UNIDAD DIDÁCTICA 1: Preprocesamiento de datos
Tema 1: Resumen de datos

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE I

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones

Índice

1. Introducción

- El examen de los datos es un paso necesario para un tratamiento adecuado de los datos
- Un análisis cuidadoso conduce a una mejor predicción y a una evaluación más precisa de la dimensionalidad.
- Examinar tanto las variables individuales como las relaciones entre ellas.
- Evaluar y solucionar los problemas en el diseño de la investigación y en la recogida de datos

2. Manipulación de Base de Datos con la librería dplyr

```
library(dplyr)
counties <- read.table("counties.csv",header = T,sep = ";")

#Visualizamos la estructura del dataset
str( )

#también lo podemos hacer con glimpse()
glimpse()

#Contar observaciones count()
count()

# Contamos por las etiquetas de una variable y ordenamos de mayor a menor
count( , sort = T)

#Distinguimos los diferentes valores/etiquetas de la variable state
distinct( )

#Selección de variables
select()

# Renombrar variables
rename( )

#Eliminar, borrar o deseleccionar variables (columnas)
select(- var)

#Seleccionar variables en función de su nombre

contains(): columnas que contengan dicho término en el nombre de la variables.
start_with(): columnas que empiecen por dicho término.
ends_with(): columnas que terminen por dicho término.
last_col(): selecciona la última columna.
```

```
#Seleccionar variables por búsqueda de caracteres
select(matches())

#Ordenar todo el dataset por una variable dada de forma ascendente
arrange()

#Seleccionar variables y ordenar de forma descendente
arrange(desc())

#Filtrado de datos
filter()
```

Ejercicio 1:

Selecciona los registros cuyos ingresos medios sean mayores a 35000 y cuyo porcentaje de población blanca sea menor al 85%. Ordena los registros (de mayor a menor) en función del número de empleados. Después de aplicar este filtro, selecciona sólo las variables state, county, employed, income y white.

```
#Filtrado de varios términos para una misma variable:
filter(var %in% c("cat1","cat2"))

#Filtrado and (&) y or (|)
filter(var1 == "cat1" & var2 == "cat2" )
```

Ejercicio 2:

Selecciona los registros de la región sur y del oeste que tengan metro.

```
#Filtrado por índice de una fila slice()
slice( )

# Renombrar etiquetas recode ()

recode()


# Crear nuevas variables (mutate()) en base a un conjunto de condiciones
mutate()
ifelse() y case_when()


#Selección y creación de nuevas variables transmute(). Combina select() + mutate():
transmute()
```

3. Eliminar datos duplicados de una base de datos

```
#Hacemos una base de datos de ejemplo:

id <- 1:200
sexo <- c(rep("hombre",100),rep("mujer",100))
pais <- c(rep("Francia",30),rep("Italia",35),rep("Portugal",45),rep("Suiza",
35),rep("Grecia",55))
edad<- c(rep("adolescente",
30),rep("joven",40),rep("adulto",100),rep("anciano",30))

datos <- data.frame (id, sexo, pais, edad)

#Hacemos uso de duplicated() y de distinct() (este último de la librería
dplyr) para explorar los duplicados del dataset.

duplicated()
distinct()

#Hacemos uso de nrow() y count() para contar los registros.

nrow()
count()

#Hacemos uso de filter() para seleccionar los registros, duplicados y no
duplicados

filter()
```

Ejercicio 3:

1. Instala la librería babynames.
2. Visualiza la estructura de la base de datos.
3. Muestra los distintos nombres de la variable name del dataset.
4. Cuenta los registros duplicados de la variable name.
5. Cuenta los registros no duplicados de la variable name.
6. Elimina del dataset los duplicados de la variable name.
7. Selecciona los registros duplicados de la variable name.

4. Resumen de los datos

```
library(dplyr)

library(psych)

counties <- read.table("counties.csv",header = T,sep = ";")

#Resumen descriptivo de las variables summarize()

summarize(Media = mean(),
          Sd = sd(),
          CV = round(sd() / mean() * 100, 2),
          Min = min(),
          Q1 = quantile(, 0.25),
          Q2 = quantile(, 0.50),
          Q3 = quantile(, 0.75),
          RIQ = IQR(),
          Max = max())

#También se puede hacer uso de describe() de la librería psych

describe()

#Resumen descriptivo de las variables haciendo uso de group_by(), permite
la agrupación de variables categóricas

group_by()

# los n valores más altos

top_n()
```

Ejercicio 4:

1. Obtén un resumen descriptivo de los ingresos per cápita de cada una de las regiones.
2. ¿Qué 3 registros poseen los ingresos per cápita más altos en cada una de las regiones? Visualiza la región, el estado, el condado y los ingresos per cápita.

Ejercicio 5:

En este ejercicio se va a realizar un resumen de los datos correspondiente al dataset "gapminder" de la librería que lleva su nombre.

1. Visualiza los 5 primeros registros de la base de datos.
2. Muestra el número de filas y de columnas del dataset, así como el nombre de las variables y su tipología.
3. Filtrar todos los datos que sean de Perú del año 2002 y selecciona la columna país, año, esperanza de vida y población.
4. Calcula la media, la desviación típica, los cuartiles, el rango intercuartílico, el mínimo, el máximo y el coeficiente de variación para la variable lifeExp en el año 2007 en cada continente. ¿En qué continente existe una mayor variabilidad en la esperanza de vida?¿En qué continente(s) la media ofrece un valor más representativo de la realidad?
5. ¿Cuáles son los 8 países con un percentil inferior en la variable lifeExp para el año 2007? Muestra el continente, el país, la esperanza de vida y el percentil (utiliza la función ntile()).
6. Muestra, por año, el número de habitantes totales de América, África y Europa.
7. Muestra, mediante dos gráficos de barras, el cambio en el PIB per cápita de España, Reino Unido, Francia, Alemania e Italia en los años 1952 y 2007.
8. Realiza un histograma, por continente, para la esperanza de vida.
9. Crea un gráfico de dispersión para las variables gdpPercap y lifeExp para los datos del año 2007. El color del punto debe variar en función del continente al que pertenezca dicho país y el tamaño del punto debe ser proporcional al número de habitantes de dicho país. Aplica el tema theme_bw().
10. Muestra, mediante gráficos de cajas, la esperanza de vida de los continentes.