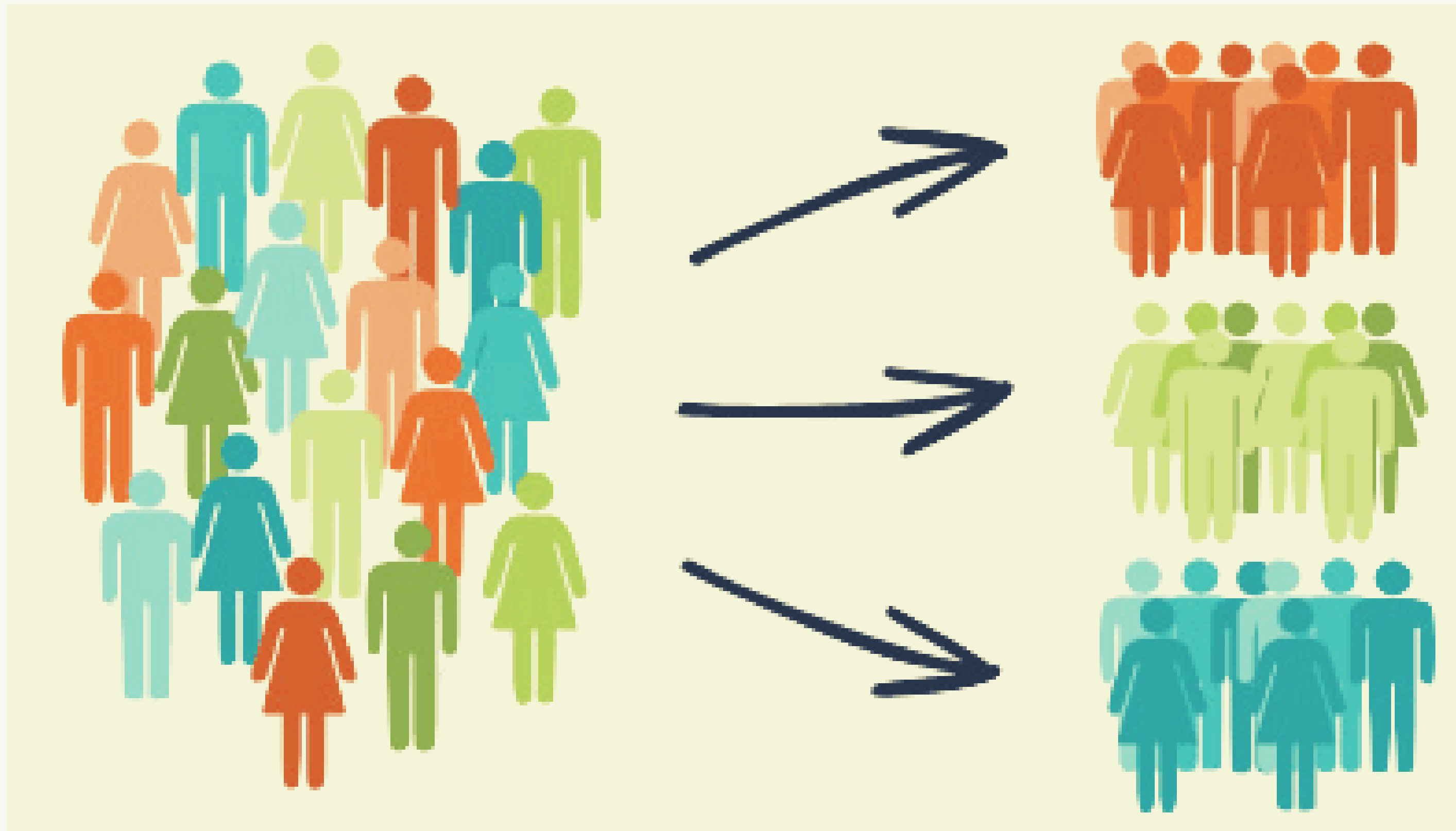


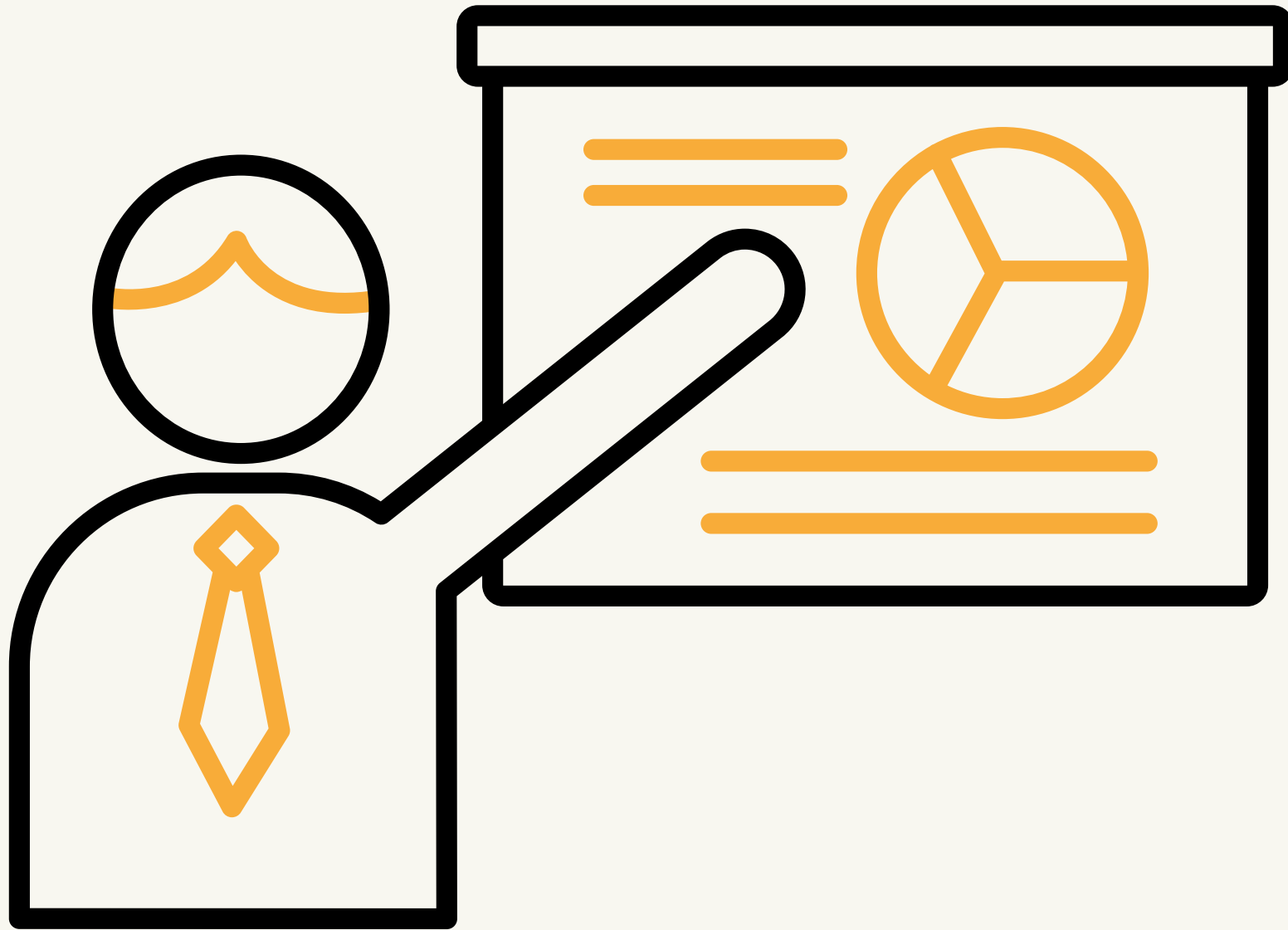
# CONSUMER SEGMENTATION USING K-MEANS CLUSTERING

*(AN INITIATIVE BY CRISA RESEARCH AGENCY)*



**Case Study: Bath Soap Households**

# PRESENTATION OUTLINE



- Introduction
- Objective & Benefits
- Exploration
- Cluster Representation
- Characteristics of Clusters
- Modeling on Clusters
- Profit Estimation
- Recommendations

# INTRODUCTION

## What is CRISA?

CRISA is an Asian market research agency that specializes in *tracking consumer purchase behavior* in consumer goods (both durable and nondurable).

## CRISA's Current Project:

CRISA wants to track consumer behaviour in the Indian market by constituting household panels in over 100 Indian towns and cities.

## CRISA's Strategy:

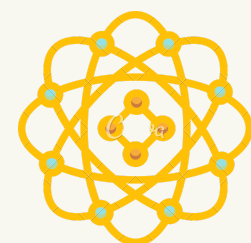
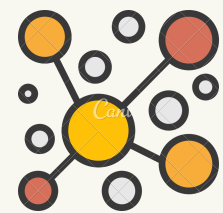
Select households using stratified sampling based on socio economic status and market.

Analyze a subset of 600 records.



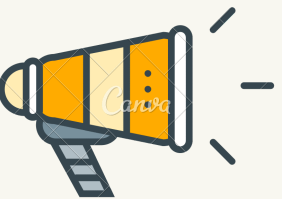
# OBJECTIVE

- **Segmentation:** Using K-means clustering to find clusters on the basis of the following factors:
  1. Purchase behavior
  2. Basis of purchase
