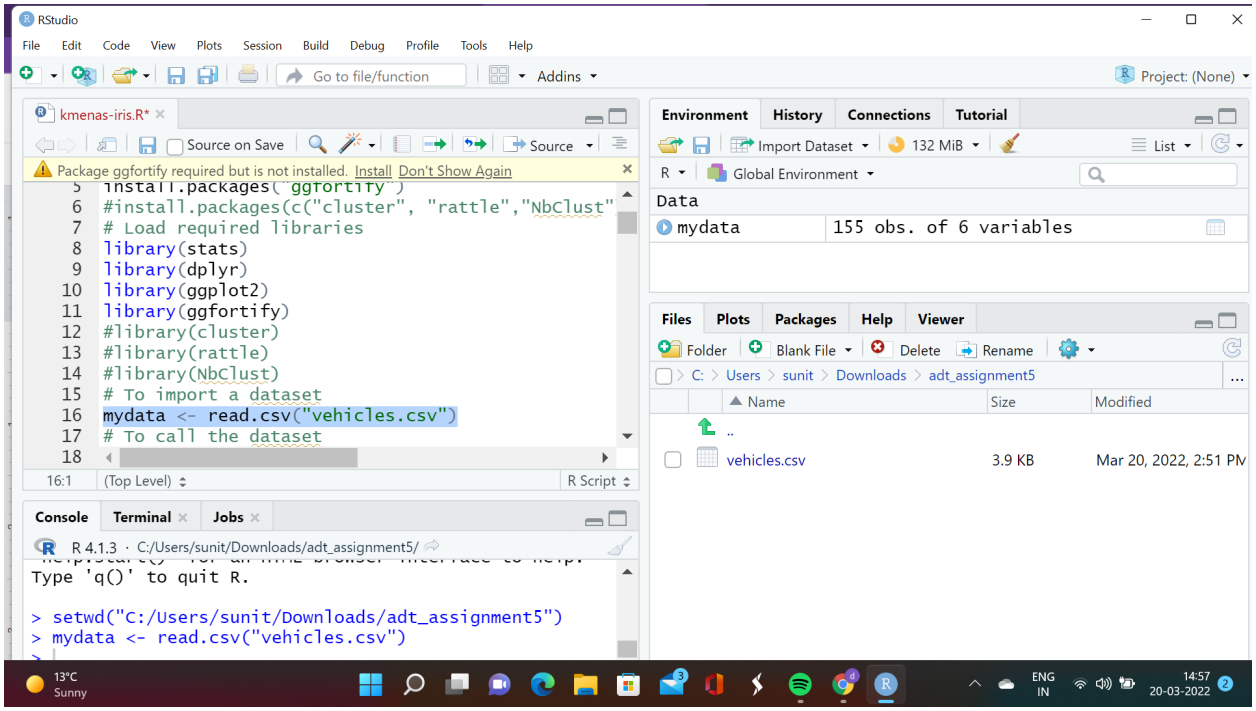
ADVANCED DATABASE TOPICS
ASSIGNMENT - 4

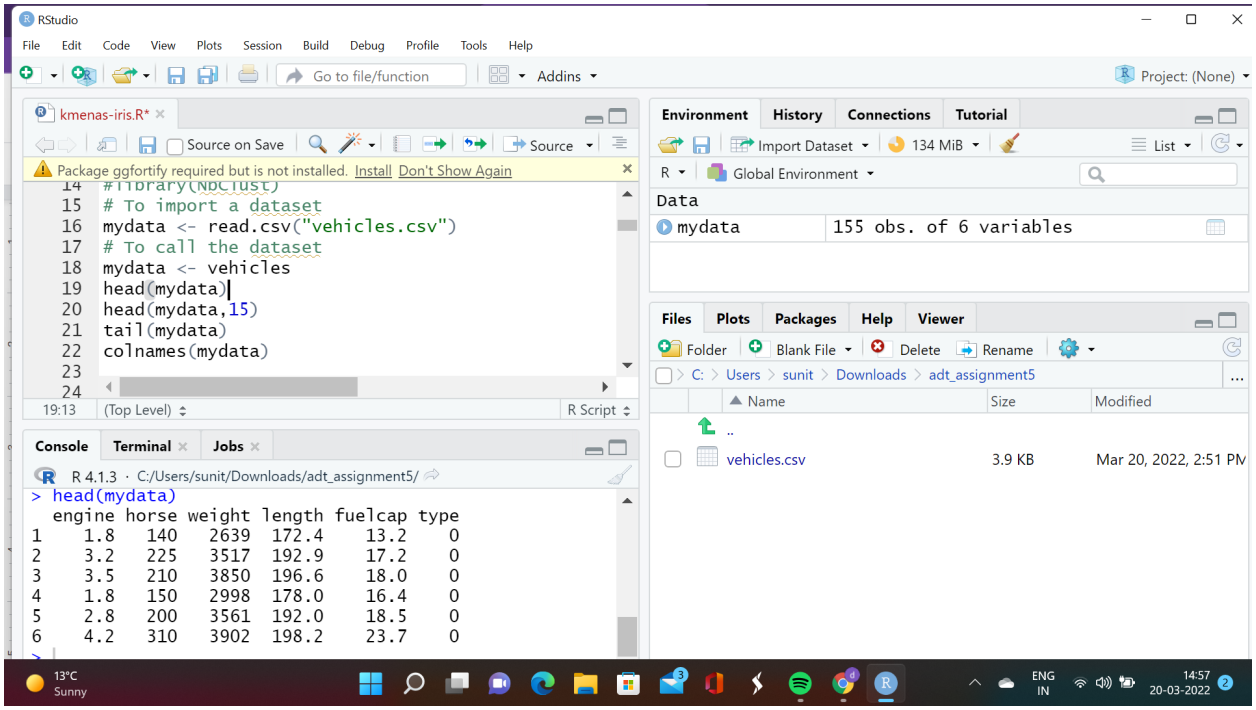
SUBMITTED BY: ABINAYA ELANCHEZHIAN
STUDENT ID: 110061220

---

1. Install, load the required packages, and import the data set (10 marks).



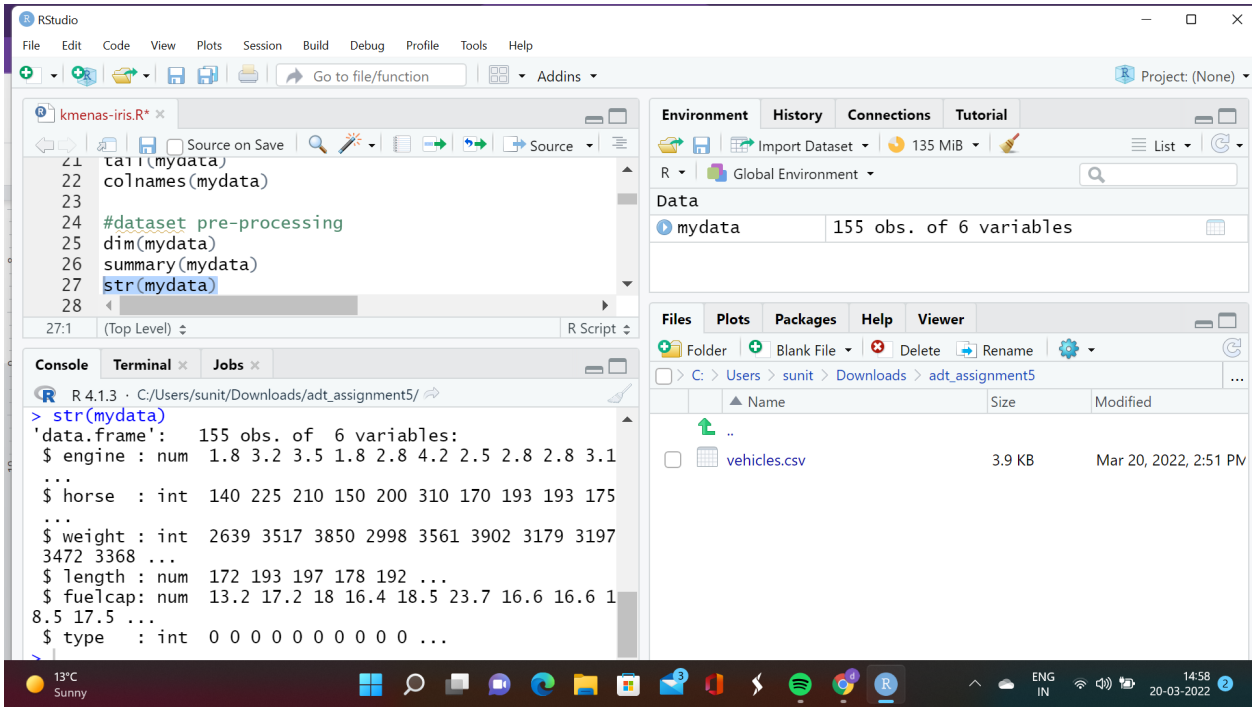
Command: mydata <- read.csv("vehicles.csv")
Explanation: Firstly, the working directory is set. The 'vehicles' dataset is imported in the R studio. The environment displays that there are 155 observations of 6 variables.



