Linear Regression

In [1]:
```
#Library
library(caTools)
```

```
Warning message:
"package 'caTools' was built under R version 3.6.3"
```

In [2]:
```
#Dataset
data=read.csv("USA_Housing.csv")
```

In [3]:
```
#Creating training and testing
#70% to train
#30% to testing
s=sample.split(data,SplitRatio=0.7)
s
#we wonot get splited data
```

```
FALSE  TRUE  TRUE  TRUE  FALSE  TRUE  FALSE
```

In [4]:
```
#To get train and test dataset
train=subset(data,train=TRUE) #Giving 70 to train
test=subset(data,train=FALSE)
##train=subset(data,split=TRUE) #Giving 70 to train
##test=subset(data,split=FALSE)
```

In [5]:
```
res=lm(Price~.,data=train)
```

In [6]:
```
#Predict
res1=predict(res,test)
```

```
Warning message in predict.lm(res, test):
"prediction from a rank-deficient fit may be misleading"
```

In [7]:
```
df1=data.frame(data["Price"],res1)
df1
```

Price	res1
1059033.6	1059033.6
1505890.9	1505890.9
1058988.0	1058988.0
1260616.8	1260616.8
630943.5	630943.5
1068138.1	1068138.1
1502055.8	1502055.8
1573936.6	1573936.6
798869.5	798869.5
1545154.8	1545154.8
1707045.7	1707045.7
663732.4	663732.4
1042814.1	1042814.1
1291331.5	1291331.5
1402818.2	1402818.2
1306674.7	1306674.7
1556786.6	1556786.6
528485.2	528485.2
1019425.9	1019425.9
1030591.4	1030591.4
2146925.3	2146925.3
929247.6	929247.6
718887.2	718887.2
743999.8	743999.8
895737.1	895737.1
1453974.5	1453974.5
1125692.5	1125692.5
975429.5	975429.5
1240763.8	1240763.8
1577017.8	1577017.8
...	...
1120943.3	1120943.3
1111307.1	1111307.1
1736401.6	1736401.6

Price	res1
1340769.8	1340769.8
801348.6	801348.6
1324382.2	1324382.2
1340343.9	1340343.9
1518478.0	1518478.0
1910585.1	1910585.1
1823498.4	1823498.4
1406865.5	1406865.5
1203850.1	1203850.1
1020095.9	1020095.9
1194357.4	1194357.4
1211899.7	1211899.7
1378937.9	1378937.9
1260241.4	1260241.4
1197073.4	1197073.4
1275143.2	1275143.2
885205.0	885205.0
479500.6	479500.6
1263720.5	1263720.5
1568700.6	1568700.6
1381830.8	1381830.8
905354.9	905354.9
1060193.8	1060193.8
1482617.7	1482617.7
1030729.6	1030729.6
1198656.9	1198656.9
1298950.5	1298950.5

Logistic Regression

In [9]:
```
df=read.csv("Train.csv")
```

In [10]:
```
#70 percent to train and 30 test
split=sample.split(df,SplitRatio=0.7)
train=subset(df,split=TRUE)
test=subset(df,split=FALSE)
```

In [12]:
```
#model
model=glm(Reached.on.Time_Y.N~.,train,family="binomial")
```

In [13]:
```
p=predict(model,test,type="response")
res=data.frame(df["Reached.on.Time_Y.N"],round(p))
res
```

Reached.on.Time_Y.N	round.p.
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
...	...
0	0
0	0

Reached.on.Time_Y.N	round.p.
0	0
1	1
0	0
0	0
0	1
0	1
0	0
0	0
1	0
1	0
0	1
1	0
1	0
1	0
1	1
1	0
1	0
0	0
0	0
0	0
1	0
0	0
0	0
1	0
0	0
0	0
0	0
0	0

K Means Clustering

In [14]:
```
# library
library(factoextra)
library(cluster)
library(ggplot2)
```

```
Warning message:
"package 'factoextra' was built under R version 3.6.3"Loading required pack
age: ggplot2
Welcome! Want to learn more? See two factoextra-related books at https://go
o.gl/ve3WBa (https://goo.gl/ve3WBa)
Warning message:
"package 'cluster' was built under R version 3.6.3"
```

