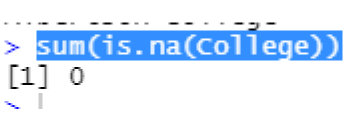
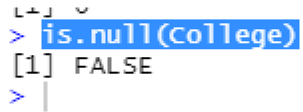
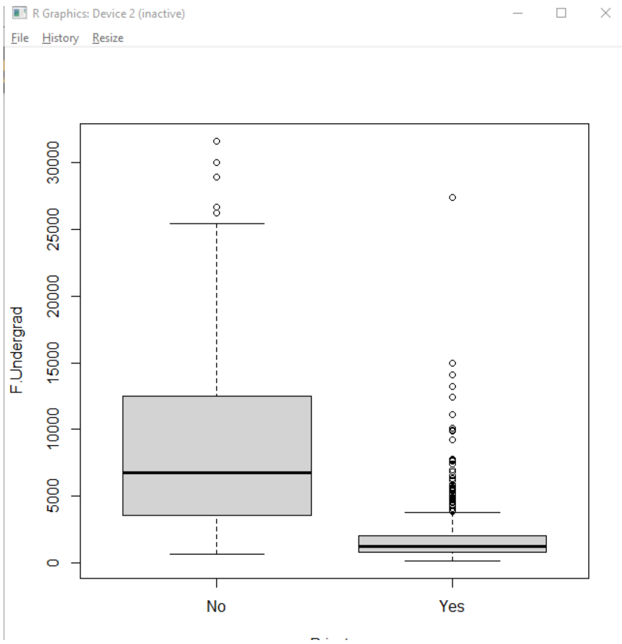
+++++++Checking for Missing Values and Null Values

