Data Manipulation: Arrange( ) & Mutate( )

Pilsung Kang

School of Industrial Management Engineering

Korea University

dplyr: Arrange( )

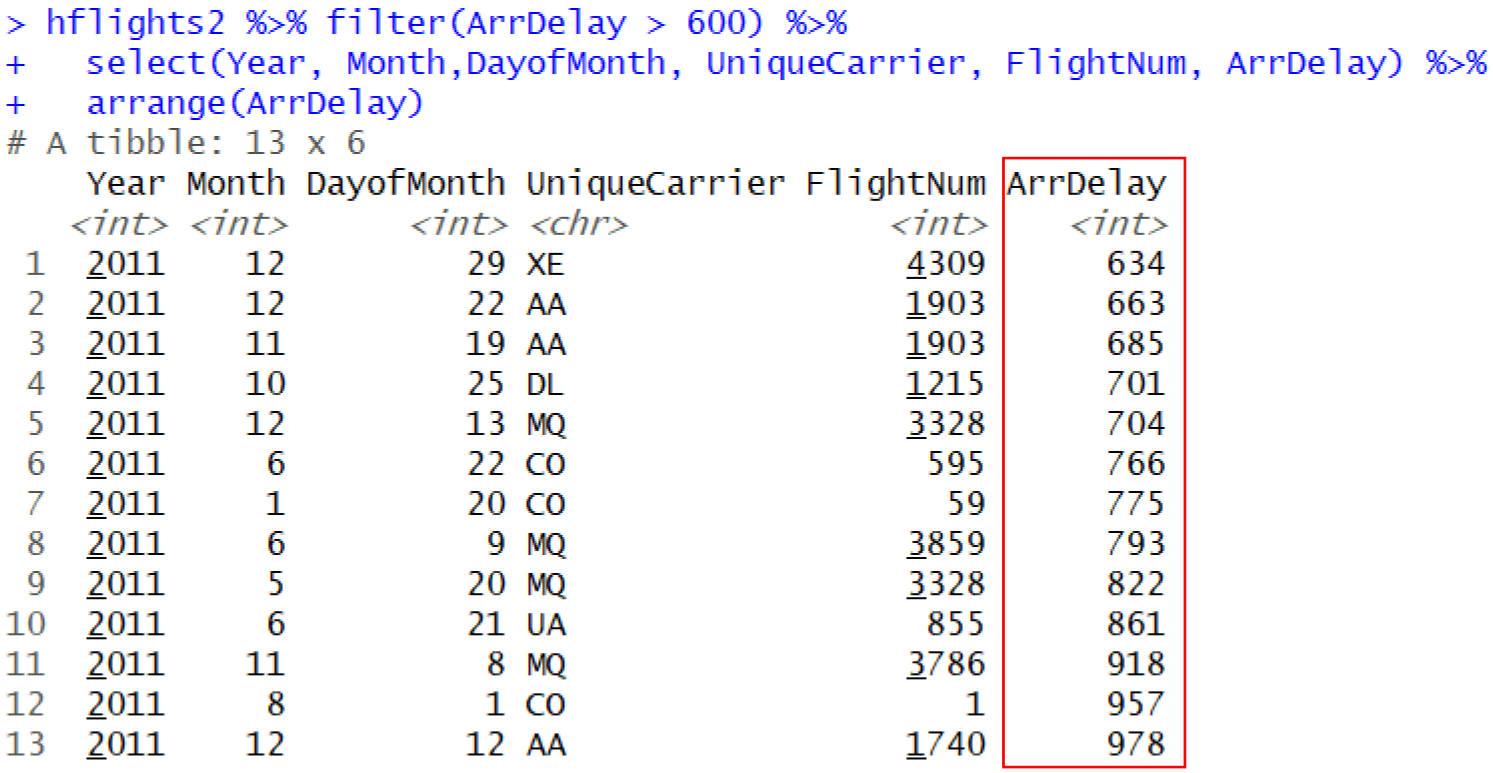
- Arrange( )
  - ✓ Q) For all flights with arrival delay greater than 10 hours, give the variables Year, Month,bDayofMonth, UniqueCarrier, FlightNum and ArrDelay
  - ✓ Sort the observations in the result according to variable ArrDelay

```r
# filter, select, and arrange
hflights2 %>% filter(ArrDelay > 600) %>%
     select(Year, Month,DayofMonth, UniqueCarrier, FlightNum, ArrDelay) %>%
     arrange(ArrDelay)

hflights2 %>% filter(ArrDelay > 600) %>%
     select(Year, Month,DayofMonth, UniqueCarrier, FlightNum, ArrDelay) %>%
     arrange(desc(ArrDelay))
```

dplyr: Arrange( )

