Alumnos:
Albert Breva Benito
Juan Matilla Varas

Práctica 1.

1. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

En este proyecto se ha decidido investigar y trabajar con datos estadísticos de la NBA. Como ya sabemos, los datos han ido cobrando cada vez más importancia en el deporte de élite y la NBA es una de las ligas profesionales que actualmente más partido saca de este recurso, con Daryl Morey como principal valedor y el primer General Manager de la liga que empezó a tomar decisiones basadas en datos. De hecho, él mismo reconocía en una entrevista que los deportes habían adoptado tarde el uso de la analítica avanzada.

"Sports are really a late adopter of using data to drive decision making. If you look at Wall Street, or you look at consumer or credit card companies, or you look at Procter & Gamble, all of these are actually quite a bit ahead in terms of using data to drive their decisions. Sports are late to the party."

Entrevista: https://insight.kellogg.northwestern.edu/article/using-data-to-call-the-shots

Teniendo en cuenta que en la NBA se juegan 1230 partidos solo en temporada regular más los partidos de Playoffs es fácil imaginar la gran cantidad de datos que se pueden extraer solamente de una temporada.

En este proyecto, se ha decidido investigar sobre la página web de estadísticas de la NBA Basketball Reference (https://www.basketball-reference.com). Dicha web recopila datos sobre multitud de aspectos de la liga, como por ejemplo: líderes en anotación, rebotes, asistencias, resultados por partido, etcétera.
Además, tiene disponible la información disponible desde la temporada 1949-1950, que fue la primera temporada de la NBA así como de tres temporadas previas, de 1946 a 1949, cuando todavía no existía la NBA y se llamaba BAA (Basketball Association of America).
Por estos motivos, nos pareció la opción idónea para llevar a cabo el web scraping y poder llevar a cabo nuestro proyecto.

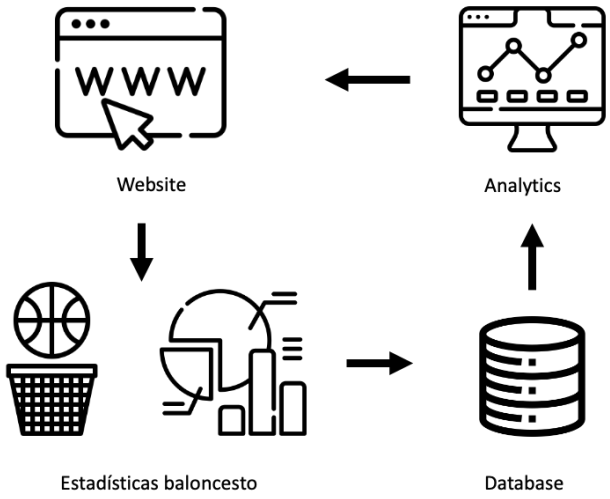
2. Título. Definir un título que sea descriptivo para el dataset.

Estadísticas de la carrera de cada jugador de la NBA.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Tal y como expresa el título, el dataset se compone de las estadísticas totales para diferentes aspectos del juego para cada jugador que ha pasado por la NBA.
Los datos han pasado un mínimo proceso de limpieza, por lo que es posible que aún puedan existir inconsistencias y que el formato no sea el más idóneo para su análisis.
La descripción de cada uno de los campos que componen el dataset se presenta en las siguientes preguntas.
Finalmente, el formato de exportación del dataset es un fichero CSV que facilita su visualización y tratamiento.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



5. Contenido. Explicar los campos que incluye el dataset, el período de tiempo de los datos y cómo se ha recogido.

Como ya se ha comentado, en el dataset presentado tenemos las estadísticas de cada jugador en cada uno de los aspectos del juego. Los campos son los siguientes:

- G: partidos jugados.
- GS: partidos jugados como titular.
- MP: minutos jugados por partido.
- FG: canastas por partido.
- FGA: lanzamientos por partido.
- FG%: porcentaje de canastas por partido.
- 3P: canastas de tres puntos.
- 3PA: lanzamientos de tres puntos.
- 3P%: porcentaje de canastas de tres puntos.
- 2P: canastas de dos puntos.

