데이터 탐색적 분석

숙명여자대학교 경영학부 오중산

함수와 패키지

함수(function)란?

- 입력(Input) – 함수(혹은 변환) – 출력(Output)
  - 함수는 입력된 데이터를 처리해서 새로운 결과로 바꾸는 과정 (도구)
- 그림의 함수(상자)는 무엇인가?

함수와 패키지

패키지(packages)란?

- 여러 함수가 들어 있는 꾸러미
  - 파이썬에서는 패키지 대신 라이브러리라고 함
  - 2021년 현재 CRAN에 등록된 패키지만 15,000개 이상
- 패키지 설치 및 불러오기
  - install.packages("패키지 이름")를 통해 설치
    - 오른쪽 아래 기타창에서 설치된 패키지 확인 가능함 (예: install.packages("dplyr"))
  - library(패키지 이름)를 통해 설치된 패키지 불러옴
    - 패키지 설치는 한 번만 하면 되지만 , RStudio 실행할 때마다 불러와야 함 (예: library(dplyr))
- 패키지 제거하기
  - remove.packages("패키지 이름")를 통해 제거

csv 파일 불러오기

csv 파일 소개

csv(comma separated value) 파일의 특징

- 엑셀파일에 비해 저장용량이 적어 대용량 데이터를 저장하는 raw data로 적합
- 호환성이 높아서 RStudio에서 불러올 때 오류를 줄일 수 있음

가급적 csv 파일로 저장하는 것이 바람직함

- 기존 raw data가 엑셀형식이라면 csv 형식으로 변경

csv 파일 불러오기

① 디폴트로 내장되어 있는 함수

csv 형식의 파일 불러오기

② 패키지

`Read R' 패키지

readr 패키지에 있는 read_csv 함수 사용

install.packages("read R")
저장
라이브러리로 불러옴
read_csv 함수 사용

- 내장함수인 read.csv보다 더 안정적임
- 기본명령문: df <- read_csv("file.csv", col_names = T, col_types = cols("i", "n", "f", "c", "D", "t"), na = "abc")
  - RStudio에서 화살표(assign) 단축키: Alt + -(마이너스)
  - RStudio에서 명령문 실행은 ctrl키를 누른 상태에서 Enter키를 누름!
  - file name은 따옴표 형식으로 확장자까지 함께 입력
  - col_names = T : 첫 번째 행은 변수로 인식하라는 의미 (F면 관측값)
  - col_types = cols("i(정수)", "n(실수)", "f(범주)", "c(문자)", "D(날짜)", "t(시각)")
  - na = "abc" : abc로 입력된 것은 결측치(missing value)로 처리하라는 의미

(결측치)

ex) 변수 20개 => 일일이 지정 힘듦

col_types = X

그냥 불러오고 검토해서 처리 해주기

데이터 프레임을 csv 형식으로 내보내기

※ write_csv건, read_csv건
프로젝트 명과 동일한 폴더 안에 있어야 함

데이터 프레임 내보내기

R스튜디오에서 작업한 데이터프레임 => CSV 파일로 내보내기

write_csv 함수를 이용하여 csv 형식으로 내보내기

- 기본명령문: write_csv(df, "file.csv", na = "NA", append = F, col_names = T)
  - => 어떤 이름으로?
  - file name은 따옴표 형식으로 확장자까지 함께 입력
  - na = NA : NA로 입력된 것은 결측치라는 의미
  - Append = F: 기존에 동일한 이름의 file(raw data)이 있다면 지금 만드는 파일로 대체하라는 의미
  - col_names = T : 변수를 포함해서 저장하라는 의미

col_names = F : 변수는 빼고 관측값 (순수 데이터)만 저장

실습용 파일 불러와서 데이터 프레임 만들기

exam 데이터 프레임 만들기

실습을 위한 exam.csv 파일 검토

- 다양한 척도로 측정된 여섯 개 변수로 구성됨
- 결측치는 na로 입력됨

변수명	address	gender	class	math	history	english
척도	문자형	범주형	범주형	정수형	정수형	정수형

exam.csv 파일 불러와서 exam 데이터 프레임 생성

정의, 만들기 등

- exam <- read_csv("exam.csv", col_names = T, col_types = cols("c", "f", "f", "i", "i", "i"), na = "na")

각 열의 첫번째는 변수명 / 결측치

exam 데이터 프레임에서 address 변수 측정값(한글)이 깨진 이유는?

