Class Activity 16

Arthur Viegas Eguia

May 01 2024

Group Activity 1

a. Scrape the first table in `List_of_NASA_missions` wiki page. Additionally, use `janitor::clean_names()` to clean the column names and store the resulting table as `NASA_missions.csv` in your working folder.

```
wiki_NASA <- "https://en.wikipedia.org/wiki/List_of_NASA_missions"

# Scrape the data and write the first table to a CSV file
bow(wiki_NASA) %>%
  scrape() %>%
  html_elements(css = "table") %>%
  html_table() %>%
  .[[1]] %>%
  write_csv("NASA_missions.csv")#Extracts just the first table
```

```
NASA_missions <- read_csv("NASA_missions.csv")
View(NASA_missions)
Error in .External2(C_dataviewer, x, title): unable to start data viewer
```

b. Now, write a code snippet to scrape all the URLs from the anchor tags () on a given Wikipedia page, convert the relative URLs to absolute URLs, and store the results in a tibble and save it as `NASA_missions_urls.csv` in your working folder.

```
# Scrape the data and write the URLs to a CSV file
wiki_NASA <- "https://en.wikipedia.org/wiki/List_of_NASA_missions"
bow(wiki_NASA) %>%
  scrape() %>%
  html_nodes("a") %>% #If you have more than one match, returns all
  html_attr("href") %>%
  url_absolute("https://en.wikipedia.org") %>%
  data.frame(url = .) %>%
  write_csv("NASA_missions_urls.csv")
```

Group Activity 2

a. Scrape player statistics from the given web page, clean and reformat the data table headers using R packages, and create a bar chart to display the top ten players by playing time starts.

