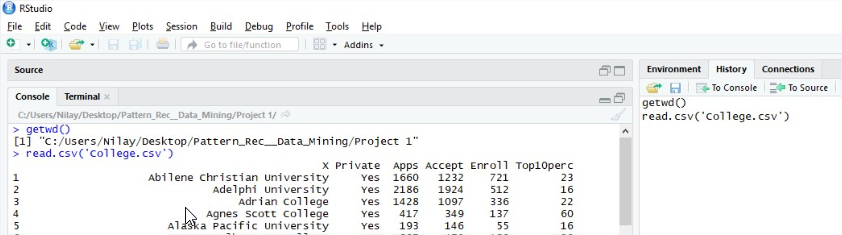
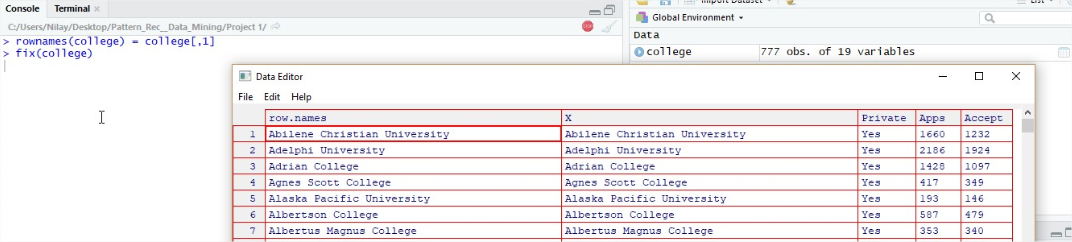
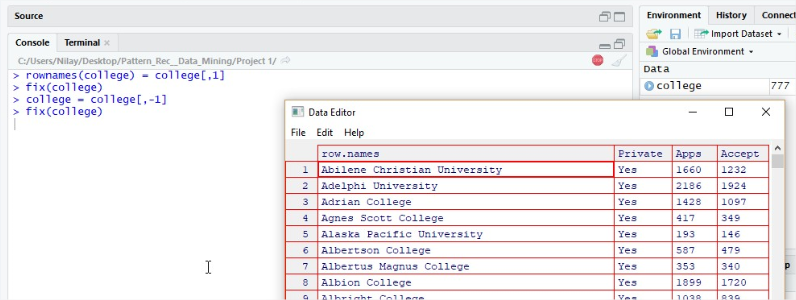
**College Data Set**

