Getting Started with R for Mass Spectrometry Data Analysis

1-Day Short Course • 2020 ASMS Annual Meeting

Ryan Benz
Seer, Inc.

Jeffrey Jones
SoCal Bioinformatics

N. Heath Patterson
Vanderbilt School of Medicine

Learn the fundamentals of R and add new data analysis skills to your scientific tool belt!

Course Overview

This course will cover the fundamentals of the R programming language and how it can be used to perform data analysis tasks, with a focus on the practical essentials to get you started on your way to becoming an effective R user

Target Audience

Anyone new to R (and perhaps coding) who wants to learn the practical fundamentals for data analysis work

Prerequisites

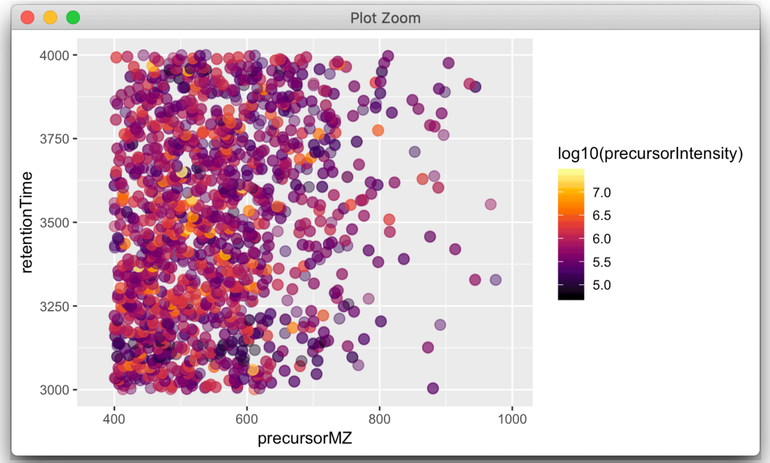
- Desire to learn one of the best data analysis tools around
- Willing to try something new and challenging
- No prior coding experience needed
- Bring a laptop for the example exercises and working sessions

