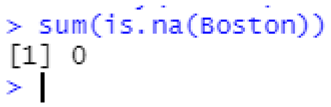
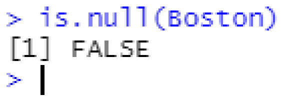
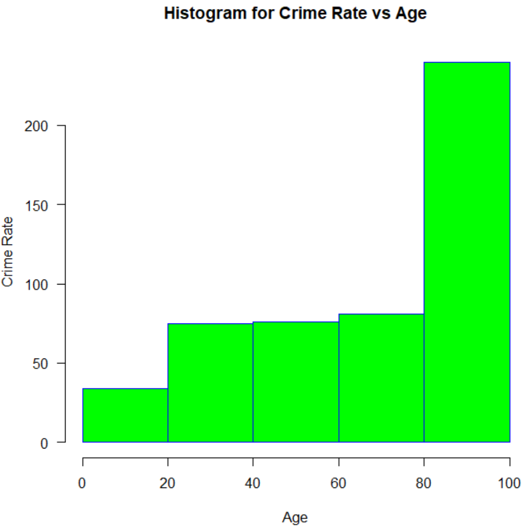
Q3) **<u>Part A – Visualization of Data</u>**



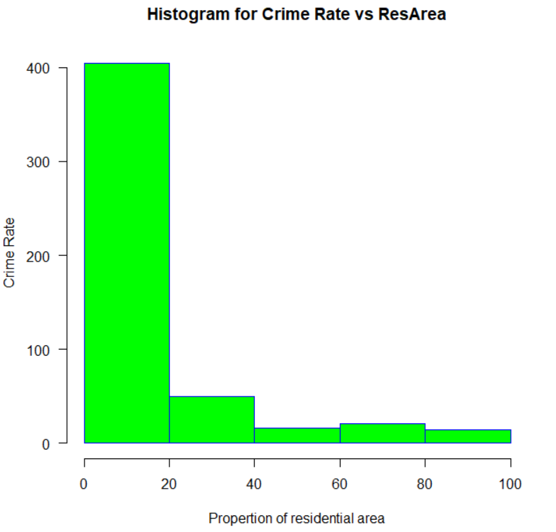
```
> sum(is.na(Boston))
[1] 0
> 
```

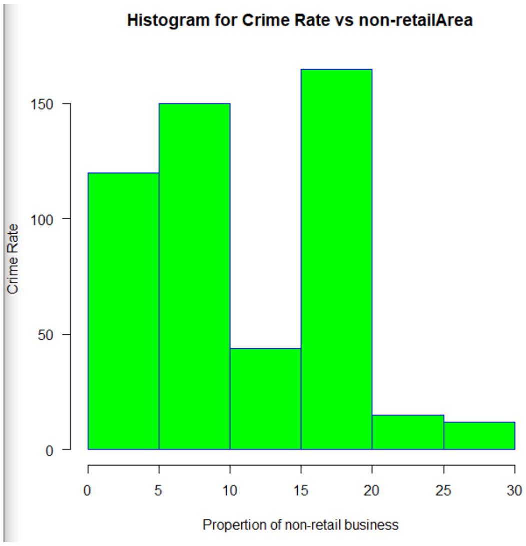


```
> is.null(Boston)
[1] FALSE
> 
```

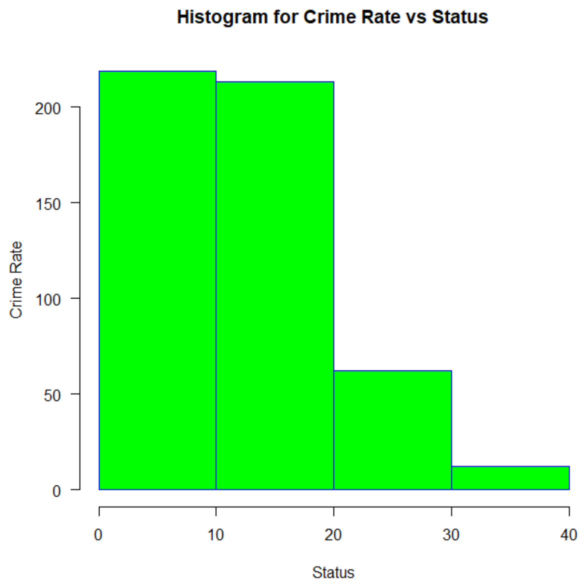


It seems the crime rate is high for the older age group that Is within 80-100.However it doesn't necessarily imply that this age group people are causing the crimes. There could be a different perspective to this ,that mostly people of this age group are the victims of this crime.

