

데이터분석 전문가를 공부하는 독자 여러분

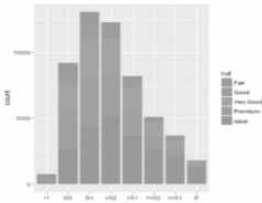
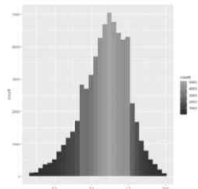
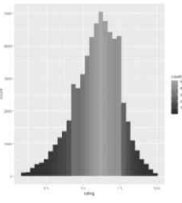
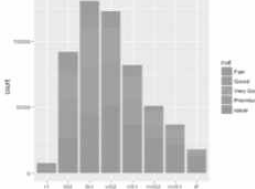
수험생 여러분들의 니즈를 반영하기 위해 많은 부분을 추가하고 수차례의 오타와 오류를 수정하였음에도 불구하고 최종 인쇄본에서 예상치 못한 오타와 오류가 발생한 점 깊이 사과의 말씀을 전해드립니다.

아래와 같이 현재까지 발견한 오타와 추가내용을 정오표로 정리하여 첨부하오니 학습에 참고하시기 바랍니다. 또한 데이터에듀([www.dataedu.kr](http://www.dataedu.kr)) 사이트의 공지사항에 정오표를 계속 업데이트 하도록 하겠습니다.

학습 도중에 추가로 오타자를 발견하시면 이메일([books@dataedu.co.kr](mailto:books@dataedu.co.kr))로 보내주시면 해당 내용을 추가하여 정오표를 업데이트하도록 하겠습니다.

2020년 3월 20일 이후 판매한 초판 3쇄본의 정오표입니다.

- 정오표 -

페이지	세부 위치	수정 전 내용	수정 후 내용
153	1단계 귀무가설	귀무가설( $H_0$ ): Enginesize... 대립가설( $H_1$ ): 적어도...	귀무가설( $H_0$ ): $\beta_1(EngineSize) = \beta_2(RPM) = \beta_3(weight)$ 대립가설( $H_1$ ): 적어도 하나의 <b>설명변수</b> <b>계수 값은 0이 아니다.</b>
	2단계 귀무가설	귀무가설( $H_0$ ): Enginesize=0 대립가설( $H_1$ ): Enginesize $\neq$ 0	귀무가설( $H_0$ ): $\beta_1(EngineSize)=0$ 대립가설( $H_1$ ): $\beta_1(EngineSize) \neq 0$
157	6단계 귀무가설	귀무가설( $H_0$ ): Education... 대립가설( $H_1$ ): 적어도...	귀무가설( $H_0$ ): $\beta_1(Education) = \dots = \beta_4(Agriculture)$ 대립가설( $H_1$ ): 적어도 하나의 <b>설명변수</b> <b>계수 값은 0이 아니다.</b>
	7단계 귀무가설	귀무가설( $H_0$ ): Education=0 대립가설( $H_1$ ): Education $\neq$ 0	귀무가설( $H_0$ ): $\beta_1(Education)=0$ 대립가설( $H_1$ ): $\beta_1(Education) \neq 0$
164	2단계	... 신뢰구간은 80~90% 사이이고...	... 신뢰구간은 <b>80~95%</b> 사이이고...
167	6단계 귀무가설	귀무가설( $H_0$ ): balance=0	귀무가설( $H_0$ ): $\beta_1(balance)=0$
168	상단 대립가설	대립가설( $H_1$ ): balance $\neq$ 0	대립가설( $H_1$ ): $\beta_1(balance) \neq 0$
	로지스틱 회귀식 밀에	StudentYes 이 1 증가할수록 졸업할 확률이 0.49배 증가	StudentYes가 <b>한단위 증가할수록</b> <b>체납할 확률이 체납하지 않을 확률의</b> <b>0.49배</b>
171	제일 잇줄	Petal.Length $\geq$ 2.45이며 Petal.Width $<$ 1.75인 54개를 versicolor로 분류하고, Petal.Length $\geq$ 2.45이며 Petal.Width $\geq$ 1.75인 46개를 virginica로 분류했다.	Petal.Length $\geq$ 2.45이며 Petal.Width $<$ 1.75인 <b>54개 중 49개</b> 를 versicolor로 분류하고, Petal.Length $\geq$ 2.45이며 Petal.Width $\geq$ 1.75인 <b>46개 중 45개</b> 를 virginica로 분류했다.
197	용어	MPP(Massive Parrallel Processing)	MPP( <b>Massively</b> Parrallel Processing)
220	1번 문제	② ... 변경된 데이터도 캡처할 수 있다.	② ... 변경된 데이터는 캡처할 수 <b>없다</b> .
482	11번 보기	②번  ④번 	②번  ④번 
506	70번 보기	④mtcars[,1:9]	<b>④mtcars[,c(1:9)]</b>
508	76번 문제	...그래프의 caret종류는...	...그래프의 <b>carat</b> 종류는...
510	객관식 정답 표	44번 ③	<b>44번 ②</b>
513	44번 해설	회귀분석의 가정은 선형성, 등분산성, 독립성, 비상관성, 정규성이 있다. 아래의 그림은 회귀모형이 비선형성을 띄므로 선형성을 위배했다고 판단할 수	회귀분석의 가정은 선형성, 등분산성, 독립성, 비상관성, 정규성이 있다. 아래의 그림은 회귀모형이 <b>등분산성을</b> 위배했다고 판단할 수 있다.

