PSYC 6136 Final Project

Kyra N. Farrelly

York University

PSY 6136: Categorical Data Analysis

AI Disclosure: I used Chat-GPT to help generate and de-bug code for generating plots before ultimately using them in excel.

Problem Description

Cannabis represents one of the most used psychoactive substances in Canada (Health Canada, 2024). However, most people do not use cannabis-only and report use of alcohol or other substances in addition to cannabis (Carlini et al., 2022). Individuals who use cannabis and other substances tend to have a higher risk of consequences than those who only use cannabis (Connor et al., 2013; Hochheimer et al., 2020; Yurasek et al., 2017). Furthermore, cannabis is often used simultaneously with other substances (i.e., so their intoxicating effects overlap; Lee et al., 2022). The combination of cannabis with other substances may have a synergistic effect, exacerbating the risk for negative consequences (Crummy et al., 2020; Yurasek et al., 2017). Thus, there is a need to consider groups of individuals who engage in simultaneous cannabis with other drug use.

Latent Class Analysis

Latent class analysis (LCA) is a person-centered modelling approach that can help to classify simultaneous cannabis with other drug use patterns. LCA allows one to identify distinct groups based on specified characteristics. This is done by using participant responses as categorical indicators and observing heterogeneity within the sample. LCA also assumes conditional independence, meaning that probabilities of indicator endorsement within a class are independent. There is a breadth of literature which have utilized LCA to establish groups of polysubstance use (for review see LeComte et al., 2025), with two studies considering simultaneous cannabis with other drug use (Bailey et al., 2019; Davis et al., 2019). However, the existing literature has included limited indicators of simultaneous use.

The Present Analysis

This brings me to the present analysis. To address this gap in the literature, I have completed an LCA identifying classes of simultaneous cannabis with other drugs use using data from the 2023 Canadian Substance Use Survey. This study consisted of an LCA done in Mplus with eight simultaneous cannabis use indicators, and considered frequency of cannabis use, severity of cannabis problems and medicinal versus recreational reasons for cannabis use as predictors of class membership (Farrelly & Wardell, in preparation). This analysis identified three latent classes of simultaneous cannabis with other drug use: 1) low overall simultaneous use with occasional simultaneous use of cannabis with alcohol, 2) simultaneous use of cannabis with alcohol and hallucinogens, and 3) high overall simultaneous use (i.e., simultaneous use of cannabis with alcohol, hallucinogens, illicit stimulants, and prescription drugs). The goal of the present study was to replicate the class selection portion of this analysis in R and try to develop better visualizations of classes and probabilities.

Data Analysis

Participants and Procedure

Data came from the 2023 Canadian Substance Use Survey (CSUS), a biannual national survey assessing substance use among Canadians aged 15+ (Health Canada, 2025a). The survey was administered from May to December 2023. Full sampling methods, procedures, and surveys are available from the Government of Canada (Health Canada, 2023b). The CSUS dataset consists of 36,180 total responses (response rate of 33.3%). The current study was limited to individuals who reported past year cannabis use (n = 12,347). An additional three people were dropped due to missing data on simultaneous use indicators (N = 12,344).

Simultaneous Use Indicators

Participants were asked "During the past 12 months, when you used cannabis, how often did you combine it with any of the following substances…" for each of these 10 listed substances: alcohol, prescription opioids, prescription stimulants, prescription sedatives, illicit opioids, cocaine, illicit amphetamines/methamphetamine, ecstasy or similar drugs like MDMA, hallucinogens, and dissociative psychedelics. Response options for each substance ranged from *Never* to *Always.*

For the LCA, I specified eight simultaneous use indicators. All simultaneous use questions were recoded to form categorical variables of presence of simultaneous use with cannabis or not. Due to low endorsement (< 2%) of simultaneous use with cannabis for some drug categories, certain indicators with similar drug classifications were combined (i.e., dissociative psychedelics were included with hallucinogens, amphetamine/methamphetamine was combined with cocaine to form an illicit stimulant indicator). Due to capabilities of R software, this analysis was unable to use a three-item ordinal value of simultaneous cannabis with alcohol use (0=never, 1=occasionally, 3=frequently) seen in Farrelly & Wardell (in preparation) and instead used a dichotomous categorical variable like the seven other indicators.

Data Analysis

The LCA was conducted in R studio version 4.1.1 using the poLCA package. I specified two to five class models, using 10,000 maximum iterations. The selection of the best class solution was guided using goodness of fit statistics: the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), likelihood ratio, chi-square test statistic, and entropy. I also considered the proportion of the sample included in the smallest class to ensure clinically meaningful groups and aid in model interpretability.