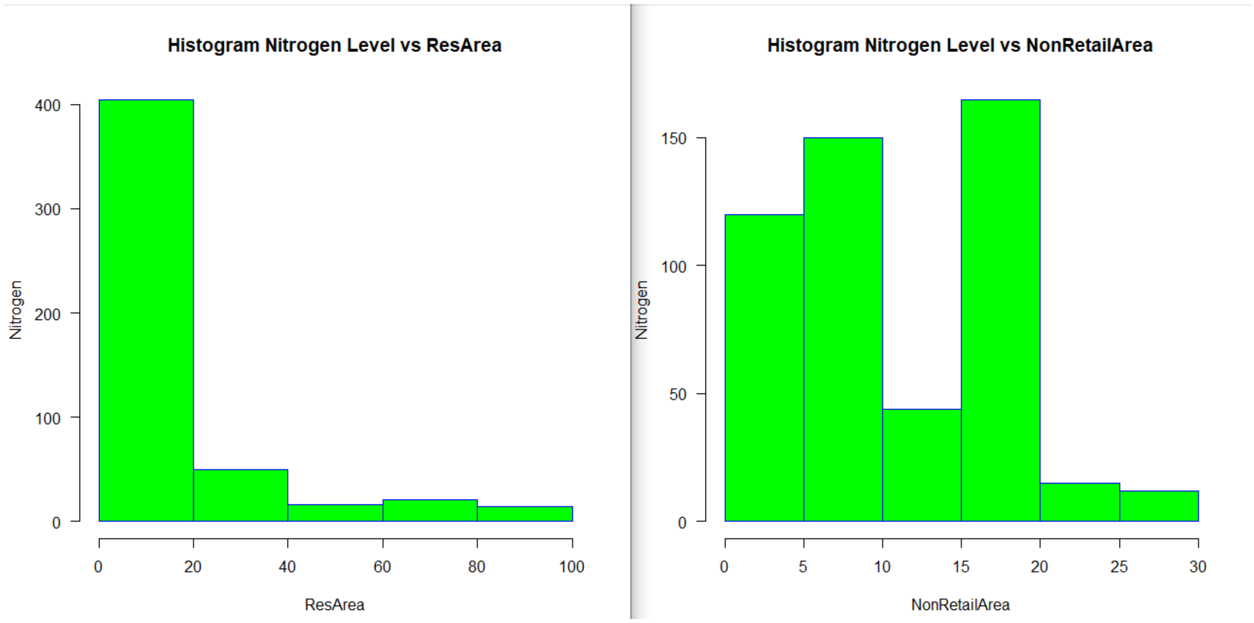




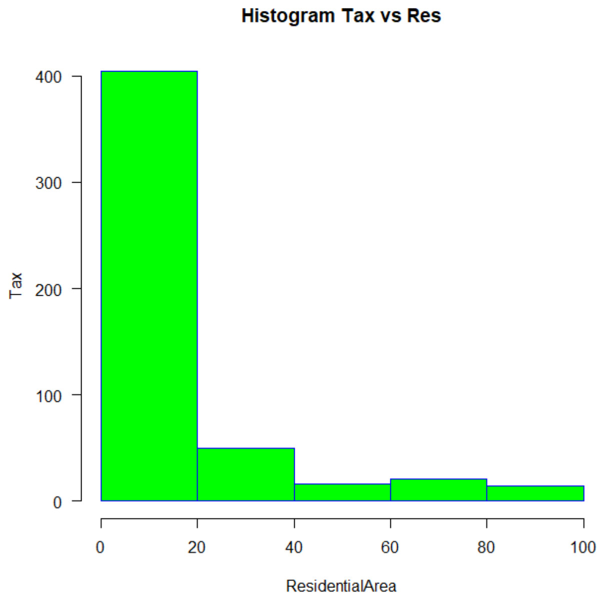
If we look at the Y axis of both plots Left one – Crime Rate vs Residential Area owned and Right One – Crime Rate vs Non-Retail Business Area, it appears that the crime rate is more in residential area of Boston Suburbs.

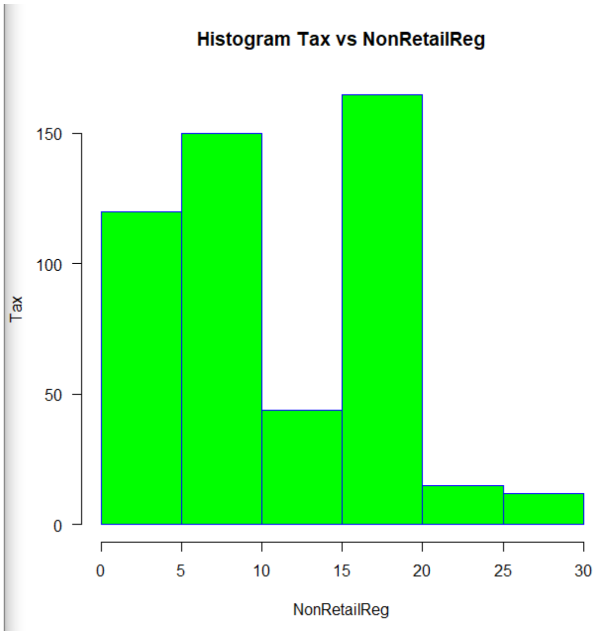


The left plot indicates crime rate vs Status, from what is observed from this histogram plot it can be stated that the crime rate for people with lower status is high ,so its not necessary but in some societies, it might be the case that people with low social status are denied privilege to white collar jobs. As a result of which they are likely to fall into trap of these unsocial activities.



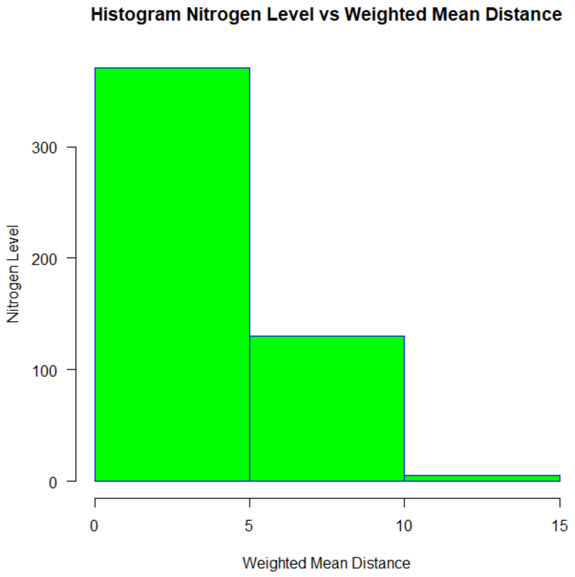
Again, we see just like crime rate the nitrogen oxide concentration seems to of higher scale that is along the Y axis in the left plot that is for residential region compared to that of non-retail business region that is in the right plot.



