Course 1 Section 3.12 - Web scraping

Jiaying Wu

18/09/2020

```r
library(tidyverse)
library(rvest)
```

```r
site <- "http://stats.espncricinfo.com/ci/engine/stats/index.html?class=10;page=1;team=289;template=resu
raw_html <- read_html(site)
tables <- html_table(raw_html, fill = TRUE)
ausw_t20 <- tables[[3]]
glimpse(ausw_t20)
```

```
## Rows: 50
## Columns: 15
## $ Player <chr> "MM Lanning", "AJ Healy", "BL Mooney", "EJ Villani", "AJ Bla...
## $ Span   <chr> "2010-2020", "2010-2020", "2016-2020", "2009-2018", "2005-20...
## $ Mat    <int> 104, 112, 52, 62, 95, 120, 64, 40, 36, 54, 67, 40, 79, 15, 3...
## $ Inns   <chr> "98", "97", "49", "58", "81", "72", "55", "40", "35", "50", ...
## $ NO     <chr> "21", "16", "11", "10", "19", "29", "10", "2", "2", "14", "2...
## $ Runs   <chr> "2788", "2060", "1452", "1369", "1314", "1218", "941", "784"...
## $ HS     <chr> "133*", "148*", "117*", "90*", "61", "60*", "68*", "61", "56...
## $ Ave    <chr> "36.20", "25.43", "38.21", "28.52", "21.19", "28.32", "20.91...
## $ BF     <chr> "2382", "1560", "1176", "1158", "1414", "1155", "875", "752"...
## $ SR     <chr> "117.04", "132.05", "123.46", "118.22", "92.92", "105.45", "...
## $ ‘100‘  <chr> "2", "1", "2", "0", "0", "0", "0", "0", "0", "0", "0", "0", ...
## $ ‘50‘   <chr> "13", "12", "9", "12", "1", "4", "3", "2", "3", "1", "3", "2...
## $ ‘0‘    <chr> "1", "10", "2", "4", "5", "4", "4", "2", "2", "4", "1", "4",...
## $ ‘4s‘   <chr> "334", "266", "188", "177", "87", "103", "82", "80", "92", "...
## $ ‘6s‘   <chr> "36", "39", "8", "12", "1", "23", "20", "11", "3", "1", "10"...
```

Modification 1: Remove empty characters

```r
ausw_t20$HS <- str_replace(ausw_t20$HS, "\\*", "")
```

Modification 2: Make the data long form

```r
ausw_t20 <- gather(ausw_t20, key = "statistic", value = "value", -Player, -Span)
```

Modification 3: Change the value column

```
ausw_t20$value <- str_replace(ausw_t20$value, "-", "")
ausw_t20$value <- as.numeric(ausw_t20$value)
```

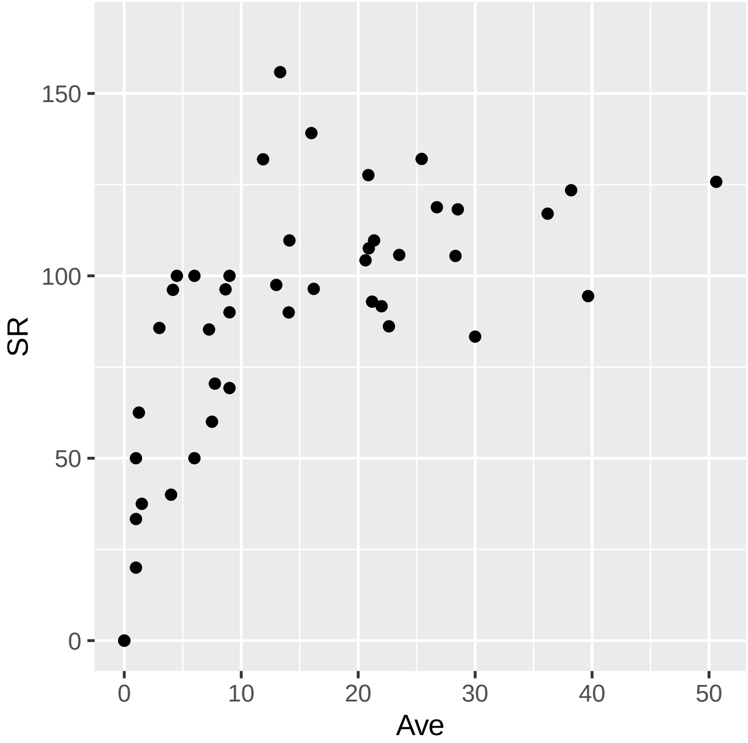
Modification 4: Spread the data set

```
ausw_t20 <- spread(ausw_t20, key = "statistic", value = "value")
```

Modification 5: Plot the batting average

```
ggplot(ausw_t20, aes(x = Ave, y = SR))+
  geom_point()
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```


SR
Ave