DSA Awards 2019 - Transformando el mundo del baloncesto en España - Liga ACB -

Marco Russo

Contents

Descripción del caso

DSA2019 - GO MOVING - Sport Analytics Liga ACB

autor: Marco Russo contact: mrusso@paradigmadigital.com date: septiembre 2019

Biography

Hola a todos, mi nombre es Marco Russo y para mí es un placer participar a este reto. De primera parto con bastante retraso por un imprevisto.

Sobre mí, tengo formación en ciencias económicas con especialización en finanza de mercado de valores y banca, sucesivamente he estudiado marketing, marketing digital , analítica de datos y finalmente con la UOC el posgrado de Business Analytics, además de formación en analítica de datos y aprendizaje automático.

Trabajo como consultor de datos y BI en Paradigma Digital en Madrid, empresa asociada con Indra en diferentes proyectos de transformación digital y área de Big Data (principalmente minería de datos y visualización). También soy formador in-company y apoyo en la formación de otros empleados en visualización de datos (las herramientas que utilizamos entre otras, Data Studio, Tableau, PowerBI principalmente), y nuestros clientes son casi la mayoría del Ibex35. A nivel interno trabajo en proyectos de apoyo al área de finanza y RRHH (business intelligence).

Por último desde hace 7 años he colaborado como docente impartiendo programas de comercio electrónico, marketing digital y datos en la Cámara de Comercio y otras escuelas de negocios como profesor de analítica

principalmente. Cómo último profesor colaborador en la UOC en la asignatura de Data Mining y en NEOLAND como profesor principal del Máster de Data Science.

Me hace más ilusión el poder compartir estos retos a mis estudiantes a que yo lo envie, porque la verdad, me ha faltado tiempo.

Que gane el mejor!

gracias, un saludo marcusRB

Transformando el mundo del baloncesto a través de Sports Analytics en España

La analítica deportiva se entiende como el uso de estadísticas históricas y relevantes, que, aplicadas correctamente, pueden proporcionar una ventaja competitiva a un equipo o deportista. A través de la colección y el análisis de estos datos, la analítica deportiva puede ayudar a jugadores y entrenadores en el proceso de la toma de decisiones previo y durante los eventos deportivos. Esta industria se popularizo masivamente tras el lanzamiento en 2011 de la película Moneyball, en la que el manager general del equipo de los Oakland Athletics de Béisbol, Billy Bean, basó la construcción de su equipo en métodos analíticos y cuantitativos. Conociendo la influencia de que los jugadores llegasen a bases para conseguir victorias, Beane se centró en fichar jugadores con un alto porcentaje de conversiones de base con la lógica de que los equipos con mayor porcentaje de conversiones de base eran más propensos a lograr carreras. Esto resultó en la construcción de un equipo tremendamente competitivo con el presupuesto más limitado de la Major League Baseball (MLB). Este éxito no pasó desapercibido para los ejecutivos de equipos profesionales de otros deportes. Hoy en día, y favorecido por el avance tecnológico, es difícil encontrar equipos profesionales que no utilizan datos para la toma de decisiones estratégicas.

Por ejemplo, el Movistar Cycling Team, el Movistar Riders, la Rafa Nadal Academy by Movistar o el Movistar Estudiantes de Baloncesto que ya implementan soluciones analíticas para impulsar la toma de decisiones deportivas basadas en datos.

Es precisamente de baloncesto sobre lo que trata este reto. Se busca encontrar ventajas competitivas para los equipos de baloncesto a partir del análisis de datos de partidos, equipos y jugadores. Al contrario que en Béisbol donde el rendimiento de cada jugador se puede cuantificar fácilmente, en el baloncesto los cinco jugadores son factores en cada jugada, y muchas de las contribuciones de algunos jugadores no se reflejan en las estadísticas tradicionales que se muestran al final de cada partido. Por ejemplo, los bloqueos o las ayudas defensivas rara vez se cuantifican en las estadísticas finales, pero ciertamente contribuyen favorablemente al equipo. Se trata por lo tanto de encontrar estadísticas avanzadas que vayan más allá de lo que se ve en las estadísticas tradicionales, con el fin de cuantificar lo más precisamente posible, el rendimiento de cada jugador, así como su impacto en el equipo.

Un claro ejemplo de cómo un equipo se ha beneficiado del poder de la analítica avanzada son los Houston Rockets. Como se observa en esta noticia, se dieron cuenta mediante la analítica, que les convenía aumentar considerablemente los intentos de tiros de 3. En reacción a este cambio de juego, muchos equipos de la NBA han tomado medidas prescriptivas al respecto y han cambiado la manera de defender a los Houston Rockets. Algunas de ellas se han hecho virales como la defensa de Ricky Rubio a James Hardem, jugador insignia de los Rockets.

Objetivo

El objetivo de este reto es descubrir ventajas competitivas para los equipos españoles de baloncesto a partir de la analítica deportiva. Como punto de partida sugerimos varias propuestas, pero el reto está abierto a otras posibilidades: * Variables más influyentes en determinar el resultado de los equipos de la Liga Endesa * Creación de KPI's para evaluar el rendimiento deportivo de jugadores de la liga Endesa * Análisis de los jugadores y equipos de la Liga Endesa durante el "clutch time" (durante el último cuarto con menos de 5 minutos para el final del partido y cuando ningún equipo tiene una ventaja de más de 5 puntos). Las estadísticas durante el "clutch" no están a priori disponibles abiertamente, pero se pueden extraer a partir del

play-by-play * Rendimiento según parámetros de estadísticas avanzadas por quintetos de todos los equipos (incluido el "clutch" de los quintetos) * Diferencias en estadísticas avanzadas del jugador y del equipo cuando se gana y cuando se pierde * Diferencias en estadísticas avanzadas del jugador y del equipo cuando juega en casa y cuando juega como visitante * Análisis espacial de las posiciones desde las que los jugadores realizan tiros (coordenadas), del posicionamiento defensivo de los rivales... etc. Posterior creación de cartas de tiro para encontrar patrones da acierto o fallo en determinadas posiciones por determinados jugadores y equipos * Categorización y clusterización de equipos y jugadores en base a su estilo de juego / estadísticas para encontrar jugadores y equipos que compartan patrones de juego * Propuestas de fichajes de jugadores y de renegociación de contratos a partir de todas las propuestas anteriores

Requisitos

Para realizar el reto existen los siguientes requisitos: * Metodología científica del problema, donde se indica los pasos necesarios para obtener la solución al problema. * Diseño e implementación de software, donde se justifican los motivos de utilización de una tecnología/software/algoritmo u otra. * Explicación analítica del proceso de selección, aprendizaje y evaluación de los modelos usados en el proyecto.

