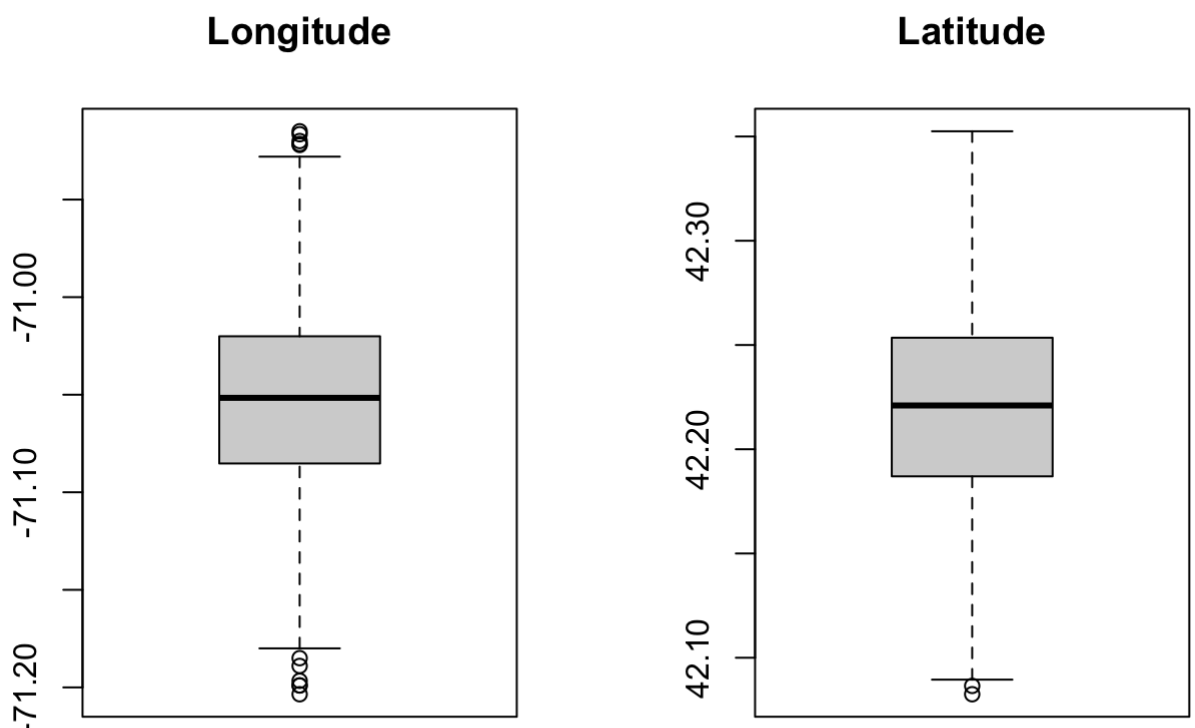
	TOWN <chr>	TRACT <int>	LON <dbl>	LAT <dbl>	MEDV <dbl>	CRIM <dbl>	ZN <dbl>	INDUS <dbl>	CHAS <int>
55	Middleton	2121	-71.0175	42.3715	18.9	0.01360	75.0	4.00	0
56	Topsfield	2141	-70.9625	42.3810	35.4	0.01311	90.0	1.22	0
57	Hamilton	2151	-70.9300	42.3740	24.7	0.02055	85.0	0.74	0
58	Wenham	2161	-70.9295	42.3715	31.6	0.01432	100.0	1.32	0
287	Norfolk	4091	-71.1950	42.0675	20.1	0.01965	80.0	1.76	0
288	Walpole	4111	-71.1480	42.0775	23.2	0.03871	52.5	5.32	0
299	Sharon	4141	-71.1200	42.0725	22.5	0.06466	70.0	2.24	0
300	Sharon	4142	-71.1000	42.0590	29.0	0.05561	70.0	2.24	0
301	Sharon	4143	-71.1035	42.0735	24.8	0.04417	70.0	2.24	0
346	Rockland	5021	-70.9470	42.0725	17.5	0.03113	0.0	4.39	0

1-10 of 11 rows | 1-10 of 17 columns Previous **1** 2 Next

```
dataset <- dataset[-which(dataset$LAT %in% outliers),]
```

```
par(mfrow=c(1,2))

boxplot(dataset$LON, main = "Longitude")
boxplot(dataset$LAT, main = "Latitude")
```



