Data story of Boston housing pricing

Nikesh Dubey

DATA DICTIONARY The Boston data frame has 506 rows and 14 columns.

This data frame contains the following columns:

crim :

per capita crime rate by town.

zn :

proportion of residential land zoned for lots over 25,000 sq.ft.

indus :

proportion of non-retail business acres per town.

chas :

Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox :

nitrogen oxides concentration (parts per 10 million).

rm :

average number of rooms per dwelling.

age :

proportion of owner-occupied units built prior to 1940.

dis :

weighted mean of distances to five Boston employment centres.

rad :

index of accessibility to radial highways.

tax :

full-value property-tax rate per $10,000.

ptratio :

pupil-teacher ratio by town.

black :

1000(Bk - 0.63)ˆ2 where Bk is the proportion of blacks by town.

lstat :

lower status of the population (percent).

medv :

median value of owner-occupied homes in $1000s.

PART 1 : Data cleaning

Importing libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Importing the data

We have assigned NA to blank values and to NA which could have been defined as string values,while reading the file itself in csv format.So that we can take into consideration of all the junk values as NA

```
df=read.csv("BostonHousing.csv",stringsAsFactors =TRUE,na.strings=c("NA",""))
```

Checking column names

To check if any renaming is necessary

```
names(df)
```

```
##  [1] "CRIM"    "ZN"      "INDUS"   "CHAS"    "NOX"     "RM"      "AGE"
##  [8] "DIS"     "RAD"     "TAX"     "PTRATIO" "B"       "LSTAT"   "MEDV"
```

Since the column names are only abbreviations, we have decided to rename them to more explainable terms

Renaming column heads

```
df1=rename(df,crime_rate=CRIM,residential_landzone=ZN,non_retail_land=INDUS,
           Charles_River=CHAS,nox_concentration=NOX,avg_no_of_rooms=RM,
           owner_age=AGE,dis_employment=DIS,access_to_highways=RAD,
           tax=TAX,PTRATIO=PTRATIO,black_people=B,
           lower_people=LSTAT,owner_occupied_homes_price=MEDV)
```

To view the renamed data frame

```
View(df1)
```

Column names look good now, we can go ahead with the cleaning of the data.There could be a lot of NA values in the data frame,which could distort our analysis.

Checking number of NA values

```
sum(is.na(df1))
```

```
## [1] 120
```

So there are a total of 120 NA values in our dataset.Now we would like to know,which of these columns have NA values , so as to get a better idea of removing NA values is feasible or not, For example, we would have to remove a whole column if all the 120 NA values are present in one single column.

To find out which columns have NA values

```
sum(is.na(df1$crime_rate))
```

```
## [1] 20
```

```
sum(is.na(df1$residential_landzone))
```

```
## [1] 20
```

```
sum(is.na(df1$non_retail_land))
```

```
## [1] 20
```

```
sum(is.na(df1$Charles_River))
```

```
## [1] 20
```

```
sum(is.na(df1$owner_age))
```

```
## [1] 20
```

```
sum(is.na(df1$lower_people))
```

```
## [1] 20
```

So, now we know that 6 columns crime_rate,residential_landzone,non_retail_land,Charles_River,owner_age,lower_people are the columns, which have NA values. Interesting thing is that, each of them have exactly 20 NA values.

Removing NA values

```
df2=na.omit(df1)
```

Again checking for NA values

```
sum(is.na(df2))
```

```
## [1] 0
```

Now since the sum is zero, all the NA values have been removed from our data frame and we have assigned the new dataframe to df2.

NA values have been removed, but there might be some duplicate values in our dataset,we need to check them to by using dublicate function from dplyr library

No. of duplicate values

```
sum(duplicated(df2))
```

```
## [1] 0
```

So there aren't any duplicate values. That's great !! Now we can finally start peeping into our dataset

Peeping into the data-Exploratory Data Analysis

This is a crucial part and usually takes up most of the time. A proper and extensive Exploratory data analysis would reveal interesting patterns which are underlying in the data. Now let's perform some exploratory data analysis to understand how the variables of the data are related to one another.

```
str(df2)
```

```
## 'data.frame':    394 obs. of  14 variables:
##  $ crime_rate               : num  0.00632 0.02731 0.02729 0.03237 0.02985 ...
##  $ residential_landzone     : num  18 0 0 0 0 12.5 12.5 12.5 12.5 12.5 ...
##  $ non_retail_land          : num  2.31 7.07 7.07 2.18 2.18 7.87 7.87 7.87 7.87 7.87 ...
##  $ Charles_River            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox_concentration        : num  0.538 0.469 0.469 0.458 0.458 0.524 0.524 0.524 0.524 0.524 ...
##  $ avg_no_of_rooms          : num  6.58 6.42 7.18 7 6.43 ...
##  $ owner_age                : num  65.2 78.9 61.1 45.8 58.7 96.1 100 94.3 82.9 39 ...
##  $ dis_employment           : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ access_to_highways       : int  1 2 2 3 3 5 5 5 5 5 ...
##  $ tax                      : int  296 242 242 222 222 311 311 311 311 311 ...
##  $ PTRATIO                  : num  15.3 17.8 17.8 18.7 18.7 15.2 15.2 15.2 15.2 15.2 ...
##  $ black_people             : num  397 397 393 395 394 ...
##  $ lower_people             : num  4.98 9.14 4.03 2.94 5.21 ...
##  $ owner_occupied_homes_price: num  24 21.6 34.7 33.4 28.7 27.1 16.5 15 18.9 21.7 ...
##  - attr(*, "na.action")= 'omit' Named int [1:112] 5 7 10 15 36 37 44 48 52 54 ...
##   ..- attr(*, "names")= chr [1:112] "5" "7" "10" "15" ...
```

```
head(df2)
```

```
##   crime_rate residential_landzone non_retail_land Charles_River
## 1    0.00632                18.0           2.31             0
## 2    0.02731                 0.0           7.07             0
## 3    0.02729                 0.0           7.07             0
## 4    0.03237                 0.0           2.18             0
## 6    0.02985                 0.0           2.18             0
## 8    0.14455                12.5           7.87             0
##   nox_concentration avg_no_of_rooms owner_age dis_employment access_to_highways
## 1             0.538           6.575      65.2         4.0900                  1
## 2             0.469           6.421      78.9         4.9671                  2
## 3             0.469           7.185      61.1         4.9671                  2
## 4             0.458           6.998      45.8         6.0622                  3
## 6             0.458           6.430      58.7         6.0622                  3
## 8             0.524           6.172      96.1         5.9505                  5
##   tax PTRATIO black_people lower_people owner_occupied_homes_price
## 1 296    15.3       396.90         4.98                       24.0
## 2 242    17.8       396.90         9.14                       21.6
## 3 242    17.8       392.83         4.03                       34.7
## 4 222    18.7       394.63         2.94                       33.4
```

```
## 6 222    18.7       394.12         5.21                       28.7
## 8 311    15.2       396.90        19.15                       27.1
```

From the str function, we can see that all the data types in all the column are numerical data type, which are correct as per the values available in them, further there are no categorical values in any of the columns, so no need to use as.factor() fuction to convert them into categorical variables

PART 2-Analysis of the data

summary to get a glimopse of each columns individually

Q1.What are the range of values in each column,are there any outliers in any columns ?

