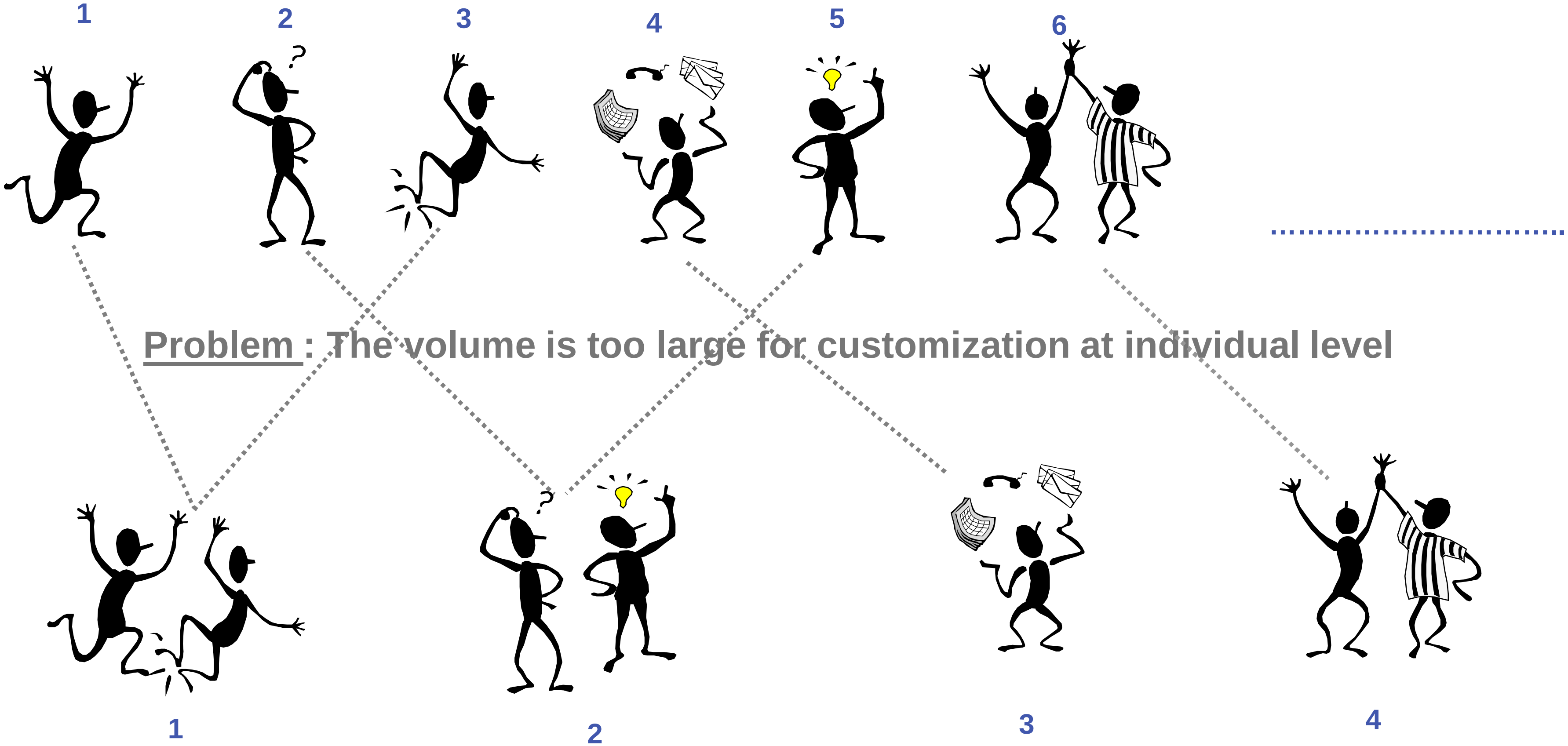
Why Segmentation ?

Each individual is so different
that ideally we would want to reach out to each one of them in a different way



<u>Problem</u> : The volume is too large for customization at individual level

<u>Solution</u> : Identify segments where people have same characters and target each of these segments in a different way

Approach to Segmentation

Segmentation is of 2 types

Objective Segmentation

Clear Objective to divide population

- Response rate
- Increase in Sales
- Conversion proportion

Objective defined Analysis. To identify the desired segment within population. Then devising strategy to tap the potential within.

↓

CHAID Analysis

Subjective Segmentation

First level analysis to see what lies within

- Who are my customers?
- Who buys what?
- When do they buy?

