Unsupervised Learning – Individual Assignment 1

Name: Yash Srivastava

Email: yash_srivastava_ampba2021s@isb.edu

PGID: 12010060

Steps for Hierarchal Cluster Analysis:

1. Reading the dataset
2. Data Preparation: Treating Missing values and normalizing the data
3. Calculating Dissimilarity matrix using euclidean distance
4. Performing Hierarchical clustering using Complete Linkage
5. Plotting the obtained dendrogram
6. Creating subclusters using CUTREE
7. Checking for optimal no. of clusters

a. Silhouette score

b. Elbow curve

8. Cluster stability: To check for cluster stability we take a random 95% sample of data. Then repeat the process of cluster analysis on it.
9. Cluster profiling: Getting insights about clusters by:

a. Plotting heatmaps

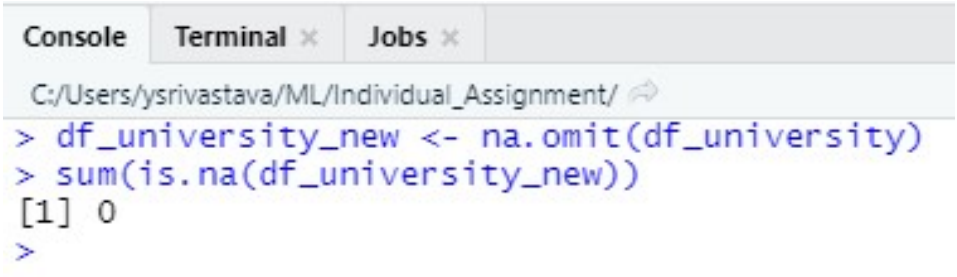
b. Parallel Coordinate plot

c. Comparing cluster statistics

10. Summarizing the categorical variables

Q1. A). Remove all records with missing measurements from the dataset.

As we can see in the code that the dataset has no missing values. This is done using omit() function.

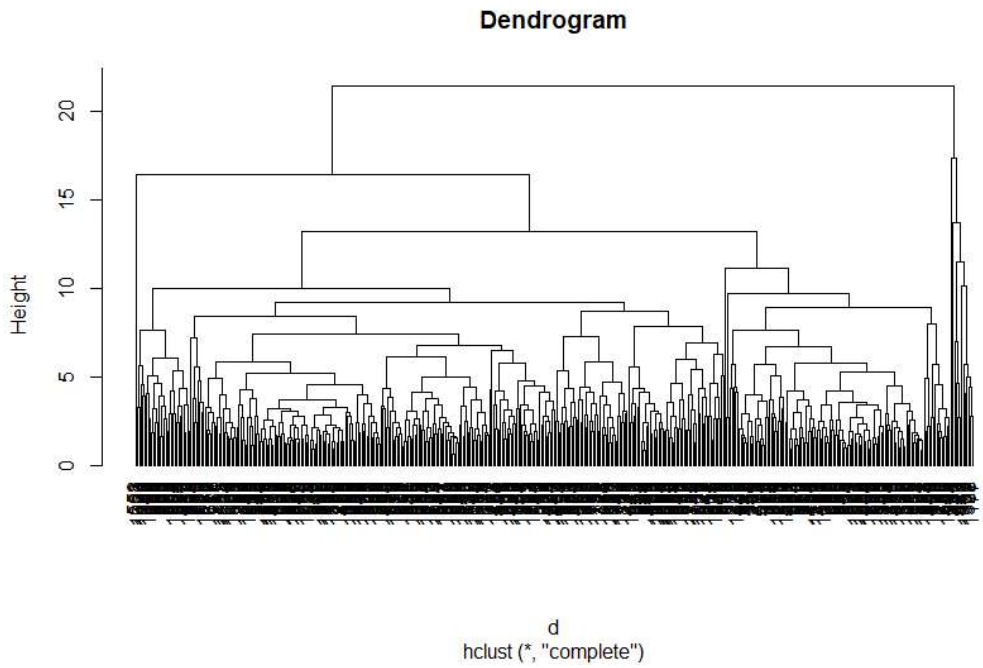


B). For all the continuous measurements, run hierarchical clustering using complete linkage and Euclidean distance. Make sure to normalize the measurements. From the dendrogram, how many clusters seem reasonable for describing these data?

