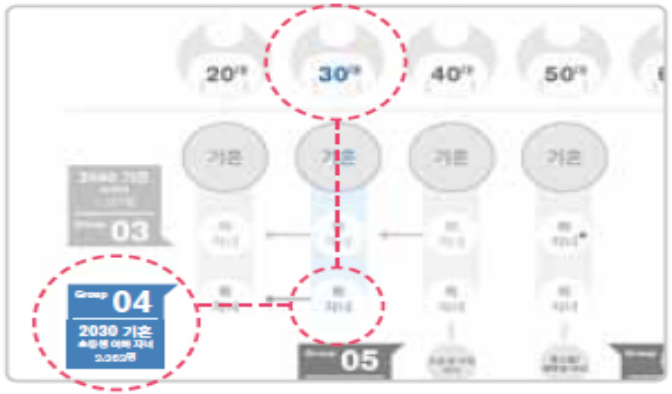
금융데이터를 활용한
“나의 금융생활 정보지수” 개발

INDEX

프로젝트 주제 및 개요

프로젝트 주제

고객금융데이터를 활용한 “나의 금융생활 정보지수” 개발

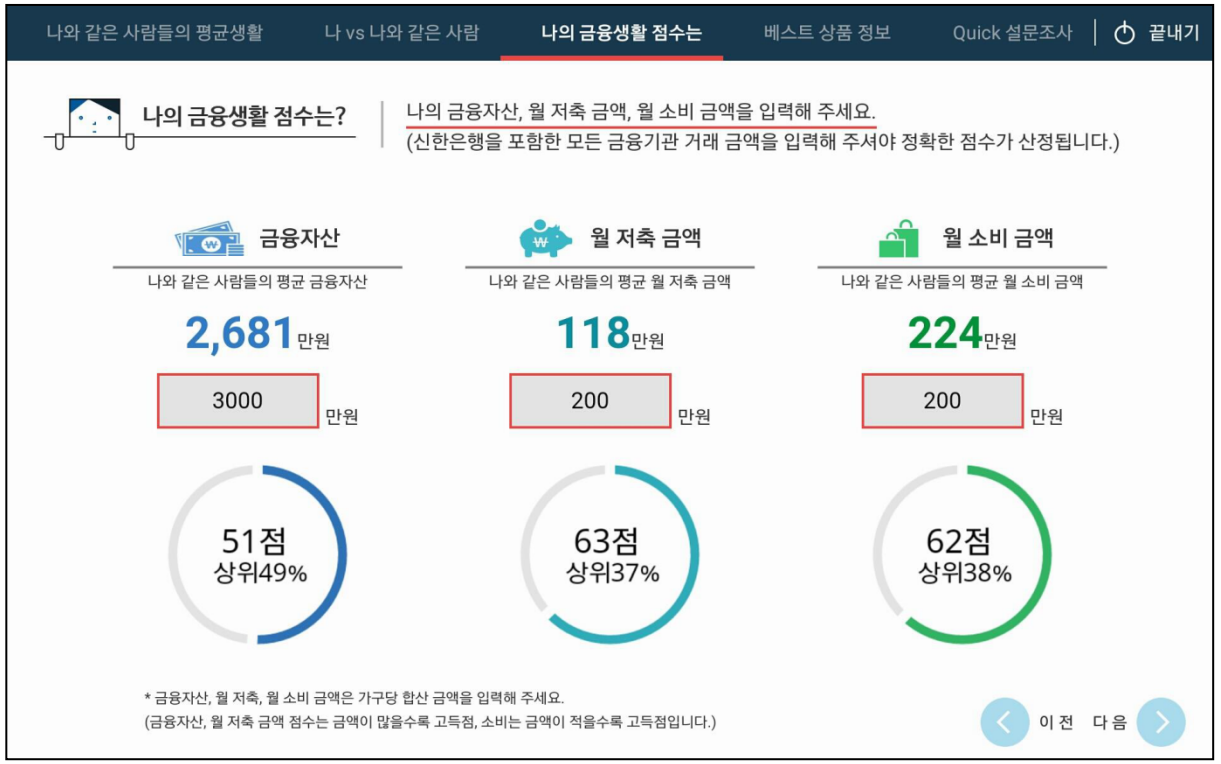


(만 원)

총자산	금융자산	부동산자산	기타자산	월총저축액	월저축액 적금
2,300	300	0	2000	40	0
7,500	7000	500	0	120	35
41,900	5900	35000	1000	120	80
123,200	16200	105000	2000	30	0
33,500	1500	30000	2000	100	0
55,800	40800	9000	6000	200	0

기존방식

제안방식



프로젝트 주제 및 개요

AS-IS

기존의 서로 가장 비슷한 사람들의 기준
☞**고객기본정보**

고객기본정보 기준으로 금융자산, 월저축금액, 월소비금액 평균수준 제시.

한계점 : 전통적인 고객정보에 따른 세분화는 고객금융정보를 제대로 반영하지 못함

예) 동일한 성별, 연령대, 거주지역이라도 금융활동(보유자산, 대출 등)은 상이한 경우가 多

TO-BE

기존의 서로 가장비슷한 사람들의 기준
☞ **고객금융정보**

1. 고객금융정보를 활용하여 **Peer group Clustering** 및 고객집단을 새롭게 정의

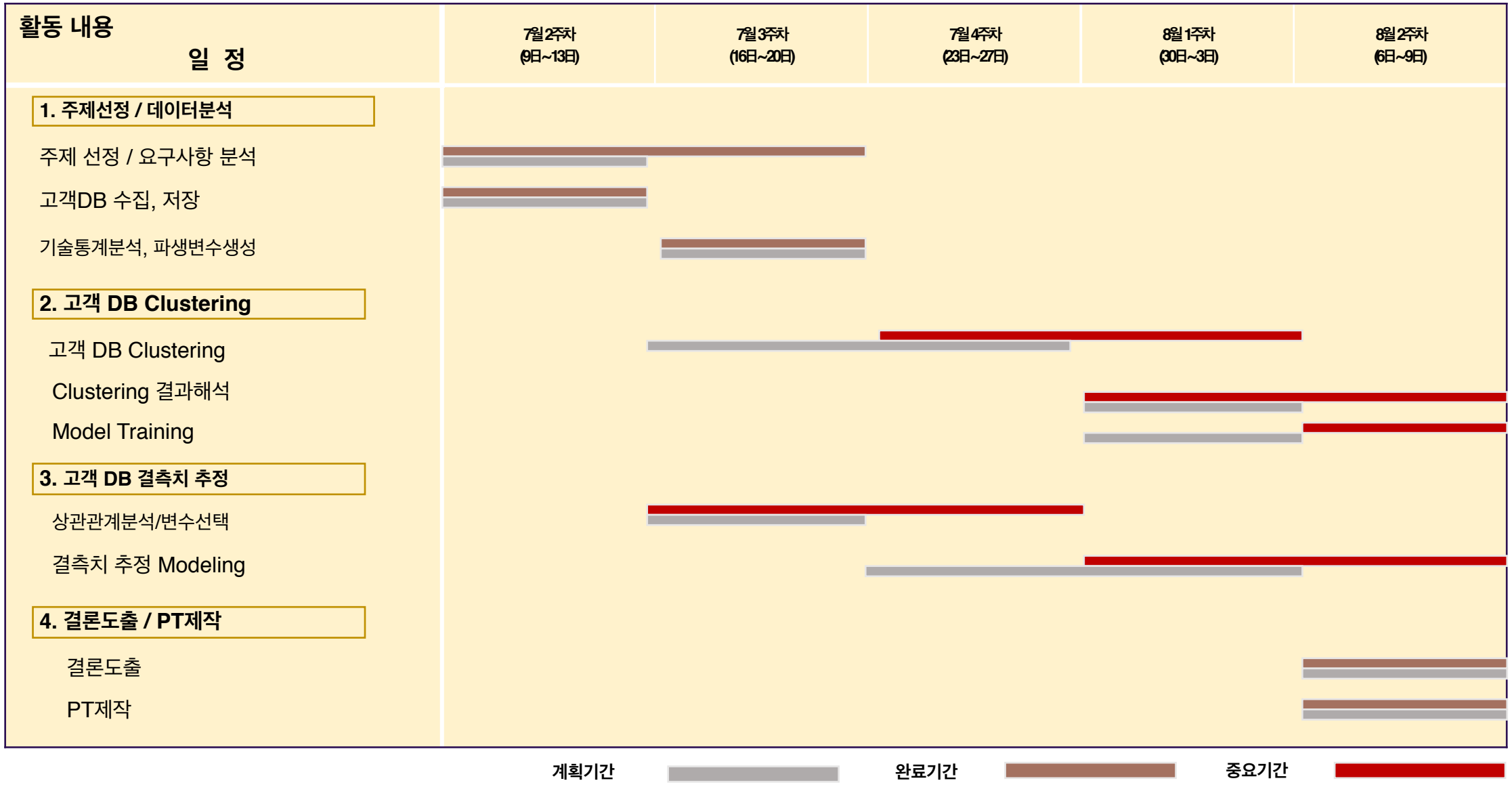
2. 고객금융정보의 **결측치(NA)에 대한 합리적인 추정**을 통한 데이터확보

3. 기본적인 고객금융정보 입력 시 해당 **Peer Group 분류 및 금융점수 제시**

기대 효과

금융기관 : Peer Group 별 더욱 효과적인 금융상품추천, 고객마케팅 전략수립
고객 : 자신의 금융생활지수를 참고하여 금융활동수준 진단

프로젝트 진행 일정


활동 내용
일 정
7월 2주차 (9日~13日)
7월 3주차 (16日~20日)
7월 4주차 (23日~27日)
8월 1주차 (30日~3日)
8월 2주차 (6日~9日)
1. 주제선정 / 데이터분석
주제 선정 / 요구사항 분석
고객DB 수집, 저장
기술통계분석, 파생변수생성
2. 고객 DB Clustering
고객 DB Clustering
Clustering 결과해석
Model Training
3. 고객 DB 결측치 추정
상관관계분석/변수선택
결측치 추정 Modeling
4. 결론도출 / PT제작
결론도출
PT제작
계획기간
완료기간
중요기간

데이터 준비

신한은행 AWS(Amazon Web Service)의 글로벌 제공데이터 활용

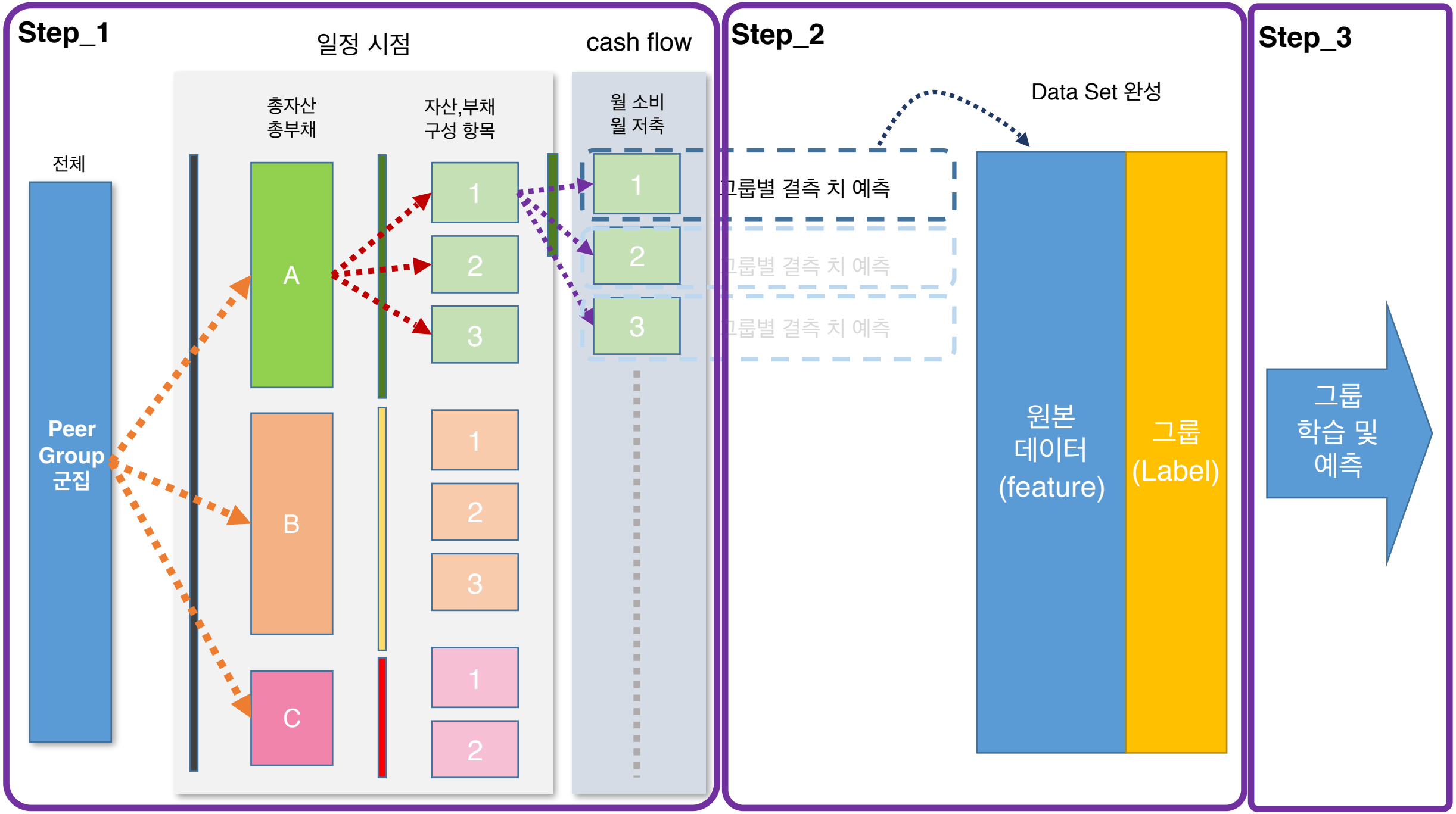
- **고객기본정보 (8개)**

  성별, 연령구분, 직업구분, 지역구분, 가구소득구간, 결혼여부, 맞벌이, 자녀 수

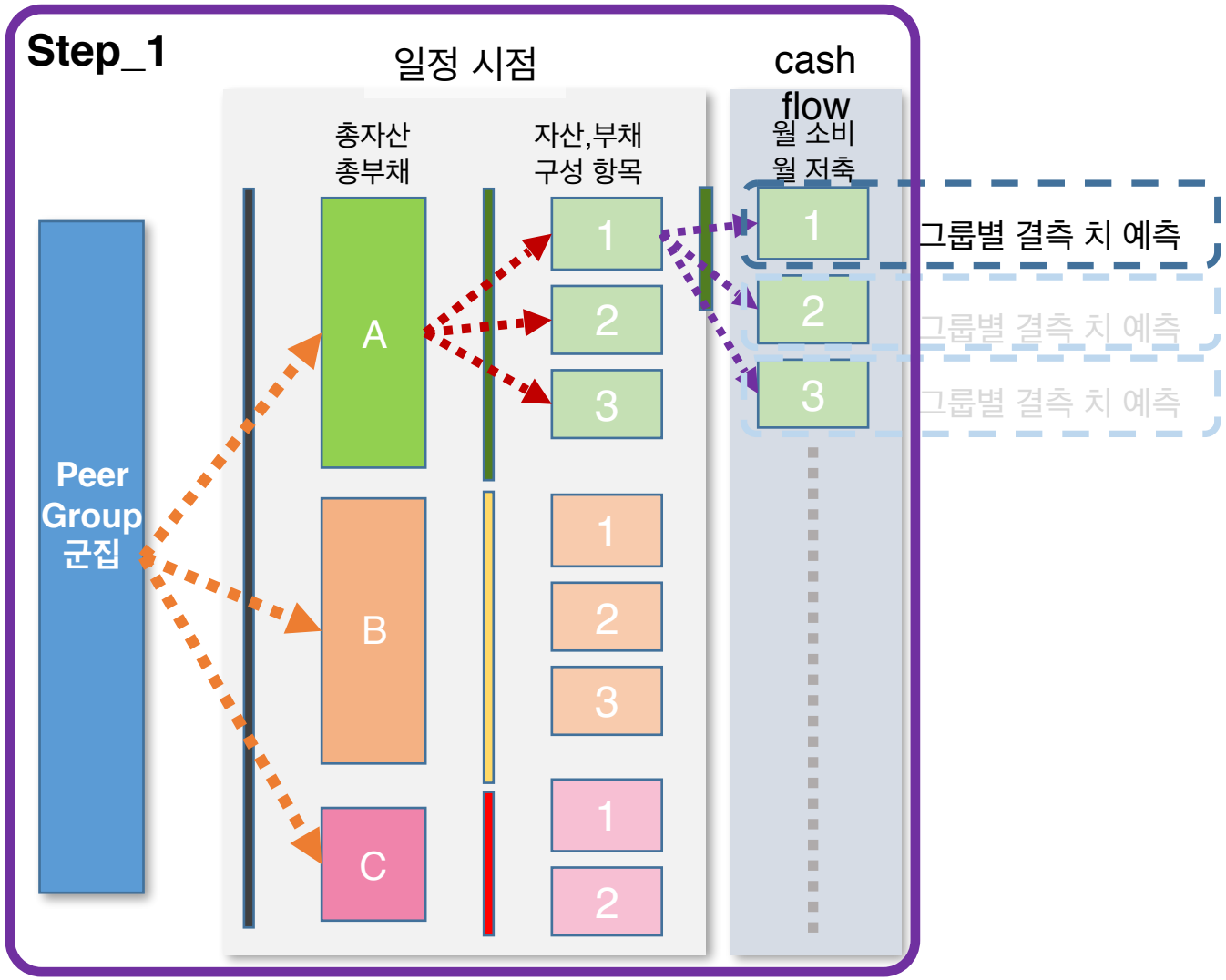
- **금융거래정보 (26개)**

  금융거래정보(26개) : 금융자산, 부동산자산, 월저축액_적금, 월저축액_펀드,
  월저축액_주식, 월저축액_저축성보험, 월저축액_청약, 부채잔액_신용대출,
  부채잔액_담보대출, 은퇴후필요자금, 노후자금융월저축액, 금융상품잔액_정기예금 등