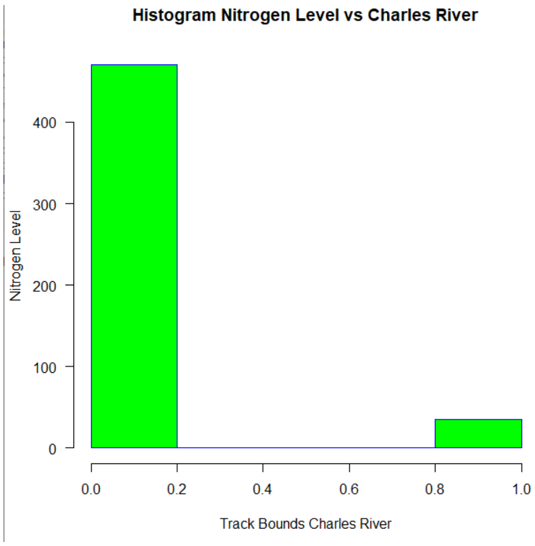


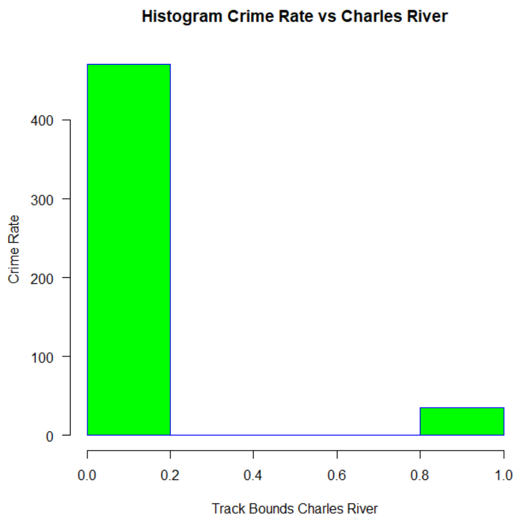
Again, we see just like crime rate ,nitrogen oxide concentration, Tax seems to of higher scale that is along the Y axis in the left plot that is for residential region compared to that of non-retail business region that is in the right plot.

Thus, we infer that for residential area three of these factors like crime rate, nitrogen oxide concentration and Tax are comparatively high.

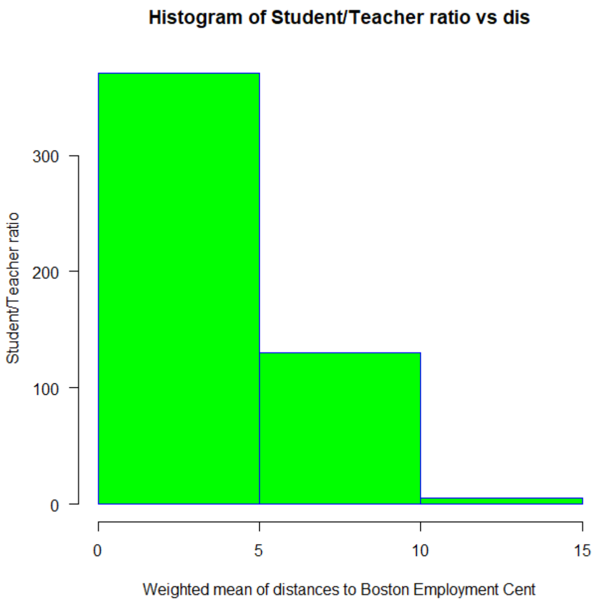


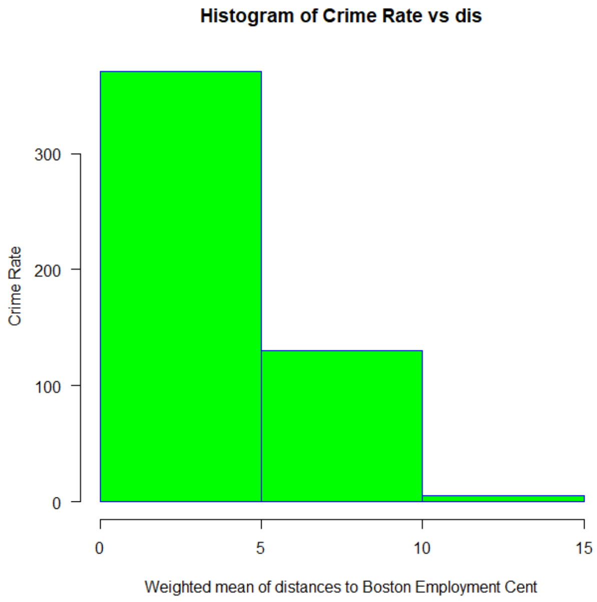
For Boston Suburbs whose weighted mean distance to five Boston Employment centers is very close seems to have higher nitrogen level.



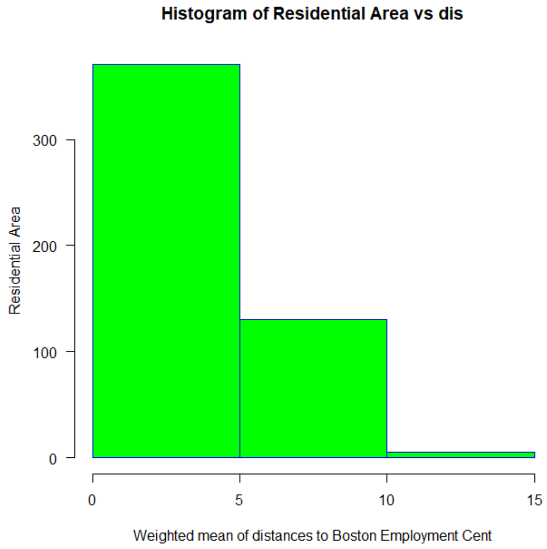


The left plot indicates nitrogen level with respect to whether the suburb region bounds Charles River or not and the right plot indicates Crime rate with respect to whether the suburb region bounds Charles River or not .Both the plot says that if the suburb region of Boston bounds the Charles River, then both the Nitrogen Level and Crime Rate are lower compared to those suburbs which do not bound the Charles River.



