Introduction to dplyr for Fast Data Manipulation in R

Note: There is a 40-minute video tutorial (https://www.youtube.com/watch?v=jWjqLW[illegible]) on YouTube that walks through this document in detail.

Why do I use dplyr?

- Great for data exploration and transformation
- Intuitive to write and easy to read, especially when using the "chaining" syntax (covered below)
- Fast on data frames

dplyr functionality

- Five basic verbs: `filter`, `select`, `arrange`, `mutate`, `summarise` (plus `group_by`)
- Can work with data stored in databases and data tables (http://datatable.r-forge.r-project.org/)
- Joins: inner join, left join, semi-join, anti-join (not covered below)
- Window functions for calculating ranking, offsets, and more
- Better than plyr (http://blog.rstudio.org/2014/01/17/introducing-dplyr/) if you're only working with data frames (though it doesn't yet duplicate all of the plyr functionality)
- Examples below are based upon the latest release (https://github.com/hadley/dplyr/releases), version 0.2 (released May 2014)

Loading dplyr and an example dataset

- dplyr will mask a few base functions
- If you also use plyr, load plyr first
- hflights is flights departing from two Houston airports in 2011

```
# 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
```