

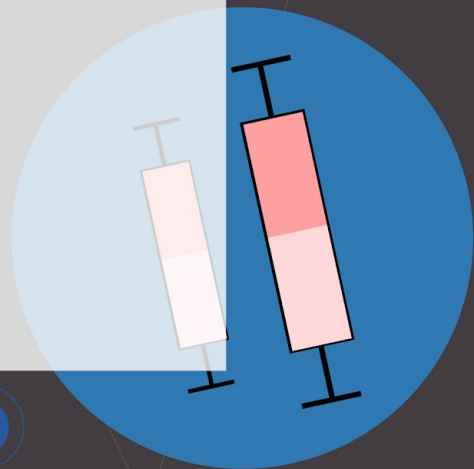
Herramientas Computacionales para Data Science

Clase 3: Introducción a Estadística con R

Ana María Alvarado y María Constanza Prado

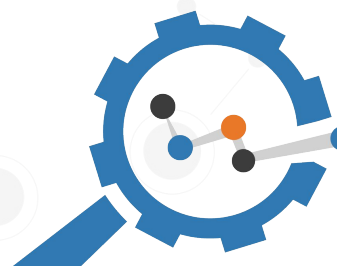


FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE



Contenido del curso (relacionado con R)

1. Introducción a R
2. Introducción a R parte II
- 3. Manipulación e importación de bases de datos**
4. Limpieza y filtrado de datos con R
5. Creación de funciones
6. Análisis descriptivo con R
7. Análisis descriptivo con R parte II
8. Introducción a la Estadística con R
9. Taller de Estructuras Estadísticas



La clase de hoy:



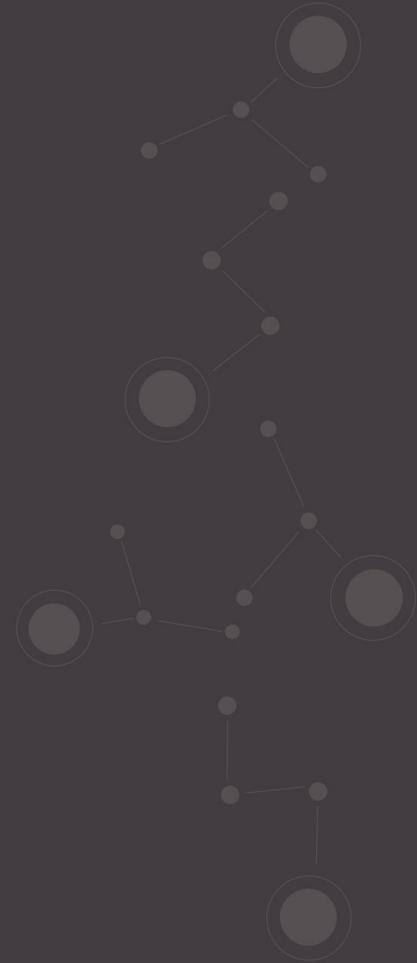
Parte I

- Repaso
- Manipulación de Objetos
- Importación y exploración de datos
- Manipulación de variables

Parte II

- Taller práctico

Repaso



Operadores

Aritméticos

+	Suma
-	Resta
*	Multiplicación
/	División
%/%	División entera

Comparación

<	menor que
>	mayor que
<=	menor o igual que
>=	mayor o igual que
==	igualdad
!=	diferencia

Lógicos

!x	negación
x & y	intersección
x y	unión

Objetos

Escalares

```
> x <- 5  
> x  
[1] 5
```

Operador de asignación

Alt + -

Vectores

```
> x <- c(3,4,5,6,7)  
> x  
[1] 3 4 5 6 7
```

```
> dias <- c("L", "M", "W", "J", "V")  
> dias  
[1] "L" "M" "W" "J" "V"
```

```
> xdias <- c(x,dias)  
> xdias  
[1] "3" "4" "5" "6" "7" "L" "M" "W" "J" "V"
```

Data Frame

```
> df <- data.frame(x,dias)  
> df  
  x dias  
1 3    L  
2 4    M  
3 5    W  
4 6    J  
5 7    V
```

Listas

```
lista <- list(x,dias,df)  
lista
```

Generación de objetos

`n1:n2`

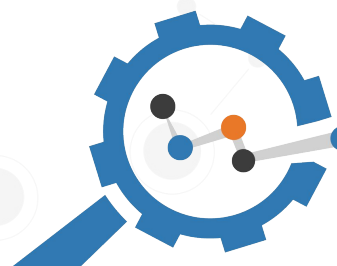
Crea datos enteros de **`n1`** hasta **`n2`**.

