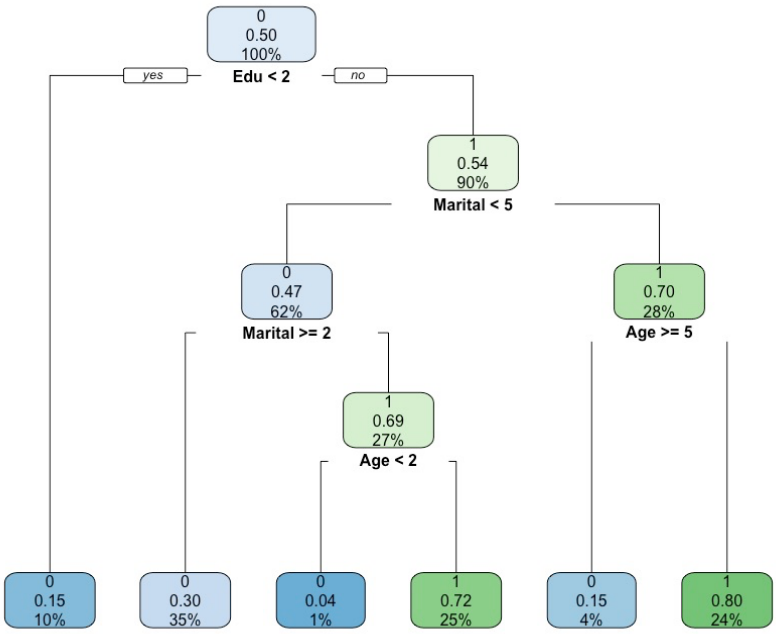
Name : Manasa Challa
UB Person ID : 50356441

REPORT HOMEWORK – 2

Question 1:

(20pointsModifiedExercise14.4inESL)
Cluster the marketing data of Table 14.1 (ESL) using a classification tree. This data is in the ISLR package, and also available on UB learns.Specifically, generate a reference sample of the same size of the training set. This can be done in a couple of ways, e.g., (i) sample uniformly for each variable, or (ii) by randomly permuting the values within each variable independently. Build a classification tree to the training sample (class 1) and the reference sample (class 0) and describe the terminal nodes having highest estimated class 1 probability. Compare the results to the results near Table 14.1 (ESL), which were derived using PRIM.




```
> summary(prediction)
       0                1         
 Min.   :0.2005   Min.   :0.03704  
 1st Qu.:0.2781   1st Qu.:0.29851  
 Median :0.7015   Median :0.29851  
 Mean   :0.5000   Mean   :0.50000  
 3rd Qu.:0.7015   3rd Qu.:0.72191  
 Max.   :0.9630   Max.   :0.79953  
> prediction
           0         1
1  0.2780948 0.72190518
2  0.2780948 0.72190518
3  0.2780948 0.72190518
4  0.2004695 0.79953052
5  0.2004695 0.79953052
6  0.2780948 0.72190518
7  0.2004695 0.79953052
8  0.7014925 0.29850746
9  0.2780948 0.72190518
10 0.2780948 0.72190518
11 0.2004695 0.79953052
12 0.7014925 0.29850746
13 0.2780948 0.72190518
14 0.2780948 0.72190518
15 0.7014925 0.29850746
16 0.2780948 0.72190518
```

