



2019 데이터 애널리틱스 랩 인턴과정 최종 발표

월별 고객 군집분석 및 이탈자 특성 파악

이동성



CONTENTS



01

개요

- 데이터 소개
- 변수 소개
- 분석 목표

02

데이터 전처리

- 변수 제거
- 관측치 제거
- 데이터 셋 생성
- 극단치 제거
- 파생변수 생성

03

데이터 분석

- 군집화
- 군집 특성 파악

04

활용 방안

- 군집 특성 제공
- 고객 소비 패턴 추적
- 군집을 통한 이탈자 예측



01. 개요

- 데이터 소개
- 변수 소개
- 분석 목표

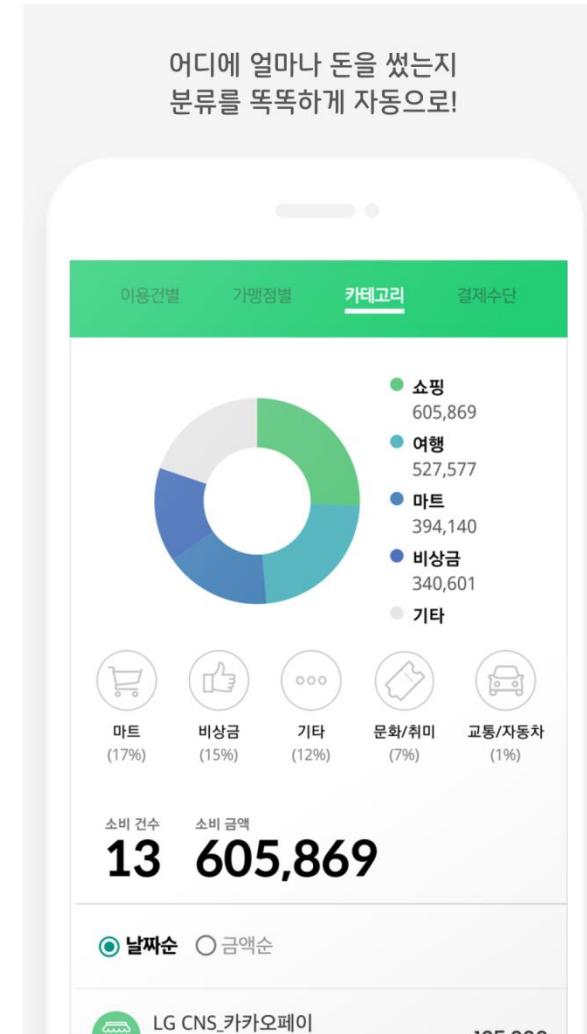
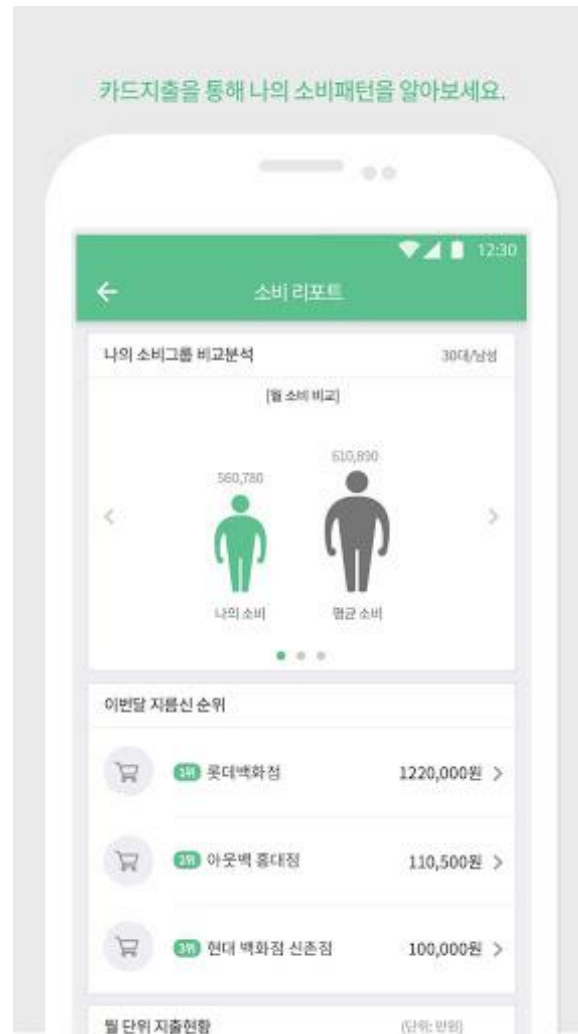
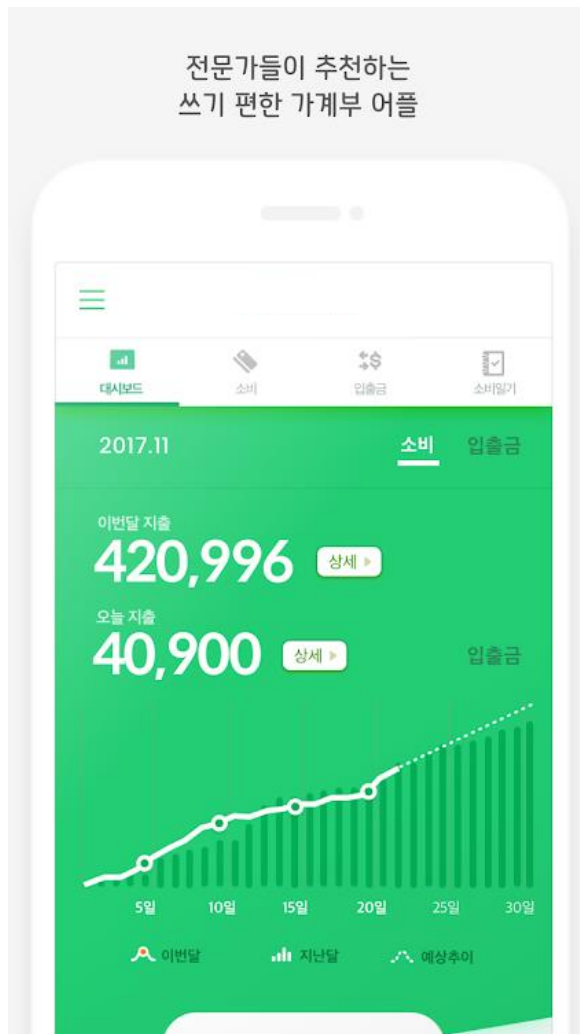
- 데이터 : HF 데이터(해빗 팩토리 가계부 어플 데이터)
- 변수 개수 : 40개
- 수집기간 2017년 11월 ~ 2018년 11월 (13개월)
- 수집 방법 : 문자/ 앱을시를 이용한 실시간 사용금액, 항목, 시간 등을 저장
- 매달 50 ~ 80 만건의 데이터 저장
- 전체 8998551건의 데이터
- 고객수 : 전체 20000명

성별	개체수 (명)
남	9028(45.1%)
여	10972(54.9%)

연령	개체수 (명)
10대이하	435(2.2%)
20대	4827(24.1%)
30대	5840(29.2%)
40대	5861(29.3%)
50대	2350(11.8%)
60대	503(2.5%)
70대 이상	184(0.9%)

타입	변수명
Key (개인 식별 ID)	<ul style="list-style-type: none"> SMS_ID USER_ID USER_SIM_NUMBER PATTERN_ID
Category (소비 항목 분류)	<ul style="list-style-type: none"> CODE CODE_USER GROUP_CODE IS_USER_CATEGORY
SMS_REGISTRATION (문자 등록 시간)	<ul style="list-style-type: none"> TIMESTAMP MONTH DATE TIME
COMPANY (이용 카드사)	<ul style="list-style-type: none"> NAME CODE

