TRABAJO FIN DE MÁSTER

Implementación del paquete de R BasketballAnalyzeR.

Máster en Estadística Aplicada.

Presentado por
Dª.:
Ana Tetuán Blanco

Escuela Internacional de Posgrado

Granada, junio de 2022.

TRABAJO FIN DE MÁSTER

Estadística Aplicada

Implementación del paquete de R BasketballAnalyzeR.

Máster en Estadística Aplicada
Escuela de Posgrado.
Autora: Ana Tetuán Blanco.
Responsable de tutorización: Yolanda Román.
Curso 2021/2022.

Agradecimientos

Quería agradecer a todas esas personas que me han introducido, de una forma u otra, al mundo de la Ciencia de Datos del Deporte:

En primer lugar, a Carla López Lares, compañera de equipo y también matemática, que me dio el empujón que necesitaba para encaminarme hacia esta 'aventura'. En segundo lugar, a Paco Ocaña Peinado, por interesarse por mi línea de TFM y por aportarme algunas ideas. En tercer lugar, a David García Sánchez, por mostrarme la labor del perfil de analista deportivo y por ofrecerme la base de datos oficial del Manuela Fundación Raca, crucial para mi estudio. Por último, y no por ello menos importante, a Víctor Vicente Palacios, por su pasión por el Baloncesto y la Estadística y su predisposición para darme ideas, interesarse por mi trabajo y ayudarme con él.

También, agradecer a la gente de mi entorno que me han ayudado a superar los baches de la vida y a no tirar la toalla.

Índice

Índice de figuras

1. Resumen.

En la sociedad actual el deporte llena el tiempo de ocio de muchas personas. Tanto en su vertiente práctica, ejercitándose de manera individual o colectiva, como en su faceta de exhibición, que genera gran cantidad de espectadores que se limitan a observar partidos o competiciones en los que participan otras personas.

A lo largo del siglo XX el deporte fue cobrando cada vez mayor importancia, aumentando la atención suscitada al mismo tiempo que su impacto económico, lo que derivó en un incremento del número de personas que dependían directa o indirectamente de la actividad deportiva. Y en lo que llevamos de siglo XXI, esa tendencia no ha disminuido, si acaso se ha acrecentado [1].

Este trabajo pretende unir estos dos mundos, en apariencia alejados. En realidad, nace por el gusto de la autora por ambas materias; soy aficionada al Deporte, por placer y como distracción, y a las Matemáticas y Estadística por interés y formación.

En particular, el objetivo del Trabajo de Fin de Máster propuesto consiste en unificar el campo del Baloncesto y la Ciencia de Datos a través del uso y análisis de la librería de R de *BasketballAnalyzeR*.

Al principio del mismo, se introducirá un capítulo relacionado con la Ciencia de Datos, haciendo hincapié en algunos aspectos clave e introduciendo dicha Ciencia al ámbito deportivo y, en concreto, al del baloncesto. En esta parte, se mencionarán numerosos artículos que se han publicado en relación al campo del Baloncesto y la Estadística.

A continuación, se hará una descripción del lenguaje que se va a utilizar para el análisis, y se va a introducir la librería de R de *BasketballAnalyzeR*, mencionando las bases de datos que posee y las funcionalidades de la misma. Dentro de estas últimas, se introducirán definiciones relacionadas con el ámbito del baloncesto, los distintos tipos de gráficos que se pueden crear utilizando la librería y las técnicas que se van a aplicar para realizar el análisis de los datos.

Especialmente, se le dedicarán unas páginas al marco teórico del Análisis de Clúster, un conjunto de técnicas multivariantes que se utilizan para clasificar datos en grupos distintos entre sí.

Finalmente, todo el marco teórico expuesto se trasladará a la práctica, aplicando las funciones de la librería al contexto de la Liga Femenina Cha-

llenge, la segunda máxima categoría del baloncesto femenino español. No está de menos mencionar que las funciones serán aplicadas con la intención de obtener conclusiones reales y llegar a un análisis más concluyente que aquél que se dedica única y exclusivamente a aplicar las funciones de la librería.

Por último, se comentarán algunas dificultades encontradas a lo largo de la realización del trabajo, y se deja una puerta abierta hacia una continuación del análisis de datos en el baloncesto.

2. La Ciencia de Datos.

La Ciencia de Datos es uno de los campos de investigación actuales que más auge están teniendo. Ahora que la tecnología moderna ha permitido la creación y el almacenamiento de cantidades cada vez mayores de información, el objetivo de generar información relevante a partir de ellos es lo que persiguen los científicos de datos.

Sin embargo, pese a que este término y otros como el de Big Data, Análisis de Datos, Inteligencia Artificial, Machine Learning, etc. están en boca de todos, en general se tratan de términos cuyo alcance no es realmente conocido por la sociedad. Aún hoy en día existe un cierno nivel de desconfianza en los algoritmos [2] que afecta a cualquier ámbito en el que se aplique esta Ciencia. Algunos expertos llegaron incluso a ser marginados sin ningún tipo de pretexto, por el simple hecho de aplicar estas técnicas. Afortunadamente, el reconocimiento de las técnicas empleadas en la Ciencia de Datos es una realidad hoy en día, y su necesidad en los distintos ámbitos obliga al desarrollo de nuevos procesimientos de análisis de la información. La Ciencia de Datos, entendida como herramienta interdisciplinar, presenta en la actualidad un gran auge en el avance científico.

Como ya hemos indicado, la Ciencia de Datos aparece junto a la capacidad de recoger información. De repente, somos capaces de acceder a grandes volúmenes de datos en tiempo real y en prácticamente todos los campos de la Ciencia. Esta 'avalancha' de información que se presenta al investigador de manera 'caótica' en muchas ocasiones, requiere de procedimentos específicos que permitan, en primer lugar, 'ordenar' los datos y proceder a continuación a su análisis obteniéndo múltiples respuestas.

La Ciencia de Datos se construye entonces como una materia interdisciplicinar que requiere de la Programación, de la Estadística, de las Matemáticas y, por supuesto, del campo específico de estudio del que provengan los datos.

Podemos resumir en los siguientes puntos algunos de los aspectos claves de la Ciencia de Datos [3]:

1. La Ciencia de Datos aspira a extraer conocimiento a partir de datos o información.

2. Interpretar resultados a partir de bases de datos es una fase extremadamente delicada. Los resultados de cualquier análisis estadístico dependen de las suposiciones que se hacen y de los datos que se tienen a disposición, así que no se pretende generalizar su significado fuera de esos límites y atribuir conclusiones que no se pueden acreditar.

3. Los resultados que muestra el análisis de datos son más estables, robustos y fiables cuanto mayor sea la base de datos de la que se parta.

4. La Ciencia de Datos se puede aplicar en cualquier ámbito (medicina, economía, finanzas, deporte, ...), así que deben existir equipos multidisciplinares de expertos, tanto del ámbito en el que se trabaje como de analistas de datos.

5. Potencialmente, no hay ninguna pregunta que la Ciencia de Datos no pueda resolver, siempre y cuando se tengan los datos adecuados. Es por ello que la base de datos de la que se disponga es crucial para el estudio. Sin embargo, a veces no existen esos datos, son difíciles o imposibles de conseguir, o incluso el proceso de limpieza de los datos consume demasiado tiempo. Por eso, la Ciencia de Datos nunca será capaz de describir todo acerca del tema analizado.

6. La Ciencia de Datos no es una bola de cristal, no aporta una certeza absoluta sino escenarios probables e indicaciones estructurales a medio/largo plazo.

7. La Ciencia de Datos no aporta decisiones, sino que es una herramienta que ayuda a la toma de éstas.

3. La Ciencia de Datos en el Deporte.

La Ciencia de Datos aplicada al deporte está creciendo muy deprisa, tal y como demuestran la gran cantidad de libros publicados recientemente al respecto [4]. Un acercamiento al análisis deportivo consiste en entender todo aquello que rodea un deporte en concreto: la industria, los negocios, lo que ocurre en el campo de juego, el desarrollo de los entrenamientos, etc.

A lo largo de la historia, la evolución de las técnicas de entrenamiento ha dependido en gran medida de la experiencia e intuición del cuerpo técnico. Sin embargo, hace poco menos de dos décadas, el análisis de datos empezó a tomar presencia en los deportes con el fin de ayudar a la toma de decisiones y de proporcionar ventajas para poder competir a un nivel aún más alto del que se hacía antes.

Es difícil determinar exactamente el punto de partida de la relación entre datos y deportes, pero el libro, y posterior película "Moneyball" [5], dieron a conocer la historia -basada en hechos reales- del equipo de béisbol estadounidense Oakland Athletics: El mánager general del club construyó una plantilla a partir de un sistema que evaluaba a jugadores poco llamativos pero que, según los datos analizados, se convertirían en piezas clave para la

formación de un equipo que no tenía mucho presupuesto para gastar en fichajes. Finalmente el equipo, conseguido de una forma pionera y muy criticada en aquel entonces, consiguió ganar 20 partidos consecutivos, estableciendo el récord de la Liga Americana. Y no solo eso; esta historia revolucionó la era moderna de béisbol y ahora cada equipo profesional en los Estados Unidos posee algún experto en el análisis deportivo; incluso algunos tienen departamentos dedicados exclusivamente a la colección, estudio e interpretación de datos con el fin de conseguir una ventaja deportiva respecto a otros equipos.

Por otra parte, la llegada del análisis deportivo a Europa fue algo más tardía que en Estados Unidos. En España, la incorporación de perfiles analíticos en los equipos deportivos llegó en torno a 2017. Afortunadamente, a día de hoy clubes y deportistas de todo el mundo, conscientes del análisis que se puede realizar a partir de sus datos de rendimiento, biométricos y todos aquellos que ayudan a la mejora, incorporan profesionales especializados en esta disciplina.

3.1. Análisis de datos en el baloncesto.

Los deportes profesionales no conllevan solo la evolución en las técnicas de juego y de entrenamiento. A día de hoy, el deporte profesional constituye un importante mercado que mueve ingentes cantidades de dinero a nivel internacional. Este hecho conlleva la inversión en tecnologías específicas para incrementar la posibilidad de ganar, de seleccionar los mejores jugadores, de reestructurar los equipos y perfeccionar movimientos específicos. Es aquí donde la Ciencia de Datos entra de lleno en todas las disciplinas deportivas y, en particular, en el Baloncesto, que es el objetivo central de nuestro trabajo.

Existe una amplia bibliografía en este campo. A modo de ejemplo, podemos citar el trabajo de Dean Oliver [6], que explica estrategias generales que han de seguir los equipos cuando están ganando o perdiendo, además de los aspectos en los que han de enfocarse en cada una de las situaciones. Asimismo, el autor estudia la interacción entre jugadores y cómo conseguir que estén en la mejor forma posible, cuantifica el valor de los distintos jugadores cuando juegan juntos, examina los equipos de la NBA -liga de baloncesto norteamericana- que fueron tan exitosos e identifica la razón de dicho éxito.

De hecho, en el caso de la NBA, incluso se ha transformado el sistema tradicional de juego: Mientras antes se apostaba por jugadores más altos que protegieran la canasta y marcaran puntos fácilmente, ahora la tendencia se basa en jugadores más bajos y versátiles porque los datos han demostrado que es más productivo anotar triples que canastas de dos puntos, aunque el propio aro no esté tan defendido [7]. En la figura 1 se puede observar la

evolución del mapa de tiro en la NBA.

Figura 1: Gráfico extraido del libro Sprawball, escrito por Kirk Goldsberry

Otros estudios realizados en relación al baloncesto consisten en predecir los resultados de un partido o de un torneo [8] [9], determinar factores discriminantes entre equipos exitosos y sin éxito [10] [11], examinar las propiedades estadísticas y los patrones de anotación durante los partidos [12], analizar la actuación de un jugador y el impacto en las posibilidades de victorias del equipo [13] con el foco en el impacto de dicho jugador en situaciones de mucha presión [14], descubrir patrones relacionados con las posiciones en las que juegan los jugadores [11], diseñar la mecánica de tiro para cada jugador en función de su eficacia en los tiros [15], representar los movimientos de un jugador y la red de acciones de pase a través del análisis de redes [16] [17], estudiar las jugadas y las tácticas de un rival para identificar las estrategias más adecuadas para conseguir la victoria [18], etc.

A pesar de este panorama tan amplio acerca del análisis deportivo del baloncesto, el conjunto de preguntas que podremos responder será cada vez más grande gracias a la disponibilidad de bases de datos cada vez más amplias y al desarrollo tecnológico.

