SEGMENTATION OF PROXIMITY MEASURES

UBCO CAPSTONE PROJECT + STATISTICS CANADA

OUR TEAM

Noman Mohammad
BSc in Computer Science, Minor in Mathematics

Data Enthusiast & programming

Ricky Heinrich
BSc Physical Sciences Minor in Business

Urban planning, sustainability & data enthusiast

Jonah Edmundson
BSc in Biology, Minor in Philosophy

Statistical Programming in R

Avishek Saha
BSc in Computer Science & Engineering

Business development & Tech Innovation

AGENDA

01 INTRODUCING THE PROJECT

02 METHODS AND RESULTS

03 DISCUSSION AND CONCLUSIONS

INTRODUCING THE PROJECT

OUR CLIENT:

STATISTICS CANADA

Statistics Canada Statistique Canada

Canada’s national statistics agency

Main contact: Jérôme Blanchet, Unit Head - Data Science Engineering - Center for Special Business Projects

PROXIMITY MEASURES DATABASE

Created by:

Data Exploration and Integration Lab (DEIL) at Statistics Canada in collaboration with the Canada Mortgage and Housing Corporation (CMHC)

Purpose:

Provide national and local policy makers granular and comparable measures of proximity to services and amenities

DISSEMINATION BLOCK (DB)

- Smallest geographic area for which population and dwelling counts are disseminated by StatCan

DISSEMINATION BLOCK (DB)

- Smallest geographic area for which population and dwelling counts are disseminated by StatCan
- Areas bounded on all sides by roads or other natural feature boundaries

DATASET

Dissemination Blocks — 489 676

10 Amenities

- Employment
- Health care
- Child care
- Primary education
- Secondary education
- Libraries
- Grocery stores
- Pharmacies
- Public transit
- Neighbourhood parks

Extra Columns — Populations, CMA types, indicators, Index of Remoteness

PROXIMITY MEASURES

- DBs that have the amenity available within a threshold distance are assigned a proximity value

*motivating example -- not true values

PROXIMITY MEASURES

- DBs that have the amenity available within a threshold distance are assigned a proximity value
- Calculated using a gravity model

*motivating example -- not true values

PROXIMITY MEASURES

- DBs that have the amenity available within a threshold distance are assigned a proximity value
- Calculated using a gravity model
  - distances

*motivating example -- not true values

PROXIMITY MEASURES

- DBs that have the amenity available within a threshold distance are assigned a proximity value
- Calculated using a gravity model
  - distances
  - mass

*motivating example -- not true values

PROXIMITY MEASURES

- DBs that have the amenity available within a threshold distance are assigned a proximity value
- Calculated using a gravity model
  - distances
  - mass

- Continuous measure normalized between 0 and 1 across Canada

*motivating example -- not true values

PROXIMITY MEASURES

- DBs that have the amenity available within a threshold distance are assigned a proximity value
- Calculated using a gravity model
  - distances
  - mass
- Continuous measure normalized between 0 and 1 across Canada

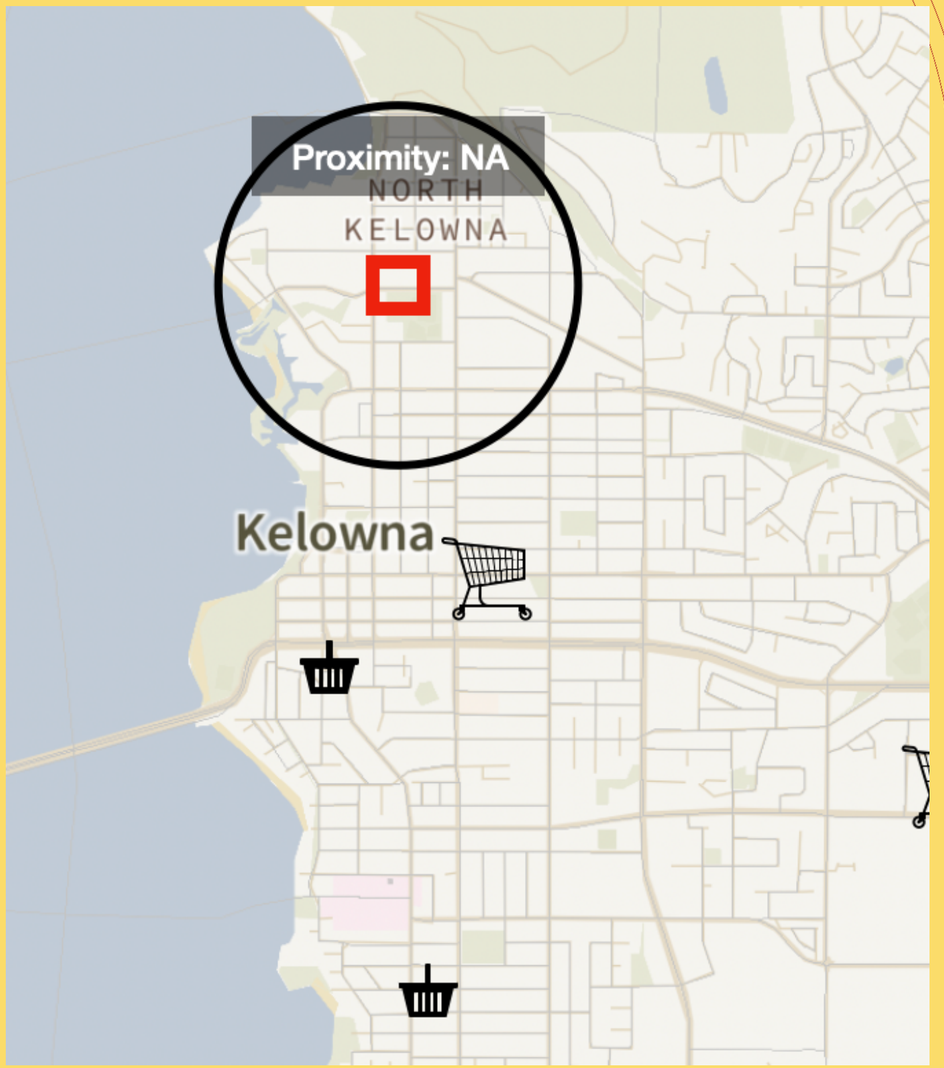
- No amenity?

*motivating example -- not true values

PROXIMITY MEASURES

- DBs that have the amenity available within a threshold distance are assigned a proximity value
- Calculated using a gravity model
  - distances
  - mass
- Continuous measure normalized between 0 and 1 across Canada

- No amenity? No value



*motivating example -- not true values

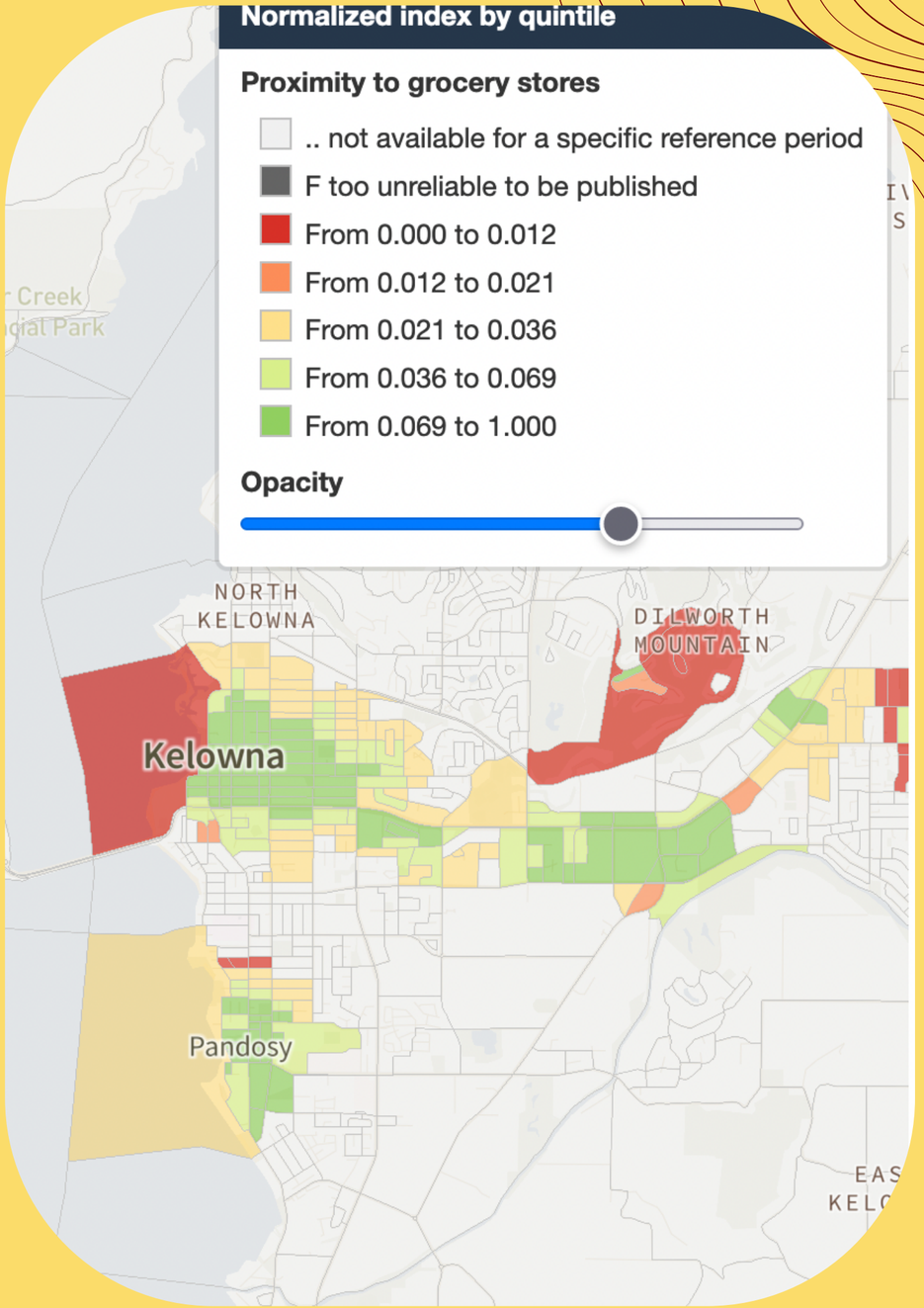
CHALLENGE:

- Continuous measure is hard to interpret

CURRENTLY:

- Visualization segments the continuous proximity measures by quintiles

- Is there a better way?



MOTIVATION

- Improved usefulness of proximity measures by converting to categorical measures

better understanding of local access to amenities

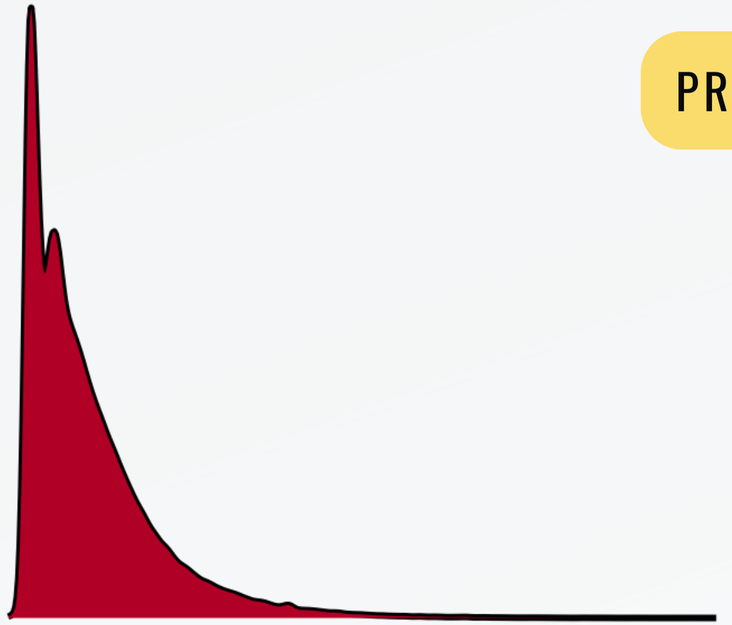
GOAL

- Segment proximity measures for each amenity, identify stable cutoffs
- Provide clustering workflow

