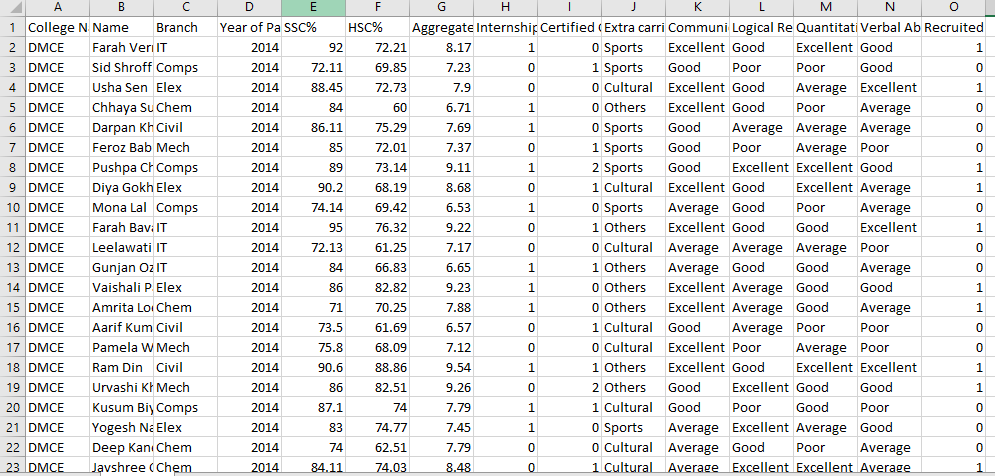
Glimpse of our dataset

The data is about the students who were recruited between the circa (2014-2018) from Datta Meghe College Of Engineering. One can access the file by converting it into .csv format.

The prediction system implemented here aims to provide placement prediction on basis of various attributes.

For viewing the columns as well as the type of data present in the dataset,Use the command-

**->****str(name of the dataset)** [i.e old.data]

>old.data<-na.omit(old.data)

>head(old.data)

>tail(old.data)

>summary(old.data$branch)

>length(old.data)

>names(old.data)

Nrow(old.data)

We have visualised the data using three ways and that is through- histogram,scatterplot,multiplot.

PRE-PROCESSING

Removing null rows and rows with incomplete data

old.data<-na.omit(old.data)

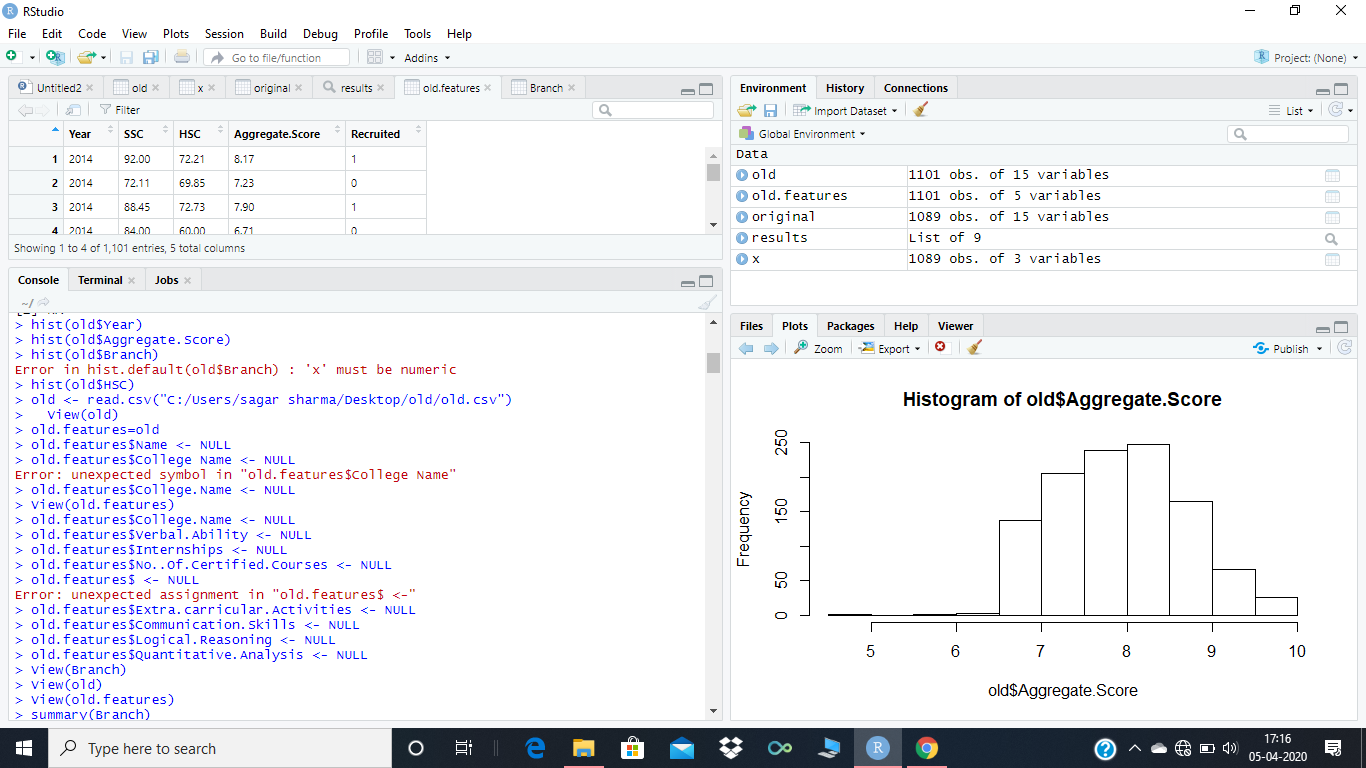
our original data was having 1101 rows after omitting null rows we are left with 1089 rows.