`seq(from, to, by, length.out)`

Crea secuencias desde un punto inicial **`from`** y un punto final **`to`**, con argumentos para agregar la distancia entre los dos valores (**`by`**), o la cantidad total de números entre ellos (**`length.out`**).

`rep(x, n)`

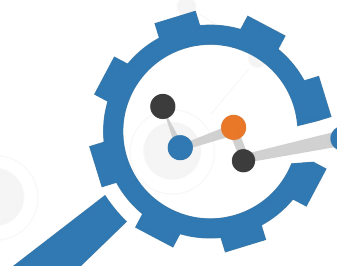
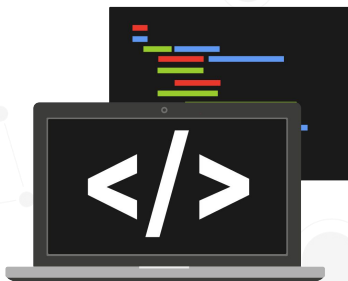
Replica el objeto **`x`** un total de **`n`** veces.



Factores

Los factores, que pueden ser ordenados o no ordenados, se utilizan para representar variables de naturaleza categórica.

```
factor(x = character(), levels, labels = levels,  
exclude = NA, ordered = is.ordered(x), nmax = NA)
```



Fechas

El formato que utiliza R para almacenar fechas es **"yyyy-mm-dd"**

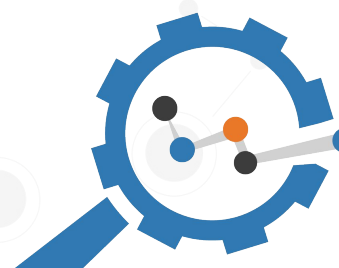
```
> Hoy <- Sys.Date()  
> Hoy  
[1] "2020-06-25"
```

Para introducir una fecha en R usamos la función **as.Date()**

```
> Fecha <- as.Date("2021-06-25")  
> Fecha  
[1] "2021-06-25"
```

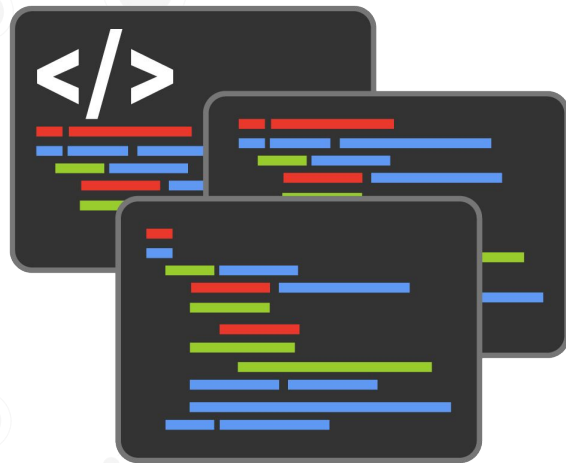
También podemos especificar el formato de entrada.

```
> Fecha2 <- as.Date("25-06-2021", format = "%d-%m-%Y")  
> Fecha2  
[1] "2021-06-25"
```



Funciones

<code>c()</code>	Concatena objetos (variables, textos, números, etc.)
<code>help()</code>	Ayuda respecto de alguna función
<code>library()</code>	Carga de librerías
<code>ls()</code>	Lista de objetos
<code>rm()</code>	Eliminar objetos
<code>abs()</code>	Valor absoluto
<code>sqrt()</code>	Raíz cuadrada
<code>exp()</code>	Exponencial
<code>log10()</code>	Logaritmo base 10
<code>log()</code>	Logaritmo natural
<code>round()</code>	Redondear
<code>mean()</code>	Promedio aritmético
<code>sum()</code>	Suma



