Data Manipulation in R with dplyr

Amar Kapote

2017-07-19

Contents

Whats Covered . . . 2
Aditional Resources . . . 2

Introduction to dplyr and tbls **3**

Section 1 - Introduction to dplyr . . . 3
Load the dplyr and hflights package . . . 3
Explore the dataset . . . 4
Section 2 - tbl, a special type of data.frame . . . 5
Convert data.frame to tibble . . . 5
Changing labels of hflight, part 1 of 2 . . . 6
Changing labels of hflights, part 2 or 2 . . . 7

Select and mutate **8**

Section 3 - The five verbs and select in more detail . . . 8
Choosing is not losing! The select verb . . . 8
Helper functions for variable selection . . . 10
Comparison to base R . . . 11
Section 4 - The second of five verbs: mutate . . . 11
Mutating is creating . . . 11
Add multiple variables using mutate . . . 12

Filter and arrange **14**

Section 5 - The third of five verbs: filter . . . 14
Logical operators . . . 14
Combining tests using boolean operators . . . 16
Blend together what you've learned! . . . 18
Recap on select, mutate and filter . . . 19
Section 6 - Almost there: the arrange verb . . . 20

Arranging your data . . . 20
Reverse the order of arranging . . . 21

Summarise and the pipe operator **23**

Section 7 - Last but not least: summarise . . . 23
The syntax of summarize . . . 23
Aggregate functions . . . 24
dplyr aggregate functions . . . 25
Section 8 - Chaining your functions: the pipe operator . . . 26
Overview of syntax . . . 26
Drive of fly? Part 1 of 2 . . . 26
Drive or fly? Part 2 of 2 . . . 27
Advanced piping exercise . . . 27

Group_by and working with databases **28**

Section 9 - get group-wise insights: group_by . . . 28
Unite and conquer using group_by . . . 28
Combine group_by with mutate . . . 29
Advanced group_by exercises . . . 30
Section 10 - dplyr and databases . . . 31
dplyr deals with different types . . . 32
dplyr and mySQL databases . . . 32
Talk with Hadley Wickham . . . 33

Whats Covered

- Introduction to dplyr and tbls
- Select and mutate
- Filter and arrange
- Summarise and the pipe operator
- Group_by and working with databases

Aditional Resources

- dplyr vignette
- Data Wrangling Cheat Sheet (dplyr and tidyr)

Introduction to dplyr and tbls

Section 1 - Introduction to dplyr

- dplyr is a grammar of data manipulation
- it provides a consistent set of vebs that help you solve the most common data manipulation challenges
- `mutate`, `select`, `filter`, `summarise`, `arrange`, and the joins
- dplyr vignette

Load the dplyr and hflights package

```
# Load the dplyr package
library(dplyr)

# Load the hflights package
library(hflights)

# Call both head() and summary() on hflights
head(hflights)
```

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

```
summary(hflights)
```

```
##       Year          Month          DayofMonth      DayOfWeek        DepTime
##  Min.   :2011   Min.   : 1.000   Min.   : 1.00   Min.   :1.000   Min.   :   1
##  1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:2.000   1st Qu.:1021
##  Median :2011   Median : 7.000   Median :16.00   Median :4.000   Median :1416
```

```
##  Mean   :2011   Mean   : 6.514   Mean   :15.74   Mean   :3.948   Mean   :1396
##  3rd Qu.:2011   3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:6.000   3rd Qu.:1801
##  Max.   :2011   Max.   :12.000   Max.   :31.00   Max.   :7.000   Max.   :2400
##                                                                  NA's   :2905
##     ArrTime     UniqueCarrier        FlightNum      TailNum
##  Min.   :   1   Length:227496      Min.   :   1   Length:227496
##  1st Qu.:1215   Class :character   1st Qu.: 855   Class :character
##  Median :1617   Mode  :character   Median :1696   Mode  :character
##  Mean   :1578                      Mean   :1962
##  3rd Qu.:1953                      3rd Qu.:2755
##  Max.   :2400                      Max.   :7290
##  NA's   :3066
##  ActualElapsedTime    AirTime         ArrDelay          DepDelay
##  Min.   : 34.0     Min.   : 11.0   Min.   :-70.000   Min.   :-33.000
##  1st Qu.: 77.0     1st Qu.: 58.0   1st Qu.: -8.000   1st Qu.: -3.000
##  Median :128.0     Median :107.0   Median :  0.000   Median :  0.000
##  Mean   :129.3     Mean   :108.1   Mean   :  7.094   Mean   :  9.445
##  3rd Qu.:165.0     3rd Qu.:141.0   3rd Qu.: 11.000   3rd Qu.:  9.000
##  Max.   :575.0     Max.   :549.0   Max.   :978.000   Max.   :981.000
##  NA's   :3622      NA's   :3622    NA's   :3622      NA's   :2905
##     Origin              Dest              Distance          TaxiIn
##  Length:227496      Length:227496      Min.   :  79.0   Min.   :  1.000
##  Class :character   Class :character   1st Qu.: 376.0   1st Qu.:  4.000
##  Mode  :character   Mode  :character   Median : 809.0   Median :  5.000
##                                        Mean   : 787.8   Mean   :  6.099
##                                        3rd Qu.:1042.0   3rd Qu.:  7.000
##                                        Max.   :3904.0   Max.   :165.000
##                                                         NA's   :3066
##     TaxiOut         Cancelled       CancellationCode      Diverted
##  Min.   :  1.00   Min.   :0.00000   Length:227496      Min.   :0.000000
##  1st Qu.: 10.00   1st Qu.:0.00000   Class :character   1st Qu.:0.000000
##  Median : 14.00   Median :0.00000   Mode  :character   Median :0.000000
##  Mean   : 15.09   Mean   :0.01307                      Mean   :0.002853
##  3rd Qu.: 18.00   3rd Qu.:0.00000                      3rd Qu.:0.000000
##  Max.   :163.00   Max.   :1.00000                      Max.   :1.000000
##  NA's   :2947
```

Explore the dataset

```
str(hflights)
```

```
## 'data.frame':    227496 obs. of  21 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 ...
```

Section 2 - tbl, a special type of data.frame

- tibble print adapts to the size of your window
- glimpse gives you a more complete view of the tibble
- if you don't like it go back to data.frame and use `str` and `head`

Convert data.frame to tibble

```
# Convert the hflights data.frame into a hflights tbl
hflights <- tbl_df(hflights)

# Display the hflights tbl
hflights
```

```
## # A tibble: 227,496 x 21
##     Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
##    <int> <int>      <int>     <int>   <int>   <int> <chr>              <int>
##  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
## # ... with 227,486 more rows, and 13 more variables: 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>
```

```
glimpse(hflights)
```

```
## Observations: 227,496
## Variables: 21
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 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...
## $ DayOfWeek         <int> 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, 1...
## $ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1...
## $ UniqueCarrier     <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "...
## $ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428,...
## $ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA",...
## $ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 6...
## $ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 4...
## $ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -...
## $ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "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, 8, 4, 6, 5, 6, 12,...
## $ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13,...
## $ Cancelled         <int> 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...
```

```
# Create the object carriers, containing only the UniqueCarrier variable of hflights
carriers <- hflights$UniqueCarrier
str(carriers)
```

```
##  chr [1:227496] "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" ...
```

Changing labels of hflight, part 1 of 2

```
# Both the dplyr and hflights packages are loaded into workspace
lut <- c("AA" = "American", "AS" = "Alaska", "B6" = "JetBlue", "CO" = "Continental",
         "DL" = "Delta", "OO" = "SkyWest", "UA" = "United", "US" = "US_Airways",
         "WN" = "Southwest", "EV" = "Atlantic_Southeast", "F9" = "Frontier",
         "FL" = "AirTran", "MQ" = "American_Eagle", "XE" = "ExpressJet", "YV" = "Mesa")

# Use lut to translate the UniqueCarrier column of hflights
hflights$UniqueCarrier <- lut[hflights$UniqueCarrier]

# Inspect the resulting raw values of your variables
glimpse(hflights)
```

```
## Observations: 227,496
## Variables: 21
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 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...
## $ DayOfWeek         <int> 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, 1...
## $ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428,...
## $ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA",...
```

```
## $ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 6...
## $ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 4...
## $ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -...
## $ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "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, 8, 4, 6, 5, 6, 12,...
## $ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13,...
## $ Cancelled         <int> 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...
```

