클린업 3주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 클린업 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실시도 퇴출이니 유의해주세요.

Chapter 1 : Data Preprocessing

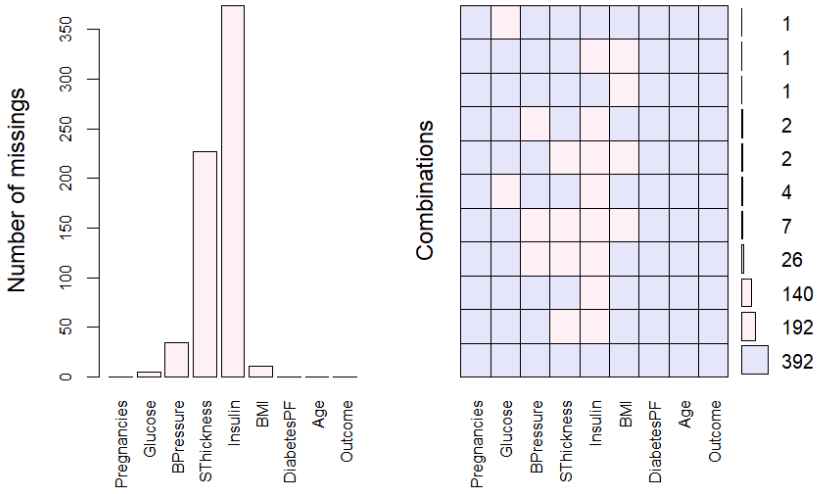
3주차 패키지의 분석 과제는 대표적인 비지도 학습 방법 중 하나인 클러스터링입니다. 클러스터링을 진행하기 전, 1~2주차에서 배운 전처리 테크닉을 이용하여 클러스터링에 적합한 데이터셋으로 만들어주는 간단한 전처리를 진행하겠습니다.

문제 1. 데이터를 불러온 후, 데이터의 구조를 파악하세요. 그 후, 데이터에서 각 변수들이 수치형 변수인지, 범주형 변수인지 판단해보고, 그 이유에 대해서 간단히 서술해 주세요.

문제 2. 데이터에서 결측치가 존재하는지 확인한 후, 이를 다음과 같이 시각화해주세요.

- 색깔은 'lavender','lavenderblush'로 지정해주세요.

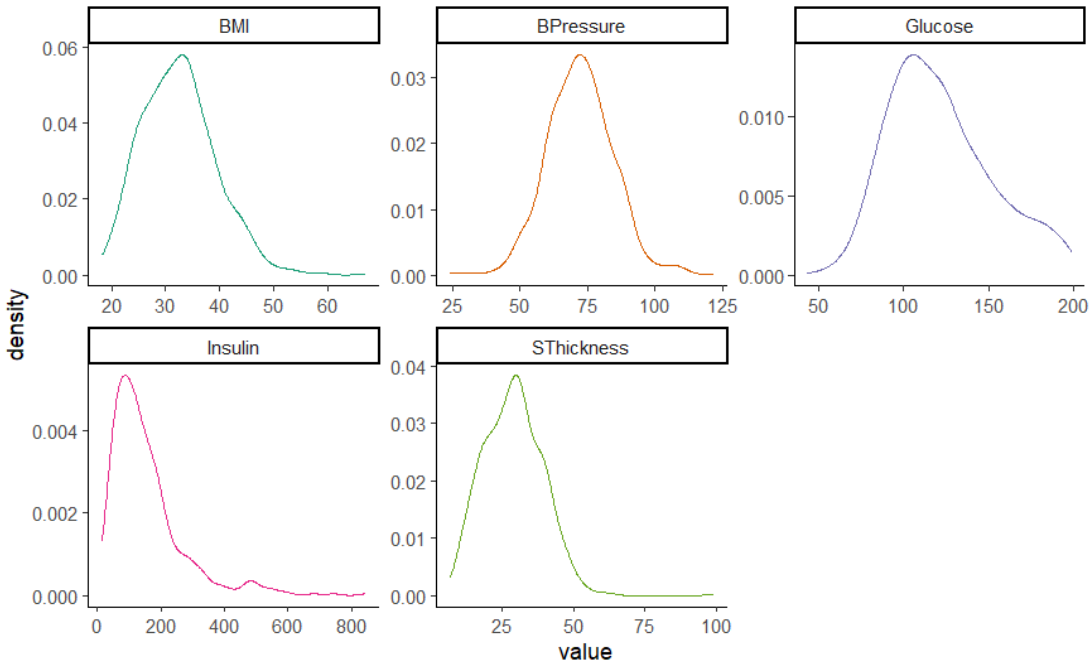
(HINT) VIM 패키지의 aggr()함수를 사용하면 결측치 개수와 발생 패턴을 시각화 가능합니다.



문제 3. 데이터에서 'Outcome' 변수는 추후에 사용할 예정이니 불러온 데이터 셋을 복사한 후에, 복사한 데이터프레임에서 'Outcome' 변수는 제거해주세요.

