제3회 L.POINT Big Data Competition

개인화 상품 추천

팀명:케세라세라

CONTENTS

1.팀원 소개 및 프로젝트 개요
2.데이터 탐색
3.데이터 정제
4.데이터 모델링
5.개선점 및 제언

팀명 : 케세라세라

케세라세라의 뜻

케세라세라는 `원하는 대로 이루어져라`
라는 뜻의 라틴어로서 세상에 대한 긍정적인 마인드를 갖자는 뜻입니다.

팀의 목표

많은 정보를 가지고 있는 **구매 내역**데이터에서 **숨겨진 가치**를 찾아 원하는 것을 이루어보자라는 목표에서 시작하였습니다.

LOTTE

L.POINT

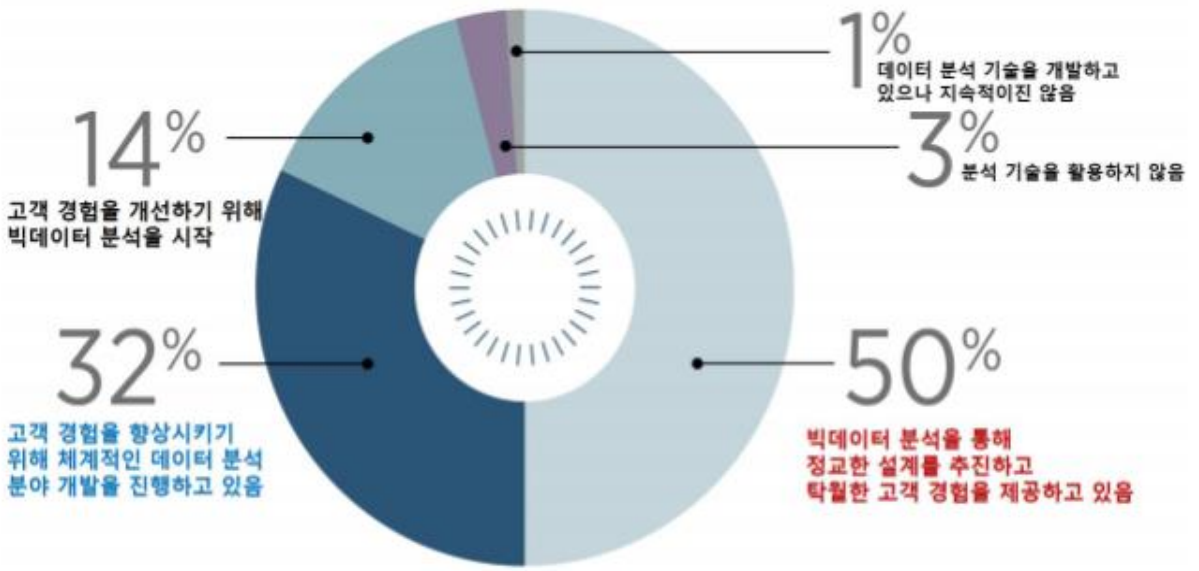
주제

고객의 **소비패턴 데이터**를 통해 구매 상품을 예측.

기획 배경

- 빅데이터 분석의 선두적 기업들은 상당한 양의 데이터를 수집, 처리 하면서 기업 경영에 필요한 Insight로 변환 시키며 높은 효과를 창출.
- 일반적 규모의 기업들은 아직 빅 데이터 이용에 미흡하지만 대규모 기업들은 고객 데이터를 수집하고 이를 폭넓게 활용함.
- 특히 고객의 소비 패턴데이터를 이용하는 비중이 높고 실제로도 정교한 데이터 분석 설계를 추진중임.

[고객 중심의 비즈니스와 고객 경험 개선을 위한 데이터 활용 여부]



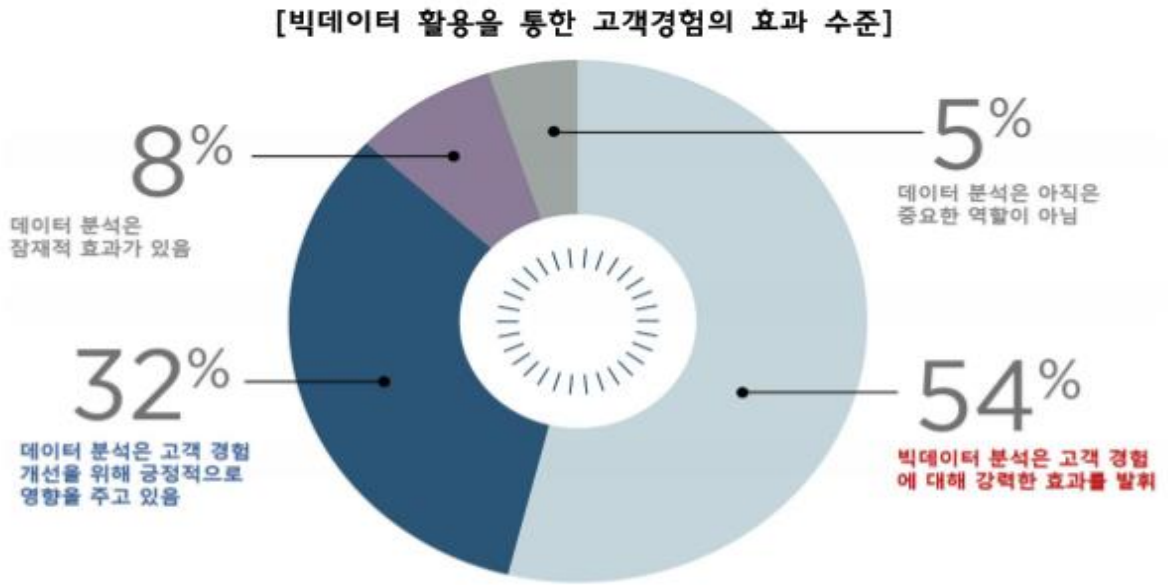
[자료] Forbes Insight : BLAZING THE TRAIL FROM DATA TO INSIGHT TO ACTION 2017

LOTTE

L.POINT

기획 배경

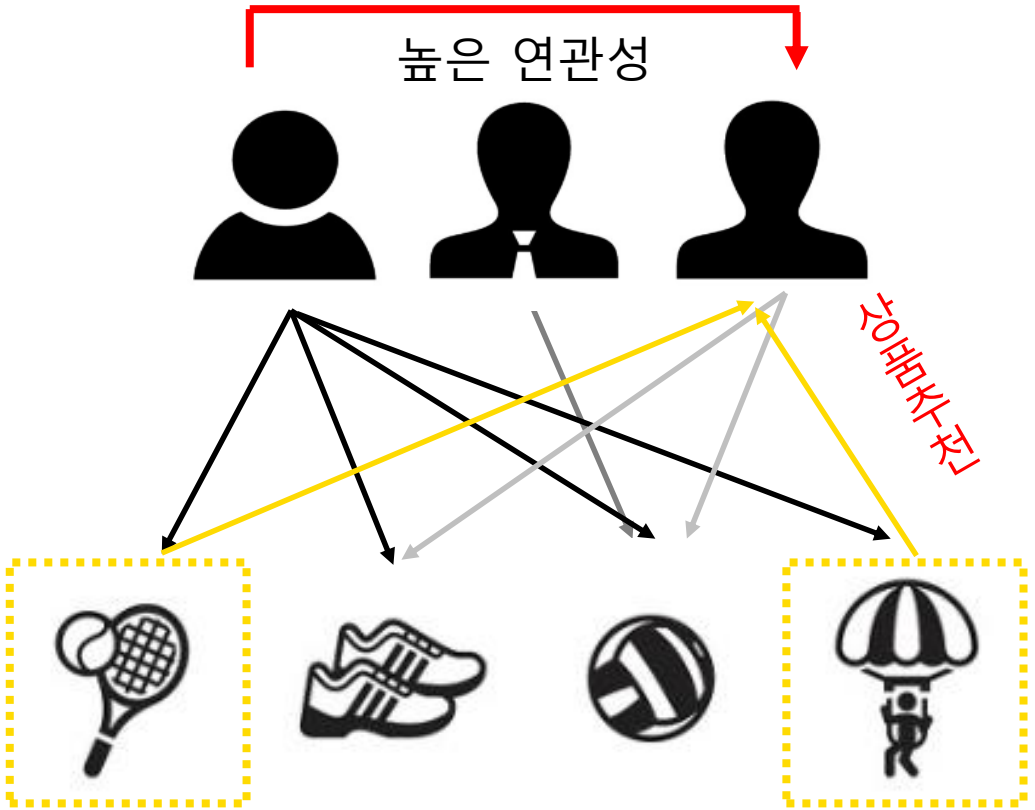
- 글로벌 기업의 54%는 빅데이터 분석을 통한 고객 경험 데이터에 대해 강력한 효과를 경험하였다고 함.



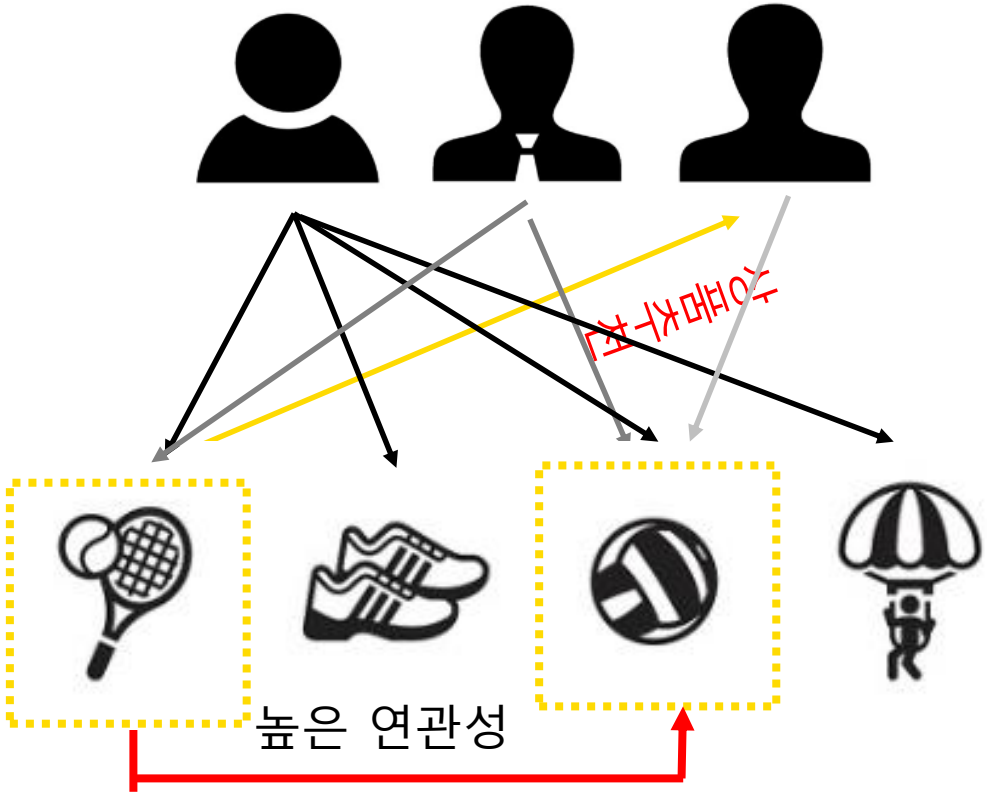
[자료] Forbes Insight : BLAZING THE TRAIL FROM DATA TO INSIGHT TO ACTION 2017

협업 필터링

- 기존 사용자 행동 정보를 분석하여 해당 물품에 대한 구매 성향을 파악하는 것.



