github.com/MarketBridge/tidyverse-training

Please visit this URL and follow the instructions there.

Preparing for success

- Please exit Outlook
- Please exit Teams

This workshop is about 1/2 day long, with a break in the middle. There is only a very short window for us to cover this material. If you are in the middle of something urgent, please feel free to keep Outlook and Teams open. Otherwise, please close it during the workshop.

Introduction to the Tidyverse

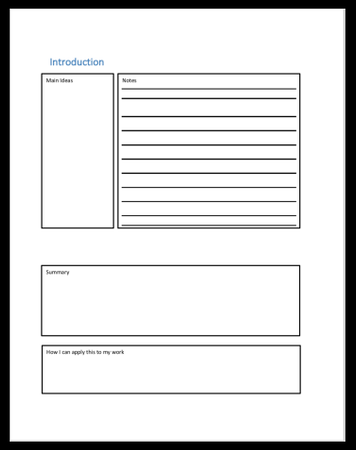
Ari Lamstein

This is originally a 2-day workshop on the Tidyverse. When I taught it at MB the first time I just did the first day, split over two half-days. This time we'll be adding in some more material, and doing 3 half-days.

Exercise: What do you already know?

- Split into groups of 2.
- Ask you partner: *What do you already know about the Tidyverse?*
- 5 minutes

There is an educational theory that people enter trainings already knowing quite a bit about the subject. Additionally, people learn better if they start out by talking about what they already know. So I'm going to split you into groups of two, and ask you to ask your partner what they already know about the Tidyverse.



Notes form - please take out

Research has shown that people retain information better if they take notes while learning. I've given everyone a notes form. It's in a specific format called "Cornell Notes" (it was created by a professor at Cornell University).

Please take it out and write notes as we go on. At the end of this module, I will give you 5 minutes to fill out the summary section.

About Me

- Ari Lamstein
- Studied Math and CS
- Former "tech worker"
- Author of the *choroplethr* package
- Certified Tidyverse Instructor
- Now do R stuff at MB

Oftentimes students want to know a bit about me at the start of a course.

Day 1

Introduction and Visualize Data	9:00 - 10:30
Morning Break	10:30 - 11:00
Transform Data	11:00 - 12:00
Lunch	12:00 - 1:00
Transform Data	1:00 - 2:30
Afternoon Break	2:30 - 3:00
Transform Data, Import Data, Q&A	3:00 - 5:00

I mentioned earlier that this is the first day of what's normally a 2 day course. We're going to be focusing on using the two most important packages in the tidyverse, and we're going to use data that's already been cleaned for us. Day 2 of this course focuses, in large part, on how to clean data using the tidyverse. Think manipulating strings, dates and so on. Those are skills that are necessary to "really" use R for "real" work.

Learning Goals

- What is the Tidyverse (and why learn it?)
- Complete a project with R Notebooks

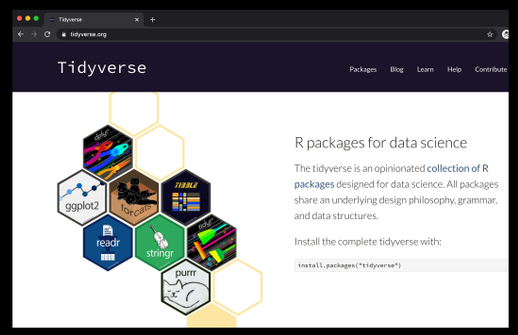
We're in the "Introduction" module of the course now. It's very short, and I have just two goals.

First, I want you to understand what, exactly, the Tidyverse is, and why it's worth learning.

Then, I want you to complete a project using R Notebooks, which I think will be new to most of you. We'll be using R Notebooks throughout the course.