타입	변수명
CARD (이용 카드 타입)	<ul style="list-style-type: none"> TYPE NAME
CARD_APPROVAL (카드 승인 정보)	<ul style="list-style-type: none"> TYPE DATE TIME METHOD PRICE STORE BALANCE CURRENCY_UNIT REAL_PRICE MEMO
Etc.	<ul style="list-style-type: none"> ORIGINATING_ADDRESS RATING LATITUDE LONGITUDE ADDRESS



- 성별/ 연령으로만 구분하여 실제 소비패턴을 비교하기 어렵다.
 - 기술통계량만 제공할 뿐 소비패턴의 변화(순/역 방향)등을 제공하지 않는다.
- ❖ 목표 : 단순 기록을 위한 어플 -> 고객에게 다양한 사용자 경험을 제공

• 목표

1. 월별 고객의 소비 데이터를 군집화 하여 특성 분석
2. 고객이 시간이 지남에 따라 소비 특성이 변하는지(군집의 변화)에 대한 추적
3. 고객의 이탈가능성을 예측해 보고 그에 맞는 마케팅 방향 추천

• 분석 과정

1. 월별 고객의 데이터의 특성을 나타낼 변수 생성
2. 주성분 분석을 통한 변수 차원 축소
3. 군집 별 특성 파악
4. 이탈 가능성 예측모형 생성

• 방법론

1. 차원 축소 : PCA
2. 군집화 : K-means / SOM(Self Organize map)
3. 예측모형 : Logistic Regression / Random Forest



02. 데이터 전처리

- 변수 제거
- 관측치 제거
- 데이터 셋 생성
- 극단치 제거
- 파생변수 생성

• 분석에 불필요한 변수 제거

1. 대다수 혹은 전부가 결측치인 경우 (제거)
 - USER_SIM_NUMBER, ADDRESS, RATING ,
 - LATITUDE, LONGITUDE, GROUP_CODE
2. DB저장에 대한 데이터로 분석에 불필요한 변수 (제거)
 - FIXED_SMS_ID, REGISTRATION_TIMESTAMP 등 9개 변수
3. 중복되는 변수
 - PRICE, CURRENCY_UNIT -> REAL_PRICE 로 사용
 - SMS_ID , USER_SIM_NUMBER , PATTERN_ID -> USER_ID를 KEY변수로 사용

타입	변수명
Key (개인 식별 ID)	<ul style="list-style-type: none"> SMS_ID USER_ID → KEY USER_SIM_NUMBER PATTERN_ID
Category (소비 항목 분류)	<ul style="list-style-type: none"> CODE CODE_USER GROUP_CODE IS_USER_CATEGORY
SMS_REGISTRATION (문자 등록 시간)	<ul style="list-style-type: none"> TIMESTAMP MONTH DATE TIME
COMPANY (이용 카드사)	<ul style="list-style-type: none"> NAME CODE

타입	변수명
CARD (이용 카드 타입)	<ul style="list-style-type: none"> TYPE NAME
CARD_APPROVAL (카드 승인 정보)	<ul style="list-style-type: none"> TYPE DATE TIME METHOD PRICE STORE BALANCE CURRENCY_UNIT REAL_PRICE MEMO
Etc.	<ul style="list-style-type: none"> ORIGINATING_ADDRESS RATING LATITUDE LONGITUDE ADDRESS

• 고객의 소비생활과 관련없는 관측치는 분석 대상에서 제거

1. 소비항목의 기타 분류 코드 재분류 (CARD_APPROVAL_STORE 이용)

- (점심,저녁,밥,간식,외식,식비 단어가 들어가 있으면 기타->식비/외식)
- (저축,적금,대출,은행 단어가 들어가 있으면 기타->금융)
- (월세 단어가 들어가 있으면 -> 주거생활)
- (입금(CARD_APPROVAL_TYPE="-- D")된 경우면 기타->입금)

대분류(CATEGORY_CODE)	
편의점(001)	문화(010)
마트(002)	의료/건강(011)
커피/디저트(003)	주거생활(012)
식비/외식(004)	레저/스포츠(013)
쇼핑(005)	여행(014)
패션/뷰티(006)	반려동물(015)
교통/자동차(007)	금융(016)
...	기타(017)

STORE	BEFORE	AFTER
월세	기타(017)	주거생활(012)
(주) 스타벅스커피 코리아	커피/디저트 (003)	커피/디저트 (003)
(주) 다이소아성산업	쇼핑(005)	쇼핑(005)
점심	기타(017)	식비/외식(004)
적금	기타(017)	금융(016)
밥	기타(017)	식비/외식(004)

- **고객의 소비생활과 관련없는 관측치는 분석 대상에서 제거**

- 2. 불필요한 관측치 제거

- 카드 승인이 취소된 경우(CARD_APPROVAL_TYPE이 “LC”, “FC” 인 경우)
 - 법인 카드의 경우 소비와 상관 없음(CARD_TYPE= “CCT”, “CCK” 인 경우)
 - 소비 항목 분류 코드가 기타, 금융, 입금인 경우 (소비 성향을 알 수 없음, **약 270만개 관측치가 제거됨**)

• 중복된 관측치 제거

- 문자와 앱PUSH를 동시에 받는 경우 중복된 정보가 생김
- 문자 발신 날짜와 시간, 승인가격이 같은 관측치들 중 문자 발신번호 (ORIGINATING_ADDRESS)가 다른 경우 그 중 하나를 제거

UIS011503	BK00002C	SIG_PUSH	201711	20171112	151034	KB국민은행	BK00002	PBK	220602-**-NA	LW	3900	카페보니또백
UIS011503	CD00004C	15881688	201711	20171112	151033	KB국민카드	CD00004	PCK	9*4*조*래	LA	3900	카페보니또 백양

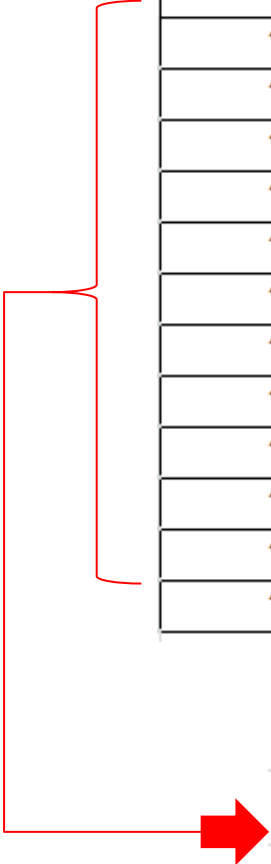


UIS011503	BK00002C	SIG_PUSH	201711	20171112	151034	KB국민은행	BK00002	PBK	220602-**-NA	LW	3900	카페보니또백	
UIS011503	CD00004C	15881688	201711	20171112	151033	KB국민카드	CD00004	PCK	9*4*	조*래	LA	3900	카페보니또 백양

분석 가능 관측치 개수 : 8998551 - > 5359497개

• 월별 고객 데이터 생성

key	Month	COMPANY_NAME	CARD_APPR OVAL_STORE	CARD_APPR OVAL_REAL_PRICE	CATEGORY	GENDER	gen
1	201711	현금	엄마용돈	400000	금융	2	30
1	201711	현금	월세	200000	주거생활	2	30
1	201711	카카오뱅크	적금(1875)	500000	금융	2	30
1	201711	카카오뱅크	비욘드전주경	1000000	패션/뷰티	2	30
1	201711	카카오뱅크	스타벅스커피코	20000	커피디저트	2	30
1	201711	카카오뱅크	엔젤홈마트	9050	마트	2	30
1	201711	카카오뱅크	이디스다이소아성산	4500	쇼핑	2	30
1	201711	카카오뱅크	세이프박스 이지	3840	입금	2	30
1	201711	카카오뱅크	이랜드파크 외	8900	식비/외식	2	30
1	201711	카카오뱅크	스프레소 플래그	51000	커피디저트	2	30
1	201711	카카오뱅크	EN SMITH TEAM	10000	커피디저트	2	30
1	201711	신한카드	키친피콜로	33000	식비/외식	2	30



