A Clustering Approach to Bottom-up Theories of Subjective Well-being:
Is Life Satisfaction Differentially Predicted by Levels of Domain Satisfaction?

Wai Yu Amanda, Ng | Department of Statistical Science | University of Toronto (St George Campus)
Sofia Panasiuk | Department of Psychology | University of Toronto (St George Campus)

Introduction

- Different individuals, at various stages of their lives, place more importance on one life domain than another
- Gain more understanding of how changes in well-being vary by hidden clusters in the population
- Here, we will discuss various clustering method used and comment on the barriers we face

The dataset: Gallup-Sharecare Well-Being Index, 2014-2017

- Sample size: 529,237 (after data cleaning)
- Demographic variables
  - Age
  - Health
  - Marital status
  - Race
  - Education
- Life domains variables
  - Purpose
  - Community
  - Physical
  - Financial
  - Social well-being
- Subjective well-being measures
  - life satisfaction
  - positive affect
  - negative affect

Research questions

Are individuals readily clustered into groups according to **life domains**?

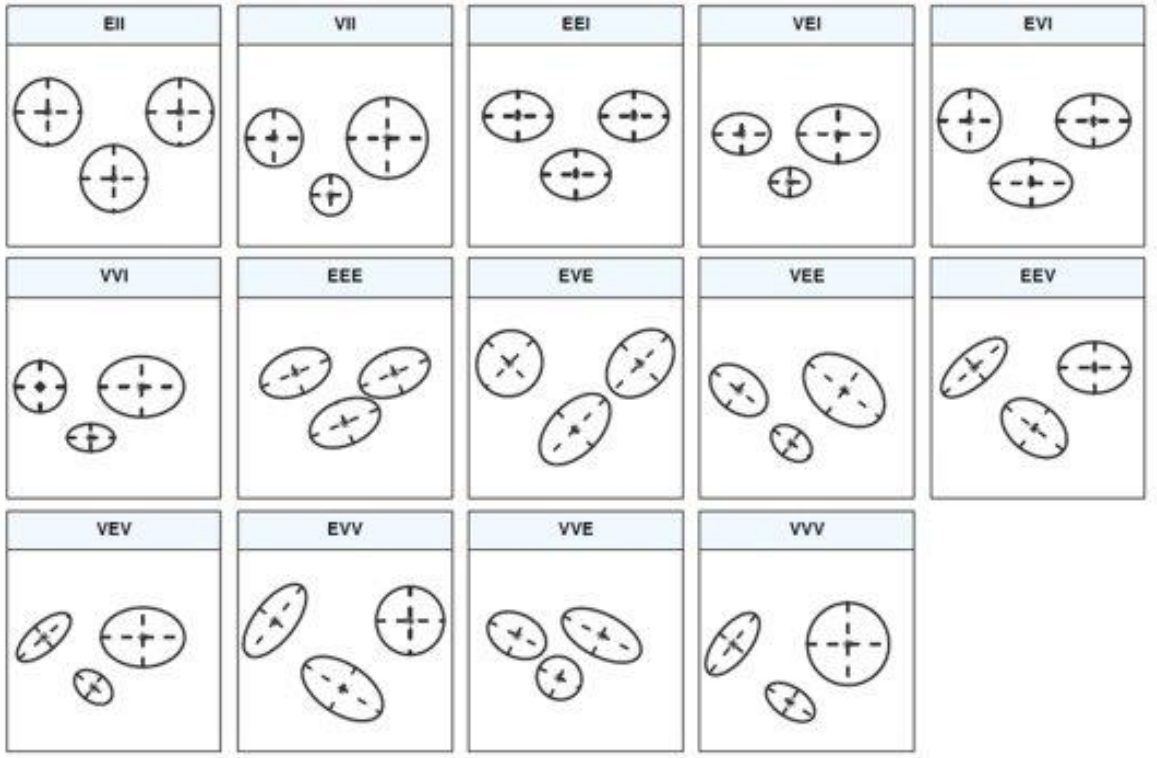
How many clusters exist and what are their qualities?

Do clusters differ from one another on **demographical variables**?

