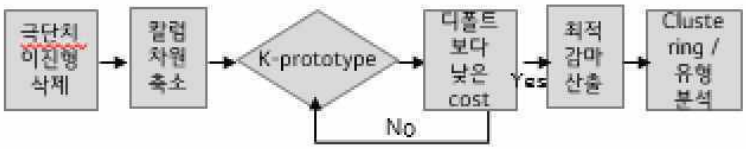
[redacted]

최종과제 명 : 설문 조사 데이터와 비식별 데이터의 결합을 통한[redacted] 구매 고객 유형 분석
- 수치형과 카테고리 혼합 자료에 대한 클러스터링 성능 향상 방법 개발을 중점으로

① 수행 업무

- 구매 제품에 대한 고객 응답 데이터와 라이프스타일 데이터를 결합, 분석하여 고객 특성 파악 및 인사이트 도출
- 기존 클러스터링 방법의 한계를 극복하기 위한 분석 방법론 개발 및 효용성 입증



- 프로젝트 진행 일정

진행 과제	1주차	2주차	3주차	4주차
데이터 파악 및 EDA				
분석 환경 구축				
칼럼 차원 축소 (PCA, MCA, MDS)				
K-prototype 클러스터링				
클러스터링 평가				
최적의 가중치 감마 찾기				
클러스터링 / 고객 유형 분석				
발표자료 작성 및 준비				

② 배운 점

- **실제 데이터 분석 수행을 통한 의미 있는 결과 도출 및 문제 해결 능력 향상**
  - ✓ 혼합형 고차원 데이터를 사용하여 고객 유형 분석을 함으로써 유의미한 인사이트를 발굴해 낼 수 있는 분석 경험을 해보았습니다.
  - ✓ 이에 이후 다른 형태의 데이터와 결합한 혼합형 데이터를 사용하게 될 때 현 프로젝트에서 경험한 분석 방법론 및 알고리즘을 응용해서 사용할 수 있습니다.
- **유의미한 결과를 위한 데이터의 다각적 이해**
  - ✓ 기존의 각각의 데이터에서 얻을 수 없었던 유의미한 인사이트를 cns의 비식별 [redacted]데이터에 서베이 데이터를 결합한 혼합형 고차원 데이터를 통해서 도출해 낼 수 있었습니다.
  - ✓ 이후 혼합형 고차원 데이터가 존재할 때 위 과정처럼 다양한 방법론을 적용해보면서 클러스터링의 성능을 높이고 유의미한 인사이트를 얻을 수 있습니다.
- **코딩 능력 향상**
  - ✓ 현업 데이터를 통해 다각적으로 방법론을 적용해 보면서 실무에서 사용하는 분석 언어를 경험할 수 있었습니다.
  - ✓ [redacted]다 이해하기 쉽게 코딩하려고 노력하여 이후에 보다 정확하면서도 파악하기 쉬운 코드를 짤 수 있을 것입니다.

[redacted]

최종과제

③ 최종과제 명 : 설문 조사 데이터와 비식별 데이터의 [redacted] 구매 고객 유형 분석
- 수치형과 카테고리 혼합 자료에 대한 클러스터링 성능 향상 방법 개발을 중점으로

➢ 분석 목표

로봇 청소기 구매 고객 유형을 발굴하는데 유용하게 사용되는 k-prototypes 클러스터가 이진형 및 카테고리형 항목이 많을 때 특정 데이터에 치우치는 문제를 개선하고 클러스터의 성능을 높이고자 함.

➢ 분석 데이터

- LG 전자에서 실시한 [redacted] 구매에 관한 설문지 데이터 및 cns가 보유한 [redacted] 지수 데이터
  (대상 : 주요 가전을 구매한 [redacted] 고객)
- 데이터 칼럼이 많고 특히 이진형 항목이 많음
- 데이터 형태 :

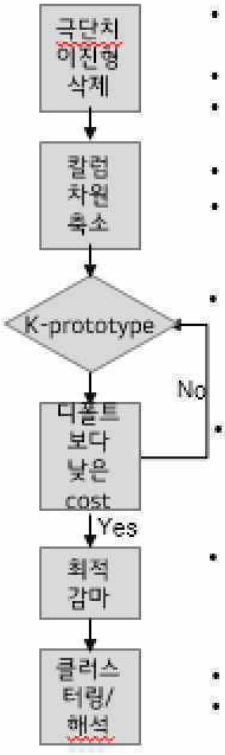
연속형 칼럼	이진형 칼럼	범주형 칼럼
[redacted]	[redacted]	[redacted]

