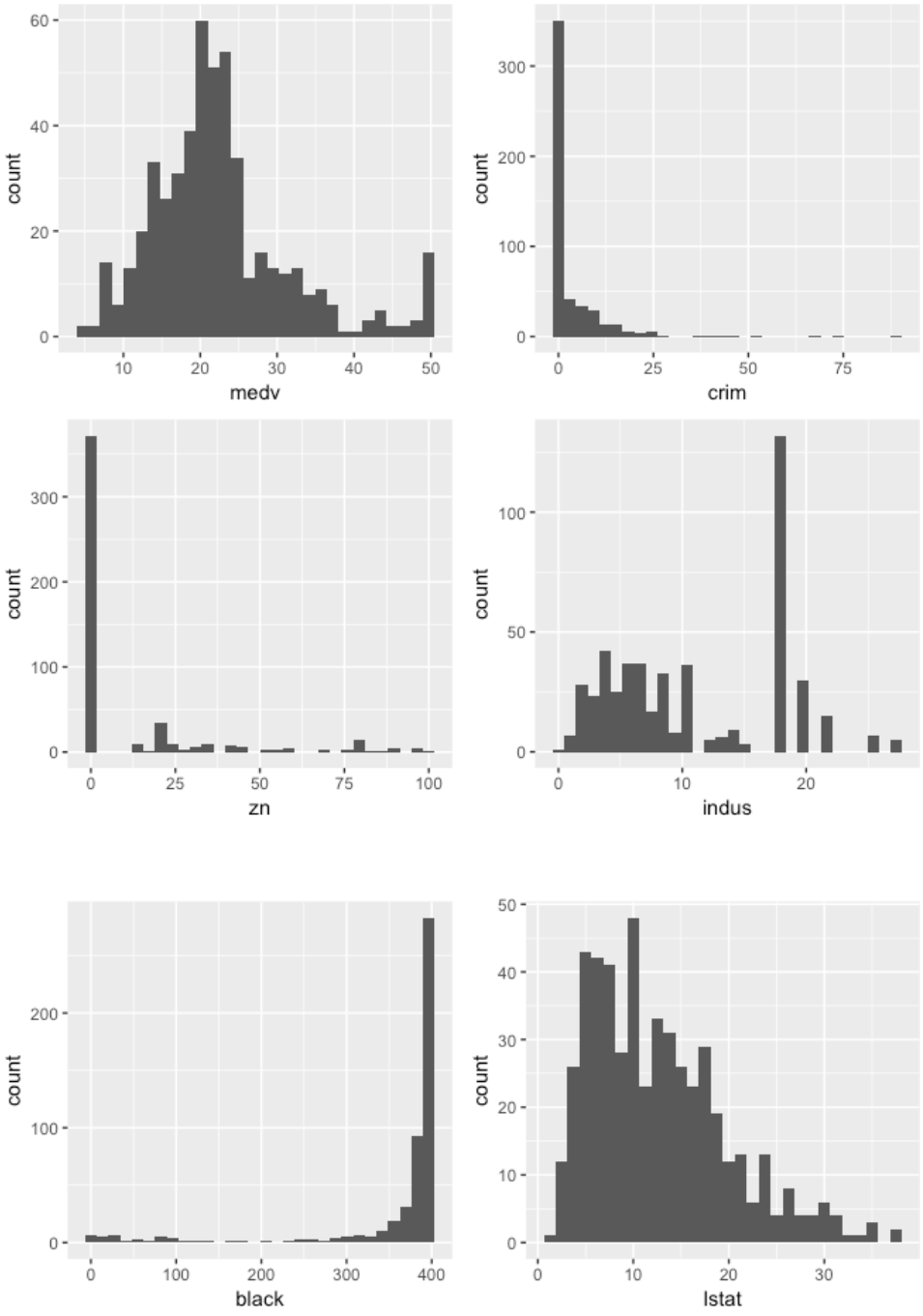
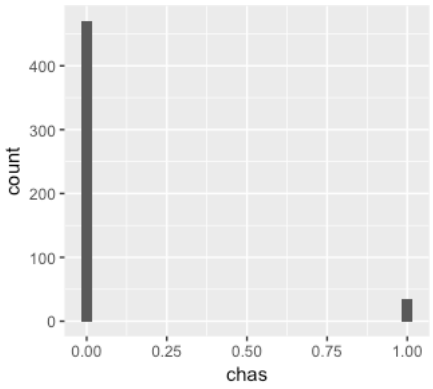
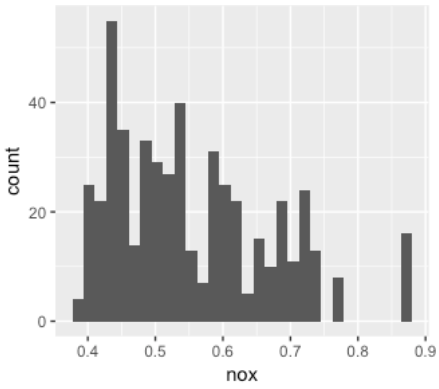
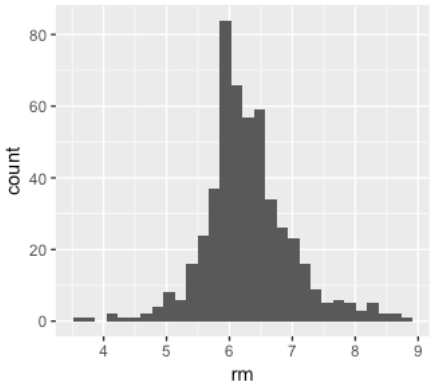
Boston Housing Data Analysis

