P9120 Final Project Presentation

Latent Class Analysis and Clustering on both Continuous and Categorical Cases

Qihang Wu | qw2331

December 15, 2022

Outline

Motivation

As one type of multivariate analysis, clustering is commonly used nowadays. It belongs to unsupervised learning family and aims to group subjects in a way that objects within the same group share more similarities than other groups.

Motivation

As one type of multivariate analysis, clustering is commonly used nowadays. It belongs to unsupervised learning family and aims to group subjects in a way that objects within the same group share more similarities than other groups.

- **Distance-based:** K-means, fuzzy k-means, hierarchical clustering
- **Distribution-based:** Clusters modeled using distributions and the expectation-maximization (EM) algorithm
- **Graph-based:** Divide the vertices in multiple groups, s.t., the numbers of edges lying between the groups is minimal

Clustering on some categorical variables is still underdeveloped compared to continuous cases and in great demand especially in social networks, psychology, etc.

Latent Class Analysis

Latent Class Analysis (LCA) is a model-based clustering. It is a Finite Mixture Model (FMM) assuming the presence of unobserved groups within the overall population.

- Use probabilistic models rather than predefined distance measures
- Used in categorical variables
- Return probabilities instead of class memberships

Customer Personality

Demographics of Customer (n=2238)

Characteristic	N = 2,238[1]
Age	52 (45, 63)
Education	
Associate degree	203 (9.1%)
Bachelor's degree	1,127 (50%)
Below some college	54 (2.4%)
Doctoral degree	484 (22%)
Master's degree	370 (17%)
Is Alone	794 (35%)
Income	51,790 (35,528, 68,307)
Number of Children	
0	638 (29%)
1	1,126 (50%)
2	421 (19%)
3	53 (2.4%)
Family Size	
1	254 (11%)
2	762 (34%)
3	889 (40%)
4	301 (13%)
5	32 (1.4%)
Number of Days Since Last Purchase	49 (24, 74)
Complaint (in last 2 yrs)	21 (0.9%)
Wine[2]	173 (23, 505)
Fruits	8 (1, 33)
Meat Products	67 (16, 232)
Fish Products	12 (3, 50)
Sweet Products	8 (1, 33)
Gold Products	24 (9, 56)
Number of Purchases made with a discount	2 (1, 3)
Customer Accepted the Offer	608 (27%)
Through the Company's Website[3]	4 (2, 6)
Using a Catelogue	2 (0, 4)
Directly in Stores	5 (3, 8)
Website Visits (in last month)	6 (3, 7)

[1] Median (IQR); n (%)
[2] Amount spent on a type of product in last two years
[3] Number of purchases made through a way

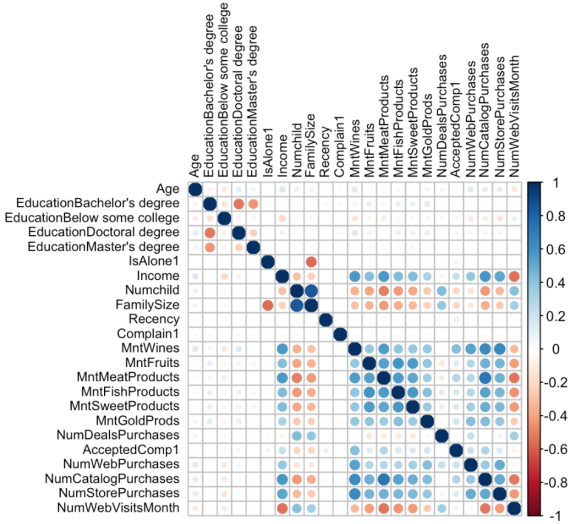


Figure: Correlation of Customers' Traits

- After cleaning, left **2238** observations with **20** variables
- Some correlations, e.g., number of kids at home and amount spent on wine, etc

