Teaching Data Science to new useRs

Mine Çetinkaya-Rundel
Duke University + RStudio

bit.ly/user2017

@minebocek
mine-cetinkaya-rundel
mine@stat.duke.edu
I'm interested in stats, but I have no idea where to get started!
I'm interested in stats, but I have no idea where to get started!

Sta 101
I'm interested in stats, but I have no idea where to get started!

Calculus
I'm interested in stats, but I have no idea where to get started!

Probability
I'm interested in stats, but I have no idea where to get started!

Regression
I'm interested in stats, but I have no idea where to get started!

Intro to data science and statistics
a course that provides a common (gateway) experience to students wanting to get started with stats, and that is

- modern
- places data front and center
- quantitative (but not mathematical)
- different than HS stats
- challenging (but not intimidating)
this course should...

emphasize modern and multivariate EDA + data visualization

start at the beginning of data analysis cycle with data collection and cleaning

encourage + enforce working collaboratively (think, code, write, present)

teach (not just expect) reproducible computation

approach statistics from a model based perspective

underscore effective communication of findings
and maybe more importantly...

- ask questions that students want to answer
- equip students with the tools to answer questions of their own choosing
this course doesn’t yet exist, but…
Better Living Through Data Science:
Exploring / Modeling / Predicting / Understanding

Combines techniques from statistics, math, computer science, and social sciences, to learn how to use data to understand natural phenomena, explore patterns, model outcomes, and make predictions. Case studies include examples from election forecasts, movie reviews, and online dating match algorithms. Discussions around reproducibility, data sharing, data privacy will accompany these case studies. Gain experience in data wrangling and munging, exploratory data analysis, predictive modeling, and data visualization, and effective communication of results. Course will focus on R statistical computing language. No computing background necessary. For students in the FOCUS Program.

Part of the What If? Explaining the Past/Predicting the Future cluster.
course overview

**curriculum:**
data gathering + wrangling, EDA + viz, multivariate modeling, basic inference, communication

**structure:**
teams: in class exercises + projects
*individual:* HW + take home midterm and final

**assessment:**
not just final work but also the process, peer evaluations and contribution diagnostics
assignments

culminating:
final open ended project on data of own choosing, team based

periodic:
semi open ended homework, individual

in (every) class:
semi open ended application exercises, team based

fast feedback:
Sakai (LMS) quizzes individual
assignments

**culminating:**
final open ended project on data of own choosing, team based

**periodic:**
semi open ended homework, individual

**in (every) class:**
semi open ended application exercises, team based

**fast feedback:**
tryr / DataCamp / etc. individual
Assignments

**Fast Feedback:**
*learnr* modules, individual

**Periodic:**
semi open ended homework, individual

**Culminating:**
final open ended project on data of own choosing, team based

**In (every) class:**
semi open ended application exercises, team based

**Fast Feedback:**
*learnr* modules, individual
Interactive Tutorials for R

Overview

The *learnr* package makes it easy to turn any R Markdown document into an interactive tutorial. Tutorials consist of content along with interactive components for checking and reinforcing understanding. Tutorials can include any or all of the following:

1. Narrative, figures, illustrations, and equations.
2. Code exercises (R code chunks that users can edit and execute directly).
3. Quiz questions.
4. Videos (supported services include YouTube and Vimeo).
5. Interactive Shiny components.

Tutorials automatically preserve work done within them, so if a user works on a few exercises or questions and returns to the tutorial later they can pick up right where they left off.

Examples

https://rstudio.github.io/learnr

# on CRAN
> install.packages("learnr")
> library(learnr)
Data Visualization Basics

Geometric objects

Geoms

How are these two plots similar?

Both plots contain the same x variable, the same y variable, and both describe the same data. But the plots are not identical. Each plot uses a different visual object to represent the data. In ggplot2 syntax, we say that they use different geoms.

A geom is the geometrical object that a plot uses to represent observations. People often describe plots by the type of geom that the plot uses. For example, bar charts use bar geoms, line charts use line geoms, boxplots use boxplot geoms, and so on. Scatterplots break the trend; they use the point geom.

As we see above, you can use different geoms to plot the same data. The plot on the left uses the point geom, and the plot on the right uses the smooth geom, a smooth line fitted to the data.

Continue
A strategy

We can add the `class` variable to the plot by mapping the levels of an aesthetic (like color) to the values of `class`. For example, we can color a point green if it belongs to the compact class, blue if it belongs to the midsize class, and so on.

