Postwork T21

Equipo 21

Postwork 1

Objetivo

El Postwork tiene como objetivo que practiques los comandos básicos aprendidos durante la sesión, de tal modo que sirvan para reafirmar el conocimiento. Recuerda que la programación es como un deporte en el que se debe practicar, habrá caídas, pero lo importante es levantarse y seguir adelante. Éxito

Desarrollo

El siguiente postwork, te servirá para ir desarrollando habilidades como si se tratara de un proyecto que evidencie el progreso del aprendizaje durante el módulo, sesión a sesión se irá desarrollando. A continuación aparecen una serie de objetivos que deberás cumplir, es un ejemplo real de aplicación y tiene que ver con datos referentes a equipos de la liga española de fútbol (recuerda que los datos provienen siempre de diversas naturalezas), en este caso se cuenta con muchos datos que se pueden aprovechar, explotarlos y generar análisis interesantes que se pueden aplicar a otras áreas. Siendo así damos paso a las instrucciones:

Importa los datos de soccer de la temporada 2019/2020 de la primera división de la liga española a R, los datos los puedes encontrar en el siguiente enlace:

```
fut<-read.csv("https://www.football-data.co.uk/mmz4281/1920/SP1.csv")
df<-data.frame(fut)
head(df)
```

```
##   Div       Date  Time   HomeTeam    AwayTeam FTHG FTAG FTR HTHG HTAG HTR HS AS
## 1 SP1 16/08/2019 20:00 Ath Bilbao   Barcelona    1    0   H    0    0   D 11 11
## 2 SP1 17/08/2019 16:00      Celta Real Madrid    1    3   A    0    1   A  7 17
## 3 SP1 17/08/2019 18:00   Valencia    Sociedad    1    1   D    0    0   D 14 12
## 4 SP1 17/08/2019 19:00   Mallorca       Eibar    2    1   H    1    0   H 16 11
## 5 SP1 17/08/2019 20:00    Leganes     Osasuna    0    1   A    0    0   D 13  4
## 6 SP1 17/08/2019 20:00 Villarreal     Granada    4    4   D    1    1   D 12 14
##   HST AST HF AF HC AC HY AY HR AR B365H B365D B365A  BWH  BWD  BWA  IWH  IWD
## 1   5   2 14  9  3  8  1  1  0  0  5.25  3.80  1.65 5.50 3.80 1.65 5.00 3.80
## 2   4  11 17 12  6  4  5  2  0  1  4.75  4.20  1.65 4.40 4.20 1.72 5.30 4.20
## 3   6   3 13 14  3  3  4  4  1  0  1.66  3.75  5.50 1.67 3.75 5.50 1.67 3.75
## 4   4   5 13 14  9  3  2  3  0  0  2.80  3.20  2.60 2.95 3.10 2.60 2.90 3.10
## 5   2   2 17 11  8  0  1  4  1  0  2.00  3.20  4.20 2.05 3.25 3.90 2.05 3.10
## 6   7   7 10 16  2  7  3  1  0  0  1.60  3.80  6.50 1.60 3.80 6.25 1.63 4.00
##    IWA  PSH  PSD  PSA  WHH WHD  WHA  VCH  VCD  VCA MaxH MaxD MaxA AvgH AvgD
## 1 1.70 5.15 3.84 1.74 5.00 3.8 1.70 5.00 3.80 1.75 5.50 3.95 1.76 5.07 3.81
## 2 1.60 4.73 4.18 1.72 5.25 4.2 1.60 4.75 4.20 1.73 5.30 4.40 1.73 4.67 4.12
## 3 5.30 1.68 3.94 5.47 1.67 3.8 5.25 1.67 3.90 5.75 1.72 3.98 5.75 1.68 3.80
## 4 2.60 2.98 3.14 2.66 2.90 3.1 2.62 2.90 3.13 2.70 3.05 3.20 2.70 2.91 3.09
```

```
