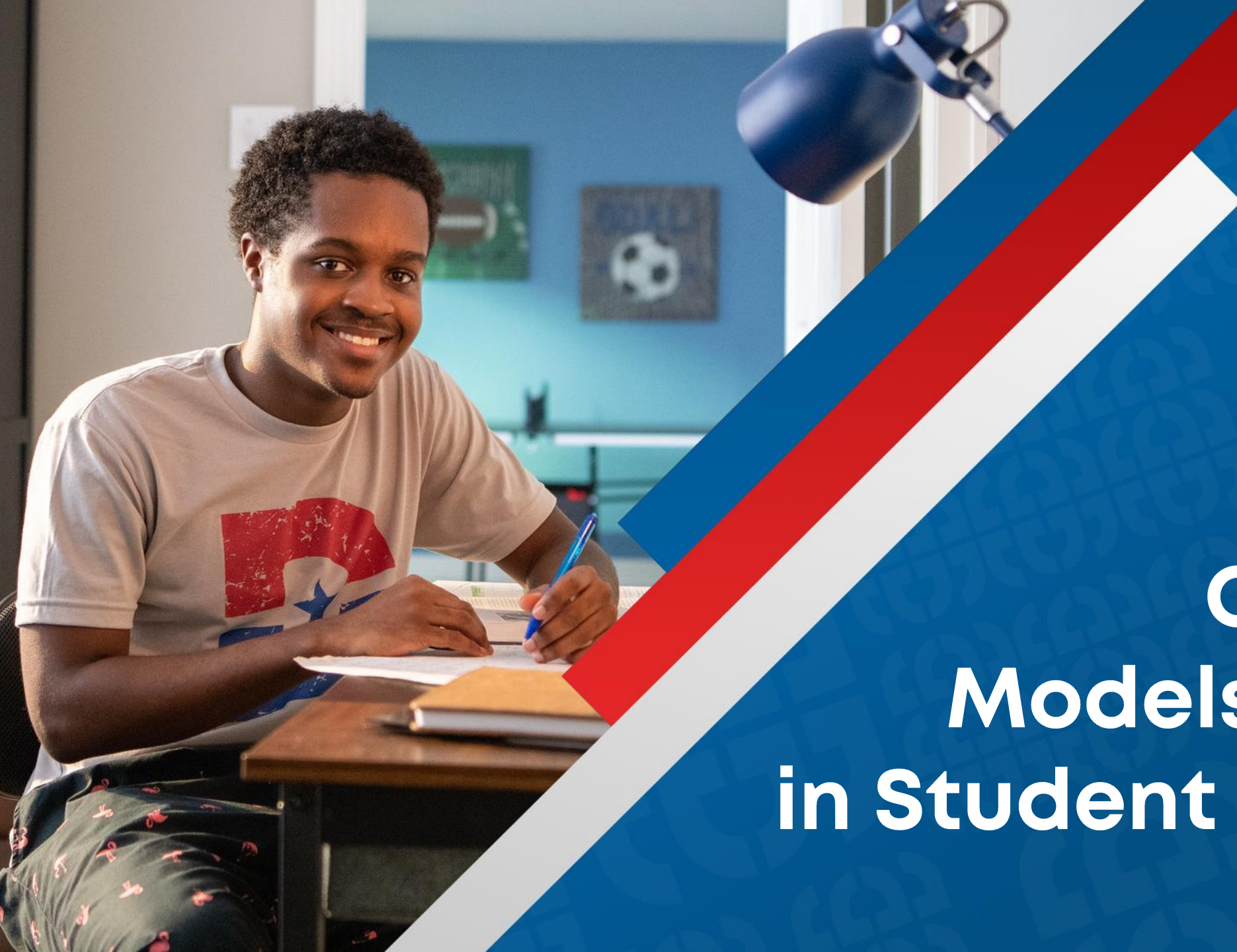
February 8, 2022



DALLAS COLLEGE

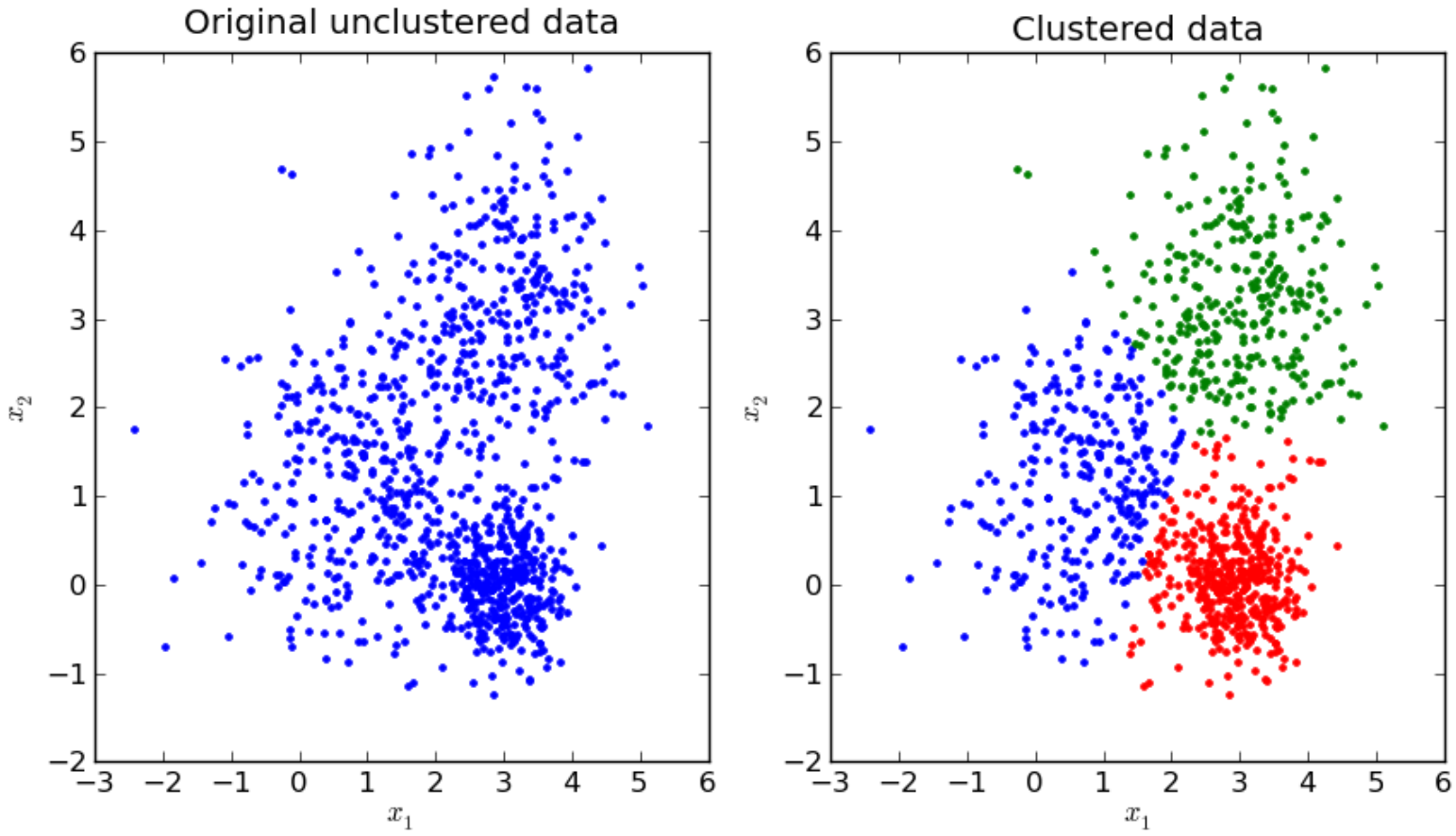
TAIR 2022

Clustering Models to Assist in Student Outreach

Session Agenda

1. Why clustering?
2. Example 1, hierarchical clustering
3. Questions
4. Overview of common algorithms and strengths/limitations
5. Example 2, K-means clustering
6. Questions

Clustering in 60 seconds



Why Clustering? Segmentation.

Reverse looking

What are characteristics of students who used/did not use a service? Did/did not persist?

Who is the prototypical student in each academic program?

Forward looking

What characteristics comprise our prospective student personas?

How do risk factors overlap in some groups of students?

Example 1

Female Student Re-enrollment Texts

Business Understanding: Female Student Re-enrollment

The challenge

Contrary to national trends, Dallas College has experienced larger than typical drops in female student re-enrollment patterns. Our goal was to stem decreases for the Fall 2021 semester through a text campaign.

The question

What messaging should be used to resonate with these students?