전체 수행 과정


Step_1
일정 시점
cash flow
총자산
총부채
자산,부채
구성 항목
월 소비
월 저축
전체
Peer
Group
군집
A
B
C
1
2
3
1
2
3
1
2
1
2
3
Step_2
Data Set 완성
그룹별 결측 치 예측
그룹별 결측 치 예측
그룹별 결측 치 예측
원본
데이터
(feature)
그룹
(Label)
Step_3
그룹
학습 및
예측

단계별 수행과제 (STEP_1 Peer Group clustering)



Step_1 군집 분석

1. 단계 및 그룹별 변수 선택
2. 데이터 표준화(Scale the data)
3. 군집의 수 결정
4. 군집 알고리즘 선택
5. 그룹별 변수 프로파일링

단계별 수행과제 (STEP_2 Estimate NA)

STEP_2 결측치 추정

141,750개 고객유형의 24개 금융거래정보항목의 결측치를 추정하여 Data Set 완성

idx2	가구소득 (구간)	연령 구분	직업 구분	성별	지역 구분	맞벌이 여부	자녀수	결혼 여부	총자산	금융자산	부동산자산
1	7	2	6	1	5	2	1	2	63,241,656	32,263,835	269,168,094
2	6	4	3	2	1	-	0	-	-	-	-
3	3	2	11	2	3	-	1	-	6,781,207	2,165,455	27,172,813
4	7	4	3	2	4	-	1	1	-	-	-
:	:	:	:	:	:	:	:	:	:	:	:
141,746	6	5	9	1	4	1	-	-	-	-	-
141,747	7	2	7	2	5	1	1	-	-	-	-
141,748	2	6	3	1	1	1	3	2	115,595,657	12,477,908	90,817,644
141,749	1	3	2	2	5	1	1	1	9,120,825	3,113,843	22,678,927
141,750	6	6	6	2	2	-	1	-	-	-	-

Estimated Data

단계별 수행과제 (STEP_3 Peer Group Prediction)

STEP_3 Peer Group 예측

금융정보를 입력한 고객의 Peer Group을 찾고, Peer Group의 정보를 토대로 고객의 '금융생활지수'를 제공한다

idx2	Peer Group No.
1	328
2	328
3	542
4	12
5	84
:	:
141,749	12
141,750	156

STEP_01　Peer Group Clustering

STEP_01 Peer Group Clustering

1. **단계 및 그룹별 변수 선택**
    1) 상관계수 확인 – cor()
    2) 분산이 0에 가까운지 확인 – caret::nearZeroVar()사용

2. **데이타 표준화(Scale the data)** : min-max 사용

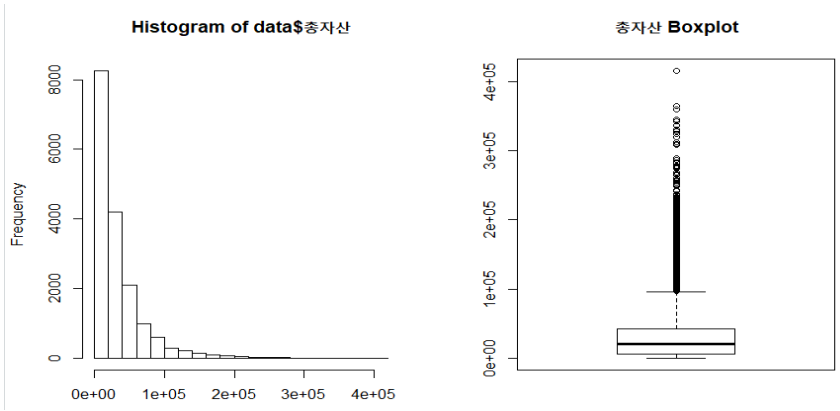
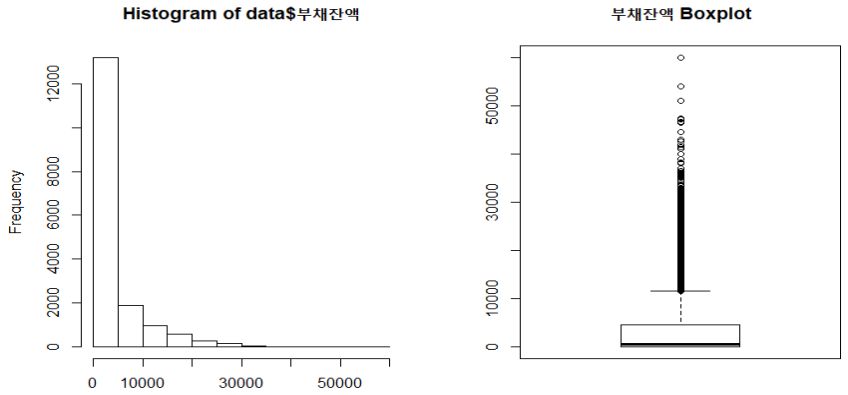
3. **군집의 갯수 결정**: NbClust () 사용

4. **군집 알고리즘 선택**
    1) K-means
    2) PAM
    3) 3그룹 이상 차이 검정 : aov(), 사후검정- TukeyHSD()
    4) silhouette-score로 알고리즘 선택

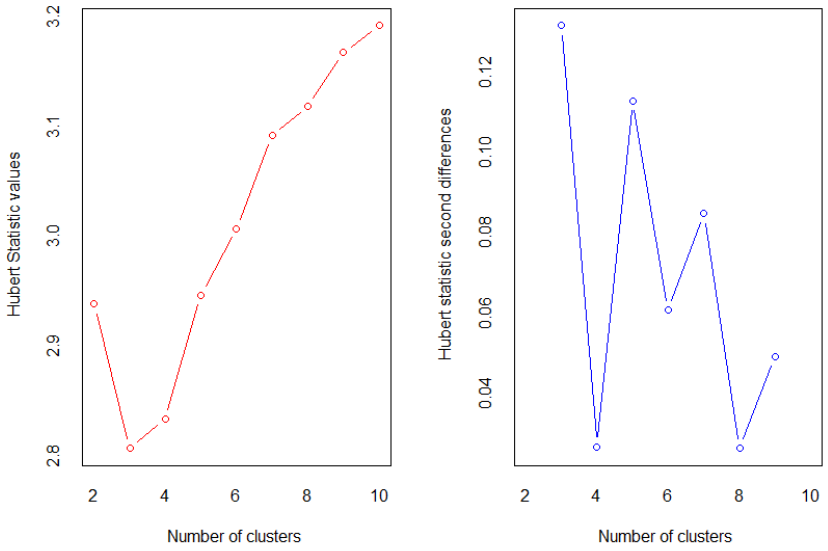
5. **그룹별 변수 프로파일링**

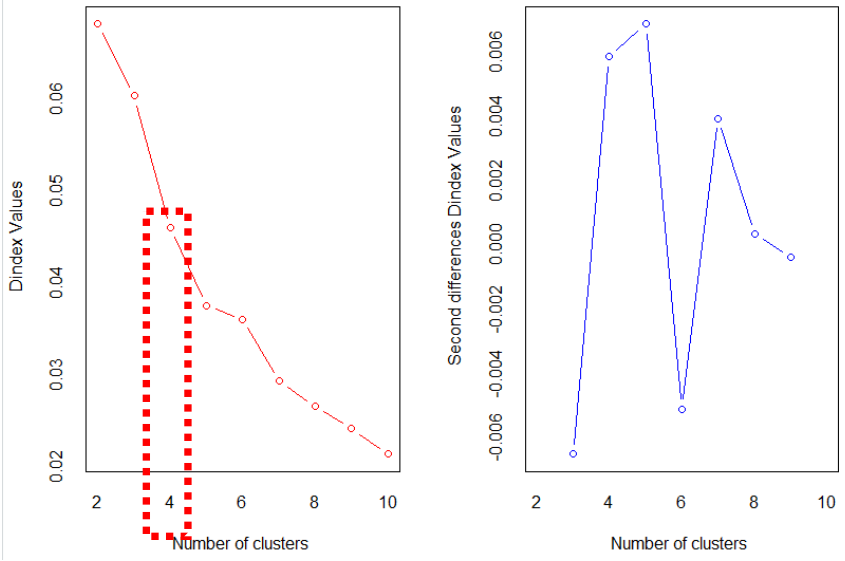
STEP_01 Peer Group Clustering

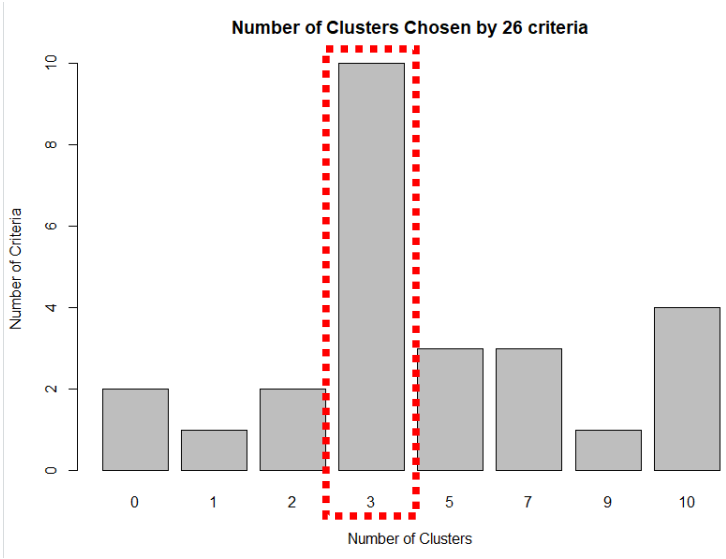
1-1 총자산, 부채 잔액 기준 군집

총자산	척도 : 연속형 단위 : 만원 NA : 0		총자산 Min. : 30 1st Qu.: 6000 Median : 21000 Mean : 31944 3rd Qu.: 42193 Max. :415500
부채 잔액	척도 : 연속형 단위 : 만원 NA : 0		부채잔액 Min. : 0 1st Qu.: 0 Median : 500 Mean : 3519 3rd Qu.: 4600 Max. :60000

1. 군집 수 K 선택







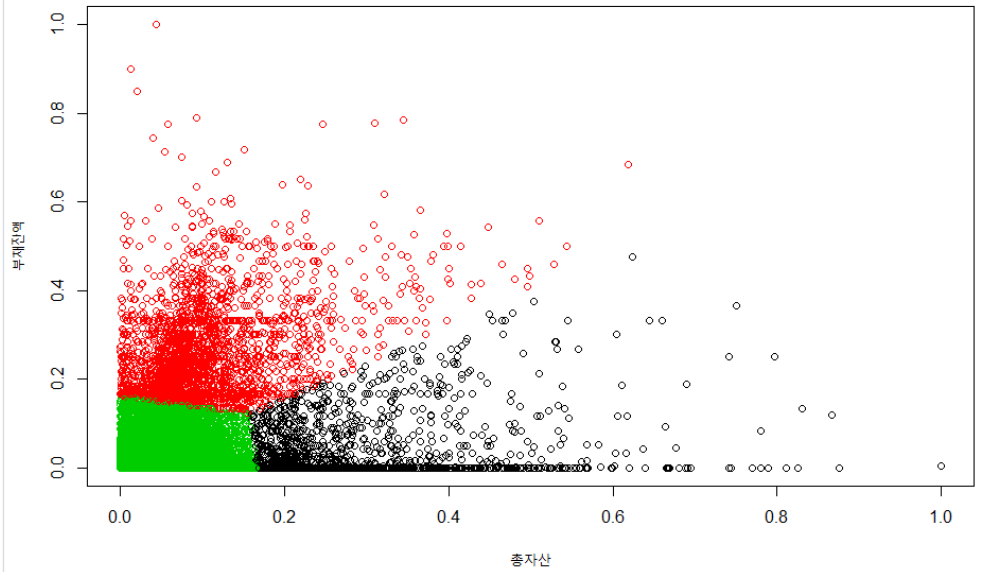

```
*****************************************************************************
* Among all indices:
* 2 proposed 2 as the best number of clusters
* 10 proposed 3 as the best number of clusters
* 3 proposed 5 as the best number of clusters
* 3 proposed 7 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 4 proposed 10 as the best number of clusters

                 ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```

Nbcluster 결과 가장 좋은 군집수 : 3개

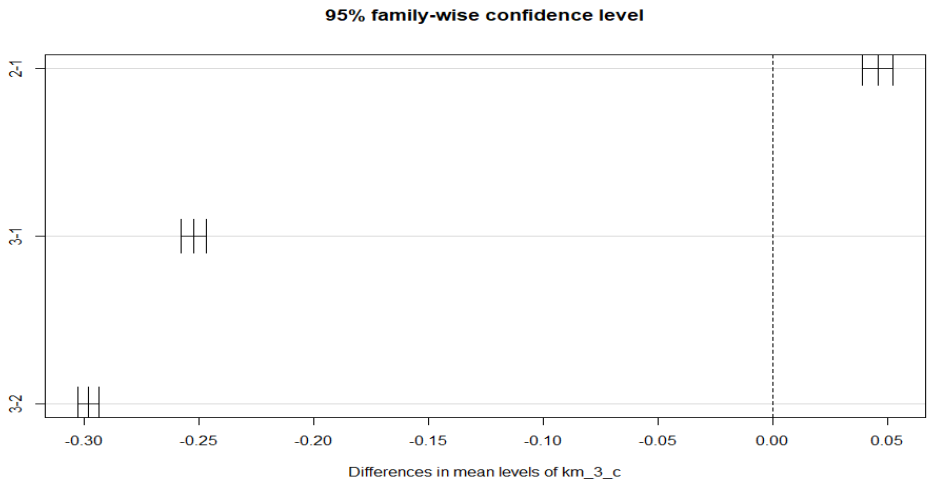
그래프 판단 : 4개

2. 군집분석_방법 1: K-means, K=3

- **K-means 결과 산점도 그래프**



- **ANOVA Tukey 검정 그래프**



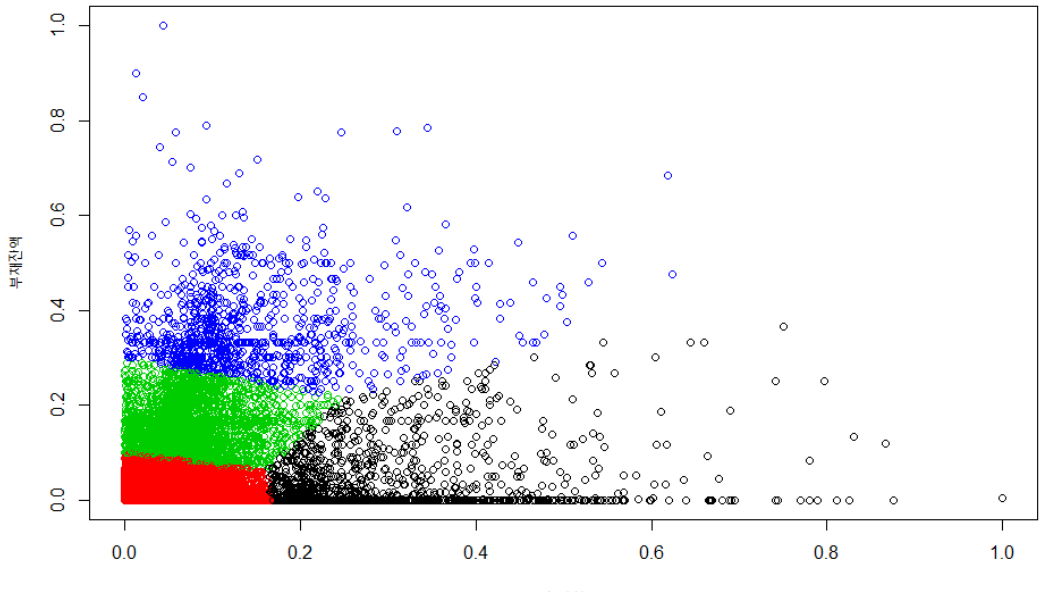