Big Five Questionnaire

- The raw data contains over 1M observations with five personality traits: **openness, conscientiousness, extraversion, agreeableness, and neuroticism** (OCEAN)
- 10 questions for each trait with scale 1-5 (s.t., 1=Disagree, 3=Neutral, 5=Agree)
- Use 0.2% observations, remove observations with any answer to 0[1] and select those IPC=1[2], finally 1191 observations

	id	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4	EST5	EST6	EST7	EST8	EST9	EST10	AGR1	AGR2	AGR3	AGR4	AGR5	AGR6
1:	1	4	2	3	3	5	2	4	3	3	2	2	4	2	4	2	3	2	2	2	2	1	4	1	5	2	5
2:	2	5	1	5	1	5	1	5	1	5	1	1	3	5	3	1	1	3	1	1	3	4	5	5	2	1	1
3:	3	5	1	5	1	4	1	4	1	5	1	2	4	2	5	2	1	2	1	1	1	1	5	4	5	1	5
4:	4	3	2	5	2	4	2	4	4	4	2	4	3	4	3	2	3	3	3	3	1	2	5	2	5	2	4
5:	5	1	2	2	3	3	2	1	4	2	5	2	4	3	4	4	1	1	1	1	2	3	4	1	4	4	4
6:	6	1	2	2	4	2	4	1	4	1	4	2	5	4	5	2	2	4	4	2	1	1	4	2	5	4	4
7:	7	3	3	3	3	2	2	1	2	5	4	3	3	5	2	4	4	3	3	5	2	1	3	1	5	3	5
8:	8	1	4	2	4	3	3	1	4	2	4	5	2	4	1	3	4	3	3	4	4	4	4	3	4	3	4
9:	9	3	4	5	3	1	4	5	4	5	5	3	5	5	3	3	3	1	1	4	1	2	5	1	5	3	5
10:	10	1	3	2	5	1	2	1	5	2	5	4	2	5	4	4	4	5	5	4	4	1	4	1	5	1	4

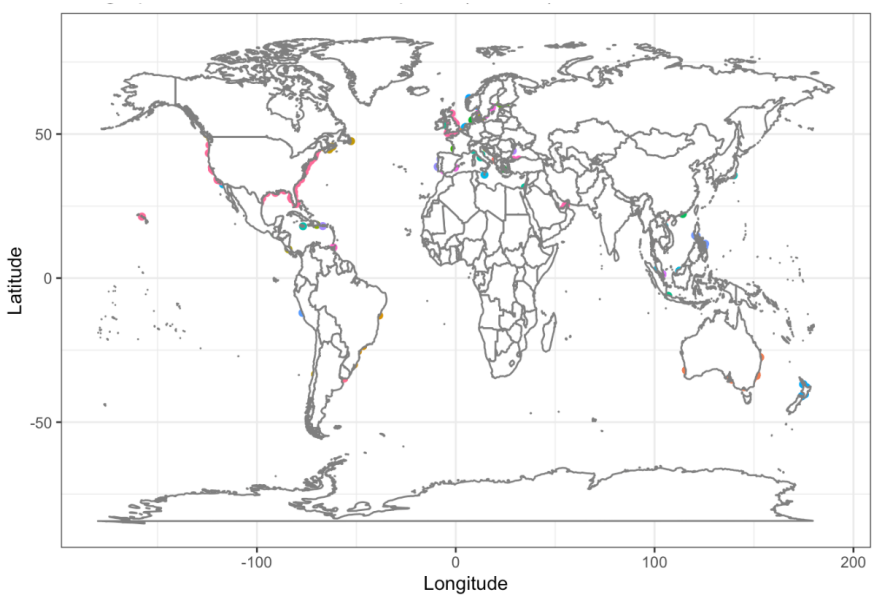
Figure: A Glimpse of Big Five Data (part)

[1] Score 0 may due to the participants didn't pick or forgot to pick an answer.