## 5 4.05 2.10 3.21 4.13 2.05 3.2 4.00 2.10 3.20 4.10 2.10 3.30 4.25 2.06 3.18
## 6 5.50 1.62 3.99 6.13 1.60 3.9 5.80 1.65 4.00 5.75 1.65 4.15 6.50 1.61 3.95
##   AvgA B365.2.5 B365.2.5.1 P.2.5 P.2.5.1 Max.2.5 Max.2.5.1 Avg.2.5 Avg.2.5.1
## 1 1.71     1.80       2.00  1.81    2.09    1.85      2.11    1.79      2.05
## 2 1.69     1.53       2.50  1.52    2.66    1.53      2.72    1.49      2.58
## 3 5.29     2.00       1.80  2.08    1.82    2.14      1.83    2.07      1.77
## 4 2.62     2.30       1.61  2.45    1.60    2.47      1.65    2.34      1.60
## 5 4.02     2.50       1.53  2.72    1.50    2.75      1.54    2.59      1.49
## 6 5.80     1.80       2.00  1.88    2.02    1.90      2.05    1.84      1.98
##     AHh B365AHH B365AHA PAHH PAHA MaxAHH MaxAHA AvgAHH AvgAHA B365CH B365CD
## 1  0.75    1.99    1.94 1.98 1.94   2.00   1.95   1.96   1.92   5.25   3.80
## 2  0.75    2.04    1.89 2.01 1.91   2.05   1.91   2.00   1.88   5.25   4.20
## 3 -0.75    1.91    2.02 1.91 2.01   1.93   2.03   1.89   1.99   1.66   3.75
## 4  0.00    2.05    1.88 2.07 1.85   2.07   1.88   2.04   1.85   2.87   3.20
## 5 -0.50    2.08    1.85 2.10 1.82   2.10   1.85   2.06   1.83   1.90   3.10
## 6 -1.00    2.05    1.75 2.11 1.81   2.14   1.85   2.07   1.80   1.53   4.00
##   B365CA BWCH BWCD BWCA IWCH IWCD IWCA PSCH PSCD PSCA WHCH WHCD WHCA VCCH VCCD
## 1   1.65 4.75 3.75 1.75 5.00 3.80 1.70 5.34 3.62 1.78 5.00  3.8 1.70 4.80 3.80
## 2   1.57 4.50 4.10 1.70 4.60 3.80 1.75 5.10 4.46 1.65 5.00  4.2 1.63 5.20 4.40
## 3   5.50 1.65 3.80 5.50 1.67 3.80 5.30 1.69 3.88 5.47 1.65  3.9 5.25 1.70 3.90
## 4   2.55 2.95 3.10 2.60 2.90 3.10 2.60 2.96 3.26 2.60 2.90  3.1 2.60 3.00 3.13
## 5   5.00 1.95 3.20 4.50 1.90 3.15 4.85 1.90 3.18 5.30 2.05  3.2 4.00 1.90 3.20
## 6   6.50 1.57 3.80 6.50 1.55 4.05 6.30 1.54 4.19 6.87 1.62  3.9 5.80 1.57 4.00
##   VCCA MaxCH MaxCD MaxCA AvgCH AvgCD AvgCA B365C.2.5 B365C.2.5.1 PC.2.5
## 1 1.80  5.80  3.90  1.81  5.03  3.66  1.76      1.90        1.90   1.98
## 2 1.65  6.00  4.52  1.75  4.93  4.26  1.65      1.44        2.75   1.49
## 3 5.50  1.72  3.95  6.20  1.68  3.82  5.37      2.00        1.80   2.06
## 4 2.63  3.05  3.29  2.72  2.93  3.14  2.59      2.20        1.66   2.20
## 5 5.20  1.95  3.26  5.30  1.90  3.16  4.91      2.75        1.44   2.84
## 6 7.00  1.58  4.20  7.30  1.54  4.05  6.66      1.90        1.90   1.95
##   PC.2.5.1 MaxC.2.5 MaxC.2.5.1 AvgC.2.5 AvgC.2.5.1  AHCh B365CAHH B365CAHA
## 1     1.93     1.99       2.11     1.86       1.97  0.75     1.93     2.00
## 2     2.76     1.51       2.88     1.47       2.63  1.00     1.82     1.97
## 3     1.85     2.08       1.98     2.00       1.82 -0.75     1.94     1.99
## 4     1.74     2.38       1.74     2.24       1.66  0.00     2.11     1.82
## 5     1.47     2.85       1.50     2.69       1.46 -0.50     1.89     2.04
## 6     1.95     1.98       2.10     1.90       1.92 -1.00     1.96     1.97
##   PCAHH PCAHA MaxCAHH MaxCAHA AvgCAHH AvgCAHA
## 1  1.91  2.01    2.02    2.03    1.91    1.98
## 2  1.85  2.07    2.00    2.20    1.82    2.06
## 3  1.92  2.00    1.96    2.12    1.89    2.00
## 4  2.09  1.83    2.12    1.88    2.07    1.83
## 5  1.90  2.01    1.95    2.06    1.90    1.99
## 6  1.96  1.96    1.98    2.12    1.93    1.95
```

