lec 1018

박효선 1585063

2018년 11월 30일

##############  
# 10 월 18일   
##############  
# 물어볼거 : gain 의미  
  
  
# 데이터  
# rm(list=ls())  
ls(pos = 2)

## [1] "airlines" "airports" "flights" "planes" "weather"

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>

#?flights  
  
# filter() : 조건에 맞는 행을 뽑음  
# 1월 1일, 1월 2일만 뽑음  
filter(flights, month == 1, day <= 2)

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

## # A tibble: 1,785 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 1,775 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>

filter(flights, month == 1, day == 1, dep\_time < 600) # 6시 이전에 출발

## # A tibble: 17 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  
## 11 2013 1 1 558 600 -2 849  
## 12 2013 1 1 558 600 -2 853  
## 13 2013 1 1 558 600 -2 924  
## 14 2013 1 1 558 600 -2 923  
## 15 2013 1 1 559 600 -1 941  
## 16 2013 1 1 559 559 0 702  
## 17 2013 1 1 559 600 -1 854  
## # ... 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>

A <- filter(flights, month == 1, day == 1, dep\_time < 600)  
  
# 수치 연산의 비교  
# near  
sqrt(2)

## [1] 1.414214

sqrt(2)^2

## [1] 2

sqrt(2)^2 == 2

## [1] FALSE

near(sqrt(2)^2, 2)

## [1] TRUE

1/49 \* 49 == 1

## [1] FALSE

near(1/49 \* 49, 1)

## [1] TRUE

# near() 오차범위 내에서 같게 만들어줌  
# Compare two numeric vectors  
# This is a safe way of comparing if two vectors of floating point numbers are (pairwise) equal. This is safer than using ==, because it has a built in tolerance  
  
1/49 \* 49 - 1 == 0

## [1] FALSE

(1/49 \* 49 - 1) == 0

## [1] FALSE

(1/49 \* 49) == 1

## [1] FALSE

round(1/49\*49 -1, 9)

## [1] 0

round(1/49\*49 -1, 9) == 0

## [1] TRUE

?round

## starting httpd help server ... done

# knitr::include\_graphics("diagrams/transform-logical.png")  
  
filter(flights, month == 11 | month == 12)

## # A tibble: 55,403 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 11 1 5 2359 6 352  
## 2 2013 11 1 35 2250 105 123  
## 3 2013 11 1 455 500 -5 641  
## 4 2013 11 1 539 545 -6 856  
## 5 2013 11 1 542 545 -3 831  
## 6 2013 11 1 549 600 -11 912  
## 7 2013 11 1 550 600 -10 705  
## 8 2013 11 1 554 600 -6 659  
## 9 2013 11 1 554 600 -6 826  
## 10 2013 11 1 554 600 -6 749  
## # ... with 55,393 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>

filter(flights, month %in% c(11, 12)) # 지정한 vector 값 안에 있는 값

## # A tibble: 55,403 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 11 1 5 2359 6 352  
## 2 2013 11 1 35 2250 105 123  
## 3 2013 11 1 455 500 -5 641  
## 4 2013 11 1 539 545 -6 856  
## 5 2013 11 1 542 545 -3 831  
## 6 2013 11 1 549 600 -11 912  
## 7 2013 11 1 550 600 -10 705  
## 8 2013 11 1 554 600 -6 659  
## 9 2013 11 1 554 600 -6 826  
## 10 2013 11 1 554 600 -6 749  
## # ... with 55,393 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>

nov\_dec <- filter(flights, month %in% c(11, 12)) # 11, 12월 데이터만 뽑아냈음  
  
  
# Missing value  
a <- 1  
NA > 5; 10 == NA; NA + 10 ; NA / 2; NA == NA # 결측치 연산은 모두 결측치

## [1] NA

## [1] NA

## [1] NA

## [1] NA

## [1] NA

x = c(1, NA, 3)  
is.na(x); !is.na(x)

## [1] FALSE TRUE FALSE

## [1] TRUE FALSE TRUE

df <- tibble(x = c(1, NA, 3))  
df

## # A tibble: 3 x 1  
## x  
## <dbl>  
## 1 1  
## 2 NA  
## 3 3

filter(df, x > 1)

## # A tibble: 1 x 1  
## x  
## <dbl>  
## 1 3

filter(df, x > 1 | is.na(x))

## # A tibble: 2 x 1  
## x  
## <dbl>  
## 1 NA  
## 2 3

filter(df, x > 1, is.na(x)) # 동시에 만족하는 행이 없기 때문에 0 x 1

## # A tibble: 0 x 1  
## # ... with 1 variable: x <dbl>

# arrange()  
# 지정한 변수의 크기 순으로 자료를 정렬하기  
arrange(flights, year, month, day)

## # 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>

arrange(flights, dep\_delay)

