2. { Data 사용하기 }

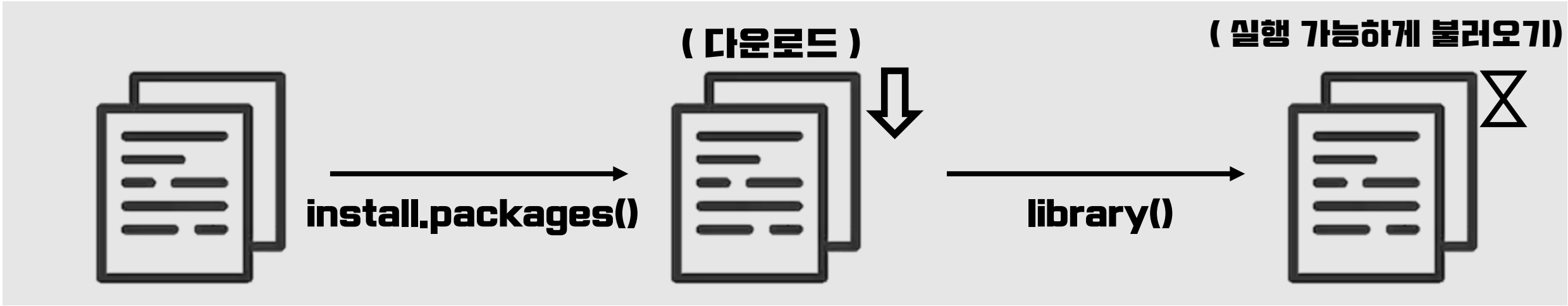
사용 가능 data

- 내부 data : R에 내장되어 있는 data ,
  컴퓨터에 저장되어 있는 data
- 외부 저장 data : Web상에 저장되어 있는 data,
  Data Base에 저장되어 있는
  data, etc.
- Web data : Web 상에 존재하는 데이터,
  ex) 크롤링, Open API 등 으로 사용

내부 data 사용하기

우선 작업

```
install.packages("tidyverse")
library(tidyverse)  # 여러 기능이 들어있는 패키지 엄청 유용함
library(readxl) # 엑셀 불러오는 패키지 (tidyverse 안에 포함되어 있음)
```



내부 data 사용하기

```
## 내부 데이터 불러오기
mpg <- data("mpg") # R 내부 데이터 불러오는 함수

## 데이터 요약해서 보기
head(mpg, 10) # mpg data의 앞에서 10줄 보여줌
tail(mpg, 10) # mpg data의 뒤에서 10줄 보여줌
str(mpg) # 데이터의 구조를 보여줌
```

기본구조

```
head(data, n)
tail(data, n)     # default = 6
```

내부 data 사용하기

기초통계 확인하기

```
mean(mpg$cty) #평균
max(mpg$cty) # 최대값
min(mpg$cty) # 최솟값
var(mpg$cty) # 분산
sd(mpg$cty) # 표준편차
```

기본구조 함수(data$변수)

내부 data 사용하기

기초통계 한눈에 보기

```
summary(mpg) # 기초통계 보여줌
summary(mpg$cty) # 변수 하나의 요약치 보기도 가능
```

기본구조 `summary(data)`
`summary(data$변수)`

내부 data 사용하기

새로운 변수 추가하기

```
mpg$mean_fuel <- (mpg$cty + mpg$hwy) / 2
mpg$mean_fuel # 잘들어갔는지 확인
```

기본구조 data$새로 만들 변수 <- 값 혹은 연산

내부 data 사용하기

Q1) mpg 데이터의 hwy 기초통계를 구하시요.

Q2) mpg 데이터의 hwy, cty의 차이를 diff에 저장하시요.

내부 data 사용하기

```
# 데이터 불러들이기
getwd() #작업공간 확인하기
setwd("C:/Users/ (본인 컴퓨터 사용자 이름)/문서")
# 작업공간 설정. 원하는 폴더 경로로 설정 가능

titanic <- read.csv("titanic.csv",
                    stringsAsFactors = F)
# 데이터 reading
# stringsAsFactors = F 문자가 포함된 변수를
  factor로  읽어드릴지 여부 (F / T)
```

내부 data 사용하기

Q3) titanic data에서 Survived 변수만 출력하시요.

Q4) titanic data의 변수들의 특성을 확인해보시요.

Q5) titanic data의 기초통계를 출력하시요.(summary 활용)

> NA ? NA's ??

+ 결측치 (Missing value)

- NULL : 값이 비어 있음
- NA(Not Avaliable Values) : 해당 없음, 이용 할 수 없음
- NaN(Not a number) : 숫자가 아님, 숫자의 범위를 초과
  ex) 0 / 0

확인방법

is.na(), is.null(), is.nan() #참일 경우 TRUE
!is.na(), !is.null(), !is.nan() #참일 경우 TRUE

summary

데이터를 읽어온 후 특성을 파악해야 원하는 값들로 **'전처리'** 작업하기에 수월하다.

데이터의 요약치를 확인함으로 **'결측치'** 유무를 확인할 수 있다.
(이상치 처리 방식은 다양함)