Del data frame que resulta de importar los datos a R, extrae las columnas que contienen los números de goles anotados por los equipos que jugaron en casa (FTHG) y los goles anotados por los equipos que jugaron como visitante (FTAG)

```
df1<-as.data.frame(cbind(goleslocal=df$FTHG,golesvisita=df$FTAG))
dflocal<-df1$goleslocal
dfvisita<-df1$golesvisita
```

```
head(df1)
```

```
##   goleslocal golesvisita
## 1          1           0
## 2          1           3
## 3          1           1
## 4          2           1
## 5          0           1
## 6          4           4
```

Consulta cómo funciona la función table en R al ejecutar en la consola ?table

```
local<-table(dflocal)
visita<-table(dfvisita)
partidos<-table(df1)
```

```
print('Local')
```

```
## [1] "Local"
```

```
local
```

```
## dflocal
##   0   1   2   3   4   5   6
##  88 132  99  38  14   8   1
```

```
print('Visita')
```

```
## [1] "Visita"
```

```
visita
```

```
## dfvisita
##   0   1   2   3   4   5
## 136 134  81  18   9   2
```

```
print('Partidos')
```

```
## [1] "Partidos"
```

```
partidos
```

```
##           golesvisita
## goleslocal  0  1  2  3  4  5
##          0 33 28 15  8  2  2
##          1 43 49 32  5  3  0
##          2 39 35 20  3  2  0
##          3 14 14  7  2  1  0
##          4  4  5  4  0  1  0
##          5  2  3  3  0  0  0
##          6  1  0  0  0  0  0
```

Posteriormente elabora tablas de frecuencias relativas para estimar las siguientes probabilidades:

La probabilidad (marginal) de que el equipo que juega en casa anote x goles (x = 0, 1, 2, . . .)

```
hg<-prop.table(local)
print('Probabilidad goles de local')
```

```
## [1] "Probabilidad goles de local"
```

```
hg
```

```
## dflocal
##           0           1           2           3           4           5
## 0.231578947 0.347368421 0.260526316 0.100000000 0.036842105 0.021052632
##           6
## 0.002631579
```

La probabilidad (marginal) de que el equipo que juega como visitante anote y goles (y = 0, 1, 2, . . .)

```
ag<-prop.table(visita)
print('Probabilidad goles de visita')
```

```
## [1] "Probabilidad goles de visita"
```

```
ag
```

```
## dfvisita
##           0           1           2           3           4           5
## 0.357894737 0.352631579 0.213157895 0.047368421 0.023684211 0.005263158
```

La probabilidad (conjunta) de que el equipo que juega en casa anote x goles y el equipo que juega como visitante anote y goles (x = 0, 1, 2, . . . , y = 0, 1, 2, . . .)

```
tg<-prop.table(partidos)
print('Probabilidad conjunta')
```

```
## [1] "Probabilidad conjunta"
```

```
tg
```

```
##           golesvisita
## goleslocal           0           1           2           3           4
##          0 0.086842105 0.073684211 0.039473684 0.021052632 0.005263158
##          1 0.113157895 0.128947368 0.084210526 0.013157895 0.007894737
##          2 0.102631579 0.092105263 0.052631579 0.007894737 0.005263158
##          3 0.036842105 0.036842105 0.018421053 0.005263158 0.002631579
##          4 0.010526316 0.013157895 0.010526316 0.000000000 0.002631579
##          5 0.005263158 0.007894737 0.007894737 0.000000000 0.000000000
##          6 0.002631579 0.000000000 0.000000000 0.000000000 0.000000000
##           golesvisita
```

```
## goleslocal          5
##          0 0.005263158
##          1 0.000000000
##          2 0.000000000
##          3 0.000000000
##          4 0.000000000
##          5 0.000000000
##          6 0.000000000
```

Postwork 2

Objetivo

- Importar múltiples archivos csv a R
- Observar algunas características y manipular los data frames
- Combinar múltiples data frames en un único data frame

Desarrollo

Ahora vamos a generar un cúmulo de datos mayor al que se tenía, esta es una situación habitual que se puede presentar para complementar un análisis, siempre es importante estar revisando las características o tipos de datos que tenemos, por si es necesario realizar alguna transformación en las variables y poder hacer operaciones aritméticas si es el caso, además de sólo tener presente algunas de las variables, no siempre se requiere el uso de todas para ciertos procesamientos.