METHODS & RESULTS

DISTRIBUTION

Problem: heavily skewed



ORIGINAL

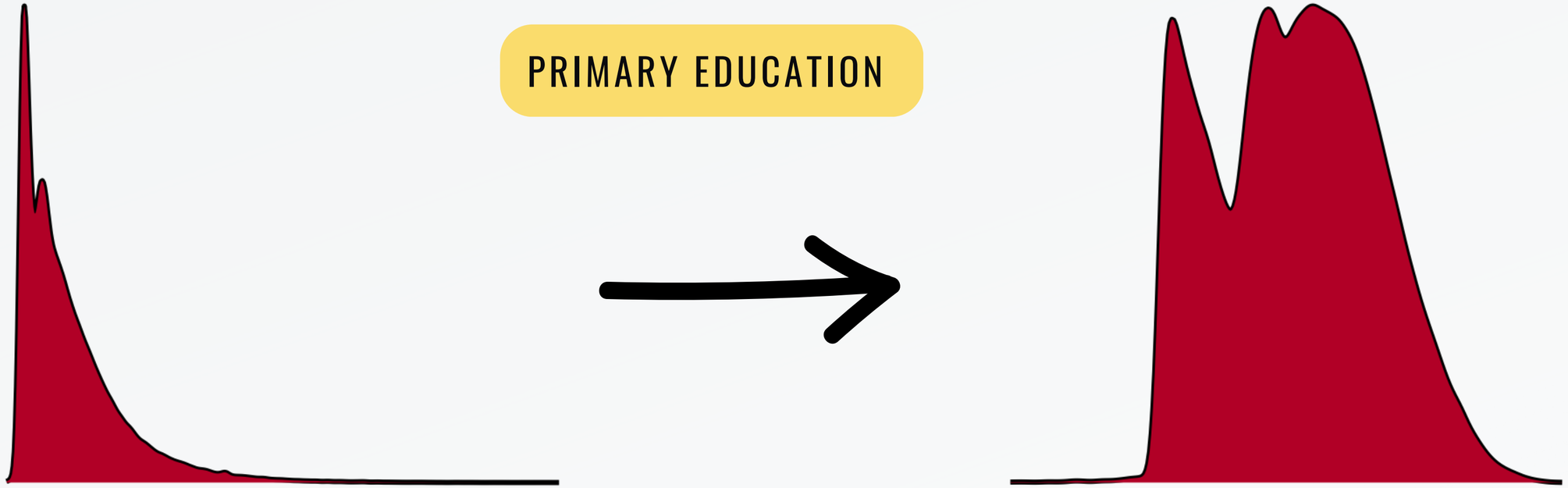
PRIMARY EDUCATION

*same trend reflected across all 10 amenities

LOG TRANSFORMATION

Reduces skewness

PRIMARY EDUCATION



ORIGINAL

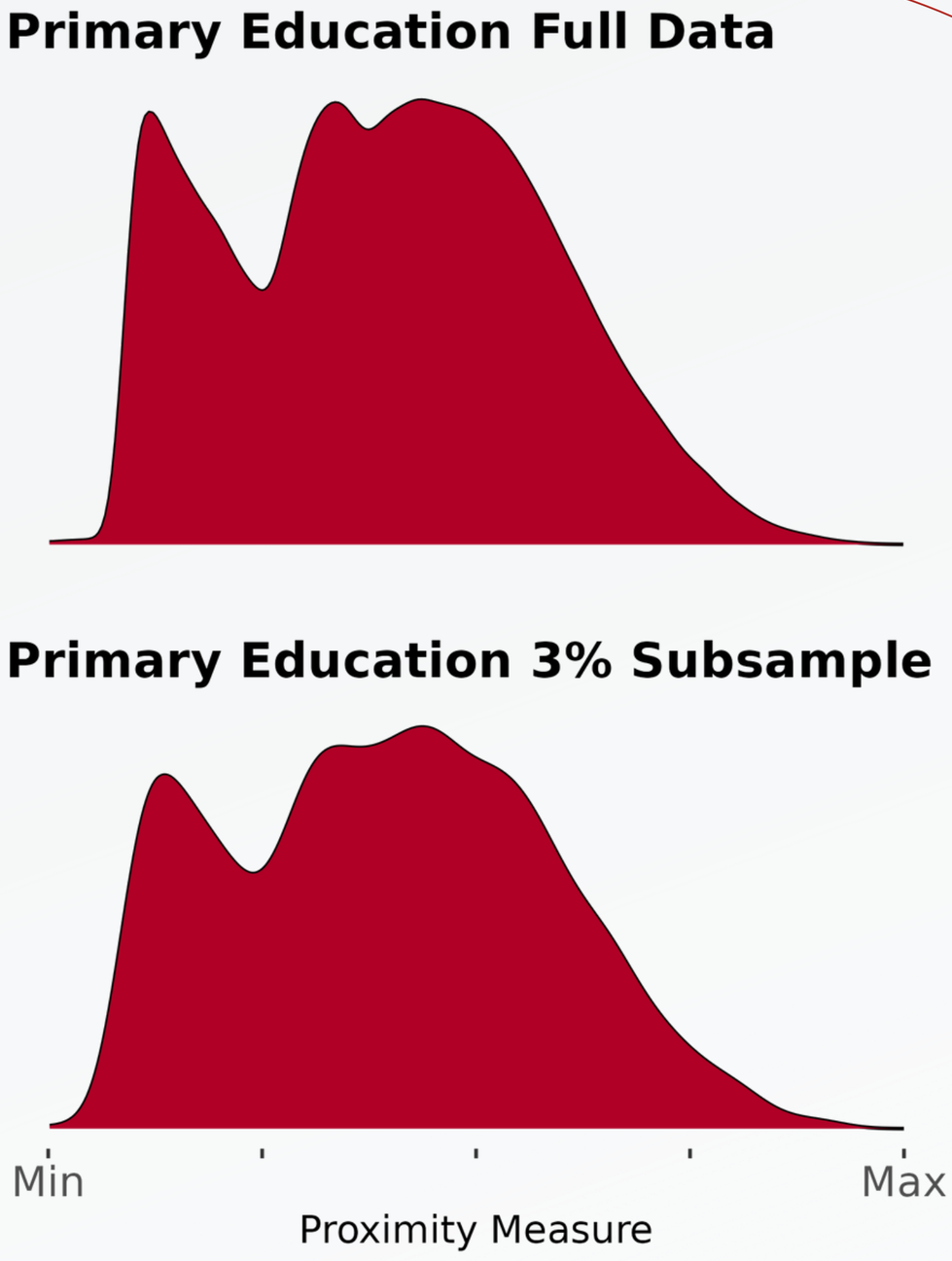
LOG-TRANSFORMED

*same trend reflected across all 10 amenities

SUBSAMPLING

Resolving computational constraints

```
Error: cannot allocate vector of
 size 893.3 Gb
```

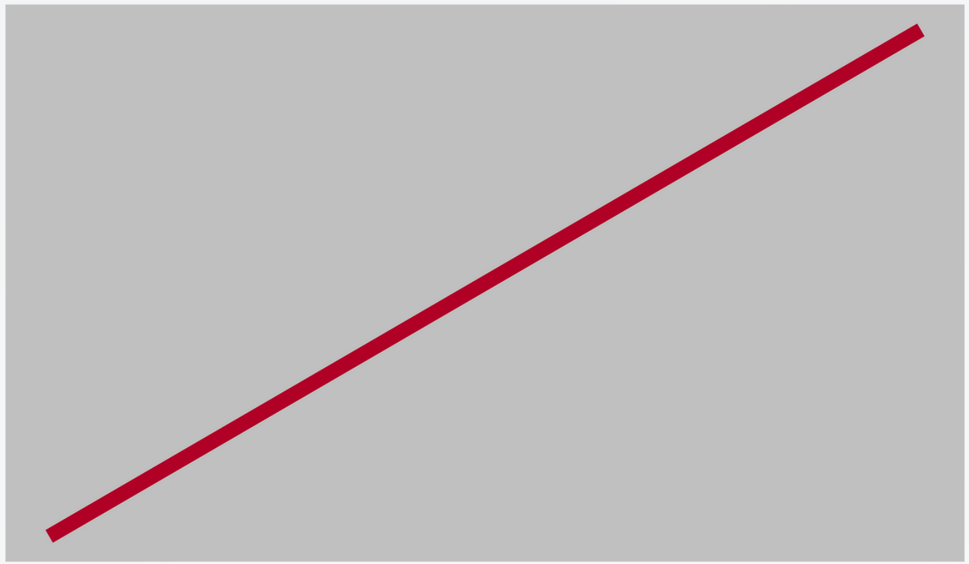


IS THE DATA EVEN CLUSTERABLE?

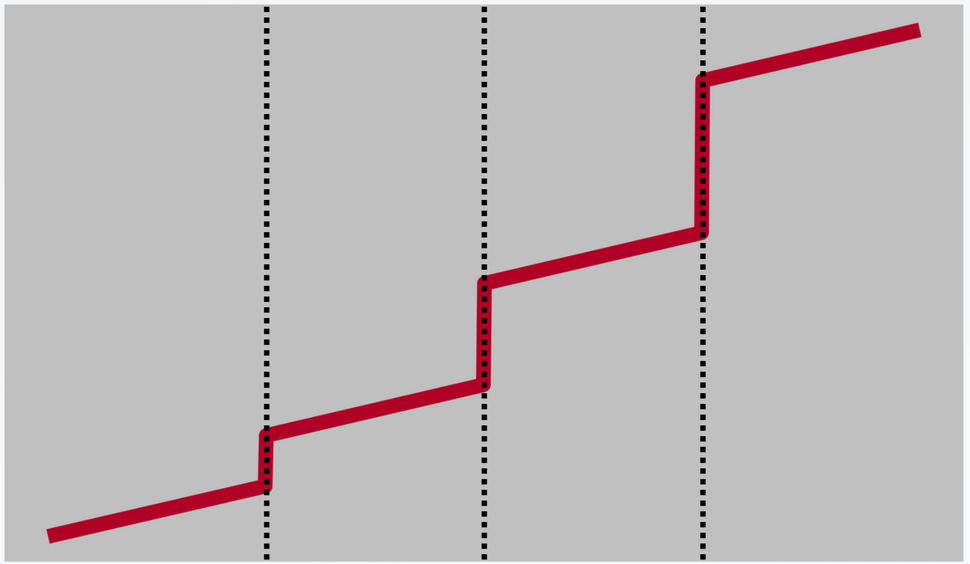
CLUSTERING TENDENCY

Check if data is clusterable

SORT PLOT EXAMPLE



NOT CLUSTERABLE

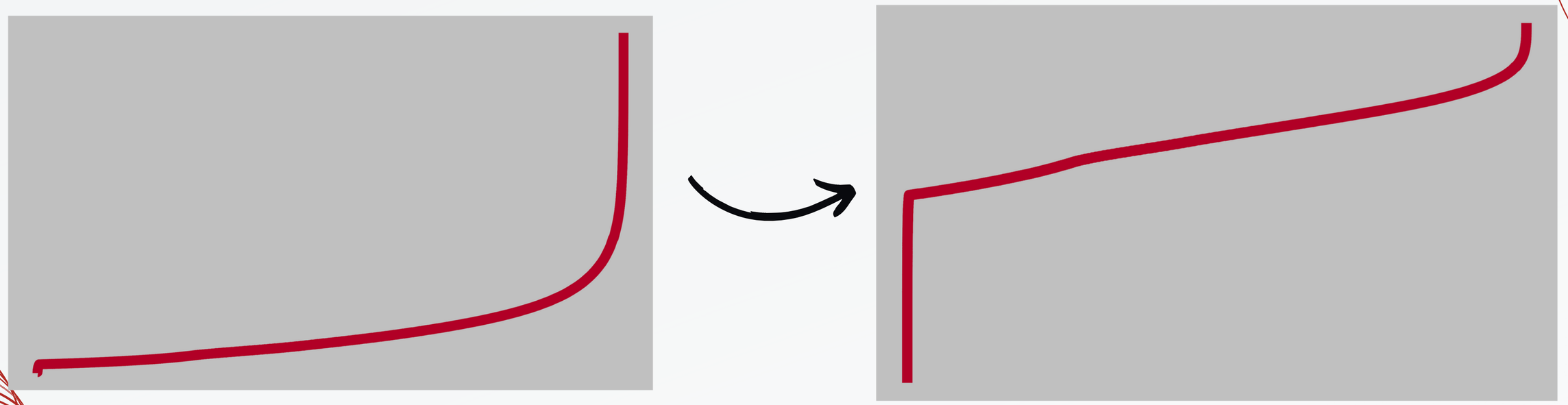


CLUSTERABLE

CLUSTERING TENDENCY

Low clustering tendency, even after log-transforming the data

SORT PLOT



ORIGINAL

LOG-TRANSFORMED

*Primary Education visualized, same trend reflected across all 10 amenities

GOAL

- Segment proximity measures for each amenity, identify stable cutoffs
- Provide clustering workflow