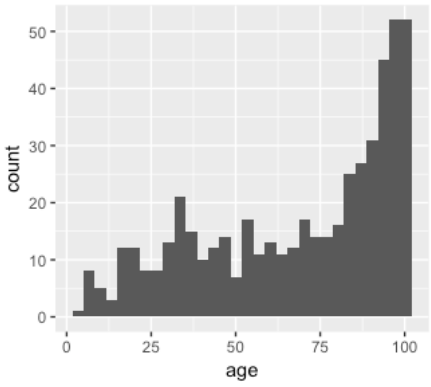
a) Visualizing the data using histograms of different variables in the data set:

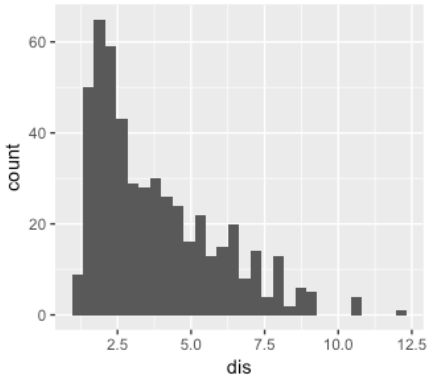


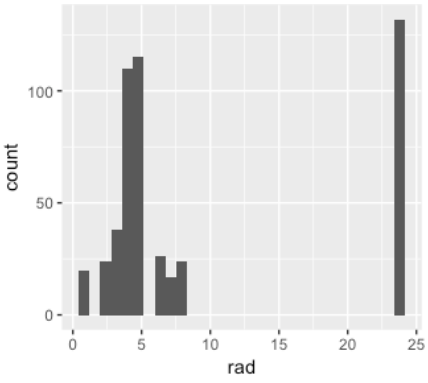

count
chas
0.00
0.25
0.50
0.75
1.00

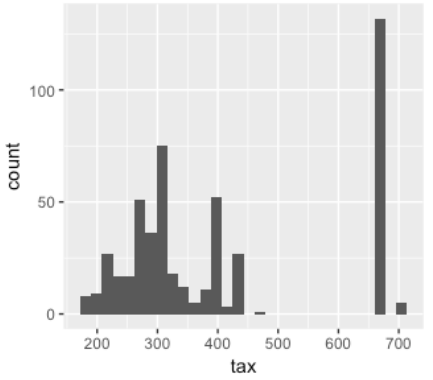

count
nox
0.4
0.5
0.6
0.7
0.8
0.9

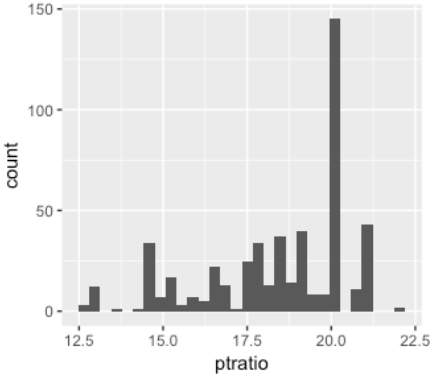

count
rm
4
5
6
7
8