Packages

```
install.packages ("nombre_paquete")
```

```
library(nombre_paquete)
```

readxl

MASS

rgl

caret

randomforest

randomForestSRC

qcc

shiny

zoo

forecast

plyr

Knitr

xtable

tidyverse

ggplot2

tibble

tidyr

readr

purrr

dplyr

stringr

forcats

gbm

xgboost

e1071

Liblinear

kernelab

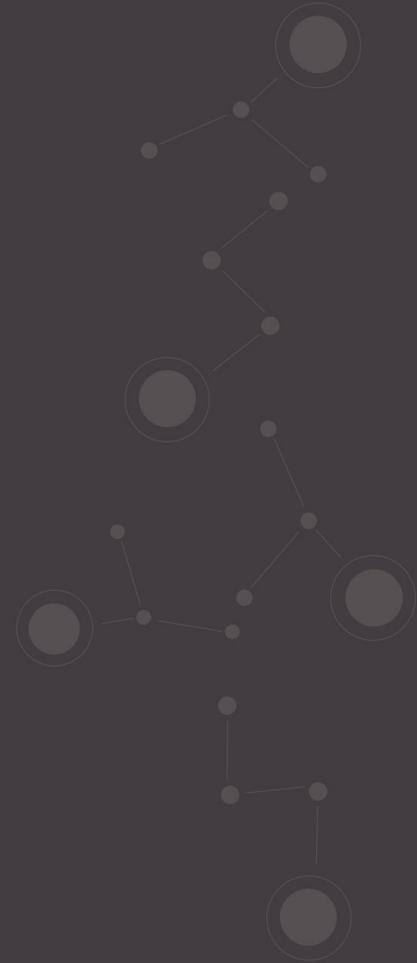
glmnet

gam

cluster

La función **search()** permite ver los *packages* actuales en funcionamiento.

Manipulación de Objetos



Manipulación de objetos

Para encontrar uno o más elementos que pertenecen a un objeto, se puede acceder a través del uso de los **paréntesis de corchete**.

Selección por posición

Un `vector[i]` entregará el valor en la posición `i`.

Un `vector[-i]` entregará todos los valores, excepto el de la posición `i`.

```
> vector <- c("A", "B", "C", "D", "E")  
> vector[2]  
[1] "B"
```

Manipulación de objetos

Selección por comparación

Se selecciona el elemento cuya condición establecida por el operador de comparación es VERDADERA

```
> vector <- c(1,2,3,4,5)
> selec <- vector > 3
> selec
[1] FALSE FALSE FALSE TRUE TRUE
> vector[selec]
[1] 4 5
```

Manipulación de Data Frames

Análogo para matriz y data frame, donde se utiliza:

`df[i, j]`



Filas

Columnas

`df <- data.frame(...)`

Manipulación de Data Frames

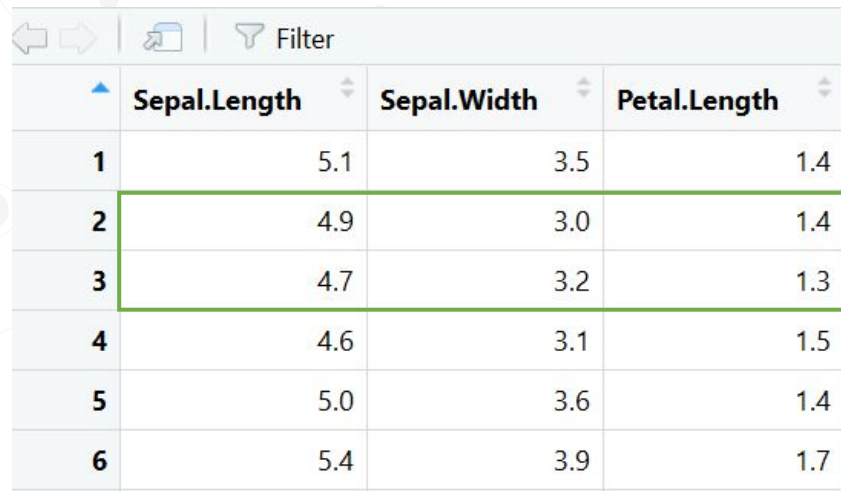
	Sepal.Length	Sepal.Width	Petal.Length
1	5.1	3.5	1.4
2	4.9	3.0	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5.0	3.6	1.4
6	5.4	3.9	1.7

Selecciona filas 2 a 5 y columna 2

`df[2:5,2]`

`df[2:5, "Sepal.Width"]`

Manipulación de Data Frames



The image shows a data frame interface. At the top, there is a light blue header bar with navigation icons (back, forward, search) and a 'Filter' label. Below this is a table with four columns: 'Sepal.Length', 'Sepal.Width', and 'Petal.Length'. The first column is an index from 1 to 6. Rows 2 and 3 are highlighted with a green border. A green arrow points from the text 'Selecciona las filas 2 y 3' to the highlighted rows.

