

**Assignment 4**

**Advanced Database Topics**

**(COMP 8157)**

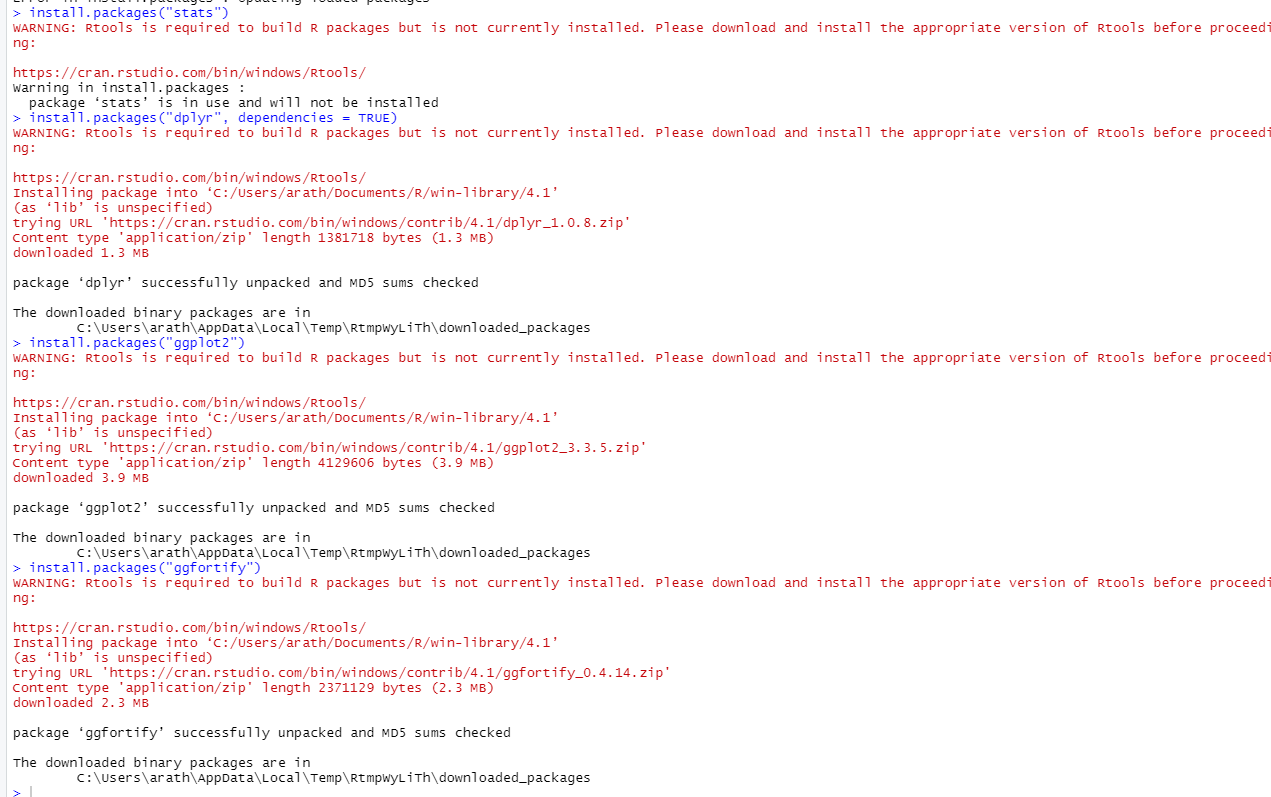
**Dr. Shafaq Khan**

**Submitted by**

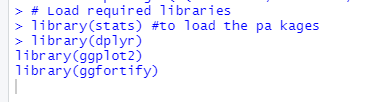
**Arathy Unni (110061760)**

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

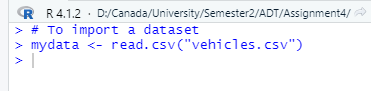
Installing the required packages.



Importing the required packages.



Importing the dataset.



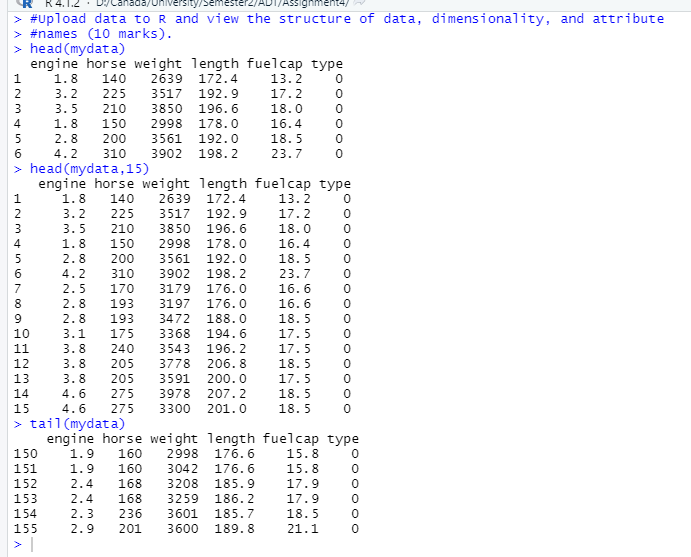
2. Upload data to R and view the structure of data, dimensionality, and attribute names (10 marks).