GOAL

- Segment proximity measures for each amenity, identify stable cutoffs

- Provide clustering workflow

SEGMENTATION TECHNIQUES

Intuitive segmentation

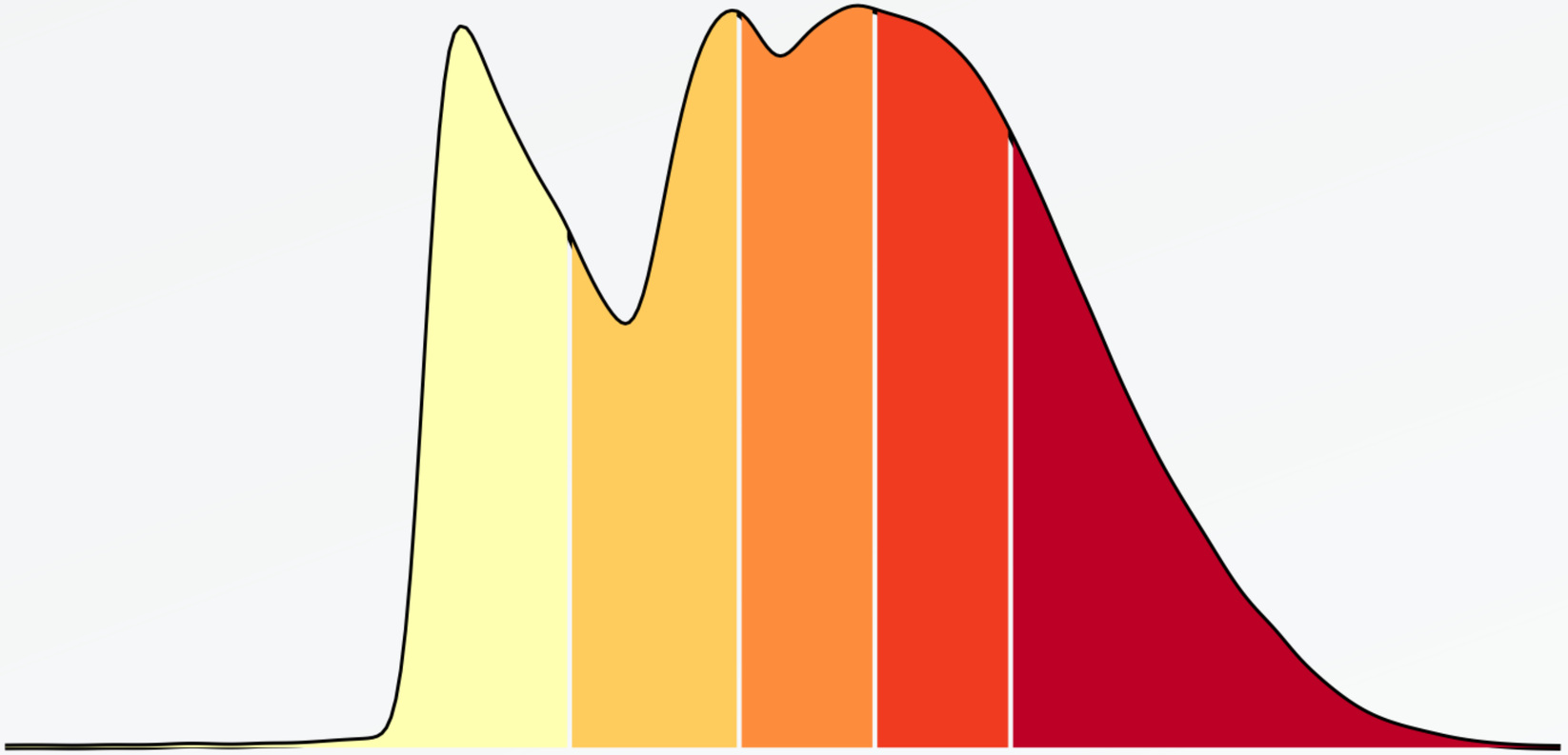
- Quintiles
- Minima Identification

Clustering algorithms

- Density Based
- Distribution Based
- Centroid Based

QUINTILES

Intuitive Segmentation

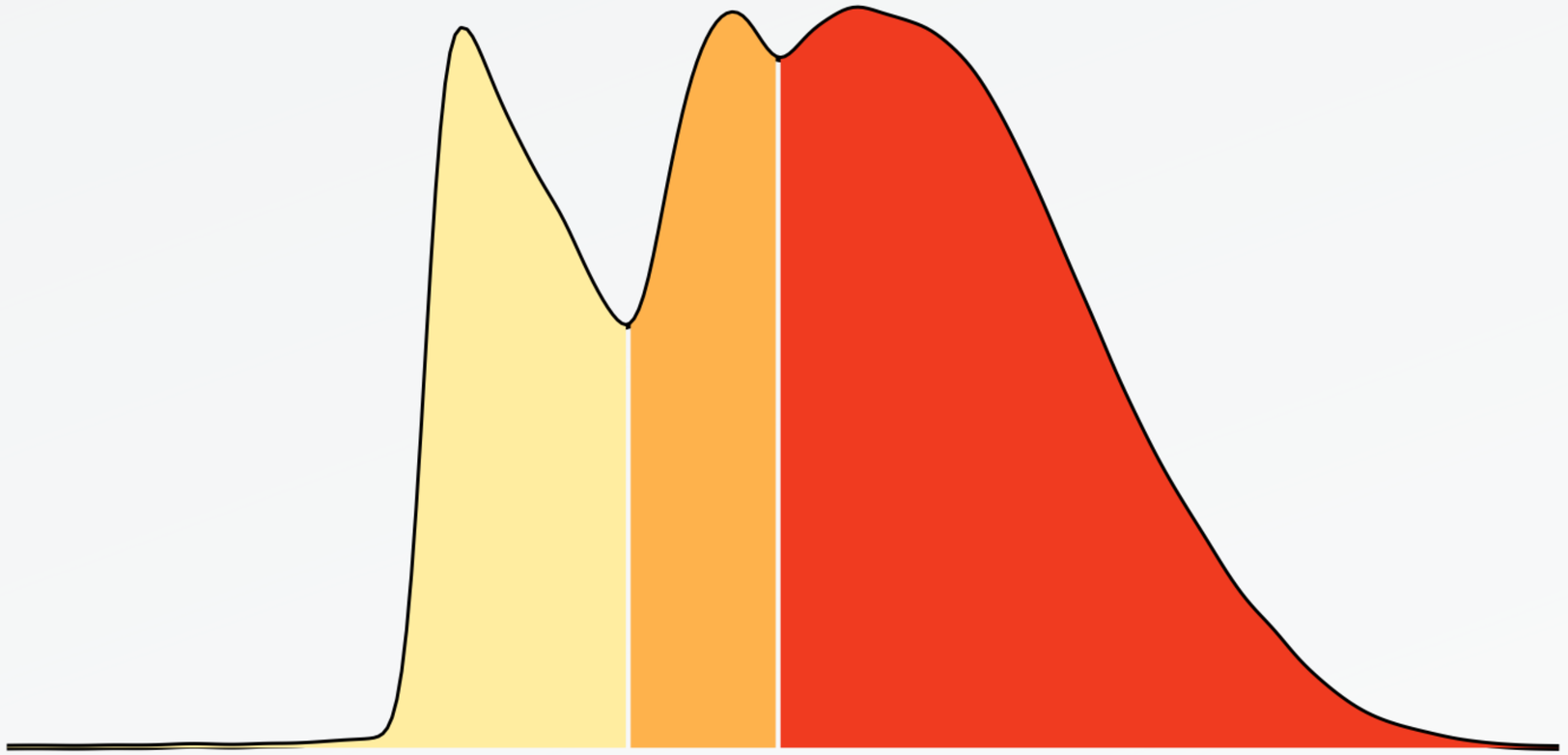


0.0416, 0.072, 0.1105, 0.172

*log transformed Primary Education visualized

MINIMA IDENTIFICATION

Intuitive Segmentation



0.04966, 0.08169

*log transformed Primary Education visualized

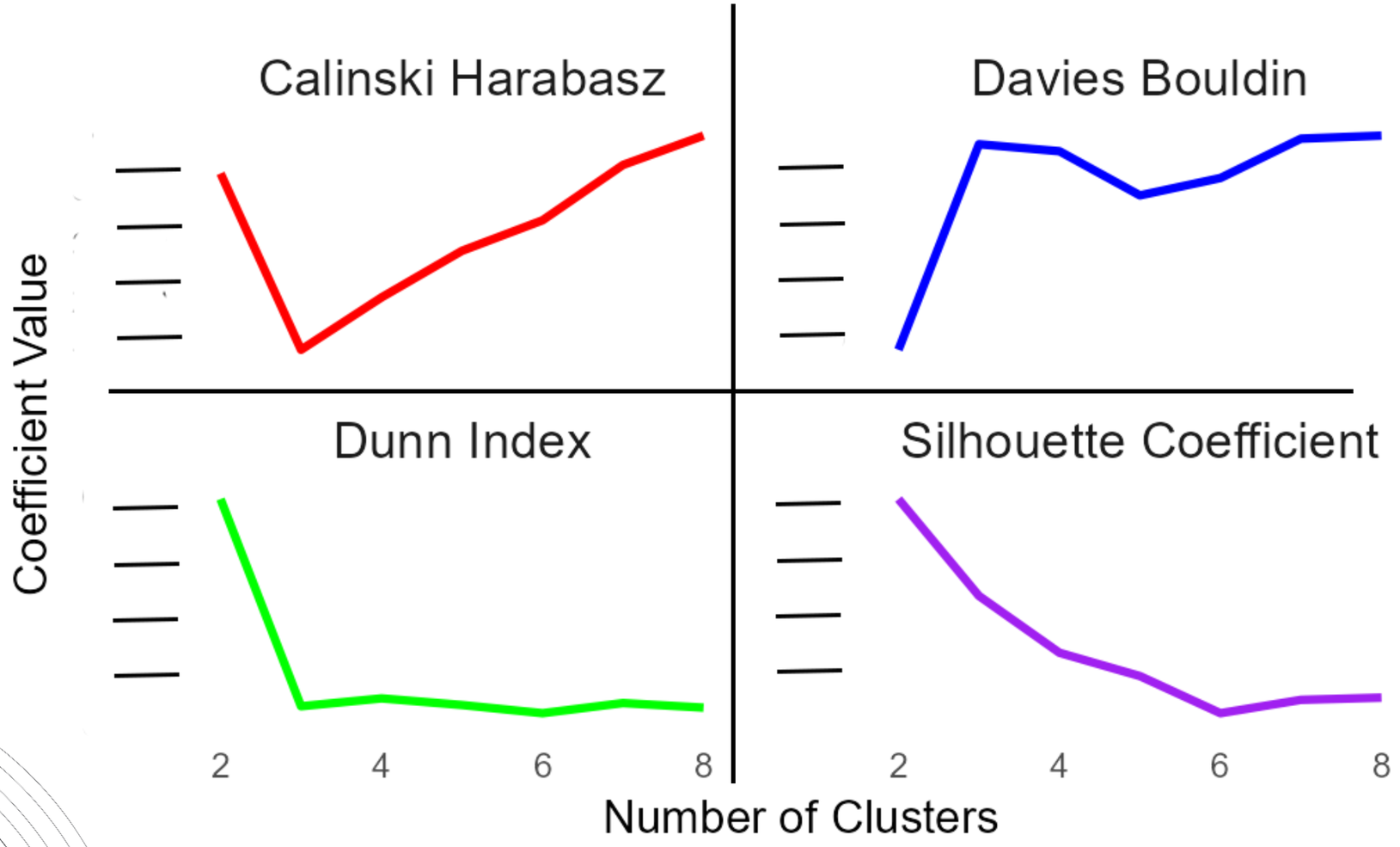
OPTIMAL NUMBER OF CLUSTERS

- Some algorithms require us to specify a number of clusters

- Validation metrics
  - Silhouette coefficient
  - Dunn index
  - Calinski-Harabasz
  - Davies-Bouldin

