Working with XML and JSON in R

Pricilla

2025-10-08

Pick three of your favorite books on one of your favorite subjects. At least one of the books should have more than one author. For each book, include the title, authors, and two or three other attributes that you find interesting.

Take the information that you've selected about these three books, and separately create three files which store the book's information in HTML (using an html table), XML, and JSON formats (e.g. "books.html", "books.xml", and "books.json").

First I create a CSV that I can use to convert to any file type just because CSV is the most beloved flexible data type. I will load xml, html and JSON files into R too.

```
books <- read.csv("https://raw.githubusercontent.com/prnakyazze94/Data_607/refs/heads/main/Books2.csv")
```

```
# View first rows
head(books)
```

```
##                                                  title
## 1                                    He's Not My Type
## 2                                          Thesaurize
## 3                                         Moral Stand
## 4                                LLC Beginner's Guide
## 5 How to Talk to Anyone and Enchant Them into Liking You
## 6                Negotiating from a Position of Weakness
##
## 1
## 2
## 3
## 4       How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Zero
## 5                   Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relati
## 6 An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Proven Techniques
##             authors
## 1      Meghan Quinn
## 2      Dakota Krout
## 3 Daniel Schinhofen
## 4      Walter Grant
## 5        Carl Wolfe
## 6   David Whitehead
##                                                                                coauthors
## 1 Connor Crais, Erin Mallon, Teddy Hamilton, Jason Clarke, J.F. Harding, Kelsey Navarro-Foster
## 2                                                                             Luke Daniels
## 3                                                                          Andrea Parsneau
## 4                                                                            John Killawee
```

```
## 5                                                                  Tim Alexander
## 6                                                                Gerhard Weigelt
##   release_date language            stars rating
## 1     11-28-23  English 5 out of 5 stars    362
## 2     11-06-23  English 5 out of 5 stars    328
## 3     11-17-23  English 5 out of 5 stars    164
## 4     11-03-23  English 5 out of 5 stars     51
## 5     11-03-23  English 5 out of 5 stars     50
## 6     11-14-23  English 5 out of 5 stars     50
```

Data not displaying correctly because of the colon in raw data

```
kable(
  head(books[, c("title", "subtitle", "authors", "release_date", "language", "rating")])
)
```

title	subtitle	authors	release_date	language	rating
He's Not My Type	Vancouver Agitators Series, Book 4	Meghan Quinn	11-28-23	English	362
Thesaurize	The Completionist Chronicles, Book 10	Dakota Krout	11-06-23	English	328
Moral Stand	Aether's Revival, Book 7	Daniel Schinhofen	11-17-23	English	164
LLC Beginner's Guide	How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Zero Experience: A Complete Up-to-Date & Easy-to-Follow Guide	Walter Grant	11-03-23	English	51
How to Talk to Anyone and Enchant Them into Liking You	Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationships and Becoming the Most Likable Person in the Room	Carl Wolfe	11-03-23	English	50
Negotiating from a Position of Weakness	An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Proven Techniques for Building Empathy, Embracing Vulnerability, and More	David Whitehead	11-14-23	English	50

I can use original csv to output.

HTML created out of csv

```
html_table <- kable(
  head(books),
  "html",
  caption = "My Top 6 Books HTML Table"
) %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))

# Save the HTML table to a file
save_kable(html_table, "books_info.html")

# open the HTML in default browser
```

```
browseURL("books_info.html")

# Print table in RStudio Viewer

webshot("books_info.html", file = "books_info.png", vwidth = 992)
```

```
## file:///C:/Users/pricc/OneDrive/Documents/books_info.html screenshot completed
```

My Top 6 Books HTML Table

title	subtitle	authors	coauthors	release_date	language	stars	ratin
He's Not My Type	Vancouver Agitators Series, Book 4	Meghan Quinn	Connor Crais, Erin Mallon, Teddy Hamilton, Jason Clarke, J.F. Harding, Kelsey Navarro-Foster	11-28-23	English	5 out of 5 stars	36
Thesaurize	The Completionist Chronicles, Book 10	Dakota Krout	Luke Daniels	11-06-23	English	5 out of 5 stars	32
Moral Stand	Aether's Revival, Book 7	Daniel Schinhofen	Andrea Parsneau	11-17-23	English	5 out of 5 stars	16
LLC Beginner's Guide	How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Zero Experience: A Complete Up-to-Date & Easy-to-Follow Guide	Walter Grant	John Killawee	11-03-23	English	5 out of 5 stars	5
How to Talk to Anyone and Enchant Them into Liking You	Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationships and Becoming the Most Likable Person in the Room	Carl Wolfe	Tim Alexander	11-03-23	English	5 out of 5 stars	5
Negotiating from a Position of Weakness	An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Proven Techniques for Building Empathy, Embracing Vulnerability, and More	David Whitehead	Gerhard Weigelt	11-14-23	English	5 out of 5 stars	5

```
#html_table failed to print directly
```

