Laboratorio de Datos

Primer Cuatrimestre 2024

Práctica N° 2: Estadística descriptiva

Estructuras de datos en Pandas

Series. Las series de Pandas son vectores similares a los arrays de NumPy, que podemos indexar usando etiquetas.

1. Crear la siguiente Series, observar qué devuelve array e index e interpretar.

```
import pandas as pd
obj = pd.Series([7,4,-5,3])
display(obj)
display(obj.array)
display(obj.index)  # Por default, los indices van de 0 a N-1.
```

2. Podemos asignar etiquetas (o índices) a cada valor de la serie.

```
obj2 = pd.Series([np.pi,0,-2,1.41], index = ["d", "b", "c", "a"])
display(obj2)
display(obj2.array)
display(obj2.index)
```

3. Al igual que con arrays de Numpy podemos acceder a los elementos por su posición, o podemos usar las etiquetas. Ejecutar los siguientes comandos.

```
obj2["a"]
obj2[3]
obj2[1:3]

obj3 = obj2[["a","b"]]
obj3
obj3.index

obj2[obj2>1]
```

4. Las operaciones que pueden aplicarse a numpy arrays pueden aplicarse también a series de Pandas, conservando los índices.

```
np.exp(obj2)
obj2 * 3
```

5. **Métodos de series.** Ejecutar los siguientes comandos e interpretar qué hace cada uno.

```
series1 = pd.Series(["a", "b", "c", "b", "a", "c"])
series1.isin(["b", "c"])
series1.valuecounts()
```

DataFrames. Un DataFrame es una representación de los datos en formato de tabla donde las columnas son vectores del mismo tamaño. Como cada columna es un vector, cada columna puede contener datos de un único tipo. Se pueden pensar como variables. Cada variable corresponde a una serie de Pandas, y todas las series de un mismo DataFrame están indexadas por los mismos índices.

6. Una forma de crear un DataFrame es utilizando un "diccionario". Todas las variables del diccionario deben ser vectores o listas de la misma longitud. Ejecutar el siguiente código.

```
data = {"nombre": ["Rodrigo", "Sergio", "Cristina", "Diana"], "altura":
    np.array([178, 172, 175, 168]), "peso": np.array([81.2, 76.1, 68.5,
    64.0])}
display(data)

pacientes = pd.DataFrame(data).set_index("nombre")
display(pacientes)
```

7. ¿Cuál es la clase del objeto `pacientes`? ¿Cuál es la clase de cada uno de los vectores columna? (para saber la clase de un objeto, utilizar el comando `type`, para saber el tipo de datos de un array de numpy, utilizar `np.dtype`)

8. Guardar en una variable nueva el vector columna `altura`. Pueden utilizar `pacientes["altura"]` o `pacientes.altura` (la primera opción es preferible, la segunda puede dar error si el nombre coincide con alguna función ya existente).

9. A diferencia de las matrices en Numpy, un DataFrame de Pandas es un conjunto de columnas, no de filas. Pensar cuál de los dos comandos será correcto antes de ejecutarlos.

```
pacientes["Rodrigo"].altura
pacientes["altura"].Rodrigo
```

10. **Gapminder.** A modo de ejemplo, vamos a explorar el dataset Gapminder que contiene datos poblacionales y de desarrollo humano de distintos países a lo largo del tiempo.

    Cargar la biblioteca `gapminder` utilizando

```
from gapminder import gapminder
```

    Si da error es posible que no esté instalado. En tal caso ejecuten primero

```
pip install gapminder
```

Esto crea un nuevo objeto `gapminder`. Pueden ver el contenido con el comando con algunos de estos comandos: `display(gapminder)`, `gapminder.info()`, `gapminder.head()`, `gapminder.tail()`.

11. ¿De qué clase es el objeto `gapminder`? ¿Qué variables tiene el dataset `gapminder` y de qué clase son? ¿Qué índices usa?

12. Explorar el tamaño del dataset gapminder usando la función `shape()`.

13. ¿De cuántos países hay datos? Ayuda: averiguar qué hacen los métodos `unique()` y `nunique()` aplicados a series.

14. Extraer la información de Argentina, Uruguay y Chile y guardarla en un nuevo DataFrame `gm_sur`. Sugerencia: recordar el método `isin()`.

    ¿Cuántas filas tiene? ¿Cuál es el primero y el último año para el cuál existen datos de Argentina en gapminder?

15. ¿Cómo está indexado el DataFrame `gm_sur`? Para acceder a una fila de un DataFrame, podemos usar los métodos `loc[]` y `iloc[]`. ¿Cómo se usan? ¿Cuál es la diferencia entre los dos comandos?

Archivos de datos

16. La biblioteca `Pandas` nos permite también trabajar con archivos de datos.

    (a) Leer el archivo `casos_coronavirus.csv`.

    (b) Graficar la curva de casos por día.