  3. Combination of both
- **Selection:** Choose the best cluster and identify its features
- **Modeling:** Develop a model that classifies the data into these segments.



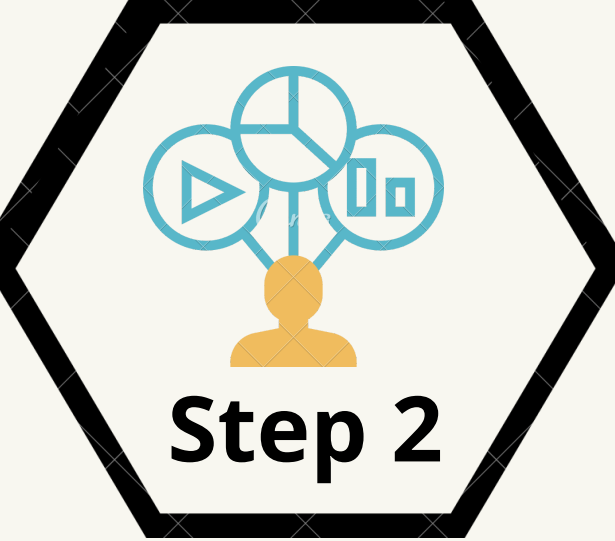
# BENEFITS

- **Information:** Gain information about demographic attributes associated with purchase behaviours and brand loyalty
- **Customized Promotions:** Design promotions suitable for each segment types
- **Reward System:** Design effective reward system and thus, increase brand loyalty
- **Cost Effective:** Saves time and money



# EXPLORATION

Uploaded data,formatted the variables, normalized and ran descriptive statistics



Selected 12 variables to create 4 clusters based on purchase behaviour

Selected 16 variables to create 4 clusters based on the basis of purchase



Combined 27 variables to create 4 clusters based on basis of purchase & purchase behaviour together

Final cluster was chosen out of the 4 clusters created in Step 4 by looking at the Elbow & insights were drawn



Performed analysis and logistic regression model on the 2 sub-clusters based on demographics such as socio economic class, age, education, gender etc.