Let's give this a try. Fill in the blank piece of code below with `color = class`. What happens? Delete the commenting symbols (`#`) before running your code. (If you prefer British English, you can use `colour` instead of `color`.)

```
1 ggplot(data = mpg) +
2 geom_point(mapping = aes(x = displ, y = hwy, color = class))
```

"Great Job! You can now tell which class of car each point represents by examining the color of the point."
Exercise 2

What does the `se` argument to `geom_smooth()` do?

- Nothing. `se` is not an argument of `geom_smooth()`
- Chooses a method for calculating the smooth line
- Controls whether or not to show errors
- Adds or removes a standard error ribbon around the smooth line

Submit Answer
computation

core: R + RStudio server

toolkit: (mostly) tidyverse

reproducibility: R Markdown + Git / GitHub
core:
R + RStudio server

why?
# Local install

- Install R: https://cran.r-project.org/
- Install RStudio: https://www.rstudio.com/products/rstudio/
- Install the following packages:
  - rmarkdown
  - knitr
  - tidyverse
  - ...
- Load these packages

vs.

# RStudio Server

- Go to smith.stat.duke.edu:8787
- Log in with your Net ID & pass
core: R + RStudio Server

goal: minimize onboarding friction and time to 1st data viz

how: avoid local installation with RStudio Server (Pro)

at the end: provide instructions for + help with local install
toolkit:
(mostly)
tidyverse

why?
# base R

```
mtcars$transmission <- ifelse(mtcars$am == 0, "automatic", "manual")
```

# tidyverse

```
mtcars <- mtcars %>%
  mutate(
    transmission = case_when(
      am == 0 ~ "automatic",
      am == 1 ~ "manual"
    )
  )
```
recoding a multi-level variable

# base R

```r
mtcars$gear_char <- ifelse(mtcars$gear == 3, "three",
                           ifelse(mtcars$gear == 4, "four",
                                  "five"))
```

# tidyverse

```r
mtcars <- mtcars %>%
mutate(
  gear_char = case_when(
    gear == 3 ~ "three",
    gear == 4 ~ "four",
    gear == 5 ~ "five"
  ))
```
visualizing multiple variables

# base R

```r
mtcars$trans_color <-
  ifelse(mtcars$transmission == "automatic",
         "green",
         "blue")

plot(mtcars$mpg ~ mtcars$disp,
     col = mtcars$trans_color)
legend("topright",
       legend = c("automatic", "manual"),
       pch = 1, col = c("green", "blue"))
```

# tidyverse

```r
ggplot(mtcars,
       aes(x = disp, y = mpg,
           color = transmission)) +
  geom_point()
```
# base R

```r
mtcars_cyl4 = mtcars[mtcars$cyl == 4, ]
mtcars_cyl6 = mtcars[mtcars$cyl == 6, ]
mtcars_cyl8 = mtcars[mtcars$cyl == 8, ]

par(mfrow = c(1, 3))
plot(mpg ~ disp, data = mtcars_cyl4,
     col = trans_color, main = "Cyl 4")
plot(mpg ~ disp, data = mtcars_cyl6,
     col = trans_color, main = "Cyl 6")
plot(mpg ~ disp, data = mtcars_cyl8,
     col = trans_color, main = "Cyl 8")
legend("topright",
       legend = c("automatic", "manual"),
       pch = 1, col = c("green", "blue"))
```

# tidyverse

```r
ggplot(mtcars,
       aes(x = disp, y = mpg,
           color = transmission)) +
geom_point() +
facet_wrap(~ cyl)
```
toolkit: (mostly) tidyverse

(closer to) human readable

consistent syntax

ease of multivariate visualizations
reproducibility:

why?

R Markdown +
Git / GitHub
R Markdown

**reproducibility:**
train new analysts whose only workflow is a reproducible one

**efficiency:**
consistent formatting + built in “show your work” = easier grading

**pedagogy:**
code + output + prose together syntax highlighting + notebooks FTW!

**key to success:**
iterative development:
knit early, and often
Git + GitHub

**version control:**
lots of mistakes along the way, need ability keep track of history (revert)

**collaboration:**
platform and interface designed to enable collaboration

**accountability:**
transparent commit history

**early intro:**
mastery takes time, start early (day 1)
marketability + discoverability
motivation

computation

interest & impact

data analysis examples

curricular considerations

syllabus
#1 Paris Paintings
element of an undergrad course that introduces students to exploratory data analysis

pairs of grad students, work with course instructor to formulate question & pathway

graduate student participants receive a travel grant

Source: http://bigdata.duke.edu/data-expeditions
Sandra Van Ginhoven
PhD, Art History