Name : Manasa Challa
UB Person ID : 50356441

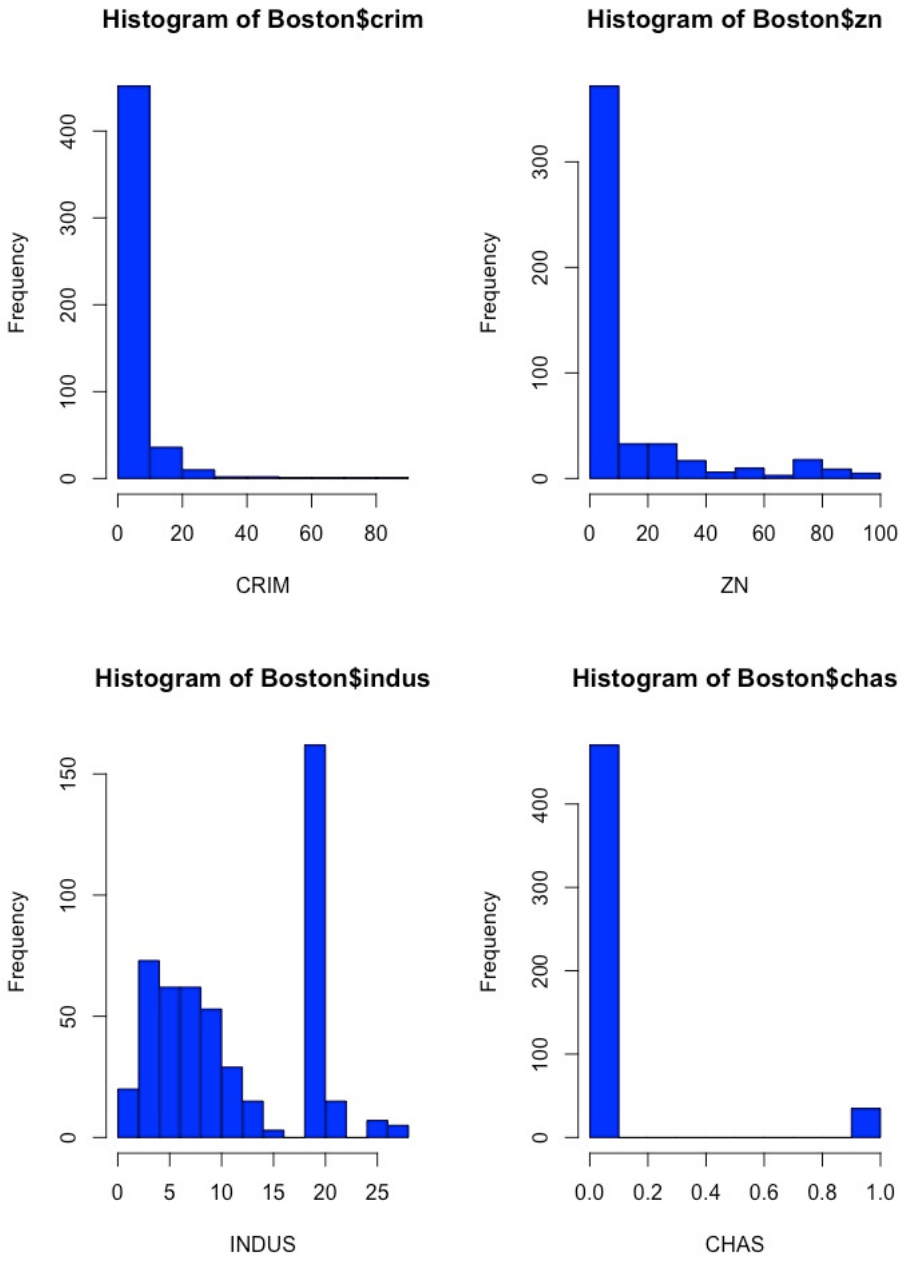
We can observe from the above that these features can predict to do a classification. To double check this we can predict the model on the training set. Therefore, confirming our assumption that they do not have any predictive power. The terminal node had percentage of 20% with a class 1 probability of 72%. Comparing the results with PRIM we notice that the values as 0.08 when the household were more than equal to 3 and similarity in the trends can be observed in case of 0.25 when the household values were lesser than 3 and language was lower than 2.

Name : Manasa Challa
UB Person ID : 50356441

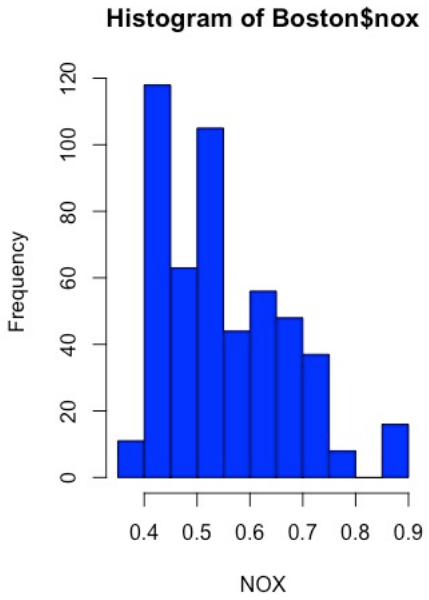
Question 2:

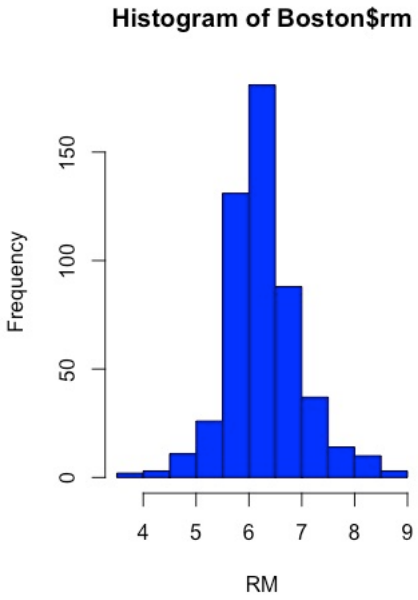
A) Visualize the data using histograms of the different variables in the data set. Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.

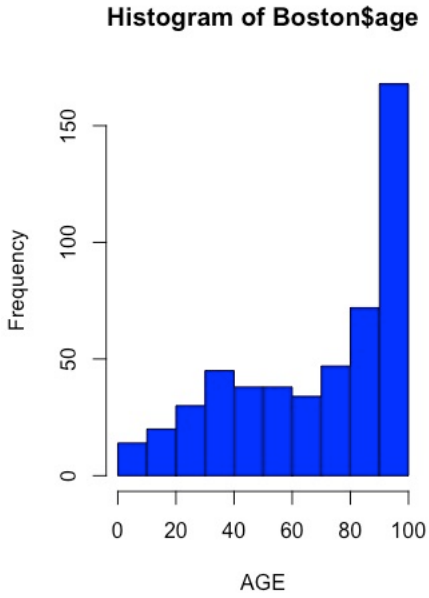
The choices I made in grouping categories was done taking the quartile, mean, minimum and maximum values of the data (summary) into account .

