Assignment_8

Mathew Katz

2022-10-17

Load in libraries needed:

```
library(XML)
library(RCurl)
library(jsonlite)
```

Read in HTML file to RStudio:

```
htmlurl <- getURL('https://raw.githubusercontent.com/MathewKatz/CUNYSPS/main/books.html')
x <- readHTMLTable(htmlurl)
df_html<-data.frame(x)
df_html
```

```
##   NULL.Book.Title     NULL.Author NULL.Originally.Published          NULL.Genre
## 1     The Martian       Andy Weir                      2011     Science Fiction
## 2          Immune Philipp Dettmer                      2021 Science Non-Fiction
## 3       The Stand    Stephen King                      1978      Horror Fiction
##   NULL.Page.Count
## 1             369
## 2             368
## 3            1152
```

For some reason, there is a 'NULL' in the column names. Let's remove it.

First let's look at the column names:

```
names(df_html)
```

```
## [1] "NULL.Book.Title"           "NULL.Author"
## [3] "NULL.Originally.Published" "NULL.Genre"
## [5] "NULL.Page.Count"
```

Change column names accordingly:

```
df_html <- setNames(df_html, c('Book_Title', 'Author','Originally_Published', 'Genre', 'Page_Count'))
df_html
```

```
##    Book_Title          Author Originally_Published               Genre
## 1 The Martian       Andy Weir                 2011     Science Fiction
## 2      Immune Philipp Dettmer                 2021 Science Non-Fiction
## 3   The Stand    Stephen King                 1978      Horror Fiction
##   Page_Count
## 1        369
## 2        368
## 3       1152
```

Read in XML file to RStudio:

```
xmlurl <- getURL('https://raw.githubusercontent.com/MathewKatz/CUNYSPS/main/books.xml')
y <- xmlParse(xmlurl)
df_xml <- xmlToDataFrame(y)
df_xml
```

```
##    Book_Title          Author Originally_Published               Genre
## 1 The Martian       Andy Weir                 2011     Science Fiction
## 2      Immune Philipp Dettmer                 2021 Science Non-Fiction
## 3   The Stand    Stephen King                 1978      Horror Fiction
##   Page_Count
## 1        369
## 2        368
## 3       1152
```

Read in JSON file to RStudio:

```
jsonurl <- getURL('https://raw.githubusercontent.com/MathewKatz/CUNYSPS/main/books.json')
df_json <- fromJSON(jsonurl)
df_json
```

```
##    Book Title          Author Originally Published               Genre
## 1 The Martian       Andy Weir                 2011     Science Fiction
## 2      Immune Philipp Dettmer                 2021 Science Non-Fiction
## 3   The Stand    Stephen King                 1978      Horror Fiction
##   Page Count
## 1        369
## 2        368
## 3       1152
```

The three dataframes are identical!