Command: head(mydata)

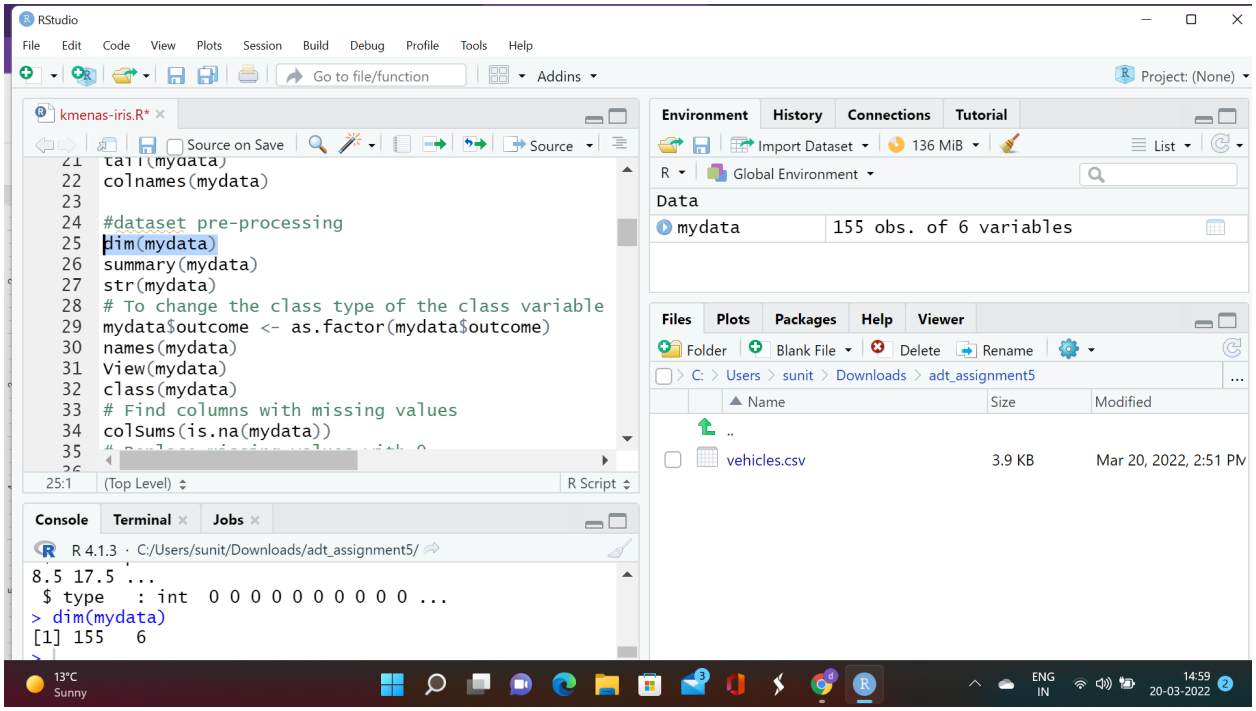
Explanation: The head() is used to display the first 6 rows of the 'vehicles' dataset. Hence, the dataset is imported successfully.

2. Upload data to R and view the structure of data, dimensionality, and attribute names (10 marks).



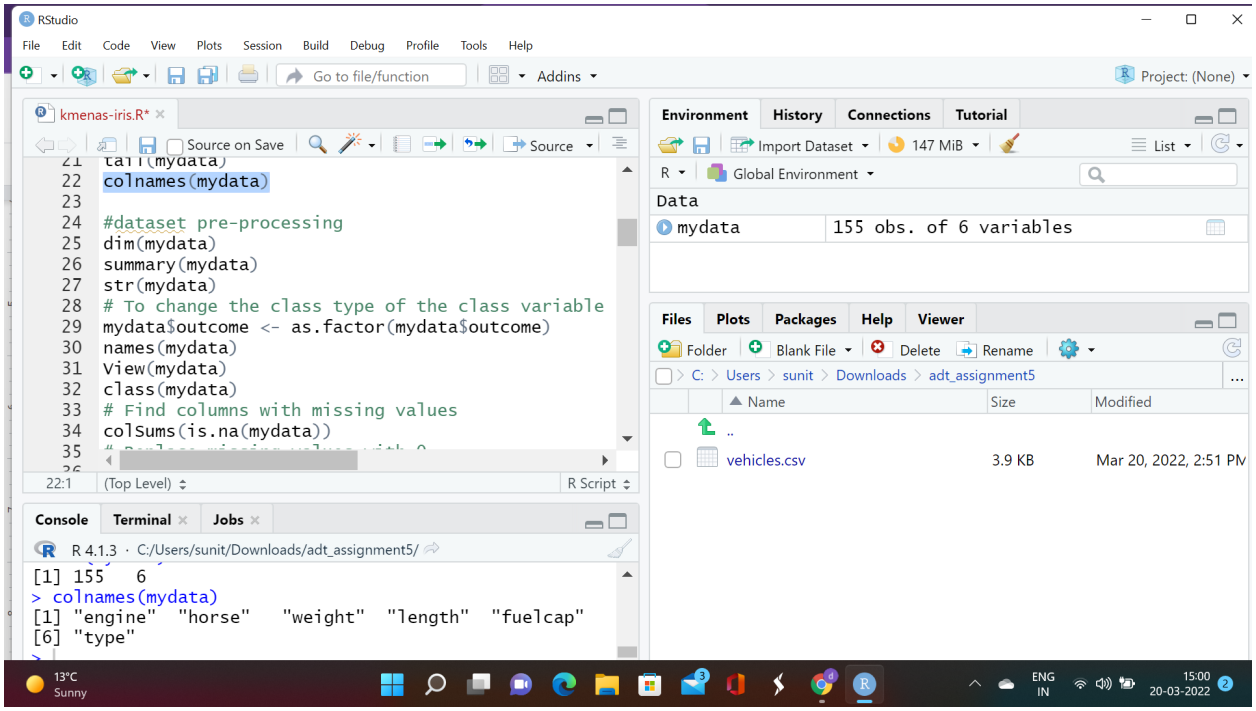
Command: str(mydata)

Explanation: The command compactly displays the internal structure of each of the attributes in the 'vehicles' dataset.



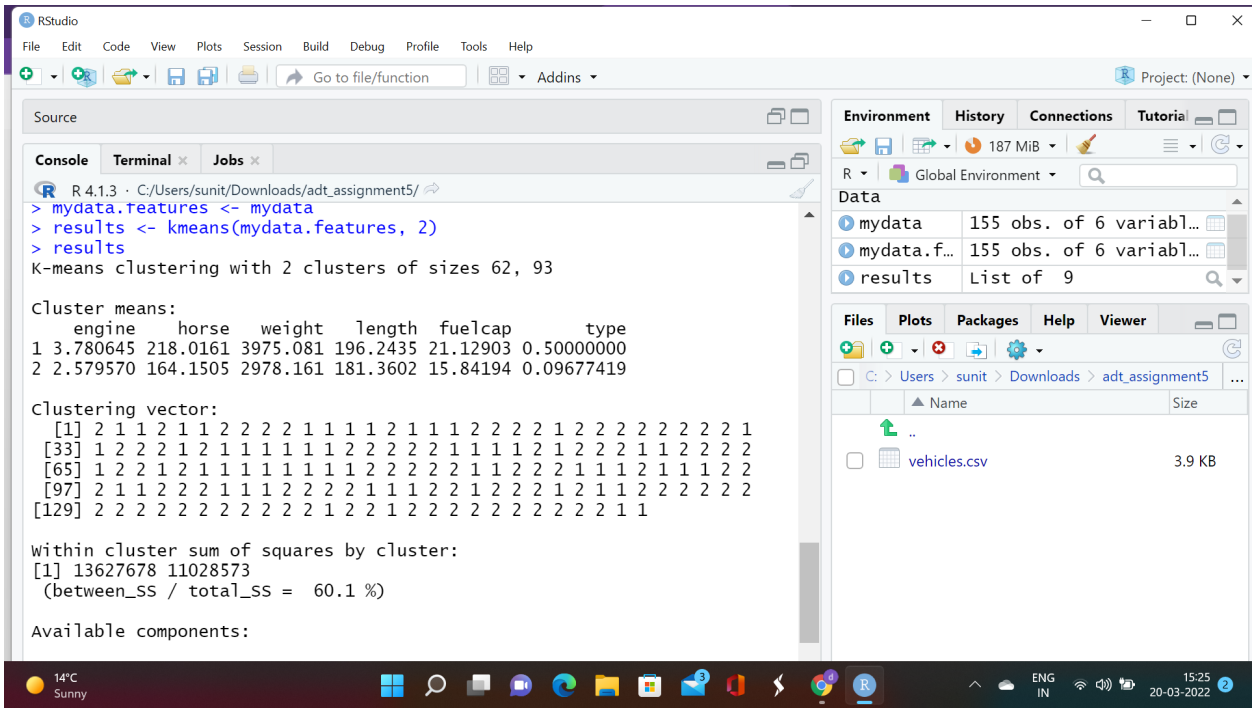
Command: dim(mydata)

Explanation: Displays the number of observations(rows) and variables(columns) available in the dataset. In the 'vehicles' dataset, there are 155 rows and 6 columns.