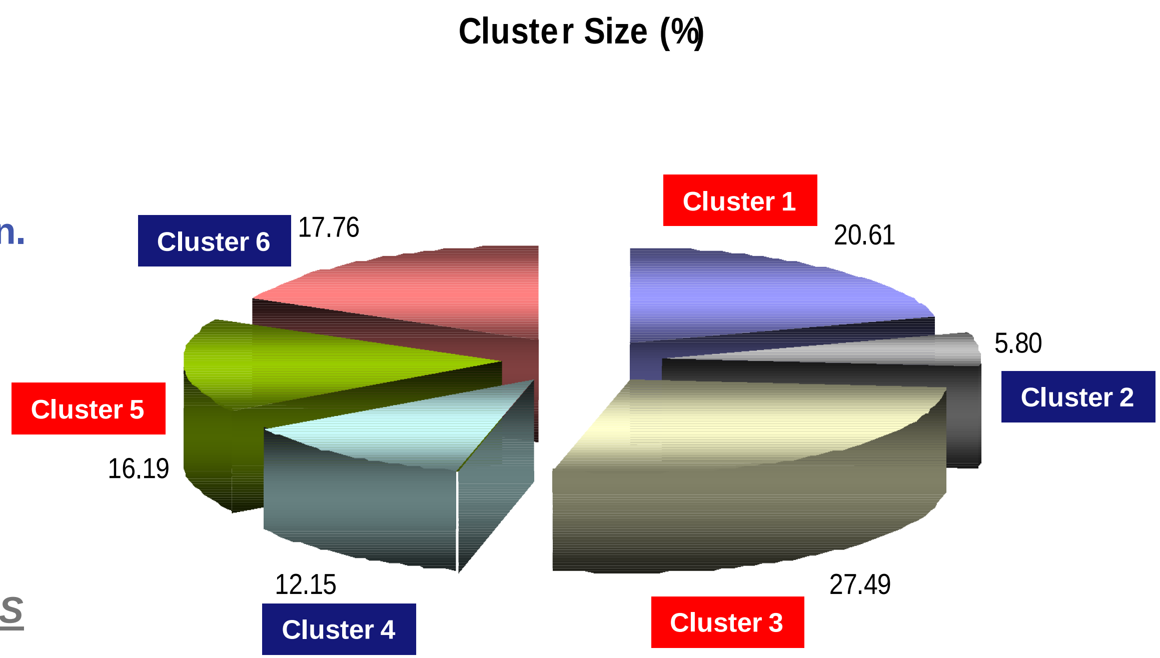
Initial Analysis to Understand & Define the Population. Based on the initial understanding – Objective Based Analysis.

↓

Cluster Analysis

Cluster Analysis

What are Clusters ?



Clusters are groups within a Population.

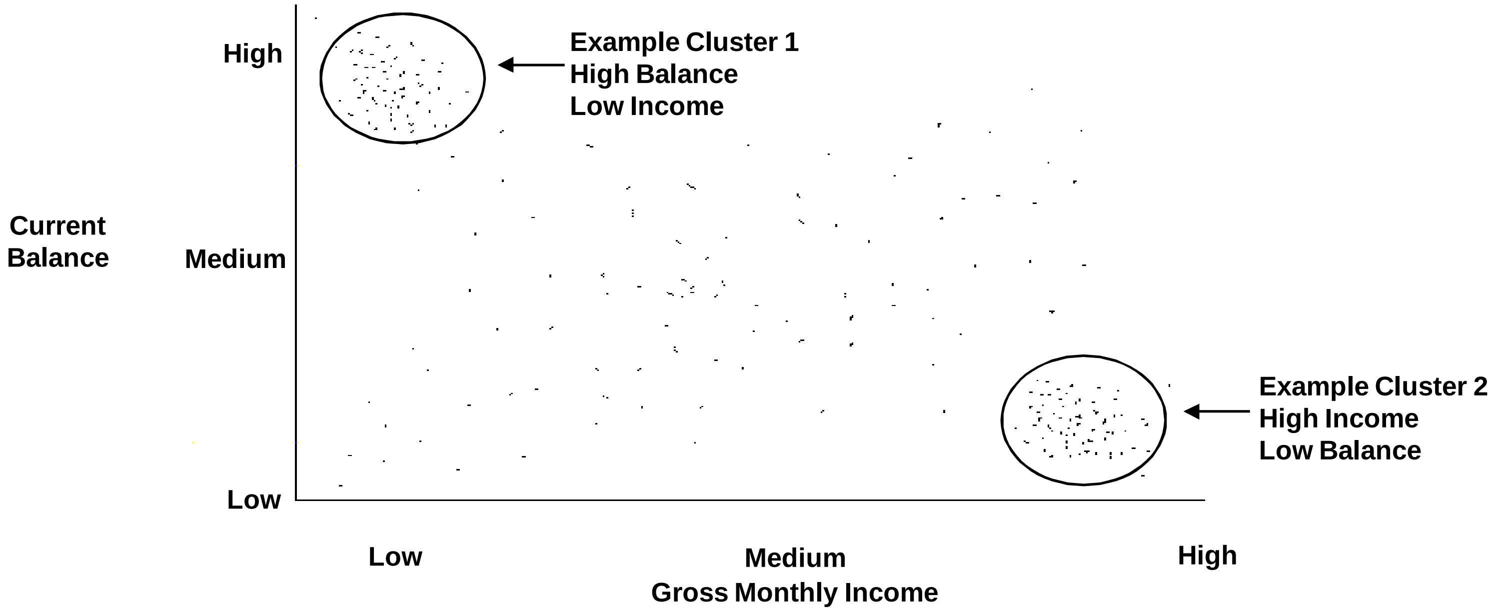
These Groups are ***HOMOGENEOUS*** within themselves.

