AIMS Skillshare

Introduction to R

Jayashree Raman & Kevin McCraney
November 2, 2018

Anyone need help?

Installations you should have right now:

- R (from https://www.r-project.org)
- RStudio (from https://www.rstudio.com)
- LaTeX (from https://www.latex-project.org/get))
- A sample knit file (either HTML/PDF)
- Project files (https://tinyurl.com/Rskillshare)

Agenda

We will:

- Introduce ourselves and talk about the goals of the session
- Give an overview of R, and why it’s useful for data science
- Clean some data
- Transform data depending on on what we need
- Visualize some data
- Show you where to go for more knowledge
- Talk about using R in industry (special guest appearance)

Who are we?

Who is this session for?

People that:

- Have some exposure to programming (1+ language(s) or so) and are looking to learn
- Are interested in learning iteratively
- People curious about doing data science statistical analysis but have limited exposure
- Don’t have any of the attributes described but aren’t afraid of confusing error messages

What are we NOT doing?

We won’t be teaching you:

- In-depth R usage (just basics to get your feet wet)
- Statistics (we’ll talk about the basics, but that’s it)
- How to program like a pro (just type in commands and try stuff; use StackOverflow)
- How to do machine learning/deep learning (if there’s enough interest, we could do a session next quarter)

An overview of today’s material.

We will be learning how to:

- Load and view datasets in R
- Clean messy data sets (this is 70% of data science in most cases)
- Use dplyr to remove outliers and unnecessary values, filter and summarize functions
- Use ggplot2 package to produce some barplots, scatterplots, and stacked bar charts
- Import .R files to your RMarkdown files
- Knit (generating an HTML/PDF report with your results)

...and some handy, nifty tricks and tons of resources for further learning.

How will we work? The 3 D’s!

Dialogue

Demonstration

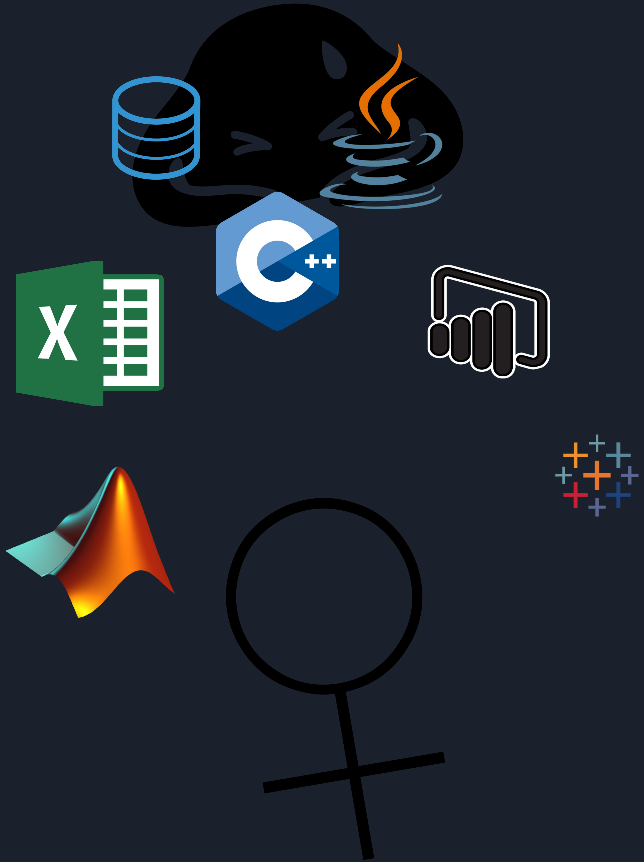
Digestion

Part I:

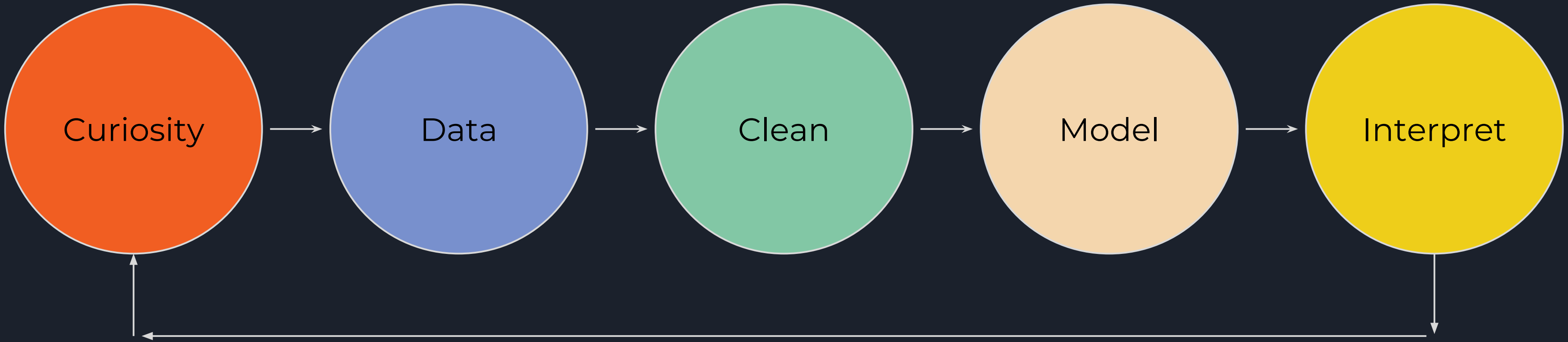
Why R?

