Sesión 1: Procesamiento y Visualización de Datos con R y Python

Mauricio Alejandro Mazo Lopera
Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín

Contenido del curso:

Inicialmente se presenta un comparativo general entre ambos programas y se tratarán los siguientes puntos tanto en R como en Python:

- Importación y exportación de datos.
- Análisis descriptivos básicos.
- Descriptivo por grupos.
- Unión de bases de datos.
- Gráficos descriptivos.

R versus Python:

Ross Ihaka

Estadístico

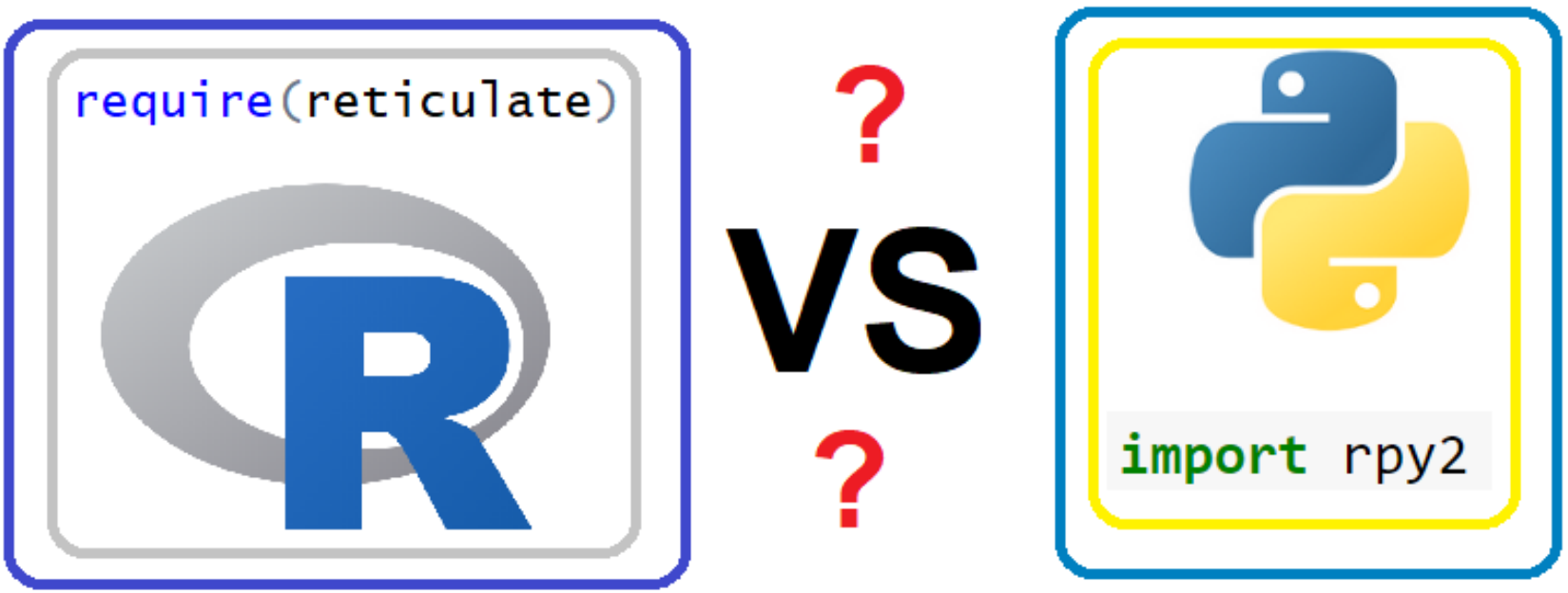
Robert Gentleman

Estadístico

Guido van Rossum

Científico de la computación

R versus Python: Enlaces



R versus Python: Integrated Development Environment (IDE)

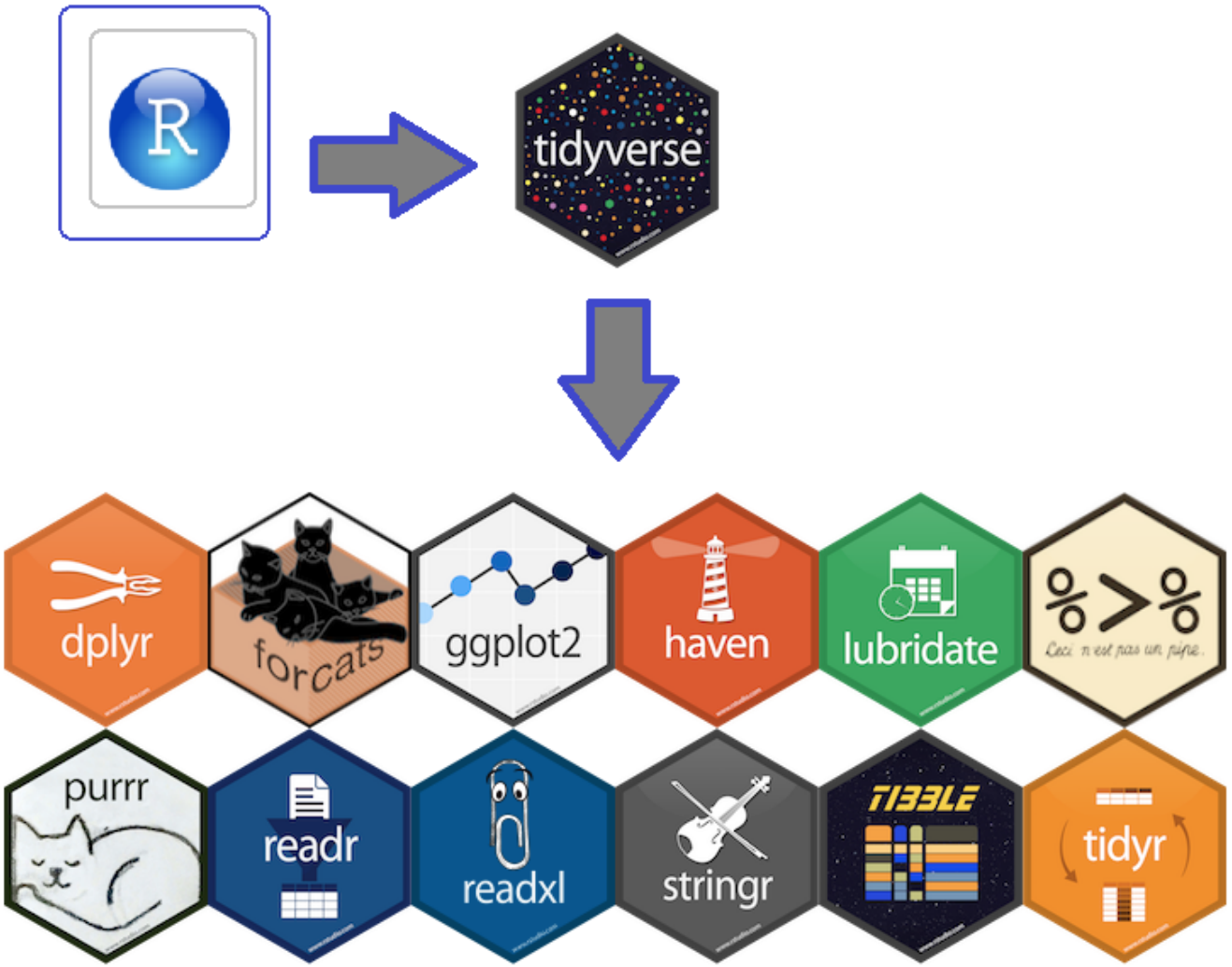
R versus Python: Integrated Development Environment (IDE)

R versus Python

- Gratuito
- Académicos
- Estadístico
- Ajuste de modelos
- Visualización gráfica
- Interacción con la nube
- Buena documentación

- Gratuito
- Desarrolladores
- Multipropósito
- En desarrollo
- En desarrollo
- Interacción con la nube
- Menos documentación

“Mundo” Tidyverse en R-Studio



"Mundo" Tidyverse en R-Studio

El Tidyverse es una colección de paquetes del R que permiten preparar, procesar y graficar bases de datos. Se destacan los siguientes:

- ggplot: permite crear visualizaciones elegantes de los datos de una manera relativamente sencilla.
- stringr: permite manipular cadenas de caracteres con el fin de realizar substitucioes, detectar duplicados, analizar patrones, etc.

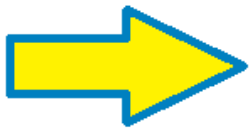
"Mundo" Tidyverse en R-Studio

- tidyr: tiene como objetivo obtener datos ordenador. Destacan funciones como **gather** para crear factores con base en nombres de columnas y **separate** para crear factores separando los caracteres de una columna.
- readr: permite importar y exportar bases de datos en diferentes formatos y tiene implementada la función **problems** que detecta problemas en nuestras bases.