- 이후 진행되는 문제에서는 복사한 데이터 셋을 사용하여 진행해주시면 됩니다.

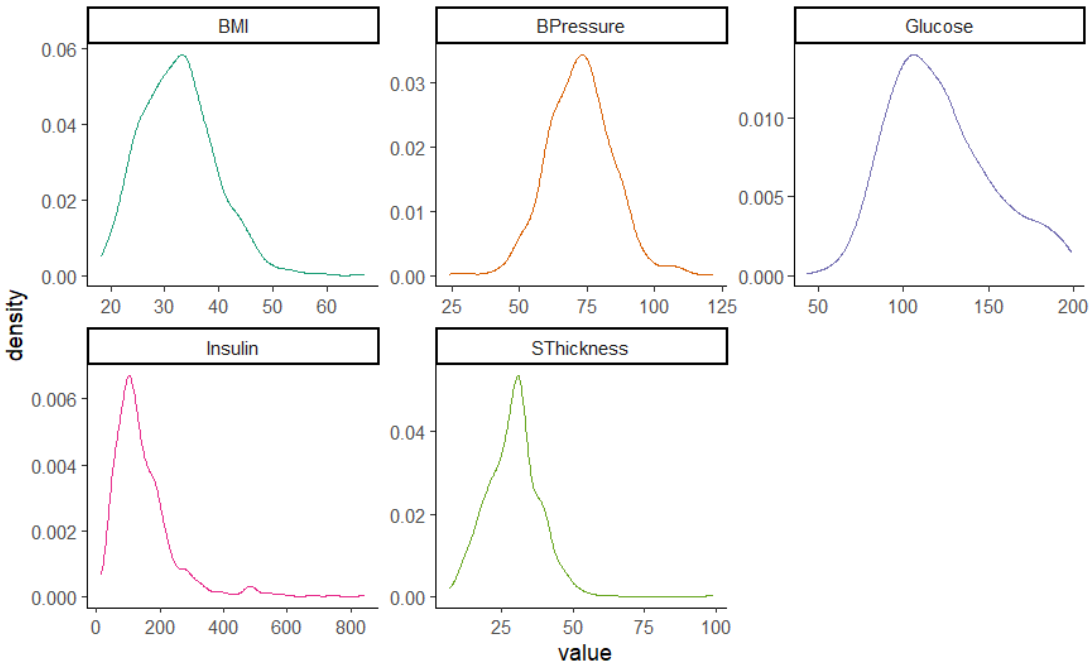
문제 4. 데이터에서 발견한 결측치를 보간하기 전, 결측치가 발생한 변수에 대해서 시각화를 통해 다음과 같이 분포를 확인해주세요.



- 각 변수 별 분포를 시각화 할 때, gather() 함수를 사용해주세요.

문제 5. KNN 알고리즘을 통하여 변수들에서 발생한 결측치를 보간할 예정입니다. 결측치를 KNN 알고리즘을 통하여 보간하고, 보간 후 결측치가 발생했던 변수에 대해 다시 한번 시각화를 진행하고, 시각화 결과를 통해 결측치 보간 전후로 각 변수들의 분포에 차이가 발생하는지 비교해보세요.

- KNN 알고리즘을 통해 데이터를 보간할 때, K=5로 설정해주세요.
- KNN 알고리즘을 사용할 때, imp_var = FALSE로 설정해주세요.

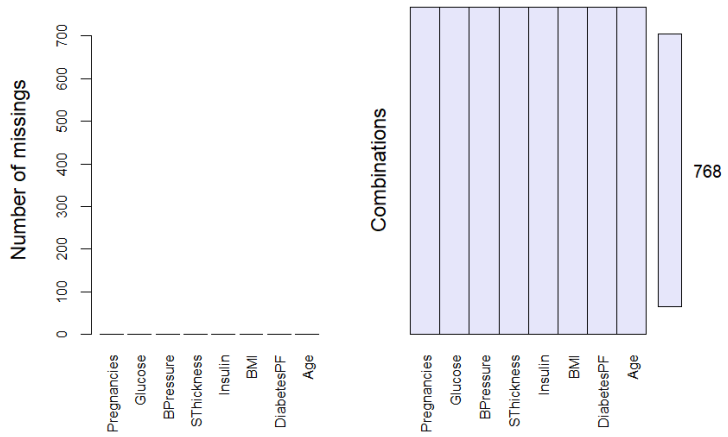


(HINT) VIM 패키지의 kNN()함수를 사용하면 KNN 알고리즘을 통해 결측치를 보간할 수 있습니다.