So what is R, exactly?

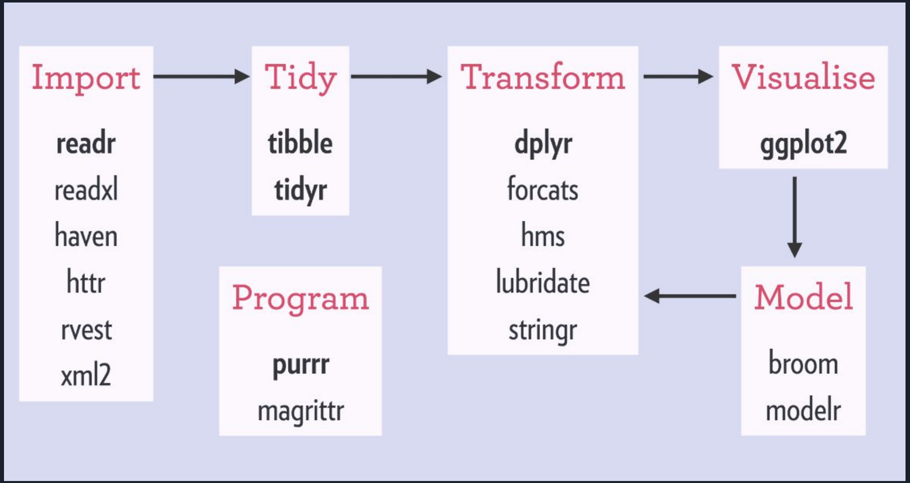
Why R and not ?



So, again, why R? The data science pipeline.



An aside: the **tidyverse**.



Import

readr

readxl

haven

httr

rvest

xml2

Tidy

tibble

tidyr

Transform

dplyr

forcats

hms

lubridate

stringr

Visualise

ggplot2

Model

broom

modelr

Program

purrr

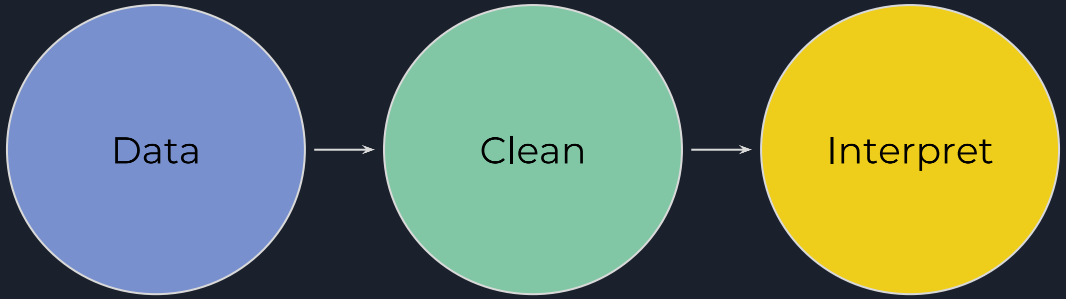
magrittr

The tidyverse is a philosophical programming system for working with R.