1. Read.csv()

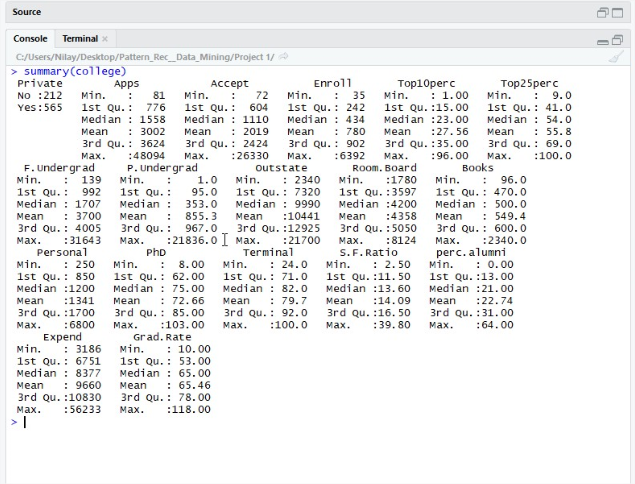


1. Fix()



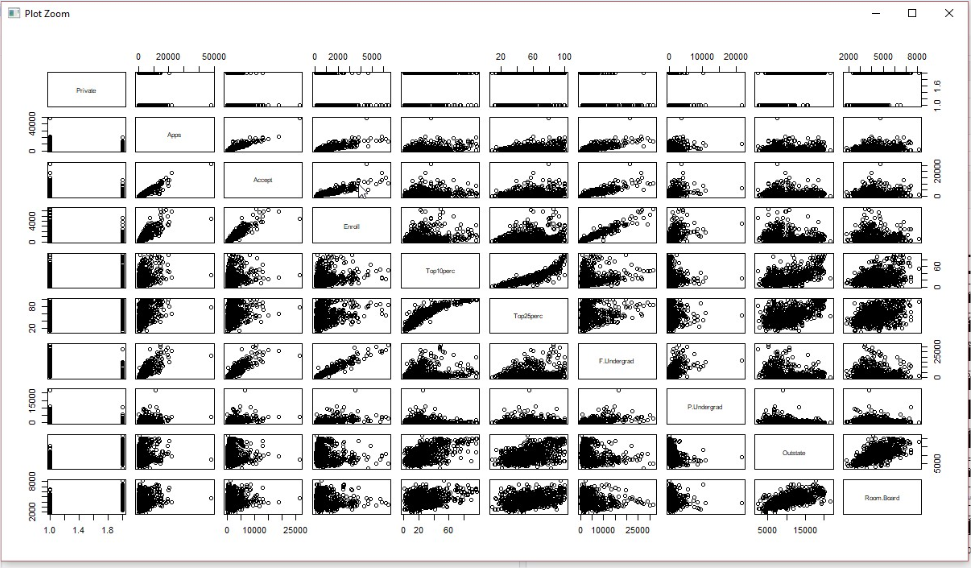


1. Using the different functions:
2. Summary()

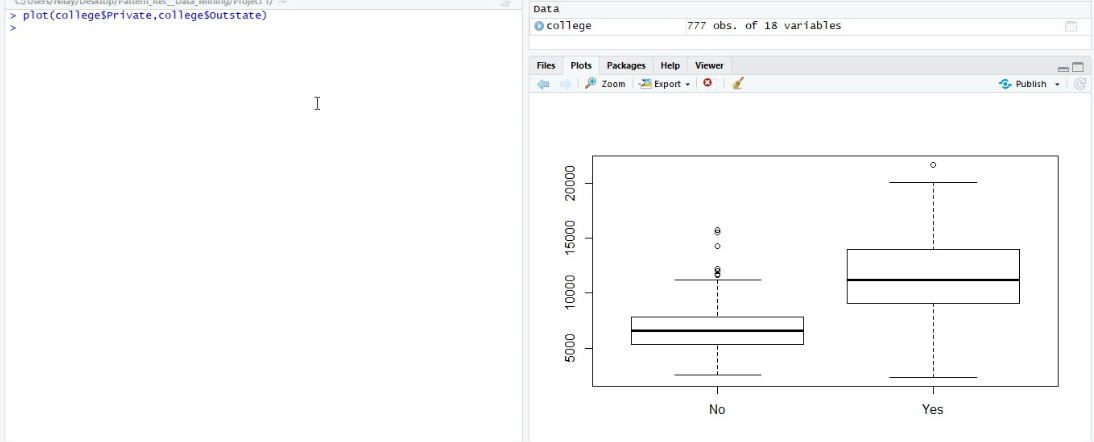


1. Pairs()

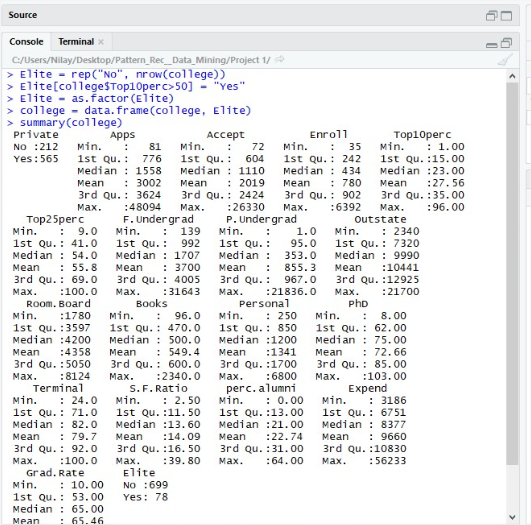


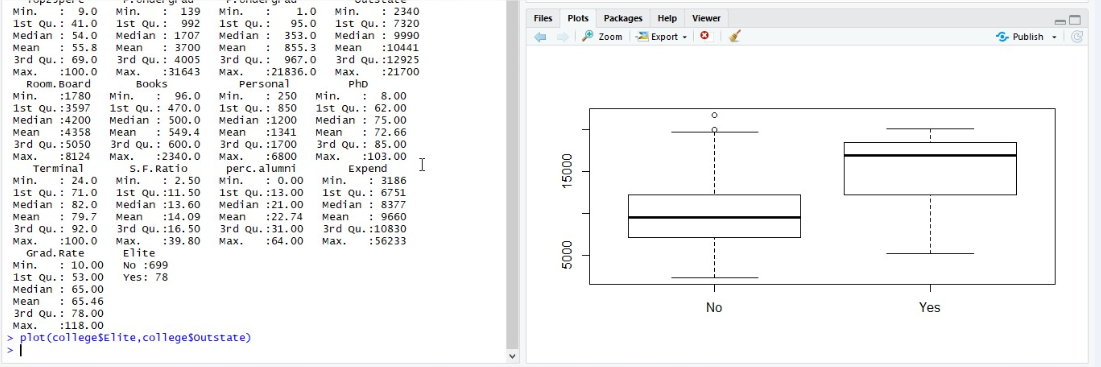


1. Plot()

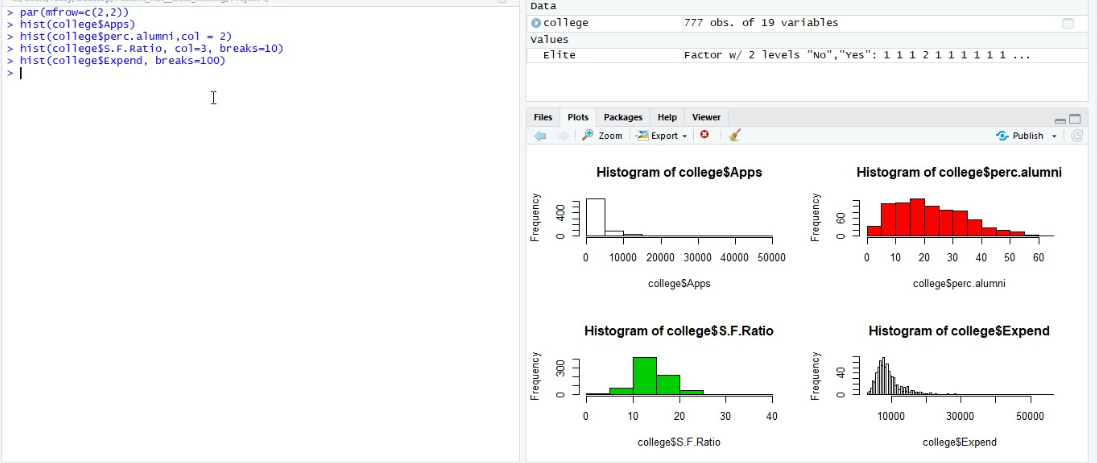


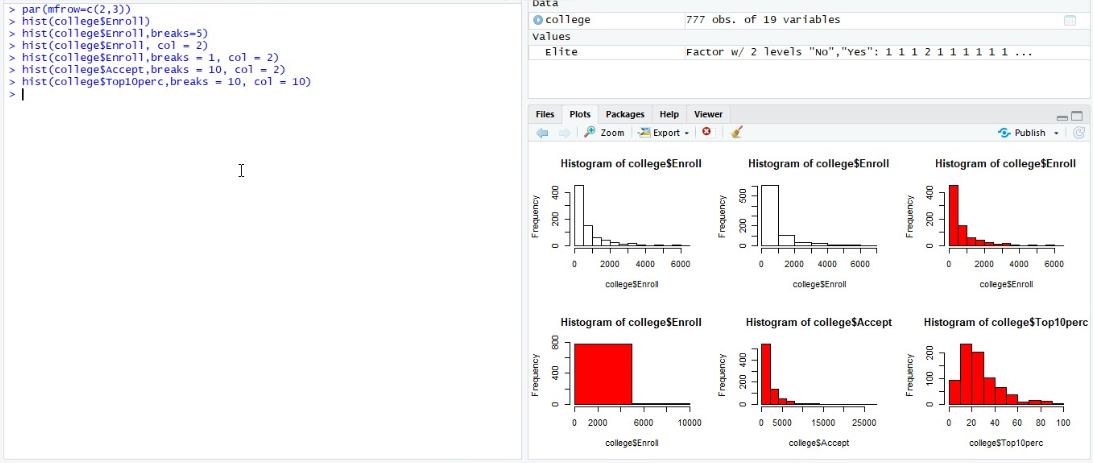
1. Elite variable





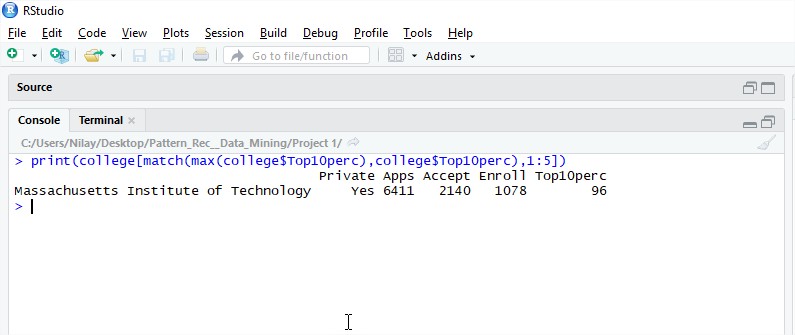
1. Hist()