What is the Tidyverse?
Answer 1: A Marketing Term

- Refers to dozens of packages created by Hadley Wickham (Chief Data Scientist at RStudio)
  - Replaces the "Hadley-verse"
- 2 main packages
  - ggplot2 (data visualization)
  - dplyr (data manipulation)



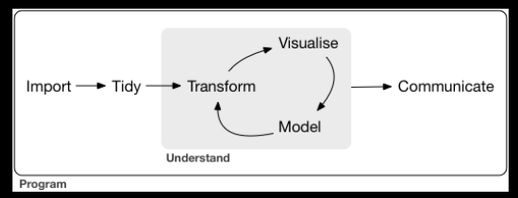
Fundamentally, the Tidyverse is a group of packages that were developed by Hadley Wickham, the Chief Data Scientist at RStudio. I used many of these packages before the term "Tidyverse" was even coined. In fact, before the Tidyverse, people who used his packages joked that they were using the "Hadley-verse"!

My only problem with the term Tidyverse is that is hides a very important point: while there might be dozens of packages in the Tidyverse, they are not all equally important. In fact, I'd argue that you really only need to know how to use two of the packages: ggplot2 and dplyr. We're going to focus on those packages today.

Throughout the course I will be explaining the Tidyverse in more nuance. But for now I really do want you to focus on your job today: learn these two R packages, ggplot2 and dplyr.

What is the Tidyverse?
Answer 2: A "Dialect of R"

- Most R packages don't "agree" with each other
  - ?View vs. ?summary
- All Tidyverse packages are consistent
- Tidyverse packages cover all aspects of the "Data Science Workflow"



Hadley once told me that he thinks of the Tidyverse as a "Dialect" of R. What did he mean by that?

Hadley co-authored a book about the Tidyverse called "R for Data Science". It had this diagram, which shows what he called the "Data Science Workflow": import, tidy, transform, etc.. The Tidyverse has packages that touch all of these aspects of the Workflow. This means that, if you are using the Tidyverse, you might not need to use "Base R" at all!

Why might this be valuable? Well, R is a package-based language. But not all the packages agree with each other. For example, the "V" in ?View is capitalized, and the "s" in ?summary is lowercase. Going back and forth between packages can sometimes be confusing.

All the packages in the Tidyverse agree on details like this, so there is less friction when moving back and forth between packages.

What is the Tidyverse?
Answer 3: A Package that Contains Packages

- "tidyverse" is a name of an actual R package
- The package itself does not do much
- Installing the package makes R install the most useful packages in the Tidyverse (such as ggplot2 and dplyr)

tidyverse: Easily Install and Load the 'Tidyverse'

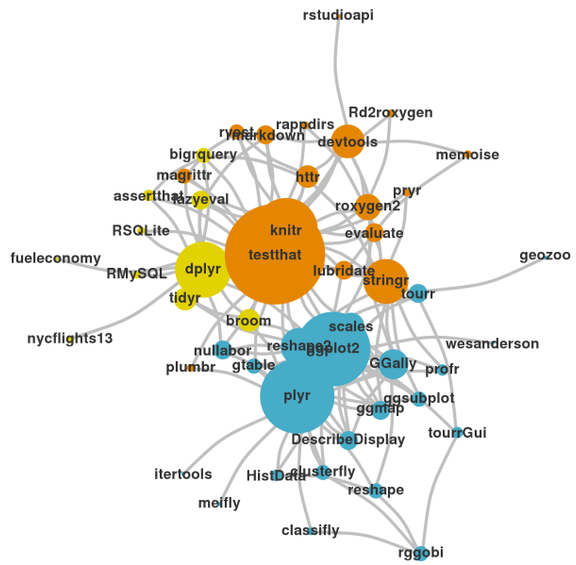
The 'tidyverse' is a set of packages that work in harmony because they sha
'tidyverse' packages in a single step. Learn more about the 'tidyverse' at <h

Version:	1.3.0
Depends:	R (≥ 3.2)
Imports:	broom (≥ 0.5.2), cli (≥ 1.1.0), crayon (≥ 1.3.4), dbplyr 1.6), lubridate (≥ 1.7.4), magrittr (≥ 1.5), modelr (≥ 0. 0.10), rvest (≥ 0.3.5), stringr (≥ 1.4.0), tibble (≥ 2.1.3)
Suggests:	covr, feather, glue, knitr, rmarkdown, testthat
Published:	2019-11-21
Author:	Hadley Wickham [aut, cre], RStudio [cph, fnd]
Maintainer:	Hadley Wickham <hadley at rstudio.com>