Data Understanding & Prep

45k students

Diverse features for the model

- **Demographics:** race/ethnicity, gender, age
- **Financial:** income, employment intensity
- **Household:** household size, dependent count
- **Academics:** last term of enrollment, credits, GPA
- **Special Pop** membership

Model Creation: SPSS Two-Step Clustering

Step 1

Starts with a case as leaf node

Each case goes into that leaf node or breaks into a new node

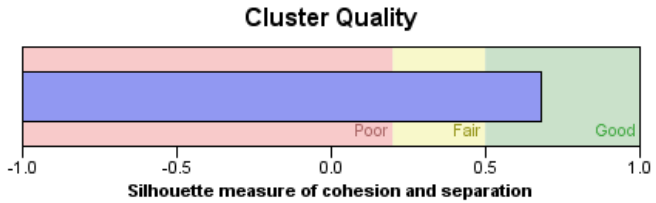
Result is a Cluster Features tree

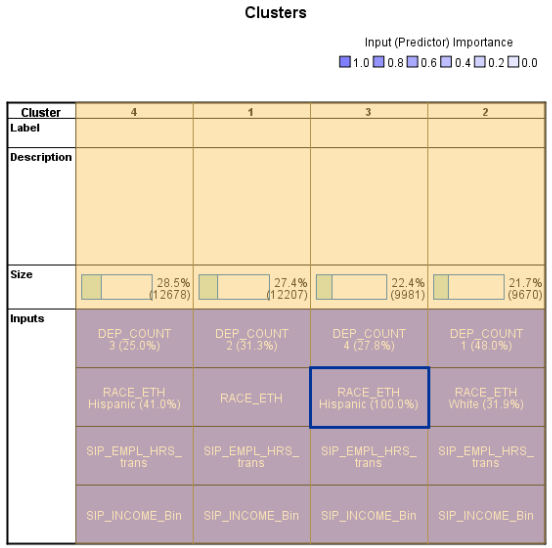
Step 2

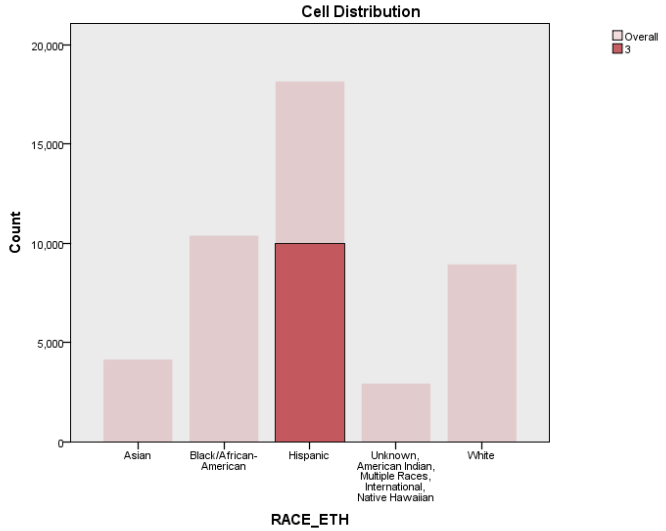
Leaf nodes are combined through agglomeration

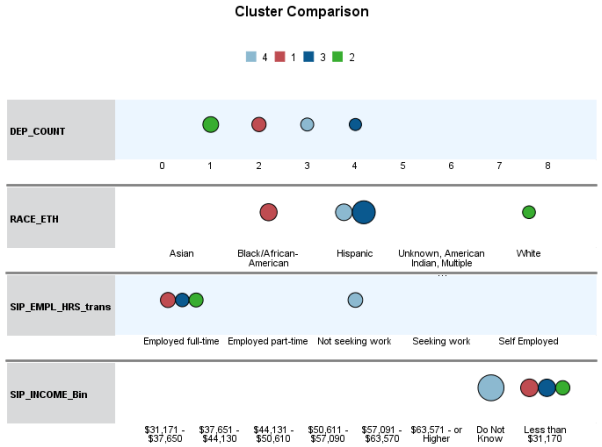
Result is several “best” clusters with a silhouette score

