Wrangling and (more) Data Verbs

Code

Instructor - Soumya Mukherjee

Content Credit- Dr. Matthew Beckman and Olivia Beck

July 17, 2023

Agenda

- Review of data verbs & summary functions
- Discussion of new functions and key concepts
- Guided example highlighting several new data verbs
- RStudio Concept Maps: https://github.com/rstudio/concept-maps (https://github.com/rstudio/concept-maps)

RMarkdown Stuff

R Markdown - Text versus R chunk

When we want to write plain text, we just write as normal.

When we want to use a header we use # before our text

When we want to write R code, we use an R chuck:

Hide

```
x <- 1:10
sum(x)
```

```
[1] 55
```

R Notebook vs R Markdown Comparison

	R Notebook	**R Markdown**
Source document	FileName.Rmd	FileName.Rmd
Typesetting syntax	markdown	markdown
YAML Header	`output: html_notebook`	`output: html_document`
compile button text	“Preview” or “Preview Notebook”	“Knit to HTML” or “Knit”
compile button icon	notebook	blue yarn ball
Output file default	FileName.nb.html	FileName.html
Code chunk handling	Run R code first	Runs R code when rendered
.Rmd source embedded?	**YES**	**NO**

Discussion/Review Task:

- **List 1: data verbs:**
  - `mutate()`
  - `filter()`
  - `select()`
  - `arrange()`
- **List 2: more data verbs:**
  - `head()` & `tail()`
  - `rename()`
  - `sample_n()`
  - `summarise()` & `group_by()`
- **List 3: summary functions:**
  - `glimpse()`
  - `str()`
  - `summary()`
  - `nrow()` & `ncol()`
  - `names()`

- `colnames()`
- `View()`

Three Important Concepts (Again)

1. Data can be usefully organized into tables with “cases” and “variables.” With “tidy data”
    - every row corresponds to a distinct case (e.g. a person, a car, a year, a country in a year)
    - columns represent variables/features of the cases
2. Data graphics and “glyph-ready” data
    - each case corresponds to a “glyph” (mark) on the graph
    - each variable to a graphical attribute of that glyph such as x- or y-position, color, size, length, shape, etc.
    - similar story applies for modeling purposes
3. When data are not yet in glyph-ready form, you can transform (i.e. wrangle) them into glyph-ready form.
    - Such transformations are accomplished by performing one or more of a small set of basic operations on data tables
    - This is the work of “data verbs” and other functions

Presidential Investigation

- Let’s take some of our new tools for a spin…
- The `presidents` data are available from the `tidyverse` package (actually `ggplot2` which was loaded by `tidyverse`)
- we’ll add a few other presidents to the data as well

Hide

```
library(tidyverse)
library(dcData)
library(tibble)     # we'll use `tibble::tribble()`  to specify new rows
# install.packages(lubridate)
library(lubridate)  # nice functions for handling dates

# data intake
data("presidential", package = "ggplot2")

## Add some more cases to the data set
# you can search the help for the `tribble` function if you want to figure it out what's happening
AddPres <-
    tibble::tribble(~name, ~start, ~end, ~party,
                     "Roosevelt", ymd("1933-03-04"), ymd("1945-04-12"), "Democratic",
                    "Truman", ymd("1945-04-12"), ymd("1953-01-20"), "Democratic",
                    "Trump", ymd("2017-01-20"), ymd("2021-01-20"), "Republican",
                    "Biden", ymd("2021-01-20"), ymd(today()), "Democratic")

# append the rows of our two "data frames"
Presidents <- bind_rows(presidential, AddPres)

head(Presidents)
```

name <chr>	start <date>	end <date>	party <chr>
Eisenhower	1953-01-20	1961-01-20	Republican
Kennedy	1961-01-20	1963-11-22	Democratic
Johnson	1963-11-22	1969-01-20	Democratic
Nixon	1969-01-20	1974-08-09	Republican
Ford	1974-08-09	1977-01-20	Republican
Carter	1977-01-20	1981-01-20	Democratic

6 rows

Guided practice

1. various summary/exploration functions to meet the Presidents
2. calculate `duration` of presidency
3. show all records for presidents named “Bush”
4. subset the data to include only `name`, `party`, `duration`
5. which 3 presidents served shortest?
6. which 3 presidents served longest?
7. which presidents served more than 4 years, but less than 8 years?
8. investigating the power balance
    - which party has most presidents?
    - which party spent most days in power?