Insight

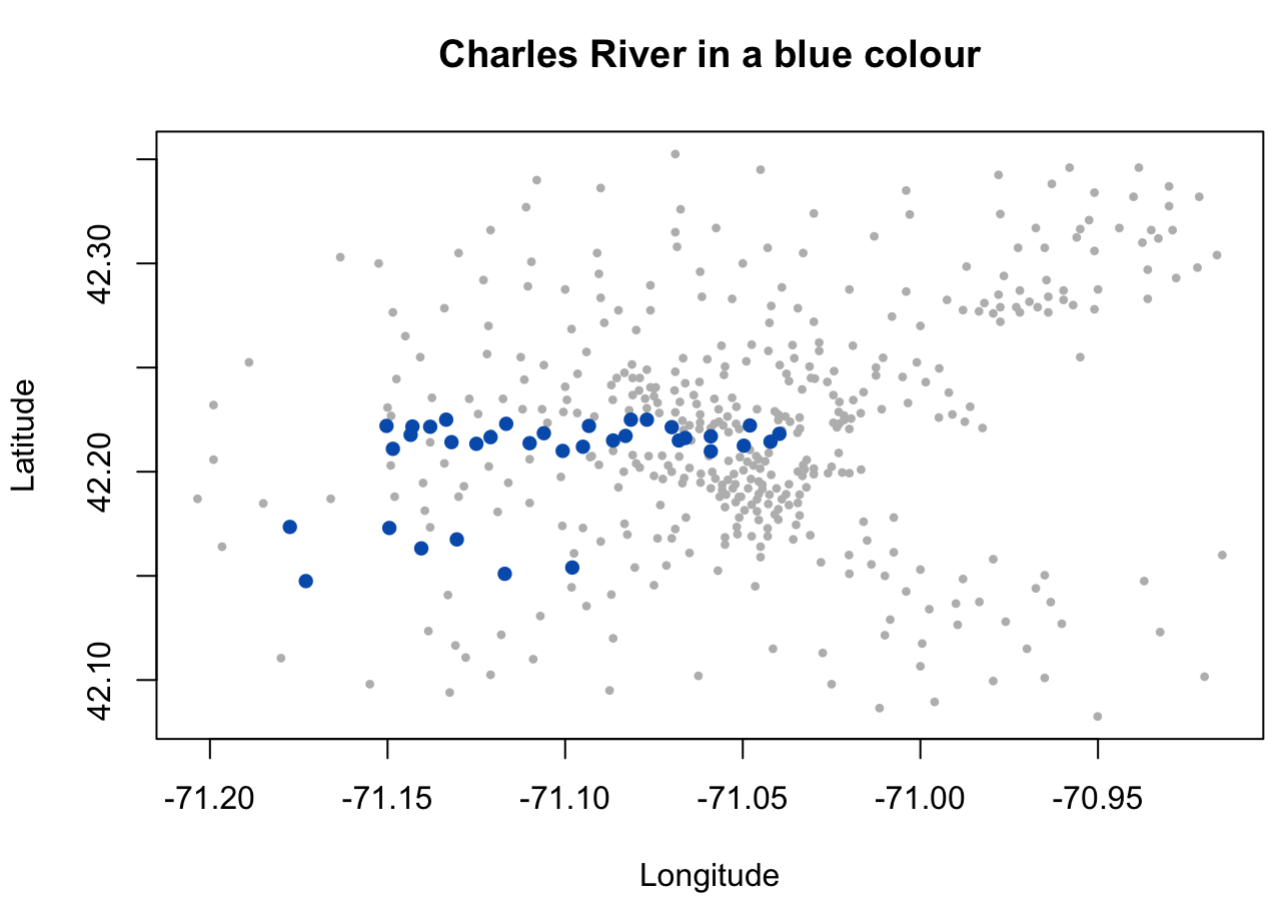
```
  -Outliers are spotted out using boxplot. Here we notice that longitude values have outlier where as latitude do
esnt have. So the outliers are firstly spotted and removed from the dataset.
```

3. Using the plot commands, plot the latitude and longitude of each of our census tracts.

4. Show all the points that lie along the Charles River in a blue colour.

```
# https://colorhunt.co/

plot(dataset$LON, dataset$LAT, main = "Charles River in a blue colour",  xlab="Longitude", ylab="Latitude", col=
"#bbbbbb", pch=20, cex= 0.7)
points(dataset$LON[dataset$CHAS==1], dataset$LAT[dataset$CHAS==1], col="#005fba", pch=16)
```



Insight

-It shows that there is no linearity and moderate relationship between latitude&longitude over house locations.
-Blue colored highlighted house locations were generally lies along the Charles River.