A command called summary gives the basic statistics of the dataset like mean, median, 1st quartile, 2nd quartile etc of each column.

```
summary(df2)
```

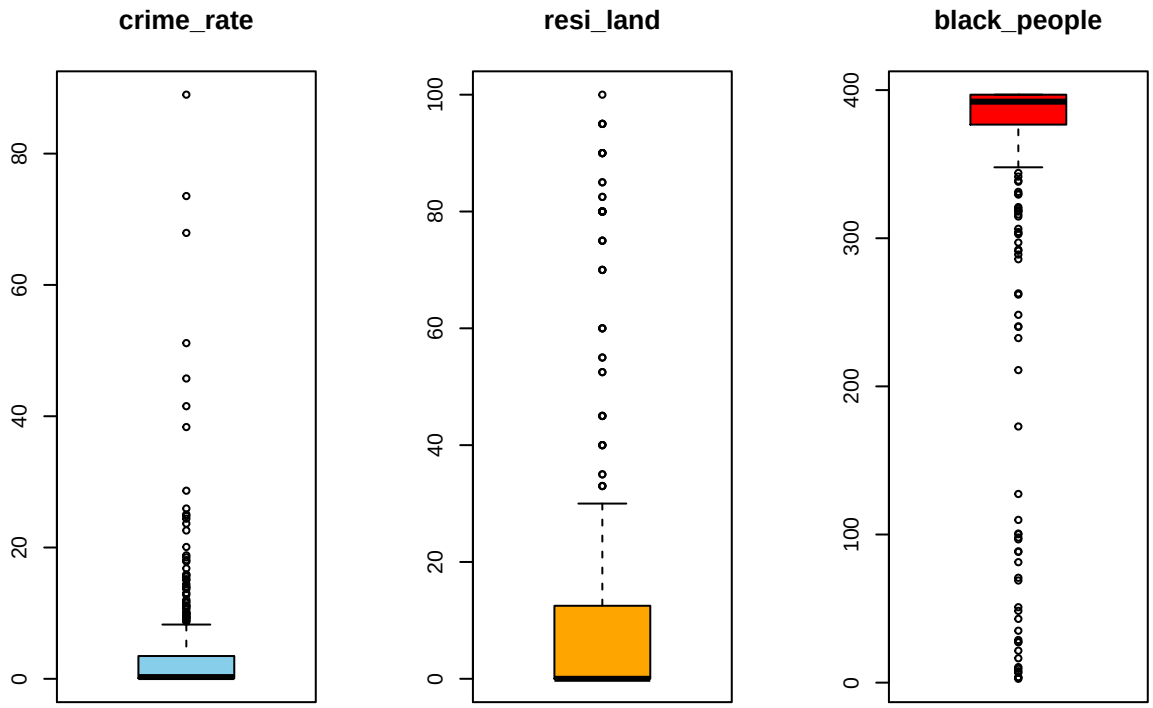
```
##    crime_rate       residential_landzone non_retail_land Charles_River
##  Min.   : 0.00632   Min.   :  0.00       Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08196   1st Qu.:  0.00       1st Qu.: 5.13   1st Qu.:0.00000
##  Median : 0.26888   Median :  0.00       Median : 8.56   Median :0.00000
##  Mean   : 3.69014   Mean   : 11.46       Mean   :11.00   Mean   :0.06853
##  3rd Qu.: 3.43597   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_concentration avg_no_of_rooms   owner_age      dis_employment
##  Min.   :0.3890    Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4530    1st Qu.:5.879   1st Qu.: 45.48   1st Qu.: 2.110
##  Median :0.5380    Median :6.202   Median : 77.70   Median : 3.199
##  Mean   :0.5532    Mean   :6.280   Mean   : 68.93   Mean   : 3.805
##  3rd Qu.:0.6240    3rd Qu.:6.606   3rd Qu.: 94.25   3rd Qu.: 5.117
##  Max.   :0.8710    Max.   :8.780   Max.   :100.00   Max.   :12.127
##  access_to_highways      tax           PTRATIO       black_people
##  Min.   : 1.000     Min.   :187.0   Min.   :12.60   Min.   :  2.6
##  1st Qu.: 4.000     1st Qu.:280.2   1st Qu.:17.40   1st Qu.:376.7
##  Median : 5.000     Median :330.0   Median :19.10   Median :392.2
##  Mean   : 9.404     Mean   :406.4   Mean   :18.54   Mean   :358.5
##  3rd Qu.:24.000     3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.9
##  Max.   :24.000     Max.   :711.0   Max.   :22.00   Max.   :396.9
##   lower_people    owner_occupied_homes_price
##  Min.   : 1.730   Min.   : 5.00
##  1st Qu.: 7.125   1st Qu.:16.80
##  Median :11.300   Median :21.05
##  Mean   :12.769   Mean   :22.36
##  3rd Qu.:17.117   3rd Qu.:25.00
##  Max.   :37.970   Max.   :50.00
```

Here we can see that variables 'crim_rate' and ,'residential_landzone','black_people' seem to be taking wide range of values.

Since these Variables have a large difference between their median and mean which indicates that these variables must have a lot of outliers in them.

Q2 Are there any outliers in the 'crim_rate' and ,'residential_landzone','black_people variables?

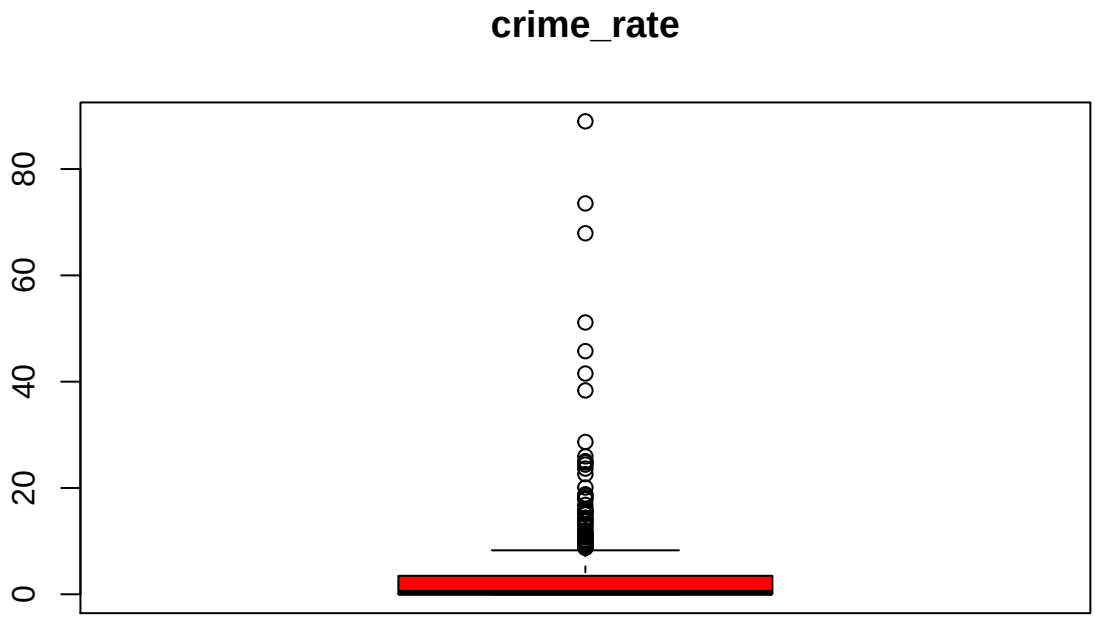
```
par(mfrow = c(1, 3))
boxplot(df2$crime_rate, main='crime_rate',col='Sky Blue')
boxplot(df2$residential_landzone, main='resi_land',col='Orange')
boxplot(df2$black_people, main='black_people',col='Red')
```



It can be seen from these boxplots that all of these variables are highly diverse.