What is the **overall subjective well-being** of individuals in each of these clusters?
Do clusters differ from one another on Subjective well-being?

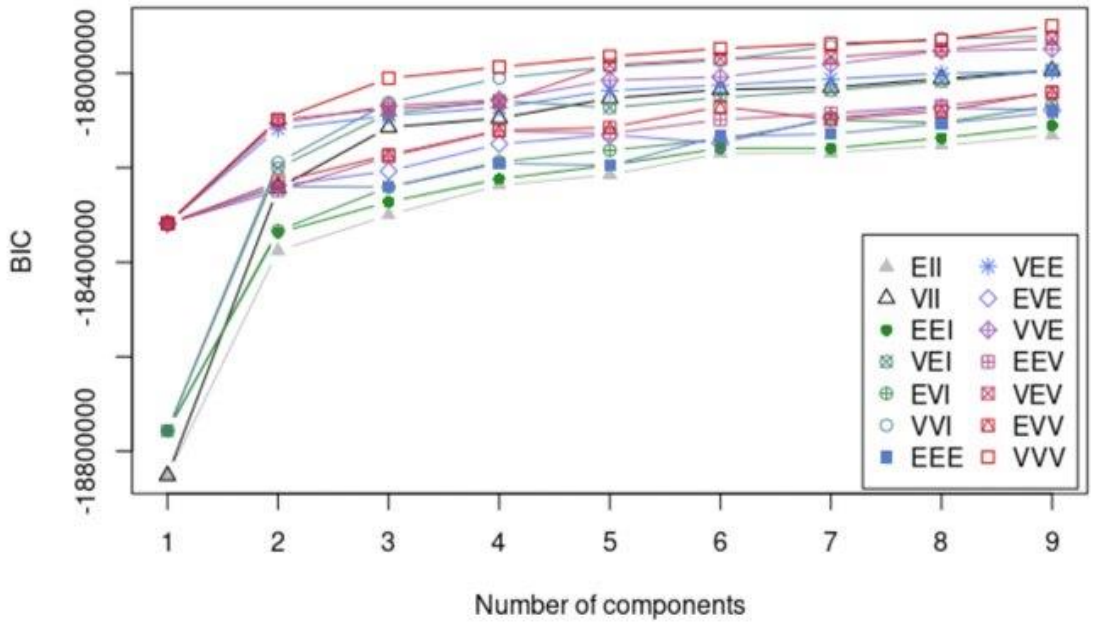
Model-based clustering using Finite Gaussian Mixture Model

- High statistical power
- Does not require us to specify the number of clusters prior to analysis
- Different models are built by varying the distribution, volume, shape and orientation.



Model-based clustering using Finite Gaussian Mixture Model

- **Expectation–maximization algorithm**
- **Bayesian information criterion (BIC)**
- Higher BIC score = better model
- Gives desired optimal number of clusters.



BIC value trend tends to increase as number of components increases!

Problems encountered and thoughts

- *BIC gave us **"finitely many clusters"***
  - Strict criterion
  - High accuracy in the parameters
  - Overlapping between clusters
  - NOT reasonable to **over-cluster data points** given limited time

- *Problem with Model-based clustering*
  - Soft clustering method: clusters are allowed to overlap with each other
  - A datapoint (participant) is allowed to fall into multiple clusters
  - May not be able to clearly identify which group does a participant is classified
  - Datapoint-cluster tracking is an important element of this research.