Snapshot of the normalized dataframe created from continuous variables:

```
> str(z1)
'data.frame':   471 obs. of  17 variables:
 $ X..appli..rec.d          : num  -0.725 -0.737 -0.575 -0.623 0.311 ...
 $ X..appl..accepted        : num  -0.766 -0.777 -0.589 -0.616 -0.225 ...
 $ X..new.stud..enrolled    : num  -0.793 -0.755 -0.539 -0.714 -0.487 ...
 $ X..new.stud..from.top.10.: num  -0.65 -1.299 2.11 -0.109 0.108 ...
 $ X..new.stud..from.top.25.: num  -0.573 -1.557 1.592 -0.426 0.214 ...
 $ X..FT.undergrad          : num  -0.71 -0.658 -0.468 -0.648 -0.569 ...
 $ X..PT.undergrad          : num  0.0463 0.6803 -0.382 -0.4344 -0.4389 ...
 $ in.state.tuition         : num  -0.335 -1.389 0.408 -0.24 -0.678 ...
 $ out.of.state.tuition     : num  -0.699 -1.241 0.252 -0.579 -1.139 ...
 $ room                     : num  -0.843 0.411 -0.24 -1.179 -1.118 ...
 $ board                    : num  0.667 0.226 0.543 0.737 -1.027 ...
 $ add..fees                : num  -0.7 -0.97 -0.728 -0.784 0.11 ...
 $ estim..book.costs        : num  1.539 -0.299 -0.912 -0.299 2.765 ...
 $ estim..personal..        : num  0.276 -0.22 -0.604 -0.311 0.129 ...
 $ X..fac..w.PHD            : num  0.1675 -2.0526 0.0475 -0.6125 -1.0325 ...
 $ stud..fac..ratio         : num  -0.52904 -1.1446 0.00958 -0.65728 0.39431 ...
 $ Graduation.rate          : num  -2.786 -1.464 0.355 -1.188 -1.078 ...
```

After running hierarchical clustering using complete linkage and Euclidean distance, we got this dendrogram:



As seen in the dendrogram these is a lot of noise and its very difficult to interpret.

However intuitively we can see that we can CUT this into 3 to 5 clusters. 5 clusters seem reasonable, but only after testing these clusters we can decide which one id optimal.

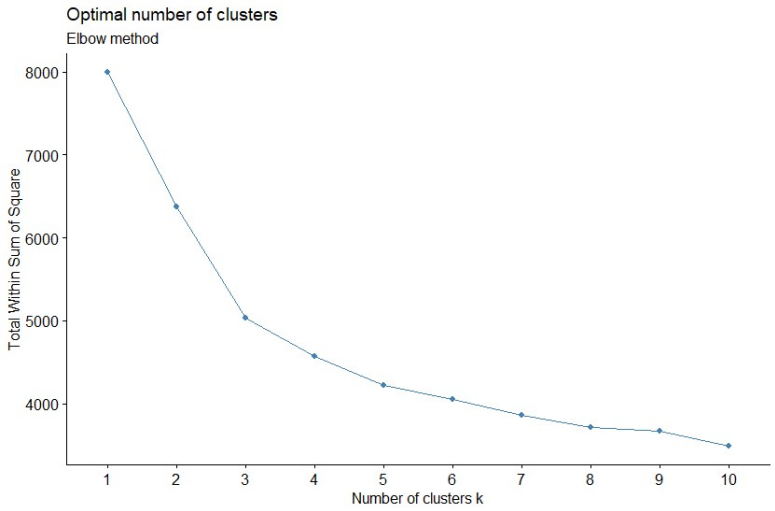
Silhouette score:

Clusters Score

3 ---------- 0.49

4 ---------- 0.49

5 ---------- 0.46



Although Silhouette score is higher for 4 clusters, but it is only through a very small margin.

5 clusters have a lower total within sum of square.

In order to check cluster stability, we took a random 95% sample from the dataset and ran the cluster analysis again. The resulting dendrogram was like the one before with complete data.

Hence 5 number of clusters is the optimal solution

C). Compare the summary statistics for each cluster and describe each cluster in this context (e.g., “Universities with high tuition, low acceptance rate…”).