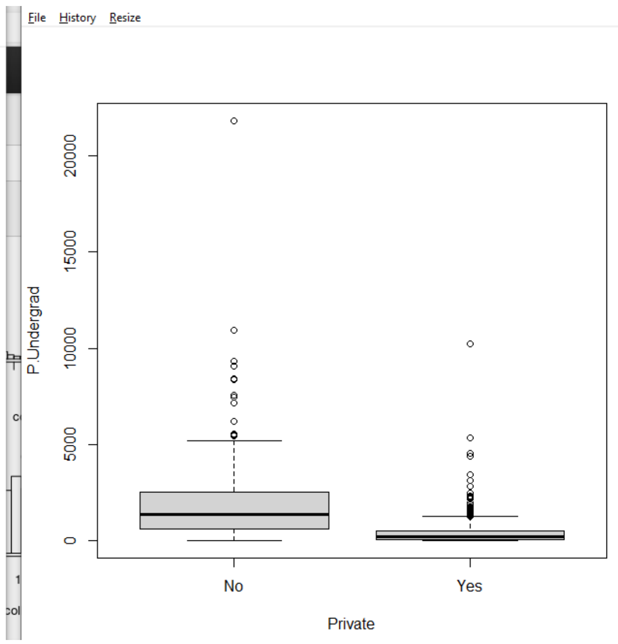




All the below visualization plots are performed without scaling the data ---

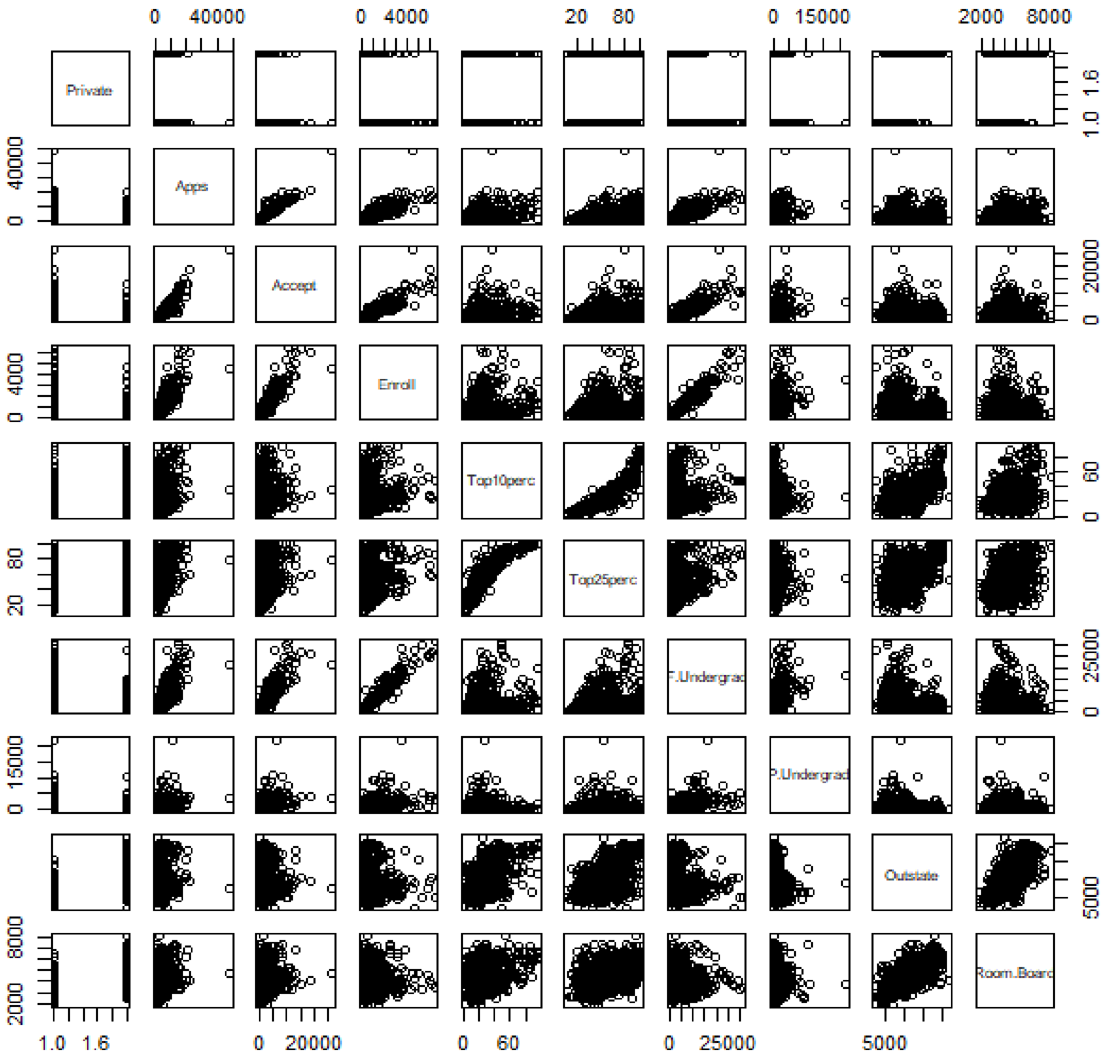
Boxplots of Full time and Part Time Undergraduates