Importa los datos de soccer de las temporadas 2017/2018, 2018/2019 y 2019/2020 de la primera división de la liga española a R, los datos los puedes encontrar en el siguiente enlace:

```
liga_19_20 <- "https://www.football-data.co.uk/mmz4281/1920/SP1.csv"
liga_18_19 <- "https://www.football-data.co.uk/mmz4281/1819/SP1.csv"
liga_17_18 <- "https://www.football-data.co.uk/mmz4281/1718/SP1.csv"
```

Cambiamos directorio de trabajo a directorio donde se encuentra el script. Creamos directorio para guardar los archivos CSV y lo usamos como directorio de trabajo, descargamos los datasets.

Cargamos los datasets

```
d<-c("liga_17-18.csv","liga_18-19.csv","liga_19-20.csv")
datasets <- lapply(d, read.csv)
```

Obten una mejor idea de las características de los data frames al usar las funciones: str, head, View y summary

```
summary(datasets)
```

```
##      Length Class      Mode
## [1,]  64    data.frame list
## [2,]  61    data.frame list
## [3,] 105    data.frame list
```

Con la función select del paquete dplyr selecciona únicamente las columnas Date, HomeTeam, AwayTeam, FTHG, FTAG y FTR; esto para cada uno de los data frames. (Hint: también puedes usar lapply).

```
datasets <- lapply(datasets, select, Date, HomeTeam, AwayTeam, FTHG, FTAG, FTR)
```

Asegúrate de que los elementos de las columnas correspondientes de los nuevos data frames sean del mismo tipo (Hint 1: usa as.Date y mutate para arreglar las fechas). Con ayuda de la función rbind forma un único data frame que contenga las seis columnas mencionadas en el punto 3 (Hint 2: la función do.call podría ser utilizada).

```
# Transformamos las fechas al formato indicado.
datasets <- lapply(datasets, mutate, Date = as.Date(Date, "%d/%m/%Y"))
```

```
# Unimos todos los dataframes en uno solo.
dataset <- do.call(rbind, datasets)
head(dataset)
```

```
##         Date   HomeTeam   AwayTeam FTHG FTAG FTR
## 1 0017-08-18    Leganes     Alaves    1    0   H
## 2 0017-08-18   Valencia Las Palmas    1    0   H
## 3 0017-08-19      Celta   Sociedad    2    3   A
## 4 0017-08-19     Girona Ath Madrid    2    2   D
## 5 0017-08-19    Sevilla    Espanol    1    1   D
## 6 0017-08-20 Ath Bilbao     Getafe    0    0   D
```

```
print('Dimensión')
```

```
## [1] "Dimensión"
```

```
dim(dataset)
```

```
## [1] 1140    6
```

Postwork 3

Objetivo

- Realizar descarga de archivos desde internet
- Generar nuevos data frames
- Visualizar probabilidades estimadas con la ayuda de gráficas

Desarrollo

Ahora graficaremos probabilidades (estimadas) marginales y conjuntas para el número de goles que anotan en un partido el equipo de casa o el equipo visitante.

Con el último data frame obtenido en el postwork de la sesión 2, elabora tablas de frecuencias relativas para estimar las siguientes probabilidades: La probabilidad (marginal) de que el equipo que juega en casa anote x goles (x=0,1,2,)

```
allocal<-table(dataset$FTHG)
plocal<-prop.table(allocal)
plocal<-as.data.frame(plocal)
plocal
```

```
##   Var1        Freq
## 1    0 0.232456140
## 2    1 0.327192982
## 3    2 0.266666667
## 4    3 0.112280702
## 5    4 0.035087719
## 6    5 0.019298246
## 7    6 0.005263158
## 8    7 0.000877193
## 9    8 0.000877193
```

La probabilidad (marginal) de que el equipo que juega como visitante anote y goles (y=0,1,2,)

```
allvisita<-table(dataset$FTAG)
pvisita<-prop.table(allvisita)
pvisita<-as.data.frame(pvisita)
pvisita
```

```
##   Var1        Freq
## 1    0 0.351754386
## 2    1 0.340350877
## 3    2 0.212280702
## 4    3 0.054385965
## 5    4 0.028947368
## 6    5 0.009649123
## 7    6 0.002631579
```