		있다.	
520	14번 보기	③ HDFS는 데이터를 청크 단위로 저장하는 시스템이다.	③ HDFS는 데이터를 <b>파일</b> 단위로 저장하는 시스템이다.
535	66번 보기	④ ... 이상값이 존재한다.	④ ... 이상값이 존재 <b>하지 않는다</b> .
541	29번 정답	②	③
543	29번 해설	채널영역은...	채널은 기업이 고객세그먼트에게 가치를 제안하기 위해 커뮤니케이션을 하고 상품이나 서비스를 전달하는 방법을 의미한다. 커뮤니케이션, 물류, 판매채널 등 기업과 고객의 인터페이스 전반이 바로 채널이다. 유통 채널을 공급하는 것은 채널영역이 아니다.
551	16번 보기	②EAI는...	② <b>ESB</b> 는...
552	22번 보기	④ 훈련용 데이터를...	④ 훈련용 데이터를 활용하여 분류, 예측, 군집 등의 모델을 <b>만들고 이를 평가, 검증한다</b> .
562	60번 보기	④다변량 자료...	<b>④이상치 탐지에 있어 활용할 수 없다.</b>
564	63번 보기	③...변수와 양의 상관... ④... hede2,...	③...변수와 <b>음</b> 의 상관... ④... <b>head2</b> ,...
568	79번 문제	②데이터 수집 ③배열	② <b>배열</b> ③ <b>데이터 수집</b>
575	60번 해설	주성분분석은...	<b>주성분분석은 서로상관성이 높은 변수들의 선형 결합으로 만들어 기존의 상관성이 높은 변수들을 요약, 축소하는 기법으로 이상치 탐지에도 사용한다.</b>
590	39번 보기	Ytest<-subset(Hitters[-train]... Ytest<-subset(Hitters[-train]...	Ytest<-subset(Hitters[-train]... <b>X</b> test<-subset(Hitters[-train]...
597	60번 보기	① 0.3/(0.7x0.45) ③ 0.4/(0.6x0.45) ④ 0.4/(0.6x0.45)	① 0.3/(0.6x0.45) ③ 0.4/(0.6x0.45) ④ 0.4/(0.7x0.45)
618	50번 보기	③... 일부 VIF가 1보다 크면....	③... 일부 VIF가 <b>4</b> 보다 크면....
	51번 보기	② speed 변수의 변동성...	② <b>위의 결과에서 회귀식을 도출하면 dist=-2.601 + 9.464*speed</b>
620	54번 보기	④ 기각역이란...	④ <b>검정력</b> 이란...
622	61번 보기	②연관성분석은...	②연관성분석은...잘 알려진 <b>사실</b> 이면서 <b>분명하고 유용한 사실</b> 이어야만 한다.
631	11회 답안	63번 ③	<b>63번 ②</b>

