05. 데이터 전처리 # src: 7_R/05_데이터전처리 참고

1) 문자셋과 인코딩 #Windows는 CP949사용 (리눅스는 UTF-8)

(1) 문자셋(charset, Character Set); 하나의 언어권에서사용하는 언어를 표현하기 위한 모든 문자

(2) 인코딩(encoding); 문자셋을 컴퓨터가 이해할 수 있는 바이트와의 매핑 규칙 Sys.getlocale()

2) 파일 입출력

(1) write.table(data, file="파일명", append=FALSE, quote=TRUE, sep=",", row.names=TRUE);

(2) data <- read.table("파일명", header=TRUE, sep=",", stringsAsFactors=FALSE,
comment.char="#", fileEncoding="UTF-8", encoding="CP949")

(3) write.csv(data, file="파일명"); csv파일 형식으로 저장시 #sep="," 등 입력 필요X

(4) data <- read.csv(file="파일명") ; csv파일 조회시 #sep="," 등 입력 필요X

(5) cat(..., file="", sep="", fill=FALSE, labels=NULL, append=FALSE); 문자열 출력 외에도 저장 가능

3) apply 계열 함수; 반복문보다 코드 최소화, 수행 속도 빠름

(1) apply(자료객체(행렬,배열), 차원, FUN, ...); 차원: 1=행별, 2=열별, 3=3차원, FUN=적용함수명

(2) lapply(객체(벡터or리스트), FUN, ...); 리스트에 지정한 함수를 적용 후 리스트로 반환

(3) sapply(객체(리스트), FUN, ...); 결과를 벡터, 행렬로 반환

(4) vapply(객체(리스트), FUN, FUN.VALUE, ..., USE.NAMES=TRUE) #FUN.VALUE: 반환 데이터 유형
; sapply()와 유사하나 FUN의 모든 값이 FUN.VALUE와 호환되는지 확인하므로 안전

(6) mapply(FUN, ...,) ; sapply()와 유사하나 다수의 인자를 함수에 넘길 수 있음
#SIMPLYFY=TRUE: 연산결과 벡터, 행렬 등으로 반환, FALSE: 리스트로 반환

(7) tapply(객체(리스트), INDEX, FUN=FULL, ..., default=NA, simplify=TRUE); **그룹별 처리**
#INDEX: 범주형 변수(factor) 목록, default:결측치일 경우 출력될 값 지정(기본값: NA)

(8) by(data, INDICES, FUN, ..., simplify=TRUE); 데이터 프레임에 적용되는 tapply를 위한 함수
#INDICES: factor #by()함수는 데이터 프레임의 여러 열에 함수 적용可 (tapply()는 한 열씩)

(9) summaryBy(formula, data, id=NULL, FUN=mean, ...) #doBy패키지 install 및 로드 필요
; 그룹별로 그룹을 특징 짓는 통계적 요약 값 계산에 사용

(10) orderBy(formula, data); 데이터 프레임의 특정 변수로 데이터 프레임의 행을 정렬

(11) sampleBy(formula, frac=0.1, replace=FALSE, data, systematic=FALSE) #replace=F: 비복원추출
; 포뮬러 변수에 따라 분할된 각각의 그룹에서 특정 비율(frac)의 샘플 추출
#replace=F: 비복원 추출, T: 복원 추출 #systematic=F: 임의 추출, T:계통 추출(동일 간격)

※포뮬러: 응답변수~예측변수: fit <- lm(Y~X)

※split(), subset(), merge(), sort(), order(), with(), within(), attach(), detach(), table(), aggregate(), which()