Data Manipulation: Filter( )

Pilsung Kang
School of Industrial Management Engineering
Korea University

dplyr: Filter( )

- Filter( )
  - ✓ Choose rows based on a specific criterion
  - ✓ Arguments:
    - Data frame
    - Relational expressions (returns true/false)
      - `x > y` is `TRUE` if `x` is greater than `y`
      - `x >= y` is `TRUE` if `x` is greater or equal than `y`
      - `x == y` is `TRUE` if `x` is equal to `y`
      - `is.na(x)` is `TRUE` if `x` is `NA`. **Warning:** never use `x == NA` to test if `x` is `NA`.
      - `x %in% c('a', 'b', 'c')` is `TRUE` if `x` is in the vector `c('a', 'b', 'c')`.
      - `!x` is `TRUE` if `x` is `FALSE` and viceversa.
  - ✓ Example: filter(surveys, year == 1995)

dplyr: Filter( )

- Filter( )
  - ✓ We keep only observations with arrival delay greater than 10 hours:

```
# We keep only observations with arrival delay greater than 10 hours:
delayed <- hflights2 %>% filter(ArrDelay > 600)
View(delayed)
```

Filter

	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	AirTime	ArrDelay	DepDelay
1	2011	1	20	4	635	807	CO	59	N74856	152	126	775	780
2	2011	5	20	5	858	1027	MQ	3328	N609MQ	89	55	822	803
3	2011	6	22	3	908	1040	CO	595	N75861	212	177	766	758
4	2011	6	21	2	2334	124	UA	855	N670UA	230	216	861	869
5	2011	6	9	4	2029	2243	MQ	3859	N6EAMQ	134	117	793	814
6	2011	8	1	1	156	452	CO	1	N69063	476	461	957	981
7	2011	10	25	2	2310	149	DL	1215	N764NC	99	92	701	730
8	2011	11	19	6	1752	1910	AA	1903	N495AA	78	40	685	677
9	2011	11	8	2	721	948	MQ	3786	N502MQ	147	120	918	931
10	2011	12	12	1	650	808	AA	1740	N473AA	78	49	978	970
11	2011	12	22	4	1728	1848	AA	1903	N580AA	80	40	663	653
12	2011	12	13	2	706	824	MQ	3328	N651MQ	78	56	704	691
13	2011	12	29	4	1928	2114	XE	4309	N16170	166	150	634	628

dplyr: Filter( )

- Filter( )
  - ✓ All flights flown by one of AA, FL, or XE

```
# Filter: All flights flown by one of AA, FL, or XE:
filter2 <- hflights2 %>% filter(UniqueCarrier %in% c("AA", "FL", "XE"))
table(UniqueCarrier)
```

```
> table(filter2$UniqueCarrier)

   AA    FL    XE
 3244  2139 73053
```

dplyr: Filter( )

- Filter( )
  - ✓ All flights where taxiing took longer than flying

```
# All flights where taxiing took longer than flying
filter3 <- hflights2 %>% filter(TaxiIn + TaxiOut > AirTime)
filter3[,c("TaxiIn", "TaxiOut", "AirTime")]
```

```
> filter3[,c("TaxiIn", "TaxiOut", "AirTime")]
# A tibble: 1,389 x 3
   TaxiIn TaxiOut AirTime
    <int>   <int>   <int>
 1     14      37      42
 2     10      40      43
 3     10      35      43
 4     27      20      45
 5      5      23      27
 6      7      25      30
 7      5      30      30
 8      5      29      32
 9      6      27      31
10     10      34      40
# ... with 1,379 more rows
```

dplyr: Filter( )

- Filter( )
  - ✓ Combining tests using boolean operators

```
# Combining tests using boolean operators
# all flights that departed before 5am or arrived after 10pm.
filter4 <- filter(hflights2, DepTime < 500 | ArrTime > 2200)
filter4[,1:7]
```

```
> filter4[,1:7]
# A tibble: 27,799 x 7
    Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier
   <int> <int>      <int>     <int>   <int>   <int> <chr>
 1  2011     1          4         2    2100    2207 AA
 2  2011     1         14         5    2119    2229 AA
 3  2011     1         10         1    1934    2235 AA
 4  2011     1         26         3    1905    2211 AA
 5  2011     1         30         7    1856    2209 AA
 6  2011     1          9         7    1938    2228 AS
 7  2011     1         31         1    1919    2231 CO
 8  2011     1         31         1    2116    2344 CO
 9  2011     1         31         1    1850    2211 CO
10  2011     1         31         1    2102    2216 CO
# ... with 27,789 more rows
```

dplyr: Filter( )

- Filter( )
  - ✓ Combining tests using boolean operators

```
# all flights that departed late but arrived ahead of schedule
filter5 <- filter(hflights2, DepDelay > 0, ArrDelay < 0)
filter5[,11:15]
```

```
> filter5[,11:15]
# A tibble: 27,712 x 5
   AirTime ArrDelay DepDelay Origin Dest
     <int>    <int>    <int> <chr>  <chr>
 1      45       -9        1 IAH    DFW
 2      44       -3        5 IAH    DFW
 3      42       -2        8 IAH    DFW
 4      46       -8        1 IAH    DFW
 5      39       -7       10 IAH    DFW
 6      44       -4       15 IAH    DFW
 7      39      -17        4 IAH    DFW
 8      37       -9        1 IAH    DFW
 9      41       -5        9 IAH    DFW
10      44       -6        1 IAH    DFW
# ... with 27,702 more rows
```

dplyr: Filter( )

- Filter( )
  - ✓ Combining tests using boolean operators

```
# all cancelled weekend flights
filter6 <- filter(hflights2, DayOfWeek %in% c(6,7), Cancelled == 1)
filter6[,c(1:4,18:21)]
```

```
> filter6[,c(1:4,18:21)]
# A tibble: 585 x 8
    Year Month DayofMonth DayOfWeek TaxiOut Cancelled CancellationCode Diverted
   <int> <int>      <int>     <int>   <int>     <int> <chr>                <int>
 1  2011     1          9         7      NA         1 B                        0
 2  2011     1         29         6      NA         1 A                        0
 3  2011     1          9         7      NA         1 B                        0
 4  2011     1          9         7      NA         1 B                        0
 5  2011     1          9         7      NA         1 B                        0
 6  2011     1          2         7      NA         1 A                        0
 7  2011     1         29         6      NA         1 A                        0
 8  2011     1          9         7      NA         1 A                        0
 9  2011     1          1         6      NA         1 A                        0
10  2011     1          9         7      NA         1 B                        0
# ... with 575 more rows
```

dplyr: Filter( )

- Filter( )
  - ✓ Combining tests using boolean operators

```
# all flights that were cancelled after being delayed
filter7 <- filter(hflights2, Cancelled == 1, DepDelay > 0)
filter7[,c(1:4,13,19)]
```

```
> filter7[,c(1:4,13,19)]
# A tibble: 40 x 6
    Year Month DayofMonth DayOfWeek DepDelay Cancelled
   <int> <int>      <int>     <int>    <int>     <int>
 1  2011     1         26         3       26         1
 2  2011     1         11         2      135         1
 3  2011     1         19         3        6         1
 4  2011     1          7         5       73         1
 5  2011     2          4         5        8         1
 6  2011     2          8         2      187         1
 7  2011     2          2         3        2         1
 8  2011     2          9         3        4         1
 9  2011     2          1         2       28         1
10  2011     3         31         4      156         1
# ... with 30 more rows
```


ANY
questions?