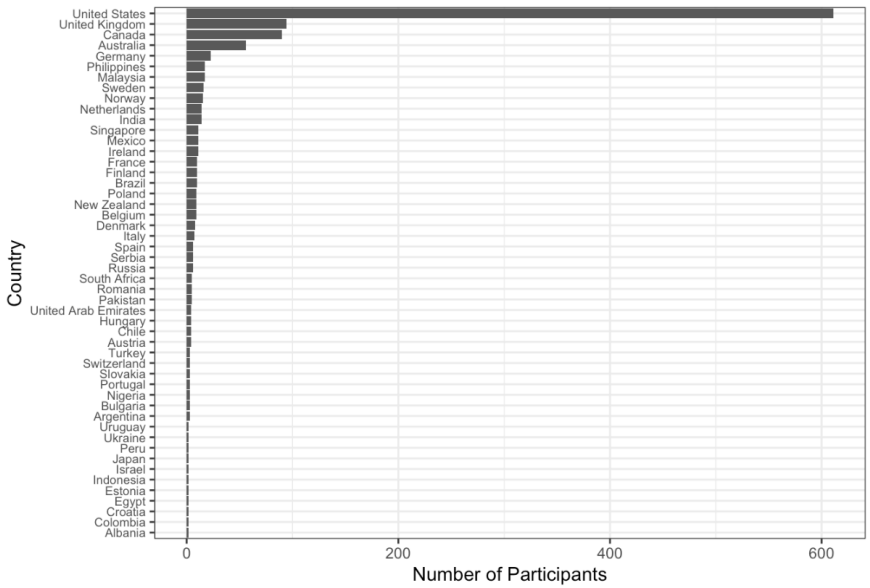
[2] IPC: number of records per user's IP address. Higher values indicate shared networks.

Big Five Questionnaire (Con't)

Other exploratory analysis:



(a) Distribution of Participants (n=1191)



(b) Number of Participants by Country

Statistical Learning

Customer personality analysis:

- **K-means** on only continuous variables with Gap statistics
- **K-medoids** with Gower distance[3] for mixed types of data with Silhouette coefficient (k=3)

Big five personality analysis:

- Latent class analysis with number of classes chosen by BIC (k=5)

[3] Scaled in a numerical range from 0 (identical) to 1 (entirely different).

Demographics of Customers

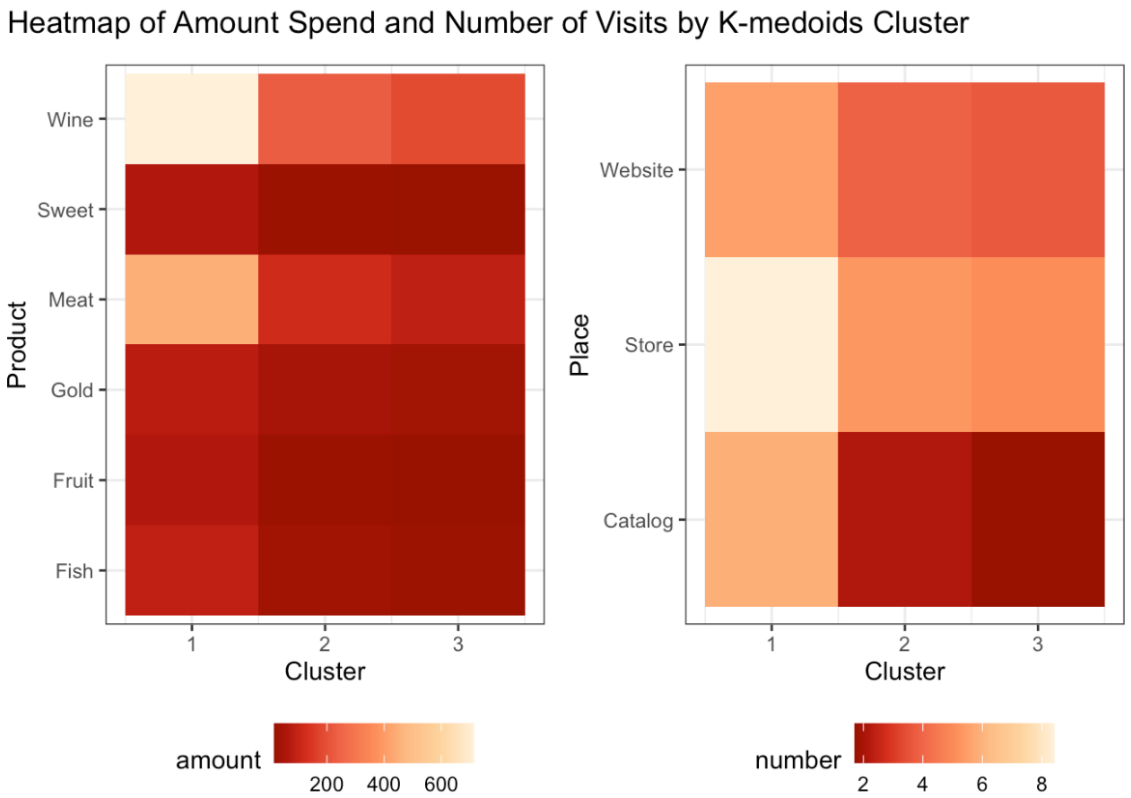
Demographics of Customers for each K-medoids Cluster (n=2238)