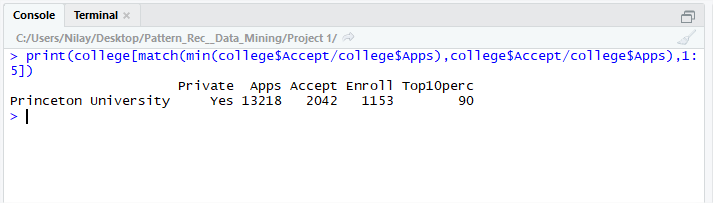


1. Summary for the criteria:

* University with the most students in the top 10% of class.

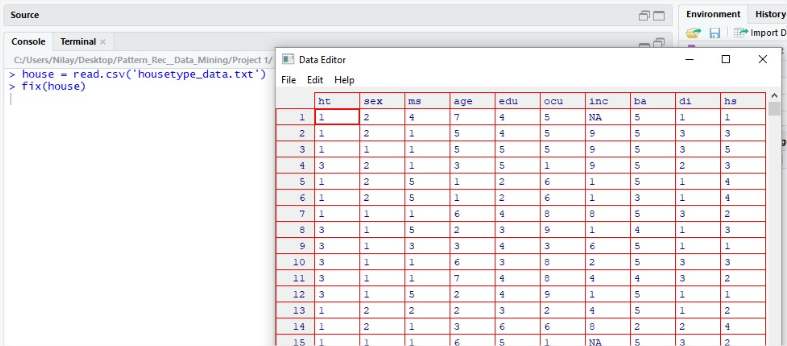


* University having the smallest acceptance rate.

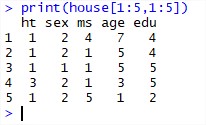


**Working on Dataset “Housetype.data”**

1. Reading data into a data frame. Keeping row and column data around. Showing matrix dimension and 5\*5 subarray.