9


count
age
0
25
50
75
100

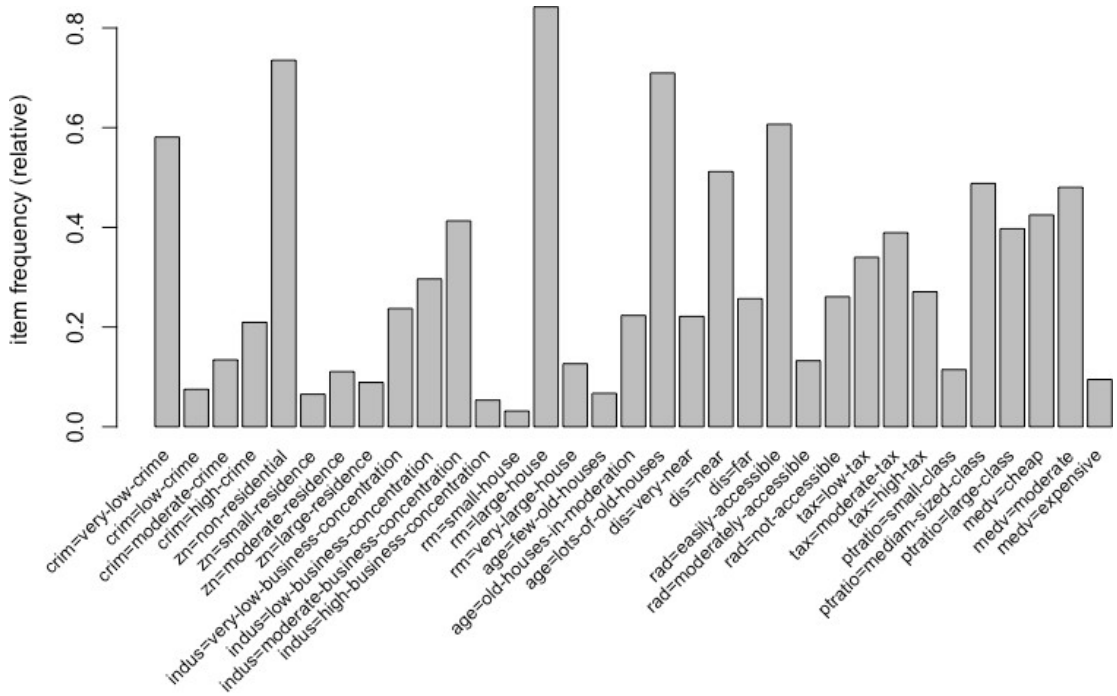

count
dis
2.5
5.0
7.5
10.0
12.5


count
rad
0
5
10
15
20
25


count
tax
200
300
400
500
600
700


count
ptratio
12.5
15.0
17.5
20.0
22.5

b) ItemFrequencyPlot:



We can infer from the above Item Frequency Plot that, 'newnox' and 'newmedv' have more frequency compared to other variables in the Boston Dataset.

