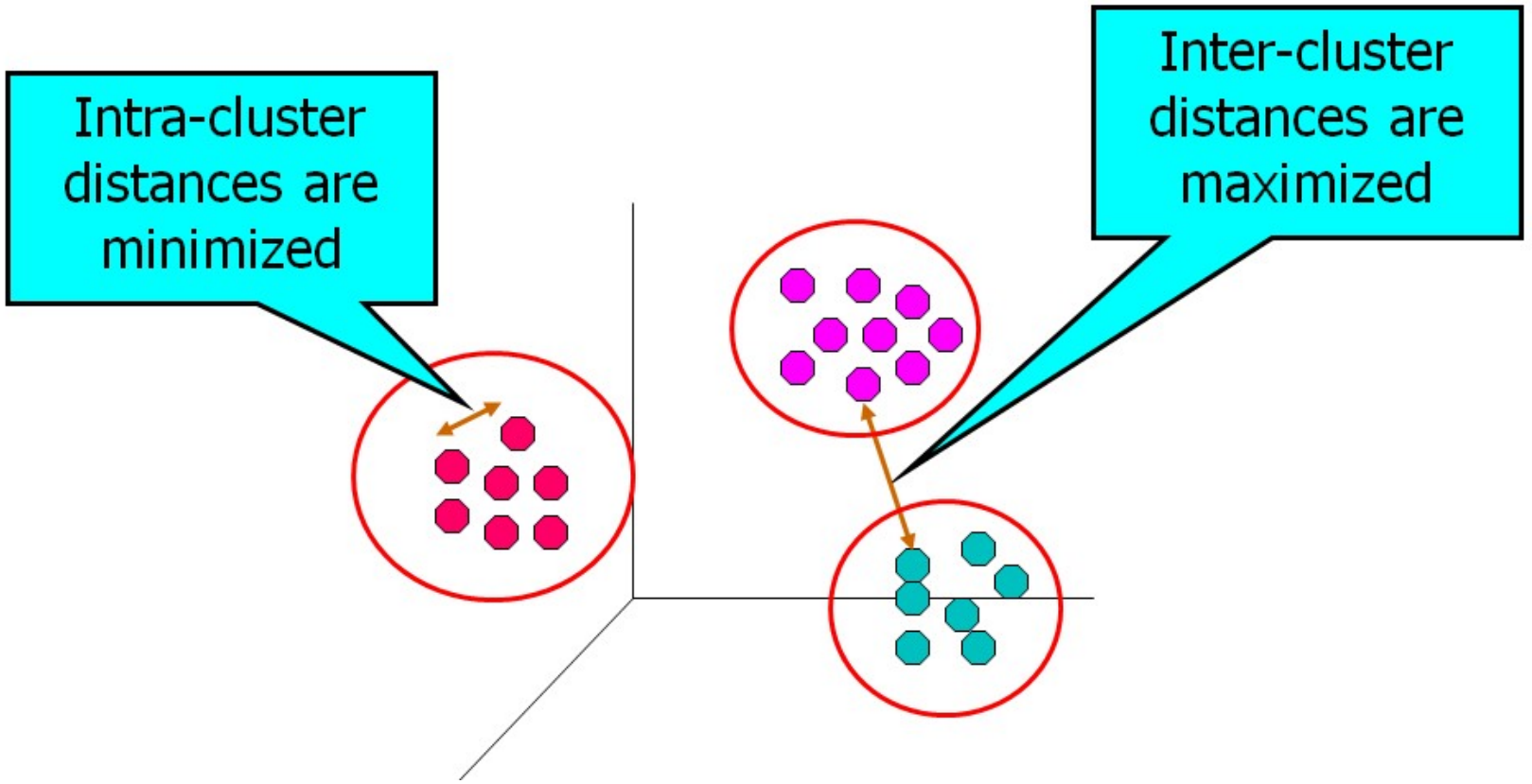
CLUSTER ANALYSIS
K MEANS METHOD

Introduction

- *Cluster analysis'* is a class of statistical techniques that can be used to classify objects or cases into groups called clusters.
- Objects can be customers, students, stores etc.
- A **cluster** is a group of relatively **homogeneous** cases or observations.
- The observations are **dissimilar** to **objects** outside the cluster, particularly objects in **other clusters**.
- Cluster Analysis is one of the unsupervised learning method. There is no concept of "Dependent" variable.

