Laboratorio de Datos

Primer Cuatrimestre 2024

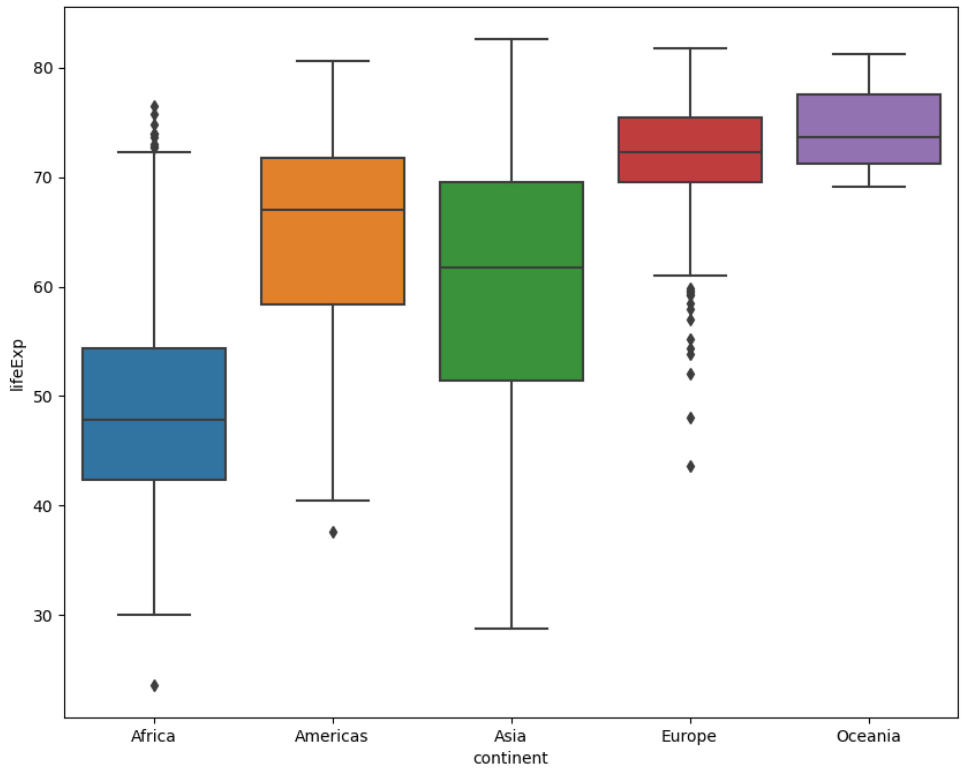
Práctica N° 3: Visualización de datos.

1. Tenés datos de una encuesta realizada en distintas provincias de Argentina y querés saber cuántas personas respondieron a la encuesta en cada provincia. ¿Hacés un gráfico de líneas, de dispersión (scatter), histograma o un gráfico de barras (bar plot)? Hacé a mano en tu cuaderno cómo esperás que se vea el gráfico.

2. Estás estudiando la relación entre altura y peso de las personas. Tenés un data-set que tiene como variables la edad, sexo y peso de cada persona. Si querés describir estas variables por separado, ¿qué gráfico harías para cada una? ¿y si querés visualizar la relación entre peso y altura? Hacé a mano en tu cuaderno cómo esperás que se vea el gráfico.

3. Hacé un gráfico de barras que muestre la cantidad de países hay en cada continente según los datos de gapminder (recordar el ejercicio 10 de la Práctica 2 para acceder a los datos de gapminder).

4. Querés investigar cómo varía la expectativa de vida entre los continentes. Para eso necesitás un gráfico como el siguiente:



Reproducí el gráfico de arriba reemplazando adecuadamente lo que falta en el siguiente código:

```
import seaborn as sns
sns.boxplot(gapminder, x=COMPLETAR, y=COMPLETAR, order=sorted(COMPLETAR
    ))
```

5. (a) Utilizando `seaborn.objects`, graficar la curva de la expectativa de vida en Argentina en función del año, completando el siguiente código. Sugerencia: recordar de la práctica anterior como filtrar datos de un dataset.

```
import seaborn.objects as so
(
    so.Plot(data = gapminder[???], x = "year", y = "???")
    .add(so.Line())
)
```

   (b) Realizar un nuevo gráfico donde puedan verse las curvas de la expectativa de vida de los países de América en función del año, una curva por cada país.

