Use the clustering methodology to segment customers into groups:
Use the following clustering algorithms:

1. K means
2. Hierarchical

• Identify the right number of customer segments.
• Provide the number of customers who are highly valued.
• Identify the clustering algorithm that gives maximum accuracy and explains robust clusters.
• If the number of observations is loaded in one of the clusters, break down that cluster further using the clustering algorithm. [hint: Here loaded means if any cluster has more number of data points as compared to other clusters then split that clusters by increasing the number of clusters and observe, compare the results with previous results.]

Here's the dataset: https://github.com/Simplilearn-Edu/Data-Science-with-R.git

Solution:

Since we want customers with high price purchases we calculate sum(Unit Price* Quantity) for each customer

```
library(dplyr)
library(cluster)

sale_info<-read.csv("D:\\Data_Analysis_Simplilearn_materials\\Data_Science_with_R\\Ecommerce.csv")

customer_sales<-sale_info%>%group_by(CustomerID)%>%summarise(total=sum(UnitPrice*Quantity))
customer_sales<-na.omit(customer_sales)
head(customer_sales)
```

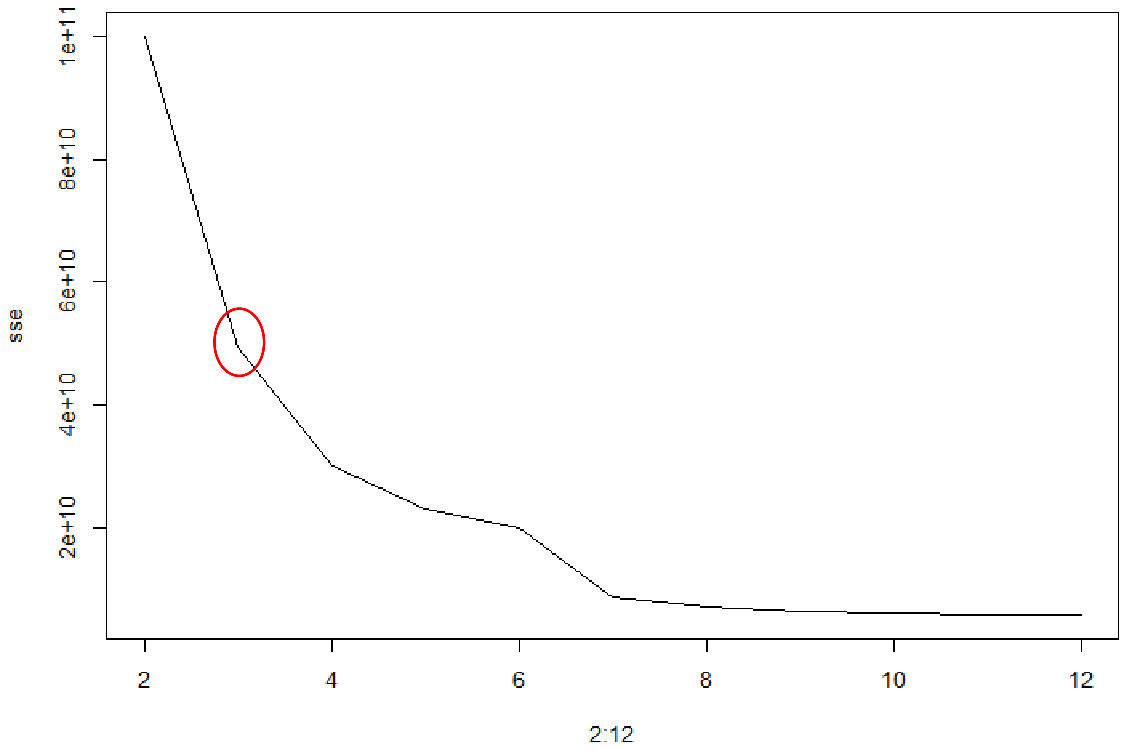
```
> library(dplyr)
> library(cluster)
>
> sale_info<-read.csv("D:\\Data_Analysis_Simplilearn_materials\\Data_Science_with_R\\Ecommerce.csv")
>
> customer_sales<-sale_info%>%group_by(CustomerID)%>%summarise(total=sum(UnitPrice*Quantity))
> customer_sales<-na.omit(customer_sales)
> head(customer_sales)
# A tibble: 6 x 2
  CustomerID total
       <int> <dbl>
1      12346     0
2      12347  4310
3      12348 1797.
4      12349 1758.
5      12350  334.
6      12352 1545.
>
```