Changing labels of hflights, part 2 or 2

```
## The "E" in my dataset is "" for some reason so I have changed it here
table(hflights$CancellationCode)
```

```
##
##             A      B      C      D
## 224523   1202   1652    118      1
```

```
hflights <- hflights %>%
  mutate(
    CancellationCode = ifelse(CancellationCode == "", "E", CancellationCode)
  )

table(hflights$CancellationCode)
```

```
##
##      A      B      C      D      E
##   1202   1652    118      1 224523
```

```
# Build the lookup table: lut
lut <- c("A" = "carrier",
         "B" = "weather",
         "C" = "FFA",
         "D" = "security",
         "E" = "not cancelled")

# Use the lookup table to create a vector of code labels. Assign the vector to the CancellationCode col
hflights$Code <- lut[hflights$CancellationCode]

# Inspect the resulting raw values of your variables
glimpse(hflights)
```

```
## Observations: 227,496
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 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...
## $ DayOfWeek         <int> 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, 1...
## $ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428,...
## $ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA",...
## $ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 6...
## $ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 4...
## $ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -...
## $ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "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, 8, 4, 6, 5, 6, 12,...
## $ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13,...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

Select and mutate

Section 3 - The five verbs and select in more detail

- 5 verbs
  - `select()` which returns a subset of the columns,
  - `filter()` that is able to return a subset of the rows,
  - `arrange()` that reorders the rows according to single or multiple variables,
  - `mutate()` used to add columns from existing data,
  - `summarise()` which reduces each group to a single row by calculating aggregate measures.
- verb focus
  - `select` and `mutate` manipulate variables
  - `filter` and `arrange` manipulate observations
  - `summarize` manipulates `groups` of observations

Choosing is not losing! The select verb

```
# Print out a tbl with the four columns of hflights related to delay
select(hflights, ActualElapsedTime, AirTime, ArrDelay, DepDelay)
```

```
## # A tibble: 227,496 x 4
##    ActualElapsedTime AirTime ArrDelay DepDelay
##                <int>   <int>    <int>    <int>
##  1                60      40      -10        0
##  2                60      45       -9        1
##  3                70      48       -8       -8
##  4                70      39        3        3
##  5                62      44       -3        5
##  6                64      45       -7       -1
##  7                70      43       -1       -1
##  8                59      40      -16       -5
##  9                71      41       44       43
## 10                70      45       43       43
## # ... with 227,486 more rows
```

```
# Print out hflights, nothing has changed!
hflights
```

```
## # A tibble: 227,496 x 22
##     Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
##    <int> <int>      <int>     <int>   <int>   <int> <chr>             <int>
##  1  2011     1          1         6    1400    1500 American            428
##  2  2011     1          2         7    1401    1501 American            428
##  3  2011     1          3         1    1352    1502 American            428
##  4  2011     1          4         2    1403    1513 American            428
##  5  2011     1          5         3    1405    1507 American            428
##  6  2011     1          6         4    1359    1503 American            428
##  7  2011     1          7         5    1359    1509 American            428
##  8  2011     1          8         6    1355    1454 American            428
##  9  2011     1          9         7    1443    1554 American            428
## 10  2011     1         10         1    1443    1553 American            428
## # ... with 227,486 more rows, and 14 more variables: 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>, Code <chr>
```

```
# Print out the columns Origin up to Cancelled of hflights
select(hflights, Origin:Cancelled)
```

```
## # A tibble: 227,496 x 6
##    Origin Dest  Distance TaxiIn TaxiOut Cancelled
##    <chr>  <chr>    <int>  <int>   <int>     <int>
##  1 IAH    DFW        224      7      13         0
##  2 IAH    DFW        224      6       9         0
##  3 IAH    DFW        224      5      17         0
##  4 IAH    DFW        224      9      22         0
##  5 IAH    DFW        224      9       9         0
##  6 IAH    DFW        224      6      13         0
##  7 IAH    DFW        224     12      15         0
##  8 IAH    DFW        224      7      12         0
##  9 IAH    DFW        224      8      22         0
## 10 IAH    DFW        224      6      19         0
## # ... with 227,486 more rows
```

```
# Answer to last question: be concise!
select(hflights, 1:4, 12:21)
```

```
## # A tibble: 227,496 x 14
##     Year Month DayofMonth DayOfWeek ArrDelay DepDelay Origin Dest  Distance
##    <int> <int>      <int>     <int>    <int>    <int> <chr>  <chr>    <int>
##  1  2011     1          1         6      -10        0 IAH    DFW        224
##  2  2011     1          2         7       -9        1 IAH    DFW        224
##  3  2011     1          3         1       -8       -8 IAH    DFW        224
##  4  2011     1          4         2        3        3 IAH    DFW        224
##  5  2011     1          5         3       -3        5 IAH    DFW        224
##  6  2011     1          6         4       -7       -1 IAH    DFW        224
##  7  2011     1          7         5       -1       -1 IAH    DFW        224
##  8  2011     1          8         6      -16       -5 IAH    DFW        224
##  9  2011     1          9         7       44       43 IAH    DFW        224
## 10  2011     1         10         1       43       43 IAH    DFW        224
## # ... with 227,486 more rows, and 5 more variables: TaxiIn <int>,
## #   TaxiOut <int>, Cancelled <int>, CancellationCode <chr>, Diverted <int>
```

Helper functions for variable selection

```
# Print out a tbl containing just ArrDelay and DepDelay
select(hflights, ends_with(c('Delay')))
```

```
## # A tibble: 227,496 x 2
##    ArrDelay DepDelay
##       <int>    <int>
##  1      -10        0
##  2       -9        1
##  3       -8       -8
##  4        3        3
##  5       -3        5
##  6       -7       -1
##  7       -1       -1
##  8      -16       -5
##  9       44       43
## 10       43       43
## # ... with 227,486 more rows
```

```
# Print out a tbl as described in the second instruction, using both helper functions and variable name
select(hflights, UniqueCarrier, ends_with(c('Num')), starts_with(c('Cancel')))
```

```
## # A tibble: 227,496 x 5
##    UniqueCarrier FlightNum TailNum Cancelled CancellationCode
##    <chr>             <int> <chr>       <int> <chr>
##  1 American            428 N576AA          0 E
##  2 American            428 N557AA          0 E
##  3 American            428 N541AA          0 E
##  4 American            428 N403AA          0 E
##  5 American            428 N492AA          0 E
```

```
##  6 American            428 N262AA          0 E
##  7 American            428 N493AA          0 E
##  8 American            428 N477AA          0 E
##  9 American            428 N476AA          0 E
## 10 American            428 N504AA          0 E
## # ... with 227,486 more rows
```

```
# Print out a tbl as described in the third instruction, using only helper functions.
select(hflights, contains(c('Time')), contains(c('Delay')))
```

```
## # A tibble: 227,496 x 6
##    DepTime ArrTime ActualElapsedTime AirTime ArrDelay DepDelay
##      <int>   <int>             <int>   <int>    <int>    <int>
##  1    1400    1500                60      40      -10        0
##  2    1401    1501                60      45       -9        1
##  3    1352    1502                70      48       -8       -8
##  4    1403    1513                70      39        3        3
##  5    1405    1507                62      44       -3        5
##  6    1359    1503                64      45       -7       -1
##  7    1359    1509                70      43       -1       -1
##  8    1355    1454                59      40      -16       -5
##  9    1443    1554                71      41       44       43
## 10    1443    1553                70      45       43       43
## # ... with 227,486 more rows
```

Comparison to base R

```
ex1r <- hflights[c("TaxiIn","TaxiOut","Distance")]
ex1d <- select(hflights, contains(c('Taxi')), Distance)

ex2r <- hflights[c("Year","Month","DayOfWeek","DepTime","ArrTime")]
ex2d <- select(hflights, Year:ArrTime, -3)

ex3r <- hflights[c("TailNum","TaxiIn","TaxiOut")]
ex3d <- select(hflights, TailNum, contains(c('Taxi')))
```

Section 4 - The second of five verbs: mutate

Mutating is creating

```
# Add the new variable ActualGroundTime to a copy of hflights and save the result as g1.
g1 <- mutate(hflights, ActualGroundTime = ActualElapsedTime - AirTime)

# Add the new variable GroundTime to a g1. Save the result as g2.
g2 <- mutate(g1, GroundTime = TaxiIn + TaxiOut)

# Add the new variable AverageSpeed to g2. Save the result as g3.
g3 <- mutate(g2, AverageSpeed = Distance / AirTime * 60)
```

```
# Print out g3
glimpse(g3)
```

```
## Observations: 227,496
## Variables: 25
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 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...
## $ DayOfWeek         <int> 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, 1...
## $ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428,...
## $ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA",...
## $ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 6...
## $ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 4...
## $ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -...
## $ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "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, 8, 4, 6, 5, 6, 12,...
## $ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13,...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
## $ ActualGroundTime  <int> 20, 15, 22, 31, 18, 19, 27, 19, 30, 25, 28, 15, 1...
## $ GroundTime        <int> 20, 15, 22, 31, 18, 19, 27, 19, 30, 25, 28, 15, 1...
## $ AverageSpeed      <dbl> 336.0000, 298.6667, 280.0000, 344.6154, 305.4545,...
```

Add multiple variables using mutate

```
# hflights and dplyr are ready, are you?

