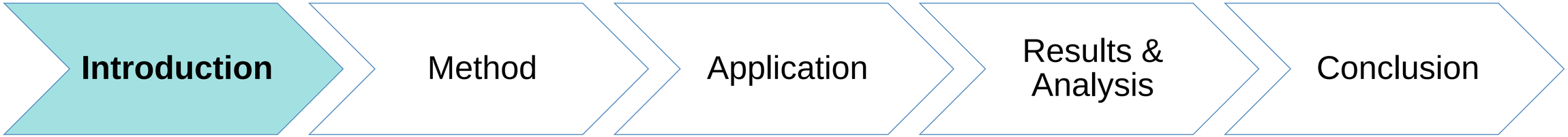
Analyzing credit card usage patterns

20151640 Youngseok Song

Contents

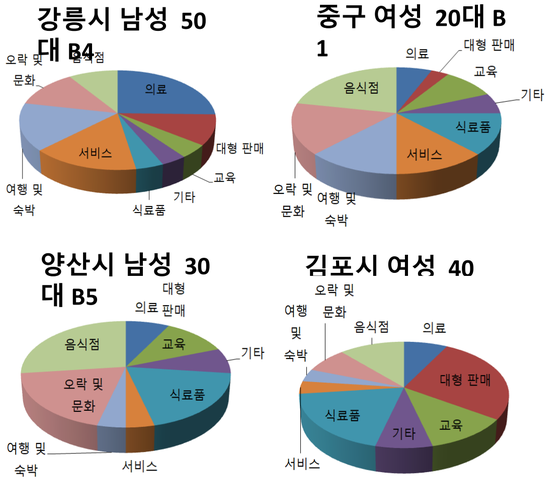
Contents

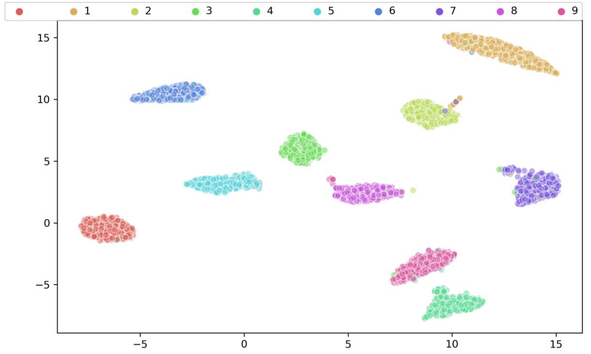


- Objective
- Dataset

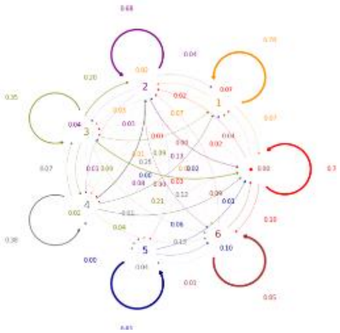
Introduction

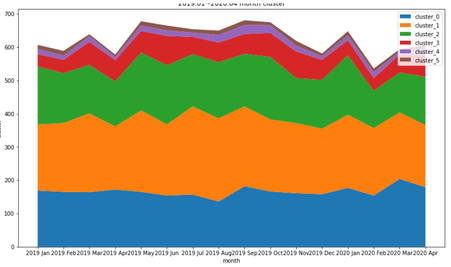
2019.01 ~ 2020.04





Clustering





- The goal of our project is to **identify credit card usage patterns** of different groups characterized by region, gender, age, income (B1~B11), etc.
- We first identify consumption patterns of different groups from raw data (e.g., 강릉시 남성 50대 B4 in the figure above) and then **cluster groups** to find how different groups have different patterns.
- After clustering, the characteristics of each cluster in the optimal number of clusters(= 7), the transition matrix, and the monthly cluster change were analyzed.

Introduction

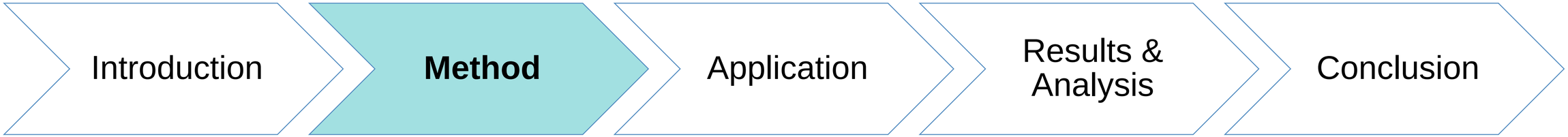
Raw data

	기준년월	고객유형	가맹점소재지1	가맹점소재지2	가맹점소재지3	업종대분류명	업종중분류명	업종소분류명	성별	연령대별	가구생애주기별	연평균소득추정	이용금액	이용건수
574784	202004	내국인-거주민	경상남도	창원시	반송등	생활	보건/위생	미용원	남성	20대	1인가구	B2	339000	22
2741922	202004	내국인-관광객	경기도	화성시	향남읍	음식	일반음식	일반한식	남성	50대	성인자녀가구	B7	4108300	109
4711523	202004	내국인-관광객	인천광역시	미추홀구	주안4등	쇼핑	유통업영리	대형할인점	남성	40대	신혼영유아가구	B5	13600	3
3710540	202004	내국인-관광객	부산광역시	서구	동대신2등	생활	연료판매	주유소	남성	30대	신혼영유아가구	B3	939916	28
151438	202004	내국인-거주민	경기도	남양주시	별내등	문화	학원	예체능학원	여성	40대	신혼영유아가구	B4	740000	4

- Credit card usage data of **16M users from BC card**
- Monthly data from **Jan 2019 to April 2020**
- Shape: 90M rows, 14 features

항목	세부항목	데이터 값
고객유형	내국인-거주민/내국인-관광객	
카드사용지역	가맹점소재지1	광역시도
	가맹점소재지2	시군구
	가맹점소재지3	행정동
고객속성	성별	남성,여성
	연령대별	20세미만,20대,30대,40대,50대 60대이상
	가구생애주기별	1인가구,신혼영유아가구,초중고 자녀가구,성인자녀가구,노인가구
	연평균소득(추정)	B1~B11

Contents



- N2D
  - Autoencoder, UMAP
  - GMM

N2D

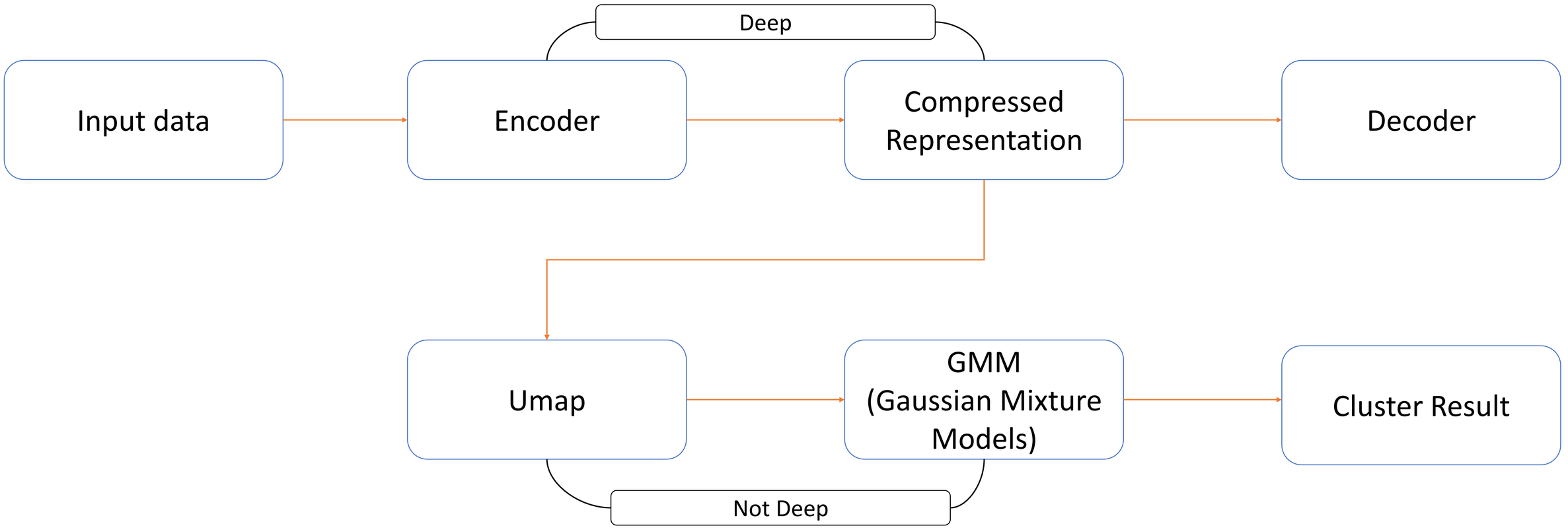
- **N2D : (Not Too) Deep Clustering** via Clustering the Local Manifold of an Autoencoded Embedding

- How does it work?
  1) Embedding the data with '**autoencoder**' (manifold learner)
  2) Learning manifold with '**UMAP**' (manifold learner)
  3) Applying traditional clustering techniques '**GMM**'

- F_C : clustering algorithm (GMM, k-means)
- F_M : manifold learner (Isomap, t-SNE, UMAP)
- F_A : autoencoder (all layers use ReLU activation function, apply Adam optimizer)

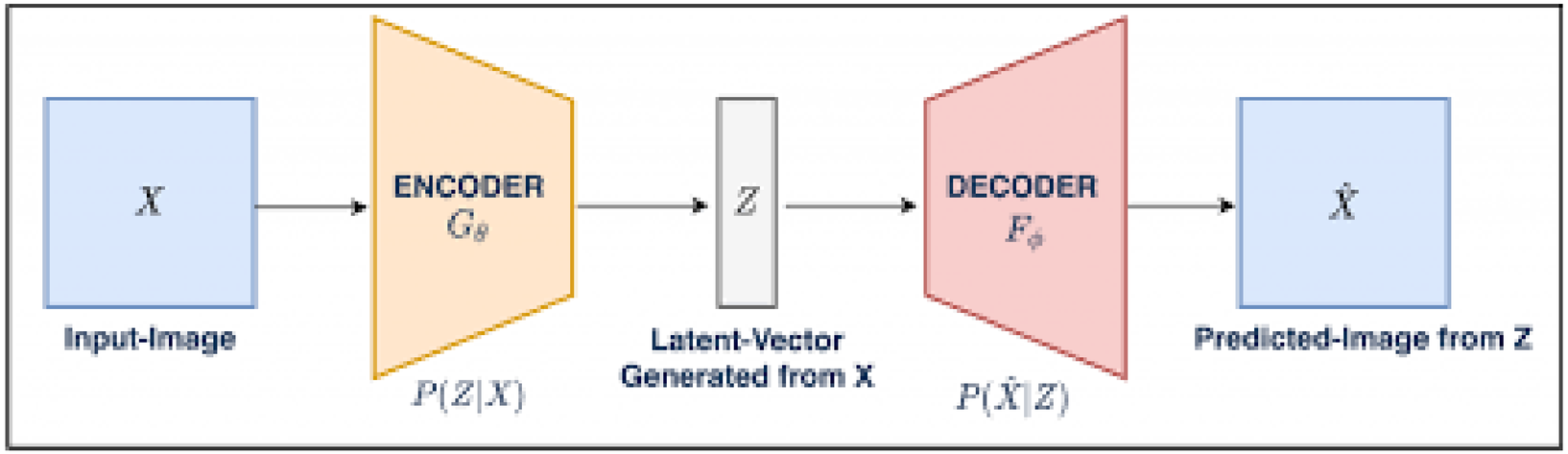
$$C = F_C(F_M(F_A(X)))$$

N2D


Deep
Input data
Encoder
Compressed Representation
Decoder
Umap
GMM (Gaussian Mixture Models)
Cluster Result
Not Deep