KEY	Month	GENDER	gen	Volume	...	N
1	201711	2	30	12	...	3

- 개인 고객을 월별로 개체(Obs)를 생성

key	Date	Spend_all	Period	GENDER	gen	age	Volume
1	201711	1824894	12	2	30	37	47
1	201801	13224540	12	2	30	37	64
1	201802	891792	12	2	30	37	44
1	201803	1770524	12	2	30	37	41
1	201712	3940070	12	2	30	37	53
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5	201804	2044849	13	2	30	33	52
5	201809	1703995	13	2	30	33	63
5	201810	875075	13	2	30	33	42
5	201811	970106	13	2	30	33	37

- 군집화를 위해 개인 고객을 월별로 나누어 분석
고객수 : 19507명 -> 140777명

- 변수 생성

Volume : 고객의 매달 관측된 거래량
Age(Gen) : 연령(세대)
Ncard : 매달 사용하는 계좌(카드) 개수

GENDER : 성별
Period : 고객의 어플 사용 기간(월)
Spend_all : 매달 소비하는 총 금액

• 극단적 관측치 제거(Outlier)

- 매달 총 소비 금액이 극단적으로 많은 경우(총 지출금액(spend_all) > 3,557,951)

```
> bx<-boxplot.stats(MonCat_rank$Spend_all)
> min(bx$out)
[1] 3557951
```

- 문제 : 지속 이용자의 소비가 급격히 증가할 때 극단적 관측치로 판단 되어 제거 되는 경우가 발생
- 해결 : 모든 이용 기간 동안 매달 전부 소비가 3,557,951원 이상인 고객에 대해서만 관측치 제거
- 고객수 19507명 -> 19332명
- 관측치 개수 : 140777개 -> 139555개

- 극단적 관측치 제거(Outlier)

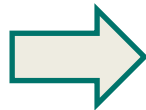
key	Month	Spend_all	period	count_rich
10471	201803	98160140	5	5
10471	201802	97970323	5	5
10471	201801	91619016	5	5
10471	201712	107034278	5	5
10471	201711	70601330	5	5
12680	201811	449076	13	5
12680	201810	225349857	13	5
12680	201809	80467024	13	5
12680	201808	57801818	13	5
12680	201807	6842860	13	5
⋮	⋮	⋮	⋮	⋮
12680	201712	1279403	13	5
12680	201711	7241066	13	5

→ 제거

→ 남김

- 카테고리 항목을 재정의 -> 필수 항목 / 선택 항목

대분류 (CATEGORY_CODE)
편의점(001)
마트(002)
커피/디저트(003)
식비/외식(004)
쇼핑(005)
패션/뷰티(006)
교통/자동차(007)
통신(008)
...
여행(014)
반려동물(015)
금융(016)
기타(017)



필수(1)		선택(10)	
대분류	소분류	대분류	소분류
주거(1)	공과금, 관리비,세금	주거(11)	경조사, 기부금, 세탁, 기타
식품(2)	식비/외식, 마트	식품(12)	편의점, 커피/디저트, 술/유흥
쇼핑(3)	가전, 가구, 패션/의류, 뷰티/미용	쇼핑(13)	온라인 쇼핑, 복합 쇼핑몰, 면세점
교통(4)	대중교통, 주유, 통행료	교통(14)	택시, 차량관리
통신(5)	통신	취미(15)	취미
교육(6)	교육	레저(16)	레저
의료(7)	의료	여행(17)	여행
		반려동물(18)	반려동물

• 특성변수 생성

- Rank 1 - 5 : 최빈 이용 항목 (5순위 이하의 절반이상이 결측이라 버림)
- Rate 1 - 5 : 최빈 이용 항목의 사용 금액 비율 (%) (5순위 이하의 절반이상이 결측이라 버림)
- JUYA : 주간,야간 구매 패턴(주간이 많으면 1, 야간이 많으면 0)
- Volume : 월별 결제 횟수 (Volumec : 결제량 카테고리화 변수)
- Ncard : 월별 사용 카드 개수
- Spendc : 월별 지출 금액 카테고리화 변수 (box.stats를 이용하여 구간 나눔)
- Mono : 월별 결제 카테고리 개수 (box.stats를이용 하여 구간 나눔)

❖ 총 15개의 특성 변수를 이용하여 군집화 실시



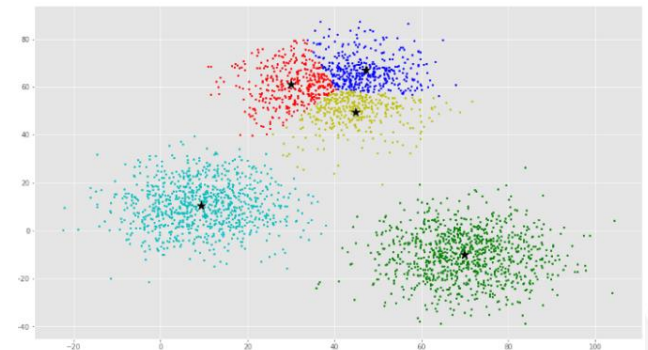
03. 군집화

- 군집화
- 군집 특성 파악

• 군집화 모형

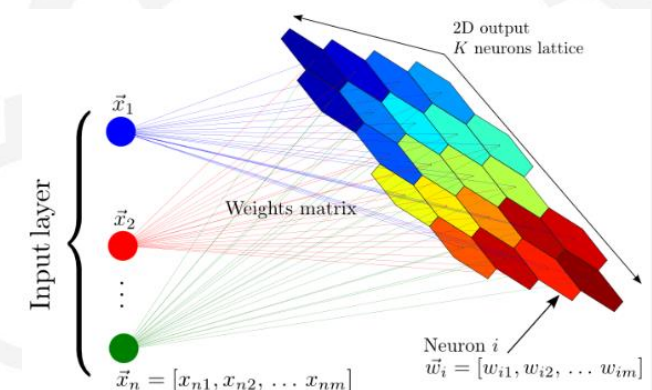
• K-means clustering

- 각 그룹 내 거리를 최소화 하는 군집의 중심을 찾는 알고리즘 (EM 알고리즘)
- Center 초기값 위치에 따라 결과가 달라질 수 있다.



• SOM(Self Organize Map)

- 차원축소와 군집분석을 동시에 실시하는 알고리즘 (인공 신경망의 한종류)
- 고차원의 데이터를 이차원 평면으로 그려줌
- 데이터가 충분히 많아야 성능이 좋아진다.