En el caso de este trabajo, se intentará seguir la línea de estas investigaciones con la ayuda del software R .

4. El lenguaje de programación R.

R es un lenguaje libre y entorno de computación de estadística que se utiliza en numerosos campos de investigación científica, siendo además una herramienta potente para aplicar técnicas estadísticas -tanto básicas como avanzadas- tales como el aprendizaje automático, la minería de datos, el análisis de series temporales, etc. Además, el hecho de ser un código abierto hace que sea un lenguaje altamente extensible para toda la comunidad.

Una de las ventajas más llamativas de R es la facilidad con la que se pueden producir gráficos bien diseñados. Por otra parte, R está formado por un amplio conjunto de herramientas de software para manipular datos, hacer cálculos y visualizar gráficos. Éste incluye un eficiente manipulador y almacén de datos, un conjunto de operadores para el cálculo con matrices y vectores, una colección de herramientas para el análisis de datos y la muestra en pantalla de gráficos. Asimismo, integra un lenguaje de programación bien desarrollado, sencillo y efectivo que incluye funciones recursivas, bucles, condicionales y habilidades para la introducción y salida de datos.

Otra de las particularidades de R es que permite a los usuarios añadir funcionalidades a través de la creación de nuevas funciones. Además, en tareas de computación, es posible vincular códigos de C, C++ o Fortran en R, lo que le convierte en un potente entorno de trabajo.

Cabe destacar que R se puede ampliar fácilmente mediante los llamados "packages" (libros). De hecho, en la actualidad, el repositorio de R (CRAN) cuenta con casi 17000 libros disponibles. Los libros son unidades de código R reproducible que incluyen funciones reutilizables, además de la documentación que describe cómo se utilizan, ejemplos de uso y conjuntos de datos. La información básica sobre un libro se puede encontrar en el *Description file*, donde se proporciona información sobre lo que hace el libro, su autor, versión, fecha, tipo de licencias que usa y dependencias del libro. Este documento puede encontrarse tanto en la web oficial de R como desde dentro del entorno en sí [19].

Todos los libros se encuentran en un repositorio, y pueden ser instalados desde ahí. Normalmente, los repositorios se encuentran en línea y disponibles para todos los usuarios, aunque algunas organizaciones pueden tener su propio repositorio. Los más conocidos para R son CRAN y Github. El oficial es el primero de ellos, pues es mantenido por la comunidad R en todo el mundo y, antes de publicarse cualquier libro, éste pasa múltiples revisiones para garantizar que cumple con las políticas de CRAN [20].

La utilización de libros de R hace que el trabajo del usuario sea más

eficiente y sencillo al mismo tiempo, ya que proporciona herramientas muy útiles para el estudio del tema que se quiera tratar. Además, entre la gran cantidad de libros disponibles se pueden encontrar algunos que se adapten a diferentes áreas del conocimiento y tengan múltiples enfoques.

En particular, el libro al que nos referiremos a lo largo de este documento se encuentra en el repositorio CRAN. Para instalarlo, se puede utilizar el comando *install.packages("BasketballAnalyzeR")* dentro de R, o descargarlo directamente desde la web del repositorio, dentro de la página oficial de R.

5. BasketballAnalyzeR. La Ciencia de los Datos para el Baloncesto con R.

El libro *BasketballAnalyzeR*, disponible tanto en CRAN como en Git, es un libro de código R abierto a la comunidad, creado en 2020 por Marco Sandri, Paola Zuccolotto y Marica Manisera para acompañar a su libro *Basketball Data Science. Applications with R* [3].

En él, se exponen métodos de Estadística y Minería de Datos, empezando por un análisis descriptivo de los datos y siguiendo con técnicas multivariantes como el clúster, la correlación, el análisis de redes o la modelización de las posibles relaciones existentes entre los datos. Además, el libro unifica estás técnicas de análisis estadístico con procedimientos gráficos que facilitan la visión de los datos. De esta manera, la acción de la persona experta en Ciencia de Datos puede apoyarse en dichos gráficos con el fin de trasladar la información de una manera más entendible para la deportista o el cuerpo técnico.

Además, en la web [21] se puede encontrar el código utilizado, novedades sobre el libro, posibles actualizaciones, discusiones sobre la preparación de los datos e información de contacto para preguntar a los desarrolladores alguna cuestión relacionada con el uso del libro de R.

5.1. Bases de datos en *BasketballAnalyzeR*.

Como la gran mayoría de los libros del repositorio R, *BasketballAnalyzeR* dispone de una serie de bases de datos que permiten la aplicación de las técnicas y procedimientos en él descritos. Concretamente, estas bases de datos corresponden a información recogida de partidos jugados en la NBA durante la temporada 2017-2018. De hecho, son cuatro data frames estructurados de la siguiente forma:

- **OBox**: Box-score en el que las filas son los equipos y las columnas muestran variables referidas a los logros de los oponentes en la NBA.
- **TBox**: Box-score en la que las filas son los equipos y las columnas muestran variables referidas a los logros del equipo en los disferentes partidos de la NBA.
- **PBox**: Box-score en el que las filas son los jugadores de la liga y las columnas muestran las variables referidas a los logros individuales durante toda la temporada.
- **Tadd**: Data frame donde las filas son los equipos analizados y las columnas son variables cualitativas.
- **PbP**: Play-by-play. Las filas son los eventos ocurridos en los partidos analizados y las columnas son las descripciones de esos eventos, tales como jugador, tiempo para acabar el partido, marcador, posición en el campo, etc.

Para aclarar mejor la disposición de los datos, en el glosario situado al final del trabajo se exponen las definiciones de box-score y play-by-play.

5.2. Funciones para el análisis de los datos.

En este apartado se van a explicar algunas de las herramientas, cálculos y definiciones básicas del análisis en baloncesto, así como las funciones disponibles del libro *BasketballAnalyzeR* que se utilizarán en el caso práctico.

5.2.1. Posesión, ritmo de juego, ataque, defensa y Four Factors.

En el artículo de Kubatko et al. (2007) [22] se detallan por primera vez estos elementos, que a día de hoy son generalmente aceptado y establecen un punto de partida en la ciencia de datos del baloncesto. En detalle, abordamos los conceptos de posesión y ritmo, así como la valoración defensiva y ofensiva y el término "Four Factors".

Siguiendo la línea de Kubatko, las posesiones y el ritmo de juego se calculan de la siguiente forma:

$$POSS = (P2A + P3A) + 0,44FTA - OREB + TOV$$

$$PACE = 5 \cdot \frac{POSS}{MIN}$$

Por otra parte, la eficacia en el ataque y la defensa se mide como los puntos encestados, o recibidos, por cada 100 posesiones. A estas dos nuevas definiciones se les denotará como $ORtg$ y $DRtg$, y se calculan como:

$$ORtg = \frac{PTS_T}{POSS_T}$$
$$DRtg = \frac{PTS_O}{POSS_O}$$

Donde los subíndices T y O denotan sobre qué tipo de datos se analiza. La T es para el equipo y la O para el oponente.

Por último, los llamados "Four Factors" son los porcentajes efectivos de tiros de campo ($eFG\,\%$), las pérdidas por posesión ($TORatio$), el porcentaje de rebotes ($REB\,\%$) y la tasa de tiros libres ($FTRate$). Estos cuatro datos se calculan tanto para el ataque como la defensa.

En la siguiente tabla se muestran los cálculos utilizados para obtener cada uno de los cuatro factores.

Factor	Ataque	Defensa
$eFG\%$	$\frac{P2M_T + 1{,}5\,P3M_T}{P2A_T + P3A_T}$	$\frac{P2M_O + 1{,}5\,P3M_O}{P2A_O + P3A_O}$
$TO\ Ratio$	$\frac{TOV_T}{POSS_T}$	$\frac{TOV_O}{POSS_O}$
$REB\%$	$\frac{OREB_T}{OREB_T + DREB_O}$	$\frac{OREB_O}{OREB_O + DREB_T}$
$FT\ Rate$	$\frac{FTM_T}{P2A_T + P3A_T}$	$\frac{FTM_O}{P2A_O + P3A_O}$

Figura 2: Fórmulas Four Factors

La función *fourfactors* del libro *BasketballAnalyzeR* calcula todos estos índices para cualquier equipo seleccionado. Además, se pueden representar gráficamente los resultados obtenidos a partir de la función *fourfactors*.

5.2.2. Diagrama de barras y de dispersión.

Una gráfica bastante útil para visualizar las diferencias entre equipos o jugadoras de acuerdo a ciertas estadísticas elegidas es el diagrama de barras. Además, se puede añadir más información relevante a través de lo que sería un polígono de frecuencias cuya escala podría leerse en otro eje que, por convención, se coloca a la derecha. Por último, también se pueden ordenar a los equipos en función de alguna otra variable. La función de la librería que realiza estos gráficos es *barline*.

Por otra parte, el diagrama de dispersión (*scatterplot* en el libro de R) ayuda a encontrar relaciones entre dos variables observadas en un conjunto de datos utilizando coordenadas cartesianas, así como encontrar casos anómalos. No obstante, se puede añadir una tercera variable que se representará por el color de los puntos dentro del diagrama.

Para estos gráficos, lo importante consiste en seleccionar bien las variables para obtener una representación que ayude realmente al análisis de datos.

5.2.3. Análisis de la variabilidad.

En baloncesto, la variabilidad puede referirse a la medida en que los jugadores de un equipo rinden de forma diferente entre sí según una estadística determinada. Que exista mucha varianza o no entre los jugadores puede ser un resultado tanto positivo como negativo. Por ejemplo, para algunas variables relacionadas con tareas más especializadas, como pueden ser el número de asistencias, una variabilidad alta puede significar que hay muy pocas jugadoras dedicadas específicamente a esa tarea. Pero, por otra parte, una variabilidad alta en el porcentaje de aciertos en tiros indica que el equipo depende demasiado en muy pocas jugadoras para anotar puntos, mientras que el resto de jugadoras están lejos de la media del resto de equipos.

La función *variability* del libro *BasketballAnalyzeR* calcula el rango, la desviación típica y el coeficiente de variación. Además, la función *plot.variability* dibuja un diagrama en el que, para cada variable, se dibuja un eje vertical y los jugadores se distribuyen en dicho eje en función del valor de cada variable. Además, se representan con una circunferencia cuyo tamaño es proporcional a otra variable. El hecho de observar cómo de lejos están unas circunferencias unas de otras da una idea de la variabilidad de la varianza. Además, en el diagrama se muestran también el rango y el coeficiente de variación de la variable respecto al conjunto de datos seleccionado.

5.2.4. Análisis de correlación lineal

. La función *corranalysis* permite llevar a cabo un estudio de la correlación lineal entre todos los pares de variables de un conjunto de datos. Lo interesante aquí es que se puede representar de manera visual la matriz de correlación y los coeficientes de correlación de Pearson. Esta función está programada para llevar a cabo una prueba con un 95 % de significación en el que la hipótesis nula es que el coeficiente de correlación de Pearson es cero.

5.2.5. Clustering

Uno de los propósitos de la minería de datos consiste en reorganizar y categorizar los datos en función de una serie de variables, con el fin de descubrir algunas relaciones dificiles de percibir a simple vista.

Los algoritmos en los que se basa la minería de datos intentan definir métodos y reglas con el objetivo de asignar unidades de observación en clases (grupos) que a priori no están definidos y supuestamente reflejan la estructura de las entidades que representan los datos.

La clasificación no supervisada de casos individuales en grupos cuyo perfil se especifica de manera espontánea a través de los datos hace referencia al Análisis Clúster, que incluye varias técnicas que se diferencian significativamente en la noción de lo que constituyen los clústeres y cómo encontrarlos de manera eficiente.

El análisis clúster en baloncesto puede aplicarse a jugadores, equipos, partidos, etc. Algunos ejemplos pueden ser clasificar a los jugadores en grupos en función de su desempeño, con el objetivo de definir nuevas posiciones, en contra de las cinco posiciones tradicionales [23], o agrupar a los jugadores en función del desempeño que han realizado en los partidos y observar cómo los salarios de los jugadores que pertenecen a un mismo clúster difieren entre sí.

El análisis clústers o análisis de conglomerados es una técnica de clasificación no supervisada dentro del análisis multivariante. El objetivo de dicha técnica consiste en agrupar datos individuales en clusters tales que los datos de cada clusters sean muy similares entre sí y distintos entre los casos del resto de clusters.