In terms of fees we can clearly found out that Universities of 3rd Cluster have the highest average fees, both in-station and outstation

sub_cl3	in.state.tuiti	out.of.state.tuition
1	9476.4705	10577.7
2	7430.6364	10845.727
3	11230	11230
4	3171	8949
5	3640	7410

In terms of acceptance rate Universities belonging to 2nd cluster have the highest acceptance rate compared to others.

sub_cl3	X..appli..rec.	X..appl..acce
1	2752.1247	1793.0788
2	14991.818	10826.545
3	601	396
4	11054	6397
5	48094	26330

Universities belonging to 2nd cluster have the highest no. of Full time and part time graduation courses.

sub_cl3	X..FT.undergr	X..PT.undergrad	
1	3084.4158	684.84245	
2	20921.636	3341.5455	
3	525	323	
4	16502	21836	
5	21401	3712	

While computing the variances we found out that 6 out of 17variables explain about **97%** of the total variance.

in.state_tuition

X._FT_undergrad

out.of.state_tuition

X._appli._rec.d

X._appl._accepted

X._PT_undergrad

Clearly while selecting the variables for modeling we can drop other variables.

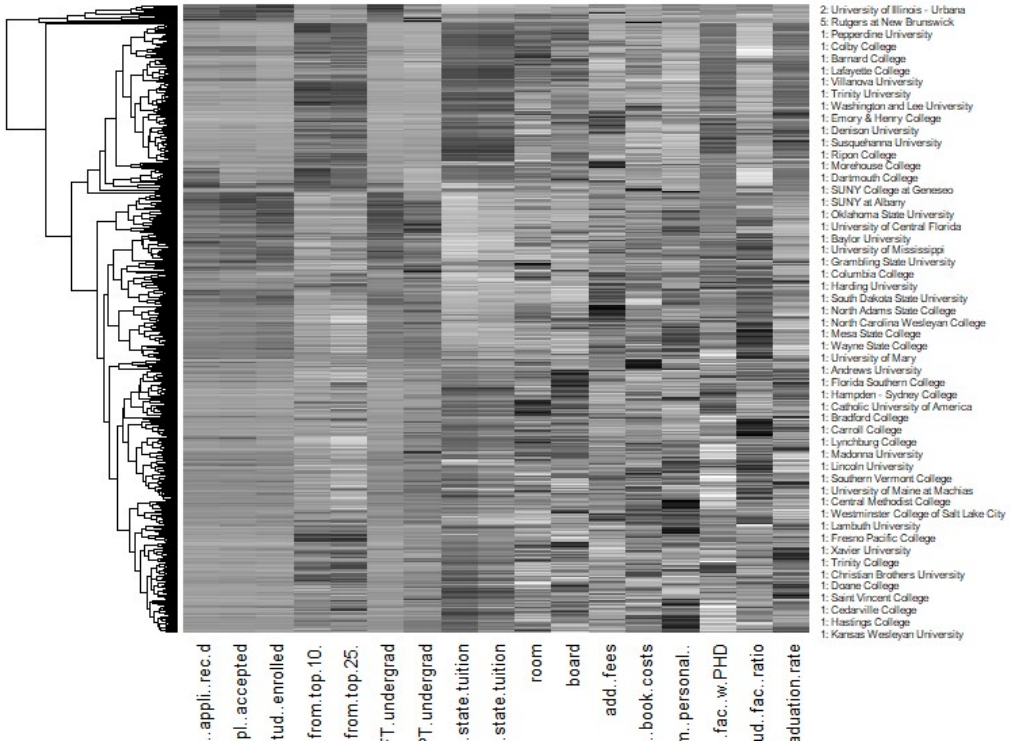
1	2	3	4	5
457	11	1	1	1

Also, while analyzing the clusters we found out that more observations are with 1st cluster.