Para más información visitar la página web:

https://www.tidyverse.org/packages/

Librerías en Python

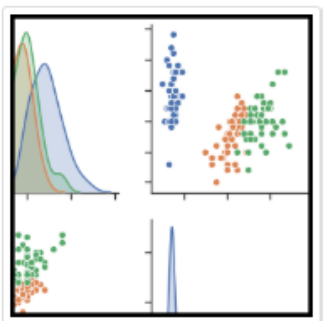


jupyter

matplotlib

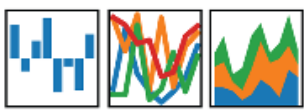
SciPy

seaborn



pandas

$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$



NumPy

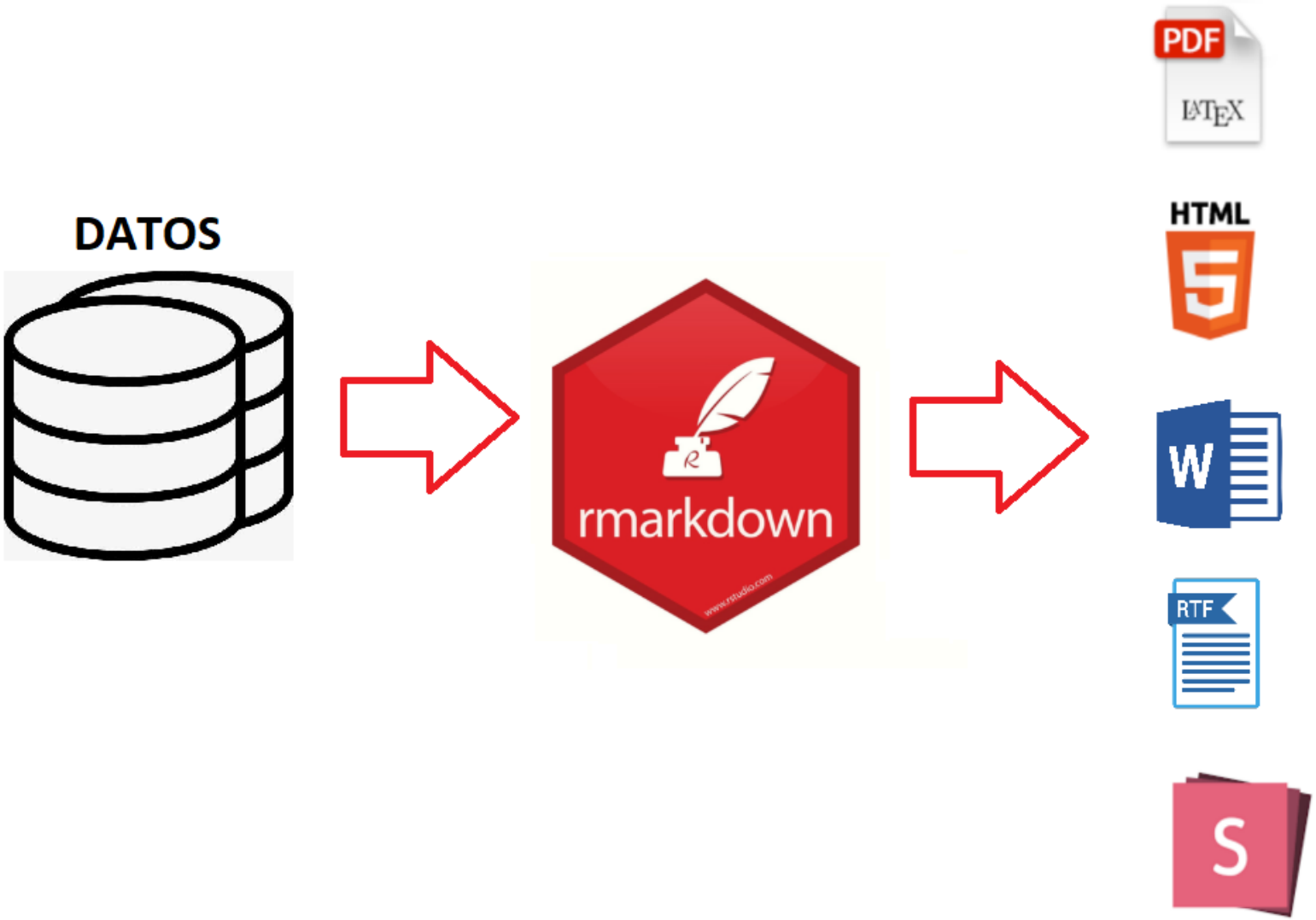
statsmodels

Librerías en Python

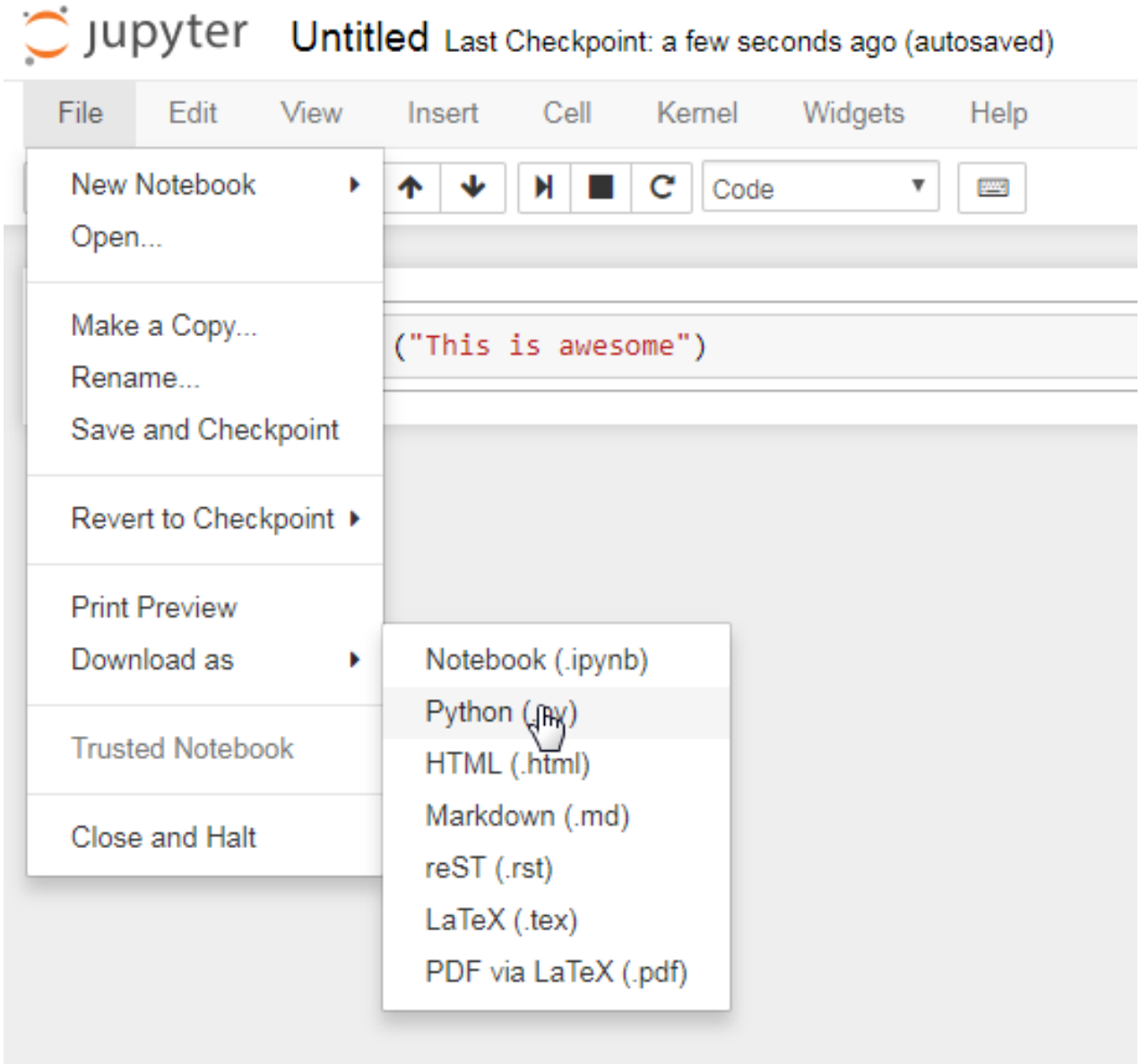
Algunas de las librerías que más se utilizan para procesamiento, visualización y análisis de bases de datos:

- pandas: importación y exportación de bases de datos, además de manipulación de las mismas.
- matplotlib: realización de gráficos.
- seaborn: realización de gráficos considerando subgrupos o categorías.
- statmodels: ajuste de modelos estadísticos.
- numpy: manejo de matrices.
- SciPy: análisis numérico.

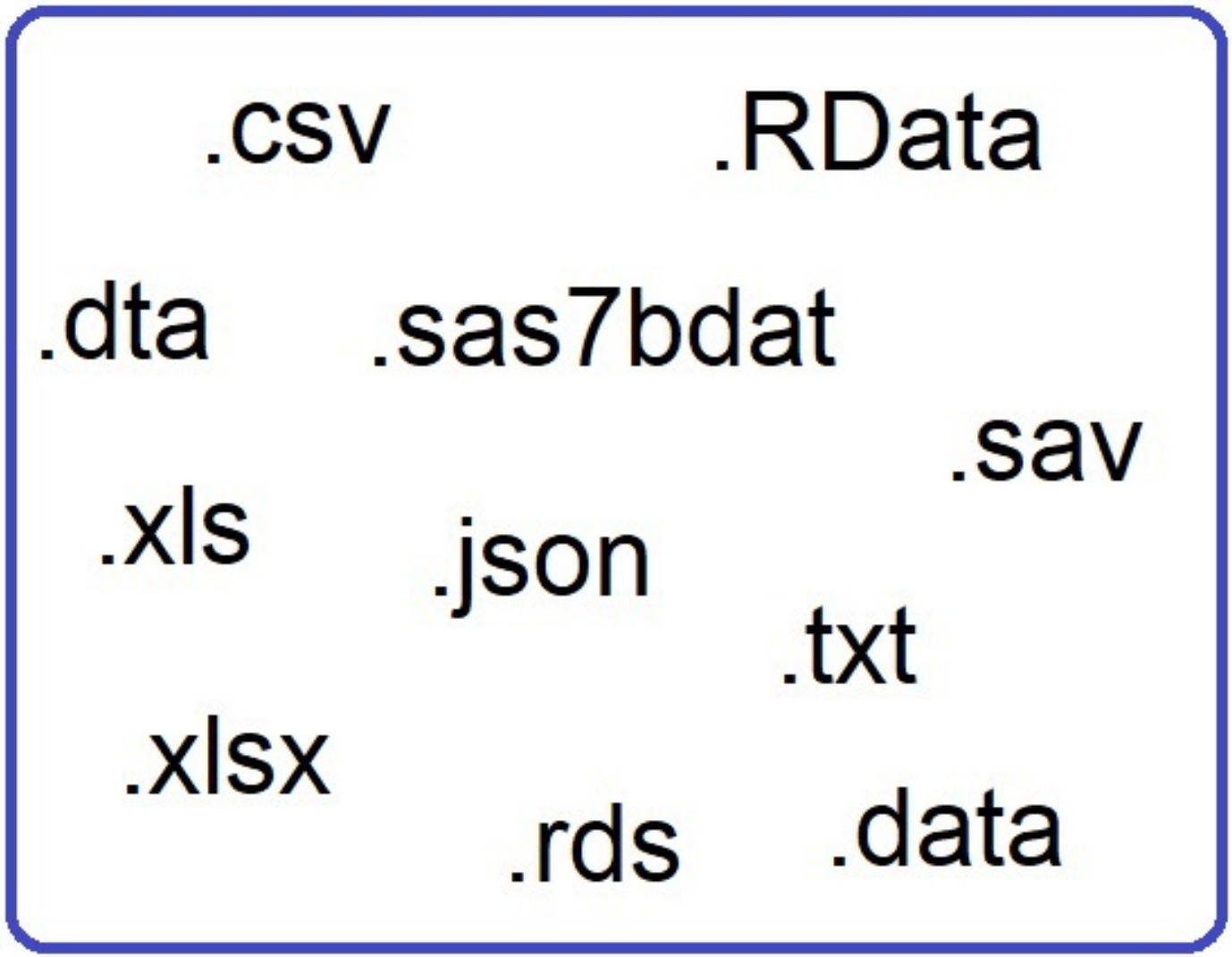
Creación de reportes en R con R Markdown:



Creación de reportes en python con Jupyter Notebook:



Extensiones - Bases de datos



Importando y exportando bases de datos en R:

Algunos paquetes utilizados para importar y exportar bases de datos en R son:

- readr: permite importar y exportar datos en varios formatos.
- readxl: Importar datos en formato "Excel".
- haven: Importar y exportar datos en "SPSS", "Stata" y "SAS".
- httr: Herramientas para trabajar con URLs y archivos HTTP.
- rvest: útil para extraer información de las páginas web
- xml2: Leer archivos HTML o XML.