## - 모의고사 서술형 답안 -

### <1회 모의고사>

1) 신용카드를 사용하는 고객의 체납 확률을 예측하기 위한 방법을 제시하시오.

→ str 함수를 통해 데이터의 구조와 summary 함수를 통해 데이터에 대한 기초 통계량을 확인했을 때, default(체납 여부)와 student(학생여부)는 범주형 변수이며, balance(카드 잔고), income(연봉)은 수치형 변수임을 확인할 수 있다. 고객의 체납 확률을 예측하기 전, 종속변수와 독립변수를 나누어 보면 종속변수는 default이며, 설명변수는 student, balance, income이다. 체납 여부인 default를 분류하기 위해서는 정형 데이터마이닝 중 분류분석을 사용하여 체납 확률을 예측해야 한다. 분류 분석의 방법으로는 로지스틱 회귀분석, 의사결정나무, 앙상블기법(배깅, 부스팅, 랜덤포레스트), 인공신경망, 나이브 베이지안, K-NN(K-Nearest Neighbor), SVM(Support Vector Machine) 등이 있다.

이 중 로지스틱 회귀분석 방법을 활용한다면 R 프로그램에서 glm함수 등을 사용하여 모델을 구축한다. 예를 들어 glm(default~student+balance+income, data=Default, family="binomial") 라는 코드로 모델을 구축하는 것은 일반화 선형 모델 구축 함수인 glm 함수를 이용해 Default 데이터를 사용하여 종속변수 default에 대해 모든 설명변수(student, balance, income)를 사용하여 모델을 구축할 수 있다. 분석결과가 나타난다면 먼저 설명변수에 대해 통계적 타당성을 가설검정하여 설명변수가 모두 유의한지를 파악한다. 유의하지 않은 변수가 포함될 수도 있으므로 step함수를 활용하여 변수선택법(전진선택법, 후진제거법, 단계선택법)으로 최적의 모형을 찾을 수 있다. 마지막으로 최종적으로 로지스틱 회귀분석의 결과를 종합하여 로지스틱 회귀식을 도출한다. 이 때, 일반 회귀분석과 다르므로

$P(X) = \frac{1}{1 + \exp(x)}$ 의 식에 맞게 회귀식을 도출하여야한다.

로지스틱 회귀분석을 진행할 때 주의사항은 분류분석이기 때문에 종속변수가 연속형 변수값이면 예측에 적당하지 않으므로 해당 종속변수를 구간화 등을 통해 범주형 변수로 변환하여 분석에 적용해야 한다.

2) 분석을 통해 얻은 결과물로부터 발견할 수 있는 인사이트를 예시로 설명하시오.

→ 예를 들어 단계적 선택법을 활용하여 로지스틱 회귀분석 결과를 얻었을 때, income(수입)의 변수는 유의하지 않아 제거되고 balance(카드 잔고)의 계수는 양수, StudentYes(학생여부 중 학생일 때)의 계수가 음수로 나타났다고 가정하자.

이 때, 로지스틱 회귀식은  $P(X) = \frac{1}{1 + \exp(-(상수 + a*balance - b*studentYes))}$ 로 나타낼 수 있으며 다른 설명변수의 조건이 동일할 때, studentYes이 1 증가할수록 졸업할 확률은 0보다 작은 값이나와 0.xx배 증가, 즉 학생일수록 체납확률이 낮아진다고 볼 수 있다. 로지스틱 회귀식을 통해 카드잔고가 증가할수록 체납여부는 증가하고 학생일수록 체납확률이 낮아질 것이라고 예측할 수 있다. 마지막으로 StudentYes처럼 해당 변수의 특징을 파악해보면 학생 여부를 0과 1로 나타낸 변수이므로 결과해석을 유의해서 해야된다.

