Lab-webscraping

Lin Pin Tzu (Ruby)

2022-07-07

3) Using the web address "https://en.wikipedia.org/wiki/ITF_Rankings" and the R coding structure presented in class to web scrape the following table found on the page.

```
wikiurl <- read_html(
  "https://en.wikipedia.org/wiki/ITF_Rankings")
datatables <- wikiurl%>%
  html_table(., fill = T)
datatables[[3]] -> dt
dt
```

```
## # A tibble: 7 x 3
##   `Opponent Nation Ranking1` `Bonus Points` `Bonus Points`
##   <chr>                      <chr>          <chr>
## 1 Opponent Nation Ranking1   Away           Home
## 2 1 to 2                     125            100
## 3 3 to 4                     112.5          90
## 4 5 to 8                     93.75          75
## 5 9 to 16                    62.5           50
## 6 17 to 32                   50             40
## 7 33 to 64                   31.25          25
```

4) Using the web address "https://www.mlb.com/stats/2018" and the R coding structure presented in class, web scrape the table found on the page.

```
wikiurl1 <- read_html("https://www.mlb.com/stats/2018")
baseballdata2018 <- wikiurl1%>%
  html_table(., fill = T)
baseballdata2018[[1]] -> BD2018
BD2018
```

```
## # A tibble: 25 x 18
##    PLAYERPLAYER      TEAMTEAM    GG  ABAB    RR    HH `2B2B` `3B3B`  HRHR RBIRBI
##    <chr>             <chr>    <int> <int> <int> <int>  <int>  <int> <int>  <int>
##   1 1MikeM TroutTrou~ LAA        140   471   101   147     24      4    39      79
##   2 2MookieM BettsBe~ BOS        136   520   129   180     47      5    32      80
##   3 3J.D.J MartinezM~ BOS        150   569   111   188     37      2    43     130
##   4 4ChristianC Yeli~ MIL        147   574   118   187     34      7    36     110
##  5 5JoseJ RamírezRa~ CLE         157   578   110   156     38      4    39    105
##  6 6NolanN ArenadoA~ COL         156   590   104   175     38      2    38    110
##  7 7AlexA BregmanBr~ HOU         157   594   105   170     51      1    31    103
##  8 8PaulP Goldschmi~ ARI         158   593    95   172     35      5    33     83
##  9 9TrevorT StorySt~ COL         157   598    88   174     42      6    37    108
## 10 10AnthonyA Rendo~ WSH         136   529    88   163     44      2    24     92
## # ... with 15 more rows, and 8 more variables: BBBB <int>, SOSO <int>,
## #   SBSB <int>, CSCS <int>, AVGAVG <dbl>, OBPOBP <dbl>, SLGSLG <dbl>,
## #   `caret-upcaret-downOPScaret-upcaret-downOPS` <dbl>
```

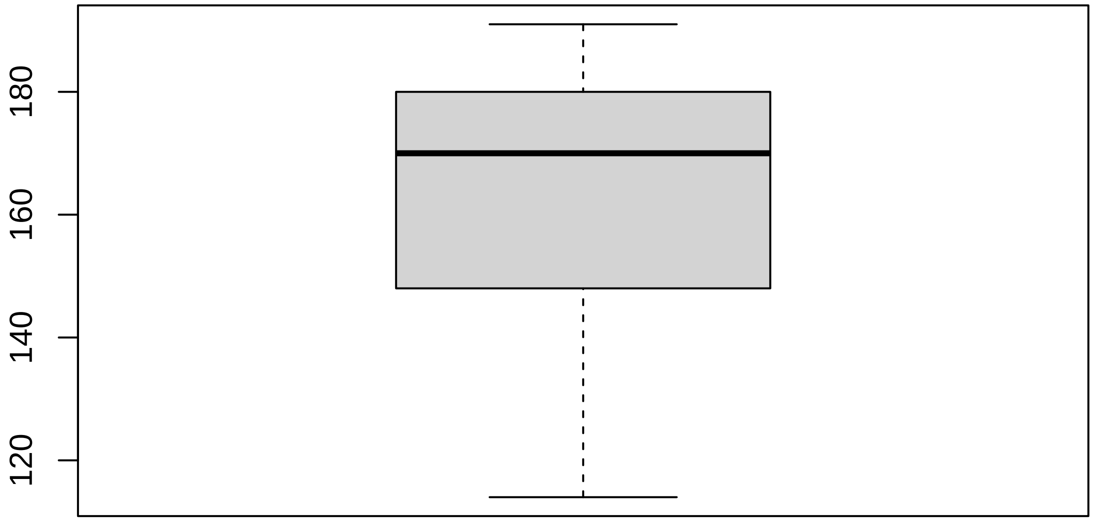
5) Use and show R code to find the average number of hits for all players in the table from number 4

```
mean(BD2018$HH)
```

