Lecture 13: Data Import

STAT 450, Fall 2021

Many of the data sets we have worked with in the class (e.g., `mpg`, `flights`) have been provided by R packages. For today's lecture we will discuss how read in a data set from a file on your computer.

`readr` is a tidyverse package that contains many useful functions for reading in data into R.

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

`read_csv()`

Comma-separated values (CSV) files are one of the most common types of data files that you will encounter. Each entry on each line of a CSV file is separated by a comma.

FiveThirtyEight Data Set

As an example, we will work with a data set called `hate_crimes.csv`. The data set was used for an article from the popular data journalism website FiveThirtyEight:

Higher Rates Of Hate Crimes Are Tied To Income Inequality

I've placed the data set on Blackboard. The authors of the article originally made it available on a repository on GitHub:

https://github.com/fivethirtyeight/data/tree/master/hate-crimes

The GitHub repository has descriptions of the variables.

A central question the article tries to address is "Why do some states see many more reported hate incidents than others?" To answer this question, the authors collected demographic data on education, diversity, and economic health for each state. They then explored the relationship between those variables and rates of hate crime incidents.

To read the hate_crimes.csv data set into the Desktop version of R (installed on you computer), follow these steps:

1. Create an R script or R Markdown file and save the file in a folder on your computer. Download the hate_crimes.csv data set and place it in the same folder.
2. In the RStudio menu go to Session → Set Working Directory → To Source File Location. This will set your working directory to the folder containing the data set, and R script or R Markdown file. You can check your working directory by typing getwd() into the console.
3. Type the following command to read in the data set from file:

```
hate_crimes <- read_csv("hate_crimes.csv")
```

```
## Rows: 51 Columns: 12

## -- Column specification ---------------------------------------------------------
## Delimiter: ","
## chr  (1): state
## dbl (11): median_household_income, share_unemployed_seasonal, share_populati...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

You'll notice that when you run read_csv() it prints out a column specification that gives the name and type of each column. We can also use the glimpse() function to preview the data set and check the different column types:

```
glimpse(hate_crimes)
```

```
## Rows: 51
## Columns: 12
## $ state                                    <chr> "Alabama", "Alaska", "Arizona~
## $ median_household_income                  <dbl> 42278, 67629, 49254, 44922, 6~
## $ share_unemployed_seasonal                <dbl> 0.060, 0.064, 0.063, 0.052, 0~
## $ share_population_in_metro_areas          <dbl> 0.64, 0.63, 0.90, 0.69, 0.97,~
## $ share_population_with_high_school_degree <dbl> 0.821, 0.914, 0.842, 0.824, 0~
## $ share_non_citizen                        <dbl> 0.02, 0.04, 0.10, 0.04, 0.13,~
## $ share_white_poverty                      <dbl> 0.12, 0.06, 0.09, 0.12, 0.09,~
## $ gini_index                               <dbl> 0.472, 0.422, 0.455, 0.458, 0~
## $ share_non_white                          <dbl> 0.35, 0.42, 0.49, 0.26, 0.61,~
## $ share_voters_voted_trump                 <dbl> 0.63, 0.53, 0.50, 0.60, 0.33,~
## $ hate_crimes_per_100k_splc                <dbl> 0.12583893, 0.14374012, 0.225~
## $ avg_hatecrimes_per_100k_fbi              <dbl> 1.8064105, 1.6567001, 3.41392~
```

As an alternative, at step 2, go to Session → Set Working Directory → Choose Directory. Then you can manually specify your work directory to the folder containing the data set.

To read `hate_crimes.csv` into R Studio Cloud, follow these steps:

1. Download the `hate_crimes.csv` data set, and place on your Desktop.
2. Click on the Upload button in the Files pane, then click on Choose File and select `hate_crimes.csv`.
3. Type the following command to read in the data set:

```
hate_crimes <- read_csv("hate_crimes.csv")
```

Exercises:

1. Use the procedure described on the previous page to read the `hate_crimes.csv` data set into R.
2. Compute some summary statistics for `avg_hatecrimes_per_100k_fbi`
3. Which states had the highest hate crime rates? Which states had the lowest hate crime rates? [Hint: use `arrange()`]
4. Use `ggplot()` to make a scatter plot with `share_unemployed_seasonal` on the x-axis and `avg_hatecrimes_per_100k_fbi` on the y-axis. Use `geom_smooth()` to add a smooth trend line to scatter plot. Label the x-axis "Seasonally adjusted unemployment", and the y-axis Average hate crimes per 100,000 residents, 2010-2015". Describe any trends in the scatter plot.