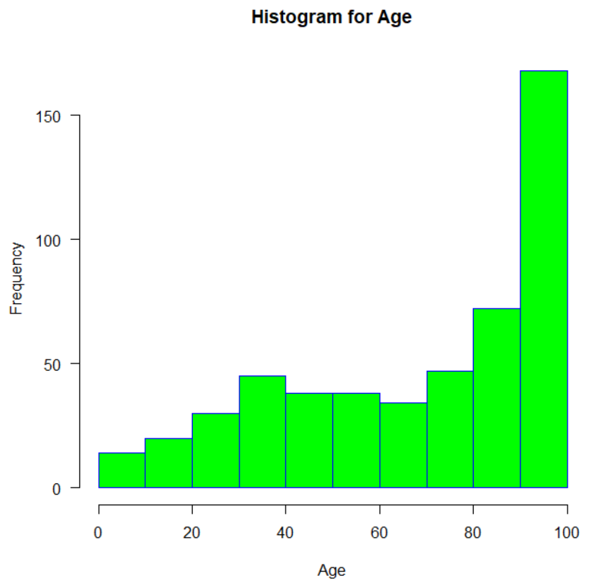


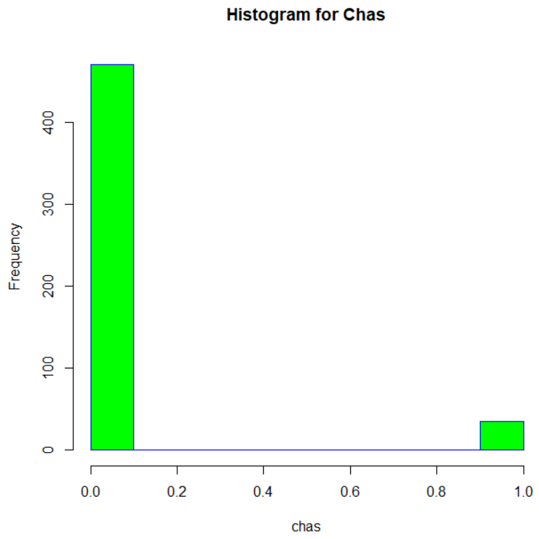
The above left plot indicates student to teacher ratio vs "weighted mean distance to 5 Boston Employment Centers" and the right plot indicates Crime Rate vs weighted mean distance to 5 Boston Employment. The right plot indicates that crime rate is high where the "weighted mean distance to 5 Boston Employment centers" is very low that is within the range[0-5]. The left plot says student to teacher ratio is also high where the weighted mean distance to 5 Boston Employment is very low that is within the range[0-5]



This particular plot says that proportion of residential area is concentrated towards lower value of "weighted mean distance to 5 Boston Employment" .So, though we still have to dig further and use Apriori Algorithm for further data analysis ,but as far as <u>safe place of living</u> and <u>closer proximity</u> to employment centers is concerned for a student , the student should probably decide to move to those places whose weighted mean distance to Boston Employment Center falls with the range of [5-10]because here the crime rate is comparatively low and also this region is considerably close to the employment centers.

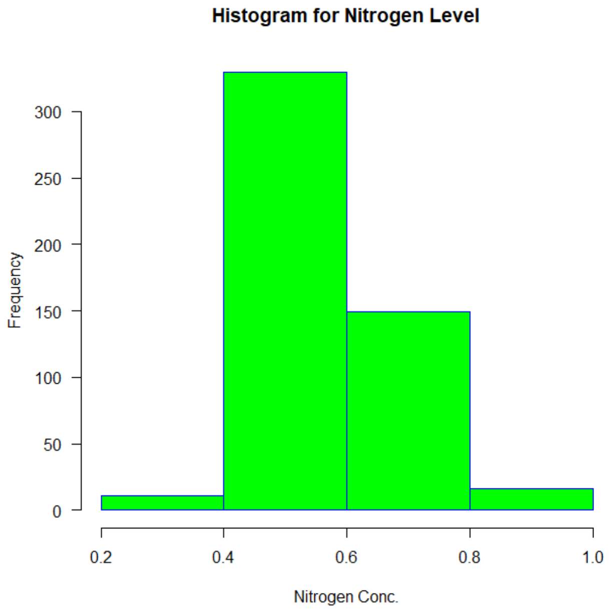
Part A) **Grouping of Data**

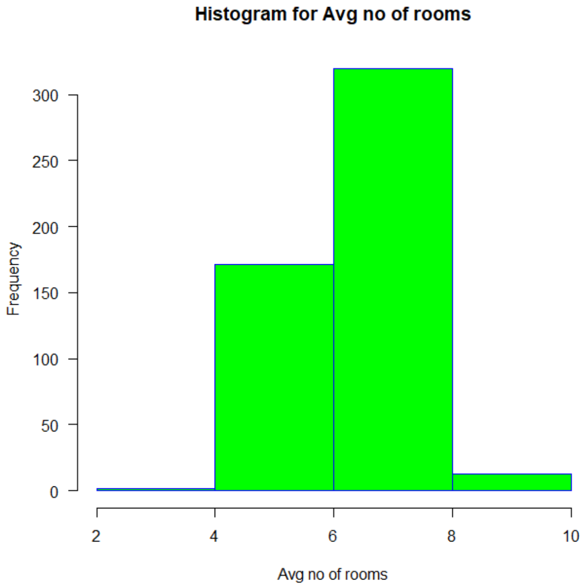




From our accumulated knowledge of age, we can group the age data into Young, Middle – Aged, Senior and Elderly.

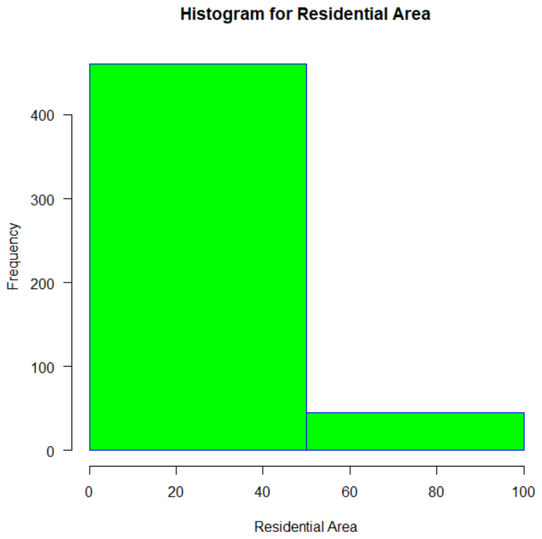
Similarly we can group the Chas River Data into two groups.

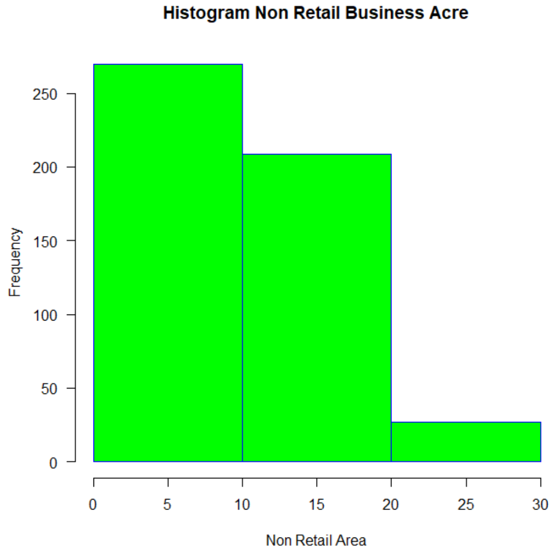




The data for Nitrogen Level can be grouped into low, Medium and high.