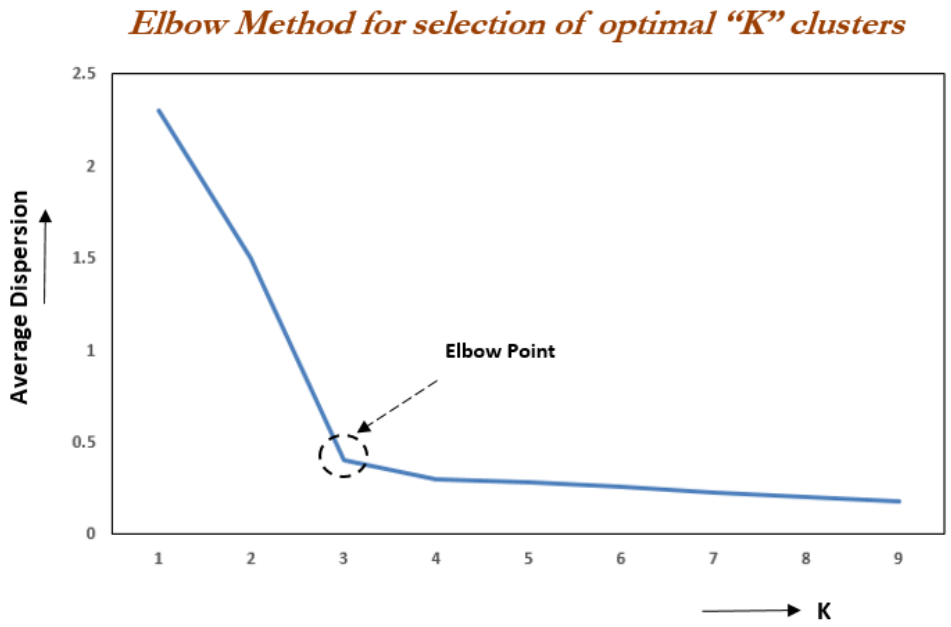
	PURPOSE	COMMUNITY	PHYSICAL	FINANCIAL	SOCIAL
PURPOSE	1.00	0.54	0.52	0.49	0.61
COMMUNITY	0.54	1.00	0.36	0.42	0.45
PHYSICAL	0.52	0.36	1.00	0.42	0.44
FINANCIAL	0.49	0.42	0.42	1.00	0.44
SOCIAL	0.61	0.45	0.44	0.44	1.00

Dimensionality reduction

- Pairwise comparison analysis
- All life domains correlates with each other
- PURPOSE is the most significant
- Caret package in R: remove redundant dimensions because domains are highly correlated with each other

K-means clustering

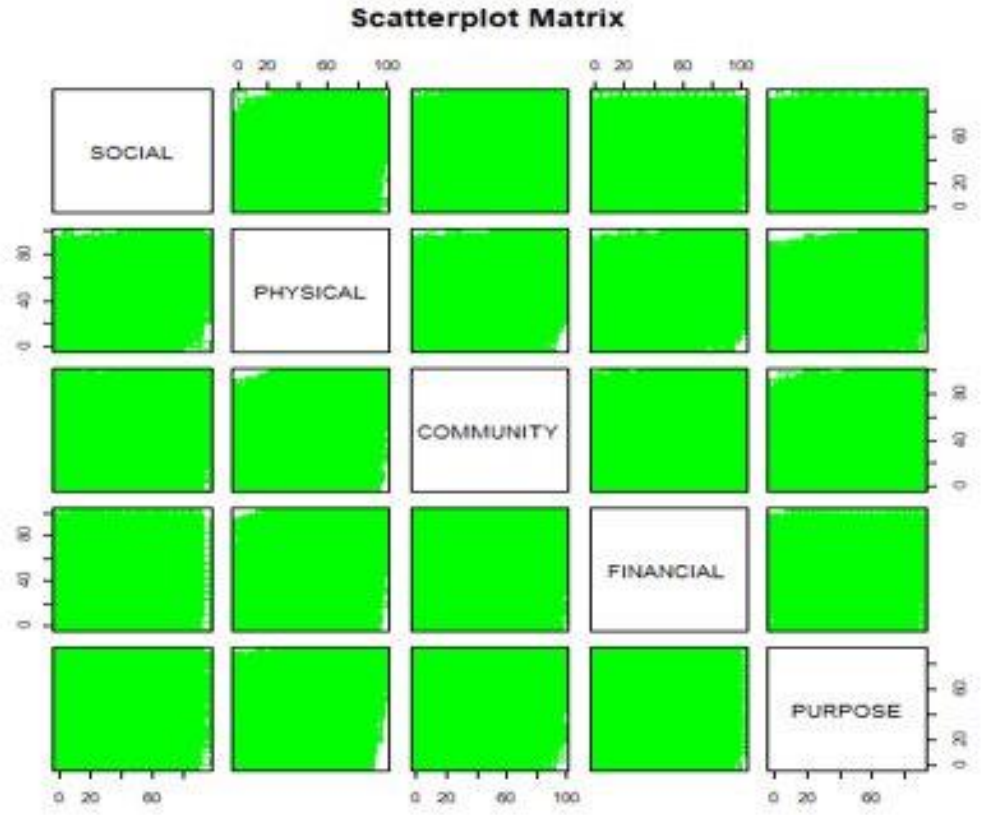
- Pre-specify the number of clusters
- Elbow method to determine the optimal cluster number
- The optimal cluster number (elbow point) is 2.