Durante el desarrollo del análisis clúster se observan **cuatro** etapas:

1. Selección de la muestra de estudio.

2. Selección de variables. La selección del número de variables se considera un punto crítico en el estudio y debe ser respaldada por un experto del campo (en nuestro caso, un experto en baloncesto), pues añadir variables al análisis no tiene por qué implicar más homogeneidad entre los datos dentro de un clúster. Por ejemplo, se pueden estar usando variables que desvirtualicen las semejanzas entre los sujetos al no estar relacionadas con el objeto de estudio. Además, una vez las variables son elegidas, en la mayoría de situaciones éstas han de ser normalizadas.

3. Realizar el cálculo de similitudes/disimilitudes entre los casos. Para ello, se necesita una medida de similitud o correspondencia (un paso importante también en el análisis de clúster es determinar la medida más apropiada), además de un algoritmo que sirva para formar los conglomerados (clústeres).

4. Validar los resultados obtenidos. Este procedimiento de evaluación y validación de los resultados del algoritmo se conoce en inglés como *cluster validity*. Este proceso es bastante complejo por la gran cantidad de opciones en el análisis de clústeres (definir la distancia, el algoritmo, el número de variables, el número de clústeres, etc.).

Los algoritmos desarrollados para el análisis clúster se clasifican en dos grandes grupos: Análisis jerárquicos o no jerárquicos, según si el procedimiento de creación de los clústeres se realice a partir de formaciones iniciales o no .

En el **clustering jerárquico** se van generando grupos en cada una de las fases del proceso, buscando el número óptimo de conglomerados. Un ejemplo de algoritmo utilizado en este caso es el método de Ward. El clustering jerárquico puede ser **aglomerativo** o **divisivo**. En el primero, se parten de tantos clústeres como datos haya (grupos individuales de casos) y, en cada uno de los pasos del algoritmo, los clústeres más cercanos se van fusionando de manera sucesiva haciendo que se forme una jerarquía en el resultado. Al final del proceso, solo queda un único clúster que aglutina todos los elementos. Por el contrario, en el clúster divisivo se parte de un único grupo que aglomera todos los datos, y éstos se van dividiendo entre sí para formar, de manera sucesiva en el algoritmo, grupos más pequeños.

Por otra parte, el **clustering no jerárquico** categoriza los datos según un número de clústeres dado. Es decir, el número de particiones está fijado a priori. El algoritmo más conocido de este tipo de análisis es el método k-means (k-medias en español).

Es importante señalar también que la asignación de un dato a un clúster determinado es un paso irreversible, así que no hay posibilidad de reasignar ese dato a otro clúster. Una manera de combinar los dos métodos mencionados consiste en utilizar el análisis de clúster jerárquico de manera exploratoria con el fin de fijar el número de clústeres. Posteriormente, se puede aplicar un análisis no jerárquico con el número de clústeres obtenido en el método jerárquico.

Al final del algoritmo, es importante evaluar la partición obtenida e identificar el número óptimo de clústeres. En general, una partición es buena cuando la homogeneidad dentro de los clústeres de bastante alta y los clústeres están separados entre ellos.

En la práctica, una partición con un número pequeño de clústeres es más fácil de interpretar, aunque esto puede hacer que la homogeneidad dentro del clúster sea difícil de conseguir.

De manera teórica, la bondad de ajuste de los resultados obtenidos se evalúa descomponiendo la desviación total (TV) como suma de la desviación dentro del clúster (WD) y la desviación entre clústeres (BD). Esto es,

$$TD = WD + BD$$

En particular, si N datos se dividen en k clústeres, para una variable X dada se tiene:

$$TD = \sum_{h=1}^{k} \sum_{i=1}^{n_k} (x_{ih} - \mu)^2$$

donde x_{ih} es el valor asumido por el i-ésimo elemento del clúster h, μ es la media de X y n_h es el número de elementos que pertenecen al clúster h.

$$BD = \sum_{h=1}^{k} (\mu_h - \mu)^2 n_h$$

donde μ_h es la media de los elementos de X que forman parte del clúster h.

$$WD = \sum_{h=1}^{k} \sigma_h^2 n_h$$

donde σ_h^2 es la varianza de los elementos de X que forman parte del clúster h.

De esta forma, BD proporciona una medida entre el grado de separación entre clústeres, mientras que WD es una medida de homogeneidad/similitud entre los elementos que pertenecen a un mismo clúster. Esto es, a mayor WD, menor homogeneidad.

El cociente BD/TD es el coeficiente de correlación de Pearson y, en este contexto, mide la calidad en el análisis de clúster respecto a la variable X porque, en la práctica, explica la dependencia media de X en la clusterización propuesta. De hecho, un BD/TD alto implica, o bien un BD alto (buena separación entre los clústeres ya que μ y μ_h difieren mucho), o bien un WD bajo (se recuerda que TD=WD+BD), es decir, una alta homogeneidad dentro del clúster.

En definitiva, la gran cantidad de opciones que han de tomarse en el análisis de clústeres introduce elementos de subjetividad en los resultados. En general, una solución puede considerarse buena cuando permanece estable a medida que los algoritmos cambian.

k-means clustering utilizando el libro *BasketballAnalyzeR*.

La función *kclustering* del libro *BasketBallAnalyzeR* es la encargada de realizar este análisis no jerárquico. El procedimiento que sigue el algoritmo consiste en

1. Elegir los centros de los k clústeres.
2. Asignar cada elemento a su clúster más cercano (utilizando la distancia eucídea).
3. Calcular los centroides de cada clúster, haciendo la media de las coordenadas de los elementos que pertenecen a él.
4. Se recalculan las distancias elementos-centroides y se reasigna cada elemento a su clúster más cercano.
5. Se continúa con el procedimiento hasta que los centroides permanezcan relativamente estables.

Además, la función *kclustering* tiene un parámetro de entrada que permite ejecutar el algoritmo un cierto número de veces y elegir la mejor opción de acuerdo al criterio de la máxima varianza. Este parámetro es *nruns*, y su valor por defecto es 10.

Por otra parte, la función devuelve un objeto de tipo *kclustering*, que se compone de una lista de data frames que contienen a los identificadores de cada clúster, su composición, la media de las variables dentro del cluster

y, para cada clúster, el índice de heterogeneidad (CHI). Este último es la medida de la variabilidad dentro del clúster, que tiene que ser la mínima posible. Se considerarán aceptables los valores menores de 0,5.

Clustering jerárquico aglomerativo.

La función *hclustering* del libro *BasketballAnalyzeR* se encarga de hacer este tipo de agrupación jerárquica. En este caso, se recuerda que, al inicio, cada elemento forma parte de su propio clúster para luego unirse, paso a paso, en un único clúster. Esta unión se realiza en función de la mínima distancia entre clústeres. El algoritmo que sigue esta función es el siguiente:

1. Buscar la distancia más pequeña entre dos clústeres a través de matrices distancias.
2. Unir esos dos clústeres en uno solo y redefinir las matrices distancias.
3. Repite los pasos 1 y 2 $N - 1$ veces, siendo N el número de elementos del dataset.

Los resultados de esta función pueden representarse a través de un dendograma, que ilustra la secuencia de esas fusiones de clústeres.

Los resultados de este tipo de clustering dependen mucho de la distancia utilizada. Los métodos más comunes de clustering jerárquico calculan esas distancias de la siguiente manera:

- Vecinos más próximos (nearest neighbour): utiliza la distancia más corta entre cualesquiera dos miembros de dos clústeres
- Vecinos más lejanos (furthest neighbour): utiliza la distancia más larga entre cualesquiera dos miembros de dos clústeres.
- Unión media: se utiliza la distancia media entre todos los pares de los dos clústeres.

Otros métodos de clústering jerárquico se basan también en el tipo de distancia que se utiliza. Por ejemplo, en el *método de Ward*, los clústeres se crean eligiendo la unión que que minimiza la desvianza dentro del clúster (WD).

La función *hclustering* utiliza el método de Ward, pero para realizar el resto de tipos de clustering se puede usar la función *hclust* de la librería stats de R.

6. Caso práctico

6.1. Base de datos a utilizar

Procedemos a continuación a realizar el análisis de un equipo de baloncesto utilizando los procedimientos descritos y el libro *BasketballAnalyzeR*. Usaremos para ellos la información recogida durante la temporada 2021/2022 de la LF Challenge, la segunda máxima categoría del baloncesto femenino español. Estos datos han sido proporcionados en formato CSV por David García, analista del Manuela Fundación RACA de Granada (RACA en adelante), cuarto equipo mejor clasificado de dicha.

En este punto de inicio, es necesario mencionar la importancia que tiene el hecho de limpiar y estructurar los datos para conseguir resultados aceptables en el estudio. En el caso de la base de datos inicial, los datos no se encontraban estructurados de la misma manera requerida por las funciones de la librería de R, por lo que hubo un trabajo previo -utilizando Excel- para conseguir el conjunto final de datos que nos permitirá ejecutar las funciones de libro sin ningún tipo de problema.

No obstante, sería de gran ayuda para la comunidad -y en especial para la del baloncesto español- contar con código de web scraping que permitiese extraer los datos necesarios de los partidos que se publican en la web de Baloncesto en vivo y que, gracias a la ejecución de funciones de la librería dplyr se consiguiese un conjunto de datos con la estructura requerida para utilizar las funciones del paquete *BasketballAnalyzeR*. Sin embargo, dado que el objetivo real del trabajo es utilizar un paquete de R, se omitirá el proceso de extracción y manipulación de esos datos y se trabajará con los ofrecidos directamente por RACA.

A continuación, concretamos un poco más la base de datos final, así como su estructura:

Box Score

En el Box-score se recopila información estructurada de algunas estadísticas recopiladas durante los partidos de cada equipo. Para nuestro conjunto de datos, se utilizarán tres box-scores diferentes:

1. Box-score del equipo: es un data frame donde las filas representan a cada uno de los equipos analizados de la liga, y las variables (columnas) se refieren a los logros totales en los partidos considerados. A este data frame lo llamaremos **TBox**.

2. Box-score de los oponentes: en este data frame, al que llamaremos

OBox, las filas representan a cada uno de los equipos de la liga, y las variables (columnas) se refieren a los logros totales de los oponentes de cada equipo en los partidos considerados.

3. Box-score de las jugadoras: las filas representan a cada una de las jugadoras de los distintos equipos, y las columnas son los logros individuales en el total de partidos considerados. Nos referiremos a este data frame como **PBox**.

4. Data frame, denotado por **TAdd** que contiene información sobre cada equipo de la liga, con su trigrama, la posición en la liga y si se han clasificado o no para los Playoffs.

A continuación, se muestran una tabla-resumen de las variables de todas las variables que se utilizarán para el análisis de los datos, así como una visión general del aspecto que tiene la base de datos OBox.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
	TEAM	GP	MIN	PTS	W	L	P2M	P2A	P2p	P3M	P3A	P3p	FTM	FTA	FTp	OREB	DREB	AST	TOV	STL	BLK	PF	PM
	MATARO	28	5600	1919	3	25	611	1302	46,9278	139	512	27,1484	280	442	63,3484	325	748	371	404	302	59	415	401
	ARDOI	30	6000	1770	17	13	459	1102	41,6515	187	640	29,2188	291	431	67,5174	299	812	393	529	253	41	491	6
	JOVENTUT	28	5675	1861	9	19	513	1107	46,3415	161	564	28,5461	352	503	69,9801	214	761	323	447	261	41	475	186
	LIMA-HORTA	30	6050	1985	14	16	540	1205	44,8133	185	648	28,5494	350	536	65,2985	367	878	406	569	263	46	548	182
	PATERNA	30	6000	2013	10	20	518	1218	42,5287	217	684	31,7251	326	505	64,5545	338	779	435	528	335	51	483	199
	MIRALVALLE	29	5800	2147	4	25	662	1337	49,5138	164	521	31,4779	331	504	65,6746	316	779	362	463	297	48	505	360
	CANOE	30	5825	1940	13	17	552	1192	46,3087	187	607	30,8072	275	412	66,7476	311	728	445	520	273	56	547	70
	IRAURGI	30	6050	2149	9	21	605	1223	49,4685	216	730	29,589	291	431	67,5174	321	741	429	464	320	58	498	208
	CACERES	30	6100	2085	11	19	520	1198	43,4057	248	736	33,6957	301	455	66,1538	268	770	420	419	255	53	545	77
	CELTA	28	5650	1778	15	13	511	1152	44,3576	168	537	31,2849	252	371	67,9245	231	707	332	504	271	50	496	-98
	RACA	30	6050	1778	21	9	479	1219	39,2945	159	567	28,0423	343	495	69,2929	344	984	330	609	282	80	483	-257
	JAIRIS	30	6025	1741	24	6	481	1208	39,8179	156	578	26,9896	311	467	66,5953	291	688	312	517	273	36	460	-372
	ALCOBENDAS	30	6000	1927	21	9	560	1293	43,3101	169	642	26,324	300	432	69,4444	290	726	414	407	288	47	517	-190
	BARCA	29	5975	1715	26	3	458	1165	39,3133	157	580	27,069	328	499	65,7315	264	712	338	545	308	54	509	-333
	ESTEPONA	30	6025	1979	18	12	509	1221	41,6871	223	789	28,2636	292	463	63,067	344	682	451	471	251	24	543	-157
	ZAMORA	29	5875	1810	21	8	548	1266	43,2859	153	565	27,0796	255	372	68,5484	225	705	299	442	265	38	507	-312