N2D

- **Autoencoder** (compress the input into a lower dimensional space)
  - Consists of two components
    - **Encoder**: mapping the input data ‘x’ to a new smaller feature vector $\mathrm{h} = f(x)$
    - **Decoder**: mapping the learned feature vector to original vector r= $g(h)$
  - Limitation: Do not preserve the distance of data well



N2D

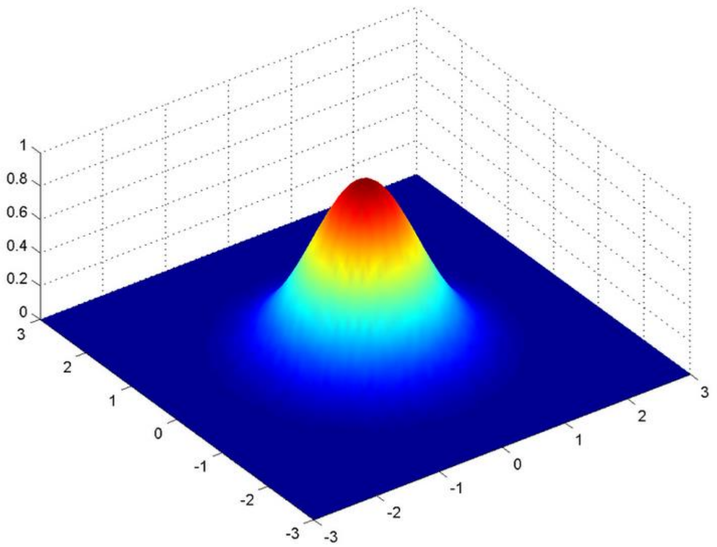
- **UMAP** (Uniform Manifold Approximation and Projection)
  - Preserve distance within global and local structure
  - Fast speed to learn
    - **Fuzzy logic**: Everything is a matter of degree. It does not have deterministic value, but analog, the infinite concentration of gray between black and white
      - Embedding is found by searching a low dimensional projection of data

N2D

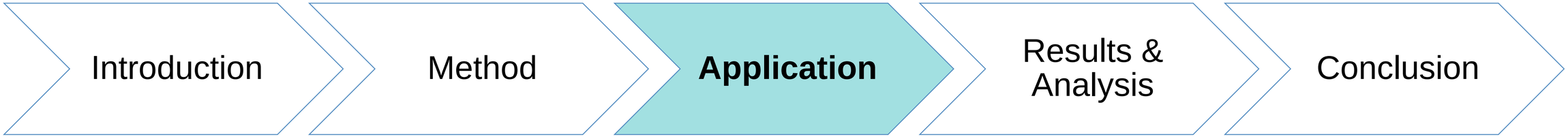
- **GMM** (Gaussian Mixture Models)
  - A **clustering algorithm** which contains multiple Gaussian distributions mixed
    - The idea is to represent probability distribution on real world into mixture of ‘K’ number of Gaussian distributions

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

-> find out proper μ_k, π_k, Σ_k while learning GMM



Contents



- Data preprocessing
- Clustering

Data preprocessing

- **Removing missing values**

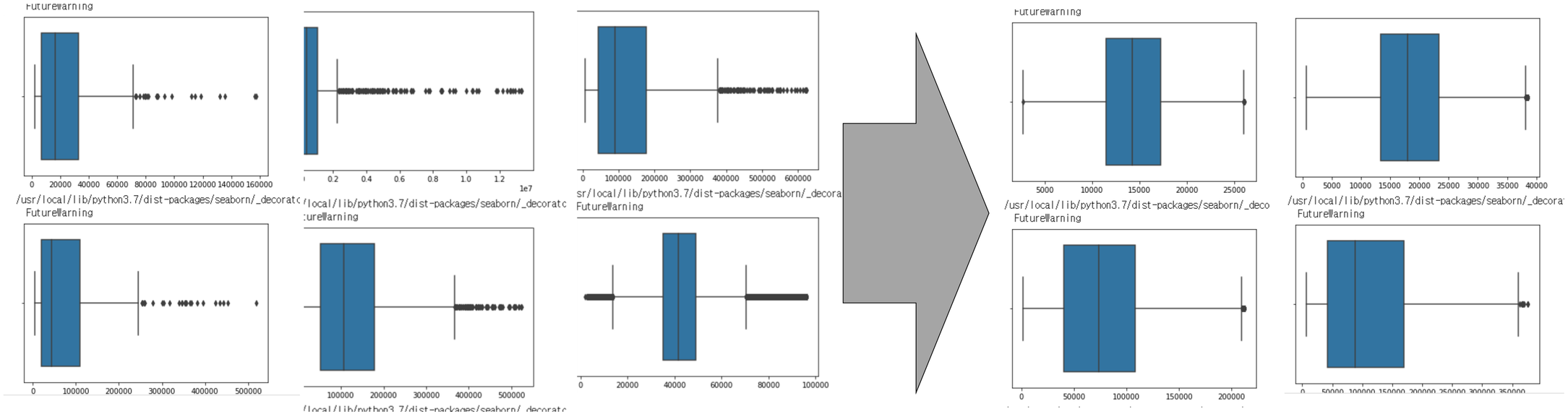
```
temp1 = df.shape


""" row 삭제: 연령대/가구/소득 x값 """
df.drop(df[df.연령대별 == 'x'].index, inplace=True)
df.drop(df[df.가구생애주기별 == 'x'].index, inplace=True)
#4월 txt에는 연령대별과 가구생애 주기별은 'x' 값 없음.
df.drop(df[df.연평균소득추정 == 'x'].index, inplace=True)

#결측치 제거

if df.isnull().sum().sum() != 0:
  df.dropna(axis=1,how='any')

temp2 = df.shape
temp3 = temp1[0]-temp2[0]

print('데이터 shape 변화: {0} -> {1}'.format(temp1, temp2))
print('삭제된 row 수: {0}개, 전체의 {1}%'.format(temp3, round(temp3/temp1[0]*100),1
```

```
데이터 shape 변화: (5469328, 14) -> (5336937, 14)
삭제된 row 수: 132391개, 전체의 2%
```

- Remove 'x' values and missing values.
- **2% were removed**.

Data preprocessing

- **Removing outliers**



Visualization of the **outliers before and after** outliers removed were confirmed through boxplot.

- **4% were removed**.

Data preprocessing

- **Category mapping**
  - Original BC card data: Divided into **‘Small Divisions, Middle Divisions, Large Divisions’**
  - Manually divided customer data into **16 categories**

New Category Name	New Category Name
음식점 (Restaurant)	의료 (Hospital)
교통 (Transportation)	식료품 (Grocery)
오락 및 문화 (Entertainment)	서비스 (Service)
의류 및 잡화 (Clothes)	자동차 (Vehicle)
보험 (Insurance)	교육 (Education)
여행 및 숙박 (Travel)	가구 및 전자제품 (Furniture and Appliance)
대형 판매점 (Mart)	종합소매 (Retail)
전자상거래 (E-Commerce)	기타 (Etc.)

Data preprocessing

음식점	의료	교통	식료품	오락 및 문화	서비스	의류 및 잡화	자동차	보험	교육	여행 및 숙박	가구 및 전자 제품	기타	대형 판매점	종합소매	전자상거래
일반음식	**의료기관**	**여행업**	**음식료품**	**레저업소**	**용역서비스**	**신변잡화**	**연료판매**	**보험**	**서적/문구**	**휴게**	**주방용품**	**기타**	**유통업영리**	**유통업비영리**	**유통업영리(생활)**
일반한식	종합병원	항 공 사	제 과 점	골프경기장	종합용역	가 방	주 유 소	생명 보험	일반 서적	칵테일바	주방 용구	성인용품점	일반백화점	구내매점(국가기관등)	인터넷 p/g
일식횟집	병원	관광여행	정 육 점	골프연습장	가례서비스	시 계	L P G	손해 보험	전문 서적	주점	주방용식기	기계공구	대형할인점	농,축협직영매장	인터넷종합 mall
중국음식	한방병원	고속 버스	주류판매점	카지노	보관창고업	귀 금 속	유류 판매	기타보험	정기간행물	스넥	정 수 기	기타업종	면세점	농협하나로클럽	인터넷 mall
서양음식	치과병원	철 도	농축수산품	스 키 장	화물 운송	악세 사리	기타연료		출판인쇄물	**숙박업**	**기타주방용구**	**보건/위생**		기타비영리유통	상품권
위탁급식업	의원	여 객 선	미곡상	볼 링 장	사무서비스	제 화 점	**자동차정비/유지**		교육테이프	특급 호텔	가전제품	기타대인서비스		**유통업영리(생활)**	전자상거래 상품권
유흥주점(음식)	한의원	택 시	기타음료식품	테니스 장	정보서비스	신 발	자동차시트/타이어		문구용 품	1급 호텔	가전 제품	**단란주점(음식)**		편의점	**유통업영리(쇼핑**
유흥주점	치과의원	택시회사	**건강식품**	수 영 장	법률회계서비스	기념품 점	자동차부품		과학기자재	2급 호텔	냉열 기기	단란주점		연쇄점	통신판매1
	제약회사	렌 트 카	홍삼 제품	헬스 크럽	부동산중개/임대	기타잡화	윤활유전문판매		완 구 점	콘 도	기타전기제품	**건축/자재**		복지	홈쇼핑
	약국	기타교통수단	인삼 제품	종합레져타운	소프트웨어	**의류**	자동차정비		기타서적문구	기타숙박업	**가구**	보일러펌프		매점	pg상품권
	한약방		기타건강식	당구장	공공 요금	정장	중장비수리		**학원**		일반 가구	건축요업품		농축수산가공품	
	산후조리원			노래방	통신서비스	아동의류	카인테리어		외국어학원		철제 가구	조명 기구			
	동물병원			기타레져업	CATV	양 품 점	세 차 장		기능학원		기타가구	페 인 트			
	건강진단			**문화/취미**	이동통신요금	내의판매점	주 차 장		컴퓨터학원		**광학제품**	유 리			
	기타의료기관및기기			골동품 점	위성방송	와이셔츠/타이	견인서비스		예체능학원		카 메 라	목재,석재,철물			
				화 랑	조세서비스	캐쥬얼의류	기타자동차서비스		보습학원		사진관	인테 리어			
				화방표구점	기타용역서비스	스포츠의류	**자동차판매**		학습지교육		기타광학품	부동산분양			
				민예공예품	**보건/위생**	단체복	국산 신차		대학등록금		**사무/통신기기**	기타건축자재			
				수 족 관	이 용 원	맞춤복 점	중고자동차		초중고교육기관		컴 퓨 터	**회원제형태업소**			
				화 원	미 용 원	기타의류	수입자동차		유치원		사무 기기	사무서비스(회원제형태)			
				애완 동물	피부미용실	**직물**	이륜차판매		유아원		통신 기기	**농업**			
				영화관	안경	옷감 직물	기타 운송		독서실		기타컴퓨터	농기계			
				티켓	화 장 품	카페트커텐천	**회원제형태업소**		유학원		기타사무용	비료/농약/사료/종자			
				문화취미기타	미용 재료	침구수예점	자동차서비스(회원제형태)		기타 교육			기타농업관련			
				레저용품	의료 용품	혼수전문점			**회원제형태업소**						
				골프 용품	**수리서비스**	기타직물			서적출판(회원제형태)						
				스포츠레져용품	레져용품수리				학원(회원제형태)						
				총포류판매	가정용품수리										
				악 기 점	신변잡화수리										
				피아노대리점	사무통신기기수리										
				DVD/음반/테이프판매	세탁소										
				보건/위생	기타수리서비스										

