Use R to fix up a data file

IN-AIR Workshop: Friday, February 16, 2023

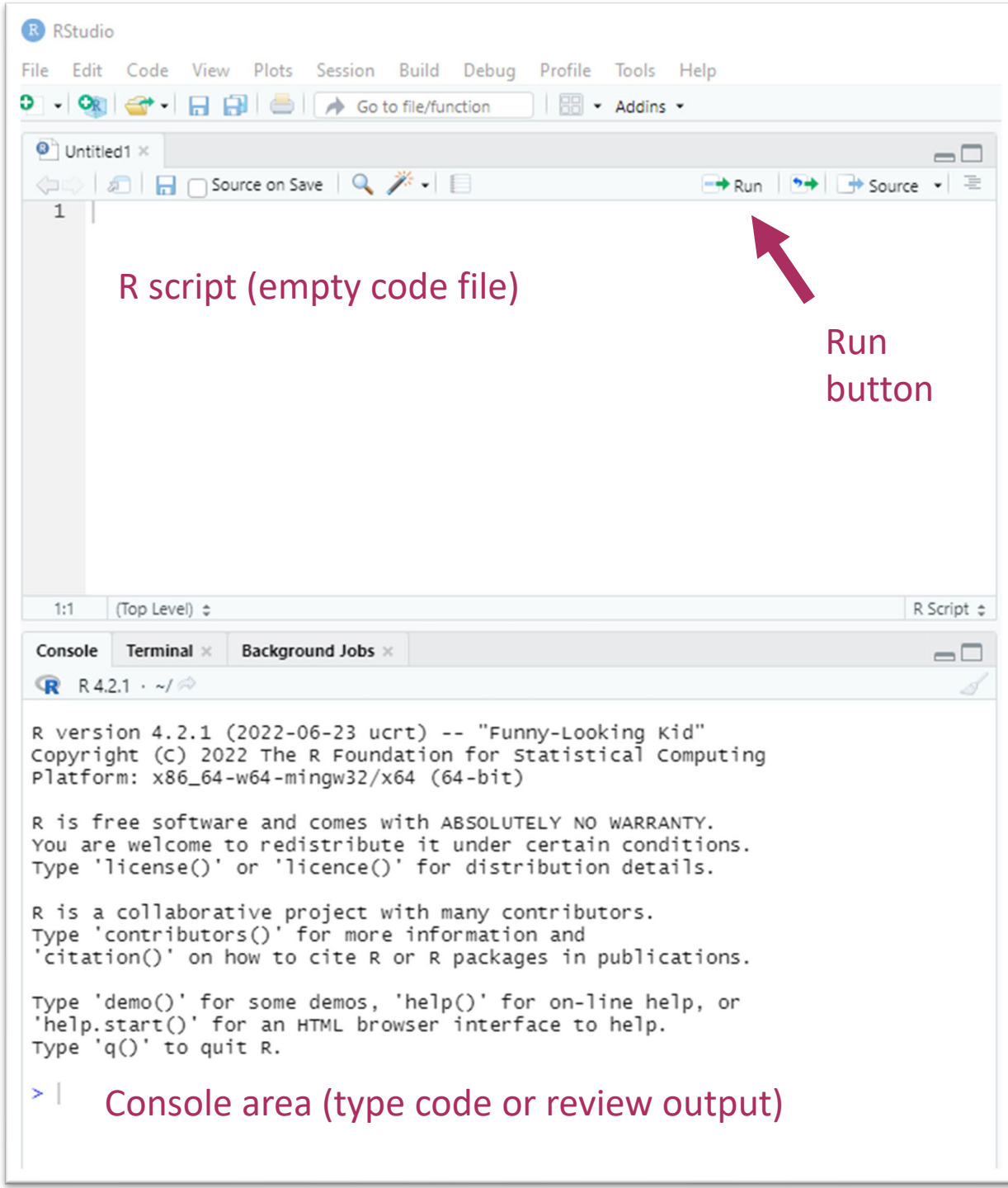
Import, reshape, and recode some data

Try a cool function to make it IPEDS-compatible

Slide deck by Alison Lanski (alanski@nd.edu)

RStudio Reminder

- Two options to run code
  - Type code in an empty R script. At the start/end of a line, hit Ctrl-Enter or Cmd-Enter. (You can also highlight and use the Run button)
  - Click into the console, type your code, and hit Enter



Open Rstudio, load packages

Note: **`this format`** is used for code that you should run

are used for comments within the code

```
#load packages
library(readr)  #for reading in files
library(dplyr)  #for manipulating data
library(IPEDSuploadables) #for prepping the upload file
```

Read in data

Situation: You have some data in a csv file that may need more preparation

Solution: Read the csv data into R; use R code to fix it

```
# your file location may be different
# include the extension .csv
# R needs this slash direction /

dat <- read_csv("C:/Users/alanski/Downloads/SampleDataPrepCOM.csv")
```

air forum

Explore: See the general structure of the data

```
# using our existing file, for now...
colnames(dat)  #very helpful when you're writing code later

# three ways to explore overall
str(dat)
glimpse(dat)
View(dat)

# can filter within "View" pane or see the top rows with
head(dat)
```

air forum

Explore: See all possible values in a column

The pipe `%>%` connects pieces of code that should be run in order

The result of each piece of code is fed into the next piece of the pipe

```
#general example
dat %>% distinct(columnname)

#specific example
dat %>% distinct(Sex)
```

If you want all possible values *and* frequencies

```
#general example
dat %>% count(columnname)

#specific example
dat %>% count(Sex)
```

air forum

To prep this data for IPEDS submission, we need to….