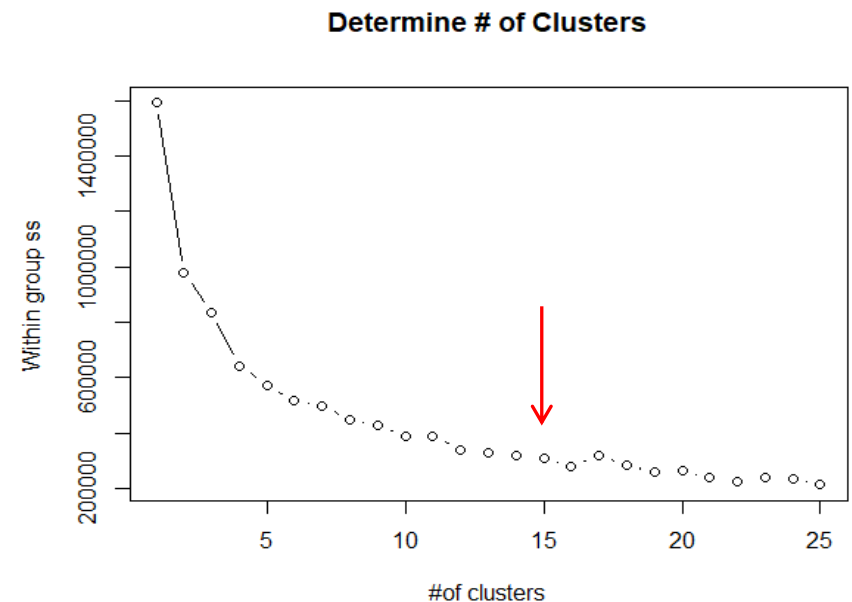
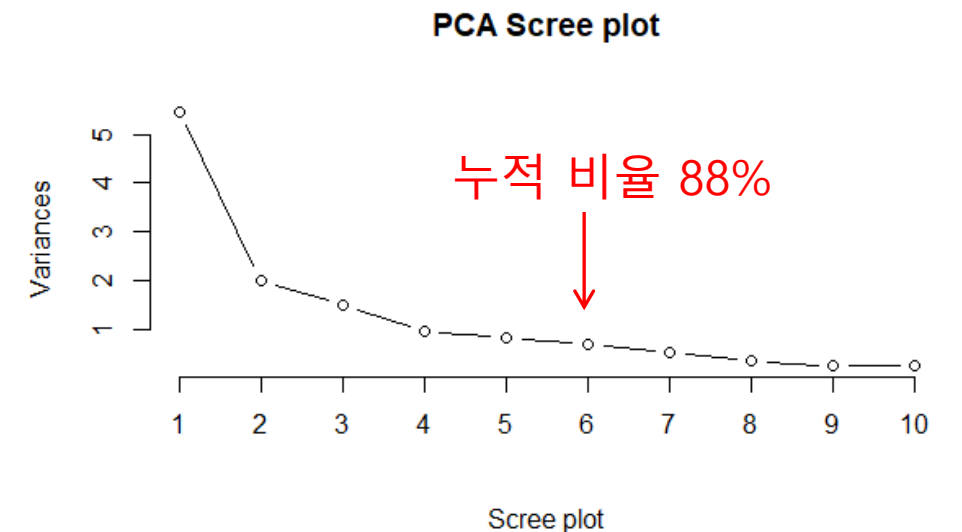


• 군집 방식

- 15개의 변수들의 차원축소(PCA)를 이용하여 군집화를 실시
- Scree plot을 보고 차원의 개수 결정 (Variance 누적 비율 90% 설명력 내외)

• 군집 개수 결정

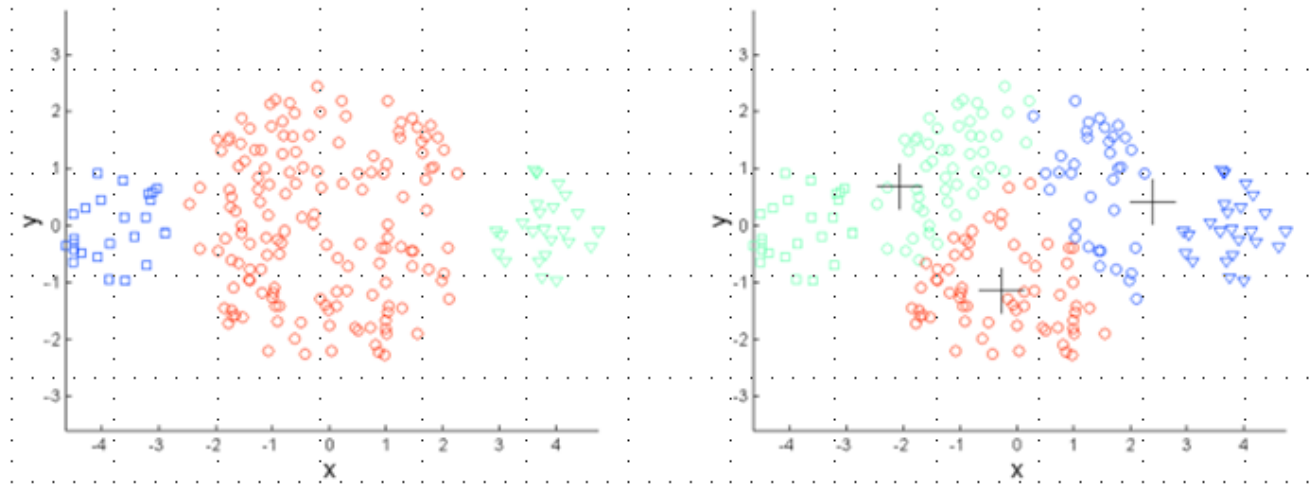
- 군집화(K-means 기준)의 그룹 내 제곱합(Within Sum of Square)가 가장 낮은 군집을 선택



❖ 주성분 개수 6개, 군집 개수 15개로 결정

• 군집 모형 선택

- 모형 비교 : 그룹간 제곱합 / 총 제곱합 (Between Sum of Square / Total Sum of Square)
- 전체 거리차이의 합 중에서 그룹간의 거리차이의 합 비중이 크면 군집이 잘 나누어 졌다.



모형	Between SS	Total SS	Ratio
K-means	1228988	1596527	76.9%
Som	1245813	1596527	78.0%

← ❖ SOM 모형 채택

• 군집화 결과

군집 별 특성 비교

군집	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
개체수	4560	7799	11716	8726	8840	10983	6704	11840	12155	19092	6078	4916	7914	6026	12206
지출	1 : 지출금액 적음 -> 7 : 지출금액 많음														
금액등급	6.533114	4.249006	4.026545	3.205019	2.803394	2.742875	2.72151	2.693497	2.568408	2.466111	1.597236	1.542311	1.481173	1.449386	1.136408
평균지출액	13,902,393	3,060,763	2,586,746	1,820,785	1,260,853	1,310,465	1,386,181	1,117,900	1,013,192	915,191	415,344	387,230	347,038	333,408	161,785
주/야															
야간	1096	184	256	0	8840	0	6704	11840	0	0	0	4916	0	6026	4811
주간	3464	7615	11460	8726	0	10983	0	0	12155	19092	6078	0	7914	0	7395
CATEGORY															
필수항목	2335	0	11716	0	0	10983	3858	11840	0	19092	0	0	7914	6026	6749
선택항목	2225	7799	0	8726	8840	0	2846	0	12155	0	6078	4916	0	0	5457
소비경향	1 : 한 개 항목 -> 5 : 여러 개 항목 소비														
	4.93	5.00	5.00	4.88	4.97	4.77	4.73	4.98	4.98	4.97	3.05	2.88	3.05	2.96	1.00
거래량	1 : 거래량 적음 -> 5 : 거래량 많음														
	3.64	3.80	3.85	2.64	3.00	2.47	2.32	3.00	2.88	2.78	1.41	1.35	1.36	1.34	1.01

• 군집화 결과

군집 별 최빈 선호 항목 비율

	1등
	2등
	3등

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
주거(1)	4%		4%			4%	2%	3%		4%			6%	5%	3%
식품(2)	21%		67%			58%	35%	66%		59%			44%	53%	15%
쇼핑(3)	9%		5%			8%	5%	6%		5%			7%	7%	3%
교통(4)	5%		8%			9%	5%	13%		14%			21%	18%	21%
통신(5)	5%		3%			6%	5%	6%		6%			10%	9%	7%
교육(6)	2%		6%			5%	3%	3%		5%			4%	6%	4%
의료(7)	5%		7%			9%	3%	4%		7%			8%	3%	3%
주거(11)	1%	1%		1%	1%		0%		1%		2%	3%			4%
식품(12)	4%	6%		4%	19%		3%		14%		14%	19%			9%
쇼핑(13)	36%	80%		82%	64%		35%		71%		69%	62%			20%
교통(14)	4%	2%		2%	2%		1%		2%		2%	2%			2%
취미(15)	1%	3%		3%	5%		1%		4%		6%	7%			8%
레저(16)	0%	1%		1%	1%		0%		1%		1%	1%			1%
여행(17)	2%	7%		8%	7%		3%		7%		5%	6%			2%
반려동물(18)	0%	0%		0%	0%		0%		0%		0%	0%			0%

지출 금액 별

최고 지출
(1)

고 지출
(2,3)

중간 지출
(4,5,6,7,8,9)

저 지출
(10,11,12)

최저 지출
(13,14,15)

시간대별

주간
(4,6,9,10,11,13)

야간
(5,7,8,12,14)

복합
(1,2,3,15)