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

does the equivalent of

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("hms")
install.packages("stringr")
install.packages("lubridate")
install.packages("forcats")
install.packages("DBI")
install.packages("haven")
install.packages("httr")
install.packages("jsonlite")
install.packages("readxl")
install.packages("rvest")
install.packages("xml2")
install.packages("modelr")
install.packages("broom")
```



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

does the equivalent of

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("hms")
install.packages("stringr")
install.packages("lubridate")
install.packages("forcats")
install.packages("DBI")
install.packages("haven")
install.packages("httr")
install.packages("jsonlite")
install.packages("readxl")
install.packages("rvest")
install.packages("xml2")
install.packages("modelr")
install.packages("broom")
```

```
library("tidyverse")
```

does the equivalent of

```
library("ggplot2")
library("dplyr")
library("tidyr")
library("readr")
library("purrr")
library("tibble")
```

?library only loads a small subset

What is the Tidyverse?

Answer 4: Packages that work with "Tidy Data"

- ggplot2 and dplyr only work with "tidy data"
- Tidy data = a data frame where:
  - Each variable is in its own column
  - Each case is in its own row
  - Each value is in its own cell

Subject (car)	**Speed** (mph)	**Distance** (ft)
1	25	85
2	24	120
3	24	93
4	24	92

Learning Goals

- What is the Tidyverse (and why learn it?)
- Complete a project with R Notebooks

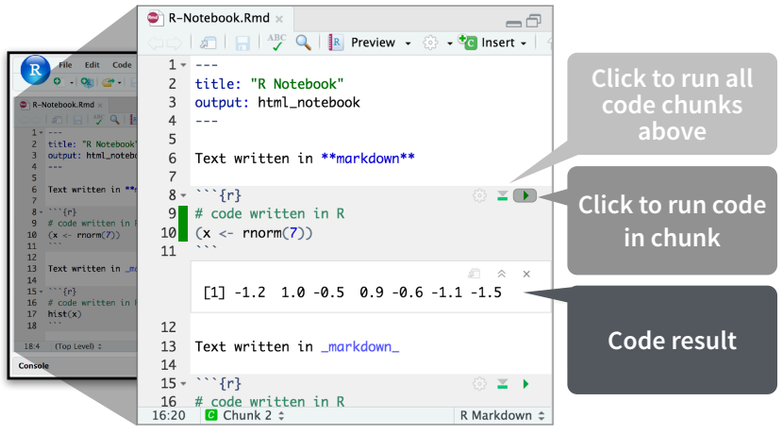
That's my explanation for what the Tidyverse is, and why you should learn it.

Before I go on: are there any questions?

Now let's start our first project with R Notebooks, which we'll be using throughout this course.

R Notebooks

An authoring format for Data Science.



The key point about R Markdown and R Notebooks is that they mix plain text and R code. The R code is in the ```{r} ... ``` parts, and it has a different background. If you hit the "play" button the code is executed, and you can see its result below.

Exercise: Your First R Notebook

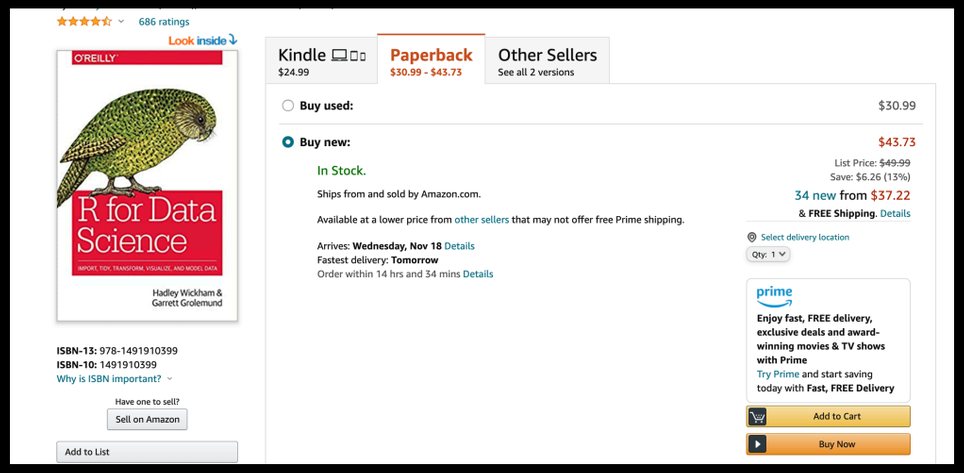
- Open up **00-Introduction.Rmd**
- Read through the notebook and follow all the instructions.
- Ask your partner (or me) if you have any questions.
- Share your result with your partner.
- 5 minutes