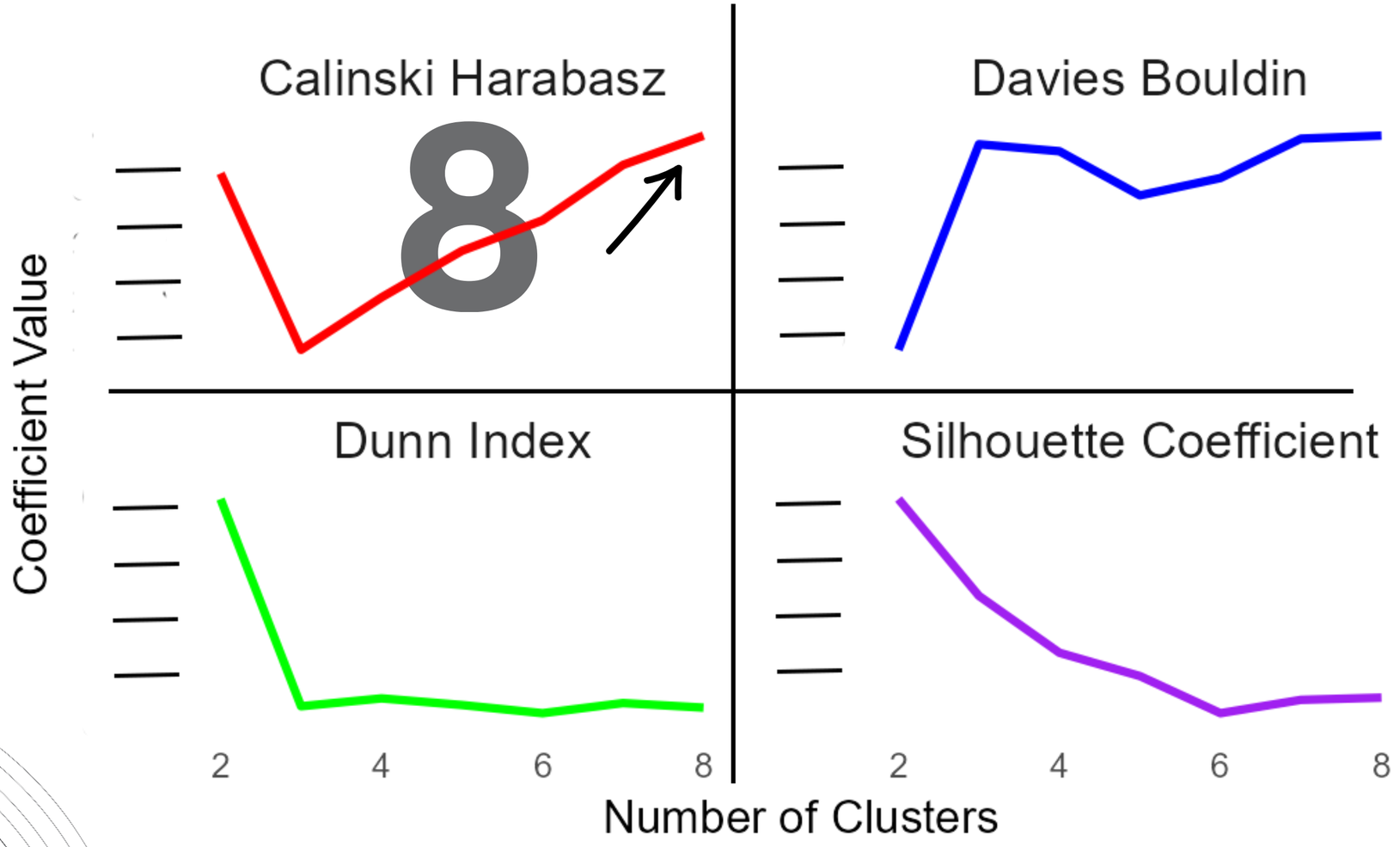
- Interval validation schemes: essentially all different measures of inter-cluster and intra-cluster spread