**Demonstration of elbow-method (not extracted from out study)*

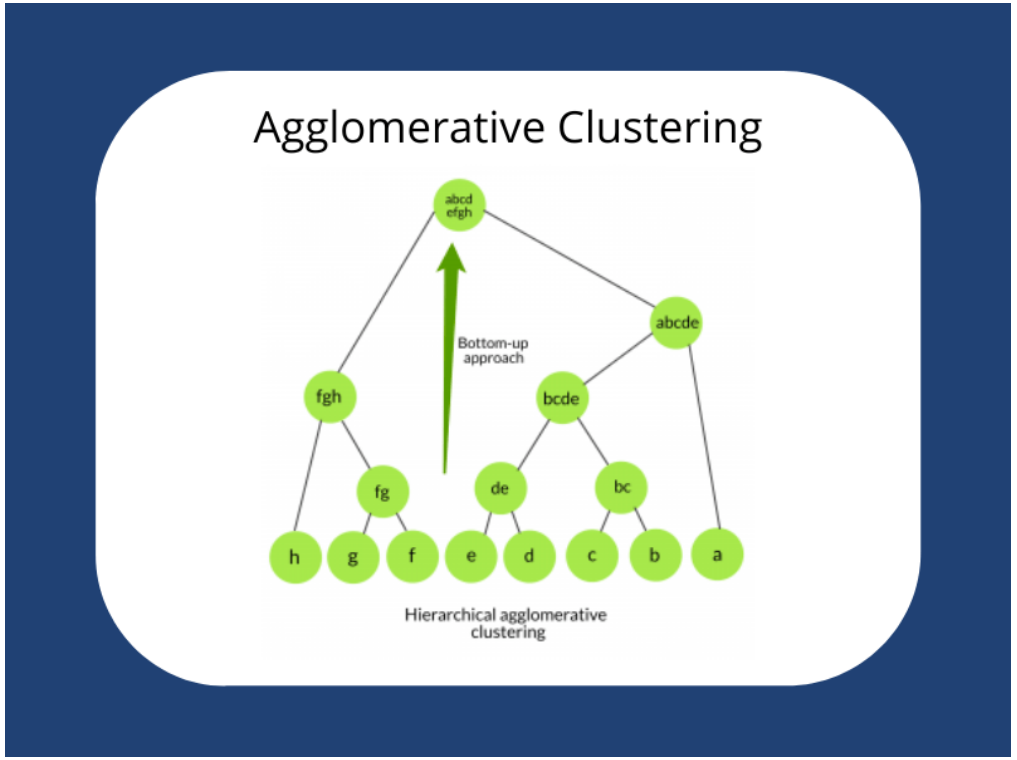
Problems encountered and thoughts

- The algorithm divided the data into 2 groups even if there are no obvious clusters.
- In the pairwise plots of each well-being domain dimension, there are **no any cluster popping out visually**
- This does not mean there are no clusters
- Other clustering methods may be better suited for us than k-means hard clustering



Agglomerative Hierarchical Clustering

- The algorithm starts by treating each object as a singleton cluster.
- Pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects
- higher the gap statistics = better number of clusters