# Add a second variable loss_percent to the dataset: m1
m1 <- mutate(hflights,
        loss = ArrDelay - DepDelay,
        loss_percent = (ArrDelay - DepDelay)/DepDelay * 100
        )
glimpse(m1)
```

```
## Observations: 227,496
## Variables: 24
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 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...
## $ DayOfWeek         <int> 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, 1...
## $ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
```

```
## $ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428,...
## $ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA",...
## $ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 6...
## $ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 4...
## $ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -...
## $ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "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, 8, 4, 6, 5, 6, 12,...
## $ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13,...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
## $ loss              <int> -10, -10, 0, 0, -8, -6, 0, -11, 1, 0, 0, -14, -7,...
## $ loss_percent      <dbl> -Inf, -1000.000000, 0.000000, 0.000000, -160.0000...
```

```
# Copy and adapt the previous command to reduce redendancy: m2
m2 <- mutate(hflights,
       loss = ArrDelay - DepDelay,
       loss_percent = loss/DepDelay * 100
       )
glimpse(m2)
```

```
## Observations: 227,496
## Variables: 24
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 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...
## $ DayOfWeek         <int> 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, 1...
## $ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428,...
## $ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA",...
## $ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 6...
## $ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 4...
## $ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -...
## $ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "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, 8, 4, 6, 5, 6, 12,...
## $ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13,...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
## $ loss              <int> -10, -10, 0, 0, -8, -6, 0, -11, 1, 0, 0, -14, -7,...
## $ loss_percent      <dbl> -Inf, -1000.000000, 0.000000, 0.000000, -160.0000...
```

```r
# Add the three variables as described in the third instruction: m3
m3 <- mutate(hflights,
       TotalTaxi = TaxiIn + TaxiOut,
       ActualGroundTime = ActualElapsedTime - AirTime,
       Diff = TotalTaxi - ActualGroundTime
       )
glimpse(m3)
```

```
## Observations: 227,496
## Variables: 25
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 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...
## $ DayOfWeek         <int> 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, 1...
## $ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428,...
## $ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA",...
## $ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 6...
## $ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 4...
## $ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -...
## $ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "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, 8, 4, 6, 5, 6, 12,...
## $ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13,...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
## $ TotalTaxi         <int> 20, 15, 22, 31, 18, 19, 27, 19, 30, 25, 28, 15, 1...
## $ ActualGroundTime  <int> 20, 15, 22, 31, 18, 19, 27, 19, 30, 25, 28, 15, 1...
## $ Diff              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

Filter and arrange

Section 5 - The third of five verbs: filter

Logical operators

```r
# All flights that traveled 3000 miles or more
filter(hflights, Distance >= 3000) %>% glimpse()
```

```
## Observations: 527
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DayofMonth        <int> 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 1...
## $ DayOfWeek         <int> 1, 7, 6, 5, 4, 3, 2, 1, 7, 6, 5, 4, 3, 2, 1, 7, 6...
## $ DepTime           <int> 924, 925, 1045, 1516, 950, 944, 924, 1144, 926, 9...
## $ ArrTime           <int> 1413, 1410, 1445, 1916, 1344, 1350, 1337, 1605, 1...
## $ UniqueCarrier     <chr> "Continental", "Continental", "Continental", "Con...
## $ FlightNum         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ TailNum           <chr> "N69063", "N76064", "N69063", "N77066", "N76055",...
## $ ActualElapsedTime <int> 529, 525, 480, 480, 474, 486, 493, 501, 489, 478,...
## $ AirTime           <int> 492, 493, 459, 463, 455, 471, 473, 464, 466, 465,...
## $ ArrDelay          <int> 23, 20, 55, 326, -6, 0, -13, 135, -15, -10, -16, ...
## $ DepDelay          <int> -1, 0, 80, 351, 25, 19, -1, 139, 1, 17, 3, 13, 1,...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "HNL", "HNL", "HNL", "HNL", "HNL", "HNL", "HNL", ...
## $ Distance          <int> 3904, 3904, 3904, 3904, 3904, 3904, 3904, 3904, 3...
## $ TaxiIn            <int> 6, 13, 4, 7, 4, 5, 5, 7, 6, 3, 6, 4, 6, 4, 5, 4, ...
## $ TaxiOut           <int> 31, 19, 17, 10, 15, 10, 15, 30, 17, 10, 19, 12, 1...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

```r
# All flights flown by one of JetBlue, Southwest, or Delta
filter(hflights, UniqueCarrier %in% c('JetBlue', 'Southwest', 'Delta')) %>% glimpse()
```

```
## Observations: 48,679
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DayofMonth        <int> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 7, 8, 9, 9, 1...
## $ DayOfWeek         <int> 6, 6, 7, 7, 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 7, 7, 1...
## $ DepTime           <int> 654, 1639, 703, 1604, 659, 1801, 654, 1608, 700, ...
## $ ArrTime           <int> 1124, 2110, 1113, 2040, 1100, 2200, 1103, 2034, 1...
## $ UniqueCarrier     <chr> "JetBlue", "JetBlue", "JetBlue", "JetBlue", "JetB...
## $ FlightNum         <int> 620, 622, 620, 622, 620, 622, 620, 622, 620, 624,...
## $ TailNum           <chr> "N324JB", "N324JB", "N324JB", "N324JB", "N229JB",...
## $ ActualElapsedTime <int> 210, 211, 190, 216, 181, 179, 189, 206, 183, 190,...
## $ AirTime           <int> 181, 188, 172, 176, 166, 165, 168, 175, 167, 166,...
## $ ArrDelay          <int> 5, 61, -6, 31, -19, 111, -16, 25, -14, -6, -17, 0...
## $ DepDelay          <int> -6, 54, 3, 19, -1, 136, -6, 23, 0, 9, -3, -6, 7, ...
## $ Origin            <chr> "HOU", "HOU", "HOU", "HOU", "HOU", "HOU", "HOU", ...
## $ Dest              <chr> "JFK", "JFK", "JFK", "JFK", "JFK", "JFK", "JFK", ...
## $ Distance          <int> 1428, 1428, 1428, 1428, 1428, 1428, 1428, 1428, 1...
## $ TaxiIn            <int> 6, 12, 6, 9, 3, 5, 9, 8, 4, 14, 7, 6, 9, 9, 3, 11...
## $ TaxiOut           <int> 23, 11, 12, 31, 12, 9, 12, 23, 12, 10, 9, 25, 10,...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

```
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

```
# All flights where taxiing took longer than flying
filter(hflights, (TaxiIn + TaxiOut) > AirTime) %>% glimpse()
```