Data preprocessing

	기준년월	고객유형	가맹점소재지1	가맹점소재지2	가맹점소재지3	업종대분류명	업종중분류명	업종소분류명	성별	연령대별	가구생애주기별	연령균소득추정	이용금액	이용건수	category	new_category_name
0	202004	내국인-거주민	강원도	강릉시	강남동	T&E	레저업소	기타레저업	남성	20대	1인가구	B2	104400	29	T&E-레저업소	오락 및 문화
1	202004	내국인-거주민	강원도	강릉시	강남동	T&E	레저업소	기타레저업	남성	20대	1인가구	B3	137500	27	T&E-레저업소	오락 및 문화
2	202004	내국인-거주민	강원도	강릉시	강남동	T&E	레저업소	노래방	남성	60대 이상	노인가구	B4	109000	3	T&E-레저업소	오락 및 문화
97	202004	내국인-거주민	강원도	강릉시	강남동	생활	유통업비영리	농축협직영매장	남성	50대	성인자녀가구	B5	2768740	97	생활-유통업비영리	종합소매
98	202004	내국인-거주민	강원도	강릉시	강남동	생활	유통업비영리	농축협직영매장	남성	50대	성인자녀가구	B6	531290	22	생활-유통업비영리	종합소매
99	202004	내국인-거주민	강원도	강릉시	강남동	생활	유통업비영리	농축협직영매장	남성	60대 이상	노인가구	B2	466130	21	생활-유통업비영리	종합소매

- New column is added on raw data called **'new_category_name'**
- 100 million rows hard to handle
  - Mapping 'str' into 'int' using Numpy

- **Before**: **90 million rows**
- **After**: **269,000 rows** ('groupby' used)
- The new categorized names go to column index
- Client's information goes to row index
- (Date, Region, Age, Sex, Income)
- Value refers to **client's payment's ratio** on each group
- Its data frame goes to the model as an input

	year	month	region	sex	age	income	Home	Education	Transporation	ETC	Mart	Insurance	Service	Grocery	Travel	Fun
117325	2019	7	화순군	female	30s	2	0.000000	17.370950	0.000000	0.263246	0.785364	0.000000	2.115391	2.415105	2.830737	4.852969
92037	2019	6	안성시	male	60s or more	5	0.000000	0.000000	0.000000	0.000000	1.168744	0.000000	0.000000	1.728933	0.390436	65.469290
247382	2020	3	용산구	male	50s	2	2.721738	0.020330	0.009632	0.080645	0.462610	5.863965	29.658817	0.286212	0.288350	0.270711
252813	2020	4	강릉시	female	20s or less	2	0.000000	0.960890	0.000000	0.000000	6.241900	0.000000	1.107160	5.769770	7.350072	2.014533
85150	2019	6	구리시	female	30s	2	0.637068	18.981087	0.000000	1.229662	13.745250	0.000000	3.453977	4.606953	2.609157	5.249358

Clustering

N2D Clustering training

- Used Hyperparameter
  - **n_cluster: 4~8**
    - Number of clusters being used
  - **epochs: 30**
    - Single learning (forward + backward) of the entire data
    - No particular error metrics to consider proper eochs
    - 30 is fixed for all trials
  - **batch_size: 32 or 64**
    - Batch size refers to the number of data belonging to one small group when training dataset is divided into small groups
    - Large batch size: use more data to calculate gradient -> make optimization easier but inappropriate to flat problem set (local minima problem)
    - Small batch size: use less data to calculate gradient -> make optimization inaccurately but allow multiple update during one update (x local minima problem)
  - **UMAP (n_components) : 2 or 3**
    - UMAP’s outcome dimension’s scale

Clustering

	0	1	2	3
0	22.6699	13.49374	5.064431	-2.82059
1	24.1604	17.29607	3.101051	-2.16607
2	16.01807	21.04624	7.471575	-3.208
3	18.78028	20.86005	10.45779	2.541709
4	19.25858	17.11965	13.0358	-2.05654
5	23.08576	29.43569	14.93757	-10.5863
6	27.91386	29.28811	24.65679	-21.708
7	42.22137	28.79487	15.40682	-15.7604
8	27.47982	23.52137	-4.33502	-1.94849
9	21.24173	22.06612	-6.55495	-2.186
10	19.84453	21.33597	-6.45077	-2.45242
11	31.93017	36.83085	-4.15624	-20.8523

Batch size:32, Cluster:4, Umap:2

	0	1	2	3	4
0	-4.63256	2.024668	1.762482	5.928173	4.046004
1	-5.58906	3.355677	2.021966	6.380584	5.599243
2	-6.45211	3.212124	3.002046	6.570167	6.633021
3	-5.23841	5.201857	2.026654	9.197597	7.54345
4	-5.58891	2.546394	1.763407	8.441706	4.182973
5	-8.27766	0.93721	4.618453	11.42687	10.829
6	-11.0587	-3.7583	4.966846	13.09876	10.01571
7	-9.73182	-1.45821	2.80231	13.994	9.998755
8	-5.81494	4.643103	1.664333	4.502744	10.07236
9	-5.79078	4.8647	2.574068	2.473558	10.5004
10	-5.68601	4.768048	3.080827	2.572995	10.93568
11	-12.7779	0.284924	5.467241	6.417719	13.3264

Batch size:32, Cluster:5, Umap:2

	0	1	2	3	4	5
0	10.51962	-7.01442	-6.58954	5.277986	-5.63606	-1.28567
1	11.27908	-6.59796	-8.79744	5.758307	-6.25938	-1.54936
2	12.89671	-7.13281	-9.79835	5.121843	-6.17727	-1.95465
3	9.164249	-8.73965	-10.3018	8.076576	-5.66613	-3.71483
4	12.031	-8.29635	-6.67663	7.109968	-5.93114	-0.25725
5	16.78115	-14.4096	-10.3138	4.332789	-6.51816	-2.89596
6	22.2362	-19.3415	-7.68105	1.015955	-9.44037	1.265381
7	17.96752	-18.4908	-7.9526	4.278263	-8.98294	1.667049
8	8.693081	-5.35279	-11.535	3.136445	-7.55928	-0.96005
9	9.38386	-3.50544	-11.7856	1.785022	-7.65715	-2.24473
10	9.869131	-3.92519	-11.8966	1.838393	-6.51506	-3.43255
11	21.14554	-11.1064	-14.3848	-2.12869	-8.32459	-0.45866

Batch size:32, Cluster:6, Umap:2

	0	1	2	3	4	5	6
0	-4.36501	-9.42221	4.895942	4.082256	-1.2648	5.474555	-0.29972
1	-3.01458	-13.4296	4.891753	3.976251	-1.02631	3.603981	-0.14488
2	-2.88485	-14.501	4.973033	4.855749	-0.98131	4.190058	0.939381
3	-3.56853	-15.5179	4.466041	3.399128	-2.99849	5.0565	-2.20241
4	-4.09215	-10.985	7.821515	3.009752	-2.77551	3.459575	-0.51819
5	-5.90623	-17.7257	6.723814	4.356152	-3.38074	7.125538	3.863184
6	-8.22128	-16.8799	11.5919	3.323482	-1.71942	6.89872	8.369391
7	-8.31484	-16.8672	10.78287	1.764939	-1.58282	6.688253	4.50081
8	-2.85433	-14.4632	1.41022	3.131589	1.648483	4.968484	0.341973
9	-2.23363	-15.1217	-0.27175	4.508301	1.916475	4.303768	1.457673
10	-1.56971	-14.7158	0.017396	5.220398	1.42794	4.758406	1.529008
11	-3.73607	-19.355	5.514733	6.9397	2.035791	7.505771	8.806317

Batch size:64, Cluster:7, Umap:2

	0	1	2	3	4	5	6	7
0	3.629201	-1.13888	4.400512	2.083998	2.001389	1.827745	6.904297	0.968093
1	2.205622	-1.39972	4.393733	4.031328	0.507414	1.909153	5.908336	2.493942
2	2.159946	-1.34648	5.767898	4.570843	0.495854	1.323788	7.245409	2.793957
3	1.52149	-3.0509	7.360979	5.091014	1.30908	2.974442	5.149931	3.559344
4	3.567103	-2.54912	8.591551	3.864717	3.463628	1.448111	7.51531	3.718048
5	3.589706	-4.35416	5.307837	5.839229	-2.14846	0.904069	9.405372	3.937433
6	7.429686	-4.89501	2.450361	4.847318	-3.74373	-0.32454	12.87794	4.484667
7	6.969608	-5.33754	3.090805	5.061178	-2.55998	1.962903	10.36024	3.966498
8	-0.14171	-1.44229	2.968124	3.685941	-1.14877	3.681858	6.631552	0.886415
9	-0.6843	0.175238	4.331214	3.913401	-2.06386	2.497396	6.712464	0.685255
10	-1.0503	0.223631	3.788231	4.011294	-1.83621	2.184187	6.807786	1.183144
11	3.015857	-1.01957	3.70127	6.334969	-3.10139	-0.67374	13.69102	1.827064

Batch size:64, Cluster:8, Umap:2

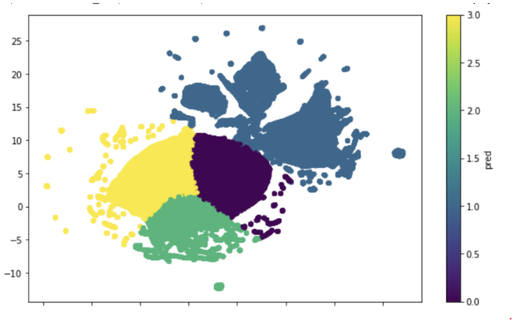
	0	1	2	3	4
0	5.283281	-0.1671	-8.26697	-8.7692	-0.49706
1	7.213083	1.310315	-8.96529	-9.29449	-1.19816
2	7.946592	1.964009	-8.26498	-10.7302	-0.81219
3	11.24403	-0.57049	-10.3894	-9.88514	-2.61356
4	5.085495	-2.67869	-11.3582	-9.71989	-0.51692
5	9.907391	-0.4472	-9.62385	-16.8675	-1.22312
6	0.74268	-2.3155	-11.2539	-20.3905	-0.65468
7	3.207336	-2.51913	-14.3988	-17.7453	-2.98573
8	11.84486	3.659879	-6.61081	-10.3859	-4.7128
9	12.06754	5.135764	-4.04803	-10.2144	-4.03412
10	11.95191	5.411531	-3.74705	-10.3108	-3.09487
11	6.953009	8.643951	-6.11969	-18.8189	-0.81036

Batch size:32, Cluster:4, Umap:3

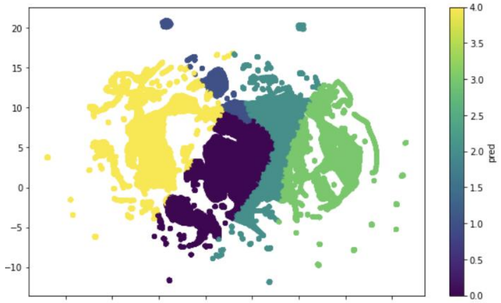
	0	1	2	3	4	5	6
0	-4.94113	1.11538	0.421675	-2.34292	6.821476	-2.23744	0.69725
1	-5.2567	-0.14677	0.247434	-2.16184	7.115843	-1.99986	0.011408
2	-4.90289	-1.07969	0.184251	-3.4565	8.362855	-1.96312	-0.69158
3	-6.68522	0.123215	-2.0331	-1.45979	8.261654	-2.1828	0.944091
4	-4.8755	0.938143	0.276074	-3.21209	7.153507	-5.07487	3.105699
5	-7.93075	-0.46284	0.599881	-7.24127	8.366812	-1.30827	0.313276
6	-8.74983	-0.61387	4.115621	-11.1669	8.724603	-1.63894	1.455382
7	-9.35746	-1.52443	2.371141	-8.35019	8.07953	-1.86892	2.677287
8	-5.24269	-1.74407	0.229054	-1.45792	6.340788	0.300455	-1.42579
9	-3.95401	-1.70854	-0.4278	-2.1444	7.123334	1.90696	-2.25275
10	-4.10868	-1.38228	0.130805	-2.06737	6.853902	0.859051	-3.1119
11	-7.25444	-0.85177	4.273735	-8.38797	9.984344	1.284606	-3.58988

