lec 1022

박효선 1585063

2018년 11월 30일

################  
# 10 월 22 일  
################  
# 물어볼거: group\_by 정보 없어지는거  
  
flights

## # A tibble: 336,776 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 542 540 2 923  
## 4 2013 1 1 544 545 -1 1004  
## 5 2013 1 1 554 600 -6 812  
## 6 2013 1 1 554 558 -4 740  
## 7 2013 1 1 555 600 -5 913  
## 8 2013 1 1 557 600 -3 709  
## 9 2013 1 1 557 600 -3 838  
## 10 2013 1 1 558 600 -2 753  
## # ... with 336,766 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # arr\_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,  
## # origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>, hour <dbl>,  
## # minute <dbl>, time\_hour <dttm>

# useful creation functions  
# • Arithmetic operators: +, -, \*, /, ^.  
# • Modular arithmetic  
# – %/% (integer division)  
# – %% (remainder), x == y \* (x %/% y) + (x %% y)  
  
32 %/% 3

## [1] 10

32 %% 3

## [1] 2

# x == y \* (x %/% y) + (x %% y)  
  
transmute(flights,   
 dep\_time,   
 hour = dep\_time %/% 100, # 시간만 남는다. (정수몫)정수 나누기 연산자  
 minute = dep\_time %% 100) # 분만 남는다. 나머지 연산자

## Warning: package 'bindrcpp' was built under R version 3.5.1

## # A tibble: 336,776 x 3  
## dep\_time hour minute  
## <int> <dbl> <dbl>  
## 1 517 5 17  
## 2 533 5 33  
## 3 542 5 42  
## 4 544 5 44  
## 5 554 5 54  
## 6 554 5 54  
## 7 555 5 55  
## 8 557 5 57  
## 9 557 5 57  
## 10 558 5 58  
## # ... with 336,766 more rows

# %% 나머지 연산자  
# %/% 정수 나누기 연산자 (정수 몫 반환)  
# %\*% 행렬 곱하기 연산자  
# %in% 벡터 내 특정 값 포함 연산자  
  
# Arithmetic operators  
# +, -, \*, /, ^  
  
# Modular arithmetic  
x <- c(1:10); y <- c(5)  
x %% y

## [1] 1 2 3 4 0 1 2 3 4 0

x %/% y

## [1] 0 0 0 0 1 1 1 1 1 2

x %in% y

## [1] FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE

y %in% x

## [1] TRUE

x == y \* (x %/% y) + (x %% y)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

# Logs  
log(2)

## [1] 0.6931472

log2(2)

## [1] 1

log10(2)

## [1] 0.30103

x<- 1:10  
x

## [1] 1 2 3 4 5 6 7 8 9 10

# Offsets  
lag(x); lead(x)

## [1] NA 1 2 3 4 5 6 7 8 9

## [1] 2 3 4 5 6 7 8 9 10 NA

# Cumulative and rolling aggregates  
cumsum(x); x

## [1] 1 3 6 10 15 21 28 36 45 55

## [1] 1 2 3 4 5 6 7 8 9 10

# Logical comparisons  
# <, <=, >, >=, !=  
y <- c(1, 2, 2, NA, 3, 4)  
y

## [1] 1 2 2 NA 3 4

rank(y) # 2 와 3의 중앙값 2.5 위

## [1] 1.0 2.5 2.5 6.0 4.0 5.0

min\_rank(y) # min\_rank() : 순위(ranking) index 반환, 동일값에 대해서는 '1, 1, 1, 4, 4,...' 처리

## [1] 1 2 2 NA 4 5

min\_rank(desc(y))

## [1] 5 3 3 NA 2 1

row\_number(y) # row\_number() : 순위(ranking) index 반환, 동일값에 대해서는 '1, 2, 3, ...' 처리

## [1] 1 2 3 NA 4 5

yy <- c(1, 2, NA, 3, 4, 2)  
yy

## [1] 1 2 NA 3 4 2

row\_number(yy)

## [1] 1 2 NA 4 5 3

dense\_rank(yy) # dense\_rank() : 순위(ranking) index 반환, 동일값에 대해서는 '1, 1, 1, 2, 2,...' 처리

## [1] 1 2 NA 3 4 2

percent\_rank(yy)