Model Evaluation



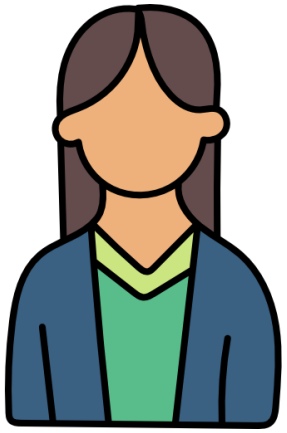






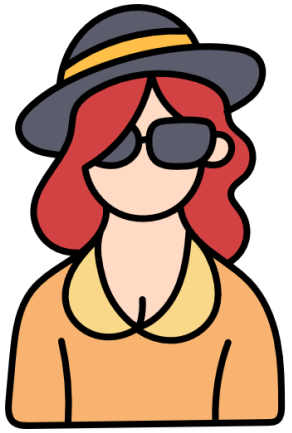
Model Output & Deployment

African Am.
3 dependents
Part-time job



Hispanic
4 dependents
Part-time job

Hispanic
2 dependents
Full-time job



White
1 dependent
Full-time job

Overview
Algorithms

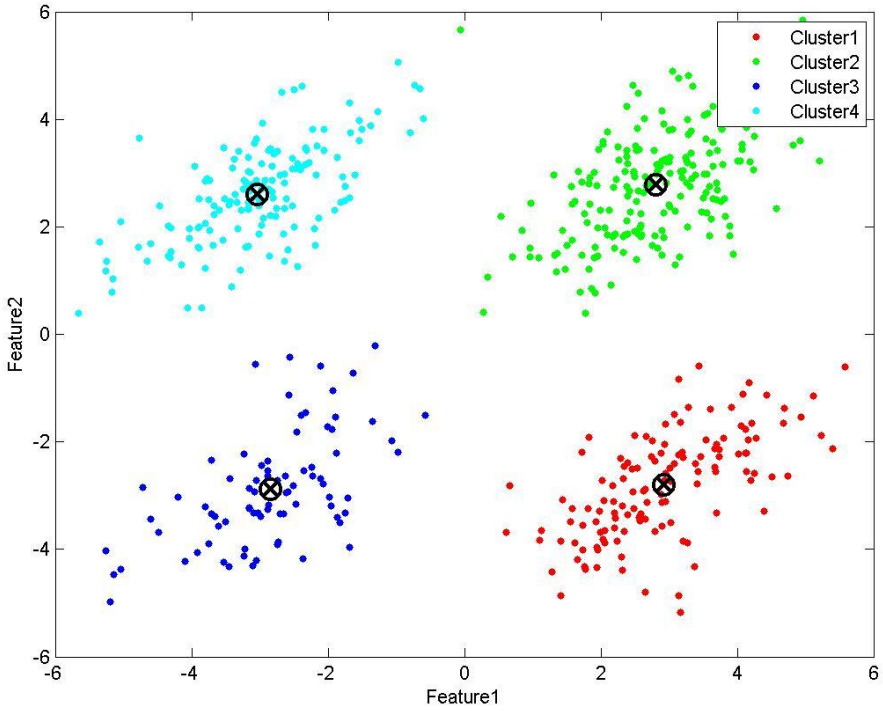
Common Algorithms

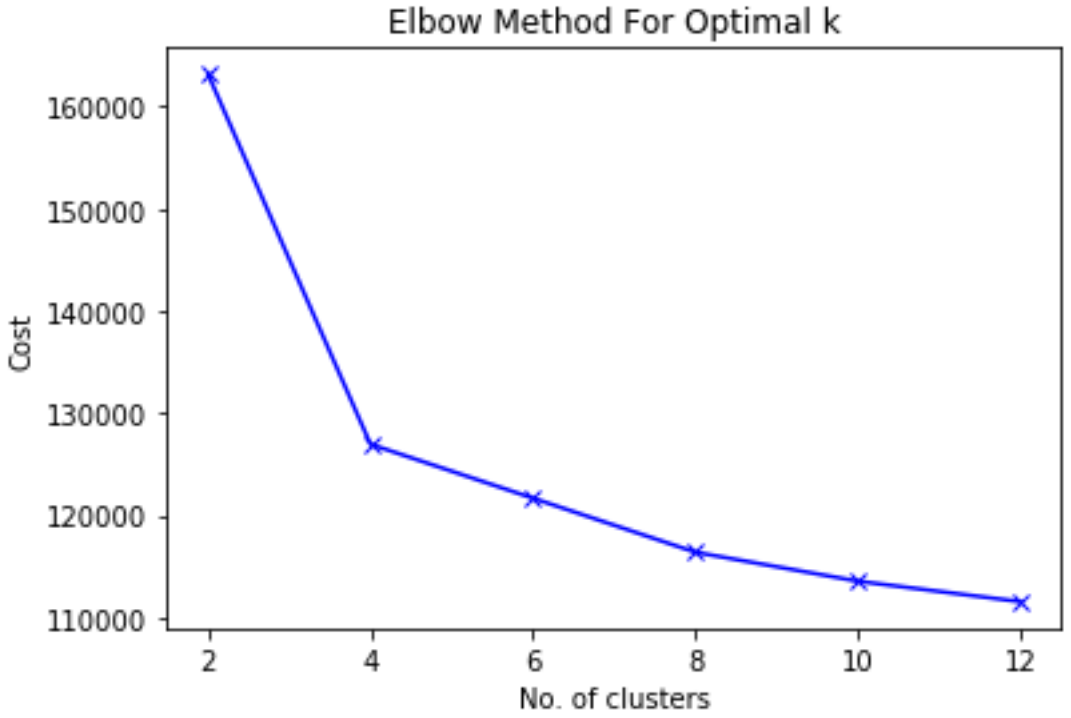
BIRCH and Two-step (SPSS proprietary; Ex 1)

