데이터 탐색

숙명여자대학교 경영학부 오중산

데이터 불러오기

- Script 파일 만들어 저장하기
  - Script 파일명을 Untitled1에서 적절한 것으로 바꾸어 저장
- CSV 형태의 데이터파일 불러오기
  - csv(comma separated value) 파일의 특성
    - 읽고 쓰기 편리하며 저장용량이 적고 호환성이 높음
    - 가급적 raw data는 csv 형태로 저장하는 게 바람직하며, MS-Excel에서 불러올 수 있음
  - read.csv 함수를 이용해서 csv 형식의 데이터파일 불러오기
    - read.csv는 내장함수로서 별도 패키지 설치가 필요 없음
    - 기본 명령문: data frame <- read.csv("파일명.csv", stringsAsFactors = F)
      - 주의사항: csv파일은 프로젝트파일이 저장된 폴더(프로젝트명과 동일)에 있어야 하며, 불러올 때는 큰 따옴표 안에 파일명과 csv 확장자까지 표시해야 함!
      - stringAsFactors = F는 csv파일에서 문자로 측정된 변수의 척도를 범주형 척도(Factor)가 아니라, 문자형 척도(Character)로 하라는 의미

명령문 실행시 Ctrl키 누른 상태에서 Enter키 눌러야 함!

데이터 파악하기: 여섯 개 함수

- 여섯 개 함수를 활용하여 데이터의 기본적인 형태 파악하기

함수	기능	비고
head()	데이터 앞부분 6행을 보여줌	head(OOO, k): k행까지 조절 가능
tail()	데이터 뒷부분 6행을 보여줌	tail(OOO, k): k행까지 조절 가능
View()	Viewer 창에서 데이터를 보여줌	
dim()	행과 열을 활용한 데이트 크기를 보여줌	
str()	데이터 속성을 보여줌	속성 확인 후 변수 척도 수정
summary()	기술통계량을 요약하여 보여줌	min, max, median, mean, Q1, Q3

데이터 파악하기: 기술통계량과 히스토그램

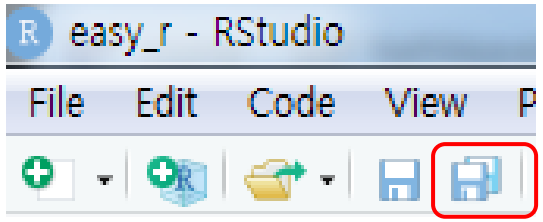
- 기술통계량(descriptive statistics) 구하기
  - 척도가 계량형(numeric이나 integer)인 변수에 대해 기술통계량
    - summary()를 통해서 여섯 가지 기술통계량을 구할 수 있음
      - 기술통계량은 개별 함수를 통해서도 구할 수 있음(예: mean(exam$english))
    - 추가로 분산과 표준편차를 아래 함수를 통해 구할 수 있음
      - var(data frame$variable) / sd(data frame$variable)
- Histogram 그리기
  - 대상 변수 척도가 반드시 계량형(numeric이나 integer)여야 함
  - 기본 명령문: hist(df$var, breaks = seq(N1, N2, by = ??))
    - hist는 내장함수
    - 대상은 변수여야 하며, N1(하한)과 N2(상한) 및 간격(??) 지정

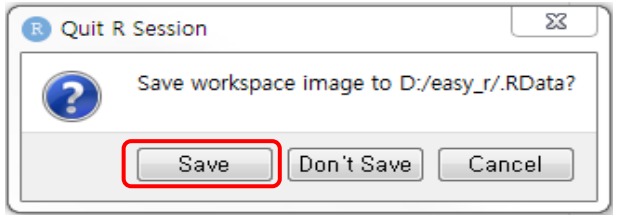
데이터 파악하기: 빈도수 파악하기

- 빈도수 파악하기1: table() 사용하기
  - 척도가 문자 혹은 범주인 변수에 대해 적용
    - 기본 명령문: table(df$var)
- 빈도수 파악하기2: qplot() 사용하기
  - 사전에 ggplot2 패키지를 설치하고 불러와야 함
    - install.packages("ggplot2")
    - library(ggplot2)
  - table()를 통해 확인한 결과를 막대 그래프 형태로 표현
    - 기본 명령문: qplot(data = df, var)
    - 두 개 변수를 동시에 고려한 명령문: qplot(data = df, var1, fill = var2)
      - “fill =” 색상을 기준으로 구분하는 것을 목적으로 함