Similarly the data for room frequency can be Grouped into small, medium, and large.



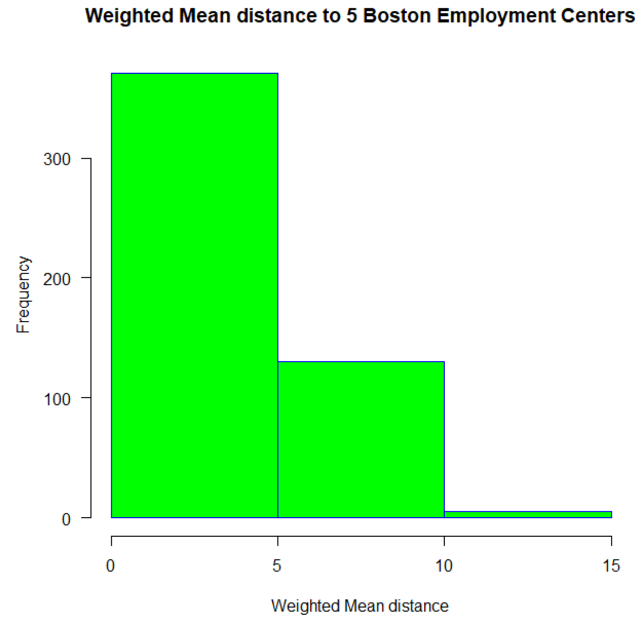


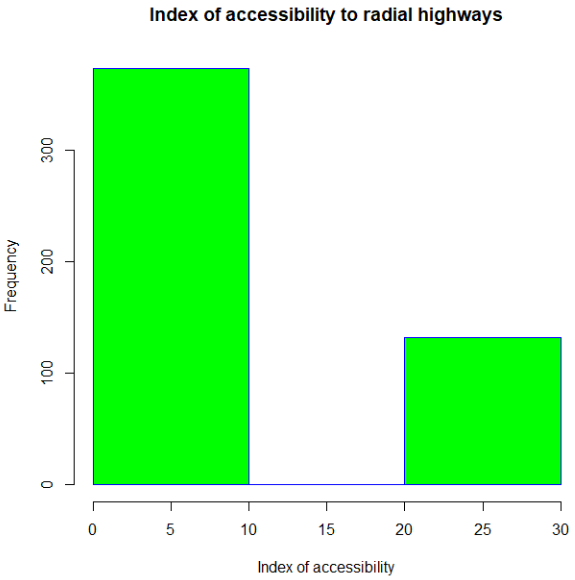
The data for residential area can be grouped into

Small, average, and large.

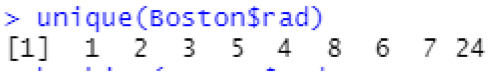
The data Non Retail Business Land can be also be

Grouped into small, average, and large





The



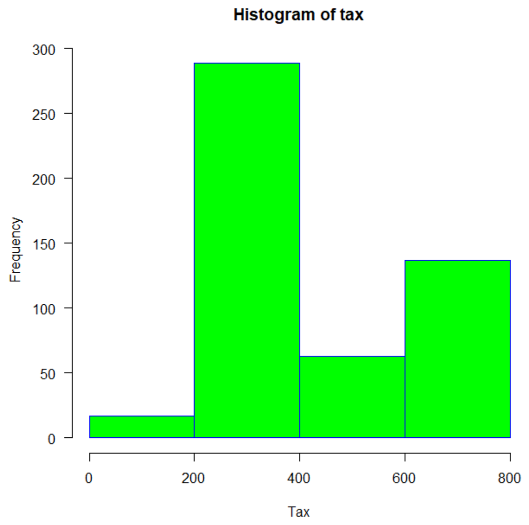
```
> unique(Boston$rad)
[1]  1  2  3  5  4  8  6  7 24
```

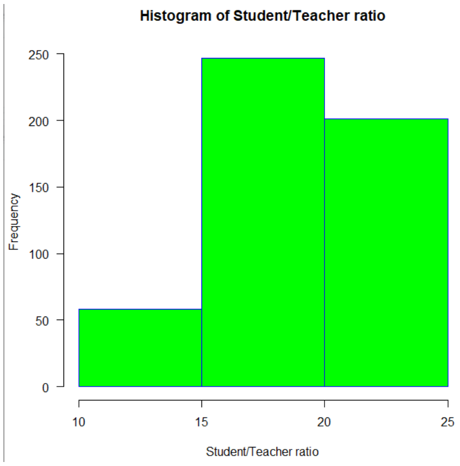
The

weighted mean distance to 5 Boston Employment
data for accessibility to radial

Centers can be grouped into close, average, and distant.

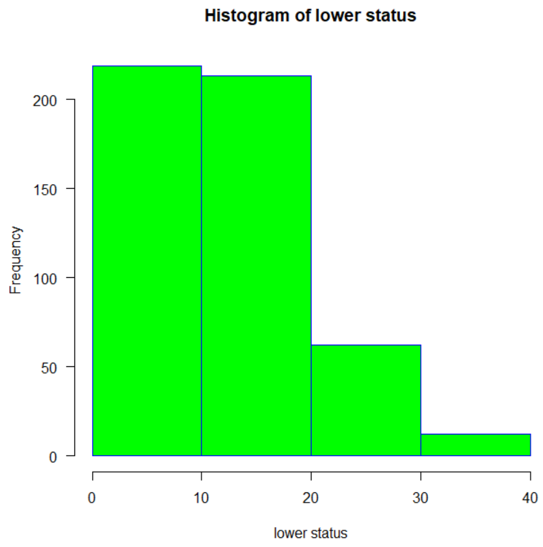
Highways can be grouped into (1,2,3,4) – Close ; (5,6,7,8) – Average ; (24) – Distant

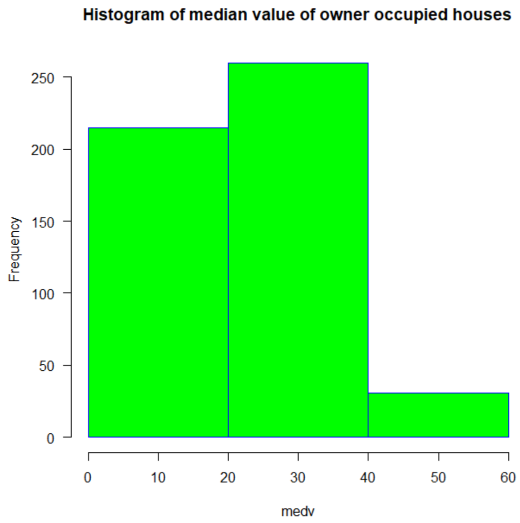




Tax data can be grouped into low, medium, high.