Figura 3: Data Frame OBox

Variable	Descripción	TBox	OBox	PBox	Tadd
Team	Equipo analizado (nombre largo)	X	X	X	X
team	Equipo analizado (nombre corto)				X
Rank	Posición en la liga regular				X
Playoff	Clasificación en los playoffs (Sí o No)				X
Player	Jugadora			X	
Position	Posición que ocupa			X	
GP	Partidos jugados	X	X	X	
MIN	Minutos jugados	X	X	X	
PTS	Puntos anotados	X	X	X	
W	Victorias	X	X		
L	Derrotas	X	X		
P2M	Tiros de dos encestados	X	X	X	
P2A	Tiros de dos intentados	X	X	X	
P2p	Porcentaje de acierto en tiros de dos	X	X	X	
P3M	Triples encestados	X	X	X	
P3A	Triples intentados	X	X	X	
P3p	Porcentaje de triples anotados	X	X	X	
FTM	Tiros libres encestados	X	X	X	
FTA	Tiros libres intentados	X	X	X	
FTp	Porcentaje de acierto en tiros libres	X	X	X	
OREB	Rebotes ofensivos	X	X	X	
DREB	Rebotes defensivos	X	X	X	
AST	Asistencias	X	X	X	
TOV	Pérdidas	X	X	X	
STL	Robos	X	X	X	
BLK	Tapones	X	X	X	
PF	Faltas personales	X	X	X	
PM	Más/Menos	X	X	X	

Cuadro 1: Estructura de la base de datos.

Para cargar estos datos no es necesario instalar la librería de R, sino que solo basta con leerlos con la función *read_excel*, disponible en la librería *readxl*. Así, el código a compilar, unido a la instalación de la librería *BasketballAnalyzeR* sería el siguiente:

```
library(readxl)
OBox <- read_excel("C:/Users/anate/Desktop/Curso 21-22/TFM/
       Baloncesto/Mi base de datos/OBox.xlsx")
TBox <- read_excel("C:/Users/anate/Desktop/Curso 21-22/TFM/
        Baloncesto/Mi base de datos/TBox.xlsx")
PBox <- read_excel("C:/Users/anate/Desktop/Curso 21-22/TFM/
        Baloncesto/Mi base de datos/PBox.xlsx")
TAdd <- read_excel("C:/Users/anate/Desktop/Curso 21-22/TFM/
        Baloncesto/Mi base de datos/Tadd.xlsx")

#Instalamos y cargamos el paquete BasketballAnalyzeR

install.packages("BasketballAnalyzeR")
library(BasketballAnalyzeR)
RNGkind(sample.kind = "Rounding")
```

6.2. Análisis de los datos.

Comenzamos el estudio de los datos con un análisis gráfico que nos permita conocer los datos que disponemos de manera visual, para que de esta manera, podamos conseguir ideas generales sobre el comportamiento de los datos y otras características. Para ello utilizaremos las funciones que describimos con detenimiento en el capítulo anterior: fourfactors y barline.

Para empezar, crearemos un objeto de tipo FourFactors con la función del mismo nombre y representaremos gráficamente los datos extraídos. Para este caso, no vamos a representar los datos de todos los equipos sino que lo haremos de los cuatro que se han clasificado para los playoffs de ascenso a Liga Femenina Endesa.

```
tm<- c("RAC","ZAM","ALC","JAI")
selTeams <- which(TAdd$team %in% tm)
selTeams
FF.sel <-fourfactors(TBox[selTeams,],OBox[selTeams,])
FF.sel
FF.1<-plot(FF.sel)
library(gridExtra)
grid.arrange(grobs=FF.1[3:4],ncol=2)
grid.arrange(grobs=FF.1[1:2],ncol=2)
```

Las tres últimas líneas de código se han utilizado para dibujar los cuatro gráficos de manera separada en lugar de en un mismo plot. Para ello, se ha

utilizado el libro *gridExtra* [24].

En primer lugar, mostraremos la variable FF.sel, y a continuación las cuatro gráficas obtenidas:

```
> head(FF.sel)
       Team POSS.Off POSS.Def  PACE.Off  PACE.Def  ORtg  DRtg F1.Off F2.Off F3.Off F4.Off F1.Def F2.Def F3.Def F4.Def
1      RACA  2290.08  2268.80 0.3785256 0.3750083 88.86 78.37  44.16  18.95  25.00  13.74  40.17  26.84  72.10  19.20
2    JAIRIS  2236.52  2217.48 0.3712066 0.3680465 94.48 78.51  47.78  21.33  34.97  13.41  40.03  23.31  73.55  17.41
3 ALCOBENDAS  2266.44  2242.08 0.3777400 0.3736800 93.41 85.95  47.61  21.88  27.18  21.03  42.04  18.15  74.22  15.50
4    ZAMORA  2204.84  2211.68 0.3752919 0.3764562 96.24 81.84  48.23  20.27  29.57  18.86  42.46  19.98  78.43  13.93
```

Figura 4: Aspecto de la variable FF.sel

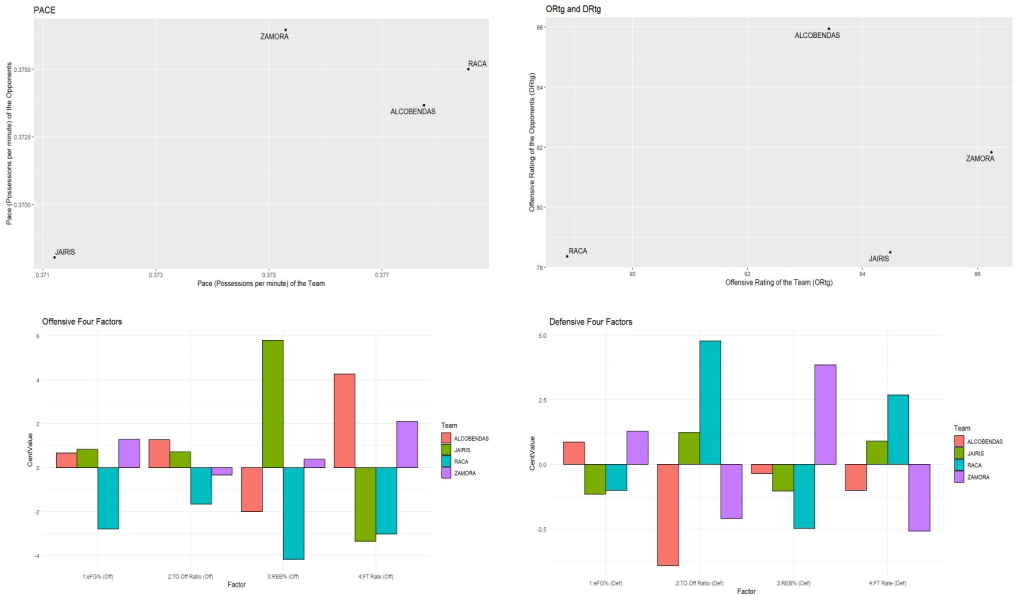

Figura 5: Gráficos Four Factors

A partir de ellos, se pueden destacar las siguientes conclusiones:

- **Ritmo (Pace):** el ritmo de los partidos ha sido lento en todos los partidos del Jairis (tanto por parte del propio equipo como por parte de sus oponentes). Por otra parte, el ritmo impuesto por el Zamora ha sido más lento que el impuesto por el de sus adversarios. Y, entre RACA y Alcobendas, los partidos del primer equipo mencionado han sido más intensos.
- **Eficacia en el ataque/defensa (ORtg y DRtg):** El equipo granadino destaca por tener valores bajos en su eficacia tanto en ataque

como defensa. El Jairis tiene eficacia en el ataque pero tambalea en la defensa, mientras que el Zamora es el equipo que más eficacia en el ataque tiene y el Alcobendas el que más tiene en defensa.

- **FourFactos en ataque y defensa:** las barras representan, para cada equipo, la diferencia entre el valor del equipo y la media de los cuatro equipos analizados. El valor positivo y negativo, así como la altura de la barra nos puede dar una idea sobre las debilidades y fortalezas de un equipo respecto a los otros. Por ejemplo, para los cuatro factores de ataque, se ve muy claro cómo Raca es el equipo más débil de los cuatro.

Otra representación gráfica que podemos utilizar y que facilita el análisis descriptivo de los datos es el diagrama de barras. En particular, vamos a hacer dos, uno que represente características de defensa y otro de ataque.

En el caso de la defensa, se tomarán como variables el número de rebotes totales, tapones y recuperaciones. Para el polígono de frecuencias, las pérdidas totales de los oponentes. Además, la disposición de los equipos se ordenará en función de los puntos recibidos en orden ascendiente (de menor a mayor).

```
X <- data.frame(TBox, PTS.O=OBox$PTS, TOV.O=OBox$TOV)
labs <- c("Robos","Tapones","Rebotes defensivos")
barline(data=X, id="Team", bars=c("STL","BLK","DREB"),
        line="TOV.O", order.by="PTS.O", decreasing = F,
        labels.bars=labs)
```

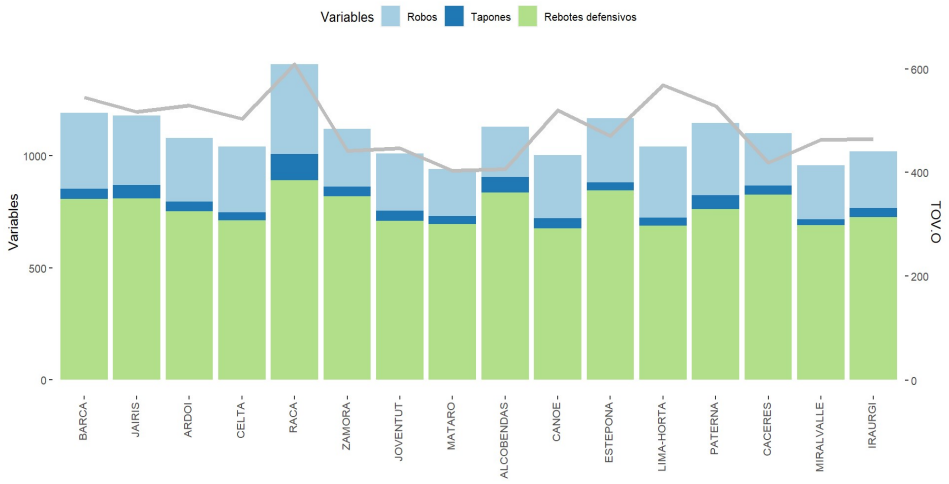

Figura 6: Gráfico de barras para la defensa.

Con este gráfico se puede observar que el equipo que mejor defiende de la liga es RACA, siendo el cuarto con menos puntos recibidos, el que más pérdidas provoca al equipo rival y el que más robos, tapones y rebotes defensivos realiza.

