Datasets

Carlos Meneses

15/4/2020

Carga de un Dataset en un Dataframe

```r
data("iris")
iris_df <- iris
```

10 primeros elementos

```r
head(iris_df, 10)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 2           4.9         3.0          1.4         0.2  setosa
## 3           4.7         3.2          1.3         0.2  setosa
## 4           4.6         3.1          1.5         0.2  setosa
## 5           5.0         3.6          1.4         0.2  setosa
## 6           5.4         3.9          1.7         0.4  setosa
## 7           4.6         3.4          1.4         0.3  setosa
## 8           5.0         3.4          1.5         0.2  setosa
## 9           4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
```

Informacion del Dataframe

```r
str(iris_df)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Nombres de las Variables

```r
names(iris_df) #ó tambien colnames(iris_df)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

Dimensiones del Dataframe

```r
nrow(iris_df)
```

```
## [1] 150
```

```r
ncol(iris_df)
```

```
## [1] 5
```

```r
dim(iris_df)
```

```
## [1] 150   5
```

Acceso a las Variables e Indexado

Para acceder a las variables se usa el simbolo "$".

```r
iris_df$Sepal.Width[1:15]
```

```
##  [1] 3.5 3.0 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 3.7 3.4 3.0 3.0 4.0
```

Indexado

Obtener 15 observaciones aleatorias, con las variables `Petal.Width`, `Sepal.Width` y `Species` del dataframe `Iris`

```r
set.seed(1341)
iris_df[sort(sample(150, 15)), c("Petal.Width", "Sepal.Width", "Species")]
```

```
##     Petal.Width Sepal.Width    Species
## 1           0.2         3.5     setosa
## 10          0.1         3.1     setosa
## 36          0.2         3.2     setosa
## 50          0.2         3.3     setosa
## 55          1.5         2.8 versicolor
## 64          1.4         2.9 versicolor
## 65          1.3         2.9 versicolor
## 68          1.0         2.7 versicolor
## 71          1.8         3.2 versicolor
## 76          1.4         3.0 versicolor
## 86          1.6         3.4 versicolor
## 105         2.2         3.0  virginica
## 125         2.1         3.3  virginica
## 136         2.3         3.0  virginica
## 146         2.3         3.0  virginica
```

Indexado 2

Acceder a los elementos del dataframe `Iris` donde `Petal.Width` <= 1.3

```
iris_df[iris_df$Petal.Length <= 1.3,]
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 3           4.7         3.2          1.3         0.2  setosa
## 14          4.3         3.0          1.1         0.1  setosa
## 15          5.8         4.0          1.2         0.2  setosa
## 17          5.4         3.9          1.3         0.4  setosa
## 23          4.6         3.6          1.0         0.2  setosa
## 36          5.0         3.2          1.2         0.2  setosa
## 37          5.5         3.5          1.3         0.2  setosa
## 39          4.4         3.0          1.3         0.2  setosa
## 41          5.0         3.5          1.3         0.3  setosa
## 42          4.5         2.3          1.3         0.3  setosa
## 43          4.4         3.2          1.3         0.2  setosa
```

Uso de Dplyr

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## 
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```

```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

Indexado 3

Acceder a los elementos del dataframe `Iris` donde `Petal.Width` <= 1.3, usando el paquete `dplyr`

```
iris_df %>% filter(Petal.Length<=1.3)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           4.7         3.2          1.3         0.2  setosa
## 2           4.3         3.0          1.1         0.1  setosa
## 3           5.8         4.0          1.2         0.2  setosa
## 4           5.4         3.9          1.3         0.4  setosa
## 5           4.6         3.6          1.0         0.2  setosa
## 6           5.0         3.2          1.2         0.2  setosa
## 7           5.5         3.5          1.3         0.2  setosa
## 8           4.4         3.0          1.3         0.2  setosa
## 9           5.0         3.5          1.3         0.3  setosa
## 10          4.5         2.3          1.3         0.3  setosa
## 11          4.4         3.2          1.3         0.2  setosa
```

Seleccionar varias columnas

Seleccione las columnas `Petal.Length`, `Sepal.Length` y `Species`