이러한 인사이트를 통해 신용카드 개설을 위한 조건 강화를 통해 카드사의 손해를 줄일 수 있는 방법을 강구해야 할 것이라고 인사이트를 도출할 수 있다. 또, 카드 잔고가 많은 고객들이 소비할 수 있도록 마케팅 전략 등을 강화하는 방안도 마련하는 인사이트도 도출할 수 있다.

## <2회 모의고사>

1) 비슷한 특징을 가진 집단으로 그룹화하기 위한 적절한 방법론을 제시하시오.

→ 먼저 summary함수를 통해 Private변수를 제외한 모든 변수는 수치형 변수임을 알 수 있다. 비슷한 특징을 가진 집단으로 그룹화 하기위한 방법은 군집분석이다. 군집분석은 계층적 군집분석과 비계층적 군집분석으로 나눌 수 있다. 계층적 군집분석은 전통적 군집분석 방법으로 군집의 개수가 제일 나중에 선정되며, 방법으로는 최단, 최장, 평균, 와드(ward) 연결법이 있다. 또, R 프로그램에서는 hclust 함수를 이용하여 계층적 군집분석을 할 수 있으며, 이 결과로 덴드로그램 시각화가 가능하다. 비계층적 군집분석은 kmeans 군집분석의 경우 군집의 모양도 계층적이지 않지만 군집의 개수를 제일 먼저 선정하고 모형을 개발하는 방식으로 kmeans, kmedoid, 혼합분포군집, SOM 등의 방법이 있고, kmeans의 경우 R 프로그램의 kmeans 함수를 활용하여 군집분석을 진행할 수 있다.

예를 들어 kmeans(College, 3)라는 코드를 실행한다면 kmeans 함수를 활용하여 College 데이터의 변수들을 활용하여 3개의 군집으로 분류할 수 있다. 앞의 코드가 결과로 나타난다면 kmeans 분석 결과에서 n개의 데이터가 3개의 군집으로 각각 몇 개씩 군집되었는지 나타날 것이며 군지비 중심정보는 \$centers로 확인할 수 있다. between\_ss/total\_SS의 값은 1에 가까울수록 군집화가 잘되었으며 좋은 model임을 알 수 있다. 인사이트를 도출한다면 군집된 변수의 크고 낮은 값들을 확인하여 특징을 따서 인사이트를 도출할 수 있다.

2) 그룹의 개수를 결정하기 위한 기준을 제시하시오.

→ 그룹의 개수를 결정하기 위한 방법은 계층적 군집의 경우 분석한 결과를 토대로 덴드로그램을 그려 군집의 개수를 결정한다. 덴드로그램은 각 단계에서 관측치의 군집화를 통해 형성된 그룹과 이들의 유사성 수준을 표시하는 트리 다이어그램이다. 유사성 수준은 수직 축을 따라 측정되거나 사용자가 거리 수준을 표시할 수 있는데, 다른 관측치는 수평 축을 따라 나열된다. 덴드로그램 시각화 결과에서 y축에 나타나는 값을 기준으로 군집을 결정할 수 있다. y값에서 수평으로 선을 그어 나뉘는 그룹을 하나의 군집으로 구성할 수 있다. 비계층적 군집의 경우 군집 수에 따른 집단 내 제곱합 그래프를 통해 그룹의 개수를 결정하는 기준을 제시할 수 있다. 군집의 집단 내 제곱합 그래프는 얼마나 군집화가 잘되었는가를 알려주는 척도로 집단내 제곱합의 합을 최소화 하는 것을 목적으로 한다. 이 그래프는 Scree Plot과 비슷한 형태로 그려진다. 해석 방법은 급격히 감소하는 지점까지만 군집으로 설정하여 최적의 군집개수를 지정한다. 또, R 프로그램에서 최적의 군집수를 정하는 함수를 직접 작성하여 만들 수도 있지만, Nbclust 패키지 함수와 Scree plot를 활용하여 최적의 군집을 정하는 방법도 있다.