Importando y exportando bases de datos en Python:

Algunas librerías utilizados para importar y exportar bases de datos en python son:

- csv: Importar y exportar datos .csv.
- pandas: Importar y exportar datos .csv, “Excel”, “SAS”, “SPSS”, “html”, “stata”, “sql”, etc.

Ejemplo 1: Desempleo en Antioquia

FUERTE: Antioquia. Departamento Administrativo de Planeación. Encuestas de Calidad de Vida 2017, Tasa de desempleo.

Ejemplo 1: Desempleo en Antioquia

```
Datos_1<-read.table("DATOS/DESEMPLEO_ECV_2017_MUNICIPIOS.txt",
                    sep=";",header=TRUE)
names(Datos_1)
head(Datos_1)
```

```
## [1] "Municipio" "Total"     "Abierto"   "Oculto"    "Urbano"    "Rural"
## [7] "Hombre"    "Mujer"     "Subregion"

##    Municipio Total Abierto Oculto Urbano Rural Hombre Mujer
## 1   Medellín  8.34    6.98   1.37   8.34  8.40   8.19  8.54
## 2    Barbosa 12.99    8.73   4.25  14.41 11.69   8.64 20.82
## 3      Bello  8.82    7.17   1.64   8.79 10.42   8.29  9.56
## 4     Caldas  9.15    6.46   2.69   8.77 10.53   8.05 10.70
## 5 Copacabana  9.42    8.14   1.28   9.58  8.26   8.66 10.56
## 6   Envigado  5.27    4.30   0.97   5.23  6.26   4.98  5.62
##         Subregion
## 1 Vallé de Aburrá
## 2 Vallé de Aburrá
## 3 Vallé de Aburrá
## 4 Vallé de Aburrá
## 5 Vallé de Aburrá
## 6 Vallé de Aburrá
```

Ejemplo 1: Desempleo en Antioquia

```
dim(Datos_1)
str(Datos_1)
```

```
## [1] 125   9
```

```
## 'data.frame':    125 obs. of  9 variables:
##  $ Municipio: Factor w/ 125 levels "Abejorral","Abriaquí",..: 71 15
##  $ Total    : num  8.34 12.99 8.82 9.15 9.42 ...
##  $ Abierto  : num  6.98 8.73 7.17 6.46 8.14 4.3 9.21 5.15 4.46 4.02
##  $ Oculto   : num  1.37 4.25 1.64 2.69 1.28 0.97 1.66 1.25 1.5 1.68
##  $ Urbano   : num  8.34 14.41 8.79 8.77 9.58 ...
##  $ Rural    : num  8.4 11.69 10.42 10.53 8.26 ...
##  $ Hombre   : num  8.19 8.64 8.29 8.05 8.66 ...
##  $ Mujer    : num  8.54 20.82 9.56 10.7 10.56 ...
##  $ Subregion: Factor w/ 9 levels "Bajo Cauca","Magdalena Medio",..:
```

Ejemplo 1: Desempleo en Antioquia

```
summary(Datos_1)
```

```
##       Municipio       Total             Abierto           Oculto
##  Abejorral :  1   Min.   : 0.420   Min.   : 0.420   Min.   :0.000
##  Abriaquí  :  1   1st Qu.: 4.575   1st Qu.: 3.297   1st Qu.:0.465
##  Alejandría:  1   Median : 6.345   Median : 4.925   Median :1.240
##  Amagá     :  1   Mean   : 7.160   Mean   : 5.545   Mean   :1.615
##  Amalfi    :  1   3rd Qu.: 8.967   3rd Qu.: 7.025   3rd Qu.:2.090
##  Andes     :  1   Max.   :22.770   Max.   :21.920   Max.   :9.240
##  (Other)   :119   NA's   :3        NA's   :3        NA's   :3
##      Urbano            Rural            Hombre           Mujer
##  Min.   : 0.420   Min.   : 0.000   Min.   : 0.180   Min.   : 0.75
##  1st Qu.: 6.035   1st Qu.: 2.272   1st Qu.: 2.947   1st Qu.: 6.57
##  Median : 8.315   Median : 4.315   Median : 4.895   Median :10.76
##  Mean   : 9.201   Mean   : 5.294   Mean   : 5.564   Mean   :11.40
##  3rd Qu.:10.940   3rd Qu.: 7.520   3rd Qu.: 7.322   3rd Qu.:13.26
##  Max.   :35.010   Max.   :28.170   Max.   :22.650   Max.   :46.22
##  NA's   :3        NA's   :3        NA's   :3        NA's   :3
##      Subregion
##  Oriente  :23
##  Suroeste :23
##  Occidente:19
##  Norte    :17
##  Urabá    :11
##  Nordeste :10
##  (Other)  :22
```

Ejemplo 1: Desempleo en Antioquia

```
with(Datos_1, summary(Urbano))
with(Datos_1, summary(Rural))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.420   6.035   8.315   9.201  10.940  35.010       3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.272   4.315   5.294   7.520  28.170       3
```

Ejemplo 1: Desempleo en Antioquia

```
with(Datos_1, summary(Urbano))
with(Datos_1, summary(Rural))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.420   6.035   8.315   9.201  10.940  35.010       3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.272   4.315   5.294   7.520  28.170       3
```

```
with(Datos_1, summary(Mujer))
with(Datos_1, summary(Hombre))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.75    6.57   10.76   11.40   13.26   46.22       3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.180   2.947   4.895   5.564   7.322  22.650       3
```

Ejemplo 1: Desempleo en Antioquia

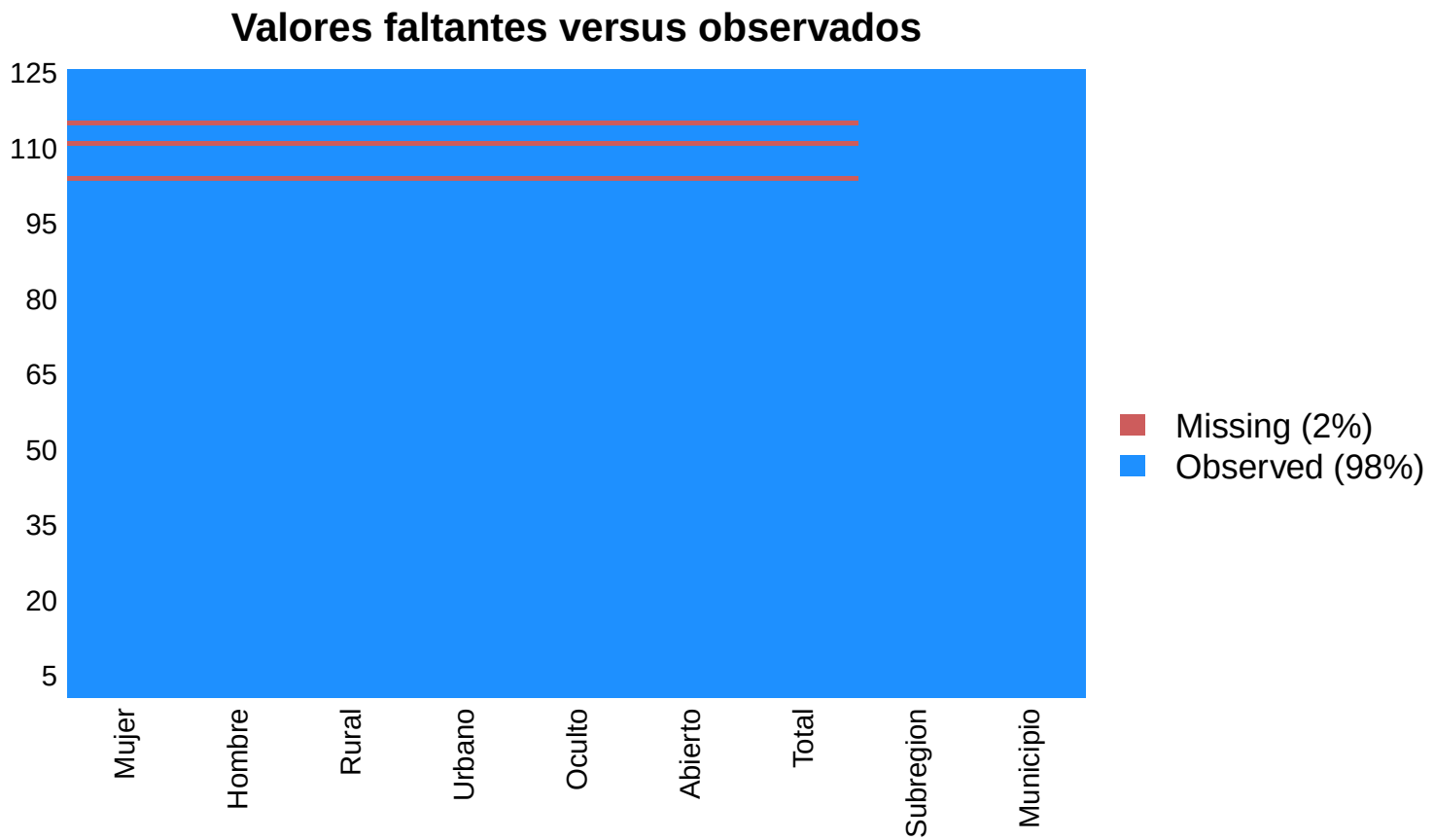
```
with(Datos_1, summary(Abierto))
with(Datos_1, summary(Oculto))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.420   3.297   4.925   5.545   7.025  21.920       3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.465   1.240   1.615   2.090   9.240       3
```