K-means, K-modes, K-medoids (Ex 2)

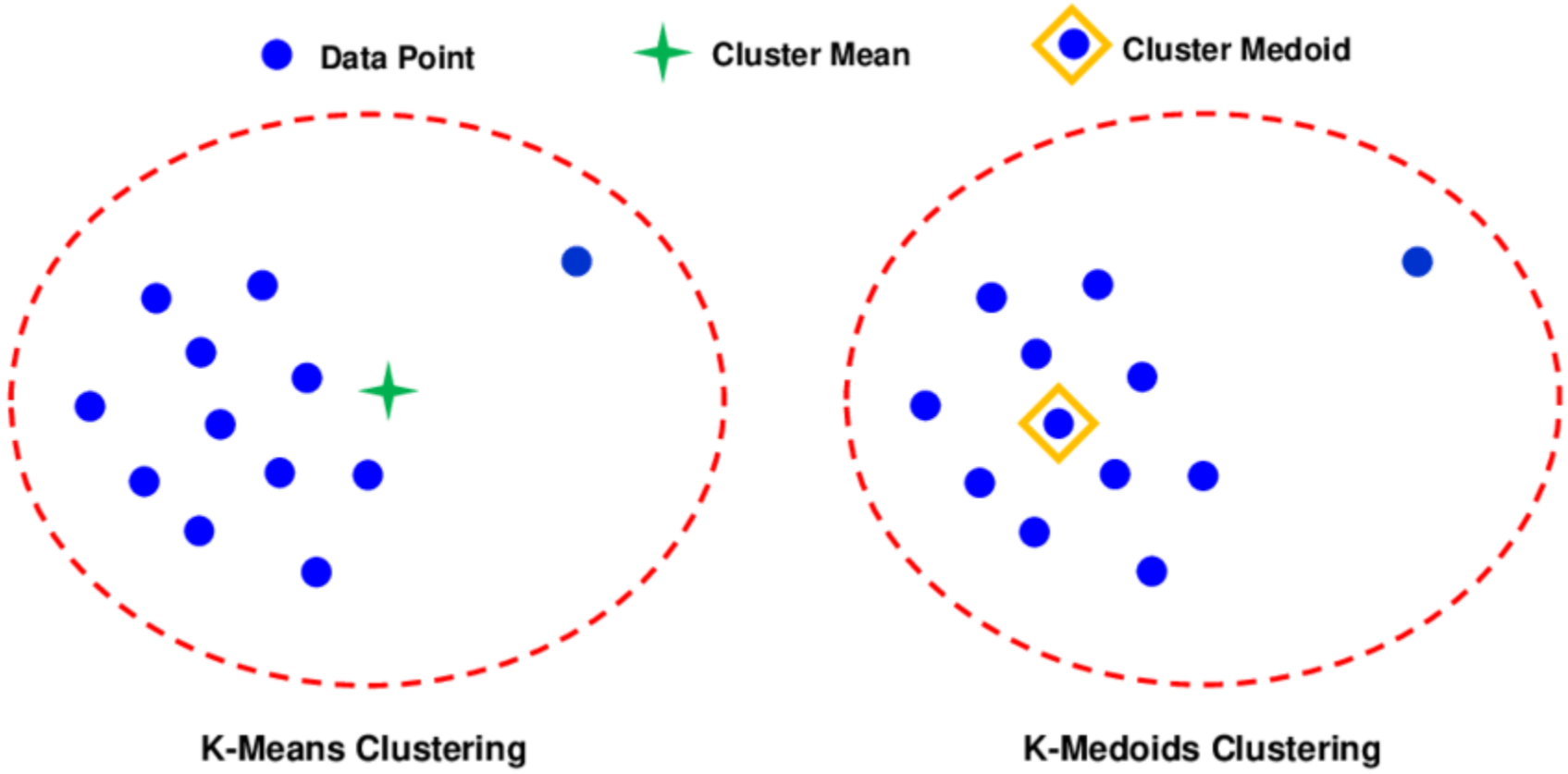
Hierarchical (Agglomerative and Divisive)