Q3. Discuss the values of outliers available in crime_rate variable

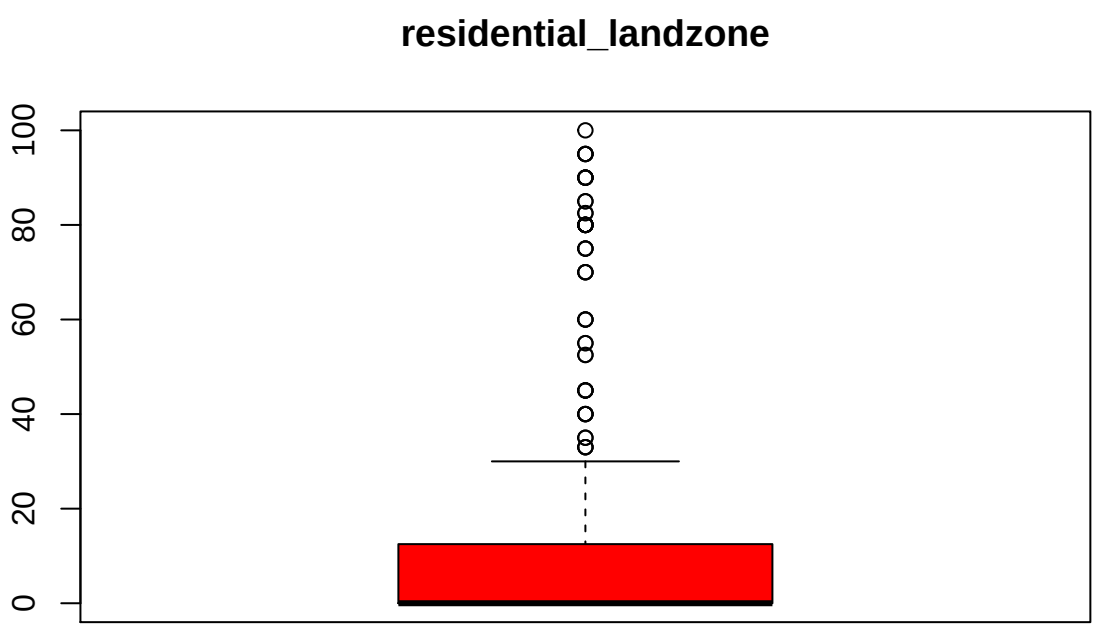
```
boxplot(df2$crime_rate,main='crime_rate',col='Red')
```



The values in this column are outlier if they cross the threshold of approximately 9-10 . So most of the values are below that.From our summary ,we also found out that the minimum value is 0.00632, maximum value is 88.97620,median is 0.26888 and mean is 3.69014

Q4. Discuss the values of outliers available in residential_landzone variable

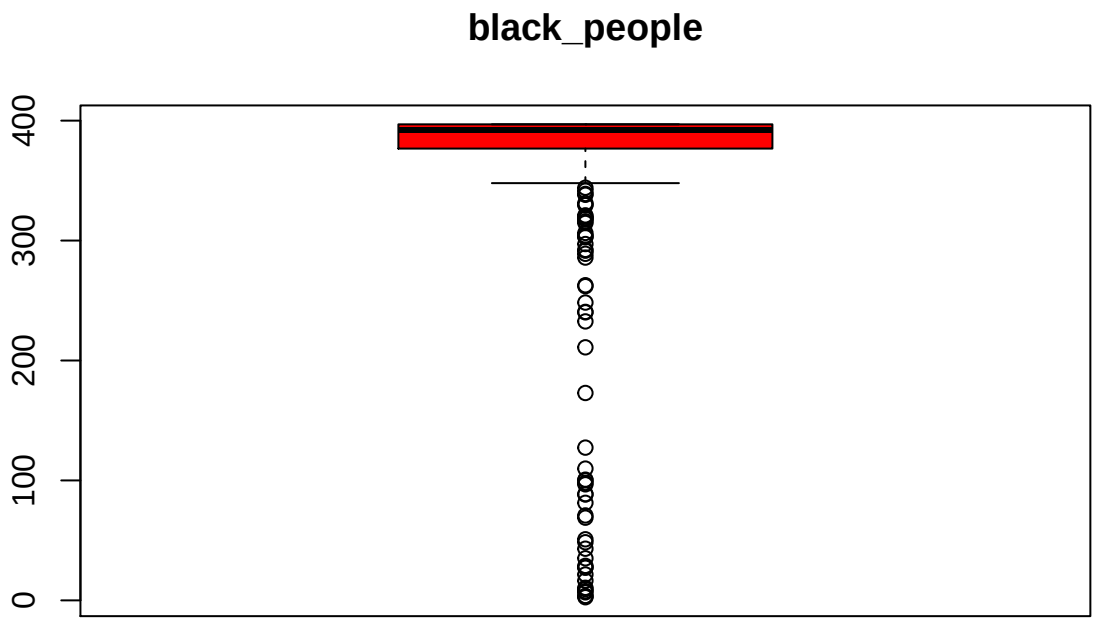
```
boxplot(df2$residential_landzone,main='residential_landzone',col='Red')
```



The values in this column are outlier if they cross the threshold of around 27 .That means most of the towns in Boston have below proportion of 27% residential land zoned that are over 25,000 sq.ft in size.Further, From our summary ,we also found out that the minimum value is 0.00 that means there are towns which have zero proportion of residential land zones which are above 25000 Sq.ft, maximum value is 100,median is also 0.00 and mean is 11.46. From this analysis we can conclude that most of the towns in boston dont have residential lands which are above 25000 sq feet in size.

Q5.Discuss the values of outliers available in black_people variable?

```
boxplot(df2$black_people,main='black_people',col='Red')
```



This columns was included into the dataset to check the mentality of people regarding the housing price they are willing to pay for the houses which have black people as neighbors.Here B = 1000(Bk - 0.63)ˆ2 where Bk is the proportion of blacks by town. The values in this column are outliers if they are below the threshold of around 350 .That means in most of the towns in Boston are having more than 350 as a proportion of black people..Further, From our summary ,we also found out that the minimum proportion is 2.6, maximum proportion is 396.9,median is 392.2 and mean is 358.8.

From the boxplot and summary it can be concluded that there aren't many outliers here, since the first,second,third quadrant and minimum, maximum values are more or less the same.

Q6.We are intrigued by the finding that, even in 1978, when this data was gathered,they included proportion of blacks by town to predict the housing prices in BOSTON. So we want to know the findings now : Does black people having as neighbors affect the housing prices in Bostom or was it just wrong in first place to include such data in survey?

To find that out, we can run simple correlation test between the two columns

```
cor.test(df2$black_people,df2$owner_occupied_homes_price)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df2$black_people and df2$owner_occupied_homes_price
## t = 7.3316, df = 392, p-value = 1.314e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2572867 0.4312570
```

```
## sample estimates:
##       cor
## 0.3472561
```

Astounding !! The correlation between the two comes out to be **Positive 0.347** , which shows that , having black people as neighbors actually increases the housing prices , but not to that margin. This finding must have something to do with the growth and immigration.The more a town grows more immigrants(black people) move into that town as the availability of resources gets better. As a result,the housing prices increase.