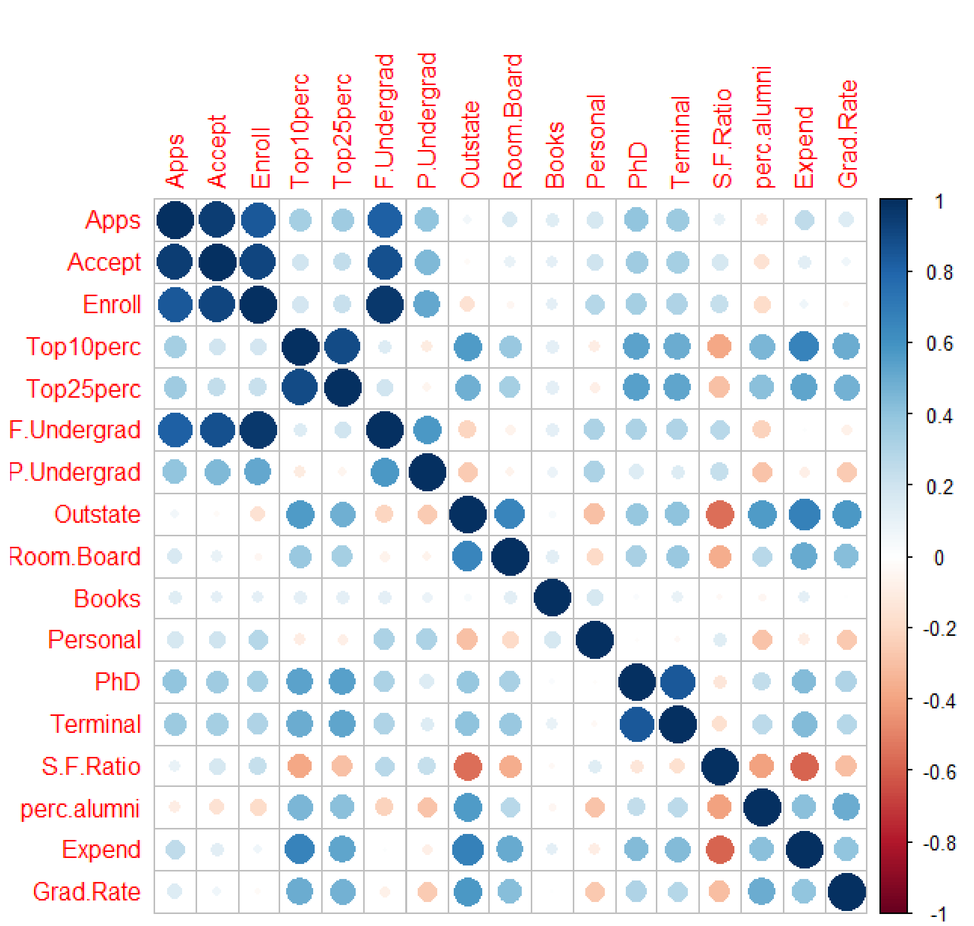


From the boxplots of part time and fulltime under graduate students versus whether they are enrolled in Private or Public schools its obvious that mostly the students whether parttime or full time are enrolled for public schools and also majority of them are Full time students

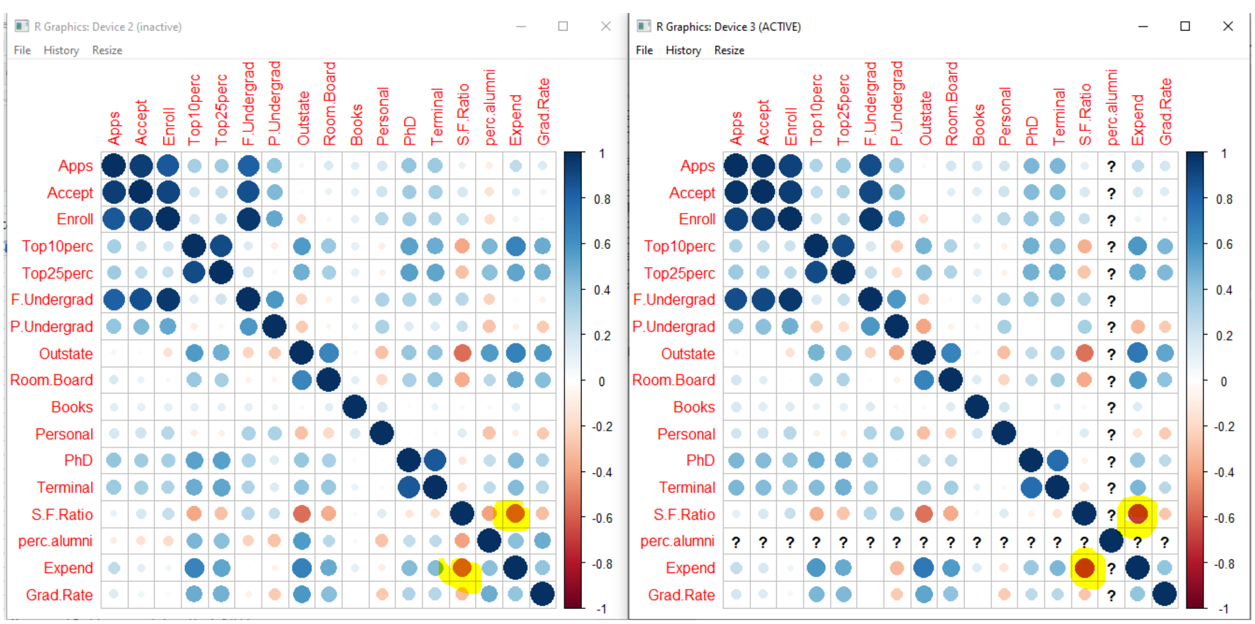
Pairs plot of College Data:



The pairs plot does not seem to infer any useful information except the high positive correlation between Enrolled students and Full time graduate students ,Out State Tuition fees and Room Board Costs. Also the same information is retrieved from the below correlation plot.



After scaling all the continuous variables to the same scale:



The left plot indicates the unscaled College data and right plot indicates the scaled college data(Log Transformation).So the only change that is getting reflected from the scaled version of college data as

highlighted in yellow is that the Student Faculty Ratio and Instructional expenditure per student has become more negatively correlated, in other words as no. of student increases (faculty assumed to be fixed) or the faculty number reduced(keeping the number of students fixed),the Instructional expenditure per student is getting reduced and vice versa.

