Web Scrapping Project

Alejandro Suarez Otero

5 de julio de 2018

Principales objetivos

La principal motivación de este trabajo es aplicar las técnicas de *web-scrapping* aprendidas durante este curso, con la intención de extraer información estadística sobre algunos de los jugadores más destacados de la liga estadounidense de baloncesto.

Aunque la variedad de estudios posibles es inmensa, dada la cantidad de datos que ofrece esta página web, el foco del análisis se centrará en unos sencillos gráficos descriptivos acerca de las estadísticas de tiro de las principales estrellas NBA.

En concreto, se analizará en qué posiciones del campo existe más predisposición al lanzamiento por parte del jugador y qué porcentajes alcanza en dichas posiciones. Para ello, se construirá una aplicación shiny que que interaccione con el usuario.

Fuentes de datos

Los datos van a ser extraídos de las páginas web: NBA de *wikipedia*. De la primera extraeremos las principales estadísticas de tiro, mientras que de la segunda obtendremos los datos biográficos básicos sobre los diferentes jugadores.

Tecnología de extracción

La página NBA tiene su información guardada en formato JSON y Wikipedia está construída en formato HTML por lo que se han utilizado diferentes técnicas de extracción medinate los paquete *rjson()*, *jsonlite()* y *rvest()*. Estos paquetes han sido utilizados mediante la lectura de apartados destacados de los siguientes artículos: (Couture-Beil 2013), (Ooms, Temple Lang, and Hilaiel 2014) y (Wickham 2015).

Búsqueda de fuentes de información

El proceso comenzó con la búsqueda de los enlaces que me permitiesen obtener la información principal de la página NBA. Como el navegador es el encargado de representar el contenido HTML, es sencillo utilizar sus herramientas de desarrollador para obtener una idea exacta de donde debemos buscar la información.

Para abrir estas opcones de desarrollador simplemente deben seguirse estos pasos: Developers Tools -> Network -> XHR. XHR es el acrónimo de XMLHttRequest, na vez se ha clickado en él, deberían aparecer diferentes entradas. Algunas de ellas, como se muestra en la imagen son APIs que nos devuelven los datos que buscamos en formato JSON.

En cuanto a los datos biográficos de cada jugador se utilizó la página específica de la Wikipedia y se extrajo el .vcard mediante el paquete rvest(), a continuación se muestran imágenes de los datos extraídos:

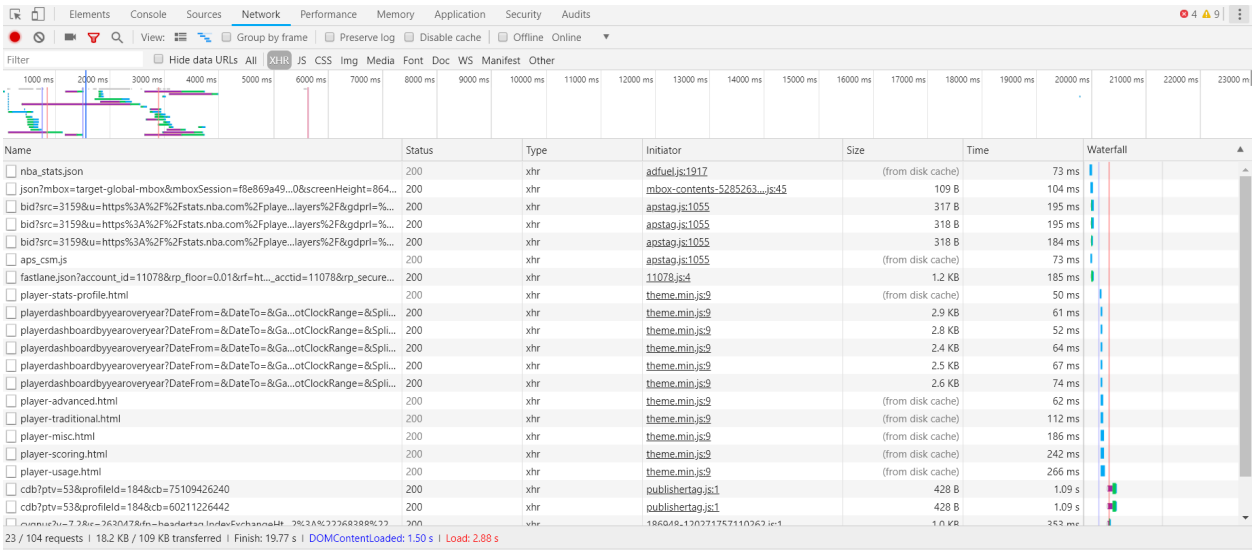


Figure 1: Developer Tools of Google Chrome

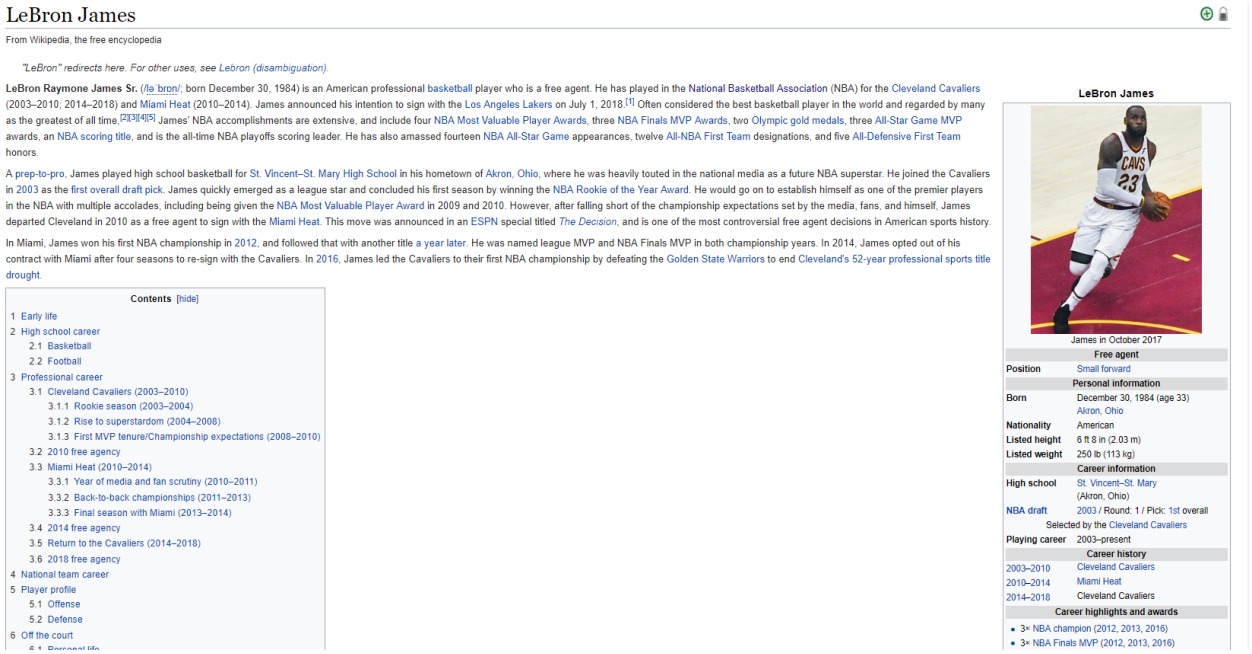


Figure 2: Wikipedia biographic information

Análisis de datos

Una vez hallados los datos relevantes he utilizado diferentes los paquetes mencionados para la extracción y procesado de la información. La primera tarea realizada consistió en localizar el ID que identifica a cada jugador dentro de la página oficial de la NBA. Los resultados obtenidos se muestran en la siguiente tabla:

Player ID	Short Name
2544	James, LeBron
201935	Harden, James
201939	Curry, Stephen
201142	Durant, Kevin

A continuación, mediante el paquete *rjson()* se obtienen los datos representados en la siguiente tabla, los cuáles muestra las localizaciones de cada tiro realizado por un determinado jugador durante toda la temporada. A modo de muestra, aquí sólo se representan los primeros diez tiros del jugador (LeBron James) durante la temporada 2014/2015.

