TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Paquetes de R en el mundo del deporte: revisión y caso aplicado del uso de modelos estadísticos.

Autor: Víctor Martínez Rech

Directores: Martí Casals Toquero y Jordi Cortés Martínez

Departamento: Estadística e Investigación Operativa

Convocatoria: Julio 2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Agradecimientos

A mis directores, Martí Casals y Jordi Cortés, por haberme ayudado y acompañado a lo largo de este proyecto, por aconsejarme, animarme y dedicar tanto tiempo a la realización de este trabajo. Quiero agradecer también a Klaus Langohr sus consejos al inicio de este trabajo.

A mis amigos, un grupo bueno de verdad, que habéis estado a mi lado, acompañándome cada día y haciéndome vivir muchas nuevas y buenas experiencias.

Finalmente, a mi familia, tanto a los que hoy están día tras día, como a aquellos que fallecieron. Muchas gracias por el apoyo y la confianza que me habéis dado siempre, por creer en mí y darme fuerzas.

Resumen

Este trabajo muestra una aplicación de la estadística al mundo del deporte. En una primera parte, se ha realizado una revisión sistemática de los paquetes de R relacionados con el deporte, para observar el impacto de esta disciplina en la estadística, su crecimiento y evolución. En una segunda parte, se han aplicado técnicas de clusterización a un conjunto de datos de jugadoras de baloncesto profesionales usando métricas avanzadas para ilustrar una de las posibles aplicaciones de la estadística en este campo. Finalmente, se ha empleado un paquete de R para mostrar el potencial exploratorio de este paquete en la visualización de datos.

Palabras clave: PRISMA, R, *clustering*, *box-score*, *Basketball Analytics*

Abstract

This work shows an application of statistics to the world of sport. In the first part, we carried out a systematic review of the R packages related to sport has been carried out, to observe the impact of this discipline on statistics, its growth and evolution. In a second part, clustering techniques have been applied to a data set of professional basketball players using advanced metrics to show one of the possible applications of statistics in this field. Finally, an R package has been used to show the exploratory potential of this package in data visualization.

Keywords: PRISMA, R, clustering, box-score, Basketball Analytics

Clasificación AMS

62H30 Classification and discrimination; cluster analysis

Índice de contenidos

Índice de gráficos

Índice de figuras

Índice de tablas

1. Introducción

La estadística en ciencias del deporte y el fenómeno de *Sports Analytics*

La estadística es actualmente una disciplina bastante utilizada en el mundo de las ciencias del deporte, la cual ha despertado interés en diferentes profesionales como managers, analistas, video analistas, periodistas de datos, médicos, preparadores físicos, fisioterapeutas, psicólogos, nutricionistas, entrenadores o incluso en los propios jugadores. A pesar de este gran interés debería contemplarse, ante todo, que ésta es una ciencia joven que aprende desde los datos, y cuyo objetivo es medir, controlar y comunicar los mismos.

Estamos ante una nueva cultura orientada a los datos que exige cambios al modelo de gestión actual de las organizaciones. En el deporte, como en otros ámbitos, esto no será inmediato. De momento se han empezado a crear y renovar departamentos de analítica del deporte en los clubes, permitiendo acercar a diferentes profesionales de un mismo cuerpo técnico (analistas, preparadores físicos, fisioterapeutas, entrenadores) con estadísticos o científicos de datos [(Barça Innovation Hub, s.f.), (Casals M. &., 2017), (Casals M. B., 2017)].

El análisis cuantitativo, y por tanto la estadística, en el mundo del deporte es una rama de la ciencia que ha crecido y está creciendo exponencialmente en los últimos años. Este crecimiento se ha producido en parte gracias a la película de *Moneyball*, las constantes innovaciones tecnológicas disponibles actualmente, conferencias abiertas de análisis del deporte, o trabajos recientes de científicos de datos del deporte en este ámbito en revistas de estadística reconocidas por la *American Statistical Association* (ASA). El interés por la estadística deportiva tanto en la academia como en la industria deportiva ha coincidido con el *boom* y reconocimiento del estadístico y sobre todo del científico de datos como la profesión más sexy del siglo XXI según *Harvard Business Review*. La profesión de científico de datos o estadístico deportivo es solicitada ya en la industria del deporte en clubs profesionales de diferentes deportes de ligas de Estados Unidos e incluso ya también en Europa. En estas ofertas y oportunidades se piden normalmente habilidades de *statistical thinking* desarrolladas en los grados o másteres de estadística o afines, y también *computational thinking* con el programa estadístico R como uno de los más útiles y solicitados.

La evolución de R y Rstudio, y su influencia en el mundo del deporte

Personas de todo el mundo están recurriendo a aprender el programa estadístico R y otros lenguajes de programación de código abierto para dar sentido a los datos. Inspirado por innovadores en ciencia, educación, gobierno e industria, RStudio desarrolla herramientas

gratuitas y abiertas para R, y productos profesionales listos para la empresa para equipos que usan R y Python, para escalar y compartir su trabajo.

Como se trata de un software gratuito y de código abierto, y debido a sus constantes mejoras que permiten un espacio de trabajo más cómodo y práctico, R ha ganado mucha popularidad dentro de la computación estadística y de gráficos. Por ello, diferentes programas estadísticos (SPSS, Jamovi, JASP, Rcommander…) han reaccionado integrando la posibilidad de editar código R en sus programas [(R-Bloggers, s.f.)].

El programa R se encuentra en constante evolución, ya que los paquetes disponibles se actualizan con el paso del tiempo, además de crearse nuevos de forma dinámica. Los paquetes se pueden encontrar e instalar desde unos lugares de almacenamiento llamados repositorios de software. Para el lenguaje R, estos son *Comprehensive R Archive Network*, conocido como CRAN, Bioconductor o GitHub. CRAN es el repositorio más conocido y popular de R, ya que, por encima de otros repositorios, es de los pocos que realiza controles de forma rutinaria y hace comprobaciones regulares de los paquetes contribuidos.

Los paquetes disponibles en este repositorio pueden ser útiles para utilizar técnicas estadísticas para la estadística descriptiva, predictiva o causal, aparte de tener otras funcionalidades como por ejemplo el desarrollo de publicaciones, libros, aplicaciones, *dashboards* y presentaciones o *webscraping*.

En los últimos años, se ha apreciado la creación cada vez mayor de tutoriales, cursos, libros e incluso aplicaciones de paquetes relacionando el programa estadístico R con el deporte [(Jovanovic, 2019), (Andres, 2014), (Albert, Handbook of statistical methods and analyses in sports., 2017), (Downie, 2019)].

Más allá de una revisión: Las revisiones sistemáticas de los paquetes de R en las ciencias del deporte

Las revisiones sistemáticas (RS) tuvieron su origen en Medicina, en donde se definió el concepto como “un estudio de las evidencias que dan respuesta a una pregunta que ha sido claramente formulada por el revisor y que utiliza métodos sistemáticos y explícitos para identificar, seleccionar, y evaluar tales evidencias; así como extraer y analizar los datos proporcionados por éstas” [(Khan, 2001)]. La RS proporciona un resumen claro y estructurado de la información que se conoce acerca de un tema. Con la revisión sistemática se pretende reducir el sesgo en las distintas etapas de una revisión, por lo que es muy importante que estas se realicen de manera cuidadosa y detallada. En la RS pueden o no utilizarse métodos estadísticos, como el meta-análisis. Además de en Medicina, las RS se han aplicado durante

años en Educación, Psicología, Deporte y otros campos de las Ciencias Sociales y de la salud, para la generación de evidencias que soporten la toma de decisiones en su práctica profesional. Sin embargo, se cuenta con pocos referentes acerca de cómo conducir una RS en el campo de la ingeniería del *software,* la investigación metodológica, y menos aún en el campo de la estadística computacional [(Padua)]. Bien establecido en salud y ciencias sociales, las RS con metaanálisis recopilan toda la investigación existente sobre un tema específico y se consideran el nivel más alto de evidencia. Algunos primeros intentos de resumir la literatura metodológica existente surgieron incluso con un metaanálisis formal de métodos para la evaluación de un determinado tipo de software en biología computacional / bioinformática Gardner [(Gardner, 2019)], alguna revisión y evaluación de su calidad reportada de algún método estadístico concreto en alguna disciplina [(Casals M. G.-F., 2014)] o alguna RS reciente del rendimiento del aprendizaje automático, también conocido como *Machine Learning* versus regresión logística [(Christodoulou, 2019)].

Dada la evolución creciente del interés de la estadística en el ámbito del deporte, y también la evolución de la popularidad del programa estadístico R se ha creído oportuno hacer una RS de los paquetes de R disponibles relacionados con la estadística y el deporte.

2. Hipótesis del trabajo

Hasta la fecha se sabe:

- Que, con el crecimiento e interés de la estadística en el deporte, se piensa que cada vez hay un mayor número de paquetes de R relacionados con este ámbito en los repositorios como CRAN.
- Se desconoce hasta la fecha la realización de RS para evaluar paquetes de R relacionados con alguna disciplina como el deporte y evaluar características relacionadas con la optimización de los paquetes, metodología estadística e información relacionada con la investigación en las ciencias del deporte.
- Los distintos paquetes de R relacionados con el deporte disponen de diferentes estructuras de datos y funciones para iniciar preguntas de investigación y también aplicar o ajustar algún modelo.

3. Objetivos del trabajo

Los objetivos de este TFG son:

- Realizar una RS de los paquetes de R actuales relacionados con el deporte.
- Realizar una clusterización de jugadoras profesionales de baloncesto en función de sus características como muestra de una aplicación de la estadística en el ámbito deportivo.

Se pretende a través de estos objetivos mostrar una visión general de la evolución de la estadística deportiva desde el punto de vista académico teniendo en cuenta las habilidades computacionales y de *statistical thinking*. Otros objetivos secundarios de este trabajo son:

- Utilizar la guía Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) para llevar a cabo la RS (PRISMA, s.f.).
- Ver el patrón de las diferentes métricas avanzadas usadas en el baloncesto mediante técnicas de clusterización.
- Utilizar un paquete de R en el ámbito de visualización de datos del deporte como ayuda para la comprensión del análisis estadístico.

4. Revisión sistemática de paquetes de R relacionados con el deporte

Se ha realizado una RS de paquetes de R relacionados con el deporte. El estudiante y directores han trabajado la revisión sistemática paralelamente, realizando la búsqueda de paquetes de R en la base de datos de CRAN y haciendo distintas verificaciones para reducir el sesgo en cada etapa.

4. 1. Metodología y estrategia de búsqueda

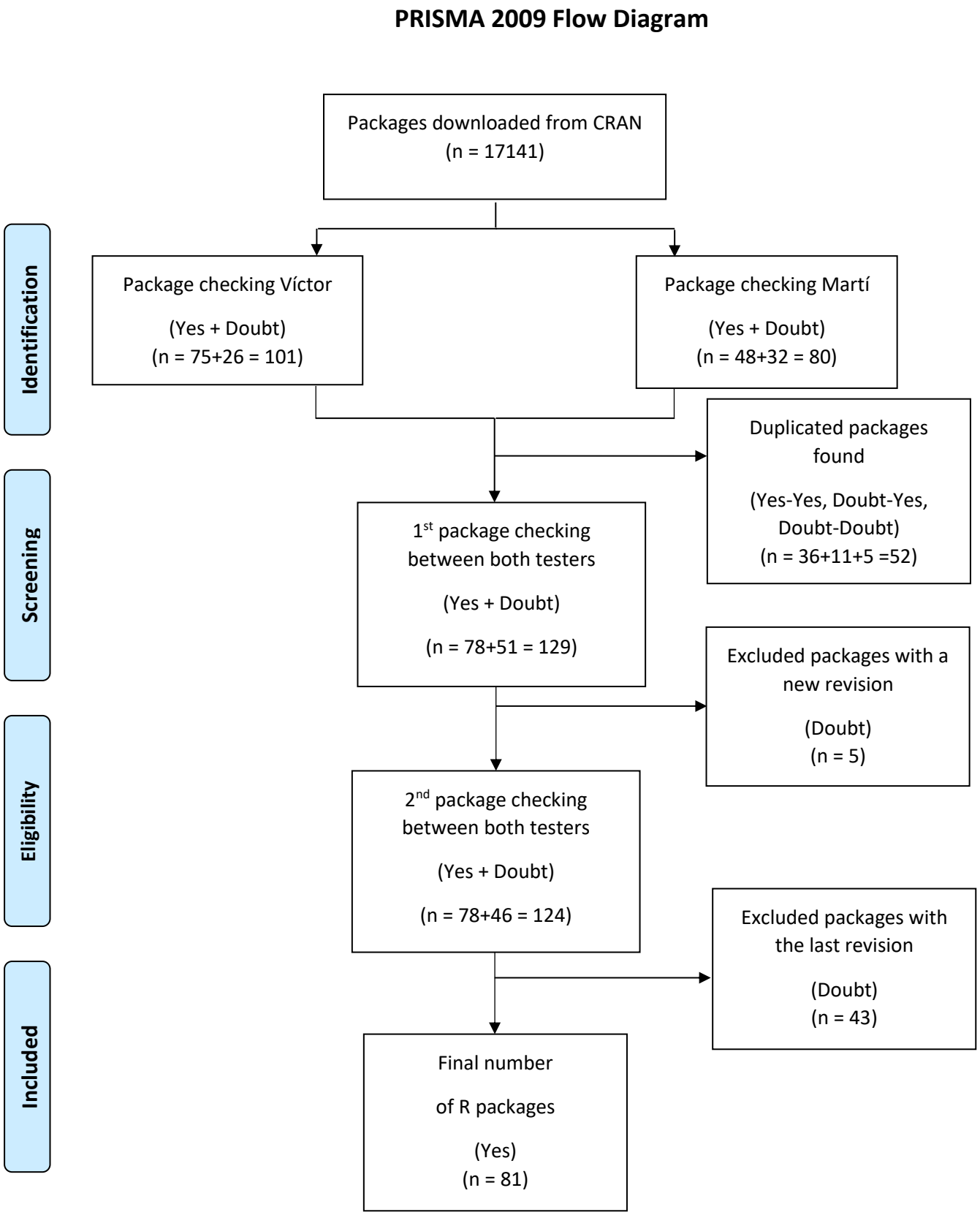
4.1.1. Selección de paquetes

R es un software de código abierto, es decir, un software cuya licencia permite el uso, modificación y distribución del mismo sin restricciones. Por este motivo, antes de empezar la revisión, se debe seleccionar el repositorio software con el que se accederá a los distintos paquetes de R. En este proyecto se ha decidido utilizar el repositorio CRAN, teniendo en cuenta todos los paquetes registrados desde el 15/03/2006 hasta el 18/02/2021.