Batch size:32, Cluster:5, Umap:3

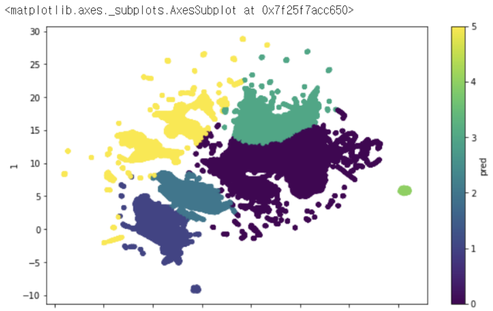
Clustering



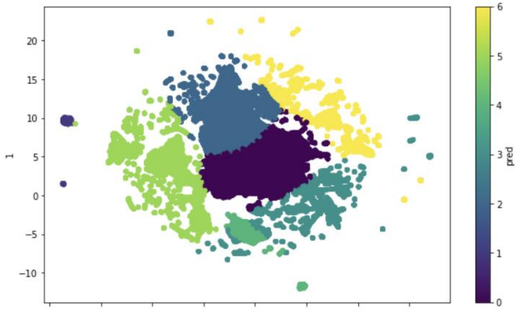
Batch size:32, Cluster:4, Umap:2



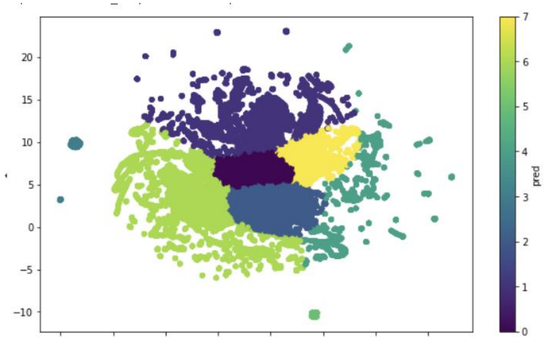
Batch size:32, Cluster:5, Umap:2



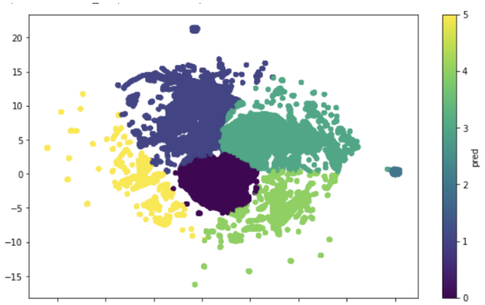
Batch size:32, Cluster:6, Umap:2



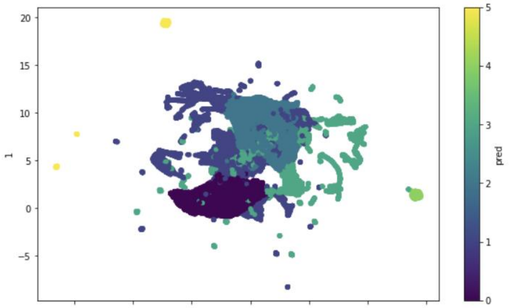
Batch size:32, Cluster:7, Umap:2



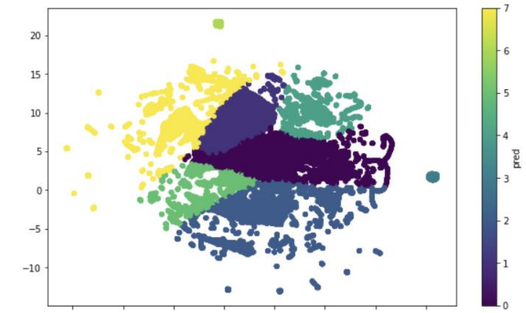
Batch size:32, Cluster:8, Umap:2



Batch size:64, Cluster:6, Umap:2

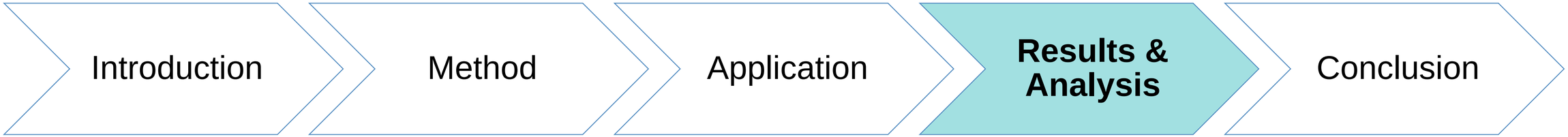


Batch size:64, Cluster:6, Umap:3



Batch size:64, Cluster:8, Umap:2

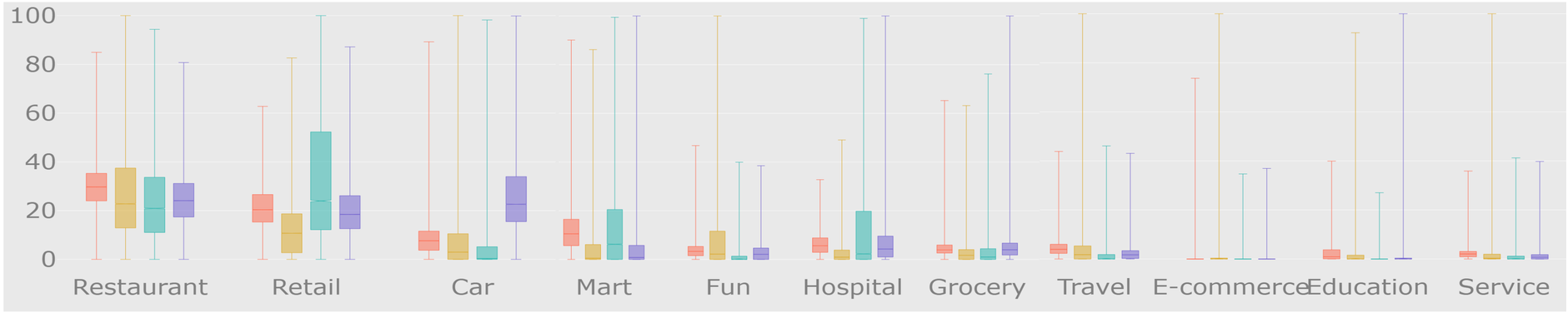
Contents



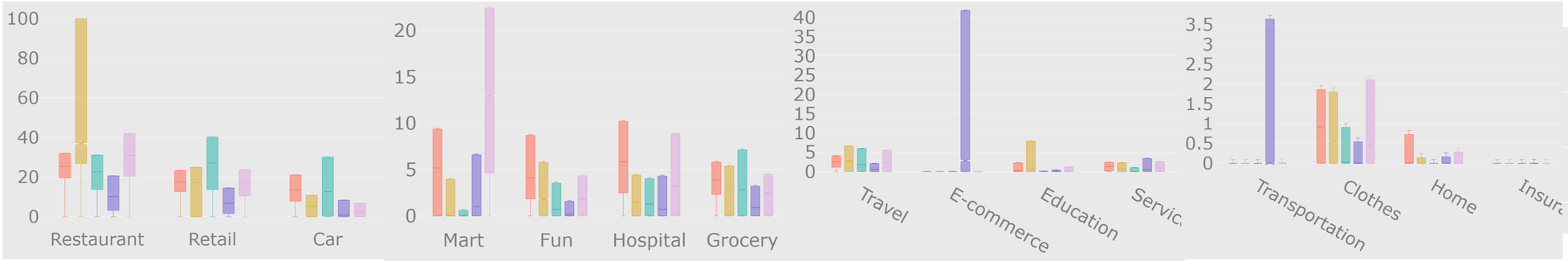
- Clustering results
- Further analysis
  - User characteristics
  - Cluster transitions
  - Monthly changes

Boxplot

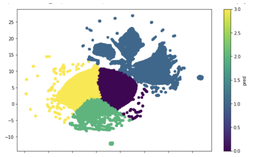
- For **capturing clusters' characteristics**
- Max cap = Q3 + 0.1
  - Changes only Max, not others (Min, Q1, median, Q3)



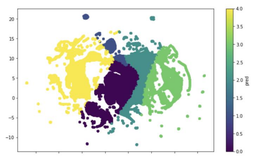
Max cap



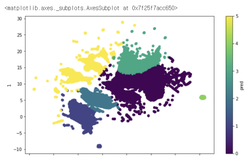
Clustering results



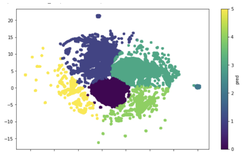
Batch size:32,
Cluster:4,
Umap:2



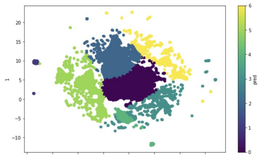
Batch size:32,
Cluster:5,
Umap:2



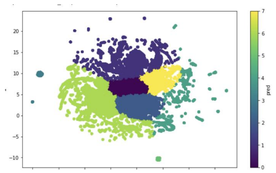
Batch size:32,
Cluster:6,
Umap:2



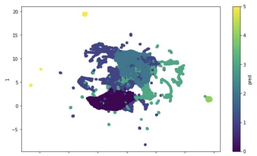
Batch size:64,
Cluster:6,
Umap:2



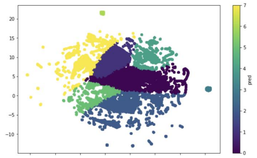
Batch size:32,
Cluster:7,
Umap:2



Batch size:32,
Cluster:8,
Umap:2



Batch size:64,
Cluster:6,
Umap:3



Batch size:64,
Cluster:8,
Umap:2

Result 1
- Cluster: 5
- Umap: 2

Result 2
- Cluster: 7
- Umap: 3

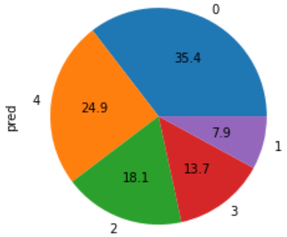
Result 3
- Cluster: 8
- Umap: 3

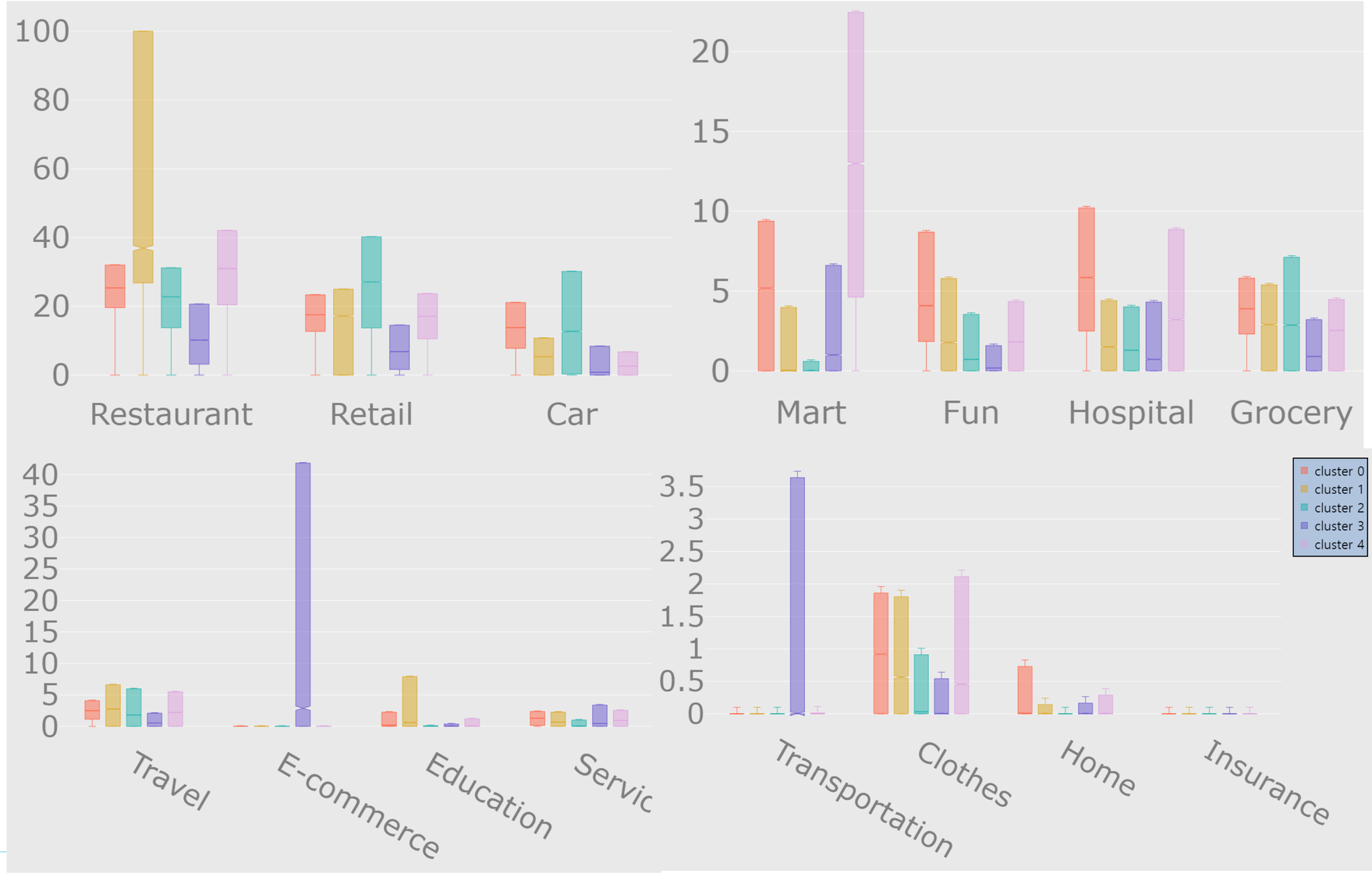
Clustering results

- **5 clusters**
- Clustered on 2-D (UMAP)