*Q7.After finding a surprising relation between proportion of black people and housing prices, we are more interested to know what are relation of other variables with the housing prices.Are there any other variables which are highly correlated with the housing prices?**

To get the relationship between all the variables with the housing prices we calculate the correlation using **cor()** function .

```
as.data.frame(cor(df2))
```

```
##                             crime_rate residential_landzone non_retail_land
## crime_rate                   1.00000000          -0.18807507      0.39155182
## residential_landzone        -0.18807507           1.00000000     -0.52125603
## non_retail_land              0.39155182          -0.52125603      1.00000000
## Charles_River               -0.05196992          -0.03335682      0.04981956
## nox_concentration            0.41615982          -0.51566046      0.76273657
## avg_no_of_rooms             -0.22716991           0.34321034     -0.40306825
## owner_age                    0.34131149          -0.56817376      0.64238703
## dis_employment              -0.36505178           0.64535889     -0.69656900
## access_to_highways           0.60866672          -0.29877294      0.59194354
## tax                          0.56084114          -0.30576760      0.73420369
## PTRATIO                      0.26542768          -0.42216416      0.39569127
## black_people                -0.38625382           0.16989420     -0.34478755
## lower_people                 0.46190578          -0.41504110      0.59815590
## owner_occupied_homes_price -0.39723006           0.40682152     -0.51082916
##                            Charles_River nox_concentration avg_no_of_rooms
## crime_rate                   -0.05196992        0.41615982     -0.22716991
## residential_landzone         -0.03335682       -0.51566046      0.34321034
## non_retail_land               0.04981956        0.76273657     -0.40306825
## Charles_River                 1.00000000        0.07666108      0.09530772
## nox_concentration             0.07666108        1.00000000     -0.31656347
## avg_no_of_rooms               0.09530772       -0.31656347      1.00000000
## owner_age                     0.07264446        0.73254019     -0.24867008
## dis_employment               -0.09503705       -0.76813683      0.21871341
## access_to_highways            0.01410209        0.62817041     -0.23605670
## tax                          -0.02651313        0.67982405     -0.32056056
## PTRATIO                      -0.10499480        0.21021622     -0.39068616
## black_people                  0.06891304       -0.38425662      0.12331954
## lower_people                 -0.03711330        0.59365548     -0.63622618
## owner_occupied_homes_price    0.17370115       -0.45905433      0.72395076
##                             owner_age dis_employment access_to_highways
## crime_rate                 0.34131149    -0.36505178         0.60866672
## residential_landzone      -0.56817376     0.64535889        -0.29877294
## non_retail_land            0.64238703    -0.69656900         0.59194354
## Charles_River              0.07264446    -0.09503705         0.01410209
```

```
## nox_concentration           0.73254019    -0.76813683         0.62817041
## avg_no_of_rooms            -0.24867008     0.21871341        -0.23605670
## owner_age                   1.00000000    -0.75354690         0.44358519
## dis_employment             -0.75354690     1.00000000        -0.47707545
## access_to_highways          0.44358519    -0.47707545         1.00000000
## tax                         0.50447249    -0.52960262         0.89999984
## PTRATIO                     0.26496758    -0.22884007         0.44194918
## black_people               -0.28198984     0.28516841        -0.44413465
## lower_people                0.60113652    -0.50503607         0.51086842
## owner_occupied_homes_price -0.40747050     0.27954693        -0.41663771
##                                   tax    PTRATIO black_people lower_people
## crime_rate                  0.56084114  0.2654277  -0.38625382    0.4619058
## residential_landzone       -0.30576760 -0.4221642   0.16989420   -0.4150411
## non_retail_land             0.73420369  0.3956913  -0.34478755    0.5981559
## Charles_River              -0.02651313 -0.1049948   0.06891304   -0.0371133
## nox_concentration           0.67982405  0.2102162  -0.38425662    0.5936555
## avg_no_of_rooms            -0.32056056 -0.3906862   0.12331954   -0.6362262
## owner_age                   0.50447249  0.2649676  -0.28198984    0.6011365
## dis_employment             -0.52960262 -0.2288401   0.28516841   -0.5050361
## access_to_highways          0.89999984  0.4419492  -0.44413465    0.5108684
## tax                         1.00000000  0.4469615  -0.43545656    0.5722177
## PTRATIO                     0.44696148  1.0000000  -0.17981583    0.3950058
## black_people               -0.43545656 -0.1798158   1.00000000   -0.3837834
## lower_people                0.57221765  0.3950058  -0.38378339    1.0000000
## owner_occupied_homes_price -0.50886427 -0.5438090   0.34725609   -0.7434496
##                            owner_occupied_homes_price
## crime_rate                                 -0.3972301
## residential_landzone                        0.4068215
## non_retail_land                            -0.5108292
## Charles_River                               0.1737012
## nox_concentration                          -0.4590543
## avg_no_of_rooms                             0.7239508
## owner_age                                  -0.4074705
## dis_employment                              0.2795469
## access_to_highways                         -0.4166377
## tax                                        -0.5088643
## PTRATIO                                    -0.5438090
## black_people                                0.3472561
## lower_people                               -0.7434496
## owner_occupied_homes_price                  1.0000000
```

As can be seen by the correlation table,

1.avg_no_of_rooms has correlation of 0.7239508, so average number of rooms in a house is highly correlated with the house's price, which makes a perfect sense as well, since a higher number of rooms will only mean a bigger house and thus higher pricing of the house.

2.lower_people has correlation of -0.7434, which means,Proportion of lower class people is highly inversely correlated with the housing prices. Which also makes sense, as if a town has high number of lower class people, it will have less pricing in the in that area

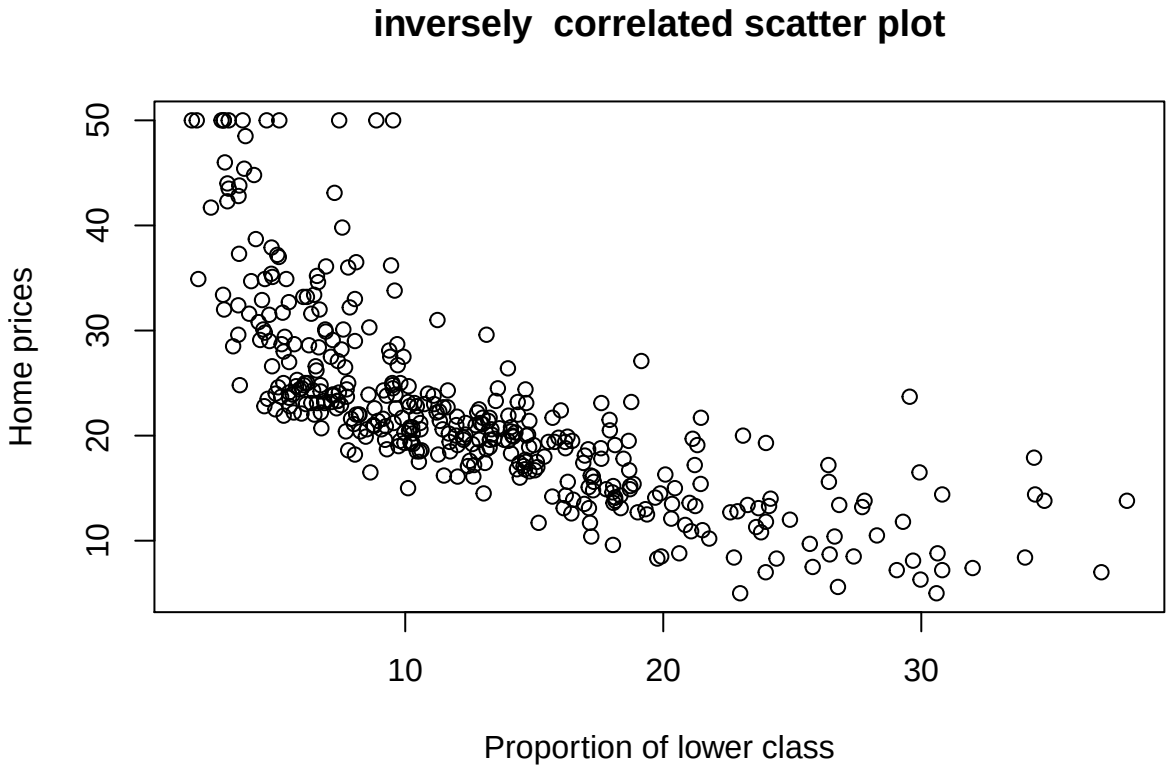
3.non_retail_land has correlation of -0.5108, which is not strong correlation, but a noticable inversely correlation. That means, if non retail land increases than the housing prices will decrease.