Answers

Hide

```
# 1. summary functions
glimpse(Presidents)  # what's the benefit?
```

```
Rows: 16
Columns: 4
$ name  <chr> "Eisenhower", "Kennedy", "Johnson", "Nixon", "For…
$ start <date> 1953-01-20, 1961-01-20, 1963-11-22, 1969-01-20, …
$ end   <date> 1961-01-20, 1963-11-22, 1969-01-20, 1974-08-09, …
$ party <chr> "Republican", "Democratic", "Democratic", "Republ…
```

Hide

```
str(Presidents)      # ?
```

```
tibble [16 × 4] (S3: tbl_df/tbl/data.frame)
 $ name : chr [1:16] "Eisenhower" "Kennedy" "Johnson" "Nixon" ...
 $ start: Date[1:16], format: "1953-01-20" ...
 $ end  : Date[1:16], format: "1961-01-20" ...
 $ party: chr [1:16] "Republican" "Democratic" "Democratic" "Republican" ...
```

Hide

```
summary(Presidents)  # not so useful?!
```

```
     name               start                end          
 Length:16          Min.   :1933-03-04   Min.   :1945-04-12  
 Class :character   1st Qu.:1963-03-08   1st Qu.:1967-10-06  
 Mode  :character   Median :1979-01-20   Median :1985-01-20  
                    Mean   :1981-08-07   Mean   :1987-07-01  
                    3rd Qu.:2003-01-20   3rd Qu.:2011-01-20  
                    Max.   :2021-01-20   Max.   :2023-07-16  
    party          
 Length:16          
 Class :character   
 Mode  :character   
```

Hide

```
#View(Presidents)     # console please!


# some wrangling
Presidents_wrangled <-
    Presidents %>%
    mutate(duration = end - start,                                        # 2
           name = if_else(name == "Bush" & start < ymd("2000-01-01"),  # 3
                          true = "Bush (Sr)",
                          false = name)) %>%
    select(name, party, duration)                                         # 4


### some results
#2 Presidency duration
Presidents_wrangled
```

name <chr>	party <chr>	duration <time>
Eisenhower	Republican	2922 days
Kennedy	Democratic	1036 days
Johnson	Democratic	1886 days
Nixon	Republican	2027 days
Ford	Republican	895 days
Carter	Democratic	1461 days
Reagan	Republican	2922 days
Bush (Sr)	Republican	1461 days
Clinton	Democratic	2922 days
Bush	Republican	2922 days

1-10 of 16 rows Previous **1** 2 Next

Hide

```
#3 The Bushes
Presidents_wrangled %>%
  filter(name %in% c("Bush","Bush (Sr)"))
```

name <chr>	party <chr>	duration <time>
Bush (Sr)	Republican	1461 days
Bush	Republican	2922 days

2 rows

Hide

```
# 6 shortest
Presidents_wrangled %>%
    arrange(duration) %>%
    head()
```

name <chr>	party <chr>	duration <time>
Ford	Republican	895 days
Biden	Democratic	907 days
Kennedy	Democratic	1036 days
Carter	Democratic	1461 days
Bush (Sr)	Republican	1461 days

name <chr>	party <chr>	duration <time>
Trump	Republican	1461 days

6 rows

Hide

```
# 7. longest
Presidents_wrangled %>%
    arrange(duration) %>%
    tail()
```

name <chr>	party <chr>	duration <time>
Eisenhower	Republican	2922 days
Reagan	Republican	2922 days
Clinton	Democratic	2922 days
Bush	Republican	2922 days
Obama	Democratic	2922 days
Roosevelt	Democratic	4422 days

6 rows

Hide

```
Presidents_wrangled %>%
    arrange(desc(duration)) %>%
    head()
```

name <chr>	party <chr>	duration <time>
Roosevelt	Democratic	4422 days
Eisenhower	Republican	2922 days
Reagan	Republican	2922 days
Clinton	Democratic	2922 days
Bush	Republican	2922 days
Obama	Democratic	2922 days

6 rows

Hide

```
# 8. more than 4 years, but less than 8 years

Presidents_wrangled %>%
    filter(duration > 1461, duration < 2922)
```

name <chr>	party <chr>	duration <time>
Johnson	Democratic	1886 days
Nixon	Republican	2027 days
Truman	Democratic	2840 days

3 rows

Hide

```
# 9. party balance

Presidents_wrangled %>%
    group_by(party) %>%
    summarise(total = n(),
              in_power = sum(duration))
```

party <chr>	total <int>	in_power <time>
Democratic	8	18396 days
Republican	8	16071 days

2 rows

Things to look through before tomorrow's class (July 18)

- Read Chapters 10 and 11 of Data Computing Ebook (No reading quiz assigned yet, but if you do read beforehand, more helpful for you)
- Start looking through the PopularNames project in the DataComputing Ebook in advance (Will be assigned as an activity some time this week). Link is https://dtkaplan.github.io/DataComputingEbook/project-popular-names.html#project-popular-names (https://dtkaplan.github.io/DataComputingEbook/project-popular-names.html#project-popular-names).