연속형 칼럼 : UTI 지수 데이터
이진형 칼럼 : [redacted] 질문에 대한 예/아니요
범주형 칼럼 : 직업군, 연령 등의 범주형 설문지

- 극단치 이진형 제거 후

연속형 칼럼	이진형 칼럼	범주형 칼럼
[redacted]	[redacted]	[redacted]

➢ 분석 과정



- 극단치 이진형 : 예/아니요라고 대답한 사람이 95% 이상인 이진형 칼럼
- 칼럼 제거했을 때 k-prototype cost가 더 낮음
- 이진형 데이터의 차원 축소가 더 잘됨
- PCA, MCA로 연속형, 이진형 차원 정도 확인
- Linear MDS, non-linear MDS로 차원 축소
- 그리드 서치 방법으로 여러 가중치 감마 값 넣어서 클러스터링 진행 (cluster : elbow graph 확인)
- 감마 [redacted] cost가 디폴트 감마를 사용한 k-prototype cost 보다 작은 것 발견
- 그 중에서 Calinski-Harabasz index와 Davies bouldin score를 사용하여 클러스터링 평가
- 찾은 최적의 감마 [redacted]로 클러스터링
- 각 클러스터 해석 후 마케팅 솔루션

[redacted]

최종과제

최적의 가중치로 k-prototypes을 사용한 로봇 청소기 구매 고객 유형 결과

	명	우세 연령	우세 성별 *	소득이 줄어도 소비를 줄이지 않을 항목 *	소득이 줄면 소비를 줄일 항목 *	우세하게 보유한 가전 제품 *	구매경로	구매 시 고민 기간
Cluster1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted] 내외
Cluster2	[redacted]	-	[redacted]	[redacted]	[redacted]	-	[redacted]	[redacted] 내외
Cluster3	[redacted]	-	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Cluster4	[redacted]	[redacted]	[redacted]	-	[redacted]	-	[redacted]	[redacted] 내외

-주요 연령층 : 20-40대 * 다른 군집 대비

➢ 해석

- Cluster1 : 쾌락형 라이프 스타일
  - ✓ 주로 [redacted] 등 자신의 외면을 가꿈
  - ✓ 최신 트렌드의 컨텐츠 소비하며 즐길 수 있는 서비스 중요.
  - ✓ 구매 고려 기간이 [redacted]
  - ✓ 득히 [redacted]대의 비율이 가장 높은 군집

- Cluster2 : 건강형 라이프 스타일
  - ✓ 주로 청소나 수납, 정리 등을 중요하게 여김.
  - ✓ 자신의 [redacted] 등을 소비

- Cluster3 : 합리형 라이프 스타일
  - ✓ 꼭 [redacted]지 않은 것에 대한 소비의 중요성이 낮음
  - ✓ 소득이 줄 때 소비를 줄일 수 있는 항목이 가장 많음.
  - ✓ [redacted]
  - ✓ 구매 고민 기간이 [redacted]

- Cluster4 : 눈에 띄는 소비 패턴이 없고 [redacted]

➢ 결론

- Cluster2는 [redacted]에 이미 관심있는 군집, Cluster4는 소비에 관심이 [redacted] 패턴이 [redacted] 군집이므로 타게팅할 군집에서 제외
- **주요 타겟팅 할 군집 파악 : 쾌락형 / 합리형 라이프 스타일**
- 각 라이프 스타일에 맞는 마케팅 솔루션 제공 :

	솔루션
쾌락형	- 전통 대중매체 보다는 최신 트렌드를 반영한 SNS 마케팅 - 이성적이면서도 감성적인 소비를 강조하는 경험 마케팅 - 구매 고민 기간이 더 길어지지 않도록 유효기간이 있는 쿠폰 제공
합리형	- 가전 제품에 대한 관심도가 높기 때문에 [redacted] 합리성을 강조한 마케팅을 진행 - 제조사 브랜드 홈페이지에서 [redacted] 일어났기 때문에 홈페이지 내에서 사용할 수 있는 전용 쿠폰을 제공 - 구매 고민 기간이 [redacted] 때문에 구매 이후 지원 서비스 기간 확대 등