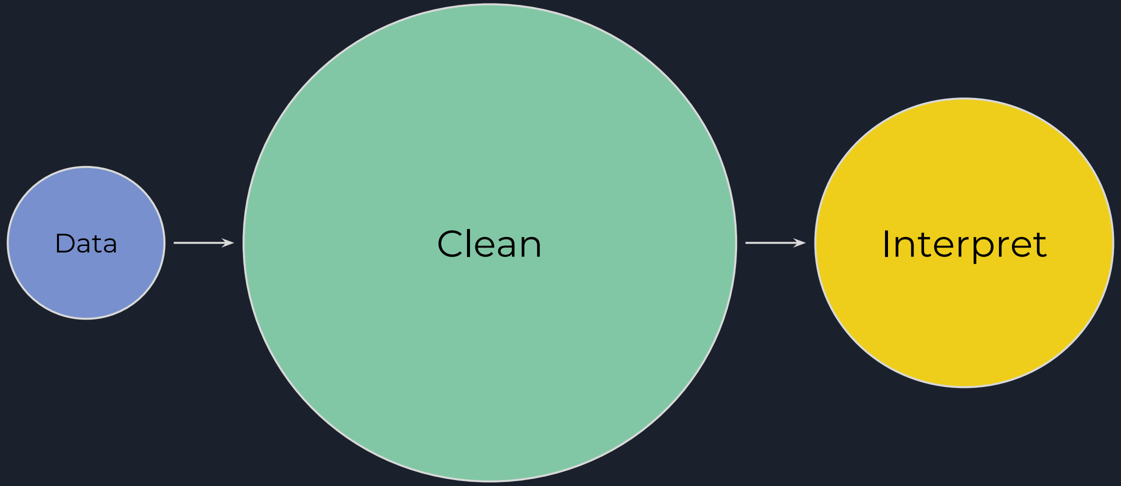
It is a series of different packages (including dplyr and ggplot2, the packages we're focusing on today) that work together to make data processing and analysis much easier.

Check out https://www.tidyverse.org for other tutorials and resources.

What we’re doing today.



Today, proportionally.



But first... conceptual structures and syntax!

- R has the same built-in `datatypes` as other languages (`str, int, float, boolean`).
- R has variables, but does `assignment` (setting a variable equal to something) weird.

Every other language:	`=`
R:	`<-`

- `Vectors` are a sequence of data `c(1, 2, 3)`. A vector is like a `list` or an `array`.
- `Dataframes` are a list of variables in row-column format, like a spreadsheet.
- Dataframes can be `indexed` (that is, changed based on a logical operation or condition).
- A `function` is a block of code that performs an operation.
- `Libraries` are collections of (usually complex) functions other people wrote.

Part II:

Data Cleaning

What’s our dataset?

It’s Halloween, so chocolate, of course!

Get it here: https://tinyurl.com/Rskillshare

How do we get our data into R?

getwd()/setwd()   or in the GUI of RStudio

```
data <- read.csv(“...chocolate.csv”,   # load CSV & assign to variable
                 header=True)          # no header (column names)
```

And now... live coding!

- We’re going to look at our data using head() , tail() , and other functions of note.
- We're going examine just one column, subset a dataframe, and look at NA values.
- Finally, we will get a count of NA values and do something with them.

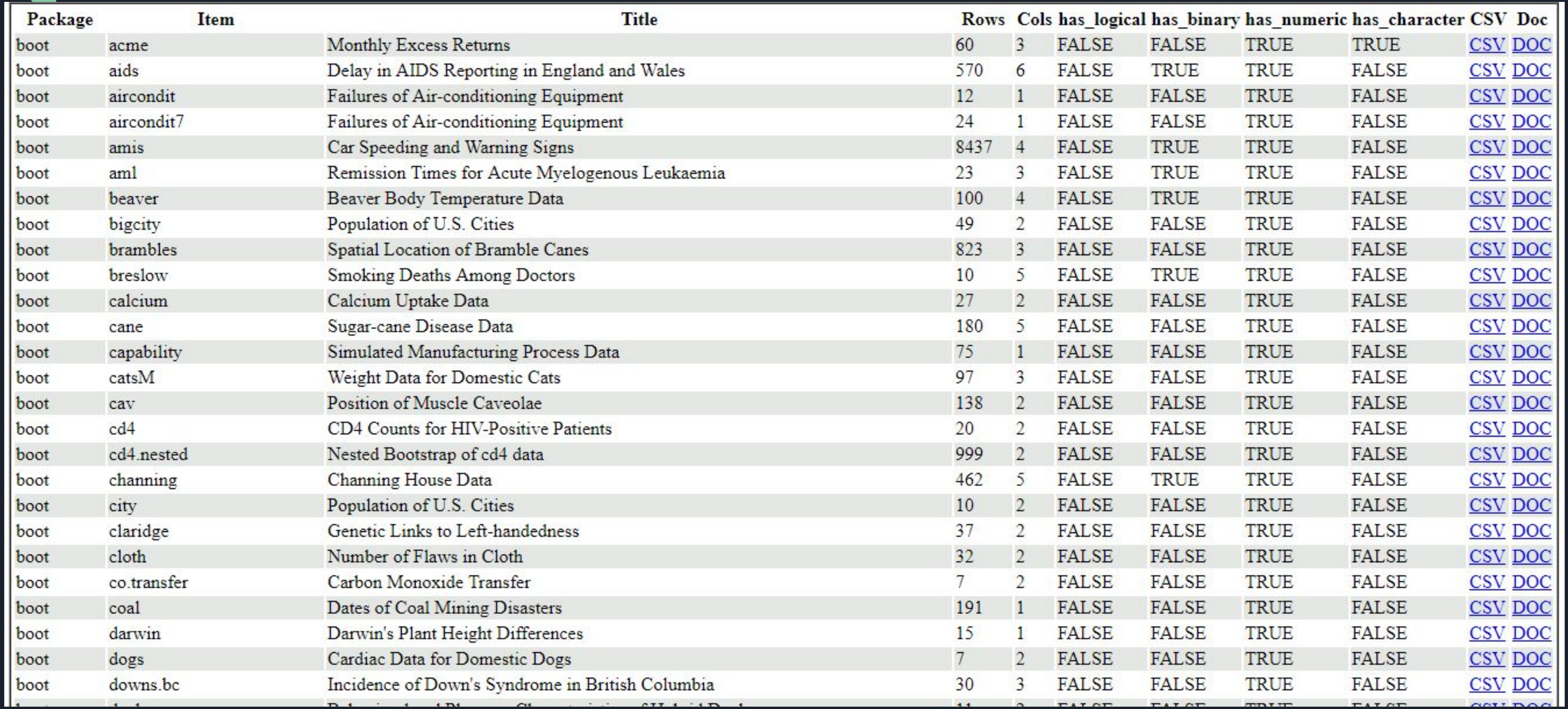
Your turn!