- Characteristics
  - 0: **Hospital**, **Home**
  - 1: **Restaurant, Education**
  - 2: **Retail**
  - 3: **e-commerce , Transportation**
  - 4: **Restaurant, Retail, Mart**

- Proportions
  - 0: 95,235
  - 1: 21,315
  - 2: 48,597
  - 3: 36,710
  - 4: 66,871
  - Total: 268,728



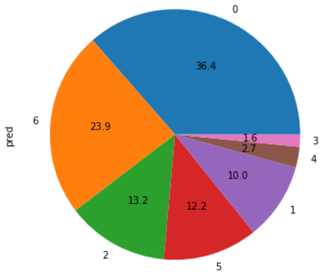


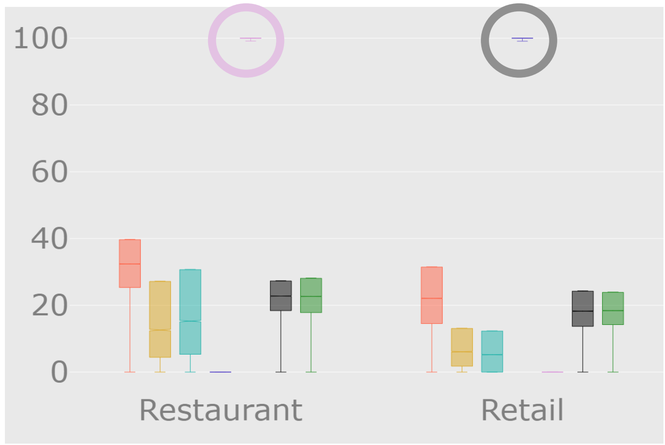
Clustering results

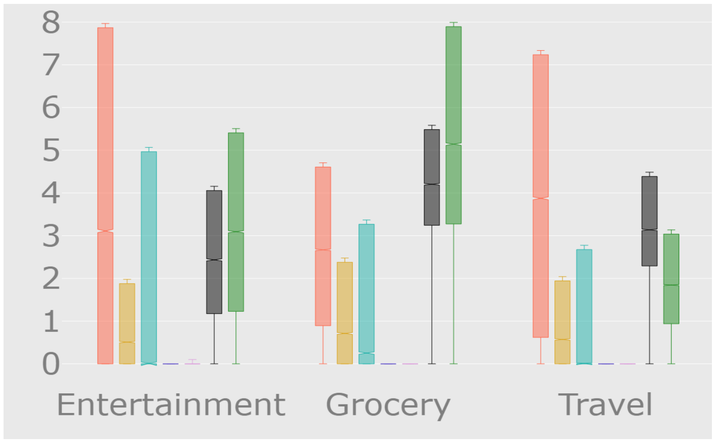
- **7 clusters**
- Clustered on 3-D (UMAP)

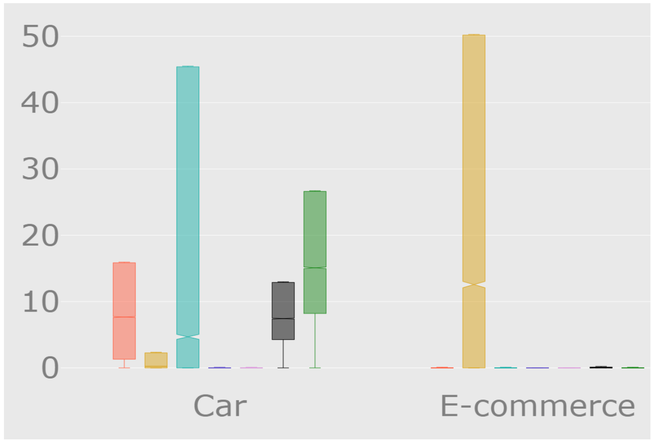
- Characteristics
  - 0: **Restaurant, Retail**
  - 1: **e-commerce, Insurance**
  - 2: **Car, Transportation**
  - 3: (Retail 100%)
  - 4: (Restaurant 100%)
  - 5: **Education, Travel**
  - 6: **Car, Hospital, Grocery**

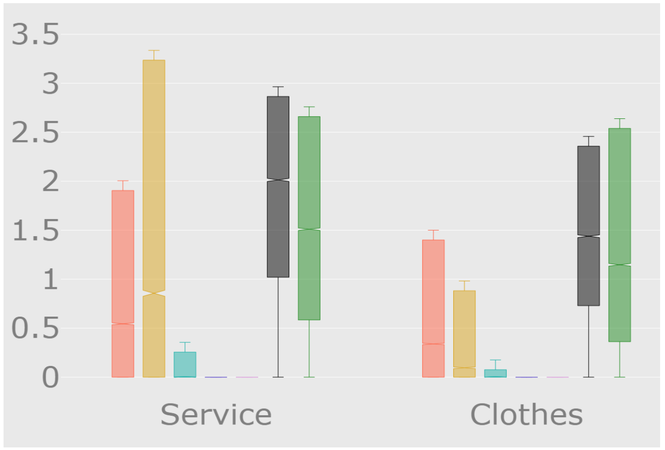
- Proportions
  - 0: 97912
  - 1: 26784
  - 2: 35473
  - 3: 4238
  - 4: 7195
  - 5: 32811
  - 6: 64315
  - Total: 268,728

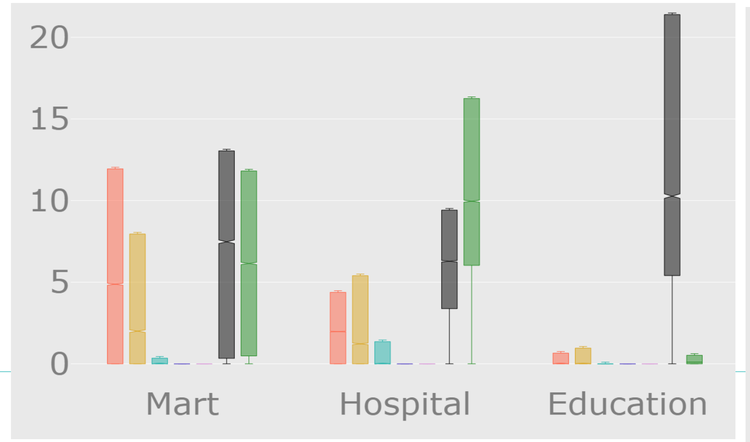


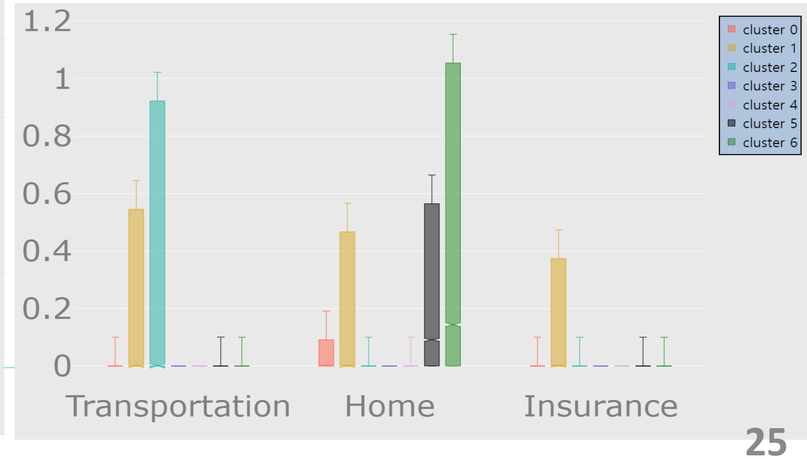










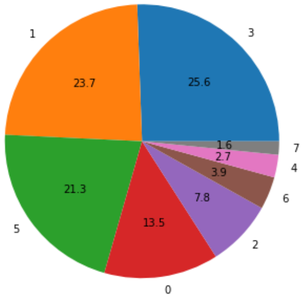


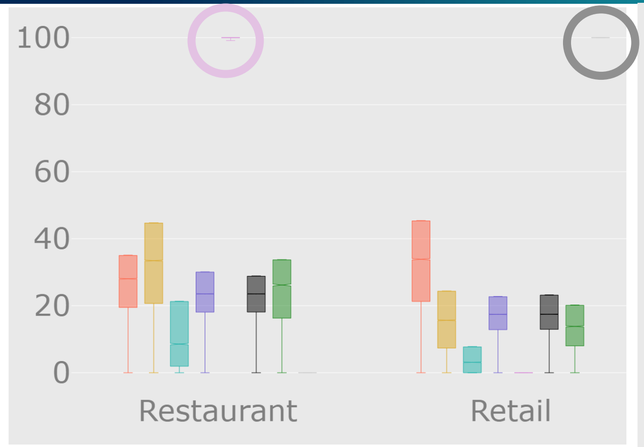
Clustering results

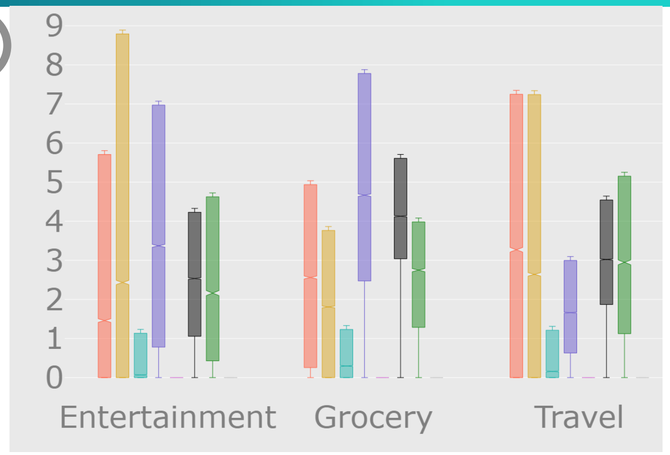
- **8 clusters**
- Clustered on 3-D (UMAP)