5.Apply Linear Regression by plotting the relationship between latitude and house prices and the longitude and the house prices.

```
library(plotly)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'plotly'
```

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

```
## The following object is masked from 'package:stats':
##
##     filter
```

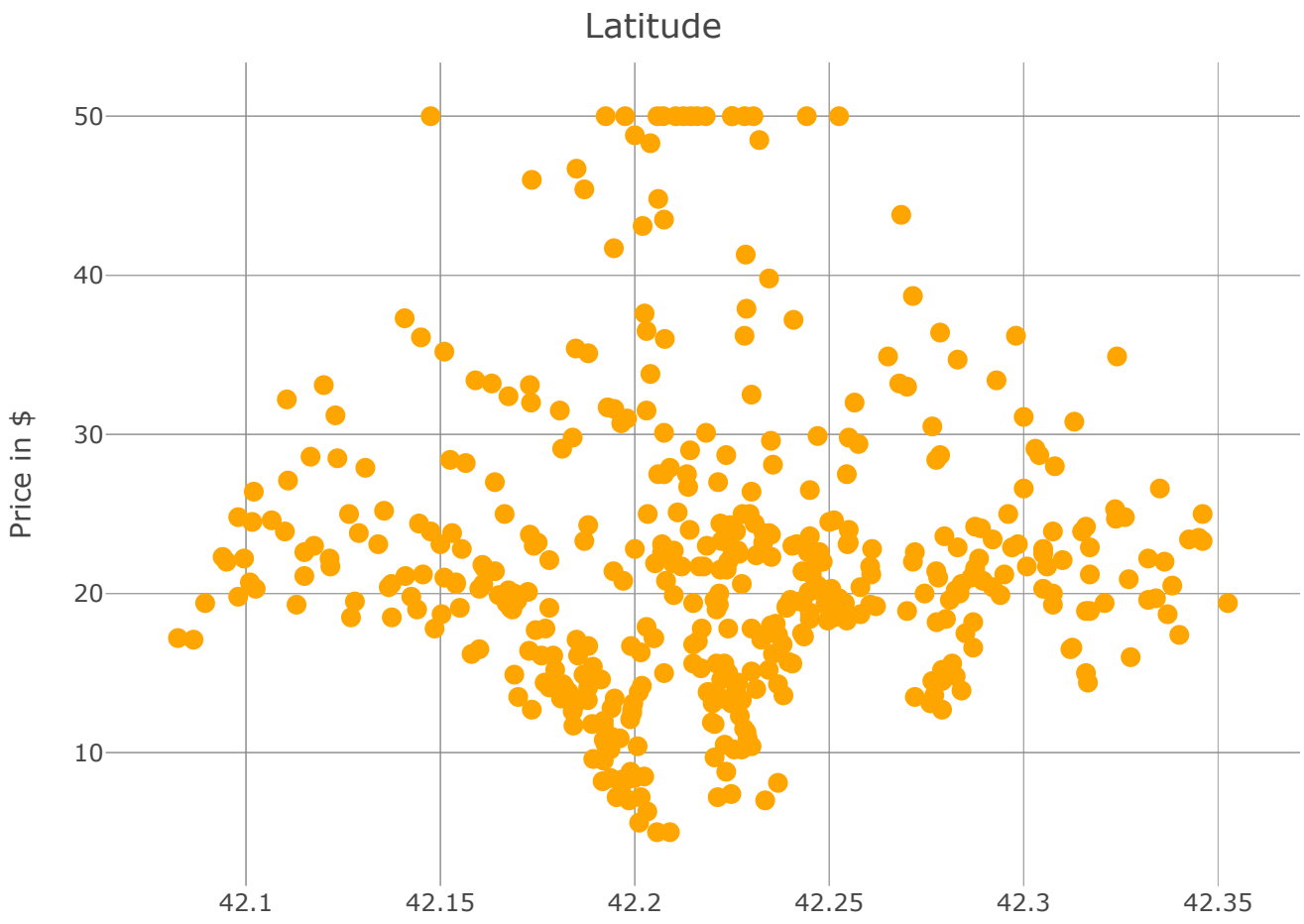
```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
