Práctica 2: Limpieza y análisis de datos

UOC - Tipología y ciclo de vida de los datos

Miguel Santos Pérez y Alejandro Arzola García

17 de junio de 2020

Índice

1 Descripción del dataset

FALTA POR HACER

2 Integración y selección de los datos de interés a analizar

FALTA POR HACER

2.1 Carga de datos originales

Procedemos a cargar el fichero csv *euroleague_stats_per_game.csv* que contiene el dataset con los datos individuales por partido:

```
data <- read.csv2('../csv/euroleague_stats_per_game.csv')
```

Hemos utilizado la función `read.csv2` ya que el fichero a cargar está en formato *csv* español, es decir, los separadores son el caracter `;`.

Mostramos los primeros registros del dataset para verificar que los datos se han cargado correctamente:

```
head(data)
```

```
##   MatchId                Team PlayerNumber          PlayerName   Min Pts X2FG
## 1       1 Khimki Moscow Region            1       SHVED, ALEXEY 30:19  22  4/7
## 2       1 Khimki Moscow Region            5       BOOKER, DEVIN 19:21   4  1/4
## 3       1 Khimki Moscow Region            6        TIMMA, JANIS 33:53  17  1/3
## 4       1 Khimki Moscow Region            8 ZAYTSEV, VYACHESLAV 10:23   0    0
## 5       1 Khimki Moscow Region            9      VIALTSEV, EGOR  4:58   0    0
## 6       1 Khimki Moscow Region           10 DESIATNIKOV, ANDREI   DNP   -    -
##   X3FG   FT O D T As St To Fv Ag Cm Rv PIR
## 1 2/10 8/10 0 2 2  6  0  3  1  0  1  5  19
## 2    0  2/2 3 3 6  0  2  1  0  0  2  1   7
## 3 5/12    0 1 3 4  1  4  1  0  0  5  2  13
## 4    0    0 1 1 2  2  1  1  0  0  2  1   3
## 5    0    0 0 0 0  1  0  0  0  0  2  0  -1
## 6    -    - - - -  -  -  -  -  -  -  -   -
```

Mostramos el número de registros transmitidos:

```
nrow(data)
```

```
## [1] 6505
```

3 Limpieza de los datos

FALTA POR HACER ## Elementos vacíos ## Identificación y tratamiento de valores extremos

4 Análisis de los datos

FALTA POR HACER

4.1 Selección de los grupos de datos que se quieren analizar/comparar

4.2 Comprobación de la normalidad y homogeneidad de la varianza

4.3 Pruebas estadísticas para comparar los grupos de datos

4.3.1 Contraste de hipótesis

4.3.2 Correlaciones

4.3.3 Regresiones

5 Exportación de datos finales

A continuación vamos a exportar nuestro *dataframe* final a un archivo csv. Este archivo se llamará *euroleague_stats_per_game_clean.csv*. Utilizamos la función `write.csv2()` para exportar el fichero en formato *csv* español:

```
write.csv2(data, row.names = TRUE, file = "../csv/euroleague_stats_per_game_clean.csv")
```

Este nuevo dataset está disponible, al igual que el resto de la práctica, en el siguiente repositorio de *GitHub*:

- https://github.com/aarzola-uoc/practica2-tycvd

6 Representación de los resultados a partir de tablas y gráficas

FALTA POR HACER

7 Resolución del problema

FALTA POR HACER

8 Contribuciones al trabajo

Contribuciones	**Firma**
Selección del dataset	MSP, AAG
Creación del repositorio GitHub	MSP, AAG
Desarrollo código en R	MSP, AAG
Redacción de las respuestas	MSP, AAG