- Add a column for the unitid
- Update Sex and GenderDetail information
- Recode student degree level and Race/Ethnnicity
- Create a combined column with a complete 6 digit CIP code
- Remove non-graduates
- Create Distance Ed 31/32 info based on this rule:
    - Location: 1 = mandatory onsite
    - Location: 2 = not-mandatory onsite
    - Location: 3 = mandatory and non-mandatory onsite
    - Location: 4 = nothing onsite

Basic column updates in R

```
# rename a column
dat <- dat %>% rename(new_column_name = old_column_name)

# change a datatype
dat <- dat %>% mutate(column_name = as.numeric(column_name))

dat <- dat %>% mutate(column_name = as.integer(column_name))

dat <- dat %>% mutate(column_name = as.character(column_name))
```

air forum

Select/Remove things

```
# remove a column
dat <- dat %>% select(-column_name)

# remove rows by rule
dat <- dat %>% filter(column_name == value)

# remove 100% duplicated rows
dat <- dat %>% distinct()  #nothing inside those ( )
```

air forum

Change/add columns with `mutate()`

```
# set a fixed value
dat <- dat %>% mutate(col_name = value)

# do arithmetic
dat <- dat %>% mutate(col_name = other_col*that_col)

# round numbers
dat <- dat %>% mutate(col_name = round(col_name, 1)
```

Change/add columns with `mutate()`

```
# replace Null values with something else
dat <- dat %>% mutate(col_name = replace_na(col_name, 0))

# combine columns into a string
dat <- dat %>% mutate(col_name = paste0(other_col,' ', 'JR'))

# take one portion of a string
dat <- dat %>% mutate(col_name = substr(other_col, 1, 3))
```

Two-prong conditions within `mutate()`

```
# ifelse function
dat <- dat %>%
     mutate(col_name = ifelse(other_col == 1,'Yes','No'))

# case_when function
Dat <- dat %>%
     mutate(col_name = case_when(other_col == 1 ~ 'Yes',
                                 TRUE ~ 'No'))
```

Multi-prong conditions with `case_when()`

```
# line breaks added for clarity
dat <- dat %>% mutate(column_name = case_when(
            other_col == 1 ~ 'PhD',
            other_col == 2 ~ 'MA',
            other_col == 3 & that_col = 'Science' ~ 'BS',
            other_col == 4 & that_col = 'Humanities' ~ 'BA',
            TRUE ~ 'No Degree'))

# I like to check when done (did I mess up?)
dat %>% count(column_name, other_column, that_col)
```

air forum

We can write multi-function multi-line code

```
dat <- dat %>%
      mutate(col20 = 'banana',
             col21 = substr(col20, 2, 3),
             col22 = ifelse(col04 == 'Yes' & col05 == 'Yes', 'UG', 'GR'),
             col23 = case_when(col22 == 'UG' ~ 5,
                               col22 == 'GR' & col06 == 'LLD' ~ 18,
                               col22 == 'GR' & substr(col06, 1, 2) == 'M' ~ 7,
                               col22 == 'GR' & substr(col06, 1, 2) == 'P' ~ 17,
                               TRUE ~ 999)) %>%
      mutate(col22 = ifelse(col22 == 'GR' & col06 == 'LLD, 'PR', col22))
```

air forum

Summary of functions so far

Exploration:

`colnames()`	`str()`	`glimpse()`	`View()`
`head()`	`count()`	`distinct()`	

Preservation/Removal:

`select()`	`filter()`	`distinct()`

Changes within a `mutate()`:

`round()`	`paste0()`	`replace_na()`
`as.numeric()`	`as.integer()`	`as.character()`
`ifelse()`	`case_when()`	

air forum

How did we fix the data?

Final code is in this file: SampleComPrepDataSolution.R

Create the final IPEDS submission file

```
# this function is part of the IPEDSuploadables package
# it will generate a popup window asking you where to
save the file
```

```
produce_com_report(new_dat)
```

Need more R help?

Some general resources:

R Cheatsheets: https://posit.co/resources/cheatsheets/

R4DS: https://r4ds.had.co.nz/

R4DS slack channel: http://r4ds.io/join

R-ladies slack channel: https://guide.rladies.org/comm/slack/

See the meeting attachment for more suggestions

IPEDSuploadables documentation: https://alisonlanski.github.io/IPEDSuploadables/

air forum