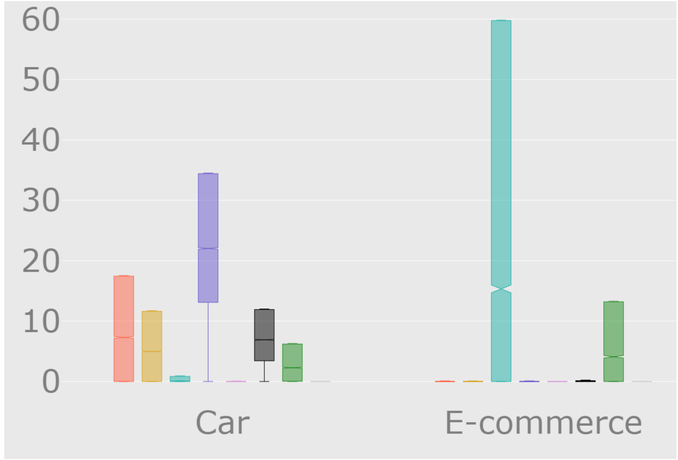
- Characteristics
  - 0: **Restaurant**, **Retail**
  - 1: **Restaurant**
  - 2: **e-commerce**, **Insurance**
  - 3: **Car**, **Grocery**
  - 4: (Restaurant 100%)
  - 5: **Hospital**, **Education**
  - 6: **Restaurant**, **Transportation**
  - 7: (Retail 100%)

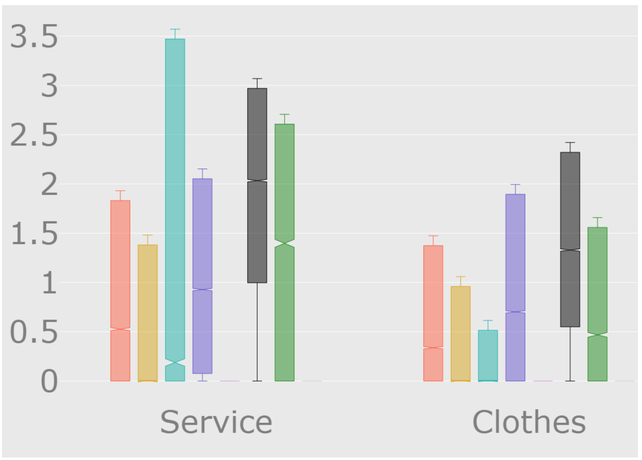
- Proportions
  - 0: 36245
  - 1: 63624
  - 2: 21050
  - 3: 68695
  - 4: 7196
  - 5: 57314
  - 6: 10370
  - 7: 4234
  - Total: 268,728

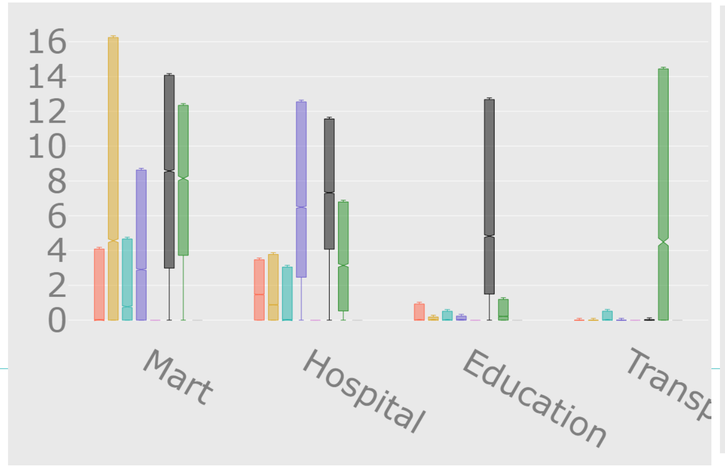


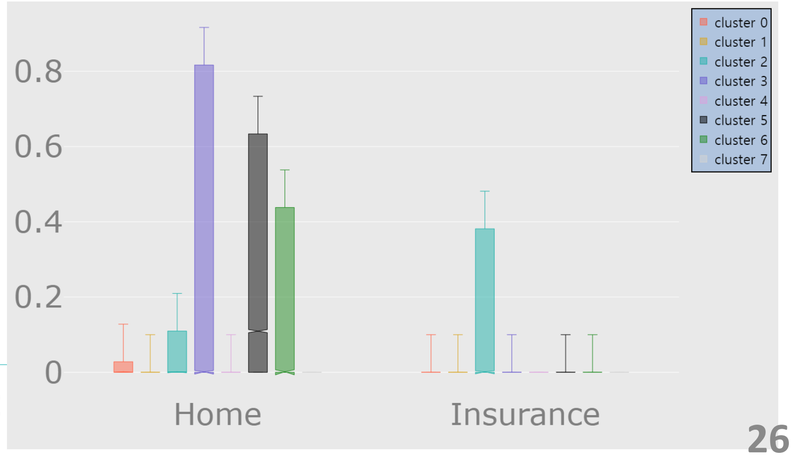




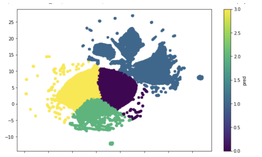




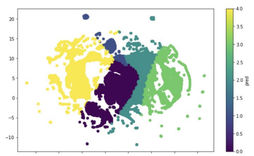




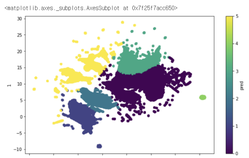
Clustering results



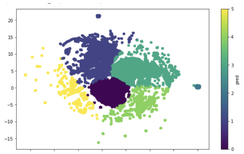
Batch size:32,
Cluster:4,
Umap:2



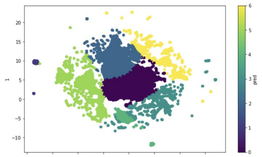
Batch size:32,
Cluster:5,
Umap:2



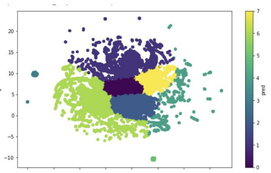
Batch size:32,
Cluster:6,
Umap:2



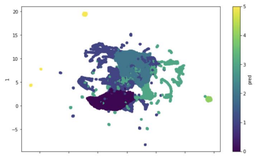
Batch size:64,
Cluster:6,
Umap:2



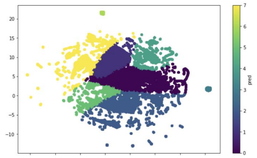
Batch size:32,
Cluster:7,
Umap:2



Batch size:32,
Cluster:8,
Umap:2



Batch size:64,
Cluster:6,
Umap:3



Batch size:64,
Cluster:8,
Umap:2

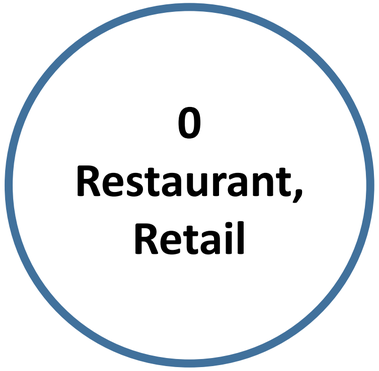
Result 1
- Cluster: 5
- Umap: 2

Result 2
- Cluster: 7
- Umap: 3

Result 3
- Cluster: 8
- Umap: 3

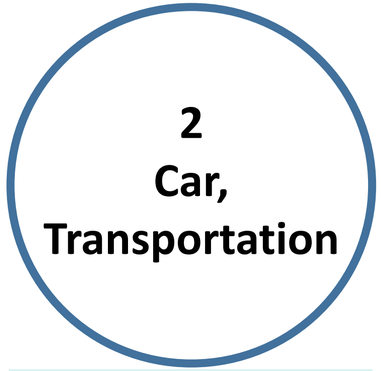
Further analysis
- User characteristics
- Cluster transitions
- Monthly changes

User characteristics



Man 70%
2030s 50%

4050s 42%
High income 46%

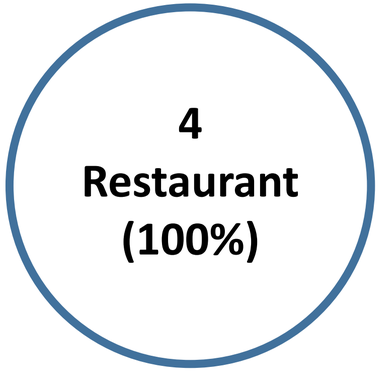


Man 68%
4050s 44%
High income 54%

2030s 57%
High income 57%

Woman 74%
60s+ 0.7%
High income 6.6%

2030s 7.3%
High income 12%



High income 69%

User characteristics

Cluster	Characters	Gender	Age	Income
0	**Restaurant, Retail**	**Man**	**2030s**	**-**
1	**E-commerce, Insurance**	**-**	**4050s**	**High**
2	**Car, Transportation**	**Man**	**4050s**	**High**
3	(Retail 100%)	-	2030s	High
4	(Restaurant 100%)	-	-	High
5	**Education, Travel**	**Woman**	**2030~4050s**	**Low**
6	**Car, Hospital, Grocery**	**-**	**4050~60s+**	**Low~Medium**

Cluster transitions

- **Making transitions**
  - **User groups**
    - Gender (Man/Woman)
    - Age (20/30/40/50/60+)
    - Income (1~11)
    - Region (207)
    - 22,770 combinations in total
  - **State**
    - Cluster belongings
  - **Transitions**
    - Month to month transition
    - e.g., cluster 1 -> cluster 4
    - 249,511 transitions in total

[transitions from Jan 2019 to April 2020]


```
{'male_20s or less_B2_가평군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 5, 0, 0],
 'male_20s or less_B2_강남구': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
 'male_20s or less_B2_강동구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_강릉시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_강북구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_강서구': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
 'male_20s or less_B2_강진군': [5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_강화군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_거제시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_거창군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_경산시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_경주시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_계룡시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_계양구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_고령군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_고성군': [0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_고양시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_고창군': [5, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0],
 'male_20s or less_B2_고흥군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_공주시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_과천시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_관악구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_광명시': [0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_광산구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_광양시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_광주시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_광진구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_괴산군': [5, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_구례군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_구로구': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
 'male_20s or less_B2_구리시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_구미시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_군산시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_군포시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_금산군': [0, 5, 5, 0, 0, 0, 5, 5, 0, 0, 0, 0, 0, 5, 0, 0],
 'male_20s or less_B2_금정구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_금천구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_기장군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 'male_20s or less_B2_김제시': [0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0],
```


[Cluster Characteristics]

0: **Restaurant, Retail**
1: **e-commerce, Insurance**
2: **Car, Transportation**
3: (Retail 100%)
4: (Restaurant 100%)
5: **Education, Travel**
6: **Car, Hospital, Grocery**

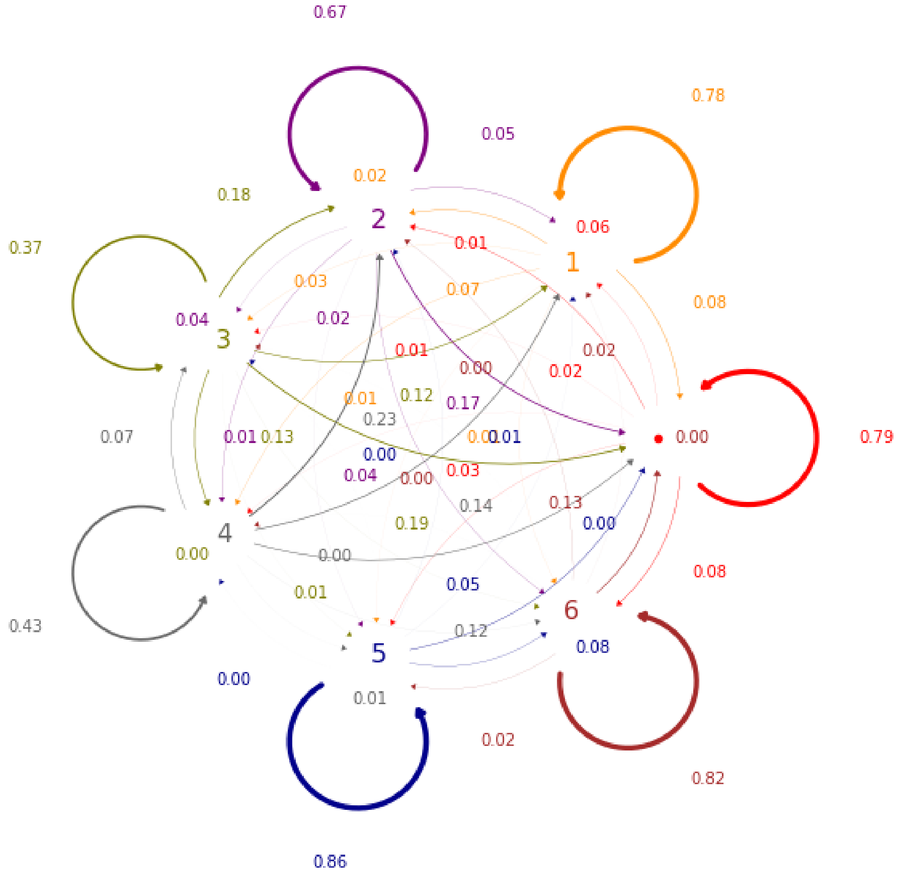
Cluster transitions

- **Findings**
  - **Cluster 3,4**
    - High probabilities to move to other clusters in next month
    - If a customer spends entirely on either Retail or Restaurant, it is likely that he will spend on other categories in the next month.
  - **Cluster 5,6**
    - High probabilities to stay in next month
    - Customers who show this patterns of spending are likely to keep showing the same patterns in the next months.