La probabilidad (conjunta) de que el equipo que juega en casa anote x goles y el equipo que juega como visitante anote y goles (x=0,1,2,, y=0,1,2,)

```
allmatch<-as.data.frame(cbind(Goles_Local =dataset$FTHG, Goles_Visita =dataset$FTAG))
allmatch<-table(allmatch)
pallmatch<-prop.table(allmatch)
pallmatch<-as.data.frame(pallmatch)
pallmatch
```

```
##    Goles_Local Goles_Visita        Freq
## 1            0            0 0.078070175
## 2            1            0 0.115789474
## 3            2            0 0.087719298
## 4            3            0 0.044736842
## 5            4            0 0.014035088
## 6            5            0 0.008771930
## 7            6            0 0.002631579
## 8            7            0 0.000000000
## 9            8            0 0.000000000
## 10           0            1 0.080701754
## 11           1            1 0.114912281
## 12           2            1 0.093859649
## 13           3            1 0.032456140
## 14           4            1 0.010526316
## 15           5            1 0.005263158
```

```
## 16          6          1 0.001754386
## 17          7          1 0.000877193
## 18          8          1 0.000000000
## 19          0          2 0.045614035
## 20          1          2 0.068421053
## 21          2          2 0.061403509
## 22          3          2 0.024561404
## 23          4          2 0.007017544
## 24          5          2 0.004385965
## 25          6          2 0.000000000
## 26          7          2 0.000000000
## 27          8          2 0.000877193
## 28          0          3 0.018421053
## 29          1          3 0.017543860
## 30          2          3 0.011403509
## 31          3          3 0.006140351
## 32          4          3 0.000000000
## 33          5          3 0.000000000
## 34          6          3 0.000877193
## 35          7          3 0.000000000
## 36          8          3 0.000000000
## 37          0          4 0.005263158
## 38          1          4 0.008771930
## 39          2          4 0.008771930
## 40          3          4 0.001754386
## 41          4          4 0.003508772
## 42          5          4 0.000877193
## 43          6          4 0.000000000
## 44          7          4 0.000000000
## 45          8          4 0.000000000
## 46          0          5 0.004385965
## 47          1          5 0.001754386
## 48          2          5 0.001754386
## 49          3          5 0.001754386
## 50          4          5 0.000000000
## 51          5          5 0.000000000
## 52          6          5 0.000000000
## 53          7          5 0.000000000
## 54          8          5 0.000000000
## 55          0          6 0.000000000
## 56          1          6 0.000000000
## 57          2          6 0.001754386
## 58          3          6 0.000877193
## 59          4          6 0.000000000
## 60          5          6 0.000000000
## 61          6          6 0.000000000
## 62          7          6 0.000000000
## 63          8          6 0.000000000
```

