Poly Substance Use and it's impact on perceived Mental Health

What We Did and Why It Matters

Mental health is one of the most pressing issues of our time, yet we still don't fully understand the complex factors that influence it. We wanted to know whether poly-substance use (using multiple substances) has any impact on perceived mental health. By exploring data from nearly 12,000 individuals, we combined substance use patterns with demographic factors like age and gender to uncover new insights.

These insights matter because they can help create more effective interventions and guide policy changes, ensuring better mental health support for those who need it most.

Dataset Information

Source: Statistics Canada

Dataset: Canadian Tobacco and Nicotine Survey 2022

Type: Survey data on the Canadian population

Size: 88 columns, 12,133 rows

Features: Demographic data, smoking history, cannabis use, vaping use, alcohol use

Study Aims

Aim 1: Cluster individuals based on substance use and demographic patterns

Aim 2: Test the association between the identified clusters and perceived mental health

Methodology

1. Data Cleaning
2. Feature Engineering
3. Modelling (Unsupervised Clustering)
4. Association Testing

1. Data Cleaning

- Treated survey responses like "valid skip" and "not stated" as separate encodings

2. Feature Engineering

- Removed redundant or unrelated variables
- Kept only relevant features: alcohol, cannabis, smoking, vaping, and demographics
- Removed variables that can cause data leakage

2. Modelling (Unsupervised)

- **Models Applied**:

1. K-Modes Clustering

Optimal number of clusters selected using Elbow Method and Silhouette Score

2. Bayesian Gaussian Mixture

Optimal number of classes determined using BIC (Bayesian Information Criterion) and Silhouette Score

- **Datasets Used:**

1. Substance use pattern only
2. Substance use pattern + demographic variables

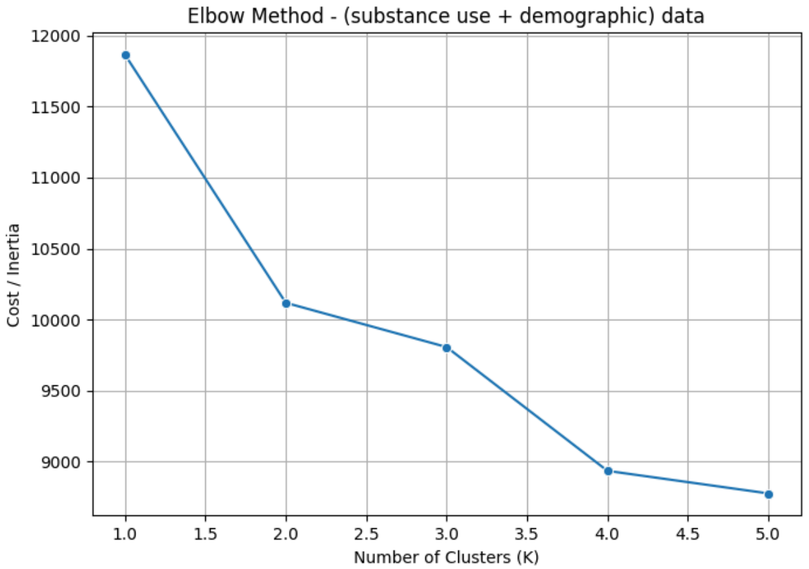
- **Cluster Visualization**:

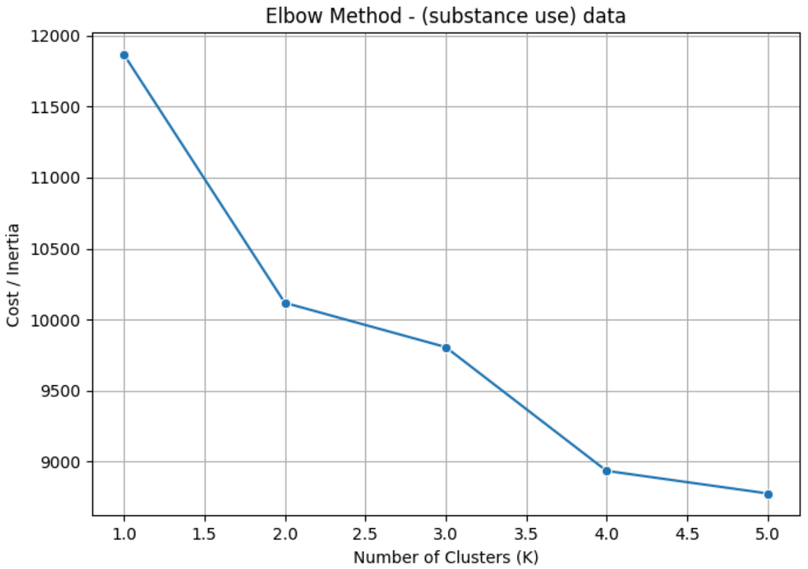
Clusters were visualized using Multiple Correspondence Analysis (MCA) for interpretability