지출 항목별

필수 지출
(3,6,8,10,13,14)

선택 지출
(2,4,5,9,11,12)

복합 지출
(1,7,15)

거래량 별

많음
(1,2,3)

중간(4,5,8,9,10)

적음
(6,7)

매우 적음
(11,12,13,14,15)





04. 활용 방안

- 군집 특성 제공
- 고객 소비 패턴 추적
- 군집을 통한 이탈자 예측

• 군집의 특성 제공

KEY	거래월	성별	나이	기간	총지출금액	주/야	최빈선헬항목	거래량	소비경향	군집(SOM)
11178	201711	남자	30대	7	50000	야간(0)	교통비(4)	매우적음(1)	집중구매(1)	15

최고 지출 (1)	주간 (4,6,9,10,11,13)	필수 지출 (3,6,8,10,13,14)	많음 (1,2,3)
고 지출 (2,3)	야간 (5,7,8,12,14)	선택 지출 (2,4,5,9,11,12)	중간 (4,5,8,9,10)
중간 지출 (4,5,6,7,8,9)	복합 (1,2,3,15)	복합 지출 (1,7,15)	적음 (6,7)
저 지출 (10,11,12)			매우 적음 (11,12,13,14,15)
최저 지출 (13,14,15)			

지출이 적고
거래량이 적은 고객

• 군집의 특성 제공

KEY	거래월	성별	나이	기간	총지출금액	주/야	최빈선헬항목	거래량	소비경향	군집(SOM)
250	201803	여자	20대	13	1113842	주간(1)	레저/스포츠(16)	많음(4)	다양함(5)	4

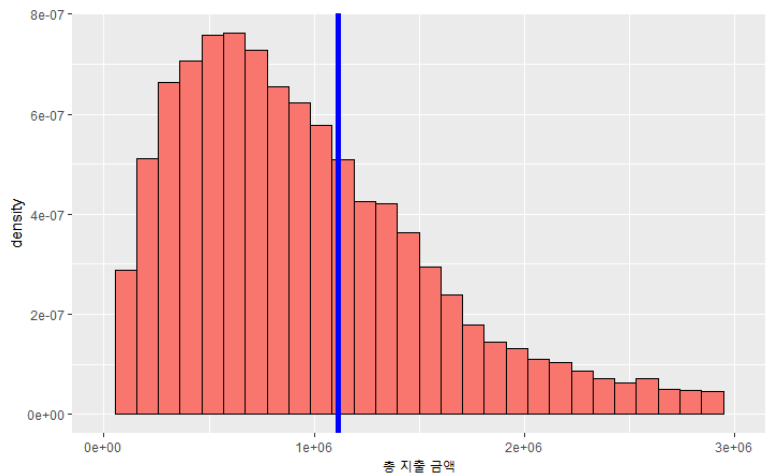
최고 지출 (1)	주간 (4,6,9,10,11,13)	필수 지출 (3,6,8,10,13,14)	많음 (1,2,3)
고 지출 (2,3)	야간 (5,7,8,12,14)	선택 지출 (2,4,5,9,11,12)	중간 (4,5,8,9,10)
중간 지출 (4,5,6,7,8,9)	복합 (1,2,3,15)	복합 지출 (1,7,15)	적음 (6,7)
저 지출 (10,11,12)			매우 적음 (11,12,13,14,15)
최저 지출 (13,14,15)			

중간지출에
주간소비가 많고
선택적 지출이 많은
다양한 소비를 하는 고객

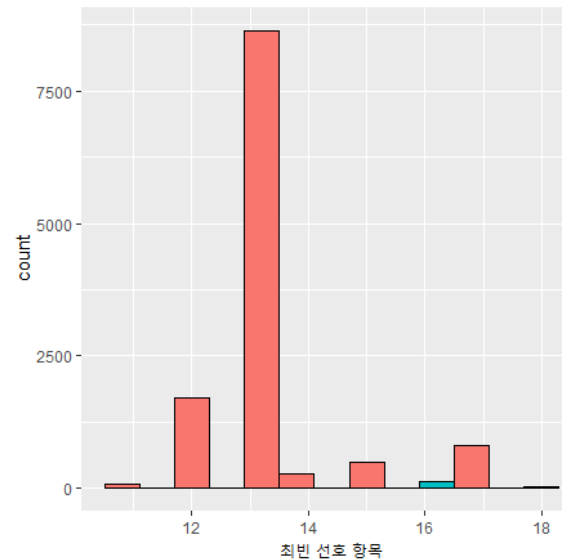
• 군집내의 자신의 위치를 보여줌

KEY	거래월	성별	나이	기간	총지출금액	주/야	최빈선포항목	거래량	소비경향	군집(SOM)
250	201803	여자	20대	13	1113842	주간(1)	레저/스포츠(16)	많음(4)	다양함(5)	4

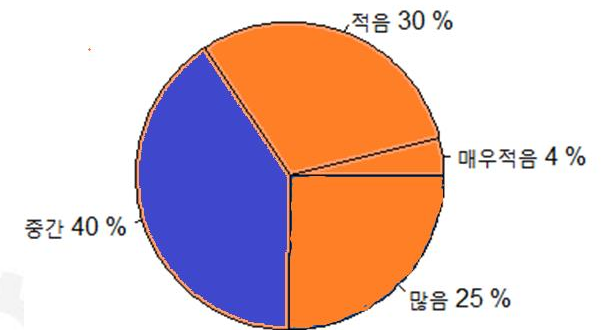
총 지출 금액



최빈 선호 항목



거래량

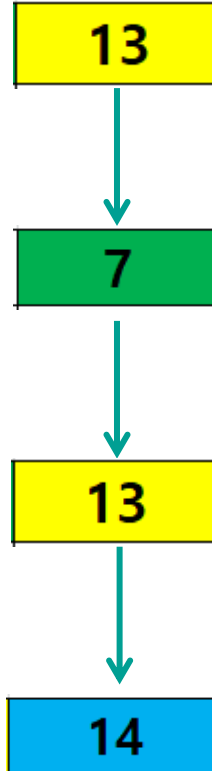


❖ 4번 군집의 특성을 보여주고, 군집 내 위치를 보여 줄 수 있다.

• 월별 소비패턴을 추적, 특성 파악

key	거래일	성별	나이	총 지출금액	주/야	최빈 선호항목	카드갯수	소비경향	거래량	SOM
1173	201711	1	30	3064546	1	2	3	5	66	13
1173	201712	1	30	1288340	1	2	2	5	41	7
1173	201801	1	30	2553500	1	1	3	5	42	13
1173	201802	1	30	4895196	1	13	4	5	93	14
1173	201803	1	30	2482660	1	6	4	5	58	13
1173	201804	1	30	5701405	1	3	4	5	106	13
1173	201805	1	30	1828810	1	3	1	5	10	7

군집	7	13	14
개체수	10983	11716	7799
금액등급	2.742875	4.026545	4.249006
평균지출액	1,310,465	2,586,746	3,060,763
주/야			
야간	0	256	184
주간	10983	11460	7615
CATEGORY			
필수항목	10983	11716	0
선택항목	0	0	7799
소비경향			
	4.77	5.00	5.00
거래량			
	2.47	3.85	3.80



• 지출금액이 감소

• 지출금액이 증가

• 지출금액이 증가

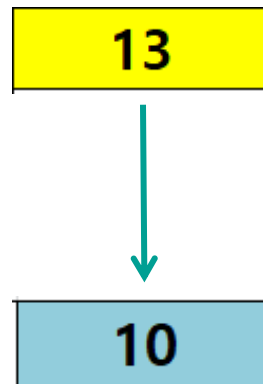
• 선호 항목이

필수(주거(1)) -> 선택(쇼핑(13))

• 월별 소비패턴을 추적, 특성 파악

key	거래월	성별	나이	총 지출금액	주/야	최빈 선호항목	카드갯수	소비경향	거래량	SOM
477	201711	1	30	2281621	1	3	3	5	94	13
477	201712	1	30	2468845	1	3	3	5	79	13
477	201801	1	30	991150	1	2	2	5	62	10
477	201802	1	30	935543	1	2	2	5	55	10
477	201803	1	30	1313002	1	4	1	5	67	10
477	201804	1	30	1440842	1	2	2	5	63	10

군집	10	13
개체수	19092	11716
금액등급	2.466111	4.026545
평균지출액	915,191	2,586,746
주/야		
야간	0	256
주간	19092	11460
CATEGORY		
필수항목	19092	11716
선택항목	0	0
소비경향		
	4.97	5.00
거래량		
	2.78	3.85



- 지출금액이 감소
- 선호항목이 쇼핑(3) -> 식품(2)
- 거래량이 2018년 1월 기준으로 대폭 감소

❖ 군집의 변화에 따른 선택적 마케팅 방안을 모색할 수 있다.

