Estadística Aplicada y Procesamiento de Datos con R

Repositorio ayudantía

Ayudantía 1: Reforzamiento tidyverse, ggplot y rmarkdown

Sofía Madariaga
<sofia.madariaga@mail.udp.cl>
s-madariaga

udp Unidad de Postgrados
FACULTAD DE CIENCIAS SOCIALES E HISTORIA

Objetivos de la ayudantía

- Repasar algunos conceptos básicos.
- Repaso y ejercicios: manejo de bases de datos en R con `dplyr`.
- Repaso y ejercicios: gráficos con `ggplot`.
- Ejercicios 5 y 6.

Introducción **A modo de repaso**

- `R` es un lenguaje enfocado en aplicaciones estadísticas y Rstudio es el entorno de desarrollo integrado más utilizado para su uso.
- Trabajar con lenguajes de programación nos permite trabajar de manera reproducible, transparente y flexible (paradigma ciencia *open source*).
- Para ello, es importante repasar en esta ayudantía:
  - Trabajar en **proyectos**.
  - Generar un script de código reproducible y con buenas prácticas. **El ideal, es registrar todos nuestros pasos en código en nuestro script**, a ser posible.

- Trabajar en proyectos es una buena práctica en R, ya que me permite organizar el trbaajo en R.
- Cuando se crea un nuevo proyecto, se genera un archivo de extensión `.Rproj` que me permite gestionar los archivos del directorio y recursos de R. En suma, nos ayuda a hacer el trabajo más reproducible y organizado.

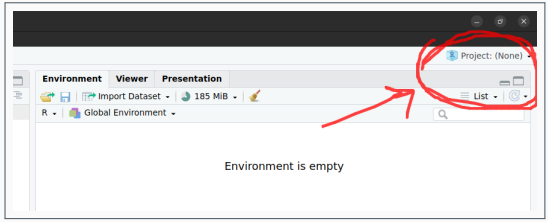


Figura 1: Menú de Proyecto

Ejercicio para iniciar

1. Cree un nuevo proyecto.

   `Menú de proyecto > New project > New Directory o Existing directory`

2. En la nueva carpeta, acomode los materiales para esta ayudantía.
   - Guía de ejercicios: link
   - Genere un nuevo script y copie el código para esta ayudantía: link
   - Genere una carpeta input e output.

Parte 1: Tidyverse

- **Página oficial:** https://www.tidyverse.org/

- Paquete que carga una **colección de paquetes** de gran utilidad para trabajar de manera ordenada y armoniosa.
- Uso de **pipe** %>% permite una sintaxis ordenada.
- **Tidyverse style guide:** https://style.tidyverse.org/

- **dplyr**
- tidyr
- **ggplot2**
- readr
- purrr
- tibble
- stringr
- lubridate
- forcats

- **mutate()** añadir variables como funciones de variables que existen.
- **select()** selección de variables.
- **filter()** selección de observaciones con base en características de una o más variables.
- **grouop_by()** agrupaciçon por valores únicos de una variable.
- **summarise()** reducción (o resumen) de una variable con una función.
- **arrange()** cordena las filas de una base de datos según los valores de una o varias variables.

A continuación, usted trabajará en la base de datos del **Estudio Longitudinal Social de Chile 2016-2022** [2]. Estas son las variables de interés:

- region
- comuna
- annio
- edad
- tendencia
- confianza_gobierno

Enlace de los datos:
https://drive.google.com/drive/u/2/folders/1AFBjiGl2Ll3dIrN1CV-pAESbTQ4ESFWw

Ejercicio I (1/2)

1. Importela basededatos y seleccione solo las variables de interés.

Ejercicio I (2/2)

1. Genere una tabla con los estadísticos descriptivos de la variable confianza_gobierno. En particular, reporte el promedio, desviación estándar, mediana, mínimo y máximo.
2. Reporte el promedio y desviación estándar de la variable confianza_gobierno, agrupando por la variable anio.
3. Reporte el promedio y desviación estándar de la variable confianza_gobierno, agrupando por la variable anio e tendencia.