JSON CREATED OUT OF CSV

```
# Select subset of columns
books_subset <- books[, c("title", "subtitle", "authors", "release_date", "language", "rating")]

# Convert to JSON (pretty format)
books_json <- toJSON(books_subset, pretty = TRUE)
```

```r
# Print JSON to console
cat(books_json)
```


```
## [
##   {
##     "title": "He's Not My Type",
##     "subtitle": "Vancouver Agitators Series, Book 4",
##     "authors": "Meghan Quinn",
##     "release_date": "11-28-23",
##     "language": "English",
##     "rating": 362
##   },
##   {
##     "title": "Thesaurize",
##     "subtitle": "The Completionist Chronicles, Book 10",
##     "authors": "Dakota Krout",
##     "release_date": "11-06-23",
##     "language": "English",
##     "rating": 328
##   },
##   {
##     "title": "Moral Stand",
##     "subtitle": "Aether's Revival, Book 7",
##     "authors": "Daniel Schinhofen",
##     "release_date": "11-17-23",
##     "language": "English",
##     "rating": 164
##   },
##   {
##     "title": "LLC Beginner's Guide",
##     "subtitle": "How to Successfully Start and Maintain a Limited Liability Company Even if You've G
##     "authors": "Walter Grant",
##     "release_date": "11-03-23",
##     "language": "English",
##     "rating": 51
##   },
##   {
##     "title": "How to Talk to Anyone and Enchant Them into Liking You",
##     "subtitle": "Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationsl
##     "authors": "Carl Wolfe",
##     "release_date": "11-03-23",
##     "language": "English",
##     "rating": 50
##   },
##   {
##     "title": "Negotiating from a Position of Weakness",
##     "subtitle": "An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Pr
##     "authors": "David Whitehead",
##     "release_date": "11-14-23",
##     "language": "English",
##     "rating": 50
##   }
## ]
```

```r
# Save JSON to file
write(books_json, "books_subset.json")
```

XML CREATED OUT OF CSV

```r
# Select subset of columns from original books csv
books_subset <- books[, c("title", "subtitle", "authors", "release_date", "language", "rating")]

# Create root XML node
books_xml <- newXMLNode("books")

# Loop through each row to add book nodes
apply(books_subset, 1, function(row) {
  book_node <- newXMLNode("book", parent = books_xml)
  newXMLNode("title", row["title"], parent = book_node)
  newXMLNode("subtitle", row["subtitle"], parent = book_node)
  newXMLNode("authors", row["authors"], parent = book_node)
  newXMLNode("release_date", row["release_date"], parent = book_node)
  newXMLNode("language", row["language"], parent = book_node)
  newXMLNode("rating", row["rating"], parent = book_node)
})
```

```
## [[1]]
## <rating>362</rating>
##
## [[2]]
## <rating>328</rating>
##
## [[3]]
## <rating>164</rating>
##
## [[4]]
## <rating> 51</rating>
##
## [[5]]
## <rating> 50</rating>
##
## [[6]]
## <rating> 50</rating>
```

```r
# Save XML to file
saveXML(books_xml, file = "books_subset.xml")
```

```
## [1] "books_subset.xml"
```

```r
# Print XML in R console
cat(saveXML(books_xml))
```

```
## <books>
##   <book>
##     <title>He's Not My Type</title>
```

```
##     <subtitle>Vancouver Agitators Series, Book 4</subtitle>
##     <authors>Meghan Quinn</authors>
##     <release_date>11-28-23</release_date>
##     <language>English</language>
##     <rating>362</rating>
##   </book>
##   <book>
##     <title>Thesaurize</title>
##     <subtitle>The Completionist Chronicles, Book 10</subtitle>
##     <authors>Dakota Krout</authors>
##     <release_date>11-06-23</release_date>
##     <language>English</language>
##     <rating>328</rating>
##   </book>
##   <book>
##     <title>Moral Stand</title>
##     <subtitle>Aether's Revival, Book 7</subtitle>
##     <authors>Daniel Schinhofen</authors>
##     <release_date>11-17-23</release_date>
##     <language>English</language>
##     <rating>164</rating>
##   </book>
##   <book>
##     <title>LLC Beginner's Guide</title>
##     <subtitle>How to Successfully Start and Maintain a Limited Liability Company Even if You've Got 
##     <authors>Walter Grant</authors>
##     <release_date>11-03-23</release_date>
##     <language>English</language>
##     <rating> 51</rating>
##   </book>
##   <book>
##     <title>How to Talk to Anyone and Enchant Them into Liking You</title>
##     <subtitle>Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationship
##     <authors>Carl Wolfe</authors>
##     <release_date>11-03-23</release_date>
##     <language>English</language>
##     <rating> 50</rating>
##   </book>
##   <book>
##     <title>Negotiating from a Position of Weakness</title>
##     <subtitle>An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Prove
##     <authors>David Whitehead</authors>
##     <release_date>11-14-23</release_date>
##     <language>English</language>
##     <rating> 50</rating>
##   </book>
## </books>
```

Write R code, using your packages of choice, to load the information from each of the three sources into separate R data frames. Are the three data frames identical?