- RStuido의 한글 인코딩은 UTF-8로 세팅됐는데, csv 파일에서는 EUC-KR로 됐기 때문
- 문제해결을 위해 조건 추가 : locale = locale('ko', encoding = 'euc-kr')

데이터 탐색: 여섯 개 함수

- 여섯 개 함수를 활용하여 데이터의 기본적인 형태 파악하기
  - exam 데이트 프레임에 대해 실습하기
    - View 함수: 오른쪽 환경창에서 exam 데이터프레임을 클릭
    - glimpse함수와 str함수: 데이터 크기, 변수명, 척도 등을 알려줌 : str보다 glimpse 많이 씀
    - summary 함수: 개별 변수에 대해 최소값, 1분위수, 중위수, 평균, 3분위수, 최대값 알려줌 : 문자는 별 정보 없음
      ↳ (밑에서 6줄) ↳ 기초통계량 보여줌 범주형은 빈도수 보여줌
    - head, tail, dim 함수는 잘 쓰이지 않음
      (위에서 6줄) ↳ 데이터 크기 Data → exam → 30 obs. of 6 variables
      (굳이 사용X)
      head가 보여주는 것 glimpse가 보여줌 (head 굳이 안씀)

<str 함수> → 오타, 결측치 존재

```
> str(exam)
spec_tbl_df [30 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ address: chr [1:30] "효창동" "청파동" "해방촌" "서계동" ...
 $ gender : Factor w/ 3 levels "Female","Male",..: 1 2 1 1 2 2 1 1 2 2 1 1
 ...
 $ class  : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 2 2 2 2 3 3
 ...
```

<summary 함수>

```
> summary(exam)
   address              gender   class      math          history
Length:30          Female:15   1:6   Min.   :20.00   Min.   :56.00
Class :character   Male  :14   2:8   1st Qu.:48.00   1st Qu.:78.00
Mode  :character   na    : 1   3:6   Median :58.00   Median :86.50
                                 4:5   Mean   :60.38   Mean   :84.13
                                 5:5   3rd Qu.:80.00   3rd Qu.:97.75
                                       Max.   :98.00   Max.   :98.00
                                       NA's   :1

    english
Min.   :12.00
1st Qu.:58.00
Median :65.00
Mean   :64.80
3rd Qu.:83.25
Max.   :98.00
```

범주형 → 빈도수 분포 / 결측치 NA's : 1 / 정량적 변수 (정수형 척도) → 최대값, 중간, 1,4분위, 결측치

데이터 탐색: 빈도 파악하기

변수의 빈도

① ↳ 내장 함수 (패키지 X)

- 빈도 파악하기1: table함수 사용하기
  - 기본 명령문: table(df$var)

```
서계동 용문동 원효로 청파동 해방촌 효창동 후암동
     5      3      5      3      5      6      2
```

    - 데이터프레임과 데이터프레임에 있는 변수는 $표시로 연결함 => glimpse, summary는 (데이터프레임 이름)
    - 사례가 많으면 문자형/범주형척도로 측정된 변수에 적용하는 게 바람직함

ex) table (exam $ math)

