**Question 3 HW2**

In this Question we are dealing With subset selection models namely

* Forward Subset Selection
* Backward Subset selection
* Exhaustive subset selection
* Linear model

In this exercise, we will predict the number of applications received using the other variables in the **cereals.csv**  data set in the ISLR package.

**Step 1 Loading the dataset**

In this step I Got to know to know about the necessary packages that are required to Excute the models in R and I imported them .And I also imported the Collage Data set That are Inbuilt in the R Studio

Graphical user interface, text, application

Description automatically generated

**Step 2 Splitting the dataset**

the seed() function is set to a random variable to get the same values when executed again. The data is spitted into 80:20 ratio so that 80% of the data is used for training and remaining 20% is used for testing

Graphical user interface, text, application

Description automatically generated

**Step 3 1 A Fitting the linear model**

After fitting the linear model to the dataset it appears that the test mean squared as around **9.890283e-14 i.e. nearly equals to zero.**

And the

**variables Protien, calories, fat sodium,fiber**, **carbo, potassium, sugars, vitamins** to be the 10 of the most significant values with lesser P value

Graphical user interface, text

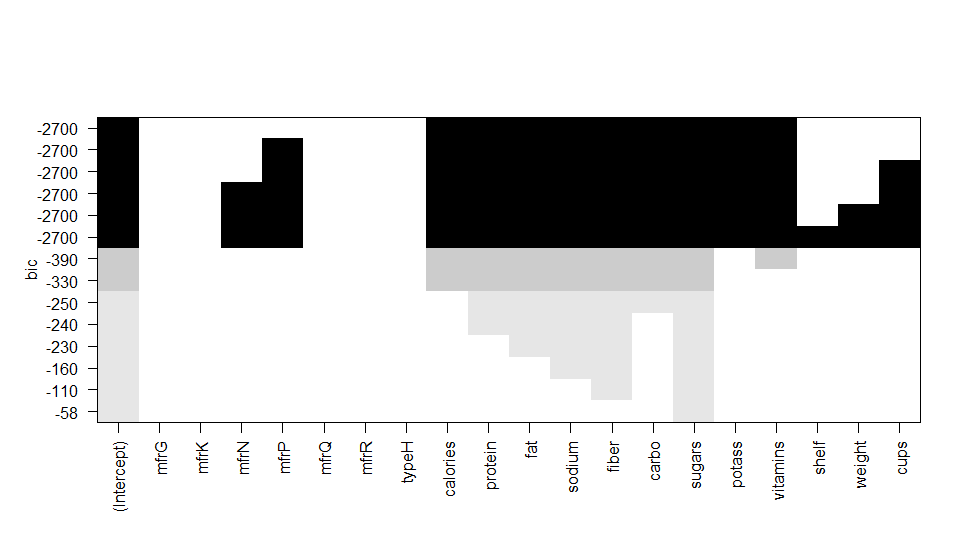
Description automatically generated

**Step 4 1 B Fitting the Forward subset selection model**

A picture containing chart

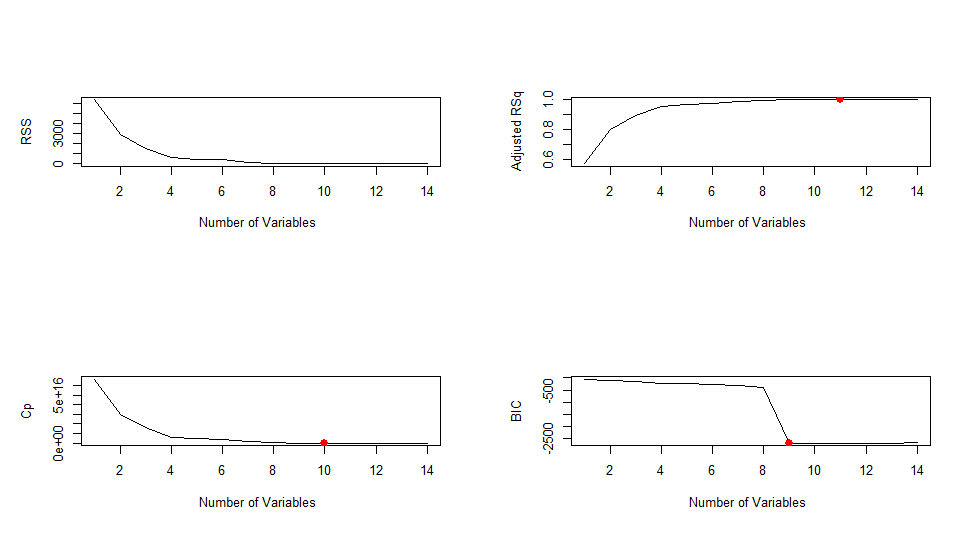
Description automatically generatedI installed the leaps package to perform some necessary functions on the dataset. By using the regsubsets() function in the leaps library I applied forward subset selection model to the dataset and set the number of variables to variable NVMAX= 14 and method as forward.

* When I look at the summary of of the model it appeard that the model is having least MSE at 10 variables.and I also plotted the results of it on the graph as shown below.

Text

Description automatically generated

When we look at the **bic** score of the variables in the graph it appears that **Protien, calories, fat sodium,fiber**, **carbo, potassium, sugars, vitamins** are presentin the model through out the dataset and contribute to the model.