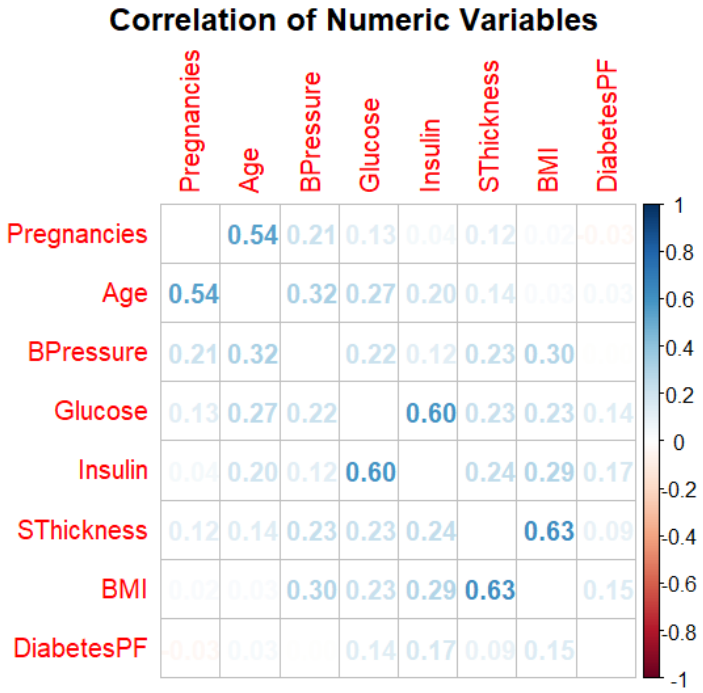
(+ 보너스 문제 1) 문제 4와 5에서 결측치 보간을 진행할 때, 결측치 보간 전후로 결측치가 발생한 변수에 대해서 분포를 다시 확인하였습니다. 이렇게 결측치 보간 전후로 결측치 보간을 진행한 변수에 대해서 분포를 시각화 하는 이유에 대해서 서술해주세요.

(+ 보너스 문제 2) 모델 기반의 결측치 보간법은 모델의 특성과 그 하이퍼파라미터에 따라 보간되는 값이 달라지게 됩니다. KNN 알고리즘의 하이퍼파라미터인 K를 조정하면서, K값에 따른 보간 성능을 확인하고, 이에 대해서 해석해보세요.

문제 6. 결측치 보간 이후에도 결측치가 존재하는지 확인한 후, 이를 다음과 같이 시각화해주세요.



문제 7. 데이터에서 수치형 변수들만을 사용하여, 수치형 변수들 간의 상관관계를 다음과 같은 상관관계 Plot 을 통해 확인해주세요. 그리고 그 결과에 대해서 간단히 해석해주세요.



Chapter 2 : Dimension Reduction & Clustering

두 번째 챕터는 클러스터링입니다. 클러스터링은 비지도학습의 일종으로써, 고객 세그먼트 분류 등 다양한 분야에 사용되는 방법론입니다. 각 데이터들 간의 거리를 계산하고, 계산한 거리를 기반으로 유저들을 군집화하는 방법입니다. 클러스터링은 이와 같이 대부분 거리 기반의 모델이기 때문에, 데이터의 차원이 높아질수록 '차원의 저주' 문제에 직면하게 됩니다.

오늘은 요인분석(Factor Analysis)를 활용한 변수선택을 통하여 클러스터링에 사용할 변수를 선택하고, 이를 바탕으로 대표적인 클러스터링 방법 중 하나인 K-means 클러스터링과 계층적(Hierarchical) 클러스터링을 진행해보겠습니다.

문제 1. 항상 동일한 결과를 얻기 위하여 Seed를 고정해주세요 (Seed: 3031)

문제 2. 데이터에서 변수들의 scale에 대한 영향을 제거하기 위해서 스케일링을 진행해주세요.

(HINT) scale()함수를 사용하면 데이터들을 스케일링 할 수 있습니다.

문제 3. 요인분석을 진행하기 전, 요인 분석에 대해서 간단히 찾아본 후, 이에 대해서 서술해주세요.

문제 4. factor=4, rotation='varimax'로 설정한 후, 요인분석을 진행해주세요.

(HINT) stats 패키지의 factanal()함수를 사용하면 요인분석을 진행할 수 있습니다

문제 5. 문제 4에서 진행한 요인분석을 바탕으로, 다음 문제들을 해결하세요.