Rules:

```
> summary(rules)
set of 220203 rules

rule length distribution (lhs + rhs):sizes
    1     2     3     4     5     6     7     8     9    10
    4   197  2554 13816 38401 61084 58510 33557 10646  1434

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     6.0     6.0     6.4     7.0    10.0

summary of quality measures:
    support         confidence          lift            count
 Min.   :0.00   Min.   :0.60   Min.   :  1   Min.   :  1
 1st Qu.:0.00   1st Qu.:0.94   1st Qu.:  1   1st Qu.:  1
 Median :0.01   Median :1.00   Median :  2   Median :  3
 Mean   :0.02   Mean   :0.94   Mean   :  3   Mean   :  8
 3rd Qu.:0.02   3rd Qu.:1.00   3rd Qu.:  3   3rd Qu.:  8
 Max.   :0.84   Max.   :1.00   Max.   :101   Max.   :426

mining info:
      data ntransactions support confidence
 Boston_tr            506   0.001        0.6
```

```
> rules <- apriori(Boston_tr, parameter = list(support = 0.001, confidence = 0.6))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.6    0.1    1 none FALSE            TRUE       5   0.001      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 0

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[34 item(s), 506 transaction(s)] done [0.00s].
sorting and recoding items ... [34 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.03s].
writing ... [220203 rule(s)] done [4.31s].
creating S4 object  ... done [0.06s].
```

c) Low crime area close to city:

```
> summary(rules_CloseToCity)
set of 2673 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5   6   7   8   9  10
  2  41 215 532 756 657 350 106  14

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     5.0     6.0     6.2     7.0    10.0

summary of quality measures:
    support         confidence        lift            count
 Min.   :0.002   Min.   :0.60   Min.   :2.7   Min.   : 1
 1st Qu.:0.004   1st Qu.:0.72   1st Qu.:3.3   1st Qu.: 2
 Median :0.008   Median :1.00   Median :4.5   Median : 4
 Mean   :0.019   Mean   :0.89   Mean   :4.0   Mean   :10
 3rd Qu.:0.024   3rd Qu.:1.00   3rd Qu.:4.5   3rd Qu.:12
 Max.   :0.132   Max.   :1.00   Max.   :4.5   Max.   :67

mining info:
      data ntransactions support confidence
 Boston_tr           506   0.001        0.6
```

```
> summary(rules_LowCrime)
set of 18187 rules

rule length distribution (lhs + rhs):sizes
   2    3    4    5    6    7    8    9   10
  14  186 1059 3058 5041 4942 2863  904  120

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     6.0     6.0     6.4     7.0    10.0

summary of quality measures:
    support         confidence        lift            count
 Min.   :0.00   Min.   :0.70   Min.   :1.20   Min.   :  1
 1st Qu.:0.00   1st Qu.:1.00   1st Qu.:1.72   1st Qu.:  1
 Median :0.01   Median :1.00   Median :1.72   Median :  3
 Mean   :0.01   Mean   :0.99   Mean   :1.70   Mean   :  7
 3rd Qu.:0.01   3rd Qu.:1.00   3rd Qu.:1.72   3rd Qu.:  7
 Max.   :0.47   Max.   :1.00   Max.   :1.72   Max.   :238

mining info:
      data ntransactions support confidence
 Boston_tr           506   0.001        0.6
```

```
> inspect(head(sort(rulesLowCrimeNearCity, by ='lift'),n = 6))
    lhs                                          rhs              support confidence lift count
[1] {zn=small-residence,
     dis=very-near}                          => {crim=low-crime}  0.0079          1   13     4
[2] {indus=very-low-business-concentration,
     dis=very-near}                          => {crim=low-crime}  0.0079          1   13     4
[3] {zn=small-residence,
     dis=very-near,
     medv=expensive}                         => {crim=low-crime}  0.0059          1   13     3
[4] {zn=small-residence,
     dis=very-near,
     ptratio=small-class}                    => {crim=low-crime}  0.0079          1   13     4
[5] {zn=small-residence,
     rm=very-large-house,
     dis=very-near}                          => {crim=low-crime}  0.0059          1   13     3
[6] {zn=small-residence,
     indus=very-low-business-concentration,
     dis=very-near}                          => {crim=low-crime}  0.0079          1   13     4
```

