DATA 607 Assignment 7

Subhalaxmi Rout

3/15/2020

Working with XML and JSON in R

1. Introduction

For this assignment, I have created 3 files manually they are:

- books.html
- books.xml
- books.json

I have stored all the file in Github repository and using different library function in R read the file.

1.1 Column names of the data set

There are total 5 columns in the book data set. They are:

- Name
- Author
- Cost
- Pages
- PublicationDate

2. Load the required libraries

```
#install.packages("XML")
#install.packages("RCurl")
#install.packages("jsonlite")
#install.packages("rvest")
#install.packages("DT")
#install.packages("htmlTable")
library(XML)
library(RCurl)
library(jsonlite)
library(DT)
library(rvest)
```

```
## Loading required package: xml2
```

```
##
## Attaching package: 'rvest'

## The following object is masked from 'package:XML':
##
##     xml
```

```
library(htmlTable)
```

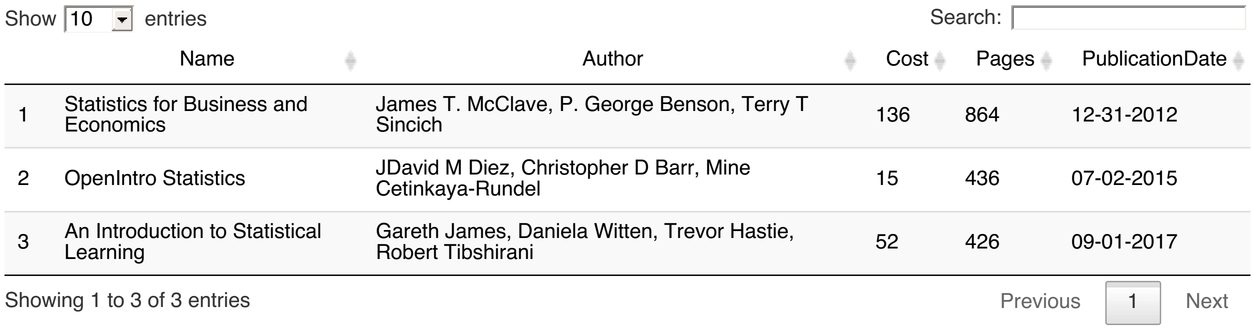
3.HTML File

3.1 Code for HTML file

```
<html>

<head>
<title>Book</title>
</head>

<body>

  <table>

  <tr>
    <th>Name</th>
    <th>Author</th>
    <th>Cost($)</th>
    <th>Pages</th>
    <th>PublicationDate</th>
  </tr>

  <tr>
  <td>Statistics for Business and Economics</td>
  <td>James T. McClave, P. George Benson, Terry T Sincich</td>
  <td>136</td>
  <td>864</td>
  <td>12-31-2012</td>
  </tr>

  <tr>
  <td>OpenIntro Statistics</td>
  <td>JDavid M Diez, Christopher D Barr, Mine Cetinkaya-Rundel</td>
  <td>15</td>
  <td>436</td>
  <td>07-02-2015</td>
  </tr>

  <tr>
  <td>An Introduction to Statistical Learning</td>
  <td>Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani</td>
  <td>52</td>
  <td>426</td>
  <td>09-01-2017</td>
  </tr>

  </table>
</body>
</html>
```

3.2 Read HTML file from Github repository

```r
# get the Github repo URL
htmlURL <- "https://raw.githubusercontent.com/SubhalaxmiRout002/DATA-607-Assignment-7/master/book.html"
#stored the URL
htmlContent <- getURLContent(htmlURL)
#read HTML table
booksHTML <- readHTMLTable(htmlContent, stringsAsFactors=FALSE)
# get all the values
booksHTML <- booksHTML[[1]]
#diaplay table data
datatable(booksHTML)
```



Show 10 entries Search:

	Name	Author	Cost	Pages	PublicationDate
1	Statistics for Business and Economics	James T. McClave, P. George Benson, Terry T Sincich	136	864	12-31-2012
2	OpenIntro Statistics	JDavid M Diez, Christopher D Barr, Mine Cetinkaya-Rundel	15	436	07-02-2015
3	An Introduction to Statistical Learning	Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani	52	426	09-01-2017

Showing 1 to 3 of 3 entries

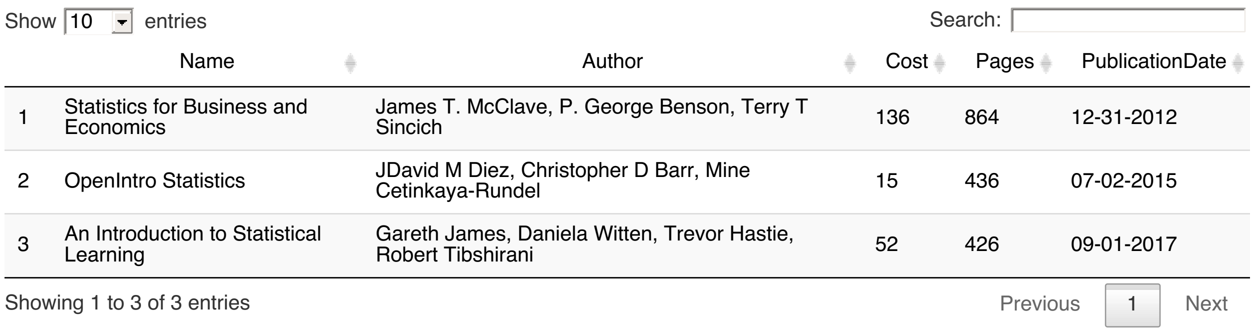
4. XML File

4.1 Code for XML file

```
<records>

   <book>
      <Name>Statistics for Business and Economics</Name>
      <Author>James T. McClave, P. George Benson, Terry T Sincich</Author>
      <Cost>136</Cost>
      <Pages>864</Pages>
      <PublicationDate>12-31-2012</PublicationDate>
   </book>

   <book>
      <Name>OpenIntro Statistics</Name>
      <Author>JDavid M Diez, Christopher D Barr, Mine Cetinkaya-Rundel</Author>
      <Cost>15</Cost>
      <Pages>436</Pages>
      <PublicationDate>07-02-2015</PublicationDate>
   </book>

   <book>
      <Name>An Introduction to Statistical Learning</Name>
      <Author>Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani</Author>
      <Cost>52</Cost>
      <Pages>426</Pages>
      <PublicationDate>09-01-2017</PublicationDate>
   </book>

</records>
```

4.2 Read XML file from Github repository

```r
library("methods")
# stored data in a var
xmlURL <- "https://raw.githubusercontent.com/SubhalaxmiRout002/DATA-607-Assignment-7/master/book.xml"
xmlContent <- getURLContent(xmlURL)
#pass the content
booksXMLparsed <- xmlParse(xmlContent)
#put in to data frame
booksXML <- xmlToDataFrame(booksXMLparsed, stringsAsFactors=FALSE)
#view data
datatable(booksXML)
```



	Name	Author	Cost	Pages	PublicationDate
1	Statistics for Business and Economics	James T. McClave, P. George Benson, Terry T Sincich	136	864	12-31-2012
2	OpenIntro Statistics	JDavid M Diez, Christopher D Barr, Mine Cetinkaya-Rundel	15	436	07-02-2015
3	An Introduction to Statistical Learning	Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani	52	426	09-01-2017