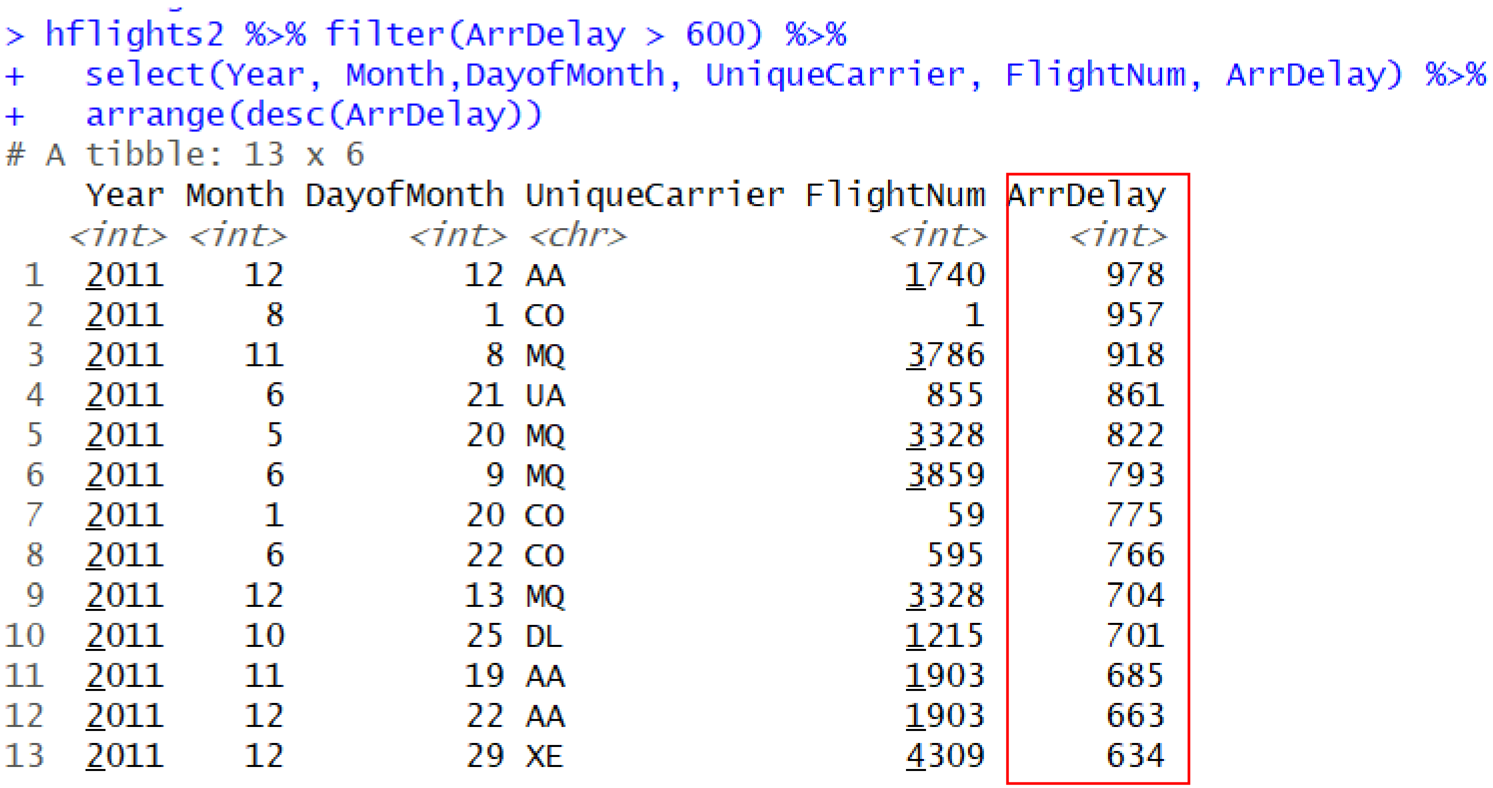
- Arrange( )
  - ✓ Arranged in an ascending order



```
> hflights2 %>% filter(ArrDelay > 600) %>%
+   select(Year, Month,DayofMonth, UniqueCarrier, FlightNum, ArrDelay) %>%
+   arrange(ArrDelay)
# A tibble: 13 x 6
    Year Month DayofMonth UniqueCarrier FlightNum ArrDelay
   <int> <int>      <int> <chr>             <int>    <int>
 1  2011    12         29 XE                 4309      634
 2  2011    12         22 AA                 1903      663
 3  2011    11         19 AA                 1903      685
 4  2011    10         25 DL                 1215      701
 5  2011    12         13 MQ                 3328      704
 6  2011     6         22 CO                  595      766
 7  2011     1         20 CO                   59      775
 8  2011     6          9 MQ                 3859      793
 9  2011     5         20 MQ                 3328      822
10  2011     6         21 UA                  855      861
11  2011    11          8 MQ                 3786      918
12  2011     8          1 CO                    1      957
13  2011    12         12 AA                 1740      978
```