```
## Observations: 1,389
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DayofMonth        <int> 24, 30, 24, 10, 31, 31, 31, 31, 30, 30, 30, 30, 3...
## $ DayOfWeek         <int> 1, 7, 1, 1, 1, 1, 1, 1, 7, 7, 7, 7, 7, 7, 7, 3, 3...
## $ DepTime           <int> 731, 1959, 1621, 941, 1301, 2113, 1434, 900, 1304...
## $ ArrTime           <int> 904, 2132, 1749, 1113, 1356, 2215, 1539, 1006, 14...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 460, 533, 1121, 1436, 241, 1533, 1541, 1583, 241,...
## $ TailNum           <chr> "N545AA", "N455AA", "N484AA", "N591AA", "N14629",...
## $ ActualElapsedTime <int> 93, 93, 88, 92, 55, 62, 65, 66, 64, 84, 80, 70, 7...
## $ AirTime           <int> 42, 43, 43, 45, 27, 30, 30, 32, 31, 40, 37, 30, 3...
## $ ArrDelay          <int> 29, 12, 4, 48, -2, 20, 15, 10, 10, 54, 16, 15, 30...
## $ DepDelay          <int> 11, -6, -9, 31, -4, 13, 4, 0, -1, 39, 2, -4, 17, ...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "DFW", "AUS", "AUS", "AUS", ...
## $ Distance          <int> 224, 224, 224, 224, 140, 140, 140, 140, 140, 305,...
## $ TaxiIn            <int> 14, 10, 10, 27, 5, 7, 5, 5, 6, 10, 6, 4, 6, 6, 3,...
## $ TaxiOut           <int> 37, 40, 35, 20, 23, 25, 30, 29, 27, 34, 37, 36, 3...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

Combining tests using boolean operators

```
# All flights that departed before 5am or arrived after 10pm
filter(hflights, DepTime < 500 | ArrTime > 2200) %>% glimpse()
```

```
## Observations: 27,799
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DayofMonth        <int> 4, 14, 10, 26, 30, 9, 31, 31, 31, 31, 31, 31, 31,...
## $ DayOfWeek         <int> 2, 5, 1, 3, 7, 7, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DepTime           <int> 2100, 2119, 1934, 1905, 1856, 1938, 1919, 2116, 1...
## $ ArrTime           <int> 2207, 2229, 2235, 2211, 2209, 2228, 2231, 2344, 2...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 533, 533, 1294, 1294, 1294, 731, 190, 209, 250, 2...
## $ TailNum           <chr> "N4XGAA", "N549AA", "N3BXAA", "N3BXAA", "N3CPAA",...
## $ ActualElapsedTime <int> 67, 70, 121, 126, 133, 290, 132, 268, 141, 134, 1...
## $ AirTime           <int> 42, 45, 107, 111, 108, 253, 107, 256, 121, 119, 1...
## $ ArrDelay          <int> 47, 69, 80, 56, 54, 78, -12, -15, -18, -10, -12, ...
```

```
## $ DepDelay          <int> 55, 74, 99, 70, 61, 73, -1, -7, 0, 8, -1, 5, 1, 6...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "MIA", "MIA", "MIA", "SEA", "MIA", ...
## $ Distance          <int> 224, 224, 964, 964, 964, 1874, 964, 1825, 1043, 8...
## $ TaxiIn            <int> 3, 5, 3, 5, 7, 5, 5, 4, 5, 6, 4, 18, 4, 7, 9, 11,...
## $ TaxiOut           <int> 22, 20, 11, 10, 18, 32, 20, 8, 15, 9, 18, 22, 17,...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

```
# All flights that departed late but arrived ahead of schedule
filter(hflights, DepDelay > 0 & ArrDelay < 0) %>% glimpse()
```

```
## Observations: 27,712
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DayofMonth        <int> 2, 5, 18, 18, 12, 13, 26, 1, 10, 12, 15, 17, 27, ...
## $ DayOfWeek         <int> 7, 3, 2, 2, 3, 4, 3, 6, 1, 3, 6, 1, 4, 7, 6, 5, 1...
## $ DepTime           <int> 1401, 1405, 1408, 721, 2015, 2020, 2009, 1631, 16...
## $ ArrTime           <int> 1501, 1507, 1508, 827, 2113, 2116, 2103, 1736, 17...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 428, 428, 428, 460, 533, 533, 533, 1121, 1121, 11...
## $ TailNum           <chr> "N557AA", "N492AA", "N507AA", "N558AA", "N555AA",...
## $ ActualElapsedTime <int> 60, 62, 60, 66, 58, 56, 54, 65, 61, 68, 64, 72, 6...
## $ AirTime           <int> 45, 44, 42, 46, 39, 44, 39, 37, 41, 44, 48, 51, 4...
## $ ArrDelay          <int> -9, -3, -2, -8, -7, -4, -17, -9, -5, -6, -9, -1, ...
## $ DepDelay          <int> 1, 5, 8, 1, 10, 15, 4, 1, 9, 1, 2, 2, 4, 5, 1, 2,...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", ...
## $ Distance          <int> 224, 224, 224, 224, 224, 224, 224, 224, 224, 224,...
## $ TaxiIn            <int> 6, 9, 7, 7, 9, 4, 9, 16, 8, 5, 5, 10, 10, 9, 9, 9...
## $ TaxiOut           <int> 9, 9, 11, 13, 10, 8, 6, 12, 12, 19, 11, 11, 13, 1...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

```
# All cancelled weekend flights
filter(hflights, Cancelled == 1 & DayOfWeek %in% c(6,7)) %>% glimpse()
```

```
## Observations: 585
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DayofMonth        <int> 9, 29, 9, 9, 9, 2, 29, 9, 1, 9, 9, 9, 9, 8, 9, 9,...
## $ DayOfWeek         <int> 7, 6, 7, 7, 7, 7, 6, 7, 6, 7, 7, 7, 7, 6, 7, 7, 7...
## $ DepTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ArrTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ UniqueCarrier     <chr> "American", "Continental", "Continental", "Delta"...
## $ FlightNum         <int> 1820, 408, 755, 8, 6726, 1629, 1590, 5229, 298, 2...
## $ TailNum           <chr> "N4XCAA", "", "", "N933DL", "N779SK", "N749SW", "...
```

```
## $ ActualElapsedTime <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ AirTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ArrDelay          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ DepDelay          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "HOU", "IAH", ...
## $ Dest              <chr> "DFW", "EWR", "ATL", "ATL", "ASE", "DAL", "ATL", ...
## $ Distance          <int> 224, 1400, 689, 689, 914, 239, 689, 469, 696, 696...
## $ TaxiIn            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ TaxiOut           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ Cancelled         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ CancellationCode  <chr> "B", "A", "B", "B", "B", "A", "A", "A", "A", "B",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "weather", "carrier", "weather", "weather", "weat...
```

```
# All flights that were cancelled after being delayed
filter(hflights, DepDelay > 0 & Cancelled == 1) %>% glimpse()
```

```
## Observations: 40
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 4, 4, 4, 4, 4...
## $ DayofMonth        <int> 26, 11, 19, 7, 4, 8, 2, 9, 1, 31, 4, 8, 21, 4, 4,...
## $ DayOfWeek         <int> 3, 2, 3, 5, 5, 2, 3, 3, 2, 4, 1, 5, 4, 1, 1, 1, 1...
## $ DepTime           <int> 1926, 1100, 1811, 2028, 1638, 1057, 802, 904, 150...
## $ ArrTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ UniqueCarrier     <chr> "Continental", "US_Airways", "ExpressJet", "Expre...
## $ FlightNum         <int> 310, 944, 2376, 3050, 1121, 408, 2189, 2605, 5812...
## $ TailNum           <chr> "N77865", "N452UW", "N15932", "N15912", "N537AA",...
## $ ActualElapsedTime <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ AirTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ArrDelay          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ DepDelay          <int> 26, 135, 6, 73, 8, 187, 2, 4, 28, 156, 42, 548, 3...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "EWR", "CLT", "ICT", "JAX", "DFW", "EWR", "DAL", ...
## $ Distance          <int> 1400, 913, 542, 817, 224, 1400, 217, 217, 689, 85...
## $ TaxiIn            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ TaxiOut           <int> NA, NA, NA, 19, 19, NA, NA, NA, 19, NA, NA, NA, 5...
## $ Cancelled         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ CancellationCode  <chr> "B", "B", "B", "A", "A", "A", "B", "B", "A", "B",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "weather", "weather", "weather", "carrier", "carr...
```