4.Tax has inverse correlation of -0.5088643, that means if tax payment on a property increses the owner will sell it at a lower price.Which is also logical.

After knowing these 4 variables which highly affect the housing prices in a town, we are interested to dwell into each of these 4 variables further and want to know more about the characteristics of them.

Q 8 Discuss the relation between proportin of lower class people and housing prices in Boston and find out if there are any cases which didn't follow the expected trend(outliers) ? A scatter plot is the first go to thing while dealing with relationship between two variables.Correlation does tell the relationship between the two, but it is always good to have a pictorial view of the setting , because it gives a more informatic picture. Here

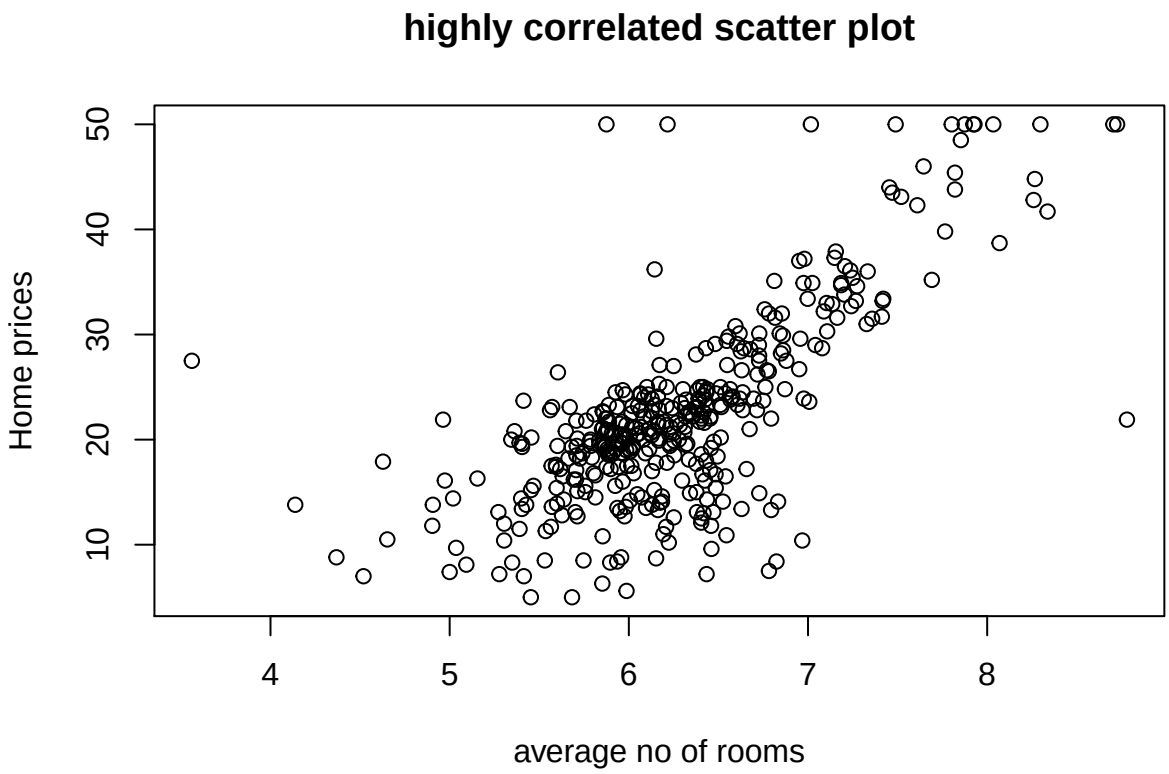
```
plot(df2$lower_people,df2$owner_occupied_homes_price,xlab = 'Proportion of lower class', ylab= 'Home pr
```



It shows how the housing prices decreases as the proportion of lower class people increases. It also shows that there aren't many outliers in the plot, which is a good thing for the prediction.

Q 9 Discuss the relation between Average number of rooms in houses and prices of the houses in Boston and find out if there are any outliers of the relation?

```
plot(df2$avg_no_of_rooms,df2$owner_occupied_homes_price,xlab = 'average no of rooms', ylab= 'Home price
```



There are a number of outliers in this setting.While the trend is that if a house has more number of rooms,it will be costly, but there are minimal number of towns which go completely aginst it. As can be seen by the scatter plot

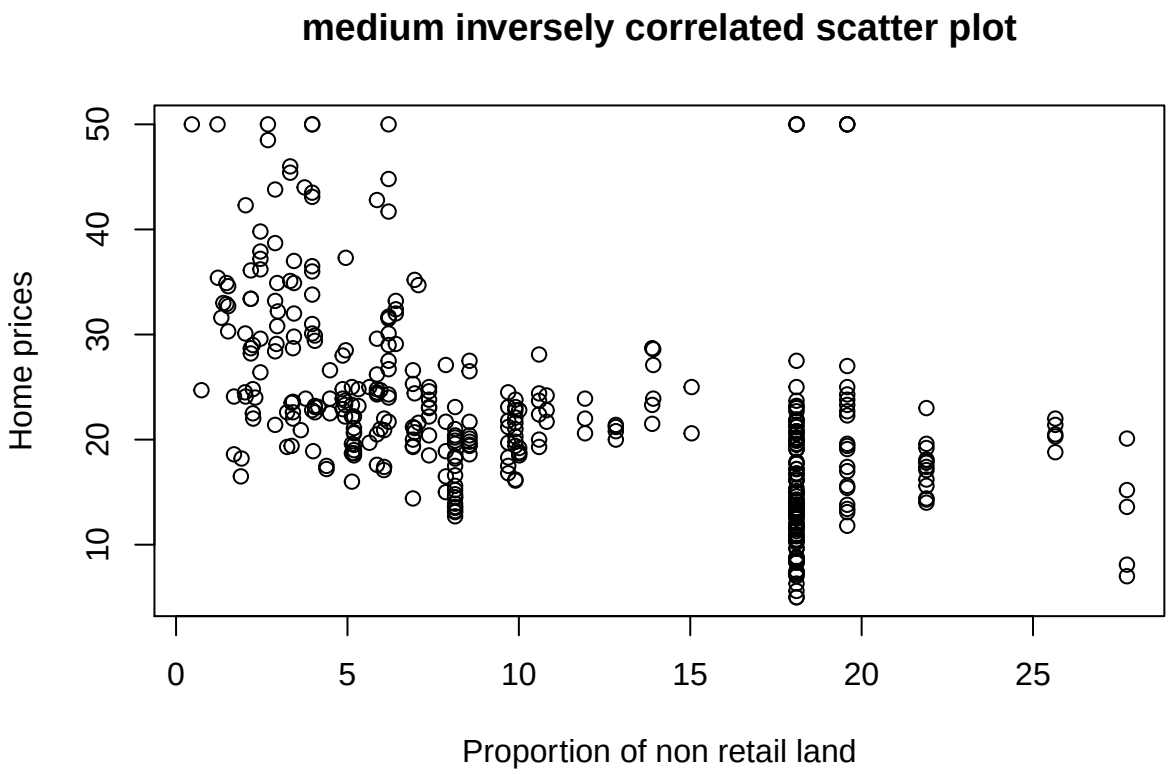
1.There is a town which has average of number of rooms of 3 only and still it has high price.