Hilary Coe Cronheim
PhD, Art History
Two paintings very rich in composition, of a beautiful execution, and whose merit is very remarkable, each 17 inches 3 lines high, 23 inches wide; the first, painted on wood, comes from the Cabinet of Madame la Comtesse de Verrue; it represents a departure for the hunt: it shows in the front a child on a white horse, a man who gives the horn to gather the dogs, a falconer and other figures nicely distributed across the width of the painting; two horses drinking from a fountain; on the right in the corner a lovely country house topped by a terrace, on which people are at the table, others who play instruments; trees and fabriques pleasantly enrich the background.
<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>AA</th>
<th>AB</th>
<th>AC</th>
<th>AD</th>
<th>AE</th>
<th>AF</th>
<th>AG</th>
<th>AH</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1777-86</td>
<td>86</td>
<td>R</td>
<td>1777</td>
<td>D/FL</td>
<td>D/FL</td>
<td>D/FL</td>
<td>0</td>
<td>620.0</td>
<td>1</td>
<td>2 femmes, enfants, paysage vu à travers une arcade</td>
<td>Bega, Cornelis Pieterszoon</td>
<td>0</td>
<td>n/a</td>
<td>Cornelle Bega</td>
<td>Lebrun</td>
<td></td>
</tr>
<tr>
<td>R1777-87</td>
<td>87</td>
<td>R</td>
<td>1777</td>
<td>D/FL</td>
<td>D/FL</td>
<td>D/FL</td>
<td>0</td>
<td>12,000.0</td>
<td>1</td>
<td>Course du halerg</td>
<td>Wouwerman, Philips</td>
<td>0</td>
<td>n/a</td>
<td>Philippe Wouwerman Donjez</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1777-88</td>
<td>88</td>
<td>R</td>
<td>1777</td>
<td>D/FL</td>
<td>D/FL</td>
<td>D/FL</td>
<td>0</td>
<td>8,000.0</td>
<td>1</td>
<td>Paysage sablonneux</td>
<td>Wouwerman, Philips</td>
<td>0</td>
<td>n/a</td>
<td>Philippe Wouwerman Lambe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1777-89a</td>
<td>89</td>
<td>R</td>
<td>1777</td>
<td>D/FL</td>
<td>D/FL</td>
<td>D/FL</td>
<td>0</td>
<td>5,300.0</td>
<td>1</td>
<td>Départ pour la chasse</td>
<td>Wouwerman, Philips</td>
<td>0</td>
<td>n/a</td>
<td>Philippe Wouwerman Langl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
paris paintings

data:
painting auction data
1764 - 1780
[3,393 x 57]

visualize:
data visualization to explore patterns and possible interactions

clean:
data cleaning and wrangling

model:
model log(price) and perform procedural and expert opinion based model selection
Calculate a similarity score between different classes of art - score between 0 and 1, higher scores reflect a greater degree of similarity among features; i.e. a score of 1 would indicate identical vectors while a score of 0 would indicate vectors with no features in common.

```
similarity = function (vec1, vec2) {
  mag1 = sqrt(vec1 %% vec1)
  mag2 = sqrt(vec2 %% vec2)
  return(vec1 %% vec2 / mag1 / mag2)
}
```

Spanish art is most notably different from the other schools (Lighter colors indicate similarities, while deep red indicates large differences).
Copper paintings, though typically small, have a notably strong interaction with surface area.
student experience