Blend together what you've learned!

```
# Select the flights that had JFK as their destination: c1
c1 <- filter(hflights, Dest == 'JFK')

# Combine the Year, Month and DayofMonth variables to create a Date column: c2
c2 <- mutate(c1, Date = paste(Year,Month,DayofMonth, sep="-"))

# Print out a selection of columns of c2
select(c2, Date, DepTime, ArrTime, TailNum)
```

```
## # A tibble: 695 x 4
##    Date     DepTime ArrTime TailNum
##    <chr>      <int>   <int> <chr>
##  1 2011-1-1     654    1124 N324JB
##  2 2011-1-1    1639    2110 N324JB
##  3 2011-1-2     703    1113 N324JB
##  4 2011-1-2    1604    2040 N324JB
##  5 2011-1-3     659    1100 N229JB
##  6 2011-1-3    1801    2200 N206JB
##  7 2011-1-4     654    1103 N267JB
##  8 2011-1-4    1608    2034 N267JB
##  9 2011-1-5     700    1103 N708JB
## 10 2011-1-5    1544    1954 N644JB
## # ... with 685 more rows
```

Recap on select, mutate and filter

- How many weekend flights flew a distance of more than 1000 miles but had a total taxiing time below 15 minutes?

```
hflights %>%
  filter(
    Distance > 1000,
    DayOfWeek > 5,
    TaxiIn + TaxiOut < 15
    ) %>%
  glimpse()
```

```
## Observations: 1,739
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DayofMonth        <int> 23, 30, 30, 29, 23, 23, 23, 22, 16, 16, 16, 16, 1...
## $ DayOfWeek         <int> 7, 7, 7, 6, 7, 7, 7, 6, 7, 7, 7, 7, 7, 7, 7, 7, 6...
## $ DepTime           <int> 1535, 851, 2234, 1220, 847, 1224, 931, 942, 848, ...
## $ ArrTime           <int> 1933, 1230, 2, 1353, 1213, 1345, 1045, 1340, 1136...
## $ UniqueCarrier     <chr> "JetBlue", "Continental", "Continental", "Contine...
## $ FlightNum         <int> 624, 1058, 1717, 1620, 1058, 1629, 1723, 1, 309, ...
## $ TailNum           <chr> "N599JB", "N39726", "N38417", "N87512", "N16709",...
## $ ActualElapsedTime <int> 178, 159, 208, 153, 146, 201, 194, 478, 288, 156,...
## $ AirTime           <int> 164, 145, 195, 139, 134, 188, 181, 465, 275, 143,...
## $ ArrDelay          <int> -27, -13, 89, 19, -30, -27, -28, -10, 12, -14, -1...
## $ DepDelay          <int> 0, -2, 94, 45, -6, -1, -5, 17, -2, -5, 3, -1, -6,...
## $ Origin            <chr> "HOU", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "JFK", "DCA", "SAN", "PHX", "DCA", "SNA", "ONT", ...
## $ Distance          <int> 1428, 1208, 1303, 1009, 1208, 1347, 1334, 3904, 1...
## $ TaxiIn            <int> 6, 3, 3, 5, 4, 4, 3, 3, 5, 3, 3, 4, 3, 6, 4, 4, 4...
## $ TaxiOut           <int> 8, 11, 10, 9, 8, 9, 10, 10, 8, 10, 9, 10, 11, 8, ...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

- In this dataset it is 1,739 flights
- In the class the answer was 155
- I think they just have the data filtered to one city (Houston)

Section 6 - Almost there: the arrange verb

Arranging your data

```
# Definition of dtc
dtc <- filter(hflights, Cancelled == 1, !is.na(DepDelay))

# Arrange dtc by departure delays
arrange(dtc, DepDelay) %>% glimpse()
```

```
## Observations: 68
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 7, 1, 12, 10, 7, 9, 2, 5, 1, 1, 2, 3, 4, 4, 5, 6,...
## $ DayofMonth        <int> 23, 17, 1, 12, 29, 29, 9, 9, 20, 17, 21, 18, 30, ...
## $ DayOfWeek         <int> 6, 1, 4, 3, 5, 4, 3, 1, 4, 1, 1, 5, 6, 7, 1, 1, 7...
## $ DepTime           <int> 605, 916, 541, 2022, 1424, 1639, 555, 715, 1413, ...
## $ ArrTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ UniqueCarrier     <chr> "Frontier", "ExpressJet", "US_Airways", "American...
## $ FlightNum         <int> 225, 3068, 282, 3724, 1079, 2062, 3265, 1177, 552...
## $ TailNum           <chr> "N912FR", "N13936", "N840AW", "N539MQ", "N14628",...
## $ ActualElapsedTime <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ AirTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ArrDelay          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ DepDelay          <int> -10, -9, -9, -8, -6, -6, -5, -5, -4, -4, -3, -3, ...
## $ Origin            <chr> "HOU", "IAH", "IAH", "IAH", "IAH", "IAH", "HOU", ...
## $ Dest              <chr> "DEN", "HRL", "PHX", "LAX", "ORD", "ATL", "DFW", ...
## $ Distance          <int> 883, 295, 1009, 1379, 925, 689, 247, 1076, 1190, ...
## $ TaxiIn            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ TaxiOut           <int> 10, NA, NA, NA, 13, NA, 11, 17, NA, 8, NA, NA, NA...
## $ Cancelled         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ CancellationCode  <chr> "A", "B", "A", "A", "A", "B", "A", "A", "A", "B",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "carrier", "weather", "carrier", "carrier", "carr...
```

```
# Arrange dtc so that cancellation reasons are grouped
arrange(dtc, CancellationCode) %>% glimpse()
```

```
## Observations: 68
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5...
## $ DayofMonth        <int> 20, 7, 4, 8, 1, 21, 9, 18, 4, 8, 21, 4, 11, 7, 30...
## $ DayOfWeek         <int> 4, 5, 5, 2, 2, 1, 3, 5, 1, 5, 4, 1, 1, 4, 6, 7, 1...
## $ DepTime           <int> 1413, 2028, 1638, 1057, 1508, 2257, 555, 727, 163...
## $ ArrTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ UniqueCarrier     <chr> "United", "ExpressJet", "American", "Continental"...
```

```
## $ FlightNum         <int> 552, 3050, 1121, 408, 5812, 1111, 3265, 109, 8, 4...
## $ TailNum           <chr> "N509UA", "N15912", "N537AA", "N11641", "N959SW",...
## $ ActualElapsedTime <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ AirTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ArrDelay          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ DepDelay          <int> -4, 73, 8, 187, 28, -3, -5, -3, 42, 548, 3, 109, ...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "HOU", ...
## $ Dest              <chr> "IAD", "JAX", "DFW", "EWR", "ATL", "AUS", "DFW", ...
## $ Distance          <int> 1190, 817, 224, 1400, 689, 140, 247, 862, 689, 23...
## $ TaxiIn            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ TaxiOut           <int> NA, 19, 19, NA, 19, NA, 11, NA, NA, NA, 5, NA, 26...
## $ Cancelled         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ CancellationCode  <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "carrier", "carrier", "carrier", "carrier", "carr...
```

```
# Arrange dtc according to carrier and departure delays
arrange(dtc, UniqueCarrier, DepDelay) %>% glimpse()
```

```
## Observations: 68
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 6, 8, 2, 10, 2, 7, 4, 4, 5, 9, 5, 9, 7, 1, 8, 7, ...
## $ DayofMonth        <int> 11, 18, 4, 12, 9, 17, 30, 10, 23, 29, 16, 26, 29,...
## $ DayOfWeek         <int> 6, 4, 5, 3, 3, 7, 6, 7, 1, 4, 1, 1, 5, 3, 4, 1, 3...
## $ DepTime           <int> 1649, 1808, 1638, 2022, 555, 1917, 612, 1147, 657...
## $ ArrTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ UniqueCarrier     <chr> "AirTran", "American", "American", "American_Eagl...
## $ FlightNum         <int> 1595, 1294, 1121, 3724, 3265, 3717, 5386, 5402, 5...
## $ TailNum           <chr> "N946AT", "N3FLAA", "N537AA", "N539MQ", "N613MQ",...
## $ ActualElapsedTime <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ AirTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ArrDelay          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ DepDelay          <int> 64, 3, 8, -8, -5, -3, -3, -3, -3, -2, -1, 220, -6...
## $ Origin            <chr> "HOU", "IAH", "IAH", "IAH", "HOU", "IAH", "IAH", ...
## $ Dest              <chr> "BKG", "MIA", "DFW", "LAX", "DFW", "ORD", "MEM", ...
## $ Distance          <int> 490, 964, 224, 1379, 247, 925, 469, 469, 696, 107...
## $ TaxiIn            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ TaxiOut           <int> 25, NA, 19, NA, 11, NA, NA, NA, NA, NA, NA, NA, 1...
## $ Cancelled         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ CancellationCode  <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "carrier", "carrier", "carrier", "carrier", "carr...
```

Reverse the order of arranging