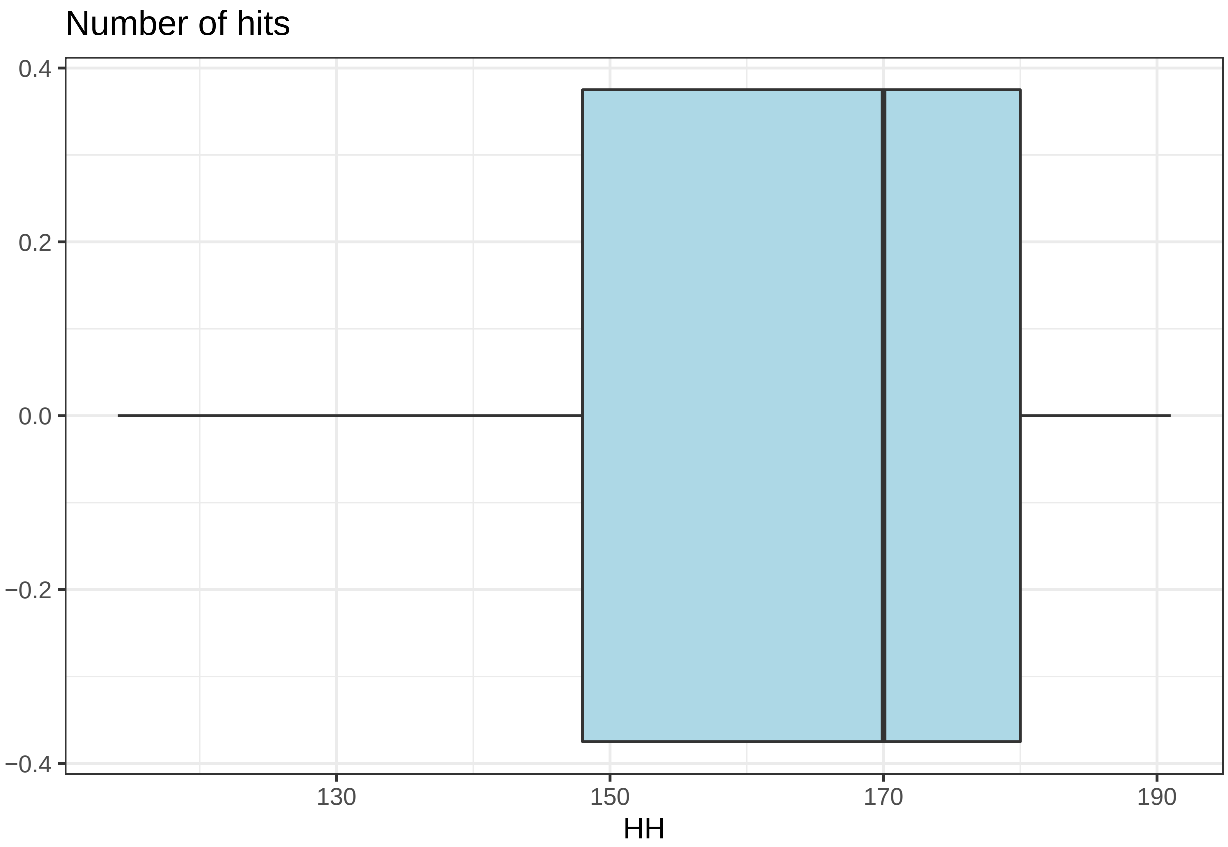
```
## [1] 163.52
```

6) Use and show R code to produce a boxplot for the number of hits (use tidyverse/ggplot coding). Use the data table from number 4

```
boxplot(BD2018$HH)
```



```
#ggplot
ggplot(data=BD2018,mapping = aes(x=HH))+
  geom_boxplot(fill="light blue")+
  ggtitle("Number of hits")+
  theme_bw()
```



7) Use and show dplyr coding to determine which player had the greatest number of strikeouts using the data table from number 4.

```
BD2018 %>%
  select(PLAYERPLAYER,SOSO) %>%
  arrange(desc(SOSO)) %>%
  slice(1)
```

```
## # A tibble: 1 x 2
##   PLAYERPLAYER             SOSO
##   <chr>                   <int>
## 1 20KhrisK DavisDavisDH20   175
```

8) Use and show dplyr coding to show the batting averages for Washington Nationals players and Colorado Rockies players using the data table from number 4

```
BD2018 %>%
  filter(TEAMTEAM =="WSH"|TEAMTEAM =="COL") %>%
  select(TEAMTEAM,AVGAVG) %>%
  group_by(TEAMTEAM)
```

```
## # A tibble: 5 x 2
## # Groups:   TEAMTEAM [2]
##   TEAMTEAM AVGAVG
##   <chr>     <dbl>
## 1 COL       0.297
## 2 COL       0.291
## 3 WSH       0.308
## 4 WSH       0.249
## 5 COL       0.291
```