[one-step transition matrix]

next_state / state	0	1	2	3	4	5	6
0	0.792627	0.021576	0.060737	0.007808	0.007567	0.025338	0.084347
1	0.082670	0.779505	0.071144	0.016315	0.033360	0.006169	0.010836
2	0.168287	0.054379	0.673119	0.019086	0.041231	0.006422	0.037475
3	0.187972	0.122603	0.176932	0.374782	0.128704	0.003196	0.005811
4	0.116605	0.142499	0.227030	0.074507	0.428834	0.003341	0.007183
5	0.076394	0.004736	0.006650	0.000487	0.000616	0.863529	0.047588
6	0.130612	0.004588	0.020528	0.000515	0.000947	0.021990	0.820820

[one-step transition diagram]

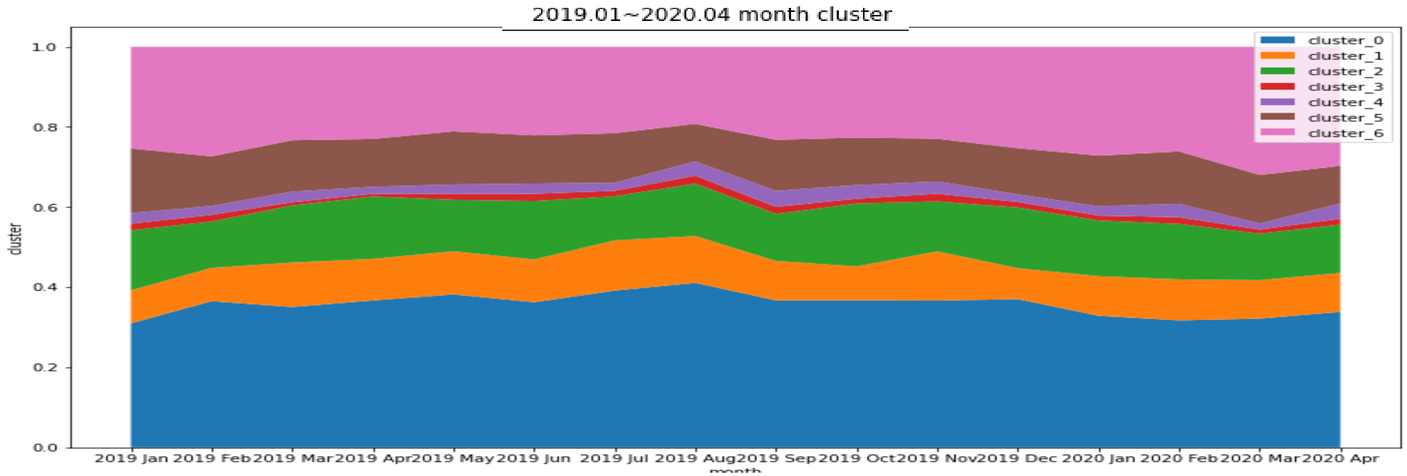


[Cluster Characteristics]

0: **Restaurant, Retail**
1: **e-commerce, Insurance**
2: **Car, Transportation**
3: (Retail 100%)
4: (Restaurant 100%)
5: **Education, Travel**
6: **Car, Hospital, Grocery**

Monthly changes

Batch size:64, Cluster:7, Umap:3

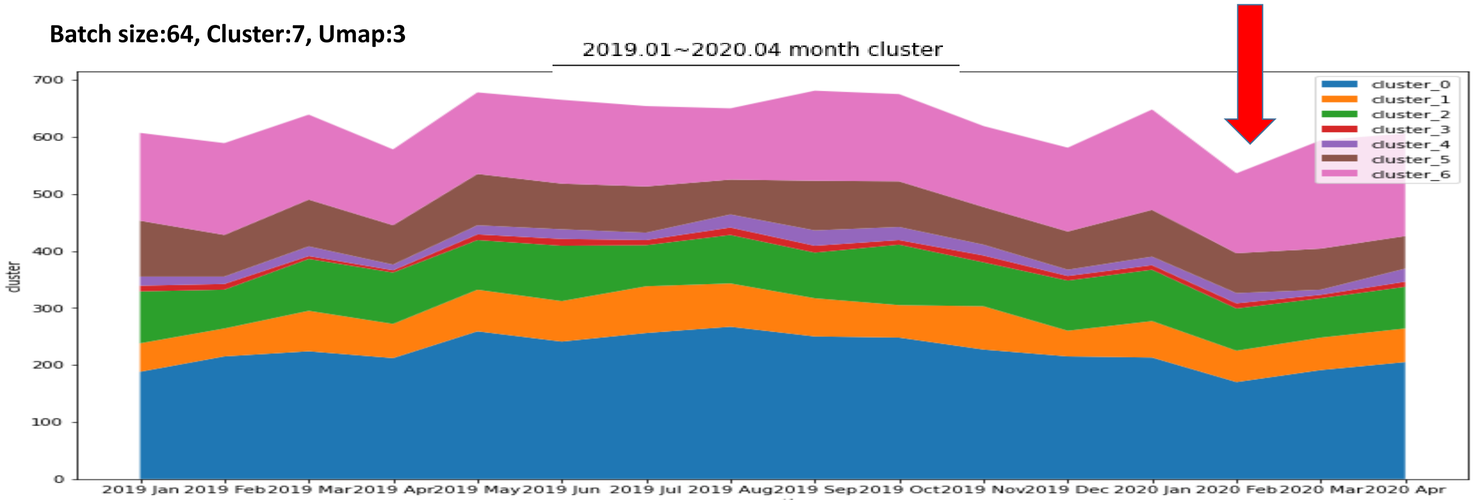


- **The percentage of clusters by month**

2019 Jan [Cluster_0 : 40%, Cluster_1 : 15%, …]
~
2020 Apr [Cluster_0 : 25%, Cluster_1 : 50%, …]

- There was **no significant difference in the proportions by month.**

Monthly changes



- **Clusters amount by month**

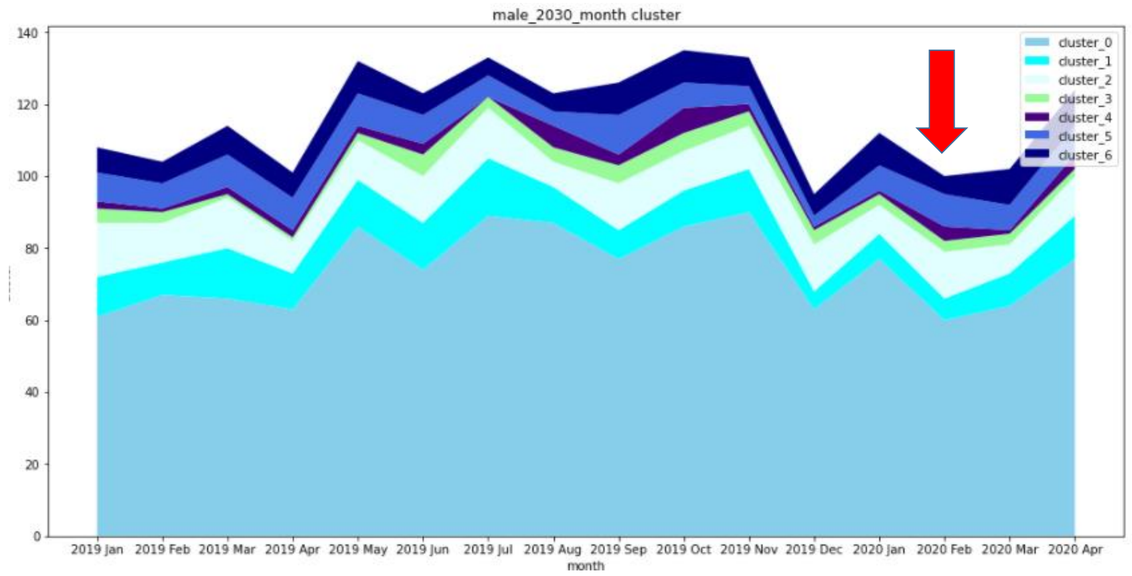
2019 Jan [Cluster_0 : 300, Cluster_1 : 25, …]
~
2020 Apr [Cluster_0 : 268, Cluster_1 : 70, …]

- In **February 2020**, when the corona spread rapidly, we could see that the **cluster was decreasing overall.**

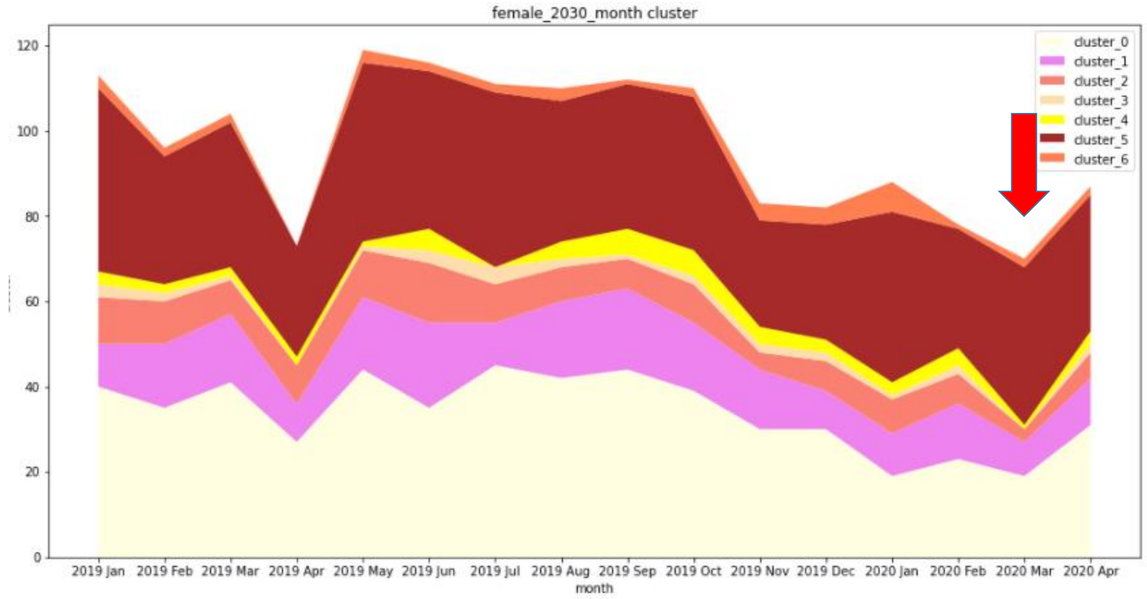
- It seems that this is more characteristic than the visualization of percent, so we also analyzed by gender & age.

