insAnalytics

Consulting. Training. Research & Development.

Insights.. Through Analytics…

Data Mining

Cluster Analysis

Key Objectives

After successful completion of the topic, participants will be able to:

- ❑ Develop awareness of the concepts of different types of **Segmentation Techniques**

- ❑ Articulate the **key difference** between of Objective Segmentation and Subjective Segmentation

- ❑ **Build**, **interpret** and draw **conclusions** from Hierarchical Segmentation using different linkage method and distance measure

- ❑ Understand the key concept of non-hierarchical segmentation techniques

- ❑ **Build**, **interpret** and draw **conclusions** from K-Means segmentation

- ❑ **Profiling the cluster solution**, and **building the cluster equation** from K-Means clustering

To learn & understand the subject matter better, participants need to be aware of the following areas:

- ❑ Descriptive Statistics
- ❑ Correlation Analysis
- ❑ Statistical Inferences

- ❑ Analysis of Variance (ANOVA)
- ❑ Data Mining – CRISP DM Methodology
- ❑ Decision Trees

Cluster Analysis

Introduction to Segmentation

Types of Segmentation

Segmentation through Hierarchical Clustering

Segmentation through Non-hierarchical Clustering

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution

Cluster Analysis

Introduction to Segmentation

Types of Segmentation

Segmentation through Hierarchical Clustering

Segmentation through Non-hierarchical Clustering

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution

Cluster Analysis

Introduction to Segmentation

Types of Segmentation

A Business Illustration

Segmentation through

What is Segmentation

Why Do We Need Segmentation

Segmentation through Non-hierarchical Clustering

Insights from Segmentation – An Illustration

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution

An Illustration – Segmentation of Retail Customers

ShopMart is one leading retail chain that has its outlets across different cities.

They know that different customers walk into their outlets for different products with different purchase behavior.

Hence, the marketing department wants to reach to their customers with customized offers, so that their marketing efforts yields maximum benefit.

For this, what they have to do?

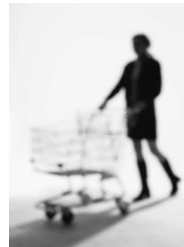## An Illustration – Segmentation of Retail Customers

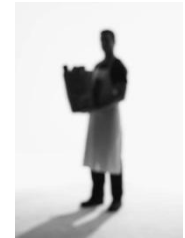ShopMart needs to identify customers with specific behaviors

Family Shoppers

High Spenders

Occasional Visitors

Internet Sales

Business Shoppers

insAnalytics

An Illustration – Segmentation of Retail Customers

ShopMart needs to identify customers with specific behaviors

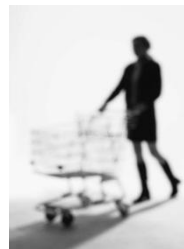Then identify the appropriate marketing strategies

Family Shoppers

High Spenders

Occasional Visitors

Internet Sales

Business Shoppers

Product Bucket
(e.g. Biscuit free
with Health Drinks)

Buy 2 Get 2 Free
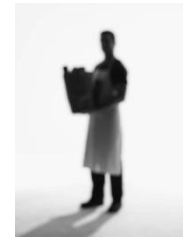
Festive
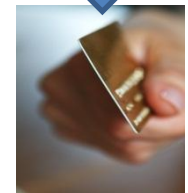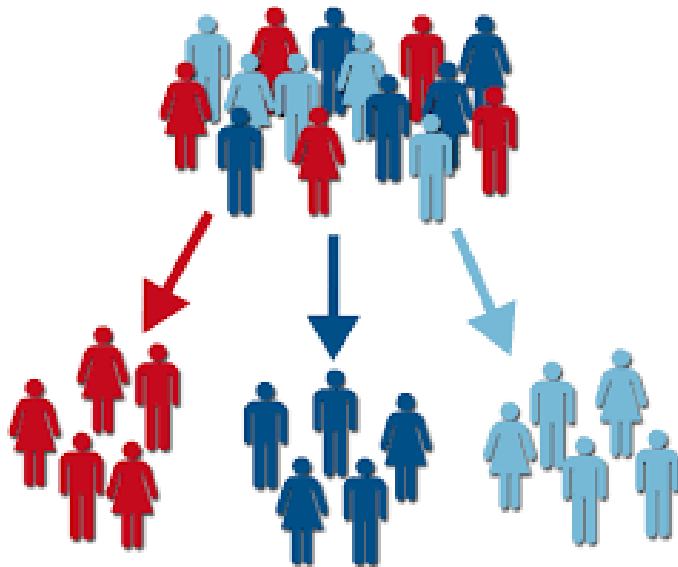Offerings

Free Home
Delivery

Easy Payment
Option

What is Segmentation?
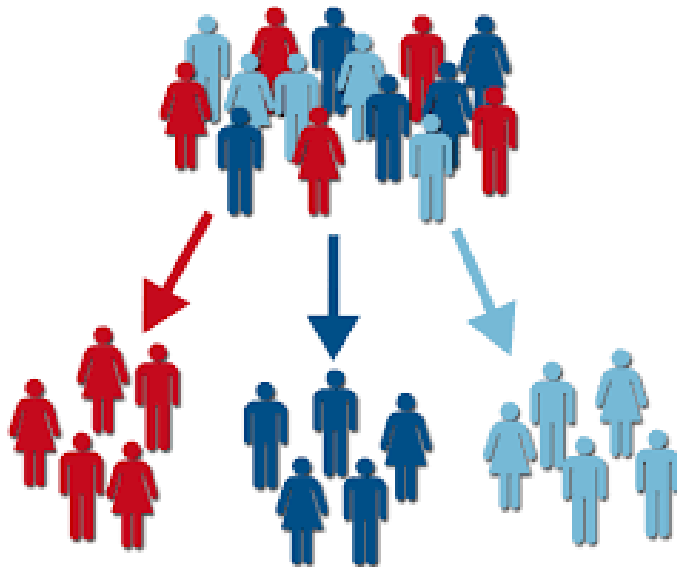
In a segmentation analysis,

We **divide the observations** into mutually exclusive and exhaustive distinct identifiable homogeneous groups.

What is Segmentation?

In a segmentation analysis,

We divide the observations into **mutually exclusive and exhaustive** distinct identifiable homogeneous **groups**.

This means

Each observation should belong **to one and only one segment**

What is Segmentation?

In a segmentation analysis,

We divide the observations into mutually exclusive and exhaustive **distinct** identifiable homogeneous **groups**.

This means

the **groups will be different from each other** with respect to some characteristics

What is Segmentation?

In a segmentation analysis,

We divide the observations into mutually exclusive and exhaustive distinct **identifiable** homogeneous **groups**.

This means

once a new observation comes in, we should be able to **identify which group it belongs** to

What is Segmentation?

In a segmentation analysis,

We divide the observations into mutually exclusive and exhaustive distinct identifiable **homogeneous groups**.
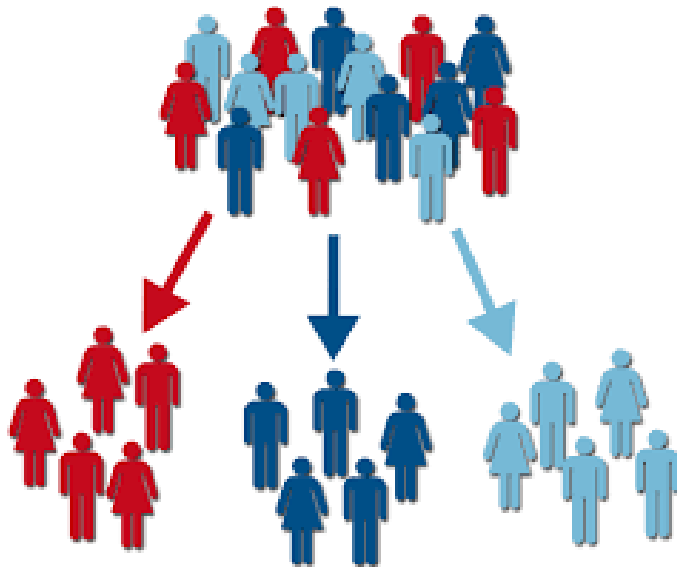
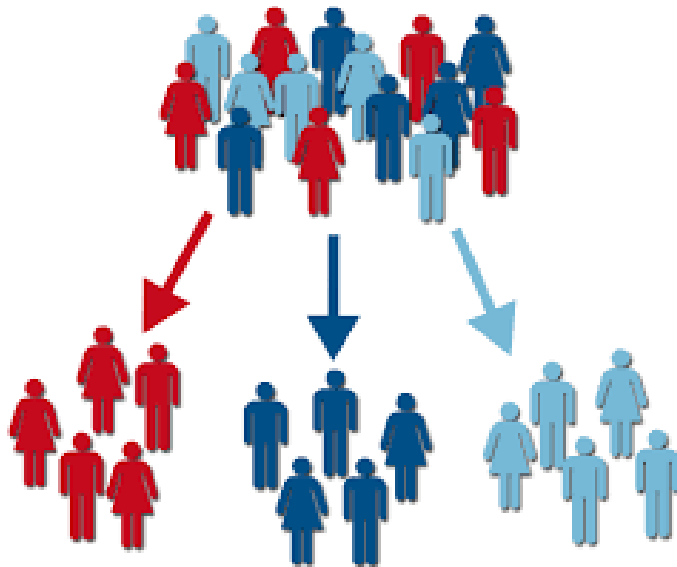This means

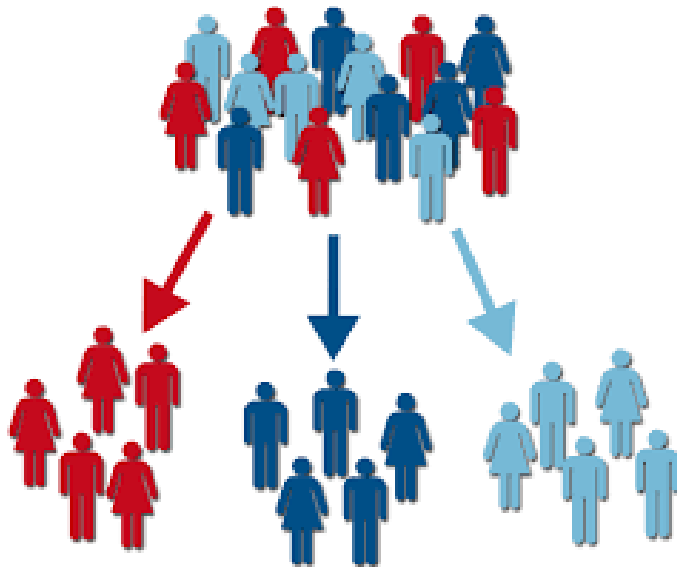**Observations in one group are similar to each other**

Why Do We Need Segmentation

Charlie, the marketing analyst of ShopMart found that

Each customer is so different that ideally she would need to reach out to each one in the way she/he buys
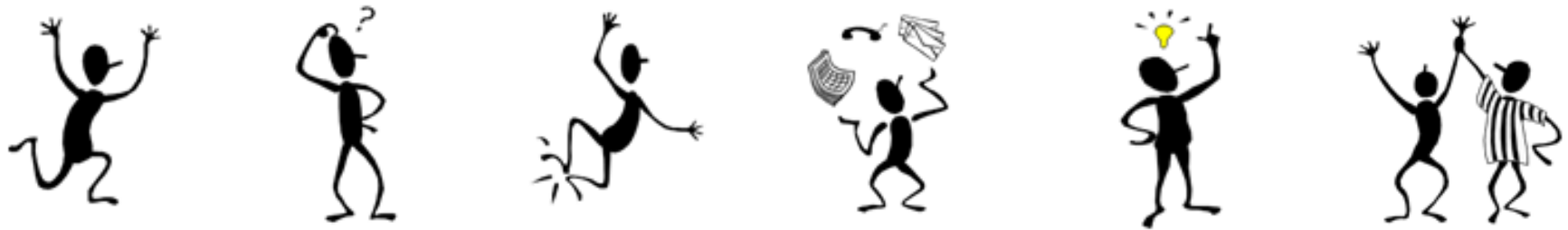
Why Do We Need Segmentation

Charlie, the marketing analyst of ShopMart found that

Each customer is so different that ideally she would need to reach out to each one in the way she/he buys
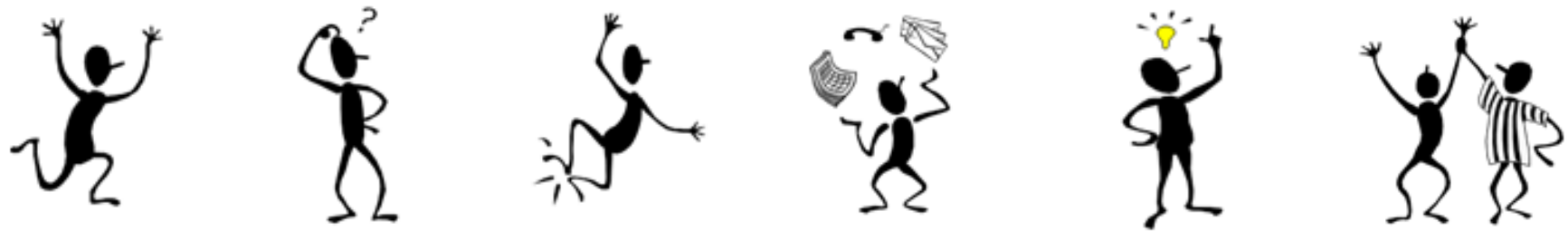
But, with approximately 20K people shopping a day at different outlets, **it is too difficult to do any customization at individual level**

Why Do We Need Segmentation

Hence, Charlie decided to

identify segments where people have same characters

and target each of these segments in a different way
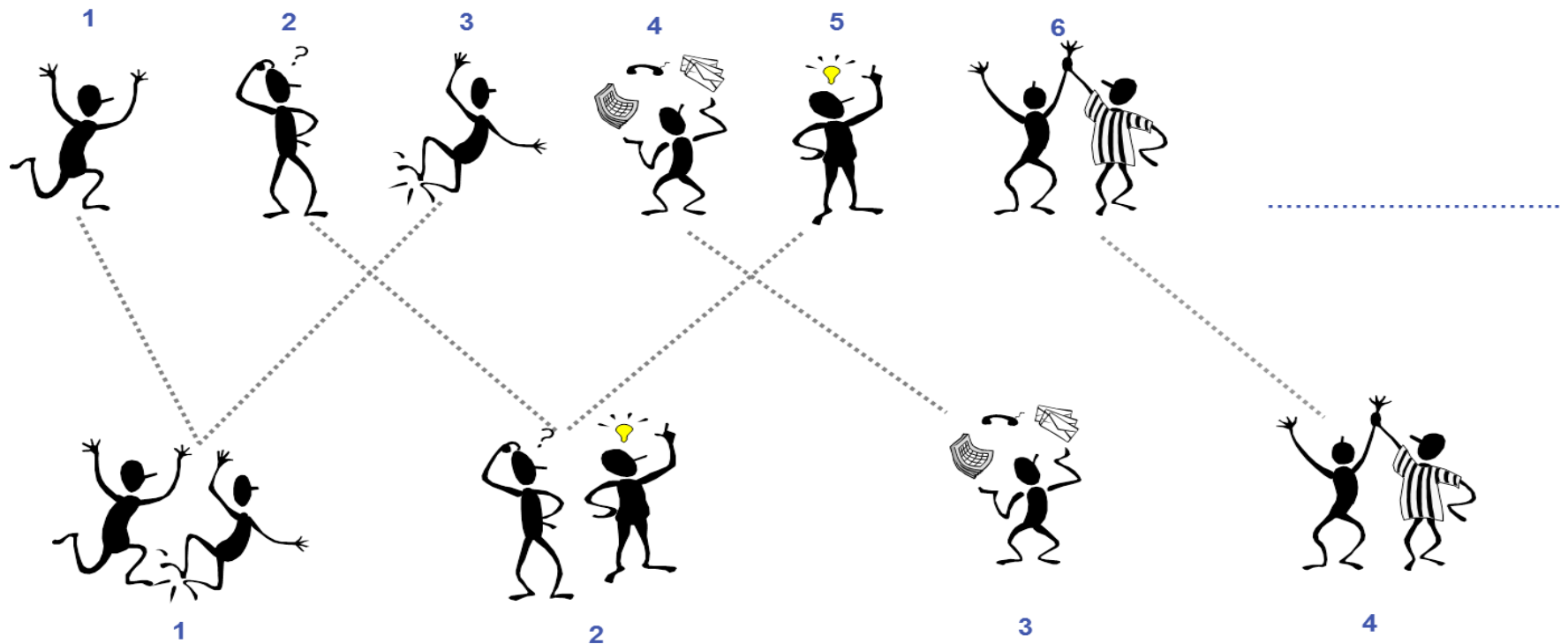
Insights from Segmentation Result

John – A Customer of ShopMart

"I am a marketing organization. My age is 34. I am married with a kid. My wife is a software professional. We often go for a family vacation on long weekends."