Crunched numbers to find out the estimated profit from the model on the selected clusters



Provided recommendations related to promotional strategy



# **CRITERIAS FOR CHOOSING THE FINAL CLUSTER**

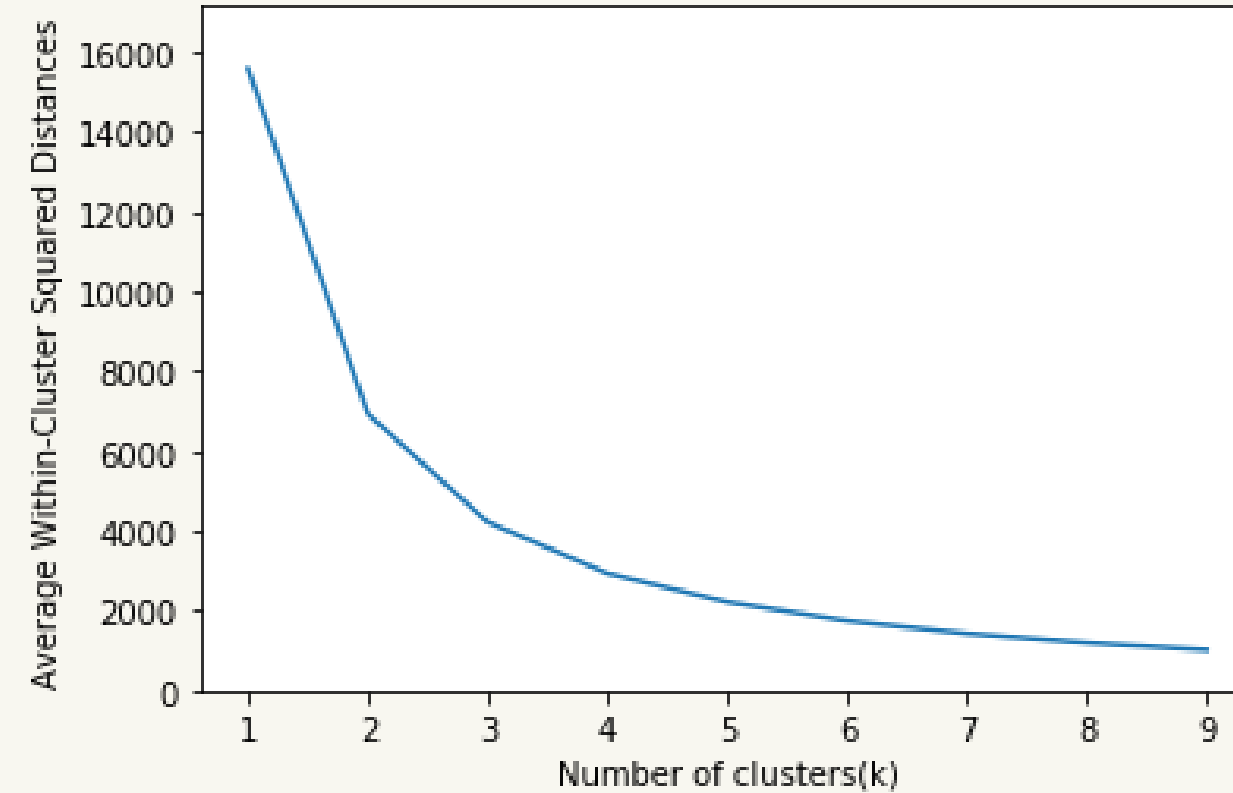
- **Use combined variables from both Purchase Behaviour and Basis of Purchase (Total Variables =27)**
- **Members <50% in each sub cluster**
- **Elbow=4**



## Chosen Variables for Final Cluster

Member Id	# of Brands	Brand Runs	# of Transactions
Transaction/brand run	Volume/Transactions	Total Volume	Value
Average Price	Purchase Vol. No Promo %	Purchase Vol. Promo 6%	Purchase Vol. Other Promo %
Pr Cat 1	Pr Cat 2	Pr Cat 3	Pr Cat 4
Prop Cat 5	Prop Cat 6	Prop Cat 7	Prop Cat 8
Prop Cat 9	Prop Cat 10	Prop Cat 11	Prop Cat 12
Prop Cat 13	Prop Cat 14	Prop Cat 15	

## Decision about number of clusters



- Elbow stretching from 2 to 4
- Ideal scenario is doing 4 main clusters
- Choose the best master cluster out of them