```
# Arrange according to carrier and decreasing departure delays
arrange(hflights, UniqueCarrier, desc(DepDelay)) %>% glimpse()
```

```
## Observations: 227,496
## Variables: 22
```

```
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 2, 3, 2, 11, 5, 5, 4, 6, 5, 7, 7, 6, 6, 7, 5, 6, ...
## $ DayofMonth        <int> 19, 14, 16, 13, 26, 26, 28, 5, 7, 25, 28, 24, 25,...
## $ DayOfWeek         <int> 6, 1, 3, 7, 4, 4, 4, 7, 6, 1, 4, 5, 6, 1, 4, 4, 1...
## $ DepTime           <int> 1902, 2024, 2349, 2312, 2353, 1922, 1045, 2207, 1...
## $ ArrTime           <int> 2143, 2309, 227, 213, 305, 2229, 1328, 52, 1256, ...
## $ UniqueCarrier     <chr> "AirTran", "AirTran", "AirTran", "AirTran", "AirT...
## $ FlightNum         <int> 298, 286, 292, 292, 296, 288, 290, 292, 290, 292,...
## $ TailNum           <chr> "N974AT", "N899AT", "N934AT", "N951AT", "N959AT",...
## $ ActualElapsedTime <int> 101, 105, 98, 121, 132, 127, 103, 105, 107, 127, ...
## $ AirTime           <int> 89, 89, 85, 99, 115, 104, 88, 91, 94, 105, 105, 9...
## $ ArrDelay          <int> 500, 483, 367, 353, 292, 290, 258, 259, 216, 216,...
## $ DepDelay          <int> 507, 493, 380, 347, 275, 274, 270, 269, 224, 212,...
## $ Origin            <chr> "HOU", "HOU", "HOU", "HOU", "HOU", "HOU", "HOU", ...
## $ Dest              <chr> "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", ...
## $ Distance          <int> 696, 696, 696, 696, 696, 696, 696, 696, 696, 696,...
## $ TaxiIn            <int> 5, 7, 4, 14, 11, 11, 7, 4, 7, 9, 6, 10, 5, 12, 12...
## $ TaxiOut           <int> 7, 9, 9, 8, 6, 12, 8, 10, 6, 13, 7, 9, 9, 8, 19, ...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

```
# Arrange flights by total delay (normal order).
arrange(hflights, (DepDelay + ArrDelay)) %>% glimpse()
```

```
## Observations: 227,496
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 7, 8, 8, 8, 8, 12, 1, 8, 8, 8, 8, 9, 12, 9, 12, 8...
## $ DayofMonth        <int> 3, 31, 21, 28, 29, 25, 30, 3, 4, 18, 26, 11, 24, ...
## $ DayOfWeek         <int> 7, 3, 7, 7, 1, 7, 7, 3, 4, 4, 5, 7, 6, 2, 6, 2, 7...
## $ DepTime           <int> 1914, 934, 935, 2059, 935, 741, 620, 1741, 930, 9...
## $ ArrTime           <int> 2039, 1039, 1039, 2206, 1041, 926, 812, 1810, 104...
## $ UniqueCarrier     <chr> "ExpressJet", "SkyWest", "SkyWest", "SkyWest", "S...
## $ FlightNum         <int> 2804, 2040, 2001, 2003, 2040, 4591, 4461, 2603, 1...
## $ TailNum           <chr> "N12157", "N783SK", "N767SK", "N783SK", "N767SK",...
## $ ActualElapsedTime <int> 85, 185, 184, 187, 186, 165, 172, 89, 191, 184, 1...
## $ AirTime           <int> 66, 172, 171, 171, 169, 147, 156, 73, 177, 172, 1...
## $ ArrDelay          <int> -70, -56, -56, -54, -54, -57, -49, -40, -49, -52,...
## $ DepDelay          <int> -1, -11, -10, -11, -10, -4, -10, -19, -10, -6, -3...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "MEM", "BFL", "BFL", "BFL", "BFL", "SLC", "SLC", ...
## $ Distance          <int> 468, 1428, 1428, 1428, 1428, 1195, 1195, 501, 142...
## $ TaxiIn            <int> 4, 3, 3, 5, 4, 4, 5, 5, 4, 4, 5, 6, 3, 4, 5, 7, 4...
## $ TaxiOut           <int> 15, 10, 10, 11, 13, 14, 11, 11, 10, 8, 10, 19, 48...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

```
# Keep flights leaving to DFW before 8am and arrange according to decreasing AirTime
hflights %>% filter(Dest == 'DFW', DepTime < 800) %>% arrange(desc(AirTime)) %>% glimpse()
```

```
## Observations: 799
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 11, 8, 10, 5, 4, 4, 6, 9, 3, 12, 4, 4, 5, 6, 6, 1...
## $ DayofMonth        <int> 22, 25, 12, 2, 4, 4, 21, 1, 14, 5, 25, 11, 11, 21...
## $ DayOfWeek         <int> 2, 4, 3, 1, 1, 1, 2, 4, 1, 1, 1, 1, 3, 2, 1, 7, 2...
## $ DepTime           <int> 635, 602, 559, 716, 741, 627, 726, 715, 729, 724,...
## $ ArrTime           <int> 825, 758, 738, 854, 949, 742, 848, 844, 917, 847,...
## $ UniqueCarrier     <chr> "American", "American_Eagle", "American_Eagle", "...
## $ FlightNum         <int> 1903, 3265, 3265, 2237, 1225, 3265, 2259, 1948, 1...
## $ TailNum           <chr> "N477AA", "N633MQ", "N632MQ", "N552AA", "N4XVAA",...
## $ ActualElapsedTime <int> 110, 116, 99, 98, 128, 75, 82, 89, 108, 83, 91, 7...
## $ AirTime           <int> 81, 74, 71, 70, 63, 62, 62, 62, 61, 61, 59, 59, 5...
## $ ArrDelay          <int> 40, 53, 33, 29, 89, 37, 9, 9, 33, 2, 20, 1, 0, 18...
## $ DepDelay          <int> 0, 2, -1, 1, 31, 27, -4, -5, -1, -1, -1, -7, -6, ...
## $ Origin            <chr> "IAH", "HOU", "HOU", "IAH", "IAH", "HOU", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", ...
## $ Distance          <int> 224, 247, 247, 224, 224, 247, 224, 224, 224, 224,...
## $ TaxiIn            <int> 11, 21, 8, 11, 6, 3, 5, 16, 11, 3, 7, 5, 7, 5, 8,...
## $ TaxiOut           <int> 18, 21, 20, 17, 59, 10, 15, 11, 36, 19, 25, 14, 1...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

Summarise and the pipe operator

Section 7 - Last but not least: summarise

The syntax of summarize

```
glimpse(hflights)
```

```
## Observations: 227,496
## Variables: 22
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2...
## $ Month             <int> 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...
## $ DayOfWeek         <int> 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, 1...
## $ ArrTime           <int> 1500, 1501, 1502, 1513, 1507, 1503, 1509, 1454, 1...
## $ UniqueCarrier     <chr> "American", "American", "American", "American", "...
## $ FlightNum         <int> 428, 428, 428, 428, 428, 428, 428, 428, 428, 428,...
```

```
## $ TailNum           <chr> "N576AA", "N557AA", "N541AA", "N403AA", "N492AA",...
## $ ActualElapsedTime <int> 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70, 56, 6...
## $ AirTime           <int> 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42, 41, 4...
## $ ArrDelay          <int> -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 29, 5, -...
## $ DepDelay          <int> 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19, -2, -...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
## $ Dest              <chr> "DFW", "DFW", "DFW", "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, 8, 4, 6, 5, 6, 12,...
## $ TaxiOut           <int> 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, 11, 13,...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CancellationCode  <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Code              <chr> "not cancelled", "not cancelled", "not cancelled"...
```

```
# Print out a summary with variables min_dist and max_dist
summarize(hflights, min_dist = min(Distance), max_dist = max(Distance))
```

```
## # A tibble: 1 x 2
##   min_dist max_dist
##      <int>    <int>
## 1       79     3904
```

```
# Print out a summary with variable max_div
hflights %>% filter(Diverted == 1) %>% summarize(max_div = max(Distance))
```

```
## # A tibble: 1 x 1
##   max_div
##     <int>
## 1    3904
```