* when we look at the errors of different types here it appears that the RSS apporaches to zero when there are **9 variables** in the model.
* and Adjusted R squared error becomes closer to 1 when there are **9 variables** in the dataset.
* Cp error becomes closer to 0 when there are **9 variables** in the dataset.
* BIC score converges to low when there are **9 variables** in the dataset.

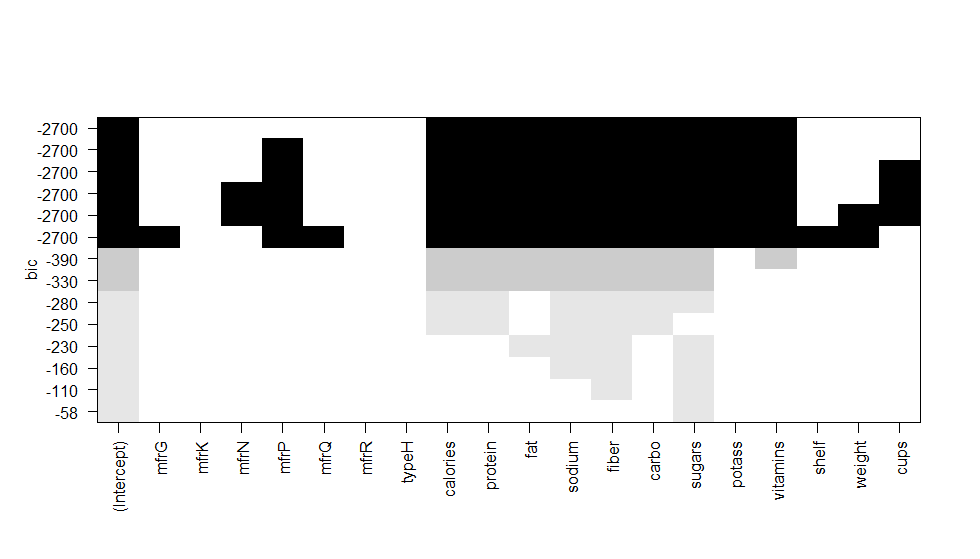
**Step 5 1 C Fitting the Exhaustive subset selection model**

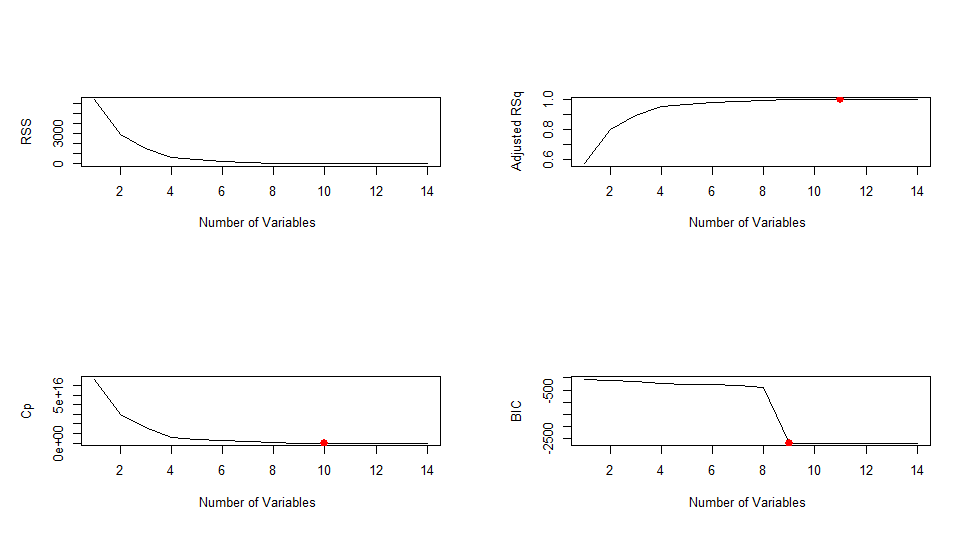
I installed the leaps package to perform some necessary functions on the dataset. By using the regsubsets() function in the leaps library I applied forward subset selection model to the dataset and set the number of variables to variable NVMAX= 14 and method as forward

Graphical user interface, text, application

Description automatically generated

* When I look at the summary of of the model it appeard that the model is having least MSE at 10 variables.and I also plotted the results of it on the graph as shown below.



* Plot of BIC Vs No of variables
* It appears same as the forward subset with littles differences.
* When we look at the **bic** score of the variables in the graph it appears that **Protien, calories, fat sodium,fiber**, **carbo, potassium, sugars, vitamins** are presentin the model throughout the dataset and contribute to the model.
* when we look at the errors of different types here it appears that the RSS apporaches to zero when there are **9 variables** in the model.
* and Adjusted R squared error becomes closer to 1 when there are **9 variables** in the dataset.
* Cp error becomes closer to 0 when there are **9 variables** in the dataset.
* BIC score converges to low when there are **9 variables** in the dataset.

**Step 6 Conclusion**

The three model appear here to be giving us more or less the same results but when we look closely to the model it can be observed that when we look at the bic score the exhaustive subset model give us better predictions by selecting the less variables.