R tidyverse 패키지

Contents.

R tidyverse 패키지?

✓tidyverse에 대하여

- 여러 데이터 분석에 유용한 패키지들의 모음집
- 공식적인 tidyverse의 시작은 2016년

  ❖ 이전 부터 각각 구성원들이 따로따로 사용되고 있었음.
  ❖ ggplot2 패키지는 2005년에 만들어짐.

핵심 구성 요소

✓ install.packages("tidyverse")
✓ library(tidyverse

- ggplot2: 데이터 시각화하기
- dplyr: 데이터 wrangling
- readr: 데이터 불러오기
- tibble: modern data frames, 티블
- stringr: 문자열 다루기
- forcats: 팩터형(factors) 데이터 다루기
- tidyr: 데이터 타이디
- purrr: 반복작업하기(functional programming)

https://www.tidyverse.org/

데이터 친해지기

기초 탐색

✓ 기초 R 강의에서 배운 함수들을 사용해서 데이터 구조를 살펴보자.

- dim( )
- head( )
- tail( )

tidyverse

✓ glimpse( )

데이터 변수 이름 바꾸기

변수 이름 설정 - 한글 코딩은 NO

```
janitor::clean_names
names()
```

모두 소문자
띄어쓰기 있을 경우 "_(언더바)"로 연결

데이터 불러오기

✓ install.packages("tidyverse")
✓ library(tidyverse)

- .csv 파일의 경우 read_csv() 함수를 사용
- .xlsx 파일의 경우 read_excel() 함수를 사용

R dplyr 기본 동사 학습하기

✓Single table 동사들

행(row) 관련 동사들

- filter( ) : 조건에 맞는 행 필터링하기
- arrange( ) : 행 정렬하기
- slice( ) : 위치 지정으로 원하는 행 필터링하기, 일부만 보기

열(column) 관련 동사들

- select( ) : 원하는 열(변수) 선택하기
- rename( ) : 변수 이름 설정.
- mutate( ) : 새로운 변수(열) 생성
- relocate( ) : 변수 이동

✓ ggplot( )

1. 하나의 ggplot( ) 함수에 여러 geom_xx( ) 함수들이 + 로 연결되어 그래프 완성

   + 기호는 반드시 명령문의 맨 **마지막**에 와야 함

2. **ggplot(사용할 데이터셋(data=xx), x, y축으로 사용할 열 이름(aes(x=x1,y=x2)))을 지정**

```
ggplot(data=xx, napping = aes(x=x1,y=x2)) +
    geom_xx( mapping = aes(매핑모음),
             stat = <스탯>,
             position = <위치>) +
    <좌표계 함수> +
    <면분할 함수>
```

3. 어떤 형태의 그래프를 그릴지 geom_xx( ) 지음 함수를 통해 지정

   ex) geom_bar( ) : 막대도표

   geom_histogram( ) : 히스토그램

   geom_boxlplt( ) : 상자 도표 등.

- **aes(x = 변수, y = 변수) :** 그래프를 그리기 위한 x축, y축 열지정
- **stat = 'identity ' :** 막대 높이는 y축에 해당하는 열의 원데이터 그대로(데이터셋내에 원데이터 포함됨)
- **width = 1 :** 막대의 폭

```
예) ggplot(r, aes( x = study1)) +          # 그래프를 작성할 데이터 지정
      geom_bar(aes(fill = gender1),        # 지옴함수로 막대 그래프 지정, 막대의 색깔 성별 변수로 구분해서 채우기 지정
        width =  0.7,                      # 막대의 폭 지정
        color = "black"                    # 막대라인의 색깔 지정
    ) +
    ggtitle("누적 막대 차트") +            # 막대 타이틀 지정
    coord_flip()                           # 수평막대
```

mpg 데이터프레임

- 예제 데이터 : mpg data
- mpg(Mile Per Gallon) 데이터는 미국 환경 보호국(US Environmental Protection Agency)에서 공개한 자료로, 1999~2008년 사이 미국에서 출시된 자동차 234종의 연비 관련 정보를 담고 있음.

변수	설명
manufacturer	자동차
model	제조사
displ	배기량
year	제조년도
cyl	실린더 수
trans	변속기
drv	구동 방식 (f = 전륜구동, r = 후륜구동, 4 = 사륜구동)
cty	도시 연비
hwy	고속도로 연비
fl	연료 종류
class	자동차 종류

diamonds 데이터프레임

- 예제 데이터 : diamonds data
- "다이아몬드(Diamonds)" 데이터는 다이아몬드의 크기와 가격, 품질 등에 대한 정보를 담고 있음.
- 53,940 레코드와 아래의 표와 같이 10개 변수로 구성됨.

변수	설명
price	가격($)($326–$18,823)
carat	무게(0.2–5.01)
cut	세공의질 (Fair, Good, Very Good, Premium, Ideal
color	컬러 (7개의 컬러:D (best) to J (worst))
clarity	투명도(범주형 타입) (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	길이
y	넓이
z	깊이
depth	깊이 z / mean(x, y) = 2 * z / (x + y) (43–79)
table	넓은 폭 대비 꼭대기의 넓이

nycflights13 데이터프레임

- 예제 데이터 : flights data
- 이 데이터프레임에는 뉴욕시에서 2013년에 출발한 336,776 편의 모든 항공편이 포함되어 있다. 데이터의 출처는 Bureau of ransportation Statistics이며,
- 2013년 미국 New York에서 출발하는 항공기의 이착륙 기록이 수록된 자료로 336,776 레코드와 아래의 표와 같이 19개 변수로 구성됨.

변수	설명
year,month,day	출발 및 도착 날짜
dep_time/arr_time	실제 출발/도착 시간 (로컬타임 적용)
sched_dep_time /sched_arr_time	예정 출발/도착 시간 (로컬타임 적용)
dep_delay/arr_delay	출발 및 도착 지연시간 (단위 분)
hour,minute	Time of scheduled departure broken into hour and minutes.
carrier	항공사 약자(영문 2글자)
tailnum	항공기등록번호 (Plane tail number)
flight	항공편명 (Flight number)
origin,dest	출발공항/도착공항 (Origin and destination)
air_time	이륙후 착륙전 까지의 시간
distance	운항거리
time_hour	예정 출발/도착시간 (POSIXct date)

Lahman 데이터프레임

- 예제 데이터 : Batting data
- 타격 데이터셋은 22개 변수에 대한 112,184개의 관측값이 포함된 야구 데이터셋입니다.

변수	설명
playerID/teamID	플레이어 ID 코드/팀ID
yearID /lgID	타격연도ID(연도) /리그ID(수준이 AA AL FL NL PL UA)
stint	정량,선수의 기간(시즌 내 출전 순서)
G	게임: 플레이어가 플레이한 게임 수
AB	유효타석수
R	실행(Runs)
H	안타(Hits): 수비진의 실수 없이 타구를 쳐서 베이스에 도달한 횟수
X2B	복식(Doubles) : 타자가 2루에 안전하게 도달한 안타
X3B	트리플 : 타자가 3루에 안전하게 도달한 안타
HR/RBI	홈런/타점득점
SB/CS	도난당한 베이스
ab	볼(안타)을 칠 기회횟수