Visualizing Clusters



K- MEANS CLUSTERING METHOD

K-Means Method

- K-Means Clustering is one of the most popular non-hierarchical clustering method.
- K-Means method is suitable for large data sets and widely used for customer segmentation in BFSI or retail domains.
- The number of clusters (k) must be known a priori (though in reality this may not be the case).
- Alternatively, cluster solutions can be observed for different k and evaluated to get the best possible cluster solution.

Steps Involved in K-means Clustering

- Define K
- Define the Distance Measure
- Select Initial Seeds
- Execute Algorithm
- Check the Output for
  - Cluster Contribution
  - R-squared
  - Within and Between Sum of Squares
- Repeat the procedure for different K if criterion not satisfied

Distance Measures

- Clustering algorithms require a mathematical measure to assess the similarity of a pair of observations or clusters.

Object	X1	X2			XP
1	a1	a2			ap
2	b1	b2			bp

- Manhattan distance - The sum of the absolute differences in value for each variable ,it is calculated as,

  $d(x , y) = | a1-b1 |+| a2-b2 |+...+| ap-bp |$

- Chebyshev distance - The maximum absolute difference in values for each variable

  $d(x , y) =Max (| ai-bi |)$

Distance Measures...

- Squared Euclidean distance- The sum of squared differences between values of each variable .

Object	X1	X2			XP
1	a1	a2			ap
2	b1	b2			bp

- $d(x , y) = (a1-b1)^2+(a2-b2)^2+....+(ap-bp)^2$

- The square root is defined as 'Euclidean Distance'.