1)Histogram

Histogram will visualise numerical data by showing the number of data points that fall within a specified range of values.Here we have used the parameter as **Aggregate.score** so as to analyse the students falling into different ranges.

Command-

**->Hist(old.data$Aggregate.score)**

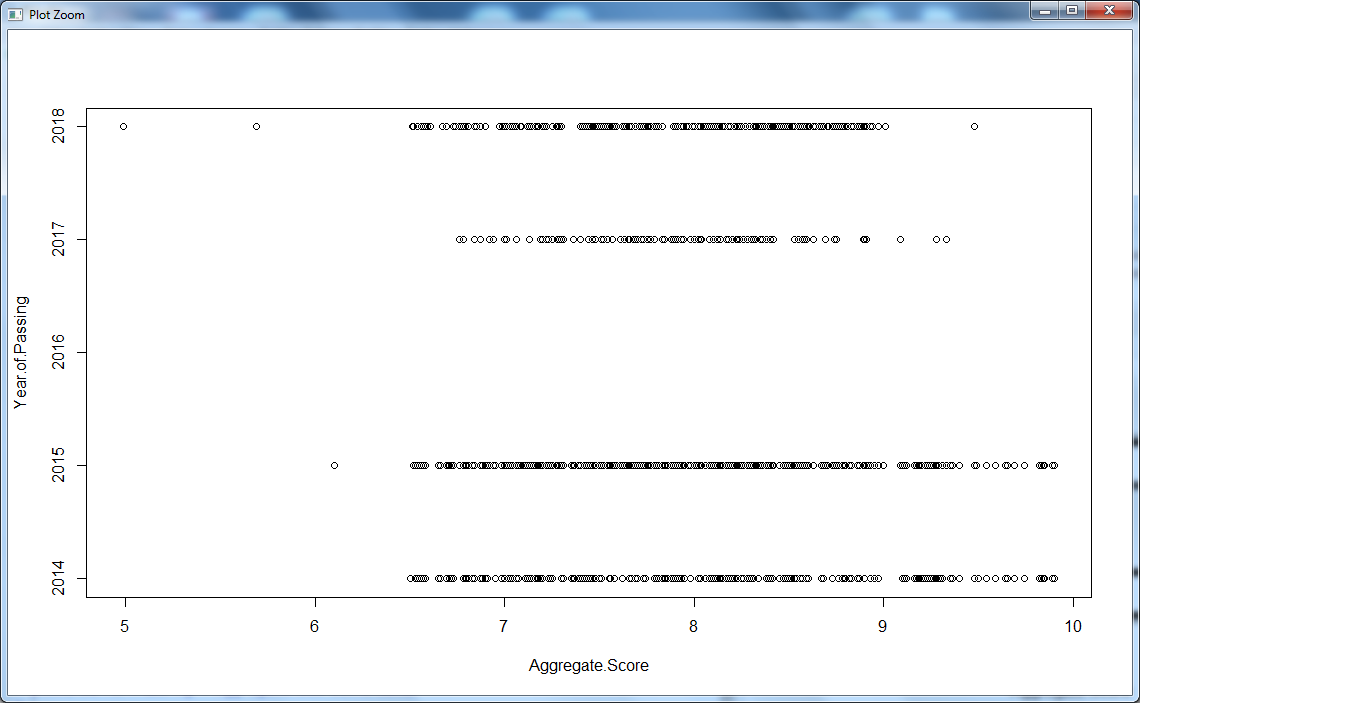


2)Scatterplot

A scatterplot will display the relationship between two numerical variables. Each member of the dataset gets plotted as a point whose ( x , y ) (x, y) (x,y)left parenthesis, x, comma, y, right parenthesis coordinates relates to its