Ejemplo 1: Desempleo en Antioquia

```
require(Amelia)
missmap(Datos_1,
        main = "Valores faltantes versus observados")
```



Ejemplo 1: Desempleo en Antioquia

```
tapply(Datos_1$Total, Datos_1$Subregion, mean)
```

```
##      Bajo Cauca Magdalena Medio        Nordeste          Norte
##              NA              NA        7.832000        6.390588
##       Occidente         Oriente       Suroeste           Urabá
##        6.671579        6.386087        5.876522        8.150909
## Vallé de Aburrá
##        8.293000
```

Ejemplo 1: Desempleo en Antioquia

```
tapply(Datos_1$Total, Datos_1$Subregion, mean)
```

```
##      Bajo Cauca Magdalena Medio         Nordeste           Norte
##              NA                NA         7.832000        6.390588
##       Occidente         Oriente         Suroeste           Urabá
##        6.671579        6.386087         5.876522        8.150909
## Vallé de Aburrá
##        8.293000
```

```
tapply(Datos_1$Total, Datos_1$Subregion, length)
```

```
##      Bajo Cauca Magdalena Medio         Nordeste           Norte
##               6                 6               10              17
##       Occidente         Oriente         Suroeste           Urabá
##              19                23               23              11
## Vallé de Aburrá
##              10
```

Ejemplo 1: Desempleo en Antioquia

```
Datos_12<-na.omit(Datos_1)
tapply(Datos_12$Total, Datos_12$Subregion, mean)
```

```
##      Bajo Cauca Magdalena Medio        Nordeste           Norte
##        8.070000       14.578000        7.832000        6.390588
##       Occidente         Oriente        Suroeste           Urabá
##        6.671579        6.386087        5.876522        8.150909
## Vallé de Aburrá
##        8.293000
```

Ejemplo 1: Desempleo en Antioquia

```
Datos_12<-na.omit(Datos_1)
tapply(Datos_12$Total, Datos_12$Subregion, mean)
```

```
##      Bajo Cauca Magdalena Medio        Nordeste          Norte
##        8.070000       14.578000        7.832000       6.390588
##       Occidente         Oriente        Suroeste          Urabá
##        6.671579        6.386087        5.876522       8.150909
## Vallé de Aburrá
##        8.293000
```

```
tapply(Datos_12$Total, Datos_12$Subregion, length)
```

```
##      Bajo Cauca Magdalena Medio        Nordeste          Norte
##               4               5              10             17
##       Occidente         Oriente        Suroeste          Urabá
##              19              23              23             11
## Vallé de Aburrá
##              10
```

Ejemplo 1: Desempleo en Antioquia

```
na_1<-which(is.na(Datos_1$Total))
na_1
```

```
## [1] 11 15 22
```

Ejemplo 1: Desempleo en Antioquia

```
na_1<-which(is.na(Datos_1$Total))
na_1
```

```
## [1] 11 15 22
```

```
Datos_1$Municipio[na_1]
```

```
## [1] Cáceres Tarazá  Yondó
## 125 Levels: Abejorral Abriaquí Alejandría Amagá Amalfi
```

```
Datos_1$Subregion[na_1]
```

```
## [1] Bajo Cauca      Bajo Cauca      Magdalena Medio
## 9 Levels: Bajo Cauca Magdalena Medio Nordeste Norte Occ
```

Ejemplo 1: Desempleo en Antioquia

```
min_1<-with(Datos_12,min(Total))
max_1<-with(Datos_12,max(Total))

with(Datos_12,Municipio[Total==min_1])
```

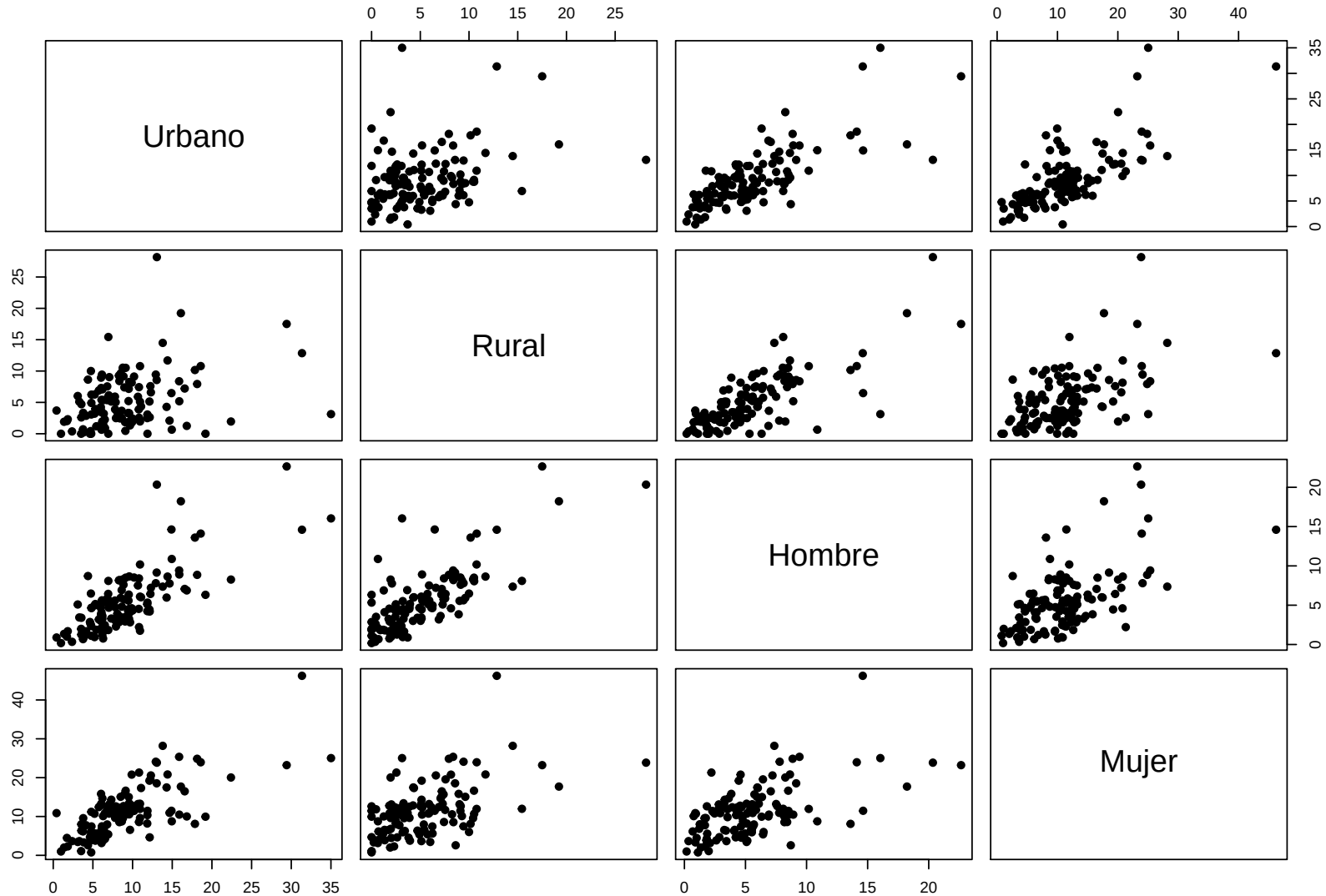
```
## [1] Concordia
## 125 Levels: Abejorral Abriaquí Alejandría Amagá Amalfi
```

```
with(Datos_12,Municipio[Total==max_1])
```

```
## [1] Puerto Nare
## 125 Levels: Abejorral Abriaquí Alejandría Amagá Amalfi
```

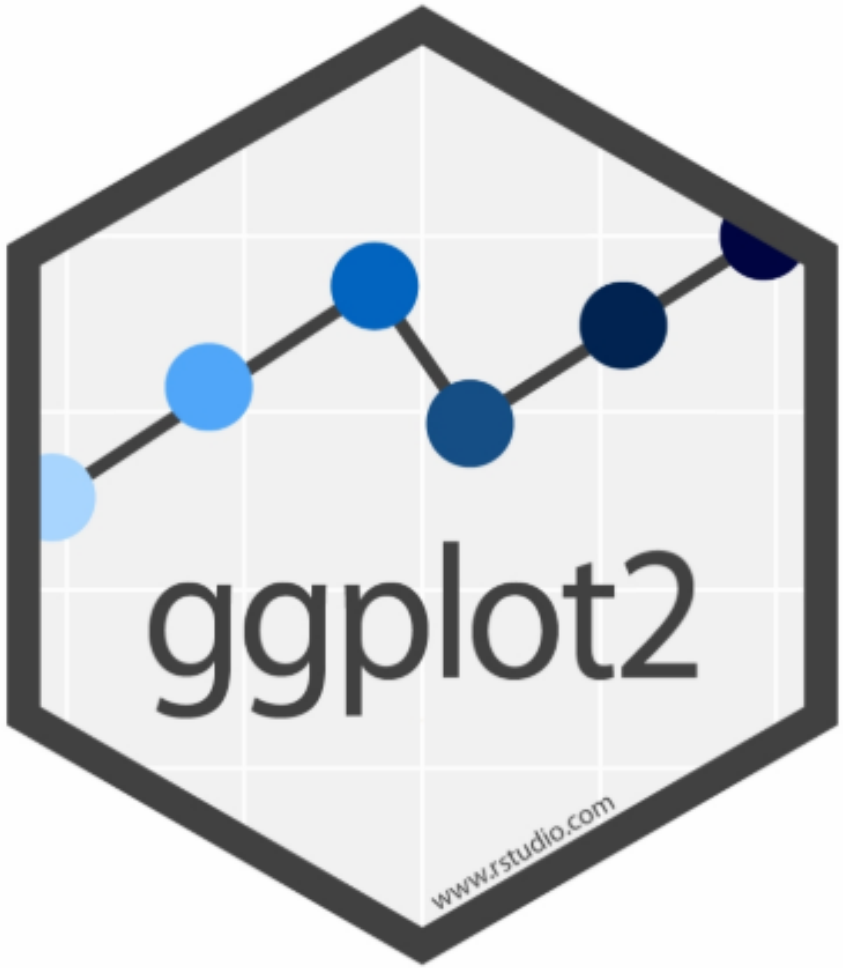
Ejemplo 1: Desempleo en Antioquia

```
plot(Datos_1[,5:8], pch=19)
```