유저 기반 (User-based)



아이템 기반 (Item-based)

비즈니스에서 협업 필터링을 이용한 추천서비스 성공사례

고객들의 영화 취향 데이터를 기반으로 사용자 기반 협업 필터링을 활용하여 가장 선호할 만한 영화를 추천

구매의 2/3가 추천으로 부터 발생

amazon

고객들의 도서 구매 이력 데이터를 바탕으로 아이템 기반 협업 필터링을 활용, 구매도서와 유사한 도서 추천

판매의 35%가 추천으로 부터 발생

watcha

내 취향을 분석하는 영화추천

개인의 영화 취향을 분석해 내용 기반, 아이템 기반 협업 필터링을 활용, 선호할 만한 영화를 추천

탁월한 서비스로 광고수익

최근 많은 분야에서 협업필터링을 활용한 추천서비스가 주목받고 있다.

LOTTE

L.POINT

내용기반 필터링

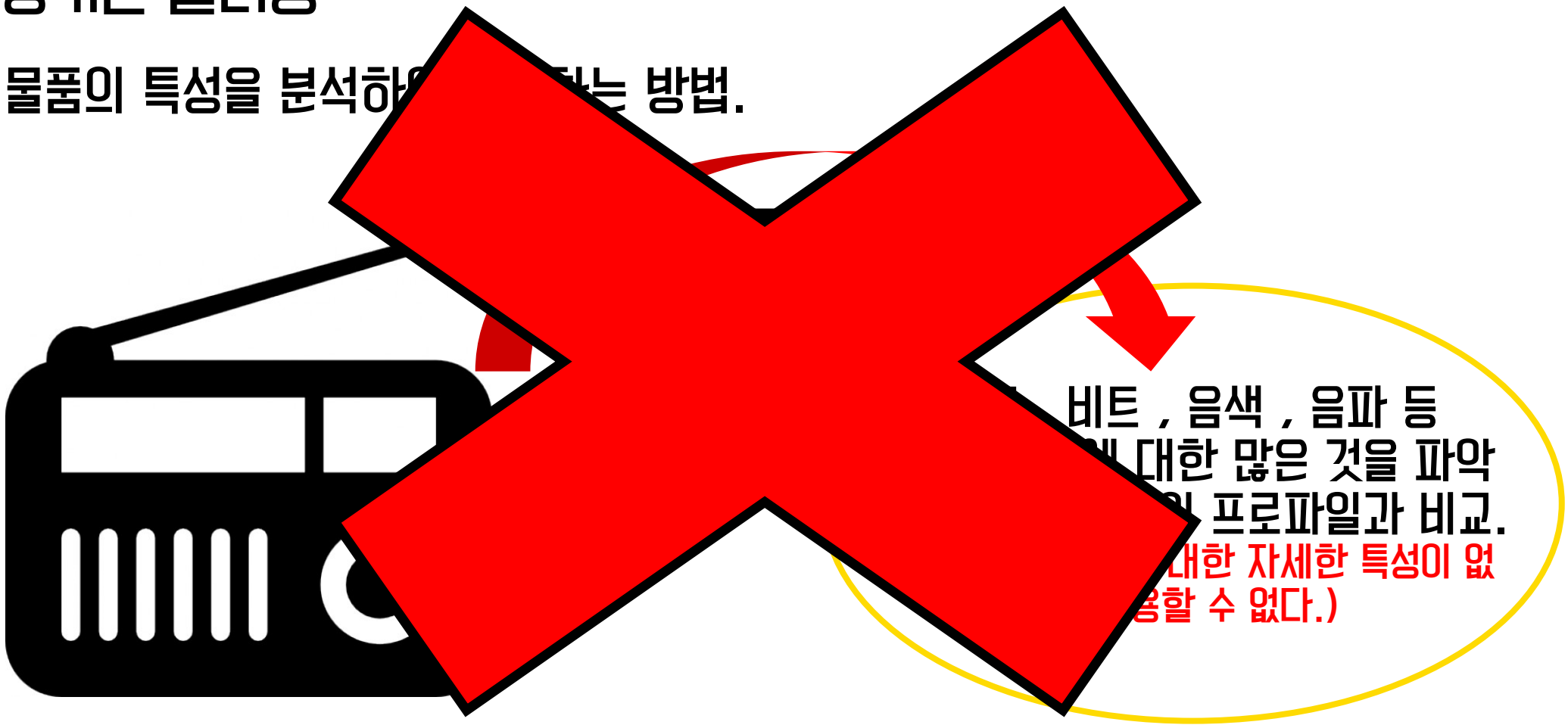
- 물품의 특성을 분석하여 추천하는 방법.

신곡 출시

장르 , 비트 , 음색 , 음파 등
한 음악에 대한 많은 것을 파악
하여 사용자의 프로파일과 비교.
(우리는 품목에 대한 자세한 특성이 없
기 때문에 사용할 수 없다.)

내용기반 필터링

- 물품의 특성을 분석하[illegible]는 방법.



협업 필터링

1. 2016년 1월은 과거의 데이터 이므로, 실제로 고객이 추천시스템을 통하여 마케팅 효과를 본 것이 아니기 때문에 고객들의 재 구매 품목을 상품들을 예측하여 추천하도록 한다.

2.유저 기반과 아이템 기반의 협업 필터링 방법 두가지를 사용하도록 한다.

프로젝트 개요

1) 분석 주제

고객 신상정보
구매 내역 데이터
상품 분류 데이터

+

경쟁사 이용 데이터
멤버쉽 여부 데이터
채널 이용 데이터

{

과제

고객이 선호하거나 선호할 것으로 예상되는 상품 추천

2) 분석 대상

2014년 1월 1일 ~ 2015년 12월 31일 간의 4개 점포에 대한 구매내역 정보 및 고객의 특성 정보 데이터

3) 분석 목표

2년간의 축적된 데이터를 통해 우리만의 새로운 예측 모델을 구축하고 이를 적용하여 간단한 마케팅 방안까지 제시하고자함.

LOTTE

L.POINT

CONTENTS

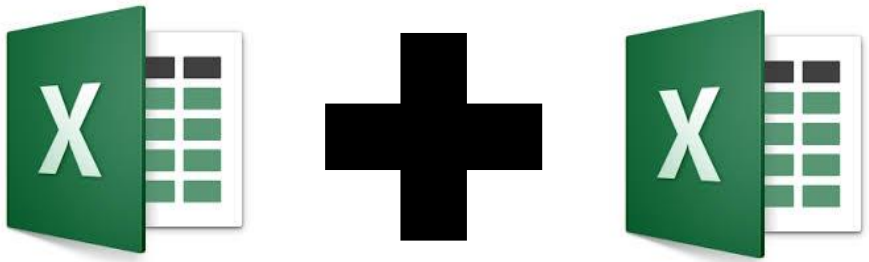
1.팀원 소개 및 프로젝트 설명
2.데이터 탐색
3.데이터 정제
4.데이터 모델링
5.마케팅 방안
6.개선점 및 제언

외부 데이터(출처)

1.주식 데이터

2014~2015년 휴일 날짜를 알아보기위해 롯데 계열사의 주식 데이터를 활용.

※출처-구글 파이낸스



2. 날씨 데이터

상품 구매 날짜를 기준으로 날씨 데이터를 불러옴.
(강수 여부,체감 온도 등)

※출처-기상청

3.지점 위경도 데이터

고객들의 우편번호와 매칭 할 수 있는 날씨 관측 지점의 위경도 데이터 활용.

※출처-기상청

파생 변수 설정

1. RFM 지수

1)RFM 지수란?

구매가능성이 높은 고객에게 집중적으로 마케팅 전략을 실행하고 불필요한 자원 낭비를 방지 함으로써 근본적인 매출액 증대보다는 수익을 창출하기 위한 모형.

R:Recency(최근성) – 가장 최근에 구매한 일자. 최근일수록 높은 점수를 부여.

F:Frequnecy(빈도성) – 아이템을 총 구매한 횟수 . 구매 횟수가 클수록 높은 점수를 부여.

M:Monetary(총구매액) – 아이템을 총 구매한 금액. 구매 금액이 클수록 높은 점수를 부여.

각각 5점 척도로 5*5*5=125의 경우의 수가 나온다.
(quantile 함수를 이용하여 균등하게 구간을 나눔.)

2)RFM 공식

RFM=(0.15*R+0.5*F+0.35*M)*0.2

※구매횟수를 가장 큰 비중을 차지한다고 생각하여 0.5를 할당하고 그 다음을 총 구매액, 그리고 최근일자순으로 하였다.

파생 변수 설정

2.고객당 구매 횟수

행	고객 번호	구매 횟수
1	1	681
2	2	676
3	3	490
4	4	533
5	5	426
6	6	821
7	7	531
8	8	716
9	9	649

3.구매 시간 데이터-3개의 변수 생성

- 출근 전 구매 변수

-구매 시간 12시 ~ 8시 기준으로 생성.

- 근무 시간 중 구매 변수

-구매 시간 9시 ~ 17시 기준으로 생성.

- 퇴근 후 구매 변수

-구매 시간 18시 ~ 23시 기준으로 생성.

4.휴일 데이터(외부 데이터)

※구매 상품TR의 구매 일자를 기준으로 하도록 함.

- 평일 구매 변수 – 평일을 기준으로 함.
- 주말 및 하루 공휴일 변수 – 주말 및 하루의 휴일 만을 기준으로 함.
- 3일 이상의 연휴– 추석,설날 등 3일 이상의 연속적 휴일을 기준으로 함.

LOTTE

L.POINT

더미 변수 설정

1.최근 온라인 사용(3개월 기준)

고객이 최근에 온라인(어플 및 인터넷)를 이용하였을 경우 1, 아니면 0 으로 할당.

2.경쟁사 이용 여부

고객이 경쟁사를 이용하였을 경우 1, 아니면 0 으로 할당.

고객번호	경쟁사 이용	최근 인터넷 사용
1	1	0
2	1	0
3	0	0
4	1	0
5	0	0
6	0	0
7	1	0
8	1	0
9	0	0
10	0	0
11	1	0
12	0	0
13	0	0
14	0	0

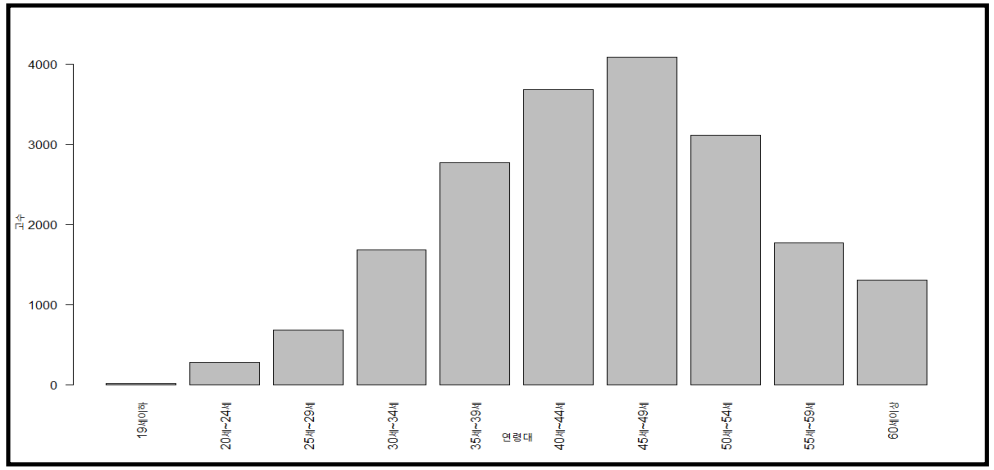
결측치 확인

데이터 탐색 결과 고객 DEMO 파일에 거주지역 정보가 결측값을 가진 것으로 확인 되었다.