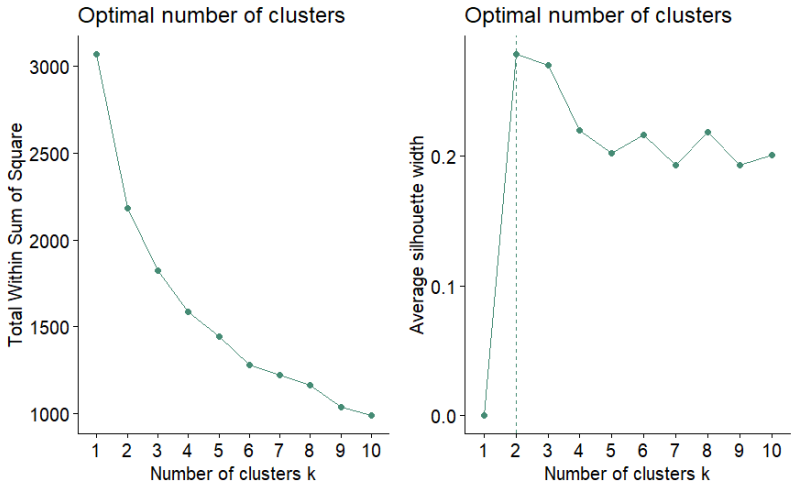
- Factor=4로 설정하여 진행한 요인 분석이 유의한지 확인해주세요
  (변수들을 4개의 factor로 표현하는 것이 유의한가?)
- Factor 1부터 Factor 4까지의 요인 적재량(Loading)을 확인하고,
  각 Factor들이 변수들에서 어떤 요인을 의미하는지 서술해주세요.
- 앞서 확인한 요인 적재량을 기준으로, 해당 Factor를 가장 잘 설명할 수 있는 변수들을
  각 Factor에서 하나씩 선택해주세요.

문제 6. 클러스터링을 위해 데이터에서 'BMI', 'Glucose', 'Age', 'BPressure' 변수만 선택하여 클러스터링 용 데이터 셋을 만들어주세요.

(+ 보너스 문제 3) 앞서 문제 5에서 선택한 변수와 클러스터링을 위해 선택한 변수가 다르다면, 왜 그 변수들 대신 다음과 같은 변수들을 사용하여 클러스터링을 진행하였는지 고민해보고, 그 이유에 대해서 서술해주세요.

문제 7. 클러스터링을 진행하기 전 최적의 클러스터 개수를 정하는 기준이 되는 Within Sum of Square(WSS)와 실루엣 계수(Silhouette Coefficient)가 무엇인지 알아본 후, 간단히 서술해주세요.

문제 8. K-means 클러스터링을 진행하기 전 fviz_nbcluster를 활용하여 Within Sum of Square플랏과 Silhouette Coefficient를 다음과 같이 시각화 한 후, 최적의 클러스터 개수를 선정해주세요.

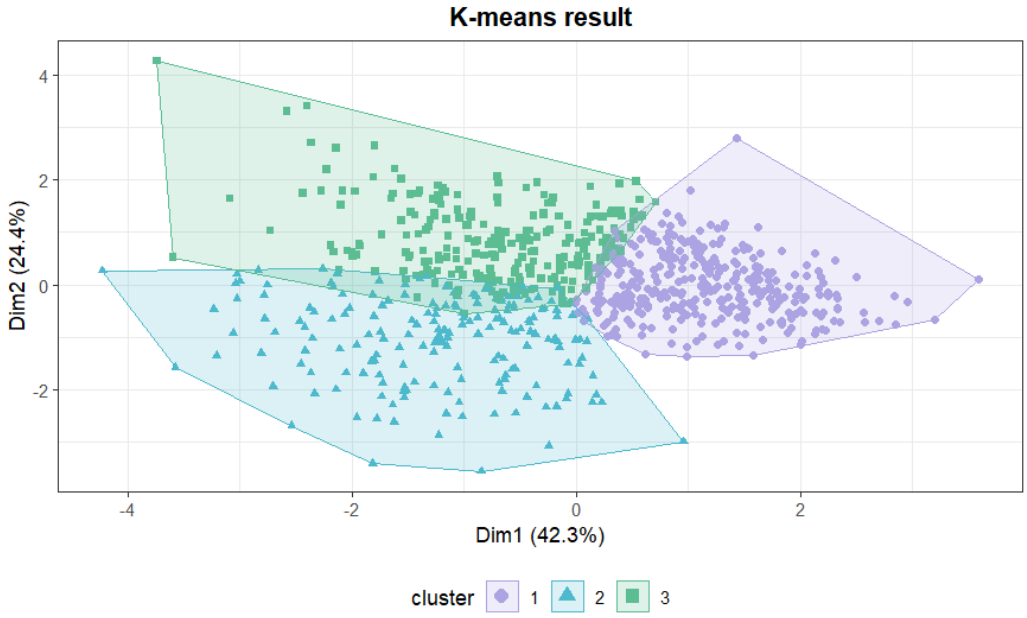


- 이때 선의 색깔은 'aquamarine4'로 설정해주세요.
- 그리고 FUNcluster=kmeans로 설정해주세요.