Data Set

Para este reto, proporcionamos algunos data sets que pueden ser utilizados por los participantes, pero aparte de estos data sets, se pueden utilizar otras fuentes adicionales. Los data sets proporcionados son los siguientes: * ACB_Players_18-19.xlsx: Data set con estadísticas avanzadas de los jugadores de la ACB durante la temporada 2018-2019. * ACB_Players_2012to2018.xlsx: Data set con estadísticas avanzadas de los jugadores de la ACB desde la temporada 2011-2012 hasta la temporada 2017-2018. * ACB_Teams_18-19.xlsx: Data set con estadísticas avanzadas de los equipos de la ACB durante la temporada 2018-2019. * Dataset-Variables-Description.docx: Documento con la descripción de las variables de los data sets.

Se sugieren páginas de baloncesto especializadas como RealGM o la página oficial de la Liga ACB para la obtención de datos abiertos sobre partidos, equipos y jugadores.

Valoración

Para afrontar el reto, se valorarán los siguientes aspectos: * El valor y la ventaja competitiva de los resultados * La creatividad para encontrar "insights" más allá de los visibles a primera vista, así como el uso de técnicas descriptivas bien ejecutadas para su correcta visualización * El uso de data sets adicionales que permiten " insights" creativos * Recomendaciones concretas para los equipos

DESCRIPCIONES VARIABLES

RealGM's Basic Stat Line

G: Games

Min: Minutes

FGM-A: Field Goals Made - Field Goals Attempts

FG%: Field Goal Percentage

3PTM-A: Three-Point Field Goals Made – Three-Point Field Goals Attempted

3PT%: Three-Point Field Goal Percentage

FTM-A: Free Throws Made – Free Throws Attempted

FT%: Free Throw

FIC (Floor Impact Counter): A formula to encompass all aspects of the box score into a single statistic. The intent of the statistic is similar to other efficiency stats, but assists, shot creation and offensive rebounding are given greater importance. Created by Chris Reina in 2007.

Formula: (Points + ORB. + 0.75 DRB + AST + STL + BLK –0.75 FGA – 0.375 FTA – TO – 0.5 PF)

FIC40 (Floor Impact Counter per 40 minutes): The FIC total presented on a per-40 minute basis.

OFF: Offensive Rebounds

DEF: Defensive Rebounds

REB: Total Rebounds

AST: Assists

STL: Steals

BLK: Blocks

TO: Turnovers

PTS: Points

Advanced/Misc. Stats

TS% (True Shooting Percentage): A measurement of efficiency as a shooter in field goal attempts, three-point field goal attempts and free throws.

Formula: (Points x 50) / [(FGA + 0,44 * FTA)]

eFG% (Effective Field Goal Percentage): A measurement of efficiency as a shooter in all field goal attempts with three-point attempts weighted fairly.

Formula: (FG + 0.5 * 3P) / FGA

ORB% (Offensive Rebound Percentage): A measurement of the percentage of offensive rebounds a player secures that are available to his team.

Formula: 100 * [Player ORB * (Team Minutes / 5)] / [Player Minutes * (Team ORB + Opponent DRB)]

DRB% (Defensive Rebound Percentage): A measurement of the percentage of defensive rebounds a player secures that are available to his team.

Formula: 100 * [Player DRB * (Team Minutes / 5)] / [Player Minutes * (Team DRB + Opponent ORB)]

TRB% (Total Rebound Percentage): A measurement of the percentage of both offensive and defensive rebounds a player secures that are available to his team.

Formula: 100 * [Total Player Rebounds * (Team Minutes / 5)] / [Player Minutes * (Team Total Rebounds + Opponent Total Rebounds)]

AST% (Assist Percentage): A measurement of the percentage of assists a player records in relation to the team's overall total while he is in the game.

Formula: 100 * Player ASTs / [((Player Minutes / (Team Minutes Played / 5)) * Team FGs) – Player FGs]

STL% (Steal Percentage): A measurement of the percentage of steals a player records in relation to the team's overall total while he is in the game.

Formula: 100 * [Player STLs * (Team Minutes / 5)] / (Player Minutes * Opponent Possessions)

BLK% (Block Percentage): A measurement of the percentage of blocks a player records in relation to the opponents two point field goal attemps.

Formula: 100 * [Player BLKs * (Team Minutes / 5)] / (Player Minutes * Opponent FGA - Opponent 3PA)

TOV% (Turnover Percentage): A measurement of the percentage of turnovers a player records in relation to the team's overall total while he is in the game.

Formula: 100 * Turnovers / (FGA + 0.44 * FTA + TOV)

Total S % (Total Shooting Percentage): The sum of a player's field goal, free throw and three-point percentage.

ORtg (Offensive Rating): The number of points a player produces per 100 possessions. Created by Dean Oliver.

DRtg (Defensive Rating): The number of points a player allows per 100 possessions. Created by Dean Oliver.

eDiff (Efficiency Differential): The difference between a team or player's ORtg and DRtg.

Formula: (ORtg - DRtg)

PER: An efficiency statistic created by John Hollinger. Click here for more information.