1) 최적회귀분석 방법에 대해 설명하고, 위의 분석에서 사용된 방법과 분석 모형의 수식을 사용하여 기술하시오.  
→ 최적회귀분석 방법은 분석 데이터에 가장 잘 맞는 모형을 찾아내는 방법으로서 R 프로그램에서는 step함수를 통해 종속변수에 대해 설명변수가 없을 경우부터 모든 설명 변수가 포함될 때의 회귀모형을 비교해 최적의 회귀방정식을 도출할 수 있다. 또, R 프로그램에서 step함수 안의 direction에서 'both'는 단계적 선택법(모든 가능한 독립변수들의 조합에 대한 회귀모형을 생성한 뒤 가장 적합한 회귀모형을 선택하는 방법), 'forward'는 전진선택법(절편만 있는 상수모형으로부터 시작해 중요하다는 생각되는 설명변수부터 차례로 모형에 추가하는 방법), 'backward'는 후진제거법(독립변수 후보 모두 포함한 모형에서 출발해 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 제거할 변수가 없을 때의 모형을 선택)을 의미한다. 위의 분석에서는 direction이 both로 입력되어 단계적 선택법을 사용했다. 위의 분석 결과를 아래와 같은 순서로 단계를 나누어 결과를 해석할 수 있다.

· 1단계 : 변수선택법을 결정하고, 초기 모형을 설정한다.

- 위의 분석 결과에서 direction이 both로 설정되어 변수선택법을 단계적 선택법으로 선정했음을 확인할 수 있다.  
또, 초기 모형은 Fertility~. 으로 설명변수가 모두 포함된 상태에서부터 시작함을 의미한다.

· 2단계 : 선택된 최적 모형의 AIC를 계산한다.

- 분석 결과에서 시작 모형은 Fertility~.이 최적 모형으로 설정되어 있으며 start에서 AIC 값이 190.69로 계산되어 있다.

· 3단계 : 선택된 모형에서 변수를 추가/삭제 할 경우의 각 모형의 AIC를 계산한다.

- Fertility~. 모형에 대해 설명변수 5개에 대한 각각의 AIC 값을 계산하여 자유도 등과 함께 나타낸다. Examination의 AIC 값이 189.86, Agriculture의 AIC 값은 195.10 등으로 나타나 있다.

그리고 모형은 Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality 으로 나타나 있다.

· 4단계 : 각 모형에서 최소의 AIC 모형을 선택하여 최적 모형으로 선정한다.

- 계산된 AIC값을 비교하여 190.69보다 작은 설명변수인 Examination을 제거하여 최적 모형으로 선정한다.

· 5단계 : 2~4단계를 반복하여 AIC가 더 이상 줄어들지 않을 때 최종모형을 최적의 모형으로 선정한다.

- 위의 과정을 반복하여 Fertility ~ Education + Catholic + Infant.Mortality + Agriculture이 최적의 모형으로 선정되고 마지막 Step에서 AIC가 189.86으로 계산되고 이 값보다 작은 값이 없어 변수를 모형에 추가, 삭제하지 않고 최적의 모형을 Fertility ~ Agriculture + Education + Catholic + Infant.Mortality으로 선정했다.

· 6단계 : 다변량 회귀분석에서 종속변수인 출산율(Fertility)에 대한 설명변수들 간의 모형에 대한 통계적 타당성을 가설 검정한다.

