Predicting Life Satisfaction from Life Domain Satisfaction: A Clustering Approach to Bottom-up Theories of Subjective Well-being

Amanda Ng[1], Sofia Panasiuk[2], Felix Cheung[2]
[1]Department of Statistical Science, University of Toronto, Canada
[2]Population Well-being Lab, Department of Psychology, University of Toronto, Canada

Introduction

- Well-being is a positive state experienced by individuals and societies and is a resource for daily life which is determined by social, economic and environmental conditions. It encompasses quality of life and the ability of people and societies to contribute to the world with a sense of meaning and purpose. (WHO, 2021)
- Aims to generate a set of well-being profiles according to different combinations of well-being domains from the Gallup-Sharecare Well-being Index: Purpose, Community, Physical, Financial and Social well-being.
- If individuals naturally cluster into well-being domain groups, it is easier to locate subpopulations with low well-being and identify them as potential recipients of additional resources to improve their well-being.
- Discusses various clustering methods we used and comment on the features and limitations of each method.
- Pre-registered on OSF: https://osf.io/fhprj

Research Questions

1. Are individuals readily clustered into groups using a combination of well-being domains?
2. What are the clusters characteristics in terms of subjective well-being (SWB) and demographic variables?

Data

- **Gallup-Sharecare Well-Being Index, 2014-2017**
- Sample size: 529,237 (after data cleaning)
- Collected from 500 American adults daily, 350 days annually

Clustering Variables

- **Well-being Domains**: Purpose, Community, Physical, Financial and Social well-being
  - ➢ Measured on a continuous scale from 0 to 100.
  - ➢ Each well-being domain is calculated by averaging answers from multiple related Likert-scale questions in the survey.
  - ➢ Financial well-being domain question example: *"Have there been times in the past twelve months when you did not have enough money to buy food that you or your family needed?"*

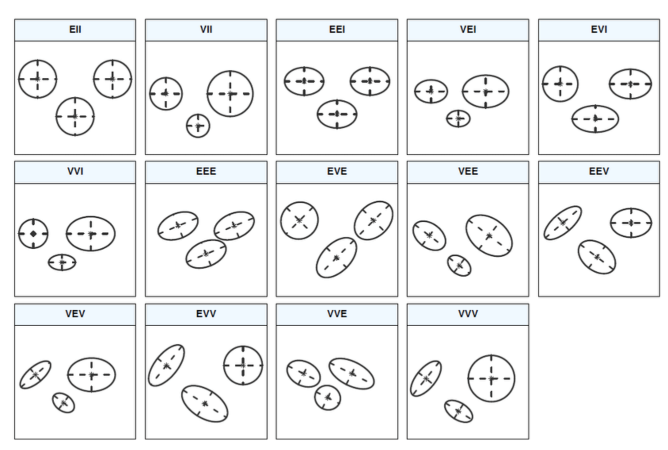
Outcome Variables

- **Demographic Variables**: age, health, marital status, race, education
  - ➢ All except age are categorical variables
- **Subjective Well-being (SWB)**: life satisfaction (LS), positive affect (PA) and negative affect (NA)
  - ➢ LS: Ordinal scale from 0 to 10

"Please imagine a ladder with steps numbered from zero at the bottom to ten at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"

