Práctica aprendizaje supervisado

Guillermo Bonafonte Criado

15/1/2017

1. Cargamos las librerías necesarias

```
library(htmltab)
library(ggplot2)
```

2. Descargamos los datos

```
myUrl <- "http://www.comoviajar.com/hoteles-
lista.cfm?idzona=149&idpob=0&idcat=0";
dsHoteles<- htmltab(doc = myUrl, which = 3, stringsAsFactors = FALSE);
head(dsHoteles)
```

```
##                                Nombre Cat Hab               Población
## 2           Posada de la Alameda, La  3*  22 28749 Alameda del Valle
## 3   AC Alcala de Henares (AC Hotels)  4*  93 28805 Alcalá de Henares
## 4    Asur Metropol Alcala (Ex. Kris)  3*  59 28802 Alcalá de Henares
## 5                          Bedel, El  3*  50 28801 Alcalá de Henares
## 6 Campanile Madrid-Alcala de Henares  3* 110 28806 Alcalá de Henares
## 7             Cisneros (Ex. Partner)  3*  42 28803 Alcalá de Henares
##                                Dirección
## 2                             Grande, 34
## 3                        Octavio Paz, 25
## 4 Fausto Elhúyar, 9 Area Empres.y de Ocio
## 5                      Plaza San Diego, 6
## 6                       Fausto Elhuyar, 3
## 7                    Paseo de Pastrana, 32
```

3. Preparamos los datos

Comprobamos los tipos

```
sapply(dsHoteles, class)
```

```
##      Nombre         Cat         Hab   Población   Dirección
## "character" "character" "character" "character" "character"
```

Cambiamos tipos

Cambio de la columna Hab

Cambiamos el tipo de la columna Hab, vamos a tener problemas con las filas en las que tenemos el valor N.D.

```
dsHoteles$Hab <- as.numeric(dsHoteles$Hab) # NAs introducidos por
coercion

## Warning: NAs introducidos por coerción
```

Comprobamos las filas que nos han dado problemas

```
which(is.na(dsHoteles$Hab))

## [1]  21  46  67 120 271 276
```

Mostramos el valor del que se ha introducido en las filas anteriores que como vemos es "NA"

```
dsHoteles$Hab[which(is.na(dsHoteles$Hab))]

## [1] NA NA NA NA NA NA
```

Volvemos a comprobar tipos

Vemos que tenemos ahora las clases de las columnas 'Cat' y 'Hab' se corresponden con tipos numéricos

```
sapply(dsHoteles, class)

##      Nombre         Cat         Hab   Población   Dirección
## "character" "character"   "numeric" "character" "character"
```