K-Means Clustering

Now we perform K-means clustering on sum(Price*Quantity)

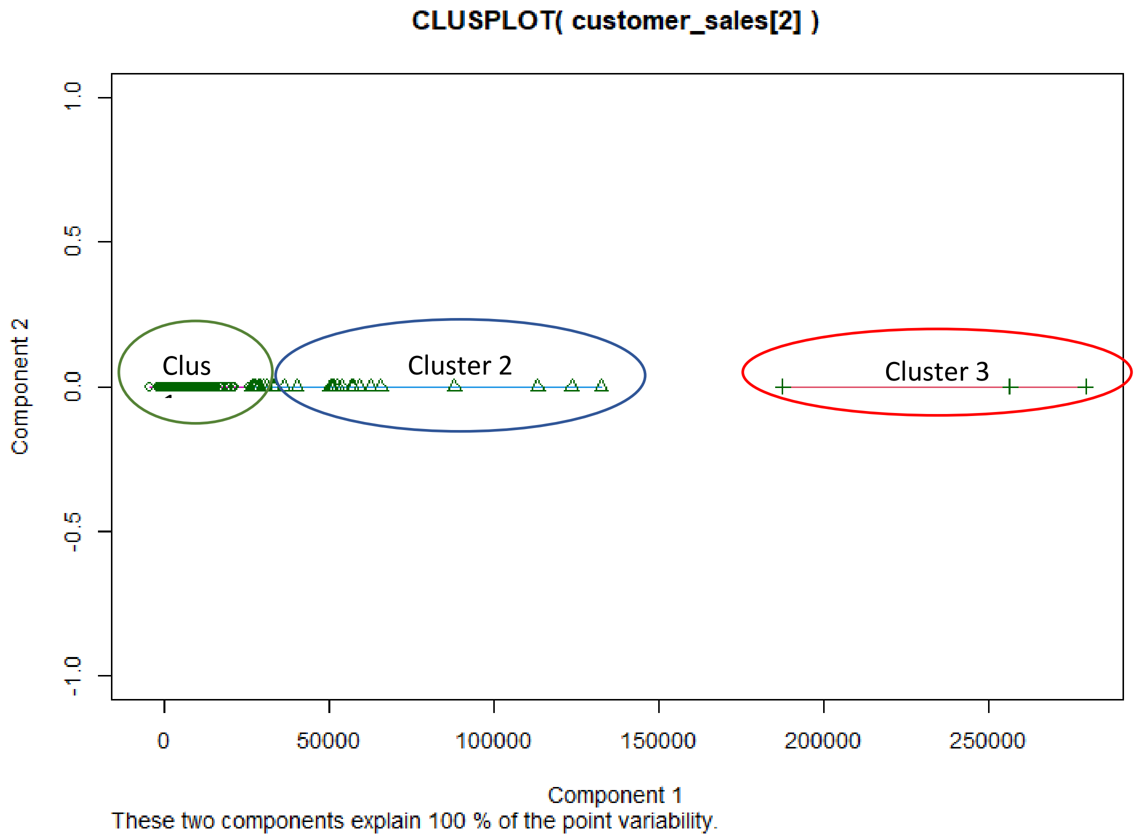
Let’s find suitable number of clusters

```r
#K-means

set.seed(101)

for(i in 2:12){
  temp.kmean<-kmeans(customer_sales$total,centers=i,nstart=10)
  cat('For ',i, ' clusters total SSE ',temp.kmean$tot.withinss,'\n')
  if(i==2){
    sse<-c(temp.kmean$tot.withinss)
  }
  else{
    sse<-c(sse,c(temp.kmean$tot.withinss))
  }
}

plot(2:12,sse,type='l')
```



Elbow bend is observed at K=3. We choose K=3.

```
K=3;
cust.kmean<-kmeans(customer_sales$total,centers=K,nstart=10)
c_sale_kmean<-customer_sales
clusplot(customer_sales[2],cust.kmean$cluster,color = T,shade=T,labels=0,lines=0)
c_sale_kmean<-cbind(c_sale_kmean,k_mean_cluster=cust.kmean$cluster)
View(c_sale_kmean)
View(customer_sales)
c_sale_kmean%>%filter(k_mean_cluster==2)
c_sale_kmean%>%filter(k_mean_cluster==3)
head(c_sale_kmean%>%filter(k_mean_cluster==1))
```