READING MY RAW FILES INTO R

HTML LOADED INTO R

LOAD HTML FILE INTO R

```r
 # Load html file from github and save it as a webpage

url <- "https://raw.githubusercontent.com/prnakyazze94/Data_607/refs/heads/main/Bookstable.html"

webpage <- read_html(url, encoding = "UTF-8")  # ensure UTF-8

# Print raw HTML to console

cat(as.character(webpage))
```

```
## <!DOCTYPE html>
## <html>
## <head>
## <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
## <meta charset="UTF-8">
## <title>Books Table</title>
## <style>
##     table { border-collapse: collapse; width: 100%; }
##     th, td { border: 1px solid black; padding: 8px; text-align: left; }
##     th { background-color: #f2f2f2; }
##   </style>
## </head>
## <body>
##   <h2>Books List</h2>
##   <table>
## <thead><tr>
## <th>Title</th>
##         <th>Subtitle</th>
##         <th>Authors</th>
##         <th>Coauthors</th>
##         <th>Release Date</th>
##         <th>Language</th>
##         <th>Stars</th>
##         <th>Rating</th>
##       </tr></thead>
## <tbody>
## <tr>
## <td>He’s Not My Type</td>
##         <td>Vancouver Agitators Series, Book 4</td>
##         <td>Meghan Quinn</td>
##         <td>Connor Crais, Erin Mallon, Teddy Hamilton, Jason Clarke, J.F. Harding, Kelsey Navarro-Fos
##         <td>11-28-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>362</td>
##       </tr>
## <tr>
## <td>Thesaurize</td>
##         <td>The Completionist Chronicles, Book 10</td>
##         <td>Dakota Krout</td>
##         <td>Luke Daniels</td>
##         <td>11-06-23</td>
##         <td>English</td>
```

```
##         <td>5 out of 5 stars</td>
##         <td>328</td>
##       </tr>
## <tr>
## <td>Moral Stand</td>
##         <td>Aether's Revival, Book 7</td>
##         <td>Daniel Schinhofen</td>
##         <td>Andrea Parsneau</td>
##         <td>11-17-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>164</td>
##       </tr>
## <tr>
## <td>LLC Beginner's Guide</td>
##         <td>How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Ze
##         <td>Walter Grant</td>
##         <td>John Killawee</td>
##         <td>11-03-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>51</td>
##       </tr>
## <tr>
## <td>How to Talk to Anyone and Enchant Them into Liking You</td>
##         <td>Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationships 
##         <td>Carl Wolfe</td>
##         <td>Tim Alexander</td>
##         <td>11-03-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>50</td>
##       </tr>
## <tr>
## <td>Negotiating from a Position of Weakness</td>
##         <td>An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Proven 
##         <td>David Whitehead</td>
##         <td>Gerhard Weigelt</td>
##         <td>11-14-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>50</td>
##       </tr>
## </tbody>
## </table>
## </body>
## </html><html>
## <head>
## <meta charset="UTF-8">
## <title>Books Table</title>
## <style>
##     table { border-collapse: collapse; width: 100%; }
##     th, td { border: 1px solid black; padding: 8px; text-align: left; }
##     th { background-color: #f2f2f2; }
```

```
##   </style>
## </head>
## <body>
##   <h2>Books List</h2>
##   <table>
## <thead><tr>
## <th>Title</th>
##         <th>Subtitle</th>
##         <th>Authors</th>
##         <th>Coauthors</th>
##         <th>Release Date</th>
##         <th>Language</th>
##         <th>Stars</th>
##         <th>Rating</th>
##       </tr></thead>
## <tbody>
## <tr>
## <td>He's Not My Type</td>
##         <td>Vancouver Agitators Series, Book 4</td>
##         <td>Meghan Quinn</td>
##         <td>Connor Crais, Erin Mallon, Teddy Hamilton, Jason Clarke, J.F. Harding, Kelsey Navarro-Fos
##         <td>11-28-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>362</td>
##       </tr>
## <tr>
## <td>Thesaurize</td>
##         <td>The Completionist Chronicles, Book 10</td>
##         <td>Dakota Krout</td>
##         <td>Luke Daniels</td>
##         <td>11-06-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>328</td>
##       </tr>
## <tr>
## <td>Moral Stand</td>
##         <td>Aether's Revival, Book 7</td>
##         <td>Daniel Schinhofen</td>
##         <td>Andrea Parsneau</td>
##         <td>11-17-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>164</td>
##       </tr>
## <tr>
## <td>LLC Beginner's Guide</td>
##         <td>How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Ze
##         <td>Walter Grant</td>
##         <td>John Killawee</td>
##         <td>11-03-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
```

```
##         <td>51</td>
##       </tr>
## <tr>
## <td>How to Talk to Anyone and Enchant Them into Liking You</td>
##         <td>Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationships a
##         <td>Carl Wolfe</td>
##         <td>Tim Alexander</td>
##         <td>11-03-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>50</td>
##       </tr>
## <tr>
## <td>Negotiating from a Position of Weakness</td>
##         <td>An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Proven T
##         <td>David Whitehead</td>
##         <td>Gerhard Weigelt</td>
##         <td>11-14-23</td>
##         <td>English</td>
##         <td>5 out of 5 stars</td>
##         <td>50</td>
##       </tr>
## </tbody>
## </table>
## </body>
## </html>
```