Entrega del reto

La entrega del reto deberá contar con los siguientes documentos entregables:

Memoria del proyecto: Ésta se presentará en formato PDF y no podrá superar las 30 páginas. La fuente empleada en el contenido será Arial Narrow de tamaño 12pt con un interlineado sencillo. Dicha memoria estará dividida en los siguientes apartados: Portada con título e identificación del concursante. Metodología y planificación. Descripción de los datos y procesamiento de los mismos. Explicación justificada del diseño e implementación de la infraestructura y componentes/servicios usados. Explicación justificada de la parte analítica (con validación analítica incluida). Explicación justificada del Backend implementado (en caso de disponer). Explicación justificada del Frontend implementado (en caso de disponer). Demostración mediante ejemplos (casos de uso). Si fuera posible, enviar link a la aplicación interactiva implementada. Ficheros que documenten el proyecto: código fuente, fuentes de datos, ... Fichero descripción.txt que enumere y describa cada uno de los ficheros presentados (obligatorio). Todos estos ficheros anteriormente descritos deberán ser almacenados (con directorios o no) en un fichero comprimido .zip, con el nombre que se desee. El fichero .zip no deberá ocupar más de 200MB, ya que el sistema no permite ficheros de tamaño superior.

Resumen del trabajo

En mi opinión, el baloncesto es un deporte maravilloso, se puede decir mucho sobre una persona por la forma en que jugaba baloncesto, cosas como cogió la pelota? ¿Presumió en la cancha? ¿La persona tenía miedo de tirar y fallar? ¿La persona mintió acerca de haber recibido una falta? Además de eso, ¡es divertido jugar y hacer un gran ejercicio!

También en mi opinión, Data Analytics / Data Science es un campo increíblemente popular y en crecimiento, tanto es así que fue nombrado "el trabajo más sexy del siglo XXI". La ciencia de datos es una mezcla de estadísticas, análisis de datos, aprendizaje automático, informática y conocimiento de los datos / negocios que tiene como objetivo proporcionar información y comprensión de los datos.

Históricamente, la recogida de datos y el análisis de datos en los diferentes deportes se centra en estadísticas acumuladas anuales para comparar el desempeño de los diferentes jugadores, tanto que la liga amricana NBA ahora ejecuta un Hackathon anual, lo que les permite obtener nuevas ideas geniales y encontrar nuevos analistas de datos con talento.

Con el gran avance que se ha producido en la recogida y procesamiento de datos, existe la posibilidad de realizar análisis más avanzados. Serán análisis que nos permitan ponderar y realizar una clasificación, aplicando los conceptos del learning to rank, de los jugadores en función de aspectos que puedan ser influyentes a la hora de comparar su desempeño.

La hipótesis en que se basa este estudio sobre el baloncesto, en particular aplicado a la LIGA ACB es que hay dos factores interrelacionados que influyen en el desempeño y que no suelen tomarse en consideración. El *primero*, es el conocimiento del juego que permite a un jugador aplicar la estrategia correcta según se plantee un problema en forma de defensa adversaria. El *segundo* es la importancia del partido, ya que varía mucho

según el momento de la temporada sea. Tomando como ejemplo en la NBA no existen descensos de categoría, la temporada regular es muy larga y en los playoffs las franquicias se juegan el trabajo de todo el año.

El objetivo de este estudio es conseguir un **análisis estadístico** que tenga en cuenta ambos factores para poder comparar los puntos fuertes y débiles de los jugadores. El resultado del estudio aportará información que permita a los entrenadores y directores deportivos realizar una rápida toma de decisiones en un mercado de fichajes muy cambiante.

Estructura del trabajo

Al disponer solo de pocos días a la semana para dedicar a este proyecto, es comenzado con una pequeña exploración de los datos proporcionados y tener un poco más la libertad de ver que hay más allá de estos dataset que se podemos concluir. Finalmente he visto muchos más trabajos y avanzados en este sentido, en la liga *NBA* y la universitaria *NCAA*. De hecho hay sub-proyectos muy interesantes a la hora de poder abordar un **PoC** con un equipo de baloncesto de la liga española.

Enumeraré los sub-proyectos que he ido enumerando que he estado desarrollando (y estaré trabajando con mis alumnos del próximo curso):

- **Data weareble**: Utilizar los datos biométricos a la hora de detectar con antelación los posibles cambios durante el partido. Se ha comprobado el mismo a través de la aplicación conocida en este mundo del deporte (y que se utiliza bastante en Movistar Cycling), **STRAVA**, además de utilizar los variables propias del jugador y así crear un nueva métrica con el fin de obtener: *%potencia*, *%cansancio*, *%lucidez*, *%lesiones*, *%respiración*, *%pulsaciones*, *%impacto*, *%estado_estrés*, etc. Además viendo muchos videojuegos utilizan exactamente un algoritmo muy similar. Aquí la noticia NBA and RDF

Acompañando este primer sub-proyecto, hablaré del **Perfomance_Analysis** que obviamente al faltar los primeros datos que creo sean muy útiles para determinar la métrica que hasta ahora se calcula de una manera, **PER**, el control de datos biométricos muy importantes para tomar decisiones basadados en tiempo real de dispositvos visto anteriormente, el atleta tendrá en todo momento incluso alertas de cuando está llegado a su límite de fuerzas.