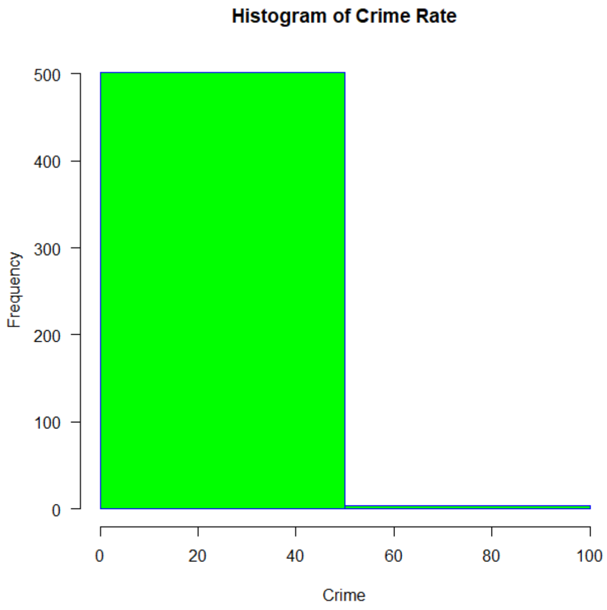
The data for Student/Teacher ratio can be Grouped into small, medium and large.





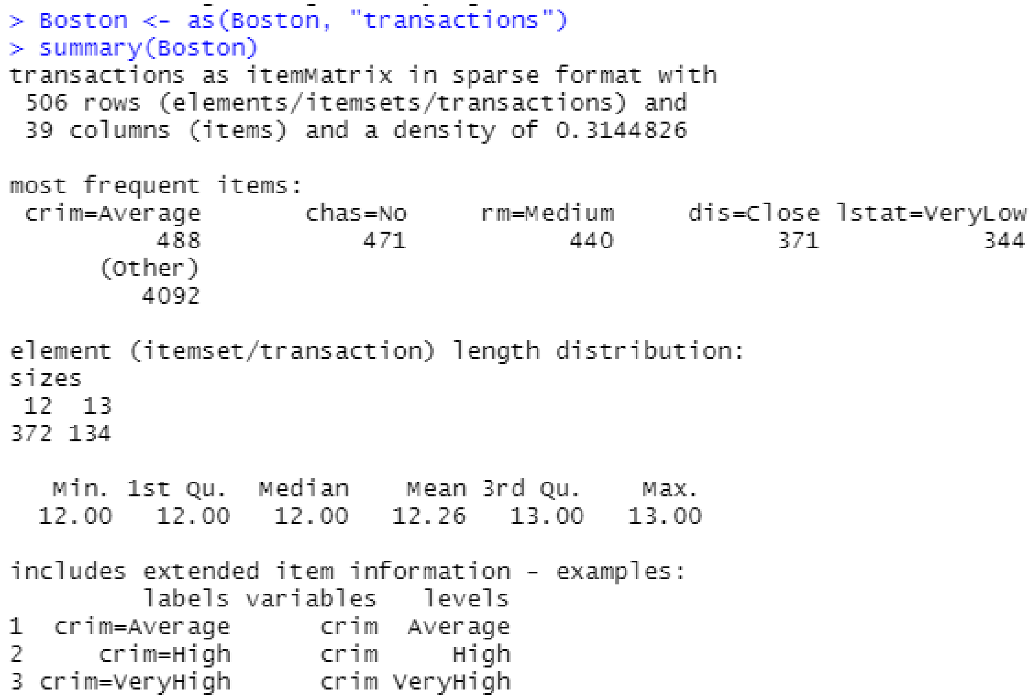
The lower status data can be further divided into Very Low, Middle, Class.

Similarly the data for median value for owner Occupied houses can be grouped into cheap, Medium and expensive.



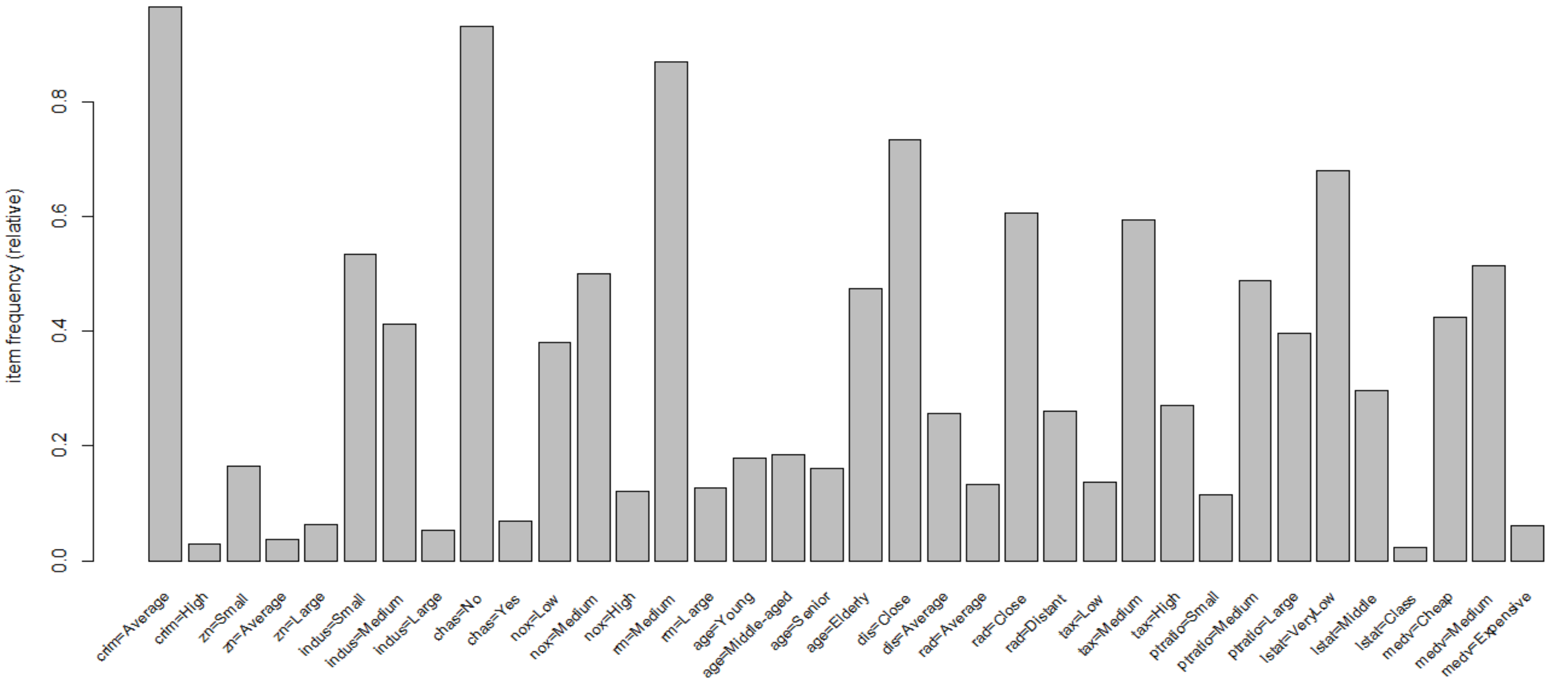
The data for crime rate in Boston suburbs likewise can be divided into Average, High and Very High.