Aggregate functions

Aggregate functions defined in R:

- `min(x)` - minimum value of vector x
- `max(x)` - maximum value of vector x
- `mean(x)` - mean value of vector x
- `median(x)` - median value of vector x
- `quantile(x, p)` - pth quantile of vector x
- `sd(x)` - standard deviation of vector x
- `var(x)` - variance of vector x
- `IQR(x)` - Inter Quartile Range (IQR) of vector x
- `diff(range(x))` - total range of vector x

```
# Remove rows that have NA ArrDelay: temp1
temp1 <- filter(hflights, !is.na(ArrDelay))

# Generate summary about ArrDelay column of temp1
summarize(temp1,
  earliest = min(ArrDelay),
```

```
  average = mean(ArrDelay),
  latest = max(ArrDelay),
  sd = sd(ArrDelay)
  )
```

```
## # A tibble: 1 x 4
##   earliest average latest    sd
##      <int>   <dbl>  <int> <dbl>
## 1      -70    7.09    978  30.7
```

```
# Keep rows that have no NA TaxiIn and no NA TaxiOut: temp2
temp2 <- filter(hflights, !is.na(TaxiIn) & !is.na(TaxiOut))

# Print the maximum taxiing difference of temp2 with summarise()
summarise(temp2, max_taxi_diff = max(abs(TaxiIn - TaxiOut)))
```

```
## # A tibble: 1 x 1
##   max_taxi_diff
##           <int>
## 1           160
```

dplyr aggregate functions

dplyr has some of its own aggregate functions:

- `first(x)` - The first element of vector x
- `last(x)` - The last element of vector x
- `nth(x, n)` - The nth element of vector x
- `n()` - The number of rows in the data.frame or group of observations that summarise() describes
- `n_distinct(x)` - The number of unique values in vector x

```
# Generate summarizing statistics for hflights
summarise(hflights,
  n_obs = n(),
  n_carrier = n_distinct(UniqueCarrier),
  n_dest = n_distinct(Dest),
  dest100 = nth(Dest, 100)
  )
```

```
## # A tibble: 1 x 4
##    n_obs n_carrier n_dest dest100
##    <int>     <int>  <int> <chr>
## 1 227496        15    116 DFW
```

```
# Filter hflights to keep all American Airline flights: aa
aa <- filter(hflights, UniqueCarrier == 'American')

# Generate summarizing statistics for aa
summarise(aa,
  n_flights = n(),
```

```
  n_canc = sum(Cancelled),
  p_canc = n_canc/n_flights * 100,
  avg_delay = mean(ArrDelay, na.rm=T)
  )
```

```
## # A tibble: 1 x 4
##   n_flights n_canc p_canc avg_delay
##       <int>  <int>  <dbl>     <dbl>
## 1      3244     60   1.85     0.892
```

Section 8 - Chaining your functions: the pipe operator

- the pipe operator %>% is probably my favorite thing in R
- it takes the output from the last function and passes it into the next function as the first argument
  - But you can also use a . to pass it into the next function in any location!
- This lets you make code easily readable from left to right and top to bottom.
  - I can't say enough about how much better this makes it to read code and see what someone is doing

Overview of syntax

```
# Write the 'piped' version of the English sentences.
hflights %>%
  mutate(diff = TaxiOut - TaxiIn) %>%
  filter(!is.na(diff)) %>%
  summarize(avg = mean(diff))
```

```
## # A tibble: 1 x 1
##     avg
##   <dbl>
## 1  8.99
```

Drive of fly? Part 1 of 2

```
# Build data frame with 4 columns of hflights and 2 self-defined columns: d
d <- hflights %>%
  select(Dest, UniqueCarrier, Distance, ActualElapsedTime) %>%
  mutate(
    RealTime = ActualElapsedTime + 100,
    mph = Distance/RealTime*60
    )

# Filter and summarise d according to the instructions
d %>%
  filter(
    !is.na(mph),
```

```
    mph < 70
    ) %>%
  summarize(
    n_less = n(),
    n_dest = n_distinct(Dest),
    min_dist = min(Distance),
    max_dist = max(Distance)
    )
```

```
## # A tibble: 1 x 4
##   n_less n_dest min_dist max_dist
##    <int>  <int>    <int>    <int>
## 1   6726     13       79      305
```

Drive or fly? Part 2 of 2

```
# Solve the exercise using a combination of dplyr verbs and %>%
hflights %>%
  mutate(RealTime = ActualElapsedTime + 100, mph = Distance / RealTime * 60) %>%
  filter(mph < 105 | Cancelled == 1 | Diverted == 1) %>%
  summarise(n_non = n(),
            p_non = n_non / nrow(hflights) * 100,
            n_dest = n_distinct(Dest),
            min_dist = min (Distance),
            max_dist = max(Distance))
```

```
## # A tibble: 1 x 5
##   n_non p_non n_dest min_dist max_dist
##   <int> <dbl>  <int>    <int>    <int>
## 1 42400  18.6    113       79     3904
```

Advanced piping exercise

```
hflights %>%
  filter(
    !is.na(ArrTime),
    !is.na(DepTime),
    ArrTime < DepTime
    ) %>%
  summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  2718
```

Group_by and working with databases

Section 9 - get group-wise insights: group_by

- Combining `group_by` with `summarize` is very powerful
  - You can also combine it with `mutate` and `arrange` to create powerful window functions

Unite and conquer using group_by

```
# Make an ordered per-carrier summary of hflights
hflights %>%
 group_by(UniqueCarrier) %>%
 summarise(
   n_flights = n(),
   n_canc = sum(Cancelled),
   p_canc = n_canc/n_flights * 100,
   avg_delay = mean(ArrDelay, na.rm=T)
   ) %>%
 arrange(avg_delay, p_canc)
```

```
## # A tibble: 15 x 5
##    UniqueCarrier      n_flights n_canc p_canc avg_delay
##    <chr>                  <int>  <int>  <dbl>     <dbl>
##  1 US_Airways              4082     46  1.13     -0.631
##  2 American                3244     60  1.85      0.892
##  3 AirTran                 2139     21  0.982     1.85
##  4 Alaska                   365      0  0         3.19
##  5 Mesa                      79      1  1.27      4.01
##  6 Delta                   2641     42  1.59      6.08
##  7 Continental            70032    475  0.678     6.10
##  8 American_Eagle          4648    135  2.90      7.15
##  9 Atlantic_Southeast      2204     76  3.45      7.26
## 10 Southwest              45343    703  1.55      7.59
## 11 Frontier                 838      6  0.716     7.67
## 12 ExpressJet             73053   1132  1.55      8.19
## 13 SkyWest                16061    224  1.39      8.69
## 14 JetBlue                  695     18  2.59      9.86
## 15 United                  2072     34  1.64     10.5
```

```
# Make an ordered per-day summary of hflights
hflights %>%
  group_by(DayOfWeek) %>%
  summarize(avg_taxi = mean(TaxiIn + TaxiOut, na.rm=T)) %>%
  arrange(desc(avg_taxi))
```

```
## # A tibble: 7 x 2
##   DayOfWeek avg_taxi
```

```
##       <int>    <dbl>
## 1         1     21.8
## 2         2     21.4
## 3         4     21.3
## 4         3     21.2
## 5         5     21.2
## 6         7     20.9
## 7         6     20.4
```