D). Use the categorical measurements that were not used in the analysis (State and Private/Public) to characterize the different clusters. Is there any relationship between the clusters and the categorical information?

```
> table(df_university_new$College.Name, sub_cl3)
                                        sub_cl3
                                         1 2 3 4 5
  Adams State College                    1 0 0 0 0
  Adrian College                         1 0 0 0 0
  Alaska Pacific University              1 0 0 0 0
  Albertson College                      1 0 0 0 0
  Albion College                         1 0 0 0 0
  Albright College                       1 0 0 0 0
  Alderson-Broaddus College              1 0 0 0 0
  Alfred University                      1 0 0 0 0
  Allegheny College                      1 0 0 0 0
  Allentown Coll. of St. Francis de Sales 1 0 0 0 0
  Alma College                           1 0 0 0 0
  Amherst College                        1 0 0 0 0
  Anderson University                    1 0 0 0 0
  Andrews University                     1 0 0 0 0
  Angelo State University                1 0 0 0 0
  Antioch University                     1 0 0 0 0
  Appalachian State University           1 0 0 0 0
  Arkansas College (Lyon College)        1 0 0 0 0
  Ashland University                     1 0 0 0 0
  Augustana College                      2 0 0 0 0
  Baker University                       1 0 0 0 0
  Baldwin-Wallace College                1 0 0 0 0
  Barnard College                        1 0 0 0 0
  Baylor University                      1 0 0 0 0
  Bellarmine College                     1 0 0 0 0
  Belmont University                     1 0 0 0 0
  Beloit College                         1 0 0 0 0
```

Public	1	119	7	0	1	1
Private	2	338	4	1	0	0



```
> table(df_university_new$State, sub_cl3)
    sub_cl3
     1  2  3  4  5
  AK  2  0  0  0  0
  AL  4  0  0  0  0
  AR  4  0  0  0  0
  AZ  1  1  0  0  0
  CA 14  1  0  0  0
  CO  6  0  0  0  0
  CT 10  0  0  0  0
  DC  4  0  0  0  0
  DE  2  0  0  0  0
  FL  8  0  0  0  0
  GA  7  0  0  0  0
  HI  1  0  0  0  0
  IA 18  0  0  0  0
  ID  2  0  0  0  0
  IL 14  1  0  0  0
  IN 15  0  0  0  0
  KS  7  0  0  0  0
  KY  6  0  0  0  0
  LA  5  0  0  0  0
  MA 19  3  0  0  0
  MD  3  0  0  0  0
  ME  6  0  0  0  0
  MI 11  1  1  0  0
  MN 10  0  0  1  0
  MO 15  0  0  0  0
  MS  5  0  0  0  0
  MT  2  0  0  0  0
  NC 23  0  0  0  0
  ND  5  0  0  0  0
  NE  7  0  0  0  0
  NH  6  0  0  0  0
  NJ 12  0  0  0  1
  NM  2  0  0  0  0
  NY 37  1  0  0  0
  OH 24  0  0  0  0
  OK  6  0  0  0  0
  OR  5  0  0  0  0
  PA 41  1  0  0  0
  RI  4  0  0  0  0
  SC  9  0  0  0  0
  SD  4  0  0  0  0
  TN 15  0  0  0  0
  TX 18  2  0  0  0
  UT  2  0  0  0  0
  VA 15  0  0  0  0
  VT  7  0  0  0  0
  WA  2  0  0  0  0
  WI  9  0  0  0  0
  WV  2  0  0  0  0
  WY  1  0  0  0  0
>
```

From the above illustrations we have inferred some results:

Maximum no of universities is clustered around 1st cluster. Of all these universities maximum no of states is PA which has 41 observations. Also, cluster 1 has a lot of private universities.

E) What other external information can explain the contents of some or all of these clusters?

We can perform PCA to analyze the variables which are contributing to the overall variance. This method will Eliminate Co-connection Features. Improves the Algorithm Performance by lessening the no of measurements: The preparation season of the calculations diminishes fundamentally with a smaller number of highlights. Lessens over-fitting of information: Overfitting predominantly happens when there are an excessive number of factors in the dataset. Based on that we can construct new components which contains projection of these datapoints.

In terms of tuition fees and acceptance rate, we have observed a lot of variances. These can be further explored with the help of some external information such as:

1. The average income of the population living near the universities.
2. Other college features such as square feet area, quality of accommodation.
3. Transport facilities offered to students.
4. Criteria for admission. This should explain the acceptance rate.

F). Consider Tufts University, which is missing some information. Compute the Euclidean distance of this record from each of the clusters that you found above (using only the measurements that you have). Which cluster is it closest to? Impute the missing values for Tufts by taking the average of the cluster on those measurements.