- Find the `class` of of each of the first 5 columns
- Rename Column #8 to “Rating”
- Find how many “Amazon Ratings” there are, and the average (mean)
- Take the Deliciousness columns and replace all the -1s with NAs
- Bonus: Make a new column, calling it whatever you like, and use a built-in function to fill it with random numbers.

Wrangling data with dplyr

Inbuilt datasets

https://vincentarelbundock.github.io/Rdatasets/datasets.html



Package	Item	Title	Rows	Cols	has_logical	has_binary	has_numeric	has_character	CSV	Doc
boot	acme	Monthly Excess Returns	60	3	FALSE	FALSE	TRUE	TRUE	CSV	DOC
boot	aids	Delay in AIDS Reporting in England and Wales	570	6	FALSE	TRUE	TRUE	FALSE	CSV	DOC
boot	aircondit	Failures of Air-conditioning Equipment	12	1	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	aircondit7	Failures of Air-conditioning Equipment	24	1	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	amis	Car Speeding and Warning Signs	8437	4	FALSE	TRUE	TRUE	FALSE	CSV	DOC
boot	aml	Remission Times for Acute Myelogenous Leukaemia	23	3	FALSE	TRUE	TRUE	FALSE	CSV	DOC
boot	beaver	Beaver Body Temperature Data	100	4	FALSE	TRUE	TRUE	FALSE	CSV	DOC
boot	bigcity	Population of U.S. Cities	49	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	brambles	Spatial Location of Bramble Canes	823	3	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	breslow	Smoking Deaths Among Doctors	10	5	FALSE	TRUE	TRUE	FALSE	CSV	DOC
boot	calcium	Calcium Uptake Data	27	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	cane	Sugar-cane Disease Data	180	5	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	capability	Simulated Manufacturing Process Data	75	1	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	catsM	Weight Data for Domestic Cats	97	3	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	cav	Position of Muscle Caveolae	138	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	cd4	CD4 Counts for HIV-Positive Patients	20	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	cd4.nested	Nested Bootstrap of cd4 data	999	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	channing	Channing House Data	462	5	FALSE	TRUE	TRUE	FALSE	CSV	DOC
boot	city	Population of U.S. Cities	10	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	claridge	Genetic Links to Left-handedness	37	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	cloth	Number of Flaws in Cloth	32	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	co.transfer	Carbon Monoxide Transfer	7	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	coal	Dates of Coal Mining Disasters	191	1	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	darwin	Darwin's Plant Height Differences	15	1	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	dogs	Cardiac Data for Domestic Dogs	7	2	FALSE	FALSE	TRUE	FALSE	CSV	DOC
boot	downs.bc	Incidence of Down's Syndrome in British Columbia	30	3	FALSE	FALSE	TRUE	FALSE	CSV	DOC

How do we get our data into R?

