Manipulación de datos - dplyr

Josman

10 de octubre de 2014

```
# load packages
suppressMessages(library(dplyr))
library(hflights)

# explore data
data(hflights)
head(hflights)
```

```
##      Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier
## 5424 2011     1          1         6    1400    1500            AA
## 5425 2011     1          2         7    1401    1501            AA
## 5426 2011     1          3         1    1352    1502            AA
## 5427 2011     1          4         2    1403    1513            AA
## 5428 2011     1          5         3    1405    1507            AA
## 5429 2011     1          6         4    1359    1503            AA
##      FlightNum TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin
## 5424       428  N576AA                60      40      -10        0    IAH
## 5425       428  N557AA                60      45       -9        1    IAH
## 5426       428  N541AA                70      48       -8       -8    IAH
## 5427       428  N403AA                70      39        3        3    IAH
## 5428       428  N492AA                62      44       -3        5    IAH
## 5429       428  N262AA                64      45       -7       -1    IAH
##      Dest Distance TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 5424  DFW      224      7      13         0                          0
## 5425  DFW      224      6       9         0                          0
## 5426  DFW      224      5      17         0                          0
## 5427  DFW      224      9      22         0                          0
## 5428  DFW      224      9       9         0                          0
## 5429  DFW      224      6      13         0                          0
```

Convertimos los datos en un local data frame. Esto hace que nuestros datos se impriman de manera más amigable.

```
# convert to local data frame
flights <- tbl_df(hflights)

# printing only shows 10 rows and as many columns as can fit on your screen
flights
```

```
## Source: local data frame [227,496 x 21]
##
##      Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier
## 5424 2011     1          1         6    1400    1500            AA
## 5425 2011     1          2         7    1401    1501            AA
## 5426 2011     1          3         1    1352    1502            AA
## 5427 2011     1          4         2    1403    1513            AA
## 5428 2011     1          5         3    1405    1507            AA
```

```
## 5429 2011     1          6         4    1359    1503            AA
## 5430 2011     1          7         5    1359    1509            AA
## 5431 2011     1          8         6    1355    1454            AA
## 5432 2011     1          9         7    1443    1554            AA
## 5433 2011     1         10         1    1443    1553            AA
## ..    ...   ...        ...       ...     ...     ...           ...
## Variables not shown: FlightNum (int), TailNum (chr), ActualElapsedTime
##   (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest
##   (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int),
##   CancellationCode (chr), Diverted (int)
```

```
# you can specify that you want to see more rows
print(flights, n=20)
```

```
## Source: local data frame [227,496 x 21]
##
##      Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier
## 5424 2011     1          1         6    1400    1500            AA
## 5425 2011     1          2         7    1401    1501            AA
## 5426 2011     1          3         1    1352    1502            AA
## 5427 2011     1          4         2    1403    1513            AA
## 5428 2011     1          5         3    1405    1507            AA
## 5429 2011     1          6         4    1359    1503            AA
## 5430 2011     1          7         5    1359    1509            AA
## 5431 2011     1          8         6    1355    1454            AA
## 5432 2011     1          9         7    1443    1554            AA
## 5433 2011     1         10         1    1443    1553            AA
## 5434 2011     1         11         2    1429    1539            AA
## 5435 2011     1         12         3    1419    1515            AA
## 5436 2011     1         13         4    1358    1501            AA
## 5437 2011     1         14         5    1357    1504            AA
## 5438 2011     1         15         6    1359    1459            AA
## 5439 2011     1         16         7    1359    1509            AA
## 5440 2011     1         17         1    1530    1634            AA
## 5441 2011     1         18         2    1408    1508            AA
## 5442 2011     1         19         3    1356    1503            AA
## 5443 2011     1         20         4    1507    1622            AA
## ..    ...   ...        ...       ...     ...     ...           ...
## Variables not shown: FlightNum (int), TailNum (chr), ActualElapsedTime
##   (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest
##   (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int),
##   CancellationCode (chr), Diverted (int)
```

```
# convert to a normal data frame to see all of the columns
data.frame(head(flights))
```

```
##      Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier
## 5424 2011     1          1         6    1400    1500            AA
## 5425 2011     1          2         7    1401    1501            AA
## 5426 2011     1          3         1    1352    1502            AA
## 5427 2011     1          4         2    1403    1513            AA
## 5428 2011     1          5         3    1405    1507            AA
## 5429 2011     1          6         4    1359    1503            AA
```

```
##      FlightNum TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin
## 5424       428  N576AA                60      40      -10        0    IAH
## 5425       428  N557AA                60      45       -9        1    IAH
## 5426       428  N541AA                70      48       -8       -8    IAH
## 5427       428  N403AA                70      39        3        3    IAH
## 5428       428  N492AA                62      44       -3        5    IAH
## 5429       428  N262AA                64      45       -7       -1    IAH
##      Dest Distance TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 5424  DFW      224      7      13         0                          0
## 5425  DFW      224      6       9         0                          0
## 5426  DFW      224      5      17         0                          0
## 5427  DFW      224      9      22         0                          0
## 5428  DFW      224      9       9         0                          0
## 5429  DFW      224      6      13         0                          0
```

