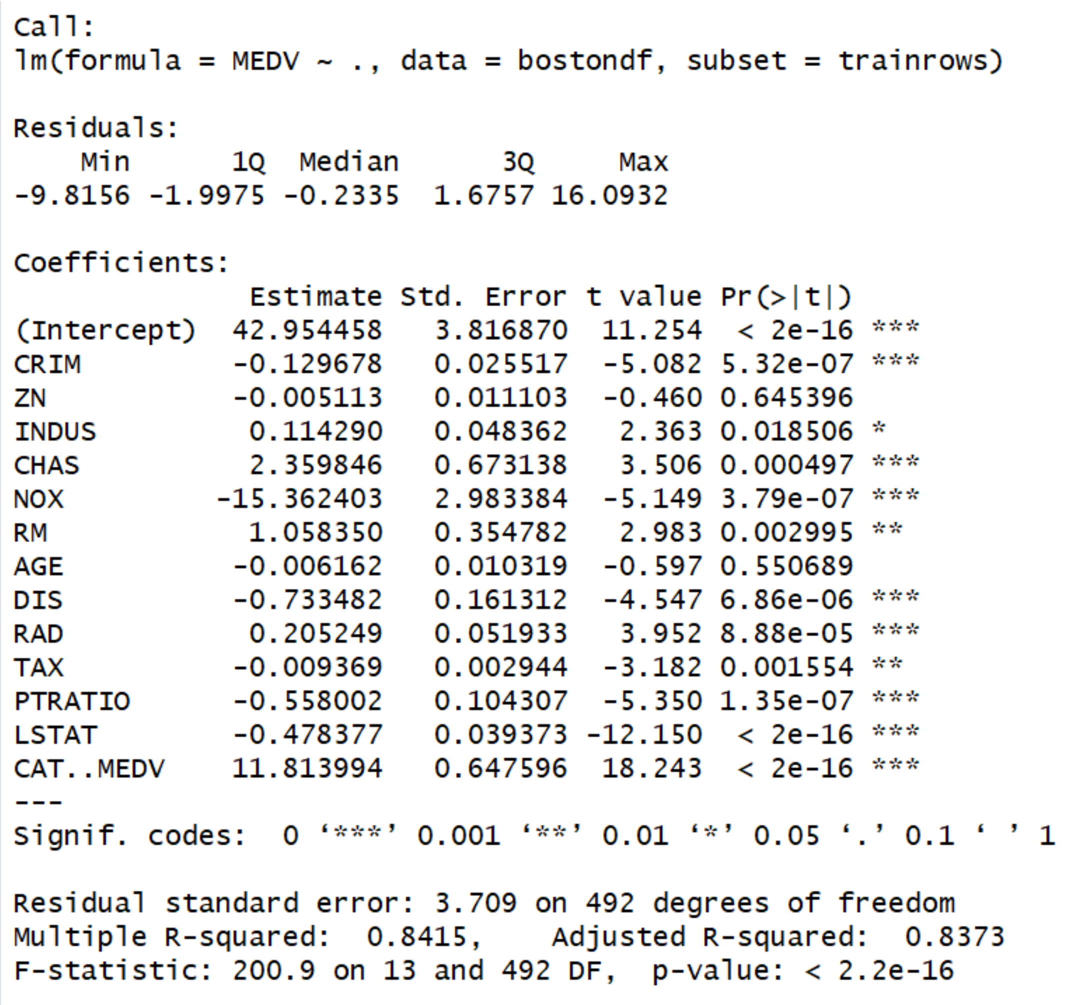
Program-6 To perform dimensionality reduction operation using PCA for Houses Data Set

Many times there are independent variables / features in the model that are codependent on each other and when one runs their correlation matrix one might see that there is a high correlation between each other. When such both variables are included in the regression model this will be like the fact that much of the variance of one of the variables has already been captured by the other variable. Let us assume that both variables are correlated as 0.69 hence when we include both of these variables in the regression model, then the 69% of the variance is already accounted for by one of the variables for the model, hence adding the other feature will not add any additional value. This is especially useful when we have a huge count of independent variables and we need to reduce the count of the model independent variables, and make our model more compact with a limited set of the independent variables.

Hence if we run the Boston housing data set using all of the variables, we will get this multiple regression output. This regression uses all of the 13 variables for the regression.