    (c) Graficar la curva de casos acumulados (utilizar la función `cum_sum` para calcularlos.

    (d) Definir y como el logaritmo de la cantidad de casos acumulados y graficar y en función de la cantidad de días transcurridos.

    Utilicen el siguiente código para leer el archivo y graficar.

```
df = pd.read_csv("casos_coronavirus.csv")    # dataFrame
df

df["confirmados_Nuevos"].plot()
```

Estadística descriptiva

17. Dar tres ejemplos de variables categóricas y numéricas.

18. En el dataset `gapminder`, una de las variables es el producto bruto per capita de los países (gdpPercap). ¿Es una variable categórica (nominal u ordinal) o numérica (discreta o continua)?

19. Supongamos que definimos una nueva variable que puede tomar los siguientes valores:

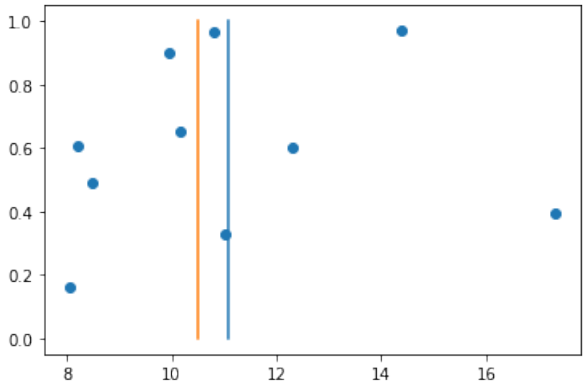
$$\text{nivelGDP} = \begin{cases} 0, & \text{si } \texttt{gdpPercap} < 1600. \\ 1, & \text{si } 1600 \leq \texttt{gdpPercap} < 6600. \\ 2, & \text{en otro caso.} \end{cases}$$

    ¿La nueva variable es categórica (nominal u ordinal) o numérica (discreta o continua)? ¿Cambia la respuesta si la variable toma valores “bajo”, “medio” y “alto” en lugar de 0, 1, 2?

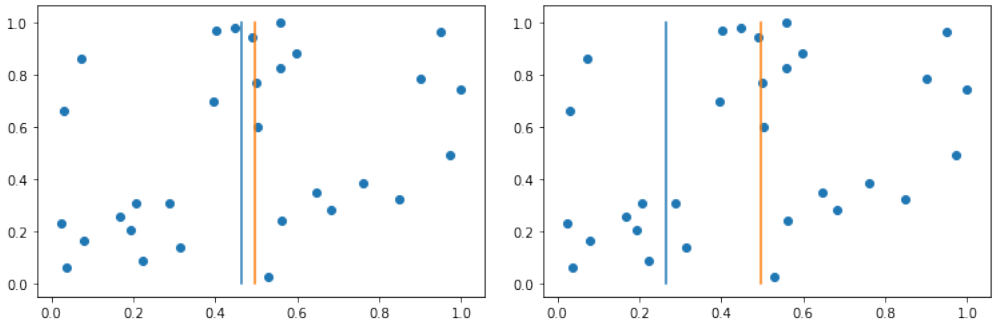
20. Filtrar el dataset de `gapminder` para el año 2007. Luego, para ese año, calcular la cantidad de países en cada continente. Explorar las funciones `groupBy`, `size`, `nunique`.

21. Con el mismo filtro que el ejercicio anterior (es decir, sólo para el año 2007), crear una variable I que valga 1 si `gdpPercap` es mayor que 2000 dólares y 0 si no lo es. Luego crear una tabla de 2 filas y 5 columnas que calcule la cantidad de países donde $I = 0$ o $I = 1$ en cada continente. Ayuda: si tenemos la información por columnas, podemos convertirla a una tabla con la función `unstack()`.

22. En el gráfico vemos 10 puntos. Consideramos la media y mediana de la coordenada x de esos puntos y graficamos dos rectas verticales $x =$ media y $x =$ mediana. ¿Cuál recta corresponde a la media y cuál a la mediana?



23. Repetimos el mismo procedimiento con otros 30 puntos. Graficamos una línea azul para la media y una línea naranja para la mediana. ¿Cuál gráfico es correcto?



24. Definir funciones que calculen la media y mediana de un vector de valores numéricos y la moda de un vector de valores categóricos. ¿Qué tiene que pasar para que existan dos modas?

25. Probar las funciones definidas con las variables numéricas y categóricas del dataset `gapminder` utilizando solo los datos del año 2007.

26. Graficar el producto bruto interno promedio en América en función del año.

27. Definir desvío estándar. ¿Por qué la diferencia en el numerador está elevada al cuadrado? Escribir una función de Python que calcule el desvío estándar. Comparar el resultado de usar la función `np.std()`.

28. Calcular el mínimo, el máximo y el desvío estandar de la expectativa de vida (`lifeExp`) entre países tomando sólo el dataset `gapminder` para el año 2007.