- 2PA: lanzamientos de dos puntos
- 2P%: porcentaje de canastas de dos puntos.
- EFG%: porcentaje de canasta efectivo. Esta estadística se ajusta al hecho de un tiro de campo de 3 puntos vale más que un tiro de campo de 2 puntos.
- FT: tiros libres anotados.
- FTA: lanzamientos de tiro libre.
- FT%: porcentaje de canastas que son tiros libres.
- ORB: rebotes ofensivos por partido.
- DRB: rebotes defensivos por partido.
- TRB: total de rebotes por partido.
- AST: asistencias por partido.
- STL: robos por partido.
- BLK: tapones por partido.
- TOV: pérdidas de balón por partido.
- PF: faltas personales.
- PTS: puntos por partido.

Todos estos datos fueron recogidos a través de web scraping en lenguaje Python sobre la tabla de estadísticas de cada jugador.
El hecho de que la página web ya presente los datos en formato tabla ha facilitado la extracción de los mismos.
Finalmente, se guardan los datos extraídos en un fichero CSV.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

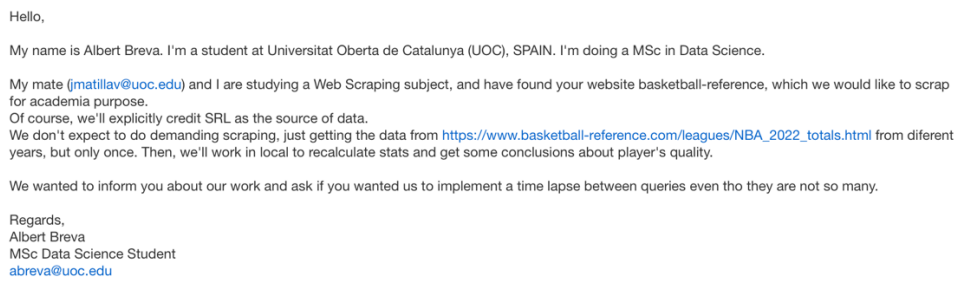
El propietario del sitio elegido es Sports Reference LLC

https://www.sports-reference.com/termsofuse.html

permitted uses of this Site and its Content. Our guiding principles are that (1) sharing, using, modifying, repackaging, or publishing data found on individual SRL webpages is welcomed, whether for commercial or non-commercial purposes, but (2) any such sharing, use, modification, repackaging, or publication should explicitly credit SRL as the source of the data to the maximum extent possible and (3) any such sharing, use, modification, repackaging, or publication must not violate any express restrictions set forth in this Section 5, especially the restrictions set forth in subparts 5(i) and 5(j) below.

i. without our express written permission, use any automated means to access or use the Site, including scripts, bots, scrapers, data miners, or similar software, in a manner that adversely impacts site performance or access; or
j. **use any material or Content from the Site, including without limitation any statistics or data, (i) to create any database, archive, or other data store that competes with or constitutes a material substitute for the services or data stores offered on the Site or by the Site's Data Providers or (ii) to provide any service that competes with or constitutes a material substitute for the services or data stores offered on the Site or by the Site's Data Providers; or**
k. attempt to or actually disrupt, impair, or interfere with the Site, or any information, data, or materials posted and/or displayed by SRL; or

Como vemos en las capturas adjuntas el propietario expresa en su página web la petición de que no se realice web scraping a su sitio, así que se ha contactado con el mismo y estamos pendientes de respuesta. Adjuntamos captura de pantalla de la consulta.



