3. { Data 전처리 }

(R Study group 자료)

(1) 결측값(missing value)

결측값 처리하는 방법은 너무나도 많음., (정해진 답은 없음 @@)
우선 결측값이 얼마나 존재하는지 확인 작업 필요 !!

```
titanic <- readxl::read_csv("titanic.csv") # data 불러옴

str(titanic)
summary(titanic)
```

참고사항

```
read.csv("data.csv", stringsAsFactors = F )
read_csv("data.csv") # stringsAsFactors 할 필요 x
```

1) 결측값(missing value) 확인

결측값이 얼마나 존재하는지 확인 작업 필요

```
is.na(titanic$Survived) # is.null( ) ,  is.nan( )
> @@ ???

table(is.na(titanic)) # NA = T
table(is.na(titanic$Survived)) # 변수 하나만 확인도 가능
table(!is.na(titanic$Survived)) # NA = F, ( ! : Not의 의미 )
```

기본구조

```
is.na(data$변수) # 변수안의 값이 결측값인지?
table(data$변수) # 변수값의 빈도 출력
```

2) 결측값(missing value) 처리

결측값이 존재하면 연산이 안됨 (ex) mean, sum, etc.)

why? 알수없음 + 5 = 알수없음

```
sum(titanic$Fare)
> NA

sum(titanic$Fare, na.rm = T) # na.rm = T
                             # NA 값을 제거하고 계산해달라
```

기본구조

```
is.na(data$변수) # 변수안의 결측값 출력
table(data$변수) # 변수값의 빈도 출력
```

2) 결측값(missing value) 처리

결측값 처리하는 방법은 너무나도 많음., (정해진 답은 없음 @@)
여러 함수들과 결측값 처리하는 방식을 함께 사용하면 더 강력함 !

```
titanic_2 <-  na.omit(titanic) # NA 값을 단순 제거
titanic_surv <- filter(titanic, !is.na(Survived)) # NA값은 빼고
                                                  # 뒤에서 자세히 .,
```

기본구조

```
na.omit(data)
na.omit(data$변수)
filter(data, function)
```

(2) 변수 선택 (select)

수많은 변수 중 필요한 변수만 선택하는 방법

```
select(titanic, Survived) # Survived 변수만 출력
select(titanic, Survived, Pclass, Name) # 여러 변수 선택
select(titanic, Survived : Name)
select(titanic, 2:4) # 변수의 자리 순번으로도 선택 가능
```

기본구조

```
select(data, 변수명, 변수명, ...)
select(data, 변수명 : 변수명)
select(data, n : n)
```

(2) 변수 선택 (select) with tidyverse

수많은 변수 중 필요한 변수만 선택하는 방법

```
titanic  %>%
    select(Survived : Name) # %>% 파이프 연산자
                            # 'and than'의 개념
```

기본구조

```
data %>%  # ctrl + shift + M
    function() %>%
    function() %>% ...
```

Q1) mpg data에서 year, manufacturer, model, cty 변수만 선택해서 mpg2에 저장하시요.

(data(mpg) 해서 mpg 불러오기)

(3) 값 추려내기 (filter)

변수의 값 중 원하는 값들만 추려내는 방법

```
filter(titanic, Survived == 1) # == 같다
filter(titanic, Survived != 1)  # != 다르다

titanic %>%
    filter(Survived == 1) # tidy code
```

기본구조 filter(data, 변수에서 값을 추려낼 조건)

(3) 값 추려내기 (filter)

변수의 값 중 원하는 값들만 추려내는 방법

```
filter(titanic, Fare >= 100 ) # 부등호 조건도 가능

titanic %>%
    filter(Fare >= 100) # tidy code
```

기본구조 filter(data, 변수에서 값을 추려낼 조건)

(3) 값 추려내기 (filter)

변수의 값 중 원하는 값들만 추려내는 방법

```
filter(titanic, Survived == 1 & Fare >= 100 )
# 여러 개 조건도 가능

titanic %>%
    filter(Survived == 1 & Fare >= 100) # tidy code
```

기본구조 filter(data, 변수에서 값을 추려낼 조건)

Q2) mpg data에서 manufacturer가 “hyundai” 이면서 cty가 20 이상인 자동차는 얼마나 있는가?

Q2) mpg data에서 manufacturer가 “audi”, “chevrolet” , “ford” 인 차량은 ?

Q2) mpg data에서 manufacturer가 “audi”, “chevrolet” , “ford” 인 차량은 ?

답안코드

```
mpg %>%
  filter(manufacturer == “audi” |
         manufacturer == “chevrolet” |
         manufacturer == “ford”)
```

간단한 코드

```
mpg %>%
  filter(manufacturer %in% c(
    “audi”, “chevrolet”, “ford”))
```

Summary.

filter : 원하는 값(조건에 부합하는 값)을 추려냄
select : 원하는 변수만 선택

%>% : 파이프 연산자 "and than" 의 개념
코드를 더 간결하게 사용 가능

%in% : 조건들 중 하나라도 부합되면 (or의 개념)