7. Manipulación de datos

Índice

1. Contar observaciones

Una vez preparados y saneados los datos, se prosigue con las operaciones propias del preprocesamiento ya encaminándose al análisis para conocer bien los datos con los que se está trabajando.

Una de estas operaciones habituales es contar el número de observaciones para saber el volumen de cuántas se disponen, tanto en total como por grupos. Esto se puede hacer fácilmente con la ayuda de `dplyr`:

```
library(dplyr)
df <- read.csv('datos_colesterol.csv')

count(df)
count(df, sexo)
```

2. Selección

Otra operación posible es seleccionar variables para generar un subconjunto de los datos, obteniendo algunas columnas concretas:

```
select(df, nombre, sexo, edad)
select(df, 1, 3)
select(df, peso:colesterol)
select(df, -edad)
```

3. Filtrado

También se pueden filtrar observaciones en función de ciertas condiciones, lo que proporciona esta vez un subconjunto por filas:

```
filter(df, sexo == "M")
filter(df, sexo == "H" & edad > 30)
filter(df, !is.na(colesterol))
```

4. Reordenación

Para reordenar las filas de un data frame / tibble:

```
arrange(df, nombre)
arrange(df, sexo, edad)
arrange(df, sexo, desc(colesterol))
```

5. Renombrado

Si se desea cambiar el nombre de las columnas:

```
rename(df, estatura = altura)
```

6. Transformación o creación de columnas

A veces es necesario mostrar los datos existentes procesados mediante alguna operación, u obtener nuevos datos dependientes de los anteriores.

En estos casos se crean fórmulas calculadas a partir de los datos previos y se asignan a una columna anterior, o se crea una nueva columna con los valores procesados:

```
mutate(df, altura = altura*100)
mutate(df, imc = round(peso/altura^2))
```

7. Resúmenes

De cara al análisis estadístico, se puede obtener un resumen del total de los datos para ciertas variables estadísticas comunes:

```
summarise(df, edad_media = mean(edad))
summarise(df, media = mean(colesterol, na.rm=TRUE), sd = sd(colesterol,
na.rm=TRUE))
```

O por grupos:

```
df.sexo <- group_by(df, sexo)

summarise(df.sexo, edad_media = mean(edad))
summarise(df.sexo, media = mean(colesterol, na.rm=TRUE), sd =
sd(colesterol, na.rm=TRUE))
```

8. Unión

Dados dos data frames, se pueden unir (al estilo de los JOINs de las bases de datos relacionales):

```
df2 <- read.csv('datos_personales.csv')

dfTotal <- merge(df, df2, by = "nombre")
```

Referencias

Preprocesamiento de datos
Función merge

dplyr documentation
merge documentation