	Sepal.Length	Sepal.Width	Petal.Length
1	5.1	3.5	1.4
2	4.9	3.0	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5.0	3.6	1.4
6	5.4	3.9	1.7

Selecciona las filas 2 y 3

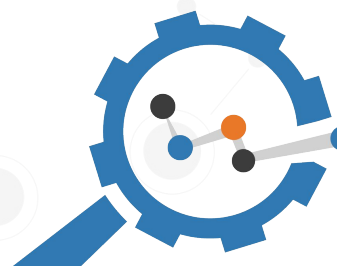
`df[2:3,]`

Manipulación de Data Frames

La función `subset()` funciona como un atajo para hacer lo mismo que hizo en los ejercicios anteriores.

```
subset(x = df, subset = Sepal.Length > 3.5,  
       select = c("Sepal.Length", "Species"))
```

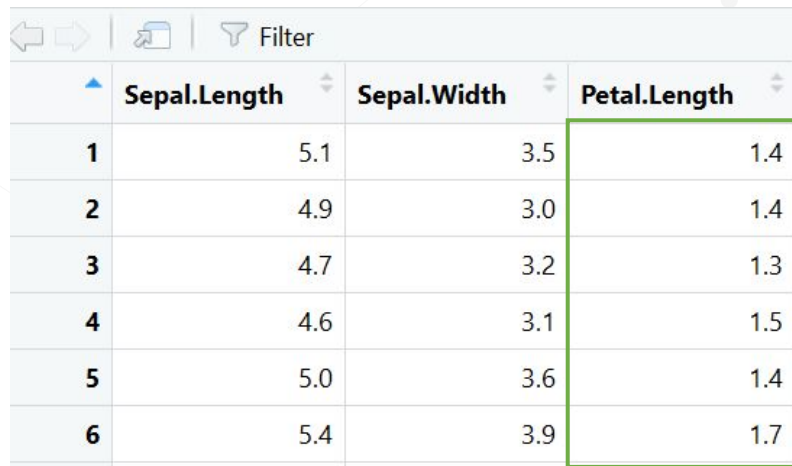
	▲ Sepal.Length ▼	Species ▼
1	5.1	setosa
2	4.9	setosa
3	4.7	setosa
4	4.6	setosa
5	5.0	setosa
6	5.4	setosa
7	4.6	setosa
8	5.0	setosa
9	4.4	setosa



Manipulación de Data Frames

El comando `attach()` reconoce cada variable del data frame como un objeto independiente.

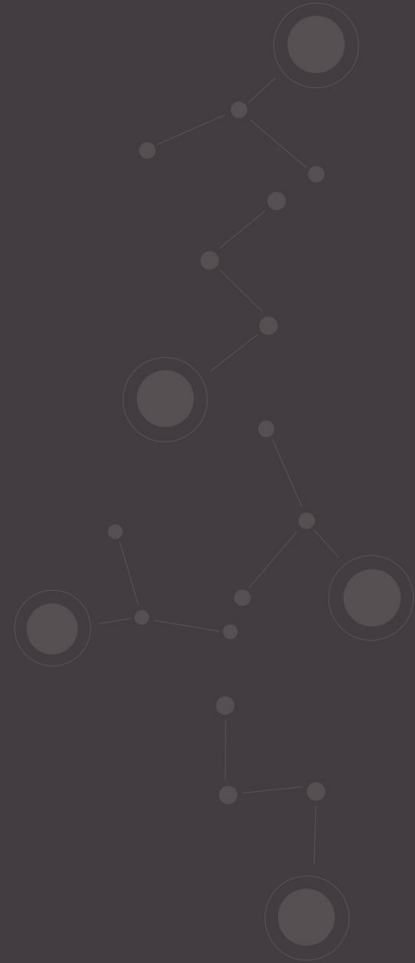
Por otra parte, los objetos se pueden conservar encadenados a través de otro objeto con el comando `$`, por ejemplo, `df$x`. De esta forma manipularemos objetos de un data frame en este curso.