Insights from Segmentation Result

John – A Customer of ShopMart

"I am a marketing organization. My age is 34. I am married with a kid. My wife is a software professional. We often go for a family vacation on long weekends."

Segment Profile:

- John and his wife both are working and hence, might be coming to stores on weekends only.
- They have a kid at home
- They often go for vacations

Insights from Segmentation Result

John – A Customer of ShopMart

"I am a marketing organization. My age is 34. I am married with a kid. My wife is a software professional. We often go for a family vacation on long weekends."

Segment Profile:
- John and his wife both are working and hence, might be coming to stores on weekends only.
- They have a kid at home
- They often go for vacations

Opportunity:
- John is a family shopper and will be more interested in product bucketing
- He should be reached out for any weekend discounts or promotions
- As he has a kid, he may be purchasing baby foods, health drinks etc.
- For his vacations times, he may be interested in travel items

Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation

Segmentation through Hierarchical Clustering

Segmentation through Non-hierarchical Clustering

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution

Cluster Analysis

Introduction to Segmentation ✓
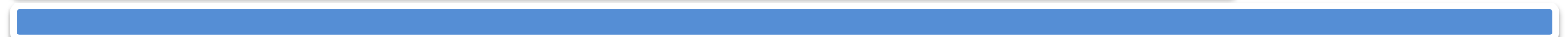
Types of Segmentation

Segmentation through Hierarchical Clustering

Segmentation through Non-hierarchical Clustering

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution

Two Broad Types of Segmentation

```
                        ┌──────────────────┐
                        │   Segmentation   │
                        └──────────────────┘
              ┌──────────────────┴──────────────────┐
     ┌──────────────────┐                  ┌──────────────────┐
     │    Objective     │                  │    Subjective    │
     │   Segmentation   │                  │   Segmentation   │
     └──────────────────┘                  └──────────────────┘
              │                                      │
     ┌─────────────────────────┐          ┌─────────────────────────┐
     │  Chai-Square Automatic  │          │  Hierarchical Clustering │
     │   Interaction Detector  │          └─────────────────────────┘
     └─────────────────────────┘
     ┌─────────────────────────┐          ┌─────────────────────────┐
     │     Classification &    │          │     Non-Hierarchical     │
     │     Regression Tree     │          │        Clustering        │
     └─────────────────────────┘          └─────────────────────────┘
```
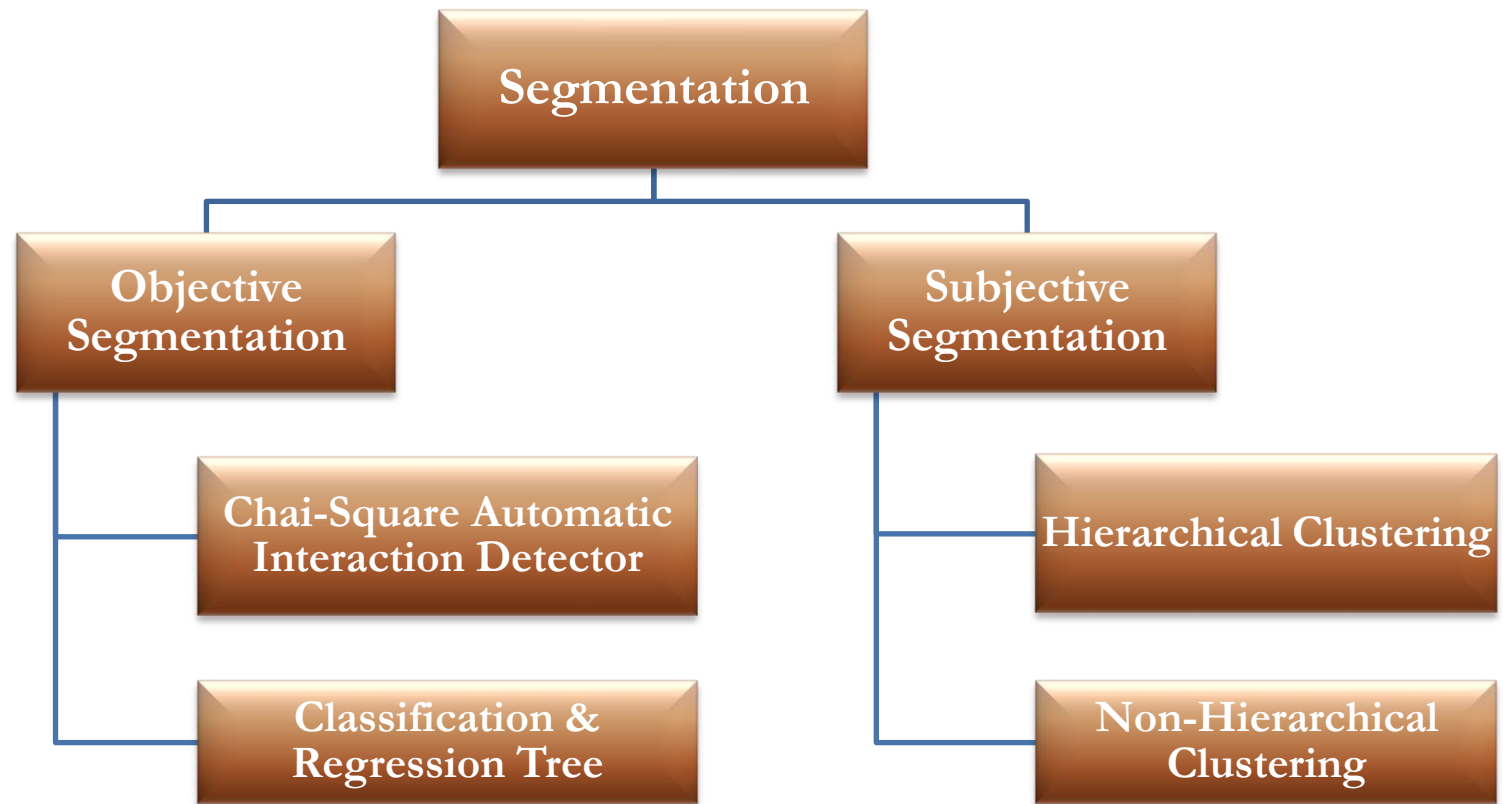
Objective Segmentation – Decision Tree
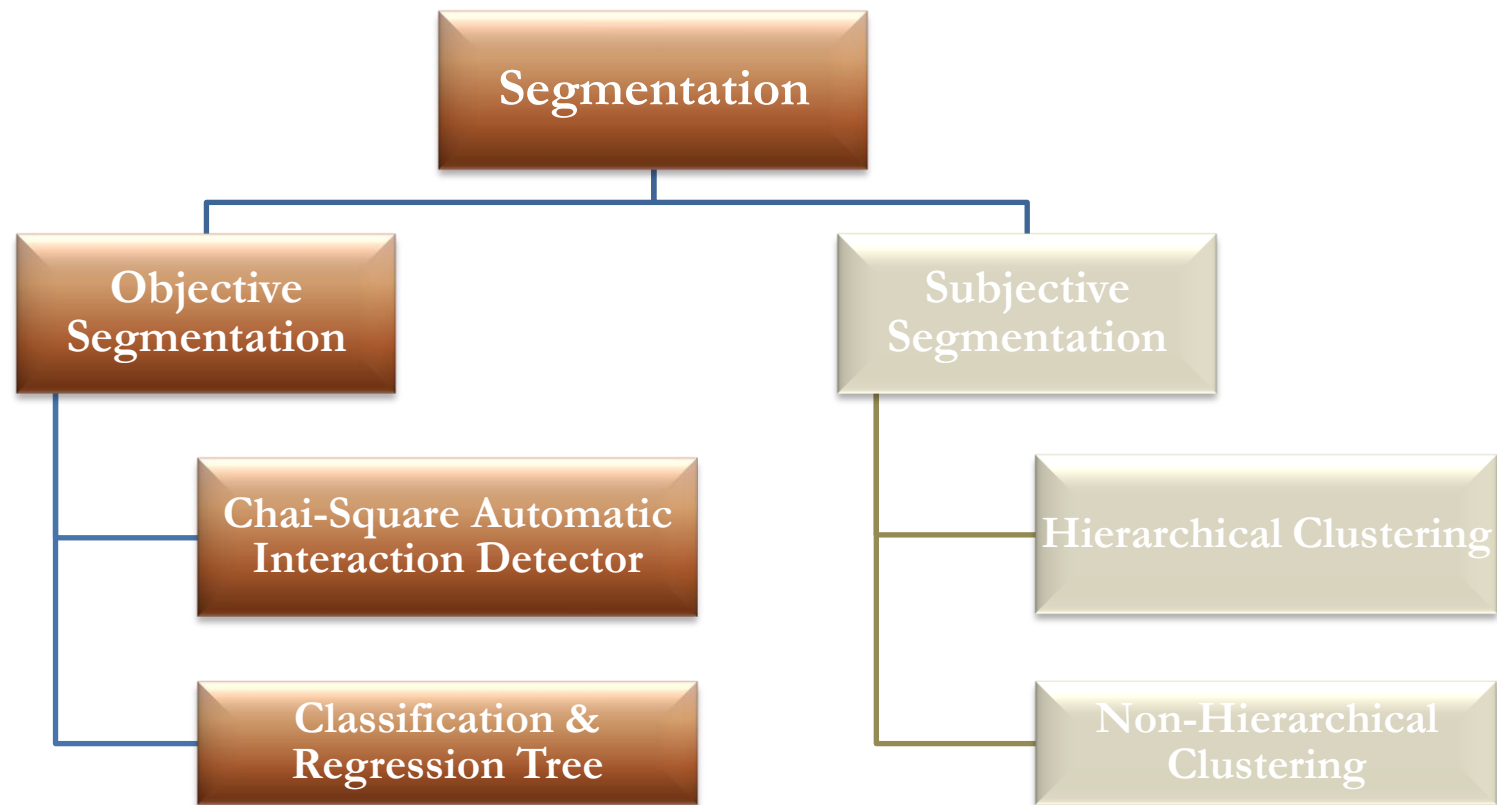
```
                            ┌──────────────────────┐
                            │    Segmentation      │
                            └──────────┬───────────┘
                    ┌──────────────────┴──────────────────┐
        ┌───────────────────────┐              ┌───────────────────────┐
        │      Objective        │              │      Subjective       │
        │    Segmentation       │              │    Segmentation       │
        └───────────┬───────────┘              └───────────┬───────────┘
                    │                                       │
        ┌───────────────────────┐              ┌───────────────────────┐
        │ Chai-Square Automatic │              │ Hierarchical Clustering│
        │  Interaction Detector │              └───────────────────────┘
        └───────────────────────┘
        ┌───────────────────────┐              ┌───────────────────────┐
        │   Classification &    │              │   Non-Hierarchical    │
        │   Regression Tree     │              │      Clustering       │
        └───────────────────────┘              └───────────────────────┘
```

When we segment a population **based on one target variable**, then it is called Objective Segmentation.

Hence, it is a **supervised learning**.

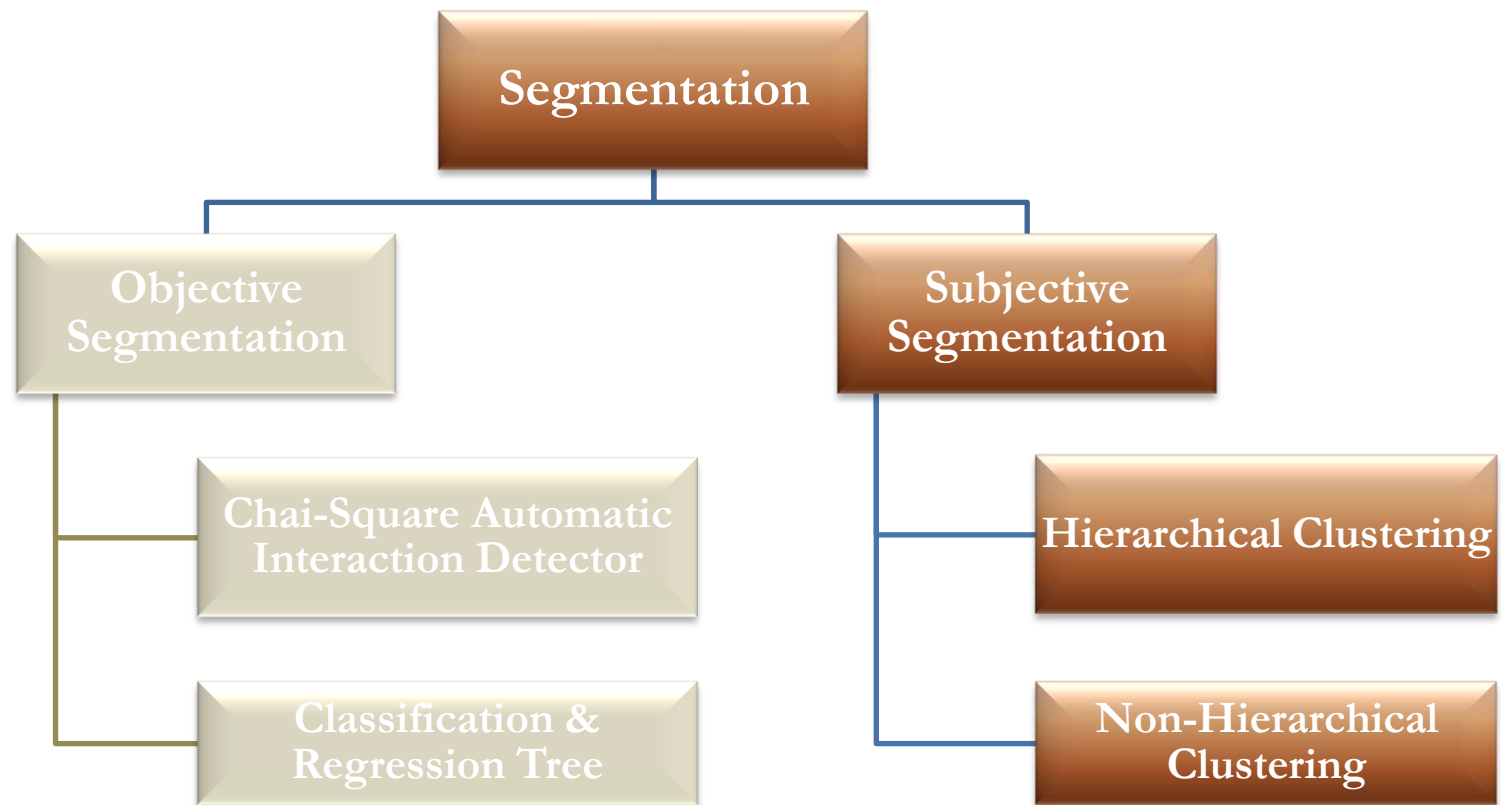It is more popularly known as **Decision Trees**

Decision Tree – Applications

- BankOnUs, a retail bank, wants to find the customers who are **unlikely to pay loan installment** in next month. So that they can contact the customer well in advance for collections

- Vivgyor Telecom, a telecom service provider, want to segment their customers based on their **potential usage in next 6 months**

- EasyBuy, a online marketplace, wants to know the which vendors **may default the service level agreement** to end the products to the consumer

Decision Tree – Applications

❑ BankOnUs, a retail bank, wants to find the customers who are unlikely to pay loan installment in next month. So that they can contact the customer well in advance for collections

❑ Vivgyor Telecom, a telecom service provider, want to segment their customers based on their potential usage in next 6 months

❑ EasyBuy, a online marketplace, wants to know the which vendors may default the service level agreement to end the products to the consumer

In all of the above cases segments are being created based on a **target variable**

Analyst knows how the end segments will look like, just she needs to find them

Subjective Segmentation – Cluster Analysis

```
                        ┌──────────────────────┐
                        │     Segmentation     │
                        └──────────┬───────────┘
              ┌────────────────────┴────────────────────┐
     ┌──────────────────┐                      ┌──────────────────┐
     │     Objective    │                      │    Subjective    │
     │   Segmentation   │                      │   Segmentation   │
     └──────────────────┘                      └──────────────────┘
              │  ┌──────────────────────┐               │  ┌──────────────────────┐
              ├──│  Chai-Square Automatic│               ├──│ Hierarchical Clustering│
              │  │  Interaction Detector │               │  └──────────────────────┘
              │  └──────────────────────┘               │  ┌──────────────────────┐
              │  ┌──────────────────────┐               └──│   Non-Hierarchical   │
              └──│   Classification &   │                  │     Clustering       │
                 │   Regression Tree    │                  └──────────────────────┘
                 └──────────────────────┘
```

When we segment a population **based on all the relevant variables** we have rather than one target variable, then it is called Subjective Segmentation.