- non-standard application piqued student interest
- "massive" data overwhelming but expert input refreshing
- unfamiliar variables made narrative challenging
- novel application pushed creativity
<table>
<thead>
<tr>
<th>Date</th>
<th>Opponent</th>
<th>Location (Venue)</th>
<th>Score (OT)</th>
<th>Attendance</th>
<th>Tip Time</th>
<th>TV</th>
</tr>
</thead>
<tbody>
<tr>
<td>11/14</td>
<td>Presbyterian</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 113-44</td>
<td>0,314</td>
<td>6 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>11/15</td>
<td>Fairfield</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 109-59</td>
<td>9,314</td>
<td>8 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>11/16</td>
<td>vs. (19) Michigan State</td>
<td>Indianapolis, Ind. (Budweiser Fieldhouse)</td>
<td>W 81-71</td>
<td>10,045</td>
<td>7 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>11/22</td>
<td>vs. Temple</td>
<td>Brooklyn, N.Y. (Barclays Center)</td>
<td>W 78-54</td>
<td>10,123</td>
<td>9:30 p.m.</td>
<td>TruTV</td>
</tr>
<tr>
<td>11/26</td>
<td>Furman</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 83-51</td>
<td>5,914</td>
<td>8 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>11/30</td>
<td>Army</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 63-79</td>
<td>3,114</td>
<td>6 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>12/3</td>
<td># at (2) Wisconsin</td>
<td>Madison, Wisc. (Campus Center)</td>
<td>W 80-70</td>
<td>17,270</td>
<td>9:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>12/15</td>
<td>Elon</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 75-62</td>
<td>9,314</td>
<td>7 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>12/18</td>
<td>vs. Connecticut</td>
<td>East Rutherford, N.J. (Prudential Center)</td>
<td>W 66-56</td>
<td>16,541</td>
<td>7 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>12/20</td>
<td>Toledo</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 86-69</td>
<td>9,314</td>
<td>7 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>12/21</td>
<td>Wofford</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 84-56</td>
<td>3,114</td>
<td>6 p.m.</td>
<td>RSN</td>
</tr>
<tr>
<td>1/3</td>
<td>* Boston College</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 85-62</td>
<td>9,314</td>
<td>4 p.m.</td>
<td>RSN</td>
</tr>
<tr>
<td>1/7</td>
<td># at Wake Forest</td>
<td>Winston-Salem, N.C. (Jedic Center)</td>
<td>W 73-65</td>
<td>12,051</td>
<td>9 p.m.</td>
<td>ACCN</td>
</tr>
<tr>
<td>1/11</td>
<td># at (9) State</td>
<td>Raleigh, N.C. (PNC Arena)</td>
<td>W 70-87</td>
<td>18,300</td>
<td>1:20 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>1/13</td>
<td>* Miami</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 74-90</td>
<td>9,314</td>
<td>9 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>1/17</td>
<td># at (5) Louisville</td>
<td>Louisville, Ky. (KFC Yum! Center)</td>
<td>W 63-52</td>
<td>22,791</td>
<td>12 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>1/19</td>
<td>* Pittsburgh</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 70-65</td>
<td>9,314</td>
<td>7 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>1/25</td>
<td>at St. Johns</td>
<td>New York, N.Y. (Madison Square Garden)</td>
<td>W 77-68</td>
<td>12,812</td>
<td>2 p.m.</td>
<td>FOX</td>
</tr>
<tr>
<td>1/29</td>
<td>* at (8) Notre Dame</td>
<td>South Bend, Ind. (Krispy Kreme Center)</td>
<td>W 73-77</td>
<td>5,140</td>
<td>7:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/1</td>
<td>* at (2) Virginia</td>
<td>Charlottesville, Va. (John Paul Jones Arena)</td>
<td>W 69-63</td>
<td>14,593</td>
<td>7 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/4</td>
<td>* Georgia Tech</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 72-66</td>
<td>9,314</td>
<td>7 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/7</td>
<td># at (19) Notre Dame</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 90-60</td>
<td>9,314</td>
<td>3 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>2/9</td>
<td># at Pitt State</td>
<td>Pullman, Wash. (Beaver Stadium)</td>
<td>W 73-70</td>
<td>14,114</td>
<td>7 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/14</td>
<td># at Syracuse</td>
<td>Syracuse, N.Y. (Carrier Dome)</td>
<td>W 80-72</td>
<td>35,446</td>
<td>6 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>2/18</td>
<td># at (15) North Carolina</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 92-50</td>
<td>9,314</td>
<td>9 p.m.</td>
<td>ESPNU/ACCN</td>
</tr>
<tr>
<td>2/21</td>
