



R Ladies DC

Open Data Day 2017

Mini Hackathon for Dataset Creation/Curation

Brought to you by:





Upcoming Events with R Ladies DC

Thursday, April 20th @ The Data Incubator DC
DuPont Circle

Shiny + Databases
Christina Brady

Upcoming Study Groups with R Ladies DC

Beginner R Study Group

POC: **Anna Yakovleva!**

- Twice a Month
- An Arlington Library?

Proposed Material:

Exploratory Data Analysis with R - Roger Peng

R Programming for Data Science - Roger Peng

Advanced R Study Group

POC: **Christina Brady!**

- Once a Month
- SE Library DC

Proposed Material:

Work through Hadley Wickham's **Advanced R** book (available for free on his website)

R and Data Science Journal Club

POC: **Kelly O'Briant (me)**

- Once a Month
- South Block Juice Co.

Proposed Material:

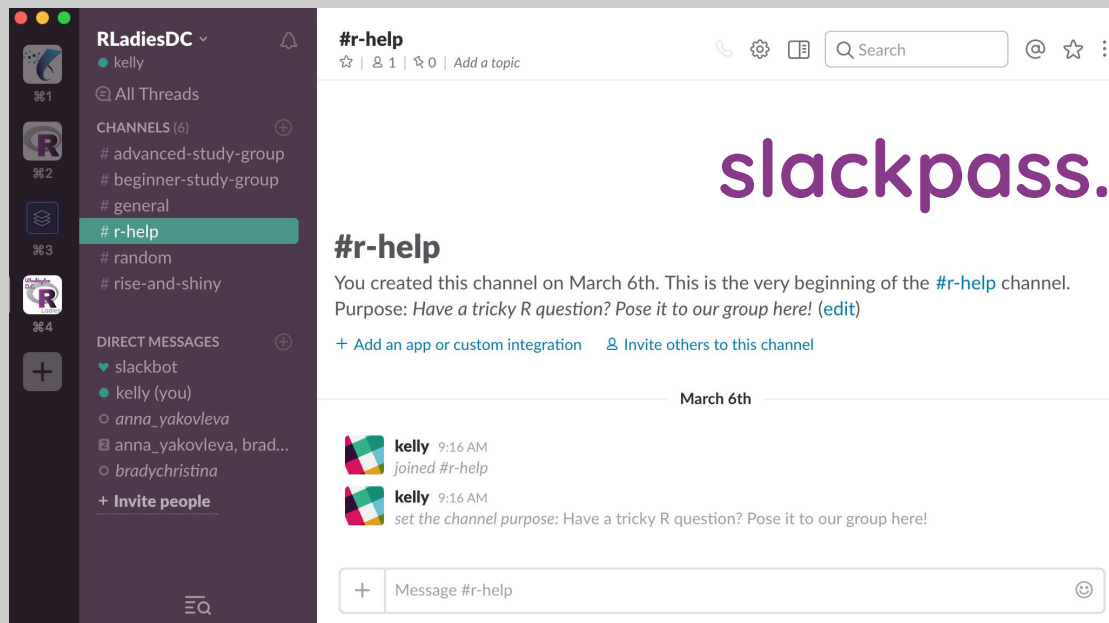
Discuss one new (or old) article from the R Data Science community

Get exposure to aspects of data science you have yet to explore!



You're Invited!

R Ladies DC Slack Team



slackpass.io/rladiesdc

Schedule for Today

- 1:00** Welcome Slide Deck (Happening Right Now)
- 1:15** Tutorial: Rvest/Web Scraping (Right After This)
- 1:45** Round Table Pitch Fest - Propose Projects
- 2:00** Form Teams based on project interest
Work in teams to creating and curate a dataset
- 3:30** Tutorial: Data.World dataset management through Github or Google Sheets
- 3:45** Work a little more
- 4:45** Drinks and sharing!

Goals For Today

- **Meet fellow R Ladies, Make friends, Have fun**
- **Seek out data that exists in a difficult to analyze format and make it tidy**
- (optional) Practice using R to collect and tidy this data
- **Describe and share your tidy data with the public on Data.World**
- (optional) Create an analysis or visualization of your new dataset to share with us over drinks

What Can I Do as a Total R Beginner?

- **Join a team with some non-beginners on it**
 - Ask questions!
 - Use alternative query tools (*More on this later...*)
 - Do internet research - find data worth scraping
 - Organize raw data in Google Sheets or Excel
 - Have your team teach you some R basics for importing, summarizing and visualizing data
- **The task is more important than the tool**

R is not required!

Transitioning into the First Tutorial

Be thinking about data topics you might like to work on

I'll end with some example data project ideas and a walk-through of data.world

We'll go straight into round table pitching and group formation after the tutorial



Rvest Web Scraping Tutorial

R-Ladies DC
March 25, 2017

Install rvest

(and possibly some other packages)

```
install.packages("rvest")  
library(rvest)
```

```
install.packages("plyr")  
library(plyr)
```

```
install.packages("dplyr")  
library(dplyr)
```

The whole shabang!

```
install.packages("tidyverse")  
library(tidyverse)
```

[Tidyverse Blog Link](#)

Fun with Warnings!

You have loaded plyr after dplyr - this is likely to cause problems.