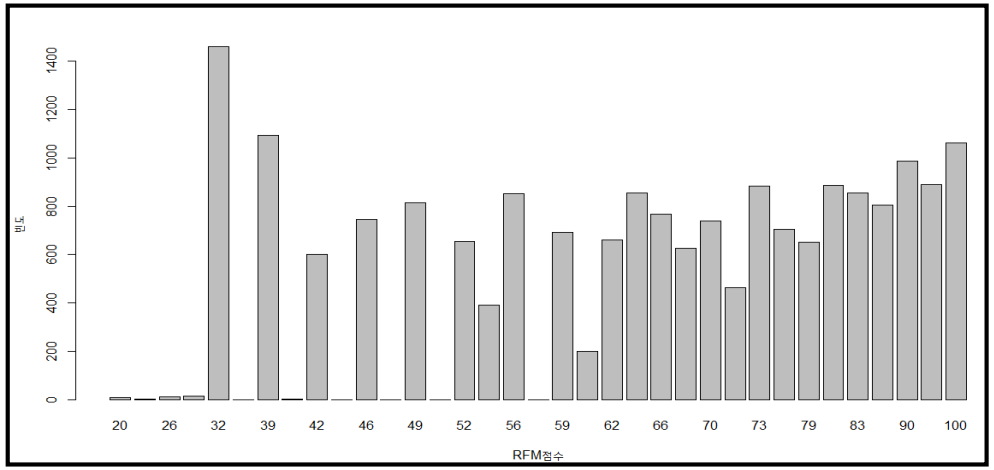
```
> summary(is.na(A1))
  고객번호          성별           연령대          거주지역
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:19383     FALSE:19383     FALSE:19383     FALSE:19205
 NA's :0         NA's :0         NA's :0         TRUE :178
                                                 NA's :0
```

```
> comple<-A1[!complete.cases(A1),]
> comple
      고객번호 성별     연령대 거주지역
22          22    F  60세이상       NA
516        516    F  60세이상       NA
940        940    F 55세~59세       NA
952        952    M 55세~59세       NA
1084      1084    F 55세~59세       NA
1108      1108    M 55세~59세       NA
1268      1268    F 55세~59세       NA
1439      1439    F 55세~59세       NA
1941      1941    F 50세~54세       NA
1983      1983    F 50세~54세       NA
2001      2001    F 50세~54세       NA
2442      2442    F 50세~54세       NA
3048      3048    F 50세~54세       NA
3116      3116    F 50세~54세       NA
3183      3183    F 50세~54세       NA
3264      3264    F 50세~54세       NA
3328      3328    M 50세~54세       NA
3931      3931    M 45세~49세       NA
3995      3995    F 45세~49세       NA
4208      4208    F 45세~49세       NA
4482      4482    F 45세~49세       NA
4815      4815    F 45세~49세       NA
5236      5236    M 45세~49세       NA
5404      5404    F 45세~49세       NA
5450      5450    F 45세~49세       NA
5662      5662    F 45세~49세       NA
5874      5874    F 40세~44세       NA
5900      5900    F 40세~44세       NA
6273      6273    F 40세~44세       NA
```

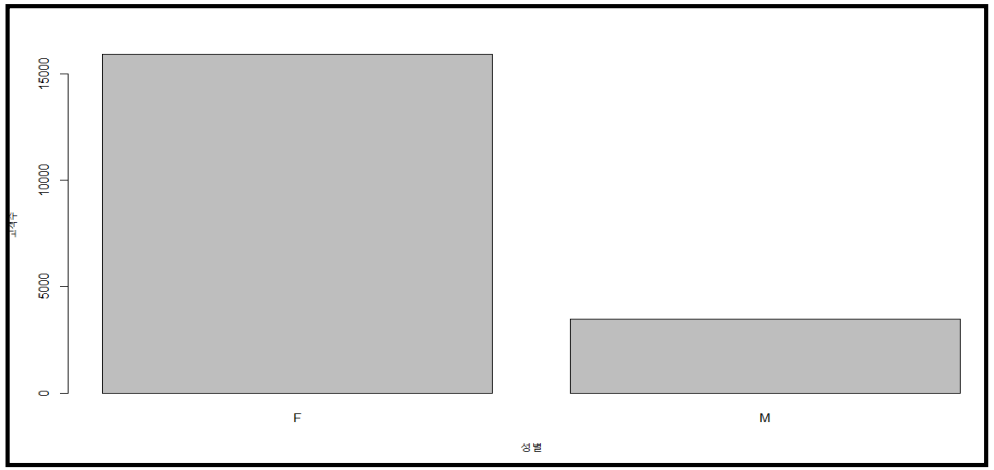
고객 특성 분포 시각화



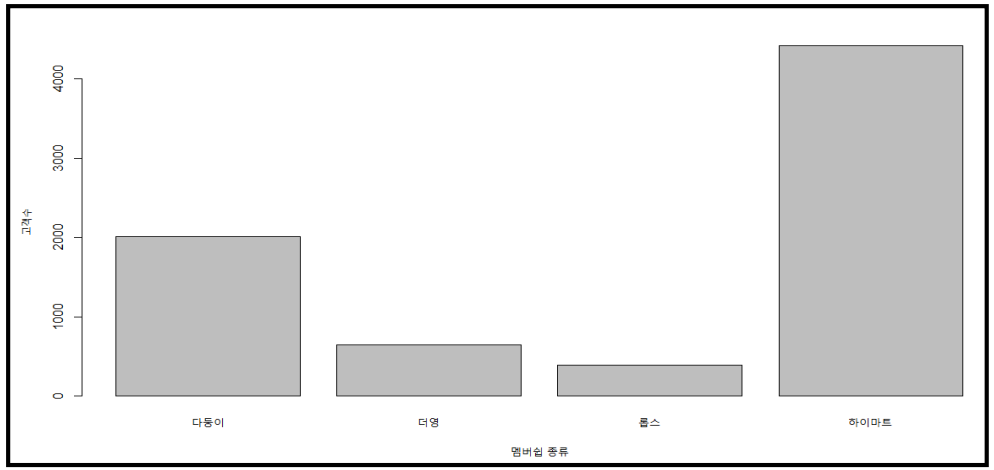
1)연령대 분포



2)RFM 점수 분포



3)성별 분포



4)멤버쉽 이용 분포

LOTTE

L.POINT

물품 구매 내역 분포

	영수증번호	고객번호	구매일자	구매시간
1	8664000	17218	20140222	20
2	8664000	17218	20140222	20
3	8664000	17218	20140222	20
4	8664000	17218	20140222	20
5	8664001	17674	20140222	22
6	8664001	17674	20140222	22
7	8664002	14388	20140222	23
8	8664002	14388	20140222	23
9	8664002	14388	20140222	23
10	8664003	15773	20140222	21
11	8664003	15773	20140222	21
12	8664003	15773	20140222	21
13	8664003	15773	20140222	21
14	8664003	15773	20140222	21
15	8664003	15773	20140222	21
16	8664003	15773	20140222	21
17	8664003	15773	20140222	21
18	8664004	17829	20140222	10
19	8664004	17829	20140222	10
20	8664004	17829	20140222	10
21	8664019	15349	20140222	22
22	8664005	17675	20140222	10
23	8664005	17675	20140222	10
24	8664005	17675	20140222	10
25	8664005	17675	20140222	10
26	8664005	17675	20140222	10
27	8664005	17675	20140222	10

중복 제거

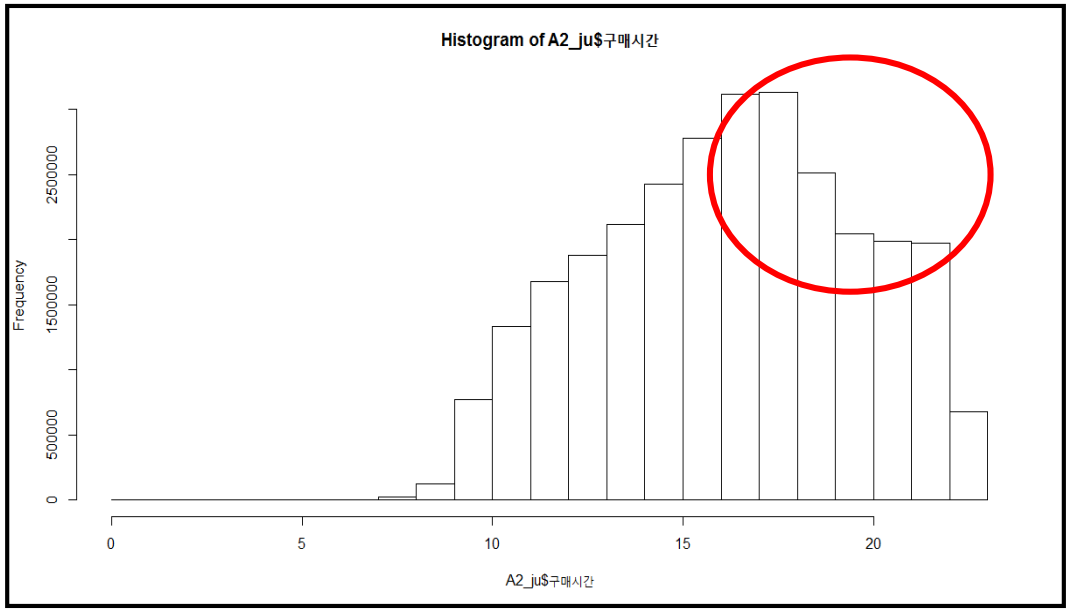
	영수증번호	고객번호	구매일자	구매시간
1	8664000	17218	20140222	20
5	8664001	17674	20140222	22
7	8664002	14388	20140222	23
10	8664003	15773	20140222	21
18	8664004	17829	20140222	10
21	8664019	15349	20140222	22
22	8664005	17675	20140222	10
29	8664006	2041	20140222	13
39	8664007	11303	20140222	13
54	8664008	15706	20140222	16
62	8664009	15706	20140222	17
63	8664010	11074	20140222	18
65	8664034	15788	20140222	18
67	8779343	6285	20140530	18
73	8779344	6285	20140530	18
76	8779345	6285	20140530	18
78	8779346	1277	20140530	19
84	6927879	7221	20150824	13
89	8744849	16643	20150615	16
91	8744853	14647	20150615	18
98	8744854	263	20150615	19
99	8744839	18255	20150615	13