- 귀무가설( $H_0$ ) : Agriculture=Education=Catholic=Infant.Mortality=0

대립가설( $H_1$ ) : 적어도 하나의 설명변수는 0이 아니다.

F-통계량은 24.42이며 p-value 값이 1.717e-10로 귀무가설의 기각역인 0.05보다 작게 나타남에 따라 유의수준 5%하에서 대립가설을 채택하게 된다. 그러므로 추정된 회귀모형은 통계적으로 매우 유의함을 알 수 있다.

· 7단계 : 다변량 회귀분석에 활용된 각 설명변수들의 계수들에 대한 통계적 타당성을 가설 검정한다.

- 첫 번째 설명변수인 Agriculture에 대한 통계적 가설검정을 실시한다.

귀무가설( $H_0$ ) : Agriculture=0

대립가설( $H_1$ ) : Agriculture  $\neq$  0

t-통계량은 -2.267이며 p-value 값이 0.02857이므로 귀무가설의 기각역인 0.05보다 작게 나타남에 따라 유의수준 5%하에서 대립가설을 채택하게 된다. 그러므로 추정된 회귀모형의 첫 번째 설명변수인 Agriculture는 통계적으로 유의함을 알 수 있다.

- 두 번째 설명변수인 Education의 경우, t-통계량은 -6.617이며 p-value 값이 5.14e-08이므로 귀무가설의 기각역인 0.05보다 작게 나타남에 따라 유의수준 5%하에서 대립가설을 채택하게 된다. 그러므로 추정된 회귀모형의 두 번째 설명변수인 Education은 통계적으로 유의함을 알 수 있다.

- 세 번째 설명변수인 Catholic의 경우, t-통계량은 4.315이며 p-value 값이 9.50e-05이므로 유의수준 5% 하에서 대립가설을 채택하고 네 번째 설명변수인 Infant.Mortality의 경우, t-통계량은 2.824이며 p-value 값이 0.00722이므로 유의수준 5% 하에서 대립가설을 채택하게 된다. 그러므로 추정된 회귀모형의 모든 설명변수는 통계적으로 유의함을 알 수 있다.

· 8단계 : 통계적으로 유의성을 확인한 다변량 회귀모형이 전체 데이터를 얼마나 잘 설명하는지 확인하기 위해 결정계수( $R^2$ )를 확인한다.

- 결정계수를 확인하기 위해 Multiple R-squared와 Adjusted R-squared:를 확인한 결과, 0.6993과 0.6707 로 나타났다으며, 이는 전체 데이터를 설계된 다변량 회귀모형이 69.93%, 67.07%를 설명하고 있다고 해석할 수 있다.

2) 농업 종사자 비율 등 5개의 변화에 따른 출산율 변화를 추정한 결과를 사용해 구체적으로 설명하시오.

→ 최종적으로 다변량 회귀분석 결과를 종합해보면 추정된 다변량 회귀식은

Fertility = 6.21 - 0.16\*Agriculture - 0.98\*Education + 0.13\*Catholic +1.08\*Infant.Mortality이다.

회귀식을 통해 Education, Agriculture이 증가할수록 출산율(Fertility)는 감소하고

Catholic, Infant.Mortality가 증가할수록 출산율(Fertility)는 증가하는 것을 확인할 수 있었다. 그리고 출산율(Fertility)에 가장 영향을 많이 끼치는 변수는 Infant.Mortality이기 때문에 다른 변수들에 비해 많은 신경을 써야하며, 출산율을 증가시키기 위해서는 Catholic, Infant.Mortality을 높이고 Education, Agriculture를 줄여야 출산율이 증가할 것이라고 말할 수 있다.

## - 기출문제 서술형 답안 -

### <10회 기출문제>

1) PC1과 PC2의 의미를 원변수와 관계의 관계를 통해 유추하시오.