dplyr: Arrange( )

- Arrange( )
  - ✓ Arranged in a descending order

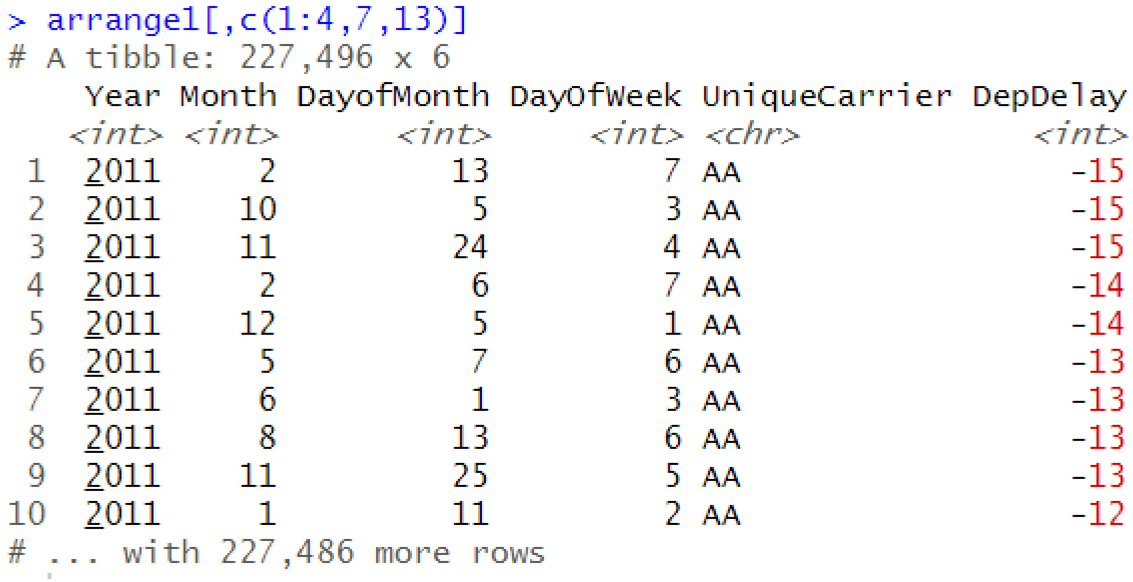


```
> hflights2 %>% filter(ArrDelay > 600) %>%
+   select(Year, Month,DayofMonth, UniqueCarrier, FlightNum, ArrDelay) %>%
+   arrange(desc(ArrDelay))
# A tibble: 13 x 6
    Year Month DayofMonth UniqueCarrier FlightNum ArrDelay
   <int> <int>      <int> <chr>             <int>    <int>
 1  2011    12         12 AA                 1740      978
 2  2011     8          1 CO                    1      957
 3  2011    11          8 MQ                 3786      918
 4  2011     6         21 UA                  855      861
 5  2011     5         20 MQ                 3328      822
 6  2011     6          9 MQ                 3859      793
 7  2011     1         20 CO                   59      775
 8  2011     6         22 CO                  595      766
 9  2011    12         13 MQ                 3328      704
10  2011    10         25 DL                 1215      701
11  2011    11         19 AA                 1903      685
12  2011    12         22 AA                 1903      663
13  2011    12         29 XE                 4309      634
```

dplyr: Arrange( )