LOTTE

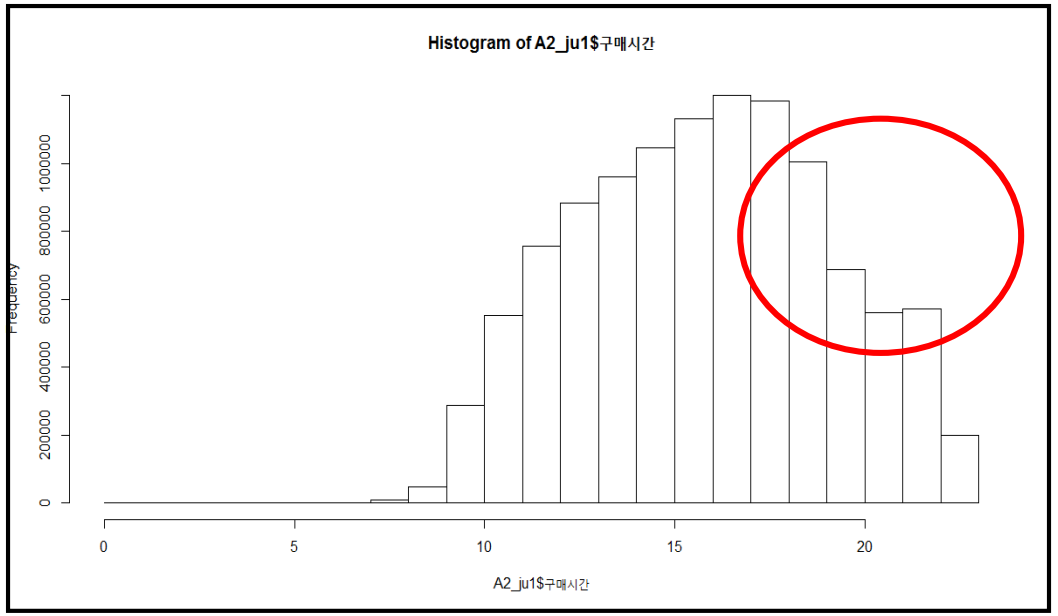
L.POINT

물품 구매 시간 비교

중복 제거전 구매 시간 빈도



중복 제거후 구매 시간 빈도

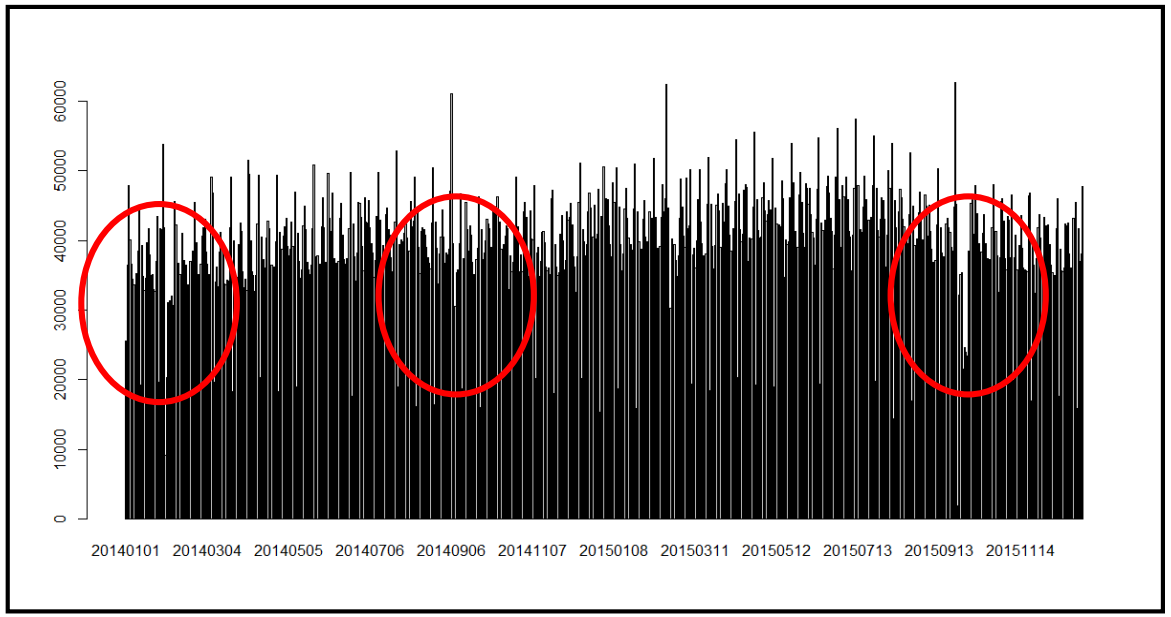


영수증 번호 중복을 제거 한 결과 퇴근 시간인 18시~22시 사이 영수증 번호 빈도수가 급격하게 줄어든 것을 알 수 있다.

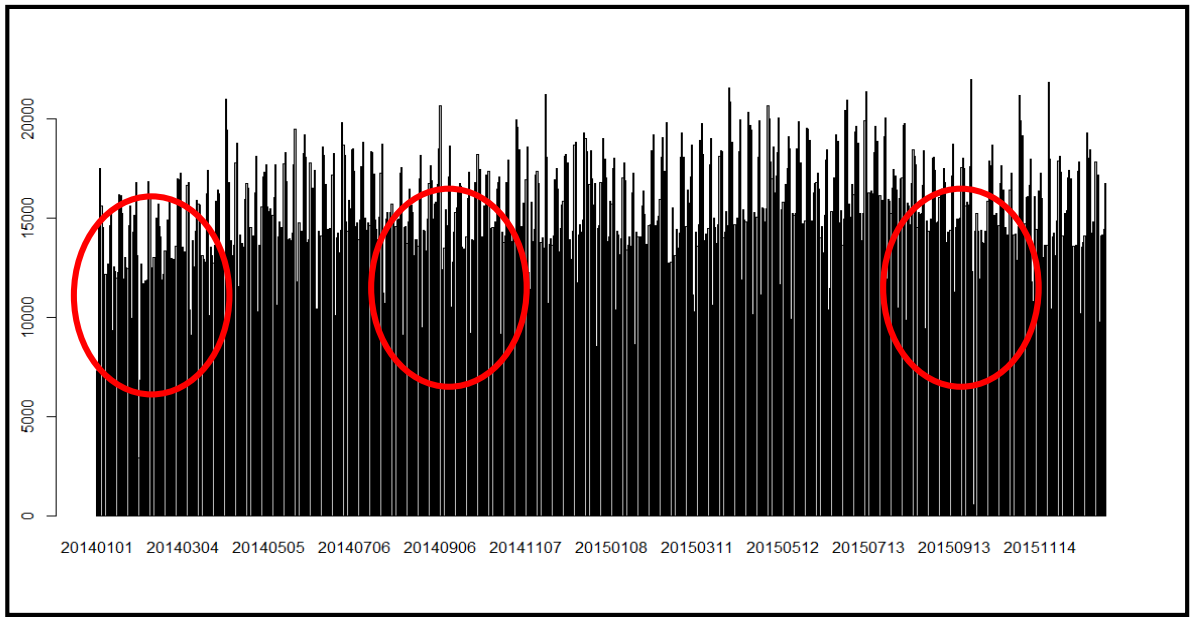
이를 통해, 퇴근 시간일 경우 다량 구매를 하는 고객이 많다는 것을 확인할 수 있다.

물품 구매 일자 비교

중복 제거전 구매 일자 빈도



중복 제거후 구매 일자 빈도



영수증 번호 중복을 제거 한 결과 설날과 추석 시즌 쯤에 큰 차이가 있다는 것을 확인 하였다.

이를 통해, 장기 연휴일 경우 사람들이 많은 물품을 구매하지 않는다고 판단 하였다. 따라서 휴일과 평일 장기 연휴 등의 외부 데이터를 설정하도록 하였다.

LOTTE

L.POINT

CONTENTS

1.팀원 소개 및 프로젝트 개요
2.데이터 탐색
3.데이터 정제
4.데이터 모델링
5.마케팅 방안

결측치 처리

1)탐색 결과 고객의 거주지역에 결측치가 있는 것이 확인 되었다.

2)기상데이터를 사용하는 것에 있어서,우편번호와 기상데이터의 관측지점을 합치기로 하였다.
(*가정-사람들은 자신의 거주지역 근처의 점포를 가장 많이 이용할 것이며,따라서 거주지역내의 날씨정보를 통해 모델링을 실시한다.)

3)따라서 고객의 거주지역에 결측치가 있으면 안된다고 판단하여 구매내역 TR의 점포코드로 역 추적을 실시 하였다.

4)점포 코드에 대한 정보는 없지만 앞선 가정(*)을 통해 역 추적을 하도록 하였다.

결측치 처리

1)탐색 결과 고객의 거주지역에 결측치가 있는 것이 확인 되었다.

```
> summary(is.na(A1))
  고객번호          성별            연령대          거주지역
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:19383     FALSE:19383     FALSE:19383     FALSE:19205
 NA's :0         NA's :0         NA's :0         TRUE :178
                                                 NA's :0
```

```
> comple<-A1[!complete.cases(A1),]
> comple
      고객번호 성별     연령대 거주지역
22          22    F  60세이상       NA
516        516    F  60세이상       NA
940        940    F 55세~59세       NA
952        952    M 55세~59세       NA
1084      1084    F 55세~59세       NA
1108      1108    M 55세~59세       NA
1268      1268    F 55세~59세       NA
1439      1439    F 55세~59세       NA
1941      1941    F 50세~54세       NA
1983      1983    F 50세~54세       NA
2001      2001    F 50세~54세       NA
2442      2442    F 50세~54세       NA
3048      3048    F 50세~54세       NA
3116      3116    F 50세~54세       NA
3183      3183    F 50세~54세       NA
3264      3264    F 50세~54세       NA
3328      3328    M 50세~54세       NA
3931      3931    M 45세~49세       NA
3995      3995    F 45세~49세       NA
4208      4208    F 45세~49세       NA
4482      4482    F 45세~49세       NA
4815      4815    F 45세~49세       NA
5236      5236    M 45세~49세       NA
5404      5404    F 45세~49세       NA
5450      5450    F 45세~49세       NA
5662      5662    F 45세~49세       NA
5874      5874    F 40세~44세       NA
5900      5900    F 40세~44세       NA
6273      6273    F 40세~44세       NA
```

결측치 처리

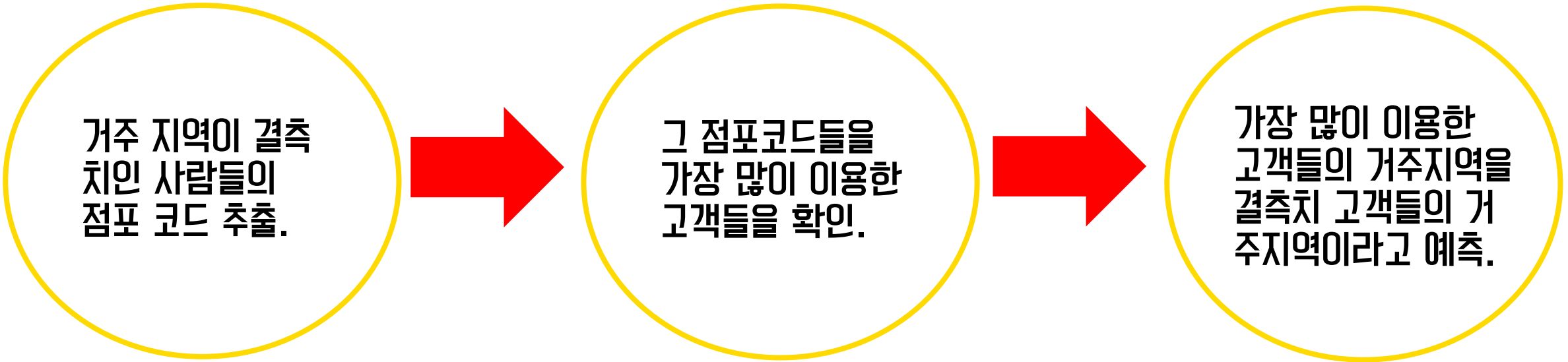
2)기상데이터를 사용하는 것에 있어서,우편번호와 기상데이터의 관측지점을 합치기로 하였다.

	고객번호	성별	연령대	거주지역
1	1	M	60세이상	60
2	2	M	60세이상	100
3	3	M	60세이상	33
4	4	F	60세이상	16
5	5	M	60세이상	100
6	6	F	60세이상	240
7	7	F	60세이상	36
8	8	M	60세이상	10
9	9	F	60세이상	100
10	10	F	60세이상	24
11	11	M	60세이상	100
12	12	M	60세이상	43
13	13	F	60세이상	210
14	14	F	60세이상	43
15	15	F	60세이상	24
16	16	F	60세이상	460
17	17	M	60세이상	43

	A	B	C	D
1	location	구매일자	rain	체감온도
2	95	20140101	0	2.391871
3	108	20140101	0	3.543739
4	112	20140101	0.1	3.896225
5	133	20140101	0	5.684926
6	140	20140101	0	5.644018
7	143	20140101	0	6.622128
8	152	20140101	0	6.576304
9	156	20140101	0	6.03687
10	159	20140101	0	6.589386
11	184	20140101	0	10.60926
12	232	20140101	0	3.812068
13	252	20140101	0	4.601305
14	264	20140101	0	5.524875
15	272	20140101	0	3.288703
16	95	20140102	0	-0.4729
17	108	20140102	0	2.013202
18	112	20140102	0	2.278214
19	133	20140102	0	4.3968
20	140	20140102	0	5.109537
21	143	20140102	0	6.355135

결측치 처리

3)따라서 고객의 거주지역에 결측치가 있으면 안된다고 판단하여 구매내역 TR의 점포 코드로 역 추적을 실시 하였다.



예시)22번 고객이 NA.

