Gathering Data from Websites Using `R` Programming language

John Karuitha

Monday, July 12, 2021

Background

Have you ever come across some valuable data on a website and wished to access and use it in your research? [1] Did you get discouraged from using data from the internet because writing down the data by hand and transferring it to a spreadsheet seemed daunting? [2] Have you ever spent a substantial amount of time trying to copy and paste data from a website with plenty of frustration and headache? If you have found yourself in any of these and similar situations, this write-up is for you. I will describe the basics of harvesting data from websites, commonly known as web scraping, using the `R` (`?`) programming language. I start with tabular data that most financial analysts, economists, and other business professionals and academics use on many occasions. In future articles, I will delve deeper into scrapping websites for text and different data types. The basics in this section should meet the needs of most users but watch out for more advanced applications of web scraping on my blog and `rpubs`.

[1] The author is at Karatina University, School of Business, P.O. Box 1957-10101, Karatina, Kenya.

[2] The R code for this project is available in my GitHub account on this link `https://github.com/Karuitha/scrapping_la_liga/blob/master/code/scraps.R`. Copy and paste the address on a new browser tab

The Target Website and Data

I have conveniently chosen the results of the premier soccer league in Spain, the `La Liga`. The data comes from two sources. The first source is wikipedia. On this site, we shall scrap data regarding the results of the Spanish La Liga from 1929 to 2021. The site gives the top 3 teams over the years together with the names of top scorers. [3]. The second source is sky news. This second source provides results of the more recent La Liga results starting 2009 to 2021. I use this site to illustrate how to scrap data that spans multiple web pages. [4].

[3] You can view this page on this address `https://en.wikipedia.org/wiki/List_of_Spanish_football_champions`

[4] Again, you can view this data on this address `https://www.skysports.com/la-liga-table/`

Objectives and Caveats

In this article, I aim to;

1. Demonstrate the basics of scrapping data tables from a website using `R` and the `rvest` package from the `tidyverse`.
2. Clean the data tables to generate an actionable set of data.

In my analysis, I assume basic knowledge of the `R` programming language and regular expressions (`Regex`). Moreover, my article aims mainly at demonstrating web scrapping and not the analysis of the