[redacted]

최종과제

고객 유형을 위한 k-prototypes 클러스터링 성능 향상 방법론 구축 과정 1) 차원 축소

➢ K-prototype 알고리즘

- 연속형과 범주형이 혼합된 데이터를 군집화 하는 방법
- 연속형 속성을 군집화 하는 k-means 알고리즘과 범주형 속성을 군집화 하는 k-modes 알고리즘 결합
- **범주형 속성 거리에 가중치 감마를 주어 연속형 속성 거리와 합한 것이 k-prototype 거리**

$$d_{k-\text{prototypes}} = d_{k-\text{means}} + \gamma\, d_{k-\text{modes}}$$

$$= d_{\text{eudid}}(X_i, C_j)^2 + \gamma\, d_{\text{simple matching}}(X_i, C_j)$$

- 한계점
  1) 클러스터 수(k)는 수동으로 결정.
  2) 범주형 데이터와 연속형 데이터 간의 비율을 조정하는 데 사용되는 감마를 수동으로 결정.

➢ 차원 확인

- 연속형 데이터
  - PCA로 차원 확인
  - 주성분 2개로 연속형 데이터의 97.85% 설명
- 이진형 데이터
  - MCA로 차원 확인
  - 주성분 22개로 이진형 데이터의 81.47% 설명

➢ MDS 차원 축소

- MDS란?

  1. linear MDS
     - 데이터가 연속형 변수인 경우 사용.
     - 각 개체들 간의 유클리드 거리 행렬을 계산하고 개체들 간의 비유사성을 공간상에 표현.

  2. Non-linear MDS
     - 데이터가 범주형 변수인 경우 사용.

- MDS 결과를 clustering에 적용 시킬 수 있는가?
  - MDS는 시각화에 확실히 유용할 수 있지만 시각화 도구일 뿐만 아니라 일반적인 차원 감소 및 잠재 변수 모델에 유용할 수 있음.
  - 몇몇의 MDS를 통한 차원 축소 후 clustering을 사용한 연구 과정을 확인할 수 있음.
- MDS 차원 축소
  - 연속형 데이터 : 2개로 차원 축소
  - 이진형 데이터 : 22개로 차원 축소

➢ 가장 성능이 좋은 차원 축소 방법 선택 (cluster4 기준)

차원 축소 안했을 때 :723560,46

	연속형 칼럼만 축소		이진형 칼럼만 축소	연속형, 이진형 칼럼 축소	
	pca	mds	nmds	Pca+nmds	Mds+nmds
cost	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

	연속형 칼럼만 축소	연속형, 이진형 칼럼 축소
	pca	Mds+nmds
Calinski harabasz score	[redacted]	[redacted]

- k-prototype의 cost는 비슷하지만, calinski_harabasz_score를 보면 이진형을 축소했을 때 성능이 더 좋아짐

[redacted]

최종과제

고객 유형을 위한 k-prototypes 클러스터링 성능 항상 방법론 구축 과정 2) 최적 가중치 산정

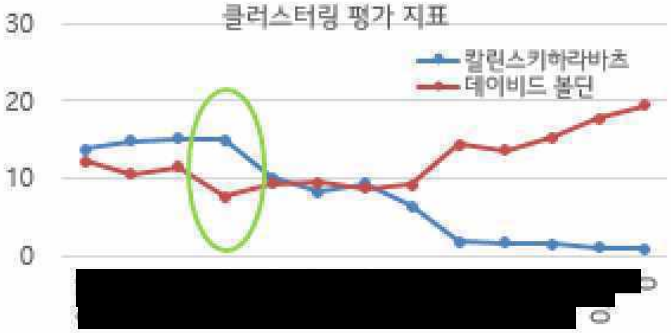
➢ 그리드 서치 방법으로 K-Prototype 클러스터링 최적 가중치 산정