To do this we have taken a subset from the data for Tufts University. While checking for missing values we found that **X..PT.undergrad** has missing value.

```
> df_university_1 <- df_university
> sub<-subset(df_university_1, College.Name =="Tufts University")
> sub
       College.Name State Public..1...Private..2. X..appli..rec.d X..appl..accepted X..new.stud..enrolled
476 Tufts University    MA                      2            7614              3605                   1205
    X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad X..PT.undergrad in.state.tuition
476                        60                        90            4598              NA             19701
    out.of.state.tuition room board add..fees estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio
476                19701 3038  2930       503               600               928             99              10.3
    Graduation.rate
476              92
> sum(is.na(sub))
[1] 1
> sapply(sub, function(x) sum(is.na(x)))
            College.Name                    State   Public..1...Private..2.           X..appli..rec.d
                       0                        0                        0                         0
       X..appl..accepted     X..new.stud..enrolled X..new.stud..from.top.10. X..new.stud..from.top.25.
                       0                        0                        0                         0
         X..FT.undergrad          X..PT.undergrad        in.state.tuition      out.of.state.tuition
                       0                        1                        0                         0
                    room                    board             add..fees         estim..book.costs
                       0                        0                        0                         0
       estim..personal..          X..fac..w.PHD          stud..fac..ratio           Graduation.rate
                       0                        0                        0                         0
```

To impute this we have to calculate Euclidean distance of these datapoints from the centroid that we found below:

	sub_cl3	X..appli..re	X..appl..ac	X..new.stud	X..new.stud	X..new.stud	X..FT.under	X..PT.under	in.state.tui	out.of.state	room	board	add..fees	estim..boo	estim..pers	X..fac..w.PH	stud..fac..ra	Graduation	e_dis
1	1	2752.1247	1793.0788	684.50328	27.868709	55.347921	3084.4158	684.84245	9476.4705	10577.7	2210.6258	2112.4245	370.19694	543.84464	1295.9081	73.004376	13.920788	65.628009	14790.912
2	2	14991.818	10826.545	4240.6364	35.909091	69.454545	20921.636	3341.5455	7430.6364	10845.727	2524.0909	2456.7273	708.36364	563.27273	1832.2727	84.727273	16.009091	65.363636	24750.516
3	3	601	396	203	1	20	525	323	11230	11230	3843	2800	130	2340	620	8	6.8	47	14984.461
4	4	11054	6397	3524	26	55	16502	21836	3171	8949	1498	2246	414	714	2910	88	12.2	45	23713.955
5	5	48094	26330	4520	36	79	21401	3712	3640	7410	2780	1986	1003	690	2009	90	19.5	77	53477.06

After comparing the distances, we found that cluster 1 had least distance from the record. Using that cluster we took the average of X..PT.undergrad and got the value of 684.8. This value was imputed in the Tufts University record.

```
> sub[is.na(sub)] = value
> sub
       College.Name State Public..1...Private..2. X..appli..rec.d X..appl..accepted X..new.stud..enrolled
476 Tufts University    MA                      2            7614              3605                   1205
    X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad X..PT.undergrad in.state.tuition
476                        60                        90            4598          684.8425             19701
    out.of.state.tuition room board add..fees estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio
476                19701 3038  2930       503               600               928             99              10.3
    Graduation.rate
476              92
```

Q2). A) Identify the categorical variables.

Id
Model
Price
Age..month.
Mfg_Month
Mfg_Year
KM
Fuel_Type
HP
Met_Color
Color
Automatic
CC
Doors
Cylinders
Gears
Quarterly_Tax
Weight
Mfr_Guarantee
Dealer_Guarantee
Guarantee_Period
ABS
Airbag_1
Airbag_2
Aircond
Automatic_aircond
Boardcomputer
CD_Player
Central_Lock
Powered_Windows
Power_Steering
Radio
Mistlamps
Sport_Model
Backseat_Divider
Metallic_Rim
Radio_cassette
Parking_Assistant
Tow_Bar

b) Explain the relationship between a categorical variable and the series of binary dummy variables derived from it.

Dummy variables also known as indicator variables represent 2 more categories of an attribute. Let us say a variable has N categories, then we can create either N or N-1 dummy variables. Each dummy variable will tell weather category is present or not.