fig <- plot_ly(data = dataset, x = ~dataset$LAT, y = ~dataset$MEDV, marker = list(size = 10,color = '#ffa500'))
fig <- fig %>% layout(title = 'Latitude', yaxis = list(title = "Price in $", zeroline = TRUE),xaxis = list(title
 = "", zeroline = TRUE))

fig
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#scatter
```

```
## No scatter mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```

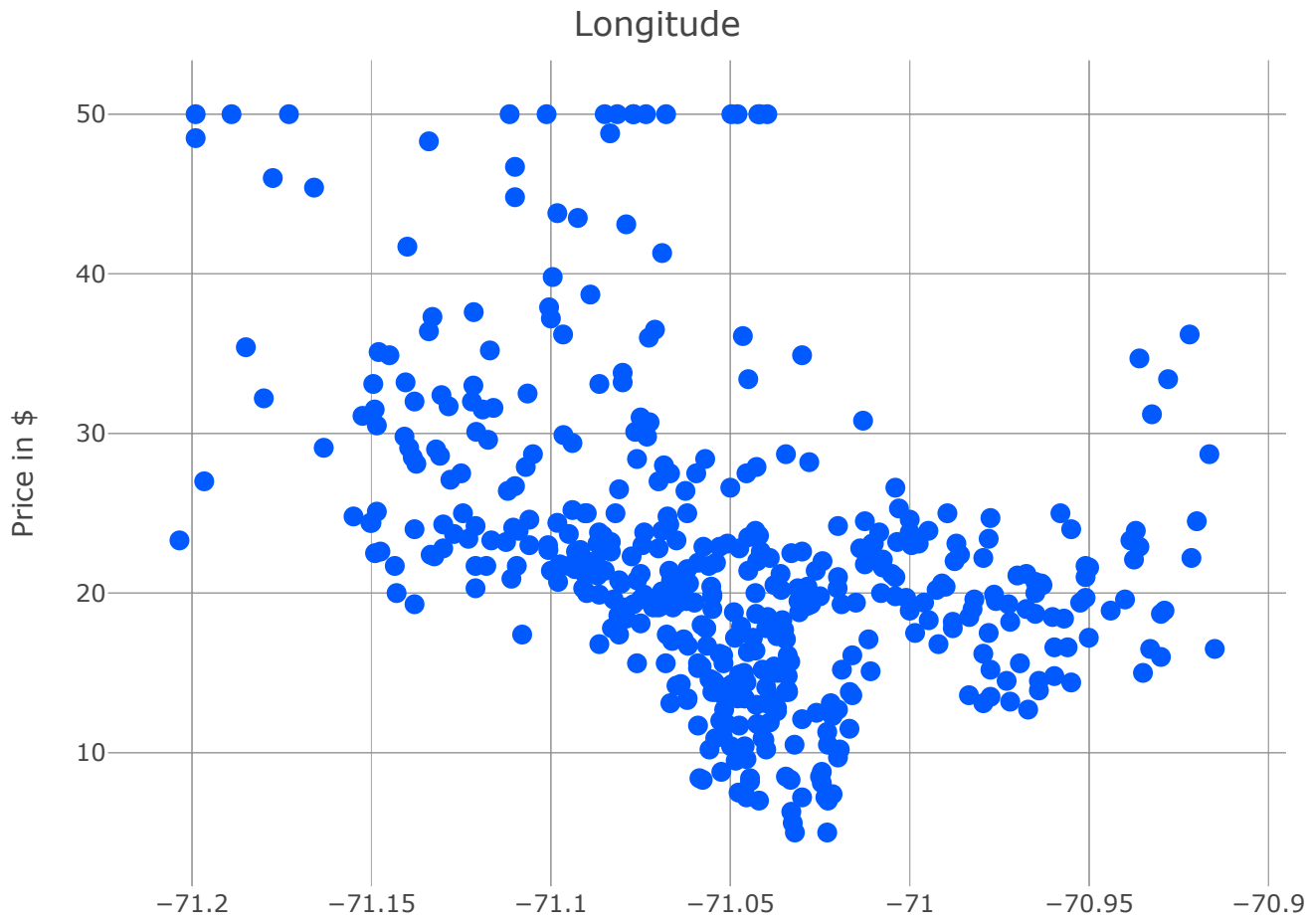