A. 감마 설정 없이 디폴트 값으로 :
- 해당 데이터의 디폴트 가중치 감마 값 = [redacted]
- Cluster 4개 선택, cost는 1[redacted]

B. 감마 설정하여 디폴트 값보다 낮은 cost 가진 감마

가중치	cluster	Calinski harabasz	Davies bouldin	cost
[redacted]				

C. 디폴트 값보다 낮은 cost 가진 감마 중 최적의 클러스터링 평가 지표를 가진 가중치 선택



➢ 결론

다음을 만족시키는 값이 최적 가중치 감마

1. 디폴트 값보다 낮은 cost 가중치
2. Calinski-Harabasz Index가 높고, Davies bouldin score가 가장 낮은 가중치
3. 각각의 cluster 내의 고객의 수가 [redacted]명 이상인 가중치 (LG 전자 요청)

[redacted]에서 최적 가중치 감마 값은 [redacted]

[redacted]

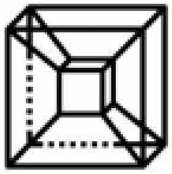
최종과제

④ 결론

혼합형 고차원 데이터를 군집화 할 때 사용하는 k-prototypes 성능 향상을 위해서는 :

극단치 제거

- 정보를 많이 주지 않는 95% 이상의 극단치 이진형 제거



차원 축소

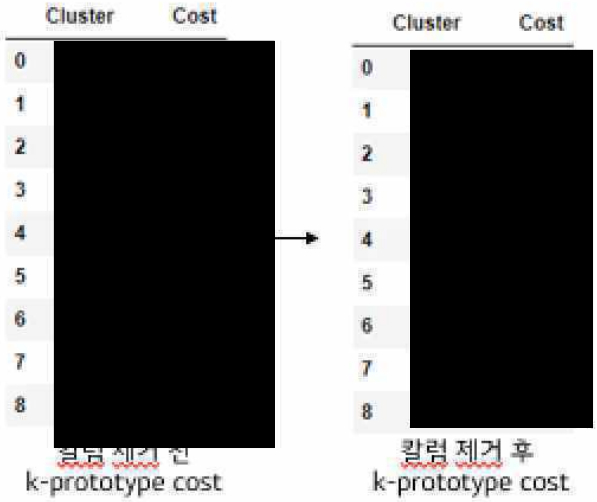
- 다양한 차원 축소 방법 중 최적의 차원 축소 방법을 사용

최적 가중치 산정

- 그리드 서치 방법으로
- 클러스터링 평가 지표 (Calinski harabas, Davies bouldin 등) 사용

별첨 - 95% 이상의 극단치 이진형 칼럼

[redacted] 낮음



- 95% 이상의 극단치 이진형 칼럼에 대해 칼럼이 제공하는 정보가 거의 없다고 판단하여 삭제

칼럼 명	0 (명)	1 (명)	극단치 퍼센트
[redacted]			0.5
			0.992
			0.994
			0.966
			0.976
			0.978
			1
			0.974