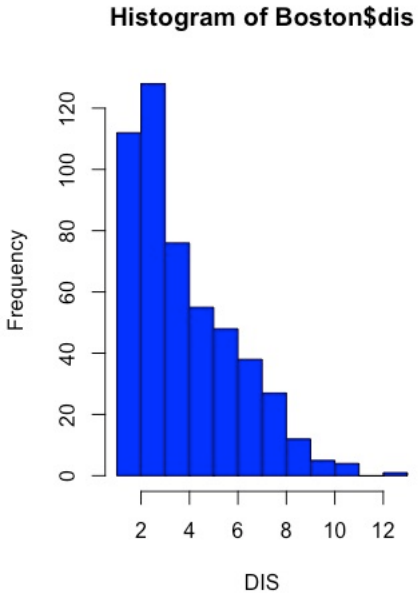


Name : Manasa Challa
UB Person ID : 50356441

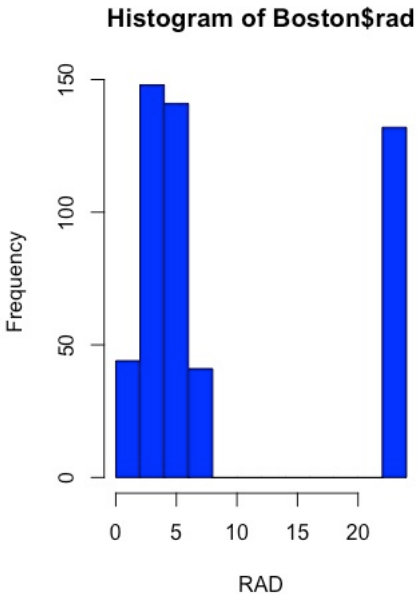


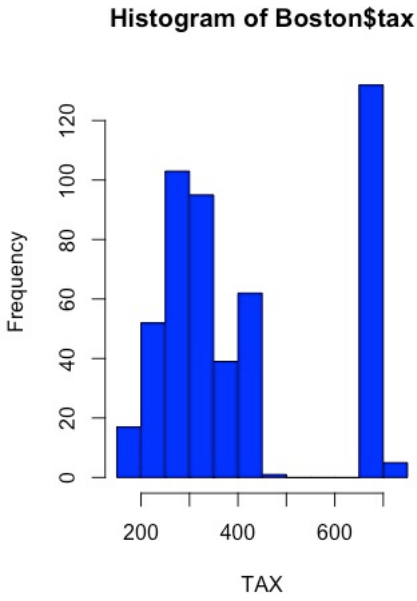


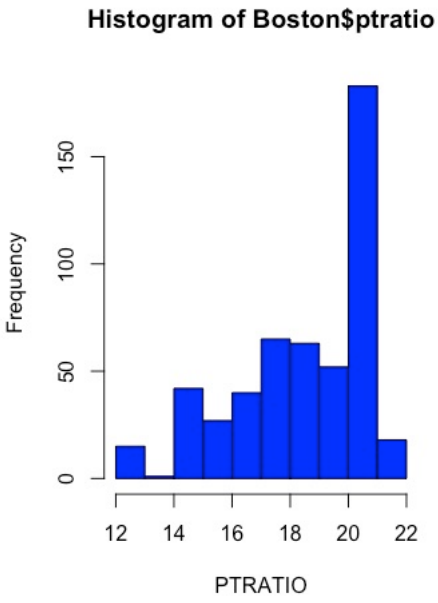


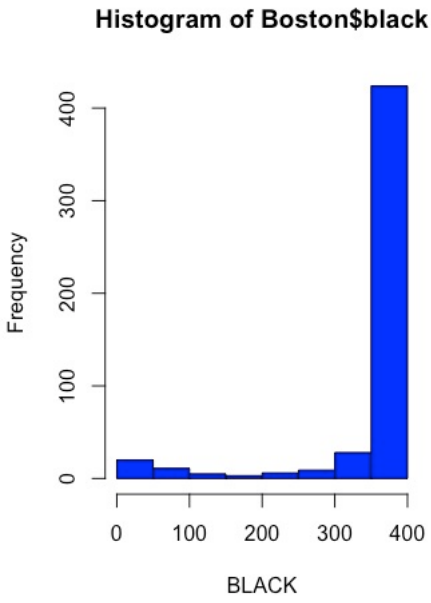


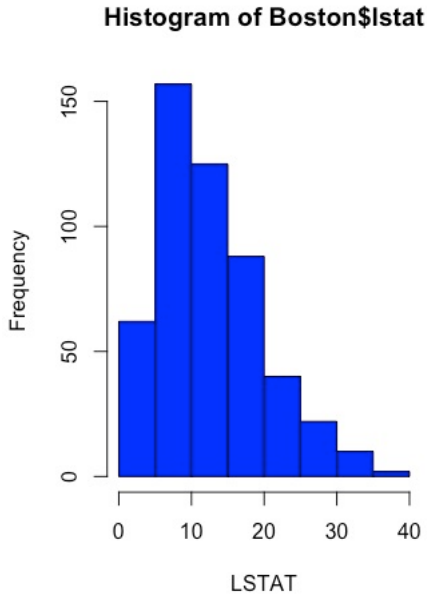
Name : Manasa Challa
UB Person ID : 50356441