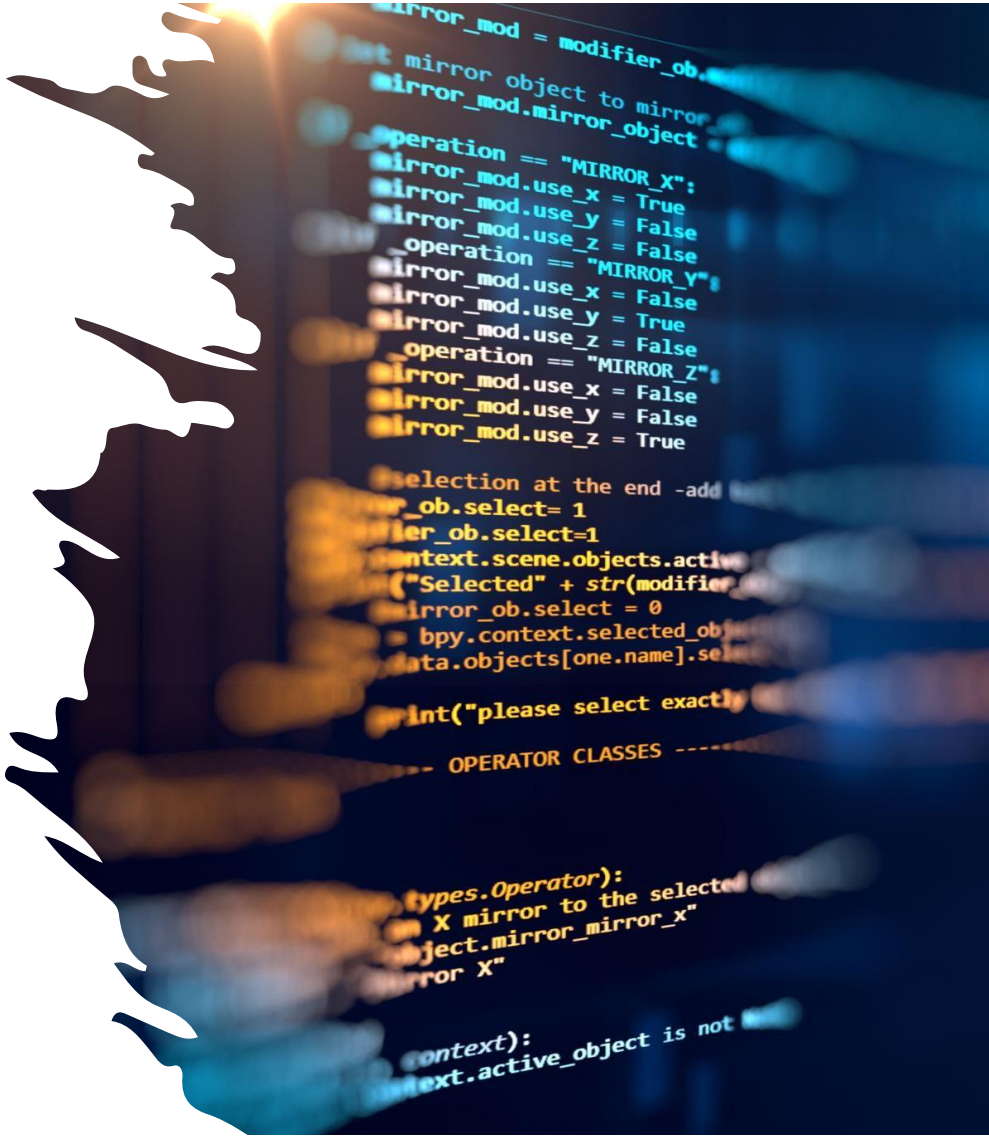
Problems encountered and solutions

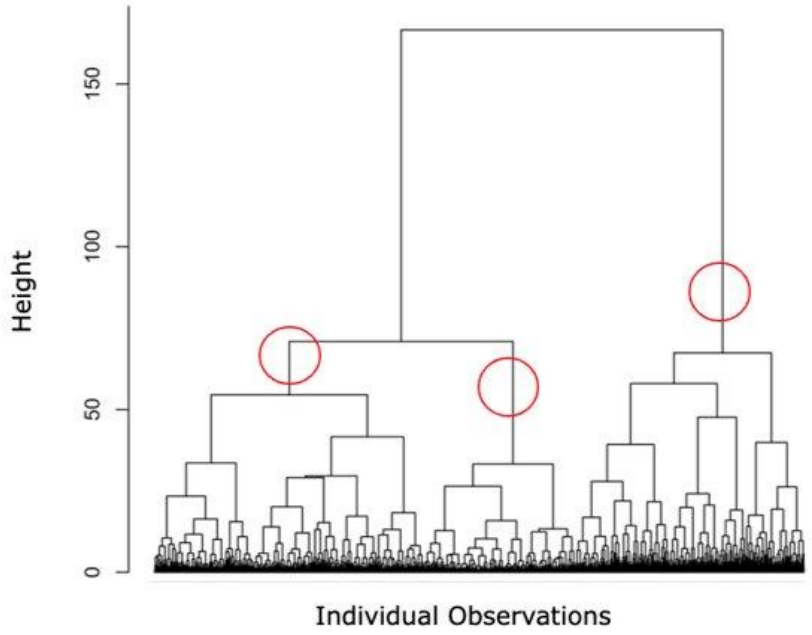
Problem:

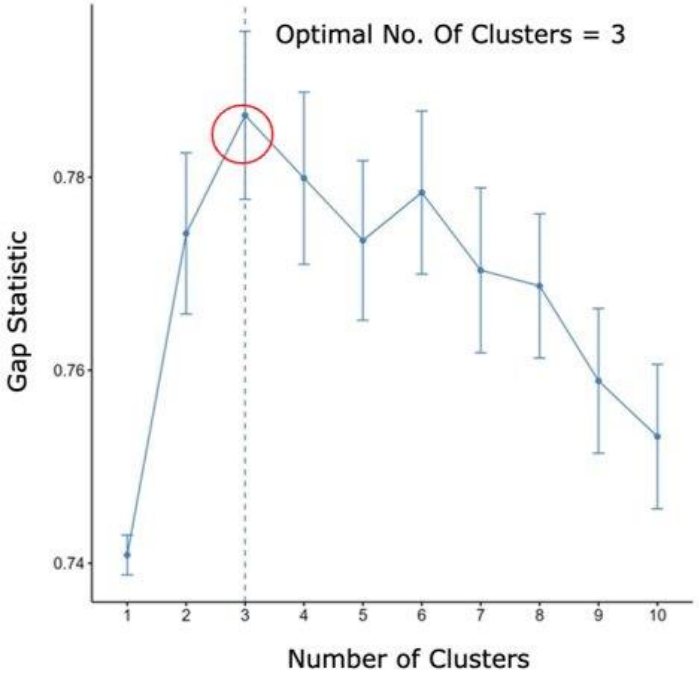
- Very computationally expensive and slow

Solution:

- Sample datapoints from the original dataset and send it to the Supercomputer to run the codes.
- Separate the original dataset into “training dataset” and “testing dataset”
- In “training dataset”, randomly select 10000 datapoints to run the algorithm