```
> summary(result)
              Df Sum Sq Mean Sq F value Pr(>F)
km_3_c         2  244.2  122.11   15682 <2e-16 ***
Residuals  17073  132.9    0.01
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #사후검정
> TukeyHSD(result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 총자산 + 부채잔액 ~ km_3_c, data = data.scaled)

$km_3_c
           diff         lwr         upr p adj
2-1  0.04583232  0.03920101  0.05246363     0
3-1 -0.25237777 -0.25782536 -0.24693017     0
3-2 -0.29821009 -0.30277828 -0.29364189     0
```

- **ANOVA 분석 결과**

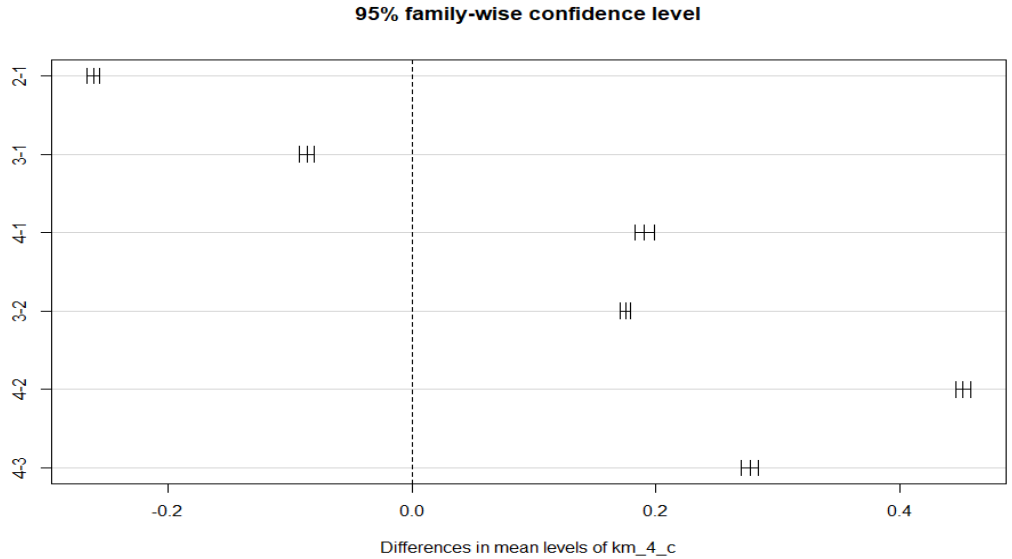
연구가설(H_0)	K-means 방법으로 나눈 3그룹은 차이가 있다.
귀무가설(H_1)	K-means 방법으로 나눈 3그룹은 차이가 없다.
유의 수준	A = 0.05
검정통계량	F = 15682 , Df=2, Sum Sq=244.2, Mean Sq= 122.11
유의 확률	P <2e-16
결과해석	유의수준 0.05 에서 귀무가설이 기각되었다. 따라서, K-means 방법으로 나눈 3그룹은 차이가 있다.

2. 군집분석_방법 1: K-means, K=4

- **K-means 결과 산점도 그래프**



- **ANOVA Tukey 검정 그래프**



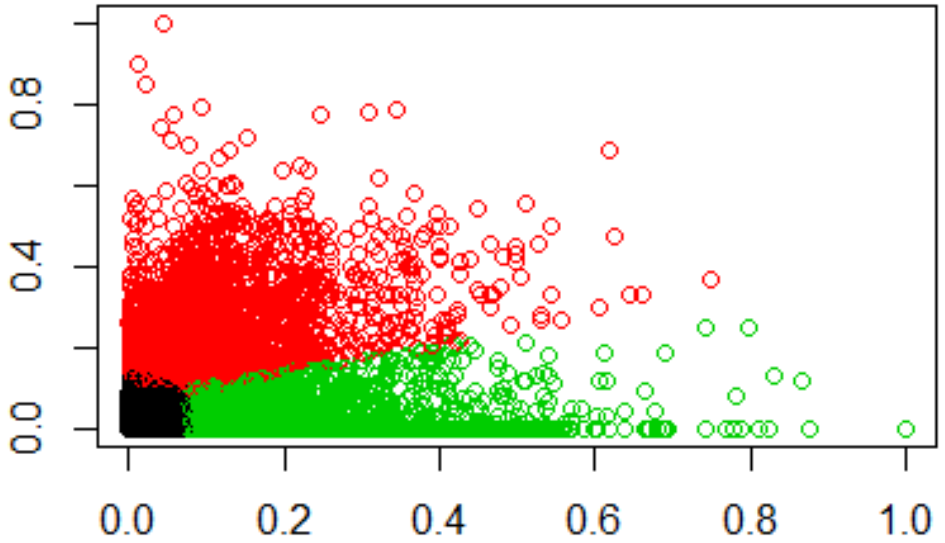

```
> summary(result)
               Df Sum Sq Mean Sq F value Pr(>F)
km_4_c          3 285.03   95.01   17604 <2e-16 ***
Residuals   17072  92.14    0.01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #사후검정
> TukeyHSD(result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 총자산 + 부채잔액 ~ km_4_c, data = data.scaled)

$km_4_c
            diff         lwr         upr p adj
2-1 -0.26121532 -0.26629266 -0.25613799     0
3-1 -0.08629654 -0.09228142 -0.08031165     0
4-1  0.19058624  0.18286327  0.19830921     0
3-2  0.17491879  0.17091009  0.17892749     0
4-2  0.45180156  0.44548531  0.45811781     0
4-3  0.27688277  0.26981633  0.28394922     0
```

- **ANOVA 분석 결과**

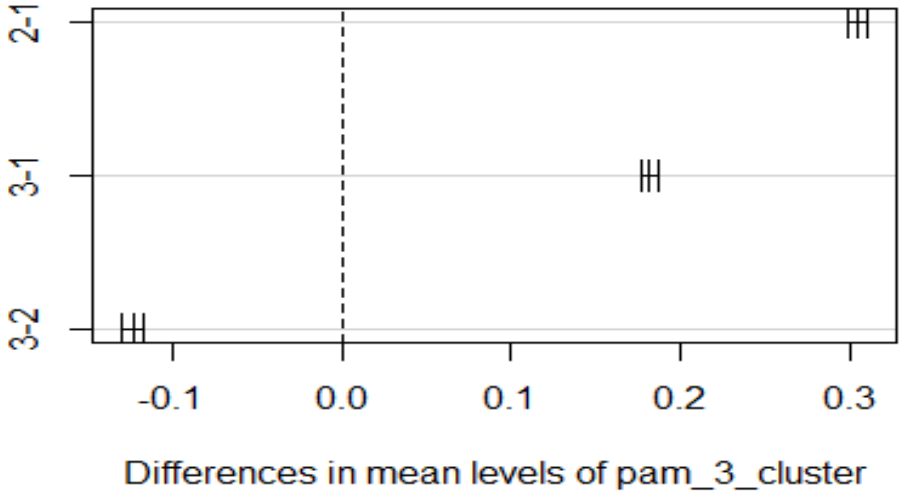
연구가설(H_0)	K-means 방법으로 나눈 4그룹은 차이가 있다.
귀무가설(H_1)	K-means 방법으로 나눈 4그룹은 차이가 없다.
유의 수준	A = 0.05
검정통계량	F = 17604 , Df=3, Sum Sq=285.03, Mean Sq= 95.01
유의 확률	P <2e-16
결과해석	유의수준 0.05 에서 귀무가설이 기각되었다. 따라서, K-means 방법으로 나눈 4그룹은 차이가 있다.

2. 군집분석_방법 3: PAM, K=3

- PAM(K=3) 결과 산점도 그래프



- ANOVA Tukey 검정 그래프



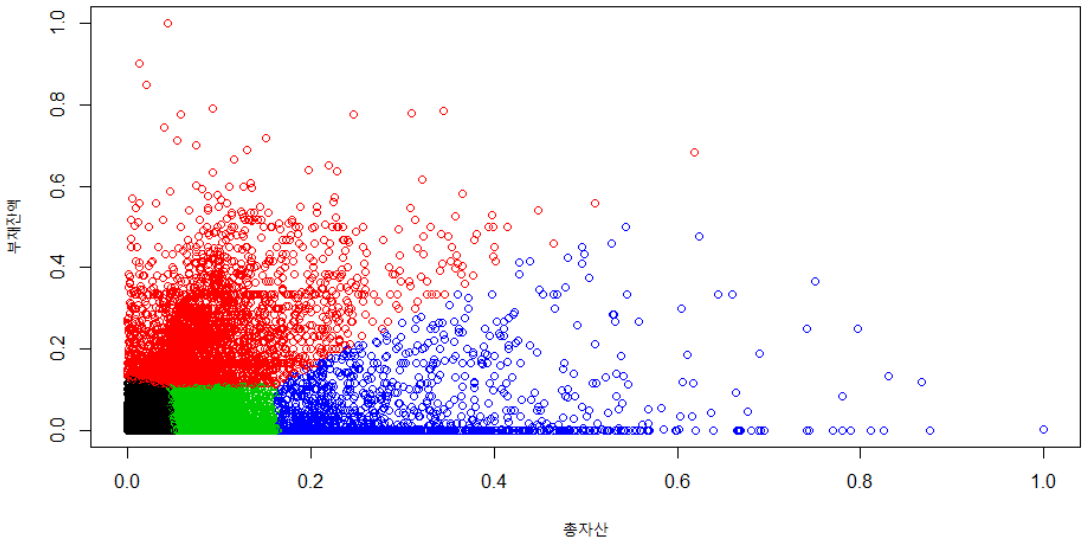

```
> summary(shinhan.total.pam3.aov)
                Df Sum Sq Mean Sq F value Pr(>F)
pam_3_cluster    2  255.5  127.76    9733 <2e-16 ***
Residuals    17073  224.1    0.01
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> tot_pam3_tukey <- TukeyHSD(shinhan.total.pam3.aov)
> print(tot_pam3_tukey)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = shinhan.total$총자산mm + shinhan.total$월총저축액mm + s

$pam_3_cluster
          diff        lwr        upr p adj
2-1  0.3043431  0.2988214  0.3098648     0
3-1  0.1813701  0.1765169  0.1862232     0
3-2 -0.1229730 -0.1291887 -0.1167574     0
```

- ANOVA 분석 결과

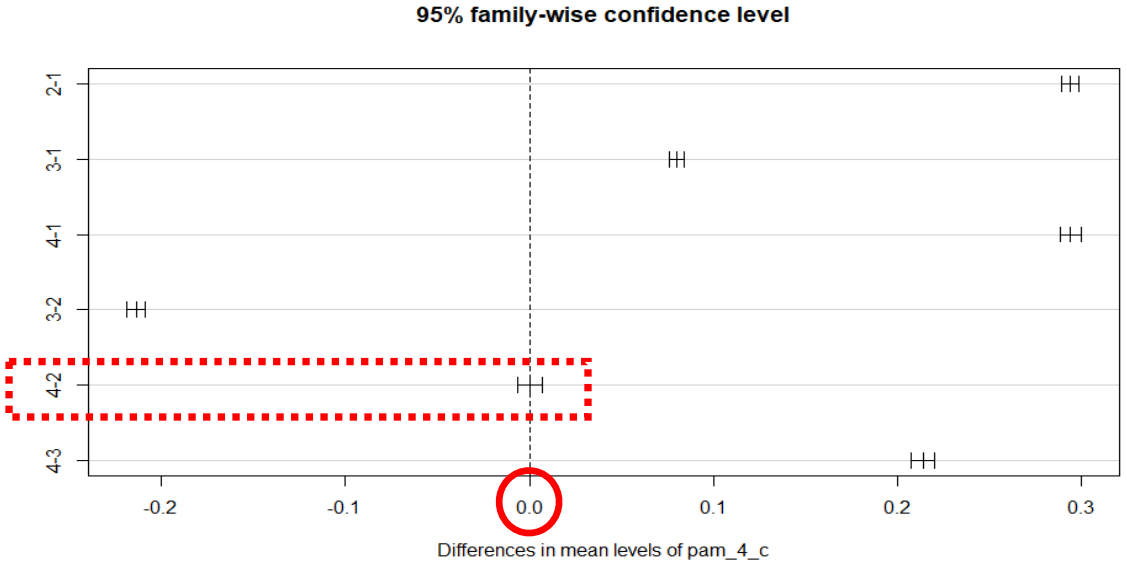
연구가설(H_0)	PAM 방법으로 나눈 3그룹은 차이가 있다.
귀무가설(H_1)	PAM 방법으로 나눈 3그룹은 차이가 없다.
유의 수준	A = 0.05
검정통계량	F = 9733 , Df=2, Sum Sq=255.5 , Mean Sq= 127.76
유의 확률	P <2e-16
결과해석	유의수준 0.05 에서 귀무가설이 기각되었다. 따라서, PAM 방법으로 나눈 3그룹은 차이가 있다.

2. 군집분석_방법 3: PAM, K=4

- **PAM(K=4) 결과 산점도 그래프**



- **ANOVA Tukey 검정 그래프**




```
> result<-aov(총자산+부채잔액 ~ pam_4_c , data=data.scaled)
> summary(result)
              Df Sum Sq Mean Sq F value Pr(>F)
pam_4_c        3  255.5   85.16   11948 <2e-16 ***
Residuals  17072  121.7    0.01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 총자산 + 부채잔액 ~ pam_4_c, data = data.scaled)