↳ 1000명 점수 numeric, integer 정량적 변수 의미 떨어짐

  - 문제: exam 데이터프레임에서 주소 빈도수 구하기
  - table 함수 대신 descr 패키지에 있는 freq 함수 사용하여 비교하기

② ③

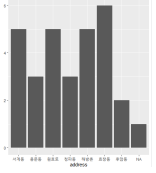
```
> freq(exam$address)
exam$address
        Frequency Percent Valid Percent
서계동          5  16.667         17.241
용문동          3  10.000         10.345
원효로          5  16.667         17.241
청파동          3  10.000         10.345
해방촌          5  16.667         17.241
효창동          6  20.000         20.690
후암동          2   6.667          6.897
NA's            1   3.333
Total          30 100.000        100.000
```

(그래프 아닌데 하지만 보여줌)

- 빈도 파악하기2: ggplot2 패키지에 있는 qplot 함수 사용하기

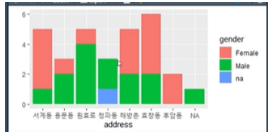
↳ 시각화

  - 기본 명령문: qplot(data = df, var) (data= exam, address)
    - 두 개 변수를 동시에 고려한 명령문 : qplot(data = df, var1, fill = var2), 그래프 전문적
      - “fill =” 색상을 기준으로 구분한다는 의미



  - 문제: exam 데이터프레임에서 주소(var1)별로 성별(var2) 빈도 파악하기

= 성별 (var2)에 따른 주소 (var1)



데이터 탐색: 기술통계량

기술통계량(descriptive statistics) 구하기

결측치 구하는 법

① summary(df $ 변수)

② table(is.na(exam $math))

: na가 있니? (함수)

- 척도가 계량형(numeric이나 integer)인 변수 대상
  - summary 함수를 통해서 여섯 가지 기술통계량을 구할 수 있음

①

- 내장함수를 통한 기술통계량 구하기
  - mean 함수(평균): mean(df$var)
  - var 함수(분산): var(df$var)
  - sd 함수(표준편차): sd(df$var)

- 문제: 내장함수로 수학점수 평균(혹은 분산이나 표준편차)을 구하면 어떻게 되는가?
  - na.rm = T 조건의 필요성

: na는 제거해줘
(함수 X, 파라미터)

데이터 탐색: 기술통계량

②

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 형태

- 문제: describe 함수를 사용하여 exam 데이터프레임의 기술 통계량을 구하시오.

describe (exam)

```
> describe(exam)                    (표준편)  (중위수)  (상위 10%, 하위 10% 뺀 평균)
   vars  n  mean(평균) sd median trimmed   mad min
X1    1 29 60.38 21.88     58   60.92 19.27  20
   max range  skew kurtosis   se (표준오차)
X1  98    78 -0.13       -1 4.06
```

1> 평균으로부터 나머지값 얼마나 떨어져 있는지

|29개 - 평균| 의 평균

데이터 탐색: 히스토그램

Histogram 그리기

qplot : 변수가 범주형, 문자형

- 대상 변수의 척도가 반드시 계량형 (numeric이나 integer)이어야 함
- 기본 명령문: hist(df$var, breaks = seq(N1, N2, by = ??))
  - hist는 내장함수
  - 대상은 변수여야 함
  - N1(하한)과 N2(상한) 및 간격 지정
- 문제: exam 데이터프레임에서 최소 0점에서 최대 100점까지 10점 간격으로 영어점수 히스토그램을 그리시오.

비교 연산자

- “같다” 혹은 “같지 않다”
  - “같다”는 ==로 표기하고, “같지 않다”는 !=로 표기
    - 주의사항: 등호(=)가 두 개!
  - 문제: exam에서 주소가 원효로인 학생과, 여성이 아닌 학생은 각각 몇 명인지 table함수와 freq함수를 이용해서 구하기
    - 조건에서 문자를 입력할 때에는 반드시 큰 따옴표 (“”)를 앞뒤로 붙여야 함!

```
freq(exam$address == "원효로")
table(exam$address == "원효로")
freq(exam$gender != "Female")
table(exam$gender != "Female")
```