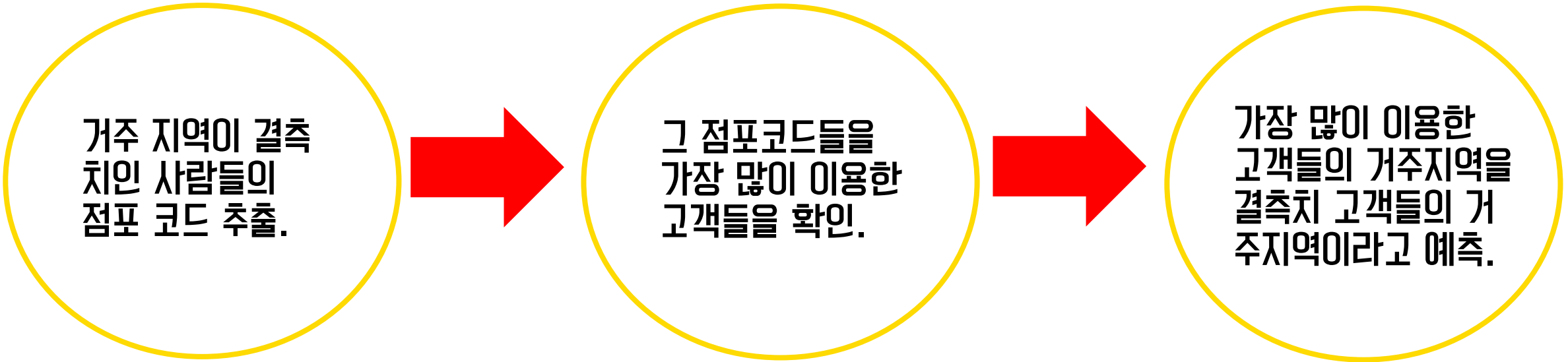
예시)22번 고객이 가장 많이 이용한 곳은 1번 점포이다.

예시)1번 점포를 가장 많이 이용한 고객의 거주 지역은 25번이다. 따라서 22번을 25번 거주 지역에 산다고 예측한다.

LOTTE

결측치 처리

3)따라서 고객의 거주지역에 결측치가 있으면 안된다고 판단하여 구매내역 TR의 점포 코드로 역 추적을 실시 하였다.



예시)22번 고객이 NA.

예시)22번 고객이 가장 많이 이용한 곳은 1번 점포이다.

예시)1번 점포를 가장 많이 이용한 고객의 거주 지역은 25번이다. 따라서 22번을 25번 거주 지역에 산다고 예측한다.

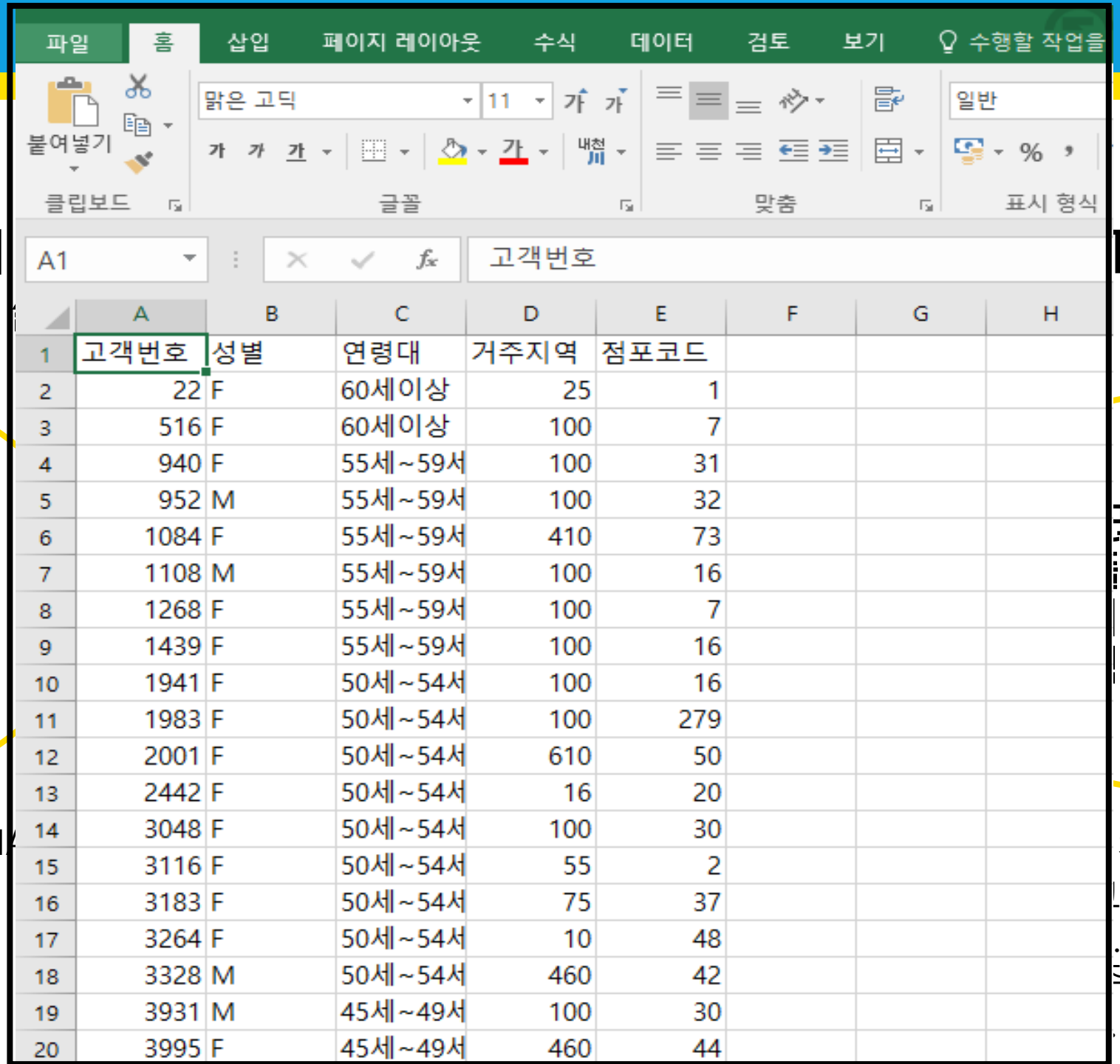
LOTTE

결측치 처리

3)따라서 고객의 코드로 역 추적을

TR의 점포

거주 지역이 결측치인 사람들의 점포 코드 추출.



파일 홈 삽입 페이지 레이아웃 수식 데이터 검토 보기 수행할 작업을

맑은 고딕 11 일반

붙여넣기 클립보드 글꼴 맞춤 표시 형식

A1 고객번호

	A	B	C	D	E	F	G	H
1	고객번호	성별	연령대	거주지역	점포코드			
2	22	F	60세이상	25	1			
3	516	F	60세이상	100	7			
4	940	F	55세~59서	100	31			
5	952	M	55세~59서	100	32			
6	1084	F	55세~59서	410	73			
7	1108	M	55세~59서	100	16			
8	1268	F	55세~59서	100	7			
9	1439	F	55세~59서	100	16			
10	1941	F	50세~54서	100	16			
11	1983	F	50세~54서	100	279			
12	2001	F	50세~54서	610	50			
13	2442	F	50세~54서	16	20			
14	3048	F	50세~54서	100	30			
15	3116	F	50세~54서	55	2			
16	3183	F	50세~54서	75	37			
17	3264	F	50세~54서	10	48			
18	3328	M	50세~54서	460	42			
19	3931	M	45세~49서	100	30			
20	3995	F	45세~49서	460	44			

많이 이용한
의 거주지역을
고객들의 거
이라고 예측.

예시)22번 고객이 N

점포를 가장 많이
객의 거주 지역은
. 따라서 22번을
지역에 산다고

LOTTE

데이터 정제(성별)

데이터 탐색에서 볼 수 있듯이, 성별이 불균형 데이터이며 남녀 별로 구매품목이 같은지 확인 해보았다.

예시)남녀별 품목

	A010101	A010102	A010103	A010104	A010105	A010106	A010201	A010202	A010203	A010204	A010205	A010206	A010207	A010301	A010302	A010303
A2_male_cust	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A2_Female_cust	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

다른 경우221/4386 =약 5%

귀무 가설:남녀간의 구매 품목은 차이가 없다.
대립 가설:남녀간의 구매 품목은 차이가 있다.
2년 동안 남자와 여자의 구매 품목이 5% 정도 같다는 것이 확인 되었다. 따라서 유의수준 5%에서 남녀간에 구매 품목에는 차이가 없다는 가설을 채택 하도록 한다. 성별 변수가 유의미 하지 않다고 판단하여 성별 변수를 제거하기로 하였다.

고객 특성 데이터 생성-구매 데이터와 고객데이터를 결합

고객번호	점포코드	구매일자	구매시간	구매금액
17218	44	20140222	20	2420
17218	44	20140222	20	1070
17218	44	20140222	20	8060
17218	44	20140222	20	6000
17674	44	20140222	22	1120
17674	44	20140222	22	1200
14388	44	20140222	23	5290
14388	44	20140222	23	5960
14388	44	20140222	23	9900
15773	44	20140222	21	970
15773	44	20140222	21	3200
15773	44	20140222	21	6600
15773	44	20140222	21	2550
15773	44	20140222	21	4480
15773	44	20140222	21	1080

+

설정한 파생 변수들

=

1	RFM지수	출근전구매	근무시간 구매	퇴근후 구매	주중 구매	2일이하의 휴일	3일 이상의 연휴
2	90	0	581	100	486	160	35
3	90	0	613	63	438	198	40
4	42	0	489	1	394	84	12
5	66	0	512	21	408	106	19
6	39	0	371	55	289	117	20
7	93	0	819	2	687	93	41
8	80	0	357	174	326	181	24
9	100	0	614	102	373	298	45
10	90	0	573	76	410	201	38
11	86	0	283	489	438	287	47
12	90	0	440	157	329	222	46
13	32	0	292	3	210	61	24
14	76	0	221	455	386	230	60
15	83	0	638	52	497	150	43
16	93	0	508	211	503	187	29
17	39	0	408	3	324	73	14
18	80	0	424	96	286	207	27
19	70	0	407	94	331	142	28
20	53	0	384	53	272	147	18

LOTTE

L.POINT

최종 데이터 셋

1)총 6개의 데이터 셋에서 파생변수 및 더미 변수등을 생성하고 불필요한 변수를 제거하였다.

2)기상 데이터를 사용하려 하였으나 강수량 및 체감온도등을 구분 짓는 기준이 애매하고 변수가 모델링 향상에 유의미 하지 않다고 판단하여 제거 하였다.

3)최종 변수 선택 (총15개의 변수)
고객 번호, 나이대,멤버쉽 종류(4가지), RFM 지수 , 구매 시간대 변수 (출근전,근무시간,퇴근-3가지)
고객 구매 일자 변수(주중, 주말 및 2일 이하의 휴일 , 3일이상의 연휴 -3가지.) , 강수량 여부, 기온

CONTENTS

1.팀원 소개 및 프로젝트 개요
2.데이터 탐색
3.데이터 정제
4.데이터 모델링
5.개선점 및 제언

모델 검증 시트

모델 검증 시트의 3가지 줄을 각각 다른 기법을 적용하여 상품을 추천하도록 한다.

유저기반 필터링

1. 1번째 줄 (Fundamental Line)

-가장 기본적인 줄로써 2016년 1월의 구매 내역을 예측하기 위해 2014,15년의 겨울에서 가장 많이 팔린 항목을 추천한다.

-구매량이 적어서 추천 상품이 나오지 않는 경우 2번째 줄의 군집분석을 통하여 같은 군집 상품을 추천 하도록 한다.

2. 2번째 줄 (Categorical Cluster Line)

- 고객 특성에 따른 자카드 계수를 통해 거리를 계산하고 군집 분석함.
- 군집당 가장 많이 팔린 상품을 추천하되 1번째 줄과 겹칠 경우 2번째로 많이 팔린 상품을 추천한다.