Characteristic	Overall, N = 2,238[1]	Cluster 1, N = 446[1]	Cluster 2, N = 671[1]	Cluster 3, N = 1,121[1]	p-value[2]
Age	52 (45, 63)	53 (43, 65)	52 (45, 62)	51 (45, 62)	0.2
Education					0.004
Associate degree	203 (9.1%)	38 (8.5%)	53 (7.9%)	112 (10.0%)	
Bachelor's degree	1,127 (50%)	222 (50%)	353 (53%)	552 (49%)	
Below some college	54 (2.4%)	0 (0%)	20 (3.0%)	34 (3.0%)	
Doctoral degree	484 (22%)	115 (26%)	142 (21%)	227 (20%)	
Master's degree	370 (17%)	71 (16%)	103 (15%)	196 (17%)	
Is Alone	794 (35%)	123 (28%)	671 (100%)	0 (0%)	<0.001
Income	51,790 (35,528, 68,307)	77,040 (69,224, 82,783)	46,734 (33,347, 61,825)	44,155 (31,535, 57,937)	<0.001
Number of Children					<0.001
0	638 (29%)	362 (81%)	141 (21%)	135 (12%)	
1	1,126 (50%)	74 (17%)	368 (55%)	684 (61%)	
2	421 (19%)	8 (1.8%)	141 (21%)	272 (24%)	
3	53 (2.4%)	2 (0.4%)	21 (3.1%)	30 (2.7%)	
Family Size					<0.001
1	254 (11%)	113 (25%)	141 (21%)	0 (0%)	
2	762 (34%)	259 (58%)	368 (55%)	135 (12%)	
3	889 (40%)	64 (14%)	141 (21%)	684 (61%)	
4	301 (13%)	8 (1.8%)	21 (3.1%)	272 (24%)	
5	32 (1.4%)	2 (0.4%)	0 (0%)	30 (2.7%)	
Number of Days Since Last Purchase	49 (24, 74)	42 (19, 71)	51 (27, 75)	50 (26, 75)	0.002
Complaint (in last 2 yrs)	21 (0.9%)	1 (0.2%)	7 (1.0%)	13 (1.2%)	0.2
Wine[3]	173 (23, 505)	710 (464, 960)	96 (17, 371)	73 (15, 292)	<0.001
Fruits	8 (1, 33)	48 (23, 93)	6 (1, 21)	4 (1, 15)	<0.001
Meat Products	67 (16, 232)	427 (258, 610)	44 (12, 142)	31 (12, 106)	<0.001
Fish Products	12 (3, 50)	72 (33, 130)	8 (2, 31)	7 (2, 21)	<0.001
Sweet Products	8 (1, 33)	49 (24, 96)	6 (1, 21)	5 (1, 16)	<0.001
Gold Products	24 (9, 56)	57 (32, 114)	22 (8, 50)	16 (6, 40)	<0.001
Number of Purchases made with a discount	2 (1, 3)	1 (1, 1)	2 (1, 3)	2 (1, 3)	<0.001
Customer Accepted the Offer	608 (27%)	342 (77%)	124 (18%)	142 (13%)	<0.001
Through the Company's Website[4]	4 (2, 6)	5 (4, 7)	3 (2, 5)	3 (2, 5)	<0.001
Using a Catelogue	2 (0, 4)	6 (4, 7)	1 (0, 3)	1 (0, 2)	<0.001
Directly in Stores	5 (3, 8)	8 (6, 11)	4 (3, 7)	4 (3, 7)	<0.001
Website Visits (in last month)	6 (3, 7)	3 (2, 5)	6 (4, 7)	6 (5, 7)	<0.001

