WEEK1

CHAcha

2021 7 1

```
library(nycflights13)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     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>
```

```
summary(flights)
```

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
##  Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##                                                  NA's   :8255
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
##  Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##  NA's   :8255      NA's   :8713                  NA's   :9430
##    carrier              flight       tailnum             origin
##  Length:336776      Min.   :   1   Length:336776      Length:336776
##  Class :character   1st Qu.: 553   Class :character   Class :character
##  Mode  :character   Median :1496   Mode  :character   Mode  :character
##                     Mean   :1972
##                     3rd Qu.:3465
##                     Max.   :8500
##
##      dest              air_time        distance         hour
##  Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 872   Median :13.00
##                     Mean   :150.7   Mean   :1040   Mean   :13.18
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##                     NA's   :9430
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :26.23   Mean   :2013-07-03 05:22:54
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # ... with 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>
```

your turn1

```
flights%>%filter(dep_delay>=120)
```

```
## # A tibble: 9,888 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      848           1835       853     1001           1950
##  2  2013     1     1      957            733       144     1056            853
##  3  2013     1     1     1114            900       134     1447           1222
##  4  2013     1     1     1540           1338       122     2020           1825
##  5  2013     1     1     1815           1325       290     2120           1542
##  6  2013     1     1     1842           1422       260     1958           1535
##  7  2013     1     1     1856           1645       131     2212           2005
##  8  2013     1     1     1934           1725       129     2126           1855
##  9  2013     1     1     1938           1703       155     2109           1823
## 10  2013     1     1     1942           1705       157     2124           1830
## # ... with 9,878 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>
```

```
flights%>%filter(dep_delay==0&arr_delay>=0)
```

```
## # A tibble: 5,400 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      600            600         0      837            825
##  2  2013     1     1      635            635         0     1028            940
##  3  2013     1     1      739            739         0     1104           1038
##  4  2013     1     1      745            745         0     1135           1125
##  5  2013     1     1      800            800         0     1022           1014
##  6  2013     1     1      805            805         0     1015           1005
##  7  2013     1     1      810            810         0     1048           1037
##  8  2013     1     1      823            823         0     1151           1135
##  9  2013     1     1      830            830         0     1018           1015
## 10  2013     1     1      835            835         0     1210           1150
## # ... with 5,390 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>
```

```
flights%>%filter(dep_delay>=60&arr_delay-dep_delay<=-30)
```

```
## # A tibble: 2,074 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1     1716           1545        91     2140           2039
##  2  2013     1     1     2205           1720       285       46           2040
##  3  2013     1     1     2326           2130       116      131             18
##  4  2013     1     3     1503           1221       162     1803           1555
##  5  2013     1     3     1821           1530       171     2131           1910
##  6  2013     1     3     1839           1700        99     2056           1950
##  7  2013     1     3     1850           1745        65     2148           2120
##  8  2013     1     3     1923           1815        68     2036           1958
##  9  2013     1     3     1941           1759       102     2246           2139
## 10  2013     1     3     1950           1845        65     2228           2227
## # ... with 2,064 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>
```

```
flights%>%filter(is.na(dep_time))
```

```
## # A tibble: 8,255 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1       NA           1630        NA       NA           1815
##  2  2013     1     1       NA           1935        NA       NA           2240
##  3  2013     1     1       NA           1500        NA       NA           1825
##  4  2013     1     1       NA            600        NA       NA            901
##  5  2013     1     2       NA           1540        NA       NA           1747
##  6  2013     1     2       NA           1620        NA       NA           1746
##  7  2013     1     2       NA           1355        NA       NA           1459
##  8  2013     1     2       NA           1420        NA       NA           1644
##  9  2013     1     2       NA           1321        NA       NA           1536
## 10  2013     1     2       NA           1545        NA       NA           1910
## # ... with 8,245 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>
```

dep_time이 결측치 일때, dep_delay,arr_time,arr_delay역시 결측치임을 확인 즉 결항을 의미

your turn2

```
dep<-flights %>%filter(!is.na(dep_time))%>% arrange(desc(dep_delay))
dep[328521,]
```

```
## # A tibble: 1 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013    12     7     2040           2123       -43       40           2352
## # ... with 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>
```

```
velocity<-flights %>% arrange(desc(distance/air_time))
velocity[1,]
```

```
## # A tibble: 1 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     5    25     1709           1700         9     1923           1937
## # ... with 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>
```

your turn3

```
flights %>% select(starts_with("dep")|starts_with("arr"))
```

```
## # A tibble: 336,776 x 4
##    dep_time dep_delay arr_time arr_delay
##       <int>     <dbl>    <int>     <dbl>
##  1      517         2      830        11
##  2      533         4      850        20
##  3      542         2      923        33
##  4      544        -1     1004       -18
##  5      554        -6      812       -25
##  6      554        -4      740        12
##  7      555        -5      913        19
##  8      557        -3      709       -14
##  9      557        -3      838        -8
## 10      558        -2      753         8
## # ... with 336,766 more rows
```

your turn4

```
flights %>% select(contains("dep_time"))
```

```
## # A tibble: 336,776 x 2
##    dep_time sched_dep_time
##       <int>          <int>
##  1      517            515
##  2      533            529
##  3      542            540
##  4      544            545
##  5      554            600
##  6      554            558
##  7      555            600
##  8      557            600
##  9      557            600
## 10      558            600
## # ... with 336,766 more rows
```

```
flight1=flights
flight1$dep_time=flight1$dep_time*0.01
flight1$sched_dep_time=flight1$sched_dep_time*0.01
flights2<-flight1%>%mutate(dt_H=floor(dep_time),dt_M=(dep_time-floor(dep_time))*100,sched_dt_H=
floor(sched_dep_time),sched_dt_M=(sched_dep_time-floor(sched_dep_time))*100)
flights2 %>% select(contains("dep_time")|contains("dt"))
```