	Sepal.Length	Sepal.Width	Petal.Length
1	5.1	3.5	1.4
2	4.9	3.0	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5.0	3.6	1.4
6	5.4	3.9	1.7

`df$Petal.Length`

Importación y exploración de datos



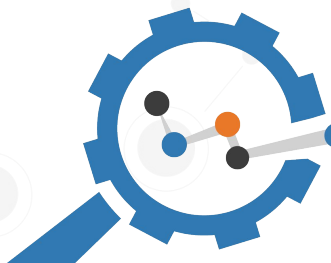
Importación de datos

Ahora, corresponde aprender a **cargar una base de datos.**

Cosas a considerar:

- Tipo de archivo
- Título en la primera fila
- Separador
- Tipo de missing
- Decimales
- Otros...

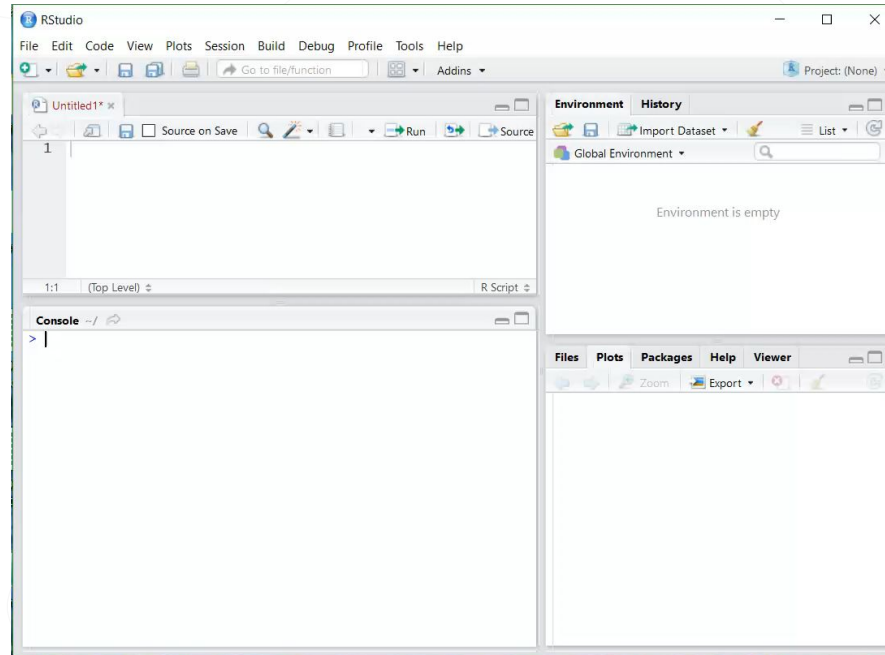
El comando usual será **read**.



Importación de datos

En esta oportunidad accederemos a través de las herramientas facilitadores de Rstudio.

**File /
Import Dataset /
From ...**



Exploración de Data Frames

Luego que la base de datos (df) esta cargada, los primeros comandos para operar son los siguientes:

head(df,k) : Muestra los primeros k registros.

tail(df,k) : Muestra los últimos k registros.

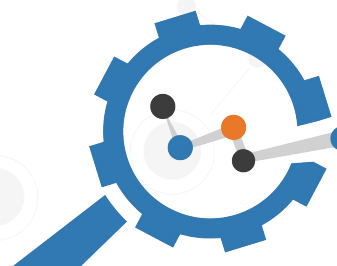
dim(df) : Filas y columnas de un objeto.

length(df) : Número de objetos dentro del objeto BD.

str(df) : Estructura de la base de datos BD.

class(df) : Naturaleza del objeto (similar a is).

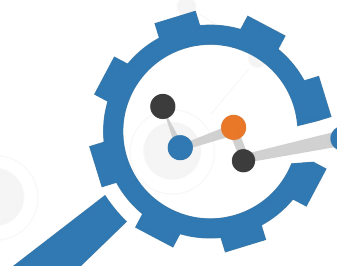
names(df) : Nombres.



