dplyr

Charlotte Wickham

April 2, 2014

Today

- Quick intro to dplyr
- Charlotte & Alix tag team programming

This intro

This intro is no substitute for reading the introduction vignette. Read it!

We'll probably also assign the databases and memory vignettes in future, but there's no harm in taking a look at them early.

Why learn dplyr?

- It's designed to be fast and avoid unneccessary memory consumption.
- It's a useful mental model, which means it reduces cognitive effort: you expend some energy learning dplyr now, in exchange for less brain power required in future for routine data analysis tasks.
- Master dplyr on data.frames and move seamlessly to databases.

dplyr is not plyr!

To run code later:

```
library(dplyr)
library(hflights)
hflights_df <- tbl_df(hflights)
```

five verbs + group_by

Five data manipulation verbs

An action on a data.frame, results in a data.frame

First argument is always a data.frame, remaining arguments specify the action, (no need for `$`):

```
verb(hflights_df, ...)
```

Verb	Action	Example
filter	subset rows	`filter(hflights_df, Dest == "PDX")`
select	subset columns	`select(hflights_df, ArrDelay, UniqueCarrier)`
arrange	reorder rows	`arrange(hflights_df, desc(ArrDelay))`
mutate	add new columns	`mutate(hflights_df, more15 = ArrDelay > 15)`
summarise	reduce to a single row	`summarise(hflights_df, avg_delay = mean(ArrDelay, na.rm = TRUE))`

group_by

Apply to a data.frame to define a "grouping" of rows based on the levels of one or more columns. The verbs know about groups:

- summarise, mutate, filter - operate within each group
- arrange - orders first by grouping variable
- select - no effect

`group_by` first:

```
carriers <- group_by(hflights_df, UniqueCarrier)
```

then use a verb:

```
summarise(carriers,
  median_delay = median(ArrDelay, na.rm = TRUE))
```

Getting good

Just a matter of learning to translate questions into a sequence verbs and grouping operations. Then writing the code is easy.

Which day in 2011 had the most delays?

- group by day
- summarise by the proportion of delayed flights
- arrange by decreasing count

There's more to learn

- the general purpose verb `do`, do some function within each group. For example, let's you do things like fit a regression model to each group and keep the results in a list.
- addtional useful dplyr functions: `n()`, `n_distinct()`, `first()`, `last()`, `nth()`,
- joins
- windowing functions: `?ranking`, `lag()`, `lead()`
- using databases