3. 3번째 줄 (Continuous Cluster+item based filtering)

- 더미 변수가 아닌 연속형 변수를 통해 유클리디안 거리 유사성 척도를 계산.
- 군집을 형성한 뒤에 아이템 상품을 추천,
- 1번줄과 2번줄의 항목이 3번 줄과 겹칠 경우 아이템 메트릭스를 통해 유사성 척도로 추천.

LOTTE

L.POINT

1번째 줄 (Fundamental Line)

1) 구매 일자를 기준으로 2014년의 1월 2월 12월을 2014년의 겨울 2015년 1월 2월 12월을 2015년의 겨울 이라고 가정 하였다.
(각각 12월들이 1월,2월과 떨어져 있지만 13년 12월의 데이터가 없고, 2년동안 추세가 급격하게 변했다고 볼 수 없기때문에 이렇게 가정 하였다.)

2) 아이템 추천 방법은 고객 번호와 구매내역 소분류코드를 **연관성 분석** 기법을 사용하였다.

예시)
```
winAR2.2 <- apriori(data=winAR2.1,
                    parameter = list(support = 0.000001, confidence = 0.005 ,minlen=2))
```

2014년과 2015년의 겨울 그리고 2015년 전체를 연관성 분석을 실시하여 minlen 값을 2로 지정 한뒤 lift 값이 가장 높은 3가지 경우를 내림차순으로 정리하여 각각의 추천 물품이 동일하게 나올 경우 추천하였다.
(*lift 값이 클 수록 고객 번호와 소분류코드가 연관성이 강한 것이다. 즉, 고객이 이 상품을 가장 많이 구매하였다고 할 수 있다.)

LOTTE

L.POINT

1번째 줄 정제 과정

STEP1

14년 겨울,15년 겨울,15년 전체 중 같은 상품이 묶일 경우 추천.

STEP2

만약 그러한 상품이 없을 경우, 15년 전체의 lift 값이 가장 높은 상품을 추천

STEP3

여기서도 걸러지지 않을 경우 2번 줄의 군집 분석을 토대로 같은 군집에 묶인 상품을 추천.

1단계에서 16000명정도의 고객들이 상품이 추천된다.

2단계에서 남은 3000명 중 2700명의 고객들에게 상품이 추천된다.

3단계에서 남은 300명의 고객들에게 모든 상품을 추천한다.

LOTTE

L.POINT

1번째 줄 -STEP1

2014년 겨울

	A	B	C
1	lhs	rhs	lift
2	{고객번호=1}	{소분류코드=A050501}	34.57253
3	{고객번호=1}	{소분류코드=A010707}	29.64597
4	{고객번호=1}	{소분류코드=A040222}	13.27411
5	{고객번호=2}	{소분류코드=A010301}	40.32628
6	{고객번호=2}	{소분류코드=A020306}	22.47879
7	{고객번호=2}	{소분류코드=A010710}	22.17269
8	{고객번호=3}	{소분류코드=C120101}	75.18322
9	{고객번호=3}	{소분류코드=C120601}	33.58477
10	{고객번호=3}	{소분류코드=C170701}	20.14663
11	{고객번호=4}	{소분류코드=A010704}	178.7646
12	{고객번호=4}	{소분류코드=A010710}	30.5244
13	{고객번호=4}	{소분류코드=C010206}	13.201
14	{고객번호=5}	{소분류코드=A010623}	86.6107
15	{고객번호=5}	{소분류코드=A010302}	27.93528
16	{고객번호=5}	{소분류코드=A010901}	26.13819
17	{고객번호=6}	{소분류코드=B010106}	25.6523
18	{고객번호=6}	{소분류코드=B540202}	21.34052
19	{고객번호=6}	{소분류코드=B050102}	20.83798

추천

2015년 겨울

	A	B	C
1	lhs	rhs	lift
2	{고객번호=1}	{소분류코드=A020208}	43.30341
3	{고객번호=1}	{소분류코드=A040209}	39.10862
4	{고객번호=1}	{소분류코드=A010606}	24.23275
5	{고객번호=2}	{소분류코드=A040902}	56.2157
6	{고객번호=2}	{소분류코드=A020306}	53.86966
7	{고객번호=2}	{소분류코드=A040601}	48.20688
8	{고객번호=3}	{소분류코드=C120601}	33.59476
9	{고객번호=3}	{소분류코드=C120101}	32.23381
10	{고객번호=3}	{소분류코드=C070101}	12.30393
11	{고객번호=4}	{소분류코드=A010704}	144.2159
12	{고객번호=4}	{소분류코드=A010647}	56.02374
13	{고객번호=4}	{소분류코드=A010710}	19.61814
14	{고객번호=5}	{소분류코드=A010302}	31.50863
15	{고객번호=5}	{소분류코드=A010901}	26.81432
16	{고객번호=5}	{소분류코드=A010402}	26.48266
17	{고객번호=6}	{소분류코드=B720103}	58.69409
18	{고객번호=6}	{소분류코드=B790310}	30.188
19	{고객번호=6}	{소분류코드=B100502}	21.09156

2015년 전체

	A	B	C
1	lhs	rhs	lift
2	{고객번호=1}	{소분류코드=A010606}	27.00721
3	{고객번호=1}	{소분류코드=A040222}	17.59293
4	{고객번호=1}	{소분류코드=A010404}	8.967757
5	{고객번호=2}	{소분류코드=A010403}	11.58735
6	{고객번호=2}	{소분류코드=A010402}	10.52102
7	{고객번호=2}	{소분류코드=A010101}	9.816732
8	{고객번호=3}	{소분류코드=C110701}	60.06404
9	{고객번호=3}	{소분류코드=C120101}	53.43559
10	{고객번호=3}	{소분류코드=C120601}	25.44148
11	{고객번호=4}	{소분류코드=A010704}	106.4218
12	{고객번호=4}	{소분류코드=A010647}	37.59601
13	{고객번호=4}	{소분류코드=A010710}	15.64681
14	{고객번호=5}	{소분류코드=A010302}	31.61294
15	{고객번호=5}	{소분류코드=A010402}	21.18782
16	{고객번호=5}	{소분류코드=A010404}	17.13709
17	{고객번호=6}	{소분류코드=A041003}	49.79141
18	{고객번호=6}	{소분류코드=B100502}	33.39398
19	{고객번호=6}	{소분류코드=A040502}	31.87213

LOTTE

L.POINT

1번째 줄 -STEP2

겹치는 상품이 하나도 없을 경우 가장 lift 값이 높은 상품 추천

2014년 겨울

	A	B	C
1	lhs	rhs	lift
2	{고객번호=1}	{소분류코드=A050501}	34.57253
3	{고객번호=1}	{소분류코드=A010707}	29.64597
4	{고객번호=1}	{소분류코드=A040222}	13.27411
5	{고객번호=2}	{소분류코드=A010301}	40.32628
6	{고객번호=2}	{소분류코드=A020306}	22.47879
7	{고객번호=2}	{소분류코드=A010710}	22.17269
8	{고객번호=3}	{소분류코드=C120101}	75.18322
9	{고객번호=3}	{소분류코드=C120601}	33.58477
10	{고객번호=3}	{소분류코드=C170701}	20.14663
11	{고객번호=4}	{소분류코드=A010704}	178.7646
12	{고객번호=4}	{소분류코드=A010710}	30.5244
13	{고객번호=4}	{소분류코드=C010206}	13.201
14	{고객번호=5}	{소분류코드=A010623}	86.6107
15	{고객번호=5}	{소분류코드=A010302}	27.93528
16	{고객번호=5}	{소분류코드=A010901}	26.13819
17	{고객번호=6}	{소분류코드=B010106}	25.6523
18	{고객번호=6}	{소분류코드=B540202}	21.34052
19	{고객번호=6}	{소분류코드=B050102}	20.83798

2015년 겨울

	A	B	C
1	lhs	rhs	lift
2	{고객번호=1}	{소분류코드=A020208}	43.30341
3	{고객번호=1}	{소분류코드=A040209}	39.10862
4	{고객번호=1}	{소분류코드=A010606}	24.23275
5	{고객번호=2}	{소분류코드=A040902}	56.2157
6	{고객번호=2}	{소분류코드=A020306}	53.86966
7	{고객번호=2}	{소분류코드=A040601}	48.20688
8	{고객번호=3}	{소분류코드=C120601}	33.59476
9	{고객번호=3}	{소분류코드=C120101}	32.23381
10	{고객번호=3}	{소분류코드=C070101}	12.30393
11	{고객번호=4}	{소분류코드=A010704}	144.2159
12	{고객번호=4}	{소분류코드=A010647}	56.02374
13	{고객번호=4}	{소분류코드=A010710}	19.61814
14	{고객번호=5}	{소분류코드=A010302}	31.50863
15	{고객번호=5}	{소분류코드=A010901}	26.81432
16	{고객번호=5}	{소분류코드=A010402}	26.48266
17	{고객번호=6}	{소분류코드=B720103}	58.69409
18	{고객번호=6}	{소분류코드=B790310}	30.188
19	{고객번호=6}	{소분류코드=B100502}	21.09156

2015년 전체

	A	B	C
1	lhs	rhs	lift
2	{고객번호=1}	{소분류코드=A010606}	27.00721
3	{고객번호=1}	{소분류코드=A040222}	17.59293
4	{고객번호=1}	{소분류코드=A010404}	8.967757
5	{고객번호=2}	{소분류코드=A010403}	11.58735
6	{고객번호=2}	{소분류코드=A010402}	10.52102
7	{고객번호=2}	{소분류코드=A010101}	9.816732
8	{고객번호=3}	{소분류코드=C110701}	60.06404
9	{고객번호=3}	{소분류코드=C120101}	53.43559
10	{고객번호=3}	{소분류코드=C120601}	25.44148
11	{고객번호=4}	{소분류코드=A010704}	106.4218
12	{고객번호=4}	{소분류코드=A010647}	37.59601
13	{고객번호=4}	{소분류코드=A010710}	15.64681
14	{고객번호=5}	{소분류코드=A010302}	31.61294
15	{고객번호=5}	{소분류코드=A010402}	21.18782
16	{고객번호=5}	{소분류코드=A010404}	17.13709
17	{고객번호=6}	{소분류코드=A041003}	49.79141
18	{고객번호=6}	{소분류코드=B100502}	33.39398
19	{고객번호=6}	{소분류코드=A040502}	31.87213

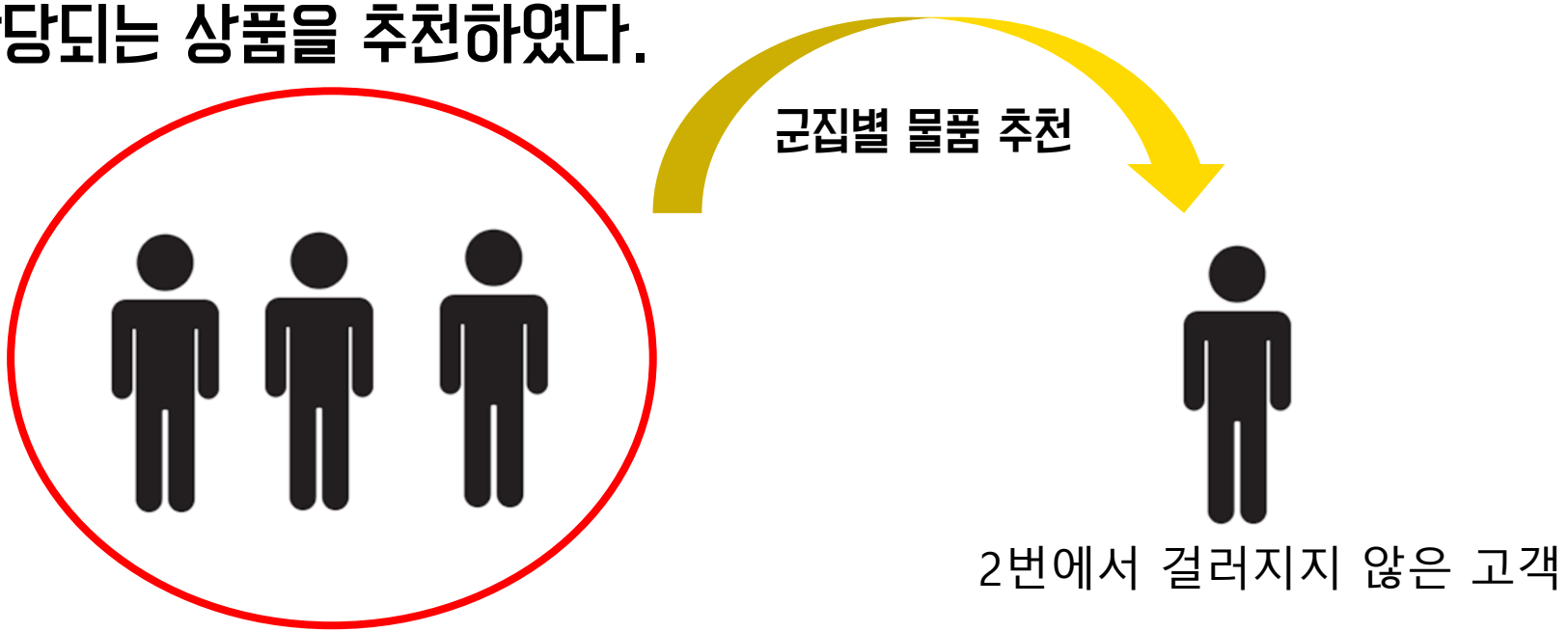
LOTTE

L.POINT

1번째 줄 –STEP3

1)2번째 과정을 해도 2015년에 물품을 구매한 횟수가 미비한 고객의 경우 상품 추천이 불가능 하다.