What next?

Using R after the workshop

Here's my promise to you. If you follow all the exercises in this workshop you'll be able to import data into R, and do basic transformation and visualizations of it. But working with "real" data in the "real" world is more complicated. Going from "workshop knowledge" to "real world knowledge" is a big jump.

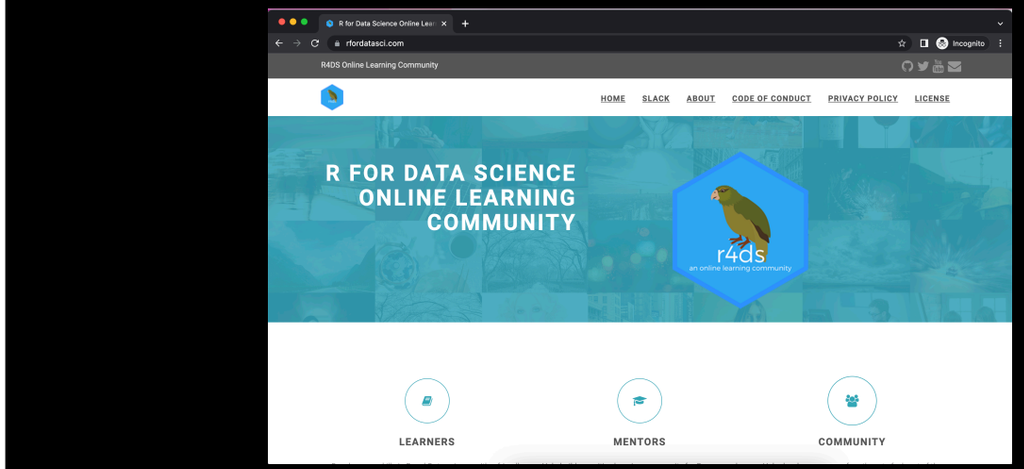
After the workshop feel free to ask me or anyone else at the company who's used R for a long time (Dan, Sam, Katrina) for help.



R for Data Science

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

I also want to point out that there are many resources for learning the Tidyverse. The most popular is the book "R for Data Science" that Hadley and Garrett wrote. The book is also freely available online.



R for Data Science Online Learning Community

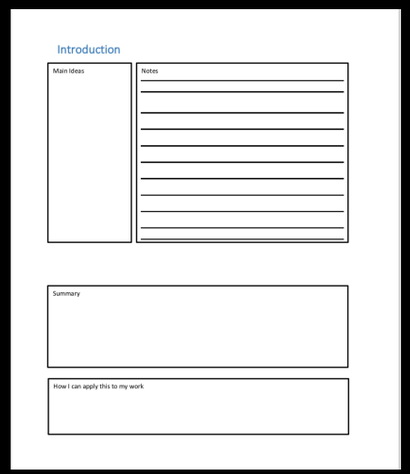
https://www.rfordatasci.com/

I've also heard of the "R for Data Science Online Learning Comunity", although I haven't used it myself. It is apparently free, and has mentors and learners going through the book together.

Closing Exercise

- Fill out the "Summary" section
- Fill out the "How I can apply to to my work"
- Share with your neighbor!

02:00



Any Questions?

In the next section, we're going to start learning to visualize data with ggplot2.

Before we do that, are there any questions?