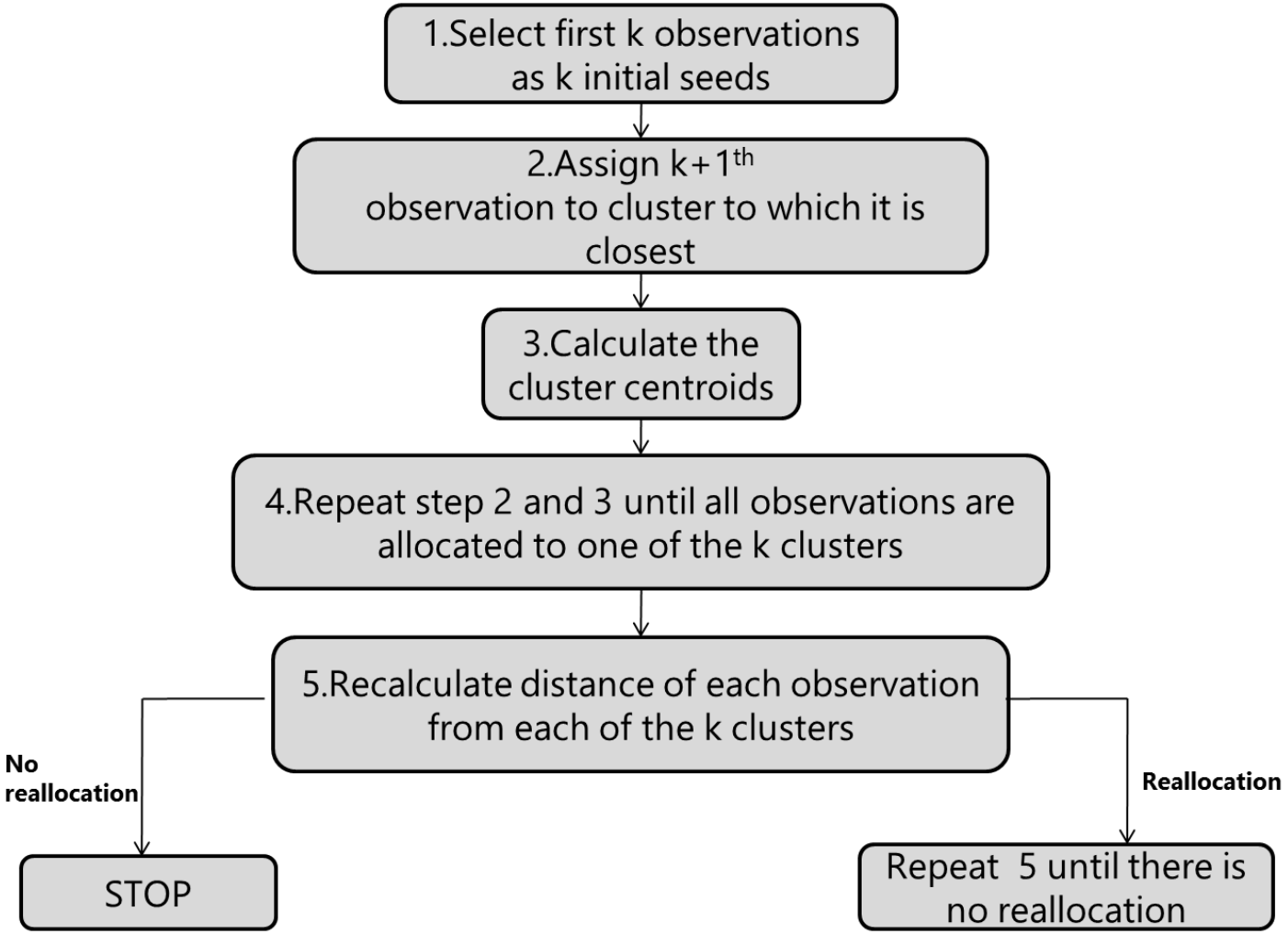
- 'Euclidean Distance' is the most widely used distance measure in cluster analysis.

Choice of Initial Seeds

There are different methods to decide initial seeds. Some of them are as follows,

- K-random observations
- First K observations
- Last K observations
- Partition the data into k partitions randomly and then use the partition mean / median as initial seeds

Algorithm



DATA SCIENCE INSTITUTE

Cluster Analysis: Example for Discussing Algorithm

- An insurance co. would like to explore its small town business and create strategies for different groups of small towns.
- Insurance data available for the towns are
  - X1: Loss Ratio
  - X2: Premium Rates
  - X3: No. of Policies from that town
- Objective is to form **two** clusters of towns

Town	Loss Ratio	Premium Rates	No. of Policies
A	1.06	9.2	151
B	0.89	10.3	202
C	1.43	15.4	113
D	1.02	11.2	168
E	1.49	8.8	192
F	1.32	13.5	111
G	1.22	12.2	175
H	1.1	9.2	245

Iteration 1

- Data:

Town	X1	X2	X3
A	1.06	9.2	151
B	0.89	10.3	202
C	1.43	15.4	113
D	1.02	11.2	168
E	1.49	8.8	192
F	1.32	13.5	111
G	1.22	12.2	175
H	1.1	9.2	245

- K=2 (It means we are forming 2 clusters of 8 towns)
- Step 1 : Initial Seeds

Initial Seeds			
Company	**X1**	**X2**	**X3**
A	**1.06**	**9.2**	**151**
B	**0.89**	**10.3**	**202**

Iteration 1 (contd.)