values for the two variables.The x and y values we have taken are-Aggregate.score and Year.of.Passing respectively.

Command

**->plot(old.data$Aggregate.score, old.data$Year.of.Passing)**

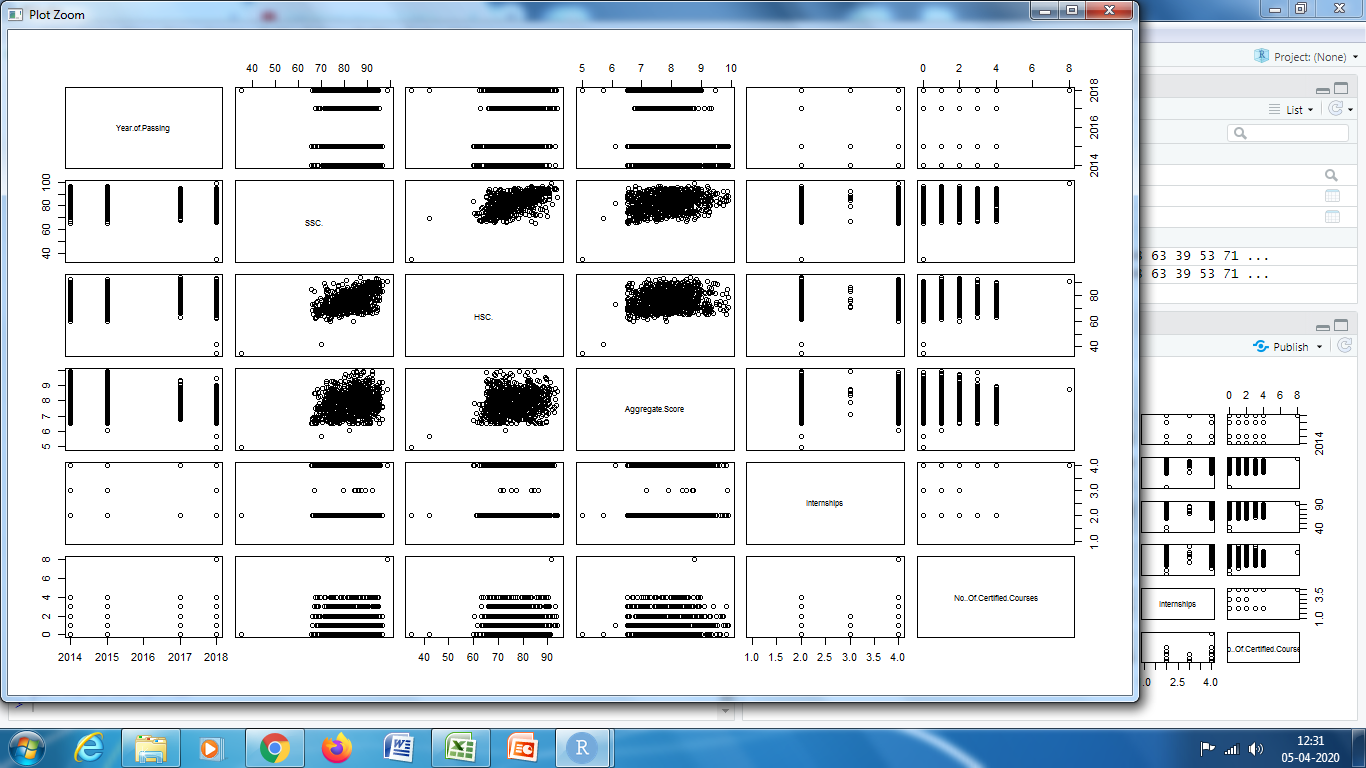


3)Multiplot

Multiplot is an easy to use function to put multiple graphs on the same page using R statistical software and ggplot2 plotting methods. This function is from Ggplot2 package.The ggplot package can be installed using the command-

**->install.packages(“ggplot2”)**

After installing the package we can use the following command to see the multiplot.