The "K" Algorithms

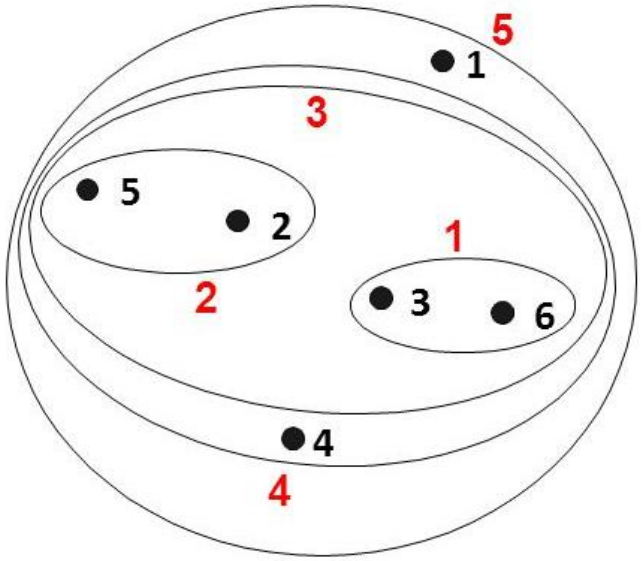




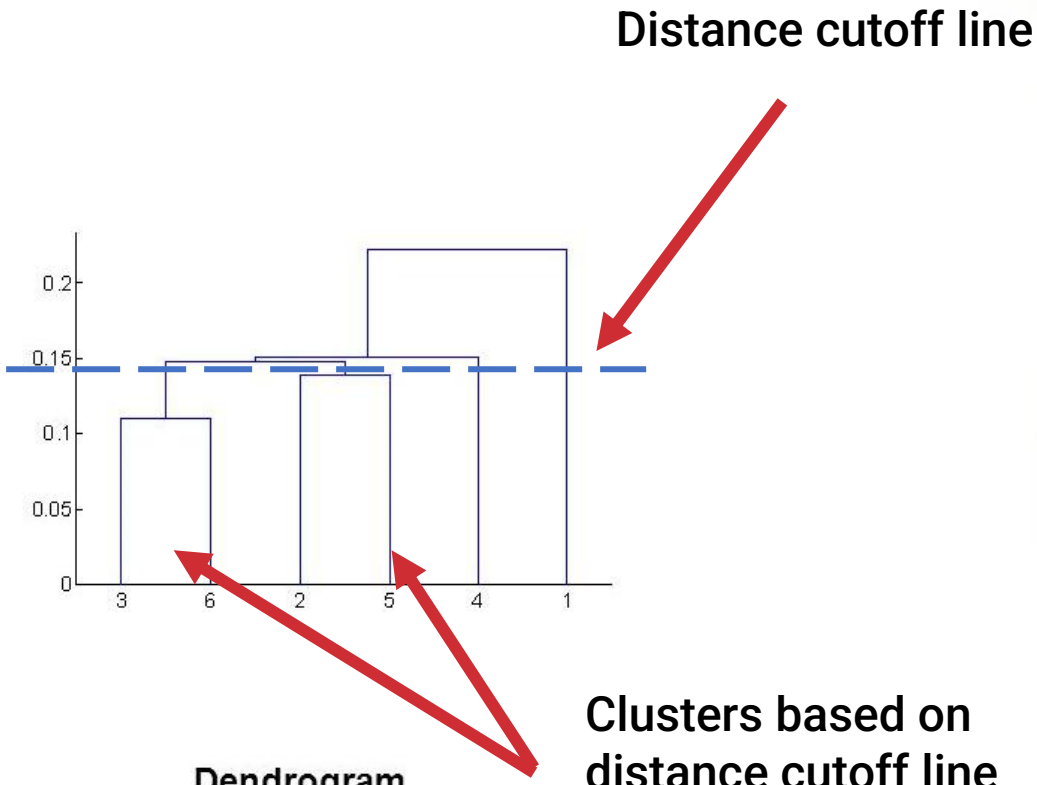
Main Difference in the “K” Algorithms



Hierarchical



Nested Clusters



Dendrogram

Matching Algorithms to Variable Types

Algorithm	Variable Types
BIRCH / 2-Step	Any, multiple at the same time
Hierarchical	Any, but stick to one kind at a time and use well-paired distance measure
K-means	Continuous
K-modes	Categorical
K-medoids	Any, multiple at the same time with the right distance measure (Gower)

Algorithms Strengths/Weaknesses

Algorithm	+/-
BIRCH / 2-Step	Flexible variable types, fast compute, large data; there is an element of a black box
Hierarchical	Flexible modeling, highly explainable; limited to small data
K-Means, K-Modes	Easy, fast, large data; sensitive to K, curse of dimensionality, clusters same sized
K-medoids	Like K-Means but more flexible variable types and more costly tuning & compute