De hecho, con tan solo observar el diagrama de barras, RACA podría enfocar más sus entrenamientos a trabajar el ataque. De hecho, según el gráfico correspondiente al ataque, podemos destacar que RACA no es un equipo que destaque tácticamente en este ámbito de juego.

```
Y <- data.frame(TBox, DREB.O=OBox$DREB, REC.O=OBox$STL)
labs <- c("Rebotes defensivos rival",
          "Recuperaciones Rival","Pérdidas")
barline(data=Y, id="Team", bars=c("DREB.O","REC.O","TOV"),
        line="AST", order.by="PTS",decreasing = F,
        labels.bars=labs)
```

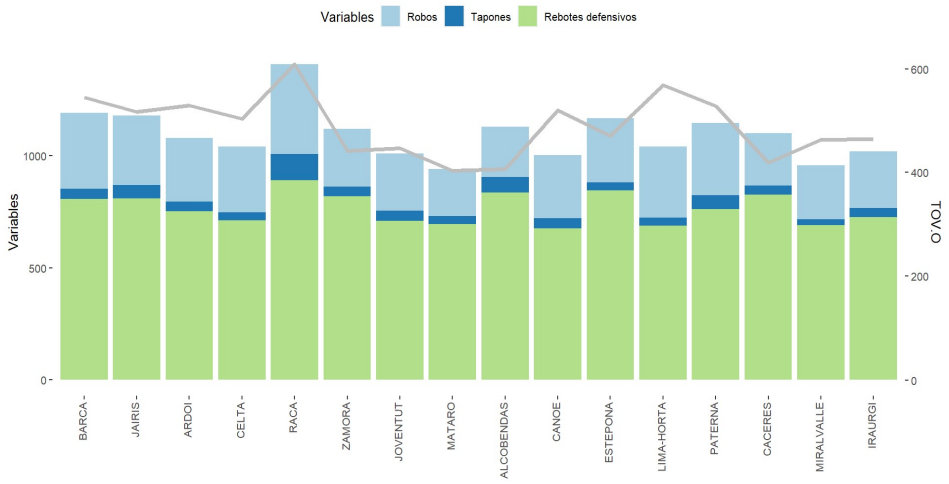

Figura 7: Gráfico de barras para el ataque.

En este caso, lo que representa la gráfica de ataque son los rebotes defensivos del rival (son los rebotes ofensivos que el equipo ha perdido), las recuperaciones del rival y el número total de pérdidas; la línea representa las asistencias y los equipos están ordenados de manera creciente en función de los puntos totales anotados.

Como conclusión de lo obtenido, se observa que todos los equipos están más o menos al mismo nivel en términos de ataque. De manera individual, RACA ha perdido más rebotes ofensivos, por lo que quizás estaría bien enfocar los entrenamientos a coger rebotes en el ataque.

Por otra parte, se puede realizar un diagrama de dispersión para investigar la relación entre el número de pérdidas y asistencias por minuto de todas las jugadoras que han participado más de 600 minutos a lo largo de la temporada. Además, los puntos del diagrama se colorearán en función de los puntos anotados por minuto de cada jugadora. El código es el siguiente:

```
PBox.sel <- subset(PBox, MIN>= 300)
attach(PBox.sel)
X <- data.frame(AST, TOV, PTS)/MIN
detach(PBox.sel)
mypal <- colorRampPalette(c("blue","yellow","red"))
scatterplot(X, data.var=c("AST","TOV"), z.var="PTS",
            labels=1:nrow(X), palette=mypal)
```

Quizás, lo lógico hubiera sido que, cuanto más mueve un jugador la pelo-

ta, más pérdidas tiene. Sin embargo, el gráfico dibujado muestra una cierta relación constante entre ambas variables, de tal forma que el número de pérdidas se encuentra en una franja horizontal y no depende del número de asistencias. Además, se observa el poco número de asistencias por minuto de prácticamente todas las jugadoras de la liga.

No obstante, se pueden observar valores fuera de la franja vertical del 0,04 y 0, 12 tales como jugadoras que tienen muchas pérdidas de balones en comparación con el resto, y más asistencias que la mayoría (31 y 89) ; y jugadoras con pocas pérdidas y pocas asistencias respecto a las demás (57 y 135). De hecho, la número 57 es una de las que más puntos anota por minuto y quizás por ello se explique su condición de tener pocas asistencias (prefiere tirar a canasta que dar un pase, porque su efectividad es mayor).

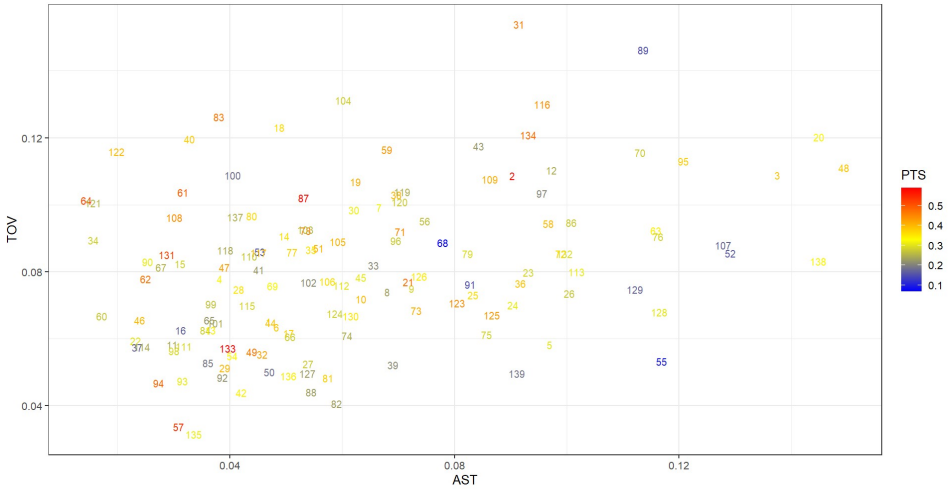

Figura 8: Diagrama de dispersión de asistencias y pérdidas.

Para continuar con la parte descriptiva del conjunto de datos,se va a realizar un análisis de variabilidad de los porcentajes de cada uno de los tres tipos de tiros (libres, de dos puntos, y triples) de las jugadoras del equipo granadino que han jugado más de 300 minutos. Para ello, se utilizará la función *variability* descrita anteriormente, y se representará de manera gráfica para extraer conclusiones.

```
Pbox.RAC <- subset(PBox, TEAM=="RACA"
                   & MIN>=300)
vrb <- variability(data=Pbox.RAC,
                   data.var=c("P2p","P3p","FTp"),
                   size.var=c("P2A","P3A","FTA"),
                   weight=TRUE)
plot(vrb, title="Diagrama de variabilidad - RACA")
```

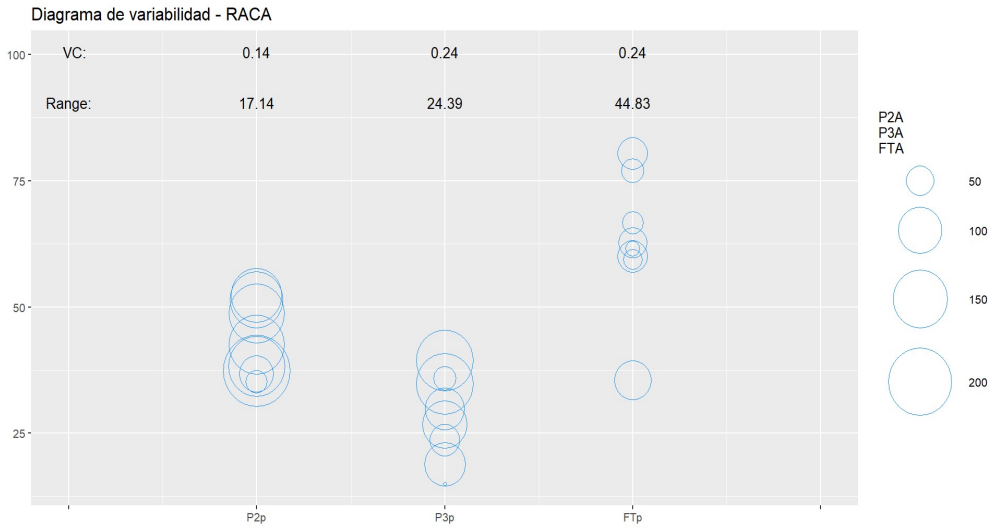

Figura 9: Diagrama de variabilidad en porcentaje de tiros.

En el gráfico, el tamaño de las circunferencias representan la proporción de tiros intentados. Como es obvio, los porcentajes de tiros de tres son los más bajos debido a la dificultad que existe a la hora de anotar desde la zona de triple.

En cuanto a la variación, de las tres variables, el porcentaje de tiros libres es el que más variabilidad se observa, aunque tiene el mismo coeficiente de variación que el porcentaje de triples. Por otra parte, destaca una jugadora con poca efectividad en tiros libres, al igual que hay dos jugadoras con una efectividad de más del 75 % que quizás pueda considerarse como un valor atípico.

También, echando un ojo a los tiros de tres, hay otra jugadora que destaca por tener un porcentaje alto de efectividad respecto a sus compañeras,

pero no es de las que más tiros intenta desde esa distancia.

A continuación, vamos a intentar observar las posibles relaciones lineales entre pares de variables a partir del análisis de correlación lineal. Para este caso, crearemos un data frame con las variables PTS, P3M, P2M, REB, AST, TOV, STL y BLK por minuto y representaremos los resultados en dos gráficos.

```
data <- merge(PBox, TAdd, by="TEAM")
data <- subset(data, MIN >= 300)
attach(data)
X <- data.frame(PTS, P3M, P2M, REB=(OREB+DREB), AST,
                TOV, STL, BLK)/MIN
X <- data.frame(X, Playoff=Playoff)
detach(data)
corrmatrix <- corranalysis(X[,1:8], threshold=0.5)
plot(corrmatrix)
```

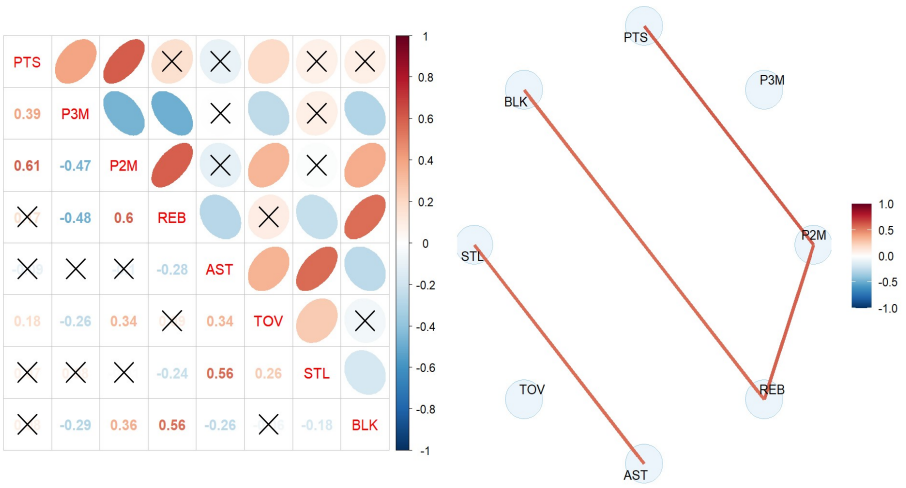

Figura 10: Gráfico que representa la matriz de correlación lineal.

En el primer gráfico se encuentra la matriz de correlación lineal. En ella, se representan las variables utilizadas en la diagonal. En la parte triangular inferior de la matriz se encuentran los coeficientes de Pearson que relacionan a cada par de variables. En la parte triangular superior se puede observar una representación por colores de esos mismos coeficientes, siendo el color más oscuro el que representa una correlación más fuerte entre dos pares de

variables.

En el grafo de la derecha dibujado, cada arista une dos variables siempre y cuando el coeficiente de correlación lineal de Pearson sea mayor que $0,5$. Así, se pueden identificar cuatro tipos de relaciones: asistencias-robos, tapones-rebotes, rebotes-tiros de dos anotados y puntos anotados-tiros de dos anotados. La relación entre los dos últimos pares de variables mencionados es mucho más fuerte que en los otros dos. Por una parte, es obvio que exista una correlación lineal entre los puntos totales y los de dos, pero es más interesante saber que existe una relación entre el número de rebotes totales con los tiros de dos anotados, pues cuantos más rebotes ofensivos coja un equipo, más tiros a canasta intentados tendrá (en este caso cerca del aro) y, por tanto, más posibilidades de anotar canasta.

