Instituto Politécnico Nacional

Centro de Investigación en Cómputo

Python en el Ámbito Científico

Profesor: Alan Badillo Salas

Tarea 4

— Problemas Científicos con Python

Hoja de Actividades y Ejercicios

Julio 12, 2025.

Actividad 7

El mapa de calor es una herramienta de visualización de datos potente que permite concentrar información entre los niveles o categorías de una serie de datos y contrastarlos contra los niveles o categorías de otra serie de datos. Para series continuas se puede hacer un proceso de discretización y para series discretas se pueden usar los valores tal cual. Sin embargo, la forma de construir un mapa de calor es en el fondo la forma en la que se generará una matriz de conteos, promedios o probabilidades.

Lo primero será identificar dos series de datos numéricas o categóricas (de valores continuos o discretos). Luego se discretizarán las series de datos en caso de que sean valores continuos, esto se puede hacer en un proceso similar a los *bins* o mediante los cuantiles (ordenar los datos y dividirlos en n grupos). Una vez realizado este proceso se deberían tener dos series de datos de valores discretos (dos series de grupos). Finalmente se contarán cuántos elementos hay en cada combinación de grupos (que pertenezcan a la categoría de la primera serie y a la categoría de la segunda serie). Con esto se podrá construir una matriz con los conteos de las categorías de la primera serie como filas y las categorías de la segunda serie como columnas, dejando los conteos asociados a cada fila y columna dentro de cada celda (en la fila y columna relacionada).

Para dibujar la matriz como mapa de calor, bastará elegir una paleta de colores, por ejemplo, rojo intenso, rosa suave, azul claro y azúl marino, para pintar cada celda de la matriz del color respectivo según se parezca más al máximo, o al mínimo, el rojo intenso podría asumir el valor máximo, el rosa suave el color cercano al máximo, el blanco el color medio entre el máximo y mínimo, el azul claro el valor debajo del valor medio y el azul marino el color más cercano al mínimo. Por ejemplo, si el máximo conteo es 80 y el mínimo es 40, el rojo intenso sería de 70 a 80, el rosa suave de 60 a 70, el azul cielo de 50 a 60 y el azul marino de 40 a 50. Y si vale exactamente 60 o muy cercano podría dejarse en color blanco para indicar que no es rojo ni azúl.

En esta actividad vamos a generar un mapa de calor dado que la matriz de conteos ya está dada, durante las próximas clases profundizaremos sobre la construcción de estas matrices de conteos o de estadísticos. Vamos a suponer que la matriz de conteos ya está dada y trata sobre diferentes tipos de personas y los tipos de productos que han comprado:

Tipo de Producto / Tipo Persona	Cerveza	Pañales	Leche	Cigarros	Pan
Mujer Embarazada	3	23	182	0	156
Hombre Embarazado*	25	45	173	23	162
Mujer con Bebé	6	314	279	2	216
Hombre con Bebé	60	218	196	33	145
Mujer Soltera	70	2	196	41	187
Hombre Soltero	425	1	125	367	165

* El hombre embarazado hace referencia a hombres cuya esposa o novia está embarazada y el se asume del mismo modo en apoyo al proceso de embarazo.

- Dibuja el mapa de calor *global* con los conteos de la tabla, usando el mínimo y máximo global.
  - Ubica el máximo global y el mínimo global de toda la matriz de conteos y determina el rango ($rango = maximo - minimo$).
  - Divide el rango entre 4 y obtén los puntos donde irá el color de la paleta formada azul marino, azul claro, blanco, rosa suave y rojo intenso. Por ejemplo, si el mínimo es 0 y el máximo es 425, entonces, los puntos serían 0, 106, 212, 319 y 425. Entonces, el azul marino irá de 0 a 106, el azul claro de 106 a 212, el rosa suave de 212 a 319 y el rojo intenso de 319 a 425.
  - Forma un mapa de calor poniendo las categoría del tipo de producto (Cerveza, Pañales, Leche, Cigarros y Pan) en el eje horizontal y los tipos de persona (Mujer Embarazada, Hombre Embarazado*, ...) en el eje vertica. Deja un cuadro suficientemente grande para que se rellene del color según el valor del conteo en esa categoría horizontal y vertical
  - Visualiza donde están los colores más rojos y los colores más azúles y qué significa
  - ¿Qué tipos de persona compran más cerveza?
  - ¿Qué tipos de persona compran más pañales?
  - ¿Qué tipos de persona compran menos cerveza?
  - ¿Qué tipos de persona compran menos pañales?
- Dibuja el mapa de calor *columna* con los conteos de la tabla, usando el mínimo y máximo de cada columna.
  - Para cada columna (tipo de producto), ubica el mínimo y el máximo.
  - Construye una paleta de colores basada en ese mínimo y máximo. Por ejemplo, para la columna de cerveza, el mínimo es 3 y el máximo es 425, entonces, la paleta tendría los puntos 3, 108.5, 214, 319.5 y 425. Las otras columnas tendrán su propia paleta de colores con