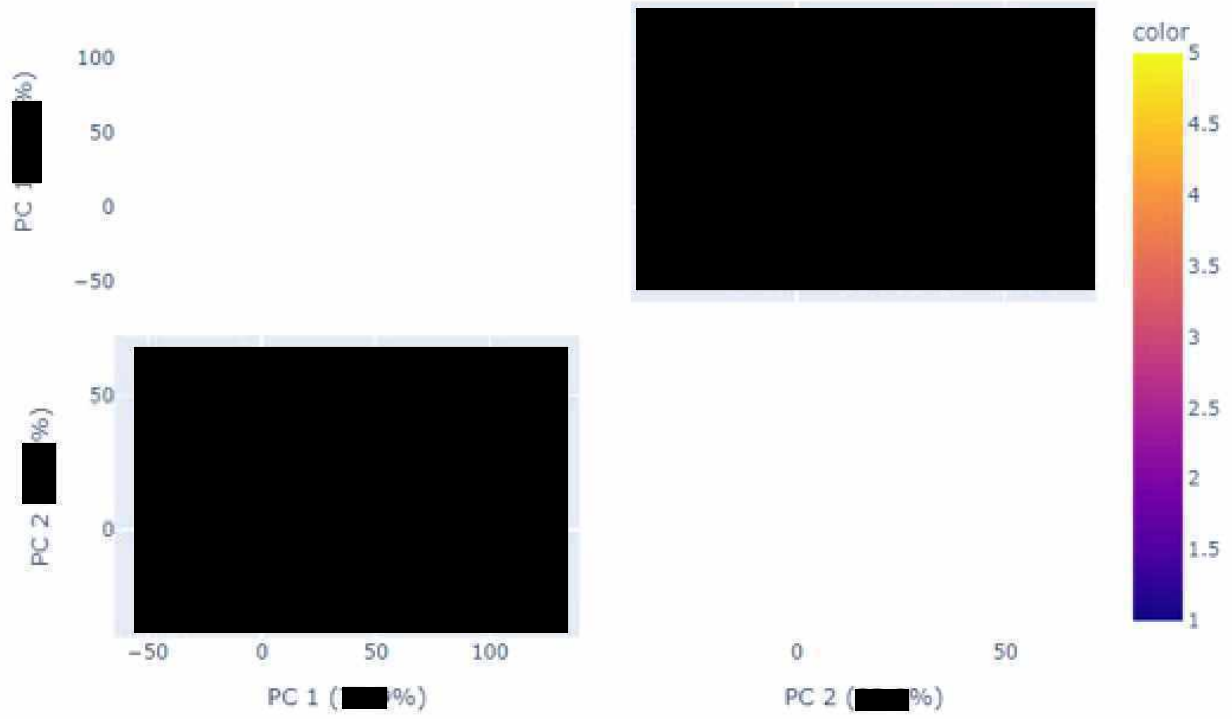
- 이진형 데이터의 차원 축소가 더 잘됨

제거	연속형 차원 크기	설명 정도	이진형 차원 크기	설명 정도
전		약 97%		약 81%
후		약 97%		약 81%

[redacted]

별첨 – PCA를 이용한 연속형 데이터의 저차원 갯수 확인

- PCA로 연속형 데이터 차원 축소 및 차원 확인



[redacted]개 주성분으로 연속형 데이터의 97.85% 설명

[redacted]

별첨 – MCA를 이용한 이진형 데이터의 저차원 갯수 확인

- MCA로 범주형 및 이진형 데이터 차원 축소 및 차원 확인
  - 범주형 데이터 + 이진형 데이터 차원 축소
    - ✓ 주성분 [redacted]로 해도 [redacted] 밖에 설명하지 못함 ➡ 차원 축소가 전혀 안됨
    - ✓ 각각의 index에 주성분을 넣었더니 값이 너무 작아서 NaN값으로 생성됨
  - 이진형 데이터 차원 축소
    - ✓ 차원 축소는 했으나 각각의 index에 주성분을 넣었더니 값이 너무 작아서 NaN값으로 생성됨
    - ✓ MCA를 사용하지 못해도 이진형 차원축소시 22개가 80% 이상을 설명한다는 것은 파악할 수 있음

[redacted]

별첨 – MCA를 이용한 이진형 데이터의 저차원 갯수 확인

- 이진형 데이터 차원 축소

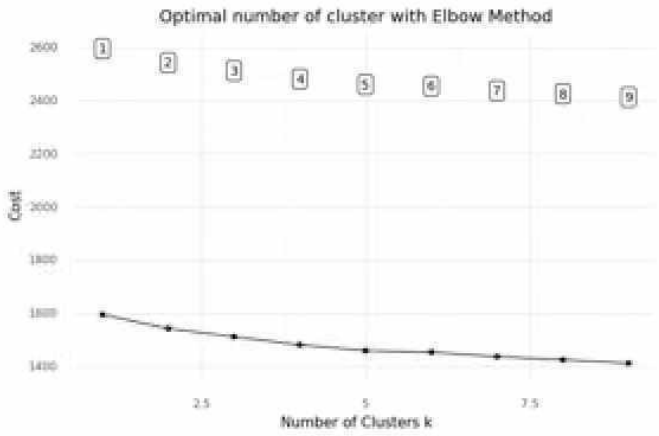
[redacted] 주성분으로 이진형 데이터의 81.47% 설명

[redacted]