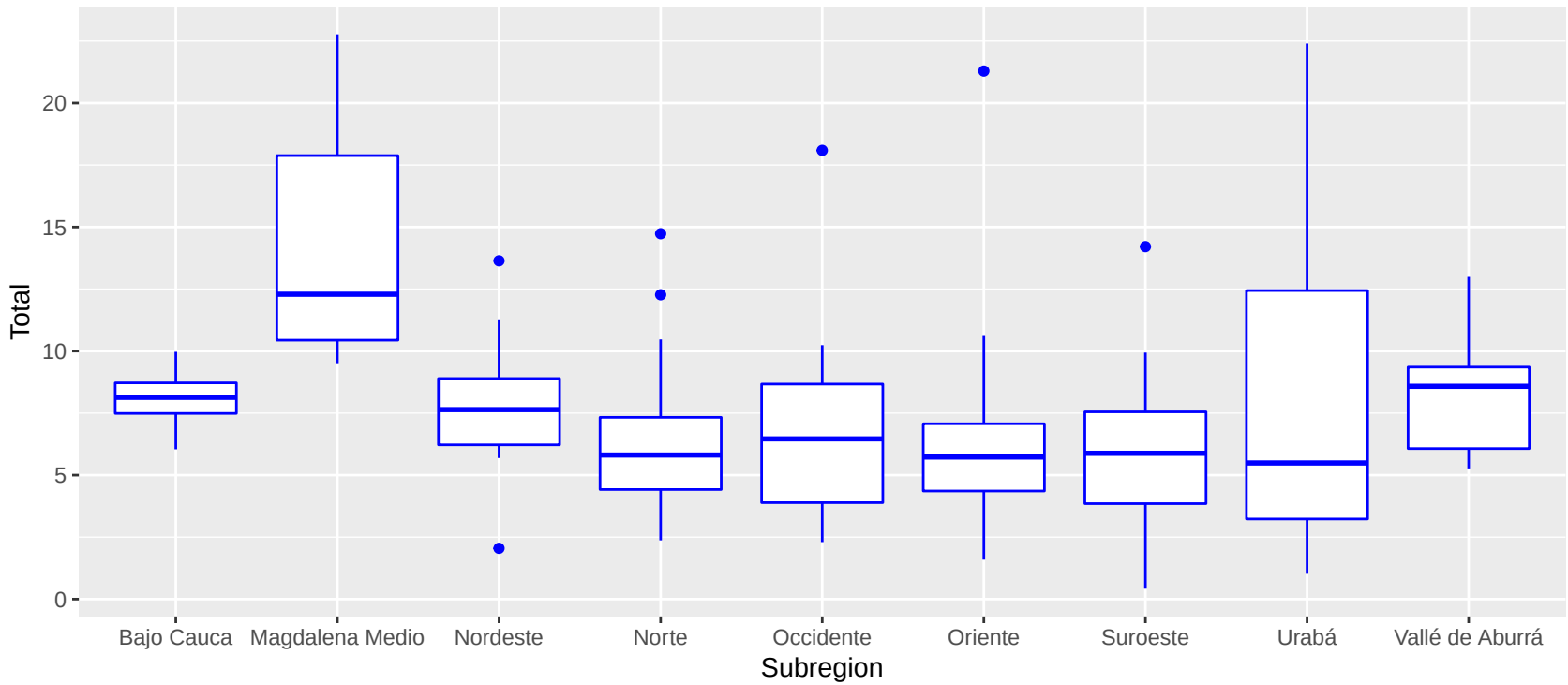
Paquete ggplot2

```
require(ggplot2)
```



Ejemplo 1: Desempleo en Antioquia

```
require(ggplot2)
ggplot(Datos_1, aes(y=Total, x=Subregion)) +
  geom_boxplot(color="blue")
```



Ejemplo 1: Desempleo en Antioquia

```
require(dplyr)

min1<-tapply(Datos_12$Total, Datos_12$Subregion, min)
BD_min1<-filter(Datos_12,Total %in% min1)

max1<-tapply(Datos_12$Total, Datos_12$Subregion, max)
BD_max1<-filter(Datos_12,Total %in% max1)

Tabla1<-data.frame(
        Subregion=BD_min1$Subregion,
        Municipio_Min=BD_min1$Municipio,
        Minimo=BD_min1$Total,
        Municipio_Max=BD_max1$Municipio,
        Maximo=BD_max1$Total,
        row.names = NULL)
```

Ejemplo 1: Desempleo en Antioquia

```
require(knitr)
kable(Tabla1)
```

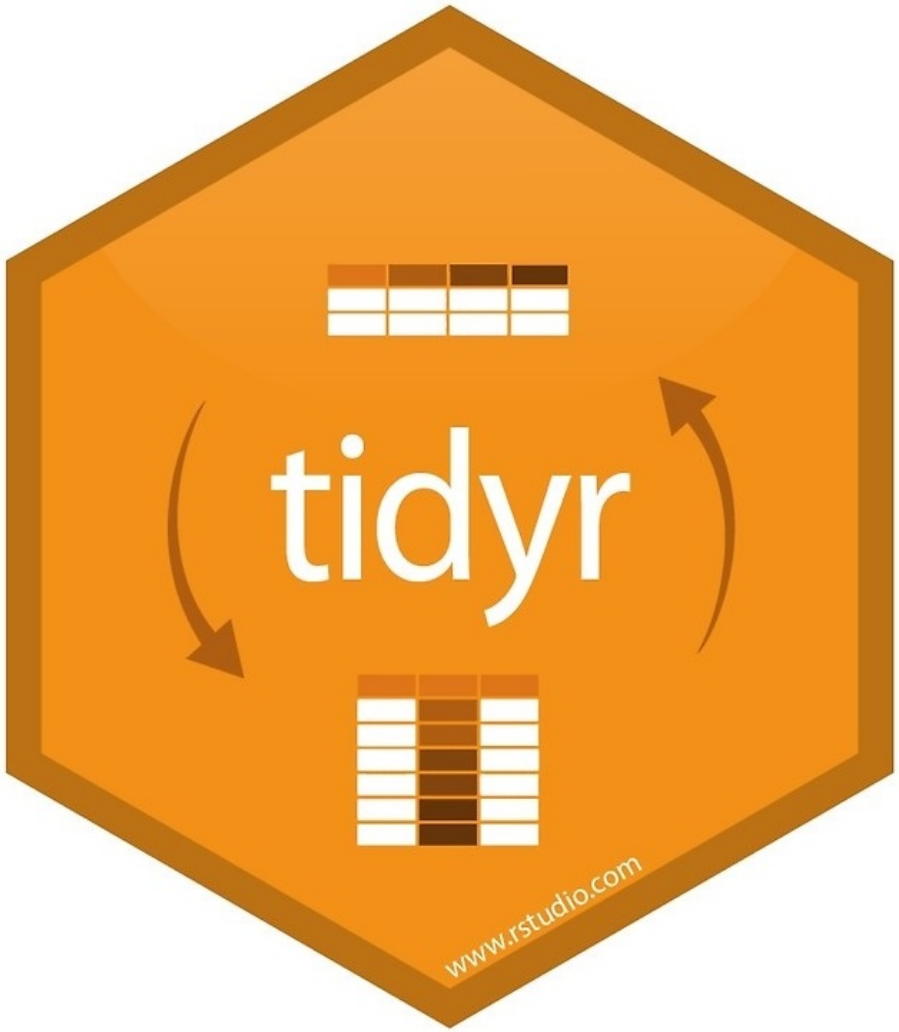
Subregion	Municipio_Min	Minimo	Municipio_Max	Maximo
Vallé de Aburrá	Envigado	5.27	Barbosa	12.99
Bajo Cauca	El Bagre	6.04	Caucasia	9.97
Magdalena Medio	Maceo	9.51	Puerto Nare	22.77
Nordeste	Anorí	2.05	Cisneros	13.64
Norte	Ituango	2.37	Yarumal	14.73
Occidente	Ebéjico	2.30	Sabanalarga	18.09
Oriente	La Unión	1.59	San Rafael	21.29
Suroeste	Concordia	0.42	Fredonia	14.21
Urabá	Murindó	1.02	Arboletes	22.40

Ejemplo 1: Desempleo en Antioquia

```
sex<-c(rep("Hombre",125),rep("Mujer",125))
psex<-c(Datos_1$Hombre,Datos_1$Mujer)

sector<-c(rep("Rural",125),rep("Urbano",125))
psector<-c(Datos_1$Rural,Datos_1$Urbano)

tipo<-c(rep("Abierto",125),rep("Oculto",125))
ptipo<-c(Datos_1$Abierto,Datos_1$Oculto)

Subregion<-rep(Datos_1$Subregion,2)

Datos_13<-data.frame(sex,psex, sector,
                     psector, tipo, ptipo,
                     Subregion)
```

Paquete `tidyr`

```
require(tidyr)
```



Ejemplo: Desempleo en Antioquia

```
require(tidyr)

bd1<-gather(Datos_1[,1:4],"Abierto","Oculto",
            key="tipo", value="ptipo")

bd2<-gather(Datos_1[,5:6],"Urbano","Rural",
            key="sector", value="psector")

bd3<-gather(Datos_1[,7:9],"Hombre","Mujer",
            key="sex", value="psex")

Datos_13<-cbind(bd1,bd2,bd3)
```