4.1.2 Diagrama PRISMA

Para introducir y resumir todas las etapas del proceso se ha utilizado la declaración PRISMA, una propuesta para mejorar la publicación de revisiones sistemáticas y metaanálisis. En ésta, se incluye un diagrama de flujo PRISMA que describe todo el proceso, desde la identificación inicial de los estudios (en nuestro caso los paquetes de R) potencialmente relevantes hasta la selección definitiva.

Figura 4.1.: PRISMA: Diagrama de flujo de la revisión sistemática de paquetes de R



En primer lugar, una vez seleccionado CRAN como repositorio de dónde se extraerían los paquetes, se debía fijar una fecha que hiciera de límite para extraer los paquetes, ya que R actualiza y añade paquetes diariamente. Se escogió el 18 de febrero de 2021, y se obtuvieron 17141 paquetes en total. El alumno y uno de sus directores del TFG (MC) realizaron una primera búsqueda por separado, con métodos diferentes, dónde se seleccionaban los paquetes que tuvieran algún tipo de relación con el deporte. Solamente se comentó tener en cuenta la mayoría de deportes que figuraban en la web *topend sports* [(Topend Sports, s.f.)] y tener en cuenta las ligas profesionales más conocidas de diferentes deportes. Se encontraron un total de 101 y 80 paquetes respectivamente, clasificados en dos categorías (*Yes* para los paquetes con una relación directa con cualquier deporte y *Doubt* para los que había dudas sobre su temática). Entre éstos, había 52 paquetes que se repetían, por lo que en la primera revisión entre los dos ensayadores se seleccionaron 129 paquetes.

De esos 129, 78 (60.47%) se encontraban en la categoría *Yes*, mientras que los 51 (39.53%) restantes se deberán revisar con más detalle para saber si están relacionados con el deporte. Finalmente, después de dos revisiones entre los ensayadores (VM y MC) y una tercera con la ayuda del otro director (JC) y Klaus Langohr (profesor de estadística de la UPC), se decidieron incorporar 3 paquetes a los 78 ya seleccionados. De esta manera, el número total de paquetes de R relacionados con el deporte en la plataforma CRAN es de 81.

4.2 Explicación detallada de la revisión sistemática

La lista inicial de paquetes de R existentes en el repositorio CRAN que se consultó (n = 17141) el 18/02/2021 fue introducida en una hoja de cálculo, que inicialmente contenía 3 columnas: la fecha de publicación, el nombre del paquete y una pequeña descripción de éste. La hoja fue repartida al alumno y al director MC para que cada uno inicialmente hiciera una selección de los paquetes relacionados con el deporte. Cabe destacar que para poder hacer esto se añadió una cuarta columna a los documentos Excel, dónde se clasificaba el paquete según si tenía relación con el deporte (*Yes*: tiene relación con el deporte, *No*: no tiene relación con el deporte, *Doubt*: posible relación con el deporte. Revisión). Por último, comentar que hay algunas etiquetas de distintas variables que se encuentran en inglés, y se ha preferido no traducirlos ya que no existe una traducción fidedigna al castellano.

4.2.1. Clasificación inicial de los paquetes R del deporte

Uno de los directores del TFG (MC) encontró un total de 80 paquetes manualmente, sin ningún tipo de código o función, de los cuales 48 estaban clasificados con un *Yes* y 32 con un *Doubt*.

Para la búsqueda de los paquetes tanto el alumno como el director tuvieron en cuenta palabras clave relacionadas con el deporte (ligas, instrumentos, o palabras afines). Éstas concretamente podían ser tanto el nombre de deportes (*Football*, *Basketball*, *Soccer*), competiciones deportivas (NBA, MLB, NFL), eSports (LoL, Dota2, Fortnite), accesorios o instrumentos deportivos (*Accelerometer*, *Racket*, *Ball*), federaciones o canales de información (ESPN, FIBA, FIFA) u otras palabras también relacionadas con el deporte (*Sport*, *Cup*, *Game*). Una vez hecha esta búsqueda, se creó un *script* en R utilizando la función *findFn* del paquete *sos*, que se puede consultar en el apéndice de este trabajo. Esta función permite introducir un *string*, en este caso la lista de palabras clave, y busca en las páginas de ayuda de R aquellos paquetes que contienen esa palabra. Esta función crea un *dataframe* que muestra, entre otras cosas, el paquete, la función y una breve descripción de esta, el enlace de la página en la que se ha buscado, etc. Gracias a esta función, se redujo el número total de paquetes sobre los que buscar a 3051.

Aun así, el nombre de paquetes a revisar seguía siendo considerable, ya que en la lista se incluían palabras genéricas como *Running*, *Ball*, *Game*, *Cup* o *Score*, y al usar una función que lee *strings*, estas pueden encontrarse dentro de otras palabras o, por ejemplo, tener un significado distinto sin relación con el deporte.

Una vez reducido el nombre de paquetes, se revisaron estos manualmente. Al tener la lista de palabras clave, también resultó más sencillo encontrar un mayor número de paquetes. Se acabaron anotando un total de 101, de los cuales 75 se clasificaron con un *Yes* y 26 con un *Doubt*.

Entre el alumno y el director hubo una primera reunión para comparar los resultados obtenidos, mirar los paquetes duplicados (en total 52) y hacer una primera lista conjunta con todos los paquetes clasificados. En esta primera revisión se obtuvo un total de 129 paquetes sin duplicar. De estos, 78 se catalogaron con *Yes* y los 51 restantes con *Doubt*.

En posteriores revisiones se siguieron descartando algunos paquetes de manera consensuada entre los dos investigadores.

Para los paquetes que seguían clasificados con *Doubt*, se hizo una última revisión para determinar si finalmente se encontraban relacionados con el tema seleccionado. Para verificar los paquetes, se contó con la ayuda del otro director (JC) y de Klaus Langohr, y se añadieron 3 últimos paquetes a la lista de 78 ya escogidos, por lo que se resolvió que el número final de librerías era de 81 (Tabla 4.1.).

Tabla 4.1.: Lista de los 81 paquetes de R finales seleccionados

Date	Package	Title
17/02/2021	GGIR	*Raw Accelerometer Data Analysis*
16/02/2021	AdvancedBasketballStats	*Advanced Basketball Statistics*
15/02/2021	nflfastR	*Functions to Efficiently Access NFL Play by Play Data*
15/02/2021	runexp	*Softball Run Expectancy using Markov Chains and Simulation*
12/02/2021	BAwiR	*Analysis of Basketball Data*
10/02/2021	squashinformr	*Politely Web Scrape Data from SquashInfo*
08/02/2021	ffscrapr	*API Client for Fantasy Football League Platforms*
03/02/2021	combinedevents	*Calculate Scores and Marks for Track and Field Combined Events*
03/02/2021	yorkr	*Analyze Cricket Performances Based on Data from Cricsheet*
28/01/2021	icdpicr	*'ICD' Programs for Injury Categorization in R*
22/01/2021	PhysicalActivity	*Process Accelerometer Data for Physical Activity Measurement*
21/01/2021	Lahman	*Sean 'Lahman' Baseball Database*
13/01/2021	SwimmeR	*Data Import, Cleaning, and Conversions for Swimming Results*
13/01/2021	tashu	*Analysis and Prediction of Bicycle Rental Amount*
12/01/2021	fitzRoy	*Easily Scrape and Process AFL Data*
10/01/2021	arctools	*Processing and Physical Activity Summaries of Minute Level Activity Data*
08/01/2021	fflr	*Collect ESPN Fantasy Football Data*
07/01/2021	nbapalettes	*An NBA Jersey Palette Generator*
06/01/2021	NFLSimulatoR	*Simulating Plays and Drives in the NFL*
04/01/2021	aRbs	*Find Arbitrage Opportunities for Sports Matches*
04/01/2021	uncmbb	*UNC Men's Basketball Match Results Since 1949-1950 Season*
14/12/2020	retrosheet	*Import Professional Baseball Data from 'Retrosheet'*
04/12/2020	chess	*Read, Write, Create and Explore Chess Games*
29/11/2020	shorts	*Short Sprints*
23/11/2020	comperes	*Manage Competition Results*
19/11/2020	matuR	*Athlete Maturation and Biobanding*
14/10/2020	qqr	*Data from Brazilian Soccer Championship*
13/10/2020	gsisdecoder	*High Efficient Functions to Decode NFL Player IDs*
26/09/2020	cyclestreets	*Cycle Routing and Data for Cycling Advocacy*
14/09/2020	PhysActBedRest	*Marks Periods of 'Bedrest' in Actigraph Accelerometer Data*
05/08/2020	bigchess	*Read, Write, Manipulate, Explore Chess PGN Files and R API to UCI Chess Engines*
05/08/2020	fivethirtyeight	*Data and Code Behind the Stories and Interactives at 'FiveThirtyEight'*
26/06/2020	BasketballAnalyzeR	*Analysis and Visualization of Basketball Data*

25/06/2020	cherryblossom	*Cherry Blossom Run Race Results*
21/06/2020	ggsoccer	*Plot Soccer Event Data*
25/05/2020	nhlapi	*A Minimum-Dependency 'R' Interface to the 'NHL' API*
17/05/2020	PAutilities	*Streamline Physical Activity Research*
15/05/2020	scuba	*Diving Calculations and Decompression Models*
03/05/2020	trackeRapp	*Interface for the Analysis of Running, Cycling and Swimming Data from GPS-Enabled Tracking Devices*
19/04/2020	bysykkel	*Get City Bike Data from Norway*
06/04/2020	Anthropometry	*Statistical Methods for Anthropometric Data*
28/03/2020	cricketr	*Analyze Cricketers and Cricket Teams Based on ESPN Cricinfo Statsguru*
24/03/2020	eddington	*Compute a Cyclist's Eddington Number*
03/03/2020	comperank	*Ranking Methods for Competition Results*
01/03/2020	PlayerRatings	*Dynamic Updating Methods for Player Ratings Estimation*
25/02/2020	nhlscrape	*Scrapes the 'NHL' API for Statistical Analysis*
22/01/2020	teamcolors	*Color Palettes for Pro Sports Teams*
07/01/2020	sport	*Sequential Pairwise Online Rating Techniques*
05/10/2019	TouRnament	*Tools for Sports Competitions*
03/07/2019	socceR	*Evaluating Sport Tournament Predictions*
27/05/2019	piratings	*Calculate Pi Ratings for Teams Competing in Sport Matches*
20/05/2019	volleystat	*Detailed Statistics on Volleyball Matches*
19/05/2019	InjurySeverityScore	*Translate ICD-9 into Injury Severity Score*
15/05/2019	trackeR	*Infrastructure for Running, Cycling and Swimming Data from GPS-Enabled Tracking Devices*
29/04/2019	NBAloveR	*Help Basketball Data Analysis*
03/01/2019	bikeshare14	*Bay Area Bike Share Trips in 2014*
15/11/2018	HMMpa	*Analysing Accelerometer Data Using Hidden Markov Models*
24/08/2018	accelerometry	*Functions for Processing Accelerometer Data*
24/08/2018	mvglmmRank	*Multivariate Generalized Linear Mixed Models for Ranking Sports Teams*
13/06/2018	ROpenDota	*Access OpenDota Services in R*
11/05/2018	Observation	*Collect and Process Physical Activity Direct Observation Data*
16/03/2018	mlbstats	*Major League Baseball Player Statistics Calculator*
18/12/2017	SpatialBall	*Spatial NBA Visualization and Analysis*
11/09/2017	opendotaR	*Interface for OpenDota API*
15/06/2017	baseballDBR	*Sabermetrics and Advanced Baseball Statistics*
08/03/2017	NHLData	*Scores for Every Season Since the Founding of the NHL in 1917*
20/02/2017	colorr	*Color Palettes for EPL, MLB, NBA, NHL, and NFL Teams*
20/02/2017	pawacc	*Physical Activity with Accelerometers*

16/12/2016	acc	*Exploring Accelerometer Data*
30/10/2016	RDota2	*An R Steam API Client for Valve's Dota2*
30/08/2016	PASenseWear	*Summarize Daily Physical Activity from 'SenseWear' Accelerometer Data*
12/08/2016	bundesligR	*All Final Tables of the Bundesliga*
12/06/2016	engsoccerdata	*English and European Soccer Results 1871-2016*
29/02/2016	tmpm	*Trauma Mortality Prediction Model*
18/01/2016	cycleRtools	*Tools for Cycling Data Analysis*
09/12/2015	pitchRx	*Tools for Harnessing 'MLBAM' 'Gameday' Data and Visualizing 'pitchfx'*
05/11/2015	rchess	*Chess Move, Generation/Validation, Piece Placement/ Movement, and Check/Checkmate/Stalemate Detection*
23/10/2013	fbRanks	*Association Football (Soccer) Ranking via Poisson Regression*
07/04/2013	SportsAnalytics	*Infrastructure for Sports Analytics*
28/02/2013	heatex	*Heat exchange calculations during physical activity*
20/01/2011	darts	*Statistical Tools to Analyze Your Darts Game*

Una vez elegidos los paquetes, el alumno y los directores se reunieron para decidir toda la información que se podía añadir para crear un conjunto de datos con información relevante de cada uno de estos paquetes. Por ello, se creó una base de datos a partir de la tabla existente, añadiendo información relacionada con la expuesta en las tablas 4.2., 4.3. y 4.4..

Tabla 4.2.: Características generales de los paquetes

Variable	**Descripción**
Contiene tutorial de uso	Clasifica con *Yes* o *No* si el paquete contiene un tutorial de uso
Fecha de la 1ª versión	Columna que muestra la fecha de la primera versión del paquete
Número de la 1ª versión	Número de la primera versión del paquete
Fecha de la última versión	Fecha de la última versión del paquete
Número de la última versión	Número de la última versión del paquete
R journal	Indica con Yes o No si el paquete ha aparecido en un artículo de la revista *R Journal* (https://journal.r-project.org/)
Citación del paquete	Muestra información sobre cómo citar el paquete de R en base a publicaciones. Se ha usado la función de R *citation*
YYYY	Año de la última versión del paquete
YYYY2	Año de la primera versión del paquete

Tabla 4.3.: Características generales del deporte

Variable	**Descripción**
Deporte	Tipo/s de deporte/s que se trabajan en el paquete
Género	Género para el que se destina el paquete (*Male, female* o *Both*)
Categoría	Indica la categoría a la que pertenecen los jugadores que practican el deporte del paquete (*Professional, Amateur* o *Both*)
JQAS	Marca con Yes o No si el paquete ha aparecido en un artículo de la revista de deportes *Journal of Quantitative Analysis in Sport* (JQAS)
Clasificación de categorías	Clasifica el paquete en grupos dependiendo de su uso. Se ha determinado cinco grandes grupos: 1) *Sports Performance Analysis*: paquetes que analizan el rendimiento deportivo 2) *Sports technology*: paquetes que estudian la tecnología dentro de un deporte. 3) *Movement integration*: paquetes que analizan el movimiento en un deporte. 4) *Athlete health*: paquetes relacionados con la salud de los deportistas. 5) *eSports*: paquetes relacionados con los deportes electrónicos.