Here we are viewing the data of the dataset using the following commands:

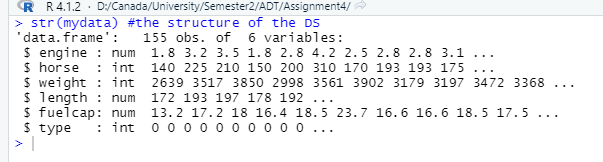
head(mydata) => which is getting the first part of the dataset mydata (6 records)

head(mydata,15) => which is getting the first 15 items of the dataset mydata

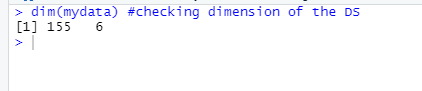
tail(mydata) => which is getting the last part of the dataset mydata



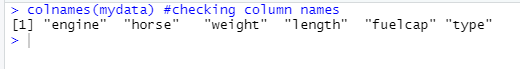
To display the structure of the dataset.



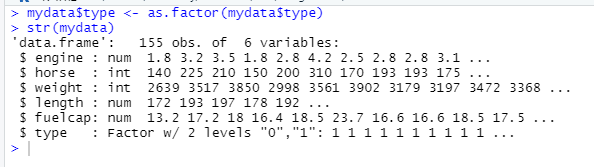
To display the dimension of the dataset.



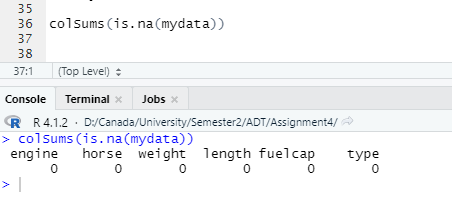
To display the attribute names of the dataset.



Additionally, I am converting the type of “type” attribute, which is basically a labelled attribute to factor from int. Because factor in R is capable of categorizing and storing data having very limited number of different values. Here the attribute “type” had only “0” and “1” as it’s values.

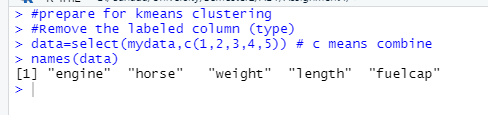


Also, as part of pre-processing I am validating if there exists any attribute with value = N/A. For the “vehicle” dataset there are no attributes with N/A.



3. Explain what attributes you are using for clustering and why? (20 marks).

The attributes chosen for clustering purposes are engine, horse, weight, length, and fuelcap, which are all unlabelled attributes. The reason is that clustering is an unsupervised learning problem and aims to find a structure in a collection of unlabelled data. And I have not chosen the attribute “type” for clustering here because it acts like a labelled attribute; hence it is best to be used for classification and regression instead of clustering.



4. Apply clustering on the dataset for several K values, e.g., 3, 4 and 5 (10 marks).

Applying clustering on dataset for the value of K=3.

Graphical user interface, text, application, email

Description automatically generated

Applying clustering on dataset for the value of K=4.

Graphical user interface, text, application, email

Description automatically generated

Applying clustering on dataset for the value of K=5.

Text, table

Description automatically generated with medium confidence

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

For K = 3, the plot and centroids between the **attribute’s “weight” and “fuelcap”**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

For K = 4, the plot and centroids between the **attribute’s “weight” and fuelcap”**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

For K = 5, the plot and centroids between the **attribute’s “weight” and fuelcap”**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

For K = 3, the plot and centroids between the **attribute’s “engine” and “weight”**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

For K = 4, the plot and centroids between the **attribute’s “engine” and “weight”**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

For K = 5, the plot and centroids between the **attribute’s “engine” and “weight”**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

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

The optimum number of clusters is found using the Elbow method. We know that the basic idea of clustering is to define clusters such that the total intra-cluster variation or, in other words, the total within-cluster sum of square (WSS) is minimized. The total WSS is used to measure the compactness of the clustering, and it is to be kept as small as possible. The Elbow method looks at the total WSS as a function of the number of clusters.

Using the Elbow method, the optimum number of clusters could be determined as below:

1. To start with, we need to compute the clustering algorithm for different values of K (K-means clustering)

2. For each value of K, compute the total WSS

3. Plot the curve of WSS

4. The location of the bend in the plot is the appropriate number of clusters.

In our case, the bend is seen at point 5 in the curve; hence the optimum number of clusters = 5.

Chart

Description automatically generated

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

Here I am plotting all the variables together so we can determine which set of attributes provide the best cluster for 3 different values of K (3,4, and 5)

For K=3 (please zoom in the picture)

Chart, scatter chart

Description automatically generated

For K=4 (please zoom in the picture)

Chart, scatter chart

Description automatically generated

For K=5 (please zoom in the picture)

Chart, scatter chart

Description automatically generated

For the values of K=3/4/5, we can see that the following attribute pairs show good clusters:

Engine – Weight

Horse – Weight

Weight – Engine

Weight – Horse

Weight – Length

Weight – Fuelcap

Length – Weight

Fuelcap – Weight

A cluster is the best if the objects in that group will be similar or related to one another and different from the objects in other group, i.e., intra-cluster distance must be minimum and inter-cluster distance should be maximum. For the above-mentioned attribute pairs, the clusters formed seem to follow these properties. Also, it is best for the objects in one cluster to not overlap with the objects in the others as well. We can see that the clusters formed by other attribute-pairs seems to overlap.