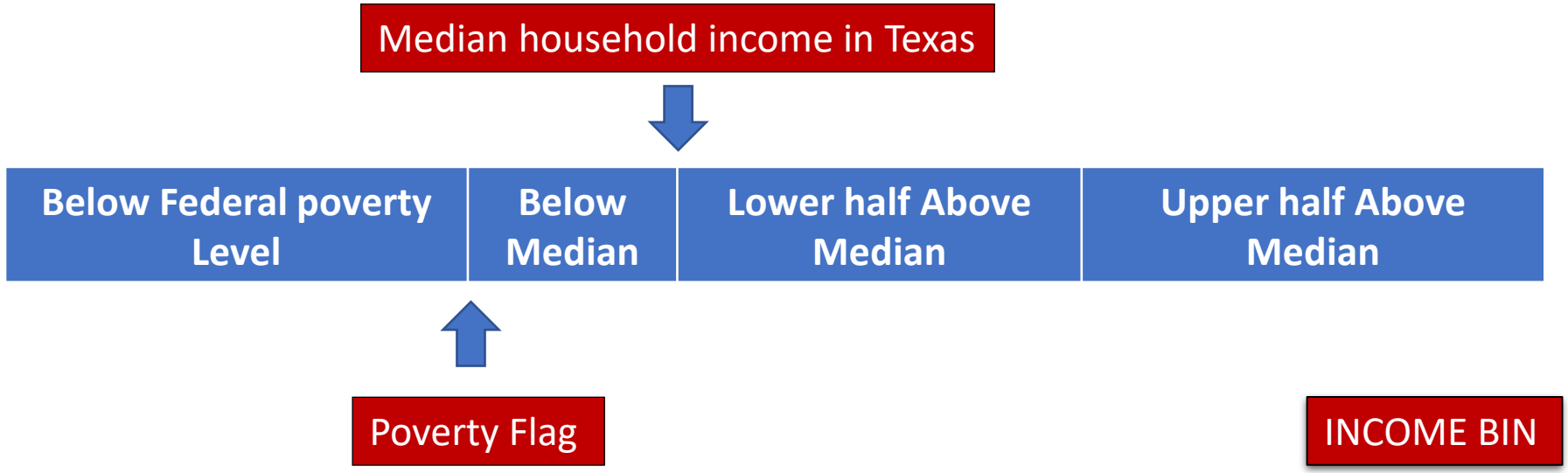
Example 2

Adult Student Persona Design

Example 2: K-means/K-mode Clustering

Data Processing

- Population: 24 years and older adults enrolled for Fall 2021
- Features : Academics, Demographic, Financial , Household and Veteran status.
- Mix of discrete and continuous variables with majority of them being categorical data.