Tabla 4.4.: Datos y metodología estadística

Variable	**Descripción**
Incluye datos	Indica con Yes o No si el paquete incluye mínimo una base de datos cuando se instala. Para poder saberlo se ha utilizado la función data de R
Contiene metodología estadística de análisis	Clasifica con con Yes o No aquellos paquetes que se utilizan para el análisis de datos
Tipo de metodología	Muestra el tipo de técnicas estadísticas y metodología que utiliza el paquete

4.3. Resultados

En este apartado, se muestran los resultados de los análisis descriptivos de las variables descritas anteriormente referentes a los paquetes de R. De la misma manera que en el apartado anterior, se analizarán las variables según el grupo de información al que pertenecen.

Tabla 4.5.: Frecuencias de las variables por grupo

Características generales de los paquetes (n = 81)		**Frecuencia absoluta (%)**
Contiene tutorial de uso	Sí	35 (43,2%)
Actualización de los paquetes	Sí	53 (65,4%)
R Journal	Sí	3 (3,7%)
Características generales del deporte		**Frecuencia absoluta (%)**
Deporte	*American Football*	9 (11,1%)
	Baseball	7 (8,6%)
	Basketball	14 (17,3%)
	Cycling	9 (11,1%)
	Football	12 (14,8%)
	Hockey	5 (6,2%)
	Physical Activity	11 (13,6%)
	Running	5 (6,2%)
	Other	32 (39,5%)
Género	*Male*	38 (46,9%)
	Female	0 (0%)
	Both	43 (53,1%)
Categoría	*Amateur*	10 (12,3%)
	Professional	48 (59,3%)
	*NS**	23 (28,4%)
JQAS	Sí	2 (2,5%)
Clasificación de categorías	*Athlete Health*	4 (4,9%)
	eSports	3 (3,7%)
	Movement integration	9 (11,1%)
	Sports performance analysis	50 (61,7%)
	Sports technology	15 (18,5%)
Datos y metodología estadística		**Frecuencia absoluta (%)**
Incluye datos	Sí	56 (69,1%)
	No	22 (27,2%)
	Sí (depende de otros paquetes)	3 (3,7%)
Contiene metodología estadística de análisis	Sí	40 (49,4%)
Tipo de metodología	*Compute*	17 (21%)
	Descriptive	7 (8,6%)
	Prediction	8 (9,9%)
	Processing	14 (17,2%)
	Scrapping	43 (53,1%)
	Simulation	7 (8,6%)
	Visualization	8 (9,9%)
	Other	17 (21%)

Other:* En *other** se han puesto las categorías que tenían una menor frecuencia: *Arbitrage, Athletics, Australian football, Billiard, Chess, Cricket, Darts, Dota2, Injury Sports Medicine, Mahjong, Swimming, Scuba diving, Softball, Speedway, Squash, Tennis, Volleyball, Walking*.

*Other**:* En *other*** se han puesto las categorías que tenían una menor frecuencia: *Categorization, Classification, Decoding, Extraction, Modeling, Preprocessing, Probabilities*.

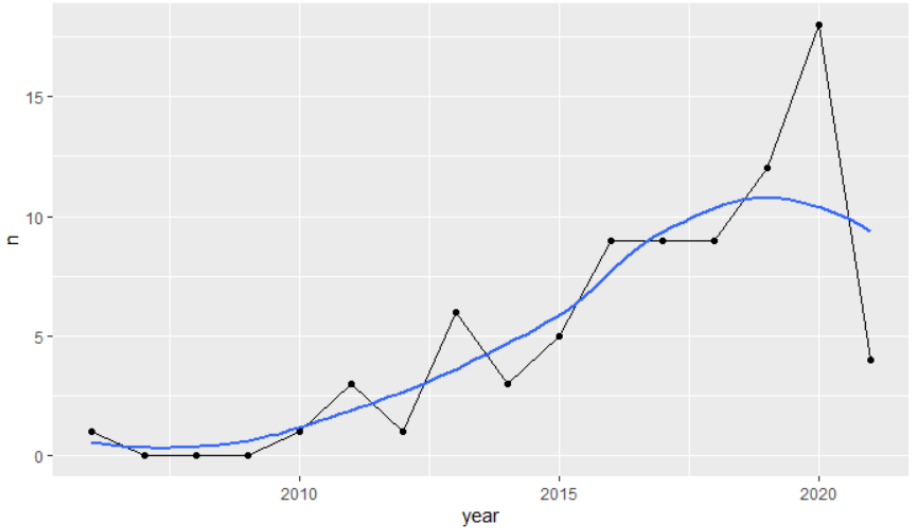
NS: Not specified*.

4.3.1. Características generales de los paquetes

Del total de 81 paquetes seleccionados, 46 no contienen tutorial de uso, que suponen un 56,8% de los paquetes totales, mientras que 35 (43,2%) utilizan *vignettes*. Únicamente 3 de los 81 paquetes (3,7%) se mencionan en artículos de *R journal*. Respecto a la versión y actualización de los paquetes revisados, se puede observar como 53 paquetes (65,4%) han sido actualizados una vez como mínimo. Por consiguiente, hay 28 paquetes (34,6%) que no han tenido una actualización (Tabla 4.5.).

En el gráfico 4.1. se observa la frecuencia de creación de paquetes de R relacionados con el deporte en el periodo (2006-2021). Se puede ver que el primer paquete fue creado en el año 2006, y no fue hasta el 2010 que se volvió a crear un paquete distinto. Luego, desde el año 2010 hay una tendencia creciente. Esto indica que, generalmente cada año, aumenta la creación de paquetes de R relacionados con el deporte. Finalmente, dada la tendencia creciente de la serie, se podría esperar que al final del año 2021 se hayan creado un mayor número de paquetes que en el año 2020. No obstante, se debe considerar que los paquetes de R creados hace más años podrían haber dejado de ser actualizados y, por tanto, no los habríamos detectado en nuestra revisión.

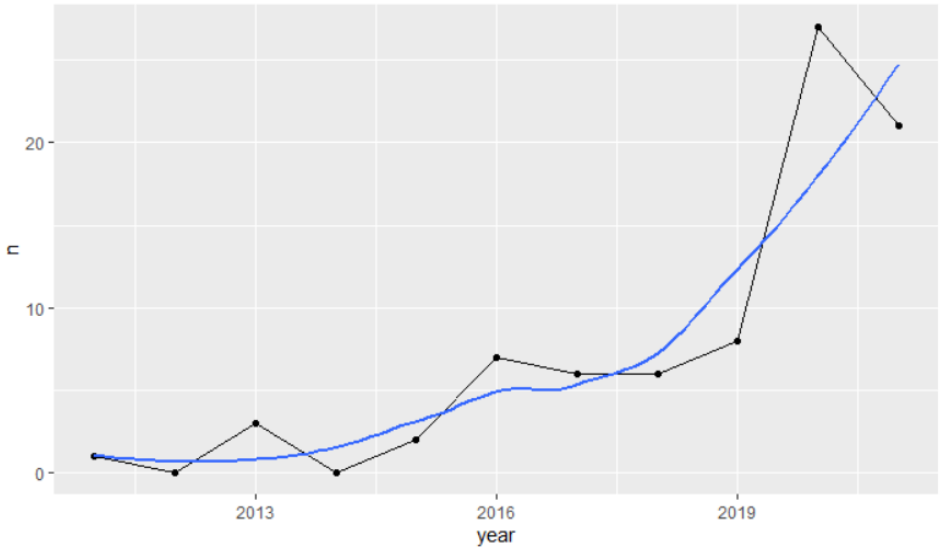
Gráfico 4.1.: Tendencia de creación de paquetes de R relacionados con el deporte por año



Nota: datos recogidos el 18/02/2021

En el gráfico 4.2. se muestra el año de la fecha de la última versión del paquete, es decir, el año de la última actualización de las librerías. De la misma forma que con el gráfico 4.1., se observa una tendencia creciente. Es decir, que con el paso de los años se realizan cada vez más actualizaciones de los paquetes. Por ejemplo, hay 48 paquetes relacionados con el deporte tienen fecha de actualización en 2020 y 2021, hecho que indica el reciente crecimiento del deporte en este campo.

Gráfico 4.2.: Año de la fecha de la última versión del paquete



Nota: datos recogidos el 18/02/2021

Por último, si nos fijamos en ambos gráficos se puede deducir que, aunque cada año se crean nuevos paquetes relacionados con el deporte, hay muchos de ellos que se actualizan con el paso del tiempo, ya que no se observa el mismo crecimiento exponencial en ambos gráficos. Aun así, se puede observar como con el paso del tiempo la creación de paquetes nuevos aumenta de una forma bastante lineal, y en los últimos años registrados la tendencia de la serie empieza a aumentar cada vez más.

4.3.2. Características generales del deporte

En primer lugar, me gustaría comentar que para trabajar con la variable *Deporte*, se han encontrado paquetes que guardan relación con muchos deportes distintos, y que por lo tanto tienen una respuesta múltiple. Por este motivo, se ha decidido tratar la variable creando *dummies* para cada deporte. Es decir, trataremos cada deporte como una variable binaria, que toma el valor 1 si el paquete trabaja con ese deporte o el valor 0 en caso contrario. Por este motivo, la suma total del número de disciplinas (n=104) supera el número de paquetes (n=81).

Como se puede observar en la tabla 4.5., el deporte con una mayor representación en los paquetes es el baloncesto con un total de 14 paquetes (17,3%). Lo siguen de cerca el fútbol con 12 (14,8%) y los paquetes relacionados con la actividad física con 11 (13,6%). También destacar el ciclismo y el fútbol americano con 9 observaciones cada uno (11,1%), y el béisbol con 7 (8,6%). Finalmente, destacar que hay deportes con poca representación, diez de ellos con una sola observación, que se han agrupado en la categoría *Other*, que cuenta con un total de 32 paquetes (39,5%).

Respecto al género, no hay paquetes deportivos dirigidos únicamente para el género femenino (Tabla 4.5.). Hay 43 paquetes (53,1%) enfocados para ambos géneros, mientras que los 38 restantes (46,9%) están destinados al deporte masculino. En cambio, la variable categoría cuenta con 48 paquetes (59,3%) que están hechos para ser utilizados en el ámbito profesional de un deporte, 23 paquetes (29,4%) que no especifican la categoría a la que va dirigido el deporte en cuestión, y 10 paquetes (12,3%) para el ámbito *amateur*.

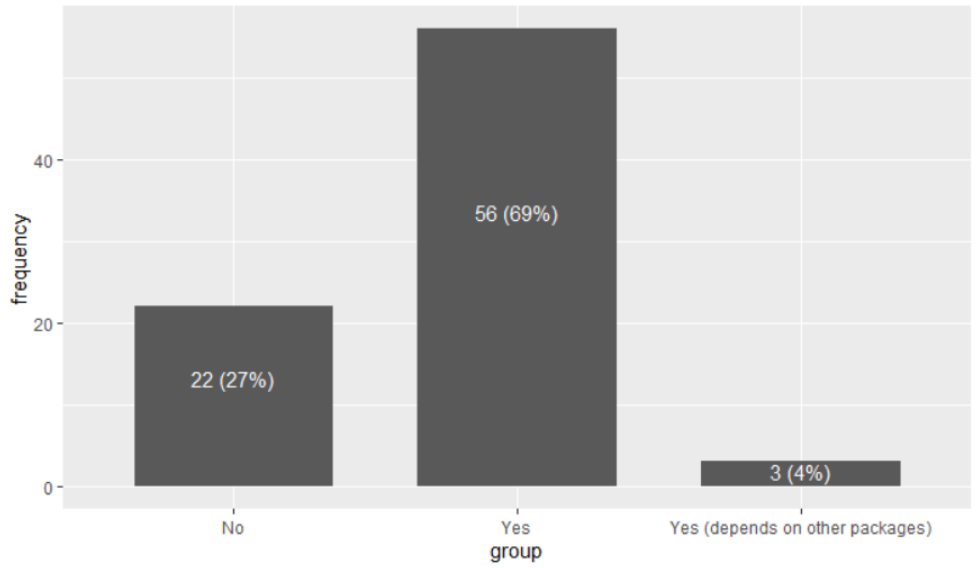
De la misma forma que con la variable *R Journal* (Tabla 4.5.), que también miraba qué librerías aparecían en artículos de dicha revista, se puede observar cómo los paquetes relacionados con el deporte tienen una representación casi nula en los artículos de la revista JQAS, ya que, de los 81 paquetes totales, solo 2 (2,5%) son mencionados en la revista, mientras que los otros 79 (97,5%) no aparecen.

Por último, la variable Clasificación de categorías cuenta con 50 paquetes (61,7%) que corresponden a la categoría *Sports performance analysis*. La siguen *Sports technology* con 15 paquetes (18,5%) y *Movement integration* con 9 (11,1%). Finalmente, las dos categorías con menor representación son eSports con 3 (3,7%) y *Athlete health* con 4 (4,9%), ya que solamente cuentan con un deporte de la variable *Deporte* que las represente (*Dota2* e *Injury_sports_medicine*, respectivamente).

4.3.3. Datos y metodología estadística

Del total de los 81 paquetes seleccionados, 56 de ellos (69,1%) incluyen base de datos, 3 paquetes (3,7%) utilizan bases de datos de otras librerías y los 22 paquetes restantes (27,2%) no tienen ninguna (Gráfico 4.3.).

Gráfico 4.3.: Frecuencias de la variable: Incluye datos



Respecto a la metodología estadística de análisis, se ha observado casi equidad en los resultados, ya que hay 41 paquetes (51,6%) que no contienen metodología (por ejemplo, paquetes que contienen colores de diferentes equipos para realizar gráficos o interfaces para acceder a datos) y 40 (49,4%) que sí que contienen. La clasificación de los paquetes para esta variable dependía de la descripción que se encontraba en el manual de referencia en CRAN. Si en dicha descripción se indicaba explícitamente que la librería contenía algún tipo de metodología para el análisis estadístico, se marcaba con un *Yes*, y en caso contrario, con *No.* Aun así, los paquetes podían contener cualquier tipo de metodología y que ésta no fuera indicada en la descripción del paquete.

Para la variable tipo de metodología, se ha indagado más en la función o distintas funciones que podía tener cada uno de los paquetes. Se ha visto que hay paquetes que utilizan más de una función estadística de análisis, es decir, que son observaciones que tienen una respuesta múltiple. Por consiguiente, se ha decidido tratar la variable creando *dummies* para cada categoría. Es decir, se trata cada categoría como una variable binaria, dónde el valor 1 corresponde a que el paquete utiliza esa metodología y el valor 0 indica el caso contrario. Por este motivo, la suma de tipo de metodologías (n=121) supera el número de paquetes (n=81).