**->pairs(old.data[4:9])**[4:9] is used to specify the range of the columns we want to analyse. 

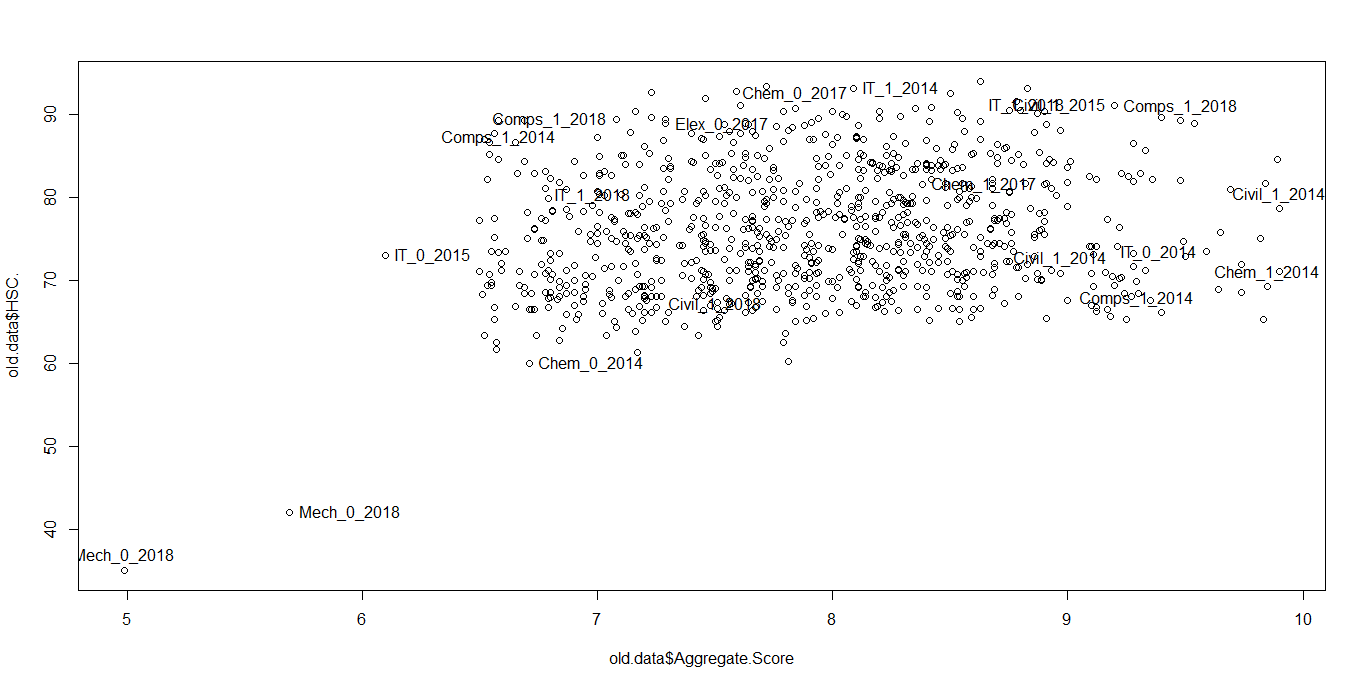
4.Plotting points based on points and showing from which department and year students are recruited

> plot(old.data$Aggregate.Score,old.data$HSC.)

> idx<-identify(old.data$Aggregate.Score,old.data$HSC.)

> plot(old.data$Aggregate.Score,old.data$HSC.)