<td>* Clemson</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 78-56</td>
<td>9,314</td>
<td>4 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/25</td>
<td># at Virginia Tech</td>
<td>Blacksburg, Va. (Maryland Fieldhouse)</td>
<td>W 81-68</td>
<td>3,114</td>
<td>4 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>2/28</td>
<td>* Syracuse</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 73-54</td>
<td>9,314</td>
<td>7 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>3/4</td>
<td># Wake Forest</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 94-51</td>
<td>9,314</td>
<td>8 p.m.</td>
<td>ACCN</td>
</tr>
<tr>
<td>3/7</td>
<td># at (12) North Carolina</td>
<td>Chapel Hill, N.C. (Dean Dome)</td>
<td>W 84-77</td>
<td>21,750</td>
<td>9 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>3/10</td>
<td>vs. (9) State</td>
<td>Greensboro, N.C. (Greensboro Coliseum)</td>
<td>W 72-53</td>
<td>22,025</td>
<td>7:45 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>3/13</td>
<td># vs. (11) Notre Dame</td>
<td>Greensboro, N.C. (Greensboro Coliseum)</td>
<td>W 64-74</td>
<td>22,025</td>
<td>9 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>3/20</td>
<td>vs. Robert Morris</td>
<td>Charlotte, N.C. (Time Warner Cable Arena)</td>
<td>W 85-56</td>
<td>16,945</td>
<td>7 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>3/22</td>
<td>vs. San Diego State</td>
<td>Charlotte, N.C. (Time Warner Cable Arena)</td>
<td>W 68-49</td>
<td>16,482</td>
<td>2 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>3/27</td>
<td>vs. (19) Utah</td>
<td>Houston, Texas (NRG Stadium)</td>
<td>W 53-47</td>
<td>18,168</td>
<td>7:45 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>3/29</td>
<td>vs. (4) Gonzaga</td>
<td>Houston, Texas (NRG Stadium)</td>
<td>W 66-62</td>
<td>20,744</td>
<td>4 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>4/4</td>
<td>vs. (23) Michigan State</td>
<td>Indianapolis, Ind. (Lucas Oil Stadium)</td>
<td>W 81-61</td>
<td>72,238</td>
<td>6 p.m.</td>
<td>TBS/TNT</td>
</tr>
<tr>
<td>4/6</td>
<td>vs. (5) Wisconsin</td>
<td>Indianapolis, Ind. (Lucas Oil Stadium)</td>
<td>W 68-63</td>
<td>71,140</td>
<td>6:15 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>Date</td>
<td>Opponent</td>
<td>Location (Venue)</td>
<td>Score (GT)</td>
<td>ACC</td>
<td>Tip time</td>
<td>TV</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------</td>
<td>---------------------------</td>
<td>------------</td>
<td>-----</td>
<td>----------</td>
<td>--------</td>
</tr>
<tr>
<td>11/14</td>
<td>Presbyterian</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 111-44</td>
<td>0:14</td>
<td>6:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>11/15</td>
<td>Fairfield</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 102-59</td>
<td>9:24</td>
<td>6:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>11/18</td>
<td>vs. (19) Michigan State</td>
<td>Indianapolis, Ind. (Bankers Life Fieldhouse)</td>
<td>W 88-71</td>
<td>10:36</td>
<td>7:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>11/21</td>
<td>vs. Temple</td>
<td>Brooklyn, N.Y. (Barclays Center)</td>
<td>W 79-54</td>
<td>10:19</td>
<td>9:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>11/22</td>
<td>vs. Stanford</td>
<td>Brooklyn, N.Y. (Barclays Center)</td>
<td>W 70-56</td>
<td>10:48</td>
<td>9:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>11/26</td>
<td>vs. Furman</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 89-54</td>
<td>9:31</td>
<td>6:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>11/30</td>
<td>Army</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 83-73</td>
<td>9:31</td>
<td>12:00 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>12/3</td>
<td>vs. (2) Wisconsin</td>
<td>Madison, Wis. (Kohl Center)</td>
<td>W 80-70</td>
<td>17:31</td>
<td>9:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>12/15</td>
<td>Elon</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 67-62</td>
<td>9:24</td>
<td>7:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>12/18</td>
<td>vs. Connecticut</td>
<td>East Rutherford, N.J. (Izod Center)</td>
<td>W 60-56</td>
<td>16:41</td>
<td>6:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>12/29</td>
<td>Toledo</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 80-69</td>
<td>9:21</td>
<td>7:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>12/31</td>
<td>Wofford</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 84-68</td>
<td>9:31</td>
<td>7:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>1/3</td>
<td>* Boston College</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 85-62</td>