Results

A four-class solution was selected as the best fitting model (see fit statistics in Table 1). The four-class solution had smaller AIC and BIC compared to the three-class solution and a lower entropy suggesting more precise class alignment. The smallest class sizes between the three and four class models were also comparable. While AIC, BIC, and entropy continued to

reduce for the five-class model, the proportion on the smallest class was too small (below 1% of the sample) to be clinically meaningful. The selection of a four-class model is counter to the analysis I ran in Mplus, which suggested a three-class solution was the best fitting model. See Figure 1 for a graphical depiction of indicator probabilities for the four-class solution.

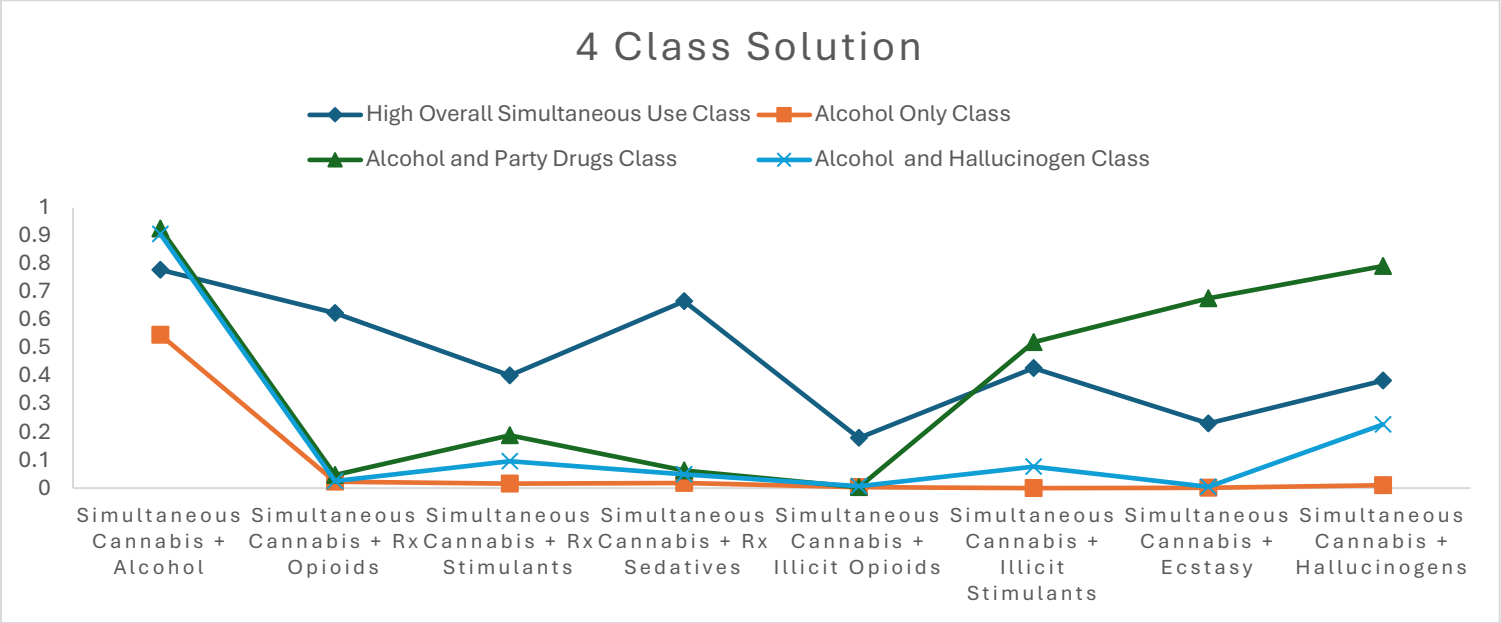
Table 1. *Fit Statistics for 2 to 5 Latent Class Models*

Number of Classes	BIC	AIC	G^2	X^2	Entropy	Proportion in smallest class
2	43889.57	43763.42	43889.57	43889.57	1.78	.098
3	43731.61	43538.66	43731.61	43731.61	1.76	.028
4	43542.15	43282.42	43542.15	43542.15	1.76	.023
5	43457.15	43130.63	43457.15	43457.15	1.75	.005

Note. BIC = Bayesian information criterion, AIC = Akaike information criterion, G^2 = Likelihood ratio, X^2 = Chi-square . Bold indicates class model selection.