별첨 – 디폴트 가중치 감마

A. 감마 설정 없이 디폴트 값으로 : 로봇청소기 데이터에서 디폴트 가중치 감마 값 = [redacted]

Cluster	Cost
0	[redacted]
1	
2	
3	
4	
5	
6	
7	
8	



index	cluster_id
1	[redacted]
2	
0	
3	

Cluster 4개 선택 , cost는 [redacted]

[redacted]

별첨 – 최적 가중치 감마 그리드 서치 (cluster 4기준)

가중치 감마	cost	가중치 감마	cost	가중치 감마	cost
90	[redacted]	1.8	[redacted]	0.15	[redacted]
80	[redacted]	1.6	[redacted]	0.14	[redacted]
70	[redacted]	1.4	[redacted]	0.13	[redacted]
60	[redacted]	1.2	[redacted]		
50	[redacted]	1.0	[redacted]		
40	[redacted]	0.8	[redacted]		
30	[redacted]	0.6	[redacted]		
20	[redacted]	0.4	[redacted]		
10	[redacted]	0.2	[redacted]		
8	[redacted]	0.19	[redacted]		
6	[redacted]	0.18	[redacted]		
4	[redacted]	0.17	[redacted]		
2	[redacted]	0.16	[redacted]		

[redacted]

별첨 – 클러스터링 평가 지표

지표	설명	식	클러스터링 잘된 정도
K-prototype cost	클러스터 중심에서의 모든 점들의 거리 합	-	작을 수록
Calinski harabasz score	클러스터 내에서의 분산과 비교하여 클러스터 전체의 분산 정도의 비율	클러스터 내 전체 분산(SS_B)분의 클러스터간 전체 분산(SS_W) $\frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$	클수록
Davies bouldin Index	클러스터 내에서 Distribution과 비교하여 다른 클러스터 간의 분리 정도의 비율	두 개의 클러스터 쌍에 대해 각 클러스터의 크기의 합을 각 클러스터의 중심 간 거리로 나눈 값	작을 수록

별첨 – 클러스터링 결과 및 클러스터 설명

	명	우세 연령	우세 성별 *	소득이 줄어도 소비를 줄이지 않을 항목 *	소득이 줄면 소비를 줄일 항목 *	우세하게 보유한 가전 제품 *	구매경로	구매 시 고민 기간
Cluster1	[redacted]							
Cluster2								
Cluster3								
Cluster4								

-주요 연령층 : 20-40대 * 다른 군집 대비

➢ 해석

- Cluster1 : 쾌락형 라이프 스타일
  - ✓ 주로 [redacted] 등 자신의 외면을 가꿈.
  - ✓ 최신 트렌드의 컨텐츠 소비하며 즐길 수 있는 서비스를 중요시 함.
  - ✓ 보유한 가전 제품이 많음
  - ✓ 구매 고려 [redacted]
  - [redacted]의 비율이 가장 높은 군집
- Cluster2 : 건강형 라이프 스타일
  - ✓ 주로 [redacted] 등을 중요하게 여김.

[redacted]에 대한 중요도가 가장 낮음

[redacted]

- Cluster3 : 합리형 라이프 스타일
  - ✓ 꼭 필요하지 않은 것에 대한 소비의 [redacted]
  - ✓ 소득이 줄 때 소비를 줄일 수 있는 항목이 가장 [redacted]
  - ✓ 보유한 가전 제품이[redacted]
  - ✓ 구매 고민 기간이 [redacted]
- Cluster4 : 눈에 띄는 소비 패턴이[redacted] 소비에 [redacted]

[redacted]