otros valores, por ejemplo, para la columna de leche la paleta sería 125, 163.5, 202, 240.5 y 279.

- Forma el mapa de calor usando la paleta de colores de cada columna, es decir, para la columna cerveza deberían poner en rojo al hombre soltero, mientras que para la columna pañales se debería poner en rojo intenso a mujer con bebé. (Ya no se usa la paleta global, sino por cada columna)
- ¿Qué tipos de persona quedan en tonos rojizos y qué tipos de persona quedan en azulados para cerveza y cigarros?
- ¿Qué tipos de persona quedan en tonos rojizos y qué tipos de persona quedan en azulados para pañales y leche?

- Dibuja el mapa de calor $fila$ con los conteos de la tabla, usando el mínimo y máximo de cada fila.
  - Para cada fila (tipo de persona), ubica el mínimo y el máximo.
  - Construye una paleta de colores basada en ese mínimo y máximo. Por ejemplo, para la fila de mujer embarazada, el mínimo es 0 y el máximo es 182, entonces, la paleta tendría los puntos 0, 45.5, 91, 136.5 y 182. Las otras filas tendrán su propia paleta de colores con otros valores, por ejemplo, para la fila de hombre embarazado la paleta sería 23, 60.5, 98, 135.5 y 173.
  - Forma el mapa de calor usando la paleta de colores de cada fila, es decir, para la fila mujer embarazada deberían poner en rojo a la leche y el pan, mientras que para la fila hombre soltero se debería poner en rojo intenso a cerveza y cigarros. (Ya no se usa la paleta global, sino por cada fila)
  - ¿Qué tipos de producto quedan en tonos rojizos y qué tipos de producto quedan en azulados para mujeres y hombres embarazados?
  - ¿Qué tipos de producto quedan en tonos rojizos y qué tipos de produto quedan en azulados para mujeres y hombres con bebés?
  - ¿Qué tipos de producto quedan en tonos rojizos y qué tipos de produto quedan en azulados para mujeres y hombres solteros?

Actividad 8

Responde las siguientes preguntas acerca de los problemas vistos en clase, indica la opción más adecuada:

1. ¿Qué es una serie de datos?

   - (a) Es una lista de números que representan una característica de la población, por ejemplo, las edades.
   - (b) Es una lista de datos numéricos o categóricos que hacen referencia a una misma característica observada sobre la muestra, por ejemplo, las edades.
   - (c) Es variable aleatoria que toma valores sobre la característica de una población es cuya media muestral es cercana a la media poblacional.

2. ¿Cuál es la mejor forma de describir en datos y visualmente una serie numérica?

   - (a) Mediante los estadísticos principales (mínimo y máximo y los cuartiles) y usando una gráfica de caja para observar la concentración de los datos sobre el eje de datos.
   - (b) Mediante los estadísticos principales (suma, promedio, desviación estándar, varanza) y usando una gráfica de histograma para mostrar la concentración de los datos sobre el eje de datos.
   - (c) Mediante los estadísticos principales (promedio y rango) y usando una gráfica de dispersión para mostrar los puntos sobre el eje y sus frecuencias.

3. ¿Cuál es la mejor forma de describir en datos y visualmente una serie categórica?