> idx<-identify(old.data$Aggregate.Score,old.data$HSC.,labels = paste0(as.character(old.data$Branch),"\_",as.character(old.data$Recruited),"\_",as.character(old.data$Year.of.Passing)),plot = TRUE)



Here the points show branch\_recruited/not recruited\_year of passing

Eg.mech\_0\_2018

4)K-means clustering

Creating an alias of our dataset as old.features

Removing rows with string as values

> old.features$College.Name<-NULL

> old.features$Name<-NULL

> old.features$Branch<-NULL

> old.features$Extra.carricular.Activities<-NULL

> old.features$Communication.Skills<-NULL

> old.features$Logical.Reasoning<-NULL

> old.features$Verbal.Ability<-NULL

> old.features$Quantitative.Analysis<-NULL

> old.features$Certified.Courses<-NULL

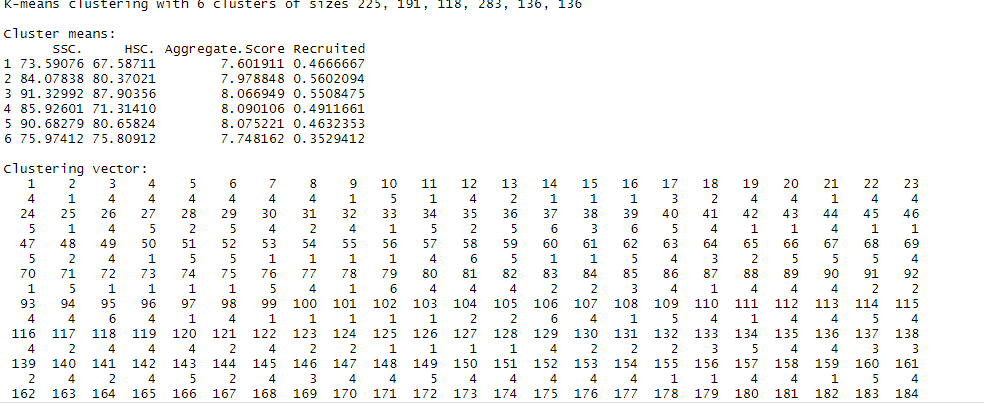
> old.features$Internships<-NULL

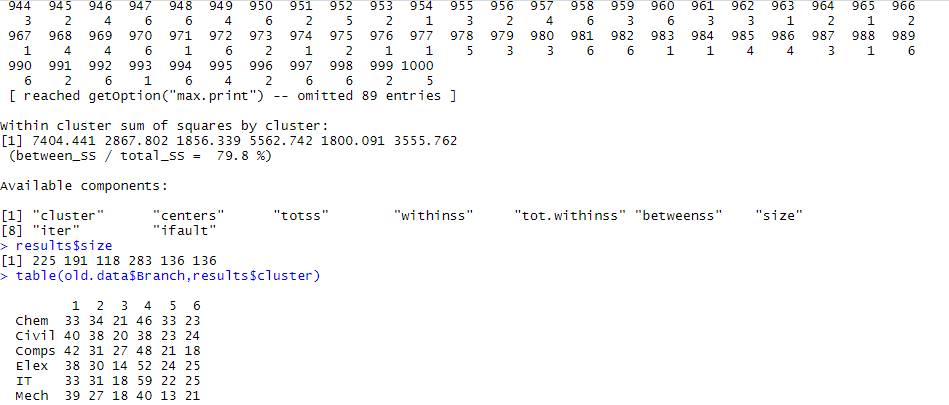
> results<-kmeans(old.features,6)

> view (results)