And these groups are ***HETEROGENOUS*** among each other.

Homogeneous segments making it possible to group people of similar characteristics.

Heterogeneous among themselves making it possible to differentiate segments within population.

Example of Clusters

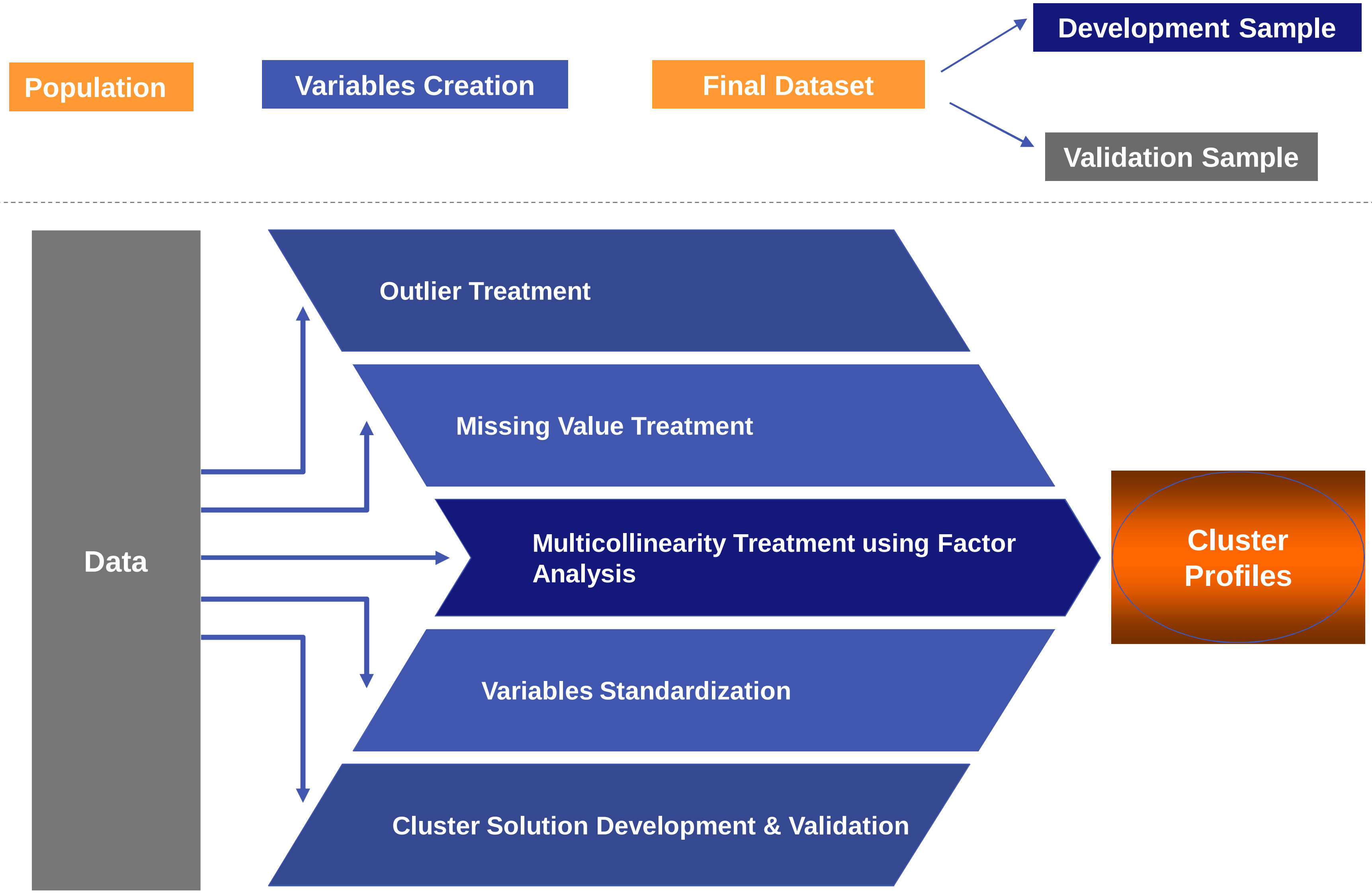


Cluster 1 and Cluster 2 are being differentiated by Income and Current Balance. The objects in Cluster 1 have similar characteristics (High Income and Low balance), on the other hand the objects in Cluster 2 have the same characteristic (High Balance and Low Income).

But there are much differences between an object in Cluster 1 and an object in Cluster 2.

Cluster Methodology

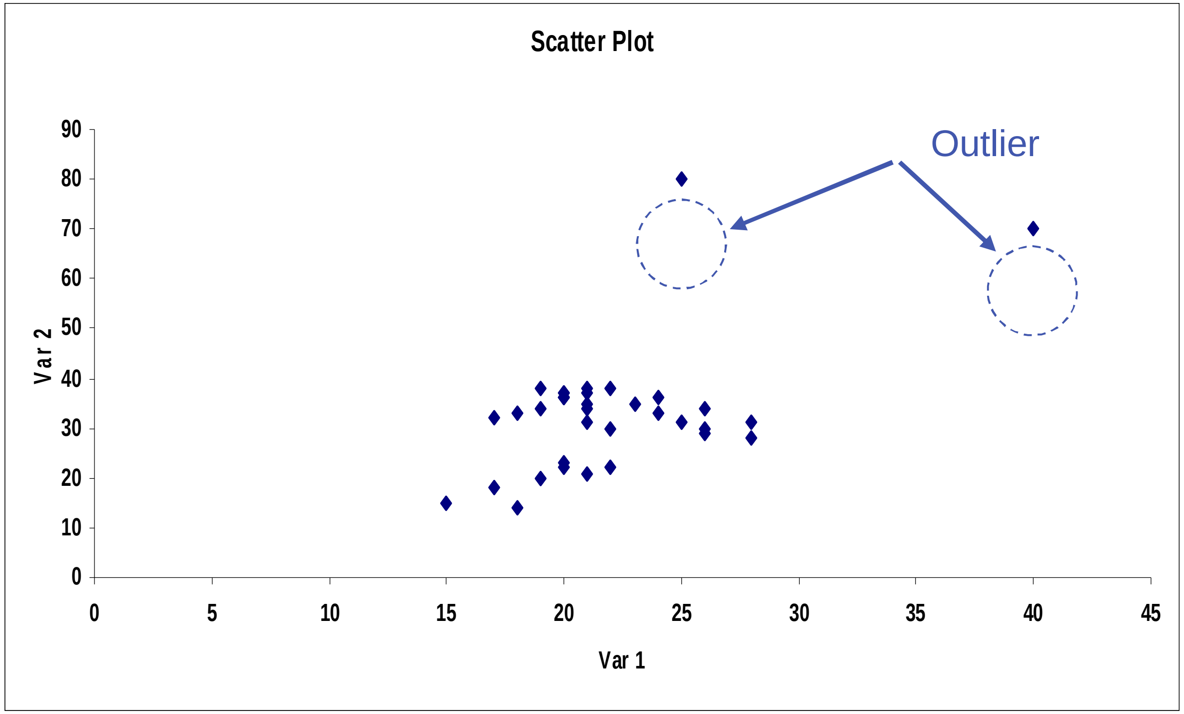
Methodology – Cluster Development



Methodology – Outlier Treatment

What is an outlier ?

An observation is said to be an outlier w.r.t. a variable if it is far away from the remaining observations.



To identify them:

- **Univariate and Frequency analysis**
- **Histogram and Box-Plot**

To tackle them:

1. **The outliers can be deleted from analysis if they are very small in number.**
2. **The variables selected can be trimmed or capped.**

Methodology – Missing Value Treatment

Variables with lot many (about 15%) missing values should not be used for clustering unless 'Missing' has a special significance and can be replaced by some meaningful number.

% of Missing	Treatments
Less than 1%	• Delete those Observations • Mean Imputation
1-5%	• Mean Imputation
5-10%	• Regression Imputation • Mean Imputation
More than 10%	• Regression Imputation • Try to use some proxy Variable

Note: - SAS does not include observations with missing values for Clustering Process

Methodology – Multicollinearity Treatment

What is 'Multi-collinearity' ?