- 크기 비교 연산자

위치주의

  - 크거나(>), 작거나(<), 이상(>=), 이하(<=)
    - 주의사항: 이상/이하에서 부등호가 앞에 위치함
  - 문제: exam에서 수학점수가 1)50점인 학생 2)50점이 아닌 학생 3)50점 이하인 학생 4)50점 미만인 학생은 각각 몇 명인가?

freq 함수

```
freq(exam$math == 50)
```

논리 연산자

- “그리고”와 “또는” and와 or
  - “그리고”는 &를 사용하여 표기하고, “또는”은 |를 사용하여 표기
    - |는 Shift + \(₩) 키를 눌러야 함
- 문제

freq(exam$english <= 50 & exam$history >= 80)

  - 영어점수가 50점 이하이고, 역사점수가 80점 이상인 학생은 몇 명인가? 6
  - 수학점수가 90점 이상이거나, 역사점수가 90점 이상인 학생은 몇 명인가? 13
  - 주소가 효창동이거나, 청파동이거나, 서계동인 학생은 몇 명인가?
    - 매치연산자(%in%)와 c()를 함께 사용하면 간단하게 표현할 수 있음 (컴바인 함수)

컴바인 함수 중에 하나라도 해당되면 빈도수 구해줘

freq(exam$address %in% c("효창동", "청파동", "서계동"))

↳ Na 못잡아내지만 True는 똑같음

데이터 탐색적 분석 실습

- imdb 데이터
  - Kaggle에 올라온 영화 1,000개에 대한 정보를 담고 있음
  - imdb.csv파일 불러와서 데이터프레임(movie) 만들기

① library(readr)
② movie <- read_csv("imbd.csv", col_names = T)

    - 변수가 많고 데이터가 크므로, col_names = T 조건만 삽입

↳ 각 열 첫번째 칸은 변수명이다.

- movie에 대한 탐색적 데이터 분석
  - View 함수를 통해 변수의 내용과 측정값의 의미 파악
  - glimpse 함수를 활용한 변수 척도 확인 및 척도 변경

: 문자 척도로 되어있는 변수지만 경우의 수가 많지 않을 때엔 범주형 척도로 바꿔주기

library(dplyr) → glimpse(movie)

character → factor (범주)

    - Certificate와 Genre의 척도를 as.factor 함수를 이용하여 factor 척도로 바꾸기 : movie $ certificate <- as.factor(movie $ certificate)
      (A, UA 등)
    - Released_Year의 척도를 as.integer 함수를 이용하여 integer 척도로 바꾸기
      (더블로 되어있음) ⇒ 정수는 double 실수이므로 굳이 안바꿔줘도 됨
    - Runtime의 척도를 as.integer 함수를 이용하여 integer 척도로 바꾸기

① library(stringr)
② movie $ Runtime <- str_replace_all(string = movie $ Runtime, pattern = "min", replacement = "")
(띄어쓰기 / 공백 x)
↳ min 문자나옴 " "

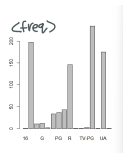
      - 미리 전처리
        - 1단계: stringr 패키지에 있는 str_replace_all 함수를 이용하여 측정값에서 공백과 문자를 제거해야 함
        (전처리를 패키지)
        ↳ min 이라는 텍스트 따문에 전처리
        - 2단계: as.integer 함수를 이용해서 척도 변경