El tipo de metodología que se encuentra en un mayor número de paquetes es *Scraping* con 43 (53,1%). Con una menor representación, le siguen *Compute or Wrangling Data* (calcular cualquier respuesta, función, estadístico, etc.) con 17 (21%) y *Processing* (procesamiento de datos) con 14 (17,2%). También se pueden destacar *Prediction* y *Visualization* con 8 observaciones cada una (9,9%), y *Descriptive* y *Simulation* con 7 (8,6%). Por último, las categorías que contaban con poca representación se han agrupado en *Other*, que cuenta con una frecuencia de 17 (21%).

También se quiso verificar la posible relación entre los paquetes que contienen tutorial e incluyen datos:

Tabla 4.6.: Tabla cruzada de frecuencias de las variables Contiene tutorial de uso e Incluye datos

		Contiene tutorial de uso		
		No	*Yes*	Total
Incluye Datos	*No*	15 (18,5%)	7 (8,6%)	22 (27,2%)
	Yes	30 (37%)	26 (32,1%)	56 (69,1%)
	Yes (depends on other packages)	1 (1,2%)	2 (2,5%)	3 (3,7%)
	Total	46 (56,8%)	35 (43,2%)	81 (100%)

Nota: los porcentajes representan el porcentaje sobre el total

En la tabla 4.6. se puede observar que ambas variables mantienen una cierta relación, ya que la gran mayoría de paquetes que tienen un tutorial de uso, también incluyen bases de datos. Concretamente, hay 26 librerías de 35 (74,3%) que tienen tutorial de uso y que incluyen datos de manera implícita. Luego, hay 7 librerías (20%) que no tienen datos y sí que tienen tutorial, y 2 (5,7%) que tienen tutorial e incluyen datos de otros paquetes. Por otra parte, de los 46 paquetes que no tienen *vignette*, 30 de ellos incluyen bases de datos (65,2%). Hay 15 (32,6%) que no tienen tutorial ni incluyen datos, y 1 paquete (2,2%) que contiene datos de otros paquetes, pero no tiene *vignette*.

Además, se encontró interesante estudiar la posible relación entre el género y la categoría de los deportes:

Tabla 4.7.: Tabla cruzada de frecuencias de las variables Categoría y Género

		Categoría			
		Amateur	*Professional*	NS	Total
Género	*Male*	2 (2,5%)	36 (44,4%)	0 (0%)	38 (46,9%)
	Both	8 (9,9%)	12 (14,8%)	23 (28,4%)	43 (53,1%)
	Total	10 (12,3%)	48 (59,3%)	23 (28,4%)	81 (100%)

Nota: los porcentajes representan el porcentaje sobre el total

En la tabla 4.7. se observa como todos los paquetes que no especifican la categoría del deporte se pueden utilizar para ambos sexos. Si la categoría es profesional, hay 36 paquetes de los 48 (75%) que son hechos para el género masculino, mientras que los 12 restantes (25%) son utilizados por ambos sexos. En cambio, cuando la categoría es amateur, 8 de los 10 paquetes en esta categoría (80%) son hechos para ambos sexos. Finalmente, si el género es masculino 36 de 38 paquetes (94,7%) se utilizan para una categoría profesional y los 2 restantes (5,3%) para un ámbito *amateur*. Mientras que con los paquetes hechos para ambos sexos la relación se encuentra algo más repartida entre los niveles de la variable categoría.

5. Clusterización aplicada a datos y métricas deportivas

Para la segunda parte del trabajo se presenta un caso práctico con datos reales con la evaluación del rendimiento deportivo como principal objetivo, y utilizando alguno de los paquetes revisados para la visualización de datos. Se ha decidido trabajar un deporte en concreto, el baloncesto. Para ello, mediante una base de datos de una liga profesional (Liga Endesa Femenina) se calcularán métricas conocidas por la literatura científica de *Basketball Analytics,* y posteriormente se hará un análisis estadístico utilizando la técnica *clustering* para conocer mejor el patrón de estas métricas en cada una de las jugadoras que conforman la liga con diferentes equipos. Aprovechando que en la primera parte se ha realizado una búsqueda de paquetes de deportes, también se usará el paquete *BasketballAnalyzeR*, fundamentalmente para el análisis exploratorio y visualización de los primeros resultados obtenidos. Todo ello se trabajará bajo la versión 4.1.0 del *software* R.

Como se ha comentado anteriormente, se usará una base de datos de la Liga Endesa Femenina de la temporada 2020-2021, que ha sido obtenida mediante la web *BueStats* (BueStats, s.f.).

BueStats

BueStats es una herramienta que permite la obtención y gestión de datos de todas las competiciones FEB (Federación Española de Baloncesto).

Esta web quiere proporcionar a los entrenadores una herramienta básica que extraiga automáticamente informes de estadísticas avanzadas, mediante factores condicionantes del juego y métricas fáciles de entender. Dado que el trabajo de los entrenadores incluye no solo la planificación de prácticas, sino también sesiones de video, reuniones individuales o incluso preparaciones físicas y de acondicionamiento, *BueStats* puede ser una herramienta muy útil para ellos.

La base de datos con la que se trabajará contiene los *box-scores* con los datos de los partidos de todas las jugadoras, ordenados por equipos y jornadas.

Box-score

El *box-score* es un resumen estadístico de los resultados de una competición deportiva. En este se anotan todas las puntaciones de cada una de las características de las jugadoras, y por tanto a nivel de equipo.

Los datos del *box-score* se obtienen de una hoja con las estadísticas de los jugadores y luego se resumen en una tabla de contingencia, como un conjunto básico de promedios. Esta información luego se correlaciona con un jugador o un equipo donde se lee para obtener una idea general de cómo se jugó el juego o cómo se desempeñó el jugador durante el juego, una temporada o su carrera.

Los deportes donde el *box-score* es utilizado con mayor frecuencia son: el béisbol, el baloncesto, el fútbol, el fútbol americano, el voleibol y el hockey.

En el baloncesto, los *box-scores* incluyen distintas métricas de este deporte, tanto básicas (puntos, rebotes, asistencias, tapones, etc.), como avanzadas (que se usarán más adelante en este trabajo).

Antes de empezar con el análisis estadístico, aprovecharemos uno de los paquetes hallados en la revisión sistemática.

5.1. Paquete *BasketballAnalyzeR*

BasketballAnalyzeR es un paquete, que se puede obtener de los repositorios CRAN y Github, que complementa el libro *Basketball Data Science – With Applications in R* (P. Zuccolotto, 2020).

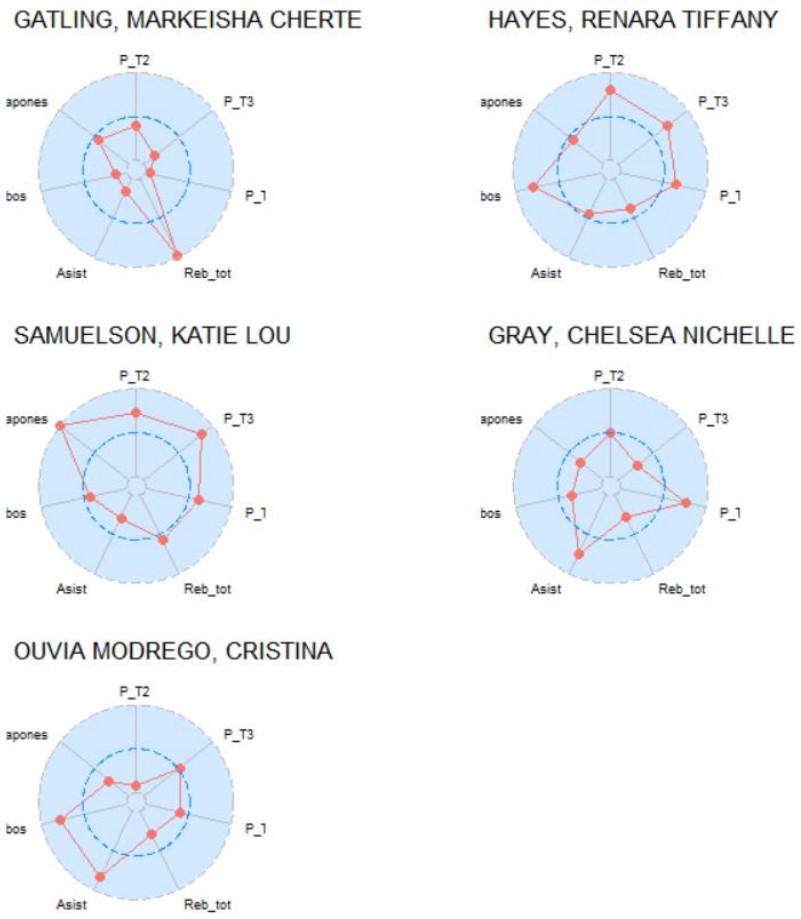
Este libro analiza una variedad de estudios de casos y ejemplos prácticos utilizando un paquete R personalizado usando datos de la temporada 2017-2018 de la NBA. Presenta herramientas para modelar gráficos y figuras para visualizar los datos, ya sea evaluando el rendimiento espacial de los tiros de un jugador o haciendo un análisis del impacto de situaciones de juego de alta presión en la probabilidad de anotar.

Los códigos de R para poder realizar los diferentes gráficos representan una herramienta muy útil para analizar datos baloncestísticos. Aunque las bases de datos que utiliza en sus ejemplos son de la NBA, los códigos son aplicables a cualquier liga profesional de la cual se tengan datos en formato *box-score*.

Por este motivo hemos decidido aprovechar la gran variedad de gráficos que se pueden obtener con este paquete para explorar el patrón y comportamiento del rendimiento de las jugadoras usando características y métricas de baloncesto. Más adelante, en el *clustering*, se usarán métricas más avanzadas conocidas en la literatura científica y del análisis del deporte.

Una vez acabada la temporada regular de la liga de baloncesto femenina, se elige el quinteto ideal de la temporada con las 5 mejores jugadoras. Este se selecciona a través de votaciones hechas por los fans, la prensa, las jugadoras y los entrenadores. En la temporada 2020-2021, el quinteto está formado por: Markeisha Gatling (jugadora del Casademont Zaragoza y MVP), Tiffany Hayes y Katie Lou Samuelson (jugadoras del Perfumerías Avenida), Chelsea Gray (jugadora del Spar Girona) y Cristina Ouviña (jugadora del Valencia B.C). El primer gráfico que se ha decidido hacer es un *radialplot*, que muestra las fortalezas y debilidades de las jugadoras, patrones y áreas con sus características básicas. Para este gráfico se han usado los porcentajes de 2 puntos, 3 puntos y tiro libre (1 punto), los rebotes totales, las asistencias, los robos y los tapones, todos ellos relativizados por minuto de juego.

Gráfico 5.1.: Radialplot del mejor quinteto de la temporada de la LF Endesa 2020-2021

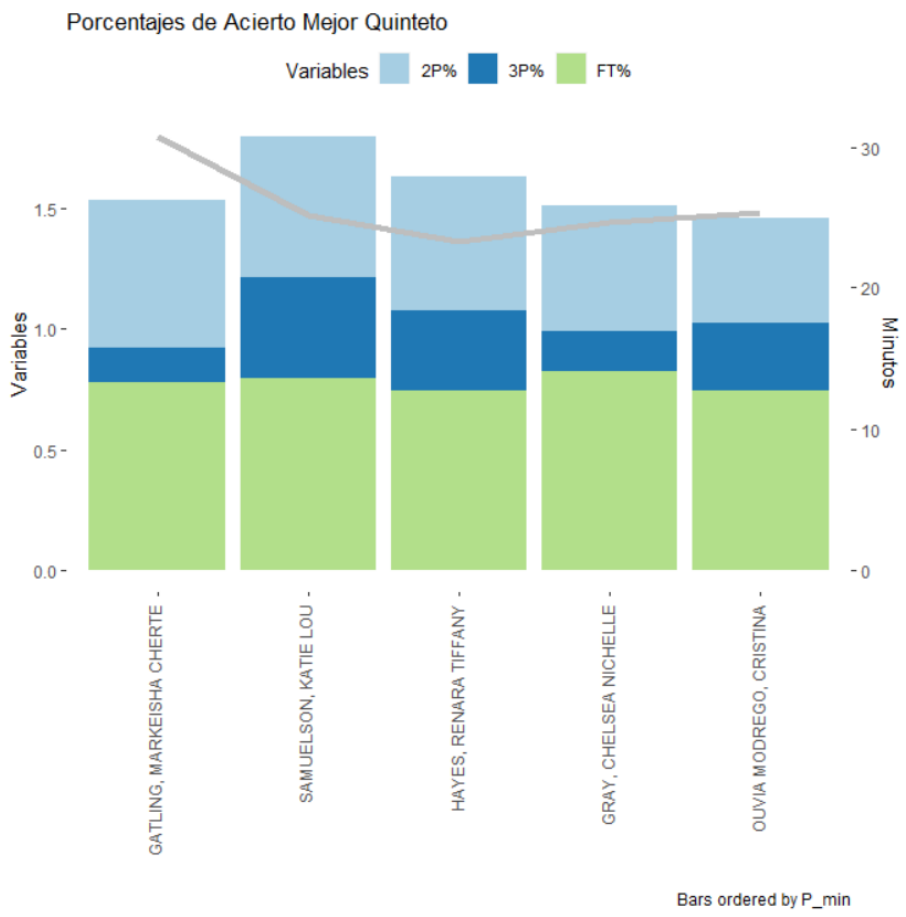


Se pueden observar las características básicas de cada jugadora y aquellas donde destacan. Cristina Ouviña juega en la posición de base, por lo que destacan sobre todo en las asistencias, ya que es la encargada de organizar el juego, y los robos. Chelsea Gray es una jugadora que juega en la posición de base y escolta, por lo que sus cualidades más destacadas son las

asistencias y el porcentaje en el tiro libre. Tiffany Hayes es una escolta y alero, por lo que en general tiene buen acierto en el tiro (se observa cómo su porcentaje en todos los tiros está por encima de la media), y realiza una buena cantidad de robos por minuto. Katie Lou Samuelson juega en la posición de alero y ala-pívot, de ahí que tenga una buena capacidad anotadora tanto desde el exterior como por dentro, además de hacer muchos tapones. Finalmente, Markeisha Gatling es una pívot, por tanto, su cualidad más destacada es el rebote. Además, está en la media en el porcentaje de tiro de 2 y en los tapones realizados por minuto.

El segundo gráfico que se ha decidido hacer es un *bar-line plot*, que permite comparar distintas características de las jugadoras, en este caso se han usado los porcentajes en el tiro de 2 puntos (2P%), 3 puntos (3P%) y tiro libre (FT%), además de poder contrastar estas características entre las jugadoras seleccionadas.