$pam_4_c
             diff          lwr          upr     p adj
2-1  0.2940298506  0.289382015  0.298677686 0.0000000
3-1  0.0802822376  0.076264495  0.084299981 0.0000000
4-1  0.2942707802  0.288395365  0.300146196 0.0000000
3-2 -0.2137476130 -0.218796853 -0.208698373 0.0000000
4-2  0.0002409297 -0.006382746  0.006864605 0.9997078
4-3  0.2139885426  0.207790724  0.220186361 0.0000000
```

- **ANOVA 분석 결과**

연구가설(H_0)	PAM 방법으로 나눈 4그룹은 차이가 있다.
귀무가설(H_1)	PAM 방법으로 나눈 4그룹은 차이가 없다.
유의 수준	A = 0.05
검정통계량	F = 9733 , Df=2, Sum Sq=255.5 , Mean Sq= 127.76
유의 확률	P <2e-16
결과해석	유의수준 0.05 에서 귀무가설이 기각되었다. ANOVA 분석 결과 PAM 방법으로 나눈 4그룹은 차이가 있다. 하지만, Tukey 사후 검정결과 2번 그룹과 4번 그룹의 차이에 대한 유의확률은 0.99로 두 그룹은 차이가 없는 것으로 판단되므로 PAM, K=4 로 나누는 방법은 사용하지 않기로 한다.

3. 모형 선택 결과

- **선택 기준 : Silhouette-score**

$$s(i) = \frac{d_{rest}(i) - d_s(i)}{\max(d_s(i), d_{rest}(i))}$$

- 포인트 군집 내 포인트 간의 평균 거리
- 포인트 타 군집의 포인트 간 최소 거리

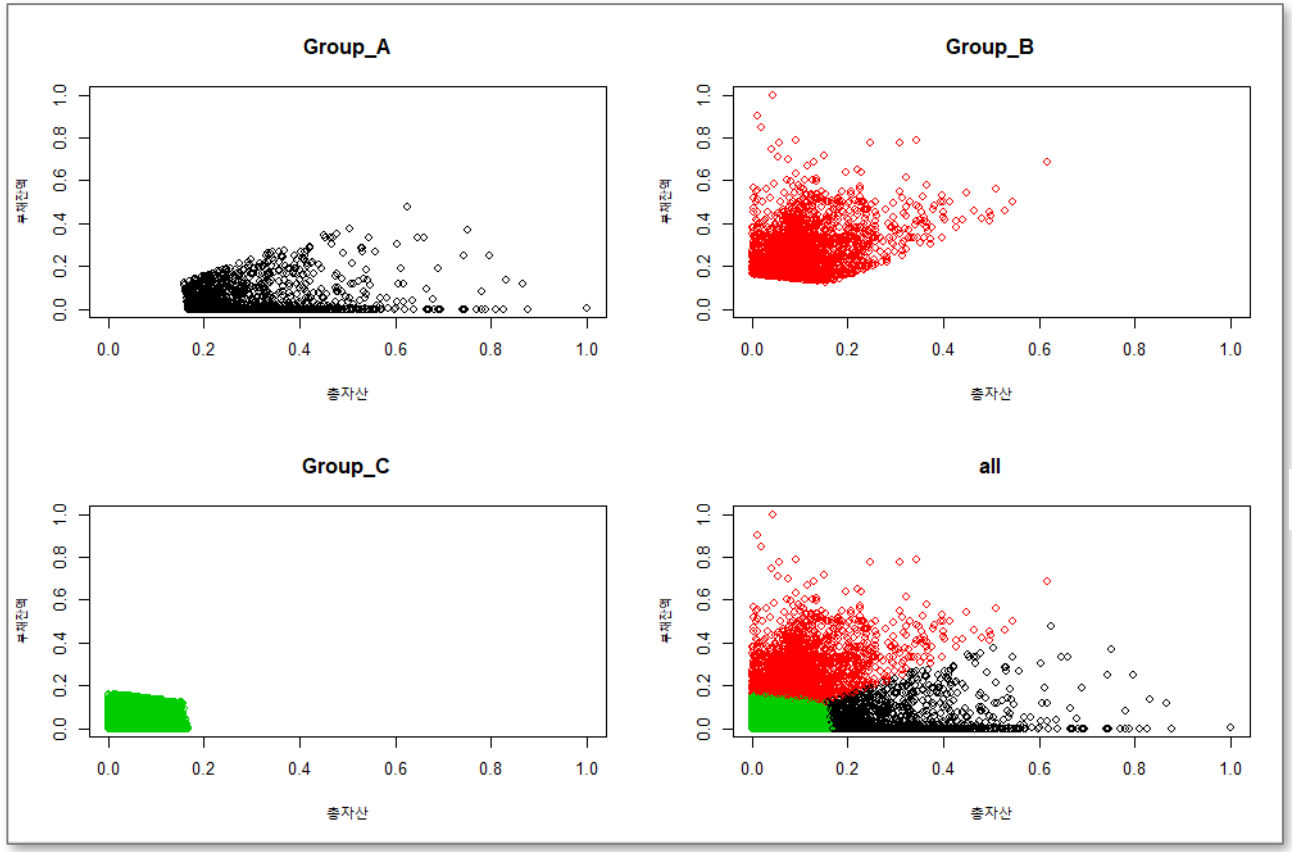
- s(i) < 0 포인트 : i 가 잘못 클러스터링 되었다.
- s(i) = 0 포인트 : i 가 두개의 클러스터링 사이에 있다.
- s(i)가 1에 가까울수록 포인트 i 가 잘 분류되었다.

- **결과**

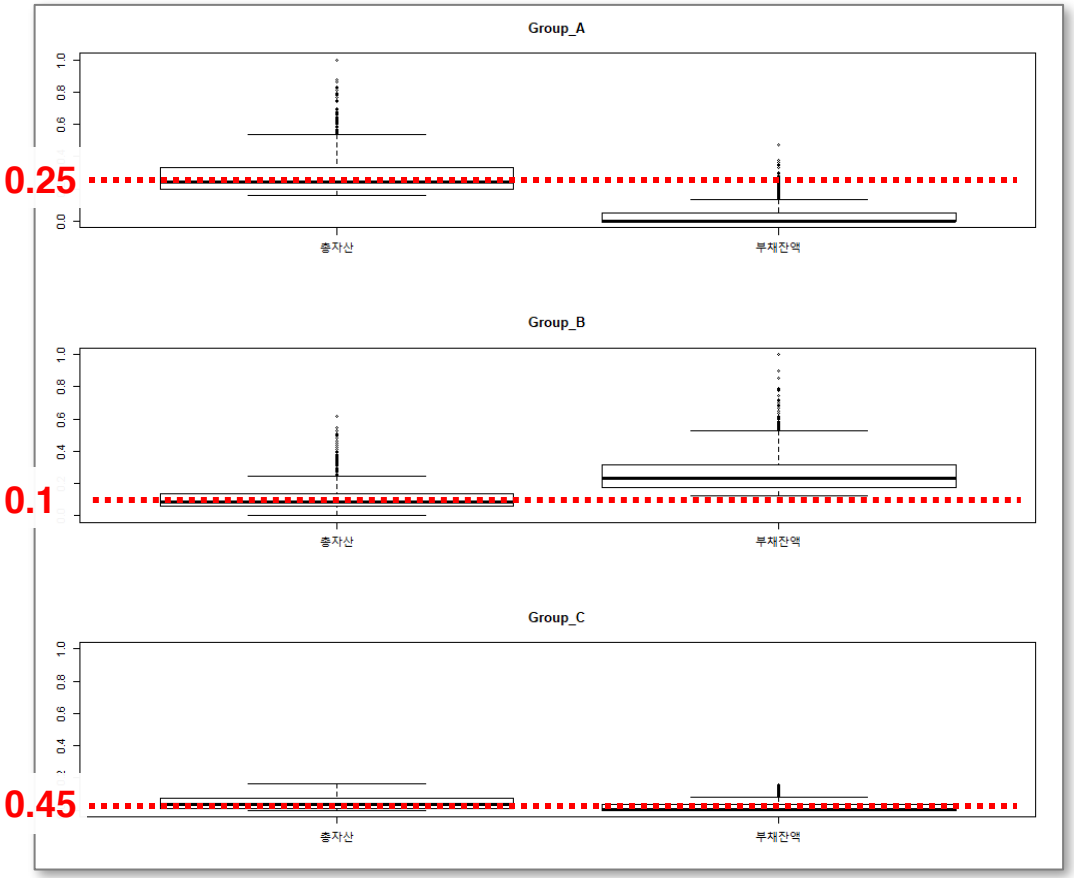
구분	선택모형	군집수	silhouette-score
방법1	**K-means**	**3**	**0.6106908**
방법2	K-means	4	0.5450471
방법3	PAM	3	0.4742634
결과	Silhouette score가 가장 큰,방법 1의 군집 결과를 채택한다.		

4. 군집 확인

- 군집 별 산점도 확인

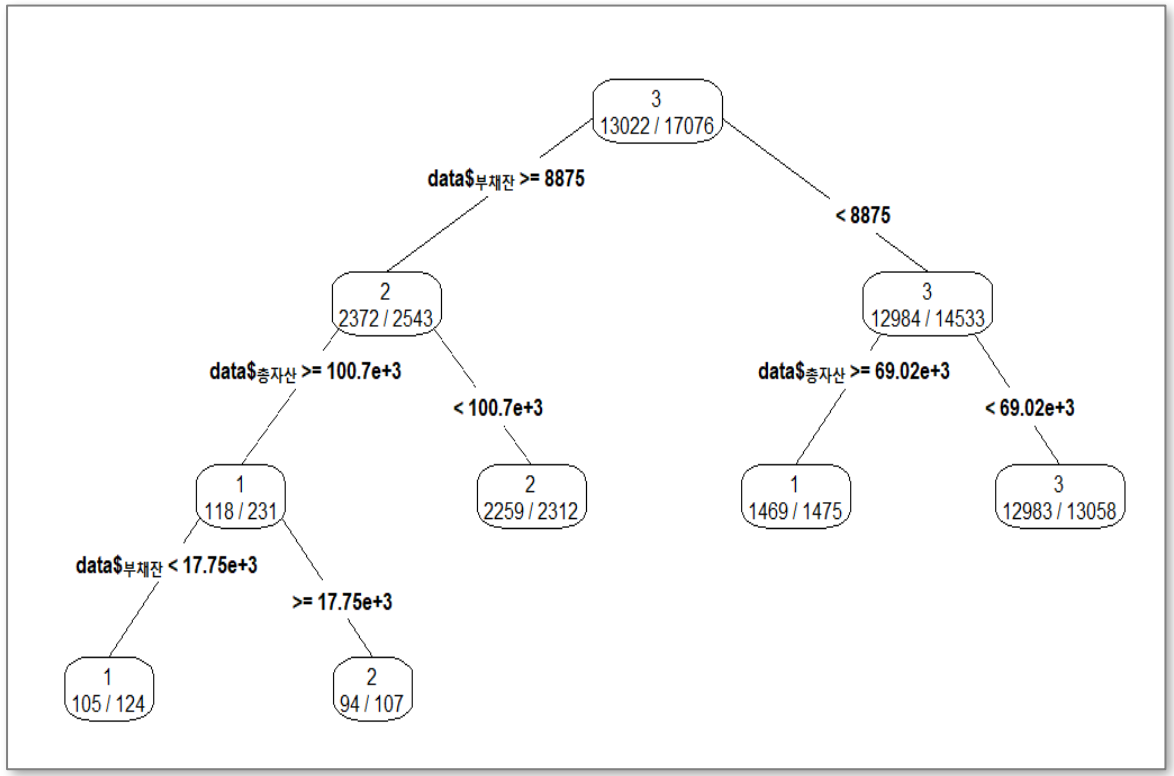


- Boxplot 확인



4. 군집 확인

- **의사결정 트리 확인**



	금액 구분		인원수	Min-Max정규화 값으로 구분
A	자산	6억9천만원 이하	16,21명	자산이 정상범위 이상이며 상대적으로 부채가 낮은 그룹
	부채	8천9백만원 이하	(9.5%)	
B	자산	10억 이하	2,433명	부채가 정상범위 이상이며 상대적으로 부채가 낮은 그룹
	부채	8천9백만원 이상	(14.25%)	
C	자산	6억9천만원 이상	13,022명	자산과 부채가 정상범위 이내에 있고 값이 비슷한 그룹
	부채	1억8천만원 이하	(76.25%)	

```
> summary(group_A)
     총자산            부채잔액
 Min.   : 66500   Min.   :    0
 1st Qu.: 81800   1st Qu.:    0
 Median :100760   Median :    0
 Mean   :117414   Mean   : 2332
 3rd Qu.:138500   3rd Qu.: 3200
 Max.   :415500   Max.   :28500
```

```
> summary(group_B)
     총자산            부채잔액
 Min.   :    55   Min.   : 7500
 1st Qu.: 24200   1st Qu.:10500
 Median : 36630   Median :14000
 Mean   : 43635   Mean   :15736
 3rd Qu.: 55610   3rd Qu.:19000
 Max.   :256800   Max.   :60000
```

```
> summary(group_C)
     총자산            부채잔액
 Min.   :   30   Min.   :    0
 1st Qu.: 4050   1st Qu.:    0
 Median :13900   Median :   50
 Mean   :19120   Mean   : 1384
 3rd Qu.:30693   3rd Qu.: 2000
 Max.   :69100   Max.   : 9600
```