Summary of <u>Boston Transaction Data –</u>



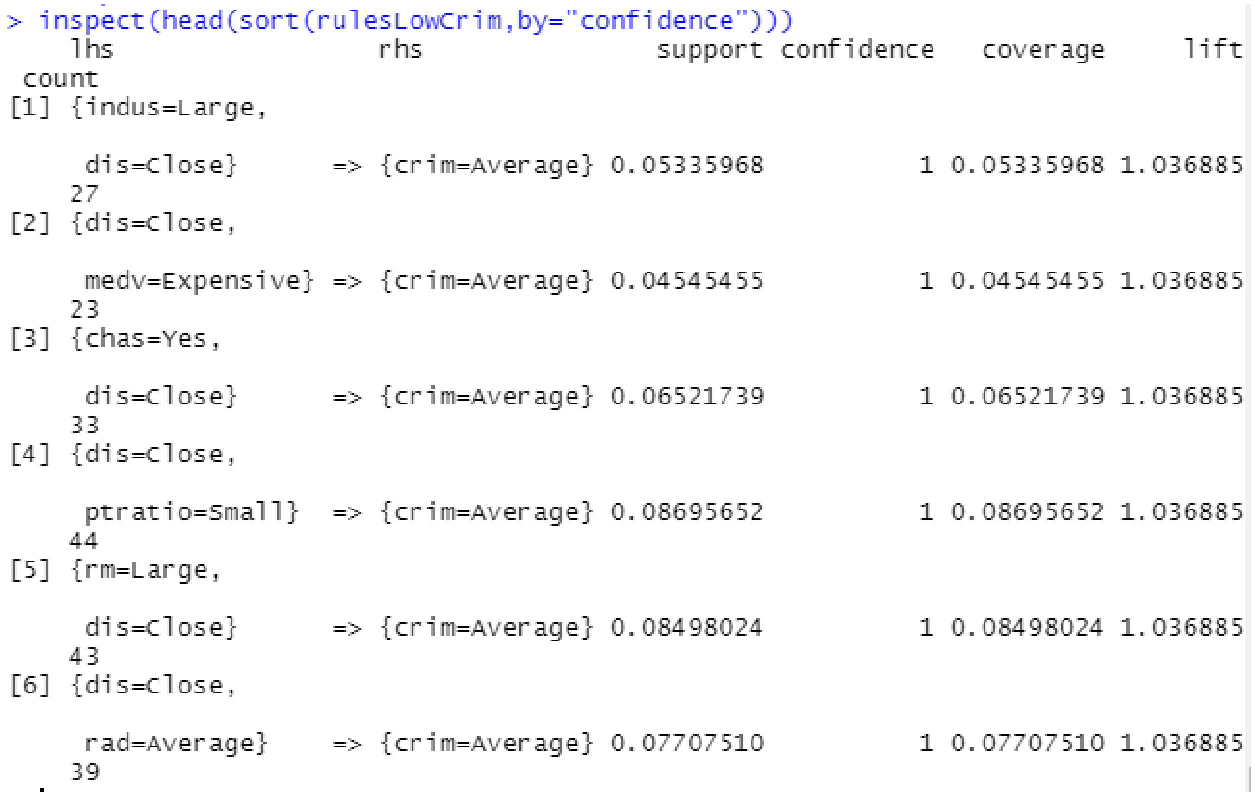
```
> Boston <- as(Boston, "transactions")
> summary(Boston)
transactions as itemMatrix in sparse format with
 506 rows (elements/itemsets/transactions) and
 39 columns (items) and a density of 0.3144826 

most frequent items:
 crim=Average       chas=No     rm=Medium     dis=Close lstat=VeryLow 
          488           471           440           371           344 
      (Other) 
         4092 

element (itemset/transaction) length distribution:
sizes
 12  13 
372 134 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.00   12.00   12.00   12.26   13.00   13.00 

includes extended item information - examples:
         labels variables   levels
1  crim=Average      crim  Average
2     crim=High      crim     High
3 crim=VeryHigh      crim VeryHigh
```

Item Frequency Plot of Boston Data –



The various rules for a Student who is interested in areas close to "weighted mean distance to 5 Boston employment centers" given that the region is also having minimum crime rate.