OPTIMAL NUMBER OF CLUSTERS



*MixAll algorithm on Primary Education

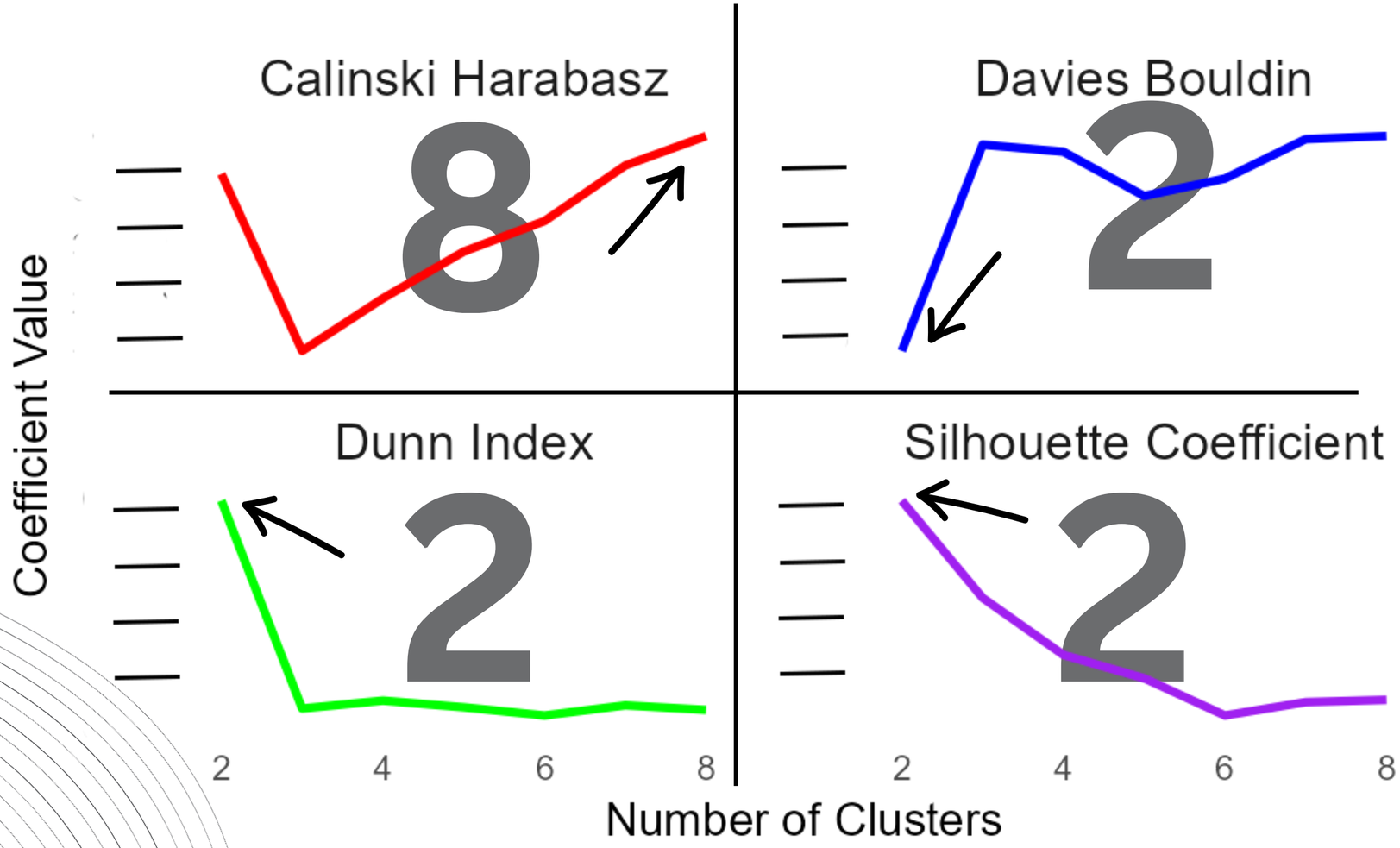
OPTIMAL NUMBER OF CLUSTERS



*MixAll algorithm on Primary Education

OPTIMAL NUMBER OF CLUSTERS

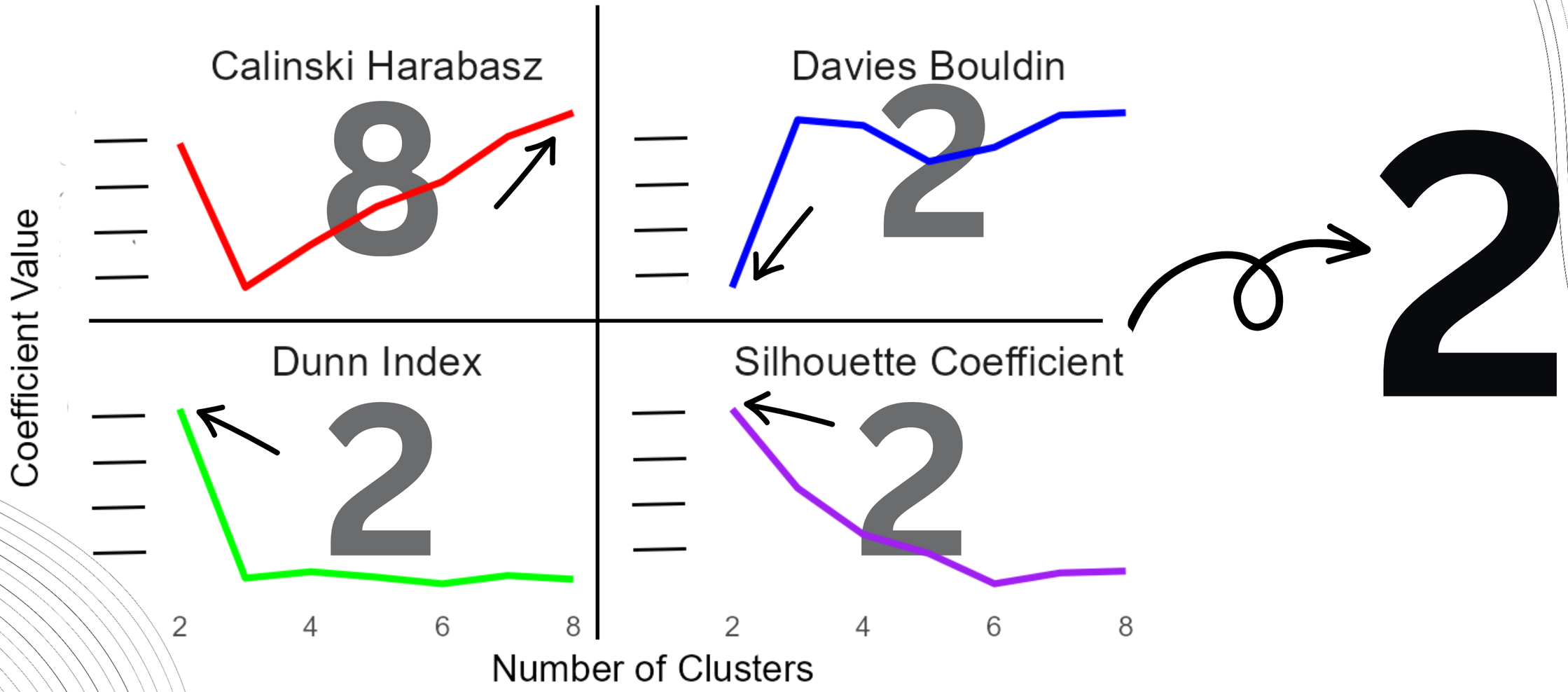
Choose the most frequent number



*MixAll algorithm on Primary Education

OPTIMAL NUMBER OF CLUSTERS

Choose the most frequent number

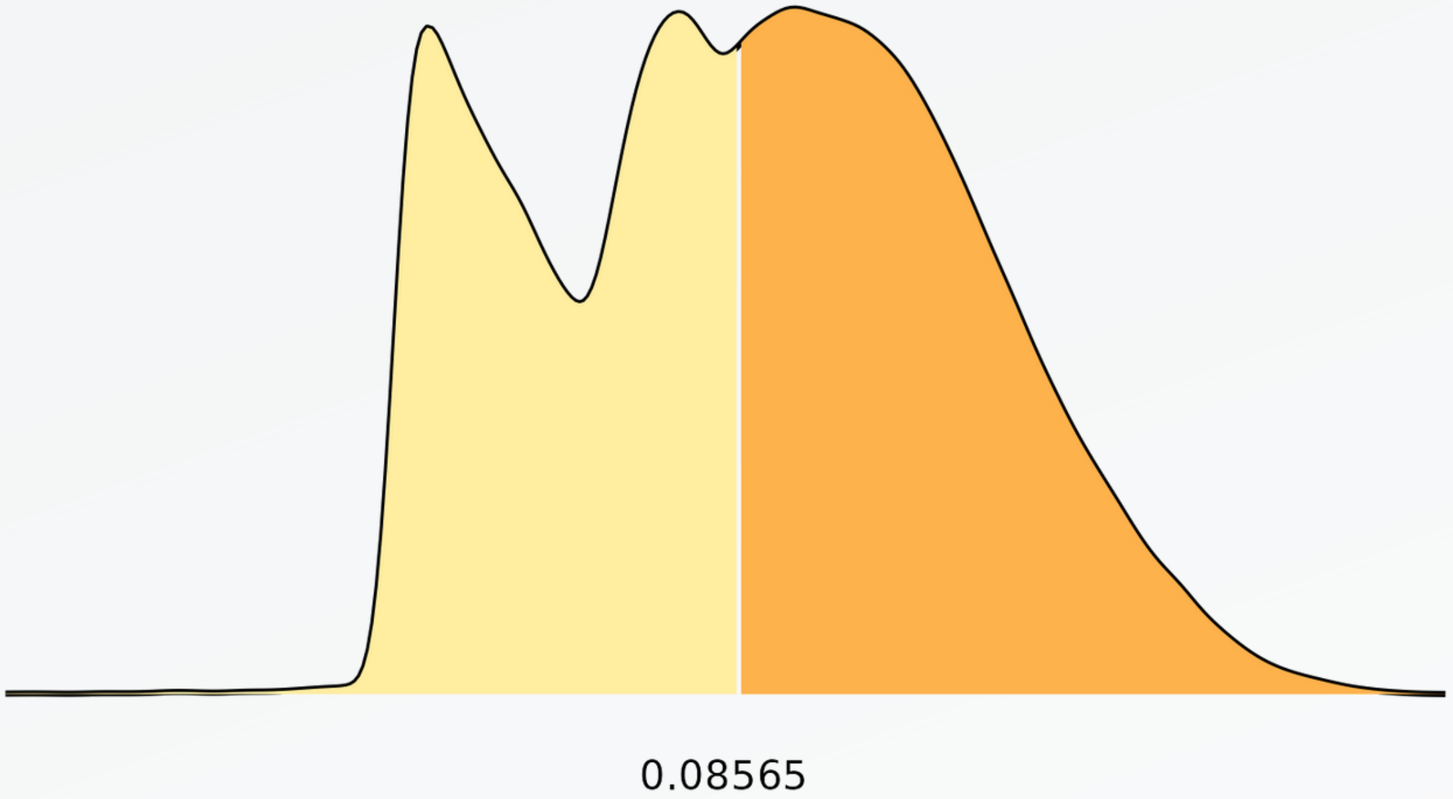


***MixAll algorithm on Primary Education**

DISTRIBUTION BASED

MixAll

- leveraging strength at handling diverse distributions

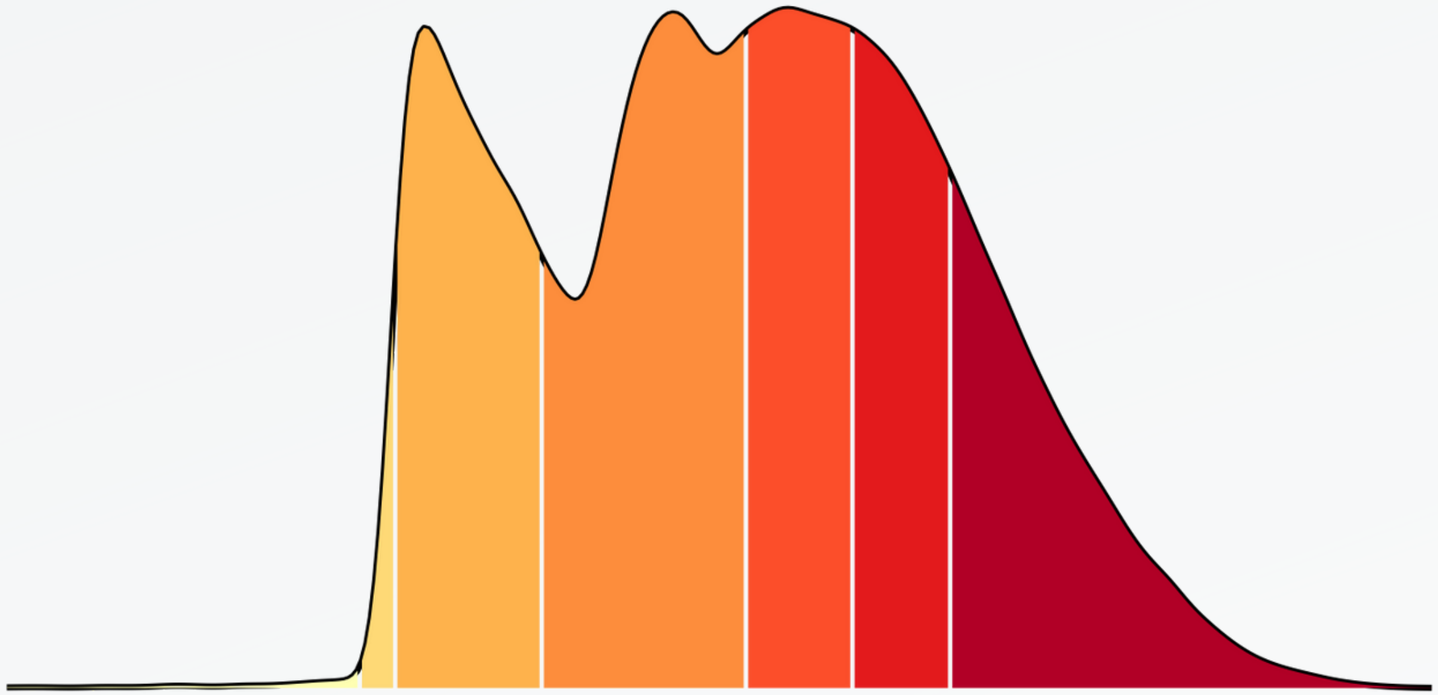


*log transformed Primary Education visualized

DISTRIBUTION BASED

MCLUST

- enhanced performance by aligning data with Gaussian distribution assumption



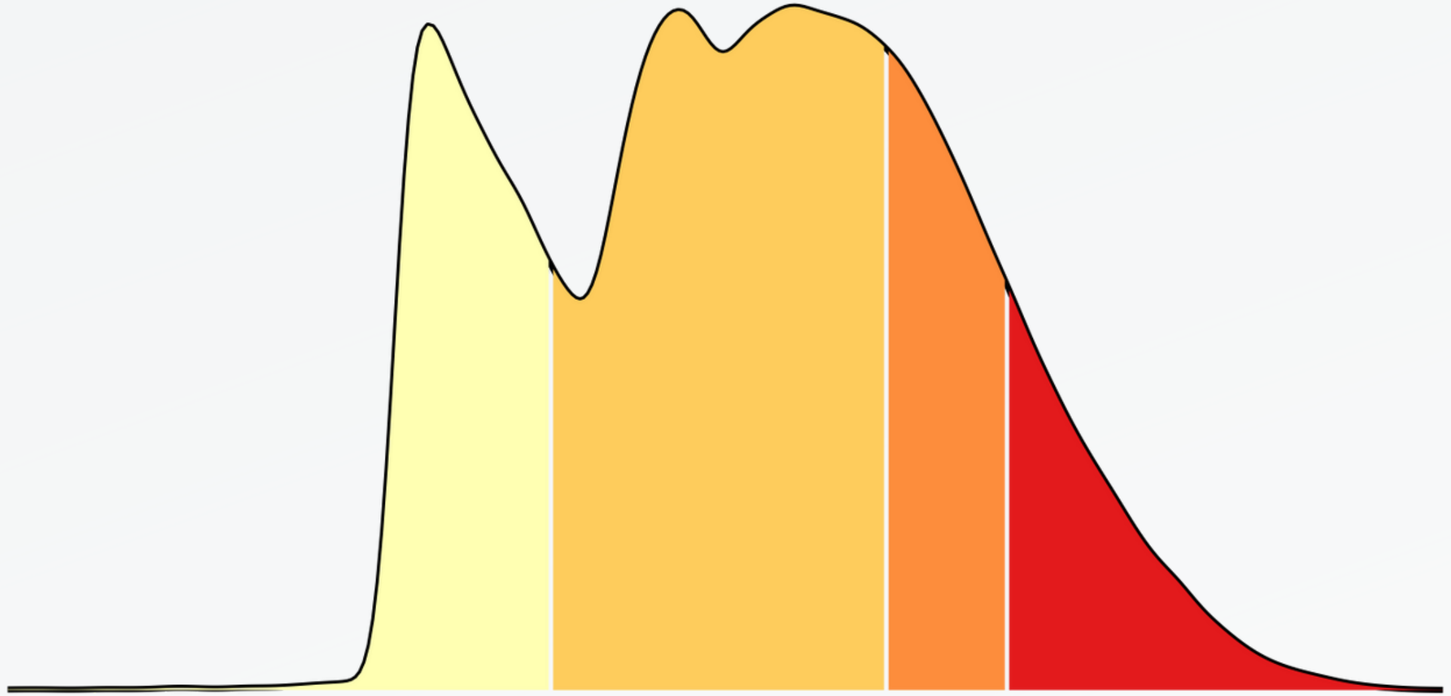
0.023, 0.027, 0.044, 0.09, 0.131, 0.185

***log transformed Primary Education visualized**

DENSITY BASED

HDBSCAN

- flexibility in identifying varying densities and handling noise



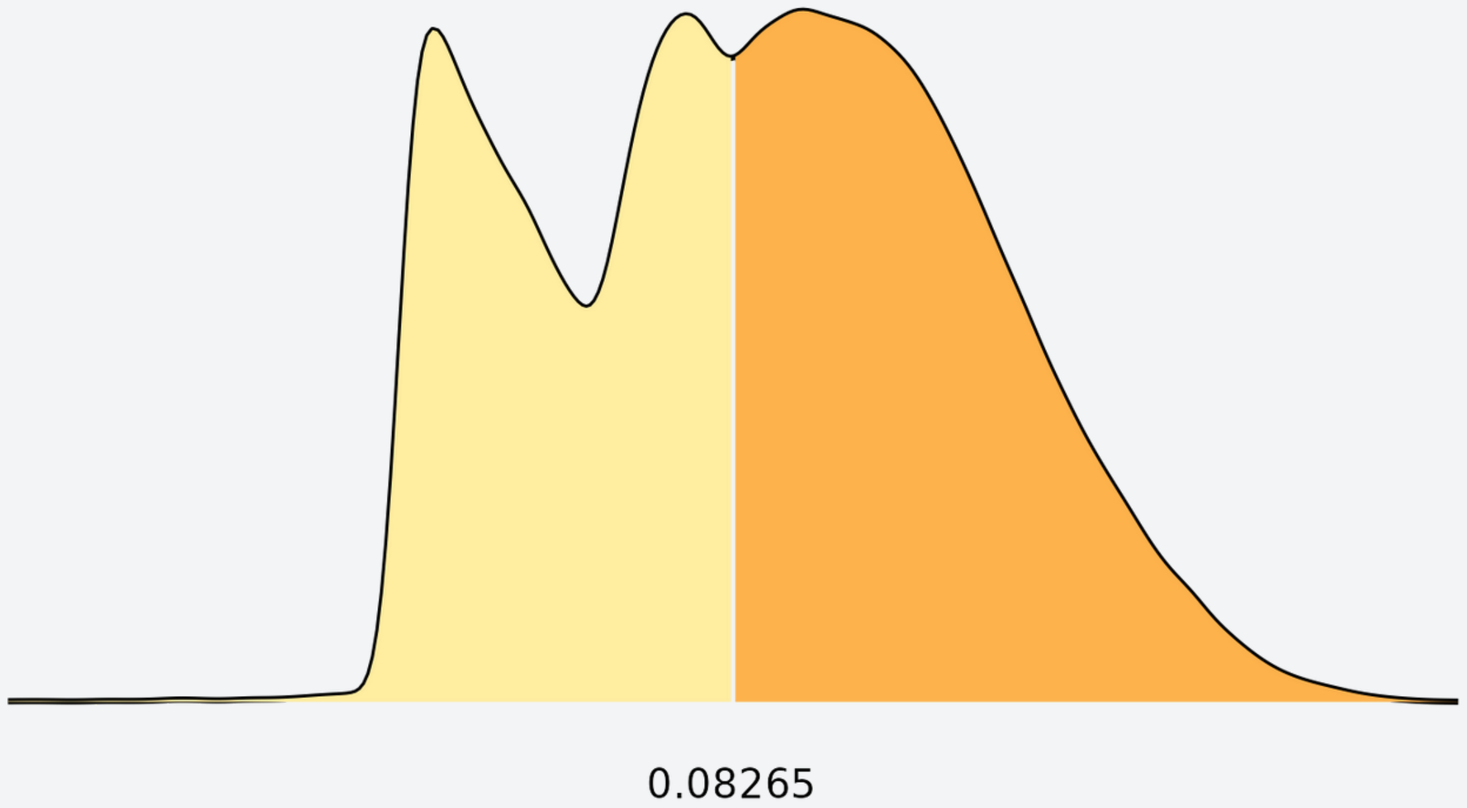
0.04495, 0.22045, 0.1449

*log transformed Primary Education visualized

CENTROID BASED

Partitioning around medioids (PAM)