- **Deep Learning aplicado a los tiros**: Otro sub-proyecto a realizar y ya estudiando en la *NCAA* es la capacidad de estudiar a través de técnicas de deep-learning y redes neuronales a estudiar y ser capaz de detectar a una distancia x con otros factores si el equipo va a canasta o no. El artículo que hace mención a esto es en FiveThirtyEight.
- **Track de Movimientos**: Una de las cosas más interesantes es un estudio desde 2009 de grabacione de partidos, que están aprovenchando con **Tensorflow**, **Keras**, **PyTorch**, para analizar cada uno de ellos y detectar patrones. Aquí el extracto :

  En 2009, la liga comenzó a utilizar un sistema de video de última generación para rastrear el movimiento de los jugadores en la cancha y la pelota. Tener este nuevo sistema de video le permitió a la NBA recopilar nuevos datos, lo que a su vez permitió a los científicos de datos utilizar el aprendizaje automático y la cartografía (la ciencia o la práctica de dibujar mapas) para evaluar mejor qué jugadores ayudaron a su equipo a ganar.
- **Rediseñando el equipo** : Será que la NBA es otro nivel que sin duda ni se acerca a cualquier europea (hasta incluso pienso que ni la NCAA), pero sin embargo algo se mueve en la dirección correcta y hay físicos, científicos, matemáticos, analistas, estadísticos tan buenos como en EEUU, así mejor aprovecharlo al máximo. Lo que se estudio lo que muestro a continuación es algo muy amplio y basado en un *método de clasificación*. Estudio completo

  ¿Se ha preguntado por qué solo hay 5 posiciones en el baloncesto o cómo se determina la posición de un jugador? Nosotros también. Pero ahora, utilizando el motor de análisis de datos patentado de Ayasdi y décadas de investigación de topología computacional en Stanford, hemos categorizado matemáticamente a los jugadores en 13 nuevas posiciones: las posiciones reales del baloncesto (que se presentará en esta presentación). Describiré esta visión revolucionaria y cómo puede agregar un gran valor para los entrenadores, propietarios, gerentes generales y fanáticos de todos los días.

Al visualizar la forma de los datos en términos de posiciones basadas en el rendimiento, podemos descubrir jugadores infravalorados, administrar decisiones en el juego, optimizar listas y redactar de manera más inteligente. Y esta mayor granularidad en las posiciones de baloncesto es solo el comienzo. También describiré cómo el análisis de datos topológicos puede abrir el camino a más evoluciones en los pensamientos en el baloncesto y otros deportes.

Métricas

Aquí unas cuantas métricas recogidas, separadas por:

- Moneyball
- Player Evaluation Metrics
- Team Evaluation Metrics

cada una están indicadas y especificadas en nbastuffer

Configuración básica

Los primeros pasos de una configuración básica son instalación de paquetes y carga de librería tanto para la exploración de los datos como las específicas de algoritmos.

Paquetes - Librerías

```
library(readxl)
```

Observación si existen hojas internas de los ficheros excel

Para poder realizar correctamente la carga de los ficheros en formato excel, nos aseguramos que no existan otras hojas a cargar y contemplar durante la fase de guardar dataset.

```
# Con la función excel_sheets observaremos si existen más de una hoja
excel_sheets("datasets/ACB_Players_18-19.xlsx")
```

```
## [1] "Sheet1"
```

```
excel_sheets("datasets/ACB_Teams_18-19.xlsx")
```

```
## [1] "Sheet1"
```

```
excel_sheets("datasets/ACB_Players_2012to2018.xlsx")
```

```
## [1] "Sheet1"
```

Perfecto!, no existen más que una hoja por fichero excel.

Carga de los datasets

```
# Efectuaremos la carga de los 3 dataset
ACB_Players_18_19 <- read_excel("datasets/ACB_Players_18-19.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
ACB_Teams_18_19 <- read_excel("datasets/ACB_Teams_18-19.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
ACB_Players_2012to2018 <- read_excel("datasets/ACB_Players_2012to2018.xlsx", )
```