```
iris_df %>% select(Petal.Length, Sepal.Length, Species) %>% head(10)
```

```
##    Petal.Length Sepal.Length Species
## 1           1.4          5.1  setosa
## 2           1.4          4.9  setosa
## 3           1.3          4.7  setosa
## 4           1.5          4.6  setosa
## 5           1.4          5.0  setosa
## 6           1.7          5.4  setosa
## 7           1.4          4.6  setosa
## 8           1.5          5.0  setosa
## 9           1.4          4.4  setosa
## 10          1.5          4.9  setosa
```

Agregar nuevas Variables

Agregue una nueva variable al dataframe `iris_df` en la cual este la relación entre la variable `Petal.Width` y `Petal.Length` llamada `Rel.Petal`

```
iris_df %>% mutate(Rel.Petal=Petal.Width/Petal.Length) %>% head(10)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species  Rel.Petal
## 1           5.1         3.5          1.4         0.2  setosa 0.14285714
## 2           4.9         3.0          1.4         0.2  setosa 0.14285714
## 3           4.7         3.2          1.3         0.2  setosa 0.15384615
## 4           4.6         3.1          1.5         0.2  setosa 0.13333333
## 5           5.0         3.6          1.4         0.2  setosa 0.14285714
## 6           5.4         3.9          1.7         0.4  setosa 0.23529412
## 7           4.6         3.4          1.4         0.3  setosa 0.21428571
## 8           5.0         3.4          1.5         0.2  setosa 0.13333333
## 9           4.4         2.9          1.4         0.2  setosa 0.14285714
## 10          4.9         3.1          1.5         0.1  setosa 0.06666667
```

Ordenar segun las columnas

Ordene las 10 primeras observaciones del dataframe `iris_df` de acuerdo a la variable `Sepal.Width`

```
iris_df %>% arrange(Sepal.Width) %>% head(15)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1           5.0         2.0          3.5         1.0 versicolor
## 2           6.0         2.2          4.0         1.0 versicolor
## 3           6.2         2.2          4.5         1.5 versicolor
## 4           6.0         2.2          5.0         1.5  virginica
## 5           4.5         2.3          1.3         0.3     setosa
## 6           5.5         2.3          4.0         1.3 versicolor
## 7           6.3         2.3          4.4         1.3 versicolor
## 8           5.0         2.3          3.3         1.0 versicolor
## 9           4.9         2.4          3.3         1.0 versicolor
## 10          5.5         2.4          3.8         1.1 versicolor
## 11          5.5         2.4          3.7         1.0 versicolor
## 12          5.6         2.5          3.9         1.1 versicolor
## 13          6.3         2.5          4.9         1.5 versicolor
## 14          5.5         2.5          4.0         1.3 versicolor
## 15          5.1         2.5          3.0         1.1 versicolor
```

Agrupar variables de acuerdo a parametros cualitativos

Agrupe una muestra aleatoria del dataframe `iris_df` de tamaño 15 de acuerdo a la variable `Species`

```
set.seed(1582)
iris_df_sample <- iris_df[sort(sample(150,15)),]
iris_df_sample %>% group_by(Species)
```

```
## # A tibble: 15 x 5
## # Groups:   Species [3]
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##  *        <dbl>       <dbl>        <dbl>       <dbl> <fct>
##  1          4.7         3.2          1.3         0.2 setosa
##  2          4.6         3.4          1.4         0.3 setosa
##  3          4.8         3            1.4         0.1 setosa
##  4          4.3         3            1.1         0.1 setosa
##  5          5.7         3.8          1.7         0.3 setosa
##  6          5.3         3.7          1.5         0.2 setosa
##  7          7           3.2          4.7         1.4 versicolor
##  8          6.1         2.9          4.7         1.4 versicolor
##  9          6.9         3.2          5.7         2.3 virginica
## 10          6.3         2.8          5.1         1.5 virginica
## 11          6.4         3.1          5.5         1.8 virginica
## 12          6.8         3.2          5.9         2.3 virginica
## 13          6.3         2.5          5           1.9 virginica
## 14          6.2         3.4          5.4         2.3 virginica
## 15          5.9         3            5.1         1.8 virginica
```

Calcule la media de las variables `Sepal.Length` y `Sepal.Width` agrupadas por especie

```
library(dplyr)
iris_df %>% group_by(Species) %>% summarise(media_PL=mean(Petal.Length),
                                            media_PW=mean(Petal.Width))
```

```
## # A tibble: 3 x 3
##   Species    media_PL media_PW
##   <fct>         <dbl>    <dbl>
## 1 setosa         1.46    0.246
## 2 versicolor     4.26    1.33
## 3 virginica      5.55    2.03
```