2.There are a number of towns where average number of houses is 6 but still the price of the house in those town is equal to the towns which have average number of rooms as 8.

Finding : There must be some towns which are highly developed,that's why even having having less number of rooms, they still demand a high price. vice versa the case with lower class towns.

Q 10 Discuss the relation between non_retail_land and prices of the houses in Boston and find out if there are any outliers of the relation?

```
plot(df2$non_retail_land,df2$owner_occupied_homes_price,xlab = 'Proportion of non retail land', ylab= '
```



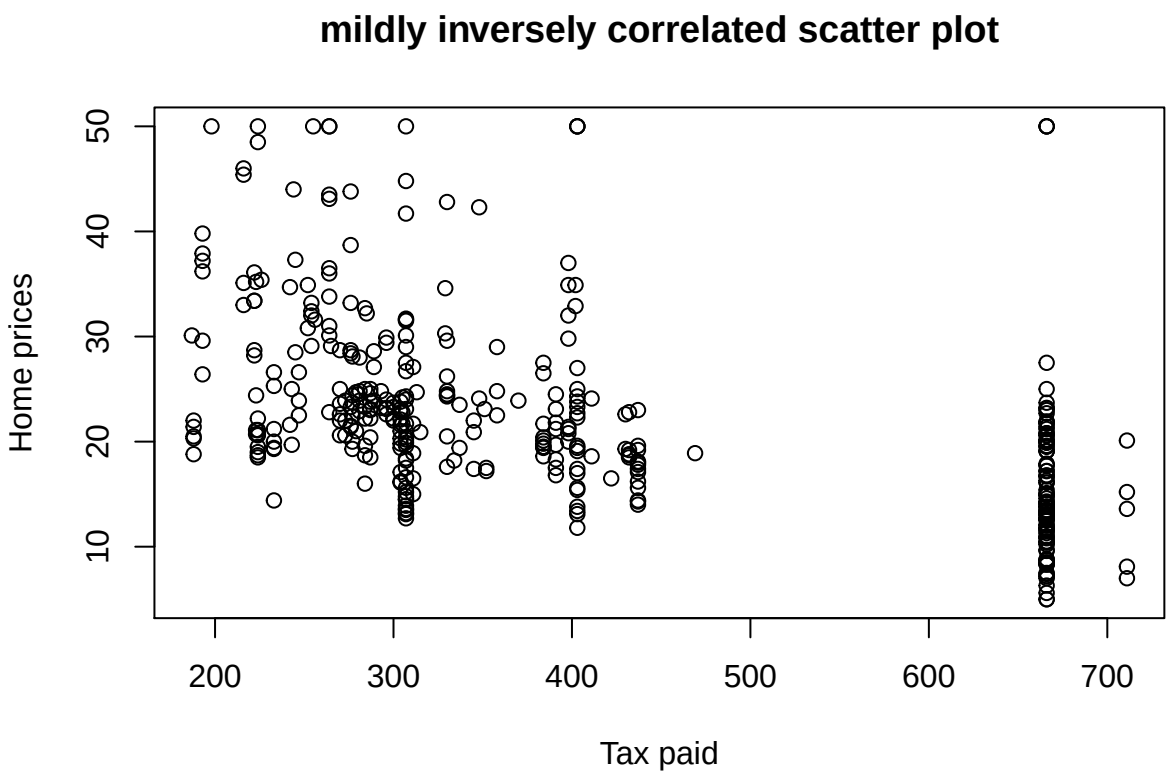
Both the variables are inversely and mildly correlated . The correlation between the two was -0.5108292.As also can be seen from the plot that,with increase in housing price, the proportion of non retail land decreases in the town. Finding from the scatter plot:

1.There are a lot outliers , as was expected from the relationship between the two variables.

2.There are two towns which have high proportion of non retail land nearly 18-19, still the prices in those towns are very high for some reason.

Q11 Discuss the relation between Tax and prices of the houses in Boston and find out if there are any outliers of the relation?

```
plot(df2$tax,df2$owner_occupied_homes_price,xlab = 'Tax paid', ylab= 'Home prices',main='mildly inverse
```



These two have inverse correlation, as the tax increases for the house the price decreases. Finding :

1. Most of the houses are segregated towrds high pricing and low tax.
2. Mostly there are not many outliers in this relationship, but still the relation between the two is mildly strong.

After looking at the 4 variables which are highly affecting the housing prices in the Boston, we are interested to get a holistic view of all the variables in the dataframe

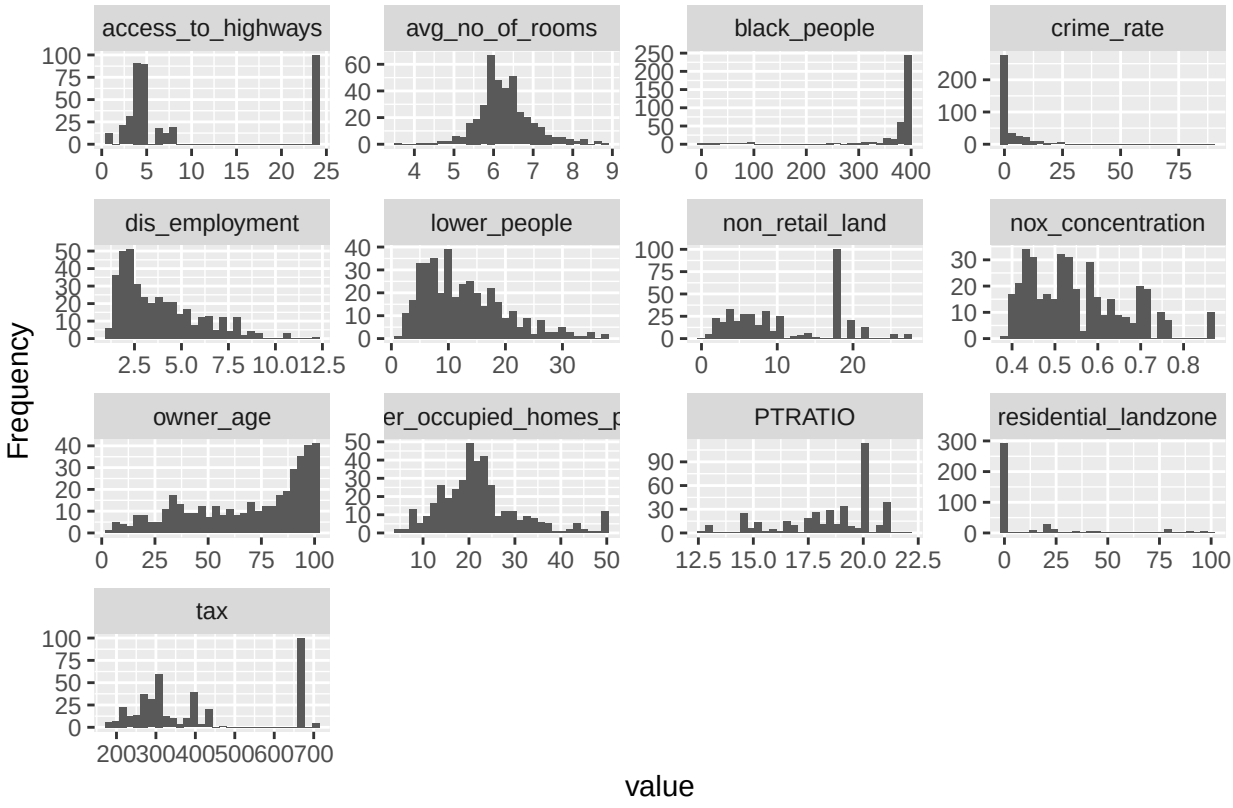
Q12 Get a holistic view of dispersion of all the variables ?

