More `dplyr` and `tidyr`

Haohan Chen[*]

October 5, 2018

Download an example dataset

```
# Just to show how the data is downloaded, in case you are curious.
# Data downloade from ICPSR: https://www.icpsr.umich.edu/icpsrweb/ICPSR/

# if (!"psData" %in% installed.packages()[, 1]){
#   install.packages("psData")
# }
#
# library(psData)
# polity <- PolityGet()
# save(polity, file = "polity.Rdata")
```

Load the data and libraries

```
# Libraries
library(dplyr)
library(tidyr)

# Data
load("polity.Rdata")
polity <- as_tibble(polity)
```

Review: What you have learned

- `select`, `filter`, `arrange`, `mutate`. What do they do?
- Exercise: Get the variables `country` and `polity` of the "Uganda" from 1990-2000, sort it form largest to smallest.

Note: Note that POLITY score captures political regime authority spectrum on a 21-pont scale ranging from -10 (hereditary monarchy) to +10 (consolidated democracy). The Polity scores can also be converted into regime categories in a suggested three part categorization of "autocracies" (-10 to -6), "anocracies" (-5 to +5 and three special values: -66, -77 and -88), and "democracies" (+6 to +10). Performance score from 0 to 100. The highest score reflects the best situation.

```
polity %>% filter(country == "Brazil") %>% select(country, year, polity2) %>% arrange(-polity2)
```

```
## # A tibble: 192 x 3
##    country  year polity2
##    <chr>   <dbl>   <dbl>
##  1 Brazil   1988       8
##  2 Brazil   1989       8
##  3 Brazil   1990       8
##  4 Brazil   1991       8
```

[*] Political Science Department, Duke University. haohan.chen@duke.edu

```
##  5 Brazil   1992       8
##  6 Brazil   1993       8
##  7 Brazil   1994       8
##  8 Brazil   1995       8
##  9 Brazil   1996       8
## 10 Brazil   1997       8
## # ... with 182 more rows
```

New

- group_by, summarise: Get the mean, maximum, minimum, median, standard deviation of `polity2` for each country from 1990-2000.
- slice: Get the first 10 rows
- reshape dataset: long <- wide: Get a wide-form dataset of 2000-2005, Each year as a column

```
polity %>%
  filter(year %in% c(2000:2005)) %>%
  select(country, year, polity2) %>%
  spread(year, polity2)
```

```
## # A tibble: 163 x 7
##    country     `2000` `2001` `2002` `2003` `2004` `2005`
##    <chr>        <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Afghanistan     -7     NA     NA     NA     NA     NA
##  2 Albania          5      5      7      7      7      9
##  3 Algeria         -3     -3     -3     -3      2      2
##  4 Angola          -3     -3     -2     -2     -2     -2
##  5 Argentina        8      8      8      8      8      8
##  6 Armenia          5      5      5      5      5      5
##  7 Australia       10     10     10     10     10     10
##  8 Austria         10     10     10     10     10     10
##  9 Azerbaijan      -7     -7     -7     -7     -7     -7
## 10 Bahrain         -9     -8     -7     -7     -7     -7
## # ... with 153 more rows
```

More information

- **Must read:** https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf
- Some tutorials with examples:
  - https://rpubs.com/bradleyboehmke/data_wrangling
  - http://garrettgman.github.io/tidying/