Step2: Find distance of Town C from A (Cluster1) and B(Cluster2)

Distance of C from A = $\sqrt{(1.43-1.06)^2+(15.4-9.2)^2+(113-151)^2}$ **=38.50**
Distance of C from B = $\sqrt{(1.43-0.89)^2+(15.4-10.3)^2+(113-202)^2}$ =89.15

Minimum Distance = 38.50

Since distance between town C and town A is minimum, town C will combined with town A.

Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C	1.245	12.3	132
2	B	0.89	10.3	202

Iteration 1 (contd.)

Step3: Find distance of Town D from Cluster1and Cluster2

Distance of D from Cluster 1	36.018
Distance of D from Cluster 2	34.012
Minimum Distance	**34.012**

Here town D will combined with town B (cluster 2)
Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C	1.245	12.3	132
2	B,D	0.955	10.75	185

Iteration 1 (contd.)

Step 4: Find distance of town E from Cluster 1 and Cluster 2

Distance of E from Cluster 1	60.102
Distance of E from Cluster 2	7.2862
Minimum Distance	**7.2862**

Here town E will combined with cluster 2(i.e. with towns B & D)
Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C	1.245	12.3	132
2	B,D,E	1.133333	10.1	187.3333

Iteration 1 (contd.)

Step 5: Find distance of town F from cluster 1 and cluster 2

Distance of F from Cluster 1	21.034
Distance of F from Cluster 2	76.409
Minimum Distance	**21.034**

Here town F will be combined with cluster 1(i.e. with towns A & C).
Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C,F	1.27	12.7	125
2	B,D,E	1.133333	10.1	187.3333

Iteration 1 (contd.)

Step 6: Find distance of town G from cluster 1 and cluster 2

Distance of G from Cluster 1	50.003
Distance of G from Cluster 2	12.511
Minimum Distance	**12.511**

Here town G will be combined with cluster 2(i.e with towns B,D & E).
Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C,F	1.27	12.7	125
2	B,D,E,G	1.155	10.625	184.25

Iteration 1 (contd.)

Step7 : Find distance of town H from cluster 1 and cluster 2

Distance of H from Cluster 1	120.05
Distance of H from Cluster 2	60.767
Minimum Distance	**60.767**

Here town H will be combined with cluster 2(i.e with towns B,D,E & G).
Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C,F	1.27	12.7	125
2	B,D,E,G,H	1.144	10.34	196.4

Since all the towns are assigned to two clusters, to verify our clusters membership we go for the next iteration.

Iteration 2

In iteration 2, initial seeds will be those two clusters which are obtained at the end of iteration 1.

Step1 :

Initial seeds				
Cluster	Members	X1	X2	X3
1	A,C,F	1.27	12.7	125
2	B,D,E,G,H	1.144	10.34	196.4

Iteration 2 (contd.)

Step 2 :- Find the Distance of town A from Cluster 1(i.e from combined towns A,C &F) and then Cluster 2(i.e from combined towns B,D,E,G & H).

Distance of A from Cluster 1 = $\sqrt{(1.06-1.27)^2+(9.2-12.7)^2+(151-125)^2}$ = 26.23536

Distance of A from Cluster 2 = $\sqrt{(1.06-1.44)^2+(9.2-10.34)^2+(151-196.4)}$ = 45.41439

Minimum Distance = 26.23536

Since distance between town A and cluster 1 is minimum, town A will be retained in cluster 1.

Iteration 2 Summary

Initial seeds				
Cluster	**Members**	**X1**	**X2**	**X3**
1	A,C,F	1.27	12.7	125
2	B,D,E,G,H	1.144	10.34	196.4

No town is reassigned to different cluster. This is final cluster solution.

Town	Distance from Cluster1	Distance from Cluster2	Cluster
A	26.24	45.41	1
B	77.04	5.61	2
C	12.30	83.55	1
D	43.03	28.41	2
E	67.11	4.67	2
F	14.02	85.46	1
G	50.00	21.48	2
H	120.05	48.61	2

Statistics Associated with Cluster Solution

Cluster solution can be assessed using ‘between clusters’ variability and
‘within clusters’ variability.