## Number of Members in each sub cluster

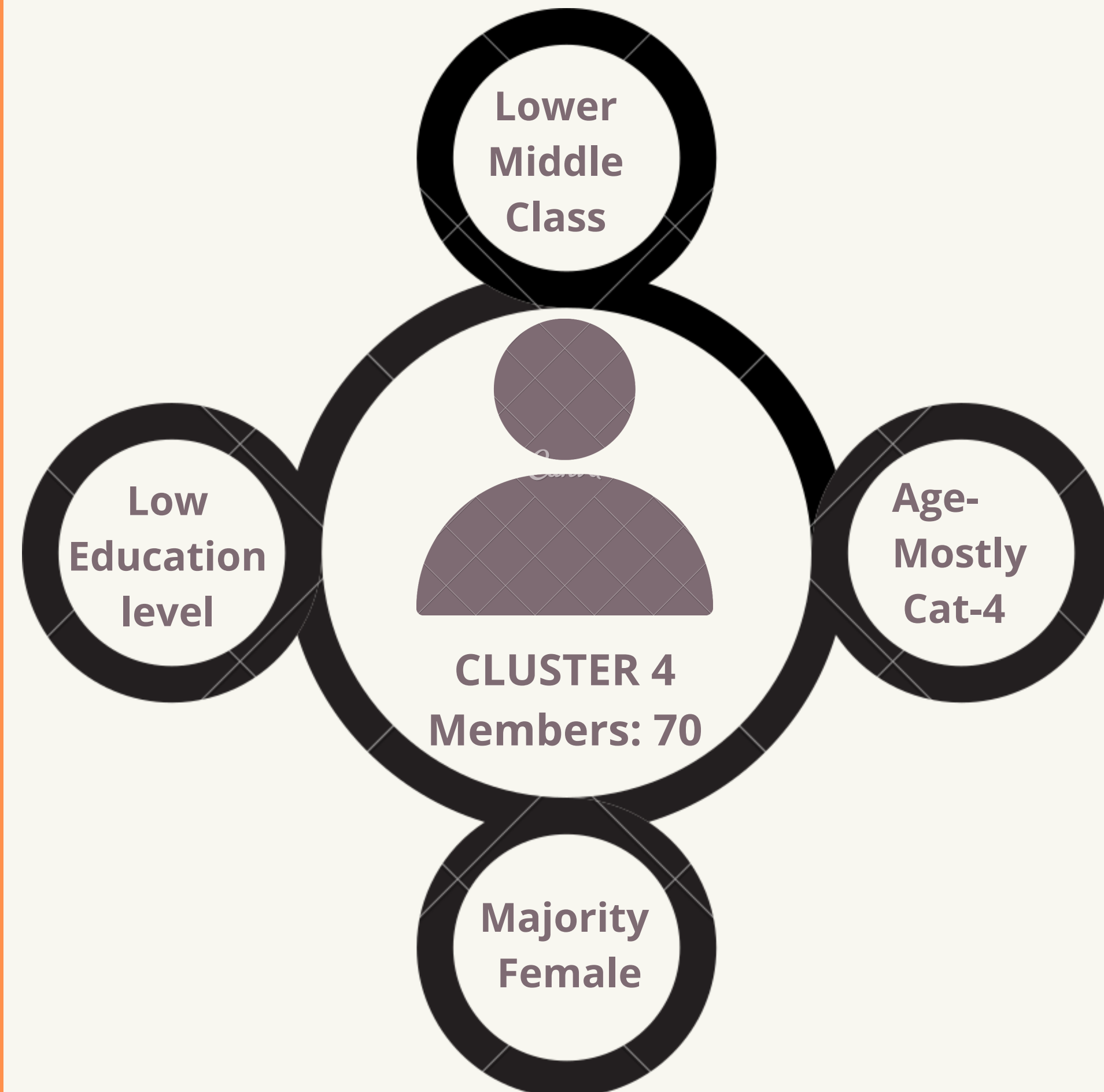
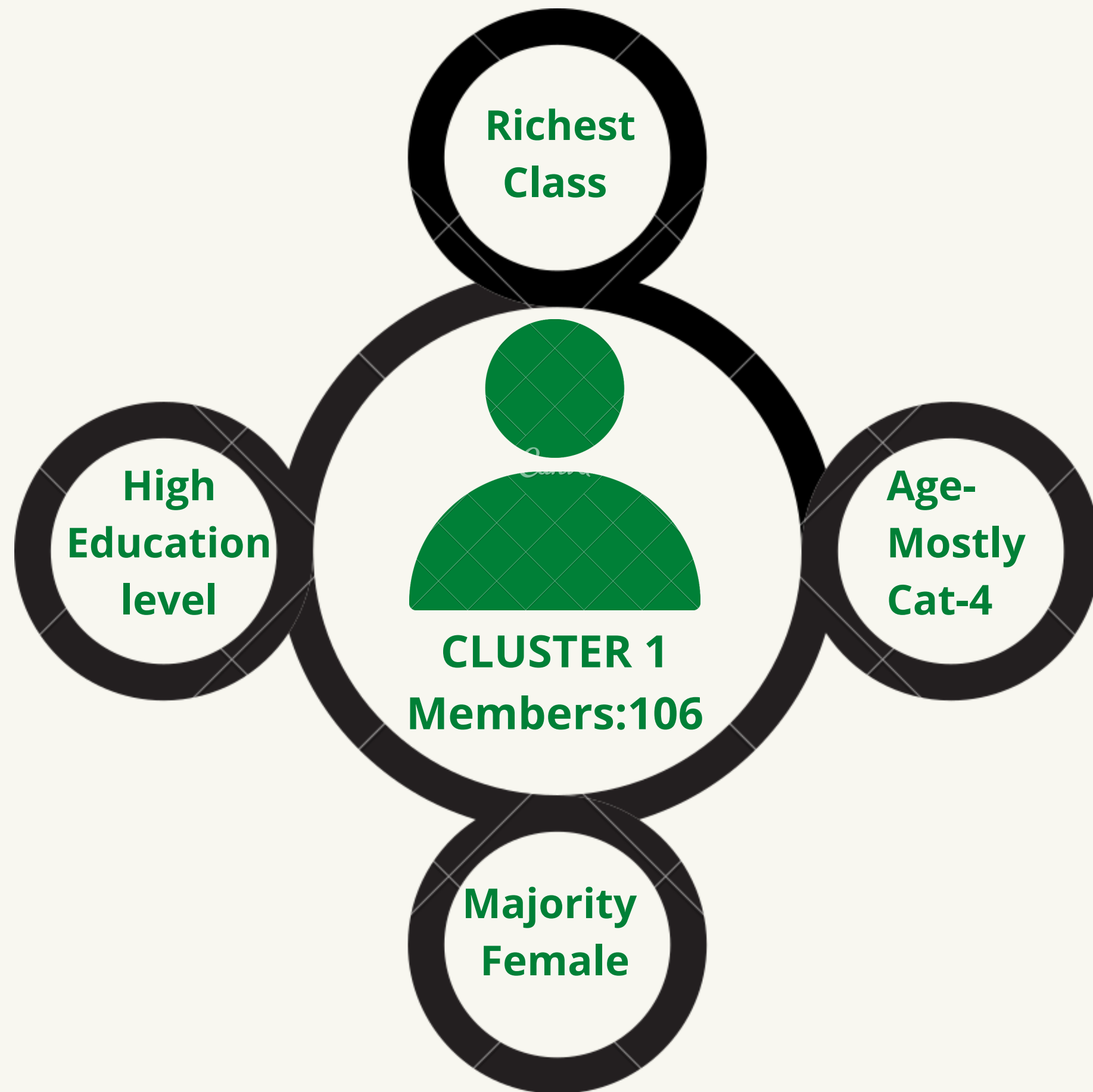
Cluster 1	Cluster 2	Cluster 3	Cluster 4
106 Members	240 Members	184 Members	70 Members