Monthly changes – 2030s

Male Age : 20 ~ 30, cluster : 7



Female Age : 20 ~ 30, cluster : 7

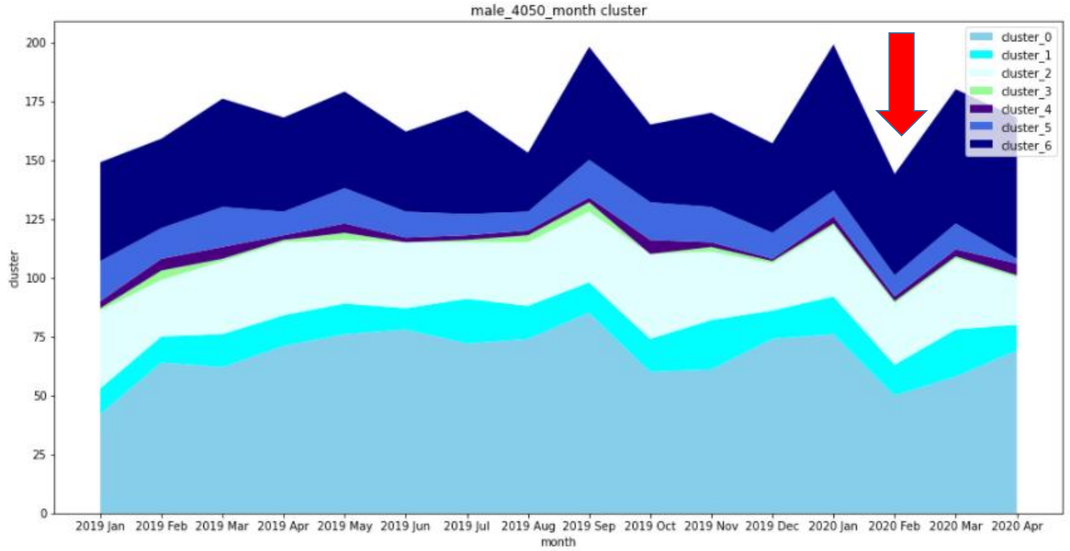


- Men mainly have high cluster_0 (restaurant, retail) for every month
- Women have high cluster_5 (travel, education) and cluster_0, and **after December 2019, cluster_5 decreases**.

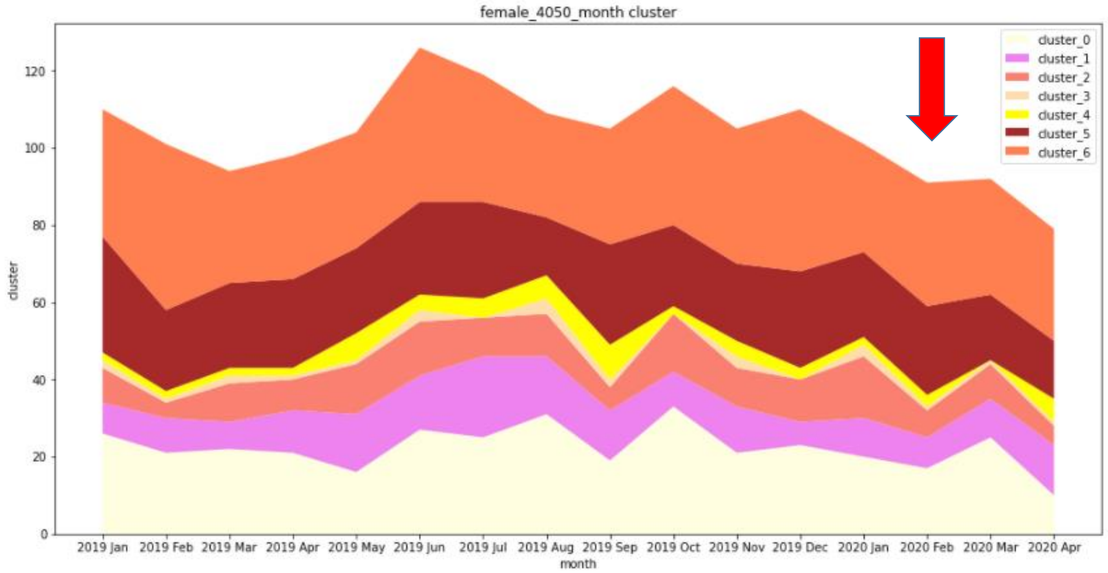
- These two graphs are relatively **similar**. **Consumption decreased around April 2019**, and it shows a sharp **decline due to the COVID-19 in February-March 2020**.

Monthly changes – 4050s

Male Age : 40 ~ 50, cluster : 7



Female Age : 40 ~ 50, cluster : 7

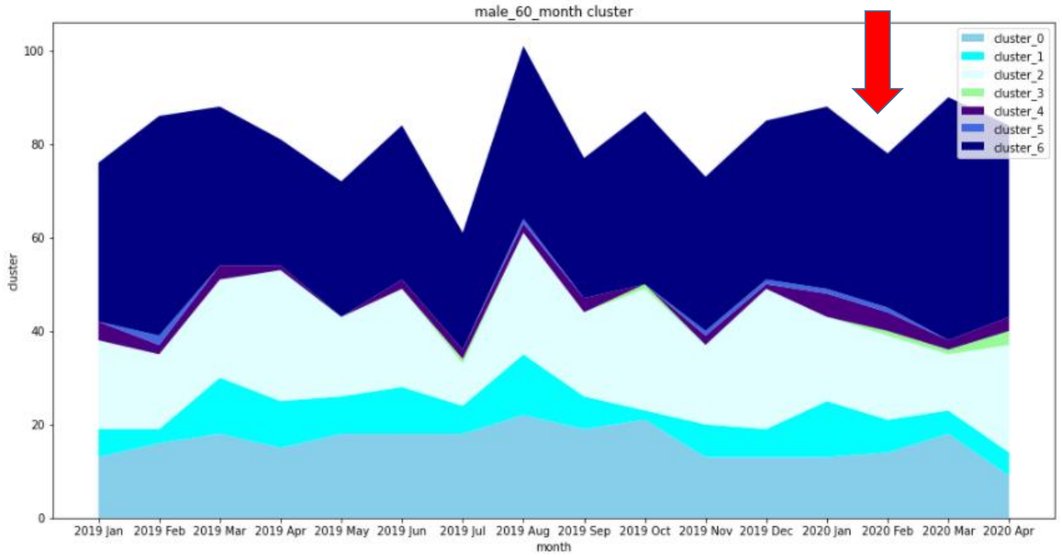


- People in their 40s and 50s show a similar pattern, with consumption going up until February 2020 and then **going down as they experience the corona virus.**

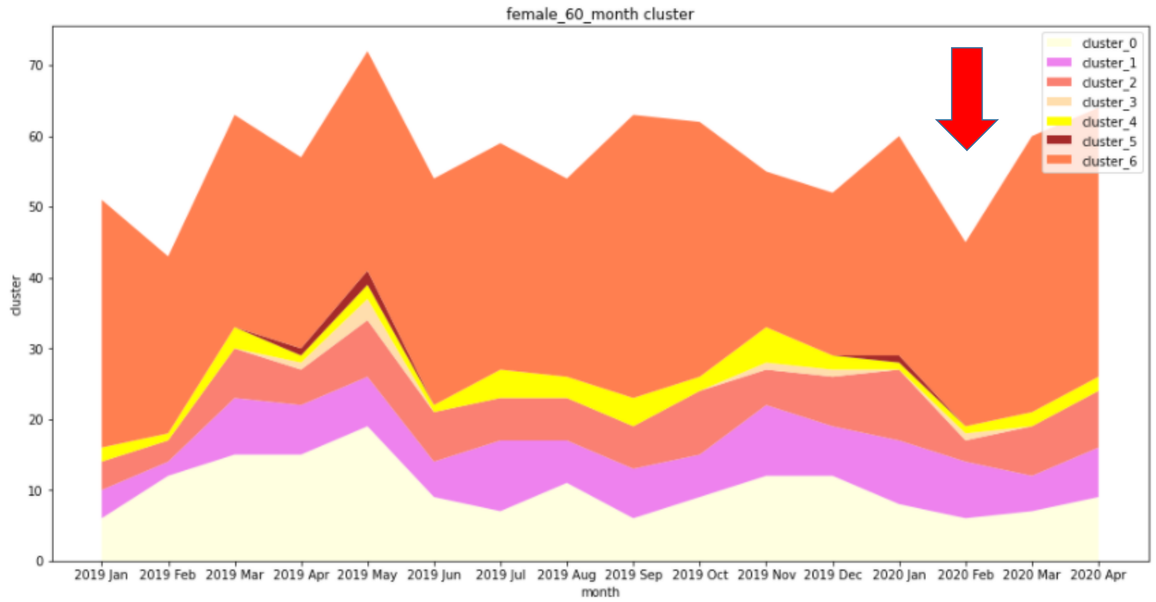
- Unlike men, **women did not appear to significantly recover** their consumption activity after February 2020.

Monthly changes – 60s+

Male Age : Over 60, cluster : 7



Female Age : Over 60, cluster : 7



- Cluster_6 (hospital,grocery,car) increases rapidly after the age of 60 and the consumption **pattern is irregular**, unlike the lower age group.

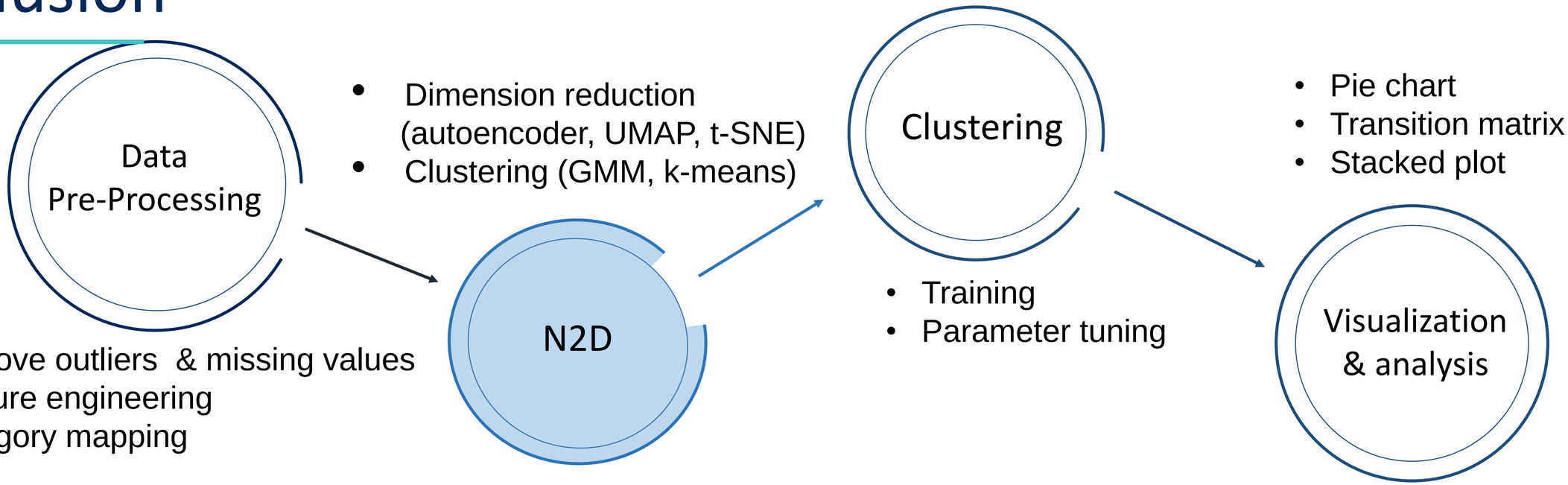
- **Relatively small affect on consumption by COVID-19**

Contents

Introduction → Method → Application → Results & Analysis → **Conclusion**

- Summary
- Limitations
- Further works

Conclusion



- ❖ Limitation
  - **Data with short period**
  - **Manual category selection** and subjective interpretation
- ❖ Follow-up research
  - Find **monthly consumption patterns** when we get more **long-term data** from more card companies.