## [1] 0.00 0.25 NA 0.75 1.00 0.25

yy

## [1] 1 2 NA 3 4 2

min\_rank(yy)

## [1] 1 2 NA 4 5 2

dense\_rank(yy)

## [1] 1 2 NA 3 4 2

percent\_rank(yy)

## [1] 0.00 0.25 NA 0.75 1.00 0.25

(min\_rank(yy) - 1) / (length(yy) - 1)

## [1] 0.0 0.2 NA 0.6 0.8 0.2

#(min\_rank(yy) - 1) / (length(yy, na.rm = TRUE) - 1)  
  
  
####  
yyy <- yy[!is.na(yy)] # NA 값이 없는 값만 인덱싱  
yyy

## [1] 1 2 3 4 2

percent\_rank(yyy)

## [1] 0.00 0.25 0.75 1.00 0.25

(min\_rank(yyy) - 1) / (length(yyy) - 1)

## [1] 0.00 0.25 0.75 1.00 0.25

c(0,1,3,4,1)/4

## [1] 0.00 0.25 0.75 1.00 0.25

cume\_dist(yyy)

## [1] 0.2 0.6 0.8 1.0 0.6

###  
# Grouped summarises   
# summarise() : 요약 통계량 계산  
flights

## # A tibble: 336,776 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 542 540 2 923  
## 4 2013 1 1 544 545 -1 1004  
## 5 2013 1 1 554 600 -6 812  
## 6 2013 1 1 554 558 -4 740  
## 7 2013 1 1 555 600 -5 913  
## 8 2013 1 1 557 600 -3 709  
## 9 2013 1 1 557 600 -3 838  
## 10 2013 1 1 558 600 -2 753  
## # ... with 336,766 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # arr\_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,  
## # origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>, hour <dbl>,  
## # minute <dbl>, time\_hour <dttm>

# flight 자료에서 delay time의 평균  
summarise(flights,   
 delay = mean(dep\_delay, na.rm = TRUE))

## # A tibble: 1 x 1  
## delay  
## <dbl>  
## 1 12.6

# flight 자료를 연, 월, 일로 그룹화를 한 뒤에 (각 날짜마다) delay 시간의 평균  
# group\_by() 와 summarise() 함께 사용할 경우 유용  
by\_day <- group\_by(flights, year, month, day);head(by\_day) # 365 개의 그룹

## # A tibble: 6 x 19  
## # Groups: year, month, day [1]  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 542 540 2 923  
## 4 2013 1 1 544 545 -1 1004  
## 5 2013 1 1 554 600 -6 812  
## 6 2013 1 1 554 558 -4 740  
## # ... with 12 more variables: sched\_arr\_time <int>, arr\_delay <dbl>,  
## # carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,  
## # air\_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,  
## # time\_hour <dttm>

AA <- summarize(by\_day,  
 delay = mean(dep\_delay, na.rm = TRUE));AA

## # A tibble: 365 x 4  
## # Groups: year, month [?]  
## year month day delay  
## <int> <int> <int> <dbl>  
## 1 2013 1 1 11.5   
## 2 2013 1 2 13.9   
## 3 2013 1 3 11.0   
## 4 2013 1 4 8.95  
## 5 2013 1 5 5.73  
## 6 2013 1 6 7.15  
## 7 2013 1 7 5.42  
## 8 2013 1 8 2.55  
## 9 2013 1 9 2.28  
## 10 2013 1 10 2.84  
## # ... with 355 more rows

# day 별 평균....[?]: day = row 수. group\_by 변수와 새 변수  
# group의미가 없어진다.  
  
# AA 자료를 ???이건 왜 이렇게 되는 거지?  
summarize(AA,   
 delay1 = mean(delay, na.rm=TRUE))

## # A tibble: 12 x 3  
## # Groups: year [?]  
## year month delay1  
## <int> <int> <dbl>  
## 1 2013 1 10.0   
## 2 2013 2 11.1   
## 3 2013 3 13.6   
## 4 2013 4 13.9   
## 5 2013 5 13.2   
## 6 2013 6 20.9   
## 7 2013 7 21.8   
## 8 2013 8 12.6   
## 9 2013 9 6.92  
## 10 2013 10 6.19  
## 11 2013 11 5.30  
## 12 2013 12 16.9

# month 별로 delay 평균. day 라는 정보 없어짐..