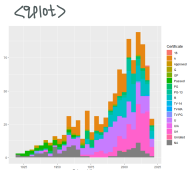
③ movie $ Runtime <- as.integer(movie $ Runtime)

ex) 120min : min이라는 텍스트

데이터 탐색적 분석 실습

- movie에 대한 탐색적 데이터 분석
  - summary 함수 이용한 검토 summary(movie)
    ↳ 평균, 일사분위, 최댓값, 최솟값 등등
  - freq 함수를 이용한 Certificate 빈도 확인하기 library(descr) → freq(movie$certificate)
    ↳ 기술통계
  - qplot 함수를 이용하여 Released_Year별 Certificate 빈도 구분하기 qplot(data=movie, Released_Year, fill=certificate)
    library(ggplot2) →
  - describe 함수를 이용하여 Meta_score와 IMDB_Rating의 기술통계량 비교 library(psych)
    describe(movie$Meta_score)
    ↳ integer ↳ numeric + 측정범위 다름
    - 변동계수(coefficient of variance)를 이용한 두 변수의 편차 비교 ⇒ 어떤게 더 많이 퍼져있나? ∴분산계수 구해야 함
      Sd/mean = 표준편차 / 평균 MS = 0.158119 IMDB = 0.03522 ⇒ MS가 더 많이 퍼져 있음
  - hist 함수를 이용한 Meat_score와 IMDB_Rating의 분포 비교 hist(movie$Meta_score, breaks=seq(0,100,1))
    - Meta_score는 0~100(1점 간격), IMDB_Rating은 0~10(0.1점 간격)
  - 비교연산자와 논리연산자를 활용하여 2019년에 출시되고 Meta_Score가 95점 초과하는 영화편수 확인
    (freq 함수 써도 됨) table(movie$Released_Year==2019 & movie$Meta_Score>95)
  - is.na 함수를 활용하여 매출(Gross)에서 NA 확인하기
    ↳ table 함수와 같이 썼었음 table(is.na(movie$Gross))





데이터 탐색적 분석 실습

변수명 바꾸기

dplyr 패키지에 있는 rename 함수를 이용하여 변수명 바꾸기

① library(dplyr)

② movie <- movie %>% rename(Title = Series_Title, Year = Released_Year)

- 기본명령문: df <- df %>% rename(var(new) = var(existing))
  - %>%는 파이프연산자(단축키: ctrl+shift+M)라고 하며, dplyr 함수와 함께 사용함
  - 주의! 변수명을 바꾼 후에는 새로운 변수를 데이터 프레임에 저장해야 함

rename 함수를 이용하여 Series_Title과 Released_Year를 각각 Title과 Year로 변경하기

참고) 변수 복사하기 및 삭제하기

- 복사하기: df1$var1 <- df2$var2 : df 2에 있는 변수2를 df 1에 var1이라는 이름으로 새롭게 복사
- 삭제하기: df1$var1 <- NULL

데이터 탐색적 분석 실습

(관측치) ex) Year 이면 1999. 2000

- 조건문(ifelse) 함수를 이용한 변수의 측정값 바꾸기
  - 기본명령문: df$var <- ifelse(df$var의 조건, 조건을 만족할 경우 값, 만족하지 않을 경우 값)
  - 예제: Runtime을 복사해서 Running이라는 변수를 새로 만들고, Running의 값이 200(분) 초과하면 Long으로 그렇지 않으면 Not Long으로 측정값 변경하기

    Movie $ Running <- Movie $ Runtime

    movie $ Running <- ifelse (movie $ Running > 200, "Long", "Not Long")

    - 주의! Long과 Not Long은 모두 문자임

      ↳ 관객의 평점 (변수)

    => Long 몇개인지 확인하고 싶으면

    table (movie $ Running == "Long")

- Gross의 NA를 Gross ÷ No_of_Votes의 평균으로 대체하기

  ↳변수 번호

  - pairs 함수를 이용한 두 변수 간의 산점도 그리기 pairs (movie [15:16] => 우상향 (한 쪽 증가 → 다른 한 쪽도 증가) => NA를 평균으로 대체 해야겠다!
  - 우선 mean 함수를 이용하여 Gross ÷ No_of_Votes의 평균 구하기 mean (movie $ Gross / movie $ No_of_Votes, na.rm = T)
  - 다음으로 ifelse 함수를 활용하여 NA를 평균으로 대체하기

    movie $ Gross <- ifelse (is.na(movie $ Gross), 217.1, movie $ Gross)