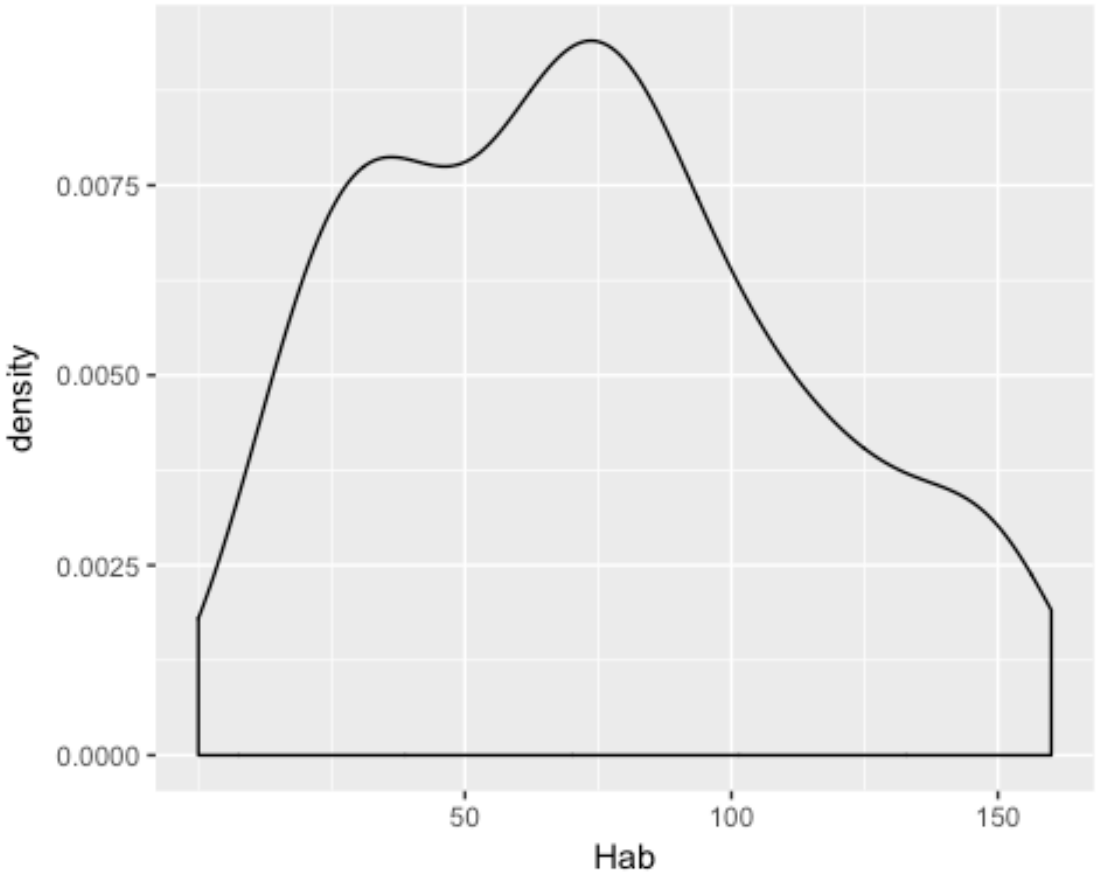
3. Obtener Función densidad

Número habitaciones

No necesitamos "na.omit" porque ggplot lo hace automaticamente

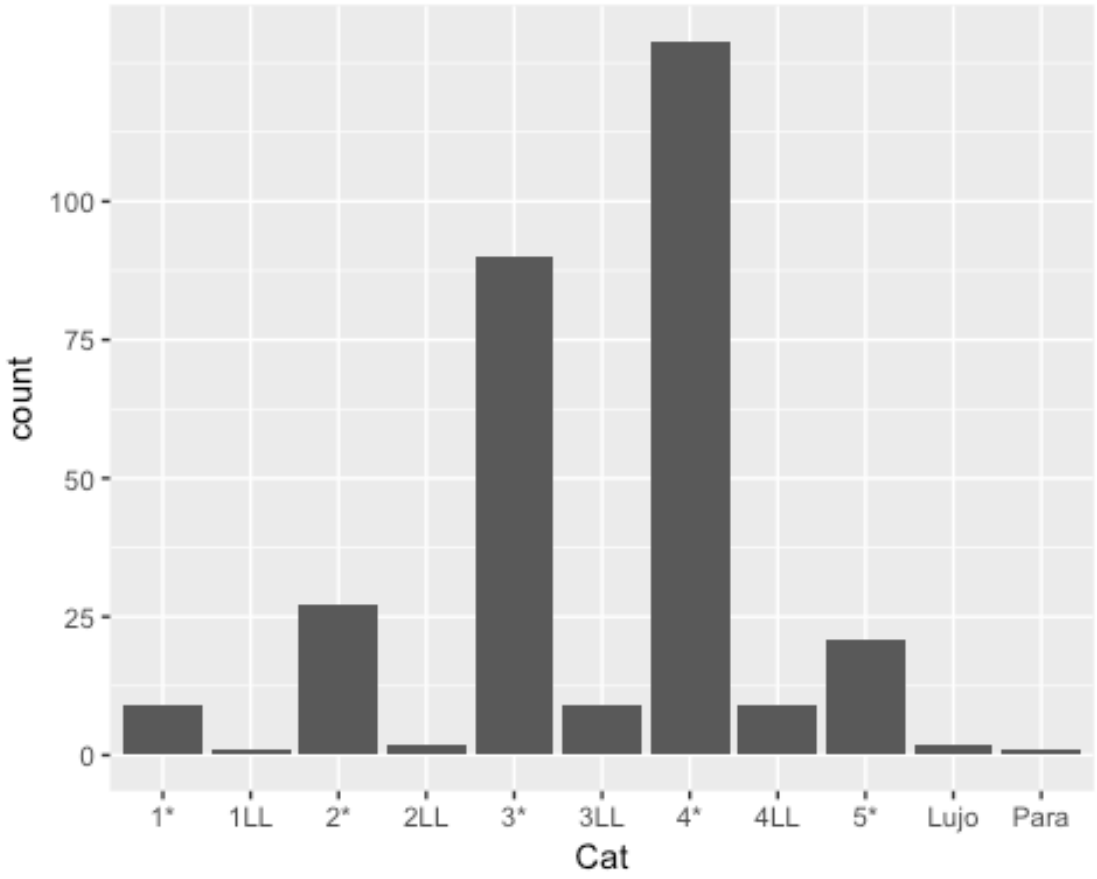
```
ggplot(na.omit(dsHoteles),
  aes(x=Hab)) +
  geom_density() +
  scale_x_continuous(breaks=c(50, 100, 150), limits = c(0,160))

## Warning: Removed 57 rows containing non-finite values (stat_density).
```