```
## # A tibble: 336,776 x 6
##    dep_time sched_dep_time  dt_H  dt_M sched_dt_H sched_dt_M
##       <dbl>          <dbl> <dbl> <dbl>      <dbl>      <dbl>
##  1     5.17           5.15     5  17            5       15.0
##  2     5.33           5.29     5  33            5       29
##  3     5.42           5.4      5  42            5       40.0
##  4     5.44           5.45     5  44.0          5       45
##  5     5.54           6        5  54            6        0
##  6     5.54           5.58     5  54            5       58
##  7     5.55           6        5  55            6        0
##  8     5.57           6        5  57            6        0
##  9     5.57           6        5  57            6        0
## 10     5.58           6        5  58            6        0
## # ... with 336,766 more rows
```

1. 소수로 바꾼 후 버림을 하여 시를 표시하고, x100을 하여 분을 나타냄

```
flights3=flights %>% mutate(gap=arr_time-dep_time)
flights3 %>% select("air_time","gap")
```

```
## # A tibble: 336,776 x 2
##    air_time   gap
##       <dbl> <int>
##  1      227   313
##  2      227   317
##  3      160   381
##  4      183   460
##  5      116   258
##  6      150   186
##  7      158   358
##  8       53   152
##  9      140   281
## 10      138   195
## # ... with 336,766 more rows
```

2.일반적으로(모든경우는 아님) gap이 airtime보다 큼을 알 수 있음. 이는 arr_time, dep_time이 이착륙을 하였을 때의 시간이라고 추측됨

```
flights4=flights %>% mutate(delay=dep_time-sched_dep_time)
flights4 %>% select("delay","dep_delay")
```

```
## # A tibble: 336,776 x 2
##    delay dep_delay
##    <int>     <dbl>
##  1     2         2
##  2     4         4
##  3     2         2
##  4    -1        -1
##  5   -46        -6
##  6    -4        -4
##  7   -45        -5
##  8   -43        -3
##  9   -43        -3
## 10   -42        -2
## # ... with 336,766 more rows
```

3.일반적으로 delay와 dep_delay가 동일한 값을 가져야 한다고 생각. 그러나 실제 데이터에서 같지 않은 것도 종종 발견됨.

your turn5

```
SD=flights %>% group_by(carrier) %>% summarise(sd1=sd(dep_delay,na.rm = T))
MEAN=flights %>% group_by(carrier) %>% summarise(mean1=mean(dep_delay,na.rm = T))
SD %>% arrange(desc(sd1))
```

```
## # A tibble: 16 x 2
##    carrier   sd1
##    <chr>   <dbl>
##  1 HA       74.1
##  2 F9       58.4
##  3 FL       52.7
##  4 YV       49.2
##  5 EV       46.6
##  6 9E       45.9
##  7 VX       44.8
##  8 WN       43.3
##  9 OO       43.1
## 10 DL       39.7
## 11 MQ       39.2
## 12 B6       38.5
## 13 AA       37.4
## 14 UA       35.7
## 15 AS       31.4
## 16 US       28.1
```

```
MEAN %>% arrange(desc(mean1))
```

```
## # A tibble: 16 x 2
##    carrier mean1
##    <chr>   <dbl>
##  1 F9      20.2
##  2 EV      20.0
##  3 YV      19.0
##  4 FL      18.7
##  5 WN      17.7
##  6 9E      16.7
##  7 B6      13.0
##  8 VX      12.9
##  9 OO      12.6
## 10 UA      12.1
## 11 MQ      10.6
## 12 DL       9.26
## 13 AA       8.59
## 14 AS       5.80
## 15 HA       4.90
## 16 US       3.78
```

평균이 가장 큰 항공사는 F9 이고, 표준편차가 가장 큰 항공사는 HA이다

```
DAY=flights %>% group_by(month,day)
DAY %>% summarise(mean1=mean(dep_delay,na.rm=T)) %>% arrange(desc(mean1))
```

```
## `summarise()` has grouped output by 'month'. You can override using the `.groups` argument.
```

```
## # A tibble: 365 x 3
## # Groups:   month [12]
##    month   day mean1
##    <int> <int> <dbl>
##  1     3     8  83.5
##  2     7     1  56.2
##  3     9     2  53.0
##  4     7    10  52.9
##  5    12     5  52.3
##  6     5    23  51.1
##  7     9    12  50.0
##  8     6    28  48.8
##  9     6    24  47.2
## 10     7    22  46.7
## # ... with 355 more rows
```

3월 8일이 제일 지연시간이 길었다.

```
flight4=flights %>% filter(month==3&day==8)
SD1=flight4 %>% group_by(carrier) %>% summarise(sd1=sd(dep_delay,na.rm = T))
SD1 %>% arrange(desc(sd1))
```

```
## # A tibble: 15 x 2
##    carrier    sd1
##    <chr>    <dbl>
##  1 F9      306.
##  2 FL      146.
##  3 EV      123.
##  4 UA       85.9
##  5 AA       83.7
##  6 MQ       81.6
##  7 DL       80.3
##  8 B6       80.2
##  9 WN       74.2
## 10 9E       69.6
## 11 US       50.7
## 12 VX       49.5
## 13 AS        1.41
## 14 HA       NA
## 15 YV       NA
```

F9가 가장 큰 표준편차를 가짐을 알 수 있음. F9 항공사는 전체 날짜를 비교했을 때 두 번째로 큰 표준편차를 가지는 항공사였음.