Within Sum of Squares (WSS) is a measure to explain homogeneity within a cluster.
WSS is calculated for each cluster and then added to get Total WSS should be small

R-squared is computed as ratio of Between Clusters Variability to R-squared should be large

DATA SCIENCE INSTITUTE

K-Means Method – Some Notes

- Standardize variables if scale differs widely. Variables with high variance tend to influence cluster solution. (Recommended always)
- Use data reduction technique like factor analysis before cluster analysis if number of variables is high.
- Run the algorithm for different choices of K and initial seeds.
- Use dummy variables for nominal scaled variables. K means algorithm is not suited for nominal scaled variables.

K-Means Method in R

Custid	nsv	n_brands	n_bills	growth
1001	2119456	7	14	-1.79
1002	1460163	12	42	-1.73
1003	147976	4	6	2.81
1004	1350474	13	30	-0.99
1005	1414461	15	29	13.56
1006	2299185	21	49	11.07
1007	1250260	6	15	1.92
1008	220072	6	4	0.58
1009	461122	7	17	4.06
1010	246484	4	7	3.45
1011	2075449	17	46	15.68
1012	1787336	15	21	1.78
1013	1669201	10	22	-0.95
1014	267064	10	5	4.74
1015	183152	5	4	0.16
1016	435751	21	14	4.17
1017	230062	5	12	5.24
1018	2213576	14	14	5.69
1019	2433971	11	25	3.71
1020	2517485	10	25	-1.48

Objective is to form clusters of FMCG company customers based on buying behavior.

nsv: Net Sales Value
n_brands: Number of unique brands purchased
n_bills: Number of bills generated
growth: Growth in net sales value

Period: One Year

K-Means Method in R

```
custsales<-read.csv(file.choose(),header=T)
custsales_cl<-subset(custsales,select=c(-Custid,-region))

#scale (standardize) all variables.(subtract mean and divide by standard deviation)

custsales_cl<-scale(custsales_cl)
CL<-kmeans(custsales_cl,4)
CL
```

```
K-means clustering with 4 clusters of sizes 210, 405, 229, 314
Cluster means:
        nsv          n_brands     n_bills        grow
1  1.0589762  1.50534917    1.6219927    1.6228
2 -0.8311329  -0.84045295  -0.7207329   -0.5315
3  1.1863778  -0.02444231   0.3044816   -0.62581250
4 -0.5014544   0.09508729  -0.3772226    0.05665368

 Within cluster sum of squares by cluster:
[1] 732.8205  166.3279  314.9123 145.5306
 (between_SS / total_SS =  70.6 %)
```

Cluster 1 looks platinum customers group

DATA SCIENCE INSTITUTE

K-Means Method in R
Append Segment Variable

```
custsales$segment<-CL$cluster
head(custsales)
```

	Custid	nsv	n_brands	n_bills	growth	region	segment
1	1001	2119456	7	14	-1.79	Mumbai	3
2	1002	1460163	12	42	-1.73	Mumbai	3
3	1003	147976	4	6	2.81	Mumbai	2
4	1004	1350474	13	30	-0.99	Delhi	3
5	1005	1414461	15	29	13.56	Delhi	1
6	1006	2299185	21	49	11.07	Delhi	1

Now segment can be used as any other variable for further analysis.