2) 따라서 2번에서 한 군집 분석을 통해 그러한 고객들의 군집 특성을 파악하고 군집에서 할당되는 상품을 추천하였다.



2번째줄 (Categorical Cluster Line)

1)앞선 고객 특성 데이터(구매상품 TR을 결합한 데이터)를 더미 변수로 나누었다.

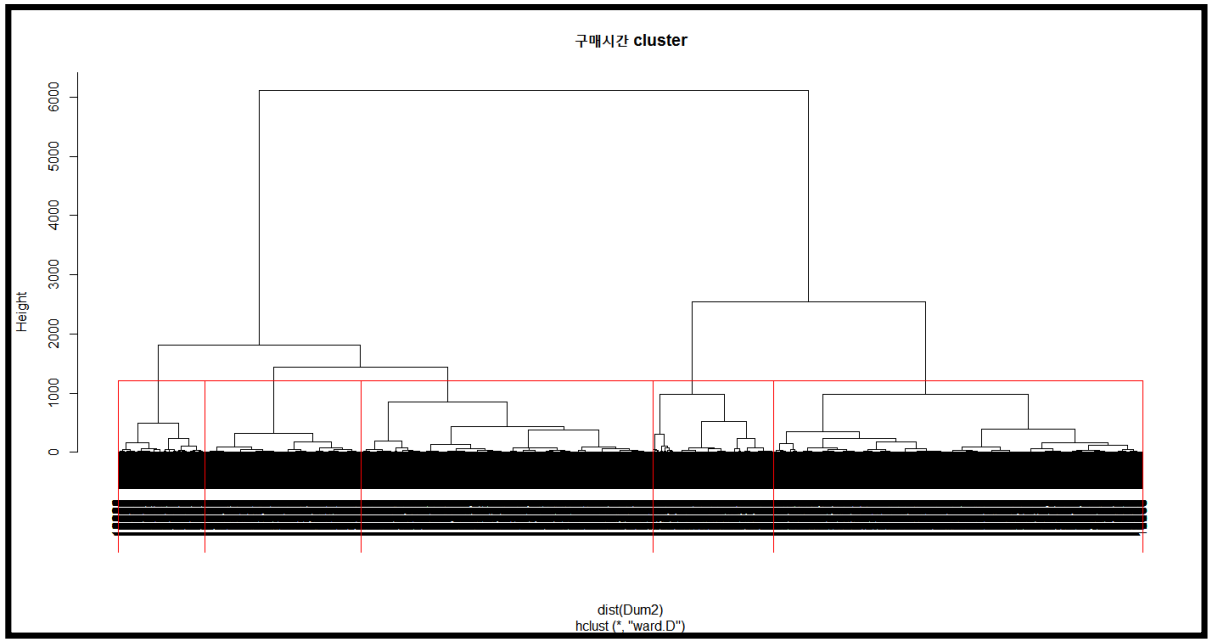
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	고객번호	gender	age1	age2	age3	age4	age5	age6	age7	age8	age9	age10	competitic	다둥이	더영	롭스	하이마트
2	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
3	2	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
4	3	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
5	4	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
6	5	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
7	6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
8	7	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
9	8	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
10	9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	10	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
12	11	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1
13	12	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
14	13	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
15	14	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
16	15	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
17	16	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
18	17	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
19	18	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
20	19	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
21	20	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0

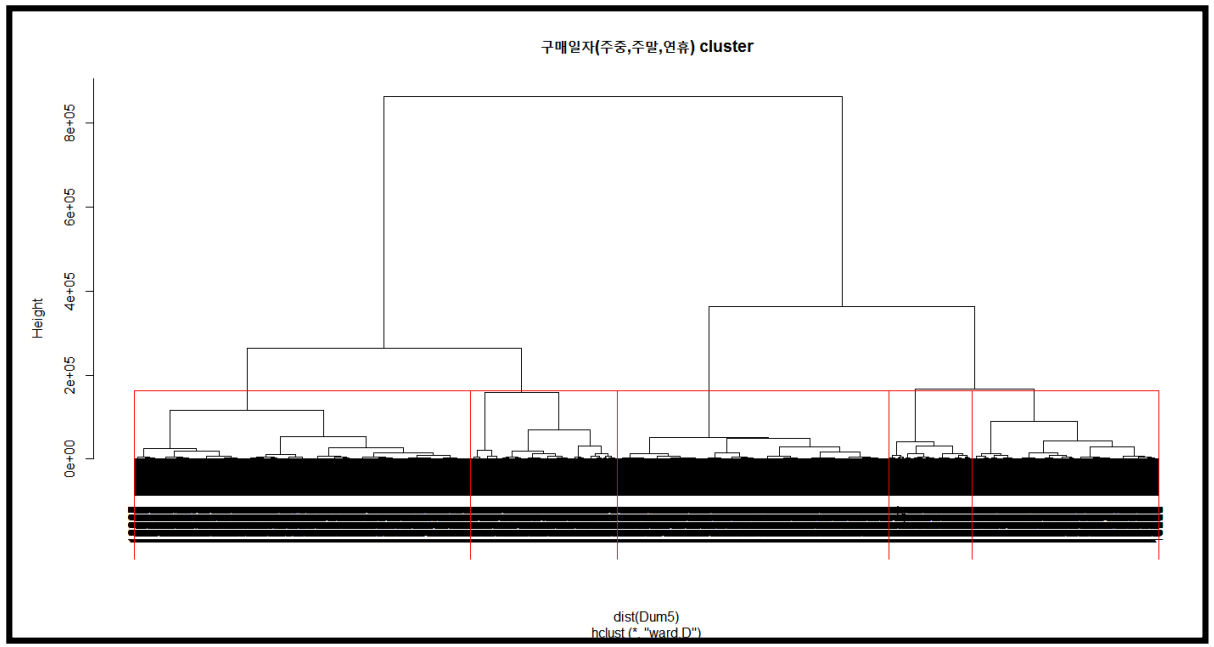
군집분석

군집분석이란? 비슷한 특성을 지닌 데이터들끼리 집단을 형성하는 것.

2번째줄 (Categorical Cluster Line)

2)구매 시간과 구매 일자 데이터는 각각 군집 분석을 통하여 5개의 군집을 형성한 뒤에 연속형 변수를 범주형 변수로 바꾸고 더미 변수를 설정하였다.





2번째줄 (Categorical Cluster Line)

2)더미 변수 속성에 맞는 거리 방법인 자카드 계수를 통해 거리를 산출.

자카드 계수(Jaccard) 기반 유사도 (1)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

- **Boolean 속성**으로 이루어진 두 개의 오브젝트 A와B에 대하여
- A와 B가 교집합으로 1의 값을 가진 속성의 개수를
- A와 B의 1의 합집합 개수로 나눈 값.

자카드 계수(Jaccard) 기반 유사도 (1)

User based 데이터 셋 (1 : 구매)

	Item1	Item2	Item3	Item4
User1	0	1	0	1
User2	0	1	1	1
User3	1	0	1	0

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

사용자 1과 사용자 2의 Jaccard 유사도를 계산하는 경우
(분모) |A∪B| 즉 두 사람이 산 상품의 합집합의 개수는 3

(분자) |A∩B| 두 사람이 산 상품의 교집합의 개수는 2
jaccard 유사도 값은 ⅔= 0.67

모든 사용자간 계산된 유사도

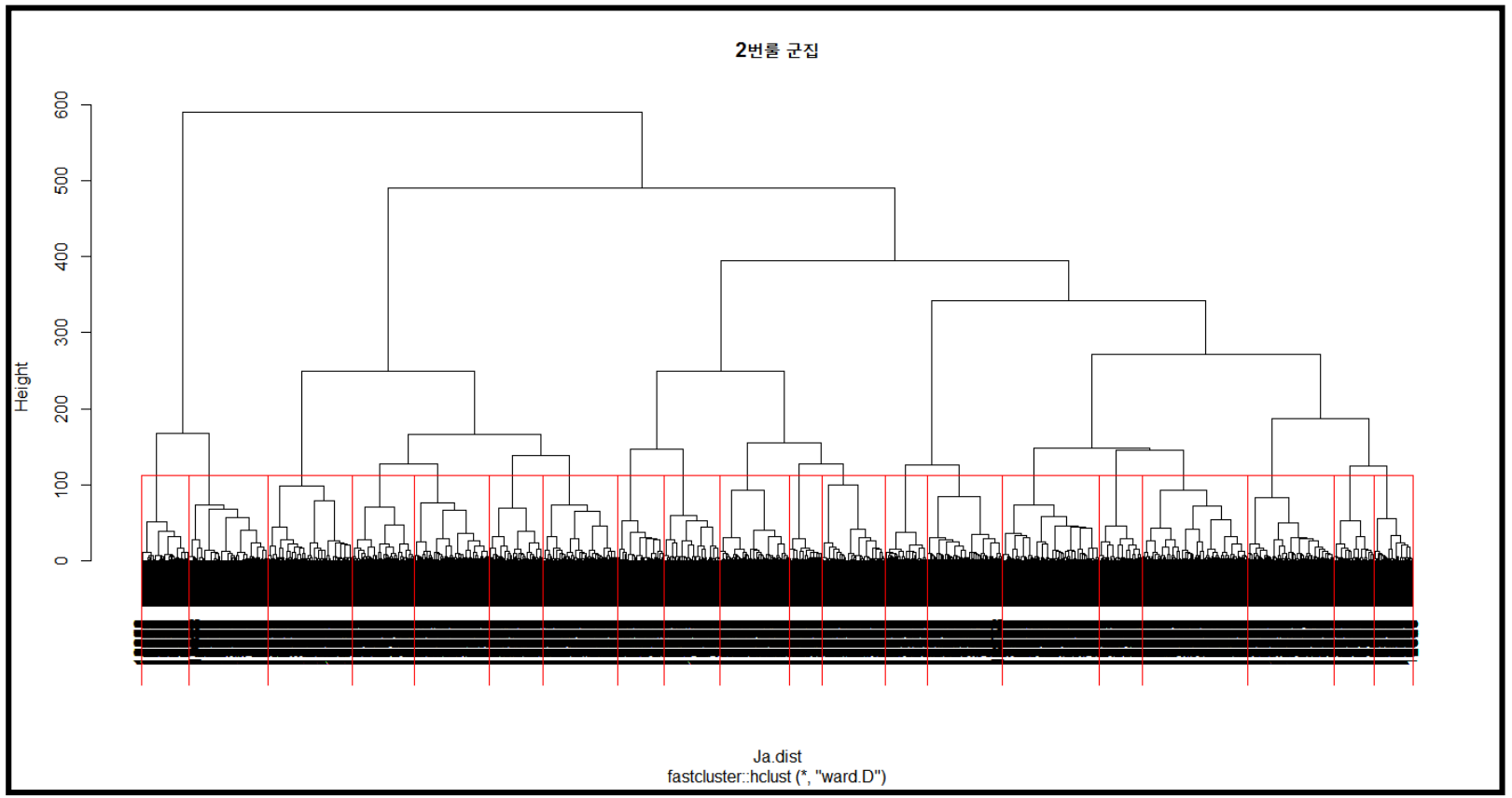
	User1	User2	User3
User1	1.0	0.67	0
User2	0.67	1.0	0.25
User3	0	0.25	1.0

출처-http://www.slideshare.net/springloops/collaborative-filtering-23732558

2번째줄 (Categorical Cluster Line)

3)자카드 계수를 통해 군집분석을 실시.