Create table from webpage. Html_node selects a single node from an HTML document that matches a CSS selector.

While table is the CSS selector, so html_node(table) selects the first table element in the

HTML LOADED INTO R TO CREATE A DF

```
books_htmltable <- webpage %>%
  html_node("table") %>%
  html_table(fill = TRUE)



# Use stringi to trim whitespace and remove non-ASCII characters
books_htmltable <- books_htmltable %>%
  mutate(across(everything(), ~ stri_trim_both(.)))  # trim spaces

# remove problematic non-ASCII characters
books_htmltable <- books_htmltable %>%
  mutate(across(everything(), ~ stri_trans_general(., "Latin-ASCII")))



head(books_htmltable)
```

```
## # A tibble: 6 x 8
##   Title          Subtitle Authors Coauthors ‘Release Date‘ Language Stars Rating
```

```
##   <chr>          <chr>    <chr>   <chr>     <chr>          <chr>    <chr> <chr>
## 1 He's Not My T~ Vancouv~ Meghan~ Connor C~ 11-28-23       English  5 ou~ 362
## 2 Thesaurize     The Com~ Dakota~ Luke Dan~ 11-06-23       English  5 ou~ 328
## 3 Moral Stand    Aether'~ Daniel~ Andrea P~ 11-17-23       English  5 ou~ 164
## 4 LLC Beginner'~ How to ~ Walter~ John Kil~ 11-03-23       English  5 ou~ 51
## 5 How to Talk t~ Proven ~ Carl W~ Tim Alex~ 11-03-23       English  5 ou~ 50
## 6 Negotiating f~ An 18 S~ David ~ Gerhard ~ 11-14-23       English  5 ou~ 50
```

```
write.csv(books_htmltable, "books_from_html.csv", row.names = FALSE)
```

SELECT ONLY A FEW COLUMNS SO I CAN TELL THE DIFFERNCE IN HTML DF

```
books_htmltable %>%
  select(Title, Authors, `Release Date`, Language, Stars, Rating) %>%
  head() %>%  # just to preview first few rows
  kable(caption = "Preview of HTML Books Data")
```

Table 2: Preview of HTML Books Data

Title	Authors	Release Date	Language	Stars	Rating
He's Not My Type	Meghan Quinn	11-28-23	English	5 out of 5 stars	362
Thesaurize	Dakota Krout	11-06-23	English	5 out of 5 stars	328
Moral Stand	Daniel Schinhofen	11-17-23	English	5 out of 5 stars	164
LLC Beginner's Guide	Walter Grant	11-03-23	English	5 out of 5 stars	51
How to Talk to Anyone and Enchant Them into Liking You	Carl Wolfe	11-03-23	English	5 out of 5 stars	50
Negotiating from a Position of Weakness	David Whitehead	11-14-23	English	5 out of 5 stars	50

```
# Preview the first few rows using kable
kable(head(books_htmltable))
```

Title	Subtitle	Authors	Coauthors	Release Date	Language	Stars	Rating
He's Not My Type	Vancouver Agitators Series, Book 4	Meghan Quinn	Connor Crais, Erin Mallon, Teddy Hamilton, Jason Clarke, J.F. Harding, Kelsey Navarro-Foster	11-28-23	English	5 out of 5 stars	362
Thesaurize	The Completionist Chronicles, Book 10	Dakota Krout	Luke Daniels	11-06-23	English	5 out of 5 stars	328

Title	Subtitle	Authors	Coauthors	Release Date	Language	Stars	Rating
Moral Stand	Aether's Revival, Book 7	Daniel Schinhofen	Andrea Parsneau	11-17-23	English	5 out of 5 stars	164
LLC Beginner's Guide	How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Zero Experience: A Complete Up-to-Date & Easy-to-Follow Guide	Walter Grant	John Killawee	11-03-23	English	5 out of 5 stars	51
How to Talk to Anyone and Enchant Them into Liking You	Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationships and Becoming the Most Likable Person in the Room	Carl Wolfe	Tim Alexander	11-03-23	English	5 out of 5 stars	50
Negotiating from a Position of Weakness	An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Proven Techniques for Building Empathy, Embracing Vulnerability, and More	David Whitehead	Gerhard Weigelt	11-14-23	English	5 out of 5 stars	50

