R for DS - 3/29/17

Web Scraping

```
library(rvest)
```

```
## Loading required package: xml2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)


year <- 2016
site_url <- "http://www.espn.com/mens-college-basketball/statistics/team/_/stat/scoring-per-game/sort/a

team_css <- "td:nth-child(2) a"

html <- read_html(site_url)

html %>%
  html_nodes(team_css) %>%
  html_attr("href") %>%
  str_match("/id/(\\d+)") %>%
  print() ->
id_matches
```

```
##       [,1]       [,2]
##  [1,] "/id/2473" "2473"
##  [2,] "/id/2643" "2643"
##  [3,] "/id/276"  "276"
##  [4,] "/id/2454" "2454"
##  [5,] "/id/2437" "2437"
##  [6,] "/id/2739" "2739"
##  [7,] "/id/2174" "2174"
##  [8,] "/id/252"  "252"
##  [9,] "/id/264"  "264"
## [10,] "/id/2057" "2057"
## [11,] "/id/84"   "84"
## [12,] "/id/153"  "153"
## [13,] "/id/2737" "2737"
## [14,] "/id/331"  "331"
## [15,] "/id/66"   "66"
## [16,] "/id/2305" "2305"
## [17,] "/id/150"  "150"
## [18,] "/id/2752" "2752"
## [19,] "/id/12"   "12"
## [20,] "/id/30"   "30"
## [21,] "/id/2617" "2617"
## [22,] "/id/2086" "2086"
## [23,] "/id/2198" "2198"
## [24,] "/id/201"  "201"
## [25,] "/id/56"   "56"
## [26,] "/id/127"  "127"
## [27,] "/id/2870" "2870"
## [28,] "/id/2250" "2250"
## [29,] "/id/96"   "96"
## [30,] "/id/250"  "250"
## [31,] "/id/314"  "314"
## [32,] "/id/156"  "156"
## [33,] "/id/309"  "309"
## [34,] "/id/2466" "2466"
## [35,] "/id/36"   "36"
## [36,] "/id/2405" "2405"
## [37,] "/id/2348" "2348"
## [38,] "/id/270"  "270"
## [39,] "/id/350"  "350"
## [40,] "/id/277"  "277"
```

```
team_ids <- id_matches[,2] %>% print()
```

```
##  [1] "2473" "2643" "276"  "2454" "2437" "2739" "2174" "252"  "264"  "2057"
## [11] "84"   "153"  "2737" "331"  "66"   "2305" "150"  "2752" "12"   "30"
## [21] "2617" "2086" "2198" "201"  "56"   "127"  "2870" "2250" "96"   "250"
## [31] "314"  "156"  "309"  "2466" "36"   "2405" "2348" "270"  "350"  "277"
```

```
team_urls <- str_c("http://www.espn.com/mens-college-basketball/team/stats/_/id/", team_ids, "/year/", y

west_virginia_url <- team_urls[40] %>% print()
```

```
## [1] "http://www.espn.com/mens-college-basketball/team/stats/_/id/277/year/2016"
```

```
html <- read_html(west_virginia_url)

stats <- html %>% html_nodes(".mod-content table") # grab both tables
game_statistics <- stats[[1]] %>% html_table() # grab season avg table
game_statistics <- game_statistics[-1:-2, ] # remove bad headers
game_statistics <- game_statistics[-nrow(game_statistics), ] # remove totals
game_statistics[-1] <- lapply(game_statistics[-1], as.numeric) # make numeric

game_statistics
```

```
##                   X1 X2   X3   X4  X5  X6  X7  X8  X9   X10   X11   X12
## 3      Jaysean Paige 35 22.5 13.7 3.5 1.2 1.5 0.2 1.5 0.455 0.785 0.323
## 4     Devin Williams 35 25.4 13.3 9.5 1.4 0.7 0.2 2.3 0.467 0.693 0.000
## 5       Jevon Carter 35 27.7  9.5 2.9 3.3 1.7 0.3 1.7 0.383 0.744 0.306
## 6   Daxter Miles Jr. 32 24.9  9.4 2.1 1.3 1.4 0.1 1.3 0.427 0.602 0.303
## 7      Tarik Phillip 35 22.3  9.3 2.5 2.8 1.5 0.3 2.0 0.417 0.705 0.409
## 8    Jonathan Holton 31 23.1  8.9 7.6 1.4 1.0 0.3 1.2 0.533 0.667 0.244
## 9          Esa Ahmad 34 18.1  4.9 2.7 1.5 0.8 0.6 1.4 0.446 0.624 0.222
## 10     Nathan Adrian 35 18.1  4.5 3.1 0.9 0.8 0.2 0.8 0.488 0.567 0.407
## 11      Elijah Macon 35 13.2  4.5 3.0 0.5 0.3 0.5 0.9 0.526 0.463 0.000
## 12      Teyvon Myers 29  8.6  2.4 0.8 0.5 0.2 0.0 0.6 0.356 0.783 0.310
## 13   Brandon Watkins 23  5.4  0.6 1.0 0.2 0.0 0.3 0.4 0.333 0.333 0.000
## 14 Richard Romeo III 10  2.1  0.6 0.2 0.0 0.0 0.0 0.2 0.667 1.000 0.000
## 15       Lamont West  1  0.0  0.0 0.0 0.0 0.0 1.0 0.0 0.000 0.000 0.000
## 16       Logan Routt  1  3.0  0.0 0.0 0.0 0.0 0.0 0.0 0.000 0.000 0.000
```