If you need functions from both plyr and dplyr, please load plyr first, then dplyr:

```
library(plyr); library(dplyr)
```

Anatomy of an rvest query (with dplyr)







```
web_page <- read_html(url)

var_name <- web_page %>%
  html_nodes("css selection") %>%
  component_reference %>%
  component_extraction
```

[Link to a pretty in-depth tutorial](#)

My Very Tables-Centric rvest Tutorial

Laureates [\[edit \]](#) VIEW

Year	Image	Laureate	Country	Category	Rationale
1903		Marie Skłodowska Curie (shared with Pierre Curie and Henri Becquerel)	Poland and France	Physics	"In recognition of the extraordinary services they have rendered by their joint researches on the radiation phenomena discovered by Professor Henri Becquerel ^[6]
1905		Bertha von Suttner	Austria–Hungary	Peace	Honorary President of Permanent International Peace Bureau, Bern, Switzerland; Author of <i>Lay Down Your Arms</i> . ^[8]
1909		Selma Lagerlöf	Sweden	Literature	"In appreciation of the lofty idealism, vivid imagination and spiritual perception that characterize her writings" ^[10]
1911		Marie Skłodowska Curie	Poland and France	Chemistry	"for her discovery of radium and polonium" ^[11]
1926		Grazia Deledda	Italy	Literature	"for her idealistically inspired writings which with plastic clarity picture the life on her native island and with depth and sympathy deal with human problems in general" ^[12]
1928		Sigrid Undset	Norway	Literature	"principally for her powerful descriptions of Northern life during the Middle Ages" ^[13]

Wikipedia Example
[List of Female Nobel Laureates](#)

JUDO RESULTS			
RESULTS MEN		RESULTS WOMEN	
U60		U48	
1	Phelipe Pelim BRA	1	Gabriela Chibana BRA
2	Bernard Azinovic CRO	2	Katelyn Bouyssou USA
3	Jose Luis Galvan ARG	3	Erin Morgan CAN
3	Yann Siccardi MON	3	Keisy Perafan ARG
5	Michael Patino PER	5	Lesley Cano PER
5	Jose Ramos GUA	5	Mary Dee Vargas Ley CHI
7	Kobe Chavez CHI	7	Loreto Montano CHI
7	Julio Molina GUA		
U66		U52	
1	Antoine Bouchard CAN	1	Eleudis Valentim BRA
2	Charles Chibana BRA	2	Abi Betsabe Cardozo Madaf ARG
3	Fernando González ARG	3	Brillith Gamarra Carbajal PER
3	Juan Hernandez COL	3	Judith Gonzalez CHI
5	Juan Perez CHI	5	Fabiola Rojas CHI
5	Ryan Vargas USA	5	Nicol Vera CHI
7	Lucas Ambrosio ARG	7	Leandra Barraza CHI
7	Gueorgui Poklitar CAN		
		U57	
		1	Gimena Laffeuillade ARG

Judo Example
[Many Tables to Scrape](#)

~Live Code~

Link to the code in R Notebook form for later reference

More Projects to Explore for Inspiration and Code Examples:

VISIT THE HACKPAD: bit.ly/rladies-rvest

- Documentation Readme for rvest
- 2017 Retail Store Closing Analysis (rvest)
- Scraping CRAN with rvest
- Twitter API data projects

...

Alternatives to rvest

SPARQL

<https://query.wikidata.org/>

Random Internet Tools?

<http://www.hongkiat.com/blog/web-scraping-tools/>

Import Directly Into Google Docs:

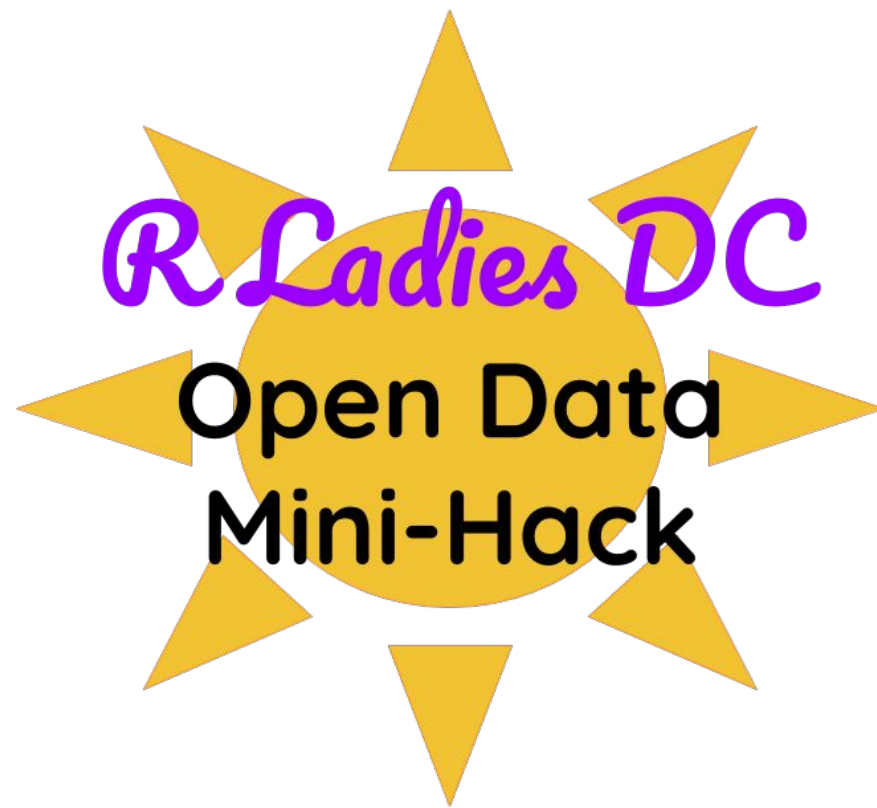
<https://www.labnol.org/internet/import-html-in-google-docs/28125/>

Data.World

[Sign up for an account](#) if you haven't already!

Data.World Legal Stuff:

“Data and who owns it and what owners want done with it can be complicated. You are responsible for figuring it out before you share or use the data.”



Hackpad Link:
bit.ly/rladies-rvest

Slack Team Sign up
Link:

slackpass.io/rladiesdc