filter: Inidicamos un criterio para seleccionar renglones

```
# base R approach to view all flights on January 1
flights[flights$Month==1 & flights$DayofMonth==1, ]
```

```
# dplyr approach
# note: you can use comma or ampersand to represent AND condition
filter(flights, Month==1, DayofMonth==1)
```

```
## Source: local data frame [552 x 21]
##
##    Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1  2011     1          1         6    1400    1500            AA       428
## 2  2011     1          1         6     728     840            AA       460
## 3  2011     1          1         6    1631    1736            AA      1121
## 4  2011     1          1         6    1756    2112            AA      1294
## 5  2011     1          1         6    1012    1347            AA      1700
## 6  2011     1          1         6    1211    1325            AA      1820
## 7  2011     1          1         6     557     906            AA      1994
## 8  2011     1          1         6    1824    2106            AS       731
## 9  2011     1          1         6     654    1124            B6       620
## 10 2011     1          1         6    1639    2110            B6       622
## ..  ...   ...        ...       ...     ...     ...           ...       ...
## Variables not shown: TailNum (chr), ActualElapsedTime (int), AirTime
##   (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr),
##   Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int),
##   CancellationCode (chr), Diverted (int)
```

```
# use pipe for OR condition
filter(flights, UniqueCarrier=="AA" | UniqueCarrier=="UA")
```

```
## Source: local data frame [5,316 x 21]
##
##    Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1  2011     1          1         6    1400    1500            AA       428
```

```
## 2  2011     1          2         7    1401    1501            AA       428
## 3  2011     1          3         1    1352    1502            AA       428
## 4  2011     1          4         2    1403    1513            AA       428
## 5  2011     1          5         3    1405    1507            AA       428
## 6  2011     1          6         4    1359    1503            AA       428
## 7  2011     1          7         5    1359    1509            AA       428
## 8  2011     1          8         6    1355    1454            AA       428
## 9  2011     1          9         7    1443    1554            AA       428
## 10 2011     1         10         1    1443    1553            AA       428
## ..  ...   ...        ...       ...     ...     ...           ...       ...
## Variables not shown: TailNum (chr), ActualElapsedTime (int), AirTime
##   (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr),
##   Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int),
##   CancellationCode (chr), Diverted (int)
```

```
# you can also use %in% operator
filter(flights, UniqueCarrier %in% c("AA", "UA"))
```

```
## Source: local data frame [5,316 x 21]
##
##    Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1  2011     1          1         6    1400    1500            AA       428
## 2  2011     1          2         7    1401    1501            AA       428
## 3  2011     1          3         1    1352    1502            AA       428
## 4  2011     1          4         2    1403    1513            AA       428
## 5  2011     1          5         3    1405    1507            AA       428
## 6  2011     1          6         4    1359    1503            AA       428
## 7  2011     1          7         5    1359    1509            AA       428
## 8  2011     1          8         6    1355    1454            AA       428
## 9  2011     1          9         7    1443    1554            AA       428
## 10 2011     1         10         1    1443    1553            AA       428
## ..  ...   ...        ...       ...     ...     ...           ...       ...
## Variables not shown: TailNum (chr), ActualElapsedTime (int), AirTime
##   (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr),
##   Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int),
##   CancellationCode (chr), Diverted (int)
```