The course will focus on 3 main topics

R Fundamentals

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
>
> dim(mtcars)
[1] 32 11
>
> mean(mtcars$mpg)
[1] 20.09062
>
> mtcars$hp / mtcars$mpg
 [1]  5.238095  5.238095  4.078947  5.140187  9.358289  5.801105 17.132867
 [8]  2.540984  4.166667  6.406250  6.910112 10.975610 10.404624 11.842105
[15] 19.711538 20.673077 15.646259  2.037037  1.710526  1.917404  4.511628
[22]  9.677419  9.868421 18.421053  9.114583  2.417582  3.500000  3.717105
[29] 16.708861  8.883249 22.333333  5.093458
>
```

Data Visualization with ggplot2

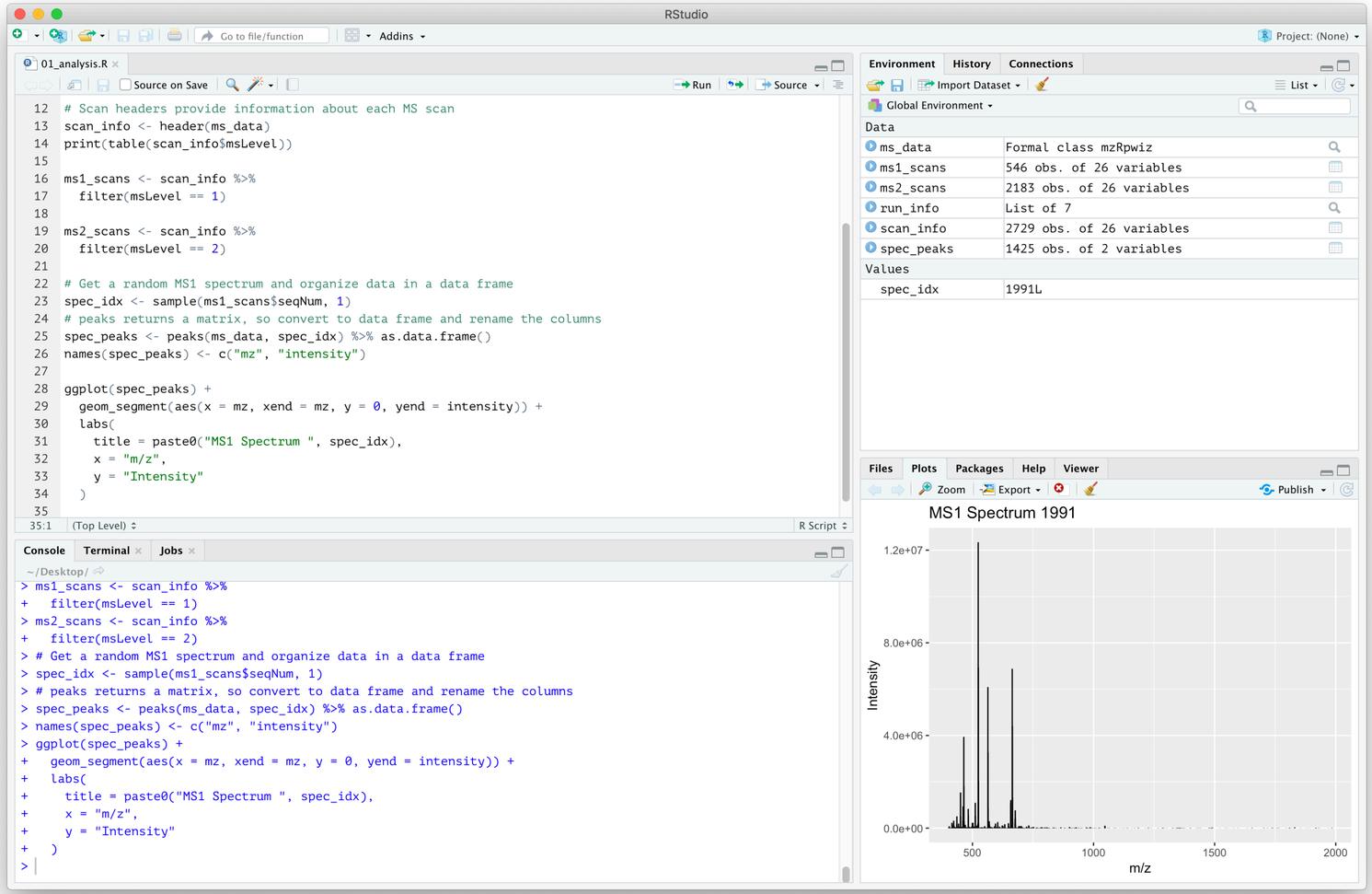


Data Analysis with the tidyverse

```
scan_info %>%
  filter(msLevel == 2) %>%
  select(precursorCharge, precursorIntensity) %>%
  group_by(precursorCharge) %>%
  summarize(mean_intensity = mean(precursorIntensity),
            sd_intensity = sd(precursorIntensity),
            median_intensity = median(precursorIntensity)) %>%
  arrange(precursorCharge)
```