Eg: the variable Gears has 2 categories 5 and 6. So its dummy variable will be Gears_5, Gears_6 and both will contain only 0 and 1 representing either of the category present.

c) How many dummy binary variables are required to capture the information in a categorical variable with N categories?

For the categorical variables with N categories we can create N or N-1 dummy variables.

For variables with only 2 categories we can create N-1 dummy variable which will contain the information of the variable.

Some some cases creating N variables will result in redundant information for the Nth variable.

Eg: for this variable "Aircond", the car either has it or it doesnt so this variable can be expressed as 0 and 1.

d). Use R to convert the categorical variables in this dataset into dummy variables, and explain in words, for one record, the values in the derived binary dummies.

To achieve this, we have used the package called Fastdummies

So, we have passed all categorical variables where not binary and converted them into dummy variables.

Let us take an example of **Fuel_Type:**

Earlier this variable had 3 categories as CNG, Diesel and Petrol. After we converted this into a dummy variable these categories got split into multiple variables. So now we have 3 variables, one for each category.

"Fuel_Type_CNG" "Fuel_Type_Diesel" "Fuel_Type_Petrol"

Each variable is binary, meaning that they can possess only 2 values, either 0 or 1.

	Id	Model	Price	Fuel_Type_CNG	Fuel_Type_Diesel	Fuel_Type_Petrol
1	1	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13500	0	1	0
2	2	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13750	0	1	0
3	3	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13950	0	1	0
4	4	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	14950	0	1	0
5	5	TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors	13750	0	1	0
6	6	TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors	12950	0	1	0

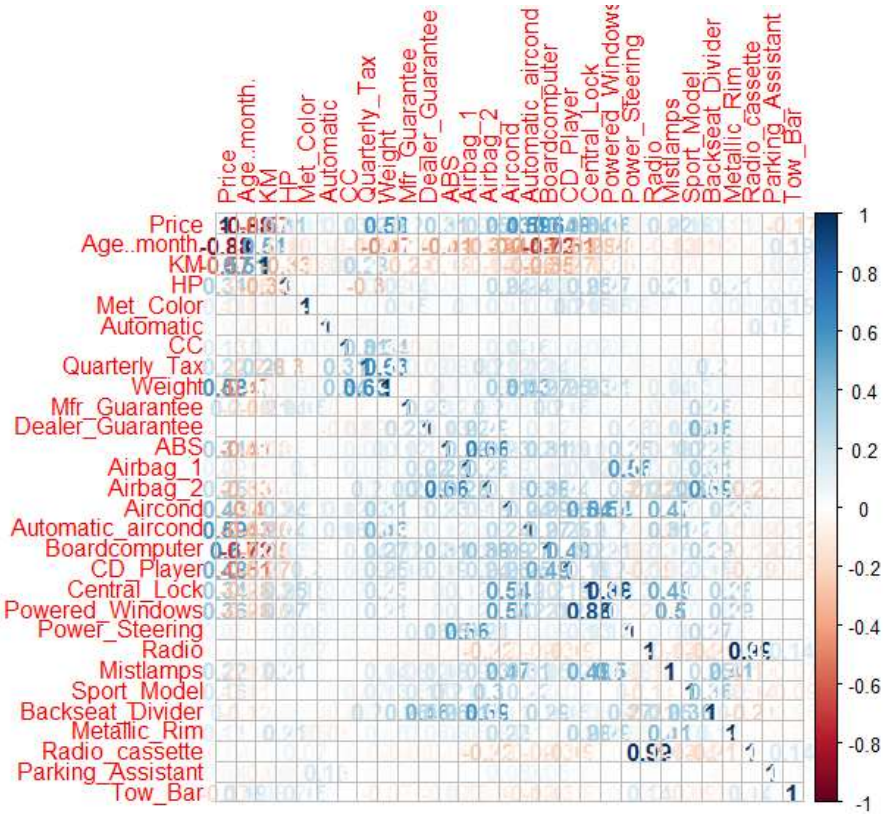
As we can see from the record these 6 observations have diesel as a fuel type hence the value 1.

e.) Correlation Matrix:

	Price	Age..month.	KM	HP	CC	Quarterly_	Weight
Price	1	-0.8765905	-0.56996	0.31499	0.126389	0.219197	0.581198
Age..month.	-0.87659	1	0.505672	-0.15662	-0.09808	-0.19843	-0.47025
KM	-0.56996	0.50567218	1	-0.33354	0.102683	0.278165	-0.0286
HP	0.31499	-0.15662202	-0.33354	1	0.035856	-0.29843	0.089614
CC	0.126389	-0.09808374	0.102683	0.035856	1	0.306996	0.335637
Quarterly_Tax	0.219197	-0.19843051	0.278165	-0.29843	0.306996	1	0.626134
Weight	0.581198	-0.47025318	-0.0286	0.089614	0.335637	0.626134	1

Inference:

- Age is adversely related with cost (- 0.88): This implies when the age of the vehicle is more the cost of the vehicle goes down.
- KM is contrarily related with cost(- 0.57): When the km builds like age the cost of the vehicle diminishes.
- Weight is decidedly corresponded with price(0.58): Weight of the Car is emphatically associated with the Price (more exorbitant cost higher weight)
- KM is emphatically corresponded with Age_08_04(0.51): Other than cost, KM and Age of Car is decidedly expressing an expanding relationship for both the elements when one increments.
- Weight is emphatically corresponded with quarterly tax(0.63) : Quarterly Road Tax gathered is decidedly identified with Weight expressing higher Car weight higher the Quarterly street charge.



Different perceptions helpful for acquiring result:

- Radio and Radio Cassette had a high relationship of 0.99 which caused me in dropping one of the variable.
- CNG faker variable was taken out as just 17 out of 1436 vehicle were shoppers of CNG.
- The quantity of Cylinders(4) were same across the dataset so was dropped while anticipating costs.
- Mfg_Month and Mfg_Year are absolute factors however were making a ton of faker factors that is 12 for quite a long time alone and numerous others for year. Thus, Age in long stretches of Car was taken to try not to cover data as it just brought about insignificant measure of loss of data.

Q3) a.) How many starting configurations are there?

Since we have been given 6 points and we need to partition them into 3 clusters, so the possible number of combinations become

nCr -> n= 6 and r=3 so …..total 20 possible combination we can achieve.

b.) What are the stable 3-partitions?

	Combination	Total_Iterations	C1	C2	C3
1	1	1	(0,3)	(8,3)	(16,3)
2	2	2	(4,0)	(16,3)	(4,6)
3	3	1	(0,3)	(12,0)	(12,6)
4	4	1	(0,3)	(8,3)	(16,3)
5	5	1	(4,0)	(16,3)	(4,6)
6	6	1	(0,3)	(16,3)	(8,3)
7	7	2	(0,3)	(12,0)	(12,6)
8	8	2	(4,0)	(4,6)	(16,3)
9	9	1	(4,0)	(4,6)	(16,3)
10	10	1	(0,3)	(8,3)	(16,3)
11	11	1	(8,3)	(16,3)	(0,3)
12	12	1	(4,0)	(16,3)	(4,6)
13	13	2	(0,3)	(8,3)	(16,3)
14	14	1	(12,0)	(0,3)	(12,6)
15	15	1	(8,3)	(0,3)	(16,3)
16	16	1	(4,0)	(4,6)	(16,3)
17	17	1	(16,3)	(0,3)	(8,3)
18	18	2	(8,3)	(0,3)	(16,3)
19	19	2	(8,3)	(0,3)	(16,3)
20	20	1	(0,3)	(8,3)	(16,3)

From the table we can find out that the 3stable partitions:

1. (0,3) (8,3) (16,3)
2. (4,0) (16,3) (4,6)
3. (0,3) (12,0) (12,6)

c) What is the number of starting configurations leading to each of the stable 3-partitions in (b) above?

This can be simply checked by viewing the table to see how many times these values are getting repeated.

(0,3) (8,3) (16,3) - 11 configurations
(4,0) (16,3) (4,6) - 6 configurations
(0,3) (12,0) (12,6) - 3 configurations

d). What is the maximum number of iterations from any starting configuration to its stable 3-partition?
From the table we checked the iteration column and found maximum number of iterations from any starting configuration to its stable 3-partition was 2.