Estructura de los datasets

Rápidamente visualizaremos la estructura de los 3 datasets y observaciones / variables en cada uno de ellos.

```
# Visualizaremos la estructura del dataset ACB_player_18_19
str(ACB_Players_18_19)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    276 obs. of  47 variables:
##  $ ...1       : chr  "1" "2" "3" "4" ...
##  $ Player     : chr  "David Jelinek" "Shayne Whittington" "David Walker" "LaDontae Henton" ...
##  $ Team       : chr  "AND" "AND" "AND" "AND" ...
##  $ Team_full  : chr  "MoraBanc Andorra" "MoraBanc Andorra" "MoraBanc Andorra" "MoraBanc Andorra" ...
##  $ Position   : chr  "SG" "C" "GF" "F" ...
##  $ altura     : num  196 211 198 198 178 201 196 188 208 213 ...
##  $ Peso       : num  86 113 91 98 75 101 89 88 100 115 ...
##  $ Nationality: chr  "Czech Republic" "United States" "United States" "United States" ...
##  $ temporada  : chr  "2018-2019" "2018-2019" "2018-2019" "2018-2019" ...
##  $ GP         : num  30 12 16 2 33 34 34 33 21 19 ...
##  $ MPG        : num  19.6 14.4 20.1 5.4 25.2 16.9 20.2 25 13.2 14.1 ...
##  $ FGM        : num  3 3.6 3.1 0.5 2.8 2.6 2.9 4.5 2.2 2 ...
##  $ FGA        : num  7.8 6.8 6.9 1.5 7.4 6 7 10.2 3.7 4.3 ...
##  $ FG%        : num  0.38 0.524 0.445 0.333 0.379 0.438 0.414 0.436 0.597 0.463 ...
##  $ 3PM        : num  1.6 0.8 0.9 0 1.7 0.5 1.4 1.6 0 0 ...
##  $ 3PA        : num  4.2 2.3 2.9 0.5 4.8 2 4 4.3 0 0 ...
##  $ 3P%        : num  0.37 0.357 0.298 0 0.348 0.265 0.338 0.366 0 0 ...
##  $ FTM        : num  1.6 1.7 1.5 0 1.1 0.5 0.8 2.9 1.6 2.4 ...
##  $ FTA        : num  1.9 2.2 1.8 0 1.4 1.1 1.4 3.9 2.1 3.6 ...
##  $ FT%        : num  0.825 0.741 0.828 0 0.761 0.5 0.553 0.738 0.75 0.676 ...
##  $ TOV        : num  1 0.8 0.9 0 2 0.6 1 1.7 0.9 1 ...
##  $ PF         : num  2.3 2.5 1.4 1 2.1 2 1.9 2 2.2 1.9 ...
##  $ ORB        : num  0.6 1.8 0.4 0 0.4 1.3 0.3 0.8 1.5 1 ...
##  $ DRB        : num  1.9 2 1.7 0.5 1.4 2.9 1.9 2.2 1.6 1.9 ...
##  $ RPG        : num  2.5 3.8 2.1 0.5 1.8 4.1 2.1 3 3.1 2.9 ...
##  $ APG        : num  1.3 0.8 1.2 0.5 5.2 0.4 2 2.7 0.5 0.7 ...
##  $ SPG        : num  0.6 0.5 0.5 0 1.2 0.5 0.5 0.8 0.6 0.3 ...
##  $ BPG        : num  0.2 0.6 0.1 0 0 0.3 0 0.2 0.1 0.6 ...
##  $ PPG        : num  9.1 9.7 8.5 1 8.3 6.3 7.9 13.4 6 6.4 ...
##  $ TS%        : num  0.525 0.618 0.554 0.333 0.52 0.489 0.52 0.561 0.649 0.614 ...
##  $ eFG%       : num  0.481 0.585 0.509 0.333 0.492 0.483 0.511 0.513 0.597 0.56 ...
##  $ Total S %  : num  157.5 162.2 157.1 33.3 148.8 ...
##  $ ORB%       : num  3.7 15 2.5 0 1.9 9.2 1.6 3.6 13.7 7 ...
##  $ DRB%       : num  12.2 18.9 10.9 10.7 7.1 21.6 11.9 11.3 16 21.5 ...
##  $ TRB%       : num  7.8 16.8 6.5 5.3 4.4 15.2 6.6 7.4 14.8 13.9 ...
##  $ AST%       : num  11.8 11.8 10.4 13.1 34.2 3.7 16.9 19.7 7.2 8.5 ...
##  $ TOV%       : num  10.1 8.7 10.9 0 20 8 11.4 12.2 16.5 20.7 ...
##  $ STL%       : num  1.6 1.9 1.4 0 2.7 1.7 1.3 1.8 2.5 0.5 ...
##  $ BLK%       : num  1.1 4.4 0.7 0 0.1 2.1 0.2 0.8 0.8 3.5 ...
##  $ USG%       : num  22.8 27.2 20.1 12.6 18.5 19.4 19.7 25.4 19.2 21 ...
##  $ PPR        : num  -0.5 -1.7 -0.7 5.8 5.8 -1.9 1.7 0.4 -4.1 -6.2 ...
##  $ PPS        : num  1.2 1.4 1.2 0.7 1.1 1.1 1.1 1.3 1.6 1.5 ...
```

```
##  $ ORtg       : num  108.4 127.7 110.9 93.3 108.9 ...
##  $ DRtg       : num  113 112 112 117 113 ...
##  $ eDiff      : num  -4.8 15.9 -1.6 -23.6 -4.1 -4.6 -6.8 0.7 11.5 -2.9 ...
##  $ FIC        : num  134.6 82.4 72.9 0.5 236 ...
##  $ PER        : num  12.8 27.3 12.6 2.4 12.6 12.9 11.1 16.7 19.1 15.5 ...
```

```
dim(ACB_Players_18_19)
```

```
## [1] 276  47
```

Contabilizamos 276 observaciones por 47 variables diferentes. Se observan variables numéricas y categorías, de la cuáles la posición del jugador, variable **Position** será nuestro factor

```
# Observaremos la estructura del dataset Team
str(ACB_Teams_18_19)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    18 obs. of  40 variables:
##  $ ...1    : chr  "1" "2" "3" "4" ...
##  $ Team    : chr  "Cafes Candelas Breogan" "Delteco GBC" "Divina Seguros Joventut" "FC Barcelona Lass
##  $ initials: chr  "BRE" "GBC" "JOV" "FCB" ...
##  $ GP      : num  34 34 34 34 34 34 34 34 34 34 ...
##  $ MPG     : num  40.3 40.4 40.1 40.1 40.1 40.9 40.1 40 40.7 40 ...
##  $ FGM     : num  28.1 27.2 29 30.9 30.2 28.6 28.1 32 28.9 28.7 ...
##  $ FGA     : num  65.6 62.1 60.2 61.6 65.7 60.3 63.1 63.4 64 64.5 ...
##  $ FG%     : num  0.428 0.438 0.481 0.501 0.46 0.474 0.446 0.505 0.452 0.446 ...
##  $ 3PM     : num  7.9 8.3 9.4 10.2 9.7 11.2 9.3 8.9 8.6 9.6 ...
##  $ 3PA     : num  24.7 24.7 25.1 24.5 27.4 29.5 26.5 23.8 25 27.6 ...
##  $ 3P%     : num  0.322 0.335 0.377 0.418 0.354 0.378 0.352 0.372 0.345 0.348 ...
##  $ FTM     : num  13.1 12.6 13.6 14.7 13 13.2 14.7 13.2 15.6 15.7 ...
##  $ FTA     : num  18.3 17.6 17.4 19.9 17.2 17.7 20.5 17.9 21.1 21.6 ...
##  $ FT%     : num  0.717 0.717 0.783 0.741 0.757 0.749 0.716 0.738 0.737 0.729 ...
##  $ TOV     : num  12.6 13.5 14.5 12.1 12 11.9 12.7 12.4 12.6 11.6 ...
##  $ PF      : num  21.5 21.3 20.9 19.9 22.1 21 21.6 19.1 22 21.7 ...
##  $ ORB     : num  10.2 8.7 8.2 9.1 9.6 8.2 9.3 8.2 8.7 10 ...
##  $ DRB     : num  23.2 21.3 22.2 24.6 20.5 20.5 22.1 24.7 20.6 21.7 ...
##  $ RPG     : num  33.4 30.1 30.4 33.8 30.1 28.6 31.4 32.9 29.3 31.7 ...
##  $ APG     : num  14.1 16.3 16.5 18.2 15.8 18 14.5 19.4 14.6 16.4 ...
##  $ SPG     : num  5.4 7 6.5 7 7.1 6.5 6.4 8 6.4 7.2 ...
##  $ BPG     : num  3.2 1.8 2.8 2.4 2.1 1.8 3.2 2.9 3.1 2.4 ...
##  $ PPG     : num  77.1 75.2 81 86.7 83.2 81.5 80.3 86.1 82 82.8 ...
##  $ TS%     : num  0.524 0.539 0.597 0.616 0.567 0.599 0.557 0.604 0.559 0.56 ...
##  $ eFG%    : num  0.489 0.504 0.56 0.584 0.534 0.566 0.52 0.574 0.519 0.52 ...
##  $ Total S%: num  147 149 164 166 157 ...
##  $ ORB%    : num  29.1 27.4 28.2 31.6 29.5 27.9 28.1 27.6 26.6 30 ...
##  $ DRB%    : num  73.4 70.8 70.1 74.8 72.5 70 70.8 72.9 65.4 69.5 ...
##  $ TRB%    : num  50.1 48.5 50 54.6 49.6 48.9 48.8 51.7 45.7 49.1 ...
##  $ AST%    : num  50.2 59.8 57 58.9 52.3 63.1 51.5 60.5 50.6 57.2 ...
##  $ TOV%    : num  14.7 16.2 17.6 14.7 14.1 14.9 15 14.8 14.7 13.6 ...
##  $ STL%    : num  7.4 9.7 8.9 9.8 9.6 9.4 8.7 10.8 8.6 9.8 ...
##  $ BLK%    : num  8.5 4.9 7.7 7.3 5.5 5.2 8.4 7.9 7.6 6.3 ...
##  $ PPS     : num  1.2 1.2 1.3 1.4 1.3 1.4 1.3 1.4 1.3 1.3 ...
##  $ FIC40   : num  47.7 47.2 54.8 66 54.1 55.9 51.6 66.8 49.8 56.1 ...
##  $ ORtg    : num  105 104 112 121 113 ...
##  $ DRtg    : num  114 114 111 106 115 ...
##  $ eDiff   : num  -9.8 -9.5 0.9 14.7 -1.8 2.6 -2.5 15.1 -9.2 1.2 ...
##  $ Poss    : num  2506 2460 2458 2437 2503 ...
```

```
##  $ Pace    : num  73.2 71.6 72 71.4 73.3 68.2 73.2 73.8 73.4 73.4 ...
```

```
dim(ACB_Teams_18_19)
```

```
## [1] 18 40
```