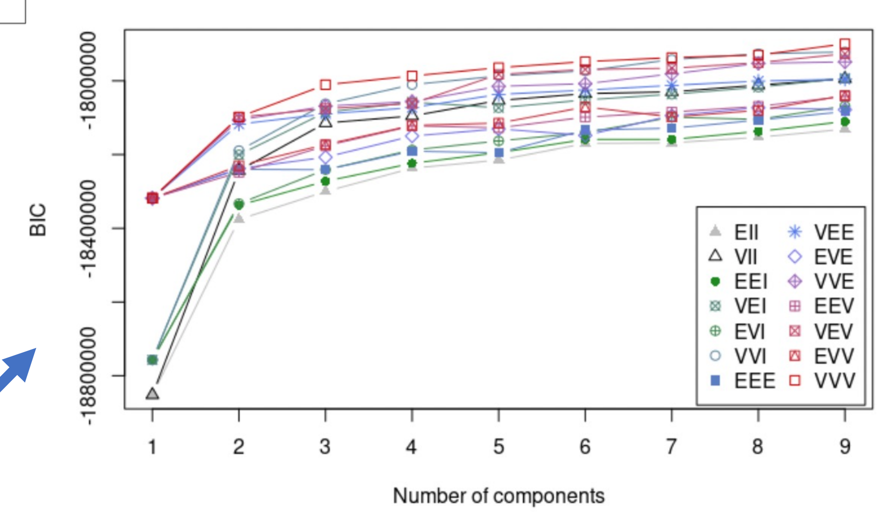
- ➢ PA: Average responses corresponding to questions asking participants whether they smiled or laughed, experienced enjoyment or experienced happiness yesterday as binary Yes/No variables.
- ➢ NA: Average responses corresponding to questions asking participants whether they experienced worry, sadness, stress or anger yesterday as binary Yes/No variables.

Overview of Analytical framework
Approach 1: Finite Gaussian Mixture Model (GMM)



1. Different models are built by **varying the distribution, volume, shape and orientation** (McLachlan et al., 2019).

2. Optimize the parameters of GMM (means, variances, weighting variables) using **expectation–maximization algorithm**.



3. Choose the model with highest **Bayesian information criterion (BIC)** value.

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

Methodological Discussion:

- GMM **does not require specification of the number of clusters** prior to analysis (McLachlan et al., 2019).
- BIC gave us **"finitely many clusters"** : a strict criterion which results in high accuracy in the parameters (Biernacki et al., 2000).
- GMM clustering is a **soft clustering method** where clusters are allowed to overlap with each other. So, a **datapoint is allowed to fall into multiple clusters**. However, we require an injective relationship from the datapoints to the clusters for tracking purposes.