    => 제대로 됐나 확인 table (is.na(movie $ Gross))

주의사항 및 알아 두면 유용한 사항

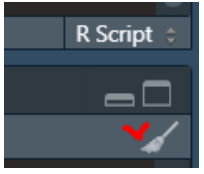
RStudio 활용시 주의 및 참고사항

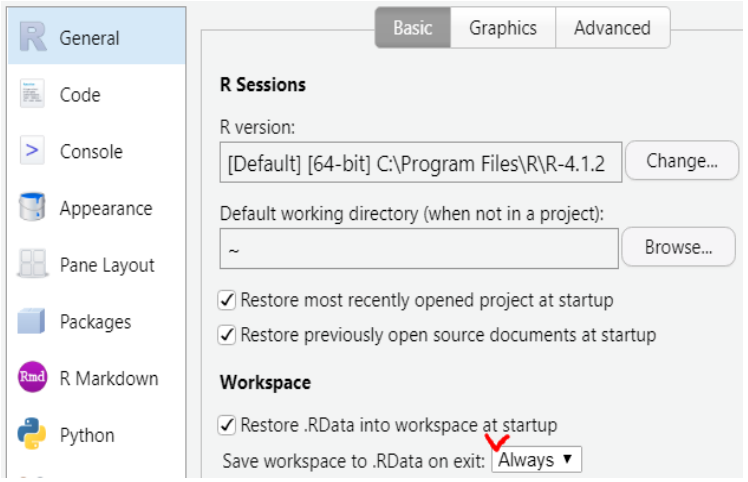
RStudio 종료시 자동저장

- Global Option에서 General의 Workspace에서 always 클릭!

Console 창 정리

- 빗자루 모양 클릭하면 Console 창이 깨끗해짐





- 주요 단축키
  - 명령문 실행: Ctrl + ENTER
  - assign 화살표(<-): Alt + -(마이너스)
  - 파이프연산자(%>%): Ctrl + Shift + M
  - 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 불러와 weather 데이터 프레임을 만드시오(단, col_types와 na 조건 입력하지 말 것!). ① library(readr) ② weather <- read_csv("weather.csv", col_names = T, locale = locale('ko', encoding = 'euc-kr'))
  - ❓ 주의! 변수가 한글로 되어 있을 때 한글 인코딩 조건 입력해 주어야 함
  - ❓ weather.csv는 2020년 서울 종로구 송현동 기상관측소에서 측정한 다양한 일별 기상 데이터 파일
    - ❓ 강수량 단위: mm / 기온 단위: 섭씨(°C) / 풍속 단위: m/s
    - ❓ 습도 단위: % / 기압 단위: hpa / 일조시간 단위: hrs / 일사량 단위: mega J / m^2
- ❓ 문제2: 변수 척도를 확인하시오. ① library(dplyr) ② glimpse(weather)
- ❓ 문제3: NA가 있는 변수는 무엇인가? summary(weather)

실습하기: weather.csv 파일

- 다음 질문에 따라 실행하고 답하시오.
  - ↳ 반드시 척도 Date여야함 ex) weekdays(as.Date("2022-10-29"))
  - weekdays 함수와 일시 변수를 이용해서 요일(월~일요일) 변수를 새로 만들어 weather에 저장하기
    - ↳ 이미 척도 Date
    - weather$요일 <- weekdays(weather$일시)
  - 일시의 척도를 범주형 척도로 변경하기
    - weather$일시 <- as.factor(weather$일시)
    - 참고: as.Date / as.character / as.numeric
  - 문제4: 요일의 척도를 범주형 척도로 변경하시오. weather$요일 <- as.factor(weather$요일)
  - 문제5: var 함수를 이용해서 '일강수량' 변수에 대해 분산을 구하시오.

var(weather$일강수량, na.rm=T)

실습하기: weather.csv 파일

* describe 함수 기술통계량 볼 때 *는 문자니까 주의

❓ 다음 질문에 따라 실행하고 답하시오.