## # 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 12 7 2040 2123 -43 40  
## 2 2013 2 3 2022 2055 -33 2240  
## 3 2013 11 10 1408 1440 -32 1549  
## 4 2013 1 11 1900 1930 -30 2233  
## 5 2013 1 29 1703 1730 -27 1947  
## 6 2013 8 9 729 755 -26 1002  
## 7 2013 10 23 1907 1932 -25 2143  
## 8 2013 3 30 2030 2055 -25 2213  
## 9 2013 3 2 1431 1455 -24 1601  
## 10 2013 5 5 934 958 -24 1225  
## # ... 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>

arrange(flights, desc(dep\_delay)) # 내림차순

## # 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 9 641 900 1301 1242  
## 2 2013 6 15 1432 1935 1137 1607  
## 3 2013 1 10 1121 1635 1126 1239  
## 4 2013 9 20 1139 1845 1014 1457  
## 5 2013 7 22 845 1600 1005 1044  
## 6 2013 4 10 1100 1900 960 1342  
## 7 2013 3 17 2321 810 911 135  
## 8 2013 6 27 959 1900 899 1236  
## 9 2013 7 22 2257 759 898 121  
## 10 2013 12 5 756 1700 896 1058  
## # ... 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>

arrange(flights, desc(day), month)

## # 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 31 1 2100 181 124  
## 2 2013 1 31 4 2359 5 455  
## 3 2013 1 31 7 2359 8 453  
## 4 2013 1 31 12 2250 82 132  
## 5 2013 1 31 26 2154 152 328  
## 6 2013 1 31 34 2159 155 135  
## 7 2013 1 31 37 2249 108 132  
## 8 2013 1 31 54 2250 124 152  
## 9 2013 1 31 453 500 -7 651  
## 10 2013 1 31 522 525 -3 820  
## # ... 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>

df <- tibble(x = c(5, 2, NA)) # NA 값은 항상 마지막 순서로 정렬  
df

## # A tibble: 3 x 1  
## x  
## <dbl>  
## 1 5  
## 2 2  
## 3 NA

arrange(df)

## # A tibble: 3 x 1  
## x  
## <dbl>  
## 1 5  
## 2 2  
## 3 NA

arrange(df, x)

## # A tibble: 3 x 1  
## x  
## <dbl>  
## 1 2  
## 2 5  
## 3 NA

arrange(df, desc(x))

## # A tibble: 3 x 1  
## x  
## <dbl>  
## 1 5  
## 2 2  
## 3 NA

# select()  
# starts\_with("abc"): “abc”로 시작하는 이름의 변수 선택  
# ends\_with("xyz"): “xyz”로 끝나는 이름의 변수 선택.  
# contains("ijk"): “ijk”가 들어있는 이름의 변수 선택.  
# num\_range("x", 1:3): x1, x2, x3 선택  
  
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>

select(flights, year, month, day)

## # A tibble: 336,776 x 3  
## year month day  
## <int> <int> <int>  
## 1 2013 1 1  
## 2 2013 1 1  
## 3 2013 1 1  
## 4 2013 1 1  
## 5 2013 1 1  
## 6 2013 1 1  
## 7 2013 1 1  
## 8 2013 1 1  
## 9 2013 1 1  
## 10 2013 1 1  
## # ... with 336,766 more rows

select(flights, year, day, month)

## # A tibble: 336,776 x 3  
## year day month  
## <int> <int> <int>  
## 1 2013 1 1  
## 2 2013 1 1  
## 3 2013 1 1  
## 4 2013 1 1  
## 5 2013 1 1  
## 6 2013 1 1  
## 7 2013 1 1  
## 8 2013 1 1  
## 9 2013 1 1  
## 10 2013 1 1  
## # ... with 336,766 more rows

select(flights, year:arr\_time)

## # A tibble: 336,776 x 7  
## 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

select(flights, year:month)

## # A tibble: 336,776 x 2  
## year month  
## <int> <int>  
## 1 2013 1  
## 2 2013 1  
## 3 2013 1  
## 4 2013 1  
## 5 2013 1  
## 6 2013 1  
## 7 2013 1  
## 8 2013 1  
## 9 2013 1  
## 10 2013 1  
## # ... with 336,766 more rows

select(flights, -(year:month))