• 이탈자를 어떻게 정의할 것인가?

- 앱 사용자는 2018년 9월까지 기준기간(3개월, 6개월) 이상 사용한 사람들 중 10월, 11월에 이탈한 사람으로 정의

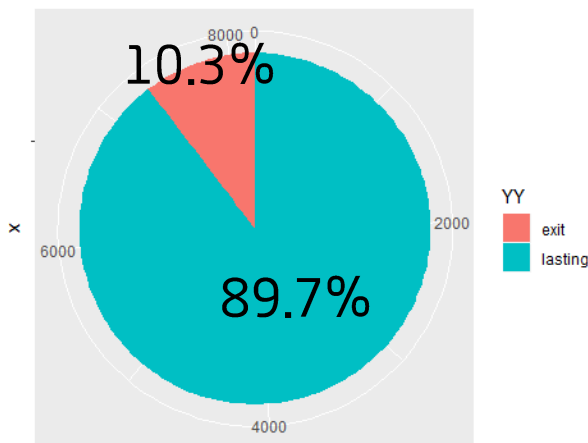
2018년 4월 or 2018년 7월 이전 사용자

2018년 9월 사용

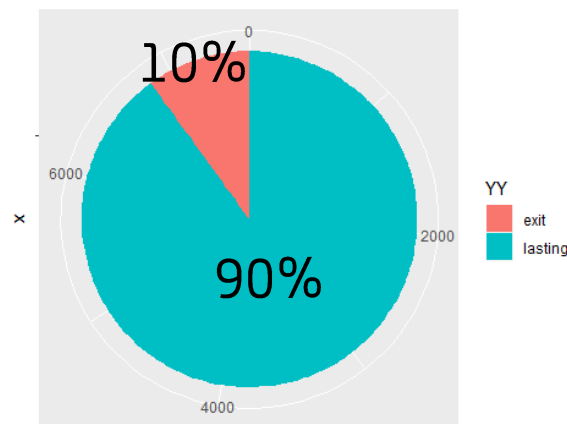
10,11월 이탈

3개월 기준 : 사용자 8193명 중 851 명 이탈 (10.3%)

6개월 기준 : 사용자 7593명 중 759 명 이탈 (10%)



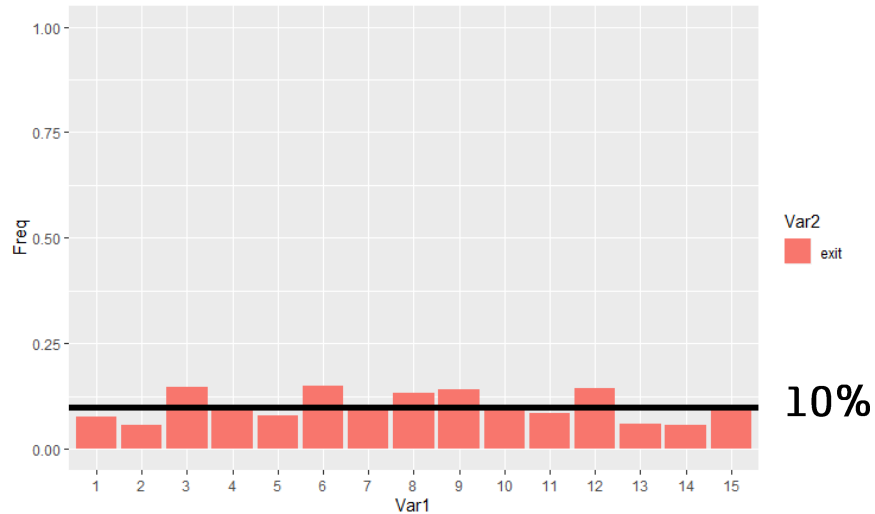
이탈자VS 비이탈자



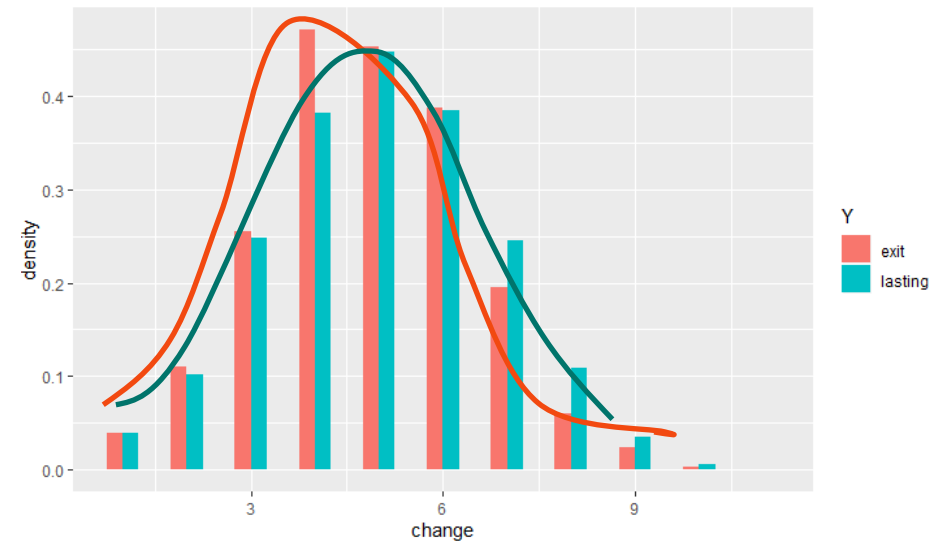
이탈자VS 비이탈자

이탈자와 비 이탈자 비교(3개월 기준)