- robustness to outliers



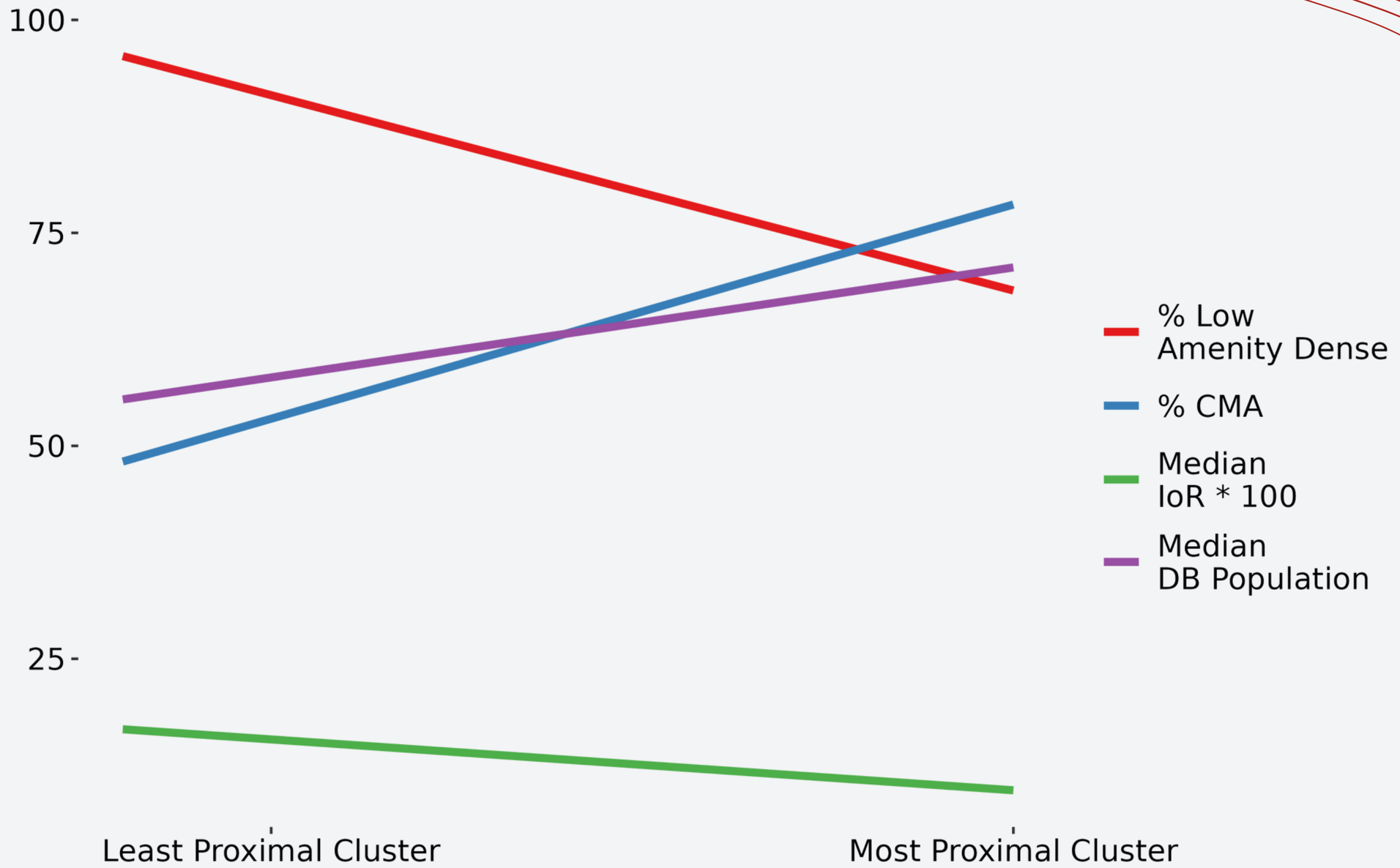
*log transformed Primary Education visualized

CLUSTER PROFILING

- Median DB population
- Median Index of Remoteness
- Percentage of DBs that are in a Census Metropolitan Area
- Percentage of DBs with low amenity density

Most of the time, different groups had distinct characteristics

CLUSTER PROFILING



*Primary Education visualized, same trend reflected across all 10 amenities

ANALYSIS

"Best Method" Fallacy

- The expectation or belief that the “true” or “best” method for solving a problem can be found using pure math
- In the unsupervised context, there is no “one true solution”
- Domain knowledge is also required

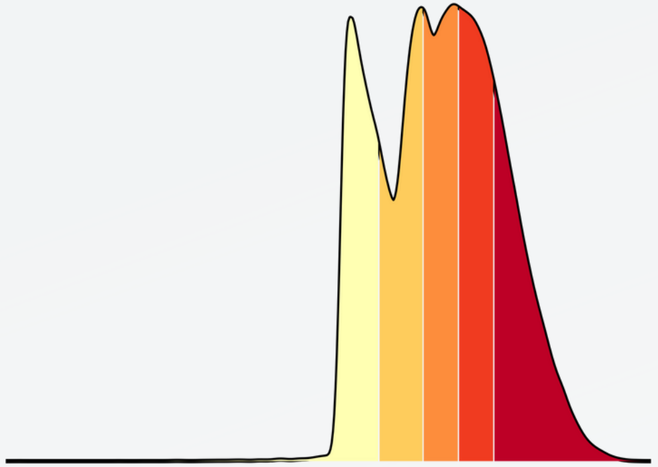
ANALYSIS

Lack of consistency in:

- Number of clusters
- Location of cutoffs

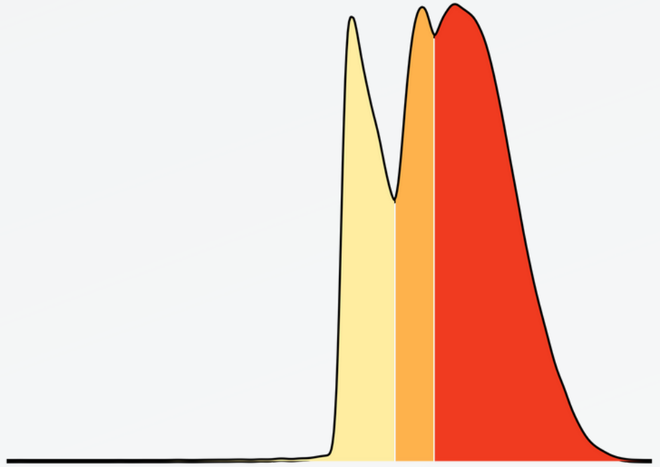
Lack of consistency

Quintiles



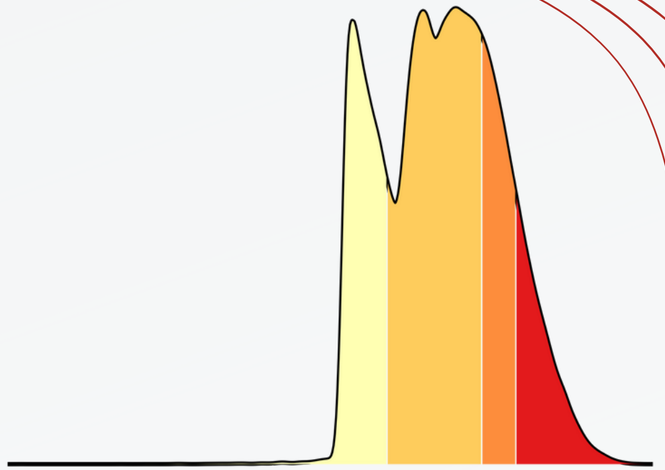
0.0416, 0.072, 0.1105, 0.172

Minima



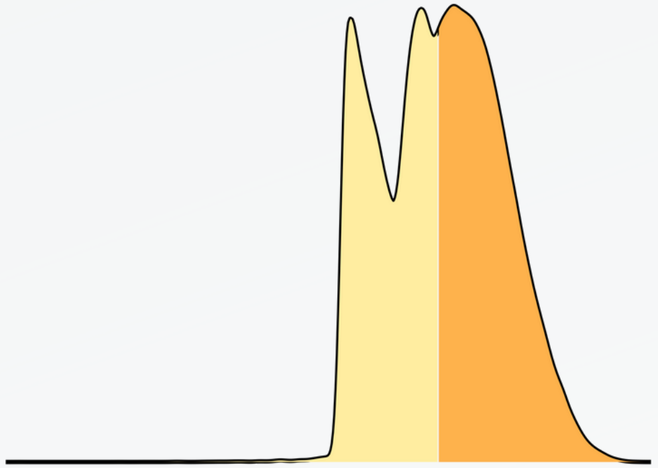
0.04966, 0.08169

HDBSCAN



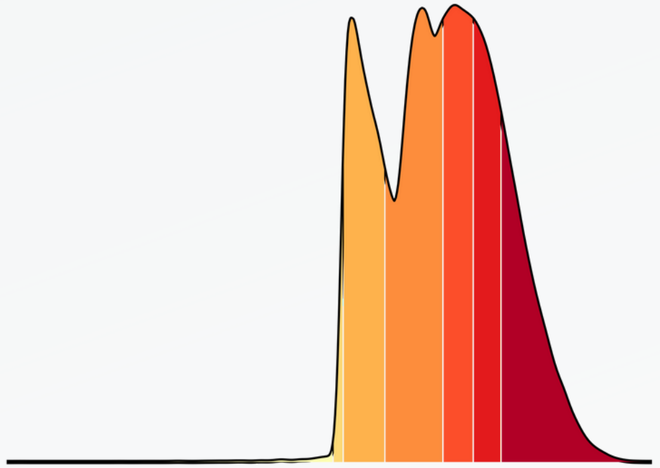
0.04495, 0.22045, 0.1449

MixAll



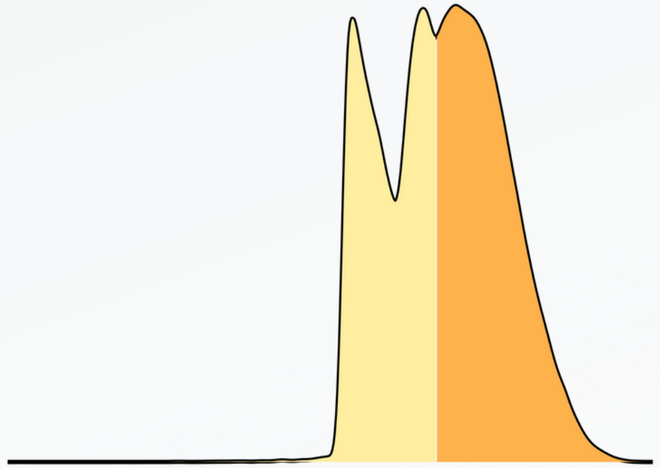
0.08565

MCLUST



0.023, 0.027, 0.044, 0.09, 0.131, 0.185

PAM k-means



0.08265

*log transformed Primary Education visualized

ANALYSIS

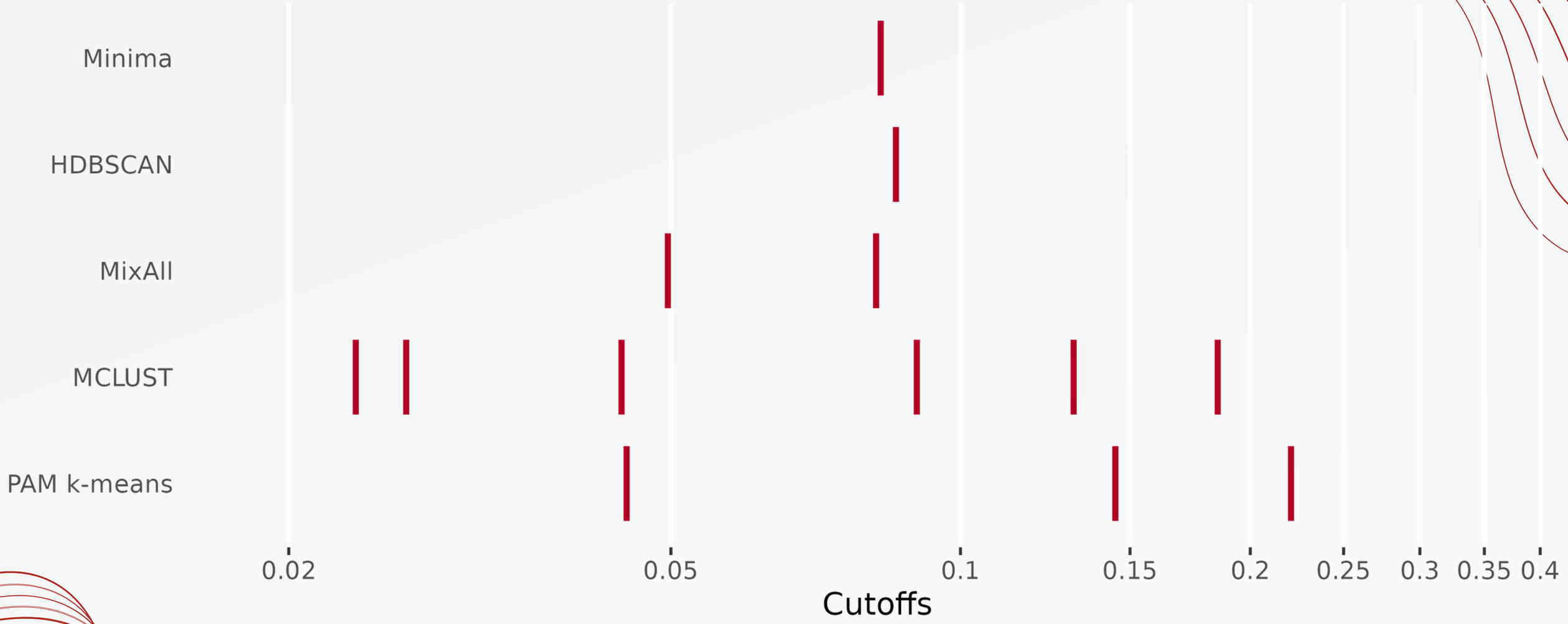
Lack of consistency in:

- Number of clusters
- Location of cutoffs

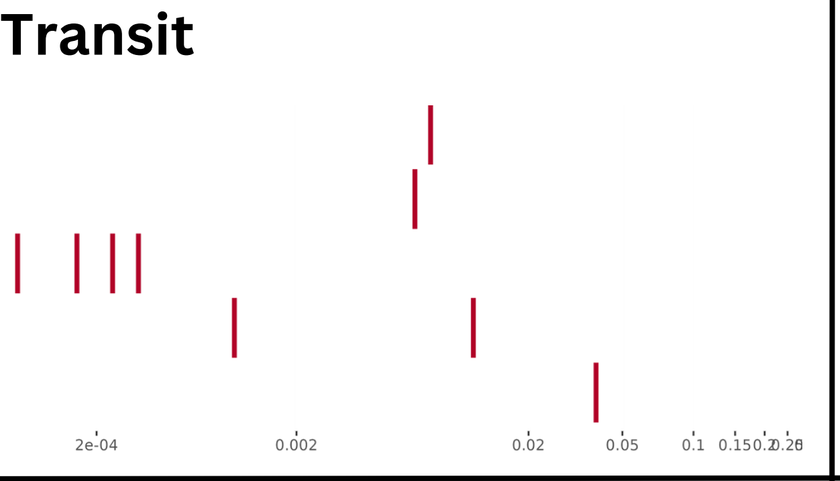
This suggests that the data:

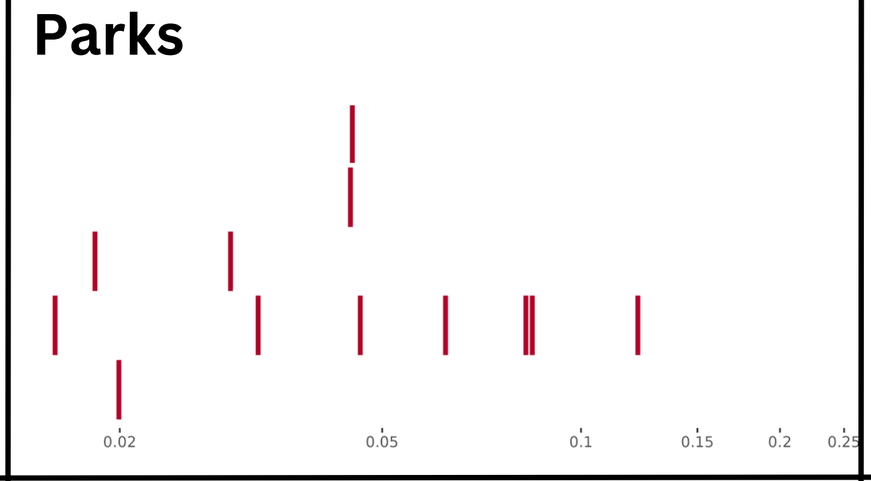
- is sensitive to the clustering algorithm used
- has a low clustering tendency

