Visualización de datos

Patricia Luengo Carretero

14/11/2023

Marimekko Diagram (Mosaic Plot)

Los gráficos de mosaico o diagramas de Marimekko son usados para mostrar la relación entre dos variables discretas, ya sean factores o cadenas de texto.

Este tipo de grafico recibe su nombre porque consiste en una cuadricula, en la que cada rectángulo representa el numero de casos que corresponden a un cruce específico de variables. Entre más casos se encuentren en ese cruce, más grande será el rectángulo.

Conjunto de datos

El dataset se ha obtenido de kaggle en el siguiente enlace y contiene datos del desempeño de los estudiantes al finalizar el año académico. Contiene 145 muestras y 33 atributos.

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
StudentsPerformance <- read_csv("C:/Users/patri/OneDrive/Documentos/MEGAsync/Ciencia de Datos Master/3 V
```

```
## New names:
## Rows: 145 Columns: 33
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): STUDENT ID dbl (32): Student Age, Sex, Graduated high-school type,
## Scholarship type, Ad...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `Reading frequency` -> `Reading frequency...19`
## * `Reading frequency` -> `Reading frequency...20`
```

```
head(StudentsPerformance)
```

```
## # A tibble: 6 x 33
##   `STUDENT ID` `Student Age`   Sex `Graduated high-school ty~1 `Scholarship type`
##   <chr>               <dbl> <dbl>                       <dbl>              <dbl>
## 1 STUDENT1                2     2                           3                  3
## 2 STUDENT2                2     2                           3                  3
## 3 STUDENT3                2     2                           2                  3
## 4 STUDENT4                1     1                           1                  3
## 5 STUDENT5                2     2                           1                  3
## 6 STUDENT6                2     2                           2                  3
## # i abbreviated name: 1: `Graduated high-school type`
## # i 28 more variables: `Additional work` <dbl>,
## #   `Regular artistic or sports activity` <dbl>, `Do you have a partner` <dbl>,
## #   `Total salary if available` <dbl>,
```

```
## #   `Transportation to the university` <dbl>,
## #   `Accommodation type in Cyprus` <dbl>, `Mother's education` <dbl>,
## #   `Father's education` <dbl>, `Number of sisters/brothers` <dbl>, ...
```

Vamos a fijarnos en los siguientes atributos: -**Sex:** (1: femenino, 2: masculino)
-**Graduated high-school type:** (1: privada, 2: estatal, 3: otra) -**Additional Work**: (1: Si, 2: No)

```
students <- select(StudentsPerformance, Sex, 'Graduated high-school type', 'Additional work')
colnames(students)[1] <-"sex"
colnames(students)[2] <-"graduated"
colnames(students)[3] <-"work"
head(students)
```

```
## # A tibble: 6 x 3
##     sex graduated  work
##   <dbl>     <dbl> <dbl>
## 1     2         3     1
## 2     2         3     1
## 3     2         2     2
## 4     1         1     1
## 5     2         1     2
## 6     2         2     2
```

Modificamos los datos categóricos de rango numérico a etiquetas más legibles:

sex

```
students <- students %>%
  mutate(sex = case_when(
    (sex == 1) ~ 'female',
    (sex == 2) ~ 'male'))
head(students$sex)
```

```
## [1] "male"   "male"   "male"   "female" "male"   "male"
```

graduated

```
students <- students %>%
  mutate(graduated = case_when(
    (graduated == 1) ~ 'private',
    (graduated == 2) ~ 'state',
    (graduated == 3) ~ 'other'))
head(students$graduated)
```

```
## [1] "other"   "other"   "state"   "private" "private" "state"
```

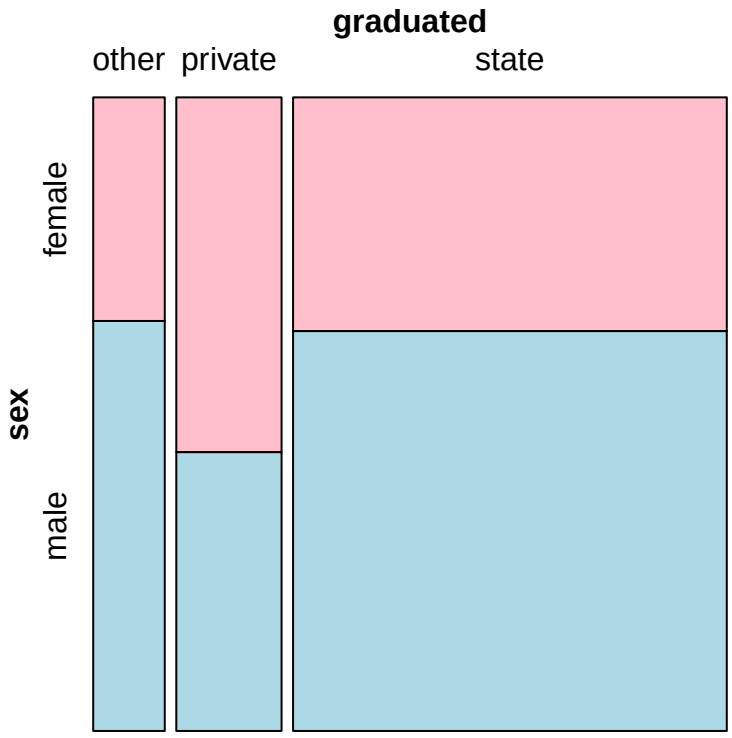
work