Hence, it is a **unsupervised learning**.

It is more popularly known as **Cluster Analysis**

Cluster Analysis – Applications

❑ ShopMart, a retail chain, wants to understand the inherent customer purchase behavior and segment the customers based on that behavior

❑ MorningToNight is a popular health drinks. The manufacturer wants to understand its consumers so that they can develop new products that can full-fill the unmet need in the market

❑ Government of India wants to design some social welfare schemes. For that, they need to segment different districts based on socio economic characteristics.

Cluster Analysis – Applications

- ☐ ShopMart, a retail chain, wants to understand the inherent customer purchase behavior and segment the customers based on that behavior

- ☐ MorningToNight is a popular health drinks. The manufacturer wants to understand its consumers so that they can develop new products that can full-fill the unmet need in the market

- ☐ Government of India wants to design some social welfare schemes. For that, they need to segment different districts based on socio economic characteristics.

Here, we do not know how the segments will look like.

Hence, appropriate actions are decided only after the segments are achieved

Cluster Analysis – Two Methods

Clusters can be created in two ways

❑ We can find the distance between observations based on different variables. Then based on closeness, we can create the segments in a step-by-step approach.
This type of clustering is called **Hierarchical Clustering**
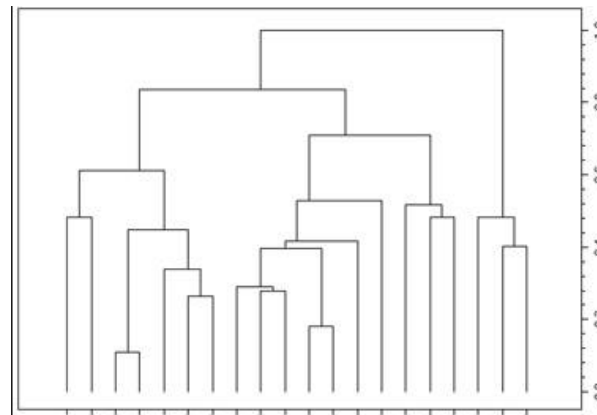
We start with all observations in different clusters

Then group observations **one by one**, until all observations are grouped into one cluster

Cluster Analysis – Two Methods

Clusters can be created in two ways

❑ Alternatively, we can try to create some boundaries based on the variables under study. Observations falling under same boundary creates cluster.
This type of clustering is called **Non-hierarchical Clustering**
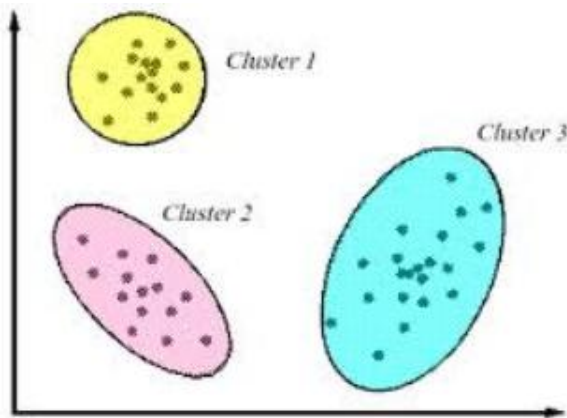
We first decide the number of clusters we want

Then **iteratively** find the best boundary conditions based on distance from cluster centroids

Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation ✓

Segmentation through Hierarchical Clustering

Segmentation through Non-hierarchical Clustering

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution

Data Mining

Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation ✓

Segmentation through Hierarchical Clustering

Segmentation through

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution
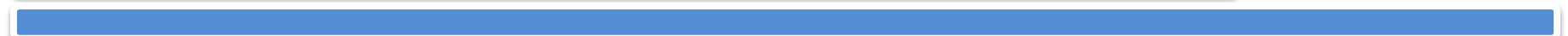
Steps to Build Hierarchical Clustering

Selecting and Standardization of Variables

Measuring Distance

Linkage Methods

Grouping of Observations

Visualizing the Clusters

Hierarchical Clustering through R

Advantages and Disadvantages

Export Market Segmentation – An Example

World Textile is an India based textile manufacturing company and wants to export its products to other countries.

To know where it can export, it wants to know two things, viz.

- ✓ Understand the demand of Indian textile in those countries
- ✓ Group the countries based on the demand and other socio-economic characteristics

Raghu, the analyst at World Textile, collected data on textile export of India to other countries during 2005 to 2014 and few other socio-economic information about these countries from different publicly available data sources.

He decided to do build a hierarchical clustering on the data

insAnalytics

Export Market Segmentation – The Data

Raghu has collected the data on following 10 variables on 20 countries

- ✓ Average export amount in last 10 years
- ✓ Average YOY change in export amount
- ✓ Maximum export amount
- ✓ No of years with positive YOY change
- ✓ Continent

- ✓ Government type
- ✓ Developmental status
- ✓ Dominant religion
- ✓ Legal system
- ✓ Primary land usage

He wanted to run a pilot run using the following 7 countries

Partner Name	Avg Export	Avg Rate of Change	Max Export	Positive Growth	Continent	Government Type	Development Status	Dominant Religion	Legal System	Land Usage
France	973831.0982	0.05827799	1225212.87	7	Europe	Republic	Developed	Christian	Civil law system	Aggriculture
Germany	1656647.423	0.08544494	2214344.41	6	Europe	Federal Republic	Developed	Protestant	Civil law system	Aggriculture
Greece	63881.0015	-0.0387414	90058.303	4	Europe	Parliamentary	Developed	Orthodox	Civil law system	Aggriculture
Romania	24737.2702	0.16139592	34353.129	5	Europe	Republic	Emerging	Orthodox	Civil law system	Aggriculture
Egypt	330738.3221	0.15246934	500590.088	6	Africa	Republic	Emerging	Muslim	Mixed legal system	Others
Qatar	23513.5918	0.15522443	41186.349	8	Africa	Emirate	Emerging	Muslim	Mixed legal system	Others
South Africa	196280.6838	0.07210792	271429.22	6	Africa	Republic	Developed	Protestant	Mixed legal system	Aggriculture

Steps to Build Hierarchical Clusters

```
┌──────────────┐      ┌──────────────┐      ┌──────────────┐
│  Select the  │ ───▶ │ Standardize  │ ───▶ │ Calculate the│
│  Variables   │      │ the Variables│      │   Distance   │
└──────────────┘      └──────────────┘      └──────────────┘
                                                    │
       ┌────────────────────────────────────────────┘
       ▼
┌──────────────┐      ┌──────────────┐      ┌──────────────┐
│  Decide on   │ ───▶ │  Group the   │ ───▶ │   Plot the   │
│   Linkage    │      │ Observations │      │   Clusters   │
│   Methods    │      │              │      │              │
└──────────────┘      └──────────────┘      └──────────────┘
```
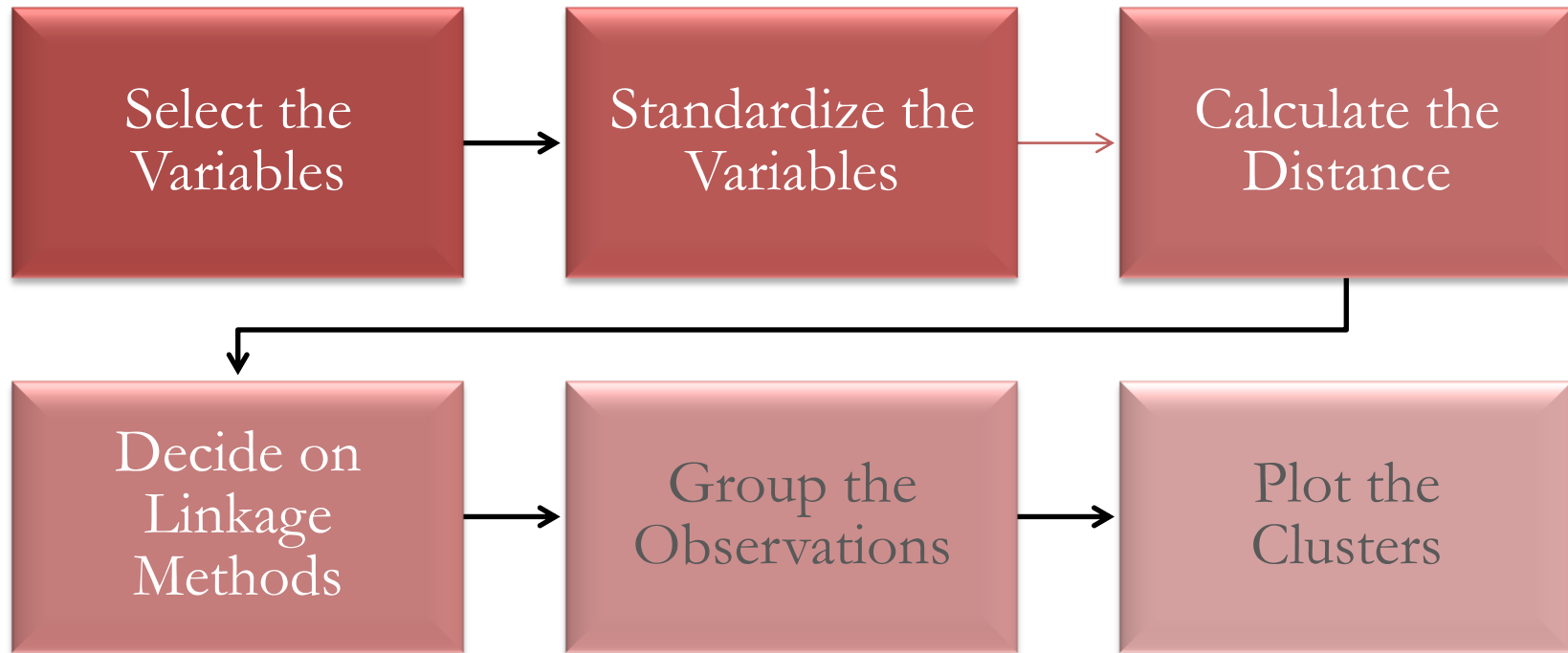
Selecting the Variables

In hierarchical clustering, we use **either all numeric variables or all categorical variables**.

We can not work with both of them together.

Beyond this, here are few more guidelines for selecting the variables for clustering

✓ Variables should provide a **clear-cut differentiation between the segments**

✓ **Correlated variables to be avoided**.

✓ Variables should be **robust outside the dataset**

✓ **Not too many variables** to be used

 Raghu wanted to do segmentation based on numeric variables in the dataset

Standardization of Numerical Variables

Note the units of variables that Raghu will be using

- ✓ Average export amount in last 10 years – unit is US thousand dollars
- ✓ Average YOY change in export amount – a fraction
- ✓ Maximum export amount – unit is US thousand dollars
- ✓ No of years with positive YOY change – a count variable

Are these variables comparable in a single scale?.... No!

Standardization of Numerical Variables

Note the units of variables that Raghu will be using

- ✓ Average export amount in last 10 years – unit is US thousand dollars
- ✓ Average YOY change in export amount – a fraction
- ✓ Maximum export amount – unit is US thousand dollars
- ✓ No of years with positive YOY change – a count variable

Are these variables comparable in a single scale?…. No!

Hence, we need to standardize **all numeric variables** first.

<p style="text-align:center; color:#9b2020;">Standardized value of X = (X – Average of X)/ SD of X</p>

This makes all variables comparable.

Standardization of Numerical Variables

After standardization, Raghu gets the following data

Partner Name	Avg Export	Avg Rate of Change	Max Export	Positive Growth
France	0.8151	-0.4724	0.7345	0.7746
Germany	1.9135	-0.0953	1.9455	-
Greece	-0.6486	-1.8192	-0.6553	-1.5492
Romania	-0.7116	0.9590	-0.7235	-0.7746
Egypt	-0.2193	0.8351	-0.1527	-
Qatar	-0.7135	0.8733	-0.7152	1.5492
South Africa	-0.4356	-0.2804	-0.4333	-

For example, "average export" variable had mean 467089.91 and SD 621661.26

Hence **for France**,

  Standardized value of average export = (973831.10 − 467089.91)/ 621661.26

$$= 0.8151$$

Measuring Distance

We start with all observations (here countries) in different clusters
and try to group them one by one based on distance between the observations

So, we need to define the distance between two countries.

Appropriate distance measures depend on the variables we are using.
If we are using numeric variables, we can use **Euclidian distance**
If we are using categorical variables, we can use **dissimilarity measure.**

All distance must be calculated on scaled or standardized data

Measuring Distance – Numeric Variables

Euclidian distance between 2 points (i, j) based on two variables (X,Y) is defined as:

$$D(i,j) = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$$

	France	Germany	Greece	Romania	Egypt	Qatar
Germany	3.4152					
Greece	11.2879	18.7006				
Romania	8.9056	15.7264	8.3268			
Egypt	4.1668	9.8172	9.8819	1.1835		
Qatar	6.8494	17.3189	16.8573	5.4074	2.9621	
South Africa	3.5649	11.2114	4.8624	2.2966	1.3699	3.8879

The above matrix is called distance matrix. Note this matrix has been created using all four selected variables

Measuring Distance – Categorical Variables

If we are using categorical variables, we can define the following **dissimilarity measure** as distance

$$D_{a(i,j)} = 1 - \frac{Number\ of\ matches}{Number\ of\ attributes}$$

Lets see what happens to the seven countries

Partner Name	Continent	Government Type	Development Status	Dominant Religion	Legal System	Land Usage
France	Europe	Republic	Developed	Christian	Civil law system	Aggriculture
Germany	Europe	Federal Republic	Developed	Protestant	Civil law system	Aggriculture
Greece	Europe	Parliamentary	Developed	Orthodox	Civil law system	Aggriculture
Romania	Europe	Republic	Emerging	Orthodox	Civil law system	Aggriculture
Egypt	Africa	Republic	Emerging	Muslim	Mixed legal system	Others
Qatar	Africa	Emirate	Emerging	Muslim	Mixed legal system	Others
South Africa	Africa	Republic	Developed	Protestant	Mixed legal system	Aggriculture

Measuring Distance – Numeric Variables

Partner Name	Continent	Government Type	Development Status	Dominant Religion	Legal System	Land Usage
France	Europe	Republic	Developed	Christian	Civil law system	Aggriculture
Germany	Europe	Federal Republic	Developed	Protestant	Civil law system	Aggriculture
Greece	Europe	Parliamentary	Developed	Orthodox	Civil law system	Aggriculture
Romania	Europe	Republic	Emerging	Orthodox	Civil law system	Aggriculture
Egypt	Africa	Republic	Emerging	Muslim	Mixed legal system	Others
Qatar	Africa	Emirate	Emerging	Muslim	Mixed legal system	Others
South Africa	Africa	Republic	Developed	Protestant	Mixed legal system	Aggriculture

$$D_{a(i,j)} = 1 - \frac{Number\ of\ matches}{Number\ of\ attributes}$$

	France	Germany	Greece	Romania	Egypt	Qatar
Germany	0.33					
Greece	0.33	0.33				
Romania	0.33	0.50	0.33			
Egypt	0.83	1.00	1.00	0.67		
Qatar	1.00	1.00	1.00	0.83	0.17	
South Africa	0.50	0.50	0.67	0.67	0.50	0.67

Grouping for the First Time

Looking at the distance matrix, we select the **two countries with least distance** and group them into one cluster.