> results

Cluster matrix and table is given below



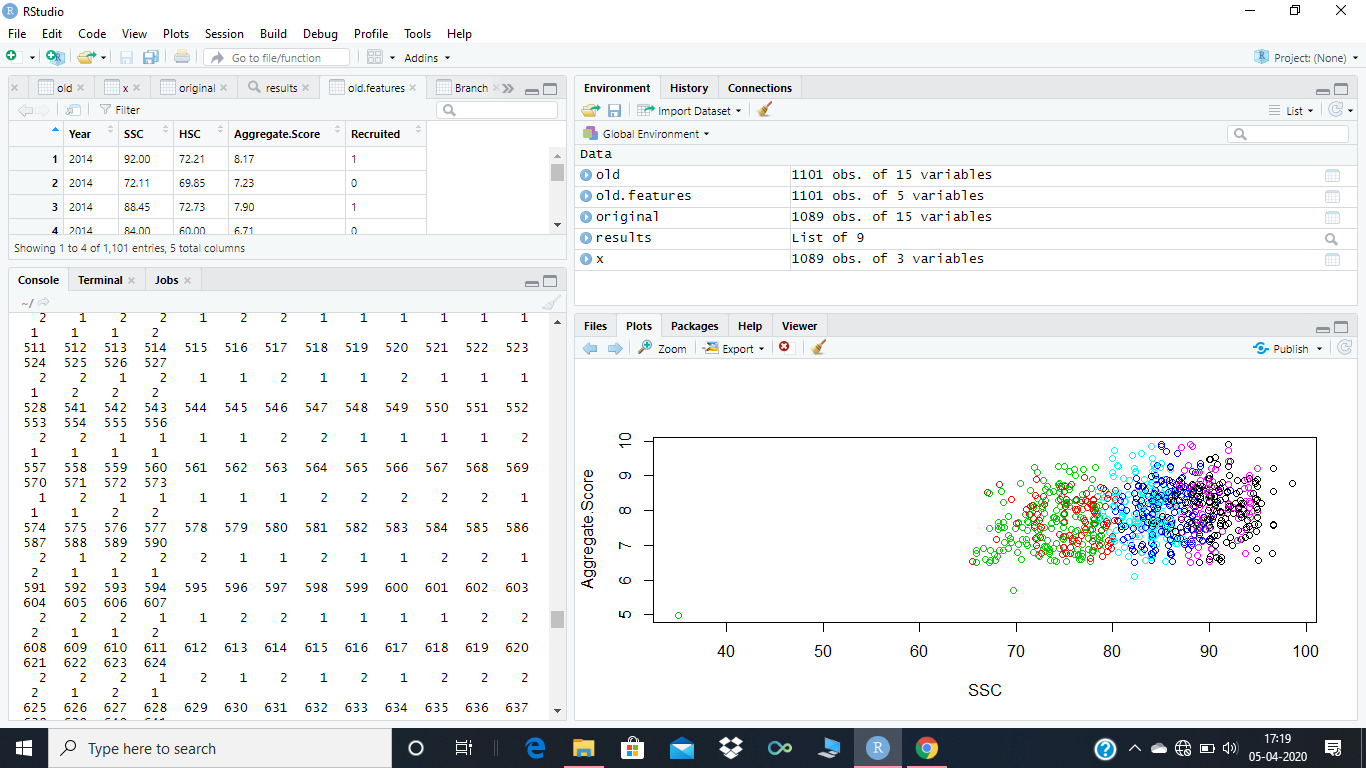


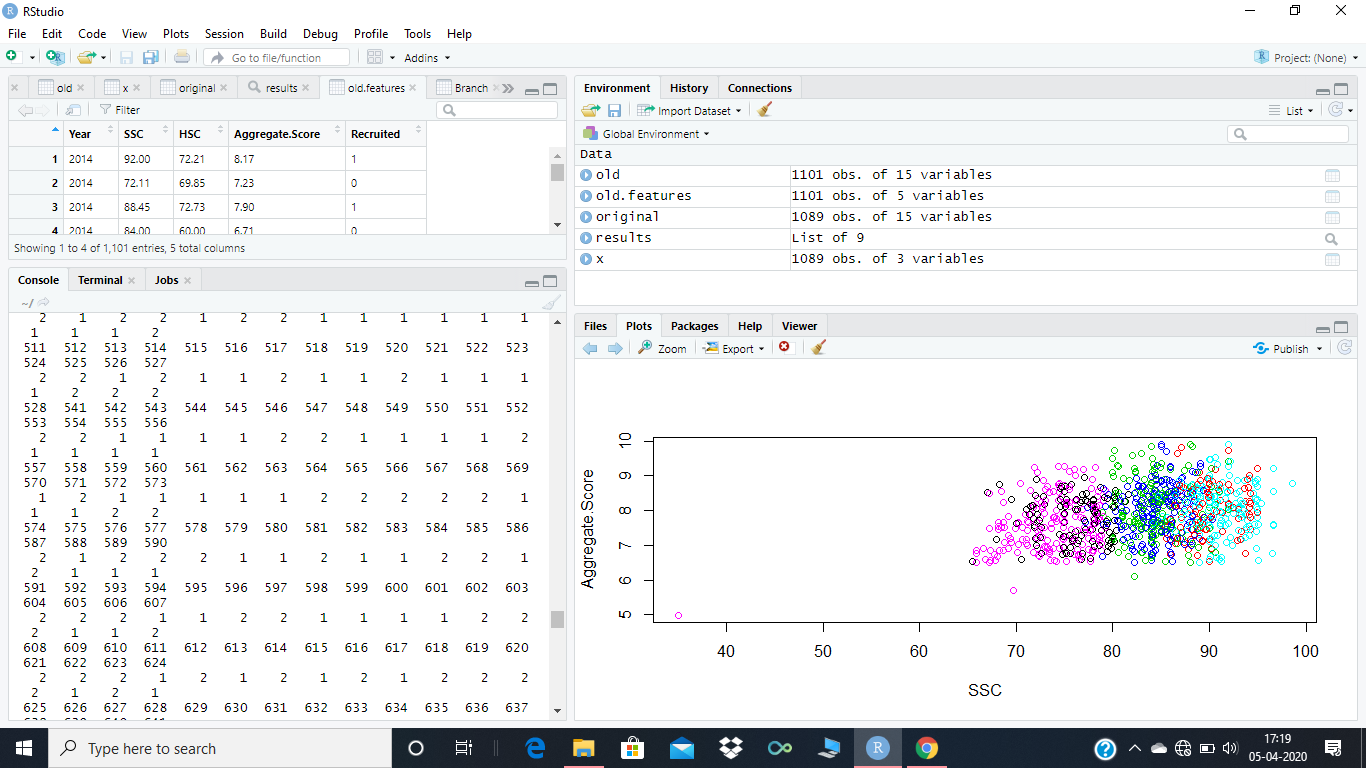
> results$size

> table(old.data$Branch,results$cluster)

> plot(old.data[c("HSC.","Aggregate.score")],col=results$cluster)