Ejemplo 1: Desempleo en Antioquia

```
head(bd1)
```

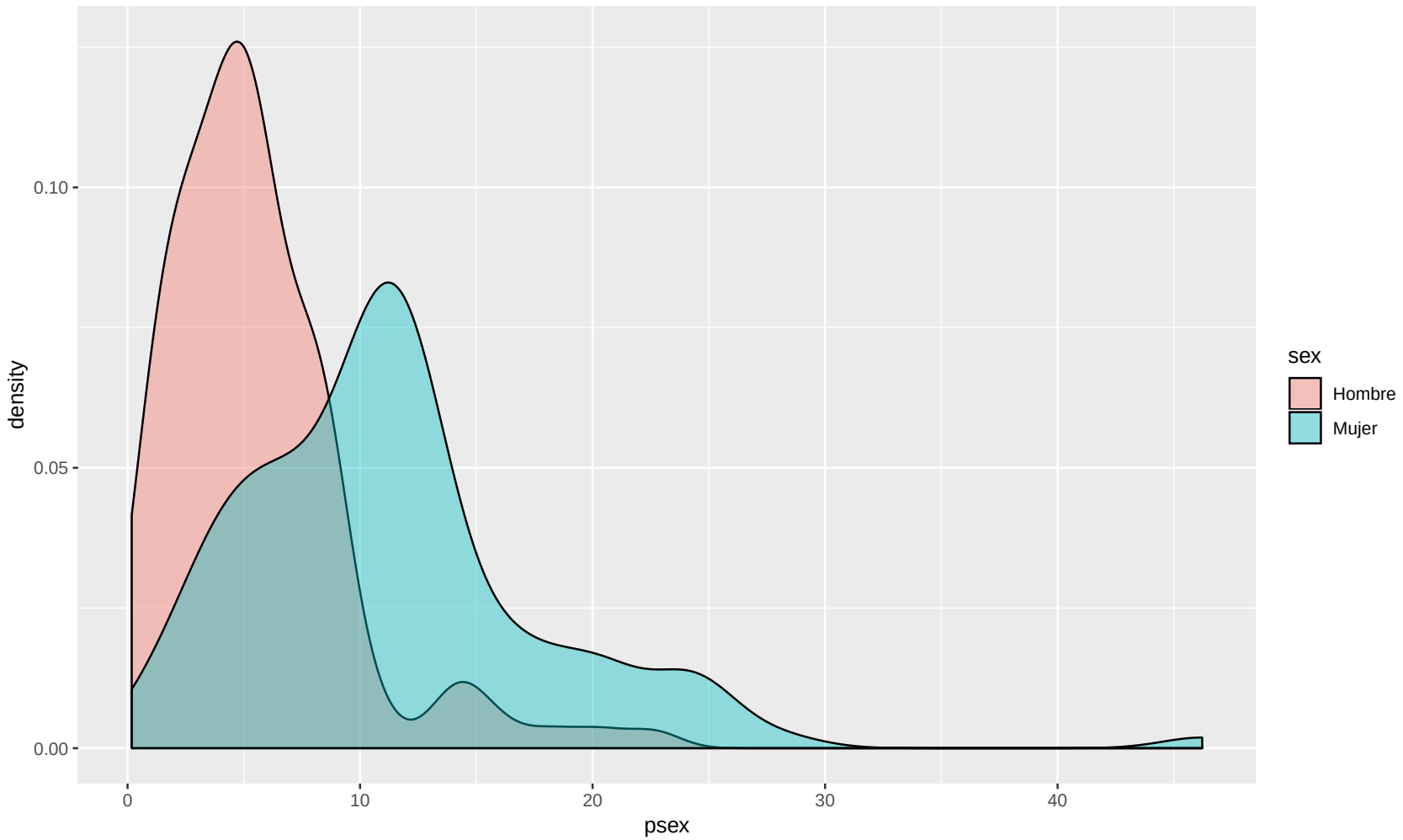
```
##    Municipio Total    tipo ptipo
## 1   Medellín  8.34 Abierto  6.98
## 2    Barbosa 12.99 Abierto  8.73
## 3      Bello  8.82 Abierto  7.17
## 4     Caldas  9.15 Abierto  6.46
## 5 Copacabana  9.42 Abierto  8.14
## 6   Envigado  5.27 Abierto  4.30
```

```
tail(bd1)
```

```
##                Municipio Total   tipo ptipo
## 245              Mutatá  3.09 Oculto  0.74
## 246             Necoclí  5.49 Oculto  0.33
## 247  San Juan de Urabá  1.43 Oculto  0.41
## 248 San Pedro de Urabá  3.69 Oculto  0.00
## 249               Turbo 18.26 Oculto  8.09
## 250   Vigía del Fuerte  3.37 Oculto  0.00
```

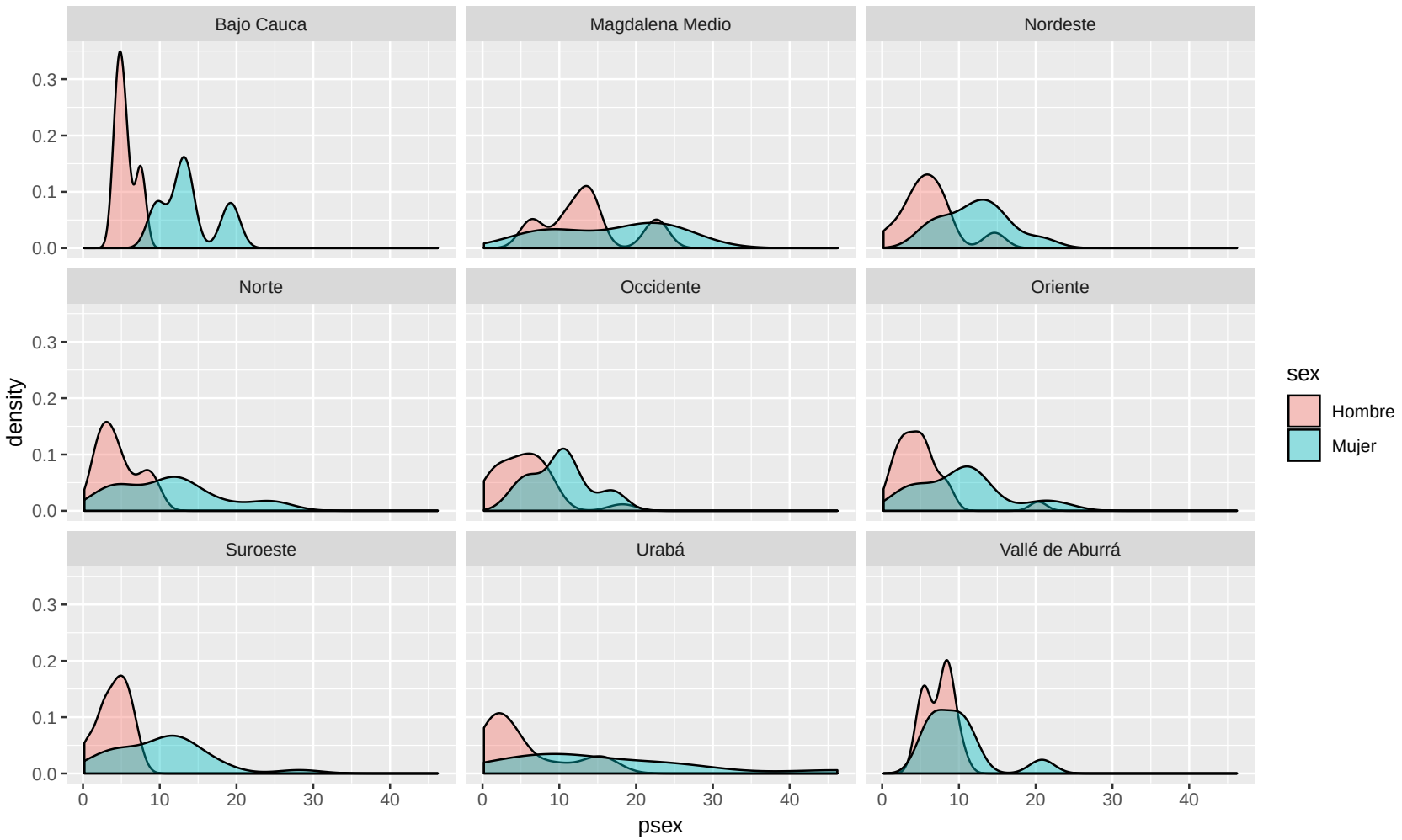
Ejemplo 1: Desempleo en Antioquia

```
ggplot(Datos_13, aes(x=psex, fill=sex)) +
  geom_density(alpha=0.4)
```