	France	Germany	Greece	Romania	Egypt	Qatar
Germany	3.4152					
Greece	11.2879	18.7006				
Romania	8.9056	15.7264	8.3268			
Egypt	4.1668	9.8172	9.8819	1.1835		
Qatar	6.8494	17.3189	16.8573	5.4074	2.9621	
South Africa	3.5649	11.2114	4.8624	2.2966	1.3699	3.8879

Hence, **we group Romania & Ezypt into one group**.

Now we have 6 clusters, viz. "Germany", "Greece", "Qatar", "South Africa", and group with "Romania" and "Egypt" (say "**R-E**")

Linkage Methods

Now,

Distance of "France" from "Romania" is 8.9056

and the same from Egypt is 4.1668

What will be the distance of "France" from the group "R-E"?

This is decided based on the linkage method we select.

Linkage method is the method the way decide the distance between two clusters are defined.

Linkage Methods – Different Methods
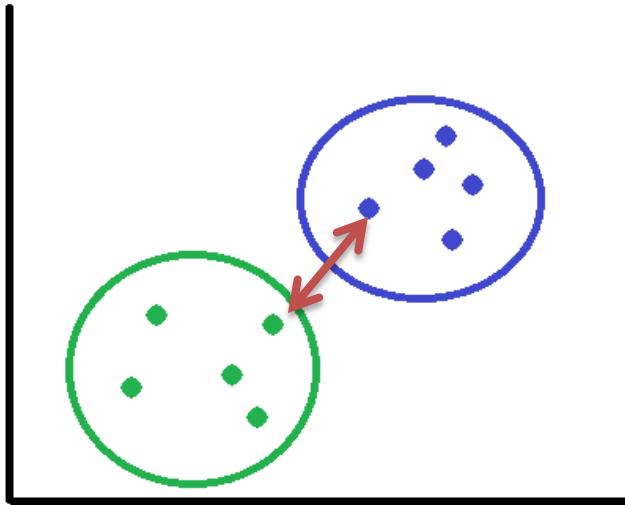
By **Single Linkage method** is distance between two clusters are defined as the **minimum** distance between two points in different clusters.

This also sometime is called **Simple Linkage method**
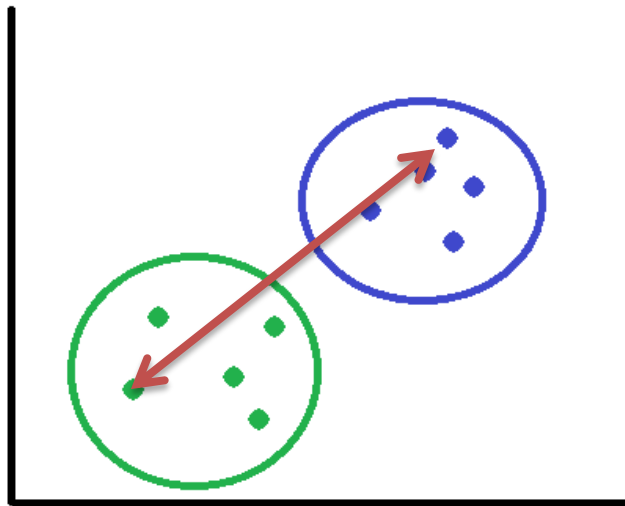
By **Complete Linkage method** is distance between two clusters are defined as the **maximum** distance between two points in different clusters.

Linkage Methods – Different Methods
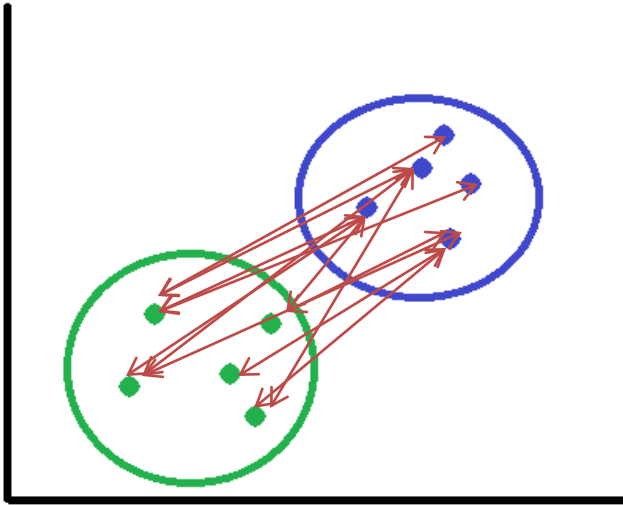
By **Average Linkage method** is distance between two clusters are defined as the **average** distance between two points in different clusters.
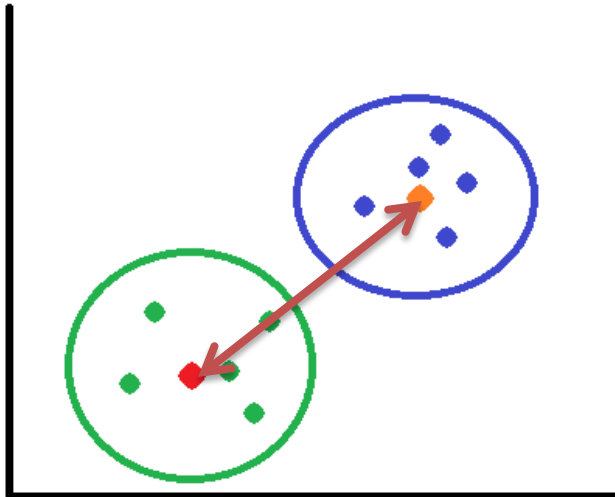
In **Centroid method** we first calculate the centre of the clusters and then find the distance between the centroids

Linkage Methods – Comparison of Measures

Let's see how the distance of other 5 countries from the cluster "R-E"

Partner Name	Distance from Romania	Distance from Egypt	Distance from Cluster "R-E"		
			Single Linkage	Complete Linkage	Average Linkage
France	8.9056	4.1668	4.1668	8.9056	6.5362
Germany	15.7264	9.8172	9.8172	15.7264	12.7718
Greece	8.3268	9.8819	8.3268	9.8819	9.1044
Qatar	5.4074	2.9621	2.9621	5.4074	4.1847
South Africa	2.2966	1.3699	1.3699	2.2966	1.8332

We will use "Single Linkage" for this illustration

Grouping of Observations – Continues

With "Single Linkage" as the method, the new distance matrix turns out to be as follows:

	France	Germany	Greece	R-E	Qatar
Germany	3.4152				
Greece	11.2879	18.7006			
R-E	4.1668	9.8172	8.3268		
Qatar	6.8494	17.3189	16.8573	2.9621	
South Africa	3.5649	11.2114	4.8624	1.3699	3.8879

This time, we group "South Africa" with "R-E".

We continue this until all observations are grouped into one single cluster.

Visualizing the Clusters – Dendogram

We started with all countries in different clusters and ended in grouping all in one cluster.

What do we gain out of it?

We create a plot that graphically identifies **how different countries clubbed one by one** and **what was the distance** before clubbing.

Visualizing the Clusters – Dendogram

We started with all countries in different clusters and ended in grouping all in one cluster.
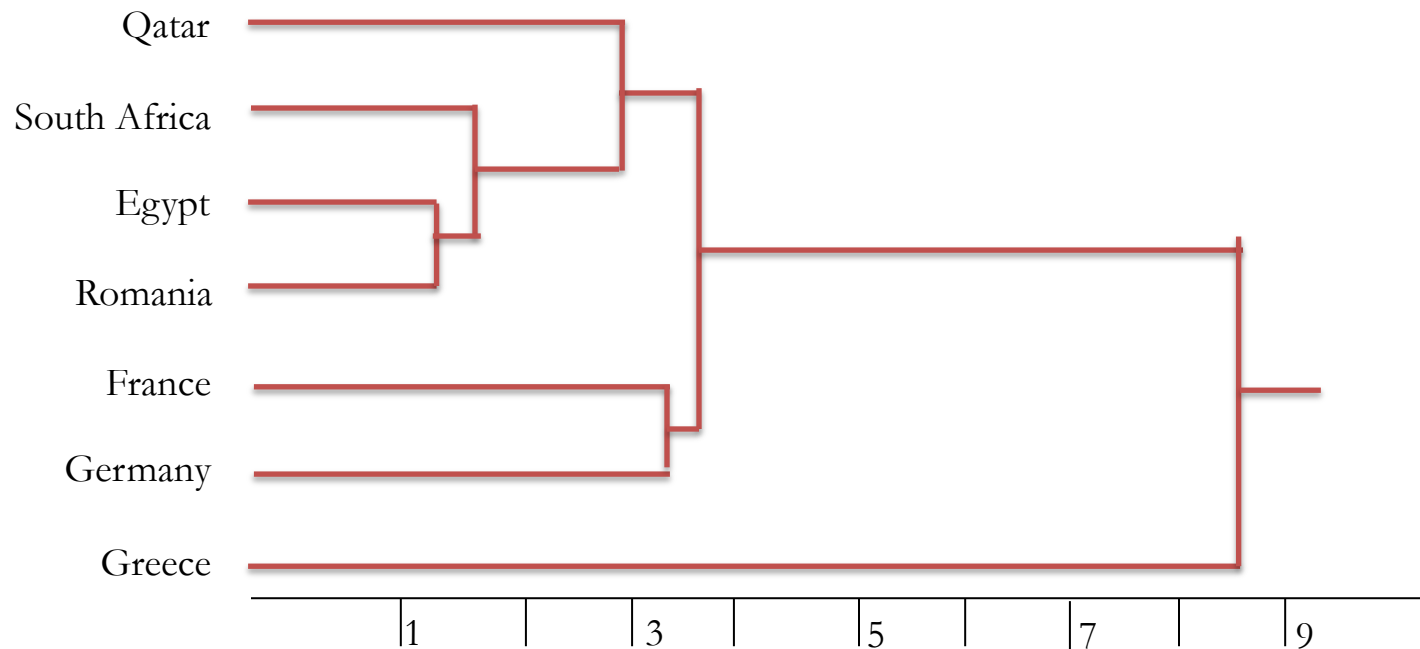
This plot is known as dendogram

Visualizing the Clusters – Dendogram

Now, we finalize

either on **number of clusters** or a **cut-off value of distance**
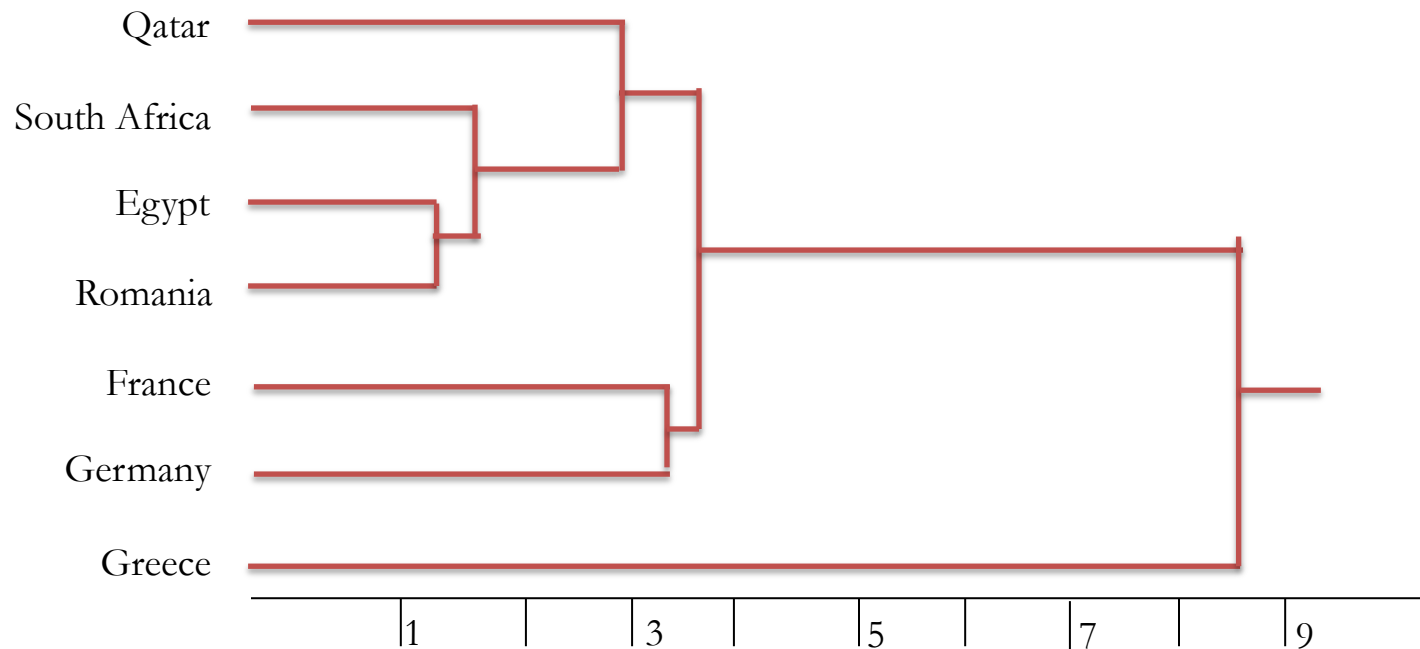
Visualizing the Clusters – Dendogram

Now, we finalize

either on **number of clusters** or a **cut-off value of distance**
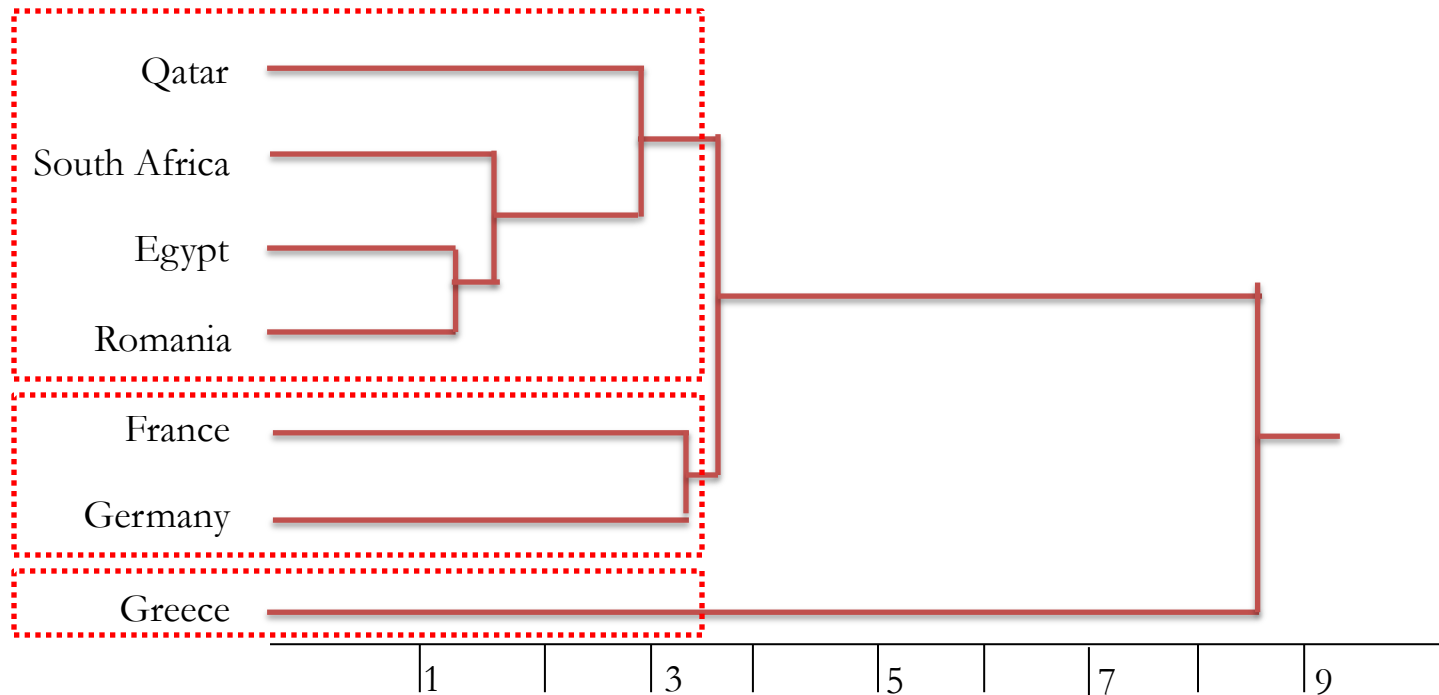
With a cut-off distance of 3.5 or 3 cluster solution gives us the above grouping

Building Clusters through R

Now, Raghu read the data for 20 countries through R and run the clustering.