Ordering by decreasing value of Apps –

a)Public College Data –

```
> Public_College_Data[1:5,]
                                         Private  Apps Accept Enroll Top10perc
Rutgers at New Brunswick                      No 48094  26330   4520        36
Purdue University at West Lafayette           No 21804  18744   5874        29
University of California at Berkeley          No 19873   8252   3215        95
Pennsylvania State Univ. Main Campus          No 19315  10344   3450        48
University of Michigan at Ann Arbor           No 19152  12940   4893        66
                                         Top25perc F.Undergrad P.Undergrad
Rutgers at New Brunswick                        79       21401        3712
Purdue University at West Lafayette             60       26213        4065
University of California at Berkeley           100       19532        2061
Pennsylvania State Univ. Main Campus            93       28938        2025
```

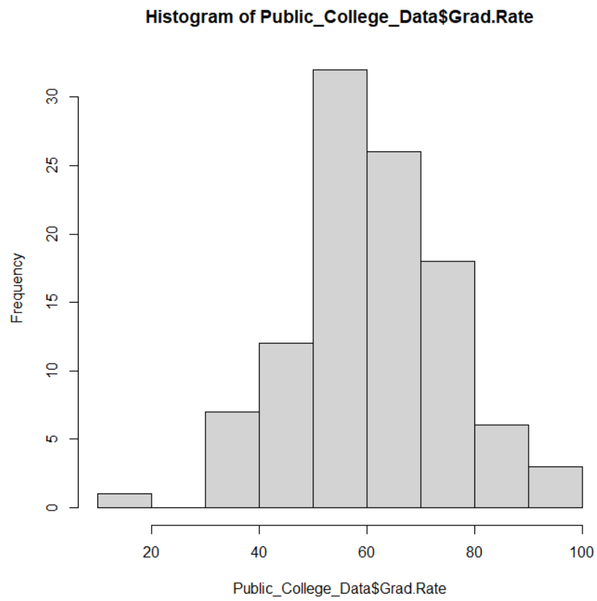
b)Private College Data

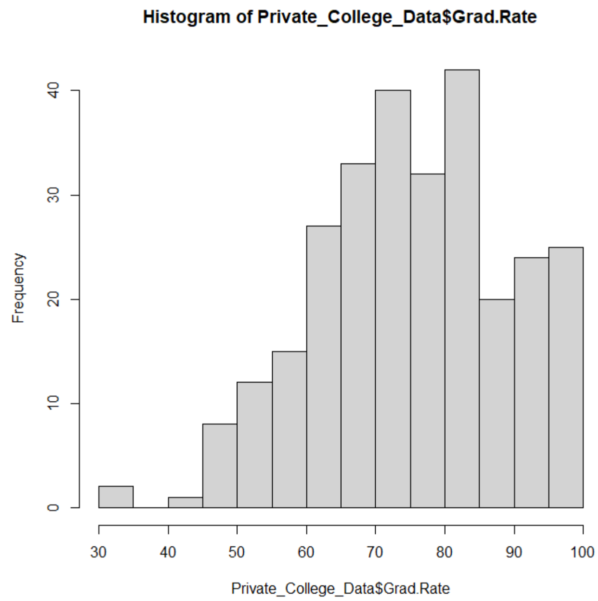
```
> Private_College_Data[1:5,]
                       Private  Apps Accept Enroll Top10perc Top25perc
Boston University          Yes 20192  13007   3810        45        80
University of Delaware     Yes 14446  10516   3252        22        57
Harvard University         Yes 13865   2165   1606        90       100
Duke University            Yes 13789   3893   1583        90        98
New York University        Yes 13594   7244   2505        70        86
                       F.Undergrad P.Undergrad Outstate Room.Board Books
Boston University            14971        3113    18420       6810   475
University of Delaware       14130        4522    10220       4230   530
Harvard University            6862         320    18485       6410   500
Duke University               6188          53    18590       5950   625
New York University          12408        2814    17748       7262   450
                       Personal PhD Terminal S.F.Ratio perc.alumni Expend
Boston University          1025  80       81      11.9          16  16836
University of Delaware     1300  82       87      18.3          15  10650
Harvard University         1920  97       97       9.9          52  37219
Duke University            1162  95       96       5.0          44  27206
New York University        1000  87       98       7.8          16  21227
```