- Arrange( )
  - ✓ Arrange with more than two variables

```r
# Arrange with more than two variables
arrange1 <- arrange(hflights2, UniqueCarrier, DepDelay)
arrange1[,c(1:4,7,13)]
```



```
> arrange1[,c(1:4,7,13)]
# A tibble: 227,496 x 6
    Year Month DayofMonth DayOfWeek UniqueCarrier DepDelay
   <int> <int>      <int>     <int> <chr>            <int>
 1  2011     2         13         7 AA                 -15
 2  2011    10          5         3 AA                 -15
 3  2011    11         24         4 AA                 -15
 4  2011     2          6         7 AA                 -14
 5  2011    12          5         1 AA                 -14
 6  2011     5          7         6 AA                 -13
 7  2011     6          1         3 AA                 -13
 8  2011     8         13         6 AA                 -13
 9  2011    11         25         5 AA                 -13
10  2011     1         11         2 AA                 -12
# ... with 227,486 more rows
```

dplyr: Mutate( )

- Mutate( )
  - ✓ Create a new column, assigns a value
  - ✓ Arguments:
    - Data frame
    - Name of new column = value
  - ✓ Example: mutate(surveys, weight_kg = weight/1000)
  - ✓ Imagine to have a data frame df with three columns: Id (the identifier), w (weight in Kg) and h (height in m)
  - ✓ We want to create a fourth variable bmi with the Body Mass Index: bmi = w/h^2. This can be easily done with the mutate() function:

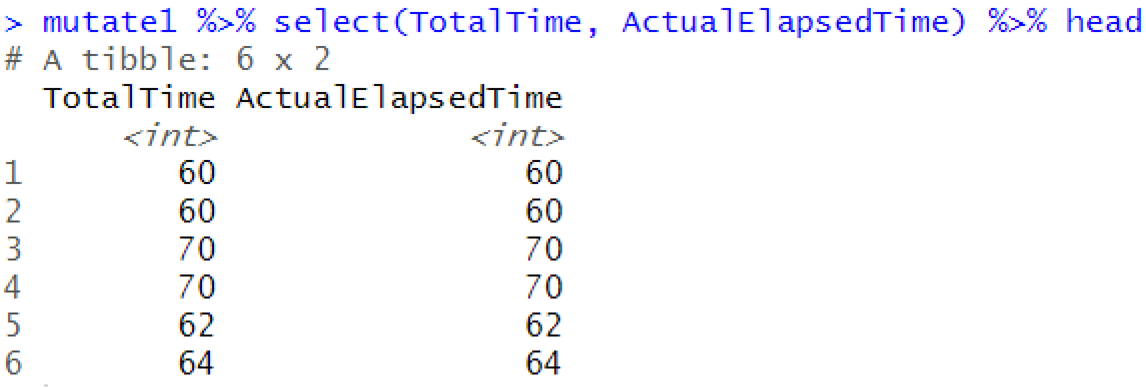
```
mutate(df, bmi = w/h^2)
```

dplyr: Mutate( )

- Mutate( )
  - ✓ Similarly, we create a new variable TotalTime measuring the total flight time, as the sum of TaxiIn (time spent on ground before taking off), TaxiOut (ground time after landing) and AirTime:

```
# Mutate example
mutate1 <- hflights2 %>% mutate(TotalTime = TaxiIn + AirTime + TaxiOut)

# Compare with the original value
mutate1 %>% select(TotalTime, ActualElapsedTime) %>% head
```