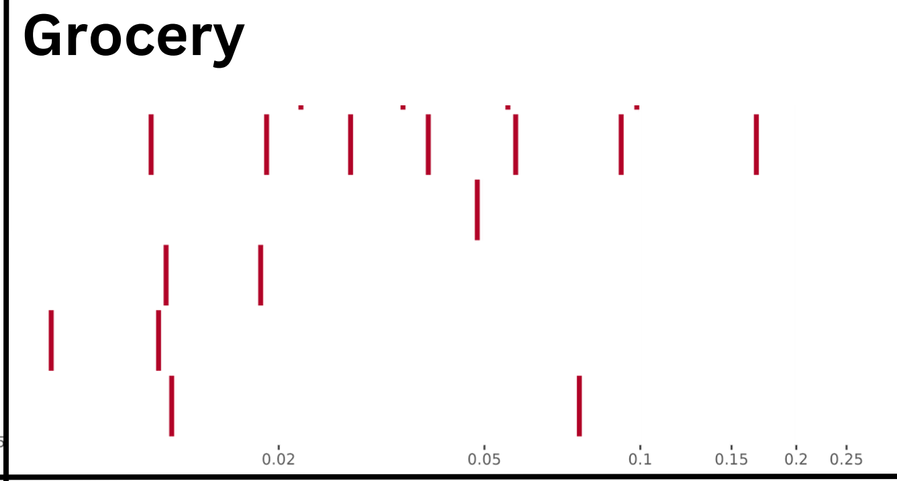
Lack of consistency

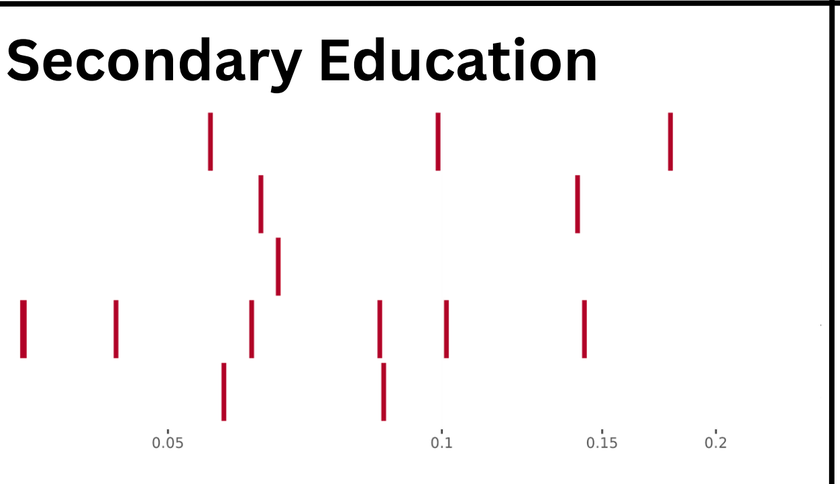


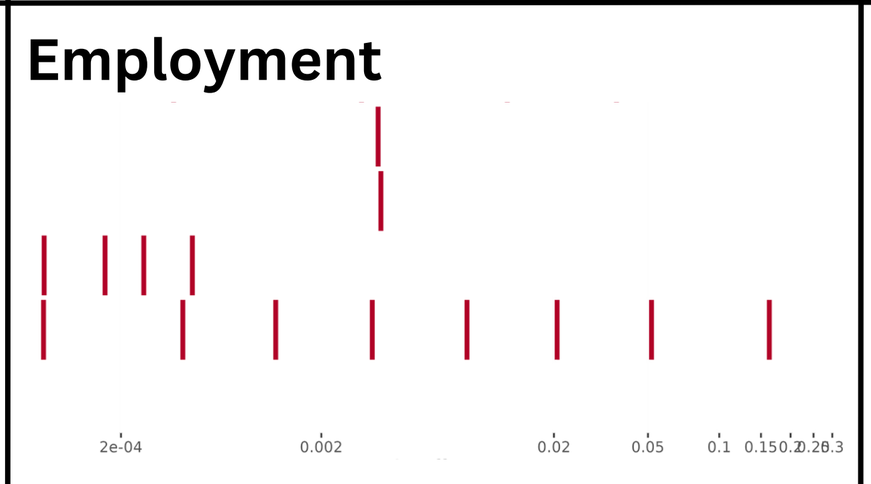
*Primary Education visualized

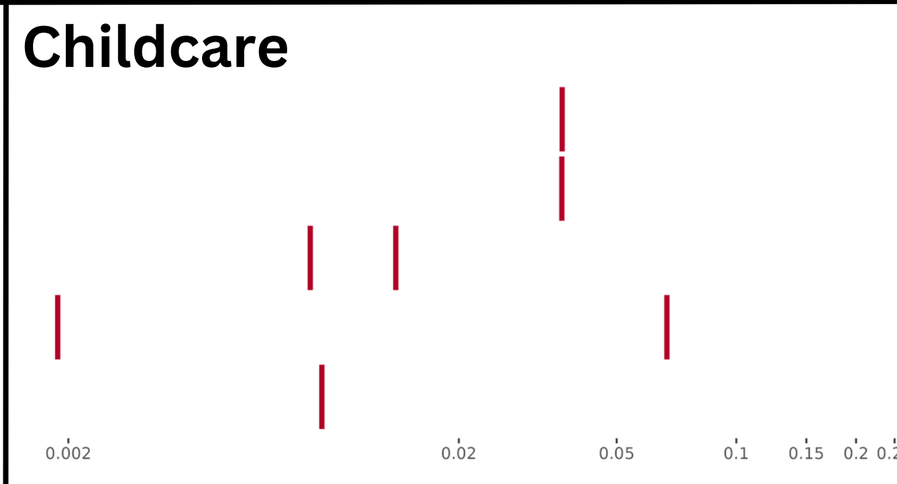

Transit
2e-04
0.002
0.02
0.05
0.1
0.15
0.2
0.25

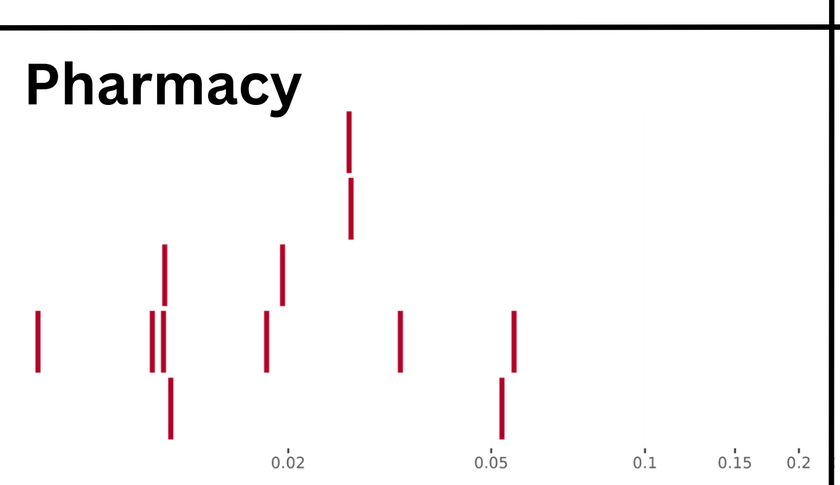

Parks
0.02
0.05
0.1
0.15
0.2
0.25

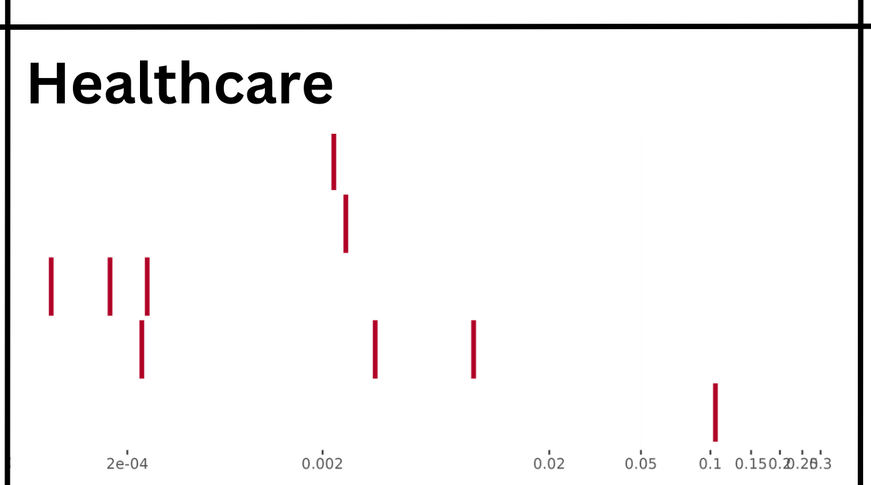

Grocery
0.02
0.05
0.1
0.15
0.2
0.25

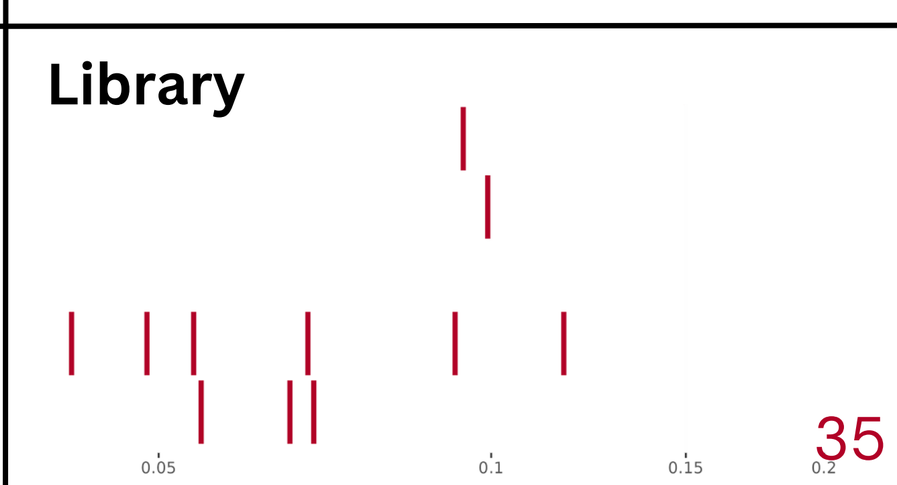

Secondary Education
0.05
0.1
0.15
0.2


Employment
2e-04
0.002
0.02
0.05
0.1
0.15
0.2
0.25
0.3


Childcare
0.002
0.02
0.05
0.1
0.15
0.2


Pharmacy
0.02
0.05
0.1
0.15
0.2


Healthcare
2e-04
0.002
0.02
0.05
0.1
0.15
0.2
0.25
0.3


Library
0.05
0.1
0.15
0.2

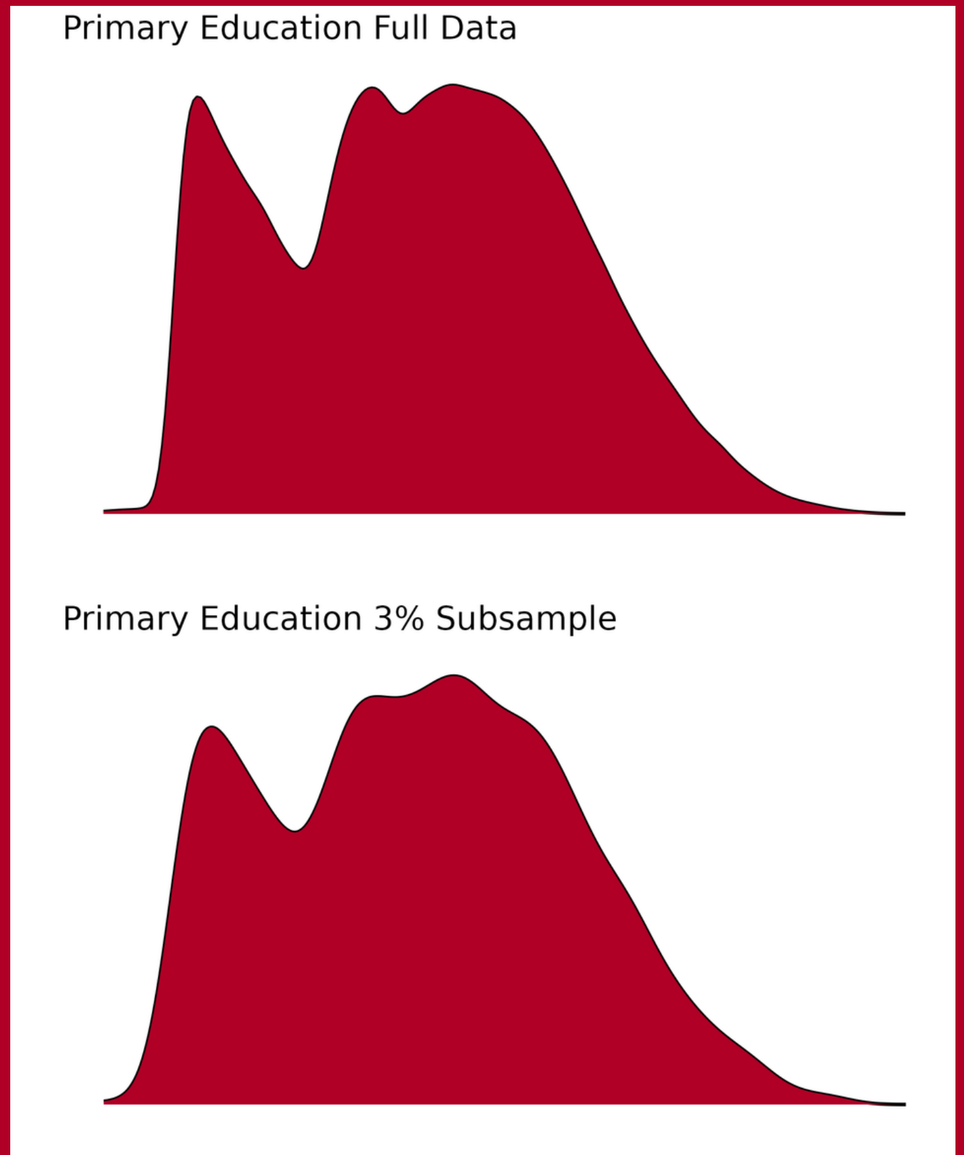
ANALYSIS

- MATHEMATICAL INDICES ARE INSUFFICIENT FOR SELECTING ONE SET OF CUTOFFS
- OPTIMAL CUTOFF SELECTION REQUIRES DOMAIN EXPERTISE

LIMITATIONS

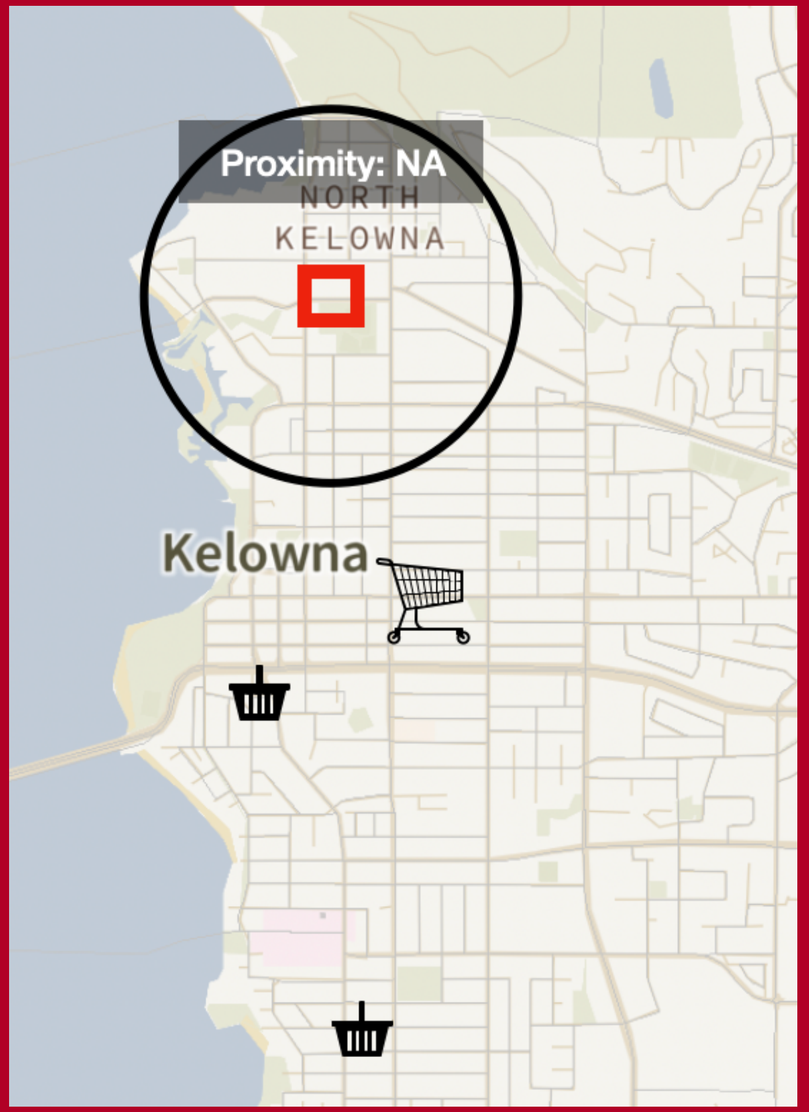
METHODOLOGY CONSTRAINTS

- 3% subsampling to prevent memory exhaustion



METHODOLOGY CONSTRAINTS

- 3% subsampling to prevent memory exhaustion
- Data limitation: distance threshold
  - ex) exclusion of low access DBs below the threshold



CONCLUSION OF PROJECT

PROJECT SUMMARY

Goal

Craft intuitive amenity categories from continuous measures, Clustering pipeline

Techniques

EDA, data transformations, clustering techniques, visualizations

Insightful Findings

PMD's low cluster tendency, Exploration of a variety of clustering algorithms

FUTURE RESEARCH DIRECTIONS

Explore multivariate clustering

Consider sub-clustering or soft assignments

Test additional clustering algorithms

Region specific clustering

CAPSTONE INSIGHTS

- project management
- teamwork and communication
- data wrangling, data visualizations, professional report writing
- clustering

ACKNOWLEDGEMENTS

Thanks to:

- Jerome, Statistics Canada industry advisor
- Firas, Irene our project instructors
- Jesse, our TA
- Jeff, consultant

REFERENCES

Alasia, A., Bédard, F., Bélanger, J., Guimond, E., & Penney, C. (2017). *Measuring remoteness and accessibility: A set of indices for Canadian communities.*Reports on Special Business Projects, Statistics Canada.https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2017002-eng.htm.

Alasia, A., Newstead, N., Kuchar, J., & Radulescu, M. (2021, February 15). *Measuring Proximity to Services and Amenities: An Experimental Set of Indicators for Neighbourhoods and Localities*. Reports on Special Business Projects, Statistics Canada. Retrieved May 4, 2023, from https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2020001-eng.htm

Statistics Canada. (2021). Dictionary, Census of Population, 2021
Dissemination block (DB). https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/definition-eng.cfm

Statistics Canada (2020a). *Proximity Measures Data Viewer*. https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2020011-eng.htm

Kassambara A, Mundt F (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7, https://CRAN.R-project.org/package=
factoextra.

Wickham, H., Averick, M., Bryan, J., et. Al, (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

Awknowledgement: These slides were created with Canva

THANKS FOR LISTENING

Questions?