After partitioning and eliminating the universities having less than the median number of HS students admitted from the top 25% of class, we have the reduced the total number of observations for both private and public datasets as shown below –

```
> dim(College)
[1] 777  18
> dim(Public_College_Data)
[1] 105  18
> dim(Private_College_Data)
[1] 281  18
> 
```

Histogram Plot of Graduation Rate –





Based on the above histogram plots of Graduation Rate the cuts for Public and Private College Data are made on the following ways:

Public_College_Data[["GradRateMod"]]=ordered(cut(Public_College_Data[["Grad.Rate"]],c(0,30,50,80,100),labels=c("Low","Medium","High","Low")))

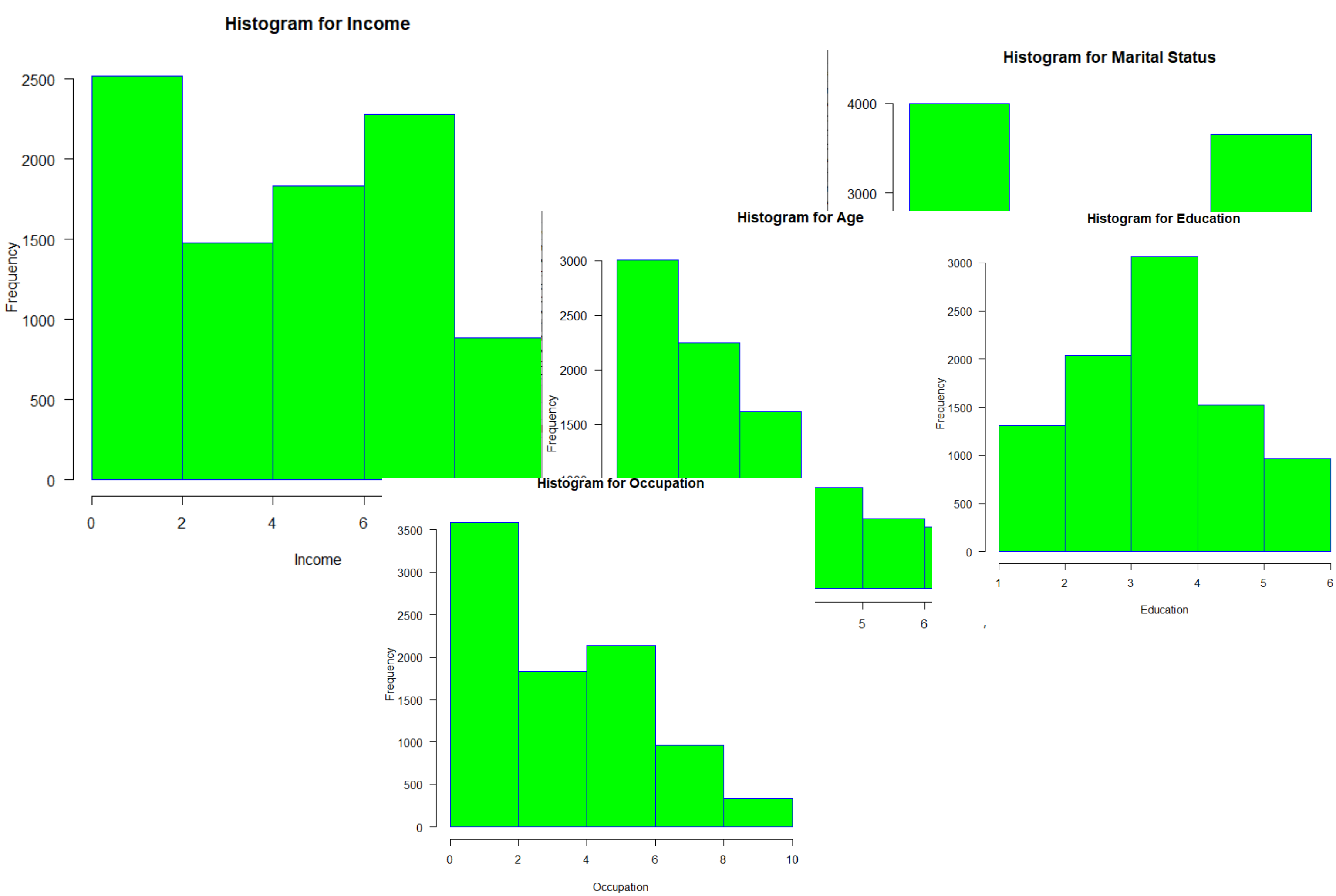
Private_College_Data[["GradRateMod"]]=ordered(cut(Private_College_Data[["Grad.Rate"]],c(0,60,85,100),labels=c("Low","High","Medium")))