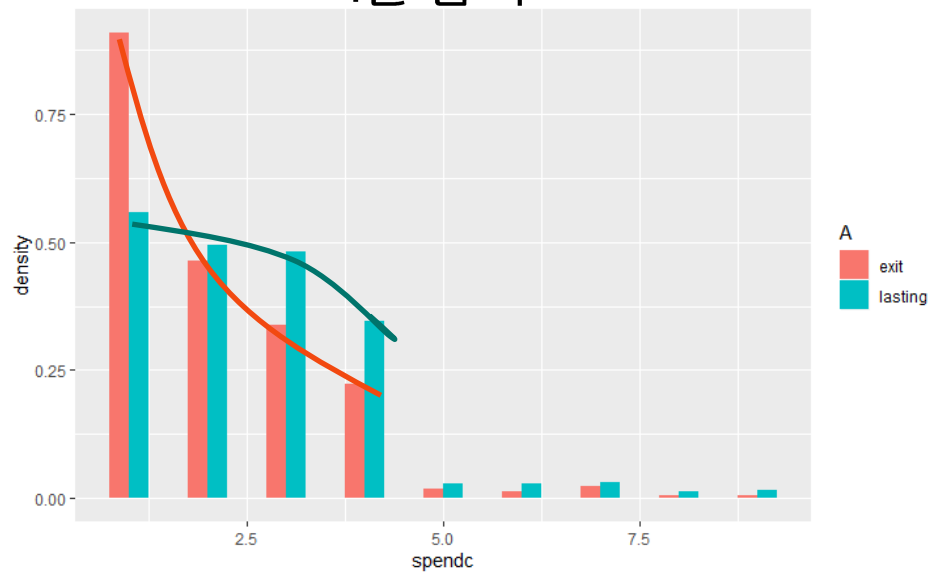
군집 별 이탈 비율



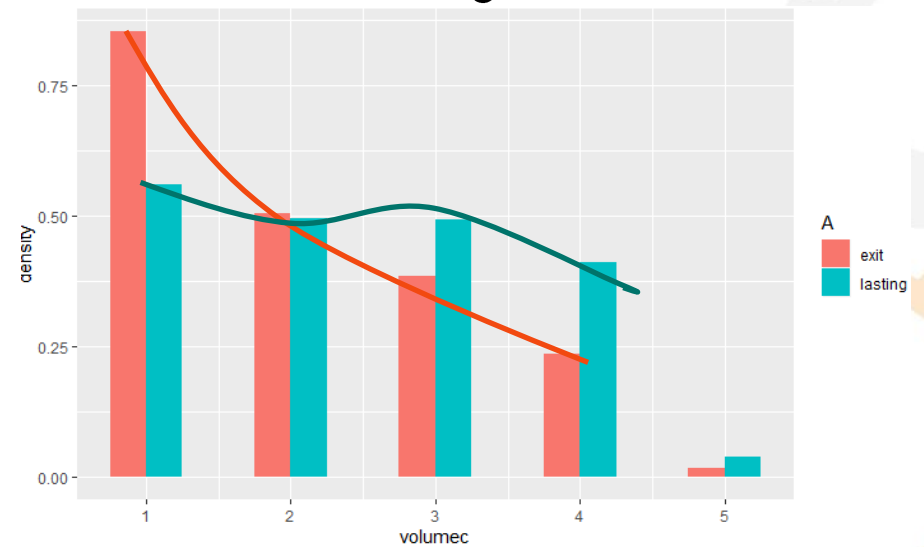
군집 변화 횟수



지출 금액

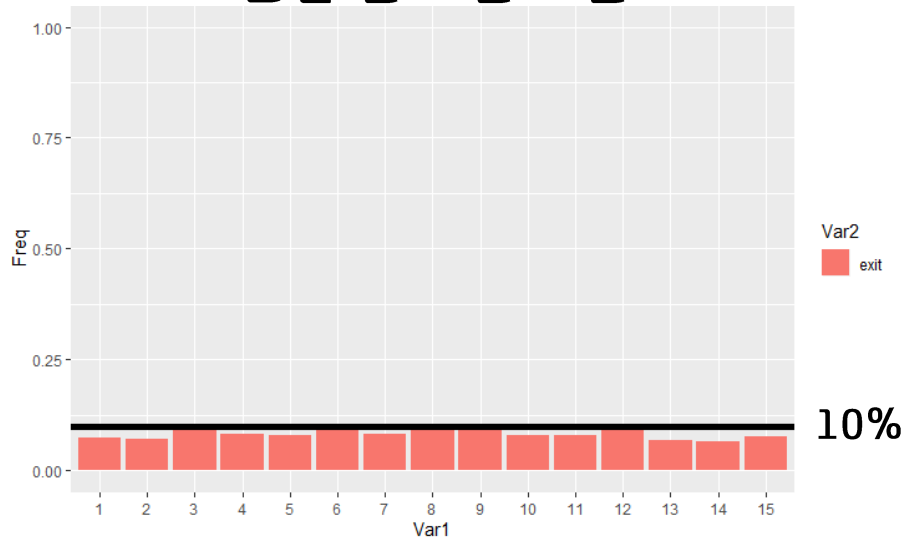


거래량

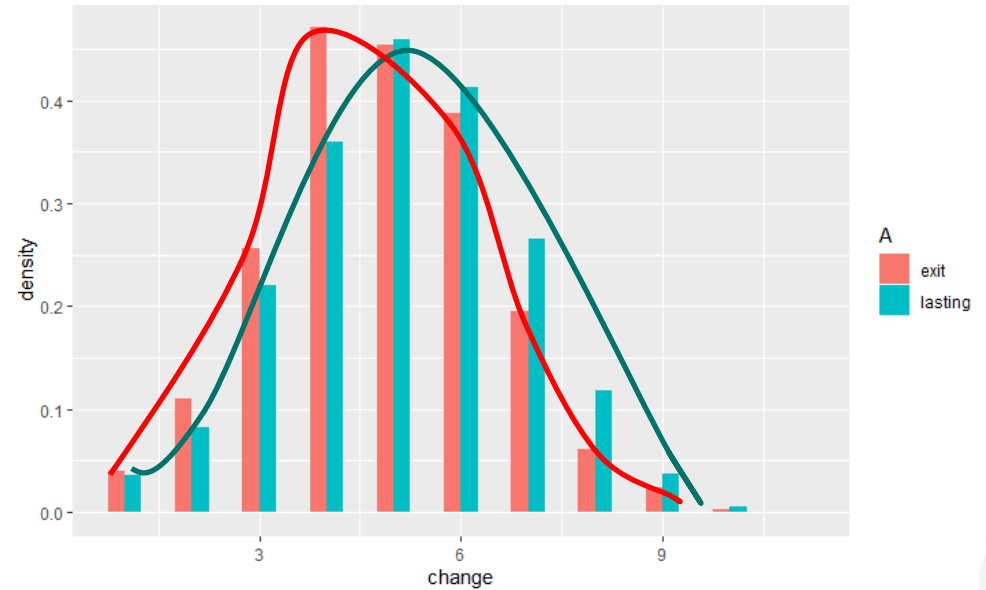


이탈자와 비 이탈자 비교(6개월 기준)

