ClusterAnalysis

Part 1

Load the mtcars dataset and check the mtcars

```
data(mtcars)
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Build kmeans object with first 3 columns,set cluster number equal to 3

```
mtcars_3<-mtcars[,1:3]
mtcars_k3<-kmeans(mtcars_3,centers=3)
```

Check the size of each cluster

```
mtcars_k3$size
```

```
## [1]  9 14  9
```

Check the average disp,wt and qsec of each cluster

```
mtcars_copy<-mtcars
mtcars_copy$cluster_id<-mtcars_k3$cluster
mtcars_cluster_mean<-setNames(aggregate(cbind(mtcars_copy$disp,mtcars_copy$wt,mtcars_copy$qsec),by=list
mtcars_cluster_mean
```

```
##   Cluster      disp       wt     qsec
## 1       1 174.52222 3.128889 18.74889
## 2       2 353.10000 3.999214 16.77214
## 3       3  96.55556 2.089222 18.62333
```

We can see each cluster have distinct mean value,Cluster 3 have the lowest mean values of disp,wt and highest values of qsec. Cluster 2 have highest value of disp and wt,and lowest qsec value.

Part 2

Lets use dummy hotel customer data to perform cluster analysis. For this dataset, we focus on following columns: Spend:How much money customer spend per year. Status:the membership of hotel:bronze,silver,gold,plantinum Stays.Per.Year:How many stays for each year Total.Days.Stayed:How many days customer stayed at hotel for each year Years.of.Loyalty:How long customers have been the membership of hotel

```
hotel<-read.csv("c:/users/Kun Hu/Desktop/hotelloyaltydata.csv")
str(hotel)
```

```
## 'data.frame':    2276 obs. of  12 variables:
##  $ Customer.Key     : int  1193 699 2491 2107 308 2882 3079 1999 272 1723 ...
##  $ First.Name       : Factor w/ 925 levels "A","AARYN","ABBASHER",..: 219 902 543 22 704 186 907 650
##  $ Last.Name        : Factor w/ 1790 levels "ABDELKADER","ABDELLA",..: 1205 478 1710 981 1701 463 55
##  $ Customer.Segment : Factor w/ 20 levels "A","B","C","D",..: 3 4 17 10 6 12 17 15 3 20 ...
##  $ Income           : Factor w/ 13 levels "A","B","C","D",..: 8 9 4 1 1 5 2 5 9 13 ...
##  $ Reedemer         : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 2 2 ...
##  $ Region           : int  1 1 10 5 1 7 10 10 6 4 ...
##  $ Spend            : num  26573 36711 46008 68501 75182 ...
##  $ Status           : Factor w/ 4 levels "Bronze","Gold",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Stays.Per.Year   : int  8 2 2 2 3 2 8 7 4 1 ...
##  $ Total.Days.Stayed: int  6 23 6 12 14 9 66 12 12 1 ...
##  $ Years.Of.Loyalty : num  0.75 11.25 2.75 6 4.75 ...
```

Use Spend,Stays.per.Year,Total.Days.Stayed,Year.of.Loyalty columns,use k-means functions, and cluster number set to 4.

```
hotel_sub<-hotel[,c(8,10,11,12)]
hotel_sub<-sapply(hotel_sub,as.numeric)
hotel_k3<-kmeans(hotel_sub,centers=4)
```

Chekc centers and size of each cluster

```
hotel_k3$centers
```

```
##       Spend Stays.Per.Year Total.Days.Stayed Years.Of.Loyalty
## 1  7829.405       4.172794          12.63971         2.995404
## 2 60294.923       3.818182          19.36364         4.795455
## 3 24069.725       4.177778          12.62222         3.050000
## 4  1250.636       4.032341          11.82803         2.940965
```

```
hotel_k3$size
```

```
## [1]  272   11   45 1948
```

From above centers and size result, we can see cluster 2 spend the most money but only have 11 size, , also we notice the years.of.loyalty and total days stayed are highest among 4 clusters, which we can define cluster as high-end customer.For cluster 1,3,4, the only significant difference is Spend category. Check each cluster to each level of status:bronze,silver,gold,plantinum

```
hotel$cluster_id<-hotel_k3$cluster
table(hotel$cluster_id,hotel$Status)
```

```
##
##     Bronze Gold Platinum Silver
##   1     12   40       18    202
##   2      0    0       10      1
##   3      1    2       28     14
```

```
##   4    740  504        8    696
```

From above table, we can see that platinum status customer tend to spend more money at hotel than other three status.