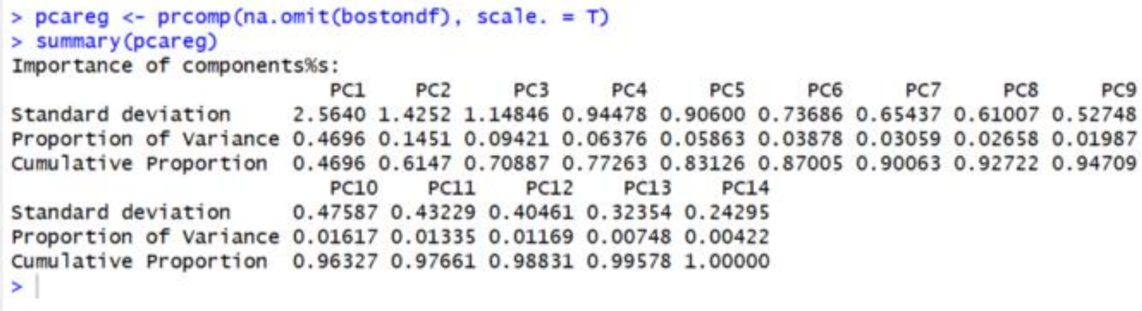
```
Call:
lm(formula = MEDV ~ ., data = bostondf, subset = trainrows)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.8156 -1.9975 -0.2335  1.6757 16.0932 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  42.954458   3.816870  11.254  < 2e-16 ***
CRIM         -0.129678   0.025517  -5.082 5.32e-07 ***
ZN           -0.005113   0.011103  -0.460 0.645396    
INDUS         0.114290   0.048362   2.363 0.018506 *  
CHAS          2.359846   0.673138   3.506 0.000497 ***
NOX         -15.362403   2.983384  -5.149 3.79e-07 ***
RM            1.058350   0.354782   2.983 0.002995 ** 
AGE          -0.006162   0.010319  -0.597 0.550689    
DIS          -0.733482   0.161312  -4.547 6.86e-06 ***
RAD           0.205249   0.051933   3.952 8.88e-05 ***
TAX          -0.009369   0.002944  -3.182 0.001554 ** 
PTRATIO      -0.558002   0.104307  -5.350 1.35e-07 ***
LSTAT        -0.478377   0.039373 -12.150  < 2e-16 ***
CAT..MEDV    11.813994   0.647596  18.243  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.709 on 492 degrees of freedom
Multiple R-squared:  0.8415,	Adjusted R-squared:  0.8373 
F-statistic: 200.9 on 13 and 492 DF,  p-value: < 2.2e-16
```

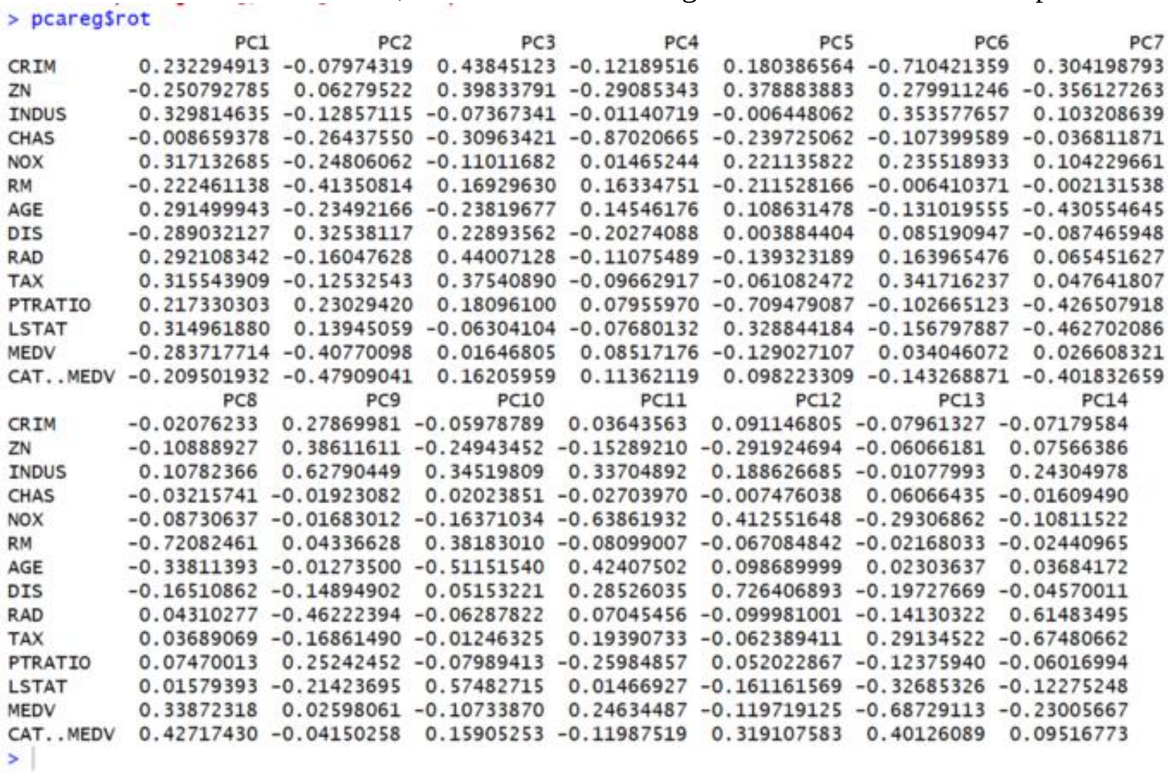
However, now let us run the PCA (Principal component analysis), and see which variables count how much of the variation and after how many features does the features stop adding any more value. We will also omit values which do not have any value in the data frame. The new variance calculation is given as below.



```
> pcareg <- prcomp(na.omit(bostondf), scale. = T)
> summary(pcareg)
Importance of components%s:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.5640 1.4252 1.14846 0.94478 0.90600 0.73686 0.65437 0.61007 0.52748
Proportion of Variance 0.4696 0.1451 0.09421 0.06376 0.05863 0.03878 0.03059 0.02658 0.01987
Cumulative Proportion  0.4696 0.6147 0.70887 0.77263 0.83126 0.87005 0.90063 0.92722 0.94709
                          PC10    PC11    PC12    PC13    PC14
Standard deviation     0.47587 0.43229 0.40461 0.32354 0.24295
Proportion of Variance 0.01617 0.01335 0.01169 0.00748 0.00422
Cumulative Proportion  0.96327 0.97661 0.98831 0.99578 1.00000
> 
```