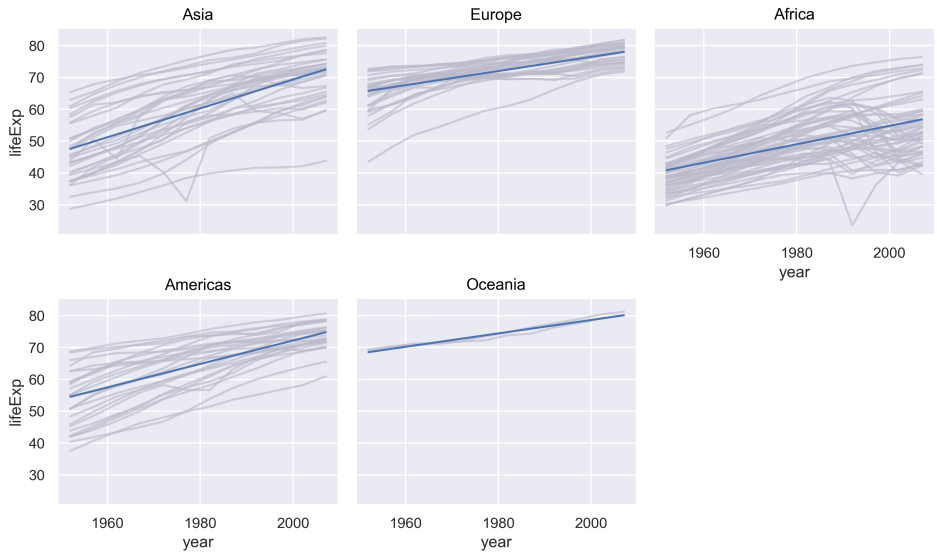
   Sugerencia: utilizar los parámetros `group = ???` o `color = ???`. ¿Cuál es la diferencia entre los dos?

   (c) Queremos agregar al gráfico del ítem anterior una curva de tendencia lineal utilizando el método `.add(so.Line(), so.PolyFit(1))`. ¿Cuál de las siguientes dos formas de agrupar los datos es la forma correcta? Explicar la diferencia entre los dos códigos.

```
# Codigo 1
(
    so.Plot(data = ???, x=???, y=???, group = ???)
    .add(so.Lines(color="#bbca"))
    .add(so.Line(), so.PolyFit(1))
)

# Codigo 2
(
    so.Plot(data = ???, x=???, y=???)
    .add(so.Lines(color="#bbca"), group = ???)
    .add(so.Line(), so.PolyFit(1))
)
```

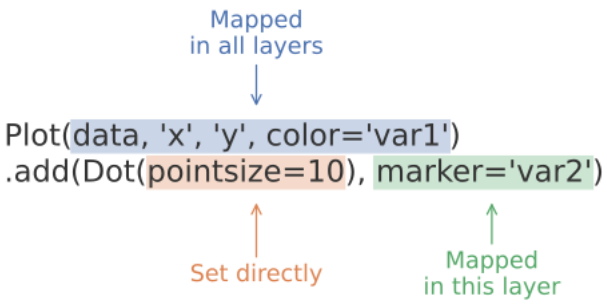
   (d) Realizar el siguiente gráfico, con las curvas de expectativa de vida agrupadas por continente. Sugerencias: ¿qué hace el método `facet()` de `seaborn.objects.Plot()`? ¿Y el parámetro `wrap = ???` de `facet()`?