Gráfico 4.2.: Bar-line plot de los porcentajes de tiro del mejor quinteto de la temporada



En el gráfico 5.2. se encuentran ordenadas por puntos por minuto las jugadoras del mejor quinteto de la temporada 2020-2021 de la Liga Endesa Femenina. También se puede entrever una línea gris que marca los minutos por partido de cada jugadora. Me gustaría remarcar la

importancia de relativizar el rendimiento de las jugadoras por minuto, ya que es necesario tener en cuenta que cada jugadora no juega los mismos minutos ni tiene las mismas oportunidades.

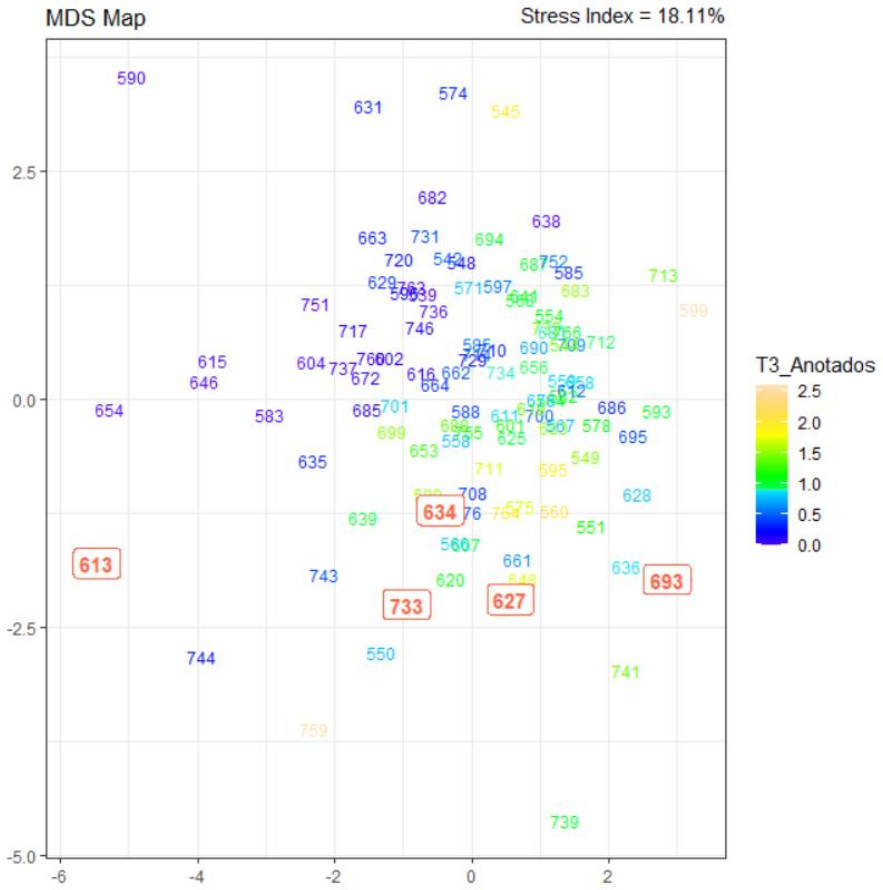
Se puede observar cómo Chelsea Gray es la jugadora con un mayor acierto desde la línea de tiro libre. Aun así, todas las jugadoras de este quinteto cuentan con un porcentaje más que decente desde esa posición.

Respecto al tiro de 3 puntos sí que se observa una mayor diferencia entre Katie Lou Samuelson, que tiene algo más del 40% de acierto, y las demás jugadoras.

Finalmente, la jugadora más dominante desde la pintura, Markeisha Gatling, es la que cuenta con un mayor porcentaje en tiros de 2 puntos.

Además de mirar estadísticas de manera individual, se puede usar un *MDS Map* (mapa de escalamiento multidimensional) donde se visualizan las métricas para ver similitudes entre jugadoras sobre el plano.

Gráfico 5.3.: MDS Map de las métricas básicas de las jugadoras que juegan un mínimo de 20 minutos por partido

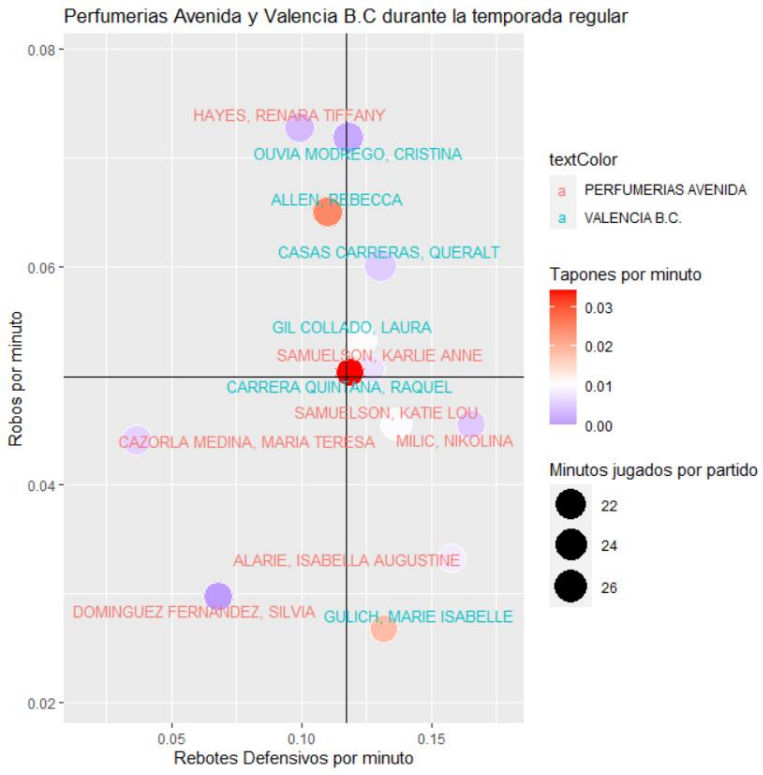


En el gráfico 5.3. se observan todas las jugadoras por su id en la base de datos, que juegan al menos 20 minutos por partido, sobre un plano bidimensional. En él se han usado las siguientes

métricas básicas: puntos, tiros de 3, tiros de 2, rebotes, asistencias, pérdidas, robos y tapones. También se puede interactuar con el gráfico, haciendo que marque las jugadoras que le indiquemos (en este caso se ha marcado con rojo a las jugadoras del mejor quinteto) o usando una escala de colores para mostrar la estadística que se quiera destacar, como, por ejemplo, los tiros de 3 anotados por partido. En este caso, las jugadoras representadas con un color amarillo claro son las que más tiros de 3 anotan por partido. Estas son: Laura Quevedo, Alexis Jones, Taylor Wurtz, Arica Carter, Angela Bjorklund, Frida Eldebrink, Maria España, Rebecca Allen, Victoria Vivians y Katie Lou Samuelson, que, aunque esté marcada de color rojo ya que es una de las jugadoras del mejor quinteto, anota una media de más de 2 triples por partido. Por último, me gustaría destacar que la mayoría de las jugadoras mencionadas se encuentran alrededor del punto (1,-1) del plano.

Con el siguiente gráfico, un *bubbleplot*, se pueden visualizar distintas estadísticas defensivas de las jugadoras de los dos mejores equipos de la temporada regular. Se trata de un *scatterplot* mejorado, ya que con este tipo de gráfico se pueden relacionar tres variables numéricas, por ejemplo, los rebotes defensivos por minuto, los robos por minuto y los tapones por minuto.

Gráfico 5.4.: Bubbleplot con las estadísticas defensivas de las jugadoras del Perfumerías Avenida y el Valencia B.C



Gracias al gráfico 5.4. se diferencian claramente las jugadoras de ambos equipos, y en qué métricas defensivas destacan. También me gustaría comentar que estas son características

muy resumidas del juego, y que sin contexto puede haber sesgo en las interpretaciones realizadas.

La jugadora que más rebotes defensivos por minuto captura es Nikolina Milic, o las jugadoras que más asistencias por minuto reparten, son las mencionadas en el gráfico 5.1., Tiffany Hayes y Cristina Ouviña. Respecto a los tapones, Raquel Carrera es la que hace más tapones por minuto de juego.

Antes de finalizar este apartado, me gustaría comentar que solo se han mostrado algunos de los gráficos y funciones que se pueden obtener gracias a este paquete. También se pueden usar funciones de *clustering*, modelización de los datos o gráficos personalizados e interactivos, por ejemplo.

Una vez hecha la visualización de los datos con los distintos gráficos que ofrece el paquete *BasketballAnalyzeR*, empezaremos a introducir las métricas avanzadas que se usarán para realizar el análisis clúster.

5.2. Métricas avanzadas en el baloncesto

Las estadísticas avanzadas generalmente llamadas en el baloncesto eran inicialmente métricas calculadas a partir de combinaciones lineales u cocientes entre diferentes características del juego, o bien teniendo en cuenta modelos de regresión con sus pesos respectivos. Estas estadísticas avanzadas han demostrado ser una herramienta crucial para los entrenadores y con la evolución de los sistemas de información y sobre todo con la nueva tecnología al alcance, se han creado departamentos de ciencia de datos altamente capacitados para manejar toneladas de información, como datos de seguimiento u otros algoritmos de visión computacional avanzados.

Sin embargo, los equipos europeos están muy por detrás en este campo, y el seguimiento de datos es completamente de otro mundo para el 97% de ellos. Lo peor de todo es que aquí no hay una tradición cultural de datos en el deporte, y se empieza a notar la necesidad de trabajar con estos (Arbues, s.f.).

Como se ha comentado anteriormente, la base de datos con la que se trabajará se trata de un box score de la temporada 2020-2021 de la Liga Endesa Femenina. Esta contiene una gran variedad de métricas de baloncesto, pero para el *clustering* se usarán las siguientes [(NBAstuffer, s.f.), (Basketball Reference)] (Para cada métrica se especifica el rango de valores dentro de nuestra muestra):

- **Puntos por minuto (Puntos/min):** puntos anotados por minuto de juego. Su rango de valores se encuentra entre 0.0434 y 0.655 puntos por minuto.

- ***Field Goal per minute*** **(FG/min):** tiros de campo anotados por minuto de juego. Incluye tiros de 2 y de 3 puntos. Los números obtenidos están entre 0.0133 y 0.274.

- ***Field Goal Attempts per minute*** **(FGA/min):** intentos de tiros de campo por minuto de juego. Incluye tanto intentos de tiros de 2 como de 3 puntos. Su rango de valores se encuentra entre 0.0365 y 0.517.

- ***Free Throws per minute*** **(FT/min):** tiros libres anotados por minuto de juego. Los datos muestran que las jugadoras anotan entre 0 y 0.131 tiros libres por minuto.

- ***Free Throw Attempts*** **(FTA/min):** intentos de tiros libres por minuto de juego. Los valores oscilan entre 0 y 0.211.

- ***Effective Field Goal Percentage per minute*** **(eFG%):** se trata de una métrica que mide la efectividad de los tiros de 2 puntos y de 3 puntos. Para esta variable, se han obtenido valores entre 0.00542 y 0.433. Su fórmula es:

$$eFG = \frac{[(All\ Field\ Goals\ Made) + 0.5 * (3P\ Field\ Goals\ Made)]}{(All\ Field\ Goal\ Attemps)}$$

- ***True Shooting Percentage per minute*** **(TSP/min)**: es una métrica que factoriza el rendimiento en la línea de tiros libres y considera la eficiencia de todos los tipos de tiros. Para esta métrica se pueden obtener valores por encima de la unidad. Con los datos trabajados, el TSP por minuto está entre 0.0576 y 5.423. Su fórmula es:

$$TSP = \frac{0.5 * (Total\ Points)}{[(Total\ Field\ Goal\ Attemps) + 0.44 * (Total\ Free\ Throw\ Attemps)]}$$

- ***Winscore per minute*** **(Winscore/min):** Métrica creada por David Berri que indica el valor relativo de los puntos, asistencias, rebotes, robos, pérdidas e intentos de tiros de campo. Se trata de una variable que puede obtener resultados negativos. En este caso, el valor mínimo obtenido es -0.516 y el valor máximo es 0.401. Su fórmula es:

$$Winscore = (Points) + (Rebounds) + (Steals) + \left(\frac{1}{2} * Assists\right) + \left(\frac{1}{2} * Blocked\ Shots\right) - (Field\ Goal\ Attempts) - (Turnovers) - \left(\frac{1}{2} * Free\ Throw\ Attempts\right) - (\frac{1}{2} * Personal\ Fouls)$$

Es muy útil para ver si un jugador juega mejor o peor que antes. Se basa en la predicción de victorias (*winsproduced*).

- **Valoracion_sin_puntos por minuto (Valoracion_without_P_min):** Se trata de una métrica que contabiliza el rendimiento de los jugadores, igual que el PIR (*Performance Index Rating*), pero sin tener en cuenta los puntos, los cuales muchas veces tienen una gran variabilidad en el rendimiento de un jugador. Igual que con la variable anterior, se pueden obtener valores negativos con esta métrica. El rango de valores va de -0.848 a 0.227.

- ***Player Total Contribution*** **(PTC/min):** Métrica para avaluar el rendimiento (producción) de los jugadores de baloncesto. Se basa únicamente en los datos de puntuación de box score validados por distintos procedimientos (Martínez, 2019).
  En definitiva, PTC es una métrica que se basa en los pesos de las variables del box score (rebotes, fallos, asistencias, pérdidas, robos, faltas personales, puntos, tapones, faltas recibidas) sobre como determinan la diferencia de marcador en un partido de baloncesto. En el PTC por minuto los números que se obtienen están entre -0.271 y 0.673.
  Su diferencia con el *winscore* es que éste último es una simplificación del *winsproduced* (que se basa en la predicción de victorias), que tiene un enfoque diferente de PTC. Aunque ambas variables tienen una alta correlación, pero parten de una base teórica distinta.

- ***Usage rate per minute*** **(UR/min):** métrica que calcula el volumen de uso de un jugador dentro de las jugadas ofensivas del equipo. También definido como "usage percentage is an estimate of the percentage of team plays used by a player while he was on the floor". Mencionar que, a veces, no tener la posibilidad de uso de la mayor parte de jugadas ofensivas en tus manos está relacionado en la producción (PTC) u otras características del juego. Esta métrica tiene valores entre 0.394 y 13.614.

$$UR = \frac{100 * ((Player's\ FGA) + 0.44 * (Player's\ FTA) + (Player's\ Turnovers)) * (Team's\ Total\ Minutes))}{((Team's\ Total\ FGA) + 0.44 * (Team's\ Total\ FTA) + (Team's\ Total\ Turnovers)) * 5 * (Player's\ Minutes)}$$

Cabe destacar que estas métricas son sobre todo ofensivas, igual que la mayoría de datos que se pueden encontrar en el box score, aunque hay algunas de ellas que utilizan tanto los datos ofensivos como defensivos de las jugadoras. Finalmente, como se muestra en las fórmulas, hay algunas que se calculan con métricas más sencillas (puntos, rebotes, asistencias, tapones, etc.).