K-Means Method in R

Summarize Clusters Using Original Variables

```
aggregate( cbind(nsv,n_brands,n_bills,growth)~segment,data=custsales,FUN=me
    an)
```

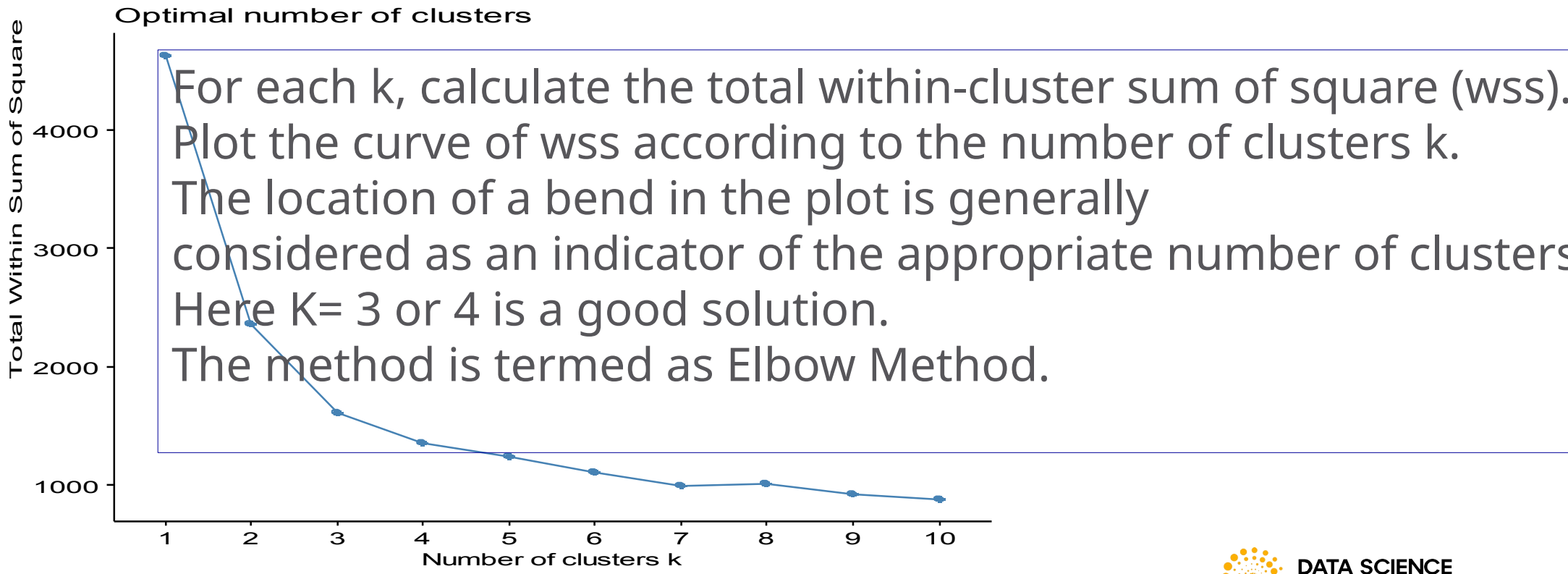
segment	nsv	n_brands	n_bills	growth
1	1875311.4	24.24	48.62	12.28
2	238729.4	4.76	5.79	2.27
3	1985624.2	11.53	24.53	1.84
4	524186.9	12.53	12.07	5.00

Cluster 1 is group of ‘Platinum’ clusters.
Cluster 2 is a group of ‘non-performers’

K-Means Method in R

How to decide number of clusters?

```
library(factoextra)
fviz_nbclust(custsales_cl, kmeans,method = "wss")
```



For each k, calculate the total within-cluster sum of square (wss).
Plot the curve of wss according to the number of clusters k.
The location of a bend in the plot is generally considered as an indicator of the appropriate number of clusters.
Here K= 3 or 4 is a good solution.
The method is termed as Elbow Method.

DATA SCIENCE INSTITUTE

kmeansruns() in "fpc" Package Finding Best K

- Package: fpc: Flexible Procedures for Clustering

- Performs K-means method for different values of 'K' and provides best value of K.

R commands:

```
CL1<-kmeansruns(custsales_cl,krange=2:10)
CL1$bestk

# Use this as only indicative.
```

DATA SCIENCE INSTITUTE

THANK YOU!!

DATA SCIENCE
INSTITUTE