A set of independent or explanatory variables are said to have 'Multi-collinearity', if there is any linear relation between them.

Device to tackle 'Multi-collinearity': -

Factor Analysis: -
By Factor Analysis select those factors, which are explaining almost 90/95 % of total variation together. Then select those variables which have high loadings towards those factors.

VIF (Variance Inflation Factor): -
Variables with VIF more than 2 should be dropped

Methodology – Variable Standardization

Why do we need 'Standardization' ?

Since the units of measurement are different for different variables, standardization is a must.

E.g.: - Consider two variables, Age and Income.
The unit of Age is 'Year' and the unit of Income is say 'Rs'.
Hence they are not comparable.
In that case there won't be an unit of measurement for the distance between two clusters.

Generally we standardize by making the mean = 0 and variance = 1 thus deunitizing the variables and bringing them on a common platform to analyze.

Post all the data treatment steps – "Cluster Development Process" is commenced upon.

Post Cluster Development – "Cluster Validation" is done on the validation sample to establish that the cluster solution is not Sample dependent.

Cluster Building

Cluster Building – Types

There are 2 ways in which Cluster solutions could be built up.

Hierarchical Clustering

Each observation is considered as an individual cluster. Distance from each observation to all others is calculated & the nearest observations are clubbed to form clusters. Intensive distance calculations required thus making it difficult to implement.

K-Means Clustering

K distinct observations are randomly selected at the highest distance from each other. Each observation is considered one by one & clubbed to the nearest Cluster. If two clusters come significantly close to each other, they are merged to each other to form a new cluster.

Hierarchical Clustering is not suitable for large datasets as the multitude of calculations involved would be impossibly huge. Thus K-Means clustering is the most used method of clustering.

Cluster Building – K-Means Clustering

K-Means Clustering SAS Code

```
rsubmit;
proc fastclus data =out.inactive maxc=200 maxiter=100 delete=25000
out=out.final;
var
CNT_LAN_MAT_TW
Loanno
NO_ADV_EMI
MONTHS_SINCE_LOAN_MATURITY
TENOR;
run;
```

Cluster Building – Cluster Solution

Cluster	Frequency	RMS Std Deviation	Max Distance - Seed to Observation	Distance Between Cluster Centroids
1	69696	0.8642	14.6487	2.7342
2	164495	0.3587	3.7355	1.7221
3	84576	0.7891	15.5323	3.2326
4	53434	0.6266	4.471	1.9309
5	111923	0.4809	8.6794	1.8346
6	171323	0.3729	2.4891	1.7221
7	61126	0.7138	12.3443	2.4533

Cluster Means

Cluster	CNT_LAN_MAT_TW	Loanno	NO_ADV_EMI	MONTHS_SINCE_LOAN_MATURITY	TENOR
1	-0.197450101	2.27108366	-1.046054301	-0.509641873	0.312811446
2	-0.375048706	-0.34924125	0.200597434	0.994222204	-0.677162416
3	2.622903928	-0.09743388	-0.435001209	-0.046890848	0.113553959
4	-0.375048706	-0.35414491	-0.97960221	1.463155948	0.777377199
5	-0.355935079	-0.28306493	1.567501549	-0.354749804	0.347933497
6	-0.375046517	-0.28265108	0.111602877	-0.724103839	-0.705347005
7	-0.363966798	0.1052424	-1.071824808	-0.629530996	1.968822045

Variable	R-Square
CNT_LAN_MAT_TW	0.923282
Loanno	0.572698
NO_ADV_EMI	0.694306
MONTHS_SINCE_LOAN_MATURITY	0.590897
TENOR	0.629882
OVER-ALL	**0.682213**