Una vez definidas las métricas, empezamos a plantear el análisis clúster.

5.3. *Clustering*

El *clustering* consiste en la agrupación automática de datos. Con esta técnica avanzada se pretende agrupar los objetos en distintos subconjuntos llamados *clusters*. Cada grupo está formado por una colección de objetos o datos que se consideran similares entre sí, pero que poseen elementos diferenciales respecto objetos de otros *clusters*. Esta técnica se enmarca dentro del aprendizaje no supervisado, ya que trata datos sin etiquetar. Por este motivo es muy útil en los casos donde se quiere agrupar y tener conocimiento a un alto nivel de cómo se han generado los datos y como están organizados, sin tener conocimiento a priori de estos (Everitt, 2011).

Aplicando la técnica de *clustering* a nuestra base de datos se pretende agrupar a las jugadoras en grupos que correspondan a distintos perfiles de jugador; por ejemplo, un grupo de jugadoras que tienen un perfil anotador u otro que reúna características más defensivas. También es importante observar cómo afectan las métricas baloncestísticas a la creación y definición de los grupos.

El *clustering* es de utilidad en distintas disciplinas, como la medicina, la biología o el marketing, por lo que hay una gran cantidad de variantes y adaptaciones de sus métodos y algoritmos. Hay tres grupos principales:

- Agrupamiento no jerárquico: Para este grupo hay que especificar de antemano el número de *clusters* que se van a crear y las observaciones se asignan a los grupos según su cercanía (*k-means* o *k-medoids*).

- Agrupamiento jerárquico: No hay que especificar el número de grupos a priori, y puede ser aglomerativo o divisivo.

- Métodos que combinan o modifican los anteriores (k-means jerárquico, *clustering* basado en modelos, *clustering* basado en densidad o *clustering* difuso)

Aunque haya distintos métodos de agrupamiento, todos ellos tienen una cosa en común. Para poder llevar a cabo las agrupaciones necesitan definir y cuantificar la similitud de las observaciones. Para ello se usa el término distancia, ya que, si se representan las observaciones en un espacio *p* dimensional, siendo *p* el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán. La característica que hace del *clustering* un método adaptable a escenarios diversos es que puede emplear cualquier tipo de distancia (distancia euclídea, distancia de Gower, correlaciones lineales, correlación Jackknife, etc.).

En este apartado se usará el método *k-means* (Burkardt, 2009) usando la distancia euclídea como medida de distancia para realizar las agrupaciones.

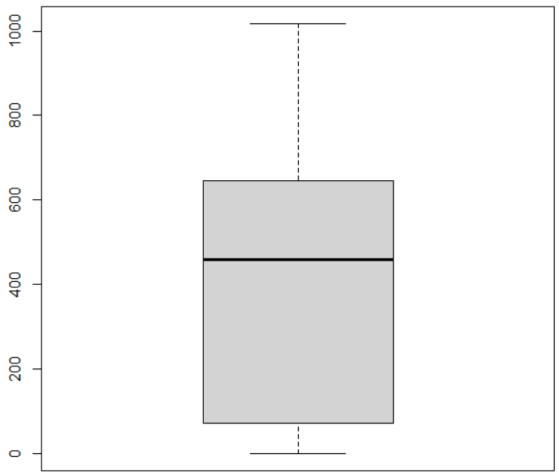
5.3.1. *Clustering* aplicado a la base de datos de las jugadoras de baloncesto de la LF Endesa

Antes de empezar con la clusterización, se han calculado las métricas baloncestísticas descritas anteriormente.

En ésta se observa cómo los datos se distribuyen según las jornadas de liga. Las jugadoras están ordenadas alfabéticamente por equipos, y se muestran las estadísticas por partido de cada una de ellas.

Para poder trabajar con la base de datos, antes se hace una pequeña descriptiva de los minutos disputados por cada jugadora.

Gráfico 5.5.: Boxplot de los minutos disputados por jugadora



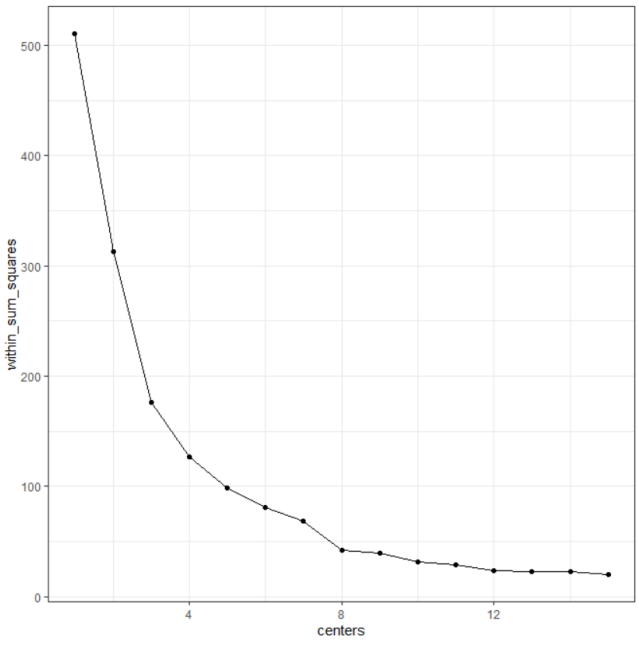
En el gráfico 5.5. se pueden observar los minutos disputados por las jugadoras en la temporada 2020-2021. Hay 16 equipos, lo que suponen un total de 30 partidos de fase regular. Los partidos de baloncesto se componen de 4 partes de 10 minutos cada una, por lo que, si una jugadora jugara el máximo de minutos posible, supondría un total de 1200 minutos. Se ha visto que el máximo de minutos disputados por una jugadora durante la fase regular de la temporada es 1016.4 minutos. También se observa que hay jugadoras que han jugado un número bajo de minutos durante la temporada, y esto hace que haya filas que no contengan datos a analizar. Por este motivo, se ha decidido filtrar arbitrariamente la base de datos por jugadoras que hayan jugado al menos 30 minutos, que supondría un mínimo de 1 minuto por partido. Con esto pasamos de tener 229 jugadoras a 187.

El siguiente paso, antes de usar el *k-means*, es coger los datos que se encuentran distribuidos por jornadas, y agruparlos por jugadoras. Para ello utilizaremos la función *mean* sobre cada una de las variables para cada jugadora durante todas las jornadas. Con esto obtendremos la base de datos con las métricas simples y avanzadas de la temporada de cada jugadora, por lo que ya podremos empezar el *clustering*.

En primer lugar, debemos fijar una *semilla* con la función *set.seed,* ya que el *clustering* hace agrupaciones que pueden variar dependiendo de la asignación aleatoria inicial de los centroides.

Como trabajamos con *k-means* (agrupamiento no jerárquico), debemos escoger el número de *clusters* a priori. Utilizaremos el método del codo para conocer el número de *clusters* óptimo. Este consiste en graficar la varianza explicada como una función del número de grupos y escoger el número de *clusters* a usar cuando la curva se dobla (codo de la curva).

Gráfico 5.6.: Gráfico del número de clusters a usar

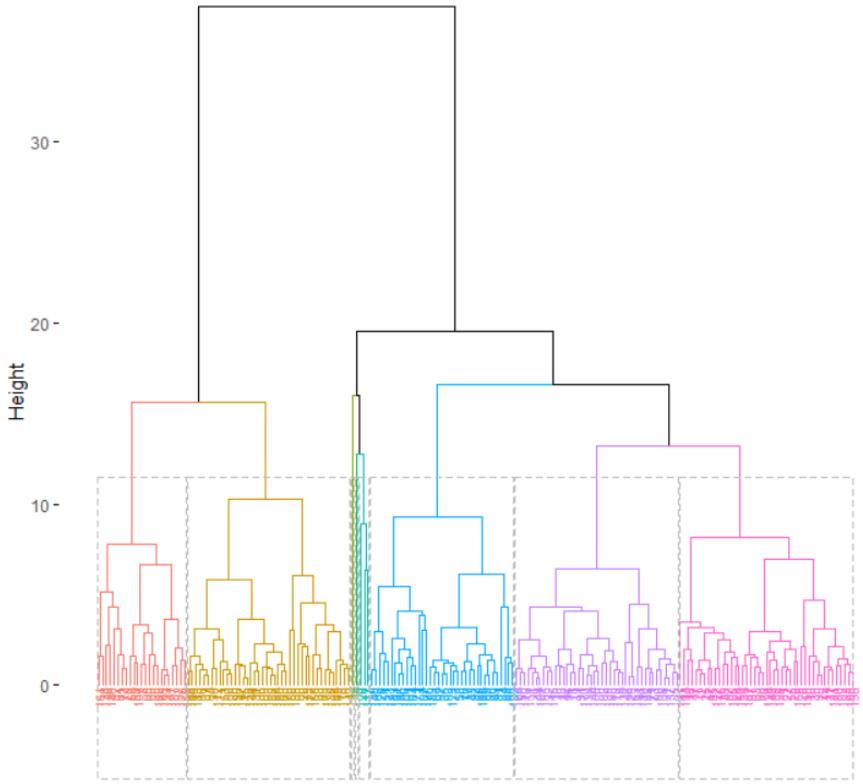


Como se puede observar, la curva se dobla al llegar a 8 *centers*, por lo que se ha decidido que este sea el número de *clusters* a usar en el *k-means*.

Una vez decidido el número de grupos, con la función de R *kmeans,* se crea un objeto que contiene, entre otras cosas, una matriz con las medias de las métricas por *clusters,* un vector que indica el *cluster* en el que se encuentra cada jugadora y el tamaño de los grupos, es decir, cuántas jugadoras hay en cada grupo. Gracias a esto se ha podido comprobar la varianza explicada con 8 grupos, que es del 91.5%.

Para empezar, se ha realizado un dendograma descriptivo, para ilustrar a gran escala como se van haciendo las subdivisiones o agrupamientos (Gráfico 5.7.).