→ 화살표는 원변수와 주성분(PC)의 상관계수를 나타내며, PC와 평행할수록 해당 PC에 큰 영향을 끼친다. 또, 화살표가 같은 방향으로 인접해 있을수록 같은 주성분으로 생성될 수 있다. PC1은 “국물”, “면” 변수가 하나로 묶여 생성되었다고 판단할 수 있으며, PC1과 두 변수는 강한 양의 상관관계가 있을 것이라고 해석이 가능하다. PC2는 “그릇” 변수가 하나로 묶여 생성이 되었다고 판단할 수 있다. 그리고 3가지 변수 중 가장 영향을 많이 끼치는 변수는 PC1과 수평을 이루고 있는 “국물”이라고 판단할 수 있다.

2) 이상치가 있다면 어떤 특징을 가지는지 서술하시오.

→ 해당 그래프에서 이상치로 판단되는 라면은 “얼큰라면”, “해물라면”과 “된장라면”으로 판단할 수 있다. “얼큰라면”의 특징은 “면”변수에 영향을 많이 받고 있는 라면이라고 해석할 수 있다. 또, “해물라면”의 특징은 “그릇”변수에 영향을 많이 받고 있는 라면으로 알 수 있으며, “된장라면”은 3가지 변수에 모두 영향을 받지 않는 라면으로 해석할 수 있다.

### <11회 기출문제>

1) 아래 1은 계층적 군집분석을 한 결과를 덴드로그램으로 표현한 것이다. 2개의 군집으로 나눌 경우와 3개의 군집으로 나눌 경우 각 군집에 포함되는 도시들을 나열하시오.

→ 덴드로그램 시각화 결과에서 Height 값(y축)을 기준으로 하여 하위 군집을 구성하는 방법은 해당 Height 값에서 수평으로 선을 그어 나뉘는 그룹을 하나의 군집으로 구성한다.

2개의 군집으로 나눌 경우 heights를 60을 기준으로 하여 나누면 {베이징, 상하이, 모스크바, 두바이}와 {파리, 스톡홀름, 도쿄, 뉴욕, 하노이, 서울, 런던}을 나눌 수 있다. 3개의 군집으로 나눌 경우 heights를 50을 기준으로 하여 나누면 {베이징, 상하이, 모스크바, 두바이}, {파리, 스톡홀름}, {도쿄, 뉴욕, 하노이, 서울, 런던}을 나눌 수 있다.

2) 아래 2는 비계층적 군집분석인 kmeans의 결과이다. 조사 자료에 대한 군집의 수를 3, 4, 5개로 군집분석을 한 결과이다. 전체 변동에서 군집 간 변동이 차지하는 비율에 대한 검토를 통해 최적 군집의 수를 정하는 방법에 대해 구체적으로 설명하시오.

→ 전체 변동에서 군집 간 변동이 차지하는 비율을 이 1에 가까울수록 잘 분류되었고 좋은 모델임을 나타낸다. 군집 간 변동이 차지하는 비율은  $\text{betweenSS}(\text{군집과 군집 간 중심의 거리 제곱합}) / \text{totSS}(\text{제곱합의 총합})$ 으로 구할 수 있다. 해당 분석 결과에 대해 확인했을 때, 3, 4개의 군집으로 나누었을 때보다 5개의 군집으로 나누었을 때 80.6%로 가장 군집이 잘 되었다고 판단할 수 있다. 또, kmeans clustering으로 5개의 군집으로 나눈 결과에서도  $[\text{between\_SS} / \text{total\_SS} = 80.6\%]$ 로 나타나 전체 변동에서 군집 간 변동이 차지하는 비율을 확인 할 수 있다.

또, 최적 군집의 수를 정하는 방법으로는 군집 수에 따른 집단 내 제곱합 그래프를 통해 Scree plot 형태로 그래프를 그려 급격히 감소하는 지점까지만 군집으로 설정하여 최적의 군집 개수를 지정하거나 R 프로그램에서 Nbclust 패키지의 Nbclust 함수를 활용하여 최적의 군집을 정하는 방법도 있다.