Cluster 3 has very high value customers, Cluster 2 has reasonably high value customers and cluster 1 has low value customers. We are interested in Cluster 3 and Cluster 2

Cluster 3: Three premium level customers

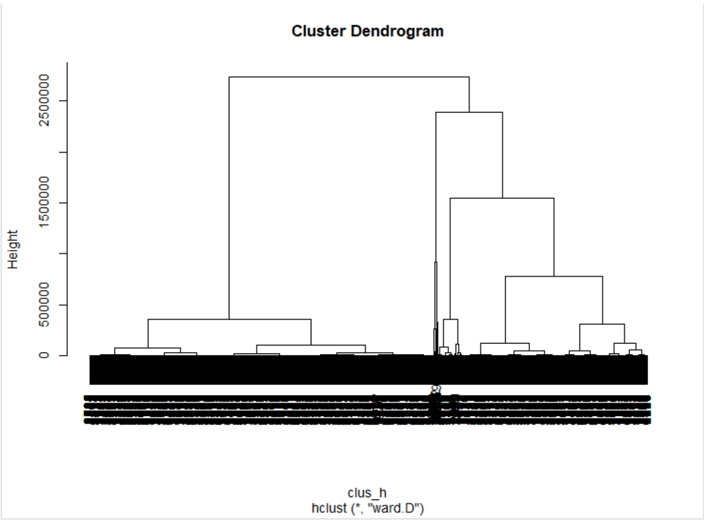
```
> c_sale_kmean%>%filter(k_mean_cluster==3)
  CustomerID    total k_mean_cluster
1      14646 279489.0              3
2      17450 187482.2              3
3      18102 256438.5              3
```

Cluster 2: Reasonably high value customers (30 in number)

```
> c_sale_kmean%>%filter(k_mean_cluster==2)
   CustomerID      total k_mean_cluster
1       12415 123725.45              2
2       12748  29072.10              2
3       12931  33462.81              2
4       13081  27964.48              2
5       13089  57385.88              2
6       13098  28658.88              2
7       13408  27487.41              2
8       13694  62653.10              2
9       13798  36351.42              2
10      14088  50415.49              2
11      14096  57120.91              2
12      14156 113384.14              2
13      14298  50862.44              2
14      14680  26932.34              2
15      14911 132572.62              2
16      15061  54228.74              2
17      15311  59419.34              2
18      15769  51823.72              2
19      15838  33350.76              2
20      16013  33366.25              2
21      16029  50992.61              2
22      16333  26626.80              2
23      16422  33805.69              2
24      16684  65892.08              2
25      17389  31300.08              2
26      17404  30300.82              2
27      17511  88125.38              2
28      17841  40340.78              2
29      17857  26763.34              2
30      17949  52750.84              2
```

Hierarchical Clustering

```
# Hierarchy
clus_h<-dist(customer_sales$total,'euclidian')
fitted_hier<-hclust(clus_h,method='ward')
c_sale_hier<-customer_sales

c_sale_hier<-cbind(c_sale_hier,hier_clus=cutree(fitted_hier,3))
View(c_sale_hier)

plot(fitted_hier)
head(c_sale_hier%>%filter(hier_clus==2))
c_sale_hier%>%filter(hier_clus==3)
head(c_sale_hier%>%filter(hier_clus==1))
```



Here cluster 3 are valued customers while cluster 1 and cluster 2 are low valued

```
> c_sale_hier%>%filter(hier_clus==3)
   CustomerID     total hier_clus
1       12415 123725.45         3
2       12748  29072.10         3
3       12931  33462.81         3
4       13081  27964.48         3
5       13089  57385.88         3
6       13098  28658.88         3
7       13408  27487.41         3
8       13694  62653.10         3
9       13777  25748.35         3
10      13798  36351.42         3
11      14088  50415.49         3
12      14096  57120.91         3
13      14156 113384.14         3
14      14298  50862.44         3
15      14646 279489.02         3
16      14680  26932.34         3
17      14911 132572.62         3
18      15061  54228.74         3
19      15311  59419.34         3
20      15769  51823.72         3
21      15838  33350.76         3
22      16013  33366.25         3
23      16029  50992.61         3
24      16333  26626.80         3
25      16422  33805.69         3
26      16684  65892.08         3
27      17389  31300.08         3
28      17404  30300.82         3
29      17450 187482.17         3
30      17511  88125.38         3
31      17841  40340.78         3
32      17857  26763.34         3
33      17949  52750.84         3
34      18102 256438.49         3
```