Programming for Professional Research Using R

Session 2

November 2, 2023

Today

- Pop quiz
- Learn how to:
  - Filter, mutate, group, and summarize data using Tidyverse functions
  - Reshape data using Tidyverse functions
  - Check for duplicates and encode missing values
- Be introduced to:
  - "Tidy" datasets and how to create them using pivot_longer() and pivot_wider()
- Practice the above!

Pop Quiz!

https://pollev.com/marcandreafiorina503

mutate() filter() select()

```r
mutate_example <- mtcars %>%
    mutate(
        heavy = case_when(
            wt > 3 ~ "Yes",
            TRUE   ~ "No"
        )
    ) %>%
    select(wt, heavy)

mutate_example %>% head()
```

```
##                      wt heavy
## Mazda RX4         2.620    No
## Mazda RX4 Wag     2.875    No
## Datsun 710        2.320    No
## Hornet 4 Drive    3.215   Yes
## Hornet Sportabout 3.440   Yes
## Valiant           3.460   Yes
```

mutate() | **filter()** | select()

```r
filter_example <- mtcars %>%
    filter(wt > 3)

filter_example %>% head()
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
```

mutate() filter() select()

```r
select_example <- mtcars %>%
    select(
        matches("mpg|carb")
    )
select_example %>% head()
```

```
##                    mpg carb
## Mazda RX4         21.0    4
## Mazda RX4 Wag     21.0    4
## Datsun 710        22.8    1
## Hornet 4 Drive    21.4    1
## Hornet Sportabout 18.7    2
## Valiant           18.1    1
```

group_by() and summarize() | pivot_longer() | pivot_wider()

```
group_by_summarize_example <- mtcars %>%
    group_by(cyl) %>%
    summarize(
        mpg = mean(mpg, na.rm = TRUE)
    )

group_by_summarize_example
```

```
## # A tibble: 3 × 2
##     cyl   mpg
##   <dbl> <dbl>
## 1     4  26.7
## 2     6  19.7
## 3     8  15.1
```

group_by() and summarize() | pivot_longer() | pivot_wider()

```
relig_income[1:6] %>% head(n = 2)
```

```
## # A tibble: 2 × 6
##   religion `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k`
##   <chr>      <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 Agnostic      27        34        60        81        76
## 2 Atheist       12        27        37        52        35
```

```
relig_income_long <- relig_income %>%
    pivot_longer(
        cols      = !religion, # Everything but religion
        names_to  = "levels",
        values_to = "num"
    )

relig_income_long %>% head(n = 4)
```

```
## # A tibble: 4 × 3
##   religion levels    num
##   <chr>    <chr>   <dbl>
## 1 Agnostic <$10k      27
## 2 Agnostic $10-20k    34
## 3 Agnostic $20-30k    60
## 4 Agnostic $30-40k    81
```

group_by() and summarize()   pivot_longer()   **pivot_wider()**

```
fish_encounters %>% head(n = 4)
```

```
## # A tibble: 4 × 3
##   fish  station  seen
##   <fct> <fct>   <int>
## 1 4842  Release     1
## 2 4842  I80_1       1
## 3 4842  Lisbon      1
## 4 4842  Rstr        1
```

```
fish_encounters_wide <- fish_encounters %>%
    pivot_wider(
        names_from  = station,
        values_from = seen
    )

fish_encounters_wide[1:6] %>% head(n = 2)
```

```
## # A tibble: 2 × 6
##   fish  Release I80_1 Lisbon  Rstr Base_TD
##   <fct>   <int> <int>  <int> <int>   <int>
## 1 4842        1     1      1     1       1
## 2 4843        1     1      1     1       1
```

Data Cleaning

From the DIME World Bank wiki:

Data cleaning is an essential step between data collection and data analysis. Raw primary data is always imperfect and needs to be prepared for a high quality analysis and overall replicability. […] [I]n the vast majority of cases, data cleaning requires significant energy and attention, typically on the part of the Research Assistant (RA).

What to look out for

Without data cleaning, you might end up with analysis that is either biased or fully inaccurate.

Thinks that RAs usually have to check for:

- **Uniquely and fully identified dataset** -- no duplicates, no missing IDs. Each row should have a unique identifier.

- **Survey codes and missing values**

    - Most survey software will make you have to code categorical answers numerically -> e.g. "yes" is 1, "no" is 0.
    - In that framework, other possible answers that we don't want to analyze (e.g. "I don't know") also need to be coded numerically. But we can't keep them that way because they'll bias mean/sum aggregations.
    - SOLUTION -- convert to missing, i.e. NA

What to look out for

Without data cleaning, you might end up with analysis that is either biased or fully inaccurate.