```
> summary(All)
     총자산            부채잔액
 Min.   :    30   Min.   :    0
 1st Qu.:  6000   1st Qu.:    0
 Median : 21000   Median :  500
 Mean   : 31944   Mean   : 3519
 3rd Qu.: 42193   3rd Qu.: 4600
 Max.   :415500   Max.   :60000
```

5. 전체 군집 구분 결과

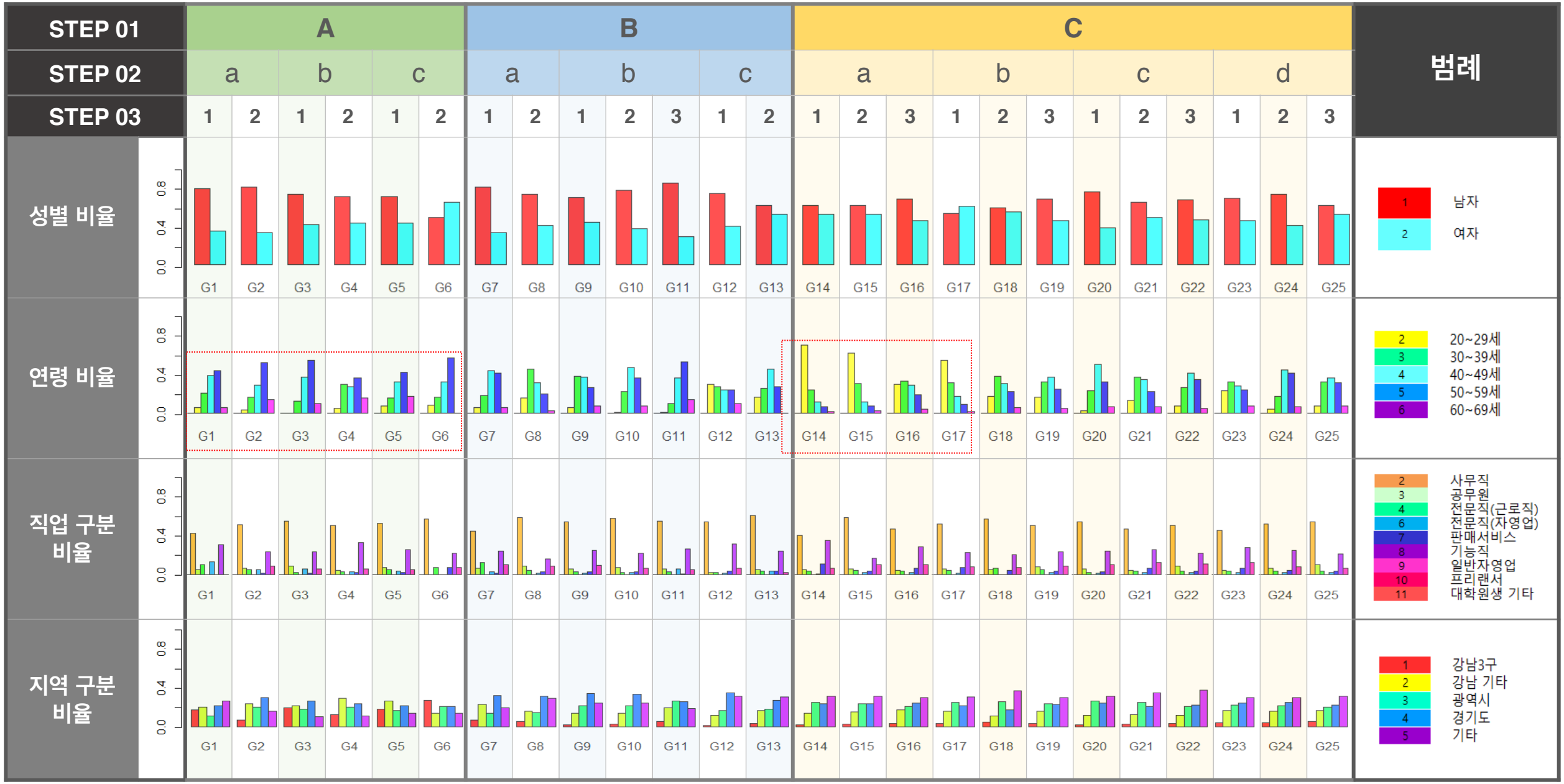
- **전체 군집 결과표**

step1	step2	step3	n
A	a	1	62
		2	432
	b	1	122
		2	280
	c	1	711
		2	14
B	a	1	83
		2	244
	b	1	928
		2	818
		3	129
	c	1	168
		2	63
C	a	1	1023
		2	699
		3	721
	b	1	2045
		2	98
		3	634
	c	1	718
		2	983
		3	401
	d	1	2780
		2	1694
		3	1226

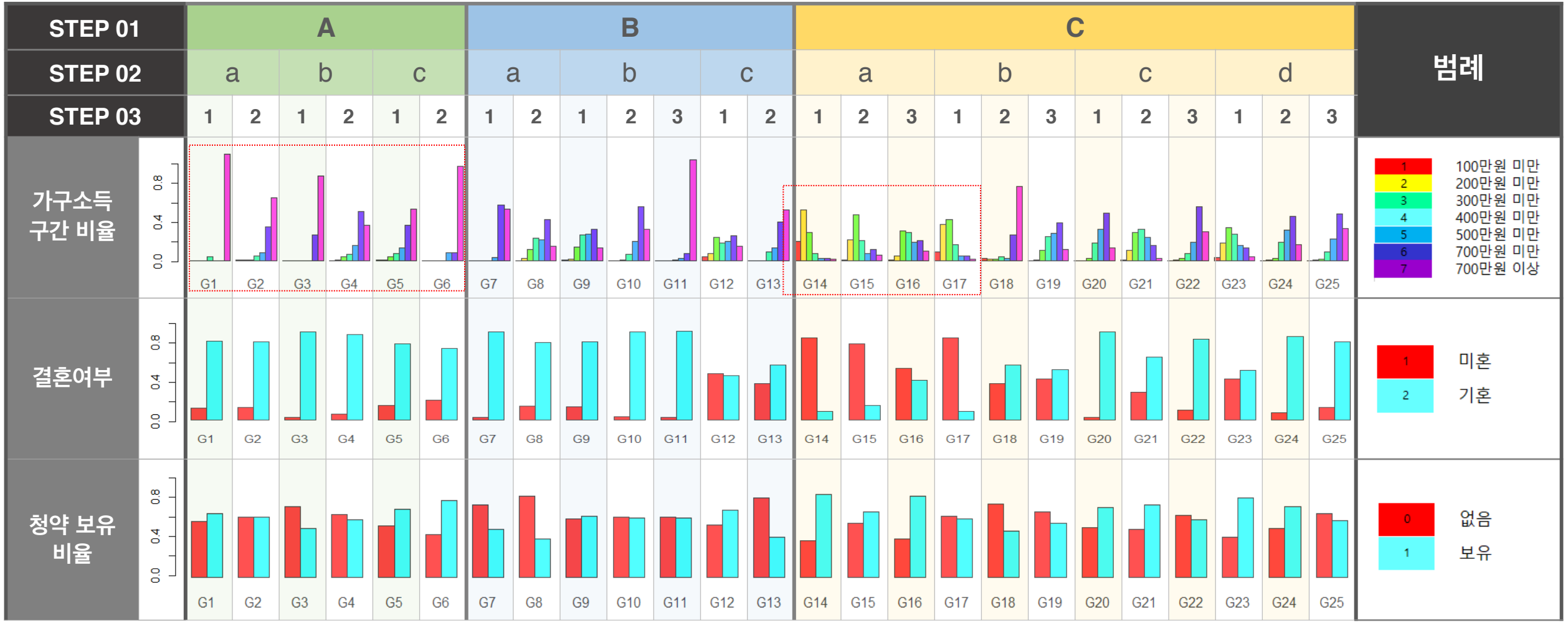
- **전체 군집 단계별 구분 방법**

단계	STEP 01			STEP 02			STEP 03		
구분방법	DATA SET	알고리즘	Sihouetee	DATA SET	알고리즘	Sihouetee	DATA SET	알고리즘	Sihouetee
	전체	K-means (k=3)	0.61	A	PAM (K=3)	0.569	a	K-means(k=2)	0.588
							b	K-means(k=2)	0.38
							c	K-means(k=2)	0.83
				B	K-means (k=3)	0.629	a	K-means(k=2)	0.46
							b	K-means(k=3)	0.475
							c	PAM(K=2)	0.361
				C	PAM (K=4)	0.583	a	PAM(K=3)	0.334
							b	K-means(k=3)	0.56
							c	PAM(K=3)	0.409
							d	PAM(K=3)	0.422
feature	총자산, 부채잔액			금융자산 비율, 부동산 자산 비율, 기타자산 비율, 부동산 담보대출 비율			월 총저축액, 월소비금액		

6-1. 고객 정보 확인

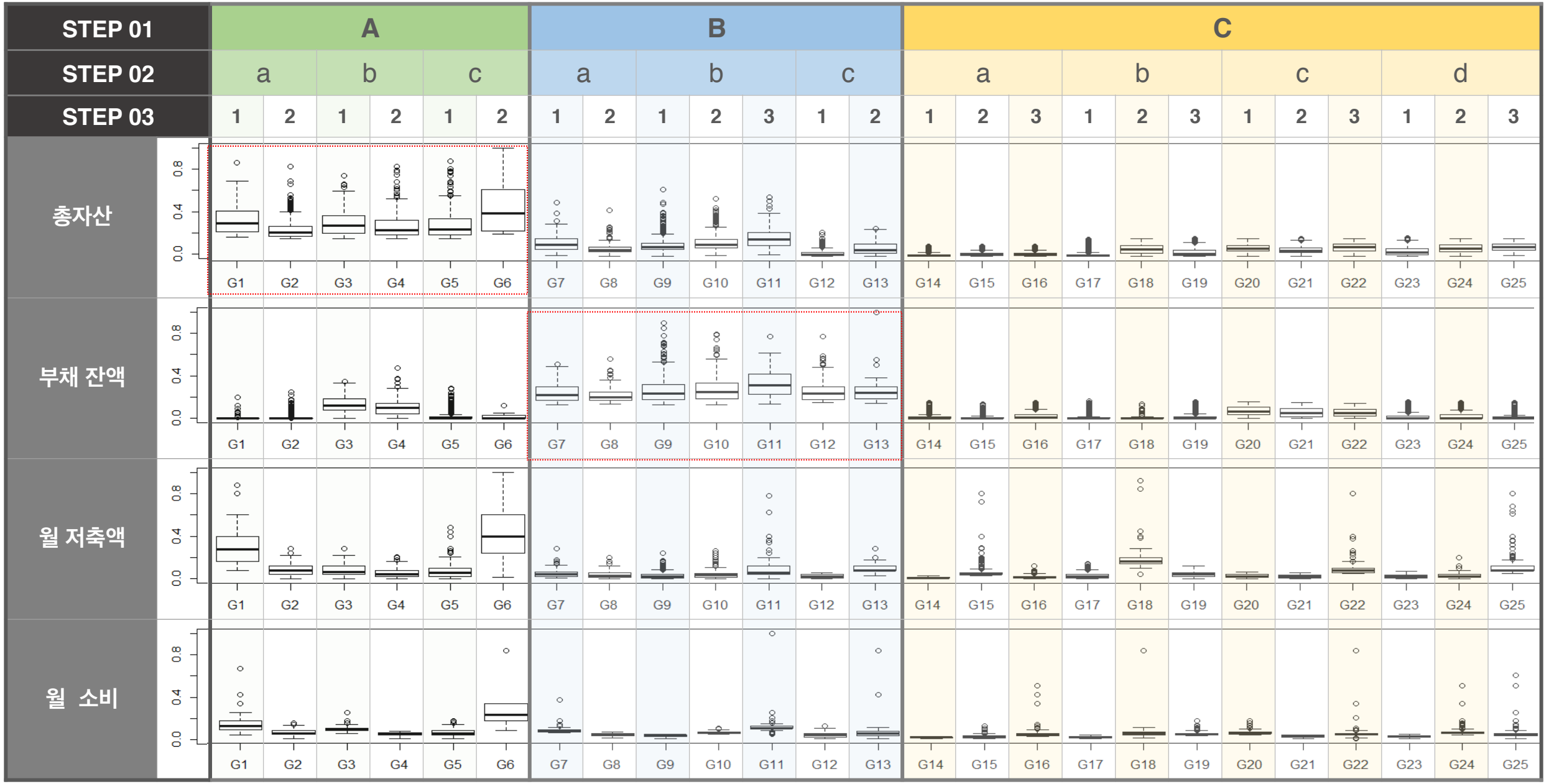

STEP 01
A
B
C
STEP 02
a
b
c
a
b
c
a
b
c
d
STEP 03
1 2 1 2 1 2 1 2 1 2 3 1 2 1 2 3 1 2 3 1 2 3 1 2 3
범례
성별 비율
0.0 0.4 0.8
G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15 G16 G17 G18 G19 G20 G21 G22 G23 G24 G25
1 남자
2 여자
연령 비율
0.0 0.4 0.8
G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15 G16 G17 G18 G19 G20 G21 G22 G23 G24 G25
2 20~29세
3 30~39세
4 40~49세
5 50~59세
6 60~69세
직업 구분 비율
0.0 0.4 0.8
G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15 G16 G17 G18 G19 G20 G21 G22 G23 G24 G25
2 사무직
3 공무원
4 전문직(근로직)
6 전문직(자영업)
7 판매서비스
8 기능직
9 일반자영업
10 프리랜서
11 대학원생 기타
지역 구분 비율
0.0 0.4 0.8
G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15 G16 G17 G18 G19 G20 G21 G22 G23 G24 G25
1 강남3구
2 강남 기타
3 광역시
4 경기도
5 기타

6-1. 고객 정보 확인


STEP 01
A
B
C
STEP 02
a
b
c
a
b
c
a
b
c
d
STEP 03
1 2 1 2 1 2 1 2 1 2 3 1 2 1 2 3 1 2 3 1 2 3 1 2 3
범례
가구소득 구간 비율
0.0
0.4
0.8
G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15 G16 G17 G18 G19 G20 G21 G22 G23 G24 G25
1 100만원 미만
2 200만원 미만
3 300만원 미만
4 400만원 미만
5 500만원 미만
6 700만원 미만
7 700만원 이상
결혼여부
1 미혼
2 기혼
청약 보유 비율
0 없음
1 보유

6_2. 금융정보 총액 비교

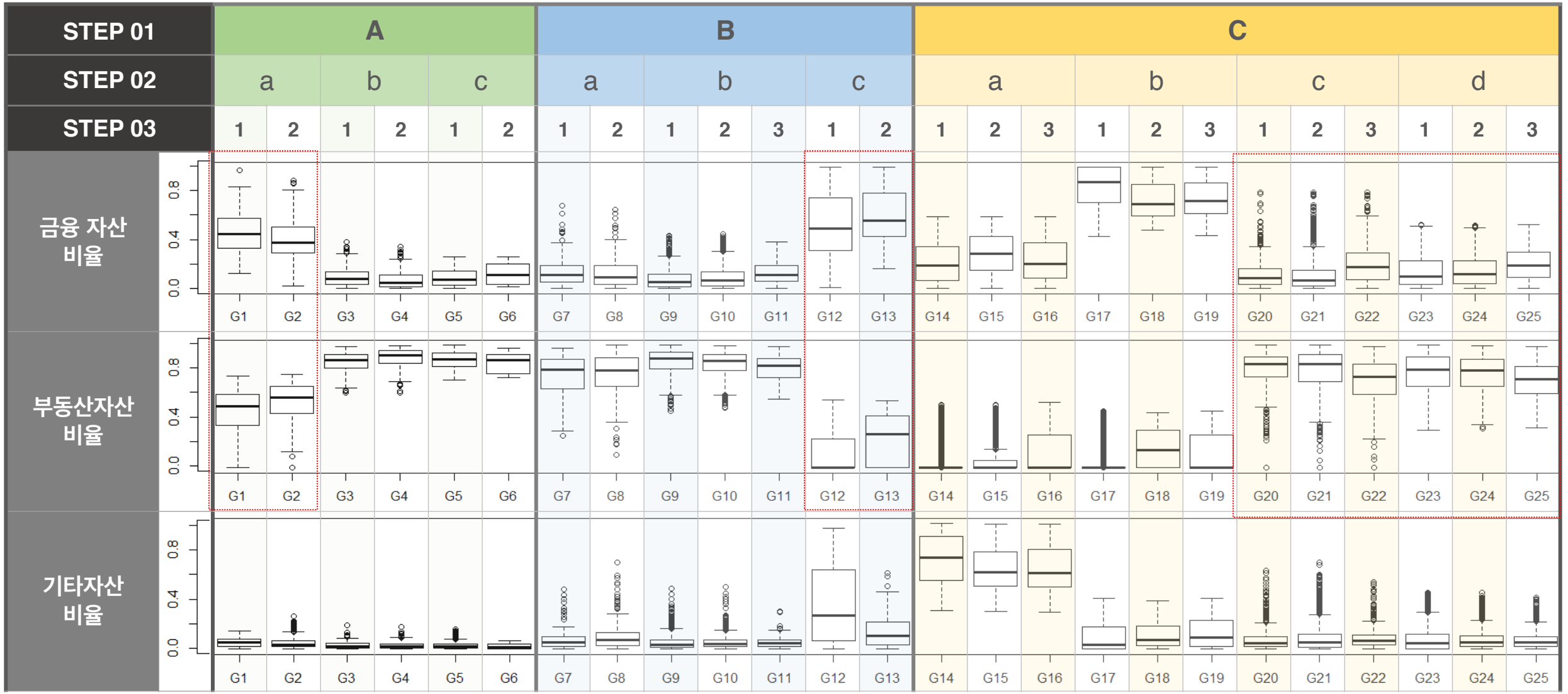
✓ STEP 01 그룹 비교



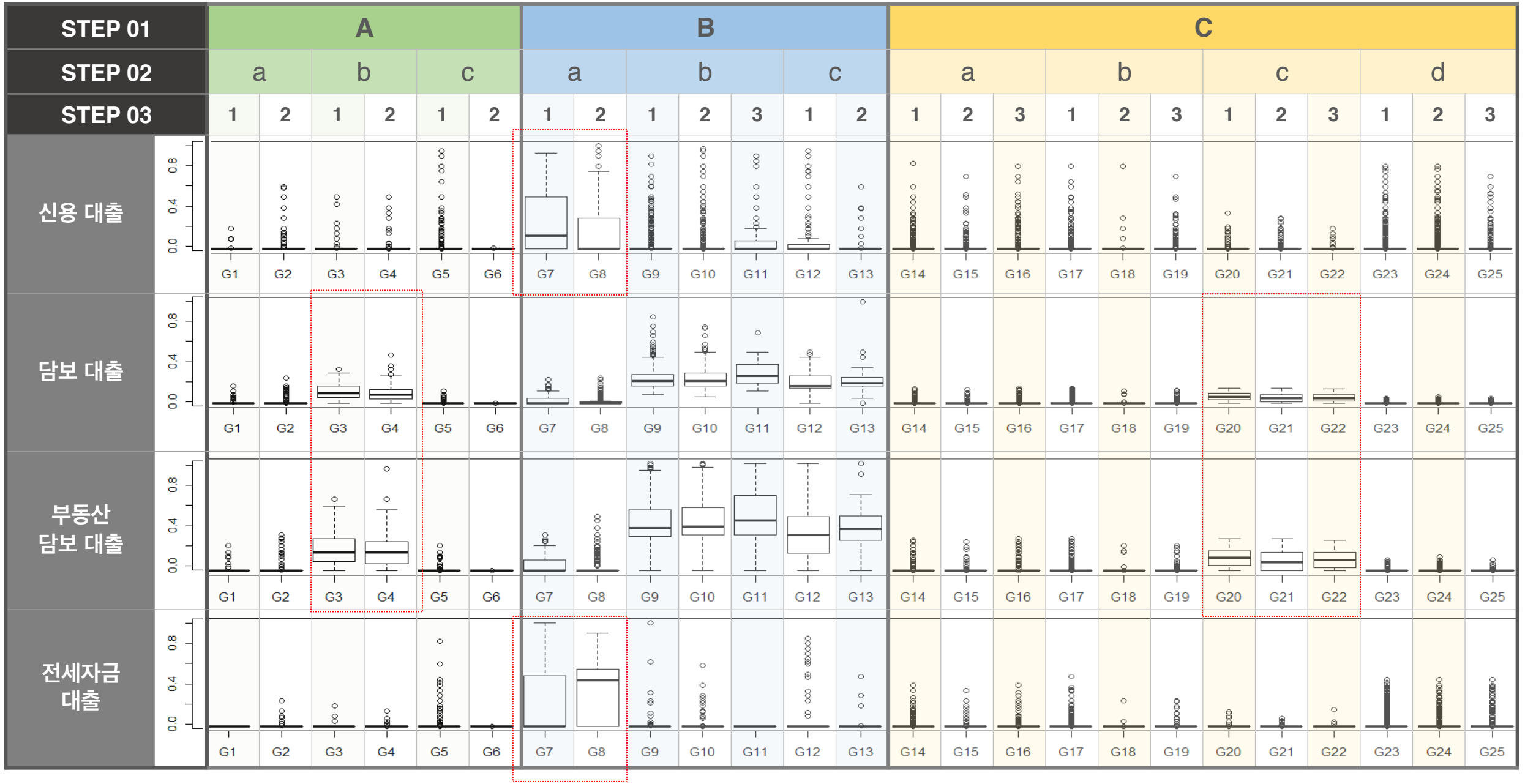
6_3. 자산 구성 항목 비교

- 자산구성 비율로 비교

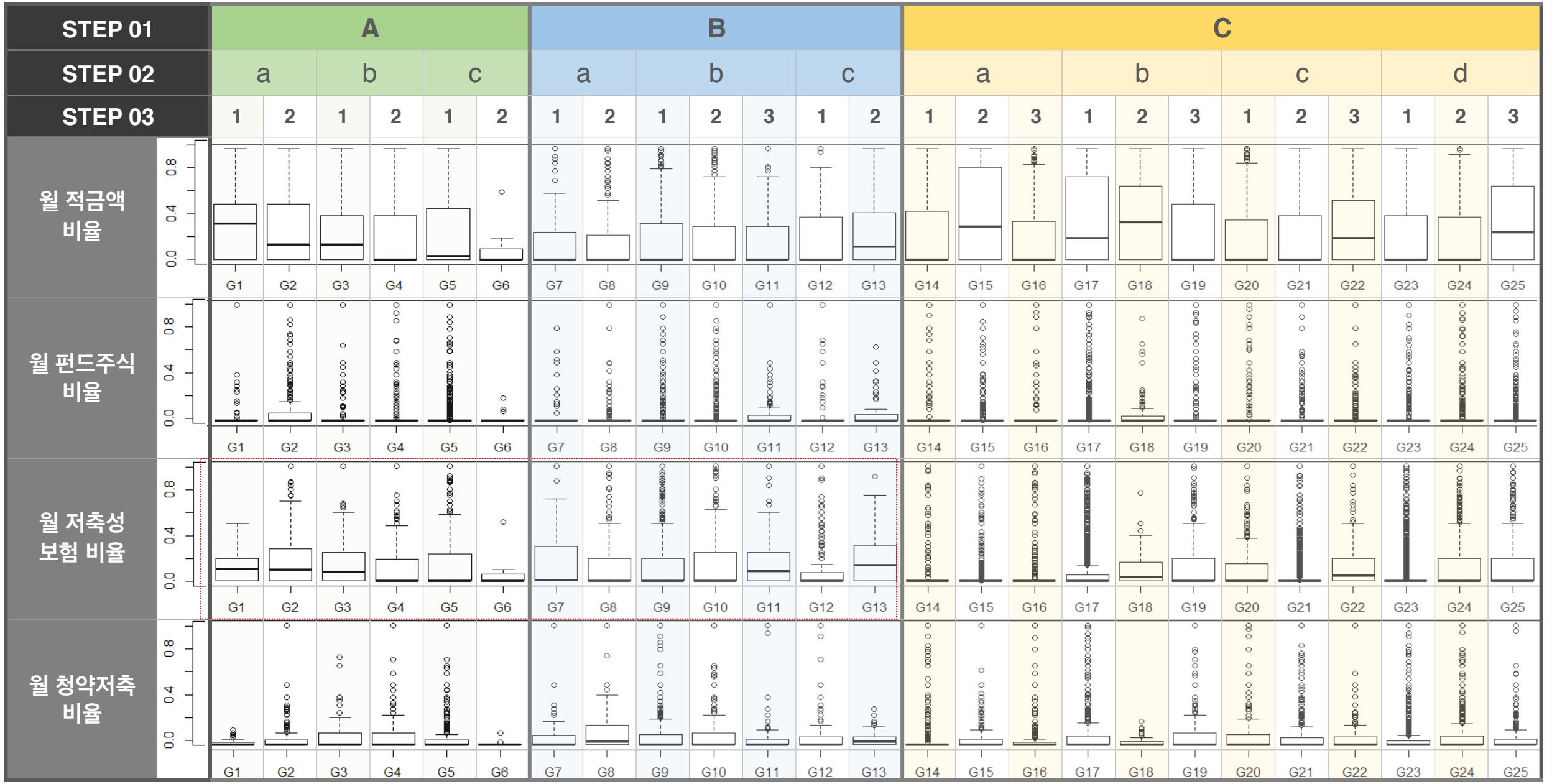
✓ STEP 02 그룹 비교



6_4. 부채 구성 항목 비교


STEP 01
A
B
C
STEP 02
a
b
c
a
b
c
a
b
c
d
STEP 03
1 2 1 2 1 2 1 2 1 2 3 1 2 1 2 3 1 2 3 1 2 3 1 2 3
신용 대출
담보 대출
부동산 담보 대출
전세자금 대출
0.0
0.4
0.8
G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15 G16 G17 G18 G19 G20 G21 G22 G23 G24 G25

6_5. 저축 비교_비율


STEP 01
A
B
C
STEP 02
a
b
c
a
b
c
a
b
c
d
STEP 03
1 2 1 2 1 2 1 2 1 2 3 1 2 1 2 3 1 2 3 1 2 3 1 2 3
월 적금액 비율
월 펀드주식 비율
월 저축성 보험 비율
월 청약저축 비율
0.0
0.4
0.8
G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15 G16 G17 G18 G19 G20 G21 G22 G23 G24 G25

7. 전체 결과 비교

step1	step2	step3	연령	가구소득	자산부채비교	저축 소비비교	자산구성항목비교	부채	저축
A	a	G1	40,50대	7	자산	저축	금융,부동산	적음	적금,저축성보험
		G2	40,50대	7	자산	-	금융,부동산	적음	적금,저축성보험
	b	G3	40,50대	7	자산	-	부동산	부동산담보	적금,저축성보험,청약
		G4	40,50대	7	자산	-	부동산	부동산담보	적금,저축성보험,청약
	c	G5	40,50대	7	자산	-	부동산	적음	적금,저축성보험
		G6	40,50대	7	자산	저축	부동산	적음	적금,저축성보험
B	a	G7	40,50대	6,7	부채	-	부동산	신용,전세자금	적금,저축성보험
		G8	30,40대	6,7	부채	-	부동산	신용,전세자금	적금,저축성보험
	b	G9	30,40대	6,7	부채	-	부동산	부동산담보	적금,저축성보험
		G10	30,40대	6,7	부채	-	부동산	부동산담보	적금,저축성보험
		G11	40,50대	7	부채	-	부동산	부동산담보	적금,저축성보험
	c	G12	고른분포	고른분포	부채	-	금융,기타	부동산담보	적금,저축성보험
		G13	30,40대	6,7	부채	-	금융	부동산담보	적금,저축성보험
C	a	G14	20대	1~3	-	-	기타,금융	적음	적금
		G15	20대	2~4	-	-	기타,금융	적음	적금,청약
		G16	고른분포	고른분포	-	-	기타,금융	적음	적금
	b	G17	20대,30대	2,3	-	-	금융	적음	적금,청약
		G18	20대~40대	6,7	자산	저축	금융	적음	적금,저축성보험
		G19	20대~40대	4~6	-	-	금융	적음	적금,저축성보험,청약
	c	G20	40대	4~6	-	-	부동산	부동산담보	적금,저축성보험,청약
		G21	30,40대	3~5	-	-	부동산	부동산담보	적금,청약
		G22	30대~50대	5~7	-	-	부동산	부동산담보	적금,저축성보험,청약
	d	G23	고른분포	3~5	자산	-	부동산	적음	적금,청약
		G24	40,50대	4~6	자산	-	부동산	적음	적금,저축성보험,청약
		G25	30대~50대	4~6	자산	저축	부동산	적음	적금,저축성보험,청약

STEP_02　Estimate NA (missing values)

STEP 2_ Estimate NA (missing values)

NA 값이 없는 Column (총 25개)
(Values : 17076)

성별, 연령, 직업구분, 지역구분,
가구소득구간, 총자산, 금융자산,
부동산자산, 기타자산,
월총저축액 등

예측

	수치 예측이 필요한 Column	NA 개수 (%)
1	은퇴후필요자금	11736개 (69%)
2	금융상품잔액_정기예금	10311개 (60%)
3	금융상품잔액_적금	8877개 (52%)
4	금융상품잔액_청약	9550개 (56%)
5	금융상품잔액_펀드	14237개 (83%)
6	금융상품잔액_ELS/DLS/ETF	16473개 (96%)

STEP 2_ Estimate NA (missing values)

수치예측 방법 A : 회귀분석

1. **독립변수(X) 선택**
   1) 상관계수 확인 – cor()
   2) 분산이 0에 가까운지 확인 – nearZeroVar()

2. **변수 정규화**
   1) 정규성 검사
   2) 정규성을 따르지 않을 경우
      최소값이 0인 변수: 범주화로 변경
      최소값이 0이 아닌 경우 : 로그함수 적용

3. **회귀분석으로 변수 선택 및 수치 예측**

수치예측 방법 B : 군집의 대표값

1. **해당 군집내 데이터 분석**
   1) NA를 제외한 데이터 및 빈도 확인
   2) 군집 내 데이터 추출 – sample()

2. **NA 값 예측 및 채우기**
   1) '금융자산' 기준 데이터 나열
   2) NA가 없는 값을 기준으로 Sample()함수에서 가능성(probability)에 따른 난수 생성
   3) 결측치 채우기

3. **데이터 검정 – var.test() 와 t.test()**

STEP 2_ Estimate NA (missing values)

수치예측 방법

A 회귀분석

A-1. 독립변수 (X) 선택

1) 상관계수 확인 - cor()
25개 예비 독립변수의 다중공선성 확인 및 제거

기준값	기준값과 0.5 이상의 상관관계를 갖는 변수
총자산	- 가구소득구간 제거 - 금융자산 제거 - 부동산자산 제거
결혼여부	- 연령_10세 단위 (결혼여부 제거) - 가구소득구간 제거
부채잔액과	- 부채잔액_담보대출 제거 - 부채잔액_아파트주택담보대출 제거
부채잔액_담보대출	- 부채잔액_아파트주책담보대출 제거
월저축액_펀드주식	- 월저축액_펀드 제거 - 월저축액_주식 제거
월총저축액	- 월저축액_적금 제거

변수간 상관관계 분석 결과,

상관관계가 높고 데이터의 의미가 중복되는 값을 가진 변수 제거 (총 9개)

- 가구소득구간
- 결혼여부
- 금융자산
- 부동산자산
- 월저축액_펀드
- 월저축액_주식
- 월저축액_적금
- 부채잔액_담보대출
- 부채잔액_아파트주택담보대출

A-1. 독립변수 (X) 선택

2) 분산이 0에 가까운지 확인 - nearZeroVar()
nearZeroVar() 함수로 분산값이 0인지 측정