Número hoteles por categoría

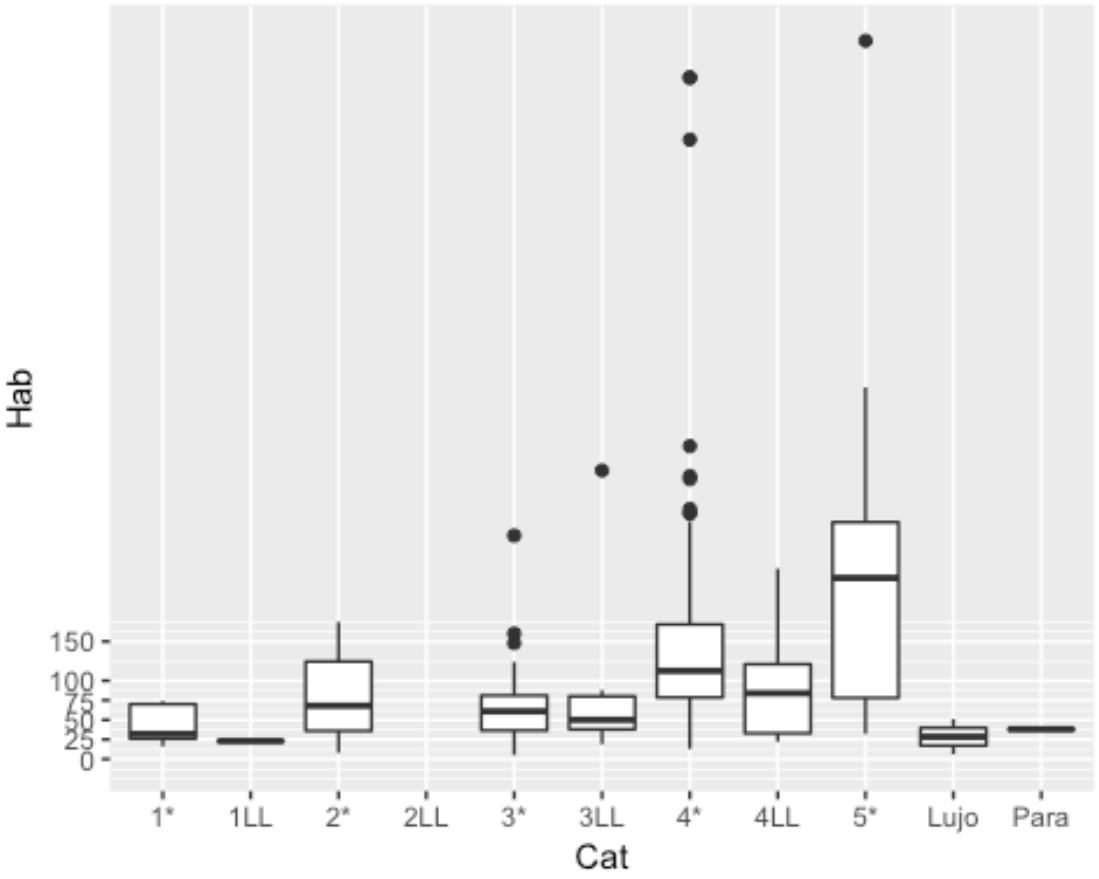
```
ggplot(data=dsHoteles, aes(Cat)) + geom_bar() +
  scale_y_continuous(breaks=c(0, 25, 50, 75, 100, 150))
```



Tablas contingencia Habitaciones por Categoría

```
ggplot(dsHoteles, aes(Cat,Hab)) + geom_boxplot() +
  scale_y_continuous(breaks=c(0, 25, 50, 75, 100, 150))

## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```


Hab
150
100
75
50
25
0
1*
1LL
2*
2LL
3*
3LL
4*
4LL
5*
Lujo
Para
Cat