Showing 1 to 3 of 3 entries

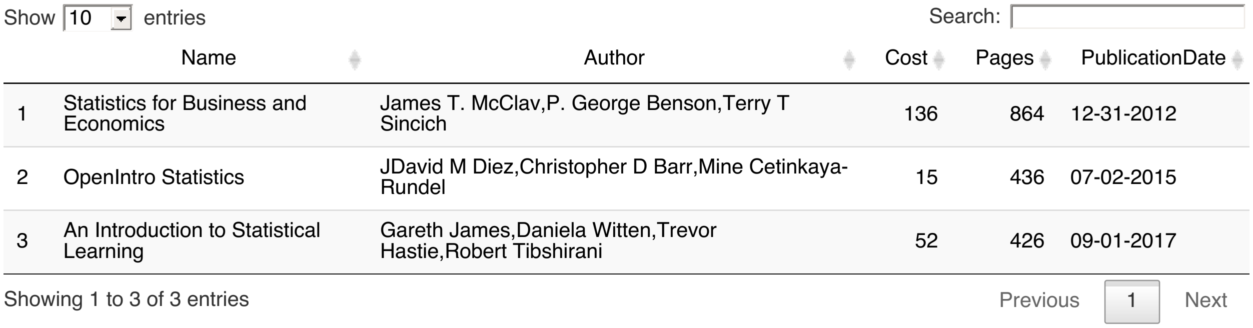
5.JSON File

5.1 Code for JSON file

```
{
"book" :[
    {
    "Name" : "Statistics for Business and Economics",
    "Author" : ["James T. McClav", "P. George Benson", "Terry T Sincich"],
    "Cost" : 136,
    "Pages" : 864,
    "PublicationDate" : "12-31-2012"
    },
    {
    "Name" : "OpenIntro Statistics",
    "Author" : ["JDavid M Diez", "Christopher D Barr", "Mine Cetinkaya-Rundel"],
    "Cost" : 15,
    "Pages" : 436,
    "PublicationDate" : "07-02-2015"
    },
    {
    "Name" : "An Introduction to Statistical Learning",
    "Author" : ["Gareth James", "Daniela Witten", "Trevor Hastie", "Robert Tibshirani"],
    "Cost" : 52,
    "Pages" : 426,
    "PublicationDate" : "09-01-2017"
    }
    ]
}
```

5.2 Read JSON file from Github repository

```
#stored url in a var
jsonURL <- "https://raw.githubusercontent.com/SubhalaxmiRout002/DATA-607-Assignment-7/master/book.json"
booksJSON <- fromJSON(jsonURL)
# load all the data
booksJSON <- booksJSON[[1]]
#view data
datatable(booksJSON)
```



	Name	Author	Cost	Pages	PublicationDate
1	Statistics for Business and Economics	James T. McClav,P. George Benson,Terry T Sincich	136	864	12-31-2012
2	OpenIntro Statistics	JDavid M Diez,Christopher D Barr,Mine Cetinkaya-Rundel	15	436	07-02-2015
3	An Introduction to Statistical Learning	Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani	52	426	09-01-2017

6.Load the information from each of the three sources into separate R data frames

```
str(booksXML)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ Name           : chr  "Statistics for Business and Economics" "OpenIntro Statistics" "An Introduc
##  $ Author         : chr  "James T. McClave, P. George Benson, Terry T Sincich" "JDavid M Diez, Chris
##  $ Cost           : chr  "136" "15" "52"
##  $ Pages          : chr  "864" "436" "426"
##  $ PublicationDate: chr  "12-31-2012" "07-02-2015" "09-01-2017"
```

```r
str(booksHTML)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ Name           : chr  "Statistics for Business and Economics" "OpenIntro Statistics" "An Introduc
##  $ Author         : chr  "James T. McClave, P. George Benson, Terry T Sincich" "JDavid M Diez, Chris
##  $ Cost           : chr  "136" "15" "52"
##  $ Pages          : chr  "864" "436" "426"
##  $ PublicationDate: chr  "12-31-2012" "07-02-2015" "09-01-2017"
```

```r
str(booksJSON)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ Name           : chr  "Statistics for Business and Economics" "OpenIntro Statistics" "An Introduc
##  $ Author         :List of 3
##   ..$ : chr  "James T. McClav" "P. George Benson" "Terry T Sincich"
##   ..$ : chr  "JDavid M Diez" "Christopher D Barr" "Mine Cetinkaya-Rundel"
##   ..$ : chr  "Gareth James" "Daniela Witten" "Trevor Hastie" "Robert Tibshirani"
##  $ Cost           : int  136 15 52
##  $ Pages          : int  864 436 426
##  $ PublicationDate: chr  "12-31-2012" "07-02-2015" "09-01-2017"
```