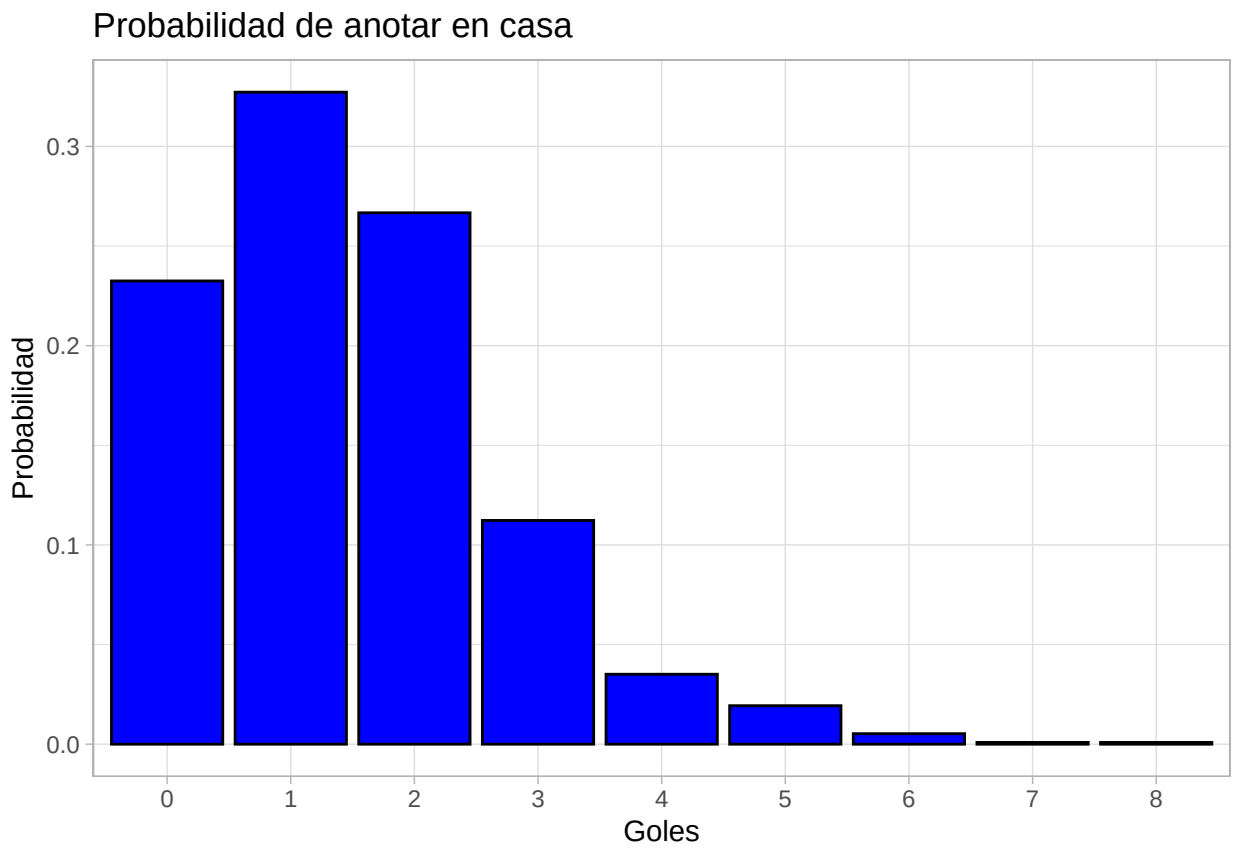
Renombramos las variables

```
plocal<-rename(plocal, Goles = Var1, Probabilidad = Freq)
pvisita<-rename(pvisita, Goles = Var1, Probabilidad = Freq)
pallmatch<-rename(pallmatch, Probabilidad = Freq)
```

Realiza lo siguiente: Un gráfico de barras para las probabilidades marginales estimadas del número de goles que anota el equipo de casa

```
ggplot(plocal, aes(x=Goles, y=Probabilidad)) +
  geom_col(binwidth = 4, col="black", fill = "blue") +
  ggtitle("Probabilidad de anotar en casa") +
  theme_light()
```

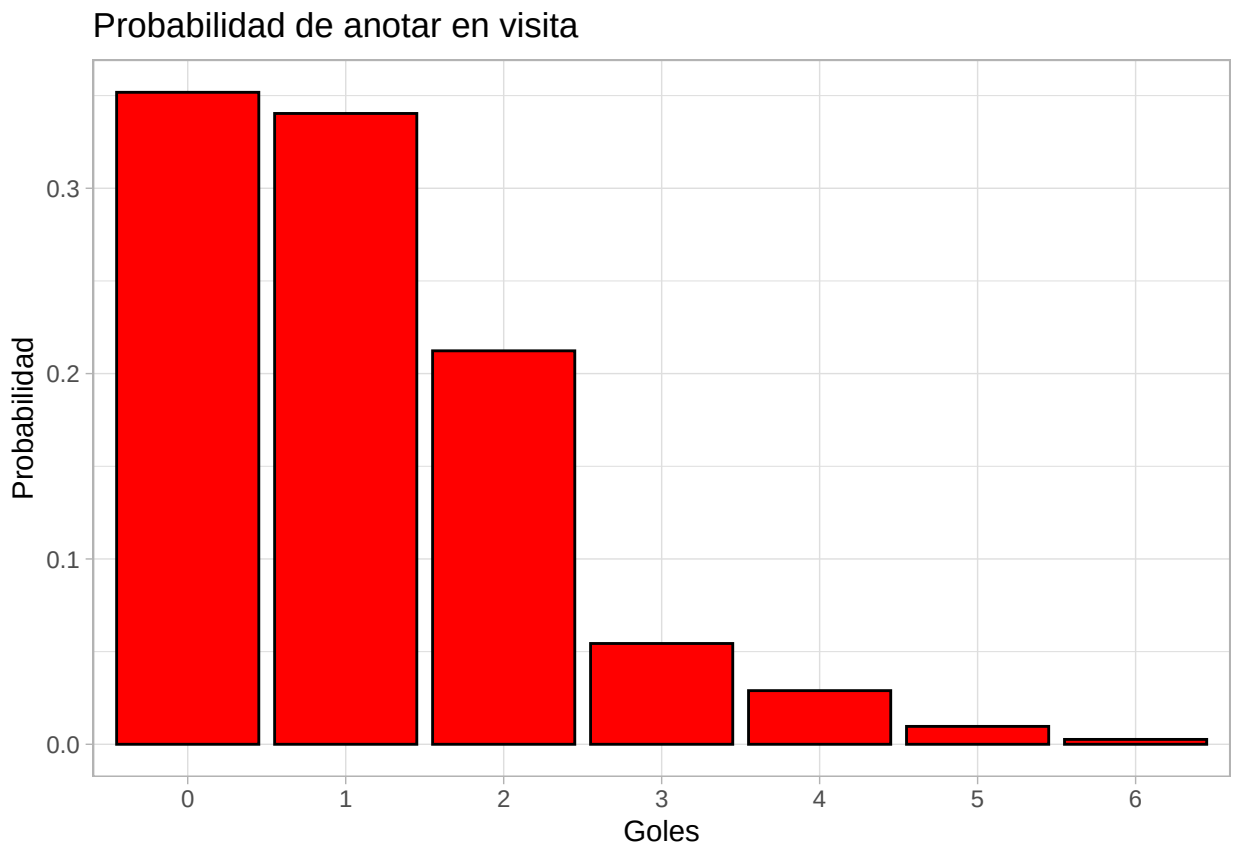
```
## Warning: Ignoring unknown parameters: binwidth
```