The largest class was called the *Alcohol Only* class, capturing 67% of the sample. Those belonging to this class were highly likely to use cannabis with alcohol, but unlikely to engage in simultaneous cannabis use with any other substance. The next class was *Alcohol and Hallucinogens* representing 28% of the sample. Those in this group were highly likely to combine cannabis with alcohol and moderately likely to combine cannabis with hallucinogens and PCP. The third class was the *Alcohol and Party Drugs* group, capturing 3.4% of the sample. Individuals were highly likely to combine cannabis with alcohol, hallucinogens and PCP, ecstasy, and illicit stimulants. The final and smallest class was the *High Overall Simultaneous Use* group (2.3%). Folks in this class were highly likely to combine cannabis with alcohol, prescription opioids, and prescription opioids, and moderately likely to combine cannabis with prescription stimulants, illicit stimulants, hallucinogens and PCP, and ecstasy.

Figure 1. *Estimated Probabilities of Indicators for the 4 Class Model*



Note. Rx = Prescription. All indicator variables reflect simultaneous use with cannabis.

Conclusion

This study sought to replicate the LCA done in Farrelly & Wardell (in preparation) in R to identify unique groups of simultaneous cannabis with other dug use in a nationally representative sample of Canadians. I identified four latent classes: an alcohol only group with simultaneous cannabis and alcohol use only, a group characterized by simultaneous use of cannabis with both alcohol and hallucinogens, a group who combined cannabis with alcohol and drugs often classified as "party drugs" (i.e., hallucinogens, ecstasy, illicit stimulants), and a group who combined cannabis with multiple other substances including alcohol, illicit drugs, and prescription drugs.

The characterization of an alcohol-only group, alcohol and hallucinogen group, and high overall simultaneous use group is consistent with the classes identified in Farrelly & Wardell's Mplus analysis. However, the selection of a four-class model is inconsistent with their use of a three-class model. The difference in model selection may reflect the different capabilities of R and Mplus when conducting an LCA. For example, I had to operationalize simultaneous cannabis and alcohol use as dichotomous, as R could not support an LCA with an ordinal variable like what was used in the Mplus analysis. The need to operationalize simultaneous use with alcohol in a different way is notable, as Farrelly & Wardell found variance in endorsement of frequent vs. occasional simultaneous use in the classes. Specifically, their class that identified individuals who only used cannabis with alcohol found that these individuals were highly likely to occasionally use cannabis with alcohol, but moderately likely to *never* use cannabis with alcohol. My analysis in R cannot make this distinction in frequency of simultaneous cannabis

and alcohol use. Overall, this analysis supports findings that there are distinct classes of simultaneous cannabis with other drugs.

Analysis Reflections: Mplus vs. R

The original data analysis I replicated was completed in Mplus, and the present analysis was done in R using the poLCA package. The use of a different statistical software revealed notable differences in not just how the analyses are run, but also in how the analysis had to be structured. A key difference between the software is the ability to use ordinal variables. In Mplus, I was able to use a three-level ordinal variable to represent simultaneous cannabis with alcohol use (i.e., Never, Occasionally, Frequently). This was done as simultaneous alcohol and cannabis use is common, and frequency of simultaneous use of alcohol and cannabis can represent different risk profiles. This was supported in in the current dataset as there was representation across all levels of simultaneous cannabis and alcohol use. In contrast, the other simultaneous use variables were highly skewed to 0, thus dichotomizing to presence of simultaneous use in the past or year or not allowed for better model estimation. However, LCA packages in R cannot support ordinal response indicators, thus simultaneous use with alcohol was dichotomized to presence of simultaneous use or not. This difference in variable operationalization appeared to have significant implications on class results, suggesting the importance of software selection when running an LCA.

There was other, albeit, less significant differences running an LCA in R versus Mplus. To start, R did not allow zero values in indicators. All variables were originally coded as 0/1 (no simultaneous use/presence of simultaneous use) and had to be recoded to 1/2. Further, there were differences in fit statistics for each software. Both generate AIC, BIC, and entropy to compare goodness of model fit. However, there are key differences in the meaning of entropy for each software. In Mplus entropy is standardized to be between 0 to 1, with values closer to 1 suggesting a more precise class alignment. However, R utilizes Shannon Entropy, with lower values suggesting better likelihood of model fit. Another key difference for testing model fit is the absence of the Lo-Mendell-Rubin adjusted likelihood test (LMR-LRT) in R. The LMR-LRT tests if class model *k* fits better than the model with *k*-1 classes, with significant LMR-RT *p* -values suggesting a better fit. The inability to generate the LMR-RT is a limitation, as it could have helped to select between a three and four class solution. It may be that the four-class solution was not significantly better than the three class, but R is unable to test this. However, R can generate other fit statistics such as likelihood ratio and chi-square.

