2021.02.04

[데이터 전처리-I]

1. 파일 입출력

(1) **Sys.getlocale**(); 시스템 인코딩 조회

(2) **write.table**(데이터명, file='파일경로/파일명'); 데이터를 파일에 저장

(3) **read.table**('파일경로/파일명'); 파일을 읽어 데이터 프레임 형태로 저장

2. apply계열 함수

(1) **apply**(대상자료, 1 or 2, 함수) : 1은 행별 함수 수행 / 2는 열별 함수 수행

(2) **lapply**(대상자료, 함수): 결과는 list로 출력

(3) **sapply**(대상자료, 함수): lapply와 유사하나, 결과는 행렬 혹은 벡터로 출력

(4) **vapply**(대상자료, 함수, FUN.VALUE = 데이터유형(결과길이))

: FUN의 모든 값이 특별 VALUE타입과 호환되는지 확인

ex – iris데이터를 list화한 irisList데이터의 평균

lapply (irisList, mean)

sapply (irisList, mean)

vapply (irisList, mean, **FUN.VALUE** = numeric(1))

(5) **mapply**(함수, 함수의 변수로 전달할 값, 함수에 전달할 다른 인자 목록)

: apply와 흡사하지만 mapply는 여러 인자를 함수에 전달 가능

3. 데이터 그룹화 함수

(1) **tapply**(대상자료, 범주형 변수, 함수)

ex – iris데이터에서 종별 꽃받침 길이의 표준편차 **tapply**(iris$Sepal.Length, iris$Species, sd)

(2) **by**( ) ; 한번에 여러 열을 집계할 수 없는 tapply의 단점 보완

(3) doBy 패키지

1) **summaryBy**(); 한번에 두 가지 FUN값을 적용할 수 없는 tapply의 단점 보완

ex – iris데이터에서 종별 꽃받침 길이와 넓이의 표준편차, 평균

summaryBy(Sepal.Length + Sepal.Width ~ Species, iris, **FUN = c(mean, sd)**)

2) **orderBy**(~정렬기준, data=데이터명); 정렬

ex – iris데이터를 꽃받침 길이 기준으로 오름차순 / 내림차순 정렬

orderBy(~Sepal.Length, **data=**iris) # 오름차순(default)

orderBy(~**-**Sepal.Length, **data=**iris) # 내림차순(정렬 기준 대상 앞에 – 붙임)

3) **samplyBy**(~추출기준, data=데이터명, frac=추출비율); 추출

ex – iris데이터에서 종별로 20%씩 표본 추출

sampleBy(~Species, **data=**iris, **frac=**0.2)