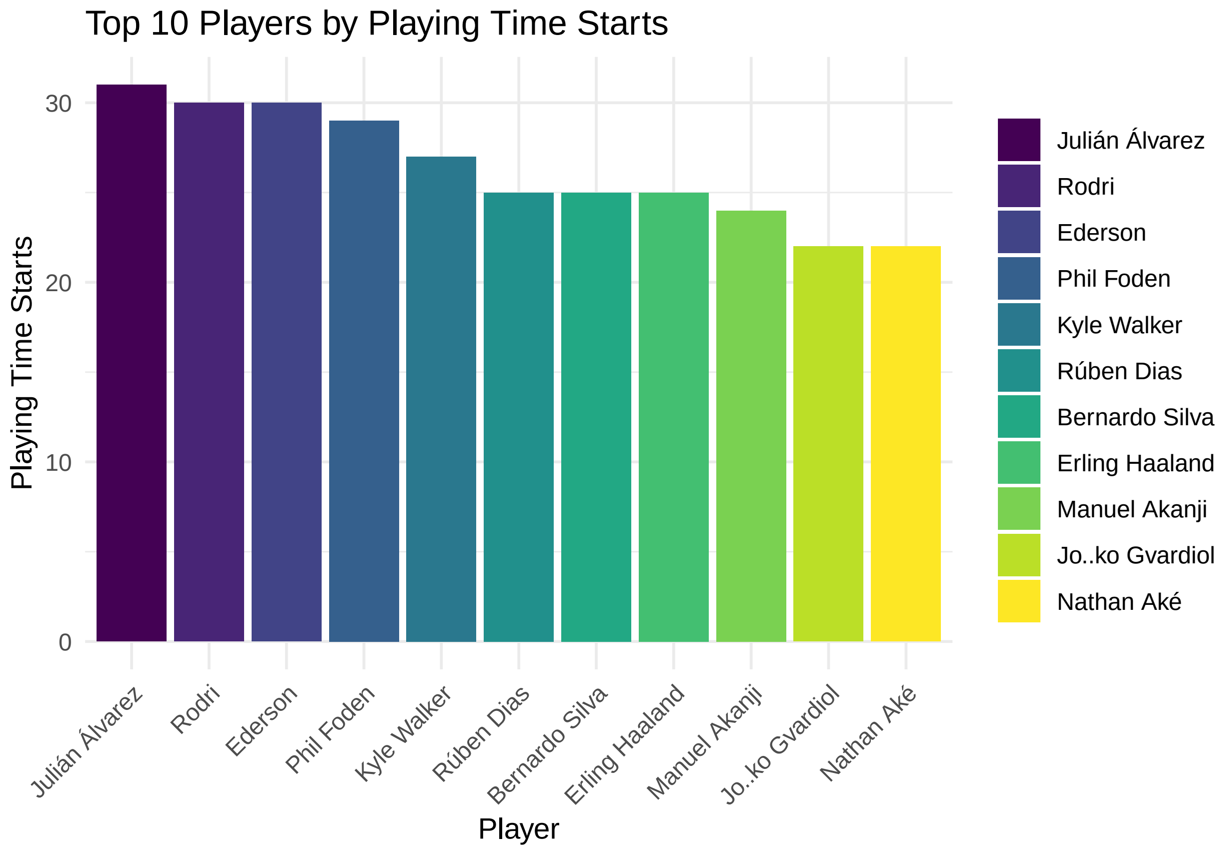
Start by extracting a table from a webpage using the `rvest` package, then clean the headers by merging them with `subheaders` and using `janitor` to standardize the names.

```
mancity <- "https://fbref.com/en/squads/b8fd03ef/Manchester-City-Stats"

data <- bow(mancity) %>%
  scrape() %>%
  html_nodes("table") %>%
  html_table() %>%
  .[[1]] -> data

data
# A tibble: 33 x 34
   ``        ``    ``    ``    ``    `Playing Time` `Playing Time` `Playing Time`
   <chr>     <chr> <chr> <chr> <chr> <chr>          <chr>          <chr>
 1 Player    Nati~ Pos   Age   MP    Starts         Min            90s
 2 Julián ~ ar A~ MF,FW 24-0~ 33    31             2,628          29.2
 3 Rodri     es E~ MF    27-3~ 30    30             2,571          28.6
 4 Ederson   br B~ GK    30-2~ 30    30             2,537          28.2
 5 Phil Fo~ eng ~ FW,MF 23-3~ 31    29             2,518          28.0
 6 Kyle Wa~ eng ~ DF    33-3~ 28    27             2,428          27.0
 7 Rúben D~ pt P~ DF    26-3~ 27    25             2,289          25.4
 8 Bernard~ pt P~ MF,FW 29-2~ 29    25             2,229          24.8
 9 Erling ~ no N~ FW    23-2~ 27    25             2,211          24.6
10 Manuel ~ ch S~ DF,MF 28-2~ 26    24             2,171          24.1
# i 23 more rows
# i 26 more variables: Performance <chr>, Performance <chr>, Performance <chr>,
#   Performance <chr>, Performance <chr>, Performance <chr>, Performance <chr>,
#   Performance <chr>, Expected <chr>, Expected <chr>, Expected <chr>,
#   Expected <chr>, Progression <chr>, Progression <chr>, Progression <chr>,
#   `Per 90 Minutes` <chr>, `Per 90 Minutes` <chr>, `Per 90 Minutes` <chr>,
#   `Per 90 Minutes` <chr>, `Per 90 Minutes` <chr>, `Per 90 Minutes` <chr>, ...
```

```
data %>% {
  subheaders <- .[1,]
  new_names <- map2_chr(names(.), subheaders, ~str_c(.x, .y, sep=" - "))
  set_names(., new_names)
} %>%
janitor::clean_names() %>%
slice(-1, -n(), -n() + 1) -> data_clean
data_clean
# A tibble: 30 x 34
   player          nation  pos   age   mp    playing_time_starts playing_time_min
   <chr>           <chr>   <chr> <chr> <chr> <chr>               <chr>
 1 Julián Álvarez ar ARG  MF,FW 24-0~ 33    31                  2,628
 2 Rodri           es ESP  MF    27-3~ 30    30                  2,571
 3 Ederson         br BRA  GK    30-2~ 30    30                  2,537
 4 Phil Foden      eng ENG FW,MF 23-3~ 31    29                  2,518
 5 Kyle Walker     eng ENG DF    33-3~ 28    27                  2,428
 6 Rúben Dias      pt POR  DF    26-3~ 27    25                  2,289
 7 Bernardo Silva  pt POR  MF,FW 29-2~ 29    25                  2,229
 8 Erling Haaland no NOR  FW    23-2~ 27    25                  2,211
 9 Manuel Akanji  ch SUI  DF,MF 28-2~ 26    24                  2,171
10 Joško Gvardiol hr CRO  DF    22-0~ 24    22                  1,968
# i 20 more rows
# i 27 more variables: playing_time_90s <chr>, performance_gls <chr>,
#   performance_ast <chr>, performance_g_a <chr>, performance_g_pk <chr>,
#   performance_pk <chr>, performance_p_katt <chr>, performance_crd_y <chr>,
#   performance_crd_r <chr>, expected_x_g <chr>, expected_npx_g <chr>,
#   expected_x_ag <chr>, expected_npx_g_x_ag <chr>, progression_prg_c <chr>,
#   progression_prg_p <chr>, progression_prg_r <chr>, ...
```