Se puede realizar también un análisis de correlación lineal entre pares de variables pero de manera en que se pueda discernir al conjunto de datos en dos partes. Por ejemplo, se puede realizar el mismo análisis de correlación pero esta vez dividiendo el conjunto de datos entre los equipos que sí se han clasificado para playoffs y los que no. En particular, la función *scatterplot* permite representar los resultados de la siguiente manera:

```
scatterplot(X, data.var=1:8, z.var="Playoff",
            diag=list(continuous="blankDiag"))
```

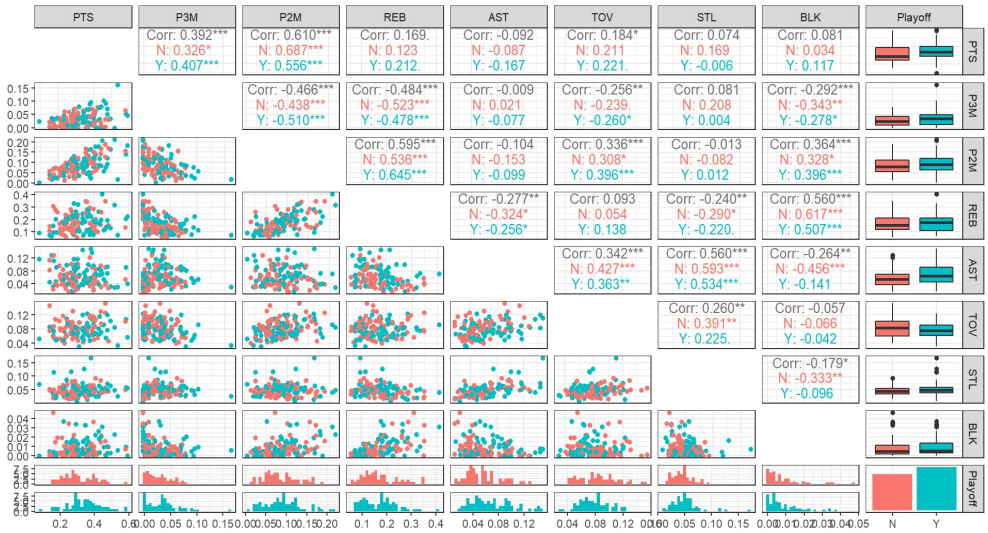

Figura 11: Diagrama de dispersión para variables correlacionadas.

donde en la parte triangular inferior se observan los diagramas de dispersión,

y en la superior los correspondientes coeficientes de correlación lineal de Pearson. Los colores indican los equipos clasificados, o no, para playoffs. Además, las dos últimas filas representan los histogramas de cada uno de los grupos y, la última columna, el diagrama de cajas y bigotes correspondiente. Este gráfico ofrece información bastante completa acerca del conjunto de equipos de la liga.

Una vez terminado el análisis descriptivo, se va a intentar clasificar tanto a los equipos como a las jugadoras en diferentes grupos en relación a una serie de variables. En efecto, vamos a llevar a cabo el **análisis de clusters**. Primero, haremos el método de las k-medias aplicado a los equipos de la liga. Como variables, se elegirán estas siete:

1. De la función *fourfactors* entre los equipos y sus oponentes:
   - La proporción $ORtg/DRtg$.
   - La proporción entre el primer factor ($eFG\,\%$) ofensivo y el defensivo.
   - La proporción entre el segundo factor ($TORatio$) defensivo y ofensivo.
   - El tercer factor ($\%REB_O$) ofensivo.
   - El tercer factor ($\%REB_D$) defensivo.
2. Del dataframe TBox el número de triples encestados $P3M$.
3. La proporción entre los robos del equipo sobre los robos de los oponentes STL_T/STL_O

El código utilizado para generar estas variables, así como el dataframe formado por las mismas y la aplicación de la función *kclustering* se muestra a continuación.

```
FF <- fourfactors(TBox,OBox)
OD.Rtg <- FF$ORtg/FF$DRtg
F1.r <- FF$F1.Off/FF$F1.Def
F2.r <- FF$F2.Def/FF$F2.Off
F3.Off <- FF$F3.Off
F3.Def <- FF$F3.Def
P3M <- TBox$P3M
STL.r <- TBox$STL/OBox$STL
data <- data.frame(OD.Rtg, F1.r, F2.r, F3.Off,
                   F3.Def,P3M, STL.r)

set.seed(29)
```

```
kclu1 <- kclustering(data)
plot(kclu1)
```

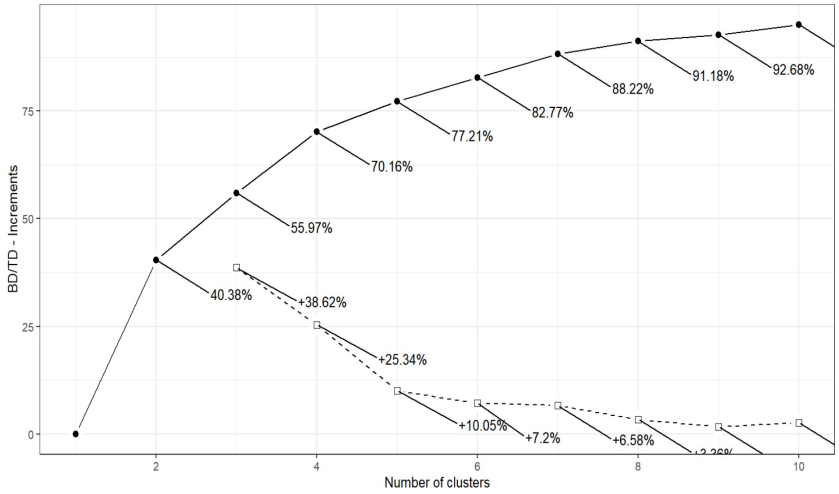

Figura 12: Diagrama de clustering k-medias.

La línea sólida del gráfico representa la relación BD/TD de la que se ha hablado en la sección 5.2.5, que mejora a la vez que el número de clusters aumenta. Valores mayores del 50 % se consideran satisfactorios, ya que la clusterización es capaz de explicar más de la mitad de la variabilidad total.

Por otra parte, la línea de puntos representa el incremento porcentual de la relación BD/TD pasando de la solución del clúster $(k-1)$ a la del clúster k.

El objetivo es conseguir un equilibrio entre esas dos líneas, de manera que haya un valor grande en cuanto a términos de BD/TD y un número de clústers pequeño (teniendo en cuenta también que son 16 equipos en total).

Por tanto, deberíamos identificar un umbral para el cual esa mejora obtenida gracias a un clúster adicional es demasiado pequeña como para justificar la mayor complejidad generada por el propio clúster. Amitiremos el criterio de aumentar el número de clústers siempre que la diferencia supere un $5-10\,\%$.

En este caso, parece que agrupar a los equipos en cuatro clústeres es la mejor opción, ya que tiene una calidad de clusterización de $BD/TD = 70,16\,\%$.

Así, los clústeres se componen de los siguientes equipos:

- **Clúster 1**: Cáceres, Canoe, Joventut, Ardoi, Celta.
- **Clúster 2**: Lima-Horta, RACA.
- **Clúster 3**: Estepona, Alcobendas, Zamora, Jairis, Barça.
- **Clúster 4**: Iraurgi, Miralvalle, Paterna, Mataró.

A continuación, vamos a representar los resultados en un gráfico radial y en un diagrama de barras con el fin de poder interpretar los datos del resultado de aplicar la función *kclustering* con $k = 4$ clústeres. Además, se comentarán las características similares de los equipos que integran cada clúster.

```
kclu2 <- kclustering(data, labels=TBox$Team, k=4)
plot(kclu2)

kclu2.PO <- table(kclu2$Subjects$Cluster, TAdd$Playoff)
kclu2.W <- tapply(TBox$W, kclu2$Subjects$Cluster, mean)
Xbar <- data.frame(cluster=c(1:4), N=kclu2.PO[,1],
                   Y=kclu2.PO[,2], W=kclu2.W)
barline(data=Xbar, id="cluster", bars=c("N","Y"),
        labels.bars=c("Playoff: NO","Playoff: YES"),
        line="W", label.line="average wins",
        decreasing=FALSE)
```

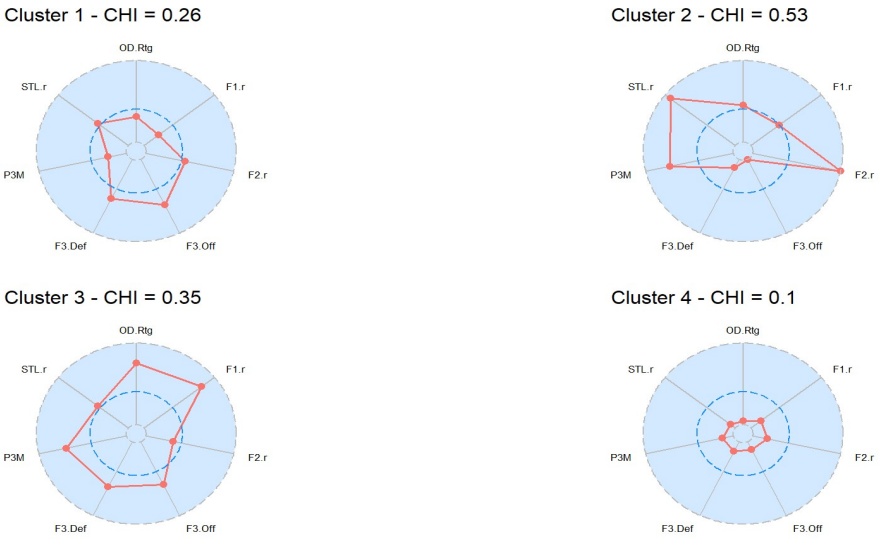

Figura 13: Diagrama radial de los clústers con el método k-medias.

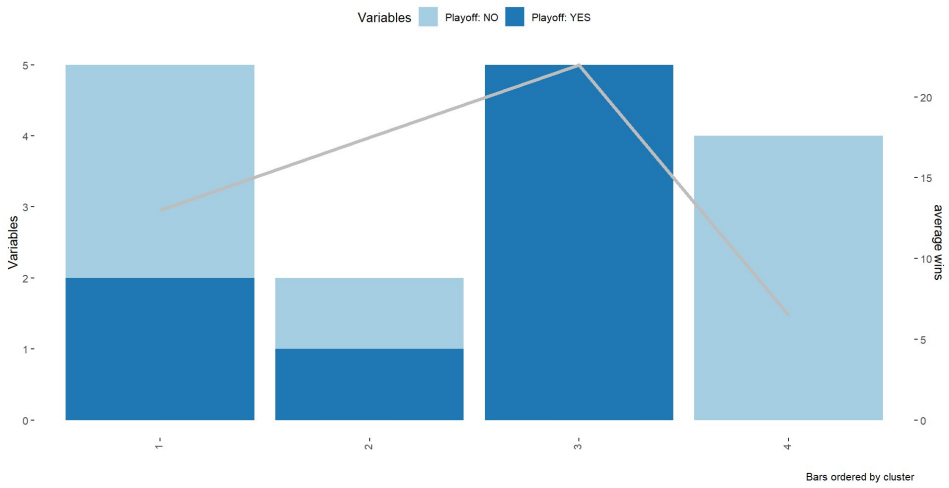

Figura 14: Cantidad de equipos que están o no en Playoffs en función del clúster al que pertenecen.

En el gráfico radial se pueden observar datos de cada uno de los clústers. Podemos observar como el índice de heterogeneidad dentro de cada clúster (CHI) nos da una pista de cómo de homogéneos son los equipos dentro de cada uno. Así, los clústeres 1, 3 y 4 están bien definidos por tener un CHI más bajo, mientras que el 2 es algo más heterogéneo.

Por otra parte, la circunferencia marcada en azul indica la media general de todas las observaciones (la media de las medias de cada una de las

variables), que han sido previamente estandarizadas.

Observando los resultados, en el clúster 4 se encuentran todas las variables por debajo de la media, mientras que en el tercero tan solo se encuentra la proporción entre las pérdidas provocadas. De hecho, mirando en el diagrama de barras se observa que los equipos que forman parte del clúster 4 no se han clasificado para los Playoffs, mientras que todos aquellos que pertenecen al clúster 3 son equipos que sí se clasificaron.

Por otra parte, en el clúster 1 se encuentran equipos en la parte mediabaja de la clasificación. El perfil de estos equipos demuestra un gran trabajo en el rebote -defensivo y ofensivo-, así como un factor 2 por encima de la media, lo que quiere decir una efectividad buena en el ataque en cuanto a pérdidas. Sin embargo, estos equipos no suelen destacar por un acierto en triples, al igual que los del clúster 4, que tienen un P3M bajo en comparación con el resto.

En el clúster 2 se observan datos muy dispares (y por eso el CHI es mayor). Los datos más pobres son el porcentaje de rebotes totales. Esto se puede interpretar como un trabajo poco efectivo en el rebote. Sin embargo, el resto de variables se mueven por encima de la media. Son equipos que sobresalen en el tiro de tres, que provocan más pérdidas de las que hacen y quien roba más balones que sus adversarios.