C:\Users\Nilay\Desktop\Pattern_Rec__Data_Mining\Project 1\Screenshots\2_2.jpg



1. Function attributeHist that produces histogram using “Age” as input parameter.

---------------------------------------------------------

function(inpt\_atr){

house\_atr = house[,grep(inpt\_atr, colnames(house))]

if(inpt\_atr %in% colnames(house)){

if(NA %in% house\_atr){

cat(sum(!complete.cases(house\_atr)),"Values with NA")

}

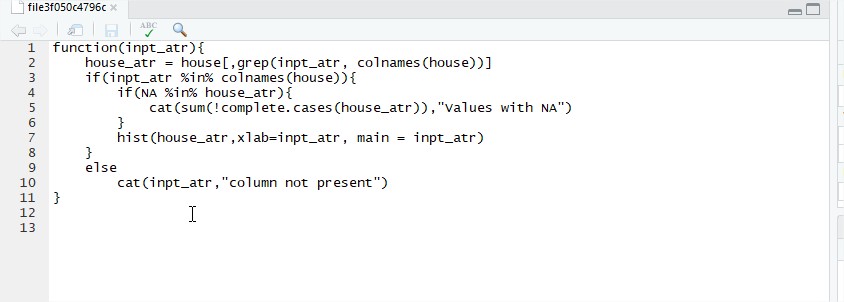
hist(house\_atr,xlab=inpt\_atr, main = inpt\_atr)