> plot(old.data[c("HSC","Aggregate.score")],col=old.data$Branch)





These clusters were of size 6 because we have six branches in our dataset as clusters here are not well separated i.e.,are overlapping

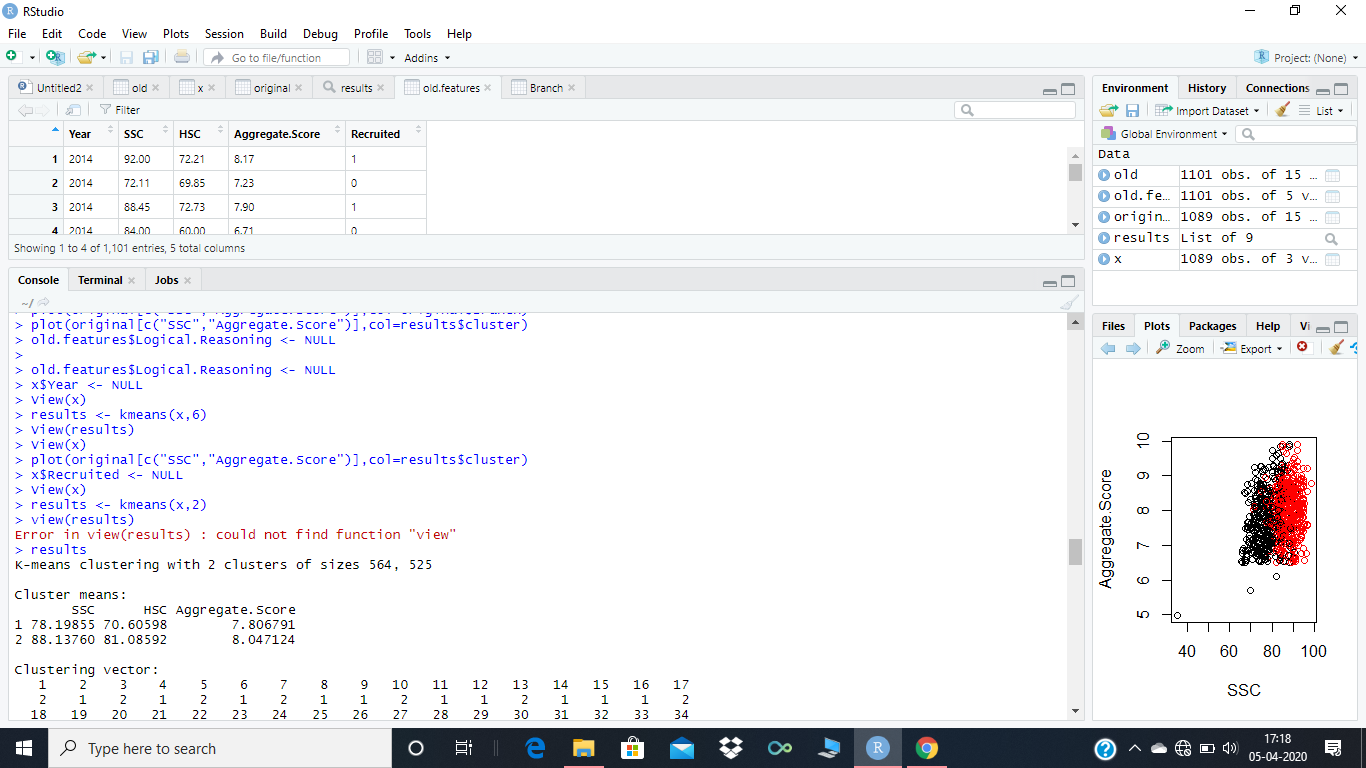
We are now taking cluster size as 2 based on recruited and non recruited.