- Total Records=600
- Members in each cluster <50% of total records (less than 300)
- So, above visual shows that our chosen master cluster satisfies this criteria





# CHARACTERISTICS OF CHOSEN CLUSTERS



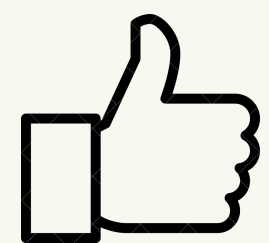
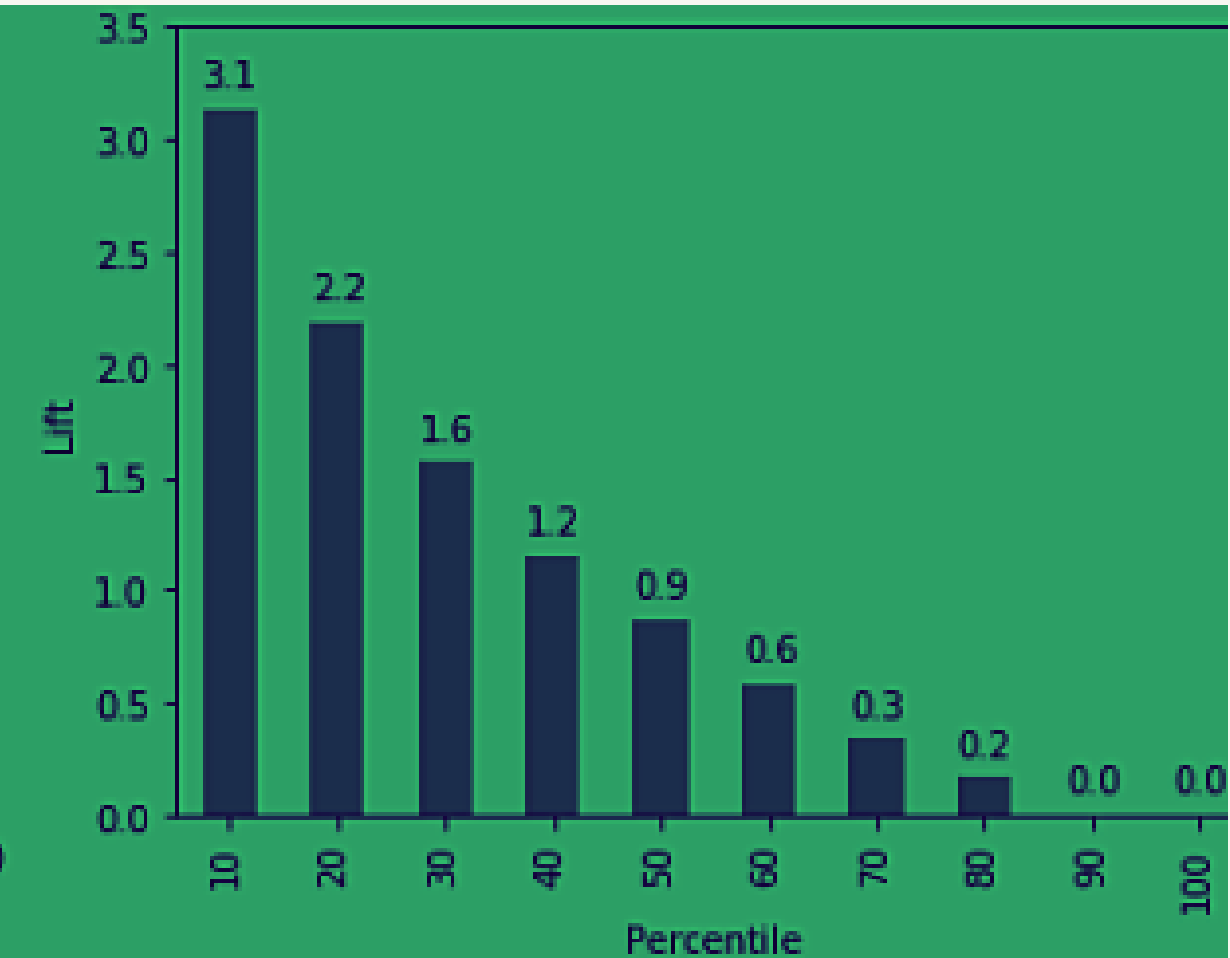
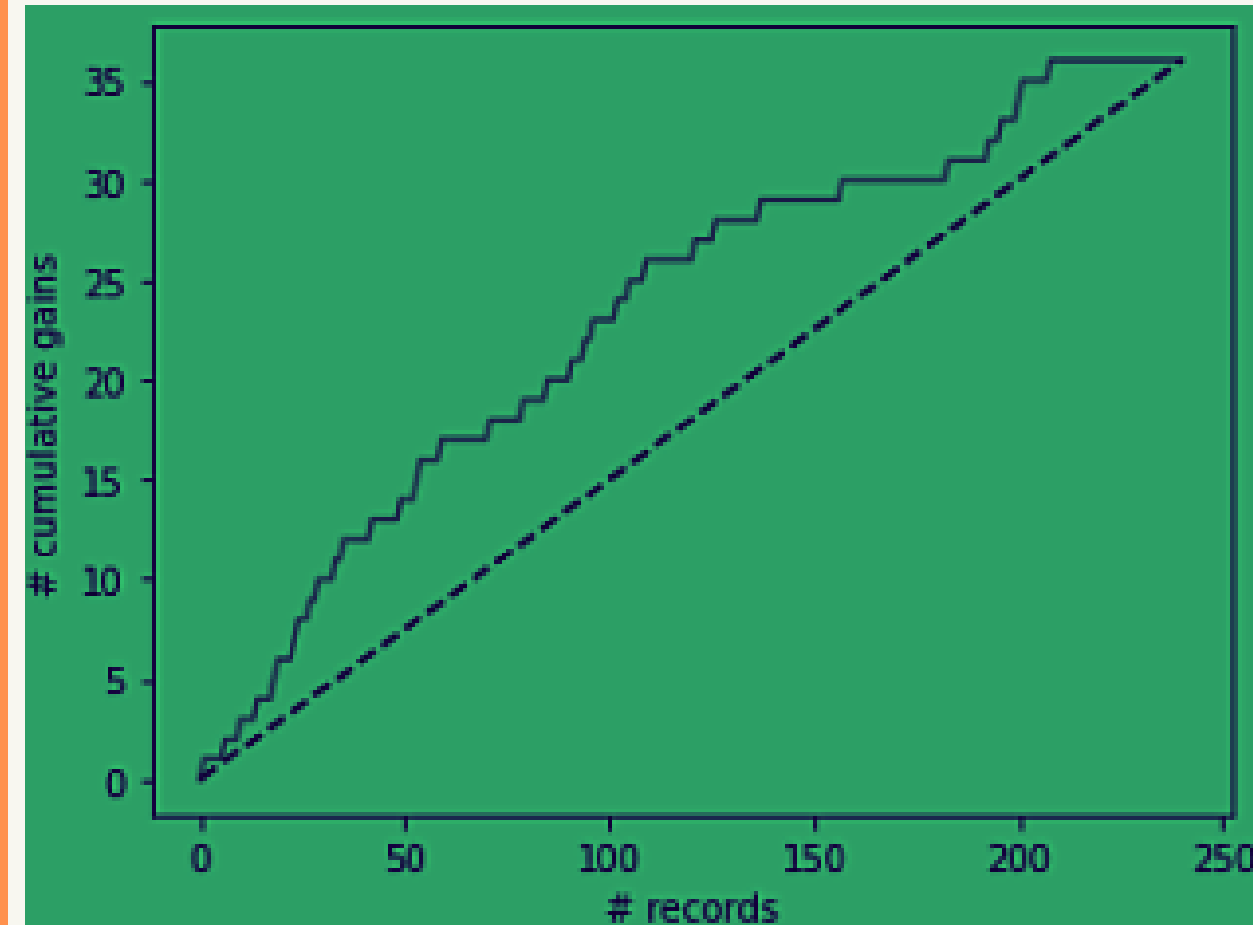
# RESULTS OF MODELING - CLUSTER 1

Confusion Matrix (Accuracy 0.8278)

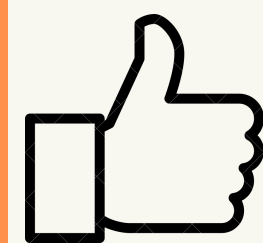
		Prediction	
Actual	0	1	
	0	280	10
1	52	18	

Confusion Matrix (Accuracy 0.8375)

		Prediction	
Actual	0	1	
	0	198	6
1	33	3	



Accuracy of 83% shows that the model on Cluster -1 is giving good results compared to the situation if there is no model.