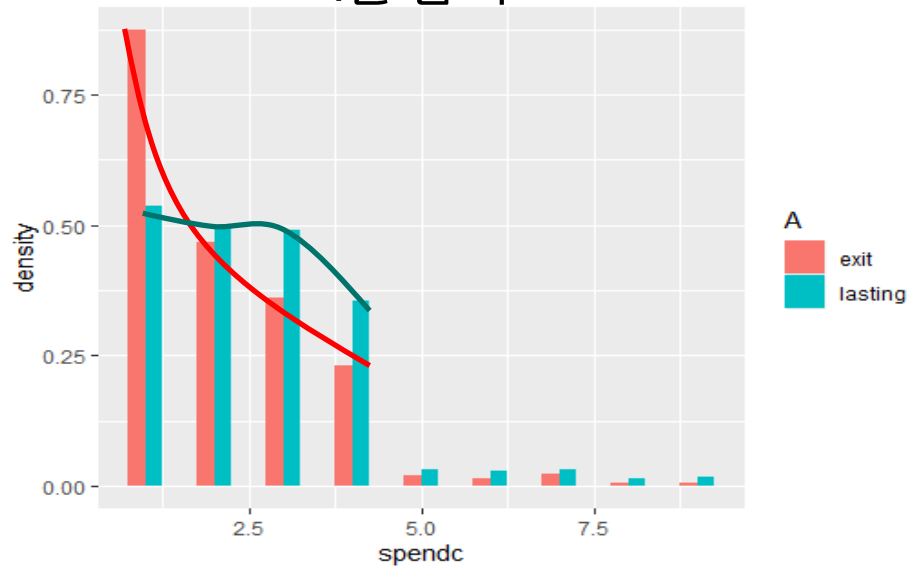
군집 별 이탈 비율



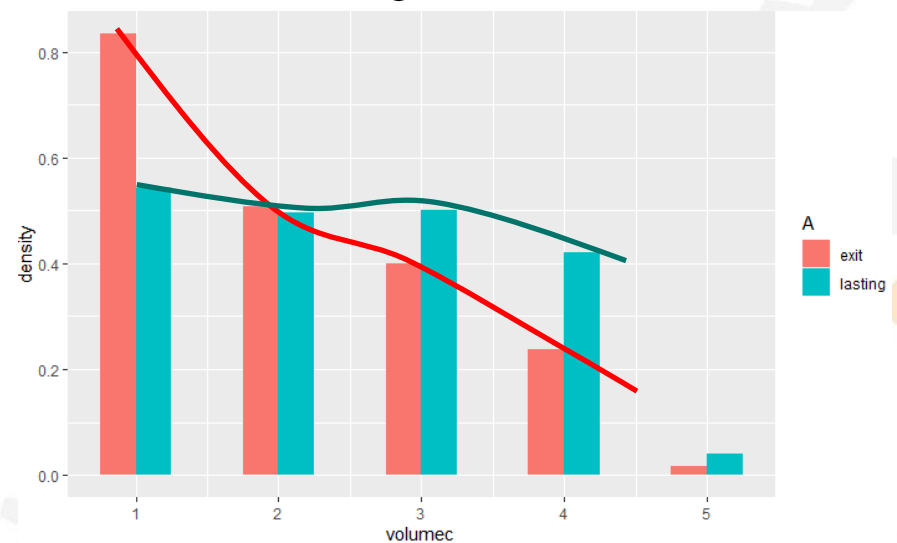
군집 변화 횟수



지출 금액



거래량

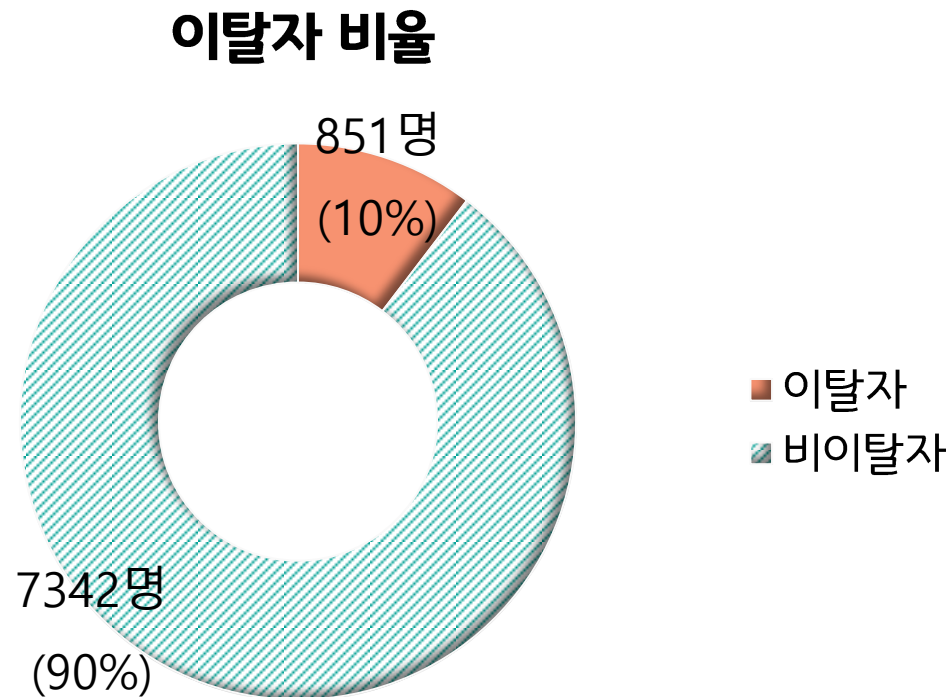


• 변수 생성

- 이용 시작과 끝을 기준으로 변화에 따른 이탈자 파악을 위한 특성 변수들 생성

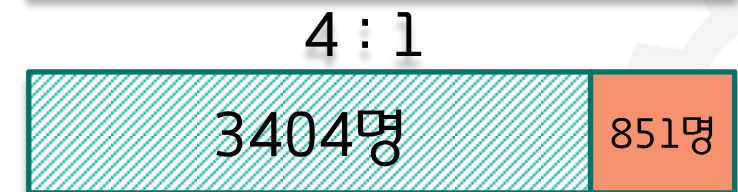
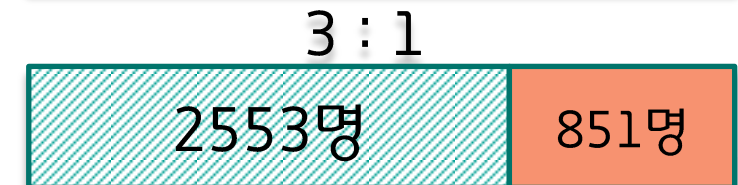
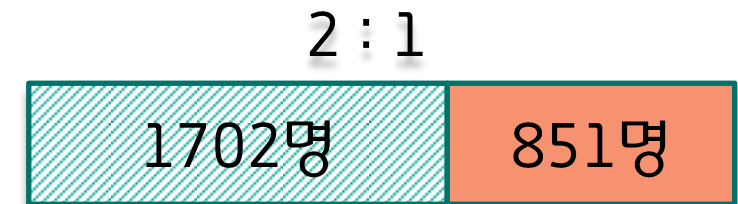
시작월	마지막 월	설명	기타	설명
Fspend	Lspend	소비금액	Period	이용 기간
Fvol	Lvol	거래량	Change	변한 군집 갯수
Fsom	Lsom	군집		
Fcat	Lcat	최빈 선호 항목		
Fmono	Lmono	소비 경향		
Fweight	Lweight	당월 이탈자 비율		

- 샘플링 (Train/ Test set 구성) (3개월 기준)



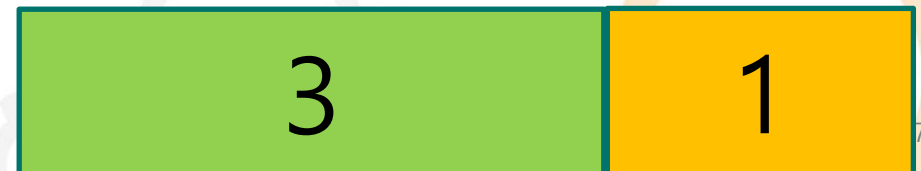
총 고객수 : 8193명

데이터셋 표본 구성 비율



트레인 셋

테스트 셋



• 모형 적합

- 한번의 모델을 각 5회씩 시행 평균 정확도로 판단.(Resampling)

표본 비율	모형	Accuracy1	Accuracy2	Accuracy3	Accuracy4	Accuracy5	평균
2:1	L.R	66.0%	67.6%	66.6%	66.9%	69.6%	67.3%
	R.F	67.9%	68.3%	66.5%	67.2%	65.7%	67.1%
3:1	L.R	73.4%	74.9%	74.0%	73.7%	74.9%	74.2%
	R.F	74.1%	74.3%	74.5%	74.0%	74.7%	74.3%
4:1	L.R	79.8%	79.7%	79.9%	79.9%	79.8%	79.8%
	R.F	79.6%	80.4%	79.4%	79.8%	79.2%	79.7%

모형 정확도가 매우 낮아 앞서 만든 군집의 변화와 관련된 변수들이 이탈자와 비 이탈자의 특성을 설명하지 못하는 것으로 생각 된다.

❖ 고객의 소비패턴의 변화가 이탈과는 연관이 없는 것으로 판단 된다. 이탈자에 대한 예측은 어플 사용에 대한 패턴 변화를 찾아야 할 것으로 생각 된다.



05. 결론

결론

1. 데이터 분석을 통한 사용자 경험을 고객에게 제공 .
2. 군집 별 특성을 파악하여 개별 군집에 대한 마케팅 방향을 고려.
3. 고객의 가계부 데이터를 통해 소비패턴의 변화를 추적하여 이를 상품개발에 이용.

한계점

1. 기타 관측치가 많아 추후 소비항목 분류를 좀더 정교하게 만들 필요가 있음.
2. 임의 추출된 표본이 아니라 실제 데이터와는 다른 결과가 나올 수 있음.
3. 군집을 이용한 이탈자 예측에서 파생변수 생성에 좀더 정교할 필요가 있음.
4. 이탈자 예측에 대해서는 고객의 소비패턴보다 어플 이용 패턴을 분석할 필요가 있음.



감사합니다