Thinks that RAs usually have to check for:

- **Illogical values** -- questionnaires should follow a specific logic but good to check that there hasn't been a breakdown. e.g. a fully empty column that should have answers, or responses that don't make sense (e.g. 2 year old child with a full-time job).
- **Multiple choice answers** -- most survey softwares store multiple-choice answers in the same value (e.g. "1 2 3 4"), which makes them hard to use in data work. Good practice to "split" out the answers into individual variables.
- **Labels** -- cleaning is also the stage at which variables are given descriptive labels, usually in a codebook.

What to look out for

Key thinking: each dataset/survey will have unique issues/problems that cannot always be predicted ahead.

One of the hardest tasks for a RA is to *think criticallly* about what could go wrong with raw data and identify errors. It takes time and thoroughness but is essential to do good data work.

Working with 'Tidy' Datasets

```
table1
#> # A tibble: 6 x 4
#>   country      year  cases population
#>   <chr>       <int>  <int>      <int>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583

table2
#> # A tibble: 12 x 4
#>   country      year type            count
#>   <chr>       <int> <chr>           <int>
#> 1 Afghanistan  1999 cases             745
#> 2 Afghanistan  1999 population   19987071
#> 3 Afghanistan  2000 cases            2666
#> 4 Afghanistan  2000 population   20595360
#> 5 Brazil       1999 cases           37737
#> 6 Brazil       1999 population  172006362
#> # … with 6 more rows
```

```
table3
#> # A tibble: 6 x 3
#>   country      year rate
#> * <chr>       <int> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583

# Spread across two tibbles
table4a  # cases
#> # A tibble: 3 x 3
#>   country     `1999` `2000`
#> * <chr>        <int>  <int>
#> 1 Afghanistan    745   2666
#> 2 Brazil       37737  80488
#> 3 China       212258 213766
table4b  # population
#> # A tibble: 3 x 3
#>   country         `1999`     `2000`
#> * <chr>            <int>      <int>
#> 1 Afghanistan   19987071   20595360
#> 2 Brazil       172006362  174504898
#> 3 China       1272915272 1280428583
```

These are all useable versions of the same data. Only one of them, however, is 'tidy'.

What makes a dataset 'tidy'? From Hadley Wickham & Garrett Grolemund, *R for Data Science* Chapter 12 -- Tidy Data:

1. Each **variable** must have **its own column**.
2. Each **observation** must have **its own row**.
3. Each **value** must have **its own cell**.

Easier to thing about when these conditions are *not* met:

- When one variable is spread across multiple columns.

- When one observation is scattered across multiple rows.

Practical Exercise -- Using the World Values Survey Dataset

World Values Survey

Background

"The survey, which started in 1981, seeks to use the most rigorous, high-quality research designs in each country. The WVS consists of nationally representative surveys conducted in almost 100 countries which contain almost 90 percent of the world's population, using a common questionnaire. [...] WVS seeks to help scientists and policy makers understand changes in the beliefs, values and motivations of people throughout the world."

Survey Contents

- Social values, attitudes & stereotypes
- Societal well-being
- Social capital, trust and organizational membership
- Economic values
- Corruption
- Migration
- Post-materialist index
- Science & technology
- Religious values
- Security
- Ethical values & norms
- Political interest and political participation
- Political culture and political regimes
- Demography

Today's practical component

1. Successfully run the code in the `session_2_template.R` script
2. Attempt the challenge at the bottom of the script: find the 5 most popular answers that people gave about what is important to teach their children.

Today's practical component

3. Create your own script and do one or more of the following:

- Clean variables Q1-Q6 or variables Q18-Q26 from the `norms_values_data` dataset. i.e. Check for duplicate IDs and convert non-relevant answers such as "don't know" or "refused to respond" as NA
- Create a 'tidy' version of a dataset containing either Q1-Q6 or Q18-Q26. This means that the values should all be in one column called "life" for Q1-Q6 or "neighbor" for Q18-Q26.
- For "life", find the 2 most important things in life for the survey's respondents. For "neighbor", find the 5 things that respondents would least want a neighbor to do.

NOTE You should refer to documentation for the dataset, which can be found in `SAIS R Course/documentation/`, for details on the variables and their given values.

Links

Syllabus: **https://mfiorina.github.io/sais_r_course/syllabus/r_course_syllabus.html**

Session 1: **https://mfiorina.github.io/sais_r_course/session_1/session_1.html**

DIME World Bank Wiki, **https://dimewiki.worldbank.org/Data_Cleaning**

Hadley Wickham & Garrett Grolemund, **R for Data Science Chapter 12 -- Tidy data**

RStudio, **RStudio Cheatsheets**