The code is as follows:

```r
textile<-read.csv("Hierarchical Clustering.csv")
# Keeping only the numeric variables and scaling them
mydata<-scale(textile[,c(5:7)])

# Finding Distance
d<-dist(mydata, method="euclidian")

# Clustering
fit<-hclust(d, method="single")

# Dendogram plot and grouping into 4 groups
plot(fit, labels=textile$Partner.Name)
groups <- cutree(fit, k=4)
rect.hclust(fit, k=4, border="red")
```

Building Clusters through R

The result is as follows:

Cluster Dendrogram
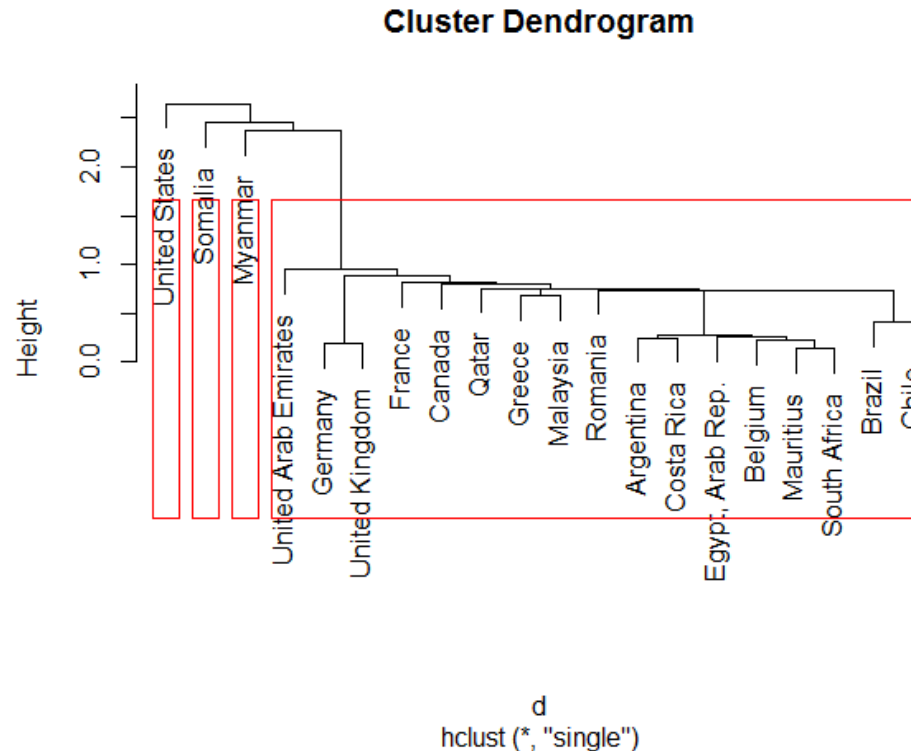
Hence, each of "United States", "Somalia" and "Myanmar" making one cluster and remaining 17 states forming another cluster

Advantages and Disadvantage

- ✓ **A method that slowly adds the observations into groups**
- ✓ Segments can be **visualized graphically**
- ✓ **A flexible method,** as number of clusters is often not known to us

- ✓ **Once a new observation comes**, we can not assign it any of the previously decided cluster. **The entire exercise needs to re-run**.

- ✓ **Computationally very heavy and can work only with limited number of observations (at most 5000)**

- ✓ Most of the scenarios, it **yields segments of disproportionate sizes**. For example, in the example before 17 countries (85% of population) went into one segment

Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation ✓

Segmentation through Hierarchical Clustering ✓

Segmentation through Non-hierarchical Clustering

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution

Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation ✓

Segmentation through Hierarchical Clustering ✓

Segmentation through Non-hierarchical Clustering

Profiling of Clusters

Building Cluster Equat...

Validation of Cluster Solution

Non-hierarchical Clustering – The Concept

Procedures

Step to Build K-means Clustering

Deciding on Number of Clusters

Criteria of a Good Cluster Solution

Non-hierarchical Clustering – sub-Agendum 6

Non-hierarchical Clustering – The Concept

Let us first try to visualize the countries with respect to the export data in a two dimensional plot
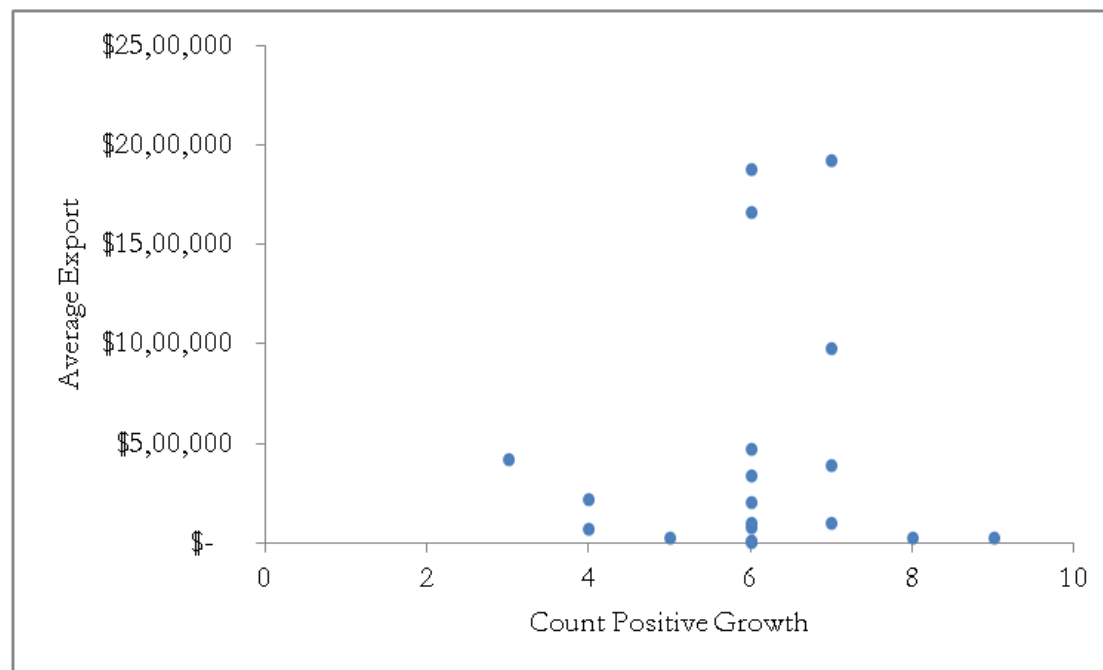
Now let's try to create 4 clusters ….

Non-hierarchical Clustering – The Concept

Probably we can create the segment in the following manner….
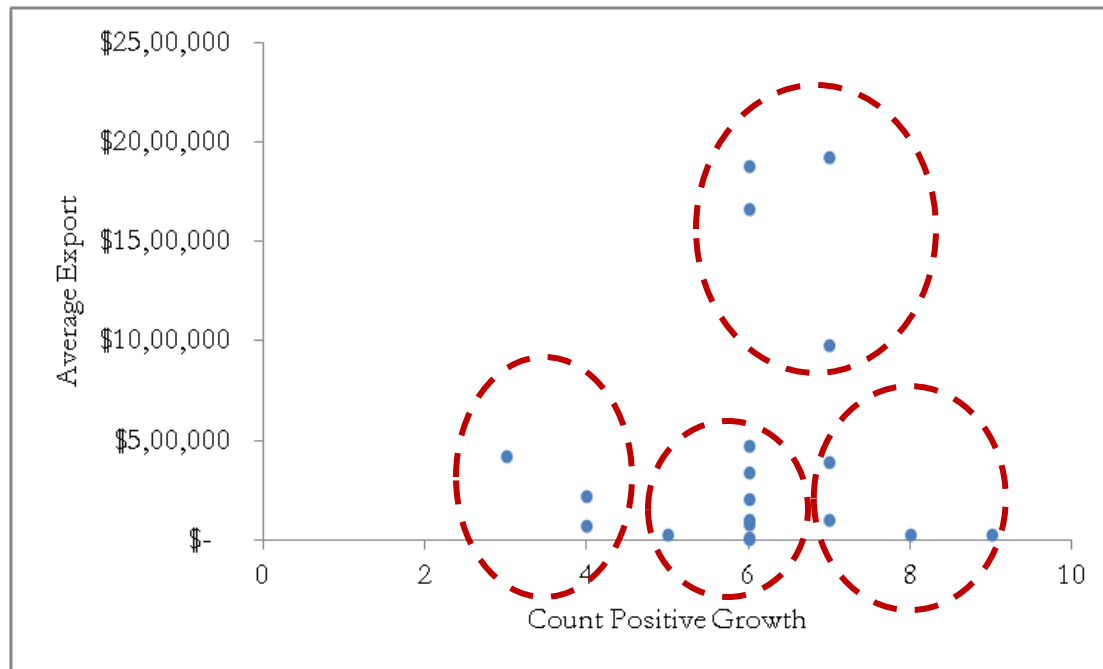
So we are finding 4 mutually exclusive zones on the graph.

Non-hierarchical Clustering – The Concept

This kind of segmentation is called **non-hierarchical clustering**.
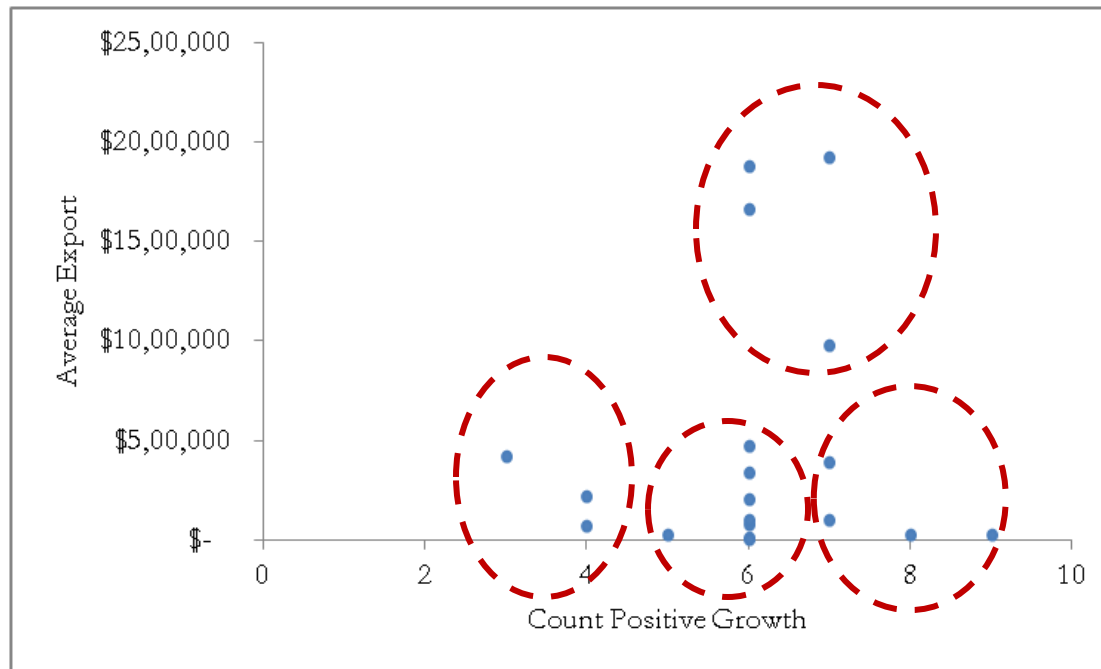
Because, as there is no hierarchy between any created groups and how the groups are formed.

Non-hierarchical Clustering – The Procedure

The procedure starts with number of clusters. Say we want 3 clusters.

The process starts with 3 random points on the co-ordinate as centroids for the 3 clusters

Randomly selected centroids
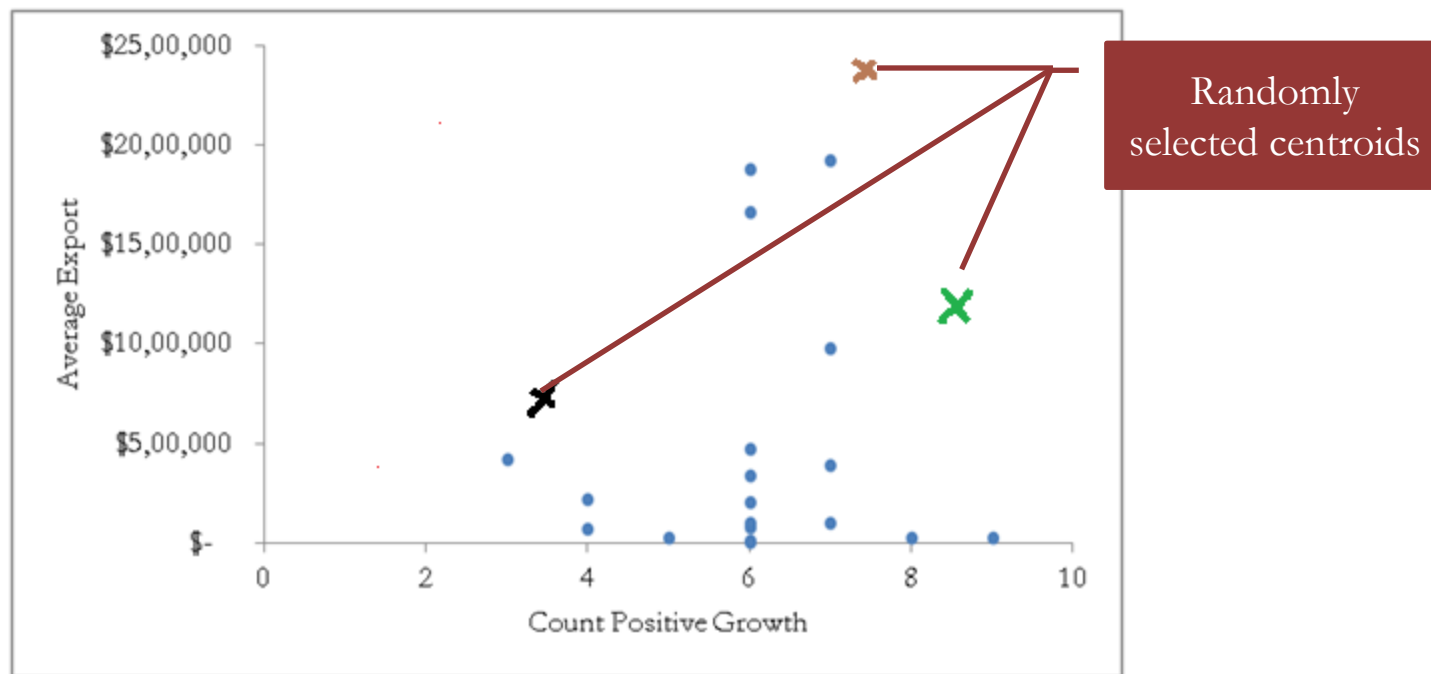
Non-hierarchical Clustering – The Procedure