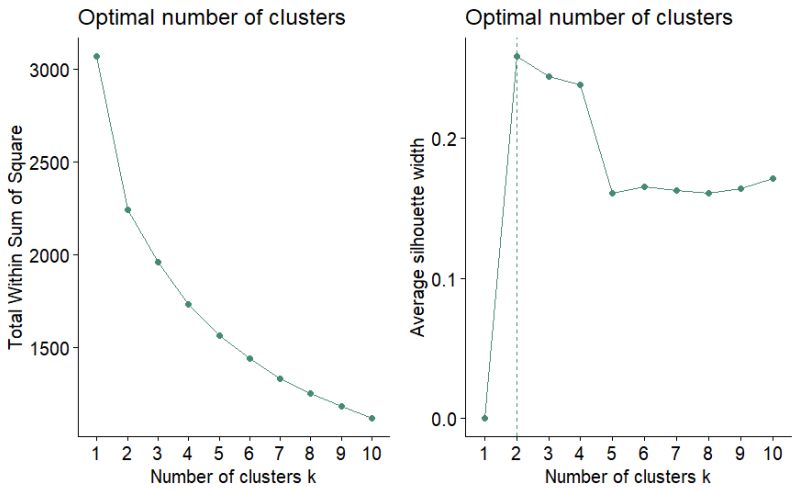
문제 9. k=3, iter.max=50, nstart=1 로 설정하여 클러스터링을 진행한 후, 다음과 같이 결과를 시각화해주세요.

- 색깔은 hcl.colors(3, palette = "cold")로 설정해주세요.

(HINT) stats 패키지의 kmeans() 함수를 사용하면 K-means 클러스터링을 진행할 수 있고, fviz_cluster를 통해 결과를 시각화할 수 있습니다.



문제 9. Hierarchical 클러스터링 또한 마찬가지로 fviz_nbcluster를 활용하여 Within Sum of Square플랏과 Silhouette Coefficient를 다음과 같이 시각화 한 후, 최적의 클러스터 개수를 선정해주세요.

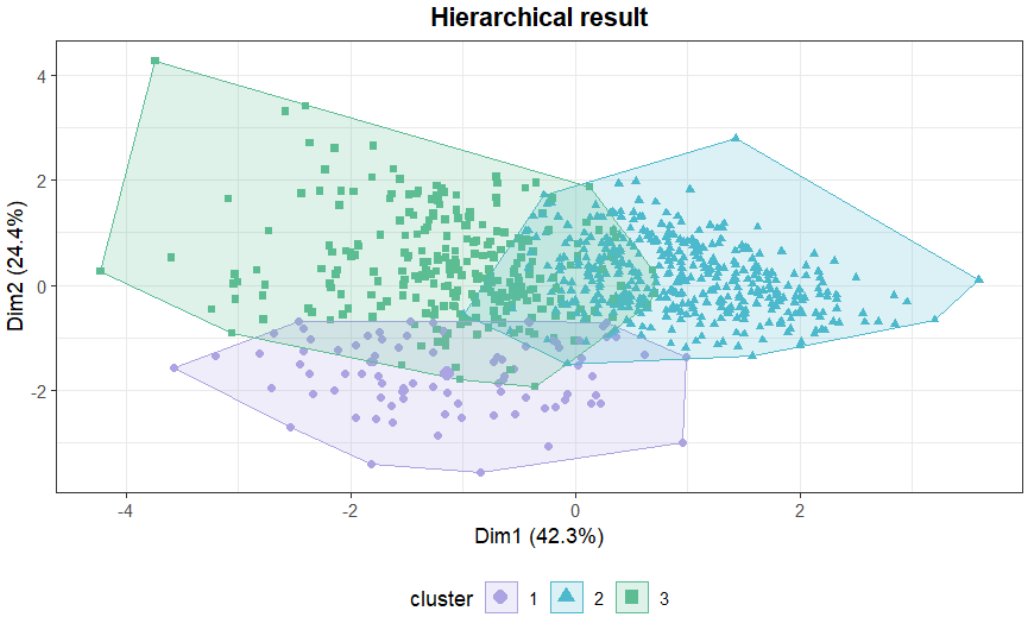


- 이때 선의 색깔은 'aquamarine4'로 설정해주세요.
- 그리고 FUNcluster=kmeans로 설정해주세요.

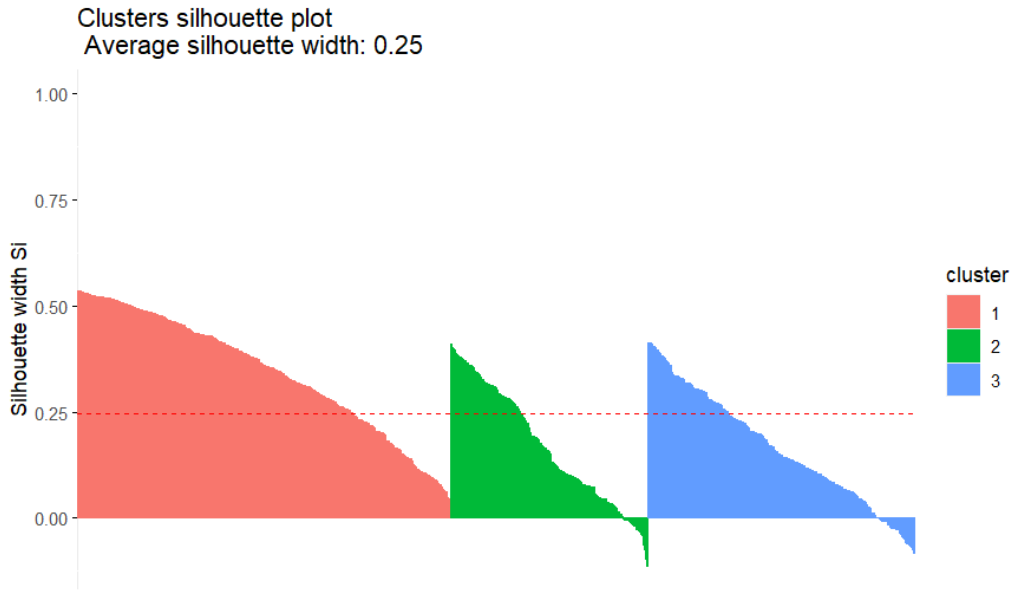
문제 10. k=3, hc_func='hclust'로 설정하여 클러스터링을 진행한 후, 다음과 같이 결과를 시각화해주세요.