Today we will look at the nassCDS data
Kevin says it’s depressing, I say it’s worth exploring to discover hidden insights.
(Or maybe I’m just a dark person ¯_(ツ)_/¯)

Loading a dataset that came with a library (i.e. ggplot2):

```
library(ggplot2)          # loads the library

data(“mpg”)               # loads dataset using data() and stores it in variable mpg
```

- **Install the DAAG package**
- **Load the DAAG package**
- **data(“nassCDS”)**
- **View(“nassCDS”)**

The DPLYR package

dplyr is a library for transforming data

- filter - filter a set of rows
- select - select a set of columns
- arrange - arrange/sort the rows
- mutate - add a “mutated” version of a column or columns
- summarize - summarize a column
- group_by - group rows by the value of some column or columns

The DPLYR package : Analogical to SQL

- filter - WHERE
- select - SELECT
- arrange - ORDER BY
- mutate - ADD a NEW COLUMN
- summarize - ‘AS’
- group_by - GROUP BY

- Which of the following are *not* equivalent?
  a. nassCDS[24,4] and nassCDS[24, "airbag"]
  b. nassCDS[24,4] and nassCDS[4, 24]
  c. nassCDS[24,4] and nassCDS$airbag[24]

- Which of the following are *not* equivalent?
  a. nassCDS[24,4] and nassCDS[24, "airbag"]
  b. nassCDS[24,4] and nassCDS[4, 24]
  c. nassCDS[24,4] and nassCDS$airbag[24]

nassCDS[24,4] => none (Levels: none, airbag)

nassCDS[4, 24] => NULL

Filter the data to find the number of accidents with male drivers and female drivers

Filter the data to find the number of accidents with males/females

Filter the data to find the number of accidents with drivers older than or aged 40 years

Wrangle your data frame so the records/rows are sorted in descending order of the age of the **male drivers/passengers.**

Select the dvcat and ageOFocc columns only

Wrangle your data frame so the records/rows are sorted in descending order of the age of the **male drivers/passengers.**

Select the dvcat and ageOFocc columns only

What was the year in which the earliest accident was recorded? And the latest?

Use the arrange() function to find the answers to both.

Display the airbag, seatbelt, yearacc columns only. Is there a noticeable trend in the seatbelt/airbag status?

Using the summarize function, find the average age of female drivers who were in an accident

Using the summarize function, find the average age of female drivers who were in an accident

Using the summarize function, find the count of passengers who did not have their seatbelts on.

What were the different injury levels that female drivers suffered, and how many in each level?

What were the different injury levels that female drivers suffered, and how many in each level?

What were the different speed ranges in which male drivers were driving and how many in each range?

What are summary statistics?

Summary or descriptive statistics are basic statistics like mean, median, standard deviation, and quartiles.

You can access these using summary(dataframe). Alternatively, describe(dataframe), which is defined in the psych package.

Functions to find these out for a list of values:

- mean(nassCDS$ageOFocc)
- median(nassCDS$ageOFocc)
- sd(nassCDS$ageOFocc)

What is an outlier?/Removing Outliers

- Outliers are observations that are wayyyyy different than the other observations.
- They occur naturally, but could be due to error as well.
- You can define outliers differently depending on the kind of work you are doing.
- The definition varies from data set to data set, but you should always ensure you remove outliers after considerable reasoning, and not just because you want your results to fit the model/your expectation.

Part III:

Visualizations

ggplot2: Concepts

ggplot2 is based on the “grammar of graphics”.

- Components of ggplot2

  ggplot(data, aes(x=, y=)) +

      geom_bar()/geom_point()/geom_boxplot()....

      +...

- https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

ggplot2: Barplots, Scatterplots, Stacked Bars

- Barplot to visualize the speed at which male drivers were driving
- Boxplot to visualize number of drivers in an accident in 1997 vs 2001
- Stacked barplot to see the distribution of injury severity by speed
- Dot Plot to visualize age vs injSeverity

Knitting your results in RMarkdown

- Source your .R file

```
source(‘analysis.R’)
```

- Call your variables/plots in code chunks

```
‘ ’ ’{r}

// your code here

‘ ‘ ‘
```

- Add comments in Markdown Syntax (https://www.markdownguide.org/cheat-sheet/)
- Once you have all the necessary content, go ahead and hit ‘Knit’

Courses that use R

Data Science I

Social Network Analysis

Population Health Informatics

Resources for future learning.

R for Data Science : https://r4ds.had.co.nz/

Software Carpentry’s Introduction to R: http://swcarpentry.github.io/r-novice-inflammation/

swiRl - Learn R, in R: https://swirlstats.com

Rstudio Tutorials: https://www.rstudio.com/online-learning/

Datasets: https://data.fivethirtyeight.com/, https://registry.opendata.aws/ , datasets()

MOOCs: https://www.coursera.org/learn/r-programming

Thank You!