7.Compare all 3 data frame

```r
identical(booksXML,booksHTML)
```

```
## [1] TRUE
```

The `booksXML` and `booksXML` are identical.

```r
identical(booksXML,booksJSON)
```

```
## [1] FALSE
```

```r
identical(booksHTML,booksJSON)
```

```
## [1] FALSE
```

Data set `booksHTML` and `booksJSON` are not identical. Dataset `booksJSON` has `char` for `Name`,`list` data type for `Author`, `numeric` for `Cost`, `numeric` for `Pages` and `char` for `PublicationDate`. However, dataset `booksHTML` has `char` for `Name`,`list` data type for `Author`, `char` for `Cost`, `char` for `Pages` and `char` for `PublicationDate`.

7.1 Try to make all 3 data frame identical

```r
# datatype change for cost and pages in HTML
booksHTML$Cost <- as.integer(booksHTML$Cost)
booksHTML$Pages <- as.integer(booksHTML$Pages)

#datatype change for Author in JSON
booksJSON$Author <- as.character(booksJSON$Author)

# datatype change for cost and pages in XML
booksXML$Cost <- as.integer(booksXML$Cost)
booksXML$Pages <- as.integer(booksXML$Pages)

str(booksHTML)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ Name           : chr  "Statistics for Business and Economics" "OpenIntro Statistics" "An Introduc
##  $ Author         : chr  "James T. McClave, P. George Benson, Terry T Sincich" "JDavid M Diez, Chris
##  $ Cost           : int  136 15 52
##  $ Pages          : int  864 436 426
##  $ PublicationDate: chr  "12-31-2012" "07-02-2015" "09-01-2017"
```

```r
str(booksXML)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ Name           : chr  "Statistics for Business and Economics" "OpenIntro Statistics" "An Introduc
##  $ Author         : chr  "James T. McClave, P. George Benson, Terry T Sincich" "JDavid M Diez, Chris
##  $ Cost           : int  136 15 52
##  $ Pages          : int  864 436 426
##  $ PublicationDate: chr  "12-31-2012" "07-02-2015" "09-01-2017"
```

```r
str(booksJSON)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ Name           : chr  "Statistics for Business and Economics" "OpenIntro Statistics" "An Introduc
##  $ Author         : chr  "c(\"James T. McClav\", \"P. George Benson\", \"Terry T Sincich\")" "c(\"JDa
##  $ Cost           : int  136 15 52
##  $ Pages          : int  864 436 426
##  $ PublicationDate: chr  "12-31-2012" "07-02-2015" "09-01-2017"
```

```r
identical(booksXML, booksHTML)
```

```
## [1] TRUE
```

```r
identical(booksXML, booksJSON)
```

```
## [1] FALSE
```

```r
identical(booksHTML, booksJSON)
```

```
## [1] FALSE
```

Here dataset `booksHTML` and `booksXML` are identical but `booksJSON` is not identical.

8.Conclusion

Created all the 3 dataset file on my own. All file has a different format and indentation. But R has many functions to read and write these files. We have seen all 3 data frames display the same values for all 3 data sets, but data type is different from one dataset to another. I have tried to change datatype for all 3 data files but in JSON file due to `Author` type list, not able to get data type character. From this assignment, got to know more about different file format and how to read and write data from these, this experience will help for data scrapping.