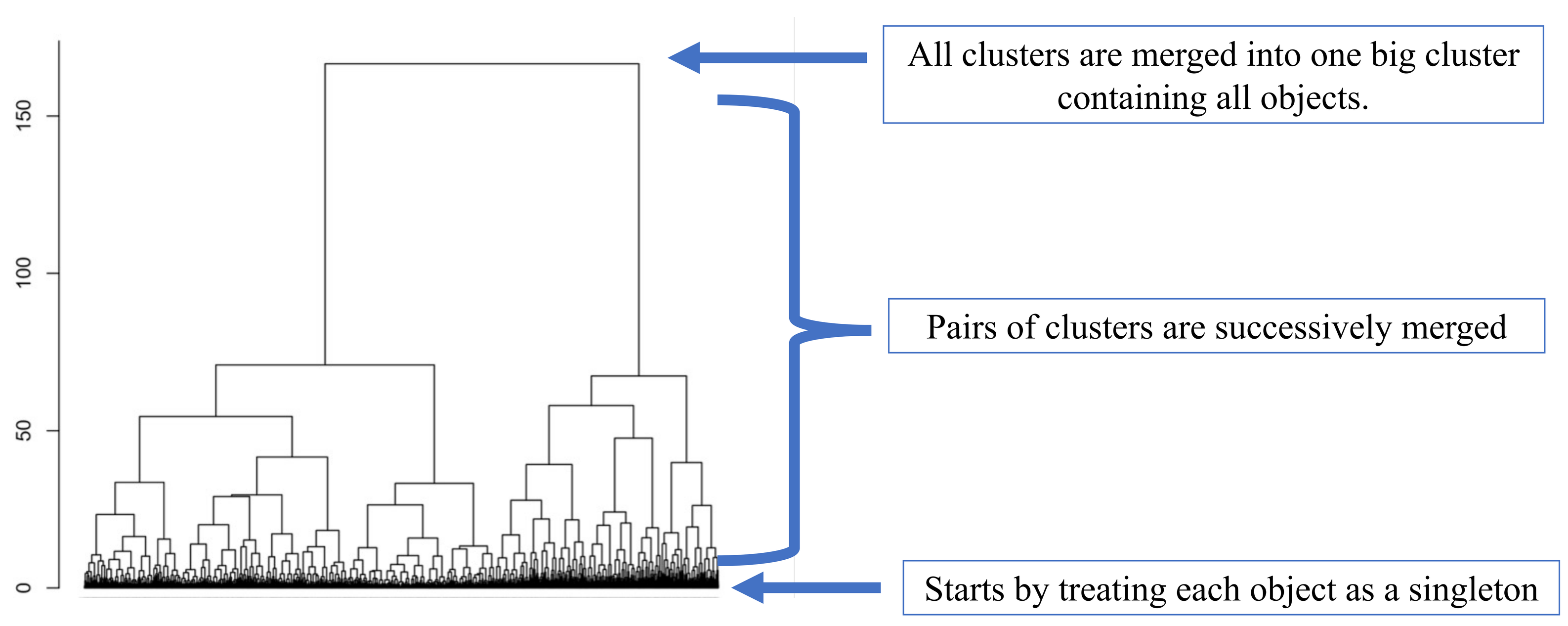
Approach 2: K-means clustering

- Conducted **pairwise comparison analysis (PCA)** on all well-being domains.
- Removed "Purpose" by **unsupervised feature selection method** because it has the highest interdomain correlations with the other domains.
- Using the elbow method, we pre-specified the number of clusters as 2.

Methodological discussion:

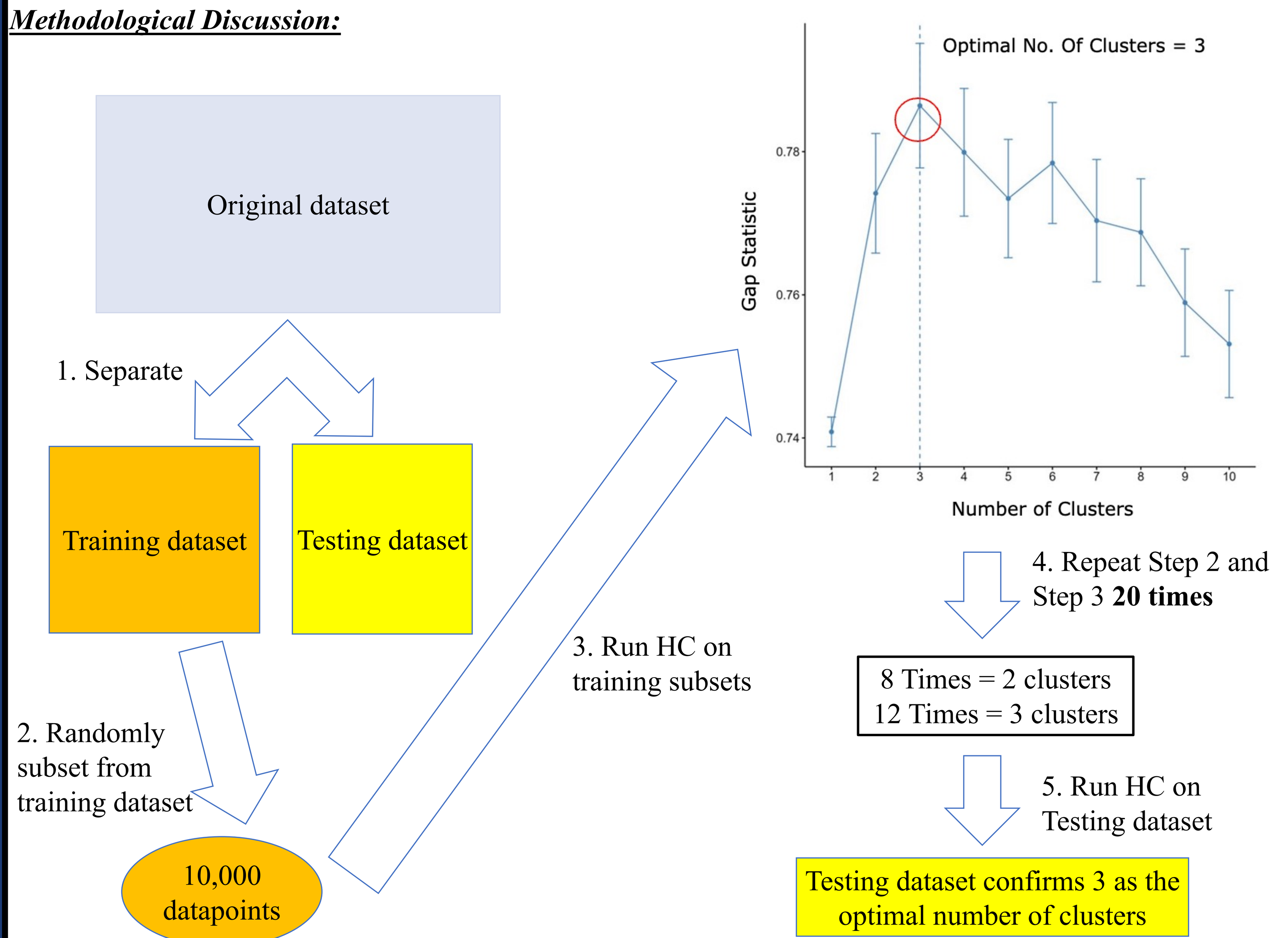
- K-means clustering relies on **linear distance** to cluster datapoints. So, we should (in theory) be able to see clusters visually if they do exist.
- However, in the PCA plots of each well-being domain dimension, there were **no cluster appearing visually**.