XML

XML FILE LOADED INTO R

```
# Force UTF-8 locale in this session because , and // are causing errors.
Sys.setlocale("LC_CTYPE", "en_US.UTF-8")
```

```
## [1] "en_US.UTF-8"
```

```
 #Load XML file from github and save it as urlxml
# Define the URL
urlxml <- "https://raw.githubusercontent.com/prnakyazze94/Data_607/refs/heads/main/BooksXML.xml"

# Read the XML file directly from the URL

books_xml <- read_xml(urlxml, encoding = "UTF-8")

# View the XML structure
print(books_xml)
```

```
## {xml_document}
## <books>
## [1] <book>\n  <title>He's Not My Type</title>\n  <subtitle>Vancouver Agitator ...
## [2] <book>\n  <title>Thesaurize</title>\n  <subtitle>The Completionist Chroni ...
## [3] <book>\n  <title>Moral Stand</title>\n  <subtitle>Aether's Revival, Book  ...
## [4] <book>\n  <title>LLC Beginner's Guide</title>\n  <subtitle>How to Success ...
## [5] <book>\n  <title>How to Talk to Anyone and Enchant Them into Liking You</ ...
## [6] <book>\n  <title>Negotiating from a Position of Weakness</title>\n  <subt ...
```

```r
# explore it more nicely
xml_structure(books_xml)
```

```
## <books>
##   <book>
##     <title>
##       {text}
##     <subtitle>
##       {text}
##     <authors>
##       <author>
##         {text}
##     <coauthors>
##       <coauthor>
##         {text}
##       <coauthor>
##         {text}
##       <coauthor>
##         {text}
##       <coauthor>
##         {text}
##       <coauthor>
##         {text}
##       <coauthor>
##         {text}
##     <release_date>
##       {text}
##     <language>
##       {text}
##     <stars>
##       {text}
##     <rating>
##       {text}
##   <book>
##     <title>
##       {text}
##     <subtitle>
##       {text}
##     <authors>
##       <author>
##         {text}
##     <coauthors>
##       <coauthor>
##         {text}
##     <release_date>
##       {text}
##     <language>
##       {text}
##     <stars>
##       {text}
##     <rating>
##       {text}
##   <book>
```

```
##     <title>
##       {text}
##     <subtitle>
##       {text}
##     <authors>
##       <author>
##         {text}
##     <coauthors>
##       <coauthor>
##         {text}
##     <release_date>
##       {text}
##     <language>
##       {text}
##     <stars>
##       {text}
##     <rating>
##       {text}
##   <book>
##     <title>
##       {text}
##     <subtitle>
##       {text}
##     <authors>
##       <author>
##         {text}
##     <coauthors>
##       <coauthor>
##         {text}
##     <release_date>
##       {text}
##     <language>
##       {text}
##     <stars>
##       {text}
##     <rating>
##       {text}
##   <book>
##     <title>
##       {text}
##     <subtitle>
##       {text}
##     <authors>
##       <author>
##         {text}
##     <coauthors>
##       <coauthor>
##         {text}
##     <release_date>
##       {text}
##     <language>
##       {text}
##     <stars>
##       {text}
```

```
##     <rating>
##       {text}
##   <book>
##     <title>
##       {text}
##     <subtitle>
##       {text}
##     <authors>
##       <author>
##         {text}
##     <coauthors>
##       <coauthor>
##         {text}
##     <release_date>
##       {text}
##     <language>
##       {text}
##     <stars>
##       {text}
##     <rating>
##       {text}
```

SAVE XML FILE INTO R data frames Extract fields into a data frame

Extract simple single-value fields (like title, subtitle, release_date, etc.) directly

```
# Force UTF-8 locale in this session because , and // are causing errors.
Sys.setlocale("LC_CTYPE", "en_US.UTF-8")
```

```
## [1] "en_US.UTF-8"
```

```
# Extract <book> nodes

book_nodes <- xml_find_all(books_xml, ".//book")

#create df
books_dfxml <- tibble(
  title = xml_text(xml_find_all(book_nodes, "title")),
  subtitle = xml_text(xml_find_all(book_nodes, "subtitle")),
  release_date = xml_text(xml_find_all(book_nodes, "release_date")),
  language = xml_text(xml_find_all(book_nodes, "language")),
  stars = xml_text(xml_find_all(book_nodes, "stars")),
  rating = xml_text(xml_find_all(book_nodes, "rating"))
)
```