Running the principal component analysis shows that after adding the 9th variables this has already accounted for 95% (0.94709) of the variance that we were expecting, and we can run the revised model with only nine parameters and we would get significant results as well for our multiple linear regression model. The option scale.=T lets us make the data normalized, which is important where some features are off scale.

The rotation matrix is as below, which shows the weights used to create the new points.



```
> pcareg$rot
                  PC1         PC2         PC3         PC4          PC5          PC6          PC7
CRIM      0.232294913 -0.07974319  0.43845123 -0.12189516  0.180386564 -0.710421359  0.304198793
ZN       -0.250792785  0.06279522  0.39833791 -0.29085343  0.378883883  0.279911246 -0.356127263
INDUS     0.329814635 -0.12857115 -0.07367341 -0.01140719 -0.006448062  0.353577657  0.103208639
CHAS     -0.008659378 -0.26437550 -0.30963421 -0.87020665 -0.239725062 -0.107399589 -0.036811871
NOX       0.317132685 -0.24806062 -0.11011682  0.01465244  0.221135822  0.235518933  0.104229661
RM       -0.222461138 -0.41350814  0.16929630  0.16334751 -0.211528166 -0.006410371 -0.002131538
AGE       0.291499943 -0.23492166 -0.23819677  0.14546176  0.108631478 -0.131019555 -0.430554645
DIS      -0.289032127  0.32538117  0.22893562 -0.20274088  0.003884404  0.085190947 -0.087465948
RAD       0.292108342 -0.16047628  0.44007128 -0.11075489 -0.139323189  0.163965476  0.065451627
TAX       0.315543909 -0.12532543  0.37540890 -0.09662917 -0.061082472  0.341716237  0.047641807
PTRATIO   0.217330303  0.23029420  0.18096100  0.07955970 -0.709479087 -0.102665123 -0.426507918
LSTAT     0.314961880  0.13945059 -0.06304104 -0.07680132  0.328844184 -0.156797887 -0.462702086
MEDV     -0.283717714 -0.40770098  0.01646805  0.08517176 -0.129027107  0.034046072  0.026608321
CAT..MEDV -0.209501932 -0.47909041  0.16205959  0.11362119  0.098223309 -0.143268871 -0.401832659
                 PC8         PC9        PC10        PC11         PC12        PC13        PC14
CRIM     -0.02076233  0.27869981 -0.05978789  0.03643563  0.091146805 -0.07961327 -0.07179584
ZN       -0.10888927  0.38611611 -0.24943452 -0.15289210 -0.291924694 -0.06066181  0.07566386
INDUS     0.10782366  0.62790449  0.34519809  0.33704892  0.188626685 -0.01077993  0.24304978
CHAS     -0.03215741 -0.01923082  0.02023851 -0.02703970 -0.007476038  0.06066435 -0.01609490
NOX      -0.08730637 -0.01683012 -0.16371034 -0.63861932  0.412551648 -0.29306862 -0.10811522
RM       -0.72082461  0.04336628  0.38183010 -0.08099007 -0.067084842 -0.02168033 -0.02440965
AGE      -0.33811393 -0.01273500 -0.51151540  0.42407502  0.098689999  0.02303637  0.03684172
DIS      -0.16510862 -0.14894902  0.05153221  0.28526035  0.726406893 -0.19727669 -0.04570011
RAD       0.04310277 -0.46222394 -0.06287822  0.07045456 -0.099981001 -0.14130322  0.61483495
TAX       0.03689069 -0.16861490 -0.01246325  0.19390733 -0.062389411  0.29134522 -0.67480662
PTRATIO   0.07470013  0.25242452 -0.07989413 -0.25984857  0.052022867 -0.12375940 -0.06016994
LSTAT     0.01579393 -0.21423695  0.57482715  0.01466927 -0.161161569 -0.32685326 -0.12275248
MEDV      0.33872318  0.02598061 -0.10733870  0.24634487 -0.119719125 -0.68729113 -0.23005667
CAT..MEDV 0.42717430 -0.04150258  0.15905253 -0.11987519  0.319107583  0.40126089  0.09516773
> 
```