For that we have decided to plot a univarate histogram for all the variables included into the data frame, to get a quick peak. For this we have used DataExplorer library.

```
library(DataExplorer)
```

```
## Warning: package 'DataExplorer' was built under R version 4.0.5
```

```
plot_histogram(df2)
```



From the univariate histograms of each variables we can conclude some of the findings:

1.Average number of rooms variable follows a normal distribution.

2.Distance with the employment centre is skewed to the right or positively skewed,which means the mean is more than the median.In simpler terms, most of the values in this variables are less as compared to the values above third quartile.

3.Owner_age variables are highly skewed to the left or negatively skewed, that means, it has higher median than the mean.So the tendency of age of owners is on higher side and the variable is very well dispersed.

After getting a holistic view of all the variables, we want to dwell into some of the categorical variables in our dataframe and find out if there any particular categories in them, which is highly relatable with the housing prices We have two categorical variables

1.Access to highways (1,2,3,4,5,6,7,8,24) here 1 shows highest accessibility to highways and 24 shows the lowest accessibility to highways.

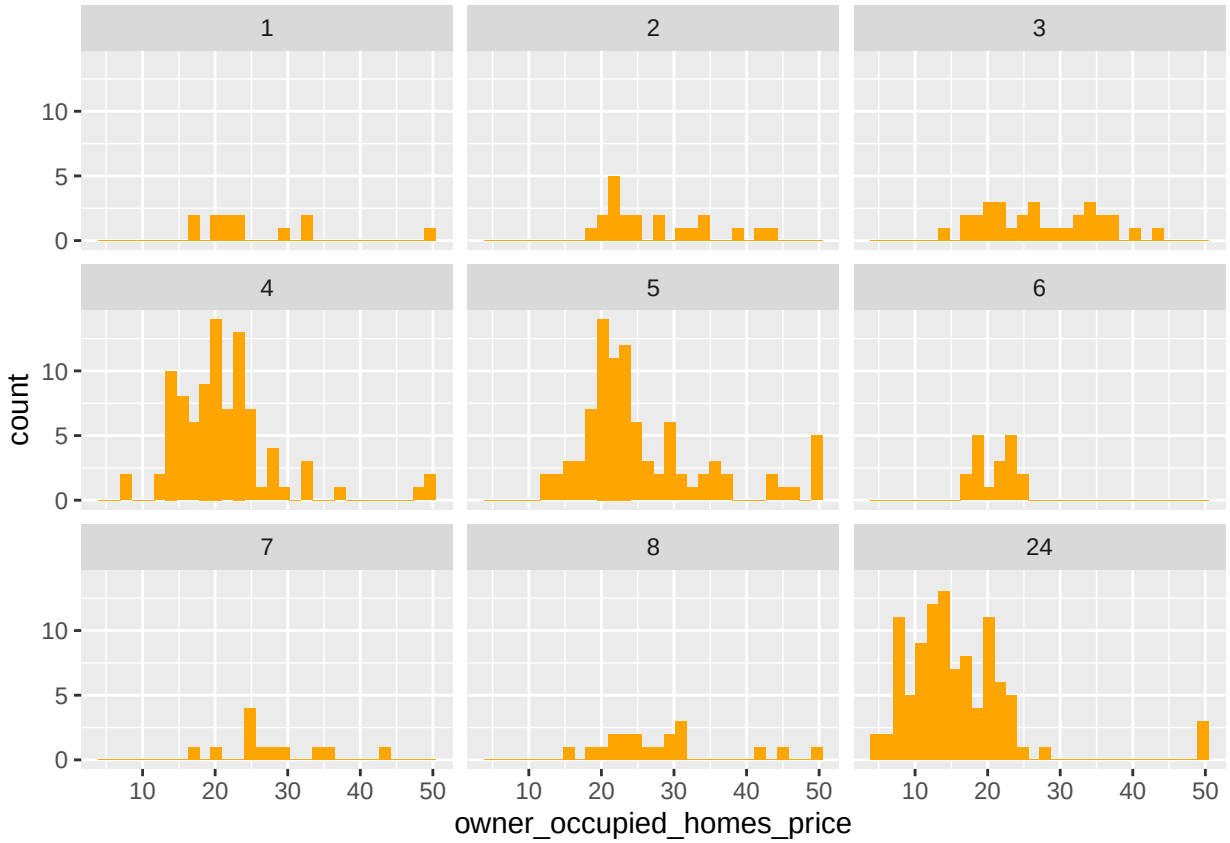
2.Charles River index (1 , 0) where 1 shows tract boundary with the river and 0 shows otherwise.

Q13 depth relation between the access to highways and housing prices.Compare each index of access to highways with the housing prices?

For this we have decided to plot histograms for each and every index of access_to_highways variable by using facet_wrap in ggplot() function

```
ggplot(df2)+aes(owner_occupied_homes_price)+geom_histogram(fill='orange')+facet_wrap(~access_to_highways
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



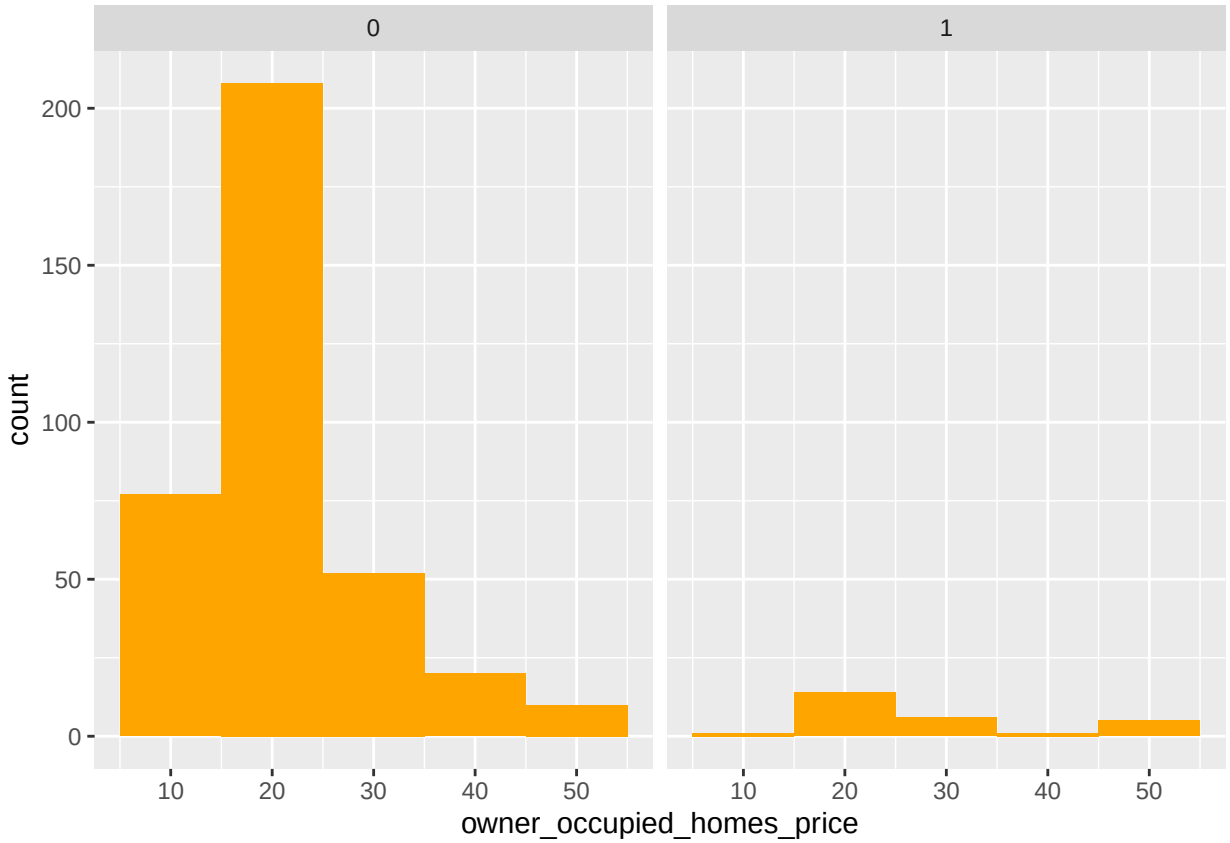
The histograms show the housing prices for each and every index. findings are :

1.Most of the values are in index 4,5,24

2. In index 24, which is the lowest accessibility with the highways, has low housing prices.Which is also logical
3. As the accessibility increases with the highway,the prices of the houses also increases.

Q14 There is one more categorical variable in our data, that is Charles_River. Which shows the proximity with the river. Index 1 shows tract bounds with the river and index 0 shows otherwise.We want to know the relation of each of these cases/indices with the housing prices