select: Seleccionamos columnas por su nombre

```
# base R approach to select DepTime, ArrTime, and FlightNum columns
flights[, c("DepTime", "ArrTime", "FlightNum")]
```

```
# dplyr approach
select(flights, DepTime, ArrTime, FlightNum)
```

```
## Source: local data frame [227,496 x 3]
##
##      DepTime ArrTime FlightNum
## 5424    1400    1500       428
## 5425    1401    1501       428
```

```
## 5426    1352    1502       428
## 5427    1403    1513       428
## 5428    1405    1507       428
## 5429    1359    1503       428
## 5430    1359    1509       428
## 5431    1355    1454       428
## 5432    1443    1554       428
## 5433    1443    1553       428
## ..       ...     ...       ...
```

```
# use colon to select multiple contiguous columns, and use `contains` to match columns by name
# note: `starts_with`, `ends_with`, and `matches` (for regular expressions) can also be used to match c
select(flights, Year:DayofMonth, contains("Taxi"), contains("Delay"))
```

```
## Source: local data frame [227,496 x 7]
##
##      Year Month DayofMonth TaxiIn TaxiOut ArrDelay DepDelay
## 5424 2011     1          1      7      13      -10        0
## 5425 2011     1          2      6       9       -9        1
## 5426 2011     1          3      5      17       -8       -8
## 5427 2011     1          4      9      22        3        3
## 5428 2011     1          5      9       9       -3        5
## 5429 2011     1          6      6      13       -7       -1
## 5430 2011     1          7     12      15       -1       -1
## 5431 2011     1          8      7      12      -16       -5
## 5432 2011     1          9      8      22       44       43
## 5433 2011     1         10      6      19       43       43
## ..    ...   ...        ...    ...     ...      ...      ...
```

Encadenando

```
# nesting method to select UniqueCarrier and DepDelay columns and filter for delays over 60 minutes
filter(select(flights, UniqueCarrier, DepDelay), DepDelay > 60)
```

```
# chaining method
flights %>%
    select(UniqueCarrier, DepDelay) %>%
    filter(DepDelay > 60)
```

```
## Source: local data frame [10,242 x 2]
##
##    UniqueCarrier DepDelay
## 1             AA       90
## 2             AA       67
## 3             AA       74
## 4             AA      125
## 5             AA       82
## 6             AA       99
## 7             AA       70
## 8             AA       61
```

```
## 9             AA       74
## 10            AS       73
## ..           ...      ...
```

arrange: reodenar renglones

```
# base R approach to select UniqueCarrier and DepDelay columns and sort by DepDelay
flights[order(flights$DepDelay), c("UniqueCarrier", "DepDelay")]
```

```
# dplyr approach
flights %>%
    select(UniqueCarrier, DepDelay) %>%
    arrange(DepDelay)
```

```
## Source: local data frame [227,496 x 2]
##
##    UniqueCarrier DepDelay
## 1             OO      -33
## 2             MQ      -23
## 3             XE      -19
## 4             XE      -19
## 5             CO      -18
## 6             EV      -18
## 7             XE      -17
## 8             CO      -17
## 9             XE      -17
## 10            MQ      -17
## ..           ...      ...
```

```
# use `desc` for descending
flights %>%
    select(UniqueCarrier, DepDelay) %>%
    arrange(desc(DepDelay))
```

```
## Source: local data frame [227,496 x 2]
##
##    UniqueCarrier DepDelay
## 1             CO      981
## 2             AA      970
## 3             MQ      931
## 4             UA      869
## 5             MQ      814
## 6             MQ      803
## 7             CO      780
## 8             CO      758
## 9             DL      730
## 10            MQ      691
## ..           ...      ...
```

mutate: Agrega nuevas variables

```
# base R approach to create a new variable Speed (in mph)
flights$Speed <- flights$Distance / flights$AirTime*60
flights[, c("Distance", "AirTime", "Speed")]
```

```
# dplyr approach (prints the new variable but does not store it)
flights %>%
    select(Distance, AirTime) %>%
    mutate(Speed = Distance/AirTime*60)
```

```
## Source: local data frame [227,496 x 3]
##
##    Distance AirTime Speed
## 1       224      40 336.0
## 2       224      45 298.7
## 3       224      48 280.0
## 4       224      39 344.6
## 5       224      44 305.5
## 6       224      45 298.7
## 7       224      43 312.6
## 8       224      40 336.0
## 9       224      41 327.8
## 10      224      45 298.7
## ..      ...     ...   ...
```

```
# store the new variable
flights <- flights %>% mutate(Speed = Distance/AirTime*60)
```