① library(psych)
② descr_result <- describe(weather)
③ descr_result $ cv <- descr_result $ sd / descr_result $ mean

평균 / sd

❓ 문제6: 변동계수가 가장 작은 변수는 무엇인가?

❓ 문제7: 왜도와 첨도가 가장 작은 변수는 각각 무엇인가?

❓ 문제8: 요일은 각각 며칠인가?

① library(descr) ② freq(weather $ 요일) 또는 table(weather $ 요일)

↳ 빈도

❓ 가나다 순서대로 된 요일의 정렬을 실제 요일 순서로 변경하기

정렬 됐는지 확인하고 싶을 때 str(weather)

❓ weather$요일 <- factor(weather$요일, levels = c("월요일", "화요일", "수요일", "목요일", "금요일", "토요일", "일요일"))

❓ 문제9: 요일별로 요일구분이 어떻게 분포하는지 qplot으로 확인하시오.

① library(ggplot2)
② qplot(data = weather, 요일, fill = 요일구분)

❓ 문제10: 평균기온에 대해 히스토그램을 그리시오.

❓ 평균기온: -20~50°C 구간에 대해 1°C 간격

hist(weather $ 평균기온, breaks = seq(-20, -50, 1))

실습하기: weather.csv 파일

❓ 다음 질문에 따라 실행하고 답하시오.

① library(descr)

② freq(weather$평균기온 >= 10 & weather$평균기온 <= 20)

❓ 문제11: 평균기온이 10°C 이상이고, 20°C 이하인 날이 며칠인가?

❓ 문제12: 일강수량 변수의 측정값이 NA(결측치)인 날은 얼마나 되는가? freq(is.na(weather$일강수량))

❓ 문제13: 월요일 혹은 화요일은 며칠인가? freq(weather$요일 %in% c("월요일","화요일"))

↳ 문자

❓ 문제14: 최고기온이 30°C 보다 높고, 평균상대습도는 80보다 높은 날은 며칠인가? freq(weather$최고기온 > 30 & weather$평균상대습도 > 80)

❓ 문제15: 최저기온이 -10°C 보다낮거나, 합계일조시간이 1시간 미만인 날은 며칠인가? freq(weather$최고기온 <-10 | weather$합계일조시간 < 1)

↳ 계량형 => %in% 안씀

❓ 문제16: 평균현지기압은 평균기압으로 변수명을 바꾸시오.

① library(dplyr)

② weather <- weather %>% rename(평균기압 = 평균현지기압)

실습하기: weather.csv 파일

❓ 다음 질문에 따라 실행하고 답하시오.

① table(weather$요일구분)으로 확인

❓ 문제17: 요일구분 출력 순서를 평일-휴일에서, 휴일-평일로 변경하시오. ② weather$요일구분 <- factor(weather$요일구분, levels=c("휴일","평일"))

❓ 문제18: 일강수량이 0으로 측정된 경우 이 값을 NA로 바꾸시오.

① table(weather$일강수량==0) => 0 몇개인지 확인

② table(is.na(weather$일강수량)) => NA 몇개인지 확인

❓ 문제19: 평균기압 NA는 며칠인가?

freq(is.na(weather$평균기압))

③ weather$일강수량 <- ifelse(weather$일강수량==0, NA, weather$일강수량)

❓ 문제20: NA를 제외한 평균기압 평균은 얼마인가 (유효숫자 소수 둘째자리)? mean(weather$평균기압, na.rm=T)

❓ 문제21: 평균기압이 NA인 경우, 문제20에서 구한 평균값으로 대체하시오.

weather$평균기압 <- ifelse(is.na(weather$평균기압), 1006.26, weather$평균기압)