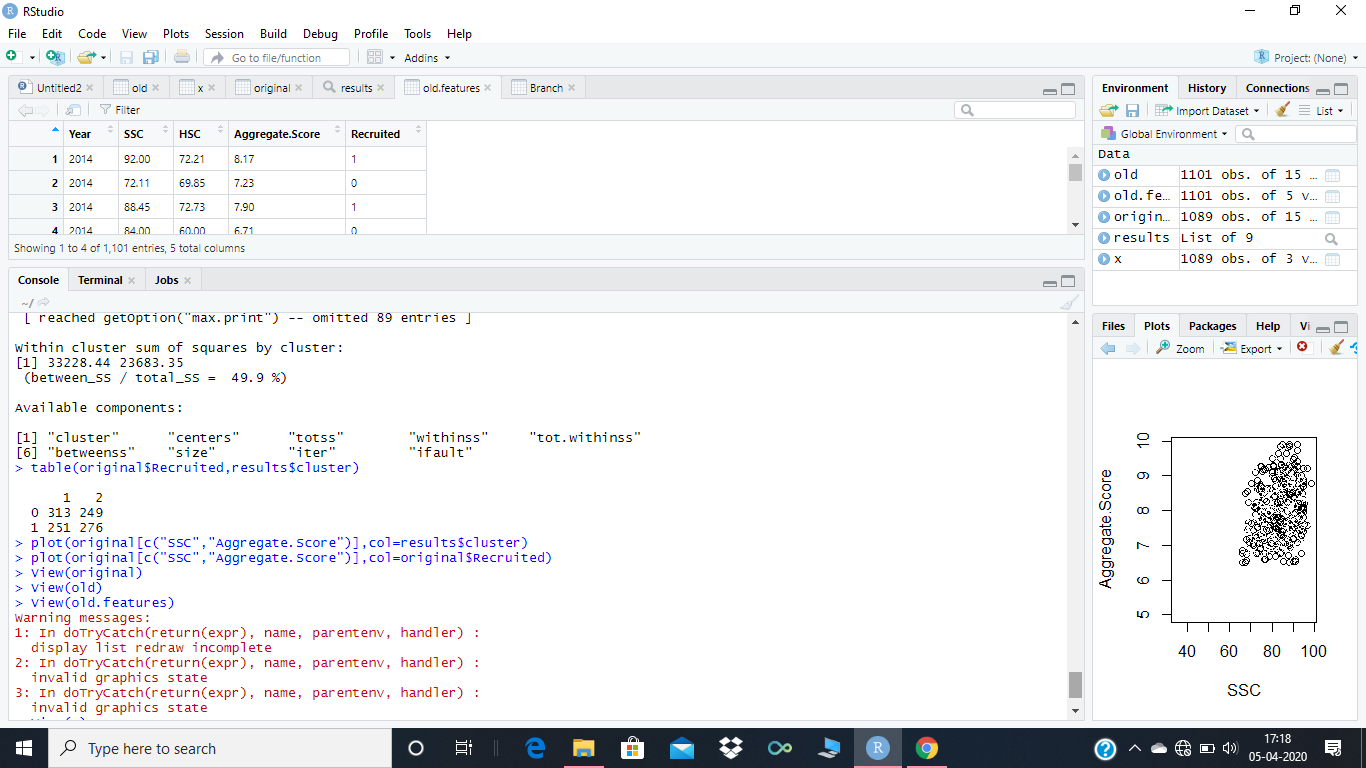
> results<-kmeans(old.features,2)

> results

> plot(old.data[c("HSC.","Aggregate.score")],col=results$cluster)

> plot(old.data[c("HSC","Aggregate.score")],col=old.data$Recruited)





So as different clusters are well separatedin two colors here so we can see patterns for our given dataset and check in which category a particular student lie.There is also some overlapping which can be ignored.