Name : Manasa Challa
UB Person ID : 50356441

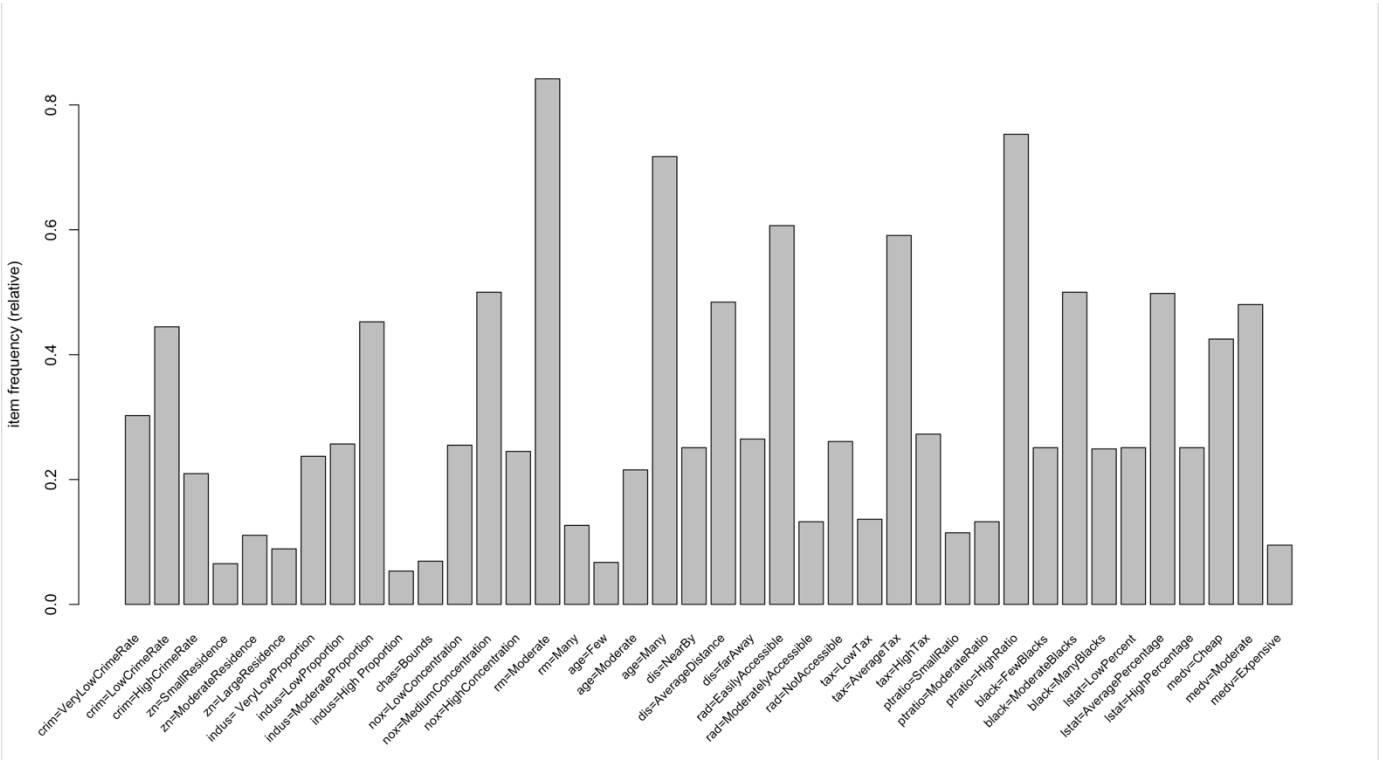
B) Visualize the data using the itemFrequencyPlot in the “arules” package. Apply the apriori algorithm (Do not forget to specify parameters in your write up).

Parameters :

support = 0.02 -> itemFrequencyPlot

Apriori Algorithm with the following parameters was using :

support = 0.02, confidence = 0.8



C) A student is interested is a low crime area, but wants to be as close to the city as possible (as measured by “dis”). What can you advise on this matter through the mining of association rules?

```
> summary(LowCrimeNearCity)
set of 588 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6   7   8   9  10
  4  38 116 175 151  78  23   3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   5.000   6.000   6.313   7.000  10.000

summary of quality measures:
    support          confidence        coverage            lift           count
 Min.   :0.02174   Min.   :0.8421   Min.   :0.02174   Min.   :1.894   Min.   :11.00
 1st Qu.:0.02372   1st Qu.:0.9200   1st Qu.:0.02569   1st Qu.:2.069   1st Qu.:12.00
 Median :0.02569   Median :0.9375   Median :0.02767   Median :2.108   Median :13.00
 Mean   :0.02976   Mean   :0.9432   Mean   :0.03155   Mean   :2.121   Mean   :15.06
 3rd Qu.:0.03162   3rd Qu.:1.0000   3rd Qu.:0.03557   3rd Qu.:2.249   3rd Qu.:16.00
 Max.   :0.08696   Max.   :1.0000   Max.   :0.09289   Max.   :2.249   Max.   :44.00
```

Name : Manasa Challa
UB Person ID : 50356441

From the above observed data it can be said that the student should be choosing a house away from work as the crime rate decrease and also an area that has low pupil teacher ratio.