Approximate Expected Over-All R-Squared = 0.54085

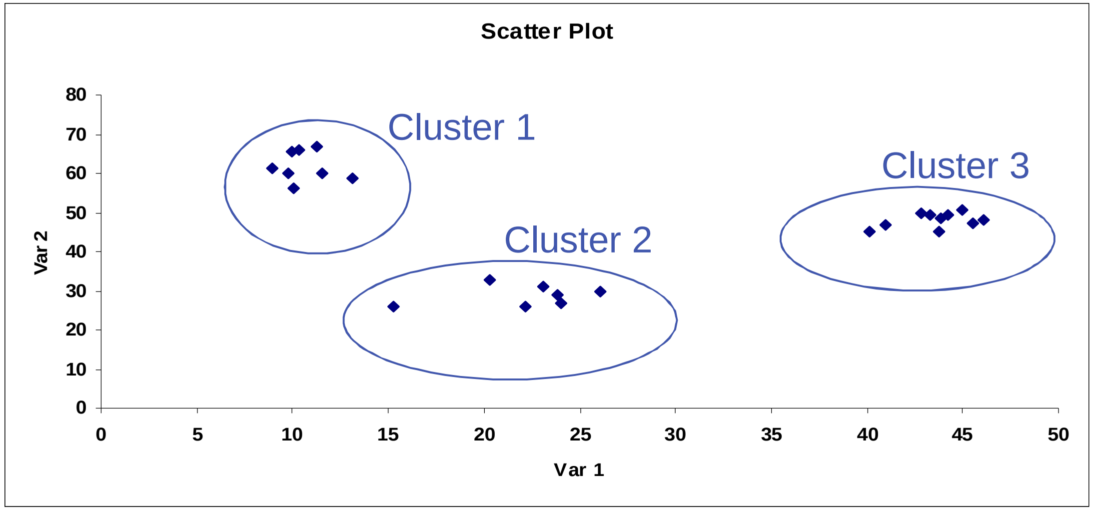
Understanding Cluster Solution: R-Square

For a given data set "Total amount of Variation" is fixed.

If there is k Clusters in the solution then Total Variation = Within Variation + Between Variation

Within Variation = (Variation within Cluster 1) + (Variation within Cluster 2) + … + (Variation within Cluster k)

Between Variation = Variation between one cluster to another (i.e. variation of cluster means).



$$\textit{R - Square} = \frac{\text{Between Variation}}{\text{Total Variation}}$$

Higher R-Square signifies high "between" variation and low "within" variation. Thus Higher the R-Square, the better it is.

Understanding Cluster Solution: Other Metrics

Approximate Expected Overall R-square

Approximate Expected Overall R-Square is calculated based on the hypothesis that all the explanatory variables used for Clustering are independent.

Hence if there is a lot of difference between Observed Overall R-square and Approximate Expected Overall R-square, we can suspect high correlation among the independent variables.

RMMSTD

RMMSTD within a cluster = Square root of Average of (Variance of variable 1 in that cluster, Variance of variable 2 in that cluster, … ,Variance of variable p in that cluster) . Assuming p variables were used for Clustering.

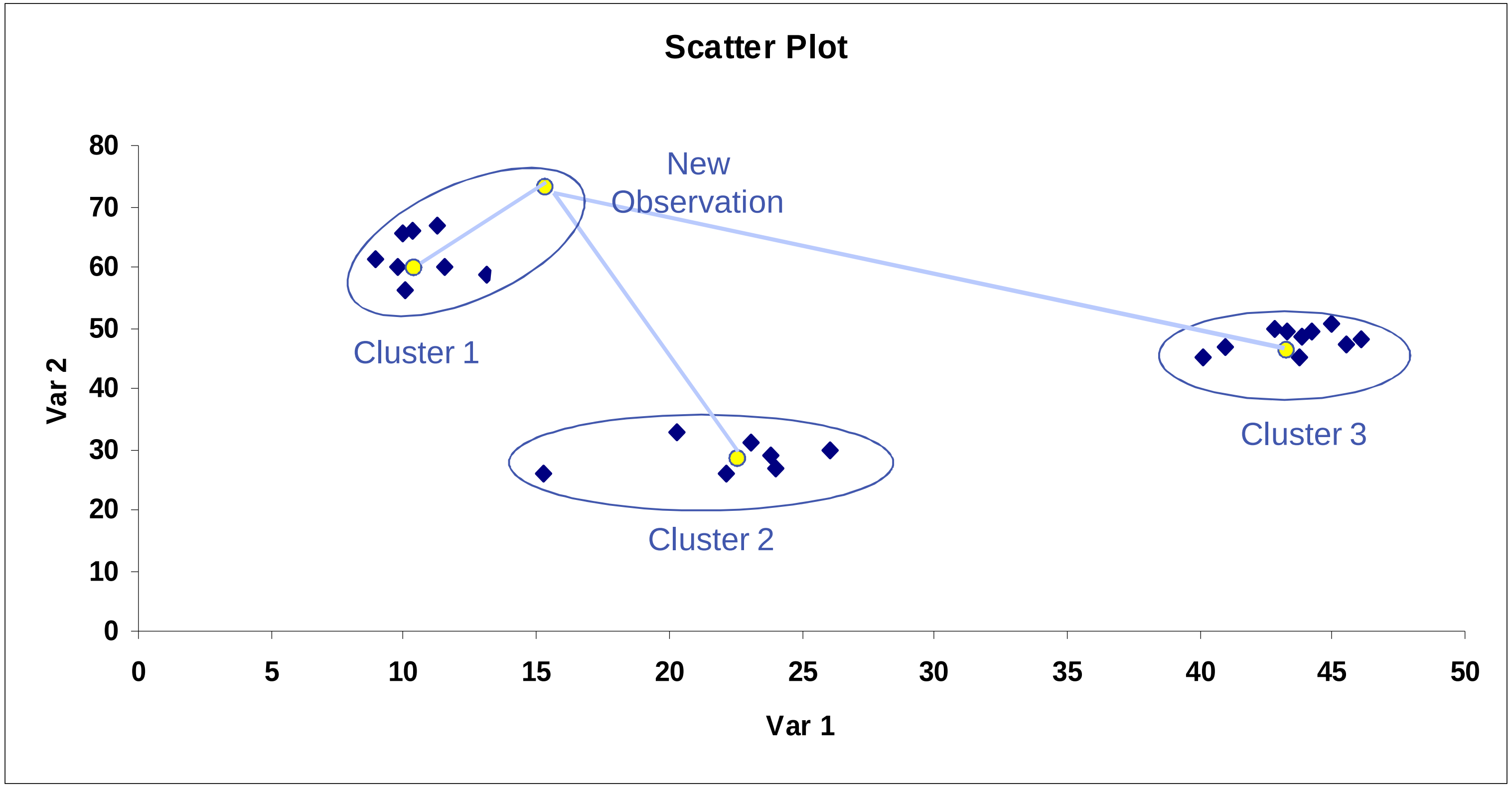
There is no restriction on the number of clusters, but it should be between 5 to 15.