6. (De acá en adelante, trabajar con el dataset `penguins` disponible en la biblioteca `seaborn`). ¿Cuántas filas y columnas hay en el dataset penguins?

```
import seaborn as sns
penguins = sns.load_dataset("penguins")
```

7. Como vimos en el Ejercicio 5c, si asignamos una codificación (o mapeo) al definir un `Plot()`, el mapeo se asigna en todas las capas de marcas (objetos mark). En cambio, si asignamos una codificación dentro del método `add()` de una marca, mapeo se realiza solo en esa capa. Por último, si asignamos un parámetro de la marca, el valor se asigna directamente (ver gráfico).



¿Qué resultado esperan para el siguiente gráfico? ¿Cuáles codificaciones se pasan de `Plot()` a `Dot()` y cuáles no pueden pasarse? ¿Cuáles codificaciones se establecen en `Dot()`? ¿Cuáles variables están asignadas directamente en `Dot()`? ¿De qué color van a pintarse los puntos?

```
(
    so.Plot(
        penguins, x="bill_length_mm", y="bill_depth_mm",
        edgewidth="body_mass_g", marker = "species",
        linestyle = "island", color = "species"
    )
    .add(so.Dot(color=".8"), edgecolor="sex")
)
```

8. (a) ¿Cuántos pingüinos hay en cada isla en la base de datos? Recordar los comandos `groupby()` y `size()` de la práctica anterior.

   (b) Realizar un grafico de barras con la cantidad de pingünos en cada isla, completando el siguiente código.

```
pinguinos_por_isla = penguins.???   # Usar el codigo del item
    anterior.
(
    so.Plot(x=pinguinos_por_isla.index, y=???)
    .add(so.Bar())
)
```

   (c) El gráfico que acabamos de hacer es un histograma categórico (usamos una variable categórica en el eje X). Podemos realizar el mismo gráfico usando la función `Hist()` para contar automáticamente las cantidades (sin definir una variable `pinguinos_por_isla`. Completar el siguiente código.

```
(
    so.Plot(data = penguins, ???)
    .add(so.Bar(), so.Hist())
)
```

   (d) ¿Por qué no especificamos ninguna variable y en el último gráfico?

   (e) Queremos ver en un gráfico cuántos pingüinos de cada especie hay en cada isla, ¿cómo podemos hacerlo? Si usan un gráfico de barras, pueden utilizar la función `dodge()` para hacer varias barras por categoría.

   (f) ¿Cómo podrían visualizar lo mismo usando `facet()`?

9. Realizar un histograma de la cantidad de pingüinos en función del tamaño del ala (variable `flipper_length_mm`). A partir del gráfico, estimar el valor mínimo, máximo, la media y la mediana. Verificar sus estimaciones utilizando los comandos apropiados.

10. (a) Hacer un `scatterplot` de `bill_depth_mm` (en el eje y) vs. `bill_length_mm` (en el eje x).

    (b) ¿Distinguen grupos distintos de puntos en el gráfico? ¿A qué puede deberse?

(c) Introducir alguna modificación en el gráfico anterior para verificar o refutar su conjetura del ítem anterior.

11. (a) Calcular distintos estadísticos de la variable `bill_depth_mm` (mínimo, máximo, media, ...).

    (b) Según lo observado en el ejercicio anterior, ¿esos valores varían según la especie? ¿Cómo podemos usar gráficos BoxPlot para para ver la relación entre `species` y `bill_depth_mm`?

12. (a) Rehacer el scatter plot del ejercicio 10, coloreando los puntos según el sexo. ¿Qué se observa?

    (b) Usando la función `facet()` separar el gráfico del item anterior en tres subgráficos, uno para cada especie.

13. (a) Rehacer el scatter plot del ejercicio 10, modificando el tamaño de los puntos según el peso de cada pingüino, utilizando el parámetro `pointsize="???"`. ¿Qué se observa?

    (b) En base a lo observado, ¿cuál es la especie con mayor peso? Verificarlo mediante alguna visualización.