}

else

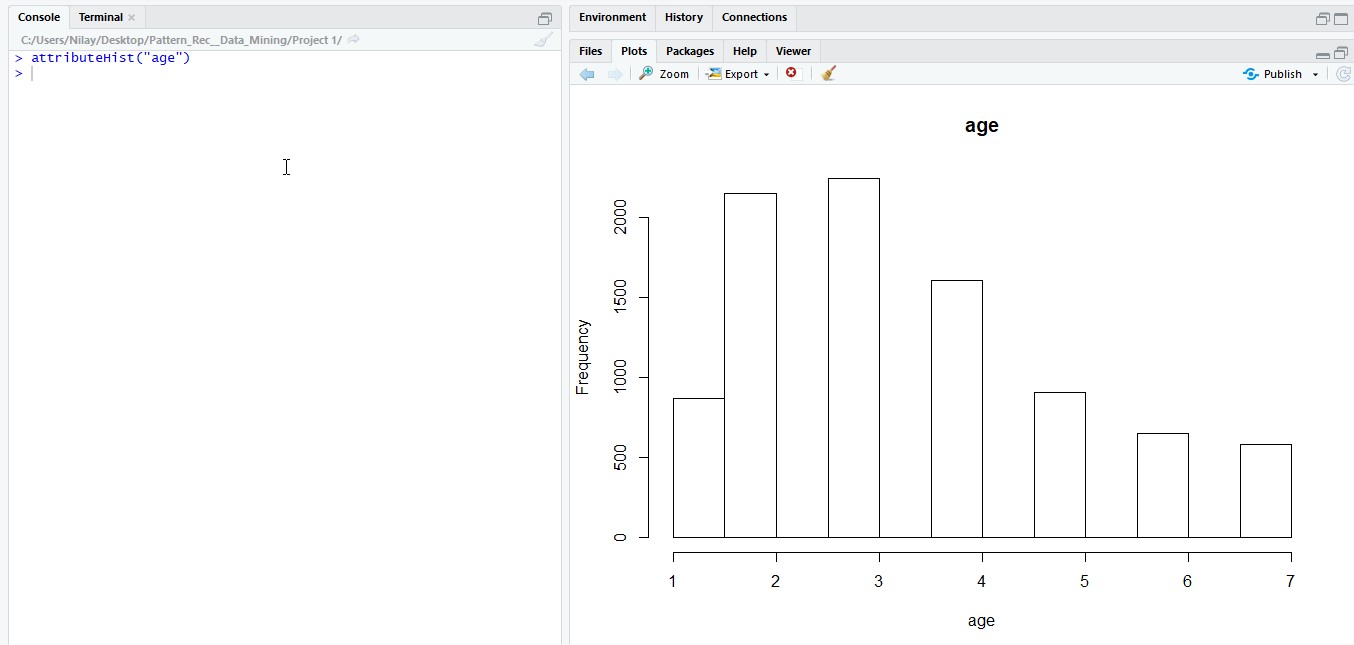
cat(inpt\_atr,"column not present")