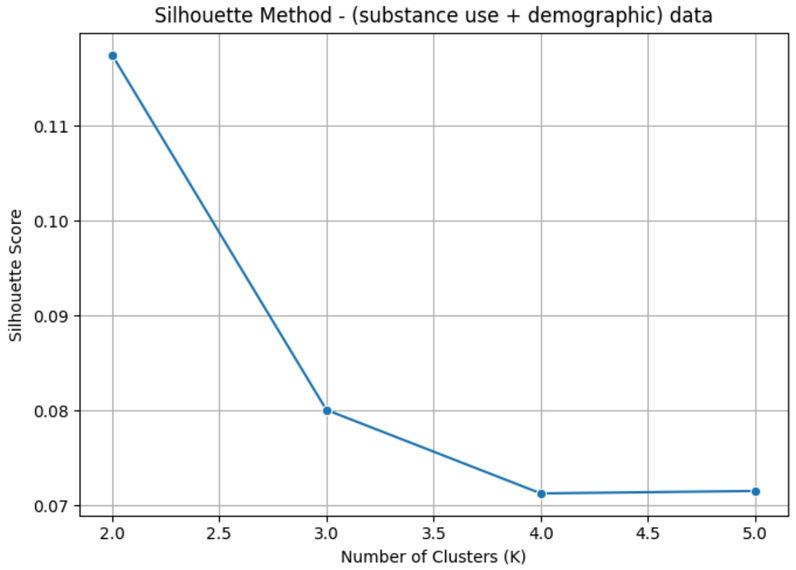
4. Association Testing

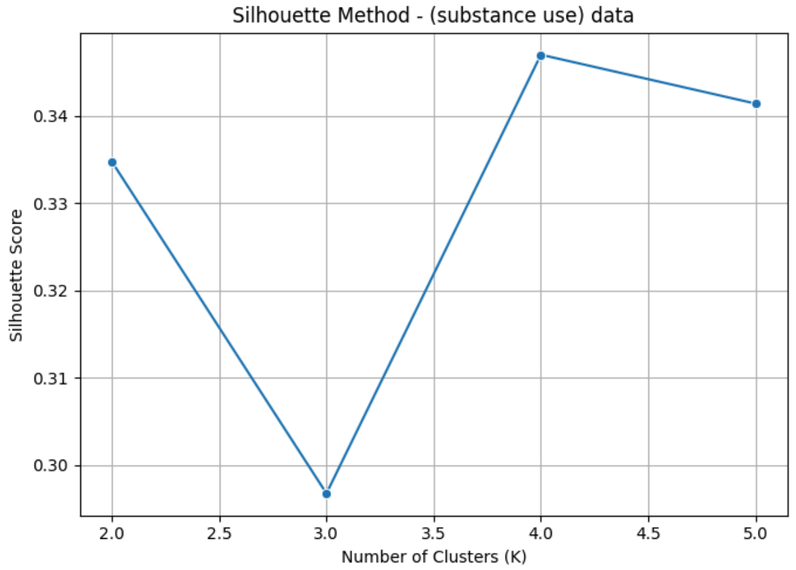
- Performed Chi-Square Test to check statistical significance between clusters and perceived mental health
- Used Cramér's V to measure the strength of the association