[1] Median (IQR); n (%)
[2] One-way ANOVA; Pearson's Chi-squared test
[3] Amount spent on a type of product in last two years
[4] Number of purchases made through a way

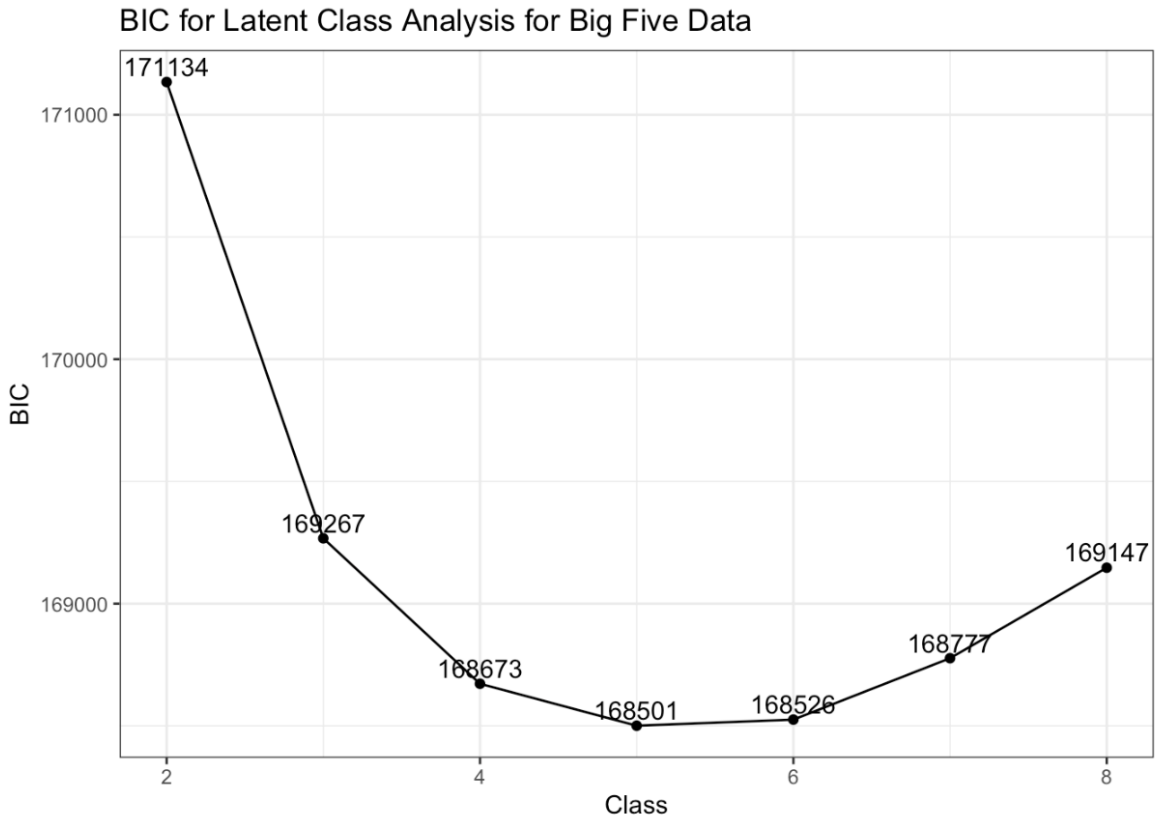
- **One-way ANOVA** for continuous and **Chi-squared test** for categorical variables
- Difference exist for many variables (i.e., $P<0.001$)