}

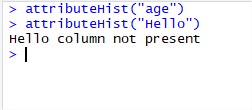


----------------------------------------------------------

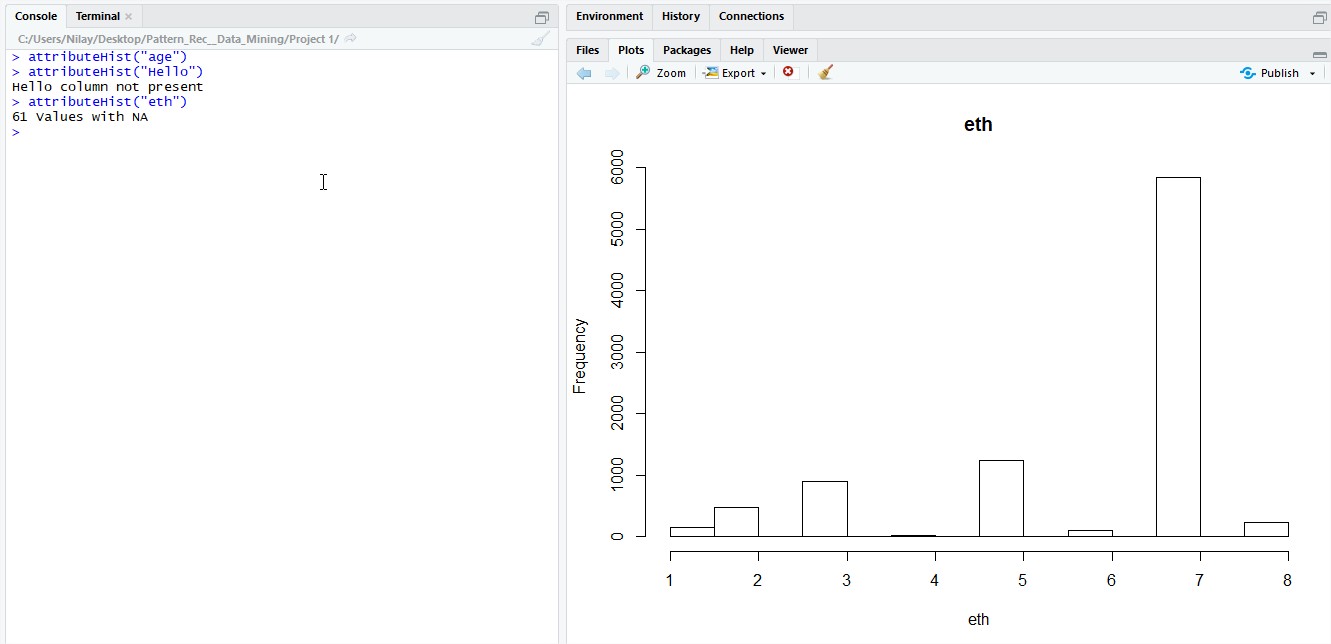
1. attributeHist(“age”)



1. attributeHist(“Hello”)

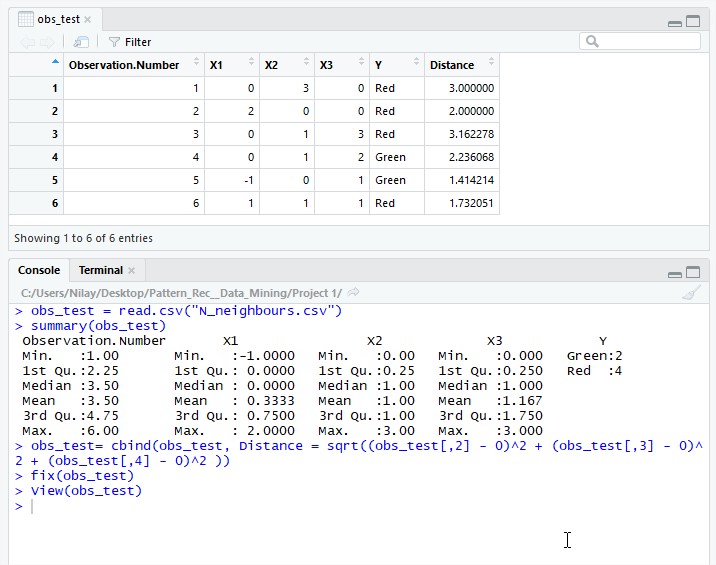


1. attributeHist(“eth”)



**K-Nearest Neighbor Classifier.** Data given in table for 6 observations.

1. Compute the Euclidean distance between each observation and the test point, *X*1 = *X*2 = *X*3 = 0.



1. What is our prediction with *K* = 1? Why?

* For K=1 the closest observation is #5 (1.41). So the prediction would be **Green**

1. What is our prediction with *K* = 3? Why?

* For K = 3 the closest observations comes to #2, 5 & 6, corresponding to Red, Green & Red. So our prediction would go to the most repeating value i.e **Red**

1. If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

* Since the Bayes decision boundary in this problem is highly nonlinear we should try to **small** value of K because a large value of K will fit more linear boundary resulting into more number of points to be taken into consideration and we need to keep K such that we get minimum number of points.