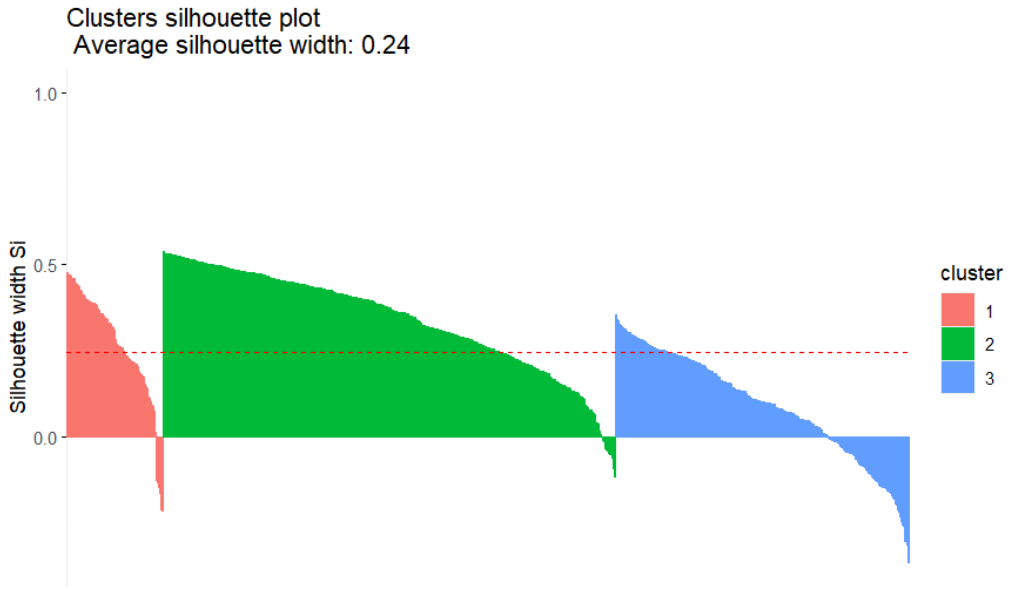
- 색깔은 hcl.colors(3, palette = "cold")로 설정해주세요.

(HINT) factoextra 패키지의 hcut() 함수를 사용하면 Hierarchical 클러스터링을 진행할 수 있고, fviz_cluster를 통해 결과를 시각화할 수 있습니다.