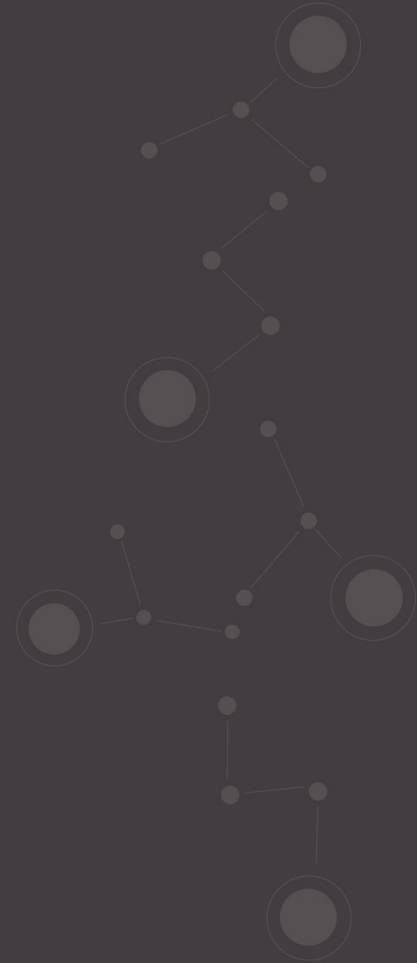
Actividad I

El archivo **Encuesta.xlsx** posee un extracto de la Encuesta Nacional de Seguridad Ciudadana, con sus principales variables.

¡Cargue la base de datos y explore!



Manipulación de variables



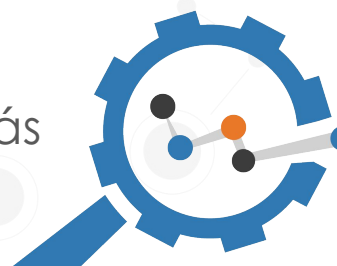
Manipulación de variables

grep(atributo, objeto): Buscar el atributo en cierto objeto, reportando las filas donde se puede encontrar

gsub(inicial, final, objeto): Reemplazar un valor inicial por uno final en cierto objeto.

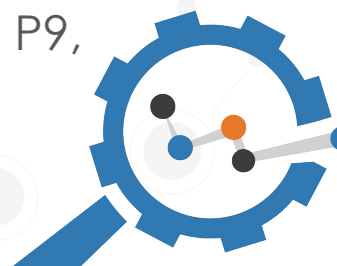
substr(objeto, inicial, final): Substraer de un objeto un carácter entre la posiciones inicial y final.

paste(objeto1,..., sep=" "): Permite unir uno o más objetos, en formato carácter.



Actividad II

- i. Transforme la variable Región, para que en cada registro ya no diga “**Región**” si no que diga “**R.**”.
- ii. Ejercicio, cambie el nombre de las columnas de la data de **Pregunta i** a **Pi**.
- iii. Cree una columna en el data frame que describa el genero y la edad de la persona.
- iv. Construya un nuevo objeto que contenga sólo Hombres de Valparaíso.
- v. Del objeto anterior conserve sólo las variables P1, P3, P8, P9, P21, P64 y P156



Exploración de variables

La función **summary()** entrega un resumen descriptivo de todas las variables.

La función **table()** construye tablas de frecuencia y tablas de contingencia para variables categóricas.

La función **prop.table()** transforma la tabla de frecuencia a porcentaje y proporciones.

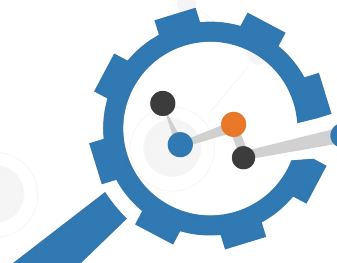
La función **aggregate()** divide los datos en subconjuntos Y, calcula alguna función estadística FUN sobre X para cada uno de ellos y devuelve el resultado en un formato apropiado.

aggregate(X~Y, FUN)



Actividad III

- i. ¿Según la encuesta, cuántos hombres de Valparaíso están casados?
- ii. Obtenga una tabla de frecuencias relativas para el estado civil de los hombres de Valparaíso.
- iii. Obtenga el promedio de edad (P8) para aquellos que creen que serán víctimas de un delito (P64).
- iv. ¿Existe relación entre la edad y percepción de seguridad?