These rules are generated with a support value of "0.01" and confidence of "0.6"



```
> inspect(head(sort(rulesLowCrim,by="confidence")))
    lhs                  rhs              support confidence   coverage     lift
 count
[1] {indus=Large,

     dis=Close}      => {crim=Average} 0.05335968          1 0.05335968 1.036885
    27
[2] {dis=Close,

     medv=Expensive} => {crim=Average} 0.04545455          1 0.04545455 1.036885
    23
[3] {chas=Yes,

     dis=Close}      => {crim=Average} 0.06521739          1 0.06521739 1.036885
    33
[4] {dis=Close,

     ptratio=Small}  => {crim=Average} 0.08695652          1 0.08695652 1.036885
    44
[5] {rm=Large,

     dis=Close}      => {crim=Average} 0.08498024          1 0.08498024 1.036885
    43
[6] {dis=Close,

     rad=Average}    => {crim=Average} 0.07707510          1 0.07707510 1.036885
    39
```

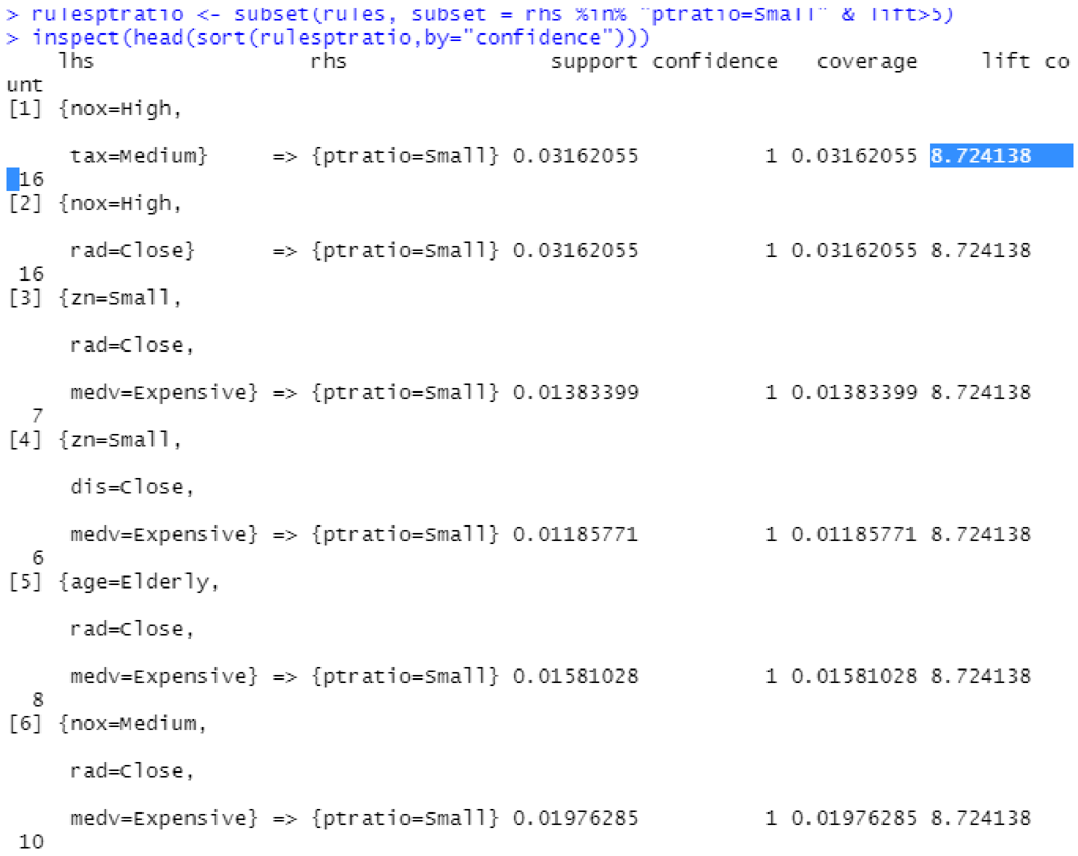
After inspecting the above data, the following inferences can be made for a student –

1)The student can move to regions having nonretail business where crime rate is at minimum level and distance is also considerable which is not enough logical though since the region is not residential

2)The student can also move to regions having student where median value of owner-occupied housing is expensive where crime rate is at minimum level and distance is also considerable but again this rule is having less clue, considering the factor the student might not be able to find housing within suitable range of cost in this region.

3)The student can move to regions having small pupil to student ratio where crime rate is at minimum level and distance is also considerable -this inference is quite conclusive of the fact that even though the student would be less able to interact with his mates, but the study environment would suit him, less of interference and more privacy for the student.

4)The student may also decide to move to regions having average distance to radial highways which would connect him to urban areas which is a quite logical inference given that the crime rate is low plus he would be able to better commute.



```
> rulesptratio <- subset(rules, subset = rhs %in% "ptratio=Small" & lift>5)
> inspect(head(sort(rulesptratio,by="confidence")))
    lhs                   rhs                   support confidence   coverage     lift co
unt
[1] {nox=High,

     tax=Medium}      => {ptratio=Small} 0.03162055          1 0.03162055 8.724138
 16
[2] {nox=High,

     rad=Close}       => {ptratio=Small} 0.03162055          1 0.03162055 8.724138
 16
[3] {zn=Small,

     rad=Close,

     medv=Expensive} => {ptratio=Small} 0.01383399          1 0.01383399 8.724138
  7
[4] {zn=Small,

     dis=Close,

     medv=Expensive} => {ptratio=Small} 0.01185771          1 0.01185771 8.724138
  6
[5] {age=Elderly,

     rad=Close,

     medv=Expensive} => {ptratio=Small} 0.01581028          1 0.01581028 8.724138
  8
[6] {nox=Medium,

     rad=Close,

     medv=Expensive} => {ptratio=Small} 0.01976285          1 0.01976285 8.724138
 10
```

The above rules are for Parents whose priority is schooling and ptratio is low.

These rules are generated with a support value of "0.01" and confidence of "0.6".

After inspecting the above data, the following valid inferences can be made for a student –

1)Rule 5 - Given that ptratio is small the student can move to regions mostly having elderly people that is rule 5 median value of owner-occupied homes is expensive.

2) Rule 4 - Also the student could move to regions where land under residential area is small but the dis value is close that is he would have easy accessibility to highways that can connect him to urban employment centers though median value of owner-occupied homes is expensive.

3) Rule 6 - Also the student could move to regions where nitrogen oxide conc level is medium though median value of owner-occupied homes is expensive.