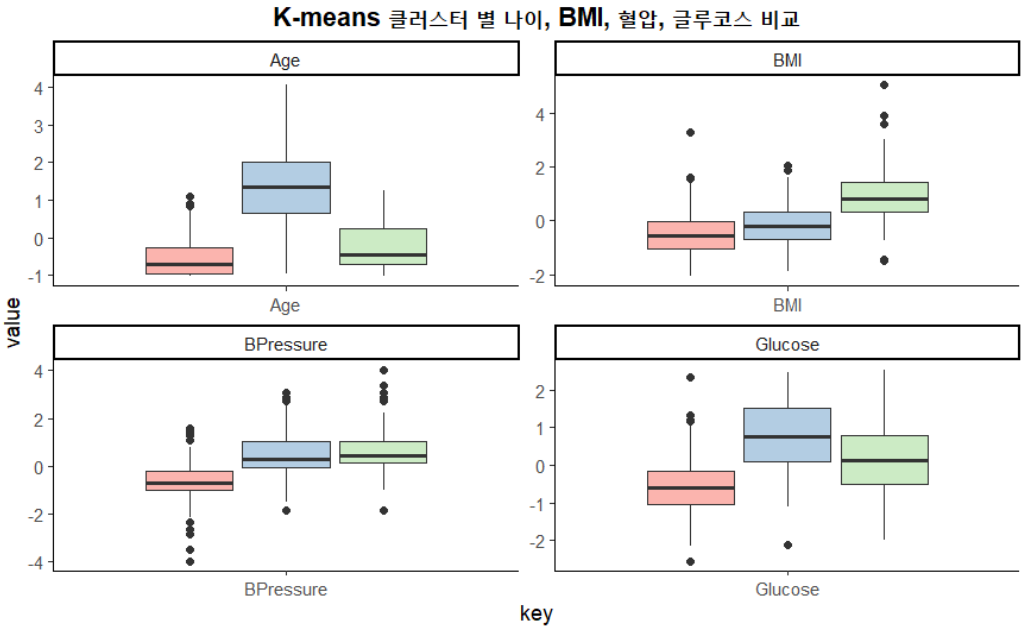


문제 11. K-means 클러스터링 결과와 Hierarchical 클러스터링의 실루엣 계수를 다음과 같이 시각화 한 후, 클러스터링 결과(각 데이터들끼리 잘 묶였는지)에 대해서 해석해주세요.





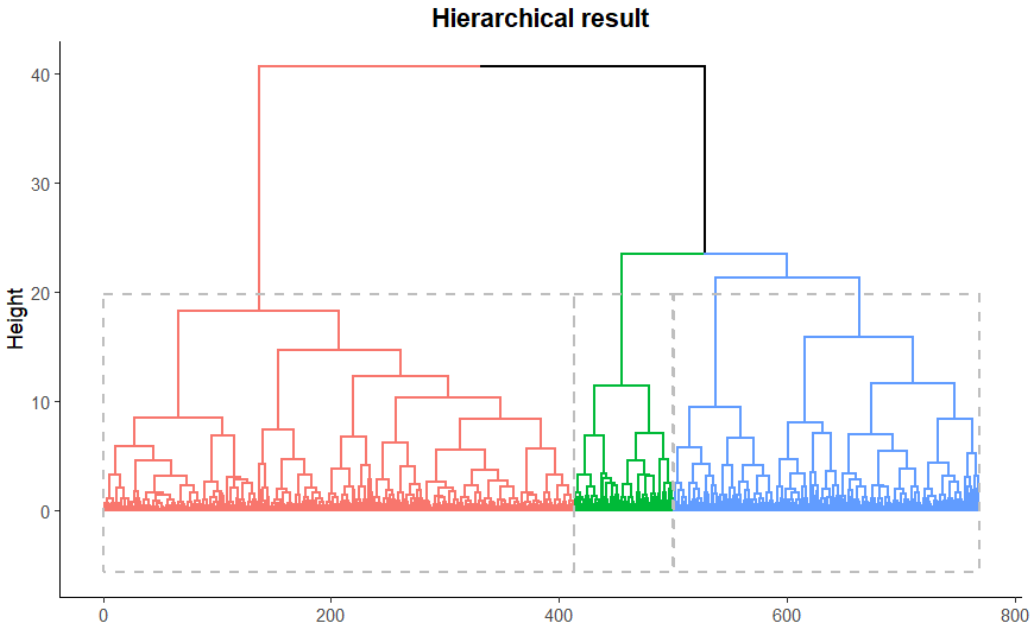
문제 12. K-means 클러스터링 결과에서 각 클러스터 별 중심점을 파악하고, 다음과 같이 각 클러스터 별 나이, BMI, 혈압, 글루코스에 대해서 비교하기 위해 박스 플랏을 그려보세요.