<td>9:31</td>
<td>4:00 p.m.</td>
<td>RSN</td>
</tr>
<tr>
<td>1/7</td>
<td>* at Wake Forest</td>
<td>Winston-Salem, N.C. (Dail Coliseum)</td>
<td>W 73-65</td>
<td>12:51</td>
<td>8:00 p.m.</td>
<td>ACCN</td>
</tr>
<tr>
<td>1/11</td>
<td>* at N.C. State</td>
<td>Raleigh, N.C. (PNC Arena)</td>
<td>L 75-67</td>
<td>15:50</td>
<td>1:30 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>1/13</td>
<td>* Miami</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 74-50</td>
<td>9:24</td>
<td>8:00 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>1/17</td>
<td>* at Louisville</td>
<td>Louisville, Ky. (KFC Yum Center)</td>
<td>W 63-52</td>
<td>12:31</td>
<td>12:00 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>1/19</td>
<td>* Cincinnati</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 79-66</td>
<td>9:31</td>
<td>7:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>1/25</td>
<td>* at St. John's</td>
<td>New York, N.Y. (Madison Square Garden)</td>
<td>W 77-68</td>
<td>19:12</td>
<td>7:00 p.m.</td>
<td>FOX</td>
</tr>
<tr>
<td>1/28</td>
<td>* at Notre Dame</td>
<td>Notre Dame, Ind. (Doyce Center)</td>
<td>L 73-77</td>
<td>0:14</td>
<td>7:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>1/31</td>
<td>* at (12) Virginia</td>
<td>Charlottesville, Va. (John Paul Jones Arena)</td>
<td>W 69-63</td>
<td>14:53</td>
<td>7:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/4</td>
<td>* at Georgia Tech</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 72-66</td>
<td>9:31</td>
<td>7:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>2/7</td>
<td>* at (10) Notre Dame</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 90-56</td>
<td>9:24</td>
<td>1:00 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>2/9</td>
<td>* at Florida State</td>
<td>Tallahassee, Fla. (Donald L. Tucker Center)</td>
<td>W 72-70</td>
<td>11:49</td>
<td>7:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>2/14</td>
<td>* at Syracuse</td>
<td>Syracuse, N.Y. (Carrier Dome)</td>
<td>W 80-72</td>
<td>35:46</td>
<td>6:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/18</td>
<td>* at (15) North Carolina</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 92-66</td>
<td>9:31</td>
<td>9:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/21</td>
<td>* vs. Clemson</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 76-56</td>
<td>9:31</td>
<td>7:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>2/25</td>
<td>* at Virginia Tech</td>
<td>Blacksburg, Va. (Cachee Coliseum)</td>
<td>W 91-66</td>
<td>9:54</td>
<td>9:00 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>2/28</td>
<td>* at Syracuse</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 73-54</td>
<td>9:31</td>
<td>7:30 p.m.</td>
<td>ESPNU</td>
</tr>
<tr>
<td>3/4</td>
<td>* Wake Forest</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
<td>W 94-51</td>
<td>9:31</td>
<td>6:30 p.m.</td>
<td>ACCN</td>
</tr>
<tr>
<td>3/7</td>
<td>* at (19) North Carolina</td>
<td>Chacei Hill, N.C. (Dean Dome)</td>
<td>W 84-77</td>
<td>21:50</td>
<td>9:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>3/12</td>
<td>vs. (3) Duke</td>
<td>Greensboro, N.C. (Greensboro Coliseum)</td>
<td>W 82-72</td>
<td>22:26</td>
<td>9:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>3/13</td>
<td>vs. (11) Notre Dame</td>
<td>Greensboro, N.C. (Greensboro Coliseum)</td>
<td>L 84-74</td>
<td>22:26</td>
<td>9:30 p.m.</td>
<td>ESPN2</td>
</tr>
<tr>
<td>3/20</td>
<td>vs. Robert Morris</td>
<td>Charlotte, N.C. (Time Warner Cable Arena)</td>
<td>W 88-56</td>
<td>16:45</td>
<td>7:30 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>3/22</td>
<td>vs. San Diego State</td>
<td>Charlotte, N.C. (Time Warner Cable Arena)</td>
<td>W 60-49</td>
<td>16:42</td>
<td>2:00 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>3/27</td>
<td>vs. (18) Utah</td>
<td>Houston, Texas (NRG Stadium)</td>
<td>W 63-57</td>
<td>21:19</td>
<td>7:45 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>3/29</td>
<td>vs. (7) Gonzaga</td>
<td>Houston, Texas (NRG Stadium)</td>
<td>W 60-52</td>
<td>20:44</td>
<td>4:00 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>4/4</td>
<td>vs. (23) Michigan State</td>
<td>Indianapolis, Ind. (Lucas Oil Stadium)</td>
<td>W 81-63</td>
<td>72:23</td>
<td>6:30 p.m.</td>
<td>TBS/FSN</td>
</tr>
<tr>
<td>4/6</td>
<td>vs. (3) Wisconsin</td>
<td>Indianapolis, Ind. (Lucas Oil Stadium)</td>
<td>W 86-63</td>
<td>71:19</td>
<td>9:30 p.m.</td>
<td>CBS</td>
</tr>
<tr>
<td>Date</td>
<td>Opponent</td>
<td>Location</td>
<td>Score</td>
<td>Date</td>
<td>Opponent</td>