```
#분산 0 확인
library(caret)
nearZeroVar(nocorr_data)
# 결과 8, 11, 12, 13 (펀드주식, 부채잔액, 부채잔액_신용대출, 부채잔액_전세자금대출)

# 분산이 0인 값 제거 후 nonzero_data 생성 (총 12개 변수 )
nozero_data <- nocorr_data[,-nearZeroVar(nocorr_data)]
str(nozero_data)
summary(nozero_data)
```

분산 값 확인 *nearZeroVar()* 코드

분산 값 확인 결과,

0으로 확인되어 독립변수에 영향을 미치지 않는 변수 제외 (총 4개)

- 펀드 주식
- 부채잔액
- 부채잔액_신용대출
- 부채잔액_전세자금대출

최종 선택된 독립변수 (12개)

성별, 연령, 직업구분, 지역구분, 총자산, 월총저축액, 월총소비금액, 기타자산
월평균카드사용금액, 노후자금융월저축액, 월저축액_저축성보험, 월저축액_청약,

A-2. 변수 정규화

독립변수(X)	데이터 유형	데이터 정제 방법
성별, 연령, 직업구분, 지역구분	범주형 데이터	
기타자산, 노후자금융월저축액, 월저축액_저축성보험, 월저축액_청약	수치형 데이터 (정규성 만족 X)	Min 값이 0인 변수 -> 0 인 값이 많아 분석이 어려움 -> 연속형 데이터를 **범주형으로 전환**
총자산,월총저축액, 월총소비금액	수치형 데이터 (정규성 만족 X)	Min 값이 0이 아닌 변수 **-> log 정규화 -> 정규성 만족 O**
월평균카드사용금액	수치형 데이터 (정규성 만족 O)	

데이터 범주화 기준

‘기타자산’
-> 이상치 제거 후(Q3 + IQR*1.5)
-> 박스플롯을 생성하여 Q1~4를 기준으로 범주화

‘노후자금융월저축액’
‘월저축액_저축성보험’,
‘월저축액_청약’,
-> YES/NO로 범주화 (0값이 다량)

```
data_gita_category <- transform(nozero_data,
    기타자산B = ifelse(기타자산 < 300, "0_300",
                    ifelse(기타자산 >= 300 & 기타자산< 1000, "300_1000",
                    ifelse(기타자산 >= 1000 & 기타자산 <2500, "1000_2500",
                    ifelse(기타자산 >= 2500 & 기타자산 <6750, "2500_6750",
                    ifelse(기타자산 >= 6759, "6750_upper", "no"
                                ))))))

category_data <- data_gita_category[c('기타자산', '기타자산B')]
head(category_data)
```

‘기타자산’리코딩 코드

A-3. 회귀분석으로 변수 선택 및 수치예측

회귀분석 수행 및 주요 변수 선택

Y = 금융상품잔액_적금 등 6개 변수 각각 대입

X = (정규화 및 리코딩 완료한) 12개의 변수

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -8.178e+03  5.872e+02 -13.927  < 2e-16 ***
X                       -1.258e-02  6.454e-03  -1.949 0.051295 .
성별                    -2.367e+02  6.448e+01  -3.672 0.000242 ***
연령_10세단위           -3.553e+01  3.482e+01  -1.020 0.307592
직업구분                 3.594e+00  1.066e+01   0.337 0.735991
지역구분                -5.780e+01  2.618e+01  -2.208 0.027251 *
총자산                   3.226e+03  3.005e+02  10.737  < 2e-16 ***
월총저축액               6.621e+02  3.749e+01  17.660  < 2e-16 ***
월총소비금액             8.102e+01  4.764e+01   1.701 0.089010 .
월평균카드사용금액       2.423e-06  4.719e-05   0.051 0.959060
기타자산B1000_2500      -3.519e+02  9.654e+01  -3.645 0.000269 ***
기타자산B2500_6750       1.252e+02  1.044e+02   1.199 0.230607
기타자산B300_1000       -3.087e+02  1.161e+02  -2.660 0.007831 **
기타자산B6750_upper      2.954e+02  1.374e+02   2.151 0.031522 *
월저축액_저축성보험BYes -3.689e+02  6.912e+01  -5.337 9.70e-08 ***
월저축액_청약BYes       -4.377e+02  6.386e+01  -6.853 7.75e-12 ***
노후자금용월저축액BYes   2.417e+02  6.653e+01   3.633 0.000282 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2800 on 8182 degrees of freedom
Multiple R-squared:  0.1277,    Adjusted R-squared:  0.126
F-statistic: 74.84 on 16 and 8182 DF,  p-value: < 2.2e-16
```

회귀 분석 결과
중요도가 별 3개 이상인 값을
독립변수로

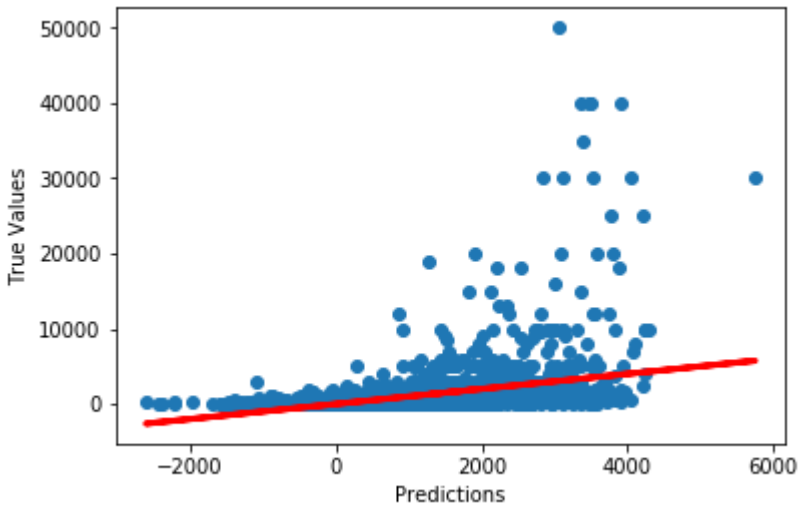
->

Linear Regression 수행

->

결과: 평균 예측률 0.1

A.회귀분석은 부적합한
수치예측 방법으로 판별

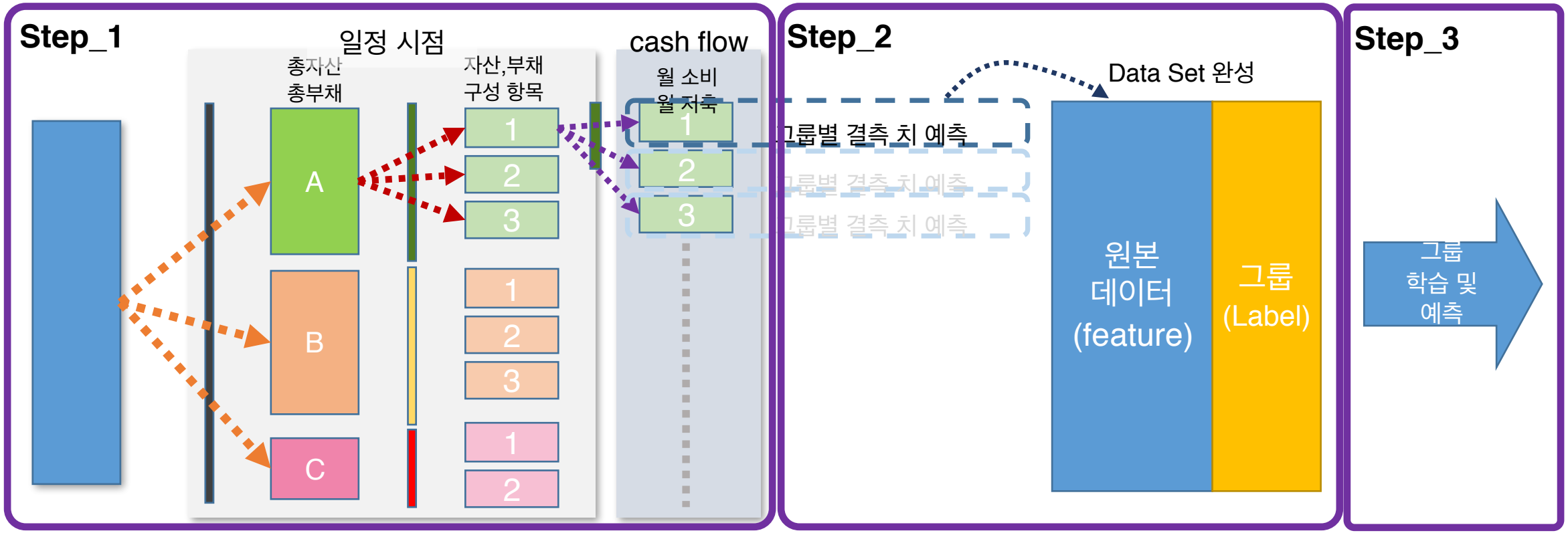


STEP 2_ Estimate NA (missing values)