Next based on distance from the centroids, each observation is assigned to one cluster.
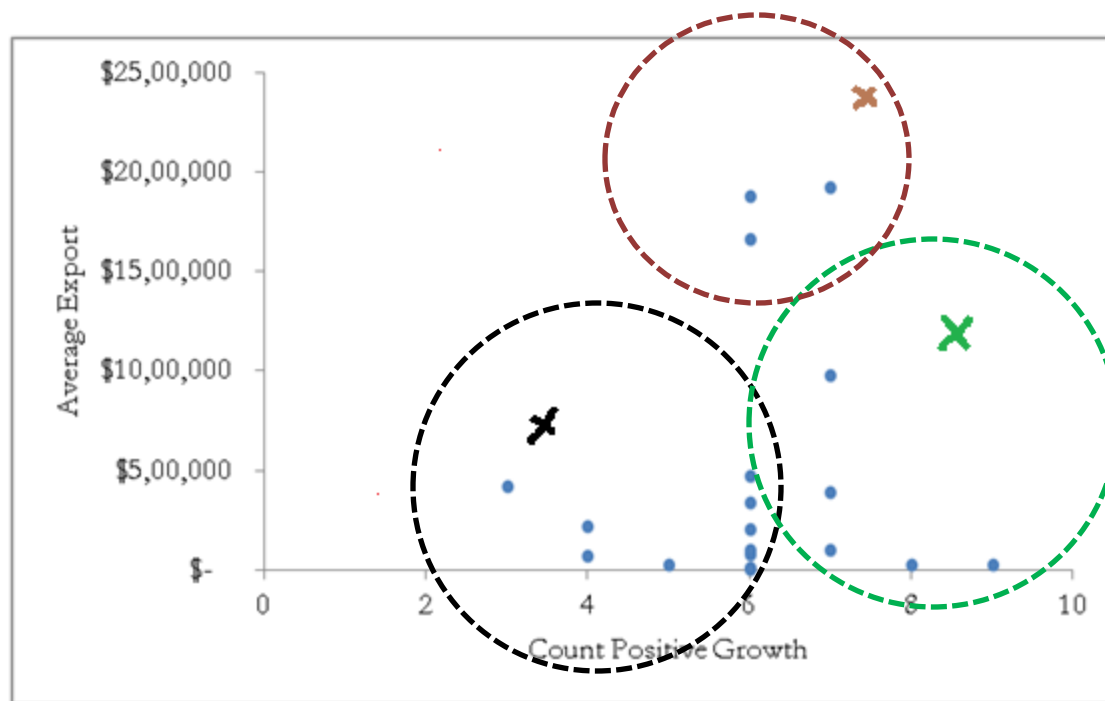
Non-hierarchical Clustering – The Procedure

Now we calculate the new centroids for the created clusters
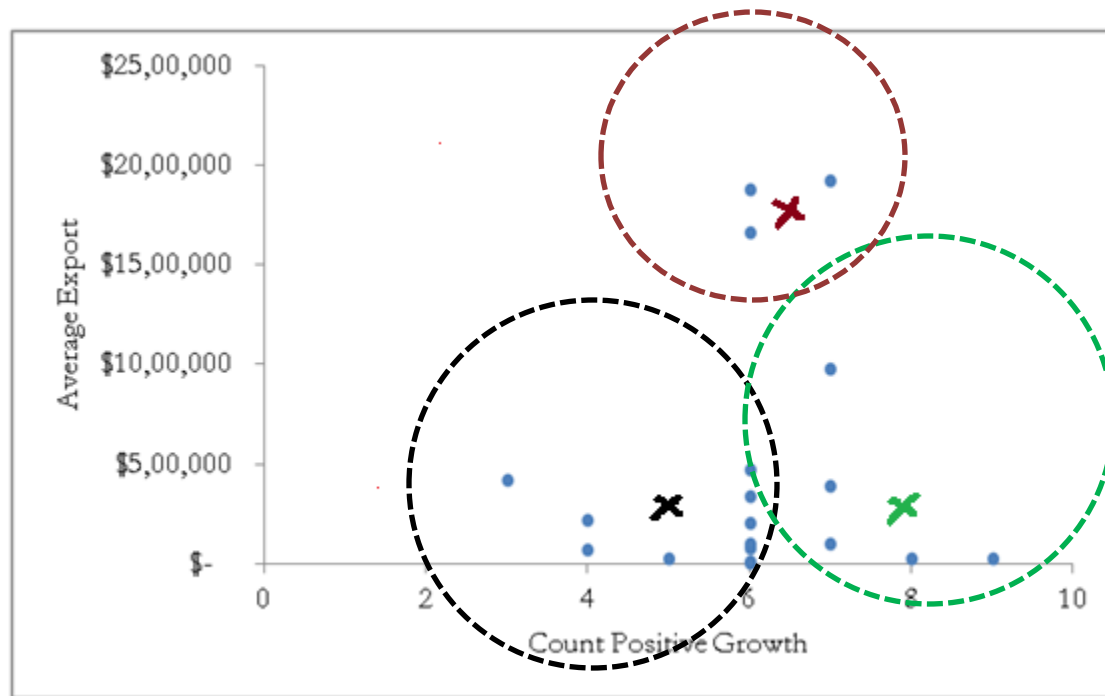
If the centroids are calculated based on mean of the coordinates, we call the techniques as **K Means clustering**.

Non-hierarchical Clustering – The Procedure

We again re-assign the observations based on new centroids.
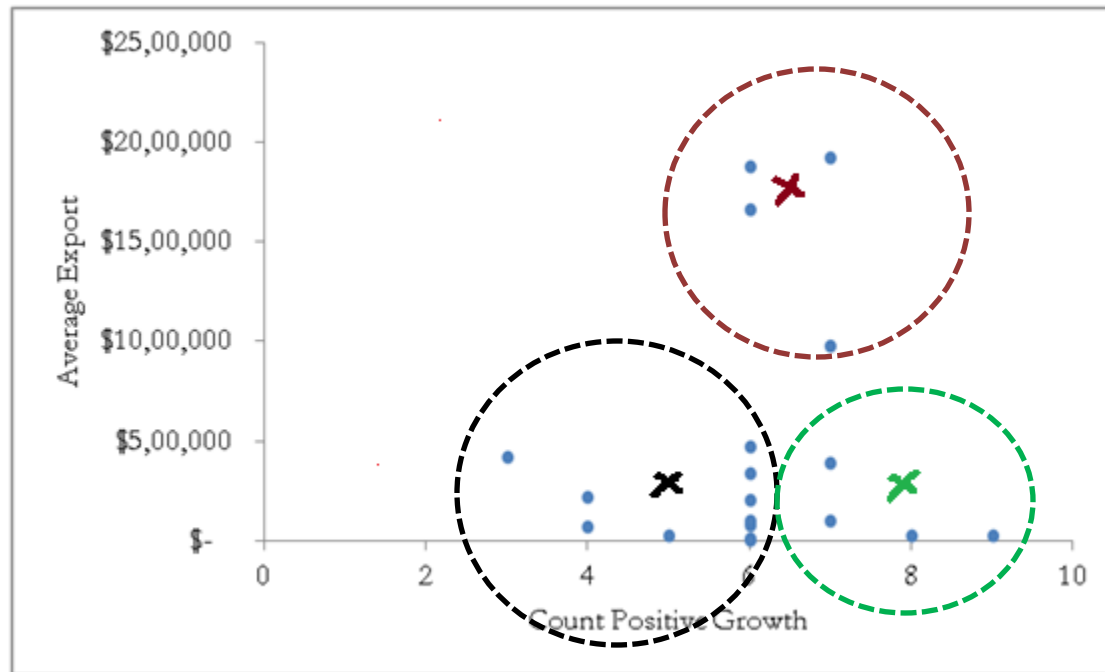
We repeat the step, until the new centroids are same as old centroids.

Steps to K-Means Clusters

```
┌──────────────┐      ┌──────────────┐      ┌──────────────┐
│  Select the  │ ───▶ │ Standardize  │ ───▶ │  Decide on   │
│  Variables   │      │ the Variables│      │  Number of   │
│              │      │              │      │   Clusters   │
└──────────────┘      └──────────────┘      └──────────────┘

┌──────────────┐      ┌──────────────┐      ┌──────────────┐
│  Build the   │ ───▶ │  Evaluate    │ ───▶ │  Build the   │
│   Cluster    │      │   Cluster    │      │   Cluster    │
│   Solution   │      │   Solution   │      │   Equation   │
└──────────────┘      └──────────────┘      └──────────────┘
```
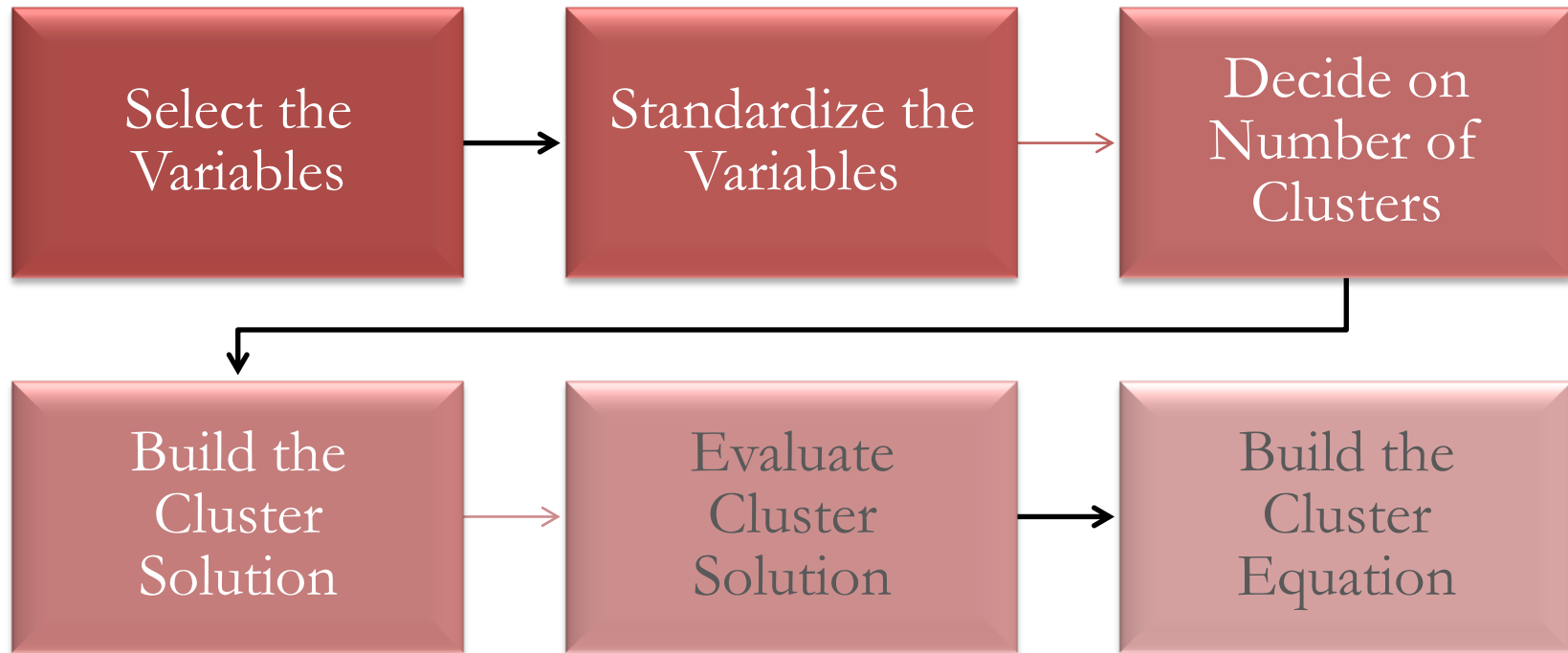
Credit Card Customer Segmentation– Illustration

BankOnUs is a retail bank that provides credit cards to different retail customers. It wants to understand the different transactional behavioral patterns of their customers.

Andrew, the analyst, has collected data on 30,000 active credit card holders. The data contains

- ✓ Purchase and utilization details on the cards (e.g. Balance, Number of transactons, Transactions under promotions, Purchase active months etc.)

- ✓ Payment details (e.g. Revolving amount, Online/Offline payment, etc.)

- ✓ External performance variables (e.g. number of credit cards, number of loans, etc.)

As **number of rows is very high**, he decided to go for K-Means clustering

Credit Card Customer Segmentation– Variable Selection

For, **K-Means segmentation we can use only numeric variables**.

Hence Andrew finalized on the following numeric variables

- ✓ Months on book
- ✓ Maximum balance amount in last 12 months
- ✓ Number of purchase active months in last 12 months
- ✓ Total finance charges billed
- ✓ Number of online transactions as a ratio of total transactions
- ✓ Average payment by balance ratio in last 12 months
- ✓ Number of offline payments as a ratio of total payments
- ✓ Number of active credit cards

Variable Scaling – Making the Variables Comparable

To make all variables comparable, like in hierarchical clustering, in K-Means clustering also we **standardized the variables**.

The R code for the same is as follow:

```
####### Data pull #####
custdata<-read.csv("customer segmentation.csv")

##### Development Validation Spitting #####
index <- c(1:nrow(custdata))
sampleIndex <- sample(index,nrow(custdata)*0.70,replace=FALSE)
custdata.dev <- custdata[sampleIndex,]
custdata.test <- custdata[-sampleIndex,]

##### Standaization of variables #####
smalldata<-custdata.dev[,c(-1)] # ID Variable is not required for
standardization
scaledata<-scale(smalldata)
```

Variable Scaling – Making the Variables Comparable

To make all variables comparable, like in hierarchical clustering, in K-Means clustering also we **standardized the variables**.

The R code for the same is as follow:

```
####### Data pull #####
custdata<-read.csv("customer segmentation.csv")

##### Development Validation Spitting #####
index <- c(1:nrow(custdata))
sampleIndex <- sample(index,nrow(custdata)*0.70,replace=FALSE)
custdata.dev <- custdata[sampleIndex,]
custdata.test <- custdata[-sampleIndex,]

##### Standaization of variables #####
smalldata<-custdata.dev[,c(-1)] # ID Variable is not required for
standardization
scaledata<-scale(smalldata)
```

Code to split into development and validation

Variable Scaling – Making the Variables Comparable

To make all variables comparable, like in hierarchical clustering, in K-Means clustering also we **standardized the variables**.

The R code for the same is as follow:

```
####### Data pull #####
custdata<-read.csv("customer segmentation.csv")

##### Development Validation Spitting #####
index <- c(1:nrow(custdata))
sampleIndex <- sample(index,nrow(custdata)*0.70,replace=FALSE)
custdata.dev <- custdata[sampleIndex,]
custdata.test <- custdata[-sampleIndex,]

##### Standaization of variables #####
smalldata<-custdata.dev[,c(-1)] # ID Variable is not required for
standardization
scaledata<-scale(smalldata)
```

Standardization of all numeric variables. We can also select only the variables we are interested in

Deciding on the Number of Clusters

Now, Andrew does not have any idea about how many clusters may be there in the population.

Charles, his boss, told that

Three or less segments are of no insights, as in this case he may not have good insights from such a less number of segments

Again, **10 or more segments are too many** as it is very difficult to design customized business strategies for so many segments.

Hence, Charles asked Andrew to build a **cluster solution that has number segments anything between 4 to 10**

But what is the ideal number of clusters?

Criterions for a Good Cluster Solution

A good cluster solution should have the following properties:

- ✓ The **smallest cluster** should not be too small (not less than 2-3% of population)
- ✓ The **largest cluster** should not be too big (not more than 35-40% of population)
- ✓ The **distance between the cluster centroids** should be as high as possible. In other words the means of each variable of clusters should be very different
- ✓ The cluster radius (the maximum distance between two points in a cluster) should be as low as possible. Alternatively we can calculate **within cluster variation as a percentage of total variation**, and this has to as low as possible
- ✓ **Between cluster variation** (as a percentage of total variation) should be as high as possible. In other words, total within cluster variation should be as low as possible

Based on the above criterions we decide the ideal number of clusters.

Finalising Cluster Solution – The R Code

Following is the R code for running K-means clustering

```
##### K-Means Cluster Analysis ####
fit <- kmeans(scaledata, center=5) # 5 cluster solution

##### Cluster Distribution #####
clustersize<-fit$size/nrow(scaledata)

##### Within Cluster Variation #####
withinvar<-fit$withinss/fit$totss # For each Cluster
wss<-fit$tot.withinss/fit$totss # Total

##### Cluster Centers #####
centroids<-fit$centers
```

Finalising Cluster Solution – Cluster Size

After different iterations, Andrew found the following distribution of customers across different clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
4 Cluster Solution	15.2%	64.9%	10.6%	9.2%						
5 Cluster Solution	9.2%	51.9%	10.7%	13.0%	15.2%					
6 Cluster Solution	26.3%	9.2%	15.1%	10.6%	12.7%	26.0%				
7 Cluster Solution	15.2%	9.2%	12.5%	18.9%	10.4%	15.0%	18.8%			
8 Cluster Solution	13.5%	15.0%	12.3%	10.1%	9.2%	16.5%	10.4%	13.0%		
9 Cluster Solution	13.2%	10.1%	12.3%	10.8%	9.2%	7.4%	7.6%	14.1%	15.2%	
10 Cluster Solution	12.4%	13.0%	7.4%	14.6%	5.6%	9.2%	7.7%	12.1%	5.3%	12.7%

For 4 cluster solution and 5 cluster solution, the biggest cluster is more that 50%. Hence, these two are ruled out.