<td>Location</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------</td>
<td>--------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>4 May</td>
<td>Presbyterian</td>
<td>Durham, N.C.</td>
<td>11-3</td>
<td>8</td>
<td>ISPMU</td>
<td>6 p.m.</td>
</tr>
<tr>
<td>15 May</td>
<td>Fairleigh Dickinson</td>
<td>Durham, N.C.</td>
<td>12-19</td>
<td>9</td>
<td>ISPMU</td>
<td>6 p.m.</td>
</tr>
<tr>
<td>18 May</td>
<td>vs. (1) Michigan</td>
<td>Columbus, IN</td>
<td>81-71</td>
<td>3</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>21 May</td>
<td>vs. Temple</td>
<td>Brooklyn, N.Y.</td>
<td>74-56</td>
<td>6</td>
<td>Television</td>
<td>9:30 p.m.</td>
</tr>
<tr>
<td>22 May</td>
<td>vs. Stanford</td>
<td>Brooklyn, N.Y.</td>
<td>73-80</td>
<td>5</td>
<td>Television</td>
<td>9:30 p.m.</td>
</tr>
<tr>
<td>25 May</td>
<td>Penn State</td>
<td>Durham, N.C.</td>
<td>93-54</td>
<td>4</td>
<td>ISPMU</td>
<td>5 p.m.</td>
</tr>
<tr>
<td>30 May</td>
<td>Army</td>
<td>Durham, N.C.</td>
<td>92-67</td>
<td>3</td>
<td>ISPMU</td>
<td>11 p.m.</td>
</tr>
<tr>
<td>1 Jun</td>
<td>vs. (2) Wisconsin</td>
<td>Madison, WI</td>
<td>72-109</td>
<td>2</td>
<td>ISPMU</td>
<td>9:30 p.m.</td>
</tr>
<tr>
<td>13 Jun</td>
<td>Elon</td>
<td>Durham, N.C.</td>
<td>93-36</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>18 Jun</td>
<td>vs. Connecticut</td>
<td>Durham, N.C.</td>
<td>14-54</td>
<td>1</td>
<td>ISPMU</td>
<td>8 p.m.</td>
</tr>
<tr>
<td>29 Jun</td>
<td>Towson State</td>
<td>Durham, N.C.</td>
<td>93-34</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>3 Jul</td>
<td>Wells College</td>
<td>Durham, N.C.</td>
<td>93-14</td>
<td>1</td>
<td>ISPMU</td>
<td>3 p.m.</td>
</tr>
<tr>
<td>9 Jul</td>
<td>Boston College</td>
<td>Boston, MA</td>
<td>78-21</td>
<td>1</td>
<td>RSN</td>
<td>4 p.m.</td>
</tr>
<tr>
<td>7 Jul</td>
<td>vs. (1) Notre Dame</td>
<td>Durham, N.C.</td>
<td>76-140</td>
<td>1</td>
<td>ISPMU</td>
<td>9:30 p.m.</td>
</tr>
<tr>
<td>11 Jul</td>
<td>vs. (3) Virginia</td>
<td>Durham, N.C.</td>
<td>74-35</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>14 Jul</td>
<td>Georgia Tech</td>
<td>Durham, N.C.</td>
<td>93-14</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>25 Jul</td>
<td>vs. (5) Notre Dame</td>
<td>Durham, N.C.</td>
<td>93-14</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>28 Jul</td>
<td>vs. Florida State</td>
<td>Durham, N.C.</td>
<td>78-48</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>1 Aug</td>
<td>vs. Syracuse</td>
<td>Syracuse, NY</td>
<td>77-43</td>
<td>1</td>
<td>ISPMU</td>
<td>6 p.m.</td>
</tr>
<tr>
<td>18 Aug</td>
<td>vs. (15) North Carolina</td>
<td>Charlotte, NC</td>
<td>73-21</td>
<td>1</td>
<td>ISPMU</td>
<td>9 p.m.</td>
</tr>
<tr>
<td>21 Aug</td>
<td>vs. Florida State</td>
<td>Durham, N.C.</td>
<td>93-36</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>25 Aug</td>
<td>vs. Virginia Tech</td>
<td>Blacksburg, VA</td>
<td>93-36</td>
<td>1</td>
<td>ISPMU</td>
<td>4 p.m.</td>
</tr>
<tr>
<td>28 Aug</td>
<td>vs. Davidson</td>
<td>Davidson, NC</td>
<td>93-14</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>4 Sep</td>
<td>Wake Forest</td>
<td>Durham, N.C.</td>
<td>93-36</td>
<td>1</td>
<td>ACCN</td>
<td>8 p.m.</td>
</tr>
<tr>
<td>7 Sep</td>
<td>vs. (19) North Carolina</td>
<td>Chapel Hill, NC</td>
<td>84-77</td>
<td>1</td>
<td>ISPMU</td>
<td>9 p.m.</td>
</tr>
<tr>
<td>12 Sep</td>
<td>vs. (9) Samford</td>
<td>Greensboro, NC</td>
<td>77-53</td>
<td>1</td>
<td>ISPMU</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>13 Sep</td>
<td>vs. (11) Notre Dame</td>
<td>Greensboro, NC</td>
<td>64-74</td>
<td>1</td>
<td>ISPMU</td>
<td>22/26 p.m.</td>
</tr>
<tr>
<td>20 Sep</td>
<td>vs. Robert Morris</td>
<td>Charlotte, N.C.</td>
<td>85-56</td>
<td>1</td>
<td>CBS</td>
<td>7 p.m.</td>
</tr>
<tr>
<td>23 Sep</td>
<td>vs. (1) Diego State</td>
<td>Durham, N.C.</td>
<td>68-80</td>
<td>1</td>
<td>CBS</td>
<td>22/24 p.m.</td>
</tr>
<tr>
<td>27 Sep</td>
<td>vs. (2) UCLA</td>
<td>Houston, Texas</td>
<td>63-37</td>
<td>1</td>
<td>CBS</td>
<td>7:45 p.m.</td>
</tr>
<tr>
<td>25 Oct</td>
<td>vs. Georgia Tech</td>
<td>Houston, Texas</td>
<td>65-92</td>
<td>1</td>
<td>CBS</td>
<td>4 p.m.</td>
</tr>
<tr>
<td>1 Nov</td>
<td>vs. (23) Marquette</td>
<td>Indianapolis, IN</td>
<td>81-31</td>
<td>1</td>
<td>ISPMU</td>
<td>6 p.m.</td>
</tr>
<tr>
<td>5 Nov</td>
<td>vs. (3) Wisconsin</td>
<td>Indianapolis, IN</td>
<td>68-32</td>
<td>1</td>
<td>ISPMU</td>
<td>9:15 p.m.</td>
</tr>
</tbody>
</table>
# Load packages -----------------------------------------------------
library(rvest)
library(stringr)
library(dplyr)