Promotion and Place



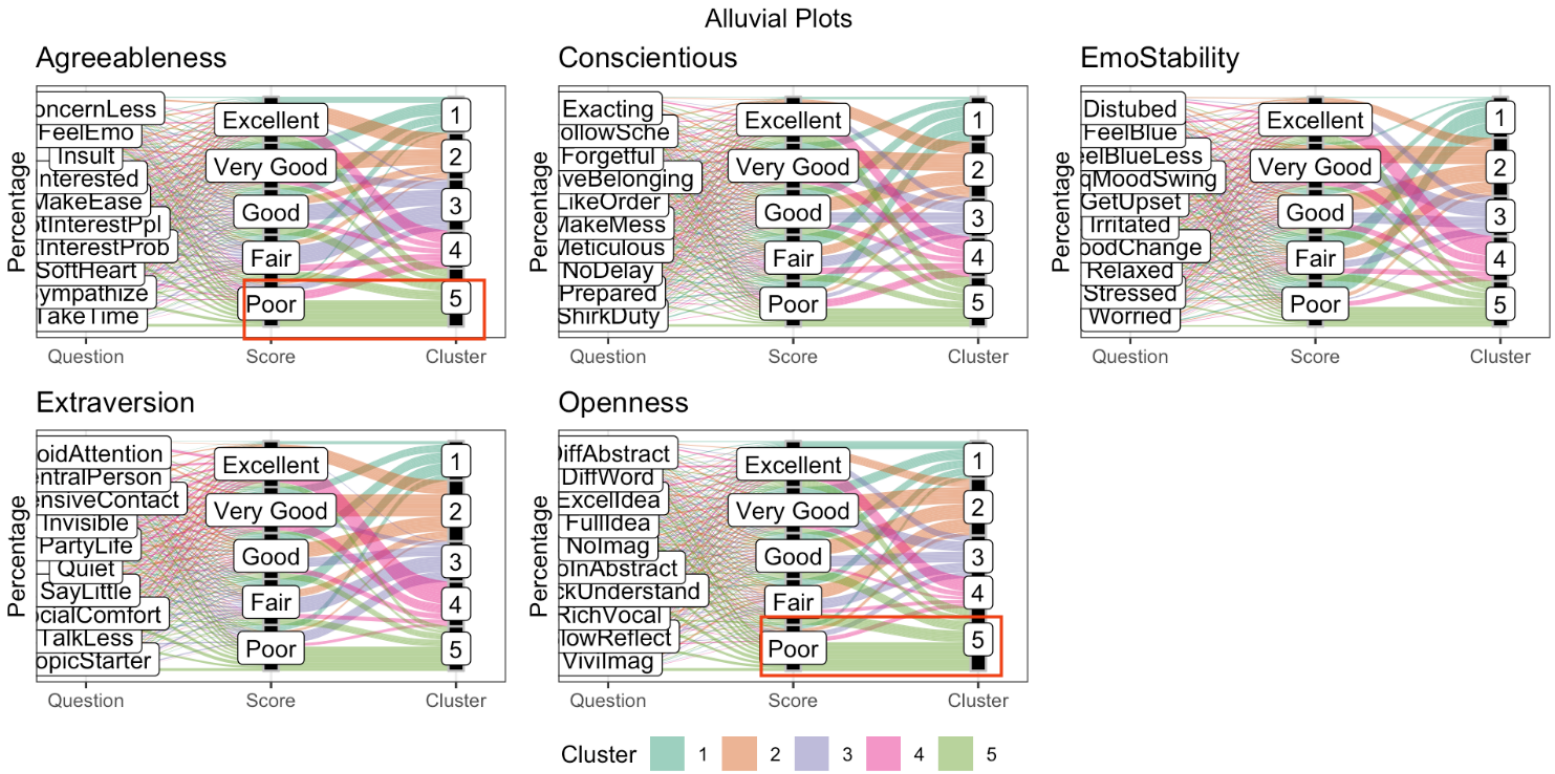
Cluster 1: higher income, few kids tend to buy more wine and meat, and more likely to shop directly in stores

BIC for LCA



Model-based method: prefer 5 classes with the lowest BIC score

Personality for each Class



- Class 2 is interested in others, prepared, relaxed, talkative, and creative
- Class 3 prefers solitude and cares less about how other people feel while Class 5 performs worse on nearly everything

Conclusion & Limitation

- K-medoids and LCA works well when categorical variables involved
- Big five questionnaire data is biased and more subjective; small data sample for analysis; lack information about the participants' demographic, e.g., sex, age, race, etc.
- Characteristics analysis for each LCA class
- Compare LCA with K-means, etc.

Thank you for listening!