Extract authors that are nested separately

```
# Extract authors and coauthors
authors_list <- lapply(book_nodes, function(x) {
  xml_text(xml_find_all(x, ".//authors/author"))
})

coauthors_list <- lapply(book_nodes, function(x) {
```

```
  xml_text(xml_find_all(x, ".//coauthors/coauthor"))
})

#  Combine all into one data frame ---
books_dfxml <- books_dfxml %>%
  mutate(
    authors = sapply(authors_list, function(x) paste(x, collapse = ", ")),
    coauthors = sapply(coauthors_list, function(x) paste(x, collapse = ", "))
  )

# View result ---
print(books_dfxml)
```

```
## # A tibble: 6 x 8
##   title            subtitle release_date language stars rating authors coauthors
##   <chr>            <chr>    <chr>        <chr>    <chr> <chr>  <chr>   <chr>
## 1 He's Not My Type Vancouv~ 11-28-23     English  5 ou~ 362    Meghan~ Connor C~
## 2 Thesaurize       The Com~ 11-06-23     English  5 ou~ 328    Dakota~ Luke Dan~
## 3 Moral Stand      Aether'~ 11-17-23     English  5 ou~ 164    Daniel~ Andrea P~
## 4 LLC Beginner's ~ How to ~ 11-03-23     English  5 ou~ 51     Walter~ John Kil~
## 5 How to Talk to ~ Proven ~ 11-03-23     English  5 ou~ 50     Carl W~ Tim Alex~
## 6 Negotiating fro~ An 18 S~ 11-14-23     English  5 ou~ 50     David ~ Gerhard ~
```

```
# Preview the first few rows using kable
kable(head(books_dfxml))
```

title	subtitle	release_date	language	stars	rating	authors	coauthors
He's Not My Type	Vancouver Agitators Series, Book 4	11-28-23	English	5 out of 5 stars	362	Meghan Quinn	Connor Crais, Erin Mallon, Teddy Hamilton, Jason Clarke, J.F. Harding, Kelsey Navarro-Foster
Thesaurize	The Completionist Chronicles, Book 10	11-06-23	English	5 out of 5 stars	328	Dakota Krout	Luke Daniels
Moral Stand	Aether's Revival, Book 7	11-17-23	English	5 out of 5 stars	164	Daniel Schinhofen	Andrea Parsneau
LLC Beginner's Guide	How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Zero Experience: A Complete Up-to-Date & Easy-to-Follow Guide	11-03-23	English	5 out of 5 stars	51	Walter Grant	John Killawee
How to Talk to Anyone and Enchant Them into Liking You	Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationships and Becoming the Most Likable Person in the Room	11-03-23	English	5 out of 5 stars	50	Carl Wolfe	Tim Alexander

title	subtitle	release_date	language	stars	rating	authors	coauthors
Negotiating from a Position of Weakness	An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength. Proven Techniques for Building Empathy, Embracing Vulnerability, and More	11-14-23	English	5 out of 5 stars	50	David Whitehead	Gerhard Weigelt

SELECT ONLY A FEW COLUMNS SO I CAN TELL THE DIFFERNCE IN XML DF

```
books_dfxml %>%
  select(title, authors, `release_date`, language, stars, rating) %>%
  head() %>%  # just to preview first few rows
  kable(caption = "Preview of XML Books Data")
```

Table 5: Preview of XML Books Data

title	authors	release_date	language	stars	rating
He's Not My Type	Meghan Quinn	11-28-23	English	5 out of 5 stars	362
Thesaurize	Dakota Krout	11-06-23	English	5 out of 5 stars	328
Moral Stand	Daniel Schinhofen	11-17-23	English	5 out of 5 stars	164
LLC Beginner's Guide	Walter Grant	11-03-23	English	5 out of 5 stars	51
How to Talk to Anyone and Enchant Them into Liking You	Carl Wolfe	11-03-23	English	5 out of 5 stars	50
Negotiating from a Position of Weakness	David Whitehead	11-14-23	English	5 out of 5 stars	50