# Read page with season data ----------------------------------------

# Harvest fields ----------------------------------------------------
date <- page %>%
  html_nodes(".stattextline b") %>%
  html_text()
opponent <- page %>%
  html_nodes(".stattextltgray2:nth-child(3)") %>%
  html_text() %>%
  str_trim()
venue <- page %>%
  html_nodes(".stattextltgray2:nth-child(5)") %>%
  html_text() %>%
  str_trim()

# Put fields into a tibble ------------------------------------------
blue_devils_1415 <- data_frame(date, opponent, venue)
<table>
<thead>
<tr>
<th>date</th>
<th>opponent</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>11/14</td>
<td>Presbyterian</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
</tr>
<tr>
<td>11/15</td>
<td>Fairfield</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
</tr>
<tr>
<td>11/18</td>
<td>vs. [19] Michigan State</td>
<td>Indianapolis, Ind. (Bankers Life Fieldhouse)</td>
</tr>
<tr>
<td>11/21</td>
<td>vs. Temple</td>
<td>Brooklyn, N.Y. (Barclays Center)</td>
</tr>
<tr>
<td>11/22</td>
<td>vs. Stanford</td>
<td>Brooklyn, N.Y. (Barclays Center)</td>
</tr>
<tr>
<td>11/26</td>
<td>Furman</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
</tr>
<tr>
<td>11/30</td>
<td>Army</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
</tr>
<tr>
<td>12/15</td>
<td>Elon</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
</tr>
<tr>
<td>12/18</td>
<td>vs. Connecticut</td>
<td>East Rutherford, N.J. (Izod Center)</td>
</tr>
<tr>
<td>12/29</td>
<td>Toledo</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
</tr>
<tr>
<td>12/31</td>
<td>Wofford</td>
<td>Durham, N.C. (Cameron Indoor Stadium)</td>
</tr>
</tbody>
</table>

Showing 1 to 13 of 39 entries
Students upset b/c website they need to scrape data from for hw assignment is down. Bad assignment or good lesson in working w/ real data?
> library(shiny)
Modeling the Distributions of Fatal Car Crashes

Select categorical variable:
- alcohol involved
- crash type
- alcohol involved
- teen driver
- older driver
- direction
duke focus: first-year undergrads modeling cluster: “What if? Explaining the Past, Predicting the Future”

interest in What If: no hard data, but “definitely significant increase in applications the last two years than previous years”

interest in DS: % of What If applicants interested in DS
2015: 76%
2016: 83%
pipeline for stats:
2014: 19% declared
2015: 31% declared
2016: ~40% expressed interest

diversity:
% female
2014: 44%
2015: 50%
2016: 35%
~25% in Probability

curricular:
basis for gateway to stats major course to be offered in Spring 2018!
Curricular Considerations

- Move away from ad-hoc computing education and/or expecting students to pick it up along the way.
- Uniformity of tools is important: choose a toolkit that works for you and stick to it throughout the curriculum.
- Teach computing early (without any prereqs) and often!
Thank you!