Care should be taken on the number of observations in each clusters. A good rule of thumb is to have >= 5% of the population in each cluster.

Cluster Validation & Profiling

Cluster Validation

The Cluster Solution is Validated on the "Validation Sample" using the Minimum Euclidean Distance Method. Validation is done by calculating the distance of each observation in the Validation sample from the Cluster Seed & assigning it to the closest cluster.

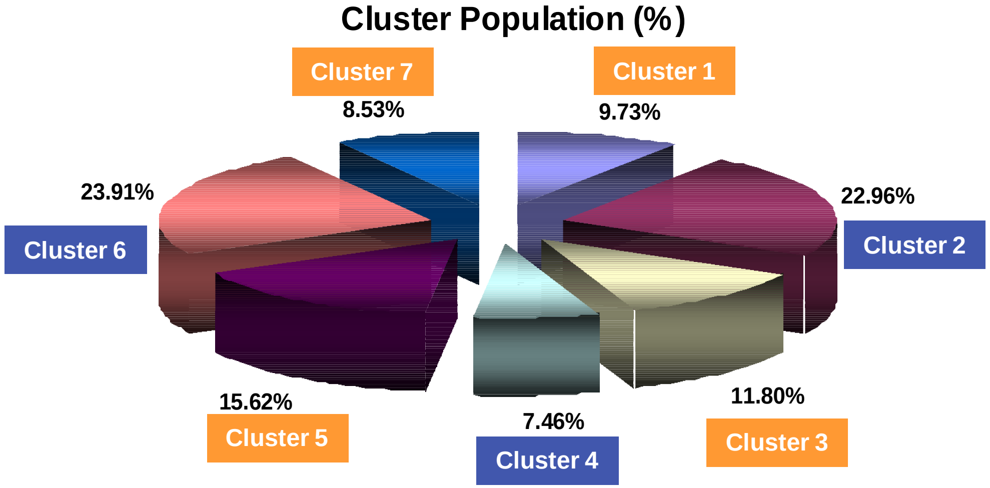


The New Observation will be a member of Cluster 1.