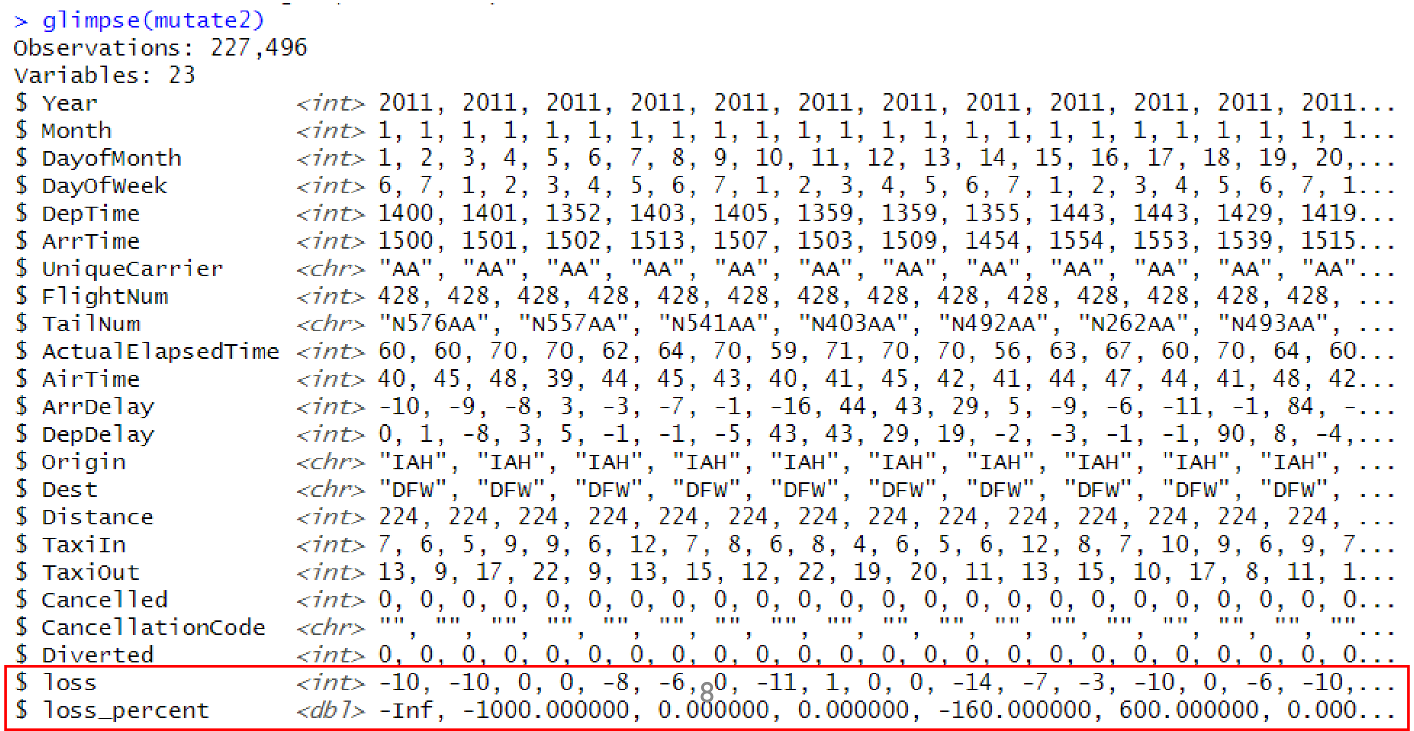


```
> mutate1 %>% select(TotalTime, ActualElapsedTime) %>% head
# A tibble: 6 x 2
  TotalTime ActualElapsedTime
      <int>             <int>
1        60                60
2        60                60
3        70                70
4        70                70
5        62                62
6        64                64
```

dplyr: Mutate( )

- Mutate( )
  - ✓ Add multiple variables using mutate

```
# Add multiple variables
mutate2 <- mutate(hflights,
                           loss = ArrDelay - DepDelay,
                           loss_percent = (ArrDelay - DepDelay)/DepDelay * 100)
glimpse(mutate2)
```



```
> glimpse(mutate2)
Observations: 227,496
Variables: 23
$ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011...
$ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ DayofMonth        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,...
$ DayOfWeek         <int> 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 1...
$ DepTime           <int> 1400, 1401, 1352, 1403, 1405, 1359, 1359, 1355, 1443, 1443, 1429, 1419...
$ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1554, 1553, 1539, 1515...
$ UniqueCarrier     <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA"...
$ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428, 428, 428, 428, 428, ...
$ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA", "N262AA", "N493AA", ...
$ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 63, 67, 60, 70, 64, 60...
$ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 44, 47, 44, 41, 48, 42...
$ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -9, -6, -11, -1, 84, -...
$ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -3, -1, -1, 90, 8, -4,...
$ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
$ Dest              <chr> "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", ...
$ Distance          <int> 224, 224, 224, 224, 224, 224, 224, 224, 224, 224, 224, 224, 224, 224, ...
$ TaxiIn            <int> 7, 6, 5, 9, 9, 6, 12, 7, 8, 6, 8, 4, 6, 5, 6, 12, 8, 7, 10, 9, 6, 9, 7...
$ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13, 15, 10, 17, 8, 11, 1...
$ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ CancellationCode  <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", ""...
$ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ loss              <int> -10, -10, 0, 0, -8, -6, 0, -11, 1, 0, 0, -14, -7, -3, -10, 0, -6, -10,...
$ loss_percent      <dbl> -Inf, -1000.000000, 0.000000, 0.000000, -160.000000, 600.000000, 0.000...
```

ANY
questions ?