GAME_EVENT_ID	PERIOD	MINUTES_REMAINING	SECONDS_REMAINING
4	1	11	20
33	1	6	30
53	1	4	45
77	1	2	31
82	1	1	51
136	2	9	18
202	2	3	16
217	2	1	55
227	2	1	16
299	3	6	54

EVENT_TYPE	ACTION_TYPE	SHOT_TYPE	SHOT_ZONE_BASIC
Missed Shot	Jump Shot	2PT Field Goal	Mid-Range
Made Shot	Layup Shot	2PT Field Goal	Restricted Area
Missed Shot	Fadeaway Jump Shot	2PT Field Goal	Mid-Range
Missed Shot	Jump Shot	3PT Field Goal	Right Corner 3
Missed Shot	Jump Shot	3PT Field Goal	Above the Break 3
Missed Shot	Jump Bank Shot	2PT Field Goal	In The Paint (Non-RA)
Missed Shot	Jump Shot	3PT Field Goal	Above the Break 3
Missed Shot	Reverse Layup Shot	2PT Field Goal	Restricted Area
Missed Shot	Turnaround Jump Shot	2PT Field Goal	Mid-Range
Made Shot	Jump Shot	3PT Field Goal	Above the Break 3

SHOT_ZONE_AREA	SHOT_ZONE_RANGE	SHOT_DISTANCE	LOC_X
Right Side Center(RC)	16-24 ft.	18	114
Center(C)	Less Than 8 ft.	0	-7
Left Side(L)	8-16 ft.	12	-105
Right Side(R)	24+ ft.	22	227
Right Side Center(RC)	24+ ft.	26	91
Right Side(R)	8-16 ft.	9	70

SHOT_ZONE_AREA	SHOT_ZONE_RANGE	SHOT_DISTANCE	LOC_X
Right Side Center(RC)	24+ ft.	26	122
Center(C)	Less Than 8 ft.	0	-8
Right Side(R)	8-16 ft.	13	135
Center(C)	24+ ft.	25	26

Además de estos datos se extraen también los promedios de cada jugador para 4 temporadas diferentes. Continuando con el ejemplo del mismo jugador, se muestra a continuación la tabla obtenida:

GROUP_VALUE	GP	W	L	REB	AST	TOV	STL	BLK	PFD	PLUS_MINUS
2014-15	69	50	19	6	7.4	3.9	1.6	0.7	6	7.8
2015-16	76	56	20	7.4	6.8	3.3	1.4	0.6	5.4	8.1
2016-17	74	51	23	8.6	8.7	4.1	1.2	0.6	5.9	6.5
2017-18	82	50	32	8.6	9.1	4.2	1.4	0.9	5.4	1.3

Por último, se extrae mediante el paquete *rvest()* la información biográfica de los jugadores. La siguiente tabla ilustra la información que se incorpora en la aplicación de Shiny:

	FIELD	PLAYER
3	Position	Small forward
5	Born	(1984-12-30) December 30, 1984 (age 33)Akron, Ohio
6	Nationality	American
7	Listed height	6 ft 8 in (2.03 m)
8	Listed weight	250 lb (113 kg)
10	High school	St. Vincent St. Mary
(Akro	n, Ohio)	
11	NBA draft	2003 / Round: 1 / Pick: 1st overall

A la par, se extraen también las fotografías de los deportistas, las cuáles se guardan en una carpeta para ser utilizadas en la aplicación.

En definitiva, y como se puede observar en el repositorio *alexsuarez94/School_project* de **GitHub**, los scrits utilizados para extraer y guardar estos datos están guardados en la carpeta data. Su estructura básica es la siguiente:

1. Para la obtención de la información que conforma los gráficos:

- player_id.R
- nba_shot_stats.R

2. Para la obtención de las estadísticas de la tabla:

- player_id.R
- lebron_stats.R
- durant_stats.R
- harden_stats.R
- curry_stats.R

3. Para la obtención de las imágenes:

- player_id.R
- save_phtos.R

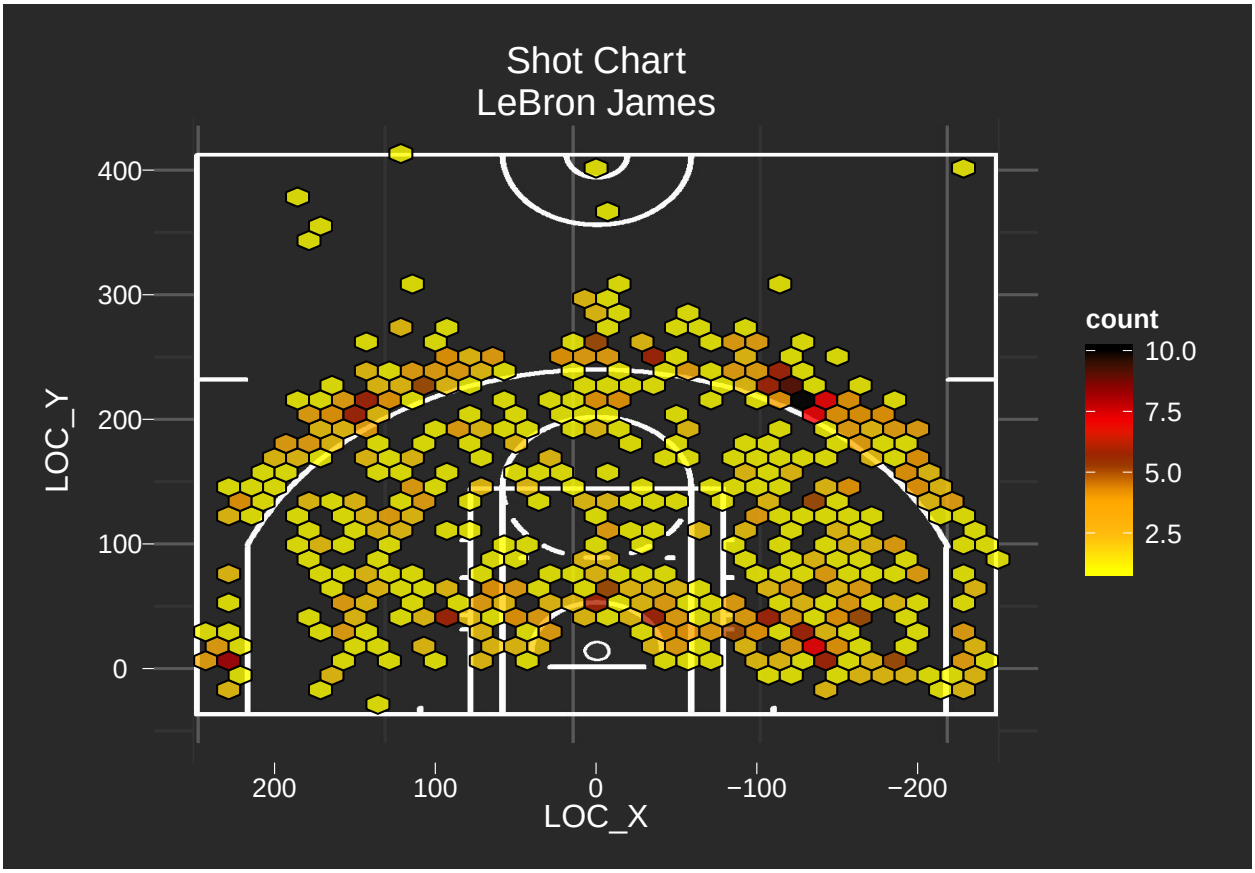
4. Para la información biográfica básica:

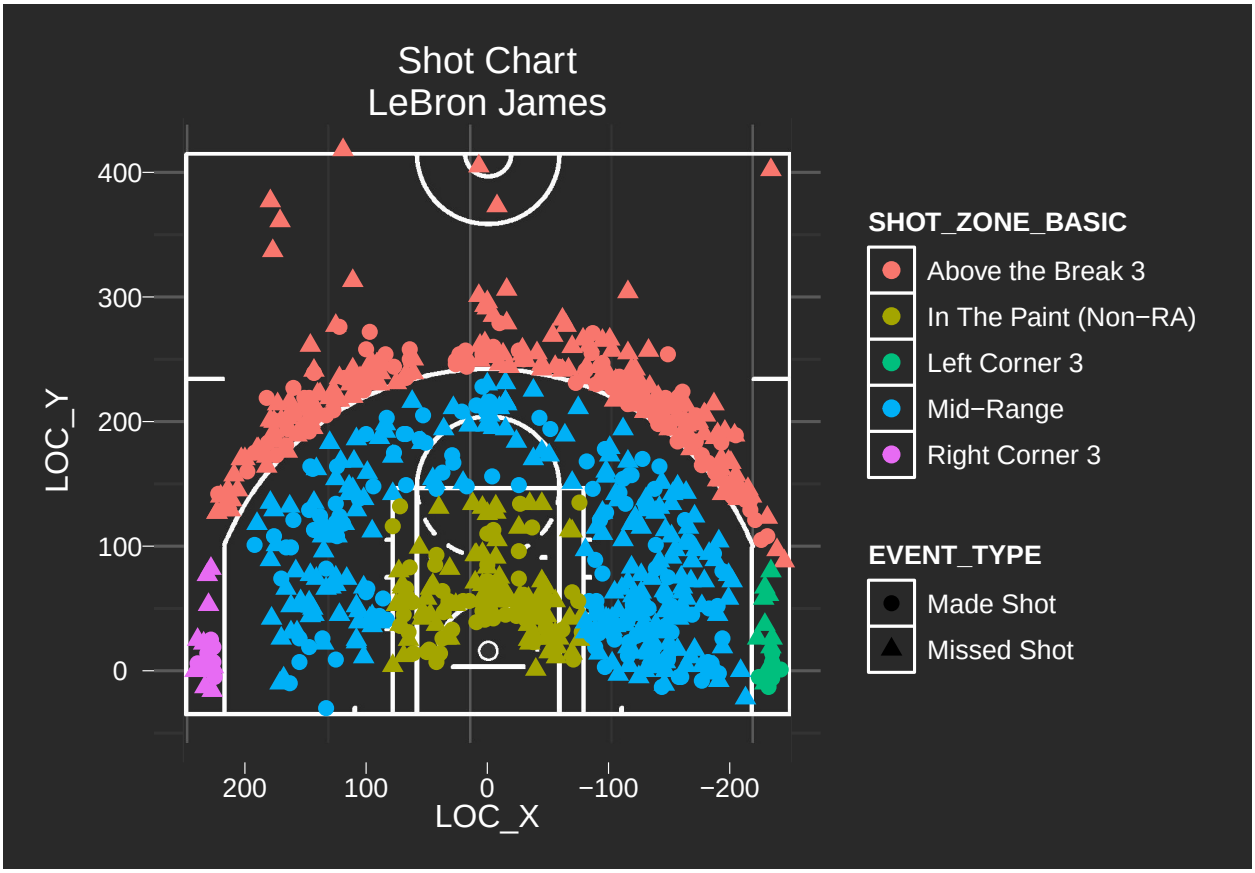
- wiki_bio.R

No realizaré aquí ninguna descripción detallada de estos archivos ya que cada *script* presenta los comentarios oportunos acerca ddel proceso llevado a cabo.

Resultado final

A modo de muestra expondré aquí dos de los gráficos obtenidos, de nuevo para el jugador LeBron James. Para observar el resto de gráficos es necesario acceder a la aplicación mediante *RStudio* y la sencilla instrucción que aparece debajo de estos gráficos.





Con esta sencilla instrucción se puede acceder a la aplicación de shiny, ubicada también en el respositorio bajo en nombre de archivo *app.R* (es importanter mencionar que la aplicación puede tardar unso segundos para cargarse al inicio, aproximadamente unos 30 segundos en el peor de los casos).

```
shiny::runGitHub( "alexsuarez94/School_projects")
```

A continuación muestro una captura tomada directamente de la app en funcionamiento:

Figure 3: Shiny App

Conclusiones

En términos generales estoy satisfecho con el trabajo realizado aunque soy consicente de lo mejorable que es el código empleado.

Debsdo a las restricciones temporales algunos aspectos a mejorar han quedado pendientes:

- Sólo he podido construir una app que refleje las estadísticas de 4 jugadores debido al coste temporal que tiene extraer los datos de la página web. Me gustaría averiguar formas más eficientes de obtener estos datos.
- No he tenido tiempo tampoco para automatizar debidamente el código y, en ocasiones, este está construido de forma muy manual.
- La extracción de HTML ha sido complicada a pesar de querer obtener únicamente una parte muy pequeña de toda la información disponible.
- Por cuestiones de fallo en la conexión con el servidor y de tiempo de espera, he tenido que guardar varios objetos obtenidos de la web para interactuar con ellos de forma local desde la aplicación de shiny.
- He intenado publicar la aplicación mediante la página web Shiny Apps, sin embargo, el coste computacional requerido para la extracción de los datos ralentiza demasiado la aplicación.

Bibliografía

Couture-Beil, Alex. 2013. "Rjson: JSON for R." *R Package Version 0.2* 13.

Ooms, Jeroen, D Temple Lang, and Lloyd Hilaiel. 2014. "Jsonlite: A Robust, High Performance Json Parser and Generator for R." *R Package Version 0.9* 13.

Wickham, Hadley. 2015. "Rvest: Easily Harvest (Scrape) Web Pages." *R Package Version 0.3* 1.