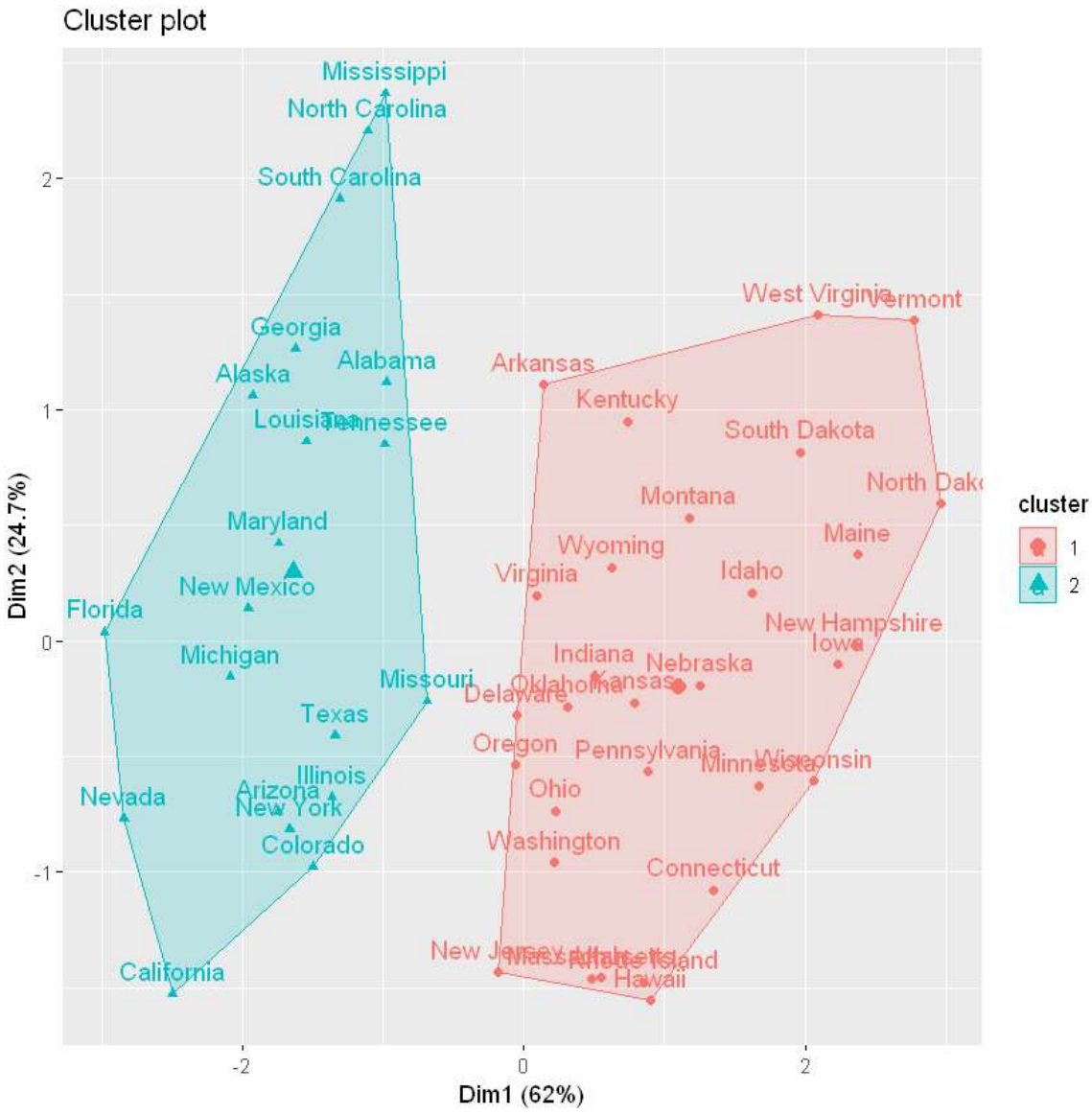
In [16]:
```
#Group the clusters
df=scale(USArrests)
head(df)
```

	Murder	Assault	UrbanPop	Rape
Alabama	1.24256408	0.7828393	-0.5209066	-0.003416473
Alaska	0.50786248	1.1068225	-1.2117642	2.484202941
Arizona	0.07163341	1.4788032	0.9989801	1.042878388
Arkansas	0.23234938	0.2308680	-1.0735927	-0.184916602
California	0.27826823	1.2628144	1.7589234	2.067820292
Colorado	0.02571456	0.3988593	0.8608085	1.864967207

In [17]:
```
#Cluster
# 2->How many clusters
#35->Random 25 values
Kclusters=kmeans(df,2,nstart=25)
Kclusters
```

```
K-means clustering with 2 clusters of sizes 30, 20

Cluster means:
     Murder    Assault   UrbanPop       Rape
1 -0.669956 -0.6758849 -0.1317235 -0.5646433
2  1.004934  1.0138274  0.1975853  0.8469650

Clustering vector:
       Alabama         Alaska        Arizona       Arkansas     California
             2              2              2              1              2
      Colorado    Connecticut       Delaware        Florida        Georgia
             2              1              1              2              2
        Hawaii          Idaho       Illinois        Indiana           Iowa
             1              1              2              1              1
        Kansas       Kentucky      Louisiana          Maine       Maryland
             1              1              2              1              2
 Massachusetts       Michigan      Minnesota    Mississippi       Missouri
             1              2              1              2              2
       Montana       Nebraska         Nevada  New Hampshire     New Jersey
             1              1              2              1              1
    New Mexico       New York North Carolina   North Dakota           Ohio
             2              2              2              1              1
      Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
             1              1              1              1              2
  South Dakota      Tennessee          Texas           Utah        Vermont
             1              2              2              1              1
      Virginia     Washington  West Virginia      Wisconsin        Wyoming
             1              1              1              1              1

Within cluster sum of squares by cluster:
[1] 56.11445 46.74796
 (between_SS / total_SS =  47.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.within
ss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

In [18]:
```
#Visualizing the cluster
fviz_cluster(Kclusters,df)
```



In [ ]:
```

```