수치예측 방법

B 군집의 대표값

B-1. 해당군집 내 데이터 분석



1. 25개의 군집된 결과를 바탕으로, 개별 군집을 나누어서 분석
2. 하나의 개별 군집의 Y 값 중 NA를 제외한 데이터의 값과 빈도(확률) 확인 후 -> 테이블로 생성
3. 테이블 값을 참조하여, NA 값 만큼 난수 생성 (오름차순 정렬)
4. NA 값에 대입 (*대입 시 NA값들의 '금융자산' 기준으로 오름차순으로 정렬한 후 대입) – sample() 함수 활용

* 기준을 '금융자산'으로 정한 이유 : 추정해야 할 Y 값과 가장 큰 상관관계를 수치적(상관계수 0.65이상) & 의미적으로 지님

B-2. NA값 예측 및 채우기

예시)

112 그룹의 '은퇴후필요자금' 값, 빈도, 확률 테이블 생성
(NA가 아닌 값을 기준으로)

```
> Frame
   Var1 Freq        prop
1   100    5 0.011574074
2   150   17 0.039351852
3   180    4 0.009259259
4   200  138 0.319444444
5   250   40 0.092592593
6   300  126 0.291666667
7   350   30 0.069444444
8   400   29 0.067129630
9   500   35 0.081018519
10  600    5 0.011574074
11  700    3 0.006944444
```

테이블 값을 기준으로
301개 (NA 값의 수)의
난수 생성 후

오름차순 정렬

112 그룹의 '금융자산' 기준
오름차순 정렬 데이터셋에,
생성된 값 대입

idx	금융자산	은퇴후필요자금	step123
23769	1520	NA	112
38079	3925	NA	112
4813	6050	NA	112
34475	8700	NA	112
1125	8900	NA	112
5500	10000	NA	112
28806	10000	NA	112
29697	10000	NA	112
14719	10500	NA	112
1990	11500	NA	112
38114	12000	NA	112
4828	12100	NA	112
18435	12300	NA	112
30320	13000	NA	112
31893	13830	NA	112
11899	14700	NA	112
32095	15000	NA	112
11060	15200	NA	112
32925	15620	NA	112

총 301개

1. 25개의 군집된 결과를 바탕으로, 개별 군집을 나누어서 분석
2. 하나의 개별 군집의 Y 값 중 NA를 제외한 데이터의 값과 빈도(확률) 확인 후 -> 테이블로 생성
3. 테이블 값을 참조하여, NA 값 만큼 난수 생성 (오름차순 정렬)
4. NA 값에 대입 (*대입시 NA값들의 '금융자산' 기준으로 오름차순으로 정렬한 후 대입) – sample() 함수 활용

* 기준을 '금융자산'으로 정한 이유 : 추정해야 할 Y 값과 가장 큰 상관관계를 수치적(상관계수 0.65이상) & 의미적으로 지님

B-2. NA값 예측 및 채우기

‘은퇴후필요자금’
‘금융상품잔액_정기예금’
‘금융상품잔액_적금’
‘금융상품잔액_펀드’
‘금융상품잔액_ELS/DLS/ETF’

앞의 예시와 동일한 방법으로
25개 군집의 5개의 항목 NA값 추정 및 대입

‘금융상품잔액_청약’

‘청약보유여부’에 따라서
청약보유를 하지 않는 사람의 값을 0으로 처리

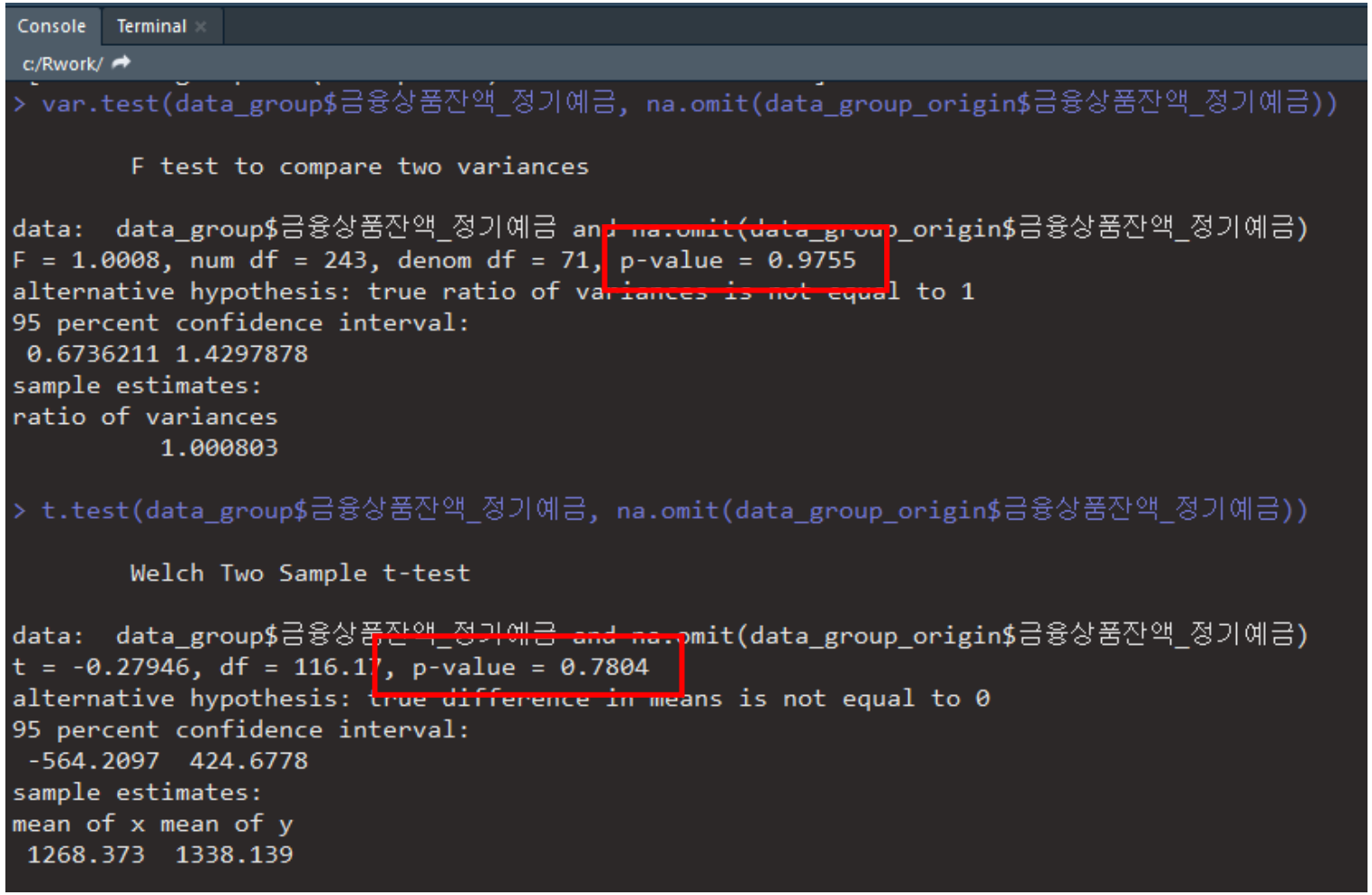
총 6개의 Column 결측치 추정 완료

idx	청약보유여부	금융상품잔액_청약	step123
5	0	NA	311
7	1	800	321
9	1	1200	222
10	1	350	222
11	0	NA	331
13	1	1000	323
16	0	NA	311
18	0	NA	312
20	1	800	341
26	0	NA	342
28	1	200	341
29	0	NA	311
30	0	NA	331
31	1	300	342
32	1	300	343
33	1	400	212
34	1	500	343
37	1	400	312
38	1	400	343
40	0	NA	111
41	0	NA	221

청약을 보유하지 않은 사람
-> ‘금융상품잔액_청약’ 값을 0으로 처리

B-3. 데이터 검정

var.test() 와 t.test()를 통한 데이터 검정



```
Console  Terminal
c:/Rwork/
> var.test(data_group$금융상품잔액_정기예금, na.omit(data_group_origin$금융상품잔액_정기예금))

        F test to compare two variances

data:  data_group$금융상품잔액_정기예금 and na.omit(data_group_origin$금융상품잔액_정기예금)
F = 1.0008, num df = 243, denom df = 71, p-value = 0.9755
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6736211 1.4297878
sample estimates:
ratio of variances
          1.000803

> t.test(data_group$금융상품잔액_정기예금, na.omit(data_group_origin$금융상품잔액_정기예금))

        Welch Two Sample t-test

data:  data_group$금융상품잔액_정기예금 and na.omit(data_group_origin$금융상품잔액_정기예금)
t = -0.27946, df = 116.17, p-value = 0.7804
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -564.2097  424.6778
sample estimates:
mean of x mean of y
 1268.373  1338.139
```

예시)
212그룹 '금융상품잔액_정기예금'의

NA 값 추정 전 데이터와
NA 값 추정 후 데이터를
바탕으로 테스트 실행

var.test() 의 p-value = 0.9755
t.test() 의 p-value = 0.7804

등분산성과, 동일한 평균 만족
(*다른 Column 값도 var, t검정결과
최소 0.2 이상의 p-value 값 확인)

B '군집의 대표값' 방법을
결측치 예측 방법으로 채택

STEP_03 Peer Group Prediction

STEP 3_ Peer Group Prediction

1. 변수 선택

1) 랜덤포레스트로 VarImp() 확인 및 주요 변수 선정

2) 상관계수 확인 및 변수 삭제

2. Data Set 만들기

3. 분류 모델 학습 및 예측 (랜덤 포레스트)

1) 교차분석 테이블

2) 모델평가지표

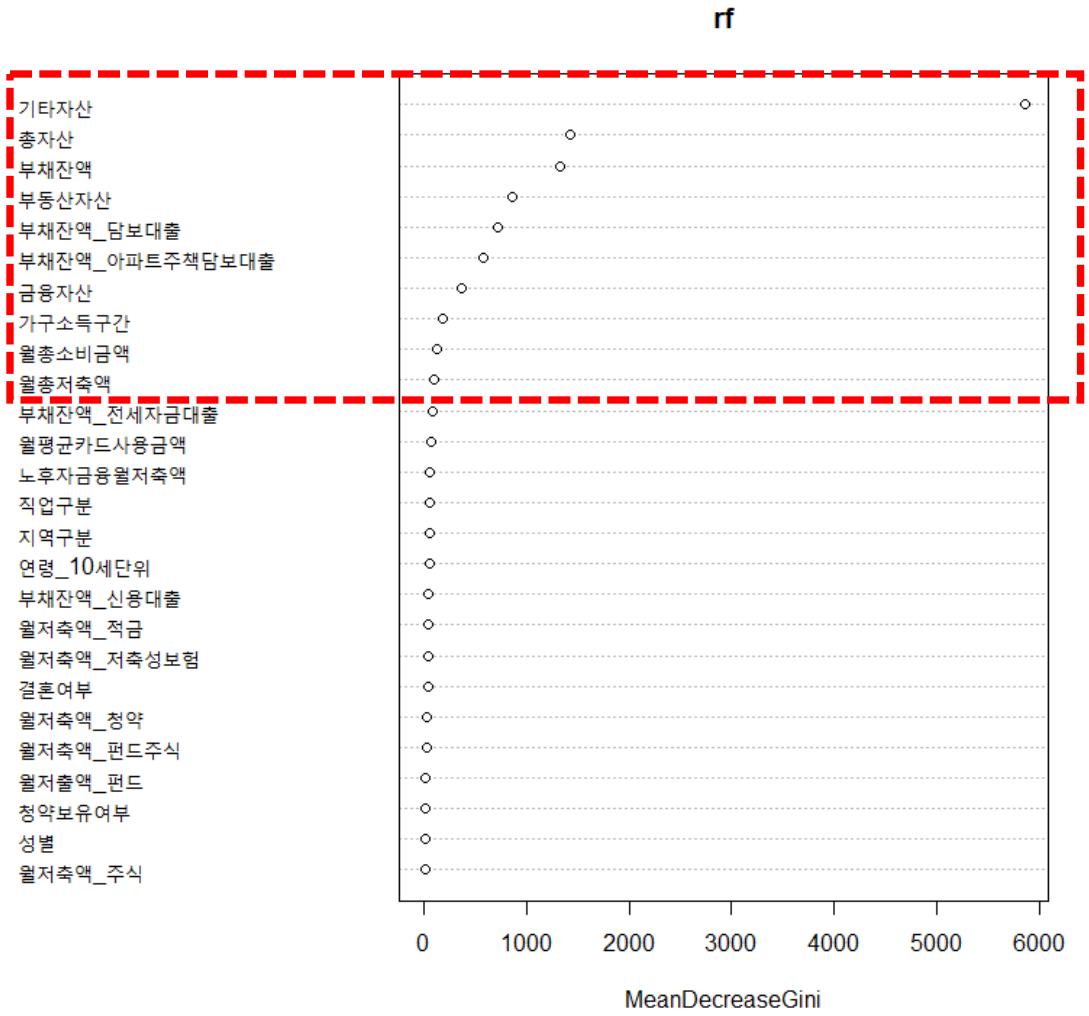
3) ROC Curve, AUC

1. 변수 선택

1) 랜덤포레스트로 varImp() 확인

```
> importance(rf)
```

	MeanDecreaseGini
성별	11.84579
연령_10세단위	50.76797
직업구분	54.26580
지역구분	54.13859
가구소득구간	183.24775
결혼여부	45.25798
총자산	1420.83937
금융자산	372.00294
부동산자산	858.46628
기타자산	5853.35689
월총저축액	98.24597
월저축액_적금	46.99702
청약보유여부	11.94438
월저축액_펀드주식	21.81224
월저출액_펀드	18.03736
월저축액_주식	11.44857
월저축액_저축성보험	45.62670
월저축액_청약	33.02853
부채잔액	1329.62547
부채잔액_신용대출	47.93526
부채잔액_담보대출	718.00401
부채잔액_아파트주책담보대출	585.45916
부채잔액_전세자금대출	80.83430
노후자금융월저축액	56.34360
월총소비금액	121.05885
월평균카드사용금액	73.22392



선택!