Size of smallest cluster is not an issue, as the smallest cluster never less than 5%

Hence, any cluster solution between 6 cluster solution to 10 cluster solution is fine from cluster size perspective

Finalising Cluster Solution – Within Cluster Variation

Now Andrew looked at the within cluster variations.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
4 Cluster Solution	3.96%	22.68%	5.52%	4.78%						
5 Cluster Solution	4.74%	10.10%	5.55%	3.35%	3.97%					
6 Cluster Solution	4.01%	4.74%	3.89%	5.54%	3.23%	3.98%				
7 Cluster Solution	2.40%	4.74%	3.10%	2.40%	5.28%	3.82%	2.45%			
8 Cluster Solution	1.27%	3.78%	3.03%	5.00%	4.74%	1.95%	2.01%	1.61%		
9 Cluster Solution	1.54%	5.05%	3.01%	2.02%	4.73%	1.33%	1.42%	1.39%	1.78%	
10 Cluster Solution	3.05%	1.35%	1.32%	1.77%	2.06%	4.74%	1.43%	1.64%	2.08%	1.29%

10 cluster solution, 9 cluster solution or 8 cluster solutions– all 3 cases the minimum variance is almost same.

Hence by increasing number of clusters to 9 or 10 we are not gaining anything .

Hence, 9 cluster solution or 10 cluster solution is rejected as they are creating extra clusters without any benefit

Finalising Cluster Solution – Scree Plot

Alternatively, we can plot "Percentage of Total Within Cluster Variation" for different solutions.
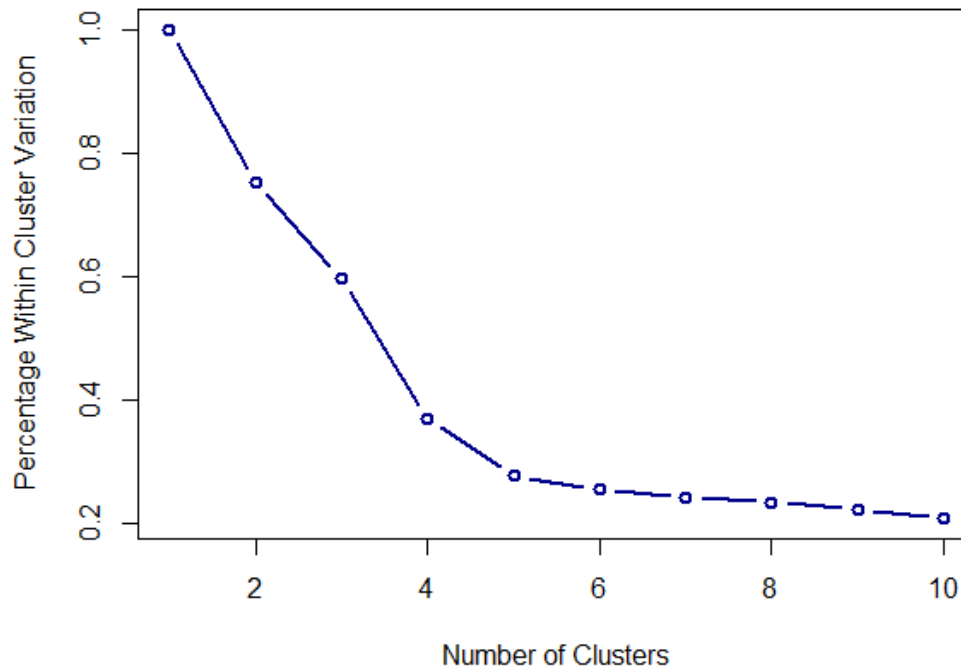
Finalising Cluster Solution – Scree Plot

Alternatively, we can plot "Percentage of Total Within Cluster Variation" for different solutions.
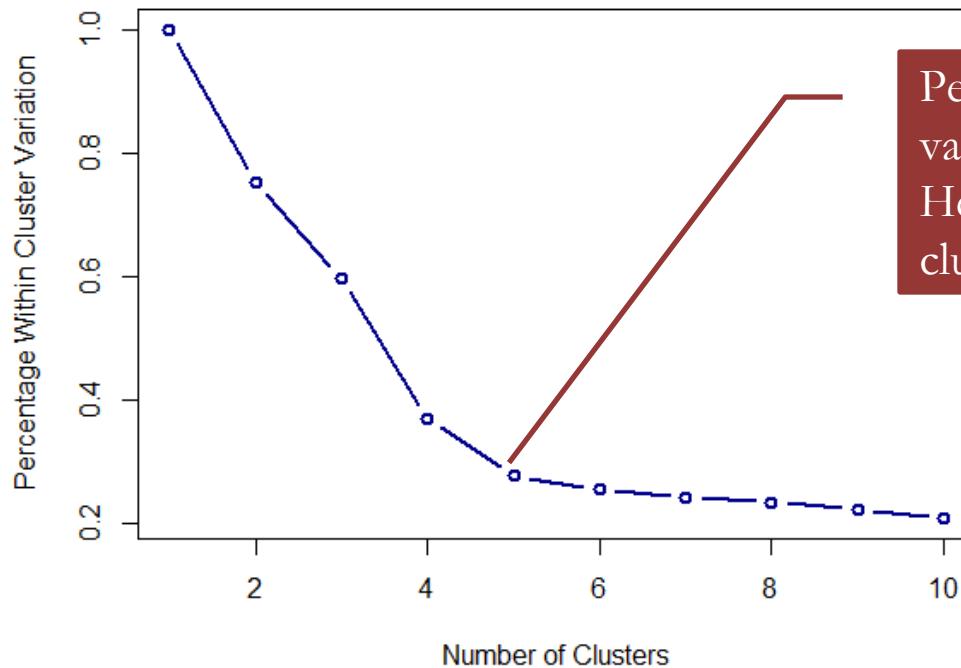
Percentage of Within cluster variation is ~27% with 5 clusters. Hence, 73% variation is between cluster

Finalising Cluster Solution – Scree Plot

Alternatively, we can plot "Percentage of Total Within Cluster Variation" for different solutions.
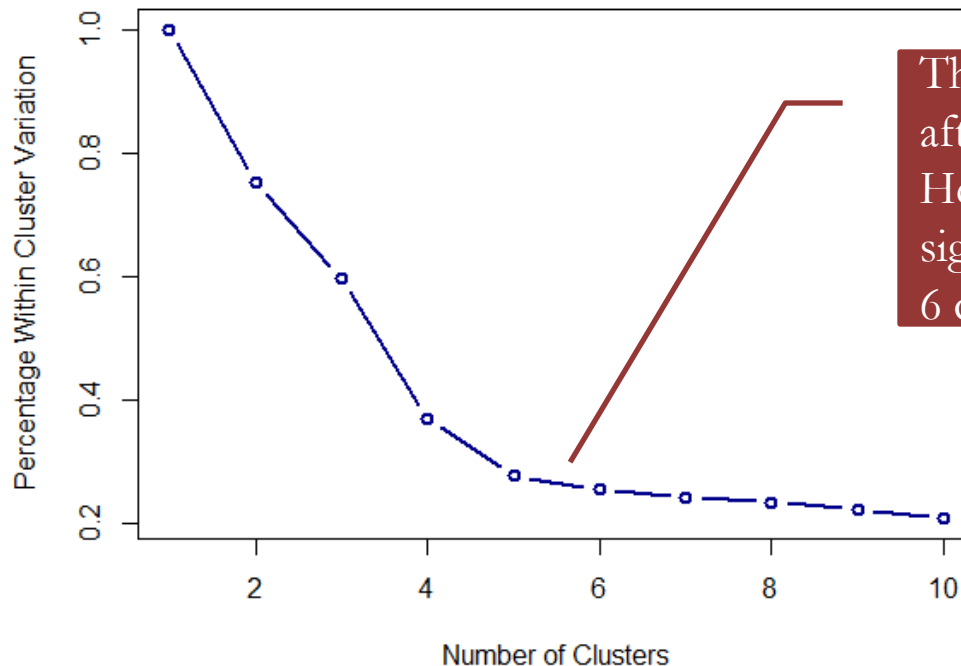
> The graph is becoming almost flat after 5 or 6 clusters.
> Hence, we are not gaining significantly by having more than 6 clusters

Hence, Andrew concluded with 6 cluster solution

Finalising Cluster Solution – Custer Centres

Now, we look at cluster centroids on the standardized data.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Months on Book	-0.43	0.38	0.78	1.09	-0.46	-0.37
Maximum Balance	-0.46	2.86	-0.20	-0.17	0.09	-0.41
Purchase Active Months	-0.65	1.36	-0.16	1.76	0.07	-0.49
Finance Charges	-0.41	2.93	-0.24	-0.37	0.08	-0.37
% Online Transactions	-0.22	0.04	-1.65	-0.10	0.37	1.02
Pay to Balance Ratio	-0.45	-0.48	0.00	2.50	-0.20	-0.30
% Offline Payment	-0.51	0.02	2.21	-0.16	-0.41	-0.51
# of Active Cards	-0.55	0.72	-0.18	-0.17	2.12	-0.56

Note, "Purchase active months" and "Months on book" variable is not very different across segment.

If we feel so, **we can repeat the previous steps by dropping those two variables.**

Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation ✓

Segmentation through Hierarchical Clustering ✓

Segmentation through Non-hierarchical Clustering ✓

Profiling of Clusters

Building Cluster Equation

Validation of Cluster Solution
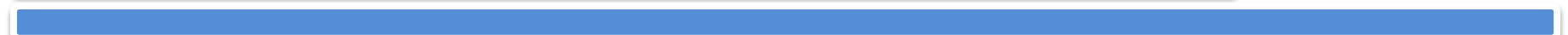
Cluster Analysis

Introduction to Segmentation	✓
Types of Segmentation	✓
Segmentation through Hierarchical Clustering	✓
Segmentation through Non-hierarchical Clustering	✓

Profiling of Clusters

Profiling of Clusters

Building Cluster Equation

Finding Key Insights – Illustration

Validation of Cluster Solution

Profiling of Clusters – Why It Is Important

Now, Andrew is done with the segmentation. But, has he been able to draw any inference out of it?

No… Still the segments are just some meaningless numbers to his boss.

Hence, he needs to profile the clusters with respect to the original variables.

Profiling is looking the properties of the clusters are trying to find key insights from the properties

Profiling of Clusters – Doing in R

We use the following R code to do the profiling:

```
##### Standaization of variables #####
smalldata<-custdata.dev[,c(-1)]
scaledata<-scale(smalldata)

##### K-Means Cluster Analysis ####
fit <- kmeans(scaledata, center=6) # 6 cluster solution

##### Get Cluster Tagging to Original Data ####
summary<-aggregate(custdata.dev, by=list(fit$cluster)
                ,FUN=mean)  # Get Cluster Means

##### Get Overall Mean ####
tot.summary<-sapply(custdata.dev, mean, na.rm=TRUE)
```

Getting cluster tagging to original data. Then calculating mean by each cluster

Profiling of Clusters – The Result

Once compared the means, we get the following tables.

	Total	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Months on Book	144.75	122.67	164.45	184.84	200.97	121.10	125.47
Maximum Balance	$ 355.59	$ 169.98	$ 1,515.53	$ 274.77	$ 287.13	$ 393.89	$ 189.58
Purchase Active Months	4.07	2.34	7.72	3.63	8.79	4.26	2.75
Finance Charges	$ 38.11	$ 9.20	$ 245.20	$ 20.85	$ 12.14	$ 43.96	$ 11.96
% Online Transactions	40.05%	36.28%	40.81%	11.62%	38.27%	46.35%	57.71%
Pay to Balance Ratio	22.93%	15.47%	14.99%	23.00%	64.10%	19.58%	18.05%
% Offline Payment	7.97%	1.33%	8.25%	36.71%	5.88%	2.62%	1.40%
# of Active Cards	3.46	2.15	5.17	3.02	3.04	8.52	2.12

Can we find any insights now? …..What are these segments?

Profiling of Clusters – The Result

We tried to compare with respect to over all mean and found the following legends

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Months on Book	Mid-Low	Medium	High	Mid-High	Mid-Low	Mid-Low
Maximum Balance	Low	High	Mid-Low	Mid-Low	Medium	Low
Purchase Active Months	Low	High	Mid-Low	High	Medium	Low
Finance Charges	Low	High	Low	Low	Mid-High	Low
% Online Transactions	Medium	Medium	Low	Medium	Mid-High	Mid-High
Pay to Balance Ratio	Low	Low	Medium	High	Mid-Low	Mid-Low
% Offline Payment	Low	Medium	High	Mid-Low	Low	Low
# of Active Cards	Low	Mid-High	Mid-Low	Mid-Low	High	Low

Where legends are:

- 🔴 Low (<70% of population mean)
- 🟤 Mid-Low (70-90% of population mean)
- 🟡 Medium (90-115% of population mean)
- 🟢 Mid-High (115-150% of population mean)
- 🟢 High (>150% population mean)

Finding Key Insights – Illustration

Looking at the comparison, Andrew could find all these insights:

Key Observations   **Name Given**

Cluster 1
- Lowest maximum balance, only $170, less than half of population average
- But payment to balance ratio is only 15.48% compared to 23% as population average
- Primarily offline transactions (~64%) but online payment (only ~98%)
- Lowest activity (active on 2.34 months on average)

Low Spending Revolvers

Cluster 2
- Highest maximum balance, $1.5K, close to 3 times of population average
- Highest finance charge paid a year($245), ~6.5 times of population average
- On average purchase active on 7 months in a year

High Spenders

Finding Key Insights – Illustration

Key Observations

Name Given

Cluster 3

- Highest off line payment (35.71% cases) 4.5 times of population average
- Lowest online transactions, only 11.6%
- Average finance charge paid is $21, almost half of population average
- Payment balance ratio is same as population average

Offline Traditional Spenders

Cluster 4

- High payment to balance ration.
- Only $12 finance charge a year, one third of population average
- Active on ~9 month on average
- But, maximum balance is on lower side $287

Active Transactors

Finding Key Insights – Illustration

Key Observations   **Name Given**

Cluster 5

- Highest number of bank cards (8.5 per individual on average)
- New customers as lowest MOB
- Low on both maximum balance and payment to balance ratio

Credit Hungry New Comers

Cluster 6

- Purchase active only once in 4 months (purchase active month 2.75)
- Highest online transactions
- Pays mostly through online (~1.4% offline payment)
- Payment to balance ratio is low, only 18%