A continuación, utilizando la función *bubbleplot*, vamos a representar gráficamente un diagrama de burbujas de los equipos. Representaremos, en el eje x, los puntos anotados por cada equipo; mientras que en el eje y estarán los anotados por sus oponentes. El tamaño de la burbuja será el porcentaje de victorias, y los clústeres se representarán por colores.

```
cluster <- as.factor(kclu2$Subjects$Cluster)
Xbubble <- data.frame(Team=TBox$Team, PTS=TBox$PTS,
                      PTS.Opp=OBox$PTS, cluster,
                      W=TBox$W)
labs <- c("PTS", "PTS.Opp", "cluster", "Victorias")
bubbleplot(Xbubble, id="Team", x="PTS", y="PTS.Opp",
           col="cluster", size="W", labels=labs)
```

El gráfico muestra cómo los equipos del clúster 3, por ejemplo, son equipos con muchos puntos anotados y pocos recibidos (y son los que se encuentran en la parte alta de la clasificación). Además, los equipos pertenecientes

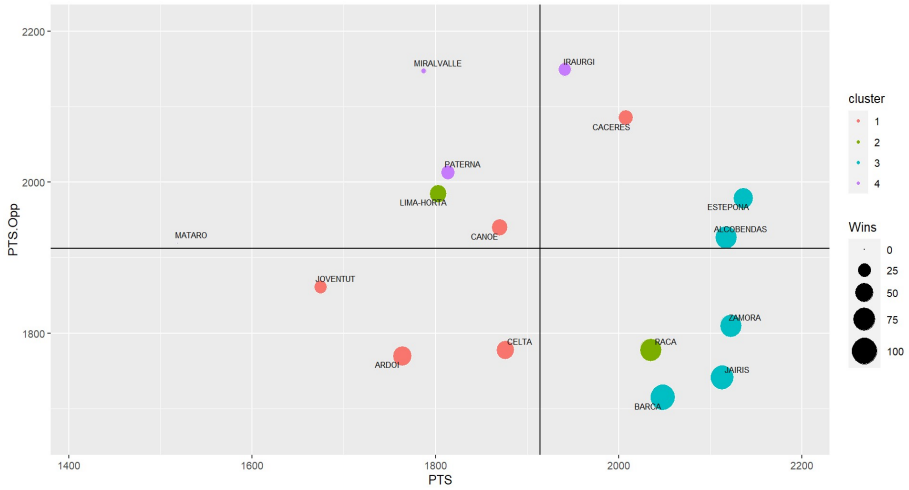

Figura 15: Puntos anotados y recibidos por cada equipo, clasificados por clústeres.

al clúster 1 están en todos los cuadrantes menos en el cuarto, lo que significa que una manera parecida de jugar puede llevar a los equipos a obtener logros diferentes. Joventut, Ardoi y Celta anotan por debajo de la media, pero también reciben menos puntos.

Además, los tres primeros equipos de la liga (Barça, Zamora y Jairis) no se caracterizan por su defensa a la hora de recibir menos puntos. De hecho, el Barça ha sido primero en la liga y ha anotado menos puntos que el resto de equipos del mismo clúster.

Después de este tipo de clustering, vamos a utilizar la función *hclustering* sobre las jugadoras de la liga que han jugado más de 300 minutos en total para agruparlas, según el método de Ward, en clústeres en los que las variables serán los puntos totales anotados, triples, rebotes totales, asistencias, pérdidas, robos, tapones y número de faltas. El código es el siguiente:

```
data <- data.frame(PBox$PTS, PBox$P3M,
                   REB=PBox$OREB+PBox$DREB, PBox$AST,
                   PBox$TOV, PBox$STL, PBox$BLK,
                   PBox$PF)
data<-subset(data, PBox$MIN>=300)
dim(data)
ID <- PBox$PLAYER[PBox$MIN>=300]

hclu1 <- hclustering(data)
plot(hclu1)
```

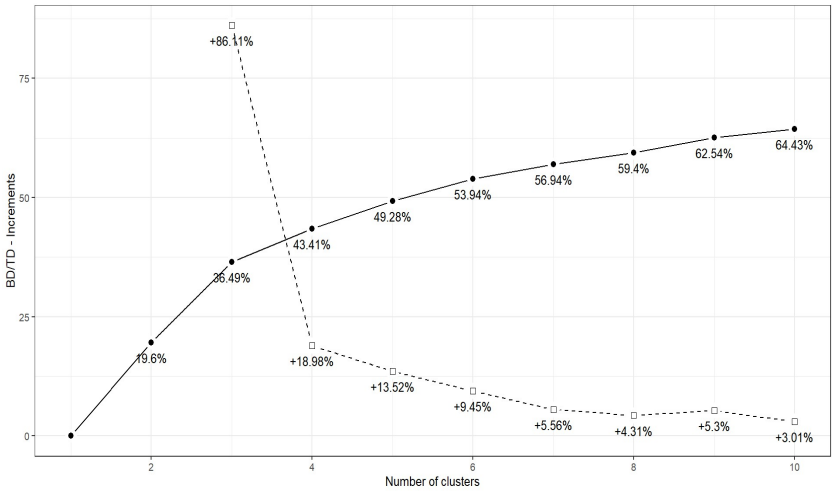

Figura 16: Diagrama de clustering aglomerativo con el método de Wald.

Según el gráfico, que ya hemos explicado antes, lo más lógico sería coger 4 clústers. Sin embargo, mirando el coeficiente de heterogeneidad CHI, el dividir a las jugadoras en cinco grupos parece mejor idea y, además, el 49,28 % de las jugadoras están bien clasificadas en dichos grupos. Para la visualización del diagrama radial con cuatro clústeres, tan solo es suficiente con ejecutar el siguiente código con k=5.

```
hclu2 <- hclustering(data, labels=ID, k=5)
plot(hclu2, profiles=TRUE)
```

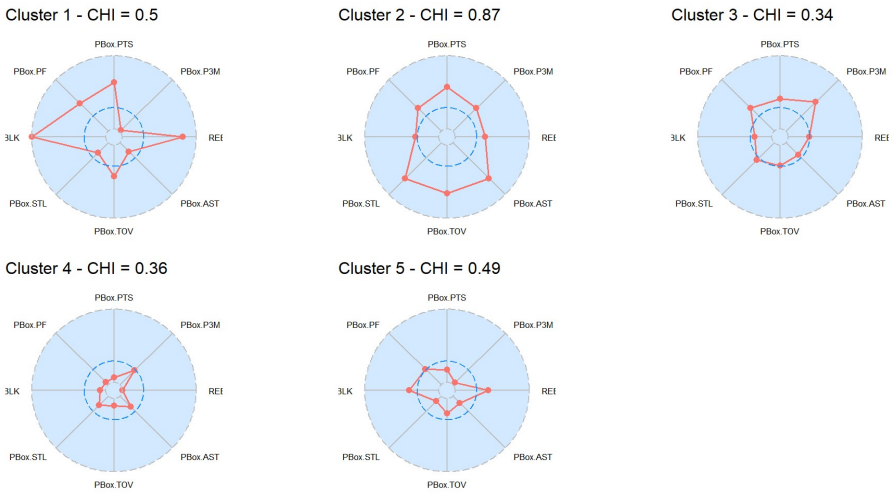

Figura 17: Diagrama radial de los clústers con el método de Wald.

Según el gráfico radial, se observa que el clúster 2 tiene un coeficiente alto de heterogeneidad, pues las jugadoras que pertenecen a dicho clúster son más dispares entre sí.

En el clúster 1 se observan a jugadoras que han anotado muchos puntos, aunque muy pocos desde la línea de tres. Además, suelen cometer bastantes faltas, hacer muchos bloqueos y coger rebotes. Estas características, junto a las bajas cifras en cuanto a robos y asistencias, hacen que en el clúster 1 se encuentren las pívots, jugadoras normalmente altas cuyo tipo de juego se realiza en el interior. En efecto, podemos comprobar fácilmente la veracidad de estos argumentos observando la posición que ocupan las jugadoras que pertenecen a este clúster.

```
datos<-PBox[PBox$MIN>=300,]
a<- hclu2$Subjects$Cluster
b<-datos[a==1, 1:3]
b
```

	TEAM	POSITION	PLAYER
1	ALCOBENDAS	INTERIOR	CLARA RODRIGUEZ
2	CÁCERES	INTERIOR	SARA ZARAGOZA
3	CELTA	INTERIOR	MAGGIE MULLIGAN
4	IRAURGI	INTERIOR	MARGUERITE EFFA
5	IRAURGI	INTERIOR	ANISHA GEORGE
6	JAIRIS	INTERIOR	ERIKA DE SOUZA
7	MIRALVALLE	INTERIOR	IJEOMA UCHENDU
8	RACA	INTERIOR	JULIETA MUNGO
9	ZAMORA	INTERIOR	NNEKA EZEIGBO

Figura 18: Jugadoras en el clúster 1.

Las jugadoras del clúster 2 parecen ser las más completas. Son buenas tiradoras; son intensas en defensa, pues roban más balones; tienen mayor número de pérdidas porque participan bastante en el juego. Y, por otra parte, no tienen números destacables en cuanto a bloqueos y rebotes. En este clúster se encuentran tan solo bases y jugadoras exteriores, que tienen un perfil más participativo que las pívots en el desarrollo del juego y que a la vez son las jugadoras que más minutos han jugado en la temporada.

En el clúster 3 se encuentran jugadoras anotadoras y con muchas faltas (también han jugado bastantes minutos). Su participación en los partidos no es muy destacable en cuanto a defensa y ataque, perí sí en cuanto a efectividad.

Dentro del clúster 4 se puede observar que prácticamente todas las variables están centro de la circunferencia. Pese a parecer que pueden ser jugadoras menos completas, la mayoría de las jugadoras han participado en todos los encuentros. Además, el número de pérdidas es muy bajo, luego son jugadoras que tienen buen control de balón. Por otra parte, el hecho de tener pocas faltas y un número bajo de rebotes y tapones hace que estas jugadoras sean exteriores o bases, y de una altura no muy grande.

Por último, se puede ver una semejanza entre el clúster 1 y el 5. En efecto, el clúster 5 contiene a jugadoras interiores pero con menos efectividad en los tiros y menos minutos.

Este tipo de clustering, aplicado a las jugadoras de todos los equipos, puede ser de gran ayuda a la hora de construir una nueva plantilla. De esta forma, si el cuerpo técnico de un equipo busca jugadoras con ciertas carac-

terísticas, definidas en función de la estrategia que se pretenda seguir a lo largo de la temporada, un analista podría agrupar a las jugadoras de la liga en función de dichas características y así el equipo técnico puede tomar la mejor decisión y fichar a aquéllas más completas, o incluso a una jugadora nueva con las mismas características que tenía otra que dejará el equipo en la siguiente temporada.

Es muy común también representar los resultados del clustering jerárquico en un dendograma. La función *plot.hclustering* del libro *BasketballAnalyzeR* dibuja un dendograma del clustering aglomerativo, que muestra la secuencia de la fusión de los clústeres y la distancia a la que cada una de dichas fusiones se llevaron a cabo.

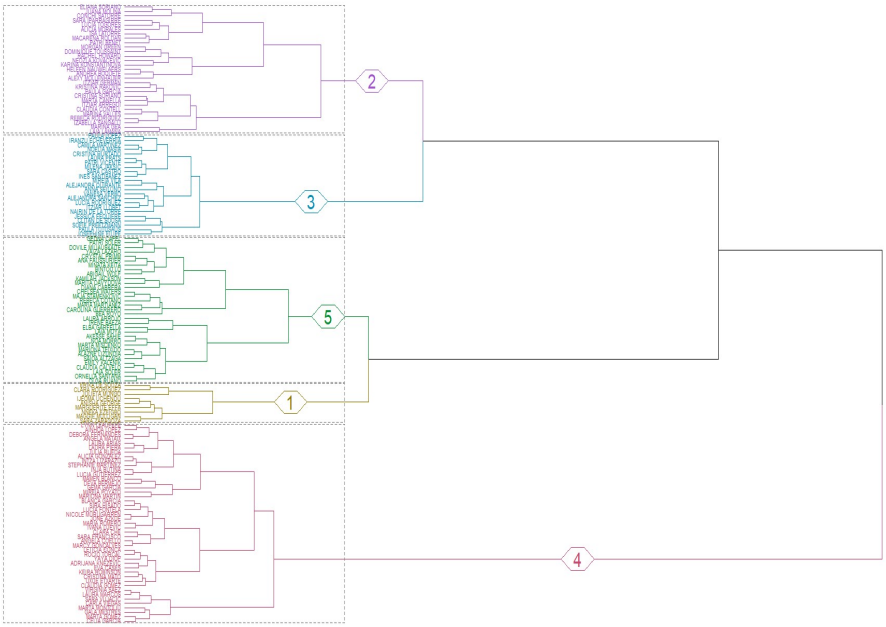

Figura 19: Dendograma utilizando el método de Wald.