summarise: reducir variables a valores

```
# base R approaches to calculate the average arrival delay to each destination
head(with(flights, tapply(ArrDelay, Dest, mean, na.rm=TRUE)))
```

```
##    ABQ    AEX    AGS    AMA    ANC    ASE
##  7.226  5.839  4.000  6.840 26.081  6.795
```

```
head(aggregate(ArrDelay ~ Dest, flights, mean))
```

```
##   Dest ArrDelay
## 1  ABQ    7.226
## 2  AEX    5.839
## 3  AGS    4.000
## 4  AMA    6.840
## 5  ANC   26.081
## 6  ASE    6.795
```

```
# dplyr approach: create a table grouped by Dest, and then summarise each group by taking the mean of A
flights %>%
    group_by(Dest) %>%
    summarise(avg_delay = mean(ArrDelay, na.rm=TRUE))
```

```
## Source: local data frame [116 x 2]
##
##    Dest avg_delay
## 1   ABQ     7.226
## 2   AEX     5.839
## 3   AGS     4.000
## 4   AMA     6.840
## 5   ANC    26.081
## 6   ASE     6.795
## 7   ATL     8.233
## 8   AUS     7.449
## 9   AVL     9.974
## 10  BFL   -13.199
## ..  ...       ...
```

```
# for each carrier, calculate the percentage of flights cancelled or diverted
flights %>%
    group_by(UniqueCarrier) %>%
    summarise_each(funs(mean), Cancelled, Diverted)
```

```
## Source: local data frame [15 x 3]
##
##    UniqueCarrier Cancelled Diverted
## 1             AA  0.018496 0.001850
## 2             AS  0.000000 0.002740
## 3             B6  0.025899 0.005755
## 4             CO  0.006783 0.002627
## 5             DL  0.015903 0.003029
## 6             EV  0.034483 0.003176
## 7             F9  0.007160 0.000000
## 8             FL  0.009818 0.003273
## 9             MQ  0.029045 0.001936
## 10            OO  0.013947 0.003487
## 11            UA  0.016409 0.002413
## 12            US  0.011269 0.001470
## 13            WN  0.015504 0.002294
## 14            XE  0.015496 0.003450
## 15            YV  0.012658 0.000000
```

```
# for each carrier, calculate the minimum and maximum arrival and departure delays
flights %>%
    group_by(UniqueCarrier) %>%
    summarise_each(funs(min(., na.rm=TRUE), max(., na.rm=TRUE)), matches("Delay"))
```

```
## Source: local data frame [15 x 5]
##
##    UniqueCarrier ArrDelay_min DepDelay_min ArrDelay_max DepDelay_max
```

```
## 1             AA          -39          -15          978          970
## 2             AS          -43          -15          183          172
## 3             B6          -44          -14          335          310
## 4             CO          -55          -18          957          981
## 5             DL          -32          -17          701          730
## 6             EV          -40          -18          469          479
## 7             F9          -24          -15          277          275
## 8             FL          -30          -14          500          507
## 9             MQ          -38          -23          918          931
## 10            OO          -57          -33          380          360
## 11            UA          -47          -11          861          869
## 12            US          -42          -17          433          425
## 13            WN          -44          -10          499          548
## 14            XE          -70          -19          634          628
## 15            YV          -32          -11           72           54
```

```
# for each day of the year, count the total number of flights and sort in descending order
flights %>%
    group_by(Month, DayofMonth) %>%
    summarise(flight_count = n()) %>%
    arrange(desc(flight_count))
```

```
## Source: local data frame [365 x 3]
## Groups: Month
##
##    Month DayofMonth flight_count
## 1      8          4          706
## 2      8         11          706
## 3      8         12          706
## 4      8          5          705
## 5      8          3          704
## 6      8         10          704
## 7      1          3          702
## 8      7          7          702
## 9      7         14          702
## 10     7         28          701
## ..   ...        ...          ...
```

```
# rewrite more simply with the `tally` function
flights %>%
    group_by(Month, DayofMonth) %>%
    tally(sort = TRUE)
```

```
## Source: local data frame [365 x 3]
## Groups: Month
##
##    Month DayofMonth   n
## 1      8          4 706
## 2      8         11 706
## 3      8         12 706
## 4      8          5 705
## 5      8          3 704
## 6      8         10 704
```

```
## 7      1          3 702
## 8      7          7 702
## 9      7         14 702
## 10     7         28 701
## ..   ...        ... ...
```

```
# for each destination, count the total number of flights and the number of distinct planes that flew t
flights %>%
    group_by(Dest) %>%
    summarise(flight_count = n(), plane_count = n_distinct(TailNum))
```

```
## Source: local data frame [116 x 3]
##
##    Dest flight_count plane_count
## 1   ABQ         2812         716
## 2   AEX          724         215
## 3   AGS            1           1
## 4   AMA         1297         158
## 5   ANC          125          38
## 6   ASE          125          60
## 7   ATL         7886         983
## 8   AUS         5022        1015
## 9   AVL          350         142
## 10  BFL          504          70
## ..  ...          ...         ...
```

```
# for each destination, show the number of cancelled and not cancelled flights
flights %>%
    group_by(Dest) %>%
    select(Cancelled) %>%
    table() %>%
    head()
```

```
##      Cancelled
## Dest     0  1
##   ABQ 2787 25
##   AEX  712 12
##   AGS    1  0
##   AMA 1265 32
##   ANC  125  0
##   ASE  120  5
```