Gains Chart shows that for the initial 150 records give us close to 30 members in Cluster -1. Without model it only gives us 20

Decile Chart shows that our top 30% records give us better results by 2.3 times in comparison to a random selection

# RESULTS OF MODELING - CLUSTER 4

Confusion Matrix (Accuracy 0.8944)

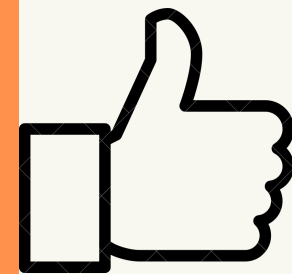
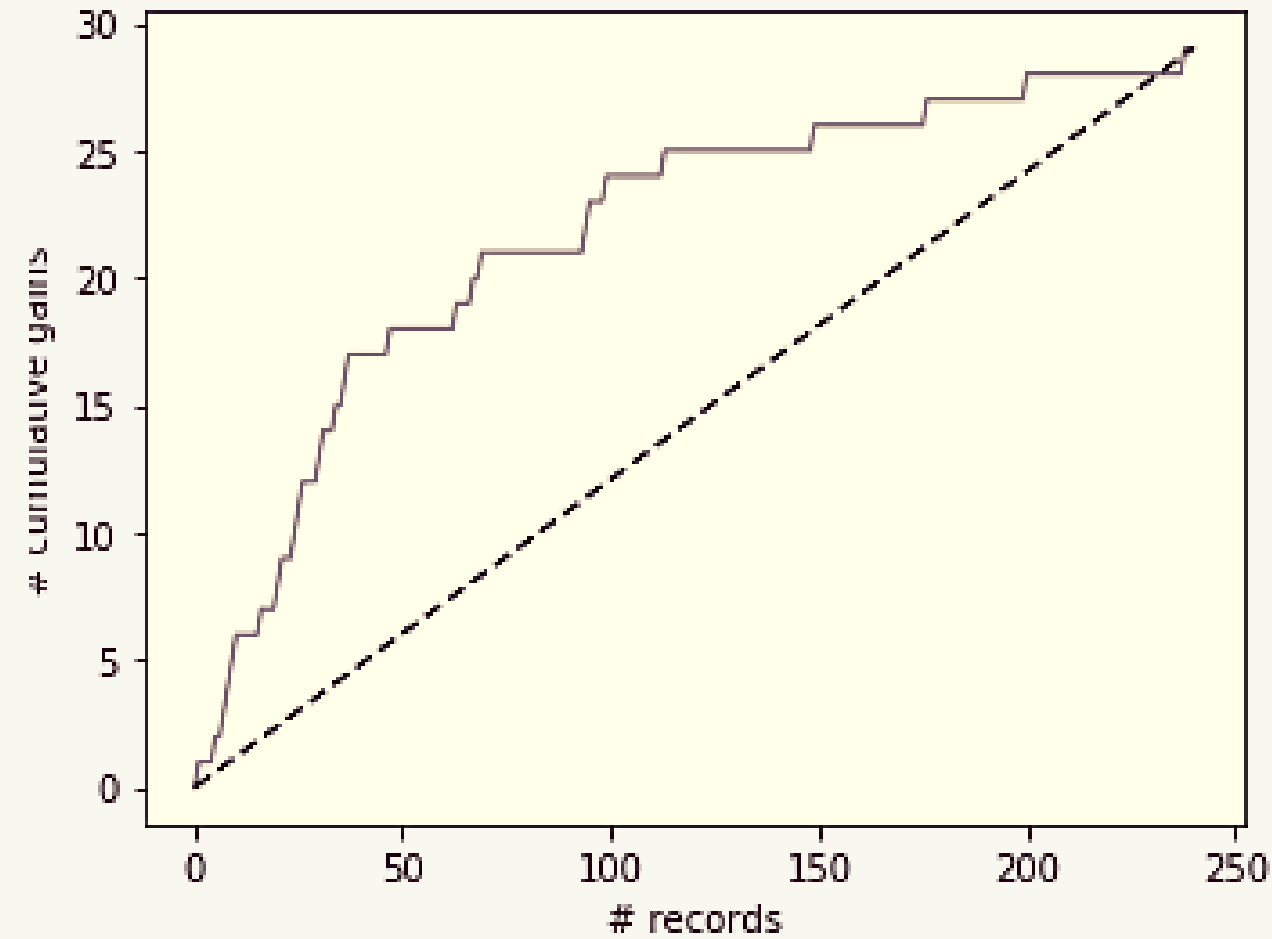
	Prediction	
Actual	0	1
0	314	5
1	33	8

Confusion Matrix (Accuracy 0.8833)