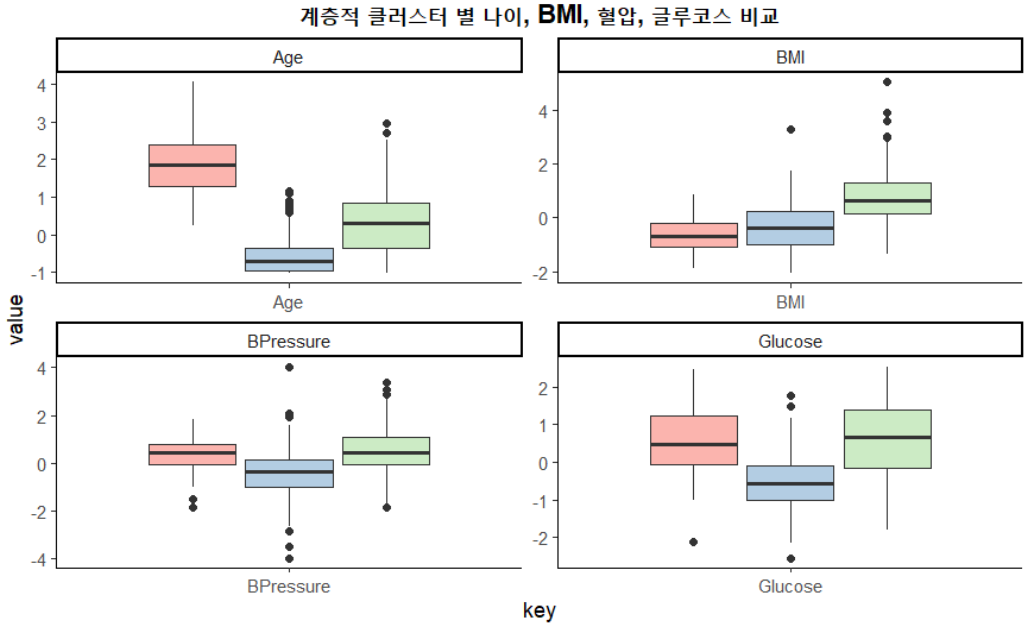


- palette는 'Pastel1'으로 지정해주세요

문제 13. Hierarchical 클러스터링 결과를 바탕으로 다음과 같이 덴드로그램을 그려보고, 각 클러스터 별 나이, BMI, 혈압, 글루코스에 대해서 비교하기 위해 박스 플랏을 그려보세요.

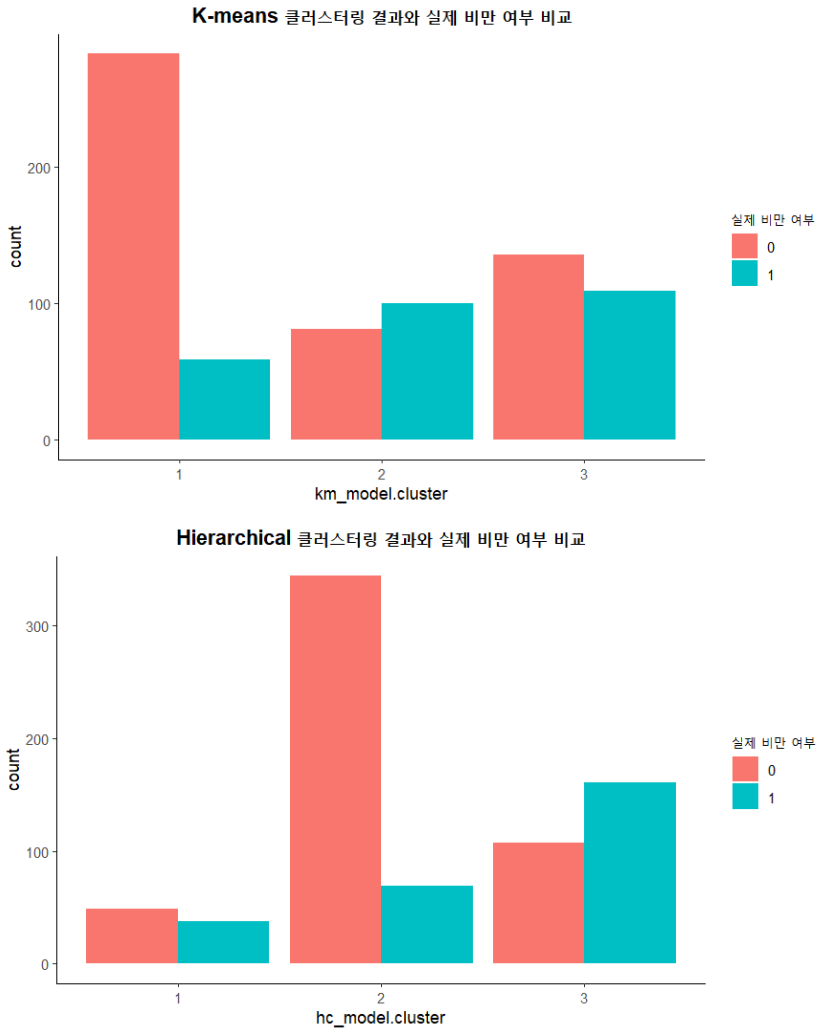
(HINT) fviz_dend함수를 사용하면 앞서 얻은 계층적 클러스터링의 덴드로그램을 그릴 수 있습니다. 이때 show_labels = FALSE, cex = 0.5, k = 3,color_labels_by_k = FALSE, rect = TRUE로 설정해주세요.





- palette는 'Pastel1'으로 지정해주세요

(+ 보너스 문제 4) 앞서 제거한 "Outcome(실제 비만 여부)"와 클러스터링 결과들을 비교하기 위해 다음과 같이 시각화를 진행한 후, 앞서 확인한 클러스터링 결과에 대한 해석과 연관지어서 생각해보세요. (시각화를 진행하지 않고 해석만 해도 괜찮습니다.)



(+ 보너스 문제 5) 고차원에서 진행된 클러스터링 결과를 시각화하기 위해 오늘 사용한 fviz_cluster()는 PCA를 통해 차원축소를 진행한 후, 클러스터링 결과를 2차원으로 표현합니다. 이처럼 클러스터링 결과의 시각화를 위해 사용되는 차원 축소 기법에는 TSNE가 존재합니다. TSNE가 무엇인지에 대해서 살펴보고, Perplexity를 50으로 설정하여 다음과 같이 클러스터링 결과에 대해서 시각화를 진행해보세요.