- Gráficos por capas.

Material recomendado

- **ggplot gallery:** https://r-graph-gallery.com/
- ***secondary*ggplot extensions:** https://exts.ggplot2.tidyverse.org/gallery/

Comienzo con mi base de datos.

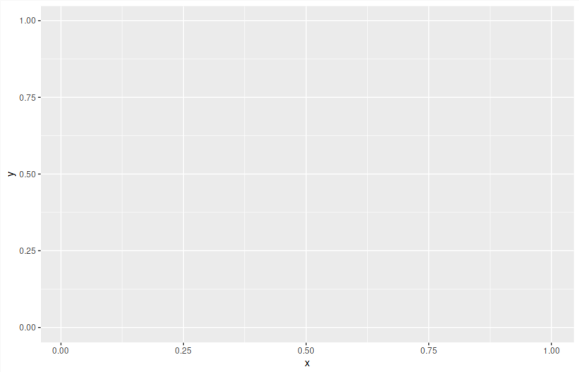
data

Solo si lo requiero, agrego algunas configuraciones a mis datos.

```
data %>%
    mutate(...) %>%
    select(...)
```

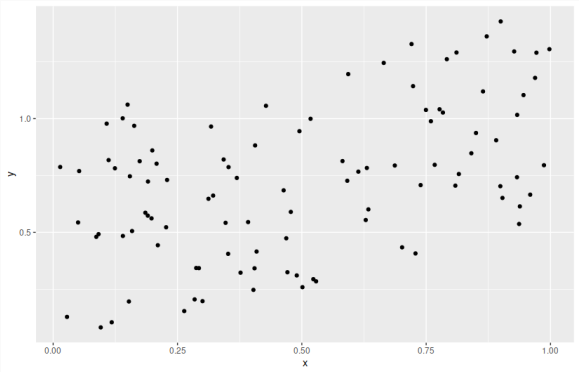
Con **aes()**, declaro cuáles son mis ejes x e y (*aesthetics*).

```
data %>%
    mutate(...) %>%
    select(...) %>%
    ggplot(aes(x, y))
```



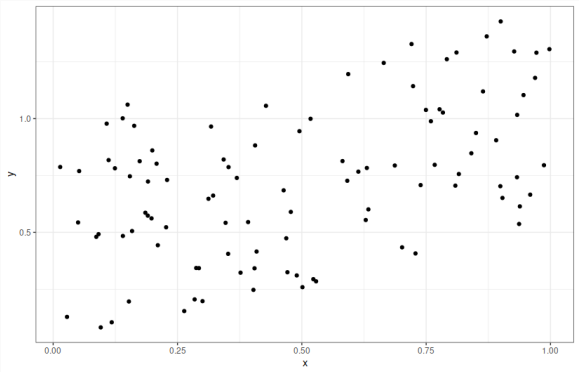
Con **geom** declaro la forma de mi gráfico (*geometry*).

```
data %>%
    mutate(...) %>%
    select(...) %>%
    ggplot(aes(x, y)) +
    geom_point()
```



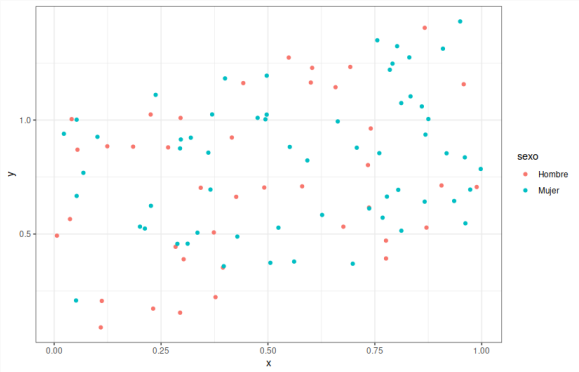
Puedo agregar un tema con **theme_**.

```
data %>%
    mutate(...) %>%
    select(...) %>%
    ggplot(aes(x, y)) +
    geom_point() +
    theme_bw()
```