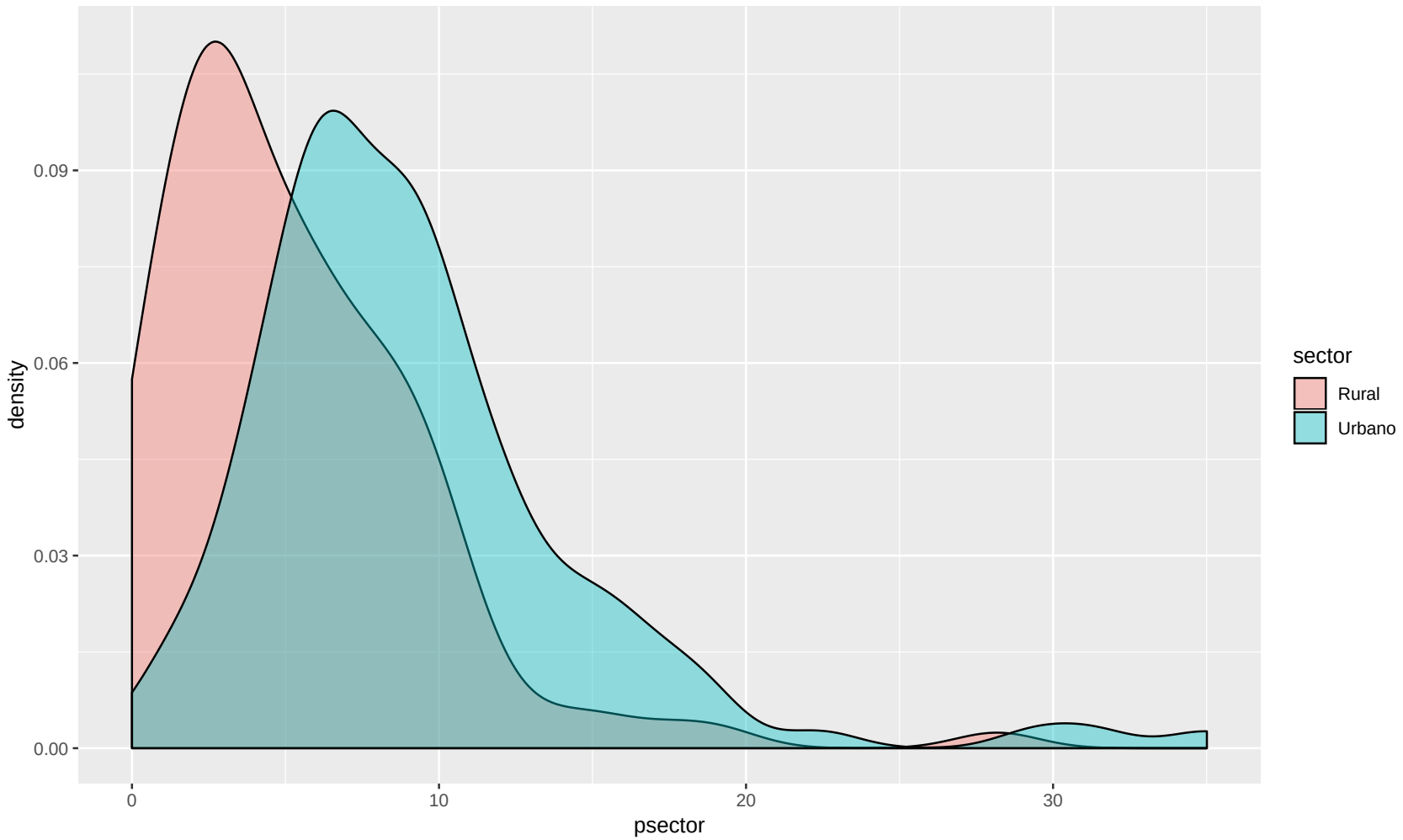
Ejemplo 1: Desempleo en Antioquia

```
ggplot(Datos_13, aes(x=psex, fill=sex)) +
  geom_density(alpha=0.4)+
  facet_wrap(~Subregion)
```



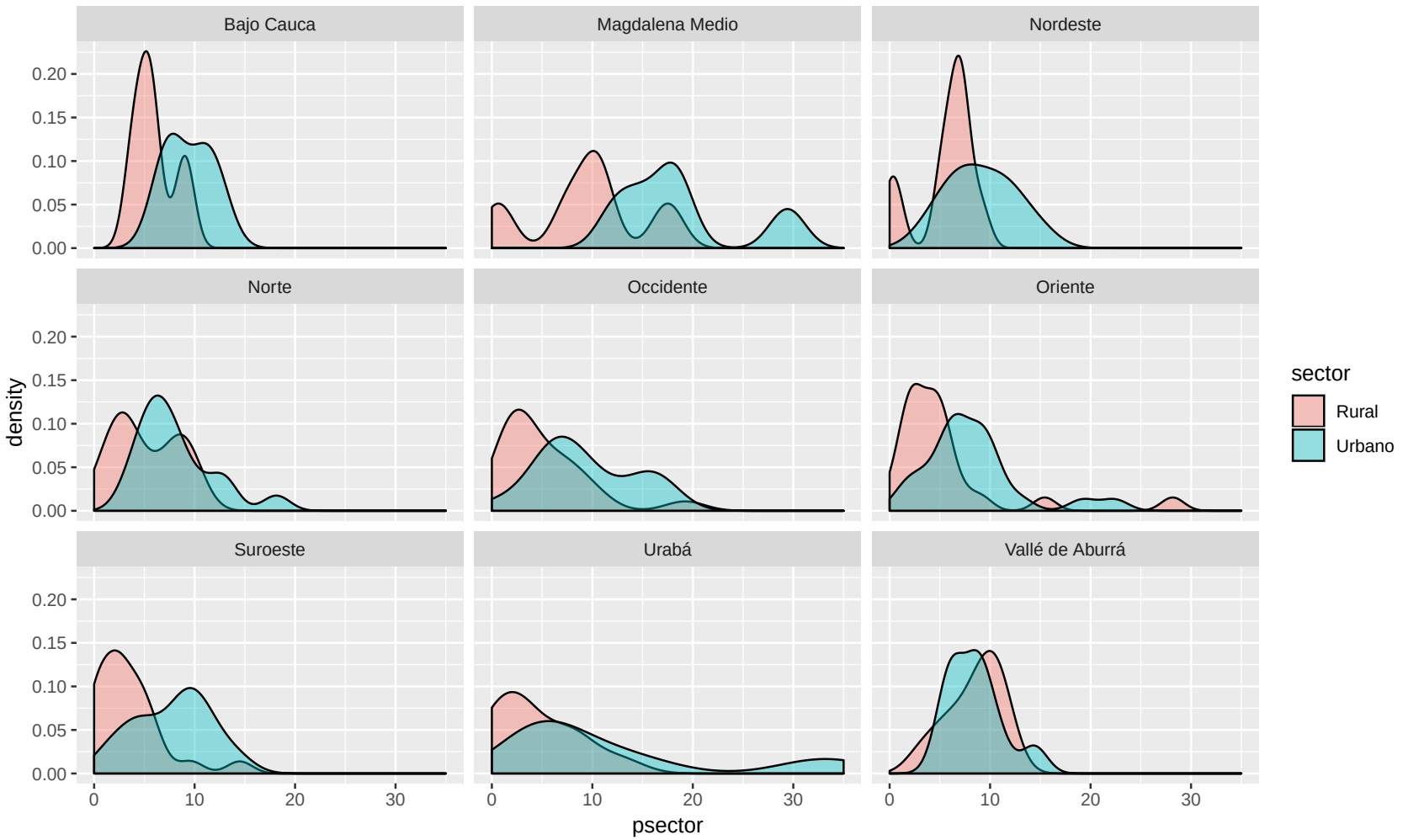
Ejemplo 1: Desempleo en Antioquia

```
ggplot(Datos_13, aes(x=psector, fill=sector)) +
  geom_density(alpha=0.4)
```



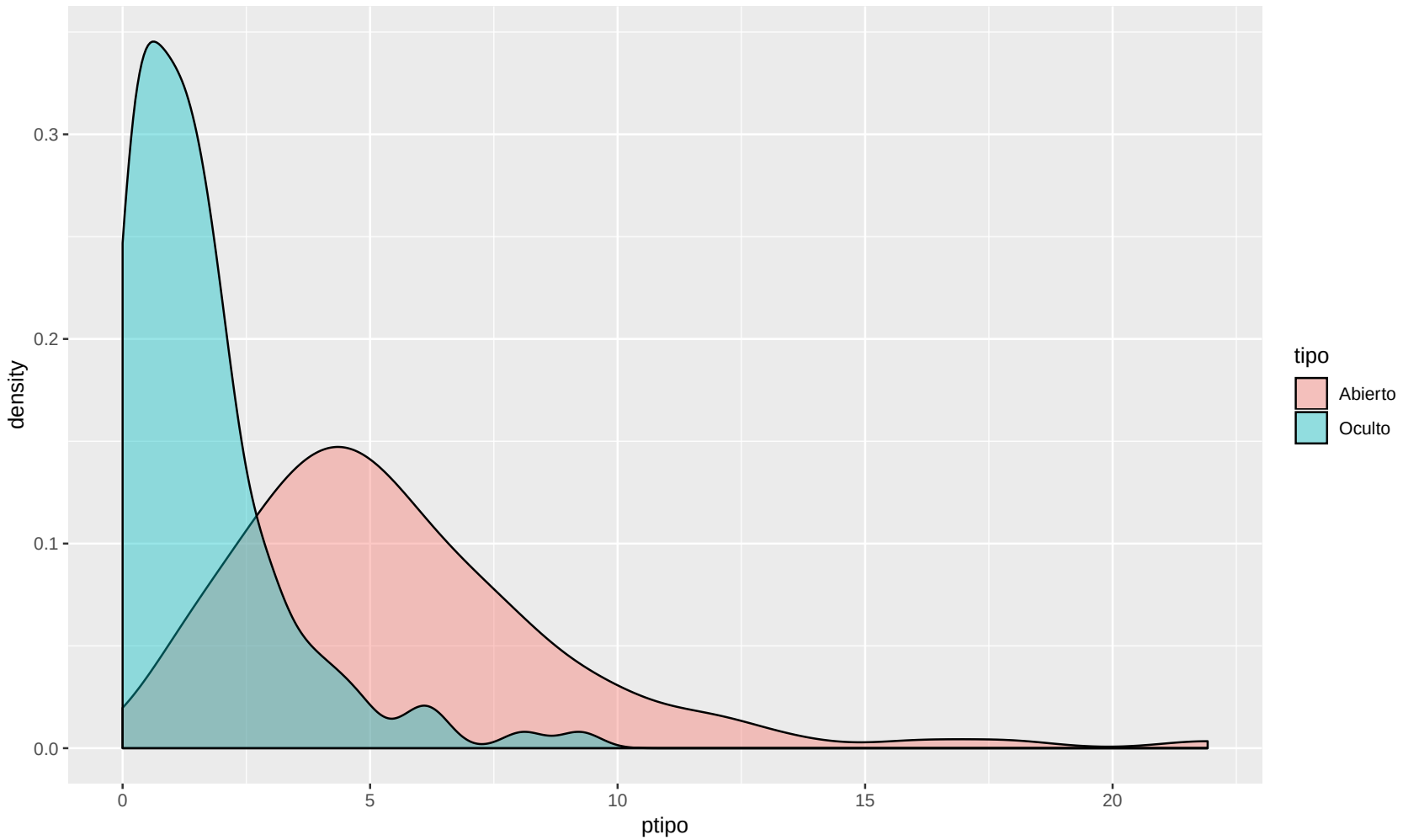
Ejemplo 1: Desempleo en Antioquia

```
ggplot(Datos_13, aes(x=psector, fill=sector)) +
  geom_density(alpha=0.4)+
  facet_wrap(~Subregion)
```