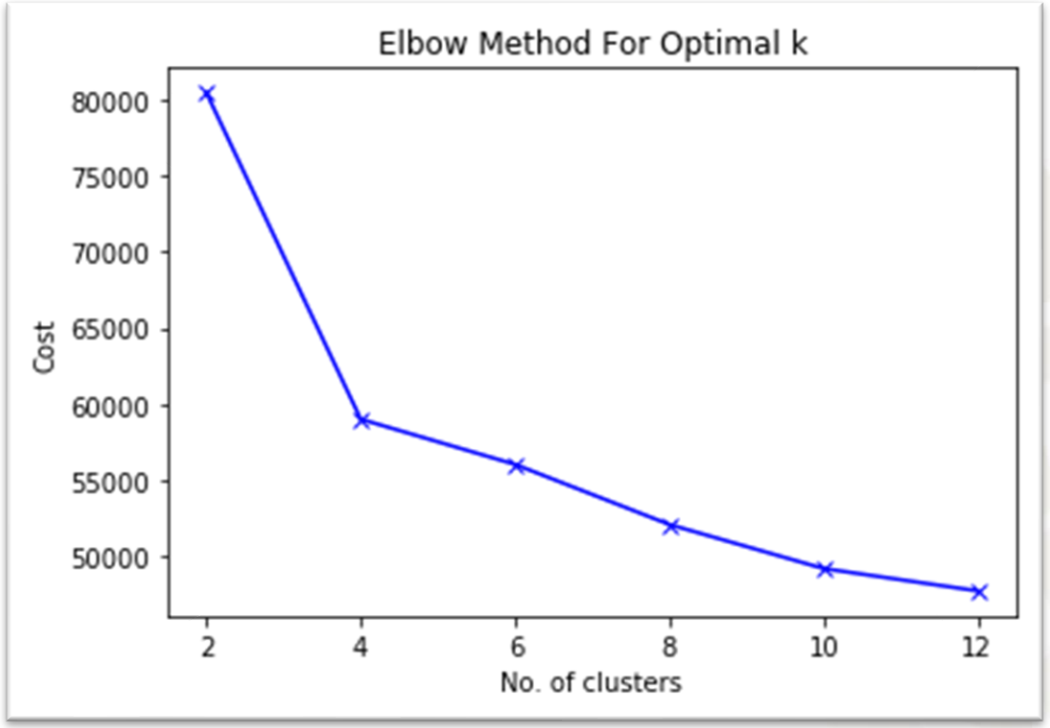


Example 2: K-means/K-mode Clustering

Python Workflow

- Randomly select the K initial centers
- Repeat
  1. Assign the samples to nearest center
  2. Update means/modes based on newly formed cluster
  3. Calculate the cost (SSE/Sum of dissimilarity)
- Stop when cluster centers converges

Minimize the cost function



Sample 1

V1	V2	V3	V4	V5
0	1	1	0	1

V1	V2	V3	V4	V5
1	0	0	0	1

Sample 2

Using frequency-based method to calculate mode instead of the mean of the sample

Dissimilarity measure

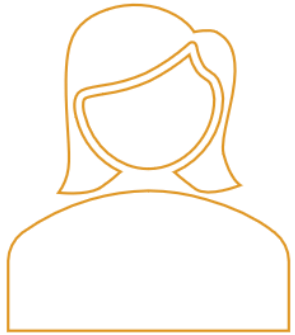
0: Mismatch 1: Match

1+1+1+1 = 4

Example 2: K-means/K-mode Clustering

Output

Student Personas



Academic Status

- Part-Time Student
- Not First Time in College
- Continuing DC Student
- Credit Student
- No Prior Dual Credit

Educational Path

- Studying Associate of Science at Richland Campus
- Have attended High School from out-of-state

Financial Status

- Have dependent children and other dependents
- Working 35+ hours/week
- Above poverty line but below median income

Demographic

- Female
- Gen Y/Millennials
- Black/African-American

Family Status

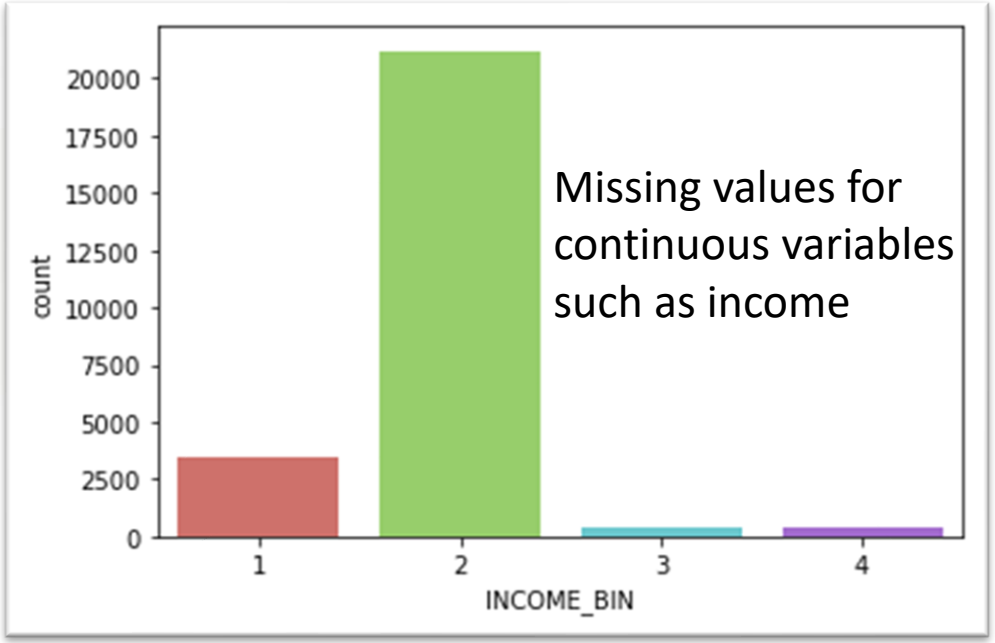
- Independent
- Living alone

Not VA

Example 2: K-means/K-mode Clustering

Challenges

Missing values for categorical variables such as employment status



Possible solutions

- Revisit our data warehouse to obtain as much information as we can to fill in the missing values
- Imputation with K-Nearest neighbors with Hamming distance for categorical data
- Imputation with K-Nearest neighbors with Euclidean distance for continuous data

Questions?

Thank you!

Jeremy Anderson, Associate Vice Chancellor of Strategic Analytics, jeremy.anderson@dcccd.edu

Dillon Lu, Data Analyst, DLu@dcccd.edu