Está compuesto de 18 observaciones y 40 variables, la mayoría de ellas numéricas.

```
# Nuevamente miraremos los estadísticos de los jugadores desde 2012 a 2018
str(ACB_Players_2012to2018)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1815 obs. of  46 variables:
##  $ Player     : chr  "Oliver Stevic" "Tomas Hampl" "Kostas Vasiliadis" "Josh Fisher" ...
##  $ Team       : chr  "BBB" "BBB" "BBB" "BBB" ...
##  $ Team_full  : chr  "RETAbet Bilbao Basket" "RETAbet Bilbao Basket" "RETAbet Bilbao Basket" "RETAbe
##  $ Pos        : chr  "FC" "C" "F" "G" ...
##  $ Height(ft) : chr  "6-10" "7-1" "6-7" "6-2" ...
##  $ Weight(lb) : chr  "220" "240" "225" "200" ...
##  $ Nationality: chr  "Serbia" "Czech Republic" "Greece" "United States" ...
##  $ temporada  : chr  "2011-2012" "2011-2012" "2011-2012" "2011-2012" ...
##  $ GP         : num  4 3 35 30 36 34 31 35 36 33 ...
##  $ MPG        : num  11.6 8.9 21.9 11.1 26.2 27.1 13.8 15.1 25.8 15.2 ...
##  $ FGM        : num  1.2 1.7 3.3 1.4 3.6 4.7 1.4 2 3.9 2 ...
##  $ FGA        : num  2.8 3.3 8.1 2.7 9 7.9 3 4.3 8.1 4.7 ...
##  $ FG%        : num  0.455 0.5 0.406 0.512 0.404 0.593 0.457 0.47 0.485 0.423 ...
##  $ 3PM        : num  0 0 1.6 0.5 1.2 0.1 0 1.2 0.9 1.2 ...
##  $ 3PA        : num  0 0 5.3 1.5 3.8 0.3 0 2.7 2.2 3.1 ...
##  $ 3P%        : num  0 0 0.299 0.341 0.304 0.222 0 0.453 0.383 0.398 ...
##  $ FTM        : num  1 1 3.9 0.1 1.6 2.1 0.9 1.1 2.2 0.5 ...
##  $ FTA        : num  1.5 1.3 4.3 0.2 2.1 2.4 1.6 1.3 2.9 0.7 ...
##  $ FT%        : num  0.667 0.75 0.907 0.5 0.727 0.855 0.569 0.844 0.757 0.625 ...
##  $ TOV        : num  0.8 0.7 1.5 0.6 2.8 1.4 1.1 0.9 1.8 0.9 ...
##  $ PF         : num  1.8 1.7 1.5 1.1 2.3 2.9 2.1 1.9 2.2 2.5 ...
##  $ ORB        : num  1.2 1.3 0.4 0.3 0.8 1.4 1.5 0.2 0.8 0.1 ...
##  $ DRB        : num  1.8 1.3 1.9 1.2 3.8 2.9 1.9 1.1 2.6 1.2 ...
##  $ RPG        : num  3 2.7 2.4 1.5 4.5 4.3 3.4 1.3 3.3 1.2 ...
##  $ APG        : num  0.8 0 1.1 0.7 2.8 0.8 0.3 1.5 2.9 0.6 ...
##  $ SPG        : num  0 0 0.7 0.5 0.7 0.6 0.4 1 0.9 0.3 ...
##  $ BPG        : num  0.5 0 0.2 0.3 0.1 0.3 0.3 0 0 0 ...
##  $ PPG        : num  3.5 4.3 12 3.4 10 11.5 3.6 6.3 10.9 5.7 ...
##  $ TS%        : num  0.513 0.553 0.603 0.603 0.503 0.642 0.494 0.655 0.581 0.564 ...
##  $ eFG%       : num  0.455 0.5 0.504 0.604 0.469 0.597 0.457 0.614 0.538 0.554 ...
##  $ TotalS%    : num  112 125 161 135 144 ...
##  $ ORB%       : num  16.3 20.4 2.8 3.9 4 7.3 15.2 1.9 4.1 0.8 ...
##  $ DRB%       : num  20.5 18.5 11.5 13.9 18.5 13.7 18.5 9.8 12.8 9.8 ...
##  $ TRB%       : num  18.5 19.4 7.3 9.1 11.6 10.6 16.9 6 8.7 5.5 ...
##  $ AST%       : num  10.2 0 9.4 11 19.6 5.4 3.2 17.3 20.6 7.3 ...
##  $ TOV%       : num  18 14.5 13.2 16.7 21.7 13.1 22.9 15.1 16 15.3 ...
##  $ STL%       : num  0 0 1.7 2.5 1.5 1.3 1.7 3.6 1.9 1.1 ...
##  $ BLK%       : num  4.6 0 0.8 2.6 0.5 1.2 2.5 0.2 0 0 ...
##  $ USG%       : num  17.7 24.5 25.7 15.1 23.8 18.7 17.1 18.5 21.2 19.2 ...
##  $ PPR        : num  -2.1 -7 -3.4 -0.9 -3.2 -3 -6.5 0.7 0.5 -3.2 ...
##  $ PPS        : num  1.3 1.3 1.5 1.2 1.1 1.5 1.2 1.5 1.3 1.2 ...
##  $ ORtg       : num  107 107.3 116.5 111.3 93.9 ...
##  $ DRtg       : num  112 112 108 106 107 ...
##  $ eDiff      : num  -4.8 -5.2 8.3 5.5 -12.7 12.6 -10.6 19.5 5.7 -7.9 ...
```

```
##  $ FIC        : num  12.2 6.5 207.5 83.5 206.9 ...
##  $ PER        : num  13.6 15.8 20.1 14.7 12.5 18.3 10.7 19.2 17.8 9.7 ...
```

```
dim(ACB_Players_2012to2018)
```

```
## [1] 1815   46
```