별첨 – 클러스터링 해석 시 각 항목 비율

각 클러스터 별 소득이 줄어도 소비를 줄이지 않을 항목에 'yes'라고 응답한 비율

패션의류

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

화장품

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

외식

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

컨텐츠

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

운동

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

여행

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

생수

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

청소도구

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

수납 정리도구

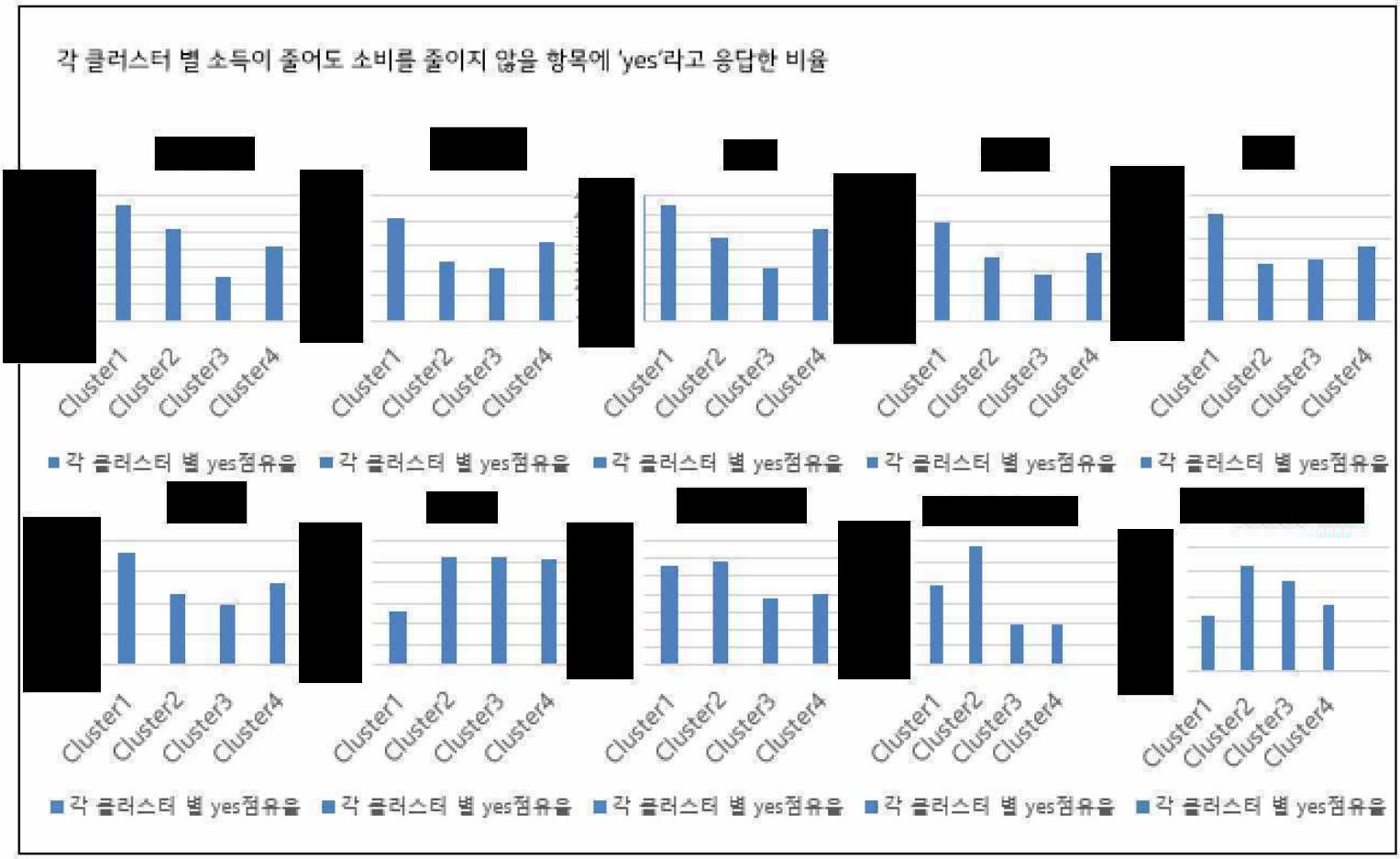
	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

건강, 영양제, 의료

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	
Cluster3	
Cluster4	

[redacted]

별첨 – 클러스터링 해석 시 각 항목 비율



별첨 – 클러스터링 해석 시 각 항목 비율

각 클러스터 별 소득이 줄면 소비를 줄일 항목에 'yes'라고 응답한 비율

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

별첨 – 클러스터링 해석 시 각 항목 비율

각 클러스터 별 우세하게 보유한 가전 제품 비율

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

	각 클러스터 별 yes점유율
Cluster1	[redacted]
Cluster2	[redacted]
Cluster3	[redacted]
Cluster4	[redacted]

[redacted]

별첨 – 클러스터링 해석 시 각 항목 비율