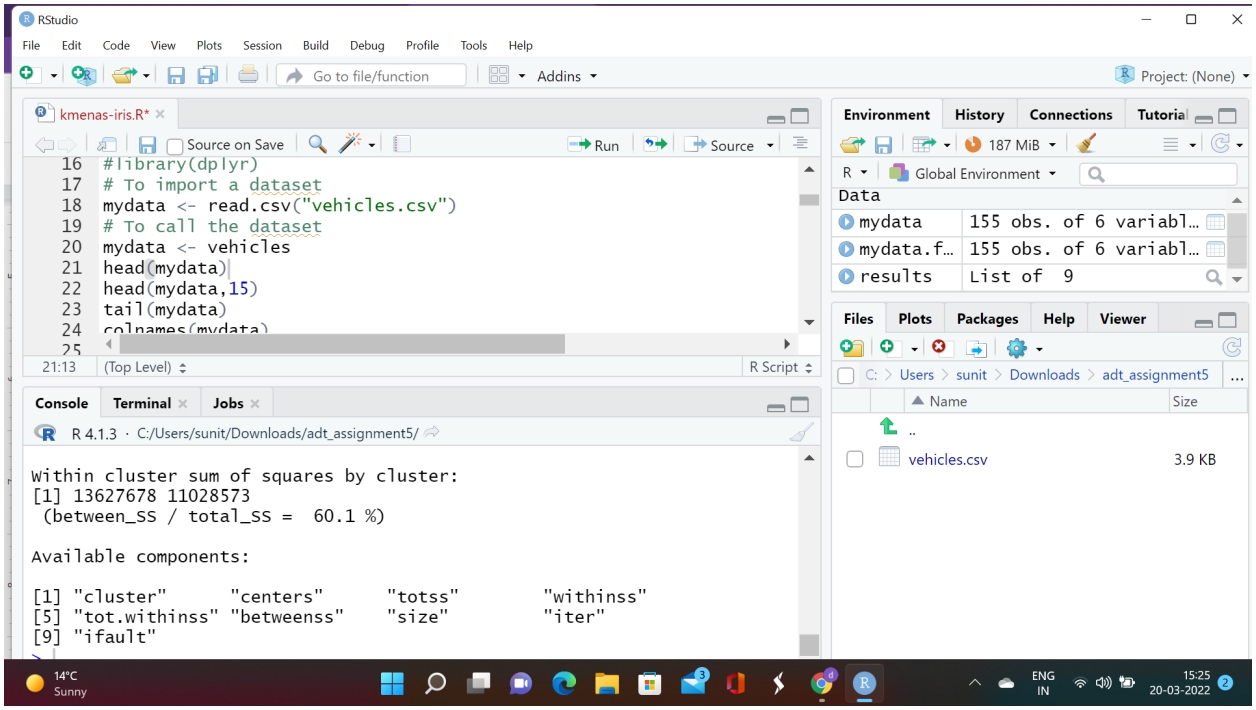


Command: colnames(mydata)
Explanation: The columns available in the 'vehicles' dataset are displayed. They are engine, horse, weight, length, fuelcap and type.

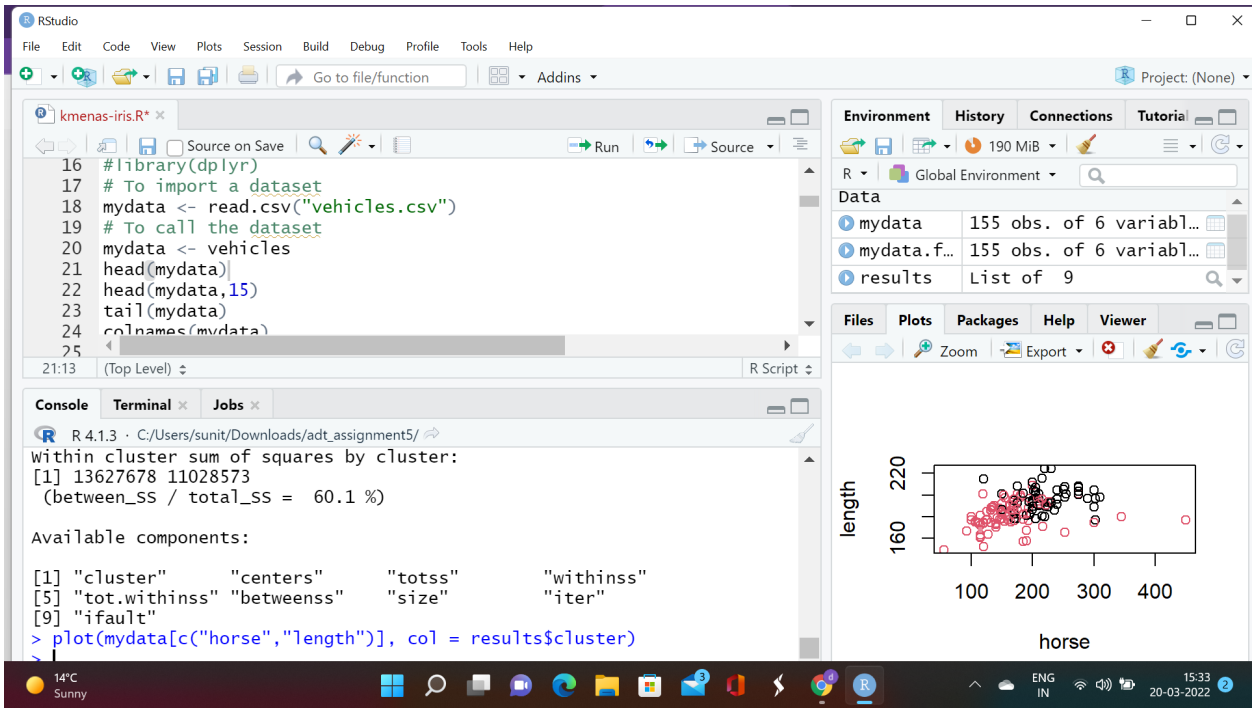
3. Explain what attributes you are using for clustering and why? (20 marks).
In simpler terms, K-means clustering minimizes the distance within a cluster and maximizes the distance between clusters. So the ideal combination of attributes will have two(considering k=2) distinct clusters that are apart.
Further, not every attribute would affect the cluster equally. Some attributes may be totally unrelated and may not add any difference to the clusters formed. As this dataset is completely new, the only way to know for sure the best combination of attributes to choose is through trying out various combinations. Once various clusters with various combinations are made, the 'Dunn index' can be used to evaluate and determine the best clusters.
To begin with, the attributes 'horse' and 'length' are used for clustering and analysis.



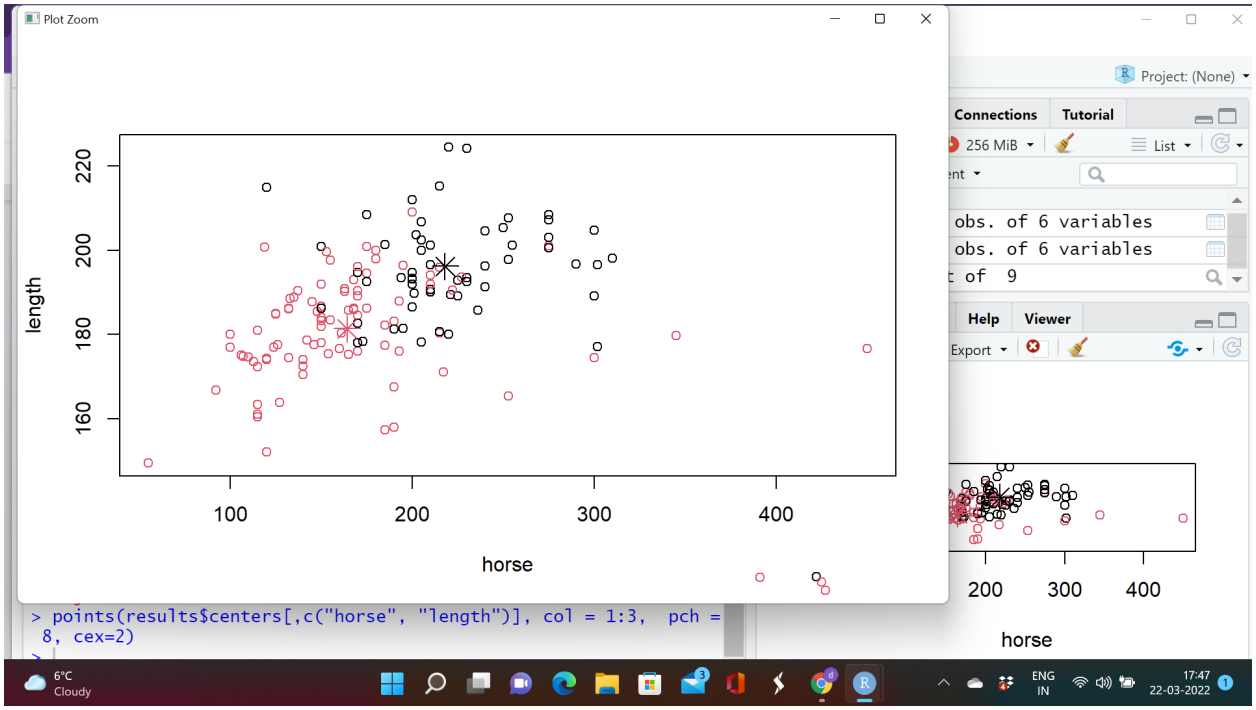
Command: mydata.features <- mydata
results <- kmeans(mydata.features, 2)
results
Explanation: A copy of mydata is stored in mydata.features and kmeans clustering is performed on mydata.features and this stored in results. In kmeans(mydata,2), the 2 denotes the number of clusters to be formed. The results displays information about the size of the clusters, clustering vector, cluster means, etc.