## # A tibble: 336,776 x 17  
## day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## <int> <int> <int> <dbl> <int> <int>  
## 1 1 517 515 2 830 819  
## 2 1 533 529 4 850 830  
## 3 1 542 540 2 923 850  
## 4 1 544 545 -1 1004 1022  
## 5 1 554 600 -6 812 837  
## 6 1 554 558 -4 740 728  
## 7 1 555 600 -5 913 854  
## 8 1 557 600 -3 709 723  
## 9 1 557 600 -3 838 846  
## 10 1 558 600 -2 753 745  
## # ... with 336,766 more rows, and 11 more variables: 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>

select(flights, -(year:arr\_delay))

## # A tibble: 336,776 x 10  
## carrier flight tailnum origin dest air\_time distance hour minute  
## <chr> <int> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 UA 1545 N14228 EWR IAH 227 1400 5 15  
## 2 UA 1714 N24211 LGA IAH 227 1416 5 29  
## 3 AA 1141 N619AA JFK MIA 160 1089 5 40  
## 4 B6 725 N804JB JFK BQN 183 1576 5 45  
## 5 DL 461 N668DN LGA ATL 116 762 6 0  
## 6 UA 1696 N39463 EWR ORD 150 719 5 58  
## 7 B6 507 N516JB EWR FLL 158 1065 6 0  
## 8 EV 5708 N829AS LGA IAD 53 229 6 0  
## 9 B6 79 N593JB JFK MCO 140 944 6 0  
## 10 AA 301 N3ALAA LGA ORD 138 733 6 0  
## # ... with 336,766 more rows, and 1 more variable: time\_hour <dttm>

select(flights, starts\_with("dep"))

## # A tibble: 336,776 x 2  
## dep\_time dep\_delay  
## <int> <dbl>  
## 1 517 2  
## 2 533 4  
## 3 542 2  
## 4 544 -1  
## 5 554 -6  
## 6 554 -4  
## 7 555 -5  
## 8 557 -3  
## 9 557 -3  
## 10 558 -2  
## # ... with 336,766 more rows

select(flights, ends\_with("time"))

## # A tibble: 336,776 x 5  
## dep\_time sched\_dep\_time arr\_time sched\_arr\_time air\_time  
## <int> <int> <int> <int> <dbl>  
## 1 517 515 830 819 227  
## 2 533 529 850 830 227  
## 3 542 540 923 850 160  
## 4 544 545 1004 1022 183  
## 5 554 600 812 837 116  
## 6 554 558 740 728 150  
## 7 555 600 913 854 158  
## 8 557 600 709 723 53  
## 9 557 600 838 846 140  
## 10 558 600 753 745 138  
## # ... with 336,766 more rows

select(flights, contains("arr"))

## # A tibble: 336,776 x 4  
## arr\_time sched\_arr\_time arr\_delay carrier  
## <int> <int> <dbl> <chr>   
## 1 830 819 11 UA   
## 2 850 830 20 UA   
## 3 923 850 33 AA   
## 4 1004 1022 -18 B6   
## 5 812 837 -25 DL   
## 6 740 728 12 UA   
## 7 913 854 19 B6   
## 8 709 723 -14 EV   
## 9 838 846 -8 B6   
## 10 753 745 8 AA   
## # ... with 336,766 more rows

select(flights, contains("arr\_"))

## # A tibble: 336,776 x 3  
## arr\_time sched\_arr\_time arr\_delay  
## <int> <int> <dbl>  
## 1 830 819 11  
## 2 850 830 20  
## 3 923 850 33  
## 4 1004 1022 -18  
## 5 812 837 -25  
## 6 740 728 12  
## 7 913 854 19  
## 8 709 723 -14  
## 9 838 846 -8  
## 10 753 745 8  
## # ... with 336,766 more rows

# rename  
# 변수 이름 바꾸기  
head(flights)

## # A tibble: 6 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  
## # ... 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>

rename(flights, tail\_num = tailnum)

## # 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>, tail\_num <chr>,  
## # origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>, hour <dbl>,  
## # minute <dbl>, time\_hour <dttm>

flights <- rename(flights, tail\_num = tailnum)  
head(flights)

## # A tibble: 6 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  
## # ... with 12 more variables: sched\_arr\_time <int>, arr\_delay <dbl>,  
## # carrier <chr>, flight <int>, tail\_num <chr>, origin <chr>, dest <chr>,  
## # air\_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,  
## # time\_hour <dttm>