D) A family is moving to the area, and has made schooling a priority. They want schools with low pupil-teacher ratios. What can you advise on this matter through the mining of association rules?

```
> LowPupilTeacherRatio <- subset(rules, subset = rhs %in% "ptratio=SmallRatio" & lift >1.2)
> summary(LowPupilTeacherRatio)
set of 671 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6   7   8   9  10
  6  48 142 211 167  76  19   2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   5.000   6.000   6.191   7.000  10.000

summary of quality measures:
    support          confidence        coverage            lift           count
 Min.   :0.02174   Min.   :0.8000   Min.   :0.02174   Min.   :6.979   Min.   :11.00
 1st Qu.:0.02372   1st Qu.:0.9375   1st Qu.:0.02372   1st Qu.:8.179   1st Qu.:12.00
 Median :0.02569   Median :1.0000   Median :0.02767   Median :8.724   Median :13.00
 Mean   :0.02822   Mean   :0.9663   Mean   :0.02934   Mean   :8.430   Mean   :14.28
 3rd Qu.:0.03162   3rd Qu.:1.0000   3rd Qu.:0.03162   3rd Qu.:8.724   3rd Qu.:16.00
 Max.   :0.05138   Max.   :1.0000   Max.   :0.05534   Max.   :8.724   Max.   :26.00

mining info:
    data ntransactions support confidence
 bmatrix           506    0.02        0.8
> inspect(head(sort(LowPupilTeacherRatio, by ='lift'),n = 6))
    lhs                                rhs                     support confidence   coverage     lift count
[1] {nox=HighConcentration,
     tax=AverageTax}           => {ptratio=SmallRatio} 0.05138340          1 0.05138340 8.724138    26
[2] {nox=HighConcentration,
     rad=EasilyAccessible}     => {ptratio=SmallRatio} 0.05138340          1 0.05138340 8.724138    26
[3] {zn=SmallResidence,
     indus= VeryLowProportion,
     medv=Expensive}           => {ptratio=SmallRatio} 0.02173913          1 0.02173913 8.724138    11
[4] {crim=LowCrimeRate,
     zn=SmallResidence,
     indus= VeryLowProportion} => {ptratio=SmallRatio} 0.02569170          1 0.02569170 8.724138    13
[5] {nox=HighConcentration,
     dis=NearBy,
     tax=AverageTax}           => {ptratio=SmallRatio} 0.04347826          1 0.04347826 8.724138    22
[6] {nox=HighConcentration,
     dis=NearBy,
     rad=EasilyAccessible}     => {ptratio=SmallRatio} 0.04347826          1 0.04347826 8.724138    22
```

From the above observed data it can be said that the family should move to a small residence where it has low proportion of non-retail business acres per town and also the medv is expensive. Also need to look at the average tax. We can observe the combinations in the lhs column in the above attached output to advise the family on what decisions to make.

Extra Credit: Use a regression model to solve part d. Are you results comparable? Which provides an easier interpretation? When would regression be preferred, and when would association models be preferred?

Name : Manasa Challa
UB Person ID : 50356441

```
> regression <- lm(ptratio~., data=boston_dataset)
> summary(regression)

Call:
lm(formula = ptratio ~ ., data = boston_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1190 -1.0126 -0.0060  0.8961  4.8945

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.484e+01  1.352e+00  18.379  < 2e-16 ***
crim        -1.578e-02  1.085e-02  -1.454  0.14661
zn          -2.473e-02  4.408e-03  -5.611 3.35e-08 ***
indus        5.722e-02  1.997e-02   2.865  0.00434 **
chas        -2.824e-01  2.846e-01  -0.992  0.32152
nox         -1.050e+01  1.187e+00  -8.848  < 2e-16 ***
rm          -7.076e-02  1.479e-01  -0.478  0.63255
age          7.198e-03  4.313e-03   1.669  0.09577 .
dis         -2.187e-02  6.883e-02  -0.318  0.75084
rad          1.177e-01  2.154e-02   5.465 7.35e-08 ***
tax          6.983e-04  1.244e-03   0.561  0.57491
black        1.573e-03  8.873e-04   1.773  0.07692 .
lstat       -3.770e-02  1.824e-02  -2.067  0.03929 *
medv        -1.021e-01  1.402e-02  -7.283 1.31e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.554 on 492 degrees of freedom
Multiple R-squared:  0.4982,    Adjusted R-squared:  0.485
F-statistic: 37.58 on 13 and 492 DF,  p-value: < 2.2e-16
```

Regression model is used when you want to predict a continuous dependent variable from a number of independent variables.