## Code::

```
#MachineLearning Lab1 (9 Jan 2019)
#Author:"Saurabh Kumar Singh"


data=read.csv("C:/Users/imsau/Desktop/6th Sem/ML/ML_lab/Lab1(9-Jan)/Wholesale
customers data.csv")
View(data)

#3.   Observe the data ( in R-through summary and str)

summary(data)



#4.   We'll need to drop the Channel and Region variables.
    #These are two ID fields and are not useful in clustering. So drop it.

cols.dont.want <- c("Channel", "Region")
data1 <- data[, ! names(data) %in% cols.dont.want, drop = F]

#5.Set some SEED value
#6.Apply the k-mean on dataset, with k=5


set.seed(0)
km<-kmeans(data1, 5)



#11.  Measure total SSE
km$totss

#10.  Measure homogeneity of each cluster (SSE)
km$withinss

km$tot.withinss
#12.  Measure the heterogeneity of cluster
km$betweenss

km$iter

#13.  Elbow measure: run the algorithm 100 time for k=2 to 20.
wss = kmeans(data1, centers=1)$tot.withinss
wss
```

```
for (i in 2:20)
  wss[i] = kmeans(data1, centers=i)$tot.withinss

#elbow plot
library(ggvis)
sse = data.frame(c(1:20), c(wss))
names(sse)[1] = 'Clusters'
names(sse)[2] = 'SSE'
sse %>%
  ggvis(~Clusters, ~SSE) %>%
  layer_points(fill := 'blue') %>%
  layer_lines()


#Plots
```

**install.packages("animation")**

```
library(animation)
kmeans.ani(data1, 5)



library(cluster)
library(fpc)
km$cluster
clusplot(data1, km$cluster, color=T, shade=F,labels=0,lines=0, main='k-Means
Cluster Analysis')
```

**Result:**

**> data=read.csv("C:/Users/imsau/Desktop/6th Sem/ML/ML_lab/Lab1(9-Jan)/Wholesale customers data.csv")**
**> View(data)**

> #3. Observe the data ( in R-through summary and str)

**> summary(data)**

```
   Channel         Region         Fresh            Milk
 Min.   :1.000   Min.   :1.000   Min.   :    3   Min.   :   55
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:  3128   1st Qu.: 1533
 Median :1.000   Median :3.000   Median :  8504   Median : 3627
 Mean   :1.323   Mean   :2.543   Mean   : 12000   Mean   : 5796
```

```
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 16934   3rd Qu.: 7190
 Max.   :2.000   Max.   :3.000   Max.   :112151   Max.   :73498
   Grocery          Frozen        Detergents_Paper    Delicassen
 Min.   :    3   Min.   :   25.0   Min.   :    3.0   Min.   :    3.0
 1st Qu.: 2153   1st Qu.:  742.2   1st Qu.:  256.8   1st Qu.:  408.2
 Median : 4756   Median : 1526.0   Median :  816.5   Median :  965.5
 Mean   : 7951   Mean   : 3071.9   Mean   : 2881.5   Mean   : 1524.9
 3rd Qu.:10656   3rd Qu.: 3554.2   3rd Qu.: 3922.0   3rd Qu.: 1820.2
 Max.   :92780   Max.   :60869.0   Max.   :40827.0   Max.   :47943.0
```

#4. We'll need to drop the Channel and Region variables. These are two ID fields and are not useful in clustering. So drop it.

```
> cols.dont.want <- c("Channel", "Region")
> data1 <- data[, ! names(data) %in% cols.dont.want, drop = F]
```

#5.Set some SEED value
#6.Apply the k-mean on dataset, with k=5

```
> set.seed(0)
> km<-kmeans(data1, 5)
```

#11. Measure total SSE

```
> km$totss
[1] 157595857166
```

#10. Measure homogeneity of each cluster (SSE)

```
> km$withinss
[1]  8521349738 14108802241 11679101316  8835879467 10060038988
```

```
> km$tot.withinss
[1] 53205171749
```

#12. Measure the heterogeneity of cluster

```
> km$betweenss
[1] 104390685416
```

```
> km$iter
[1] 4
```

#13. Elbow measure: run the algorithm 100 time for k=2 to 20.

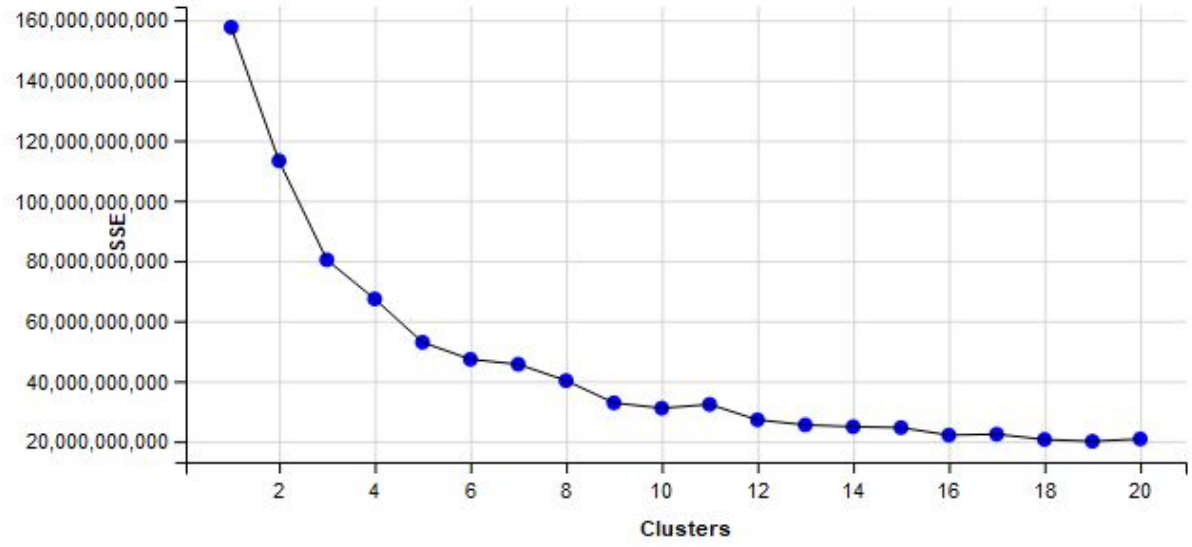
```
wss = kmeans(data1, centers=1)$tot.withinss
wss
```

[1] 157595857166

```
for (i in 2:20)
  wss[i] = kmeans(data1, centers=i)$tot.withinss
```

#elbow plot

```
library(ggvis)
sse = data.frame(c(1:20), c(wss))
names(sse)[1] = 'Clusters'
names(sse)[2] = 'SSE'
sse %>%
  ggvis(~Clusters, ~SSE) %>%
  layer_points(fill := 'blue') %>%
  layer_lines()
```




#Plots

**>install.packages("animation")**
**> library(animation)**
**> kmeans.ani(data1, 5)**
This will result an Animation of Clutering Iterationwise

**>library(cluster)**
**>library(fpc)**
**>km$cluster**
**>clusplot(data1, km$cluster, color=T, shade=F,labels=0,lines=0, main='k-Means Cluster Analysis')**