Gráfico 5.7.: Dendograma del k-means

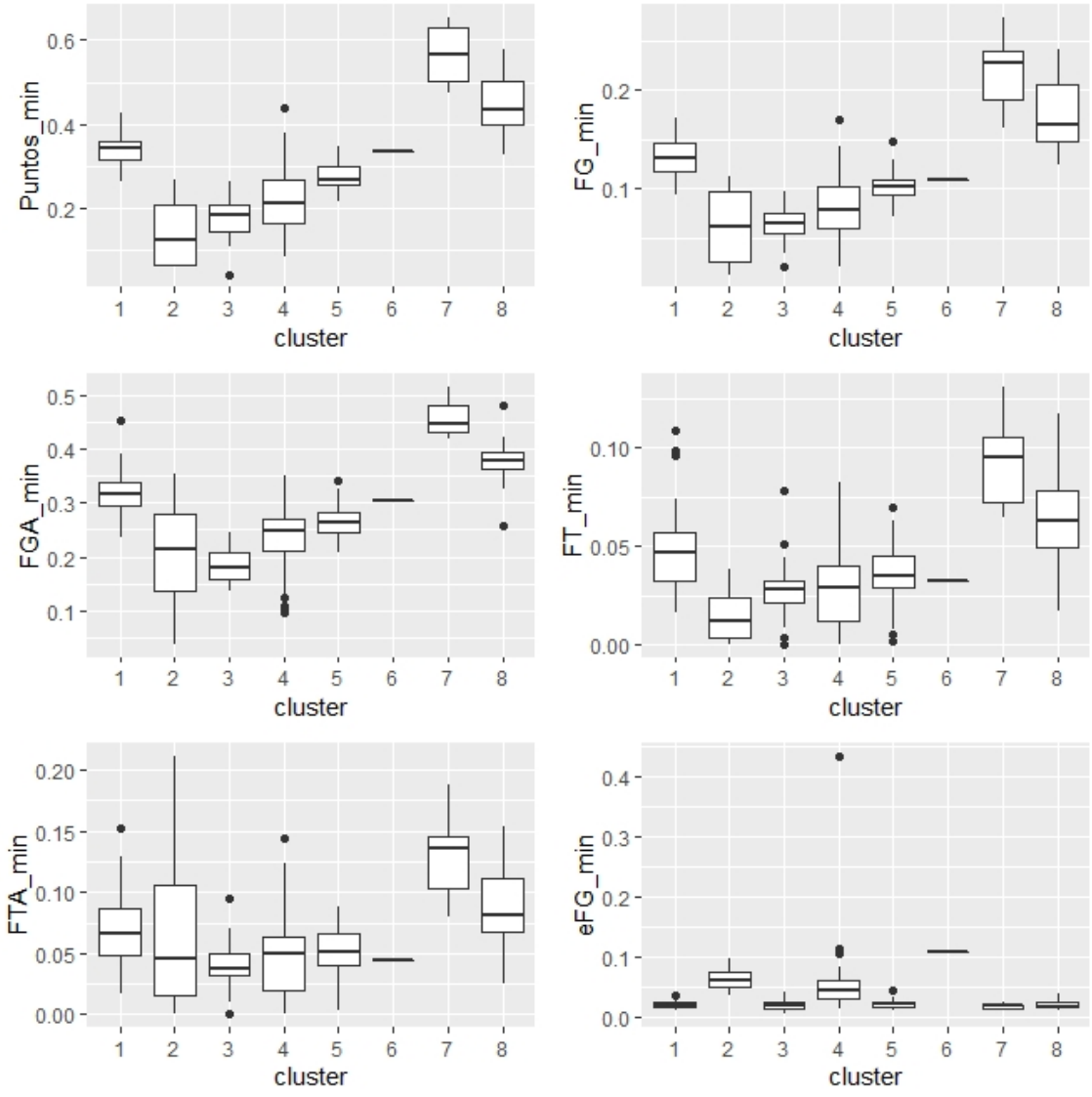


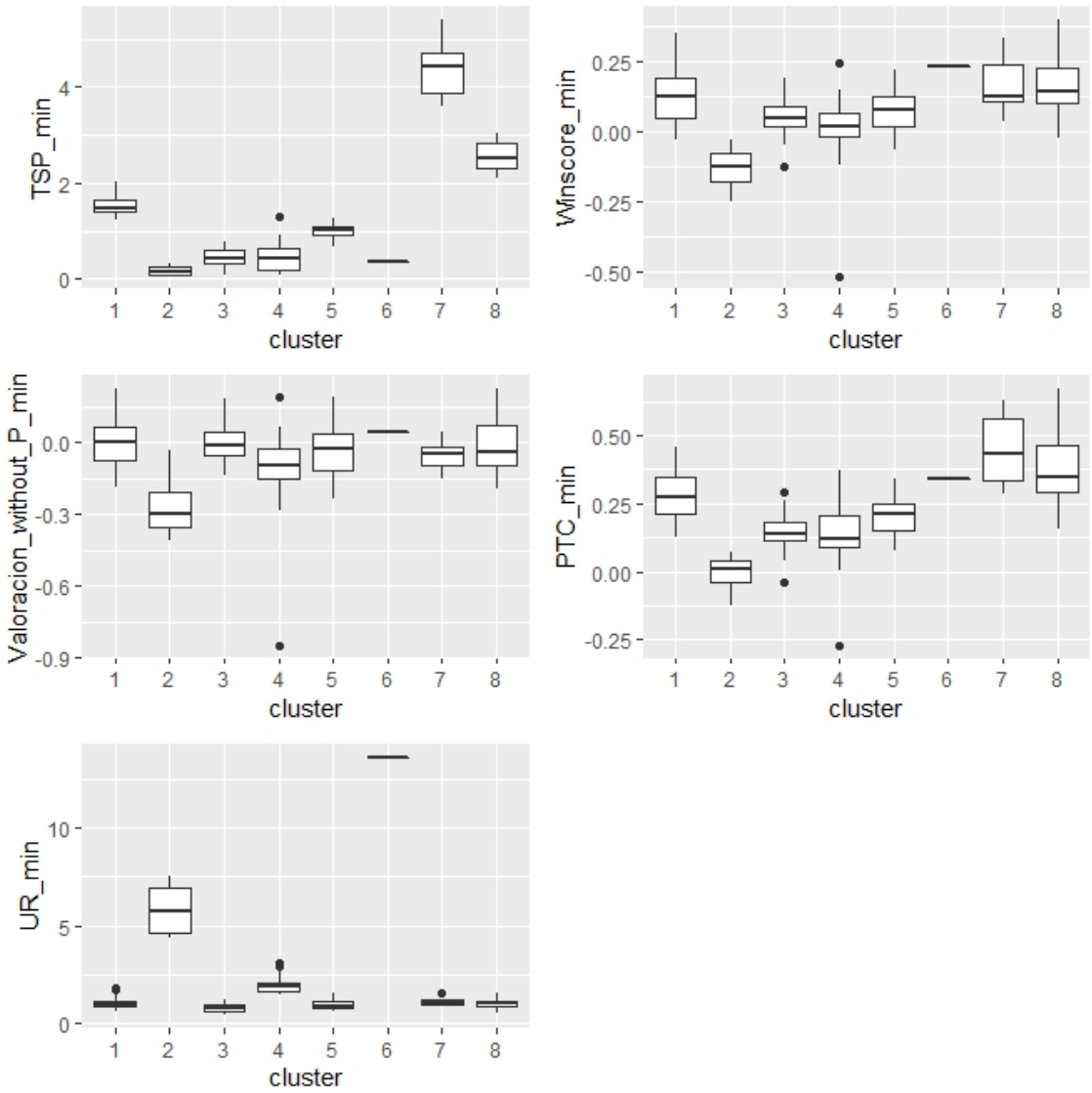
Se puede ver visualmente las particiones que se han hecho hasta obtener los 8 grupos. Se observa como en el centro del dendograma hay una rama que se divide en 3 partes, donde cada una contiene muy pocas jugadoras, por lo que se puede pensar que serán perfiles de jugadoras bastante únicos.

Es importante comentar que la división de las jugadoras en grupos no es equitativa, por lo que los *clusters* tienen tamaños diferentes entre ellos. El *cluster* 1 tiene 49 jugadoras, el 2 tiene 4, el grupo 3 cuenta con 30 jugadoras, el *cluster* 4 tiene 32, el 5 tiene 37, el 6 está formado por una única jugadora, el 7 tiene 9 y el grupo 8 cuenta con 32 jugadoras.

También se ha hecho un resumen visual de las distintas métricas usadas en el *clustering*, según los distintos grupos obtenidos. De esta manera podremos definir las características de cada *cluster* y hallar los distintos perfiles de jugadoras (Gráfico 5.8.).

Gráfico 5.8.: Distribución de las distintas métricas en cada clúster





Nota: el gráfico 5.8. se ha tenido que dividir en dos páginas por sus dimensiones

Con este gráfico se pretende observar algunos posibles perfiles de jugadoras antes de empezar a visualizar las métricas avanzadas una por una. A simple vista, se detecta que los mejores perfiles son los de los grupos 7 y 8. Estas parecen tener el perfil de ser jugadoras bastante anotadoras, pero también más productivas como determina la métrica PTC. Les siguen los *clusters* 1, 5 y 6, que parecen mostrar un perfil de jugadora bastante equilibrada. Finalmente, los *clusters* 2, 3 y 4, parecen mostrar el perfil de jugadoras con peor rendimiento.

Puntos por minuto

En referencia a esta variable, se puede observar como las jugadoras del *cluster* 7 anotan un mínimo de 0.5 puntos por minuto, y las jugadoras del *cluster* 8 un mínimo de 0.4 puntos por

minuto de media. Los grupos con peores anotaciones son el 2, el 3 y el 4. En estos también se pueden ver algunas observaciones atípicas. Finalmente, los grupos 1, 5 y 6, parecen tener un perfil de anotación por minuto medio, y sus valores oscilan entre 0.25 y 0.35 puntos por minuto.

Field Goal per minute

Se puede observar cómo los grupos mantienen perfiles bastante similares con la variable anterior, puesto que ambas variables se encuentran bastante correlacionadas, ya que cuantos más tiros de campo se anotan, más puntos se consiguen. De la misma forma que con la variable de puntos por minuto, las jugadoras del *cluster* 7 son las que anotan más tiros de campo por minuto, seguidas de las jugadoras del grupo 8 (que son los *clusters* que más destacan). El grupo 2 sigue siendo el que tiene un peor perfil anotador, ya que es el que menos tiros de campo anota. Los grupos 3, 4 y 5 tienen algún *outlier*. Por último, el *cluster* 1 parece ser el que más destaca dentro de los grupos menos anotadores.

Field Goal Attempts per minute

Se observa mucha más dispersión en los grupos que con las variables anteriores. El *cluster* 7 es el que hace más intentos de tiros de campo por minuto. Le siguen los grupos 8 y 1, pero en estos se observan valores atípicos. También se destaca el *cluster* 2 entre los grupos que menos intentos de tiros de campo hace, ya que abarca a jugadoras que intentan menos de 0.15 tiros de campo por minuto, a jugadoras que casi hacen 0.3 intentos de tiros de campo por minuto. Por último, comentar que los grupos 4 y 5 también presentan *outliers*.

Free Throws per minute

Para esta métrica se pueden observar unos rangos de valores menores que los anteriores. Esto se debe a la menor tendencia de lanzar tiros libres respecto a los tiros de campo. Aún así, el grupo 7 destaca como el que más tiros libres anota. Excepto los *clusters* 1 y 8, los otros grupos no tienen a ninguna jugadora que anote más de 0.05 tiros libres por minuto, siendo el grupo 2 el que menos tiros libres anota por minuto. Los grupos 1, 3 y 5 tienen valores atípicos.

Free Throw Attempts per minute

Si observamos los intentos de tiros libres por minuto, el *cluster* 7 sigue destacando como el grupo que más intentos realiza, aunque comparando la escala de valores con la de tiros libres

anotados por minuto, se puede ver que la mayoría de las jugadoras tienen un buen porcentaje de acierto en el tiro libre. Junto con el grupo 8, son los dos únicos *clusters* que tienen jugadoras que intentan una media de 0.1 tiros libres por minuto. Pese a no intentar muchos tiros libres por minuto, los grupos 1 y 5 parecen tener también un buen acierto en la línea de 4.6 metros. El *cluster* 2 destaca por su bajo acierto en el tiro libre, ya que por la cantidad de intentos de tiro libre que realizan, su anotación es bastante baja. Finalmente, los grupos 1, 3 y 4 tienen *outliers*.

Effective Field Goal Percentage per minute

Los grupos 1, 3, 5, 7 y 8 tienen un bajo porcentaje de tiros de campo efectivos respecto a los otros grupos. El *cluster* 6, que está formado por una jugadora, es el único con un valor mayor a 0.1. Los grupos 2 y 4 tienen un perfil medio, con valores que superan el 0.05. Los *clusters* 1, 4 y 5 tienen *outliers*.

True Shooting Percentage per minute

Se puede ver como el *cluster* 7 es el que cuenta con un TSP más elevado, con un valor por encima de 4 en casi todas sus jugadoras. Los siguientes grupos tienen valores bastante menores. El grupo 8 se encuentra en el rango de valores entre 2 y 3, seguido del grupo 1 que cuenta con valores entre 1 y 2. Los otros *clusters* tienen un TSP inferior a 1, a excepción de algunas jugadoras del *cluster* 5 y de una jugadora del grupo 4 que resulta ser un *outlier.*

Winscore per minute

Esta variable presenta valores positivos y negativos. El *cluster* con un mayor *winscore* es el 6, seguido muy de cerca por el 1, el 7 y el 8. Los grupos 3, 4 y 5 tienen un *winscore* cercano al valor 0. Me gustaría destacar que todas las jugadoras del *cluster* 2 presentan un valor negativo para esta variable, aunque también hay jugadoras de los grupos 3 y 4 (los únicos con *outliers*) con valores por debajo de 0.

Valoracion_sin_puntos por minuto

Dado que los puntos son uno de los mayores contribuidores en la valoración de las jugadoras, al no contar con estos, los valores que se obtienen son bastante bajos en general. Se observa cómo los únicos *clusters* con una media por encima del valor 0 son el 1 y el 6. Todos los demás tienen medias negativas. Todas las jugadoras de los grupos 2 y 7 tienen una valoración por

minuto negativa. Los otros *clusters*, en cambio, cuentan con jugadoras que tienen valoraciones con valores positivos y negativos.

Player Total Contribution

En el gráfico 5.8. se puede ver como el *cluster* 7 es el que cuenta con las jugadoras con una mayor contribución total. Lo siguen los grupos 8 y 6, que tienen todos sus valores por encima del 0.3. Los únicos grupos que contienen jugadoras con una contribución total negativa son el 2, el 3 y el 4, estos dos últimos a causa de valores atípicos.

Usage rate per minute

Para esta variable se pueden observar unos perfiles bastante parecidos a los de la métrica de porcentaje de tiro de campo efectivos por minuto, aunque se encuentren en una escala de valores distinta. El grupo con una mayor ratio de uso es el 6, con un valor por encima del 13. Bastante por debajo se encuentra el *cluster* 2, con valores entre 4 y 7. A excepción del grupo 4, los otros *clusters* presentan medias de valores por debajo del 2, y la mayoría de ellas se encuentran muy cerca del 1. Por último, los *clusters* 1, 4 y 7 tienen valores atípicos, por lo que hay que mirar a posteriori en detalle el contexto, por si se ven rendimientos muy buenos o malos respecto al resto en ciertas características del juego.

Una vez vistas las variables, se ha querido representar los *clusters,* en un plano de dos dimensiones, mediante componentes principales. También se muestran en una tabla la correlación de las métricas respecto a la 1ª y 2ª componente principal.

Gráfico 5.9.: Clusters definidos sobre las componentes principales 1 y 2

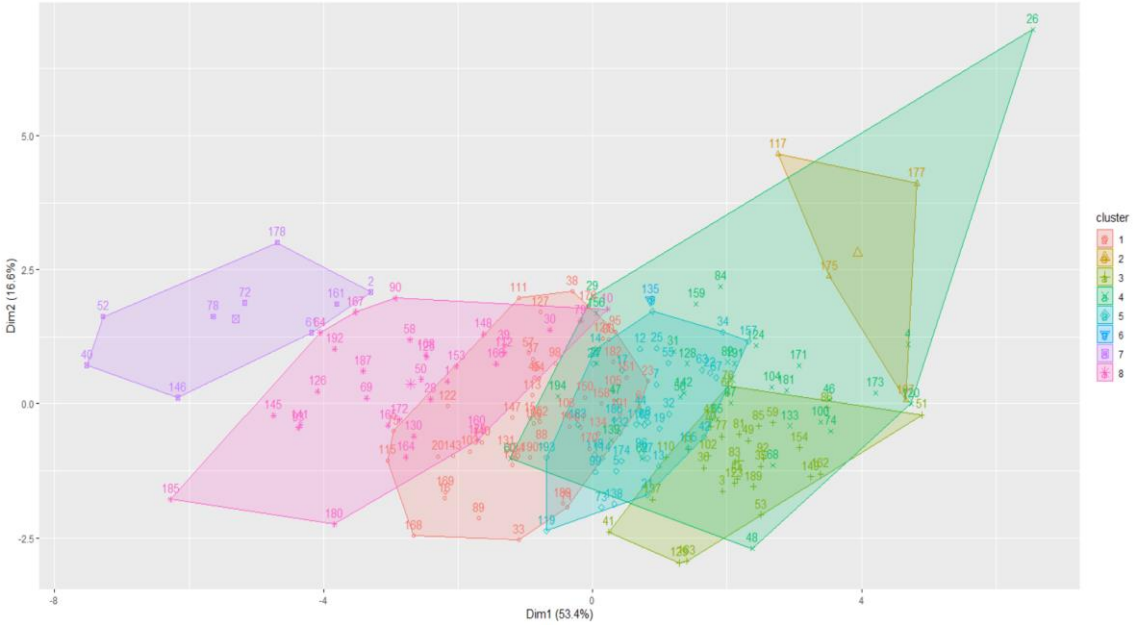


Tabla 5.1.: Correlaciones entre las métricas y la 1ª y 2ª componente principal

	PC1	**PC2**
Puntos_min	-0.9464243	0.22318031
FG_min	-0.9273027	0.19111318
FGA_min	-0.7949511	0.44095246
FT_min	-0.7908956	0.09288799
FTA_min	-0.6881376	0.18195192
eFG_min	0.2321888	0.09787325
TSP_min	-0.8913202	0.24932089
Winscore_min	-0.7115498	-0.62779334
Valoracion_without_P_min	-0.3829838	-0.88033789
PTC_min	-0.9187679	-0.28907431
UR_min	0.2055536	0.42664065

En primer lugar, se puede ver como la mayoría de variables tiene una fuerte correlación negativa respecto a la 1ª componente, a excepción de la valoración sin puntos, la ratio de uso y el porcentaje de tiros de campo efectivos. Por este motivo, las jugadoras que se encuentran más a la izquierda, son las que pertenecen a los *clusters* con unas mejores características. Para la segunda componente, solamente las variables *winscore* y valoración sin puntos muestran una fuerte correlación negativa, mientras que las otras variables, excluyendo la contribución total de la jugadora, muestran una correlación leve positiva. Por tanto, las jugadoras que se encuentren alrededor del valor 0 del eje de las ordenadas tendrán unas mejores estadísticas.