Task 2 – Visualizing and Exploring the market dataset

```
> summary(marketing)
     Income            Sex            Marital           Age       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000  
 Median :5.000   Median :2.000   Median :3.000   Median :3.000  
 Mean   :4.895   Mean   :1.547   Mean   :3.031   Mean   :3.415  
 3rd Qu.:7.000   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:4.000  
 Max.   :9.000   Max.   :2.000   Max.   :5.000   Max.   :7.000  
                                 NA's   :160                    
      Edu          Occupation        Lived         Dual_Income   
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:1.000   1st Qu.:4.000   1st Qu.:1.000  
 Median :4.000   Median :4.000   Median :5.000   Median :1.000  
 Mean   :3.835   Mean   :3.788   Mean   :4.198   Mean   :1.545  
 3rd Qu.:5.000   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.:2.000  
 Max.   :6.000   Max.   :9.000   Max.   :5.000   Max.   :3.000  
 NA's   :86      NA's   :136     NA's   :913                    
   Household      Householdu18        Status        Home_Type    
 Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.000  
 Median :3.000   Median :0.0000   Median :2.000   Median :1.000  
 Mean   :2.852   Mean   :0.6669   Mean   :1.837   Mean   :1.856  
 3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:3.000  
 Max.   :9.000   Max.   :9.0000   Max.   :3.000   Max.   :5.000  
 NA's   :375                      NA's   :240     NA's   :357    
     Ethnic          Language    
 Min.   :1.000   Min.   :1.000  
```

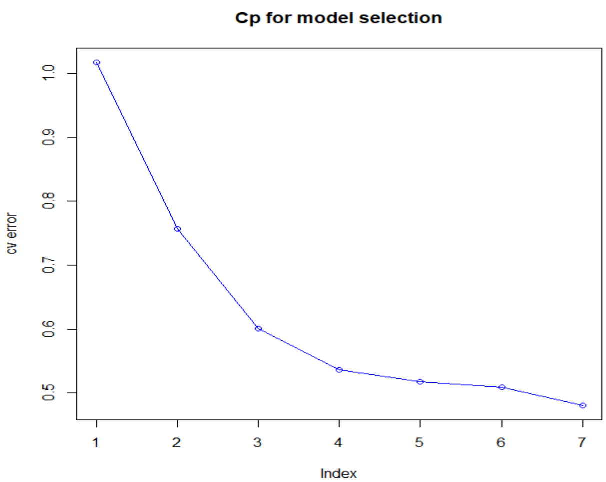
It's a transaction data and highlighted regions shows quite a lot of missing values for each of column values, it's quite natural be it in online marketing or in a physical shop, it's not always possible to capture all the details of a customer due to several reasons.

<u>Histogram plots of all the variables –</u>



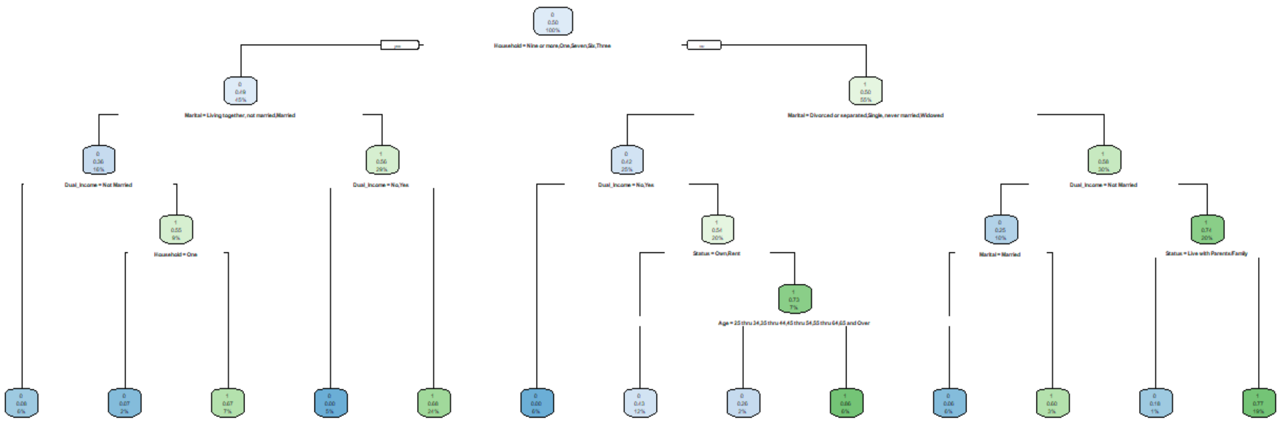
Recoded the marketing data as per the ESL textbook.