Analyze the 'playing__time__starts' to find the top ten players and visualize this data in a bar chart using ggplot2, ensuring the chart is both informative and aesthetically pleasing.

```
data_clean_plot <- data_clean %>%
  mutate(
    playing_time_starts = readr::parse_number(playing_time_starts),
    player = factor(player, levels = player)
  ) %>%
  arrange(desc(playing_time_starts)) %>%
  top_n(10, playing_time_starts)

ggplot(data_clean_plot, aes(x = reorder(player, -playing_time_starts),
                            y = playing_time_starts, fill = player)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 10 Players by Playing Time Starts",
       x = "Player",
       y = "Playing Time Starts") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.title = element_blank()) +
  scale_fill_viridis_d()
```



Group Activity 3

In this activity, you'll scrape web data using `rvest` and tidy up the results into a well-formatted table. Start by extracting job titles from a given URL, then gather the associated company names, and trim any leading or trailing whitespace from the location data. Next, retrieve the posting dates and the URLs for the full job descriptions. Finally, combine all these elements into a single dataframe, ensuring that each piece of information aligns correctly. Your task is to produce a clean and informative table that could be useful for job seekers. To facilitate the selection of the correct CSS selectors, you may find the `SelectorGadget` Chrome extension particularly useful.

```r
url <- "https://realpython.github.io/fake-jobs/"
```

```r
title <- bow(url) %>%
  scrape() %>%
  html_elements(".is-5") %>%
  html_text()  # part 1
company <- bow(url) %>%
  scrape() %>%
  html_elements(".company") %>%
  html_text()# part 2
location <- bow(url) %>%
  scrape() %>%
  html_elements(".location") %>%
  html_text()  %>%
  str_trim()# part 3
```

```
time <- bow(url) %>%
  scrape() %>%
  html_elements(".has-text-grey") %>%
  html_text() %>%
  str_trim()# part 4
html <- bow(url) %>% scrape() %>%
  html_elements(".card-footer-item+ .card-footer-item") %>%
  html_attr("href") # part 5

# Create a dataframe
tibble(title = title,
       company = company,
       location = location,
       time = time,
       html = html) # part 6
# A tibble: 100 x 5
   title                        company                    location     time  html
   <chr>                        <chr>                      <chr>        <chr> <chr>
 1 Senior Python Developer      Payne, Roberts and Davis   Stewartbury~ 2021~ http~
 2 Energy engineer              Vasquez-Davidson           Christopher~ 2021~ http~
 3 Legal executive              Jackson, Chambers and Levy Port Ericab~ 2021~ http~
 4 Fitness centre manager       Savage-Bradley             East Seanvi~ 2021~ http~
 5 Product manager              Ramirez Inc                North Jamie~ 2021~ http~
 6 Medical technical officer    Rogers-Yates               Davidville,~ 2021~ http~
 7 Physiological scientist      Kramer-Klein               South Chris~ 2021~ http~
 8 Textile designer             Meyers-Johnson             Port Jonath~ 2021~ http~
 9 Television floor manager     Hughes-Williams            Osbornetown~ 2021~ http~
10 Waste management officer     Jones, Williams and Villa  Scotttown, ~ 2021~ http~
# i 90 more rows
```