En el gráfico 5.9. se observan los *clusters* sobre el plano de las dos primeras componentes, que muestran un 70% de la varianza explicada. Comentar que los números que aparecen no son el identificador con el que se ha definido a las jugadoras en la base de datos, sino el número de fila en el que se encuentra cada jugadora dentro de la base de datos.

El grupo que más destaca es el 7, que es el que se encuentra más a la izquierda del plano, y el peor sería el 2 y un *outlier* del grupo 4.

5.3.2. Interpretación de los *clusters*

Centrándonos en los *clusters* de la temporada 2020-2021, se pueden observar distintos perfiles de jugadoras:

- *Cluster* 7: son jugadoras con una gran capacidad de anotación. Son las que aportan un mayor número de puntos por minuto, además de tener unos buenos porcentajes de tiro y tiro libre. También es muy importante destacar su alto PTC, que indica que son jugadoras muy productivas en general. Son también las que lanzan un mayor número de tiros de campo, además de ser las que juegan un mayor número de minutos. Su peor virtud se encuentra en la ratio de uso, la valoración sin puntos y el porcentaje de tiros de campo efectivos. Una jugadora de este *cluster* sería la MVP de la temporada, Markeisha Gatling.

- *Cluster* 8: son jugadoras con una buena capacidad de anotar puntos. Aportan un buen número de puntos por minuto, cuentan con buenos porcentajes en el tiro y tiro libre. Igual que con el grupo anterior, lanzan un buen número de tiros de campo, y sus peores métricas son la ratio de uso, la valoración sin puntos y el porcentaje de tiros de campo efectivos. En este grupo se encuentra la jugadora Tiffany Hayes.

- *Cluster* 6: este perfil solo pertenece a una jugadora, Begoña de Santiago, que aporta una cantidad decente de puntos por minuto. Lo más curioso de este perfil es que tiene un gran porcentaje de tiros de 3 y un mal porcentaje en los tiros de 2. Esto se podría explicar a partir de las variables porcentaje de tiros de campo efectivos y ratio de uso, dónde destaca muy por encima de los otros grupos.

- *Cluster* 1: parece mostrar un perfil de jugadora bastante equilibrado. Aporta un buen número de puntos, con una valoración sin puntos y *winscore* bastante buenos. Su TSP y contribución total también es bastante decente. Su mayor falla se encuentra en el porcentaje de tiros de campo, donde no destacan mucho. Es el *cluster* con un mayor número de jugadoras, como, por ejemplo, Marie Chatrice White.

- *Cluster* 5: también parece mostrar un perfil de jugadora más o menos equilibrado, aunque destaca menos que los del grupo 1. Cuentan con buenos porcentajes en tiros libres. También tienen un *winscore* y un TSP decente, aunque no destacan por su porcentaje de tiros de campo efectivos. En este *cluster* se encuentra la jugadora internacional española seleccionada en el draft de 2021 de la WNBA, Raquel Carrera.

- *Cluster* 3: es un grupo que cuenta con algún *outlier* para muchas de las variables trabajadas, por lo que habrá jugadoras que no se asemejen mucho al perfil definido por este grupo. No aportan muchos puntos por minuto, pero cuentan con porcentajes de tiro decentes. No destacan mucho en las otras métricas definidas. Una jugadora de este *cluster* es Alba Prieto.

- *Cluster* 4: sus jugadoras parecen definir un perfil bastante amplio, debido a los valores atípicos que muestran en casi todas las métricas. Además, es el *cluster* más extenso en el plano bidimensional de las componentes. Generalmente, no aportan muchos puntos, y además no cuentan con buenos porcentajes de tiro, sobre todo exterior. Tienen un porcentaje de tiros de campo efectivos y una ratio de uso decentes, comparado con los otros grupos. Una jugadora destacada de este grupo es Sarah Imovbioh.

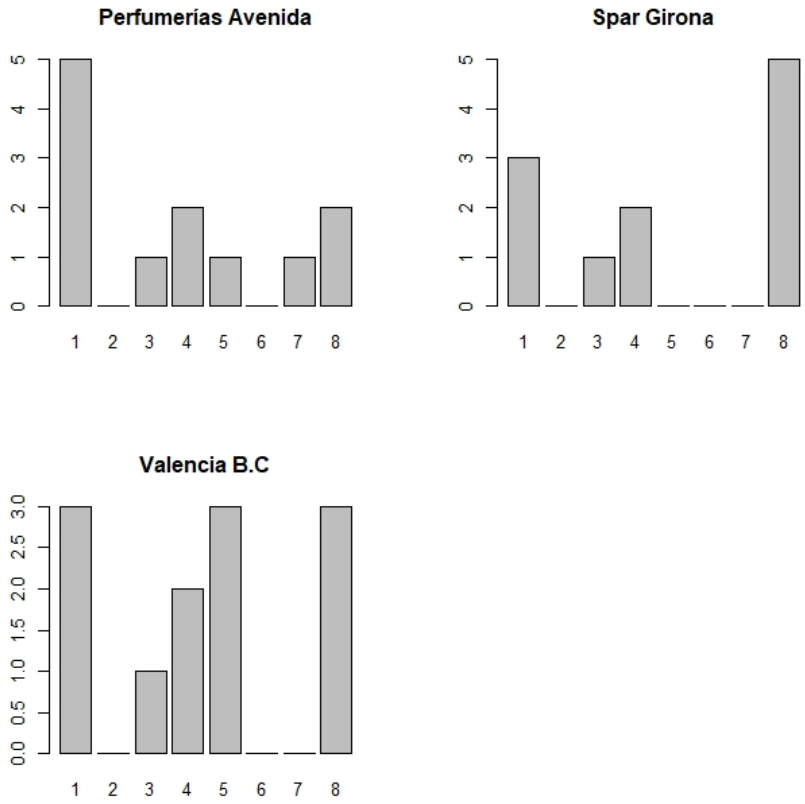
- *Cluster* 2: este grupo es el que aporta menos puntos por minuto, además tiene una efectividad baja en el tiro de campo. Destaca en la ratio de uso y en el porcentaje de tiros de campo efectivos, pero es el peor grupo en variables como la contribución total, la valoración sin puntos o el *winscore*. Una jugadora de este *cluster* es Sokhna Adji Fall.

5.4. Conclusiones

Este tipo de análisis puede tener distintas utilidades en las competiciones de baloncesto. Por ejemplo, cuando una jugadora ficha por un equipo, el entrenador puede encargar al departamento de *scouting* del club fichar una jugadora de su mismo *cluster*, ya que tendrían perfiles parecidos. En caso que un equipo busque reforzar su plantilla con una jugadora que destaque en distintas cualidades (anotación desde la línea de 3 puntos, capacidad defensiva, etc.) pueden buscar jugadoras con estas características utilizando técnicas de clusterización.

Como antes se ha comentado, las jugadoras del *cluster* 7 son las que más anotan. Pero se trata de un grupo con pocas jugadoras. De los tres mejores equipos de la fase regular de la Liga Endesa Femenina, solo el Perfumerías Avenida tiene una jugadora de este *cluster*. Por este motivo, aunque en el grupo 7 se encuentren las mejores anotadoras, los equipos necesitan de otros grupos con también buena capacidad de anotación.

Gráfico 5.10.: Número de jugadoras por cluster en los equipos: Perfumerías Avenida, Spar Girona y Valencia B.C



Como se observa en el gráfico 5.10., los *clusters* con una mayor presencia en estos clubes son el 1 y el 8, dos *clusters* que agrupan a jugadoras con una buena capacidad de anotación.

Hemos visto como la estadística puede ayudar a extraer información de un conjunto de datos con información agregada basada en *box-scores*. Otros conjuntos de datos con información de los eventos a lo largo del partido (*eventing data*) o con información posicional de las jugadoras a lo largo del partido en cada instante (*tracking data*) hubiese permitido realizar análisis más complejos. Por otra parte, los paquetes de R relacionados con el ámbito del deporte proporcionan una gran versatilidad, sobre todo en cuanto a la visualización de datos, que puede ser muy útil a nivel técnico/táctico para los equipos.

Por último, me gustaría añadir que este trabajo se continuará desarrollando en un futuro. Debido al tiempo consumido por la revisión sistemática de los paquetes, no se ha dispuesto del tiempo necesario para realizar un análisis estadístico a mayor escala. Por este motivo, se seguirá analizando distintas métricas baloncestísticas dentro de la Liga Endesa Femenina. Uno de los objetivos futuros de este trabajo es contactar con algún entrenador de algún equipo de la LF Endesa y presentarle un análisis que le pueda servir de ayuda en la planificación de la temporada para añadir contexto baloncestístico al análisis.

6. Bibliografía

Adèr, H. J. (2008). Advising on Research Methods: a consultant's companion. (J. v. Publishing, Ed.) *Modelling*, 271-304.

Albert, J. G. (n.d.).

Albert, J. G. (2017). *Handbook of statistical methods and analyses in sports.* CRC Press.

Andres, A. (2014). Sabermetrics 101: Module 4, R and RStudio. *SABR101x*.

Arbues, A. (n.d.). *BueStats*. Retrieved from https://www.upf.edu/web/adria-arbues/buestats

Barça Innovation Hub. (n.d.). Retrieved from https://barcainnovationhub.com/es/bioestadistica-ciencias-del-deporte/

Basketball Reference. (n.d.). Retrieved from https://www.basketball-reference.com/about/glossary.html

BueStats. (n.d.). Retrieved from https://github.com/arbues6/BueStats

Burkardt, J. (2009). K-means clustering. Virginia Tech, Advanced Research Computing, Interdisciplinary Center for Applied Mathematics.

Casals, M. &. (2017). Sports Biostatistician: a critical member of all sports science and medicine teams for injury prevention. Injury prevention. *BMJ*.

Casals, M. B. (2017). Infographic: Sports Biostatisticians as a critical member of all sports science and medical teams for injury prevention. *BMJ*.

Casals, M. G.-F. (2014). Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000–2012): a systematic review. *PloS one*.

Christodoulou, E. M. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 12-22.

Downie, T. (2019). Analyzing Baseball Data with R. *Journal of Statistical Software*, 1-4.

Everitt, B. L. (2011). *Cluster analysis.*

Gardner, P. P. (2019). Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ*.

Jovanovic, M. (2019). *Statistical Modelling for Sports Scientists: Practical Introduction Using R (part 1).*

Khan, K. S. (2001). *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews.*

Martínez, J. A. (2019). A more robust estimation and extension of factors determining production (FDP) of basketball players. *International journal of physical education, sports and health*, 81-85.

NBAstuffer. (n.d.). Retrieved from https://www.nbastuffer.com/analytics-101/player-evaluation-metrics/

P. Zuccolotto, M. M. (2020). *Basketball Data Science - With Applications in R.*

Padua, A. G. (n.d.). *Propuesta de un proceso de revisión sistemática de experimentos en Ingeniería del Software.*

PRISMA. (n.d.). Retrieved from http://prisma-statement.org/

R-Bloggers. (n.d.). Retrieved from What's the Best Statistical Software? A Comparison of R, Python, SAS, SPSS and STATA: https://www.r-bloggers.com/2019/07/whats-the-best-statistical-software-a-comparison-of-r-python-sas-spss-and-stata/

Topend Sports. (n.d.). Retrieved from https://www.topendsports.com/sport/list/index.htm

7. Apéndice

7.1. Justificación del trabajo

El motivo por el que he decidido realizar este trabajo se debe a la atracción que he tenido siempre por el mundo del deporte, tanto practicando distintos deportes, como siendo un mero espectador de estos. Gracias a la estadística he podido descubrir la modelización y sus aplicaciones, con la que he podido realizar un acercamiento a los deportes de una manera analítica.

Con este trabajo pretendo repasar conceptos ya aprendidos en el grado de Estadística y mejorar mis capacidades con estos; utilizar nuevos programas y técnicas no estudiados en la carrera; y finalmente conocer la aplicación de los modelos en el mundo del deporte.

7.2. Palabras clave utilizadas en la revisión de paquetes

Deportes:

- Football
- Basketball
- Soccer
- Tennis
- Volleyball
- Handball
- Water polo
- Hockey
- Baseball
- Martial
- Judo
- Sumo
- Taekwondo
- Karate
- Aikido
- Jujutsu

- Kendo
- Capoeira
- Athletics
- Running
- Jumping
- Climbing
- Swimming
- Throwing
- Walking
- Cycling
- Riding
- Surfing
- Racing
- Diving
- Archery
- Rugby
- Badminton
- Cricket
- Croquet
- Softball
- Chess
- Futsal
- Golf
- Squash
- Polo
- Paddle
- Gymnastics
- Alpinism
- Skiing
- Wrestling
- Boxing
- Calisthenics
- Curling
- Javelin
- Marathon
- Darts

Ligas/Competiciones:

- NBA
- NFL
- MLB
- NHL
- MLS
- ACB
- Liga
- League (Euroleague and Premier League)
- Ligue1
- Bundesliga
- Calcio
- SerieA
- F1
- MotoGP
- Nascar
- Champions
- UEFA
- Olympics

eSports:

- LoL (League of Legends)
- CS:GO (counter strike: global offense)
- Fortnite
- FIFA
- NBA2k
- PUBG (player unknowns battleground)
- COD (call of duty)
- Minecraft
- Valorant
- Dota2
- WoW (World of Warcraft)
- GTA (grand theft auto)
- Final Fantasy
- Skyrim
- Pokémon
- Sims

- Clash Royale
- Apex (Apex Legends)

Otras palabras:

- Sports
- Player
- Accelerometer
- Ball
- Team
- Cup
- Score
- Tournament
- Kayak
- Canoe
- Racket
- FIFA
- FIBA
- ESPN
- Game
- Athlon

Tabla de deportes completa:

Deporte	
Deporte	Count
American_Football	9
Arbitrage	1
Athletics	2
Australian_Football	2
Baseball	7

Basketball	14
Billiard	1
Chess	4
Cricket	2
Cycling	9
Darts	1
Dota2	3
Football	12
Hockey	5
Injury_Sports_Medicine	4
Mahjong	1
Physical_Activity	11
Running	5
Swimming	3
Scuba_diving	2
Softball	1
Speedway	1
Squash	1
Tennis	1
Volleyball	1
Walking	1

Tabla de Tipo de metodología completa:

Tipo de metodología	
Tipo	Count
Categorization	1
Classification	5
Compute	17
Decoding	1
Descriptive	7
Extraction	1
Modeling	3
Prediction	8
Preprocessing	4
Probabilities	2
Processing	14
Scraping	43
Simulation	7
Visualization	8

7.3. Código R

Link Github: https://github.com/VictorMartinezRech/TFG_R