"[https://hastie.su.domains/ElemStatLearn/datasets/marketing.info.txt (blackboardcdn.com)](https://hastie.su.domains/ElemStatLearn/datasets/marketing.info.txt)"

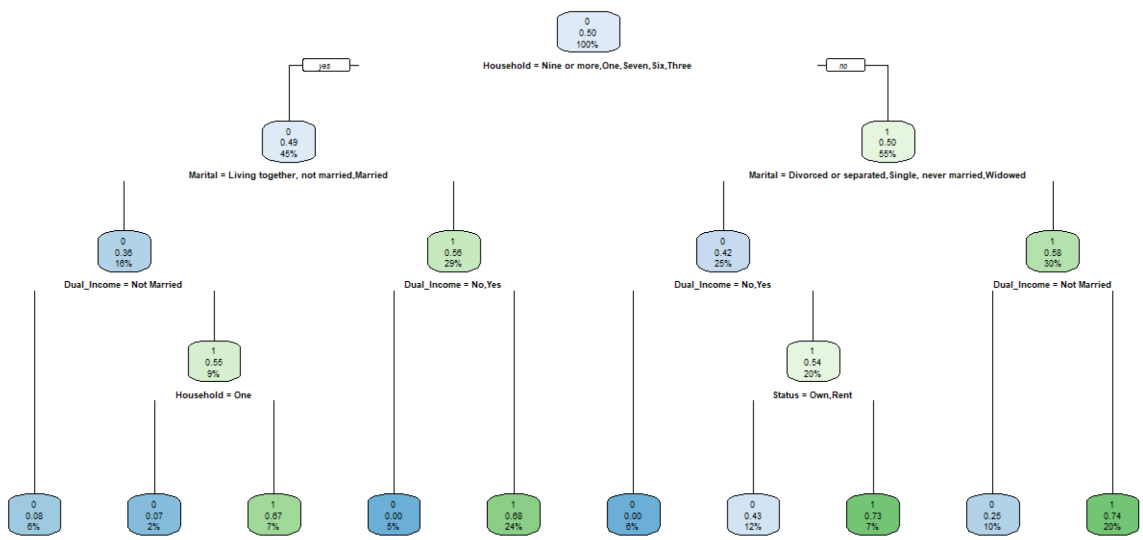


The cp table there is a significant decrease in the modelling error with just four variables ,so we should stick to these four models rather then making it more and complex.

Classification Tree for unsupervised learning - Marketing Data Before Pruning



Classification Tree for unsupervised learning - Marketing Data After Pruning



We have pruned the original tree to a smaller tree using lesser number of variables were crucial in terms of importance and generalized the model at the same time thus keeping the model simple.

<u>Dual Income has got three categories Yes, No and Not Married ,so for people who are under Dual Income has got three categories =[Yes, No] and might be married or living together represent the</u>

```
x
   Y                                                                                                                                                                                cover
0.00 when Household is Nine or more or One or Seven or Six or Three & Marital is Divorced or separated or Single, never married or Widowed & Dual_Income is No or Yes               5%
0.00 when Household is Five or Four or Two & Marital is Divorced or separated or Single, never married or Widowed & Dual_Income is No or Yes                                       6%
0.07 when Household is One & Marital is Living together, not married or Married & Dual_Income is    No or Yes                                                                      2%
0.08 when Household is Nine or more or One or Seven or Six or Three & Marital is Living together, not married or Married & Dual_Income is Not Married                              6%
0.25 when Household is Five or Four or Two & Marital is Living together, not married or Married & Dual_Income is Not Married                                                      10%
0.43 when Household is Five or Four or Two & Marital is Divorced or separated or Single, never married or Widowed & Dual_Income is Not Married & Status is Own or Rent            12%
0.67 when Household is Nine or more or Seven or Six or Three & Marital is Living together, not married or Married & Dual_Income is No or Yes                                       7%
0.68 when Household is Nine or more or One or Seven or Six or Three & Marital is Divorced or separated or Single, never married or Widowed & Dual_Income is Not Married           24%
0.73 when Household is Five or Four or Two & Marital is Divorced or separated or Single, never married or Widowed & Dual_Income is Not Married & Status is Live with Parents/Family  7%
0.74 when Household is Five or Four or Two & Marital is Living together, not married or Married & Dual_Income is    No or Yes                                                     20%
```

maximum probability of belonging to the actual marketing dataset. Again, this conclusion is based on the pruned tree version of the original dataset.