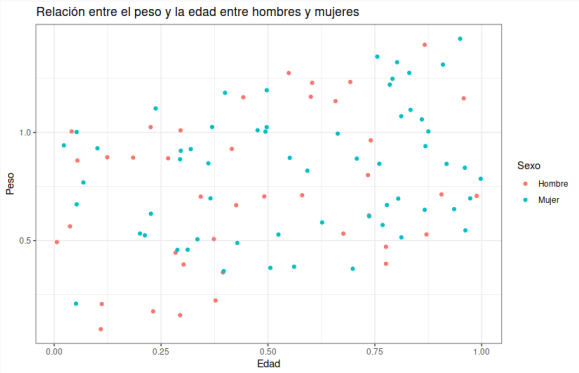
En **aesthetics**, puedo especificar intrucciones en relación a los datos.

```
data %>%
    mutate(...) %>%
    select(...) %>%
    ggplot(aes(x, y,
               color = sexo)) +
    geom_point() +
    theme_bw()
```



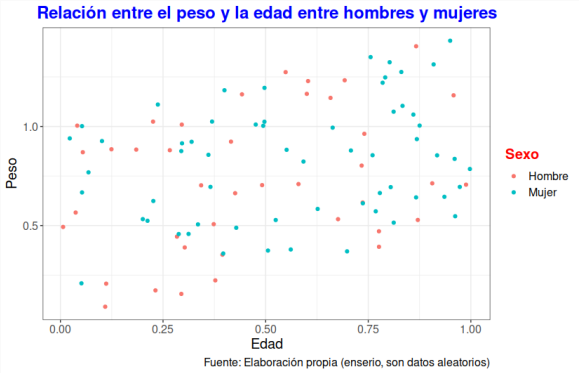
Con **labs** puedo modificar algunas etiquetas y texto de mi gráfico.

```
data %>%
    mutate(...) %>%
    select(...) %>%
    ggplot(aes(x, y,
               color = sexo)) +
    geom_point() +
    theme_bw() +
    labs(...)
```

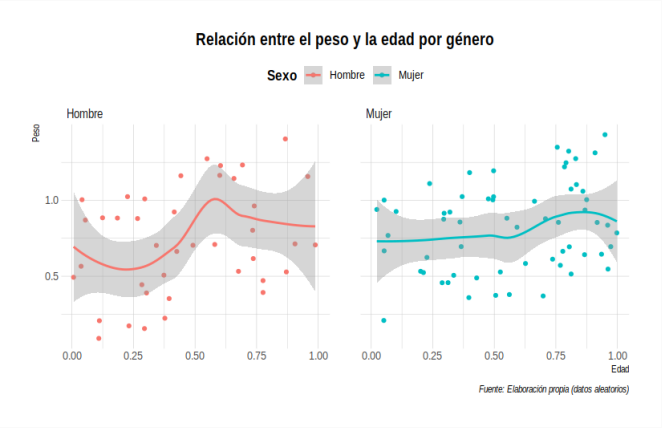


Con **theme**, puedo modificar tamaño y colores de los elementos de texto y fondo de mi gráfico.

```
data %>%
    mutate(...) %>%
    select(...) %>%
    ggplot(aes(x, y,
               color = sexo)) +
    geom_point() +
    theme_bw() +
    labs(...) +
    theme(...)
```



¡Y mucho más!


Relación entre el peso y la edad por género
Sexo
Hombre
Mujer
Hombre
Mujer
Peso
1.0
0.5
0.00
0.25
0.50
0.75
1.00
Edad
Fuente: Elaboración propia (datos aleatorios)

Ejercicio II (1/2)

Usando la base de datos generada, resuelva los siguientes ejercicios:

1. Elabore un histograma de la variable ‘edad‘.
2. Elabore un gráfico de barras con el porcentaje de encuestados ‘tendencia‘. ¿Es posible anteponer un gráfico de líneas? Luego, agrupe los datos según los años.
3. De gran interés es observar cómo ha variado la confianza en el gobierno a lo largo de los años. Elabore un gráfico de líneas del promedio de ‘confianza_gobierno‘ por ‘anio‘. Luego, genere la mejor visualización para agrupar estos promedios por ‘tendencia‘.