El tercer dataset se consta de 1815 observaciones y 46 variables, al igual que el anterior necesitamos factorizar la variable posición **Position**.

```
# Observación de los primeros resultados de los tres datasets
head(ACB_Players_18_19, n=100)
```

```
## # A tibble: 100 x 47
##    ...1  Player Team  Team_full Position altura  Peso Nationality temporada
##    <chr> <chr>  <chr> <chr>     <chr>     <dbl> <dbl> <chr>       <chr>
##  1 1     David~ AND   MoraBanc~ SG          196    86 Czech Repu~ 2018-2019
##  2 2     Shayn~ AND   MoraBanc~ C           211   113 United Sta~ 2018-2019
##  3 3     David~ AND   MoraBanc~ GF          198    91 United Sta~ 2018-2019
##  4 4     LaDon~ AND   MoraBanc~ F           198    98 United Sta~ 2018-2019
##  5 5     Andre~ AND   MoraBanc~ PG          178    75 France      2018-2019
##  6 6     Reggi~ AND   MoraBanc~ F           201   101 United Sta~ 2018-2019
##  7 7     Miche~ AND   MoraBanc~ G           196    89 Italy       2018-2019
##  8 8     Dylan~ AND   MoraBanc~ G           188    88 United Sta~ 2018-2019
##  9 9     Olive~ AND   MoraBanc~ FC          208   100 Serbia      2018-2019
## 10 10    Jerom~ AND   MoraBanc~ C           213   115 Jamaica     2018-2019
## # ... with 90 more rows, and 38 more variables: GP <dbl>, MPG <dbl>,
## #   FGM <dbl>, FGA <dbl>, `FG%` <dbl>, `3PM` <dbl>, `3PA` <dbl>,
## #   `3P%` <dbl>, FTM <dbl>, FTA <dbl>, `FT%` <dbl>, TOV <dbl>, PF <dbl>,
## #   ORB <dbl>, DRB <dbl>, RPG <dbl>, APG <dbl>, SPG <dbl>, BPG <dbl>,
## #   PPG <dbl>, `TS%` <dbl>, `eFG%` <dbl>, `Total S %` <dbl>, `ORB%` <dbl>,
## #   `DRB%` <dbl>, `TRB%` <dbl>, `AST%` <dbl>, `TOV%` <dbl>, `STL%` <dbl>,
## #   `BLK%` <dbl>, `USG%` <dbl>, PPR <dbl>, PPS <dbl>, ORtg <dbl>,
## #   DRtg <dbl>, eDiff <dbl>, FIC <dbl>, PER <dbl>
```

```
head(ACB_Teams_18_19, n=20)
```

```
## # A tibble: 18 x 40
##    ...1  Team  initials    GP   MPG   FGM   FGA `FG%` `3PM` `3PA` `3P%`
##    <chr> <chr> <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 1     Cafe~ BRE         34  40.3  28.1  65.6 0.428   7.9  24.7 0.322
##  2 2     Delt~ GBC         34  40.4  27.2  62.1 0.438   8.3  24.7 0.335
##  3 3     Divi~ JOV         34  40.1  29    60.2 0.481   9.4  25.1 0.377
##  4 4     FC B~ FCB         34  40.1  30.9  61.6 0.501  10.2  24.5 0.418
##  5 5     Herb~ HGC         34  40.1  30.2  65.7 0.46    9.7  27.4 0.354
##  6 6     Iber~ TEN         34  40.9  28.6  60.3 0.474  11.2  29.5 0.378
##  7 7     ICL ~ MAN         34  40.1  28.1  63.1 0.446   9.3  26.5 0.352
##  8 8     KIRO~ BKN         34  40    32    63.4 0.505   8.9  23.8 0.372
##  9 9     Mont~ FUE         34  40.7  28.9  64   0.452   8.6  25   0.345
## 10 10    Mora~ AND         34  40    28.7  64.5 0.446   9.6  27.6 0.348
## 11 11    Movi~ EST         34  40.1  29.6  64.5 0.459   9.2  25.9 0.355
## 12 12    Real~ RMA         34  40.1  31.3  63.6 0.492  10.4  27.7 0.377
## 13 13    Rio ~ OBR         34  40.3  26.5  59.9 0.443  10.9  29.5 0.369
## 14 14    San ~ BUR         34  40    29.9  62.6 0.478   8.7  24.3 0.358
## 15 15    Tecn~ ZAR         34  40.3  31    67.2 0.461   7.8  21.9 0.358
## 16 16    UCAM~ MUR         34  40.3  28    63.1 0.444   9.2  25.5 0.359
## 17 17    Unic~ UNI         34  40.3  29.1  62.1 0.469  10.8  28.9 0.374
```

```
## 18 18    Vale~ VAL         34  40    29.4  61   0.482  10.6  27.6 0.384
## # ... with 29 more variables: FTM <dbl>, FTA <dbl>, `FT%` <dbl>,
## #   TOV <dbl>, PF <dbl>, ORB <dbl>, DRB <dbl>, RPG <dbl>, APG <dbl>,
## #   SPG <dbl>, BPG <dbl>, PPG <dbl>, `TS%` <dbl>, `eFG%` <dbl>, `Total
## #   S%` <dbl>, `ORB%` <dbl>, `DRB%` <dbl>, `TRB%` <dbl>, `AST%` <dbl>,
## #   `TOV%` <dbl>, `STL%` <dbl>, `BLK%` <dbl>, PPS <dbl>, FIC40 <dbl>,
## #   ORtg <dbl>, DRtg <dbl>, eDiff <dbl>, Poss <dbl>, Pace <dbl>
```

```
head(ACB_Players_2012to2018, n=100)
```

```
## # A tibble: 100 x 46
##    Player Team  Team_full Pos   `Height(ft)` `Weight(lb)` Nationality
##    <chr>  <chr> <chr>     <chr> <chr>        <chr>        <chr>
##  1 Olive~ BBB   RETAbet ~ FC    6-10         220          Serbia
##  2 Tomas~ BBB   RETAbet ~ C     7-1          240          Czech Repu~
##  3 Kosta~ BBB   RETAbet ~ F     6-7          225          Greece
##  4 Josh ~ BBB   RETAbet ~ G     6-2          200          United Sta~
##  5 Alex ~ BBB   RETAbet ~ SF    6-7          220          Spain
##  6 Marko~ BBB   RETAbet ~ SF    6-8          250          Croatia
##  7 Dimit~ BBB   RETAbet ~ C     6-10         265          Greece
##  8 Raul ~ BBB   RETAbet ~ PG    6-0          175          Spain
##  9 Aaron~ BBB   RETAbet ~ G     6-4          185          United Sta~
## 10 Janis~ BBB   RETAbet ~ SG    6-2          190          Latvia
## # ... with 90 more rows, and 39 more variables: temporada <chr>, GP <dbl>,
## #   MPG <dbl>, FGM <dbl>, FGA <dbl>, `FG%` <dbl>, `3PM` <dbl>,
## #   `3PA` <dbl>, `3P%` <dbl>, FTM <dbl>, FTA <dbl>, `FT%` <dbl>,
## #   TOV <dbl>, PF <dbl>, ORB <dbl>, DRB <dbl>, RPG <dbl>, APG <dbl>,
## #   SPG <dbl>, BPG <dbl>, PPG <dbl>, `TS%` <dbl>, `eFG%` <dbl>,
## #   `TotalS%` <dbl>, `ORB%` <dbl>, `DRB%` <dbl>, `TRB%` <dbl>,
## #   `AST%` <dbl>, `TOV%` <dbl>, `STL%` <dbl>, `BLK%` <dbl>, `USG%` <dbl>,
## #   PPR <dbl>, PPS <dbl>, ORtg <dbl>, DRtg <dbl>, eDiff <dbl>, FIC <dbl>,
## #   PER <dbl>
```