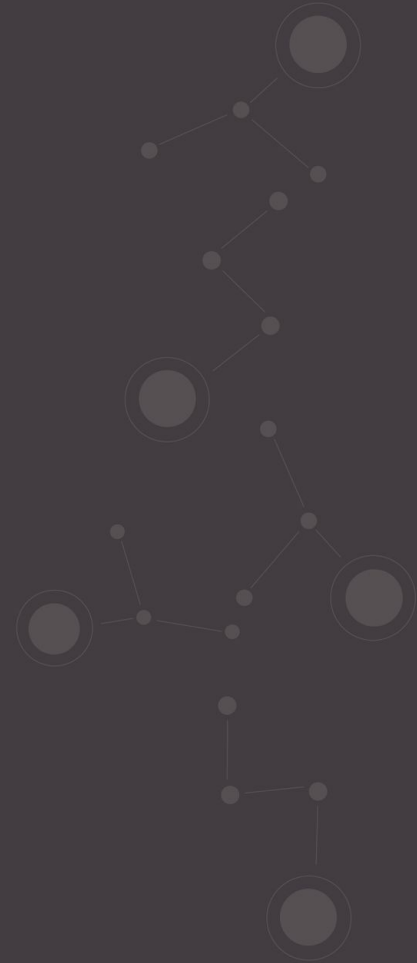




Volvemos en...



Taller Práctico



Taller práctico

La base de datos “Defunciones2016.xlsx” contiene información sobre las defunciones ocurridas el año 2016. Genere un código R que reporte indicadores por región en una única tabla (o data frame) con las siguientes columnas:

- REGION: Región (Escrito como: “Región de Tarapacá”, ...)
- EDAD_PROM: Edad promedio de defunción por región.
- N_DEF: Número de defunciones por región
- INDICE_MASC: índice de masculinidad de defunción.
- P_DEF15: Porcentaje de defunciones de menores a 15 años.
- P_DEFJ: Porcentaje de defunciones asociadas a la causa de muerte J.

Taller práctico

- i. Importe la base de datos DEF (Defunciones 2016).
- ii. Explore el contenido de la base de datos, sus dimensiones, nombres, etc. con el fin de identificar las columnas de su tabla final o la forma de construirla.
- iii. Construya una variable que represente el año de nacimiento de la persona.
- iv. Construya una variable que represente el año de fallecimiento de la persona.
- v. Obtenga la Edad promedio de defunción por región.
- vi. Calcule el número de defunciones ocurridos el 2016 por región.
- vii. Calcule la cantidad de mujeres fallecidas por región, calcule la cantidad de hombres fallecidos por región.

Taller práctico

- viii. Obtenga el índice de masculinidad de defunción.
- ix. Genere un indicador de los fallecidos menores de 15 años. Calcule la proporción de defunciones de menores de 15 años.
- x. Calcule el porcentaje de defunciones asociados a la causa de muerte J, para eso obtenga la primera letra del Diag1, y exponga el porcentaje de muertes por esa causa.
- xi. Asigne nombres a las columnas de su tabla final.

Referencias y material complementario

- <https://cran.r-project.org/web/packages/janitor/vignettes/janitor.html>
(Ejemplos de funciones del paquete janitor) este paquete contiene funciones para examinar y limpiar datos. (La función `clean_names()` es muy útil cuando importamos un set de datos y los nombres de las variables no tienen un formato adecuado).
- <https://cran.r-project.org/web/views/> Paquetes de R organizados por tema.

Referencias y material complementario

Conjunto de símbolos para definir formato de fecha
(los ejemplos corresponden al 13 de enero de 1982):

- %Y: 4-digit year (1982)
- %y: 2-digit year (82)
- %m: 2-digit month (01)
- %d: 2-digit day of the month (13)
- %A: weekday (Wednesday)
- %a: abbreviated weekday (Wed)
- %B: month (January)
- %b: abbreviated month (Jan)

Los siguientes comandos R crearán el mismo objeto Date para el día 13 de enero de 1982:

```
as.Date("1982-01-13")  
as.Date("Jan-13-82", format = "%b-%d-%y")  
as.Date("13 January, 1982", format = "%d %B, %Y")
```