```
fig <- plot_ly(data = dataset, x = ~dataset$LON, y = ~dataset$MEDV, marker = list(size = 10,color = '#005AFF'))
fig <- fig %>% layout(title = 'Longitude', yaxis = list(title = "Price in $", zeroline = TRUE),xaxis = list(title
= "", zeroline = TRUE))

fig
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#scatter
```

```
## No scatter mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```




```
#linear model
linear_model = lm(MEDV ~ LAT + LON, data=dataset)
summary(linear_model)
```

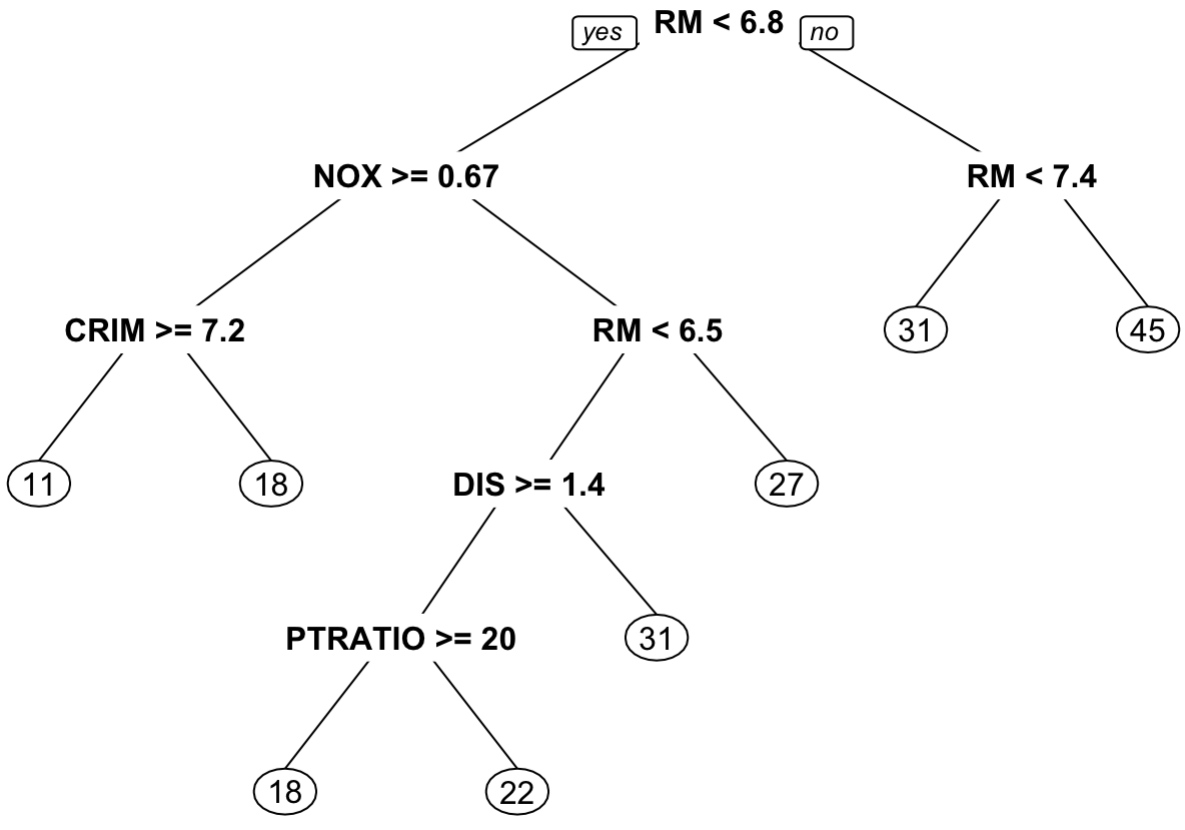
```
##
## Call:
## lm(formula = MEDV ~ LAT + LON, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.616  -5.677  -1.039   3.807  28.677
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5481.508    651.768  -8.410 5.27e-16 ***
## LAT            13.983      7.311   1.913   0.0564 .
## LON           -69.151      7.214  -9.585  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.521 on 457 degrees of freedom
## Multiple R-squared:  0.1674, Adjusted R-squared:  0.1638
## F-statistic: 45.95 on 2 and 457 DF,  p-value: < 2.2e-16
```

Insight

6.Apply Regression Tree to the problem and draw conclusions from it.

```
library(rpart)
library(rpart.plot)
```

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



Insight

-The variables RM, NOX, AGE, CRIM, DIS are the variables used to predict the correspoding output- MEDV. According to Variable Importance, the variables in decreasing order of the priority: RM, NOX, CRIM, INDUS, PTRATIO, DIS, TAX, RAD, ZN, LON, CHAS. Among them the first 6 variables are taken into consideration. The set of approximate values that are predicted are 11,18,26,22,27,32,46.