Desde los tres dataset limpiamos la primera columna correspondiente a la numeración de filas.

Data Cleaning

Durante la fase de limpieza nos centraremos en la observación de valores que podrían distorsionar nuestros análisis

Eliminar variables

```
# Utilizamos la técnica de selección de todas las observaciones y solo incluyendo desde la 2 columna has
ACB_Players_18_19 <- ACB_Players_18_19[,2:47]
ACB_Teams_18_19 <- ACB_Teams_18_19[,2:40]
ACB_Players_2012to2018 <- ACB_Players_2012to2018[,2:46]
```

Detectar valores nulos

Observaremos con las funciones is.na, is.null, is.nan si existen valores a tratar

```
# Utilizamos las tres funciones al primer dataset
table(is.null(ACB_Players_18_19))
```

```
##
```

```
## FALSE
##     1
```

```
table(is.na(ACB_Players_18_19))
```

```
##
## FALSE
## 12696
```

```
# Utilizamos las tres funciones al segundo dataset
table(is.null(ACB_Teams_18_19))
```

```
##
## FALSE
##     1
```

```
table(is.na(ACB_Teams_18_19))
```

```
##
## FALSE
##   702
```

```
# Utilizamos las tres funciones al segundo dataset
table(is.null(ACB_Players_2012to2018))
```

```
##
## FALSE
##     1
```

```
table(is.na(ACB_Players_2012to2018))
```

```
##
## FALSE
## 81675
```