-군집의 개수는 20개가 적당하다고 판단하여 20개로 나누고 군집별로 가장 구매율이 높은 상품을 추천하였다.



3번째줄 (Continuous Cluster + item based filtering)

1)앞선 고객 특성 데이터를 군집분석을 실시 하였다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	고객번호	gender	age	competitic	최근온라인	멤버쉽사용합	date	freq	sum	RFM지수	출근전구매	근무시간 구매	퇴근후 구매	주중 구매	2일이하의 휴일	3일 이상
2	1	1	10	1	0	0	5	4	5	90	0	581	100	486	160	
3	2	1	10	1	0	0	5	4	5	90	0	613	63	438	198	
4	3	1	10	0	0	0	5	2	1	42	0	489	1	394	84	
5	4	0	10	1	0	0	5	3	3	66	0	512	21	408	106	
6	5	1	10	0	0	0	5	1	2	39	0	371	55	289	117	
7	6	0	10	0	0	0	5	5	4	93	0	819	2	687	93	
8	7	0	10	1	0	0	5	3	5	80	0	357	174	326	181	
9	8	1	10	1	0	0	5	5	5	100	0	614	102	373	298	
10	9	0	10	0	0	0	5	4	5	90	0	573	76	410	201	
11	10	0	10	0	0	0	5	5	3	86	0	283	489	438	287	
12	11	1	10	1	0	1	5	4	5	90	0	440	157	329	222	
13	12	1	10	0	0	0	5	1	1	32	0	292	3	210	61	
14	13	0	10	0	0	0	5	4	3	76	0	221	455	386	230	
15	14	0	10	0	0	0	5	4	4	83	0	638	52	497	150	
16	15	0	10	1	0	0	5	5	4	93	0	508	211	503	187	
17	16	0	10	1	0	0	5	1	2	39	0	408	3	324	73	
18	17	1	10	1	0	0	5	3	5	80	0	424	96	286	207	
19	18	0	10	0	0	0	5	2	5	70	0	407	94	331	142	
20	19	0	10	0	0	0	5	1	4	53	0	384	53	272	147	
21	20	1	10	1	0	0	5	4	4	83	0	564	106	473	164	

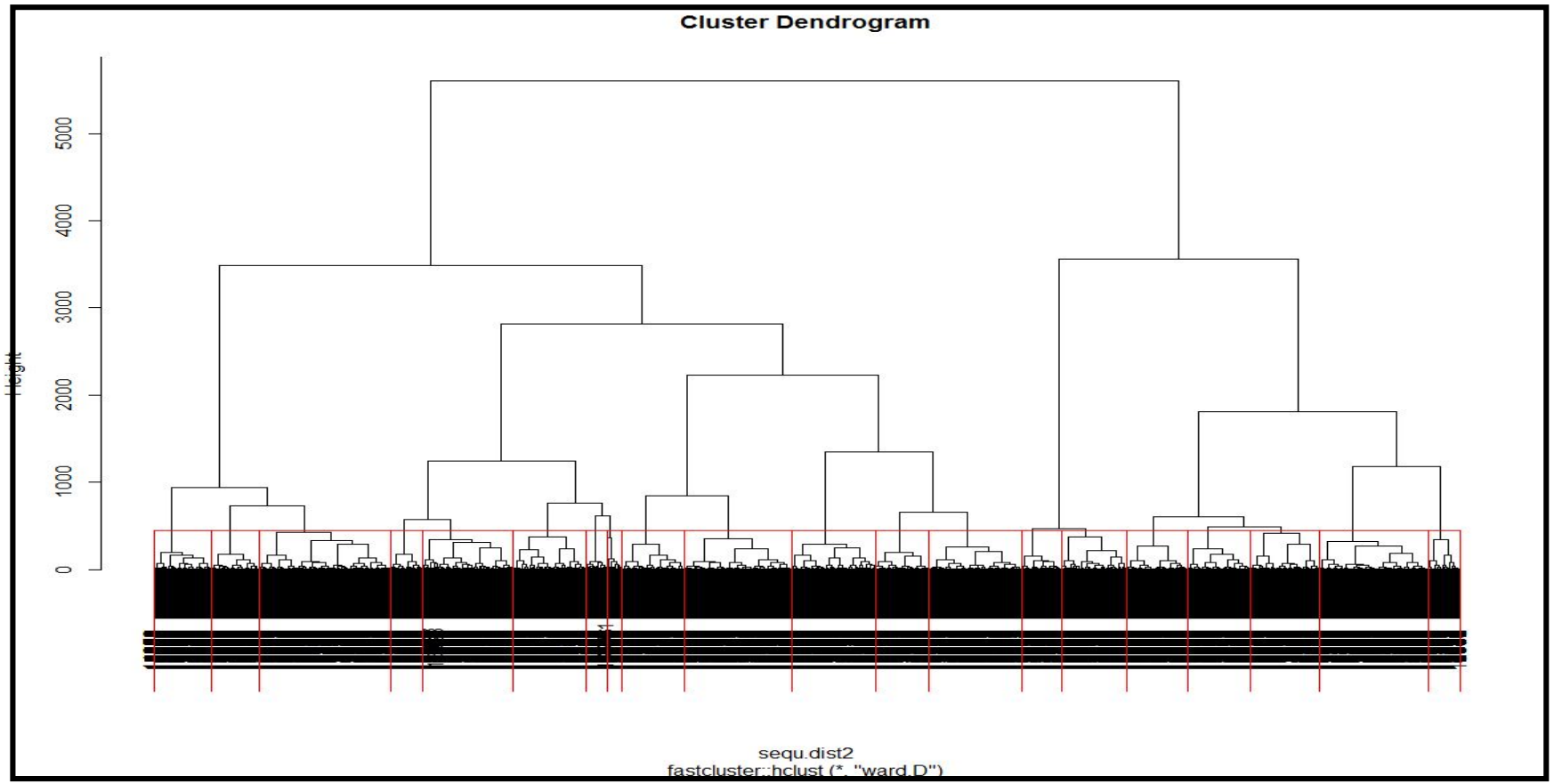
※범주형 변수들이 있지만 나이대 같은 경우에 결국 나이가 많을 수록높은 값을 갖도록 설정하였고, 경쟁사 이용 , 멤버쉽 사용의 합등은 사용할수록 좋다라는 가중치의 개념으로 판단하여 scale을 통해 연속형 변수들로 군집분석을 하였다.

LOTTE

L.POINT

3번째줄 (Continuous Cluster + item based filtering)

2)유클리디안 거리를 사용하였고 앞선 자카드 계수를 통한 군집 분석과 같이 군집의 개수를 20개로 형성하여 아이템을 추천한다.



3번째줄 (Continuous Cluster + item based filtering)

3) 단, 1번 줄과 2번줄의 상품이(*이 2개의 상품은 겹치는 것이 없도록 조절하였다.) 3번줄의 상품과 겹치는 경우 아이템 기반 협업 필터링을 사용하였다.

1. 각 고객 번호당 상품 분류 코드에 대한 테이블을 실시하였다.
2. 행렬곱을 통해 아이템의 연관성을 계산하였다(4386*4386)
3. 상대성 유사 척도로 만들기 위해 거리값을 (아이템거리값/1+아이템거리값) 산정하였다.

* 상대성 유사척도 값이 클수록 연관성이 큰것이다.

	A010101	A010102	A010103	A010104	A010105	A010106	A010201	A0102
A010101	0.9998788	0.9974160	0.9998351	0.9997881	0.9960784	0.9997474	0.9998449	0.999
A010102	0.9974160	0.9975062	0.9968454	0.9967532	0.9473684	0.9960159	0.9969512	0.996
A010103	0.9998351	0.9968454	0.9998663	0.9997473	0.9955556	0.9997172	0.9998162	0.999
A010104	0.9997881	0.9967532	0.9997473	0.9998003	0.9943820	0.9996300	0.9997655	0.999
A010105	0.9960784	0.9473684	0.9955556	0.9943820	0.9963370	0.9936306	0.9954955	0.995
A010106	0.9997474	0.9960159	0.9997172	0.9996300	0.9936306	0.9997851	0.9997269	0.999
A010201	0.9998449	0.9969512	0.9998162	0.9997655	0.9954955	0.9997269	0.9998668	0.999
A010202	0.9998233	0.9966443	0.9997906	0.9997421	0.9950980	0.9996925	0.9998237	0.999
A010203	0.9996962	0.9935065	0.9996394	0.9995925	0.9923077	0.9994689	0.9996947	0.999
A010204	0.9917355	0.7500000	0.9901961	0.9868421	0.9166667	0.9824561	0.9919355	0.990
A010205	0.9982175	0.9583333	0.9979167	0.9973545	0.9756098	0.9967320	0.9982906	0.998
A010206	0.9995493	0.9838710	0.9994840	0.9993610	0.9892473	0.9991857	0.9995495	0.999
A010207	0.9998233	0.9968153	0.9997964	0.9997350	0.9952153	0.9996083	0.9998168	0.999

	V1	V2	V3	V4	V5
A010101	A010101	A011004	A010401	A010302	A010608
A010102	A010102	A010401	A010608	A010302	A010101
A010103	A010103	A011004	A010401	A020302	A010302
A010104	A010104	A010401	A010302	A010608	A011004
A010105	A010105	A011004	A010302	A010401	A010608
A010106	A010106	A010401	A011004	A010302	A020302
A010201	A010201	A010401	A010402	A011004	A010302
A010202	A010202	A010401	A010402	A011004	A010302
A010203	A010203	A010401	A010402	A010404	A010302
A010204	A010204	A010401	A010402	A010404	A020302
A010205	A010205	A010401	A020302	A011004	A010302
A010206	A010206	A010401	A011004	A010402	A010608
A010207	A010207	A010401	A011004	A010302	A010402
A010208	A010208	A010401	A010302	A010402	A011004

CONTENTS

1.팀원 소개 및 프로젝트 개요
2.데이터 탐색
3.데이터 정제
4.데이터 모델링
5.개선점 및 제언

참고 사이트 및 문헌

-RFM 기반 점진적 마이닝을 이용한 개인화 추천기법 – 조영성
-Apriori 알고리즘 기반의 개인화 정보 추천시스템 설계 및 구현에 관한 연구 – 김용
-www.google.co.kr
-www.naver.com

분석 한계점

1. 구매데이터이기때문에 샀던 물건을 그대로 추천해줄 수 밖에 없었다.
2. 구매내역이 적은 고객에 대해서 Rule이 표현이 되지 않아 비슷한 부분으로 묶을 수 밖에 없었다.
3. 컴퓨터의 성능부족으로 인해 연산량이 많은 작업을 수행할 수 없었다.
4. 물건자체에 대한 정보가 없어 콘텐츠 기반 필터링을 하지 못했다.

감사합니다.