Explanation: The various components available in the cluster are displayed. The clusters can be differentiated into different colors by using any of these components.



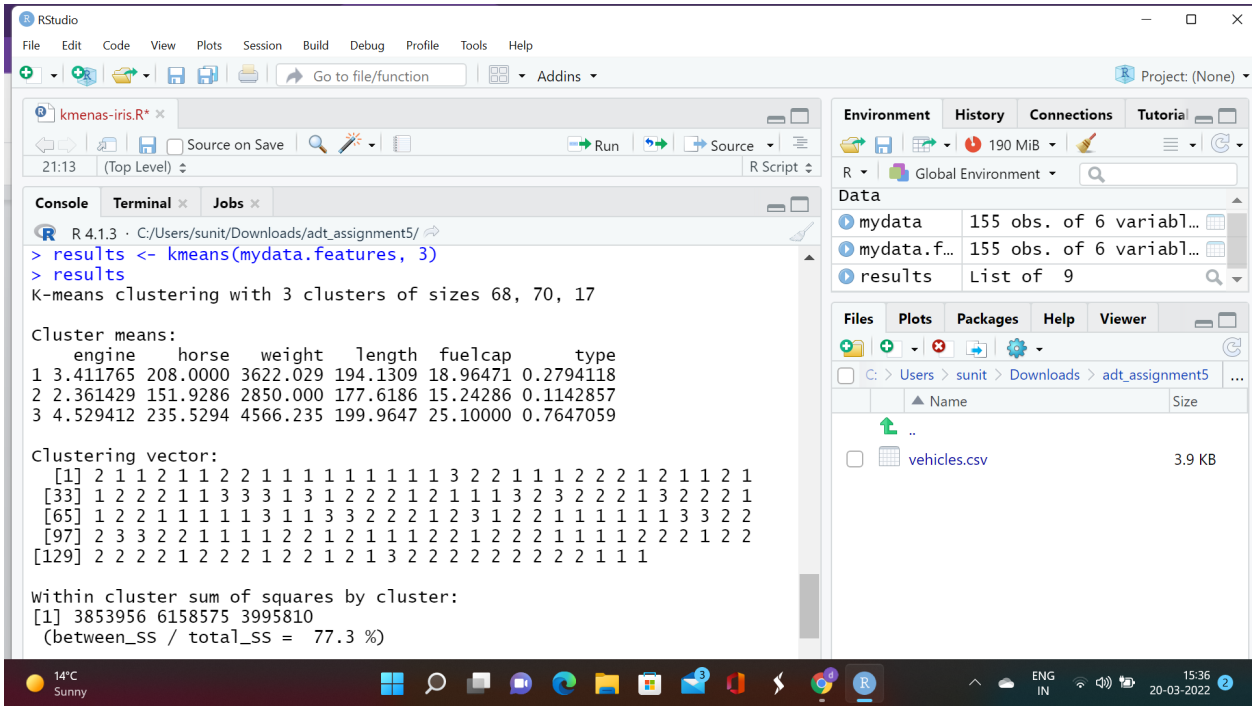
Command: plot(mydata[c("horse", "length")], col = results$cluster)
Explanation: The cluster is plotted by choosing the horse and length attributes. The clusters are differentiated in different colors using the 'col' command.

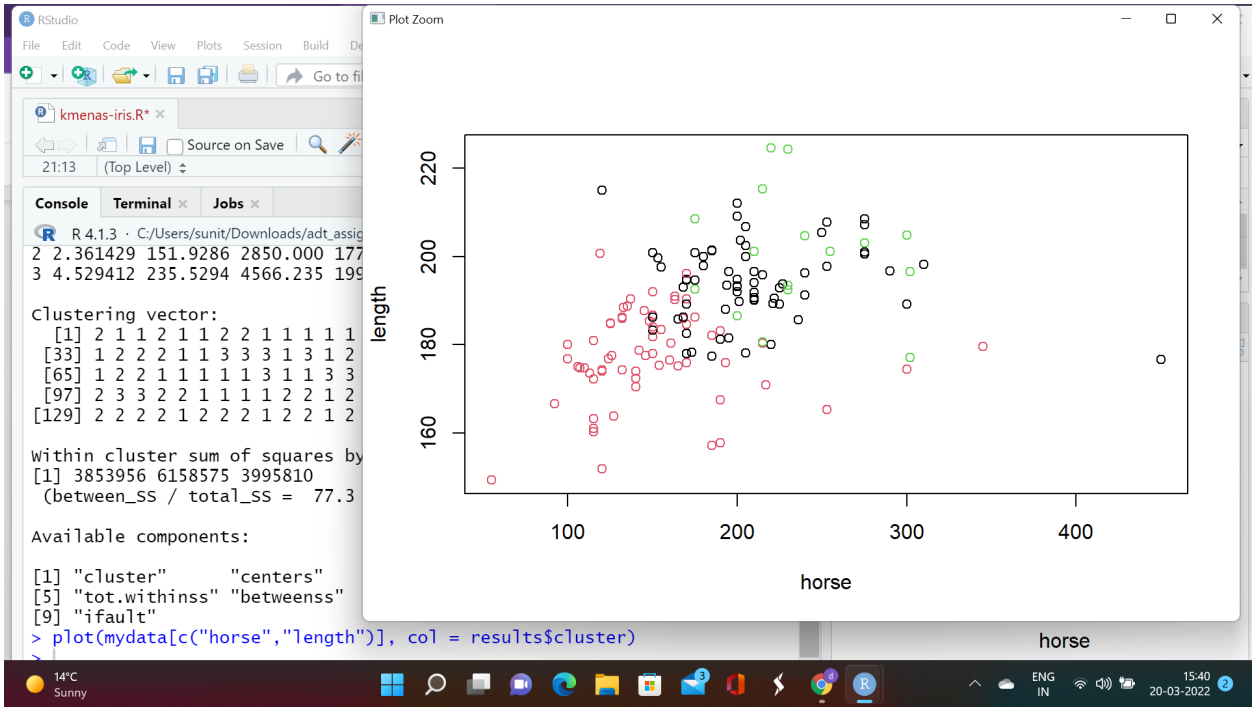


Command: points(results$centers[,c(“horse”, “length”)], col = 1:3, pch = 8, cex = 2)
Explanation: The centroids of both the clusters are plotted.

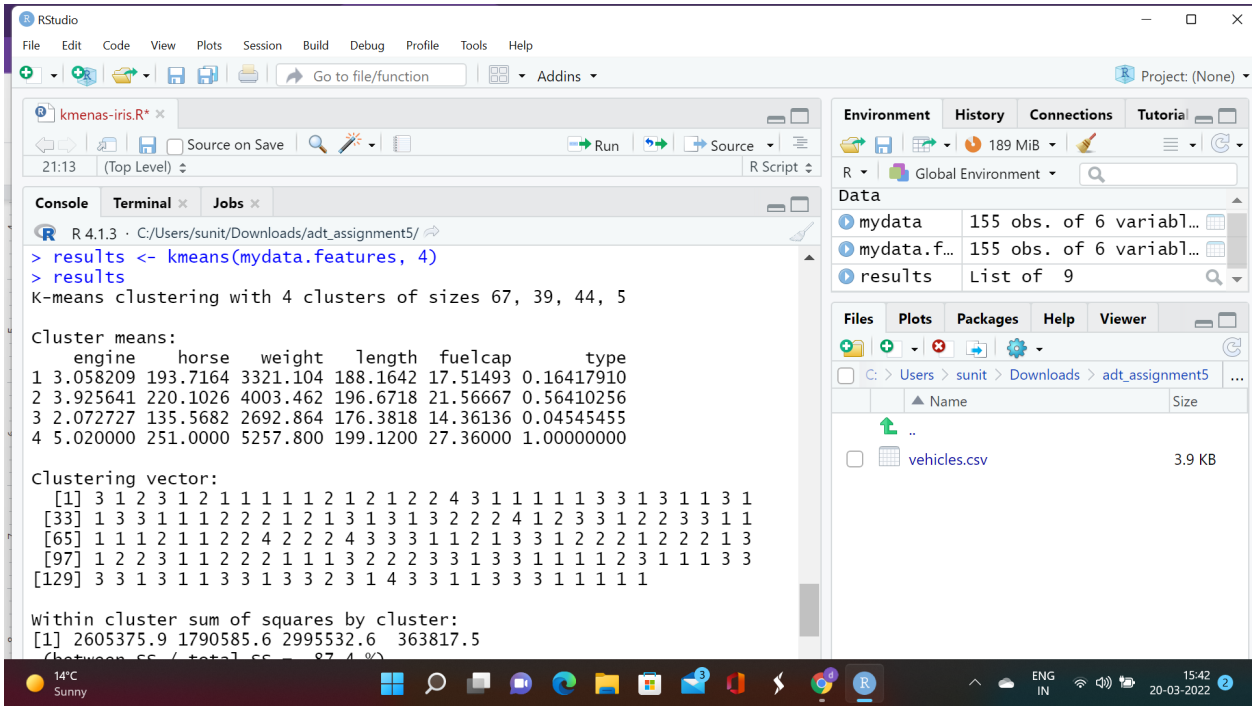
4. Apply clustering on the dataset for several K values, e.g., 3, 4 and 5 (10 marks).
Command: results <- kmeans(mydata.features, k) where k can be 3,4,5, etc.
results
plot(mydata[c(“horse”, “length”)], col = results$cluster)
Explanation: As similar to question 3, k-means clustering is performed on mydata.features for different k values where k denotes the number of clusters to be formed. The results are displayed which displays various information like cluster sizes, cluster means, cluster vectors and available components. The plot() is used to plot the clusters and the clusters are differentiated in various colors.