Un gráfico de barras para las probabilidades marginales estimadas del número de goles que anota el equipo visitante.

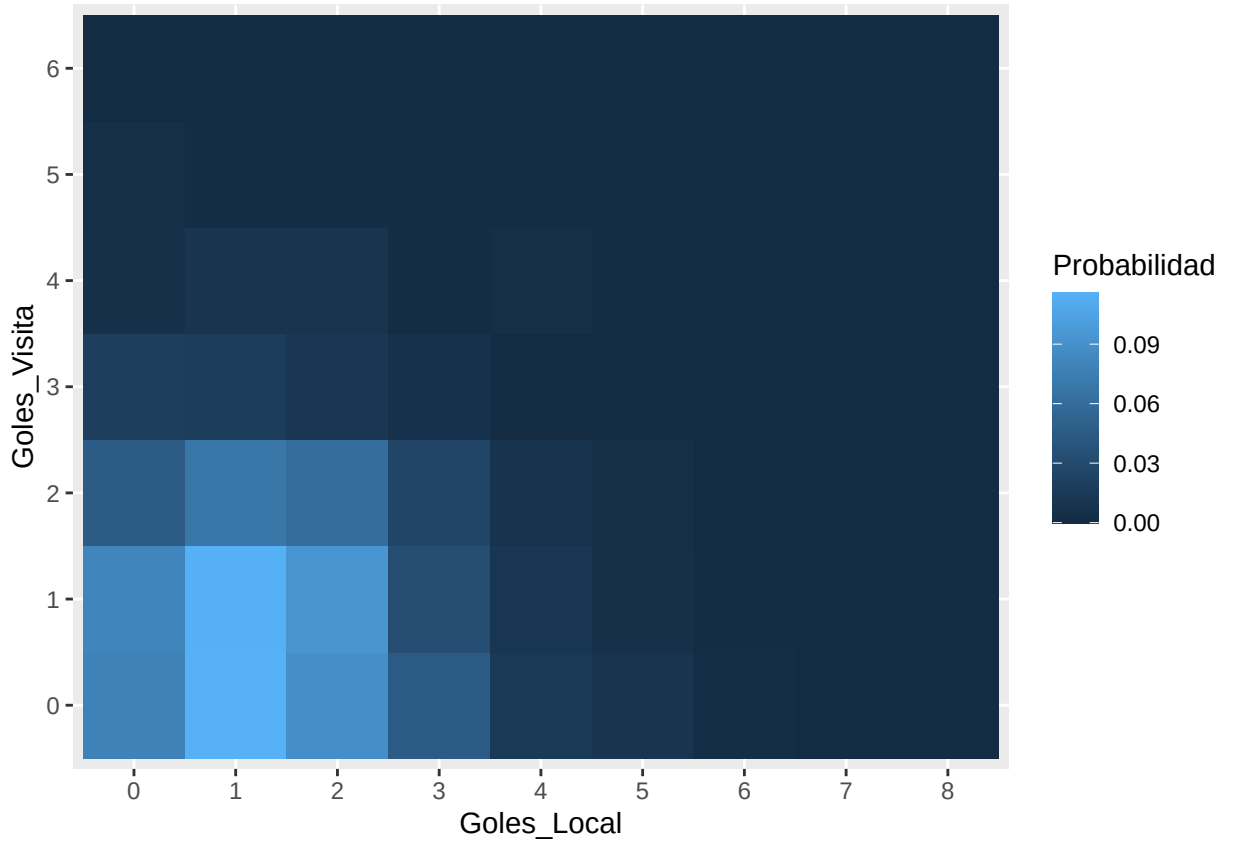
```
ggplot(pvisita, aes(x=Goles, y=Probabilidad)) +
  geom_col(binwidth = 4, col="black", fill = "red") +
  ggtitle("Probabilidad de anotar en visita") +
  theme_light()
```

```
## Warning: Ignoring unknown parameters: binwidth
```