Another difference between software, and a limitation for R, is how they model the inclusion of class predictors. Class predictors are added with a one-step approach in R. A one-step approach combines the LCA model and a latent regression model with the predictors into a joint model (Asparouhov & Muthén. 2014). However, this approach can be limited as it risks the validity of the class formation. Thus, when using a one-step approach the class structure may shift when adding predictors, adding an additional decision point in the analysis of selecting

model fit before or after including predictors. The use of a one step model is made even more complicated when using many predictors, which was the case in the analysis I was trying to replicate. The instability of including class predictors is why this replication analysis only considered the model selection portion. The inclusion of class predictors led to too much variability in the class structure and could not be meaningfully compared to the analysis I replicated. In contrast, three step modelling in Mplus first models the regular LCA, then allows for class assignment, and lastly estimates the relationships between classes and predictors (Asparouhov & Muthén. 2014).). This three-step approach helps to reduce bias and ensue model stability.

However, R does have unique functions that are more user friendly, namely greater ease for both inputting data and comparing fit. Mplus requires dataset names to be specified in the syntax file, which does add extra cleaning steps. In contrast, R can use the csv, excel, or dat file as is. Further, Mplus generates each LCA model as a separate output file, which can add additional burden when comparing goodness of fit statistics, while R allows for model comparisons in the code and can hold all models in the same file allowing for easier comparison. While I had initially planned to generate better plots within R, it turns out that the ability to plot class probabilities in R was much more tedious than anticipated, and for the sake of time and my own sanity was dropped from this project. Instead, I utilized a line graph in excel, same as I did for Mplus output, to create a visual depiction of model fit.

Considering both software, Mplus appears to a more sophisticated way to conduct an LCA. Despite R allowing for easier data input and comparison across models, Mplus is better equipped to handle complex data and can generate more comprehensive goodness of fit statistics to aid in model selection. The inability to allow for a 3-step model regression is a significant limitation for R and further supports the utility of Mplus for conducting an LCA.

References

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329-341. https://doi.org/10.1080/10705511.2014.915181

Bailey, A. J., Farmer, E. J., & Finn, P. R. (2019). Patterns of polysubstance use and simultaneous co-use in high risk young adults. *Drug and Alcohol Dependence*, *205*, 107656.. https://doi.org/10.1016/j.drugalcdep.2019.107656

Carlini, B. H., & Schauer, G. L. (2022). Cannabis-only use in the USA: Prevalence, demographics, use patterns, and health indicators. *Journal of Cannabis Research*, *4*(1), 39. https://doi.org/10.1186/s42238-022-00143-y

Connor, J. P., Gullo, M. J., Chan, G., Young, R. M., Hall, W. D., & Feeney, G. F. (2013). Polysubstance use in cannabis users referred for treatment: Drug use profiles, psychiatric comorbidity and cannabis-related beliefs. *Frontiers in Psychiatry*, *4*, 79. https://doi.org/10.3389/fpsyt.2013.00079

Crummy, E. A., O'Neal, T. J., Baskin, B. M., & Ferguson, S. M. (2020). One is not enough: Understanding and modeling polysubstance use. *Frontiers in Neuroscience*, *14*, 569. https://doi.org/10.3389/fnins.2020.00569

Davis, C. N., Slutske, W. S., Martin, N. G., Agrawal, A., & Lynskey, M. T. (2019). Identifying subtypes of cannabis users based on simultaneous polysubstance use. *Drug and Alcohol Dependence, 205,* 107696. https://doi.org/10.1016/j.drugalcdep.2019.107696

Health Canada, 2024. Canadian Cannabis Survey 2024: Summary. Retrieved from https://www.canada.ca/en/health-canada/services/drugs-medication/cannabis/research-data/canadian-cannabis-survey-2024-summary.html. Accessed December 11, 2025.

Hochheimer, M., Sacco, P., & Ware, O. D. (2020). Latent classes of lifetime drug use disorder in national epidemiological survey on alcohol and related conditions–III. *Addictive Behaviors*, *106*, 106379. https://doi.org/10.1016/j.addbeh.2020.106379

LeComte, R., Skandan, N., Hochheimer, M., Pattillo, E., White, J., Huhn, A., & Ellis, J. (2025). A systematic review of latent class analyses of adult polysubstance use patterns. *Experimental and Clinical Psychopharmacology, 33*(6), 531–575. https://doi.org/10.1037/pha0000791

Lee, C. M., Calhoun, B. H., Abdallah, D. A., Blayney, J. A., Schultz, N. R., Brunner, M., & Patrick, M. E. (2022). Simultaneous alcohol and marijuana use among young adults: A scoping review of prevalence, patterns, psychosocial correlates, and consequences. *Alcohol Research: Current Reviews*, *42*(1), 08. https://doi.org/10.35946/arcr.v42.1.08

Yurasek, A. M., Aston, E. R., & Metrik, J. (2017). Co-use of alcohol and cannabis: A review. *Current Addiction Reports*, *4*(2), 184-193. https://doi.org/10.1007/s40429-017-0149-8