Kmode Clustering Results

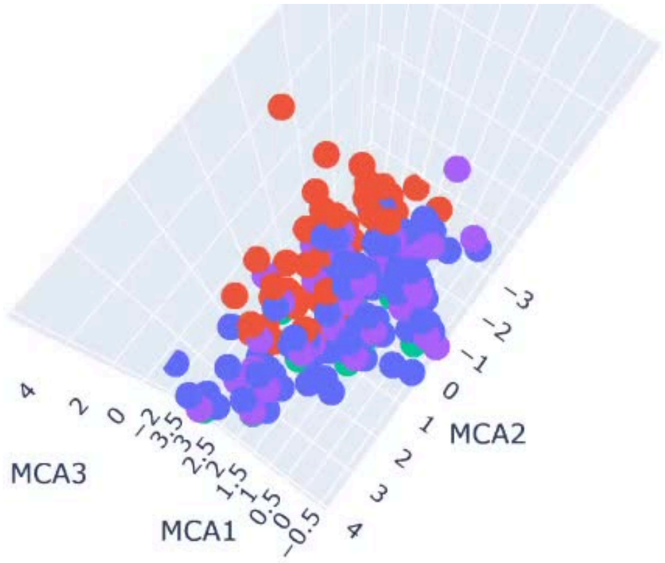

Elbow Method - (substance use + demographic) data
Cost / Inertia
Number of Clusters (K)


Elbow Method - (substance use) data
Cost / Inertia
Number of Clusters (K)

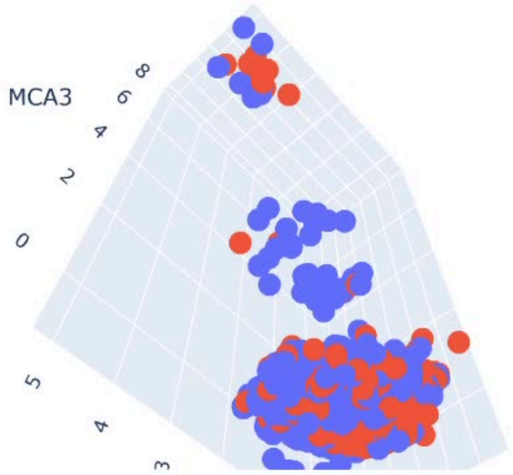

Silhouette Method - (substance use + demographic) data
Silhouette Score
Number of Clusters (K)


Silhouette Method - (substance use) data
Silhouette Score
Number of Clusters (K)