Support Calculated for Y=1 that is for the original dataset:

```
)"
> probability_mat
                                                 nodeprob
[1,] 2  437  905 0.3256334 0.6743666 0.07461359
[2,] 2 1386 2996 0.3162939 0.6837061 0.24363394
[3,] 2  148  880 0.1439689 0.8560311 0.05715557
[4,] 2  243  372 0.3951220 0.6048780 0.03419326
[5,] 2  754 2584 0.2258838 0.7741162 0.18558879
> x11()
> rpart.plot(fit.overall_mar,main="Classification Tree for unsupervised learning - M
arketing Data Before Pruning")
> probability_mat[,5]
[1] 0.6743666 0.6837061 0.8560311 0.6048780 0.7741162
> probability<-as.vector(probability_mat[,5])
> support<-(probability*n)/size
> support
[1] 0.10063383 0.33314800 0.09785389 0.04136551 0.28733459
```

After calculating support for terminal nodes where Y=1 we would be tracing back to the rpart.fit summary to calculate the lift and confidence for the above observations:Below are the plots of parents of the root nodes for which Y=1,these plots would help us to get the information of lift and confidence

```
> path.rpart(fit.overall_mar, 19)

 node number: 19
   root
   Household=Nine or more,One,Seven,Six,Three
   Marital=Living together, not married,Married
   Dual_Income=No,Yes
   Household=Nine or more,Seven,Six,Three
>
```

```
> path.rpart(fit.overall_mar, 11)

 node number: 11
   root
   Household=Nine or more,One,Seven,Six,Three
   Marital=Divorced or separated,Single, never married,Widowed
   Dual_Income=Not Married
>
```

```
> path.rpart(fit.overall_mar, 55)

 node number: 55
   root
   Household=Five,Four,Two
   Marital=Divorced or separated,Single, never married,Widowed
   Dual_Income=Not Married
   Status=Live with Parents/Family
   Age=14 thru 17,18 thru 24
```

```
> path.rpart(fit.overall_mar, 29)

 node number: 29
   root
   Household=Five,Four,Two
   Marital=Living together, not married,Married
   Dual_Income=Not Married
   Marital=Living together, not married
>
```

```
> path.rpart(fit.overall_mar, 31)

 node number: 31
   root
   Household=Five,Four,Two
   Marital=Living together, not married,Married
   Dual_Income=No,Yes
   Status=Own,Rent
```

```
      var    n   wt  dev yval  complexity ncompete nsurrogate     yval2.V1
19 <leaf> 1342 1342  437    2 0.008284221        0          0 2.000000e+00
11 <leaf> 4382 4382 1386    2 0.003335928        0          0 2.000000e+00
55 <leaf> 1028 1028  148    2 0.000000000        0          0 2.000000e+00
29 <leaf>  615  615  243    2 0.005559880        0          0 2.000000e+00
31 <leaf> 3338 3338  754    2 0.003224730        0          0 2.000000e+00
       yval2.V2     yval2.V3     yval2.V4     yval2.V5 yval2.nodeprob
19 4.370000e+02 9.050000e+02 3.256334e-01 6.743666e-01   7.461359e-02
11 1.386000e+03 2.996000e+03 3.162939e-01 6.837061e-01   2.436339e-01
55 1.480000e+02 8.800000e+02 1.439689e-01 8.560311e-01   5.715557e-02
29 2.430000e+02 3.720000e+02 3.951220e-01 6.048780e-01   3.419326e-02
31 7.540000e+02 2.584000e+03 2.258838e-01 7.741162e-01   1.855888e-01
   support_percent
19       10.063383
11       33.314800
55        9.785389
29        4.136551
31       28.733459
```

It appears that the support for terminal node 11 and 31 are very high .So for terminal nodes 11 and 31 if we get the following association rule:

Association Rule 1 –Support 33.31%

No. of persons in house -One, Three, Six, Seven, Nine or more

Marital Status -Single, Never Married, Widowed, "Divorced or Separated"

Dual Income –"Not married"

Association Rule 2 –Support 28.74%

No. of persons in house -Two, Four, Five

Marital Status -Living Together, Married, Not Married

Dual Income –No, Yes

Status = Own , Rent

]