주의 사항 및 알아두면 유용한 사항

- RStudio 활용시 주의 및 참고사항
  - RStudio 종료시 저장해야 함
    - 디스켓 두 개 겹친 것 클릭
    - 종료시 Save 클릭





      - ❖ 다시 실행시 환경창에서 변수나 데이터 프레임이 저장된 것 확인
  - Console 창 정리
    - 빗자루 모양 클릭하면 Console 창이 깨끗해짐
  - 주요 단축키
    - 명령문 실행: Ctrl + ENTER
    - 정의 표시 화살표(<-): Alt + -(마이너스)
    - 연산자(%>%): Ctrl + Shift + M
    - 문장 자동 완성: Tab
    - 화명 크기 조정: Ctrl + -(마이너스) 혹은 +(플러스 / Shift 함께 눌러야 함)
    - Alt + =를 누르면 글자가 이상한 모양이 되는데, 다시 똑같이 누르면 해결됨

실습하기: weather.csv 파일

- 다음 질문에 따라 실행하고 답하시오.
  - 문제1: weather.csv를 불러와서 weather라는 데이터 프레임을 만드시오.
    - weather.csv는 2020년 서울 종로구 송현동 기상관측소에서 측정한 다양한 일별 기상 데이터 파일
      - 강수량의 단위: mm / 기온의 단위: 섭씨(˚C) / 풍속의 단위: m/s / 습도의 단위: %
      - 기압의 단위: hpa / 일조시간의 단위: hrs / 일사량의 단위: mega J / m^2
  - 문제2: 측정 척도가 문자인 변수는 무엇인가?
  - 문제3: ‘일시’ 변수에 대해 측정 척도를 문자에서 날짜(date)로, ‘요일.구분’에 대해 범주(factor)로 바꾸시오.
    - 힌트: 요일로 변경 - df$var <- as.Date(df$var) / 범주로 변경 - df$var <- as.factor(df$var)
  - 문제4: 일시 변수를 이용해서 요일(월~일요일)을 파악한 후 이를 ‘요일’이라는 이름의 변수에 저정하시오. 그리고 ‘요일’변수의 척도를 문자에서 범주로 바꾸시오.
    - 힌트: 내장함수인 weekdays()를 활용함

실습하기: weather.csv 파일

- 다음 질문에 따라 실행하고 답하시오.
  - 문제5: 14개 변수에 대해 summary()를 통해 검토해 보시오.
    - 일강수량과 평균.현지기압에서 NA로 표기된 것은 무엇일까?
  - 문제6: '일강수량' 변수에 대해 분산을 구해 보시오.
    - 이 문제를 어떻게 해결해야 할까?
  - 문제7: 요일과 요일.구분에 대해 빈도수를 각각 구하시오.
    - 요일의 경우, 가나다 순서대로 된 정렬을 요일별로 바꿀 수는 없을까?
  - 문제8: 요일과 요일.구분을 동시에 고려한 qplot을 그려 보시오.
  - 문제9: 평균기온과 평균.상대습도에 대해 히스토그램을 그리시오.
    - 평균기온: -20~50°C 구간에 대해 1°C 간격
    - 평균.상대습도: 0~100% 구간에 대해 1% 간격

기술통계량 관련 추가 확인

- pastecs 패키지에 있는 stat.desc() 활용하기
  - 기본 명령문: stat.desc(df$var)
    - nbr.val(null, na): 사례개수, NULL값 개수, NA 개수
    - range = max –min
    - SE.mean = std.dev(s) / $\sqrt{n}$
    - coef.var = s / mean
- psych 패키지에 있는 describe() 활용하기
  - 기본 명령문: describe(df$var)
    - trimmed: 상하위 10%를 제외한 평균값
    - mad(mean absolute deviation): '측정값 – 평균값' 절대값의 평균
    - skew(왜도): 정규분포를 가정했을 때 좌/우로 기울어진 정도를 보여주는 통계량으로, +값이 크면 왼쪽으로, - 값이 작아지면 오른쪽으로 기울어지며, 0이면 좌우대칭
    - kurtosis(첨도): 정규분포를 가정했을 때 뾰족한 정도를 보여주는 통계량으로, +값이 크면 뾰족하고, -값이 작으면 평평하며, 0이면 적절한 형태임