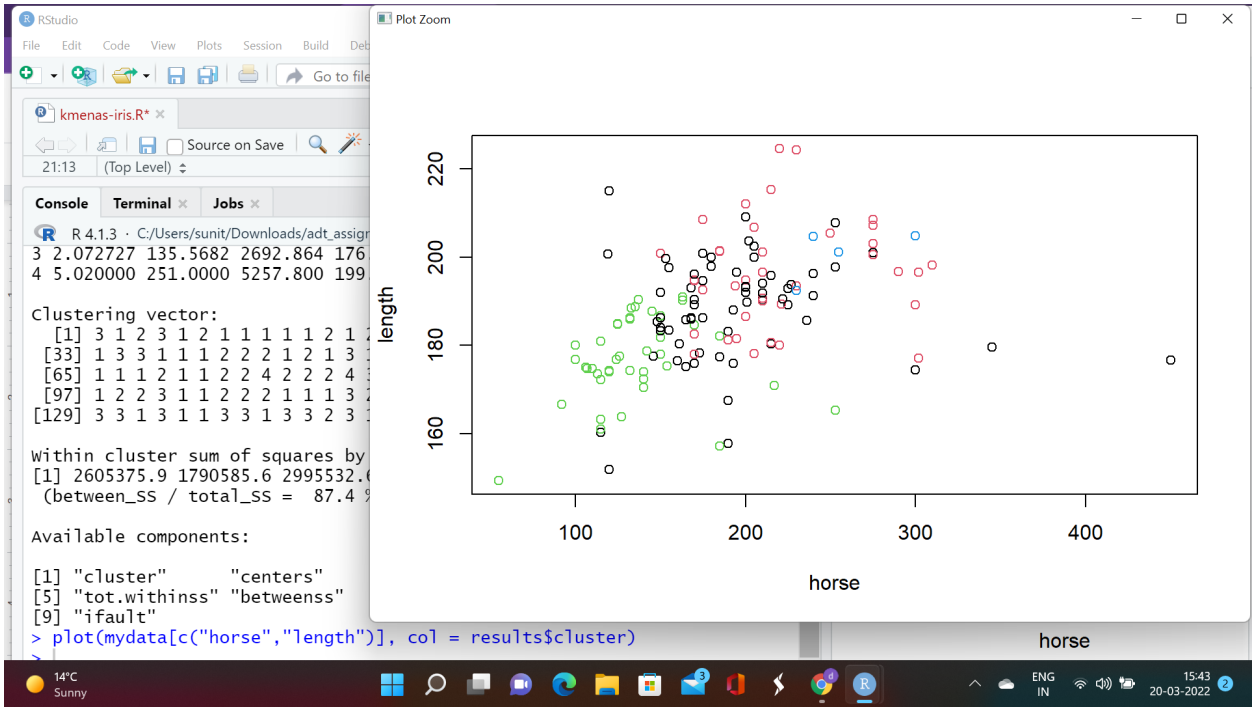
(i) K = 3:


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
kmenas-iris.R*
Source on Save
Run
Source
21:13 (Top Level)
R Script
Console Terminal Jobs
R 4.1.3 · C:/Users/sunit/Downloads/adt_assignment5/
> results <- kmeans(mydata.features, 3)
> results
K-means clustering with 3 clusters of sizes 68, 70, 17
Cluster means:
engine horse weight length fuelcap type
1 3.411765 208.0000 3622.029 194.1309 18.96471 0.2794118
2 2.361429 151.9286 2850.000 177.6186 15.24286 0.1142857
3 4.529412 235.5294 4566.235 199.9647 25.10000 0.7647059
Clustering vector:
[1] 2 1 1 2 1 1 2 2 1 1 1 1 1 1 1 1 3 2 2 1 1 1 2 2 2 1 2 1 1 2 1
[33] 1 2 2 2 1 1 3 3 3 1 3 1 2 2 2 1 2 1 1 1 3 2 3 2 2 2 1 3 2 2 2 1
[65] 1 2 2 1 1 1 1 1 1 3 1 1 3 3 2 2 2 1 2 3 1 2 2 1 1 1 1 1 1 3 3 2 2
[97] 2 3 3 2 2 1 1 1 1 1 2 2 1 2 1 1 1 2 2 1 1 2 2 2 1 1 1 1 2 2 2 1 2 2
[129] 2 2 2 2 1 2 2 2 1 2 2 1 2 1 3 2 2 2 2 2 2 2 2 2 1 1 1
Within cluster sum of squares by cluster:
[1] 3853956 6158575 3995810
(between_SS / total_SS = 77.3 %)
Environment History Connections Tutorial
Global Environment
Data
mydata 155 obs. of 6 variabl...
mydata.f... 155 obs. of 6 variabl...
results List of 9
Files Plots Packages Help Viewer
C: Users sunit Downloads adt_assignment5
Name Size
vehicles.csv 3.9 KB
14°C Sunny
ENG IN
15:36 20-03-2022

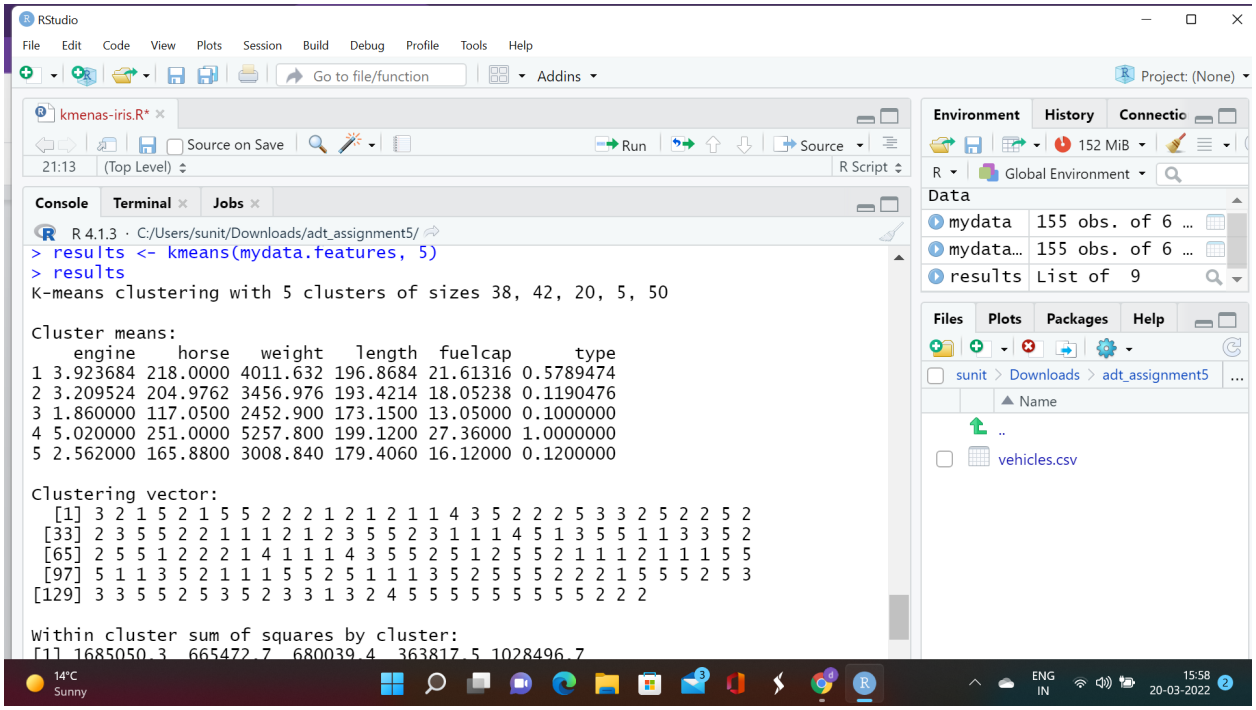

RStudio
kmenas-iris.R*
Source on Save
21:13 (Top Level)
Console Terminal Jobs
R 4.1.3 · C:/Users/sunit/Downloads/adt_assig
2 2.361429 151.9286 2850.000 177
3 4.529412 235.5294 4566.235 199
Clustering vector:
[1] 2 1 1 2 1 1 2 2 1 1 1 1 1
[33] 1 2 2 2 1 1 3 3 3 1 3 1 2
[65] 1 2 2 1 1 1 1 1 1 3 1 1 3 3
[97] 2 3 3 2 2 1 1 1 1 1 2 2 1 2
[129] 2 2 2 2 1 2 2 2 1 2 2 1 2
Within cluster sum of squares by
[1] 3853956 6158575 3995810
(between_SS / total_SS = 77.3
Available components:
[1] "cluster" "centers"
[5] "tot.withinss" "betweenss"
[9] "ifault"
> plot(mydata[c("horse","length")], col = results$cluster)
Plot Zoom
length
220
200
180
160
100
200
300
400
horse
horse
14°C Sunny
ENG IN
15:40 20-03-2022