Combine group_by with mutate

```
# Solution to first instruction
hflights %>%
  filter(!is.na(ArrDelay)) %>%
  group_by(UniqueCarrier) %>%
  summarize(p_delay = sum(ArrDelay > 0)/n()) %>%
  mutate(rank = rank(p_delay)) %>%
  arrange(rank)
```

```
## # A tibble: 15 x 3
##    UniqueCarrier      p_delay  rank
##    <chr>                <dbl> <dbl>
##  1 American             0.303     1
##  2 AirTran              0.311     2
##  3 US_Airways           0.327     3
##  4 Atlantic_Southeast   0.368     4
##  5 American_Eagle       0.370     5
##  6 Delta                0.387     6
##  7 JetBlue              0.395     7
##  8 Alaska               0.437     8
##  9 Southwest            0.464     9
## 10 Mesa                 0.474    10
## 11 Continental          0.491    11
## 12 ExpressJet           0.494    12
## 13 United               0.496    13
## 14 SkyWest              0.535    14
## 15 Frontier             0.556    15
```

```
# Solution to second instruction
hflights %>%
  filter(
    !is.na(ArrDelay),
    ArrDelay > 0
    ) %>%
  group_by(UniqueCarrier) %>%
  summarize(avg = mean(ArrDelay)) %>%
  mutate(rank = rank(avg)) %>%
  arrange(rank)
```

```
## # A tibble: 15 x 3
##    UniqueCarrier        avg  rank
```

```
##    <chr>              <dbl> <dbl>
##  1 Mesa                18.7     1
##  2 Frontier            18.7     2
##  3 US_Airways          20.7     3
##  4 Continental         22.1     4
##  5 Alaska              22.9     5
##  6 SkyWest             24.1     6
##  7 ExpressJet          24.2     7
##  8 Southwest           25.3     8
##  9 AirTran             27.9     9
## 10 American            28.5    10
## 11 Delta               32.1    11
## 12 United              32.5    12
## 13 American_Eagle      38.8    13
## 14 Atlantic_Southeast  40.2    14
## 15 JetBlue             45.5    15
```

Advanced group_by exercises

```
# Which plane (by tail number) flew out of Houston the most times? How many times? adv1
adv1 <- hflights %>%
  group_by(TailNum) %>%
  summarize(n = n()) %>%
  filter(n == max(n))
adv1
```

```
## # A tibble: 1 x 2
##   TailNum     n
##   <chr>   <int>
## 1 N14945    971
```

```
# How many airplanes only flew to one destination from Houston? adv2
adv2 <- hflights %>%
  group_by(TailNum) %>%
  summarize(n_dest = n_distinct(Dest)) %>%
  filter(n_dest == 1) %>%
  summarize(nplanes = n())
adv2
```

```
## # A tibble: 1 x 1
##   nplanes
##     <int>
## 1    1526
```

```
# Find the most visited destination for each carrier: adv3
adv3 <- hflights %>%
  group_by(UniqueCarrier, Dest) %>%
  summarize(n = n()) %>%
  group_by(UniqueCarrier) %>%
  mutate(rank = rank(desc(n))) %>%
  filter(rank == 1) %>%
```

```
  arrange(UniqueCarrier, rank)
adv3
```

```
## # A tibble: 15 x 4
## # Groups:   UniqueCarrier [15]
##    UniqueCarrier      Dest      n  rank
##    <chr>              <chr> <int> <dbl>
##  1 AirTran            ATL    2029     1
##  2 Alaska             SEA     365     1
##  3 American           DFW    2105     1
##  4 American_Eagle     DFW    2424     1
##  5 Atlantic_Southeast DTW     851     1
##  6 Continental        EWR    3924     1
##  7 Delta              ATL    2396     1
##  8 ExpressJet         CRP    3175     1
##  9 Frontier           DEN     837     1
## 10 JetBlue            JFK     695     1
## 11 Mesa               CLT      71     1
## 12 SkyWest            COS    1335     1
## 13 Southwest          DAL    8243     1
## 14 United             SFO     643     1
## 15 US_Airways         CLT    2212     1
```

```
# Find the carrier that travels to each destination the most: adv4
adv4 <- hflights %>%
  group_by(Dest, UniqueCarrier) %>%
  summarize(n = n()) %>%
  group_by(Dest) %>%
  mutate(rank = rank(desc(n))) %>%
  filter(rank == 1)
adv4
```

```
## # A tibble: 116 x 4
## # Groups:   Dest [116]
##    Dest  UniqueCarrier     n  rank
##    <chr> <chr>         <int> <dbl>
##  1 ABQ   Southwest      1019     1
##  2 AEX   ExpressJet      724     1
##  3 AGS   Continental       1     1
##  4 AMA   ExpressJet     1297     1
##  5 ANC   Continental     125     1
##  6 ASE   SkyWest         125     1
##  7 ATL   Delta          2396     1
##  8 AUS   Continental    2645     1
##  9 AVL   ExpressJet      350     1
## 10 BFL   SkyWest         504     1
## # ... with 106 more rows
```

Section 10 - dplyr and databases

- dplyr can connect to a database

  - You can manipulate the data in the dabase (query essentially) and then only pull back the result into R
  - This lets you work with much larger datasets stored in a relational database than you could on your local machine or having to augment R with hadoop

- I used this functionality a lot at work.
  - Getting comfortable with this also help when using spark and sparklyr.
  - It really nice to have one consistent way to manipulate data where ever its stored:
    * locally in the workspace,
    * a relational database
    * in HDFS
    * any file system accessed with spark

dplyr deals with different types

```
library(data.table)

# Convert hflights to a data.table
class(hflights)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
hflights2 <- as.data.table(hflights)
class(hflights2)
```

```
## [1] "data.table" "data.frame"
```

```
# Use summarise to calculate n_carrier
s2 <- hflights2 %>%
  summarize(n_carrier = n_distinct(UniqueCarrier))
s2
```

```
##   n_carrier
## 1        15
```

dplyr and mySQL databases

```
library(RMySQL)
library(dbplyr)

# Set up a connection to the mysql database
my_db <- src_mysql(dbname = "dplyr",
                   host = "courses.csrrinzqubik.us-east-1.rds.amazonaws.com",
                   port = 3306,
                   user = "student",
                   password = "datacamp")

# Reference a table within that source: nycflights
```

```
nycflights <- tbl(my_db, "dplyr")

# glimpse at nycflights
glimpse(nycflights)
```

```
## Observations: ??
## Variables: 17
## Database: mysql 5.6.44-log [student@courses.csrrinzqubik.us-east-1.rds.amazonaws.com:/dplyr]
## $ id        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17...
## $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 201...
## $ month     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ day       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ dep_time  <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, 55...
## $ dep_delay <int> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1, ...
## $ arr_time  <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849, 8...
## $ arr_delay <int> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -14,...
## $ carrier   <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "AA...
## $ tailnum   <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N39463...
## $ flight    <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 49,...
## $ origin    <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA", "...
## $ dest      <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD", "...
## $ air_time  <int> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 158...
## $ distance  <int> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, 10...
## $ hour      <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, ...
## $ minute    <int> 17, 33, 42, 44, 54, 54, 55, 57, 57, 58, 58, 58, 58, 58, 5...
```

```
# Ordered, grouped summary of nycflights
nycflights %>%
   group_by(carrier) %>%
   summarise(n_flights = n(), avg_delay = mean(arr_delay)) %>%
   arrange(avg_delay)
```

```
## # Source:     lazy query [?? x 3]
## # Database:   mysql 5.6.44-log
## #   [student@courses.csrrinzqubik.us-east-1.rds.amazonaws.com:/dplyr]
## # Ordered by: avg_delay
##    carrier n_flights avg_delay
##    <chr>       <dbl>     <dbl>
##  1 AS            714    -9.86
##  2 HA            342    -6.92
##  3 AA          32729     0.356
##  4 DL          48110     1.63
##  5 VX           5162     1.75
##  6 US          20536     2.06
##  7 UA          58665     3.50
##  8 9E          18460     6.91
##  9 B6          54635     9.36
## 10 WN          12275     9.47
## # ... with more rows
```

Talk with Hadley Wickham

- Two goals

  - Make it easier to think about data manipulation. What are the fundamental verbs
  - Compute efficiently with the data. It uses C++. It can generate SQL for you and send to database

- plyr was about using split, apply, combine
  - dplyr focuses on just data frames, but thats what most people use anyways

- Learn about tidy data
- Get a dataset you are motivated by and start playing with it
- Get familiar with window functions
  - There are a wide class a problems that can be solved by window functions