Funciones de ventana

```
# for each carrier, calculate which two days of the year they had their longest departure delays
# note: smallest (not largest) value is ranked as 1, so you have to use `desc` to rank by largest value
flights %>%
    group_by(UniqueCarrier) %>%
    select(Month, DayofMonth, DepDelay) %>%
    filter(min_rank(desc(DepDelay)) <= 2) %>%
    arrange(UniqueCarrier, desc(DepDelay))
```

```
# rewrite more simply with the `top_n` function
flights %>%
    group_by(UniqueCarrier) %>%
    select(Month, DayofMonth, DepDelay) %>%
    top_n(2) %>%
    arrange(UniqueCarrier, desc(DepDelay))
```

```
## Selecting by DepDelay
```

```
## Source: local data frame [30 x 4]
## Groups: UniqueCarrier
##
##    UniqueCarrier Month DayofMonth DepDelay
## 1             AA    12         12      970
## 2             AA    11         19      677
## 3             AS     2         28      172
## 4             AS     7          6      138
## 5             B6    10         29      310
## 6             B6     8         19      283
## 7             CO     8          1      981
## 8             CO     1         20      780
## 9             DL    10         25      730
## 10            DL     4          5      497
## 11            EV     6         25      479
## 12            EV     1          5      465
## 13            F9     5         12      275
## 14            F9     5         20      240
## 15            FL     2         19      507
## 16            FL     3         14      493
## 17            MQ    11          8      931
## 18            MQ     6          9      814
## 19            OO     2         27      360
## 20            OO     4          4      343
## 21            UA     6         21      869
## 22            UA     9         18      588
## 23            US     4         19      425
## 24            US     8         26      277
## 25            WN     4          8      548
## 26            WN     9         29      503
## 27            XE    12         29      628
## 28            XE    12         29      511
## 29            YV     4         22       54
## 30            YV     4         30       46
```

```
# for each month, calculate the number of flights and the change from the previous month
flights %>%
    group_by(Month) %>%
    summarise(flight_count = n()) %>%
    mutate(change = flight_count - lag(flight_count))
```

```
## Source: local data frame [12 x 3]
##
```

```
##    Month flight_count change
## 1      1        18910     NA
## 2      2        17128  -1782
## 3      3        19470   2342
## 4      4        18593   -877
## 5      5        19172    579
## 6      6        19600    428
## 7      7        20548    948
## 8      8        20176   -372
## 9      9        18065  -2111
## 10    10        18696    631
## 11    11        18021   -675
## 12    12        19117   1096
```

```
# rewrite more simply with the `tally` function
flights %>%
    group_by(Month) %>%
    tally() %>%
    mutate(change = n - lag(n))
```

```
## Source: local data frame [12 x 3]
##
##    Month     n change
## 1      1 18910     NA
## 2      2 17128  -1782
## 3      3 19470   2342
## 4      4 18593   -877
## 5      5 19172    579
## 6      6 19600    428
## 7      7 20548    948
## 8      8 20176   -372
## 9      9 18065  -2111
## 10    10 18696    631
## 11    11 18021   -675
## 12    12 19117   1096
```