```
students <- students %>%
  mutate(work = case_when(
    (work == 1) ~ 'Yes',
    (work == 2) ~ 'No'))
head(students$work)
```

```
## [1] "Yes" "Yes" "No"  "Yes" "No"  "No"
```

Dos variables

```
mosaic( ~ sex + graduated, data = students,
       highlighting = "sex", highlighting_fill = c("pink", "lightblue"),
       direction = c("h", "v"))
```

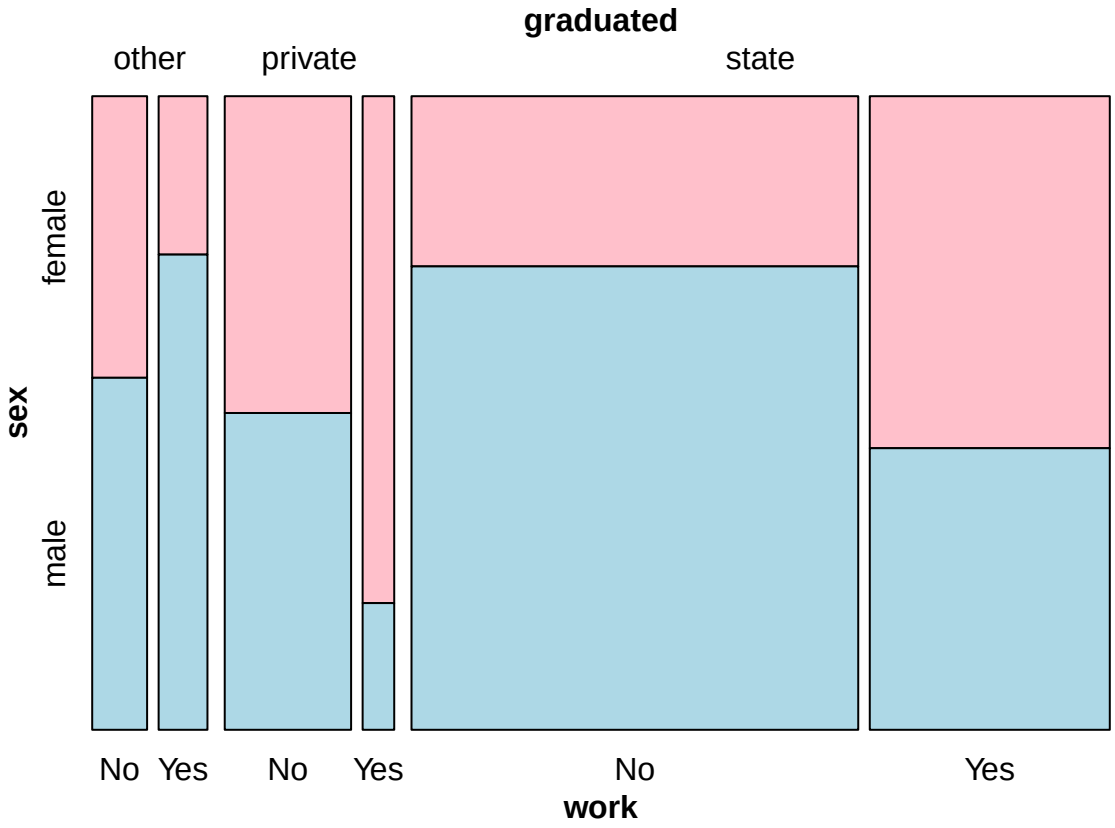


Conclusion

Observamos que los hombres en su mayoria se han graduado un universidades estatales.

Tres variables

```
mosaic( ~ sex + graduated + work, data = students,
       highlighting = "sex", highlighting_fill = c("pink", "lightblue"),
       direction = c("h", "v", "v"))
```



Conclusion

La mayor parte de los hombres que se graduaron en universidades estatales no trabajaban.

Choropleth Map

Choropleth Map proporcionan una manera fácil de visualizar cómo varía una variable en un área geográfica o muestran el nivel de variabilidad dentro de una región. La diferencia con un mapa de calor es que en un Choropleth Map utiliza las regiones dibujadas según el patrón de la variable.

Conjunto de datos

El dataset se ha obtenido de kaggle en el siguiente enlace y contiene datos de las precipitaciones por provincia en España en los útimos años, en concreto, nosotros hemos seleccionado los del año 2021.