Approach 3: Agglomerative Hierarchical Clustering (HC)



- The **Gap statistic** compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data (Tibshirani, R.,2001).
- The optimal clusters is determined by choosing the value that maximize the gap statistic.

Methodological Discussion:



Results

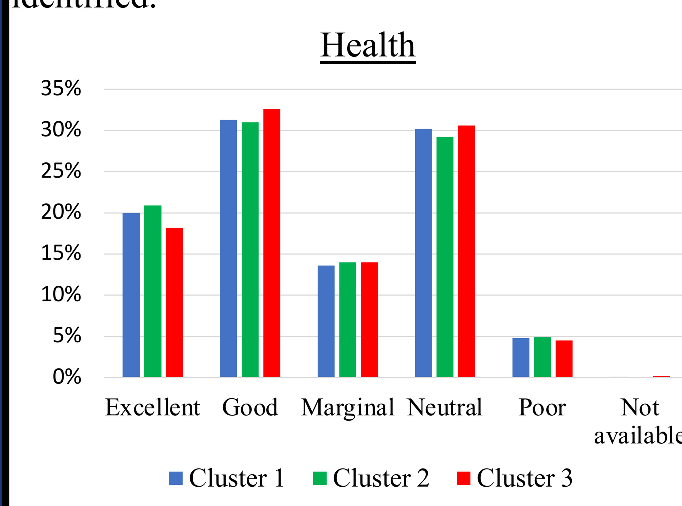
There is **no significant differences** between clusters on any outcome variable.

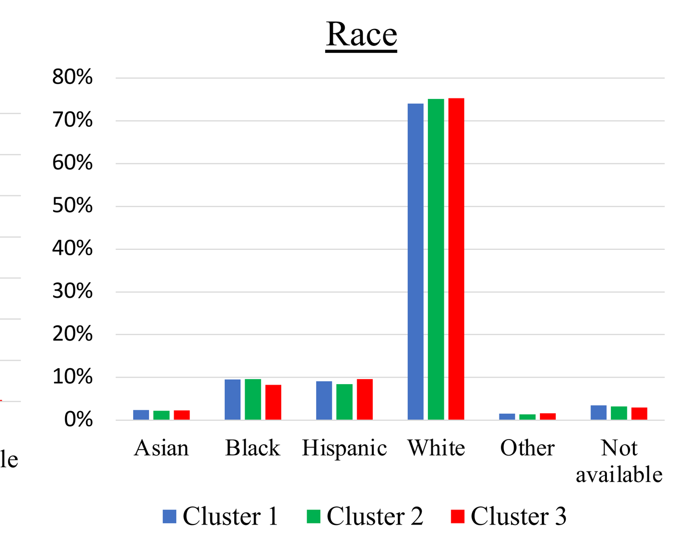
Continuous Variables:

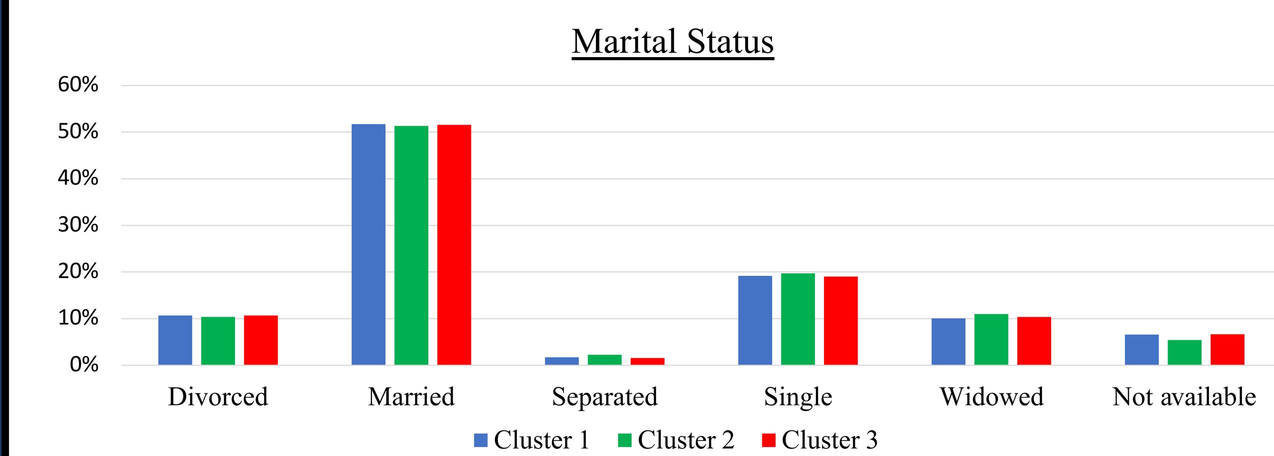
- Age and SWB (LS, PA and NA)
- We ran T-tests to confirm that we cannot reject the null hypothesis that there is **no significant differences** in the continuous variables among the clusters identified.

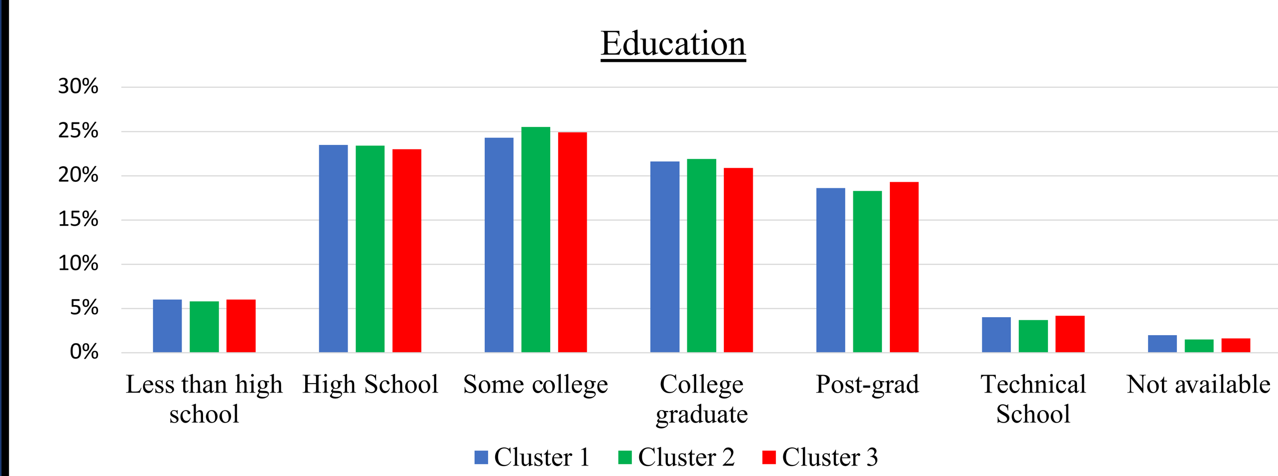
Categorical variables:

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

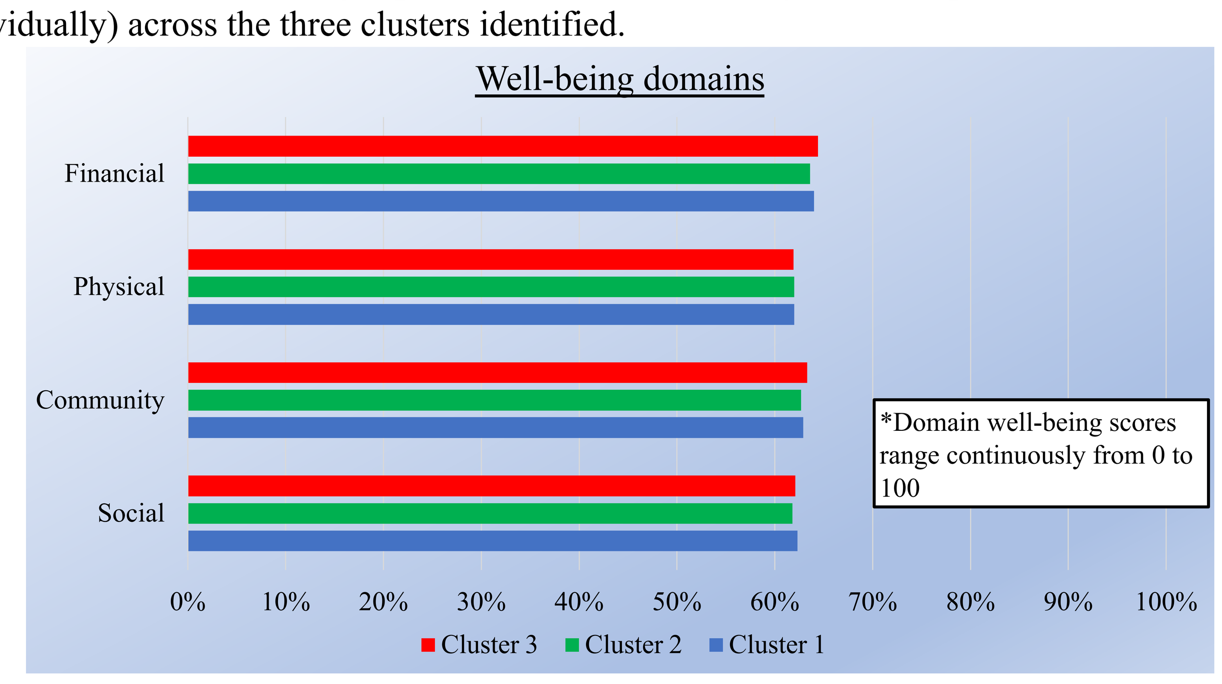








We also found **no statistically significant differences** in any well-being domain variables (individually) across the three clusters identified.



Conclusion

Theoretical implications:

- Successfully identified **3 sub-populations** using the well-being domain variables.
- However, the clusters **do not differ significantly** from one another on well-being domains, demographic variables and SWB.

1. Failed to find unique clusters
2. The problem of identifying sub-populations with higher/lower well-being is harder than we initially thought.

Null effects among clusters potentially reflects that people **do not neatly fit** into well-being clusters. It provokes thoughts on the nature of well-being and hence tends to support existing theories about **well-being existing on a continuous scale** (Franken, K., 2018).

Methodological implications:

1. Clustering Method: Hierarchical clustering

- Hard clustering
  - ➢ Each datapoint is only allowed to fall into one cluster.
  - ➢ But people in real life may carry characteristics that can fit into multiple clusters.
- Inflexible
  - ➢ Bottom-up iterative approach
  - ➢ Once a datapoint is clustered, it cannot be moved to another cluster even if it may fit better to another clusters.

2. Clustering variables

- Not all variables are conducive to clustering: Well-being domain variables may just not cluster naturally.
- There may be other undiscovered variables outside of our research scope that can be used to define our clusters.

For more information, please feel free to contact Amanda Ng by email: waiyuamanda.ng@mail.utoronto.ca or scan QR code