Otras funciones muy útiles

```
# randomly sample a fixed number of rows, without replacement
flights %>% sample_n(5)
```

```
## Source: local data frame [5 x 22]
##
##        Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier
## 102369 2011     6         26         7    1241    1434            WN
## 170194 2011     9          8         4    2146      38            OO
## 112284 2011     6          4         6    1448    1738            FL
## 115492 2011     7         20         3    1718    2051            CO
## 209829 2011    12         26         1    1047    1241            CO
## Variables not shown: FlightNum (int), TailNum (chr), ActualElapsedTime
```

```
##   (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest
##   (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int),
##   CancellationCode (chr), Diverted (int), Speed (dbl)
```

```
# randomly sample a fraction of rows, with replacement
flights %>% sample_frac(0.25, replace=TRUE)
```

```
## Source: local data frame [56,874 x 22]
##
##        Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier
## 187730 2011    10         14         5     734     821            OO
## 71599  2011     4         23         6     632     832            OO
## 14169  2011     1         22         6    1625    1928            XE
## 187970 2011    10         13         4    1533    1642            OO
## 121932 2011     7         25         1    2233    2322            WN
## 199317 2011    11         29         2     728     817            WN
## 145894 2011     8         12         5     601     720            UA
## 126398 2011     7         19         2    1904    2003            XE
## 153473 2011     8          1         1    1706    1824            CO
## 21776  2011     2         16         3    1731    1852            CO
## ..      ...   ...        ...       ...     ...     ...           ...
## Variables not shown: FlightNum (int), TailNum (chr), ActualElapsedTime
##   (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest
##   (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int),
##   CancellationCode (chr), Diverted (int), Speed (dbl)
```

```
# base R approach to view the structure of an object
str(flights)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    227496 obs. of  22 variables:
##  $ Year             : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ Month            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ DayofMonth       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ DayOfWeek        : int  6 7 1 2 3 4 5 6 7 1 ...
##  $ DepTime          : int  1400 1401 1352 1403 1405 1359 1359 1355 1443 1443 ...
##  $ ArrTime          : int  1500 1501 1502 1513 1507 1503 1509 1454 1554 1553 ...
##  $ UniqueCarrier    : chr  "AA" "AA" "AA" "AA" ...
##  $ FlightNum        : int  428 428 428 428 428 428 428 428 428 428 ...
##  $ TailNum          : chr  "N576AA" "N557AA" "N541AA" "N403AA" ...
##  $ ActualElapsedTime: int  60 60 70 70 62 64 70 59 71 70 ...
##  $ AirTime          : int  40 45 48 39 44 45 43 40 41 45 ...
##  $ ArrDelay         : int  -10 -9 -8 3 -3 -7 -1 -16 44 43 ...
##  $ DepDelay         : int  0 1 -8 3 5 -1 -1 -5 43 43 ...
##  $ Origin           : chr  "IAH" "IAH" "IAH" "IAH" ...
##  $ Dest             : chr  "DFW" "DFW" "DFW" "DFW" ...
##  $ Distance         : int  224 224 224 224 224 224 224 224 224 224 ...
##  $ TaxiIn           : int  7 6 5 9 9 6 12 7 8 6 ...
##  $ TaxiOut          : int  13 9 17 22 9 13 15 12 22 19 ...
##  $ Cancelled        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CancellationCode : chr  "" "" "" "" ...
##  $ Diverted         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Speed            : num  336 299 280 345 305 ...
```

```r
# dplyr approach: better formatting, and adapts to your screen width
glimpse(flights)
```

```
## Variables:
## $ Year              (int) 2011, 2011, 2011, 2011, 2011, 2011, 2011, 20...
## $ Month             (int) 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, 1...
## $ DayOfWeek         (int) 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6,...
## $ DepTime           (int) 1400, 1401, 1352, 1403, 1405, 1359, 1359, 13...
## $ ArrTime           (int) 1500, 1501, 1502, 1513, 1507, 1503, 1509, 14...
## $ UniqueCarrier     (chr) "AA", "AA", "AA", "AA", "AA", "AA", "AA", "A...
## $ FlightNum         (int) 428, 428, 428, 428, 428, 428, 428, 428, 428,...
## $ TailNum           (chr) "N576AA", "N557AA", "N541AA", "N403AA", "N49...
## $ ActualElapsedTime (int) 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, ...
## $ AirTime           (int) 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, ...
## $ ArrDelay          (int) -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29,...
## $ DepDelay          (int) 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, ...
## $ Origin            (chr) "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "I...
## $ Dest              (chr) "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", "D...
## $ Distance          (int) 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...
## $ TaxiOut           (int) 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11...
## $ Cancelled         (int) 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,...
## $ Speed             (dbl) 336.0, 298.7, 280.0, 344.6, 305.5, 298.7, 31...
```