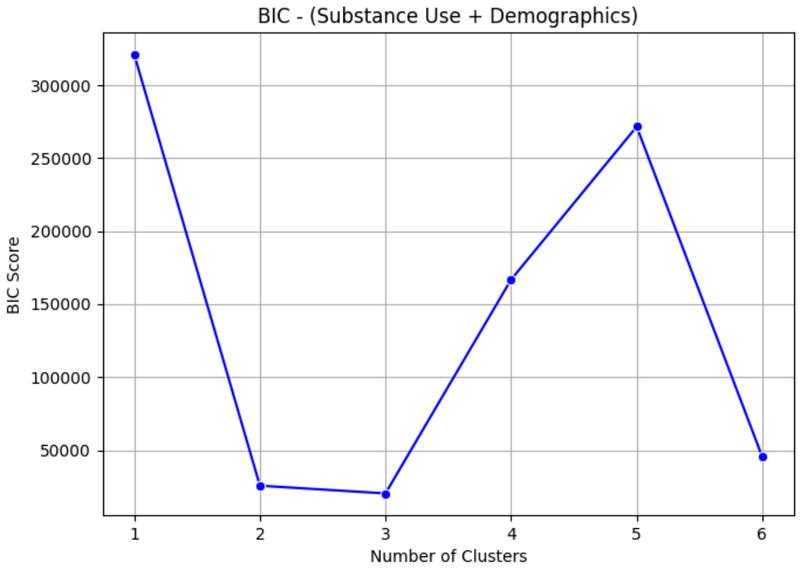
Substance use only

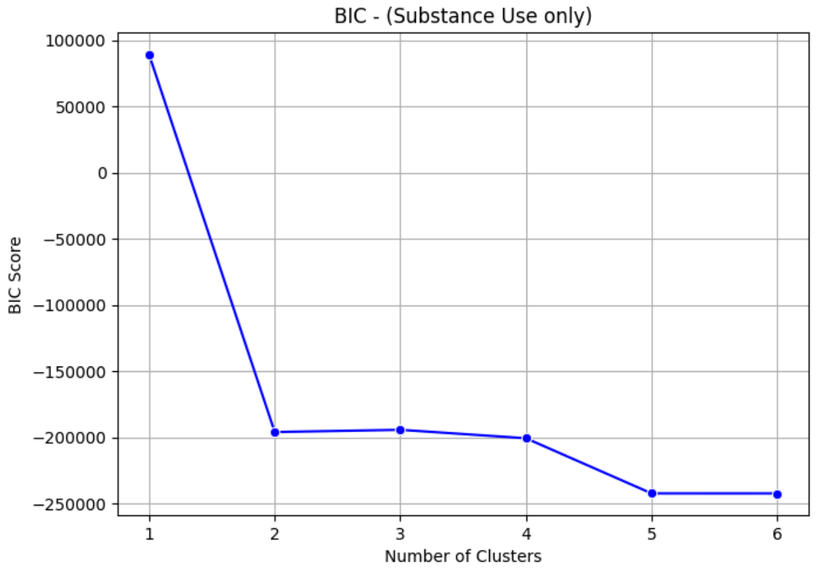


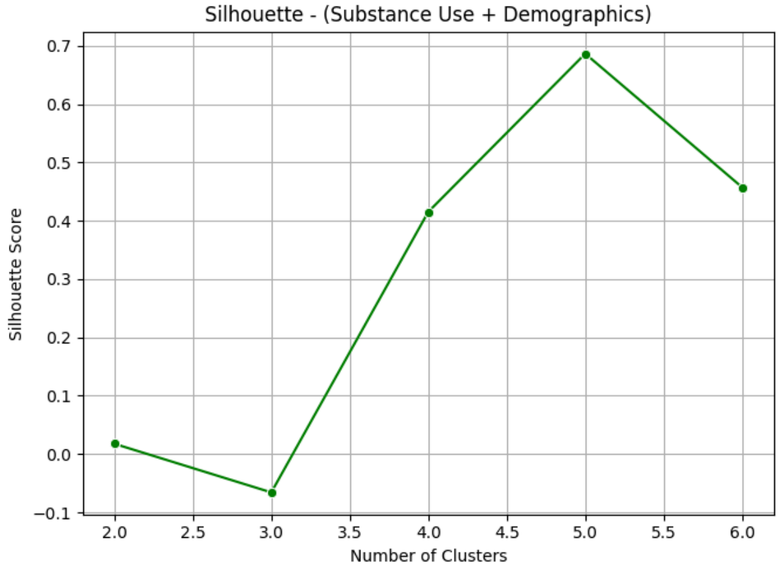
Substance use +Demographic

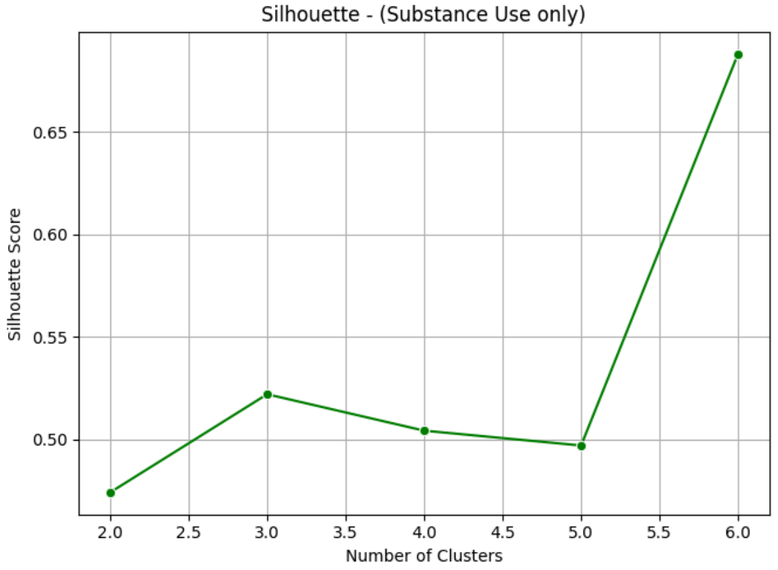


Bayesian Gaussian Mixture

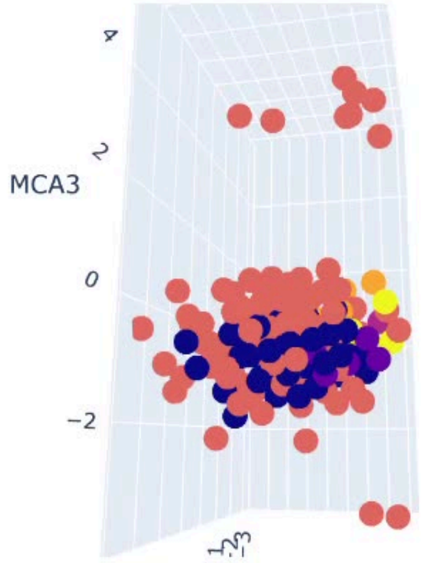

BIC - (Substance Use + Demographics)
BIC Score
Number of Clusters
300000
250000
200000
150000
100000
50000
1
2
3
4
5
6


BIC - (Substance Use only)
BIC Score
Number of Clusters
100000
50000
0
−50000
−100000
−150000
−200000
−250000
1
2
3
4
5
6

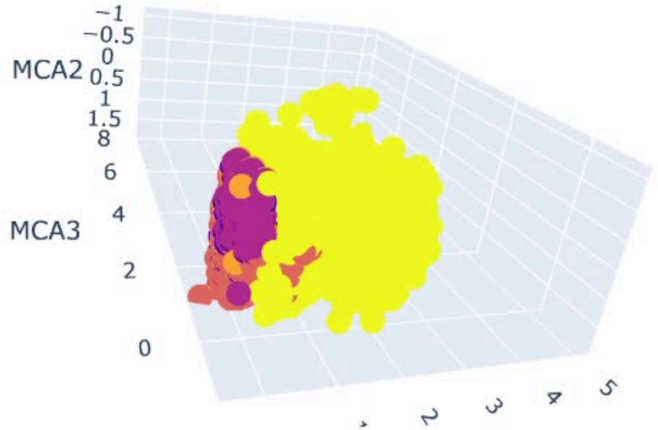

Silhouette - (Substance Use + Demographics)
Silhouette Score
Number of Clusters
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0


Silhouette - (Substance Use only)
Silhouette Score
Number of Clusters
0.65
0.60
0.55
0.50
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0

Substance use only



Substance use +Demographic



Association Testing Results

MODEL	*DATA*	*CHI SQUARE VALUE*	*CRAMER'S VALUE*
KMODE CLUSTER	*SUBSTANCE USE*	*CHI-SQUARE STATISTICS=233.2648319342576* *P-VALUE~0*	*0.0801*
KMODE CLUSTER	*SUBSTANCE USE* *+* *DEMOGRAPHIC*	*CHI-SQUARE=93.80484331520209* *P-VALUE~0*	*0.0880*
BAYESIAN GAUSSIAN MIXTURE	*SUBSTANCE USE*	*CHI-SQUARE STATICS=530.4896188764197* *P-VALUE~0*	*0.104657954655493 67*
BAYESIAN GAUSSIAN MIXTURE	*SUBSTANCE USE+DEMOGRA PHIC*	*CHI SQUARE STATISTICS=288.78732904001754* *P-VALUE~0*	*0.0892*

Conclusion

- Distinct clusters were identified using K-Modes and Bayesian Gaussian Mixture across different variable sets (substance use only, and substance use + demographics).
- While the Cramér's V value is relatively low, this is common in social science and public health research where many factors influence mental health.
- Even weak associations can be important when the sample size is large and the variables are complex or multidimensional.
- Overall, the significant statistical associations ($p < 0.05$) across all models validate the relevance of clustering for understanding mental health patterns based on substance use and demographics.