Tech-Savvy Occasional Users

Finding Key Insights – The Final Look
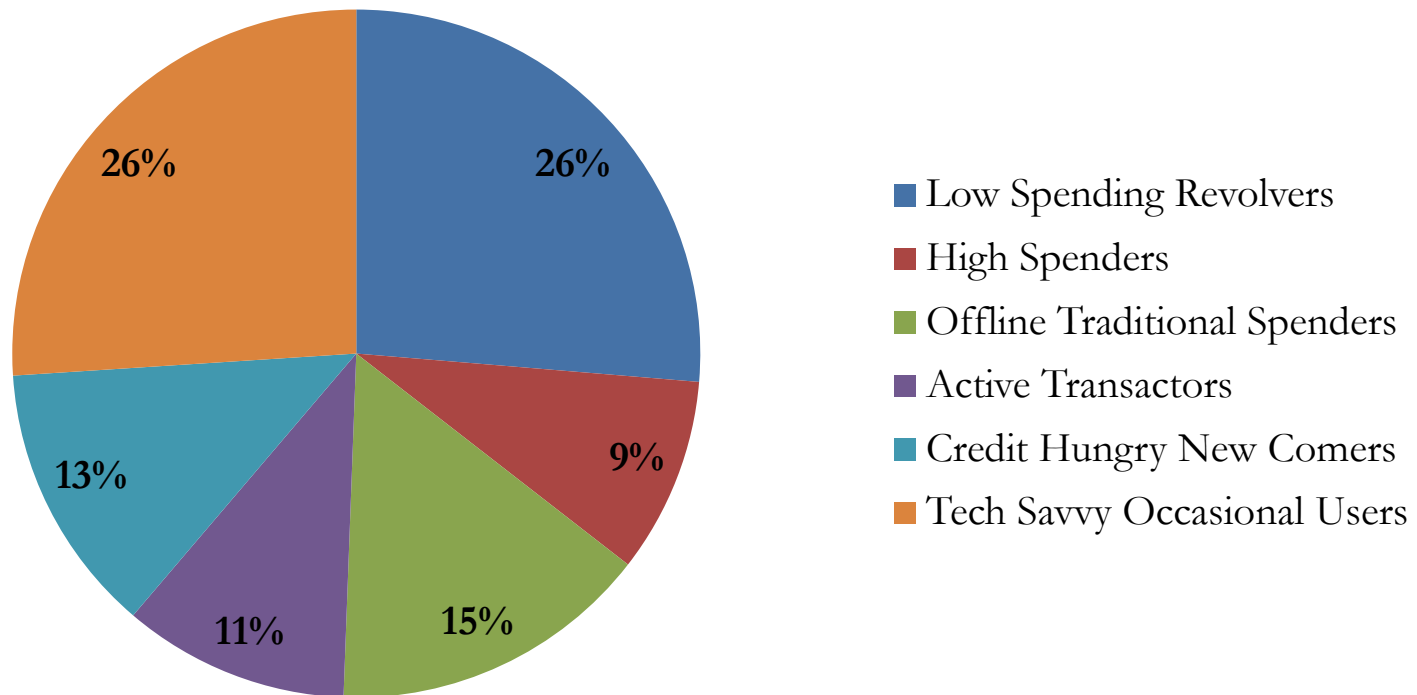
Legend:
- Low Spending Revolvers
- High Spenders
- Offline Traditional Spenders
- Active Transactors
- Credit Hungry New Comers
- Tech Savvy Occasional Users

Pie chart values: 26%, 9%, 15%, 11%, 13%, 26%

Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation ✓

Segmentation through Hierarchical Clustering ✓

Segmentation through Non-hierarchical Clustering ✓

Profiling of Clusters ✓

Building Cluster Equation

Validation of Cluster Solution
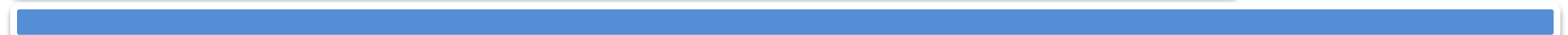
Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation ✓

Segmentation through Hierarchical Clustering ✓

Segmentation through Non-hierarchical Clustering ✓

Profiling of Clusters ✓

Building Cluster Equation

Validation of Cluster Solution

Assigning Cluster to a New Observation

Cluster Equation – Steps

Building Cluster Equation – Illustration

© insAnalytics

Assigning Cluster to a New Observation

Once the cluster solution is finalized, clusters are profiled then comes the question that

How do we know which observation falls in which cluster?

For the observations in the development sample, we know where do they fall.

But what happens when a new observations comes, for example a new customer comes on book on BankOnUs?

Hence, **we need a mathematical rule** that assigns every observation to any one of the identified clusters.

Cluster Equation – Steps

Here is the logic of building the logic.

Step1: Standardize the variables through the sample mean and standard deviation of development sample

Step2: Find the centroids for each cluster

Step3: Find the distance of the observations from cluster centroids (from development sample)

Step4: Find the minimum of all distances. The assigned cluster is the cluster for which the centroid is closest to the observation

Building Cluster Equation – Illustration

Here is how Andrew created the cluster equation to get the clusters on validation sample

Step1: Andrew calculated mean and standard deviation of variables from development sample and standardized them

```
##### Mean of variables #####
means<-sapply(custdata.dev,mean, na.rm=TRUE)

##### SD of variables #####
sds<-sapply(custdata.dev,sd, na.rm=TRUE)

##### Standardization of Month On Book ####
custdata.test$scale_mob<-(custdata.test$MOB-144.74862591)/51.4559529
```

Building Cluster Equation – Illustration

Step2: Andrew calculated centroids for 6 clusters

```
##### Getting centroids of clusters #####
summary<-aggregate(as.data.frame(scaledata),by=list(fit$cluster),FUN=mean)
```

Building Cluster Equation – Illustration

Step2: Andrew calculated centroids for 6 clusters

```
##### Getting centroids of clusters #####
summary<-aggregate(as.data.frame(scaledata),by=list(fit$cluster),FUN=mean)
```

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Months on Book	-0.43	0.38	0.78	1.09	-0.46	-0.37
Maximum Balance	-0.46	2.86	-0.20	-0.17	0.09	-0.41
Purchase Active Months	-0.65	1.36	-0.16	1.76	0.07	-0.49
Finance Charges	-0.41	2.93	-0.24	-0.37	0.08	-0.37
% Online Transactions	-0.22	0.04	-1.65	-0.10	0.37	1.02
Pay to Balance Ratio	-0.45	-0.48	0.00	2.50	-0.20	-0.30
% Offline Payment	-0.51	0.02	2.21	-0.16	-0.41	-0.51
# of Active Cards	-0.55	0.72	-0.18	-0.17	2.12	-0.56

Building Cluster Equation – Illustration

Step3: Andrew calculated distances from each cluster

```
##### Getting distance from Cluster 1#####
custdata.test$d1<- sqrt((custdata.test$scale_mob+0.4290699)**2
                       +(custdata.test$scale_maxbal+0.45739706)**2
                       +(custdata.test$scale_puract+0.6460292)**2
                       +(custdata.test$scale_fc+0.40889449)**2
                       +(custdata.test$scale_ot+0.21825897)**2
                       + (custdata.test$scale_pb+0.453948665)**2
                       +(custdata.test$scale_op+0.51125691)**2
                       +(custdata.test$scale_ac+0.5479844)**2

)
```

Building Cluster Equation – Illustration

Step4: Finally he found out cluster for each observation

```
##### Finding the Minimum Distance #####
custdata.test$mind<-min(custdata.test$d1,custdata.test$d2,custdata.test$d3,
                        custdata.test$d4,custdata.test$d5,custdata.test$d6)
##### Assigning the Clusters #####
custdata.test$cluster<-0
custdata.test$cluster[custdata.test$mind==custdata.test$d1]<-1
custdata.test$cluster[custdata.test$mind==custdata.test$d2]<-2
custdata.test$cluster[custdata.test$mind==custdata.test$d3]<-3
custdata.test$cluster[custdata.test$mind==custdata.test$d4]<-4
custdata.test$cluster[custdata.test$mind==custdata.test$d5]<-5
custdata.test$cluster[custdata.test$mind==custdata.test$d6]<-6
```

Cluster Analysis

Introduction to Segmentation	✓
Types of Segmentation	✓
Segmentation through Hierarchical Clustering	✓
Segmentation through Non-hierarchical Clustering	✓
Profiling of Clusters	✓
Building Cluster Equation	✓
Validation of Cluster Solution	

© insAnalytics

Data Mining
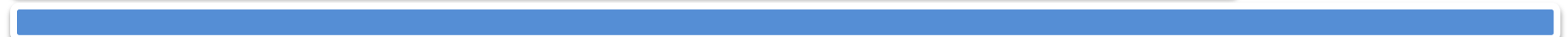
Cluster Analysis

Introduction to Segmentation	✓
Types of Segmentation	✓
Segmentation through Hierarchical Clustering	✓
Segmentation through Non-hierarchical Clustering	✓
Profiling of Clusters	✓
Building Cluster Equation	✓
Validation of Cluster Solution	

Validation of Cluster Solution

Validation of cluster solution can be done in two way

Option 1:

Use the same set of variables and runs cluster analysis. Check if the solution is comparable.

Note, " Cluster 1" in development sample may not be "Cluster 1" in validation sample. But after segmenting validation sample, **we should get similar segment** (in size, radius, and profile) of "Cluster 1" in development.

This is because K-means is an iterative approach

Validation of Cluster Solution

Validation of cluster solution can be done in two way

Option 2:

Implement the clustering equation in validation sample and then profile the clusters. Check if the profiles are comparable with that of development sample

Validation of Cluster Solution – Illustration

After implementing the cluster equation, Andrew got the clusters for each observations in validation sample. (Described in previous section)

Then he profiled the segment from validation data.

```
##### Finding the Cluster Size in Validation Data #####
table(custdata.test$cluster)

##### Profiling the Clusters Again #####
library(doBy)
summary_val<-
  summaryBy(MOB+MaxBalanceAmt+CntPurchActMth+TotFinCharge+PctOnlineTrans
        +PayToBalRatio+PctOfflinePymt+CntActiveCards~cluster,
        data=custdata.test, FUN = function(x)
        { avg=round(mean(x),digits=5)
        })
```

Validation of Cluster Solution – Illustration

We first look at the cluster distribution

- Low Spending Revolvers
- High Spenders
- Offline Traditional Spenders
- Active Transactors
- Credit Hungry New Comers
- Tech Savvy Occasional Users
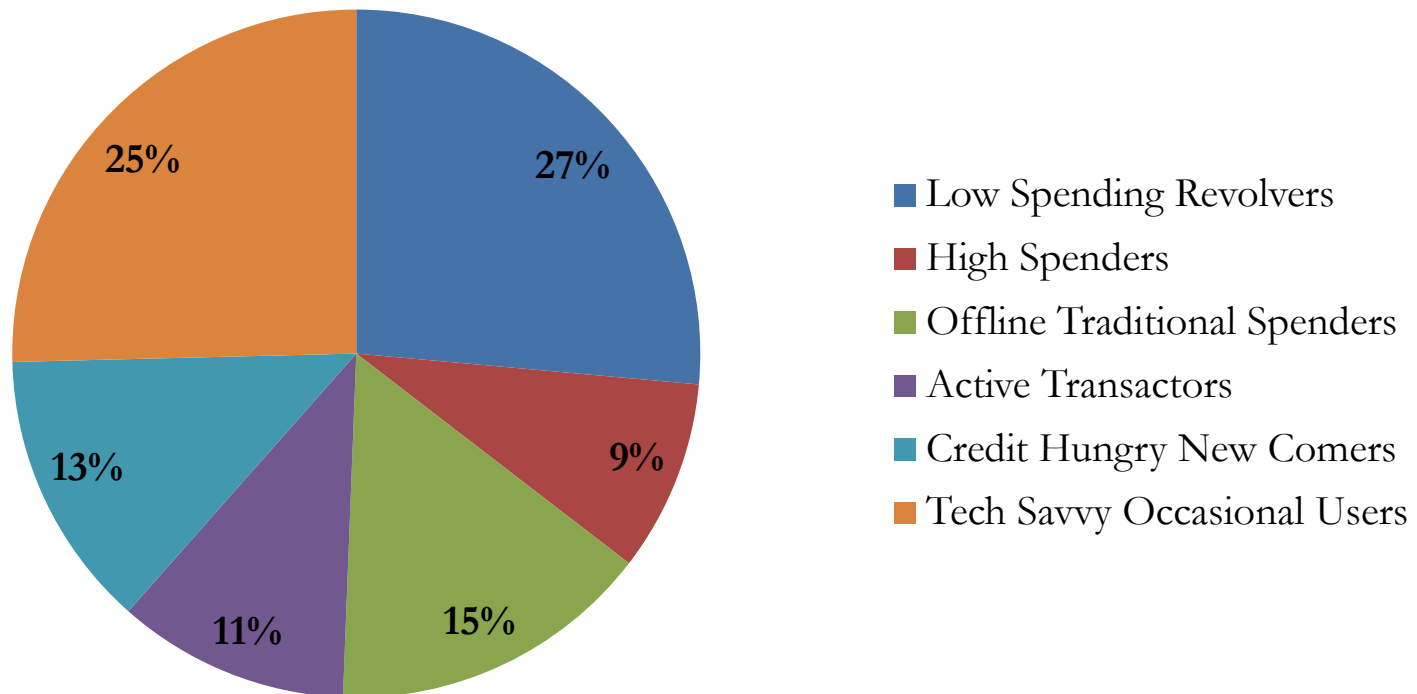
The cluster distribution is almost same as that from development sample

Validation of Cluster Solution – Illustration

Next we look at the cluster profile and compare with that of development sample

	Cluster 1		Cluster 2		Cluster 3	
	Development	**Validation**	**Development**	**Validation**	**Development**	**Validation**
Months on Book	122.67	123.42	164.45	161.99	184.84	185.51
Maximum Balance	$ 169.98	$ 170.31	$ 1,515.53	$ 1,522.59	$ 274.77	$ 272.62
Purchase Active Months	2.34	2.36	7.72	7.77	3.63	3.55
Finance Charges	$ 9.20	$ 9.48	$ 245.20	$ 242.14	$ 20.85	$ 20.71
% Online Transactions	36.28%	36.20%	40.81%	40.72%	11.62%	11.57%
Pay to Balance Ratio	15.47%	15.89%	14.99%	15.34%	23.00%	23.00%
% Offline Payment	1.33%	1.33%	8.25%	8.31%	36.71%	37.10%
# of Active Cards	2.15	2.15	5.17	5.18	3.02	3.06

Validation of Cluster Solution – Illustration

Next we look at the cluster profile and compare with that of development sample

	Cluster 4		Cluster 5		Cluster 6	
	Development	Validation	Development	Validation	Development	Validation
Months on Book	200.97	198.76	121.10	119.74	125.47	125.33
Maximum Balance	$ 287.13	$ 289.41	$ 393.89	$ 396.34	$ 189.58	$ 185.72
Purchase Active Months	8.79	8.86	4.26	4.29	2.75	2.70
Finance Charges	$ 12.14	$ 11.93	$ 43.96	$ 44.18	$ 11.96	$ 11.79
% Online Transactions	38.27%	38.57%	46.35%	46.41%	57.71%	57.88%
Pay to Balance Ratio	64.10%	62.57%	19.58%	19.42%	18.05%	17.67%
% Offline Payment	5.88%	6.02%	2.62%	2.57%	1.40%	1.34%
# of Active Cards	3.04	3.07	8.52	8.52	2.12	2.13

We find that the profiles of the clusters has not changed in validation sample.

Hence, we conclude that the segmentation has got validated in validation sample

Cluster Analysis

Introduction to Segmentation ✓

Types of Segmentation ✓

Segmentation through Hierarchical Clustering ✓

Segmentation through Non-hierarchical Clustering ✓

Profiling of Clusters ✓

Building Cluster Equation ✓

Validation of Cluster Solution ✓

Recapitulation & Key Takeaways

- ❑ In segmentation analysis, we divide the observations into **mutually exclusive and exhaustive distinct identifiable homogeneous groups**.
- ❑ When we segment a population based on one target variable, then it is called **Objective Segmentation**. Hence, it is a supervised learning and is more popularly known as **Decision Trees**
- ❑ When we segment a population based on all the relevant variables we have rather than one target variable, then it is called **Subjective Segmentation**. Hence, it is a unsupervised learning and is more popularly known as **Cluster Analysis**
- ❑ In **hierarchical clustering**, we group different observations **one by one based on the distance between two observations**. Different linkage methods are available, viz. simple, complete, average, etc.
- ❑ **Dendogram** helps us to visualize the hierarchical clustering
- ❑ In **k-means clustering**, we create some boundaries based on the variables and divide into k clusters
- ❑ We can use **SCREE plot** to decide the optimal number of clusters.
- ❑ We decide on the cluster solution based on the following criteria:
  - ❑ The **smallest cluste**r should not be less than 2-3% of the population
  - ❑ The **biggest cluster** should not be more than 40-45% of the population
  - ❑ The **percentage of within cluster variation** should be as high as possible
  - ❑ The **means of the segmenting variables** should be significantly different
- ❑ Once cluster solution is finalized, we **profile the clusters** to find insight from the clusters
- ❑ Once segments are profiled, we build **cluster equation** that helps a new observation to be classified into any one of the decided clusters

Data Mining

Cluster Analysis

insAnalytics

Consulting. Training. Research & Development.

insAnalytics

Insights.. Through Analytics…