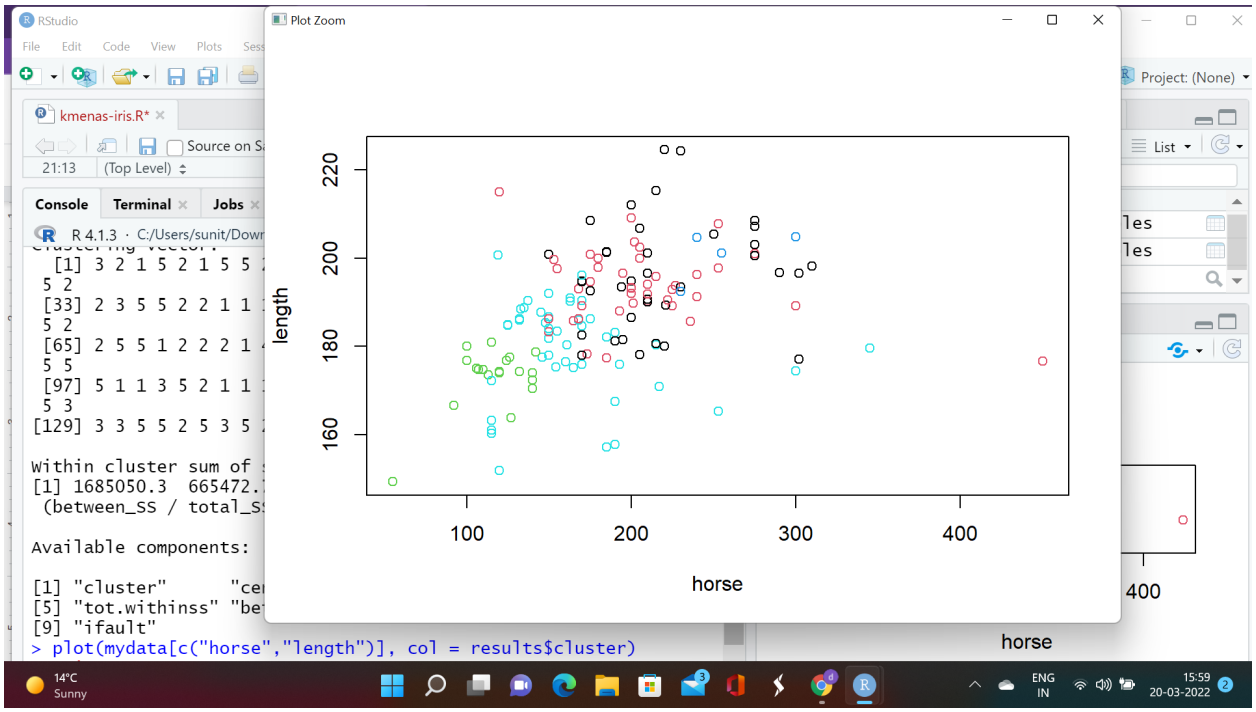
(ii) K = 4:


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
kmenas-iris.R*
Source on Save
Run
Source
21:13 (Top Level)
R Script
Console Terminal Jobs
R 4.1.3 · C:/Users/sunit/Downloads/adt_assignment5/
> results <- kmeans(mydata.features, 4)
> results
K-means clustering with 4 clusters of sizes 67, 39, 44, 5
Cluster means:
engine horse weight length fuelcap type
1 3.058209 193.7164 3321.104 188.1642 17.51493 0.16417910
2 3.925641 220.1026 4003.462 196.6718 21.56667 0.56410256
3 2.072727 135.5682 2692.864 176.3818 14.36136 0.04545455
4 5.020000 251.0000 5257.800 199.1200 27.36000 1.00000000
Clustering vector:
[1] 3 1 2 3 1 2 1 1 1 1 1 2 1 2 1 2 2 4 3 1 1 1 1 1 3 3 1 3 1 1 3 1
[33] 1 3 3 3 1 1 1 2 2 2 1 2 1 3 1 3 1 3 2 2 2 4 1 2 3 3 1 2 2 3 3 1 1
[65] 1 1 1 2 1 1 2 2 4 2 2 2 4 3 3 3 1 1 2 1 3 3 1 2 2 2 1 2 2 2 1 3
[97] 1 2 2 3 1 1 2 2 2 1 1 1 3 2 2 2 3 3 1 3 3 1 1 1 1 2 3 1 1 1 3 3
[129] 3 3 1 3 1 1 3 3 1 3 3 2 3 1 4 3 3 1 1 3 3 3 1 1 1 1 1
Within cluster sum of squares by cluster:
[1] 2605375.9 1790585.6 2995532.6 363817.5
Environment History Connections Tutorial
189 MiB
Global Environment
Data
mydata 155 obs. of 6 variabl...
mydata.f... 155 obs. of 6 variabl...
results List of 9
Files Plots Packages Help Viewer
C: Users sunit Downloads adt_assignment5
Name Size
..
vehicles.csv 3.9 KB
14°C Sunny
ENG IN
15:42 20-03-2022


RStudio
kmenas-iris.R*
Source on Save
21:13 (Top Level)
Console Terminal Jobs
R 4.1.3 · C:/Users/sunit/Downloads/adt_assign
3 2.072727 135.5682 2692.864 176
4 5.020000 251.0000 5257.800 199
Clustering vector:
[1] 3 1 2 3 1 2 1 1 1 1 1 2 1
[33] 1 3 3 3 1 1 1 2 2 2 1 2 1 3
[65] 1 1 1 2 1 1 2 2 4 2 2 2 4
[97] 1 2 2 3 1 1 2 2 2 1 1 1 3
[129] 3 3 1 3 1 1 3 3 1 3 3 2 3
Within cluster sum of squares by
[1] 2605375.9 1790585.6 2995532.6
(between_SS / total_SS = 87.4 %
Available components:
[1] "cluster" "centers"
[5] "tot.withinss" "betweenss"
[9] "ifault"
> plot(mydata[c("horse","length")], col = results$cluster)
Plot Zoom
length
horse
100 200 300 400
160 180 200 220
horse
14°C Sunny
ENG IN
15:43 20-03-2022

(iii) K = 5:



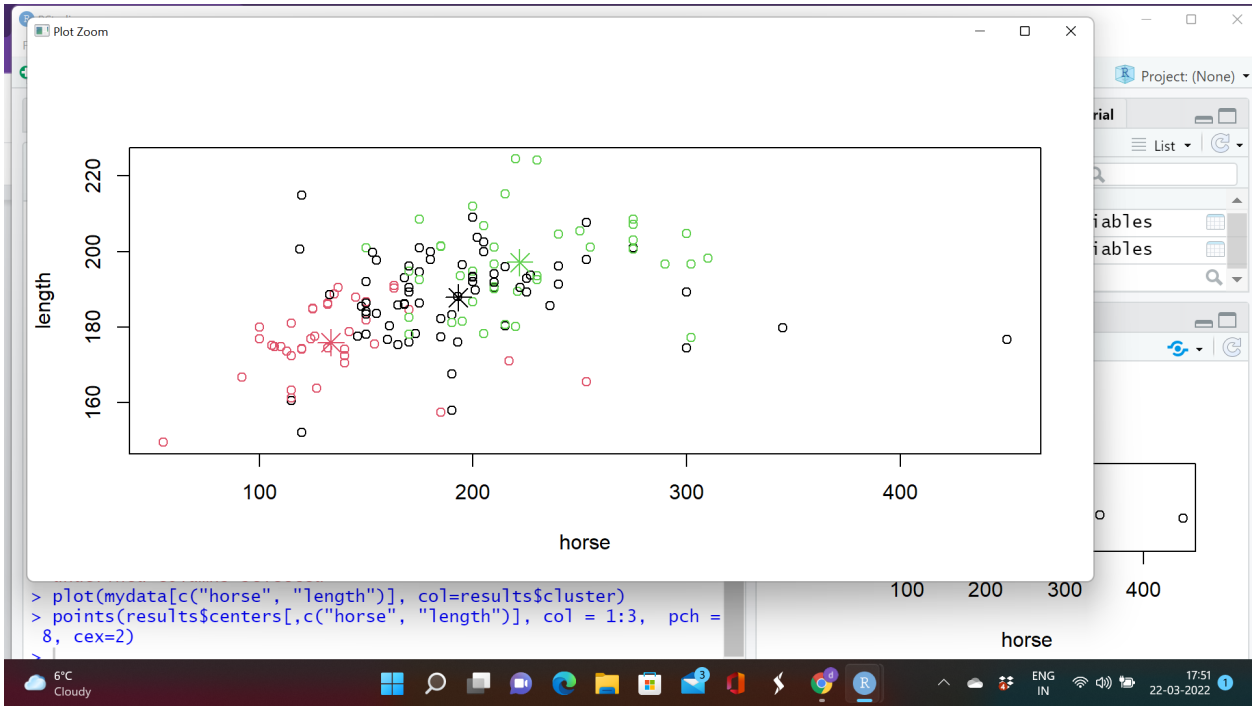


5. Plot the model and show the centroids for each K selected in the previous question (10 marks).

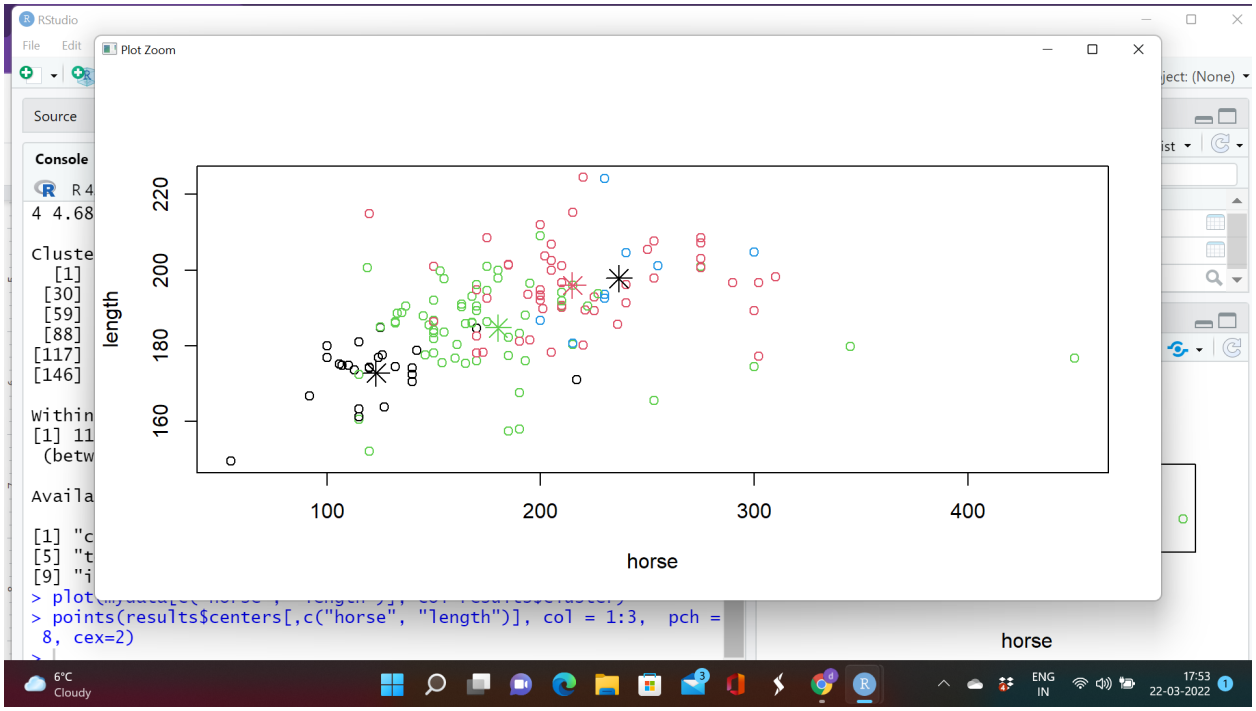
Command: points(results$centers[,c("horse", "length")], col = 1:3, pch = 8, cex = 2)

Explanation: The centroids of all the clusters are plotted. They are denoted by * symbol.

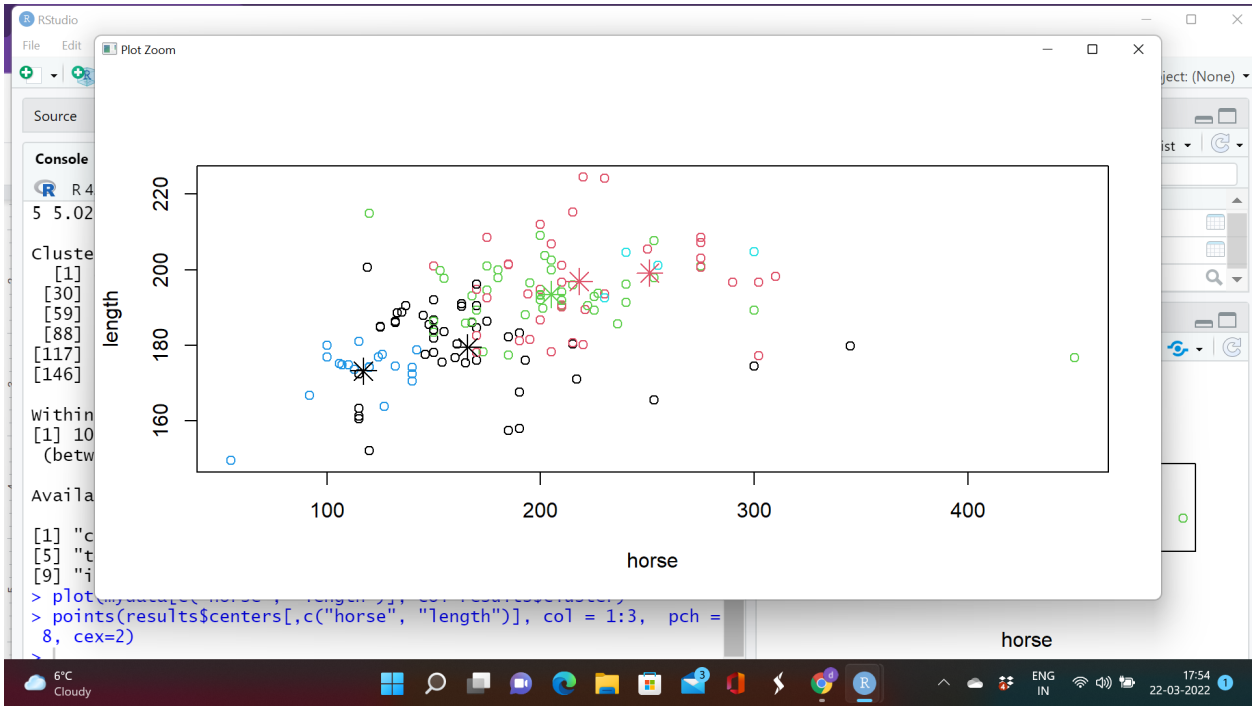
(i) K=3:


Plot Zoom
length
horse
100
200
300
400
160
180
200
220
> plot(mydata[c("horse", "length")], col=results$cluster)
> points(results$centers[,c("horse", "length")], col = 1:3, pch =
8, cex=2)

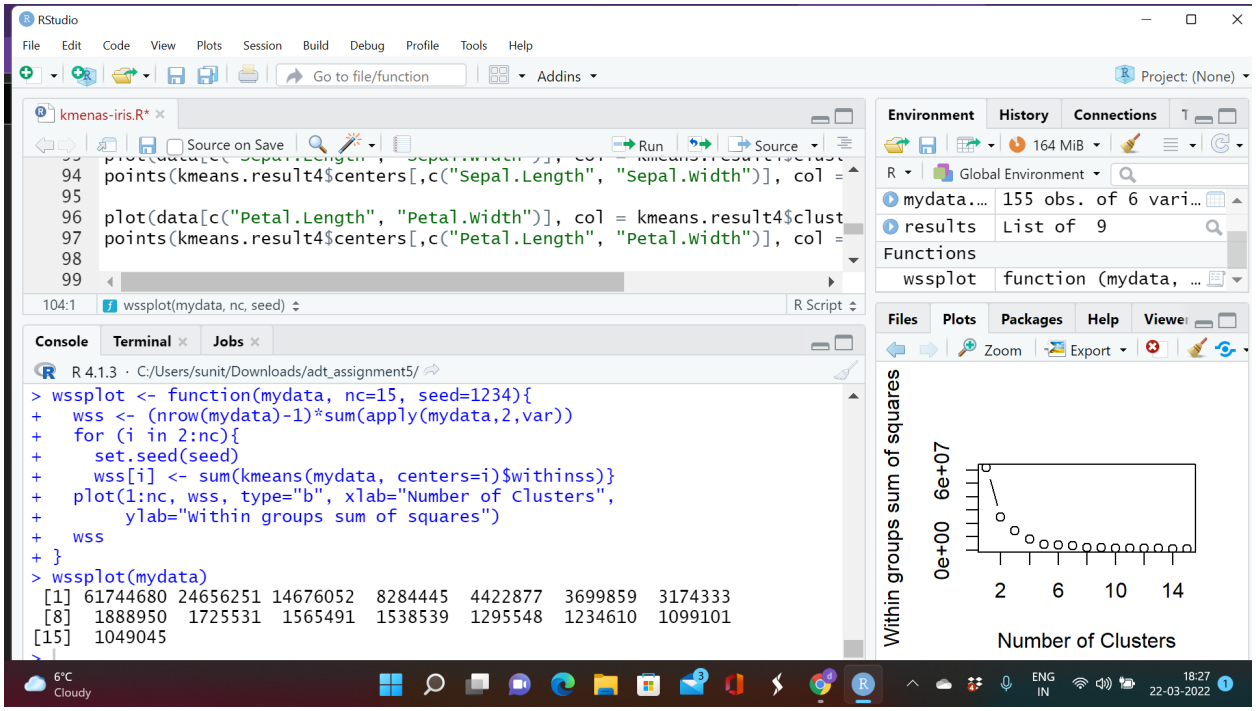
(ii) K=4:


RStudio
Plot Zoom
length
horse
100
200
300
400
160
180
200
220
> points(results$centers[,c("horse", "length")], col = 1:3, pch =
8, cex=2)

(iii)K=5:



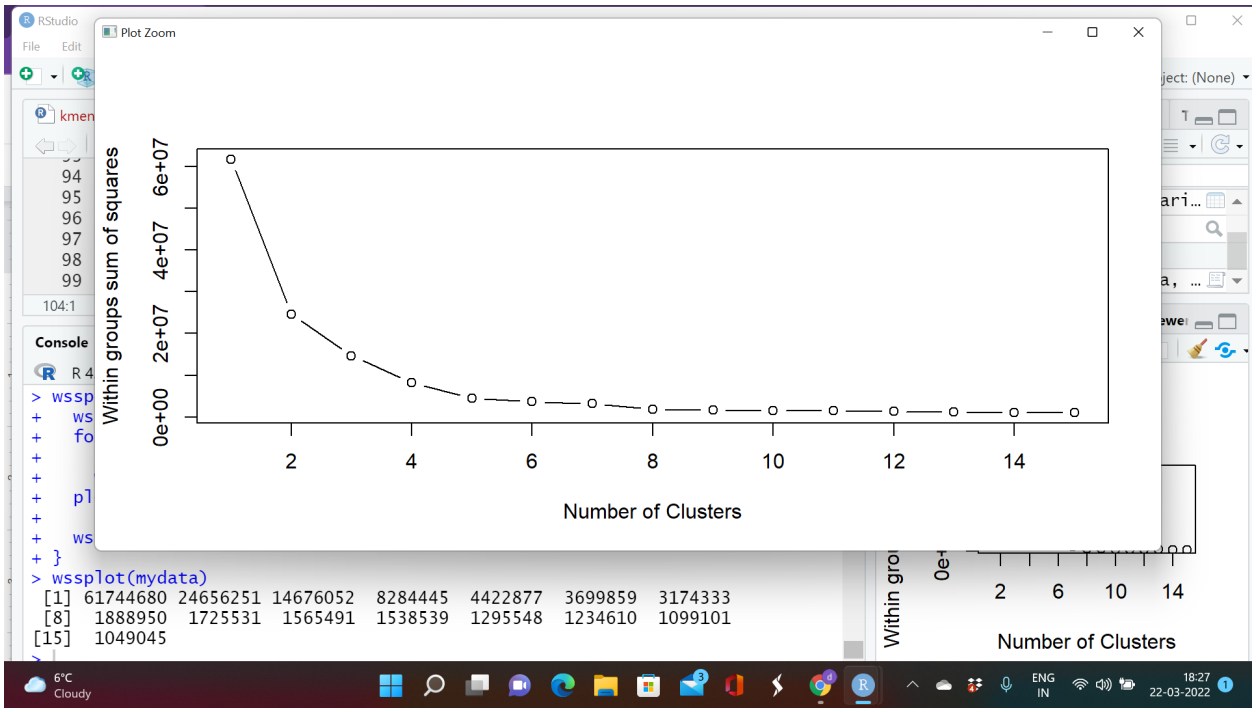
6. Show the optimum number of clusters (20 marks).



Command: wssplot <- function(mydata, nc=15, seed=1234){
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:nc){
set.seed(seed)
wss[i] <- sum(kmeans(mydata, centers=i)$withinss)}

```
    plot(1:nc, wss, type="b", xlab="Number of Clusters",
         ylab="Within groups sum of squares")
    wss
  }
wssplot(mydata)
```

Explanation: The command displays the optimum number of clusters for the ‘vehicles’ dataset. It is an elbow curve and the value at the elbow displays the optimum number of clusters that can be formed.

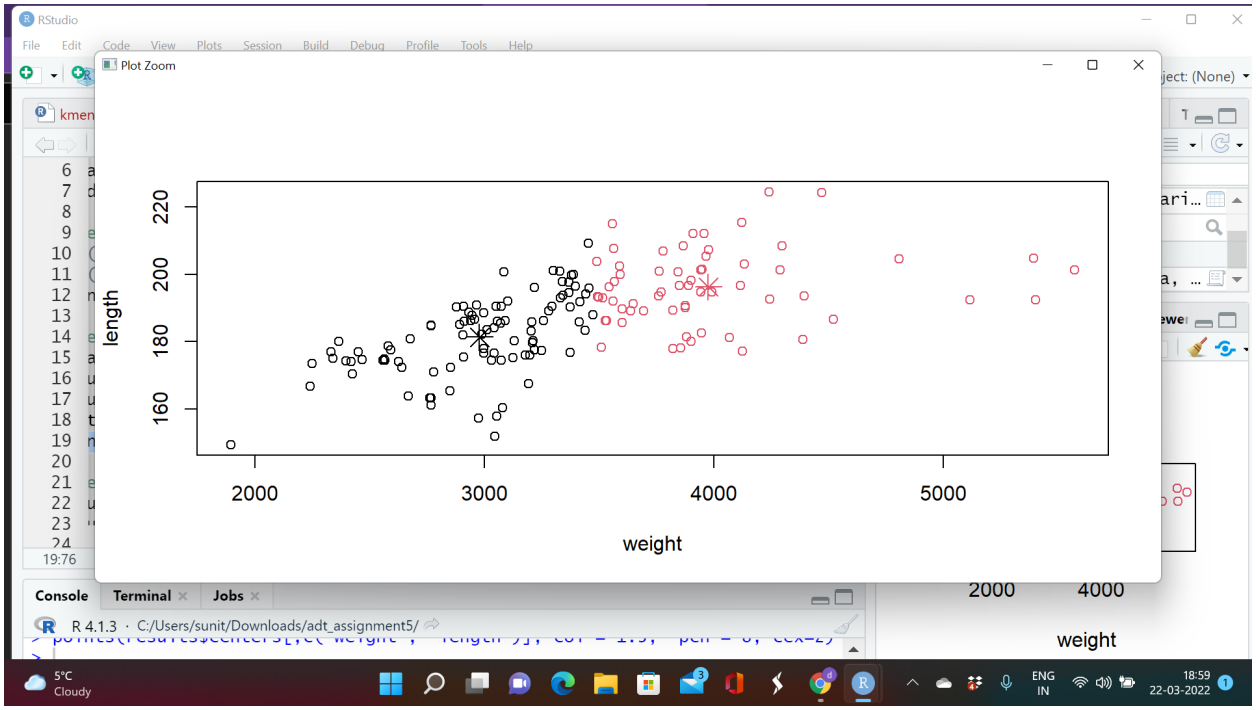


7. Select different attributes and explain which ones show good clusters (20 marks).

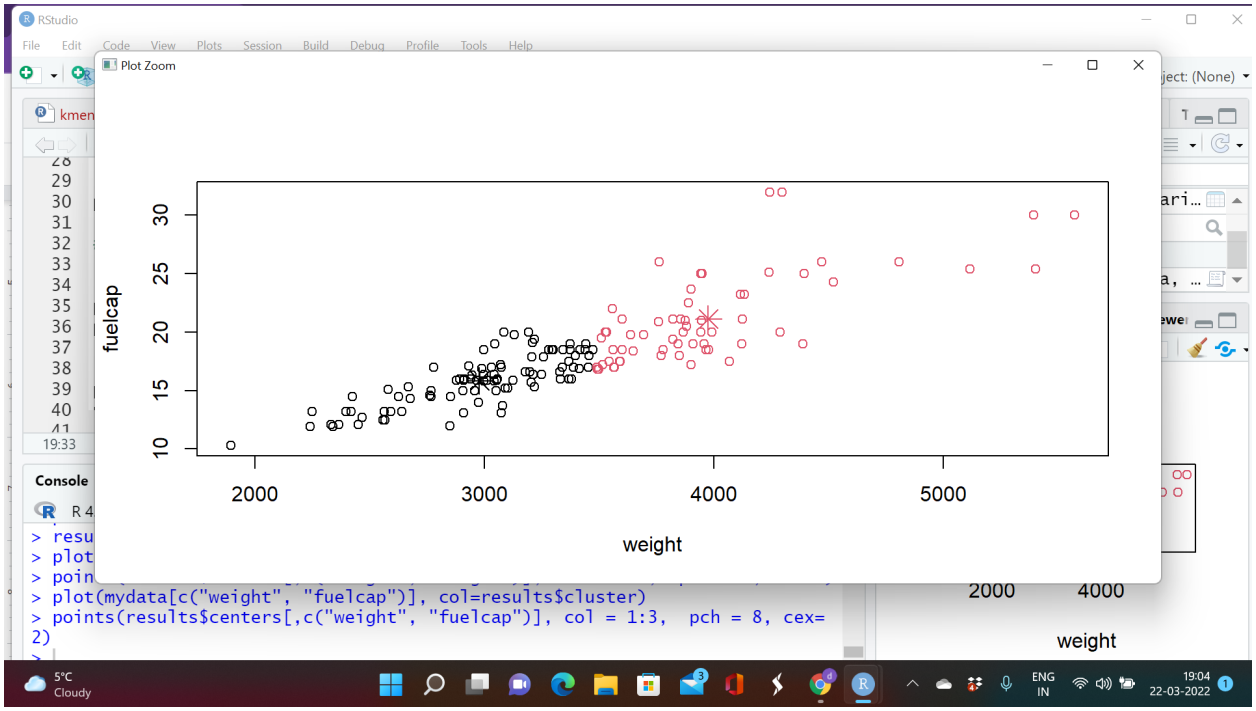
Command: plot(mydata, col = results$cluster)
Explanation: The command plots clusters with various combinations of the available attributes.

In my opinion, good cluster combinations are shown by (weight, length) and (weight, fuelcap) combinations. The reasons are:

1. Distance inside the cluster should be minimal and distance between clusters should be maximum. Among the available cluster combinations, most combinations have overlapping clusters. So these clusters are not ideal.
2. From the other available cluster combinations, the clusters with well different centroids that are distinct and far from each other are chosen as ideal.



Explanation: K-means clustering for weight and length attributes from the ‘vehicles’ dataset.



Explanation: K-means clustering for weight and fuelcap attributes from the ‘vehicles’ dataset.