Cluster Validation: Sample Example

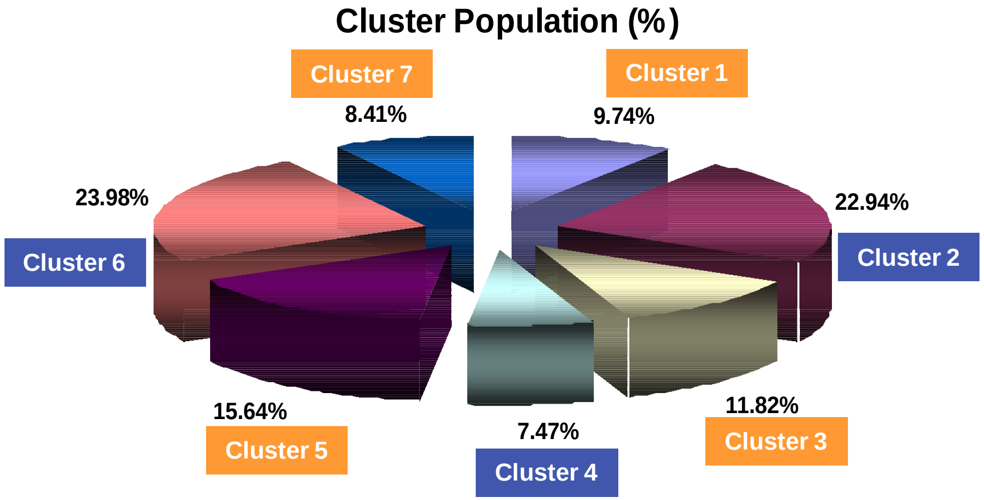
Development Sample

Cluster	Frequency	%
1	69,696	**9.73**
2	164,495	**22.96**
3	84,576	**11.80**
4	53,434	**7.46**
5	111,923	**15.62**
6	171,323	**23.91**
7	61,126	**8.53**
Total	**716,573**	**100**



Validation Sample

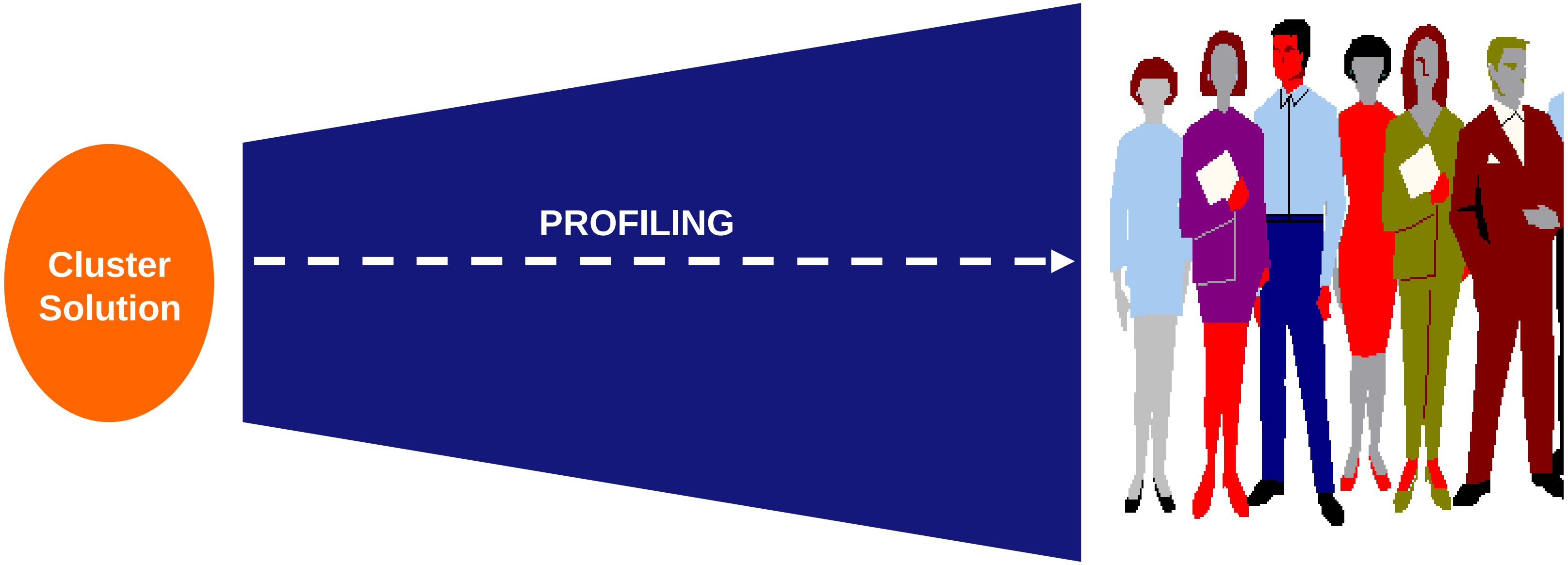
Cluster	Frequency	%
1	69,899	**9.74**
2	164,653	**22.94**
3	84,837	**11.82**
4	53,625	**7.47**
5	112,250	**15.64**
6	172,084	**23.98**
7	60,320	**8.41**
Total	**717,668**	**100**

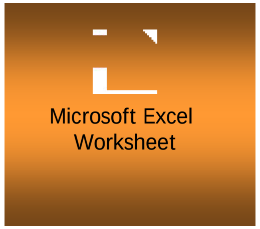


The Validation sample was scored using the cluster solution. The frequency plot shows a <u>similar distribution</u> on the Validation sample as in the Development sample.

Cluster Profiling with Example

Cluster Solution is profiled against Variables to identify and assign the character of individual clusters.





Data file Continuous Numeric Variables.

Data file Categorical Variables.