```
library(readr)
precipitaciones <- read_delim("C:/Users/patri/OneDrive/Documentos/MEGAsync/Ciencia de Datos Master/3 Vis
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 50 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr  (4): Parametro, region, cpro, codauto
## dbl (13): enero, febrero, marzo, abril, mayo, junio, julio, agosto, septiemb...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(precipitaciones)
```

```
## # A tibble: 6 x 17
##   Parametro     region cpro  codauto enero febrero marzo abril  mayo junio julio
##   <chr>         <chr>  <chr> <chr>   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Precipitacion ALMER~ 04    01       56.3     5.2  35.8  43.1  38.8  16.9  10.2
## 2 Precipitacion CADIZ  11    01      148.     75.4  65    35.5   7.5  12.3   0
## 3 Precipitacion CORDO~ 14    01       88.6    57.5   9.6  64.5  11.6  20.7   0.3
## 4 Precipitacion GRANA~ 18    01       72.2    32.9  17.2  51    38.1  15.5   1.3
## 5 Precipitacion HUELVA 21    01       50.7   102.   35.4  51.5  15.1  14.5   0
## 6 Precipitacion JAEN   23    01      101.     66.8  10    53.2  18.5  31.2   0.1
## # i 6 more variables: agosto <dbl>, septiembre <dbl>, octubre <dbl>,
## #   noviembre <dbl>, diciembre <dbl>, anual <dbl>
```

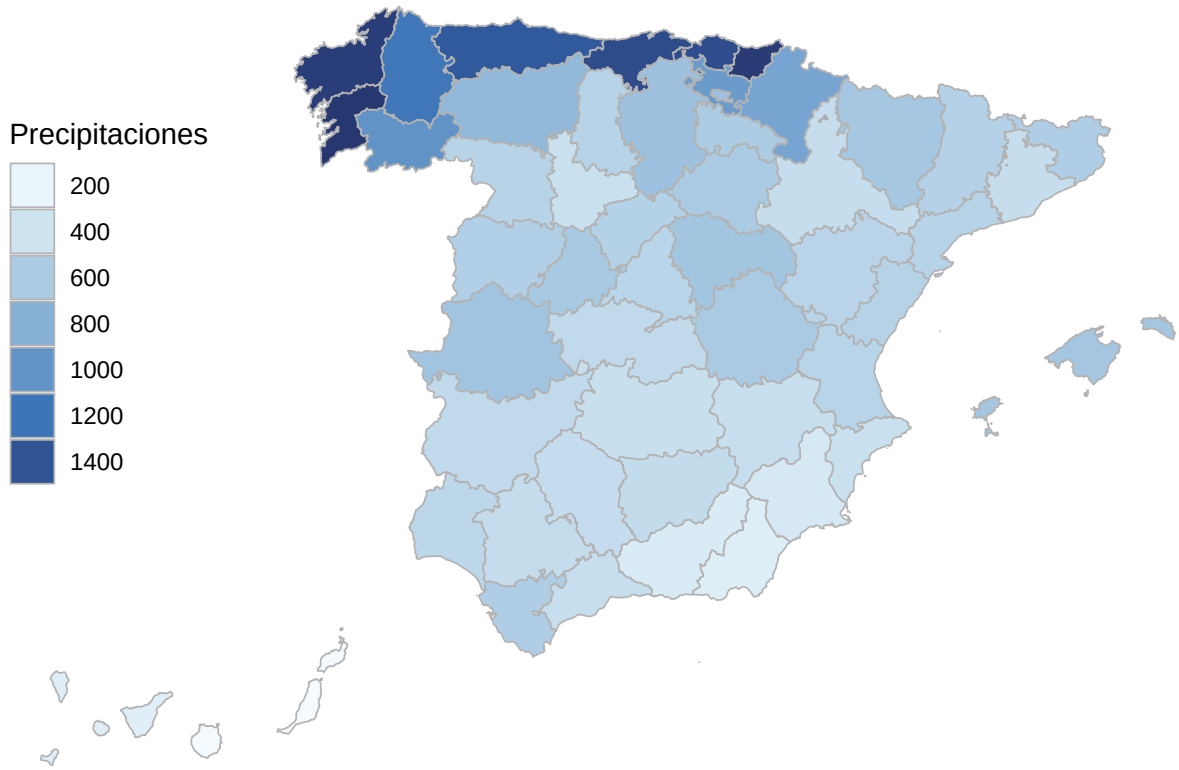
```r
library(mapSpain)
library(sf)

