

데이터분석 기초

숙명여자대학교 경영학부 오중산

Excel 파일 불러오기 & 데이터 프레임 내보내기

- Excel 형식의 파일 불러오기

- ◆ readxl 패키지에 있는 read_excel 함수 사용

- 기본 명령문: `df <- read_excel("파일명.xlsx", sheet = ??)`

- ❖ 주의사항: data sheet가 여러 개 있을 경우 지정해 주어야 함!

- 데이터 프레임 내보내기

- ◆ csv 형식으로 내보내기

- 내장함수를 이용한 기본 명령문: `write.csv(df, file = "파일명.csv")`

- ◆ Excel 형식으로 내보내기

- writexl 패키지에 있는 write_xlsx 함수 사용
 - 기본 명령문: `write_xlsx(df, path = "파일명.xlsx")`

비교 연산자

- “같다” 혹은 “같지 않다”

- ◆ “같다”는 ==를 사용하여 표기

- 주의사항: 등호(=)가 두 개!

- ◆ “같지 않다”는 !=를 사용하여 표기

- 주의사항: 등호(=)가 한 개!

- ◆ 문제: exam에서 주소가 원효로인 학생수와 성별이 여성이 아닌 학생수는 각각 얼마인가?

- 크기 비교 연산자

- ◆ 크거나(>) 작거나(<)

- ◆ 이상(>=) 혹은 이하(<=)

- ◆ 문제: exam에서 수학점수가 1)50점인 학생수 2)50점이 아닌 학생수 3)50점 이하인 학생수 4)50점 이상인 학생수는 각각 얼마인가?

논리 연산자

- “그리고”와 “또는”

- ◆ “그리고”는 &를 사용하여 표기
- ◆ “또는”은 |를 사용하여 표기
 - |는 Shift + \(\W) 키를 눌러야 함

- 관련 문제

- ◆ 영어점수가 50점 이하이고, 역사점수가 80점 이상인 학생은 몇 명인가?
- ◆ 수학점수가 90점 이상이거나, 역사점수가 90점 이상인 학생은 몇 명인가?
- ◆ 주소가 효창동이거나, 청파동이거나, 서계동인 학생은 몇 명인가?
 - 매치연산자(%in%)와 c()를 함께 사용하면 간단하게 표현할 수 있음

변수명 바꾸기

- mpg 데이터

- ◆ ggplot2 패키지에 내장된 데이터

- head(mpg)를 통해 데이터 탐색
 - mpg라는 이름의 데이터 프레임 만들기
 - mpg에 대한 탐색적 데이터 분석: str(), summary(), table() 사용

- 변수명 바꾸기

- ◆ dplyr 패키지에 있는 rename()를 이용하여 변수명 바꾸기

- 기본명령문: `df <- rename(df, var(new) = var(existing))`
 - ❖ 변수명을 바꾼 후에는 바꾼 변수명을 데이터 프레임에 저장해야 함

- ◆ mpg에서 fl(연료)를 fuel로, cty(도심연비)를 city로, hwy(고속도로연비)를 highway로 바꾸시오.

- ◆ 주의사항: 만약 mpg 원본을 다시 가져오려면 다음과 같은 명령문 입력

- `mpg <- ggplot2::mpg`

측정값 바꾸기

- 조건문(ifelse) 함수

- ◆ 기본명령문: ifelse(조건, 부합할 경우 값, 그렇지 않을 경우 값)
 - MS-Excel에서 if 함수와 동일함

- drv 변수의 측정값 바꾸기

- ◆ f는 forward, r은 rear로 바꾸기
 - 어떤 변수에 대해 ifelse 함수를 사용하여 측정값을 바꾼 경우, 그 결과를 해당 변수에 반드시 저장해야 함

- fuel 변수의 측정값 바꾸기

- ◆ c: CNG / e: ethanol / d: diesel / p: premium / r: regular

추가 예제1

- weather 데이터 프레임을 대상으로 다음 질문에 답하십시오.
 - ◆ 문제1: 평균기온이 27°C 이상인 날이 몇일인가?
 - ◆ 문제2: 평균기온이 10°C 이상이고, 20°C 이하인 날이 몇일인가?
 - ◆ 문제3: 일강수량이 0인 날은 얼마나 되는가?
 - ◆ 문제4: 일강수량 변수의 측정값이 NA(결측치)인 날은 얼마나 되는가?
 - ◆ 문제5: 월요일, 화요일, 수요일은 몇일인가?
 - ◆ 문제6: 최고기온이 30°C 보다 높고, 평균.상대습도는 80보다 높은 날은 몇일인가?
 - ◆ 문제7: 최저기온이 -10°C 보다 낮거나, 합계.일조시간이 1시간 미만인 날은 몇일인가?

추가 예제2

- weather의 복사본인 weather_new 데이터 프레임을 만들고, 이를 대상으로 다음 질문에 답하십시오.
 - ◆ 문제8: 요일.구분은 요일구분으로, 평균.현지기압은 평균기압으로 변수명을 바꾸시오.
 - ◆ 문제9: 새로운 변수 요일구분과 관련해서 출력 순서를 기존의 평일-휴일에서, 휴일-평일로 변경하십시오.
 - ◆ 문제10: 일강수량이 0으로 측정된 경우 이 값을 NA로 바꾸시오.
 - ◆ 문제11: 평균기압에 대한 결측치는 몇 개인가?
 - ◆ 문제12: 결측치를 제외한 평균기압 평균은 얼마인가(유효숫자 소수 둘째자리)?
 - ◆ 문제13: 평균기압이 NA인 경우, 이를 문제12에서 구한 평균값으로 대체하십시오.