JSON FILE LOADED INTO R

```
# Define the GitHub raw URL
urljson <- "https://raw.githubusercontent.com/prnakyazze94/Data_607/refs/heads/main/Booksjson.json"

# Read the JSON directly from the URL
books_json <- fromJSON(urljson)

# View the structure
str(books_json)
```

```
## 'data.frame':    7 obs. of  15 variables:
##  $ title        : chr  "He's Not My Type" "Thesaurize" "Moral Stand" "LLC Beginner's Guide" ...
##  $ subtitle     : chr  "Vancouver Agitators Series, Book 4" "The Completionist Chronicles, Book 10" "
##  $ authors      :List of 7
##   ..$ : chr "Meghan Quinn"
##   ..$ : chr "Dakota Krout"
##   ..$ : chr "Daniel Schinhofen"
##   ..$ : chr "Walter Grant"
##   ..$ : chr "Carl Wolfe"
```

```
##   ..$ : chr "David Whitehead"
##   ..$ : chr "Henry Matthias"
##  $ narrators    :List of 7
##   ..$ : chr  "Connor Crais" "Erin Mallon" "Teddy Hamilton" "Jason Clarke" ...
##   ..$ : chr "Luke Daniels"
##   ..$ : chr "Andrea Parsneau"
##   ..$ : chr "John Killawee"
##   ..$ : chr "Tim Alexander"
##   ..$ : chr "Gerhard Weigelt"
##   ..$ : chr "KC Wayman"
##  $ series       : chr  "The Vancouver Agitators" "The Completionist Chronicles" "Aether's Revival" "
##  $ length       : chr  "Length: 11 hrs and 40 mins" "Length: 11 hrs and 40 mins" "Length: 13 hrs and
##  $ release_date : chr  "Release date: 11-28-23" "Release date: 11-06-23" "Release date: 11-17-23" "Re
##  $ language     : chr  "Language: English" "Language: English" "Language: English" "Language: Englis
##  $ rating       : chr  "5 out of 5 stars" "5 out of 5 stars" "5 out of 5 stars" "5 out of 5 stars" .
##  $ no_of_ratings: chr  "362 ratings" "328 ratings" "164 ratings" "51 ratings" ...
##  $ regular_price: chr  "$24.95" "$24.95" "$24.95" "$14.95" ...
##  $ sales_price  : chr  "" "" "" "" ...
##  $ category     : logi  NA NA NA NA NA NA ...
##  $ genres       :List of 7
##   ..$ : list()
##   ..$ : list()
##   ..$ : list()
##   ..$ : list()
##   ..$ : list()
##   ..$ : list()
##   ..$ : list()
##  $ url          : chr  "https://www.audible.com/pd/Hes-Not-My-Type-Audiobook/B0CMJSQSJF" "https://ww
```

CONVERT JSON FILE INTO DF

```r
#  convert to a data frame
books_dfjson <- as.data.frame(books_json)

# View the first few rows
head(books_dfjson)
```

```
##                                                 title
## 1                                 He's Not My Type
## 2                                       Thesaurize
## 3                                      Moral Stand
## 4                                LLC Beginner's Guide
## 5 How to Talk to Anyone and Enchant Them into Liking You
## 6               Negotiating from a Position of Weakness
##
## 1
## 2
## 3
## 4        How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Zero
## 5                   Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relati
## 6 An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength: Proven Techniques
##            authors
## 1     Meghan Quinn
```

```
## 2      Dakota Krout
## 3 Daniel Schinhofen
## 4      Walter Grant
## 5        Carl Wolfe
## 6   David Whitehead
##                                                                                                narrators
## 1 Connor Crais, Erin Mallon, Teddy Hamilton, Jason Clarke, J.F. Harding, Kelsey Navarro-Foster
## 2                                                                                             Luke Daniels
## 3                                                                                           Andrea Parsneau
## 4                                                                                             John Killawee
## 5                                                                                             Tim Alexander
## 6                                                                                           Gerhard Weigelt
##                          series                     length
## 1      The Vancouver Agitators Length: 11 hrs and 40 mins
## 2 The Completionist Chronicles Length: 11 hrs and 40 mins
## 3             Aether's Revival  Length: 13 hrs and 3 mins
## 4                               Length: 3 hrs and 8 mins
## 5                              Length: 3 hrs and 21 mins
## 6                               Length: 7 hrs and 7 mins
##             release_date          language           rating no_of_ratings
## 1 Release date: 11-28-23 Language: English 5 out of 5 stars   362 ratings
## 2 Release date: 11-06-23 Language: English 5 out of 5 stars   328 ratings
## 3 Release date: 11-17-23 Language: English 5 out of 5 stars   164 ratings
## 4 Release date: 11-03-23 Language: English 5 out of 5 stars    51 ratings
## 5 Release date: 11-03-23 Language: English 5 out of 5 stars    50 ratings
## 6 Release date: 11-14-23 Language: English 5 out of 5 stars    50 ratings
##   regular_price sales_price category genres
## 1        $24.95                   NA   NULL
## 2        $24.95                   NA   NULL
## 3        $24.95                   NA   NULL
## 4        $14.95                   NA   NULL
## 5        $14.95                   NA   NULL
## 6        $19.95                   NA   NULL
##
## 1                                        https://www.audible.com/pd/Hes-Not-My-Type-Audiobook/B0CMJS
## 2                                             https://www.audible.com/pd/Thesaurize-Audiobook/B0CMR1
## 3                                           https://www.audible.com/pd/Moral-Stand-Audiobook/B0CNKV
## 4                                    https://www.audible.com/pd/LLC-Beginners-Guide-Audiobook/B0CMFJ
## 5 https://www.audible.com/pd/How-to-Talk-to-Anyone-and-Enchant-Them-into-Liking-You-Audiobook/B0CMF9
## 6                https://www.audible.com/pd/Negotiating-from-a-Position-of-Weakness-Audiobook/B0CNB1
```

```
# Preview the first few rows using kable
kable(head(books_dfjson))
```