Un HeatMap para las probabilidades conjuntas estimadas de los números de goles que anotan el equipo de casa y el equipo visitante en un partido.

```
ggplot(pallmatch, aes(x=Goles_Local, y=Goles_Visita, fill= Probabilidad)) +
  geom_tile()
```



Postwork 4

Objetivo

Investigar la dependencia o independecia de las variables aleatorias X y Y, el número de goles anotados por el equipo de casa y el número de goles anotados por el equipo visitante.

Desarrollo

Ahora investigarás la dependencia o independencia del número de goles anotados por el equipo de casa y el número de goles anotados por el equipo visitante mediante un procedimiento denominado bootstrap, revisa bibliografía en internet para que tengas nociones de este desarrollo.

Ya hemos estimado las probabilidades conjuntas de que el equipo de casa anote X=x goles (x=0,1,... ,8), y el equipo visitante anote Y=y goles (y=0,1,... ,6), en un partido. Obtén una tabla de cocientes al dividir estas probabilidades conjuntas por el producto de las probabilidades marginales correspondientes.

```
pallmatch<-prop.table(allmatch)
plocal<-prop.table(allocal)
pvisita<-prop.table(allvisita)

pallmatch/outer(plocal, pvisita, "*")
```

```
##            Goles_Visita
## Goles_Local         0         1         2         3         4         5
```

```
##           0 0.9547829 1.0200350 0.9243724 1.4570907 0.7821612 1.9554031
##           1 1.0060639 1.0318952 0.9850885 0.9859033 0.9261516 0.5556910
##           2 0.9351621 1.0341495 1.0847107 0.7862903 1.1363636 0.6818182
##           3 1.1327151 0.8493073 1.0304752 1.0055444 0.5397727 1.6193182
##           4 1.1371571 0.8814433 0.9421488 0.0000000 3.4545455 0.0000000
##           5 1.2922240 0.8013121 1.0706236 0.0000000 1.5702479 0.0000000
##           6 1.4214464 0.9793814 0.0000000 3.0645161 0.0000000 0.0000000
##           7 0.0000000 2.9381443 0.0000000 0.0000000 0.0000000 0.0000000
##           8 0.0000000 0.0000000 4.7107438 0.0000000 0.0000000 0.0000000
##            Goles_Visita
## Goles_Local         6
##           0 0.0000000
##           1 0.0000000
##           2 2.5000000
##           3 2.9687500
##           4 0.0000000
##           5 0.0000000
##           6 0.0000000
##           7 0.0000000
##           8 0.0000000
```

Mediante un procedimiento de boostrap, obtén más cocientes similares a los obtenidos en la tabla del punto anterior. Esto para tener una idea de las distribuciones de la cual vienen los cocientes en la tabla anterior. Menciona en cuáles casos le parece razonable suponer que los cocientes de la tabla en el punto 1, son iguales a 1 (en tal caso tendríamos independencia de las variables aleatorias X y Y).

```
simul<-c()
for (i in 1:1000){
  set.seed(i+132)
  sm <- sample(1:dim(dataset)[1], size = 760, replace = F)
  df<-dataset[sm,]
  al<-table(df$FTHG) ; pl<-prop.table(al)
  av<-table(df$FTAG) ; pv<-prop.table(av)
  am<-as.data.frame(cbind(Goles_Local =df$FTHG, Goles_Visita =df$FTAG)); am<-table(am); pm<-prop.table(a
  simul[[i]]<-  pm/outer(pl, pv, "*")
}
```

```
simul<-lapply(simul, as.data.frame)
simul<-lapply(simul,rename, Cocientes = Freq)
```