En la imagen, se puede observar también cómo el clúster 4 es el que más diferencias tiene junto con el resto (es el último en fusionarte con los demás), mientras que las parejas 2-3 y 1-5 son las que menos diferencias tienen entre ellas. Por otra parte, si se hiciera zoom en el dendograma, se podrían observar las parejas de jugadoras con menos diferencias utilizando este tipo de clustering.

6.3. Gráficos de tiro.

Para concluir con esta última sección, dedicaremos una parte a la función *shotchart*, que proporciona diferentes tipor de gráficos basados en las coordenadas de tiro.

Sin embargo, como la librería *BasketballAnalyzeR* está diseñada para utilizarlas con equipos y jugadores de la NBA, y el tamaño de la cancha es distinto al de Europa, no podemos aplicarla a nuestra base de datos. Además, tampoco tenemos los datos del Play-by-play de la Liga Femenina Challenge. Sin embargo, haremos una excepción y mostraremos algunos gráficos de tiro utilizando la base de datos *PbP* proporcionada por la propia librería. Así, por ejemplo, con este código se muestran los porcentajes efectivos de tiro de Kevin Durant (jugador de los Golden State Warriors) en función de la zona de tiro. Además, los colores muestran la media de duración cada una de sus jugadas que acaban en tiro.

```
PbP <- PbPmanipulation(PbP.BDB)
subdata <- subset(PbP, player=="Kevin Durant")
subdata$xx <- subdata$original_x/10
subdata$yy <- subdata$original_y/10-41.75


shotchart(data=subdata, x="xx", y="yy", z="playlength",
          num.sect=5, type="sectors", scatter=FALSE,
          result="result")
```

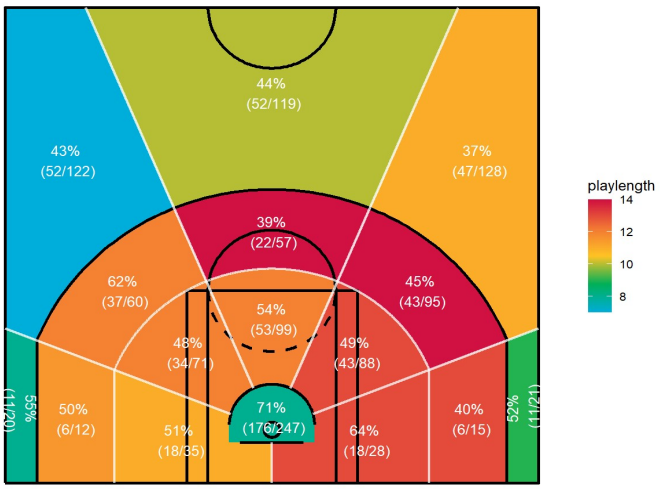

Figura 20: Porcentaje de acierto en tiros de Kevin Durant.

La función *shotchart* es una herramienta bastante potente a la hora de representar los datos de manera gráfica. Además, posee la facilidad de poder utilizarlo para cualquier conjunto de datos siempre y cuando se tengan las coordenadas de tiros e incluso de asistencias, por ejemplo. De hecho, se podría utilizar también para representar los tiros anotados y fallados de los jugadores/equipos en función de los clústeres a los que pertenecen, y así poder describir más el método de clasificación empleado.

7. Conclusiones

En este trabajo se ha puesto de manifiesto la utilidad del software R y de la librería *BasketballAnalyzeR* a la hora de realizar un análisis estadístico a nivel deportivo en el ámbito del baloncesto profesional.

La utilidad de la librería nos ha permitido sacar conclusiones acerca del estilo de juego de los equipos, así como de las estadísticas de muchas de las jugadoras de la Liga Femenina Challenge, aunque bien es cierto que las funciones propuestas se pueden aplicar para otros fines de interés. Es decir, las funciones están ahí para ser utilizadas, pero es trabajo del analista deportivo saber para qué se quieren utilizar y sobre qué variables y datos poder aplicarlas para sacar resultados óptimos y que sirvan de respuesta al equipo técnico de cada equipo.

En nuestro caso, cabe destacar la posibilidad que tiene la librería para realizar un anáisis de clústeres, lo que permite al analista encontrar diferencias y similitudes en su base de datos. Además, este tipo de agrupaciones no solo tienen sentido descriptivo, sino que también pueden servir de herramienta útil para, por ejemplo, contratar a jugadoras con unas características similares a un perfil en concreto.

Por otra parte, me gustaría dedicar unas líneas de las conclusiones a lo que se mencionó en el punto 5 de los aspectos claves de la Ciencia de Datos (Apartado 2). Y es que es cierto que disponer de una buena base de datos permite a los analistas poder llegar a conclusiones más precisas y, por tanto, más útiles.

Actualmente, la propia organización de la NBA toma las estadísticas de los partidos a través de un sistema que cuenta con barias cámaras en cada cancha, además de un software informático específico que permite realizar un seguimiento detallado de todos los jugadores y del balón 25 veces por segundo [26].

En comparación con España, las estadísticas aún se siguen tomando 'a

mano' por parte de la Federación Española de Baloncesto (FEB), lo que hace que la base de datos de la que disponemos sea más pobre y más susceptible a errores comparado con la de la NBA. En las páginas web de Estadísticas WNBA y Estadísticas LF Endesa se pueden ver claramente las diferencias en cuanto a recogida de datos entre la WNBA - Liga de Baloncesto Femenino Americano - y la Liga Endesa -Liga de Baloncesto Femenino Español-.

Es por ello que aún queda un largo camino para colocar la Ciencia de Datos en el Baloncesto en el lugar que se merece estar.

8. Continuación del trabajo.

Para terminar, me gustaría reiterar que el tema del TFM tiene también una parte vocacional. De hecho, ha sido mi primera toma de contacto entre la Estadística y el Deporte. En este caso, el hecho de haber trabajado en ello ha abierto una nueva 'pasión' que, a mi parecer, puede seguir dando nuevas herramientas que perfeccionen, cada vez más, el mundo del baloncesto español. Especialmente el femenino que, como es bien sabido, tiene menos recursos y se le dedica menos tiempo que al deporte masculino.

Como ampliación de este trabajo, a lo largo de estos meses han ido apareciendo dificutades a las que he tenido que enfrentarme de una forma particular. En general, el hecho de contar con un documento -proporcionado por el scouting de Raca- me ha permitido acceder a los datos sin necesidad de recurrir a la web (aunque reitero que sí que hubo mucho trabajo de limpieza y reestructuración de datos). Sin embargo, como me gustaría que este trabajo fuera útil y sirviera para algo, creo que sería muy conveniente poder utilizar una API [25] para poder realizar web scraping de cualquier partido de baloncesto y poder extraer así una base de datos útil. A partir de ahí, la idea sería utilizar nuevas funciones de R de la librería *dplyr* que estructuraran, de manera automática, la base de datos tal y como se quiere para poder utilizar las funciones de nuestra librería de *BasketballAnalyzeR*.

Por otra parte, hemos visto que la librería utilizada en este trabajo tiene funciones que muestran las gráficas de tiro de los jugadores a partir del Play-by-play de los partidos. No obstante, sabemos que dichas gráficas son las correspondientes al formato de juego de la NBA y, por tanto, no servirían para representar datos extraídos de partidos FIBA.

Así, otra forma de continuar con el trabajo sería construir unas funciones análogas a las de la librería utilizada con el fin de poder representar los gráficos de tiro a una base de datos en la que se juegue con las normas propuestas por la FIBA.

Finalmente, la idea sería subir los resultados en un repositorio de Git destinado a una comunidad abierta de analistas que permitiese realizar directamente el estudio estadístico sin la necesidad de perder el tiempo extrayendo, ordenando y limpiando los datos.

9. Glosario

- **DRtg**: Puntos permitidos por cada 100 posesiones.
- **ORtg**: Puntos producidos por cada 100 posesiones.
- **Box-score**: Es un resumen detallado de las estadísticas cuantitativas de un partido. En un box-score de baloncesto, se pueden encontrar datos referidos al número de lanzamientos intentados, lanzamientos que acaban en canasta, el número de tapones, rebotes, pérdidas o robos, etc.
- **Clasificación para los playoffs**: En una liga, se clasifican para playoffs los ocho primeros equipos de la clasificación.
- **FIBA**: Federación Internacional de Baloncesto. Es el organismo que se dedica a regular las normas del baloncesto a nivel mundial.
- **Play-by-play**: Es una base de datos donde se especifica cada una de las acciones de un evento deportivo. En baloncesto, un ejemplo de acción podría ser una asistencia de un jugador a otro, o un tiro a 3m de distancia, etc.
- **WNBA**: Del inglés, Women National Basketball Association. Es la liga femenina de baloncesto profesional estadounidense.

Referencias

[1] Ogando, P. G. (2019). *Las estadísticas avanzadas en el baloncesto.* Suma: Revista sobre Enseñanza y Aprendizaje de las Matemáticas, (91), 33-40.

[2] Walter Frick, 2015. Here's Why People Trust Human Judgment Over Algorithms. Here's Why People Trust Human Judgment Over Algorithms.

[3] Zuccolotto, P., & Manisera, M. (2020). *Basketball data science: with applications in R.* CRC Press.

[4] Severini, T. A. (2020). *Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports.* Chapman and Hall/CRC.

[5] Lewis, M. (2003). *Moneyball.* New York: W.W. Norton.

[6] Oliver, D. (2004). *Basketball on paper: rules and tools for performance analysis.* Potomac Books, Inc.

[7] Vázquez, A. (2 de Abril de 2020). Medium

[8] Manner, H. (2016). *Modeling and forecasting the outcomes of NBA basketball games.* Journal of Quantitative Analysis in Sports, 12(1), 31-41.

[9] Jones, E. S. (2016). *Predicting outcomes of NBA basketball games.* Doctoral dissertation, North Dakota State University.

[10] García, J. I. (2013). *Identifying basketball performance indicators in regular season and playoff games.* Journal of human kinetics, 36, 161.

[11] Sampaio, J., Janeira, M., Ibáñez, S., & Lorenzo, A. (2006). *Discriminant analysis of game-related statistics between basketball guards, forwards and centres in three professional leagues.* European journal of sport science, 6(3), 173-178.

[12] Schwarz, W. (2012). *Predicting the maximum lead from final scores in basketball: A diffusion model.* Journal of Quantitative Analysis in Sports, 8(4).

[13] Page, G. L., Fellingham, G. W., & Reese, C. S. (2007). *Using box-scores to determine a position's contribution to winning basketball games.* Journal of Quantitative Analysis in Sports, 3(4).

[14] Zuccolotto, P., Manisera, M., & Sandri, M. (2018). *Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions.* International journal of sports science & coaching, 13(4), 569-589

[15] Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008). *Action anticipation and motor resonance in elite basketball players.* Nature neuroscience, 11(9), 1109-1116.

[16] Skinner, B. (2010). *The price of anarchy in basketball.* Journal of Quantitative Analysis in Sports, 6(1).

[17] Clemente, F. M., Martins, F. M. L., Kalamaras, D., & Mendes, R. S. (2015). *Network analysis in basketball: Inspecting the prominent players using centrality metrics.* Journal of Physical Education and Sport, 15(2), 212.

[18] Skinner, B., & Goldman, M. (2017). *Optimal strategy in basketball.* In Handbook of statistical methods and analyses in sports (pp. 245-260). Chapman and Hall/CRC..

[19] The R Project for Statistical Computing.

[20] Álvarez, A. *R Packages: A Beginner's Guide. An introduction to R packages based on 11 of the most frequently asked user questions.* Tutorials, 2019.

[21] BasketballAnalyzeR

[22] Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). *A starting point for analyzing basketball statistics.* Journal of quantitative analysis in sports, 3(3).

[23] Bianchi, F., Facchinetti, T., & Zuccolotto, P. (2017). *Role revolution: towards a new meaning of positions in basketball.* Electronic Journal of Applied Statistical Analysis, 10(3), 712-734.

[24] Auguie, B., Antonov, A., & Auguie, M. B. (2017). *Package 'gridExtra'. Miscellaneous Functions for "Grid" Graphics.*

[25] API de Sergio Olmos para realizar web scraping

[26] How NBA Analytics is Changing Basketball.