PRA 2. Limpieza y análisis de datos

Contents

Descripción del dataset. **1**

Integración y selección de los datos de interés a analizar **2**

Limpieza de datos **2**
Elementos vacíos y cero. 2
Valores extremos. 3

Análisis de los datos. **3**
Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). 3
Comprobación de la normalidad y homogeneidad de la varianza. 3
Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes. 3

Representación de los resultados a partir de tablas y gráficas. **3**

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? **3**

Descripción del dataset.

El dataset elegido corresponde al generado en la PRAC1. El dataset contiene los resultados obtenidos por distintos equipos profesionales de fútbol en diversas competiciones a lo largo de la última década.

Nos serviremos de estos datos para anlizar los diferentes clubs de fútbol a nivel global así como para buscar patrones entre ellos.

```
df <- read.csv(file = 'football_competitions.csv')
head(df)
```

```
##                       competition       pl_team pl_pi pl_w pl_d pl_l pl_f pl_a
## 1 African Nations - Group A 2008       Ghana *     3    3    0    0    5    1
## 2 African Nations - Group A 2008      Guinea *     3    1    1    1    5    5
## 3 African Nations - Group A 2008     Morocco *     3    1    0    2    7    6
## 4 African Nations - Group A 2008     Namibia *     3    0    1    2    2    7
## 5 African Nations - Group B 2008 Ivory Coast *     3    3    0    0    8    1
## 6 African Nations - Group B 2008     Nigeria *     3    1    1    1    2    1
##   pl_gd pl_pts
## 1     4      9
## 2     0      4
## 3     1      3
## 4    -5      1
## 5     7      9
## 6     1      4
```

```r
str(df)
```

```
## 'data.frame':    6623 obs. of  10 variables:
##  $ competition: Factor w/ 825 levels "Africa Cup of Nations - Group A 2010",..: 19 19 19 19 20 20 20
##  $ pl_team    : Factor w/ 993 levels "1. FC Union Berlin",..: 398 419 616 624 493 643 579 103 298 171
##  $ pl_pi      : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ pl_w       : int  3 1 1 0 3 1 1 0 2 2 ...
##  $ pl_d       : int  0 1 0 1 0 1 1 0 1 0 ...
##  $ pl_l       : int  0 1 2 2 0 1 1 3 0 1 ...
##  $ pl_f       : int  5 5 7 2 8 2 1 1 8 10 ...
##  $ pl_a       : int  1 5 6 7 1 1 3 7 3 5 ...
##  $ pl_gd      : int  4 0 1 -5 7 1 -2 -6 5 5 ...
##  $ pl_pts     : int  9 4 3 1 9 4 4 0 7 6 ...
```

El fichero contiene una tabla con diez columnas, siendo las dos primeras de tipo cadena y el resto de tipo entero.

Los campos son:

- competition: Nombre de la competición, año y grupo dentro de los cuales el equipo indicado obtuvo los resultados indicados.
- pl_team: Nombre del equipo.
- pl_pi: Número de partidos jugados.
- pl_w: Número de partidos ganados.
- pl_d: Número de partidos empatados.
- pl_l: Número de partidos perdidos.
- pl_f: Número de goles a favor.
- pl_a: Número de goles en contra.
- pl_gd: Diferencia de goles.
- pl_pts: Puntos obtenidos por el equipo.

Integración y selección de los datos de interés a analizar

todo: hablar que análisis vamos a hacer para poder hacer este apartado

Limpieza de datos

Elementos vacíos y cero.

Primero vamos a ver si hay valores vacíos en el dataset.

```r
sapply(df, function(x) sum(is.na(x)))
```

```
## competition     pl_team       pl_pi        pl_w        pl_d        pl_l
##           0           0           0           0           0           0
##        pl_f        pl_a       pl_gd      pl_pts
##           0           0           0           0
```

No hay, buscamos valores 0.

```r
sum(df==0)
```

```
## [1] 4384
```

Hay muchos datos que son igual a cero. Debido a la naturaleza del dataframe no es necesario quitarlos, porque puede ser que un equipo tenga 0 puntos, victorias, derrotas, etc.

Valores extremos.

Para la detección de outliers o valores extremos utilizamos la función boxplots.stats(), en concreto el atributo "out" que nos devuelve los valores que distan mucho del rango intercuartílico que se dibujan en los diagramas de caja.

```
boxplot.stats(df$pl_pi)$out
```

```
## integer(0)
```

```
boxplot.stats(df$pl_w)$out
```

```
## [1] 34 33 34 33
```

```
boxplot.stats(df$pl_d)$out
```

```
## integer(0)
```

```
boxplot.stats(df$pl_l)$out
```

```
## [1] 35 38
```

```
boxplot.stats(df$pl_f)$out
```

```
## integer(0)
```

```
boxplot.stats(df$pl_a)$out
```

```
## [1] 143 140
```

A penas hay valores outliers y tras comprobarlos, vemos que efectivamente hay equipos que han ganado más de 30 partidos en una comoptición y lo mismo ocurre con las derrotas y los goles en contra.

Análisis de los datos.

Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

P_w, P_l, p_d

Comprobación de la normalidad y homogeneidad de la varianza.

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Aquí puede ir estadísticas de los equipos más destacados y un clustering de equipos

Representación de los resultados a partir de tablas y gráficas.

plots de los equipos y un plot del clustering con sus fronteras de decisión

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?