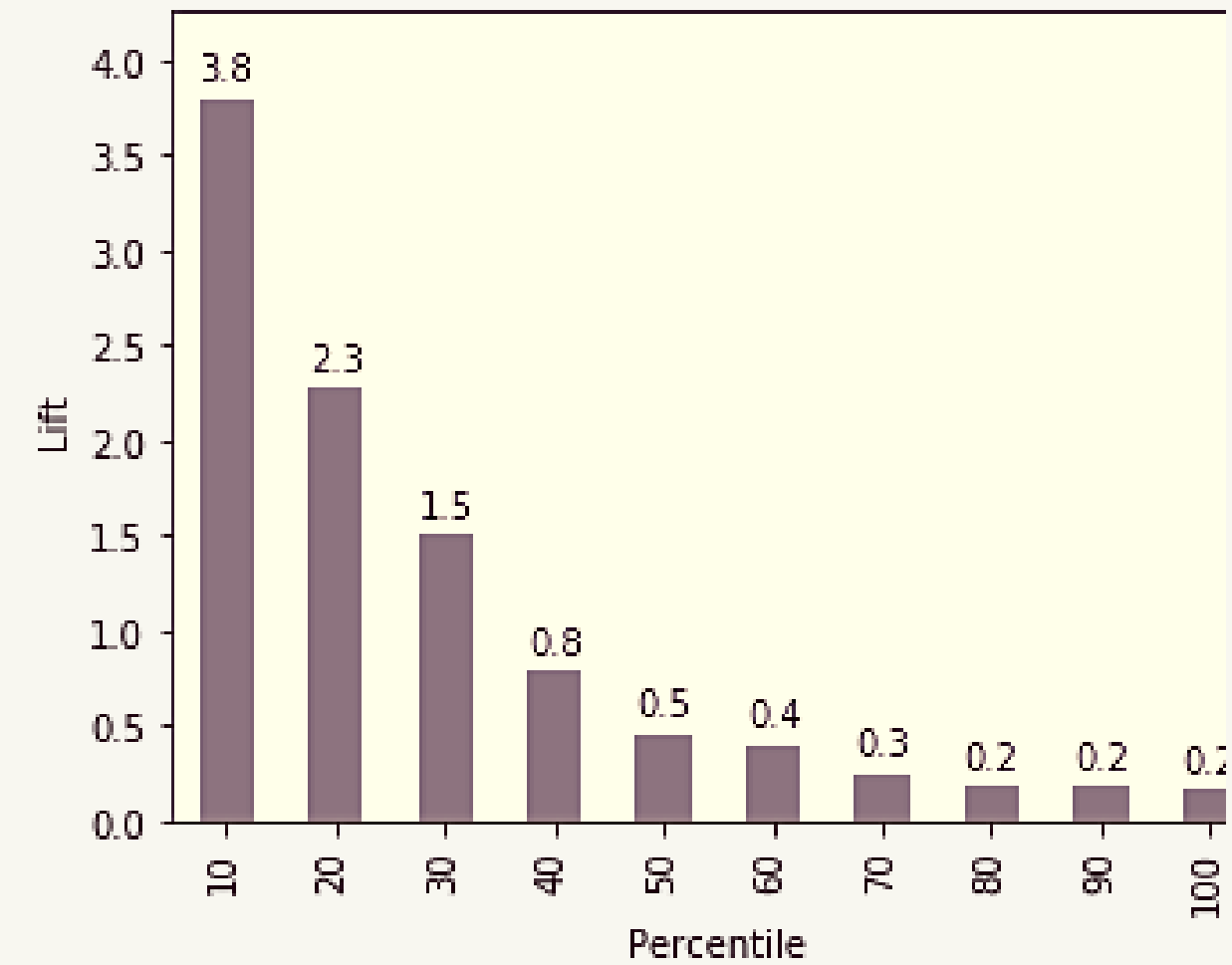
	Prediction	
Actual	0	1
0	207	4
1	24	5



Accuracy of 89% shows that the model on Cluster -4 is giving good results compared to the situation if there is no model.



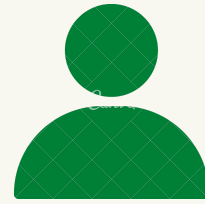
Gains Chart shows that for the initial 125 records, the model give us close to 25 members in Cluster- 4. Without the model it only gives us 15



Decile Chart shows that our top 30% records give us better results by 2.5 times in comparison to a random selection

# ESTIMATED PROFIT FROM THE MODELS

## Cluster -1



Total Members = 106

Sample Size = 27 (From the confusion matrix)

Sample Cost =  $27 \times 5 = \$135$

No. of members accepting promo = 9

No. of members rejecting promo = 18

Decile Chart shows:

Top 10% gives lift of 3

No. of records from random selection = 3

Profit from random selection =  $250 \times 3 = \$750$

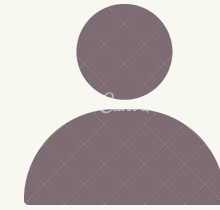
***Estimated Profit from random selection:***  
 ***$\$750 - \$135 = \$615$***

Profit from the model =  $9 \times 250 = \$2250$

***Estimated Profit from the Model:***  
 ***$\$2250 - \$135 = \$2115$***



## Cluster -4



Total Members = 70

Sample Size = 9 (From the confusion matrix)

Sample Cost =  $9 \times 5 = 45$

No. of members accepting promo = 5

No. of members rejecting promo = 4

Decile Chart shows:

Top 10% gives lift of 4

No. of records from random selection = 1

Profit from random selection =  $250 \times 1 = \$250$

***Estimated Profit from random selection:***  
 ***$\$250 - \$45 = \$205$***

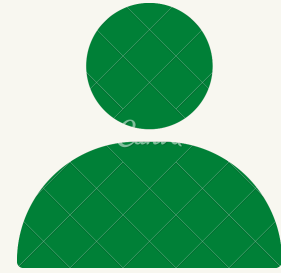
Profit from the model =  $5 \times 250 = \$1250$

***Estimated Profit from the Model:***  
 ***$\$1250 - \$45 = \$1205$***



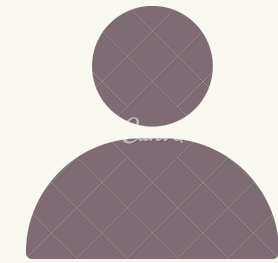
# RECOMMENDATIONS

## Cluster -1



- Cluster 1 has mostly rich people which means that they have the ability to spend more so promotions should be offered from that perspective
- Cluster 1 should be given more offers related to Price Category 1, Proposition Category 10 & 13

## Cluster -4



- Cluster 4 has lower middle and poor people so more affordable options can be highlighted in the promotional offers. Schemes like 'Buy 2 Get 1' will be really useful.
- Cluster 4 should be given more offers related to other promotions, Price Category 3 & Proposition Category 14

- Women are common in both clusters so offers can be designed to catch their attention with attractive offers.

# THANK YOU

---

*Questions?*