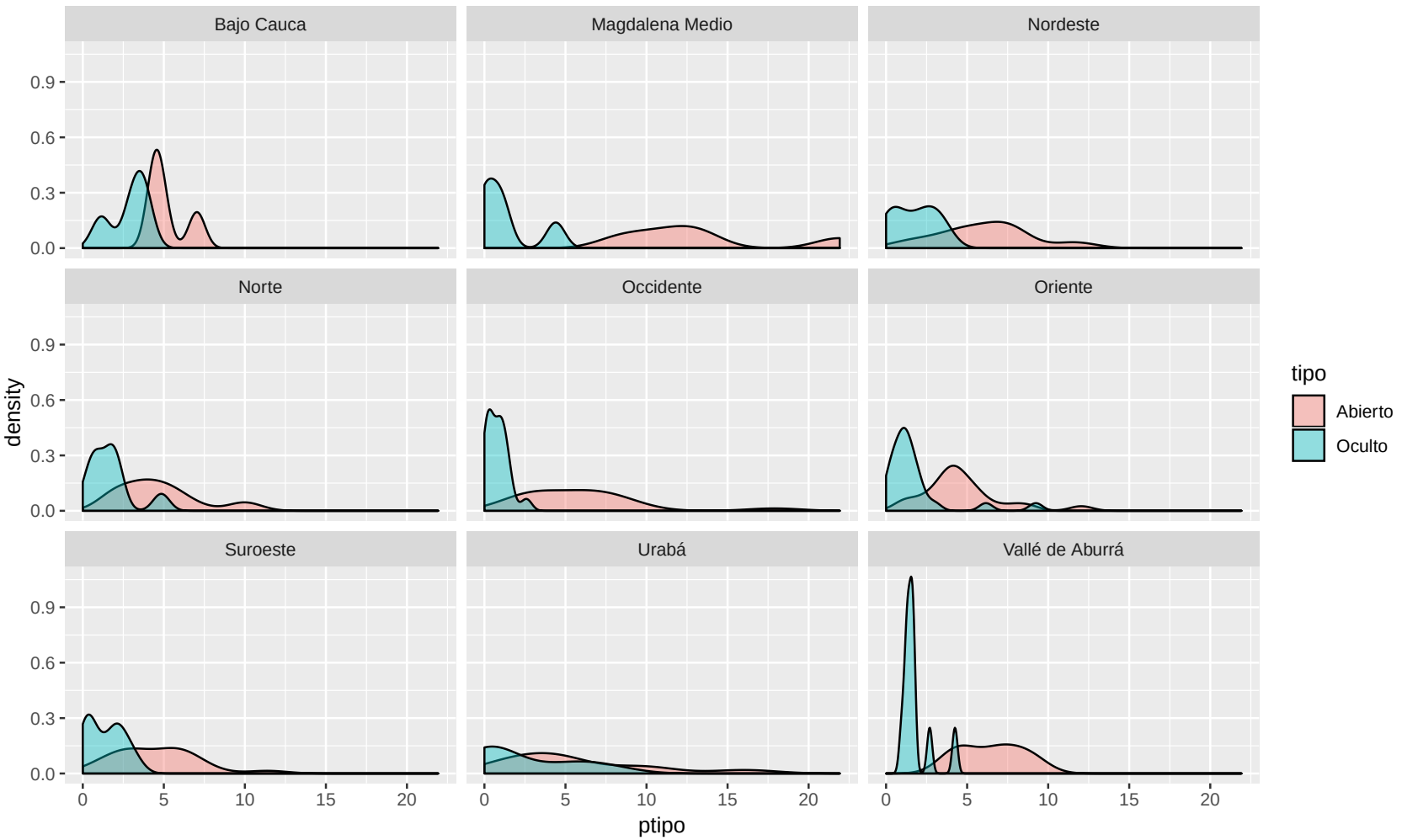
Ejemplo 1: Desempleo en Antioquia

```
ggplot(Datos_13, aes(x=ptipo, fill=tipo)) +
  geom_density(alpha=0.4)
```



Ejemplo 1: Desempleo en Antioquia

```
ggplot(Datos_13, aes(x=ptipo, fill=tipo)) +
  geom_density(alpha=0.4)+
  facet_wrap(~Subregion)
```

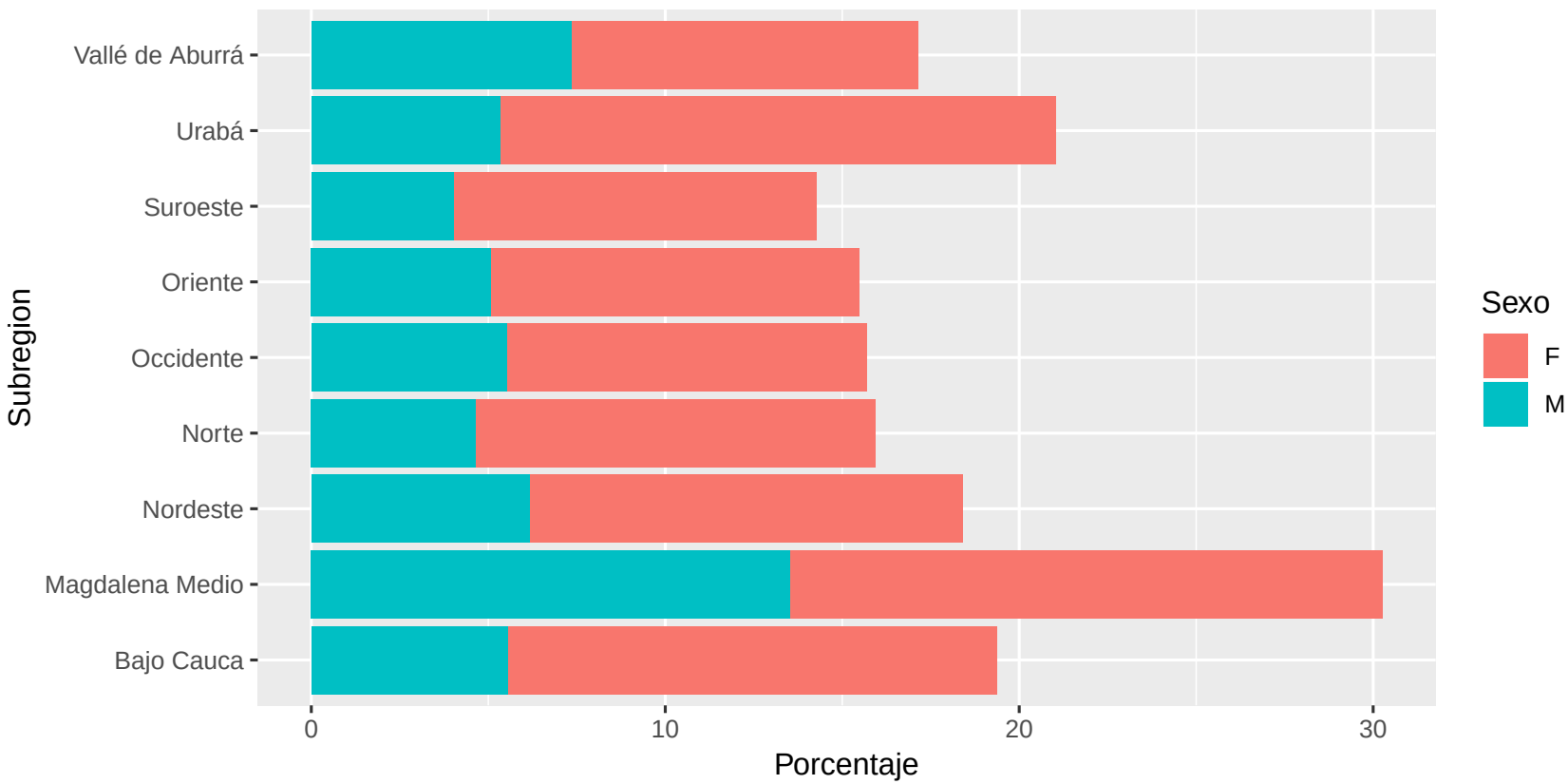


Ejemplo 1: Desempleo en Antioquia

```
require(tibble)
Datos_12<-na.omit(Datos_1)
Porc_M<-tapply(Datos_12$Hombre, Datos_12$Subregion,
               mean)
Porc_F<-tapply(Datos_12$Mujer, Datos_12$Subregion,
               mean)
Subr1<-levels(Datos_12$Subregion)
n1<-length(Subr1)

reg1<-tibble(
Porcentaje=c(Porc_M,Porc_F),
Sexo=c(rep("M",n1),rep("F",n1)),
Subregion=c(Subr1,Subr1)
)
```

Ejemplo 1: Desempleo en Antioquia

```
require(ggplot2)
ggplot(reg1, aes(x = Subregion, y = Porcentaje,
                 fill = Sexo)) +
 geom_bar(stat = "identity")+
    coord_flip()
```



Ejemplo 1: Desempleo en Antioquia

```
subset(Datos_1,subset=(Hombre>Mujer),
       select=c(Municipio, Subregion))
```

```
##                    Municipio       Subregion
## 17                  Caracolí Magdalena Medio
## 21           Puerto Triunfo Magdalena Medio
## 25                  Cisneros        Nordeste
## 35                   Briceño           Norte
## 38                 Donmatías           Norte
## 43  San Andrés de Cuerquia           Norte
## 64              Sabanalarga       Occidente
## 65            San Jerónimo       Occidente
## 75                 El Peñol         Oriente
## 79                   Guarne         Oriente
## 106             Pueblorrico        Suroeste
## 119                 Murindó           Urabá
```

Ejemplo 1: Desempleo en Antioquia

```
subset(Datos_1,subset=(Urbano<(Rural-2)),
       select=c(Municipio, Subregion))
```

```
##                  Municipio       Subregion
## 8                   Itagüí Vallé de Aburrá
## 35                Briceño           Norte
## 38               Donmatías           Norte
## 39              Entrerríos           Norte
## 43 San Andrés de Cuerquia           Norte
## 63                   Peque       Occidente
## 64              Sabanalarga       Occidente
## 74   El Carmen de Viboral         Oriente
## 89              San Rafael         Oriente
```

Ejemplo 1: Desempleo en Antioquia

```
subset(Datos_1,subset=(Oculto>Abierto),
       select=c(Municipio, Subregion))
```

```
##          Municipio Subregion
## 54         Caicedo Occidente
## 87  San Francisco   Oriente
## 93           Andes  Suroeste
## 105    Montebello  Suroeste
```