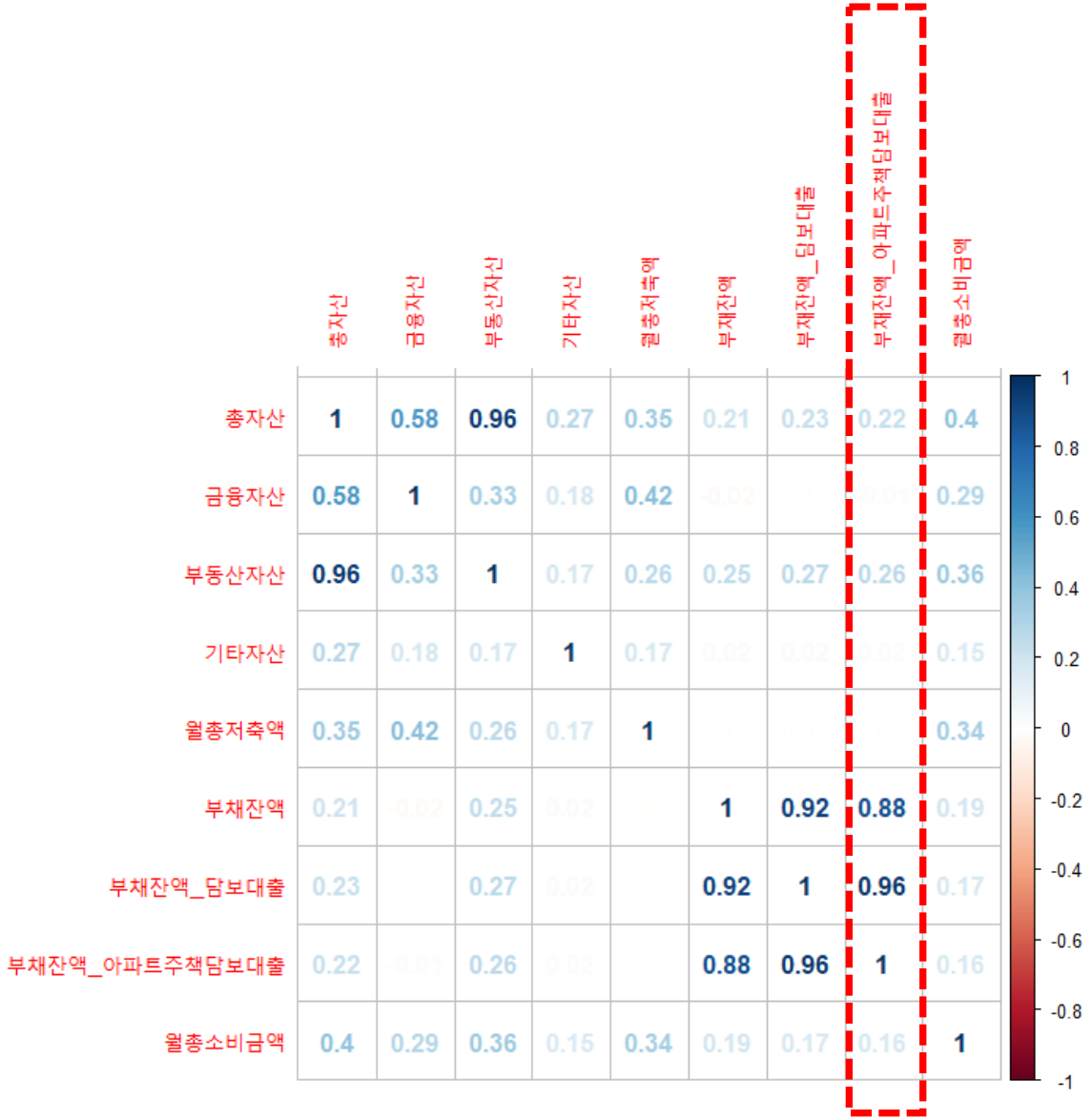
1. 변수 선택

2) 상관계수 확인

변수 삭제 : 아파트 주택담보대출
(다량의 변수와 상관계수가 높음)

최종 변수 선택 (8개 column)

- ✓ 총자산
- ✓ 부채 잔액
- ✓ 금융자산
- ✓ 부동산 자산
- ✓ 기타자산
- ✓ 부동산담보대출
- ✓ 월저축액
- ✓ 월 소비금액



2. Data set 만들기

학습데이터 생성

1. Train dataset, Test dataset 분할
   1) 층화 임의 추출
   2) Train : 50 % / Test = 50 %

2. Train dataset UP Sampling
   1) 분류 된 그룹별 데이터 개수 편차가 크다.
      (MAX: 2780, MIN:14)
   1) 가장 큰 그룹의 개수에 맞추어 UP Sampling

step1	step2	step3	n
1	1	1	62
		2	432
	2	1	122
		2	280
	3	1	711
		2	14
2	1	1	83
		2	244
	2	1	928
		2	818
		3	129
	3	1	168
		2	63
3	1	1	1023
		2	699
		3	721
	2	1	2045
		2	98
		3	634
	3	1	718
		2	983
		3	401
	4	1	2780
		2	1694
		3	1226

111	31
112	216
121	61
122	140
131	356
132	7
211	42
212	122
221	464
222	409
223	65
231	84
232	32
311	512
312	350
313	361
321	1023
322	49
323	317
331	359
332	492
333	201
341	1390
342	847
343	613

3. 분류모델 학습 및 예측

랜덤포레스트 결과 1 : 교차분석 테이블

Prediction \ Reference	111	112	121	122	131	132	211	212	221	222	223	231	232	311	312	313	321	322	323	331	332	333	341	342	343
111	21	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
112	7	195	3	2	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
121	2	0	52	1	12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
122	0	5	4	126	7	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
131	0	16	0	4	327	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
132	1	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
211	0	0	0	0	1	0	26	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
212	0	0	0	0	0	0	1	101	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1	4	0
221	0	0	0	0	0	0	0	10	443	1	1	3	0	0	0	0	0	0	0	1	4	0	0	0	0
222	0	0	1	4	0	0	6	6	0	388	9	3	8	0	0	0	0	0	0	3	0	0	0	0	0
223	0	0	1	0	0	0	2	0	0	7	52	0	2	0	0	0	0	0	0	1	0	0	0	0	0
231	0	0	0	0	0	0	2	1	5	1	0	75	2	0	0	0	2	0	1	3	1	0	0	0	0
232	0	0	0	0	0	0	0	0	3	0	0	1	17	0	0	0	1	0	0	0	0	0	0	0	0
311	0	0	0	0	0	0	0	0	0	0	0	0	0	498	1	2	9	0	0	0	6	0	3	0	0
312	0	0	0	0	0	0	0	0	0	0	0	0	0	2	322	12	9	1	1	0	2	2	6	0	0
313	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	318	7	0	6	2	8	0	7	8	0
321	0	0	0	0	0	0	0	0	0	0	0	0	0	6	6	2	972	5	6	0	15	0	20	0	3
322	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	36	4	0	0	0	0	0	1
323	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	7	6	6	277	0	0	2	7	9	4
331	0	0	0	3	0	0	0	1	3	5	1	0	0	0	0	4	0	0	2	338	0	0	1	39	0
332	0	0	0	0	0	0	0	0	6	0	0	0	0	5	3	3	3	0	3	0	449	1	48	0	0
333	0	0	0	0	0	0	1	0	1	5	1	0	1	0	5	2	0	0	3	8	2	195	1	1	13
341	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	5	11	0	2	0	4	0	1286	1	6
342	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4	0	0	4	3	0	0	6	767	2
343	0	0	0	0	0	0	0	1	0	0	0	0	0	0	4	0	1	1	8	0	0	0	4	15	583

3. 분류모델 학습 및 예측

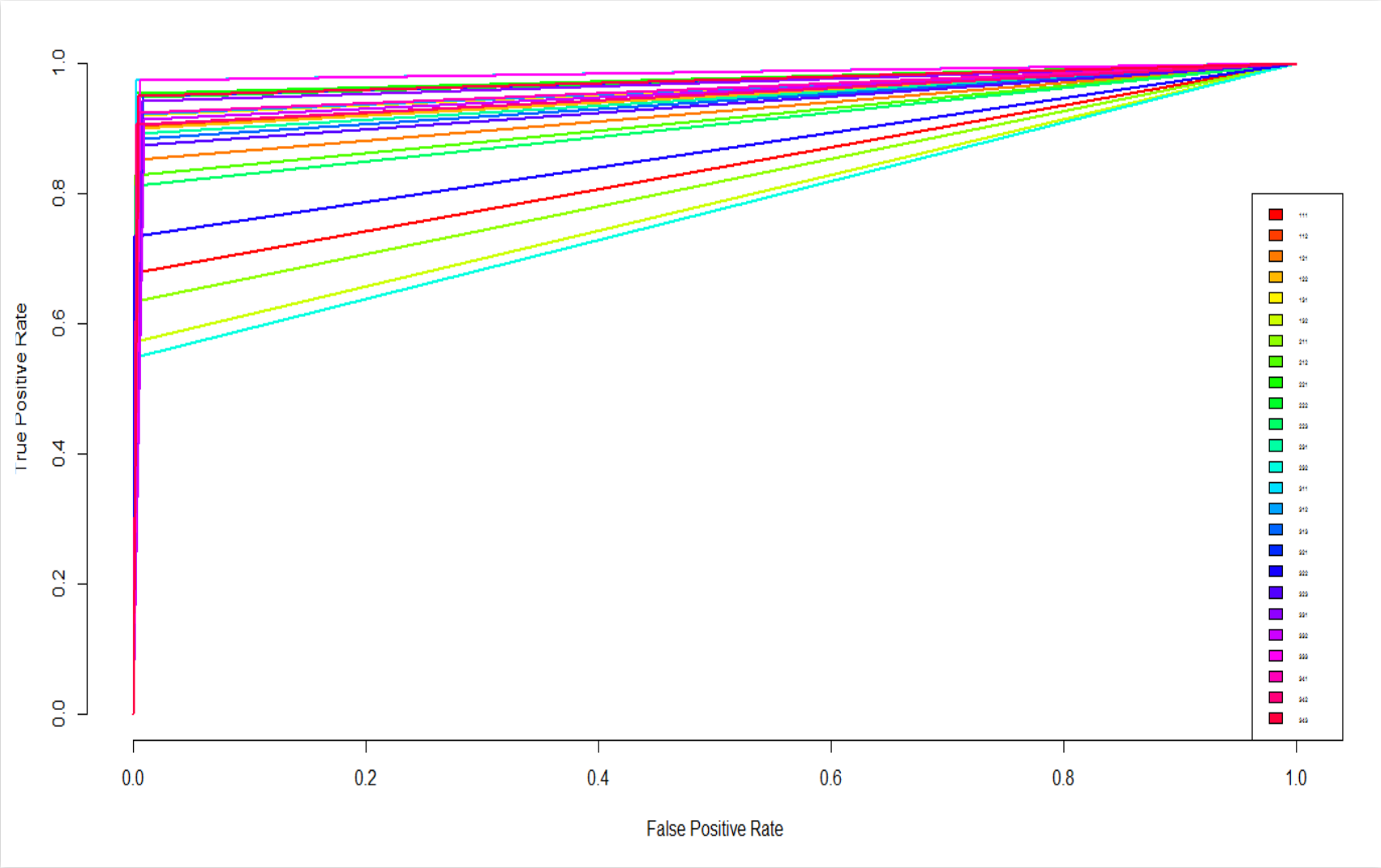
랜덤포레스트 결과 2 : 모델 평가 지표

```
Overall Statistics

               Accuracy : 0.9221
                 95% CI : (0.9162, 0.9277)
    No Information Rate : 0.1629
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9156
```

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1
Class: 111	0.6774194	0.9996471	0.8750000	0.9988248	0.8750000	0.6774194	0.7636364
Class: 112	0.9027778	0.9975953	0.9069767	0.9974754	0.9069767	0.9027778	0.9048724
Class: 121	0.8524590	0.9981114	0.7647059	0.9989368	0.7647059	0.8524590	0.8062016
Class: 122	0.9000000	0.9976171	0.8630137	0.9983307	0.8630137	0.9000000	0.8811189
Class: 131	0.9211268	0.9966985	0.9237288	0.9965766	0.9237288	0.9211268	0.9224260
Class: 132	0.5714286	0.9998827	0.8000000	0.9996482	0.8000000	0.5714286	0.6666667
Class: 211	0.6341463	0.9996467	0.8965517	0.9982361	0.8965517	0.6341463	0.7428571
Class: 212	0.8278689	0.9989300	0.9181818	0.9975068	0.9181818	0.8278689	0.8706897
Class: 221	0.9547414	0.9975214	0.9568035	0.9973978	0.9568035	0.9547414	0.9557713
Class: 222	0.9486553	0.9950763	0.9065421	0.9974090	0.9065421	0.9486553	0.9271207
Class: 223	0.8125000	0.9984650	0.8000000	0.9985829	0.8000000	0.8125000	0.8062016
Class: 231	0.8928571	0.9978696	0.8064516	0.9989336	0.8064516	0.8928571	0.8474576
Class: 232	0.5483871	0.9994119	0.7727273	0.9983551	0.7727273	0.5483871	0.6415094
Class: 311	0.9745597	0.9973822	0.9595376	0.9983778	0.9595376	0.9745597	0.9669903
Class: 312	0.9226361	0.9957234	0.9019608	0.9966977	0.9019608	0.9226361	0.9121813
Class: 313	0.8833333	0.9952282	0.8907563	0.9948630	0.8907563	0.8833333	0.8870293
Class: 321	0.9510763	0.9916123	0.9391304	0.9933316	0.9391304	0.9510763	0.9450656
Class: 322	0.7346939	0.9991749	0.8372093	0.9984688	0.8372093	0.7346939	0.7826087
Class: 323	0.8738170	0.9947663	0.8656250	0.9951297	0.8656250	0.8738170	0.8697017
Class: 331	0.9415042	0.9927820	0.8513854	0.9974189	0.8513854	0.9415042	0.8941799
Class: 332	0.9144603	0.9910470	0.8618042	0.9947579	0.8618042	0.9144603	0.8873518
Class: 333	0.9750000	0.9947198	0.8158996	0.9993972	0.8158996	0.9750000	0.8883827
Class: 341	0.9251799	0.9953801	0.9749810	0.9855836	0.9749810	0.9251799	0.9494278
Class: 342	0.9055490	0.9973979	0.9745870	0.9896721	0.9745870	0.9055490	0.9388005
Class: 343	0.9510604	0.9957071	0.9448947	0.9962102	0.9448947	0.9510604	0.9479675

3. 분류모델 학습 및 예측

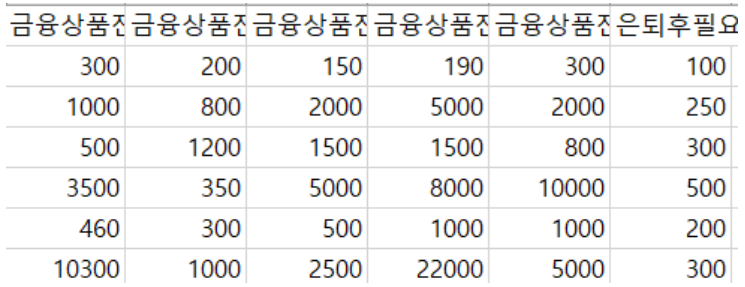
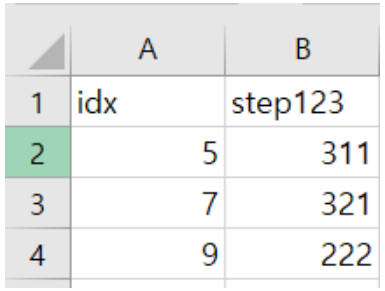
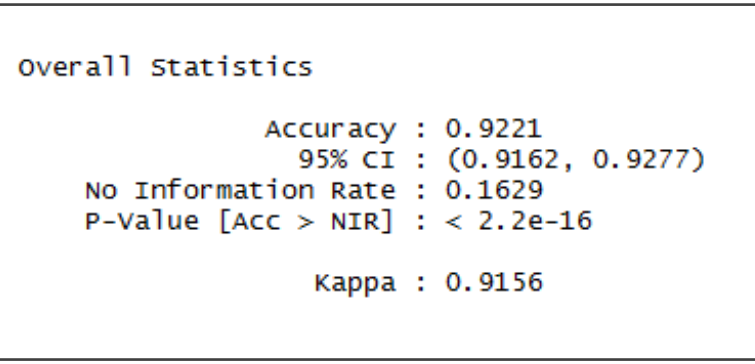
랜덤포레스트 결과 3 : ROC Curve, AUC




```
111 AUC : 0.838533248343843
112 AUC : 0.950186532269916
121 AUC : 0.925285221133454
122 AUC : 0.948808530918623
131 AUC : 0.958912609922189
132 AUC : 0.785655641566972
211 AUC : 0.81689653389704
212 AUC : 0.913399412556936
221 AUC : 0.976131378712057
222 AUC : 0.971865786904448
223 AUC : 0.905482494981698
231 AUC : 0.94536335661025
232 AUC : 0.773899499927911
311 AUC : 0.985970942920667
312 AUC : 0.959179732905355
313 AUC : 0.939280761858151
321 AUC : 0.971344311448232
322 AUC : 0.866934397521385
323 AUC : 0.934291672170021
331 AUC : 0.96714308497696
332 AUC : 0.952753644182705
333 AUC : 0.984859894395776
341 AUC : 0.96027997425663
342 AUC : 0.951473431354206
343 AUC : 0.973383714798886
```

결과 정리

과제 별 수행 내역

과제 1	금융거래정보항목의 결측치를 추정하여 Data Set완성	
분류 된 그룹들의 분포에 맞춰 결측 치 추정 하여 dataset 완성		
과제 2	Data Set의 금융거래정보를 이용하여 유사한 집단 Peer Group으로 묶어, 결과 Mapping Table를 완성	
3단계 군집분석으로 총 25개 군집 생성		
과제 3	고객이 정보를 입력 하면 소속된Peer Group 예측	
랜덤포레스트를 이용한 소속 그룹 예측		