Hello,

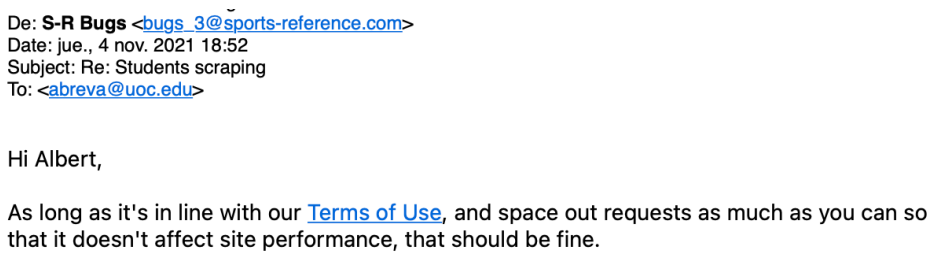
My name is Albert Breva. I'm a student at Universitat Oberta de Catalunya (UOC), SPAIN. I'm doing a MSc in Data Science.

My mate (jmatillav@uoc.edu) and I are studying a Web Scraping subject, and have found your website basketball-reference, which we would like to scrap for academia purpose.
Of course, we'll explicitly credit SRL as the source of data.
We don't expect to do demanding scraping, just getting the data from https://www.basketball-reference.com/leagues/NBA_2022_totals.html from diferent years, but only once. Then, we'll work in local to recalculate stats and get some conclusions about player's quality.

We wanted to inform you about our work and ask if you wanted us to implement a time lapse between queries even tho they are not so many.

Regards,
Albert Breva
MSc Data Science Student
abreva@uoc.edu

Recibimos una respuesta afirmativa del propietario del sitio, siempre y cuando no perjudicásemos el funcionamiento de la página y le acreditsemos como propietario original de los datos.



De: **S-R Bugs** <bugs_3@sports-reference.com>
Date: jue., 4 nov. 2021 18:52
Subject: Re: Students scraping
To: <abreva@uoc.edu>

Hi Albert,

As long as it's in line with our Terms of Use, and space out requests as much as you can so that it doesn't affect site performance, that should be fine.

Algunos análisis similares encontrados en internet son:

- https://sebasdarius.medium.com/part-1-advanced-box-score-6ce33e638fce
- https://towardsdatascience.com/part-2-shot-quality-5ab27fd63f5e
- https://towardsdatascience.com/predicting-2020-21-nbas-most-valuable-player-using-machine-learning-24aaa869a740

Los dos primeros son análisis descriptivos con datos muy similares a los nuestros y obtenidos también de Basketball Reference.
El tercer link utiliza un conjunto de datos diferente (https://github.com/dribbleanalytics/ml-mvp-predict/blob/master/2018-19-season/mid-season/final-csv-data/historical-mvps.csv) e intenta predecir el MVP de la temporada 2020-2021, algo que efectivamente consigue. Las métricas que utiliza también existen en nuestro dataset así que podría ser interesante tomarlo como ejemplo para intentar predecir el MVP de la temporada actual.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El interés en analizar este conjunto de datos se debe como ya se ha mencionado al inmenso potencial que tiene la analítica avanzada en el mundo del deporte.
El objetivo es poder detectar qué jugadores cuentan con la confianza de su entrenador en base a titularidades, cuales han ido progresando, que jugadores pueden sufrir un estancamiento, etcétera.

Un uso práctico de este análisis permitiría detectar qué jugadores están sobrevalorados o infravalorados por sus entrenadores. Se asume que los mejores jugadores son los que juegan más minutos por partido y, tienen mejores "números" (Rebotes por partido o por minuto jugado, puntos por partido o minuto jugado, etc.). Por tanto, podríamos detectar jugadores que participan muchos minutos con valoraciones mediocres o, por el contrario, jugadores que se les conceden pocos minutos por partido pero que los saben aprovechar y consiguen objetivos.

8. Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección.

Una licencia que encajaría bien con este conjunto de datos puede ser CC BY-SA 4.0 License debido a los siguientes motivos:

- Se provee el nombre de la compañía que ha creado los conjuntos de datos sobre los que se realiza web scraping, por lo que se reconoce el trabajo de terceros.
- En los términos y condiciones de la web se específica que se permite el uso comercial o no comercial de los datos siempre que se acredite a SRL como la fuente originaria de los datos.

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	AB, JM
Redacción de las respuestas	AB, JM
Desarrollo del código	AB, JM