title	subtitle	authors	narrators	series	length	release_date	language	rating	number_of_ratings	price	number_of_pages	publication_year	product_url
He's Not My Type	Vancouver Agitators Series, Book 4	Meghan Quinn	Connor Crais , Erin Mallon , Teddy Hamilton , Jason Clarke , J.F. Harding , Kelsey Navarro-Foster	The Vancouver Agitators	Length: 11 hrs and 40 mins	Release date: 11-28-23	English	5 out of 5 stars	362 ratings	$24.95	NA	NULL	https://www.audible.com/pd/Hes-Not-My-Type-Audiobook/B0CMJSQSJF
Thesaurize	The Completionist Chronicles, Book 10	Dakota Krout	Lake Daniels	The Completionist Chronicles	Length: 11 hrs and 40 mins	Release date: 11-06-23	English	5 out of 5 stars	328 ratings	$24.95	NA	NULL	https://www.audible.com/pd/Thesaurize-Audiobook/B0CMR1XPQY
Moral Stand	Aether's Revival, Book 7	Daniel Schinhofen	Andrea Parsneau	Aether's Revival	Length: 13 hrs and 3 mins	Release date: 11-17-23	English	5 out of 5 stars	164 ratings	$24.95	NA	NULL	https://www.audible.com/pd/Moral-Stand-Audiobook/B0CNKVXCJ6
LLC Beginner's Guide	How to Successfully Start and Maintain a Limited Liability Company Even if You've Got Zero Experience: A Complete Up-to-Date & Easy-to-Follow Guide	Walter Grant	John Killawee		Length: 3 hrs and 8 mins	Release date: 11-03-23	English	5 out of 5 stars	51 ratings	$14.95	NA	NULL	https://www.audible.com/pd/LLC-Beginners-Guide-Audiobook/B0CMFJ7Y64

title	subtitle	authors	narrators	series	length	release_date	language	rating	no_of_ratings	price	sale_price	publisher	image	audio_samples
How to Talk to Anyone and Enchant Them into Liking You	Proven Techniques to Become a People-Magnet by Building Positive, Lasting Relationships and Becoming the Most Likable Person in the Room	Carl Wolfe	Tim Alexander		Length: 3 hrs and 21 mins	Release date: 11-03-23	English	5 out of 5 stars	50 ratings	$14.95	NA	NULL	https://www.audible.com/pd/How-to-Talk-to-Anyone-and-Enchant-Them-into-Liking-You-Audiobook/B0CMF9B353
Negotiating from a Position of Weakness	An 18 Step Comprehensive Negotiation System to Turn Vulnerability into Strength: Proven Techniques for Building Empathy, Embracing Vulnerability, and More	David White	Gerhard Weigelt		Length: 7 hrs and 7 mins	Release date: 11-14-23	English	5 out of 5 stars	50 ratings	$19.95	NA	NULL	https://www.audible.com/pd/Negotiating-from-a-Position-of-Weakness-Audiobook/B0CNB17GP4

SELECT ONLY A FEW COLUMNS SO I CAN TELL THE DIFFERNCE IN JSON DF

```
books_dfjson %>%
  select(title, authors, `release_date`, language, rating,no_of_ratings ) %>%
  head() %>%  # just to preview first few rows
  kable(caption = "Preview of JSON Books Data")
```

Table 7: Preview of JSON Books Data

title	authors	release_date	language	rating	no_of_ratings
He's Not My Type	Meghan Quinn	Release date: 11-28-23	Language: English	5 out of 5 stars	362 ratings
Thesaurize	Dakota Krout	Release date: 11-06-23	Language: English	5 out of 5 stars	328 ratings
Moral Stand	Daniel Schinhofen	Release date: 11-17-23	Language: English	5 out of 5 stars	164 ratings
LLC Beginner's Guide	Walter Grant	Release date: 11-03-23	Language: English	5 out of 5 stars	51 ratings
How to Talk to Anyone and Enchant Them into Liking You	Carl Wolfe	Release date: 11-03-23	Language: English	5 out of 5 stars	50 ratings

title	authors	release_date	language	rating	no_of_ratings
Negotiating from a Position of Weakness	David Whitehead	Release date: 11-14-23	Language: English	5 out of 5 stars	50 ratings

Are the three data frames identical?

No, the three data frames are not identical.

JSON retains nested structures (authors/narrators arrays), XML and HTML have flattened text. JSON requires text cleaning, XML and HTML are more ready for analysis

XML and HTML capture similar core fields but lose some detail and may have formatting inconsistencies.

For analysis, I would likely need to standardize column names and formats before comparing or combining them. For example I have title and Title.Language: English and English.

I personally like the xml file format the best when converted into a df, which comes as a suprise. I have not had a lot of experience with XML files.