# View(flights)  
  
# everything() : 나머지 모든 변수를 의미  
select(flights, time\_hour, everything()) # time\_hour 제일 앞으로 나옴

## # A tibble: 336,776 x 19  
## time\_hour year month day dep\_time sched\_dep\_time dep\_delay  
## <dttm> <int> <int> <int> <int> <int> <dbl>  
## 1 2013-01-01 05:00:00 2013 1 1 517 515 2  
## 2 2013-01-01 05:00:00 2013 1 1 533 529 4  
## 3 2013-01-01 05:00:00 2013 1 1 542 540 2  
## 4 2013-01-01 05:00:00 2013 1 1 544 545 -1  
## 5 2013-01-01 06:00:00 2013 1 1 554 600 -6  
## 6 2013-01-01 05:00:00 2013 1 1 554 558 -4  
## 7 2013-01-01 06:00:00 2013 1 1 555 600 -5  
## 8 2013-01-01 06:00:00 2013 1 1 557 600 -3  
## 9 2013-01-01 06:00:00 2013 1 1 557 600 -3  
## 10 2013-01-01 06:00:00 2013 1 1 558 600 -2  
## # ... with 336,766 more rows, and 12 more variables: arr\_time <int>,  
## # sched\_arr\_time <int>, arr\_delay <dbl>, carrier <chr>, flight <int>,  
## # tail\_num <chr>, origin <chr>, dest <chr>, air\_time <dbl>,  
## # distance <dbl>, hour <dbl>, minute <dbl>

# mutate() : 새로운 변수를 만들어 기존 자료에 추가  
df

## # A tibble: 3 x 1  
## x  
## <dbl>  
## 1 5  
## 2 2  
## 3 NA

df$y <- LETTERS[1:3] # "A","B","C" 를 y 변수 안에 저장  
df

## # A tibble: 3 x 2  
## x y   
## <dbl> <chr>  
## 1 5 A   
## 2 2 B   
## 3 NA C

flights\_sml <- select(flights, year:day, ends\_with("delay"), distance, air\_time)  
flights\_sml %>% mutate(gain = arr\_delay - dep\_delay,   
 hours = air\_time/60,  
 gain\_per\_hour = gain/hours)

## # A tibble: 336,776 x 10  
## year month day dep\_delay arr\_delay distance air\_time gain hours  
## <int> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2013 1 1 2 11 1400 227 9 3.78   
## 2 2013 1 1 4 20 1416 227 16 3.78   
## 3 2013 1 1 2 33 1089 160 31 2.67   
## 4 2013 1 1 -1 -18 1576 183 -17 3.05   
## 5 2013 1 1 -6 -25 762 116 -19 1.93   
## 6 2013 1 1 -4 12 719 150 16 2.5   
## 7 2013 1 1 -5 19 1065 158 24 2.63   
## 8 2013 1 1 -3 -14 229 53 -11 0.883  
## 9 2013 1 1 -3 -8 944 140 -5 2.33   
## 10 2013 1 1 -2 8 733 138 10 2.3   
## # ... with 336,766 more rows, and 1 more variable: gain\_per\_hour <dbl>

A <- mutate(flights\_sml,   
 gain = arr\_delay - dep\_delay, #### 예정비행시간 보다 순수 지연된 시간??  
 hours = air\_time/60, # 비행시간  
 gain\_per\_hour = gain/hours) # 시간당 얼마나 늦어졌는가  
A

## # A tibble: 336,776 x 10  
## year month day dep\_delay arr\_delay distance air\_time gain hours  
## <int> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2013 1 1 2 11 1400 227 9 3.78   
## 2 2013 1 1 4 20 1416 227 16 3.78   
## 3 2013 1 1 2 33 1089 160 31 2.67   
## 4 2013 1 1 -1 -18 1576 183 -17 3.05   
## 5 2013 1 1 -6 -25 762 116 -19 1.93   
## 6 2013 1 1 -4 12 719 150 16 2.5   
## 7 2013 1 1 -5 19 1065 158 24 2.63   
## 8 2013 1 1 -3 -14 229 53 -11 0.883  
## 9 2013 1 1 -3 -8 944 140 -5 2.33   
## 10 2013 1 1 -2 8 733 138 10 2.3   
## # ... with 336,766 more rows, and 1 more variable: gain\_per\_hour <dbl>

# transmute() : 새로 만든 변수만을 가지는 자료를 만듦  
transmute(flights\_sml,  
 gain = arr\_delay - dep\_delay,   
 hours = air\_time/60,   
 gain\_per\_hour = gain/hours)

## # A tibble: 336,776 x 3  
## gain hours gain\_per\_hour  
## <dbl> <dbl> <dbl>  
## 1 9 3.78 2.38  
## 2 16 3.78 4.23  
## 3 31 2.67 11.6   
## 4 -17 3.05 -5.57  
## 5 -19 1.93 -9.83  
## 6 16 2.5 6.4   
## 7 24 2.63 9.11  
## 8 -11 0.883 -12.5   
## 9 -5 2.33 -2.14  
## 10 10 2.3 4.35  
## # ... with 336,766 more rows

```