```
ggplot(df2)+aes(owner_occupied_homes_price)+geom_histogram(binwidth = 10,fill='orange')+facet_wrap(~Cha
```



Findings are :

1.Most of these towns have no proximity with the charles river,

2.towns with no proximity with river also show lower home prices, however there are some towns still have high home pricing, which must be due to some other reason.

3.Only a small number of towns are there which have proximity with the river and still having very low prices. So proximity with the river does effect the housing prices in Boston,

Q15 Find out if all the variables are independent of each other, if not, which of these variables are highly correlated to each other? We formed this question keeping in mind , if this data is further used for regression analysis, than as per the assumption in regression analysis, the variables must be independent of each other , so as to knowing which of these variables are highly correlated , we can actually form a good regression model.

To check the inter-relation between the variables, we can calculate correlation for each of these variables with each other, whuch we have already done above. Further we can also plot scatter plot between these variables to get a quick view of the correlation.

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.5
```

```
## corrplot 0.84 loaded
```

```
cor.test(df2$tax,df2$access_to_highways)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df2$tax and df2$access_to_highways
## t = 40.88, df = 392, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8793965 0.9172384
## sample estimates:
##       cor
## 0.8999998
```

Here we find very strong correlation of 0.89999984 between access to highways and tax rate. One possibility could be that, the towns which have very good access to highways are well funded ,well managed and have better connectivity.Hence to raise those funds tax rates are also high.

Another high correlation can be seen between nox_concentration and non_retail_land variables

```
cor.test(df2$non_retail_land,df2$nox_concentration)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df2$non_retail_land and df2$nox_concentration
## t = 23.351, df = 392, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7180493 0.8011610
## sample estimates:
##       cor
## 0.7627366
```

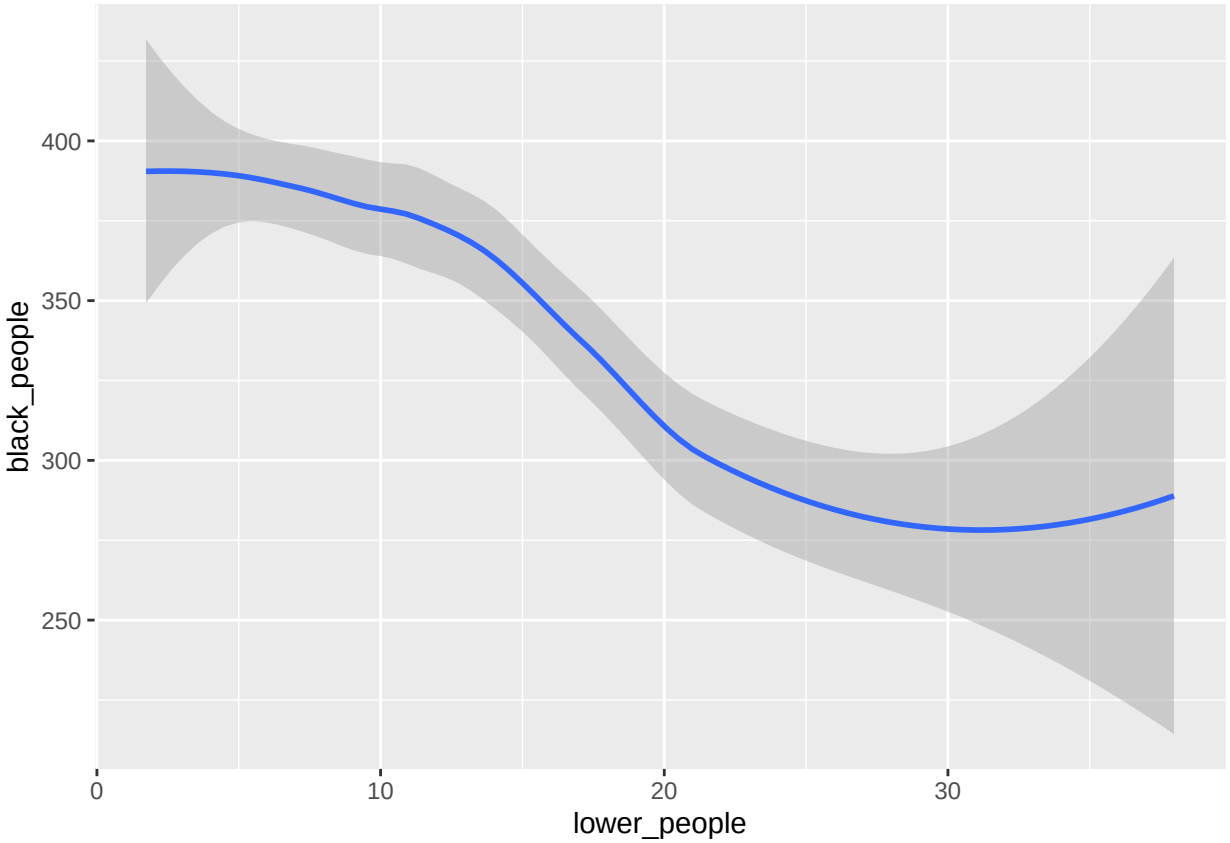
There is a high correlation between nox_concentration and non_retail_land.The correlation comes to be 0.7627366 .One Reasons could be most of these land is for industries, thats why they emit larger concentration of nitrogen oxide.

Q16 We are trying to identify a relation between proportion of black people in town and proportion of lower class people in that town by using geom_smooth.

```
ggplot(df2)+aes(x=lower_people,y=black_people)+geom_smooth(binwidth = 50)
```

```
## Warning: Ignoring unknown parameters: binwidth
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Here, it can be seen from the graph that, as the proportion of black people reduces , the proportion of lower class people increase upto a certain point,which is nearly 30 in case of lower class proportion, after this point, we see a sharp increase in number of black people.

Conclusion :

1.First astonishing finding was that having black people as neighbors not really decrease the pricing of houses , on the contrary it increases it.

2.Average number of rooms in a house is the most affecting factor for the housing prices in Boston.

3.Proportion of lower class of people in a town is the second most prominent factor to inersely affect the housing prices.

4.Tax and access to highways perform collinearity, so it's advisable to use only one of these while using these data for regression. Same is the case with non retail land and nitrogen concentration.