codelist <- mapSpain::esp_codelist

prep <- unique(merge(precipitaciones, codelist[, c("cpro", "codauto")], all.x = TRUE))

prov <- esp_get_prov()
prov_sf <- merge(prov, prep)

ggplot(prov_sf) +
  geom_sf(aes(fill = anual),
    color = "grey70",
    linewidth = .3
  ) +
  scale_fill_gradientn(
    colors = hcl.colors(10, "Blues", rev = TRUE),
    n.breaks = 10,
    guide = guide_legend(title = "Precipitaciones")
  ) +
  theme_void() +
  theme(legend.position = c(0.1, 0.6))
```



Conclusion

Como era de esperar la mayor cantidad de precipitaciones en el acumulado anual se producen en el Norte de España.

Histograms

Su origen se debió al matemático Karl Pearson en 1895. Se utiliza para representar variables cuantitativas continuas y lo que suele hacerse es realizar una agrupación por franjas.

Conjunto de datos

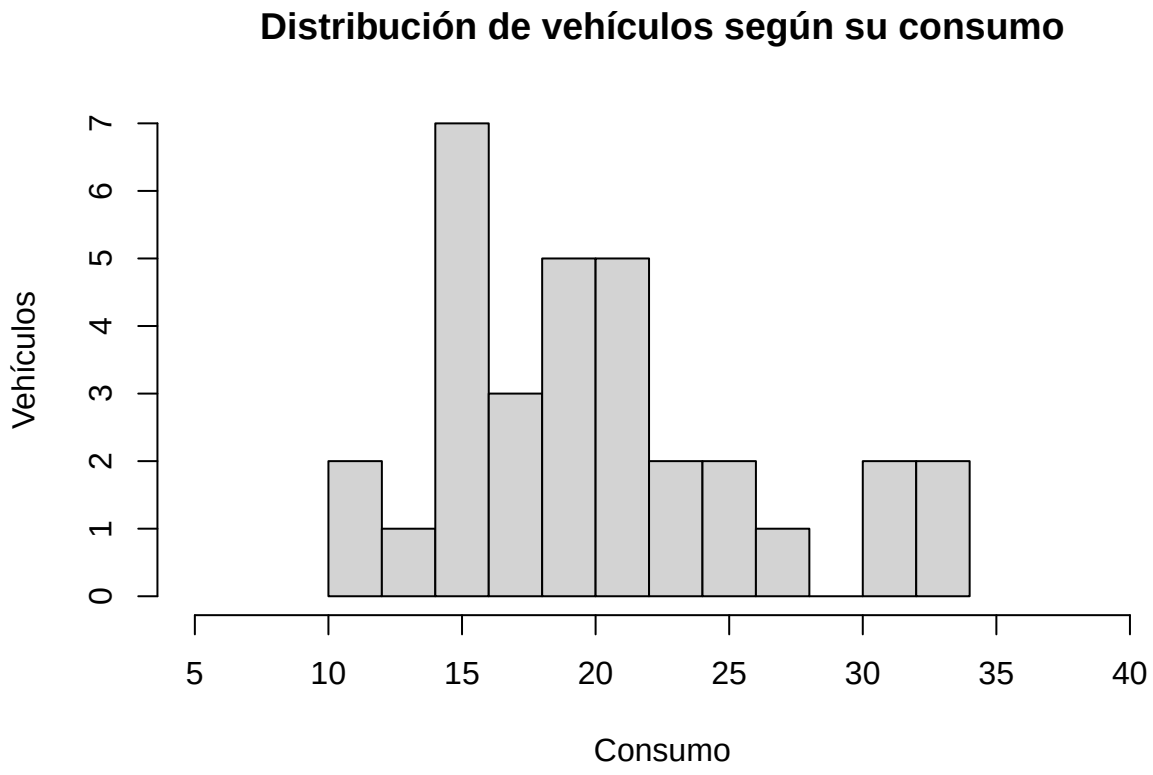
Utilizamos un dataset que contiene datos relacionados con vehículos, por ejemplo cilindrada, consumo, etc.

```
mtcars
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
```

```
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

```
hist(mtcars$mpg,main="Distribución de vehículos según su consumo",breaks=15,xlab="Consumo",ylab="Vehícul
```



Conlusión

Vemos que hay una gran cantidad de vehículos cuyo consumo se encuentra entre 14 y 22 galones/milla.