As observed from the above analysis, my advice to the student would be he should opt for a place that is far from his work place as the crime rate decreases as we go away from it. Also the chances of getting a large house in low crime rate is low so he might end up in a small residence.

d) Schools with low pupil-teacher ratios:

```
> summary(rulesLowPupil_TeacherRatio)
set of 3340 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6   7   8   9  10
 26 190 590 959 903 501 152  19

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    3.0     6.0     6.0     6.4     7.0    10.0

summary of quality measures:
    support         confidence        lift            count
 Min.   :0.002   Min.   :0.60   Min.   :5.2   Min.   : 1.0
 1st Qu.:0.004   1st Qu.:1.00   1st Qu.:8.7   1st Qu.: 2.0
 Median :0.006   Median :1.00   Median :8.7   Median : 3.0
 Mean   :0.010   Mean   :0.94   Mean   :8.2   Mean   : 4.9
 3rd Qu.:0.010   3rd Qu.:1.00   3rd Qu.:8.7   3rd Qu.: 5.0
 Max.   :0.059   Max.   :1.00   Max.   :8.7   Max.   :30.0

mining info:
      data ntransactions support confidence
 Boston_tr           506   0.001        0.6
```

```
> inspect(head(sort(rulesLowPupil_TeacherRatio, by ='lift'),n = 6))
    lhs                                                  rhs                    support confidence lift count
[1] {crim=low-crime,zn=small-residence}                  => {ptratio=small-class} 0.0237  1          8.7  12
[2] {zn=small-residence,dis=very-near}                   => {ptratio=small-class} 0.0079  1          8.7   4
[3] {crim=low-crime,indus=very-low-business-concentration} => {ptratio=small-class} 0.0237  1          8.7  12
[4] {crim=low-crime,tax=low-tax}                         => {ptratio=small-class} 0.0237  1          8.7  12
[5] {indus=very-low-business-concentration,dis=very-near} => {ptratio=small-class} 0.0079  1          8.7   4
[6] {rm=small-house,dis=very-near,tax=moderate-tax}      => {ptratio=small-class} 0.0040  1          8.7   2
```

Advice to the family would be for getting a school with low pupil-teacher ratios they would have to sacrifice on the size of the house but the house would be very near to the school. Also, the business concentration around the residence near to these school is low. But the good thing is there are many industries near the school so if a member of family is looking for a new job nearby this would be ideal setting. There's 79% chance of this happening.

Regression Model:

```
Call:
lm(formula = newptratio ~ ., data = subset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97694 -0.07909  0.07719  0.32456  0.63415

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.30758    0.09197   3.344 0.000887 ***
newcrim      0.12589    0.04585   2.746 0.006255 **
newindus    -0.12126    0.05040  -2.406 0.016505 *
newnox      -0.15629    0.05520  -2.831 0.004823 **
newage       0.12954    0.04666   2.776 0.005706 **
newdis       0.07302    0.04677   1.561 0.119119
newtax       0.69147    0.05135  13.466  < 2e-16 ***
newblack     0.05416    0.03792   1.428 0.153800
newmedv     -0.35566    0.04769  -7.457 3.96e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3696 on 497 degrees of freedom
Multiple R-squared:  0.4631,    Adjusted R-squared:  0.4544
F-statistic: 53.58 on 8 and 497 DF,  p-value: < 2.2e-16
```

Regression follows the similar trend where black proportion is less, crime is less, and distance to employment center is less. When the features are mostly non-categorical regression is generally preferred.