```
# A tibble: 3 x 4
  precursorCharge mean_intensity sd_intensity median_intensity
            <int>          <dbl>        <dbl>            <dbl>
1               2         873799.     1589582.           457223.
2               3         953256.     1619545.           514319.
3               4         784594.     1068287.           413015.
```

Get started with R and RStudio, learn best practices for data analysis with R

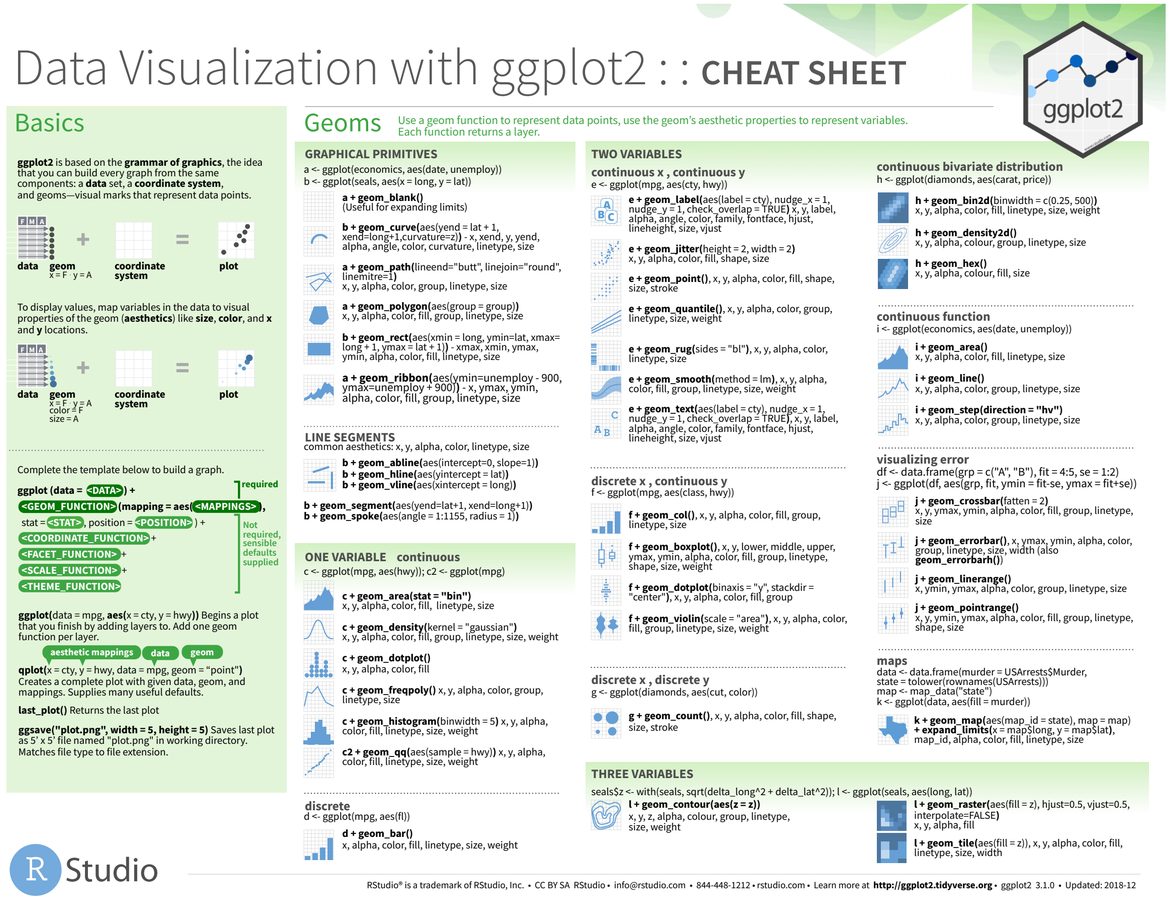


- RStudio provides an integrated environment for coding
  - code editor
  - interactive terminal
  - variables, plots, documentation
  - version control
  - *and much more*

- Streamlines all aspects of the data analysis process

- Learn best practices for organizing analyses into projects and make your work reproducible

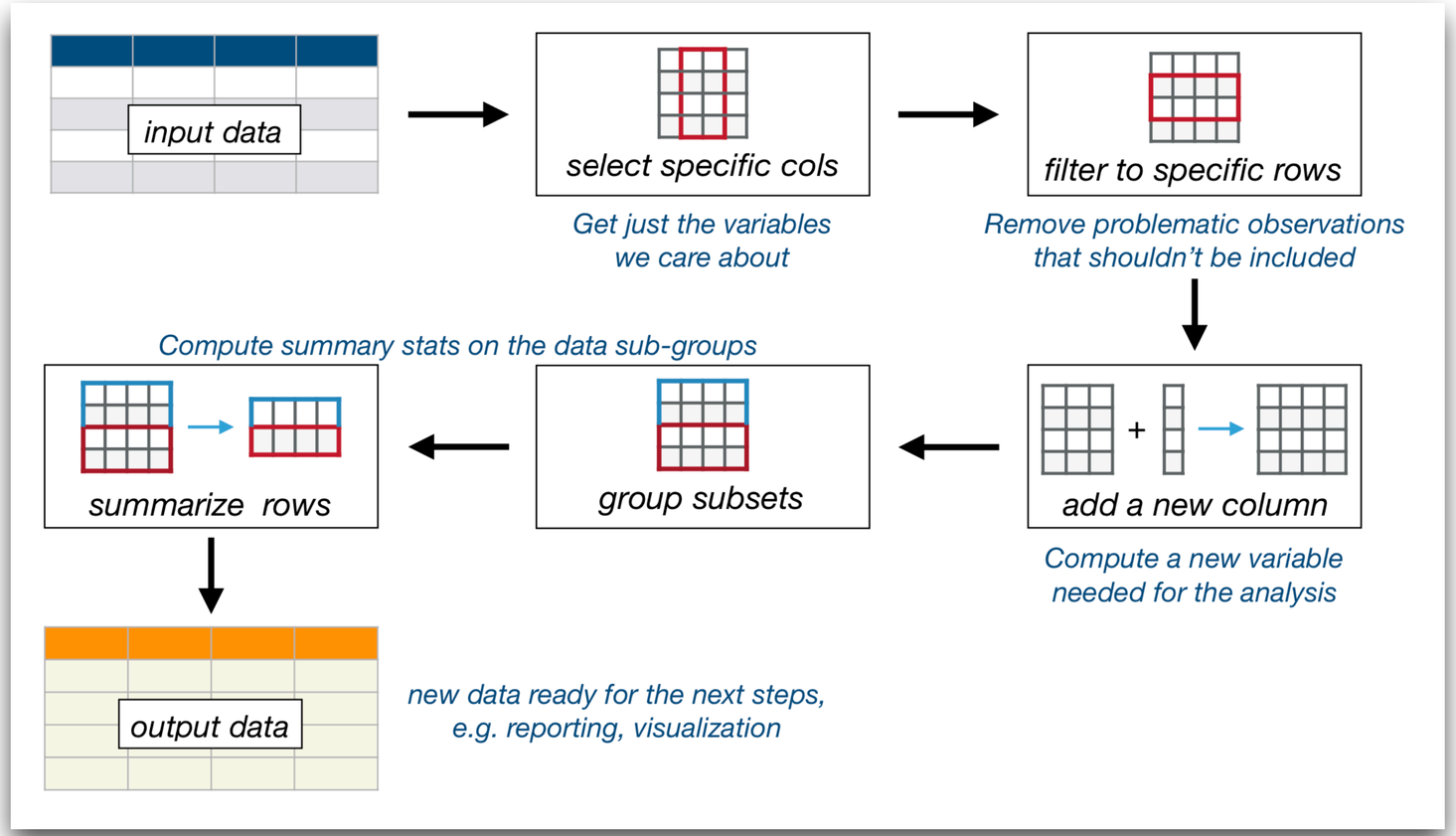
Start making plots and visualizations from your own data



https://github.com/rstudio/cheatsheets/blob/master/data-visualization-2.1.pdf

- ggplot2 is a powerful plotting package in R that allows you to easily make plots from your data
- Provides the fundamental pieces to make all kinds of visualizations, from simple to complex
- Quickly make plots for exploratory data analysis
- Fine tune the look and style of your plots, make them ready for presentation and publication

Learn about tidy data and how the tidyverse collection of packages can streamline your analyses



- Learn how tidy data and the tidyverse can make your data analysis code easier to read, write and understand
- Learn about dplyr, an R package that allows you to create streamlined data analysis pipelines
- The tidyverse covers all steps of the analysis process, from data import & cleaning to visualization and final reporting

Additional course info…

- This course is designed to help participants get over the initial hurdles new R users often face
- Mass spectrometry and related data sets / examples will be used during the course to help put the material into context
- Information on any required pre-course set-up (i.e. software to install) will be provided in advance
- If you've been thinking about trying to learn R, this course is a great way to get started