Introduction to R for Data Management and Analysis

Marcel Ramos, MPH

Tuesday, June 14, 2016

Notes on last Thursday's lecture

- Examples with pipes
- Reshaping your data

Using the `nycflights13` *dataset*

```
library(nycflights13); library(dplyr)
flights %>% group_by(carrier) %>%
  summarise(avg_depdelay = mean(dep_delay, na.rm = TRUE),
            count = n()) %>% left_join(airlines) %>%
    arrange(avg_depdelay) %>% head
```

```
## Source: local data frame [6 x 4]
##
##   carrier avg_depdelay count                  name
##     (chr)        (dbl) (int)                 (chr)
## 1      US     3.782418 20536       US Airways Inc.
## 2      HA     4.900585   342 Hawaiian Airlines Inc.
## 3      AS     5.804775   714   Alaska Airlines Inc.
## 4      AA     8.586016 32729 American Airlines Inc.
## 5      DL     9.264505 48110   Delta Air Lines Inc.
## 6      MQ    10.552041 26397             Envoy Air
```

Reshaping data using ***gather***

```
data(iris); library(tidyr)
longdata <- gather(tbl_df(iris), key = measure, n,
  Sepal.Length:Petal.Width) %>% separate(measure, c("type",
    "dimension"))
longdata %>% group_by(Species, type, dimension) %>%
  summarise(avg_dim = mean(n, na.rm = TRUE))
```

```
## Source: local data frame [12 x 4]
## Groups: Species, type [?]
##
##      Species  type dimension avg_dim
##       (fctr) (chr)     (chr)   (dbl)
## 1     setosa Petal    Length   1.462
## 2     setosa Petal     Width   0.246
## 3     setosa Sepal    Length   5.006
## 4     setosa Sepal     Width   3.428
```

Pew example

```
library(readr)
(pew <- read_csv("../Data/pew.csv"))
```

```
## Source: local data frame [18 x 11]
##
##                   religion <$10k $10-20k $20-30k $30-40k $4
##                      <chr> <int>   <int>   <int>   <int>
## 1                 Agnostic    27      34      60      81
## 2                  Atheist    12      27      37      52
## 3                 Buddhist    27      21      30      34
## 4                 Catholic   418     617     732     670
## 5       Don't know/refused    15      14      15      11
## 6         Evangelical Prot   575     869    1064     982
## 7                    Hindu     1       9       7       9
## 8  Historically Black Prot   228     244     236     238
## 9         Jehovah's Witness    20      27      24      24
```

Gather dataset

```
pew %>% gather(income, n, -religion) %>% head
```

```
## Source: local data frame [6 x 3]
##
##                religion income     n
##                   <chr>  <chr> <int>
## 1             Agnostic  <$10k    27
## 2              Atheist  <$10k    12
## 3             Buddhist  <$10k    27
## 4             Catholic  <$10k   418
## 5 Don't know/refused  <$10k    15
## 6   Evangelical Prot  <$10k   575
```

income, religion : variables to gather n : variable in cells -religion means all except religion

Using group_by

```
pew %>% gather(income, n, -religion) %>%
  group_by(income) %>% summarise(totals = sum(n))
```

```
## Source: local data frame [10 x 2]
##
##                  income totals
##                   <chr>  <int>
## 1                  <$10k   1930
## 2                  >150k   2608
## 3                $10-20k   2781
## 4              $100-150k   3197
## 5                $20-30k   3357
## 6                $30-40k   3302
## 7                $40-50k   3085
## 8                $50-75k   5185
## 9               $75-100k   3990
```

Using group_by

```
pew %>% gather(income, n, -religion) %>%
  group_by(religion) %>% summarise(totals = sum(n))
```

```
## Source: local data frame [18 x 2]
##
##                   religion totals
##                      <chr>  <int>
## 1                 Agnostic    826
## 2                  Atheist    515
## 3                 Buddhist    411
## 4                 Catholic   8054
## 5       Don't know/refused    272
## 6         Evangelical Prot   9472
## 7                    Hindu    257
## 8  Historically Black Prot   1995
## 9         Jehovah's Witness    215
```

Plotting and Graphing

- Exploratory Data Analysis
- Base graphics
- Intro `ggplot2`
- Saving graphics

Exploratory Data Analysis

- Informal representation data
- Looking for patterns, outliers, etc.

Types of graphs

- Historgram
- Scatterplot
  - Scatterplot matrix
- Boxplots
- Violin plots (`ggplot2`)
- Q-Q plots

par function

- Check parameters for graphing

ggplot2 - Grammar of Graphics

- Different syntax
- Powerful operations

Saving output to file

- Formats
  - PDF
  - SVG
  - PNG/TIFF

End in `dev.off()`

`ggplot2` may require a `print` before it gets rendered in the file.