Ejercicio II (2/2)

Usando la base de datos generada, resuelva los siguientes ejercicios:

1. Aplique los siguientes ajustes al gráfico anterior:
    - Aplique el tema `bw` a su gráfico.
    - Cambie el tamaño de la letra.
    - Añada el título: "Confianza en el gobierno".
    - Añada una etiqueta correcta de los ejes x e y.
    - Añada una leyenda: *"Fuente: Elaboración propia a partir de los datos de ELSOC 2016-2022"*.

Ejercicios finales **Parte 2: Rmarkdown**

Rmarkdown es una herramienta que me permite **generar documentos** con código ejecutables para diversos fines:

- Informes automatizados
- Documentación de código
- Blogs
- Entre otros

Podemos generar todo tipo de documentos: **informes** (artículo) y **presentaciones**. También, podemos generarlos en diferentes formatos **pdf**, **html** y **word**.

udp UNIVERSIDAD DIEGO PORTALES

Texto: markdown

```
# Título de primer orden
## Titulo de segundo orden
### Título de segundo orden

- **Texto en negrita**
- *Texto en cursiva*

1. Lista enumerada (elemento 1).
2. Elemento 2.
3. Elemento 3.
```

Código: chunks (ctrl + alt + I)

````
```{r, ...}
print("Hola mundo!")
```
````

Opciones

- **echo: false/true** - mostrar el código o no.
- **eval: false/true** - evaluar el código o no.
- **message: false/true** - imprimir mensajes o no.
- **warning: false/true** - imprimir advertencias o no.

Ejercicio 5 (enunciado)

Despliegue el siguiente ejercicio en un markdown en formato .html.

- De algunas de las base de datos de permisos de circulación pagados y tramitados en la Municipalidad de Cochamó el 2016 (https://datos.gob.cl/dataset/permisoscirculacion2016cochamo) [1].
    1. Obtenga el porcentaje por columna, según corresponda al tipo de variable y nivel de medición.
    2. Obtenga la media y la mediana, según corresponda al tipo de variable y nivel de medición.
    3. EXTRA: Obtenga una tabla de 2 vías, según corresponda al tipo de variable y nivel de medición.

Ejercicio 6 (enunciado)

De los datos sobre interrupción voluntaria del embarazo, genere un gráfico de lineas en que el eje x sea el AÑO y las líneas sean la frecuencia. Cada línea debe representar cada causal (rojo= Causal 1: Peligro para la vida de la mujer; azul=Causal 2: Inviabilidad fetal de carácter letal; morado= Causal 3: Embarazo por violación), **utilizando tidyverse**.

Base de datos

```
#https://deis.minsal.cl/#tableros
#notese, que no escribimos con ñ por notación
data_df <- data.frame(
  ANIO = c(2018, 2018, 2018, 2019, 2019, 2019, 2020, 2020, 2020, 2021, 2021, 2021, 2022,
           2022, 2022, 2023, 2023, 2023),
  Frecuencia = c(262, 346, 124, 267, 414, 137, 160, 348, 154, 250, 442, 130, 254, 368,
                 209, 103, 162, 142),
  CAUSAL = c("Causal 1", "Causal 2", "Causal 3", "Causal 1", "Causal 2", "Causal 3",
             "Causal 1", "Causal 2", "Causal 3",
             "Causal 1", "Causal 2", "Causal 3", "Causal 1", "Causal 2", "Causal 3",
             "Causal 1", "Causal 2", "Causal 3")
)
```

Gráfico

```
#Ejemplo de un gráfico con el total, sin división en causales
data_df %>%
    group_by(ANIO) %>%
    summarise(total=sum(Frecuencia, na.rm=T)) %>%
    ggplot(aes(ANIO, total, group=1))+
    geom_point()+
    theme_minimal()
```

Referencias I

[1] M. de Cochamó.
Permisos circulación 2016.
Enlace.

[2] C. f. S. C. Reproducible Research and C. S. COES.
Estudio Longitudinal Social de Chile 2016-2022.
2023.