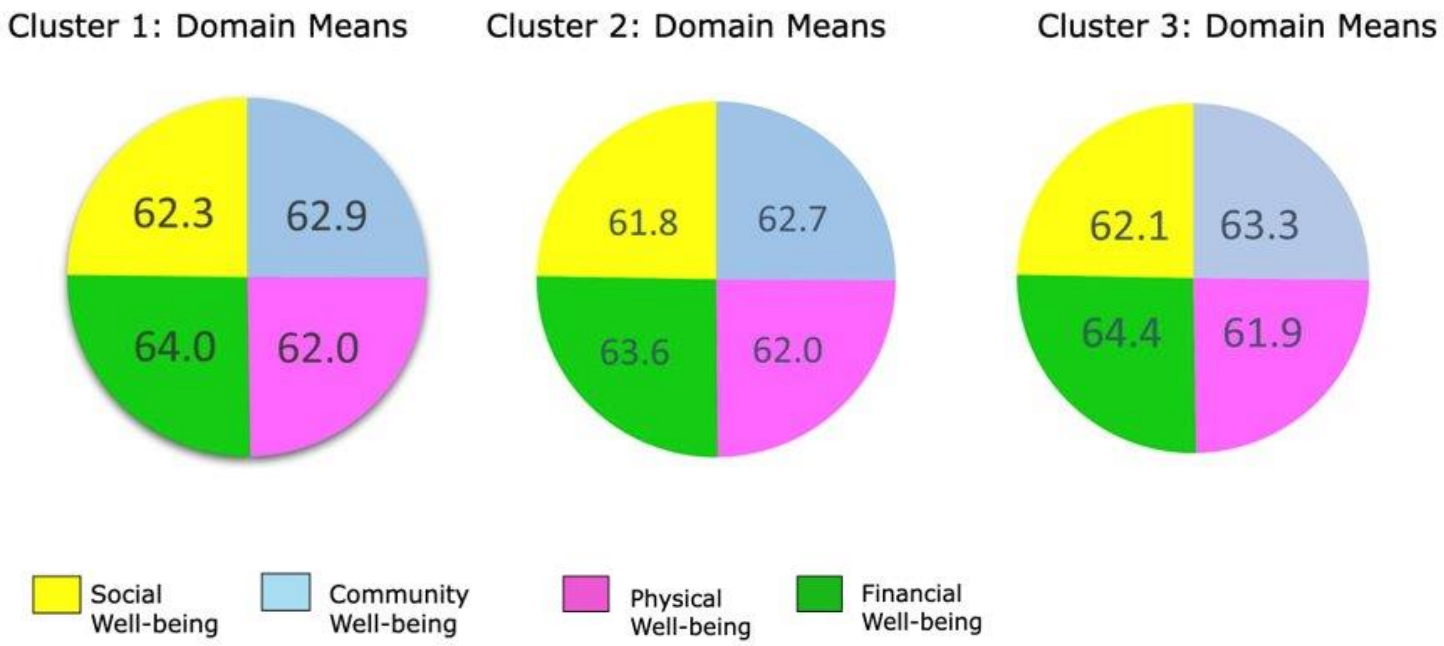




Agglomerative Hierarchical Clustering

- The algorithm gave us **3** as the optimal number of clusters
- We confirmed it by using “for-loop” to randomly reselect 10000 datapoints from the “training dataset” and run the algorithm on the subsets.
- We also tested the result with the ”testing dataset”.

Clusters characteristics analysis



*Domain well-being scores range from 0 to a maximum score of 100

Continuous variables

- All life domains (community, physical, financial and social well-being)
- Age
- All Subjective well-being measures (life satisfaction, positive affect and negative affect)

*T-test: CANNOT reject the null hypothesis that there is **no significant differences** in the variables among the clusters identified.*

Clusters characteristics analysis

Categorical variables

Health	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Excellent	20.0	20.9	18.2
Good	31.3	31.0	32.6
Marginal	13.6	14.0	14.0
Neutral	30.2	29.2	30.6
Poor	4.8	4.9	4.5
Not Available	0.1	0.0	0.2

Race	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Asian	2.4	2.2	2.3
Black	9.5	9.6	8.3
Hispanic	9.1	8.4	9.6
Other	1.5	1.4	1.6
White	74.0	75.1	75.3
Not Available	3.5	3.2	3.0

Marital Status	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Divorced	10.7	10.4	10.7
Married	51.7	51.3	51.6
Separated	1.7	2.3	1.6
Single	19.2	19.7	19.0
Widowed	10.1	11.0	10.4
Not available	6.6	5.4	6.7

Education	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Less than high school	6.0	5.8	6.0
High School	23.5	23.4	23.0
Some college	24.3	25.5	24.9
College graduate	21.6	21.9	20.9
Post-grad	18.6	18.3	19.3
Technical School	4.0	3.7	4.2
Not available	2.0	1.5	1.6

Similarly, there is **no significant differences** in the variables among the clusters identified

Conclusion

Successfully identified 3 sub-populations using the life domain variables.

The clusters do not differ significantly from one another on life domains, demographical variables and subjective well-being.

Failed to identify the clusters unique characteristics.

The problem of identifying sub-populations with higher/lower well-being is harder than we initially thought

Discussion

Null effect in cluster unique characteristics:

- Unsuitable clustering method
- No cluster exist
- The variables we tested may not describe the clusters well. Perhaps, there are other undiscovered variables that can be used to define our clusters

People **DO NOT neatly fit** into well-being domain clusters

Well-being exists on a continuous scale