   - (a) Mediante los probabilidades de pertenecer a cada categoría y usando una gráfica de mapa de calor, para ver qué categorías tienen un color más intenso y cuales un color más suave.
   - (b) Mediante los totales acumulados de cada categoría y usando una gráfica de pastel, para observar las categorías más pesadas y las más pequeñas, sobre todo cuando hay muchas categorías.
   - (c) Mediante los conteos y proporciones en cada categoría y usando una gráfica de barras para ver la uniformidad de los datos en cada categoría.

Hoja 16 de Ejercicios

Carga el conjunto de datos llamado `student_habits_performace.csv`.

```
import pandas

students = pandas.read_csv("https://github.com/dragonnomada/" +
                           "ipn-cic-pycien-junio-2025/" +
                           "raw/refs/heads/main/datasets/" +
                           "student_habits_performance.csv")

students.head()
```

- ¿Cuántas columnas observas?
- ¿Cuál será la columna de principal interés?

Hoja 17 de Ejercicios

Analiza y visualiza la serie de datos de las edades de los estudiantes.

```
from matplotlib.pyplot import subplots

age = students["age"]

print(age.describe())

figure, axis = subplots(nrows=2,
                gridspec_kw={"height_ratios": [1,3]})

age.plot.box(ax=axis[0], vert=False)
age.plot.hist(ax=axis[1], density=True, bins=8)
age.plot.density(ax=axis[1])
```

- ¿Cuál es la media de edades?
- ¿Cuál es la mediana de edades?
- ¿Cuál es la edad mínima y máxima?
- Según la caja, ¿Hacia donde hay más edades, arriba o debajo de 20 años?
- ¿Se podría decir que la distribución de edades es uniforme?

Hoja 18 de Ejercicios

Contrasta las horas de estudio contra las horas viendo netflix y separa los datos por los que sacaron más de 70 en el examen y los que no sacaron más de 70 en su examen.

```
import seaborn

seaborn.jointplot(students,
    x="study_hours_per_day", y="netflix_hours",
    hue=students["exam_score"] >= 70)
```

- ¿Cómo se ven los puntos después antes de 4 horas de estudio?
- ¿Cómo se ven los puntos después después de 4 horas de estudio?
- ¿Qué proporción visual de alumnos sacaría más de 70 puntos en su examen si estudia menos de 2 horas?
- ¿Qué proporción visual de alumnos sacaría más de 70 puntos en su examen si estudia más de 6 horas?
- ¿Se puede confirmar que al ver más netflix reprueban más personas?
- Observando las distribuciones horizontales (horas de estudio), ¿Dónde está la máxima área de confusión?
- Observando las distribuciones verticales (horas viendo netflix), ¿Cuál es el porcentaje aproximado que sacaría más de 70 al ver menos de 2 horas de netflix?

Hoja 19 de Ejercicios

Para las columnas categóricas, verifica la creencia sobre si la puntación promedio del examen es mayor o similar en alguna de las categorías. Define una lista de las columnas categóricas y recorre cada columna. Luego agrupa los estudiantes mediante las categorías de esa columna y selecciona la puntuación de exam, finalmente aplica el promedio a la puntación del examen e imprime el reporte. Esto generará en cada impresión un reporte de la puntuación promedio del examen para cada categoría.

```
columnas = [
    "gender",
    "part_time_job",
    "diet_quality",
    "internet_quality",
    "exercise_frequency",
    "parental_education_level",
    "extracurricular_participation",
    "mental_health_rating"
]

for columna in columnas:
    print(students.groupby(columna)["exam_score"].mean())
    print("-" * 40)
```

- ¿Se puede sospechar que la calificación promedio de las mujeres es mayor a la de los hombres o es similar?
- ¿Se puede sospechar que los estudiantes que no tienen un trabajo de medio tiempo, tienen un mejor puntaje en su examen?
- ¿Se puede sospechar que la puntuación promedio en el examen de los estudiantes que tienen mejor alimentación (Fair) es mayor al de estudiantes que tienen peor alimentación (Poor)?
- ¿Se puede sospechar que los alumnos que tienen buen internet (Good) son peores que los que tienen mal internet (Poor)? ¿Cómo explicarías esto?
- ¿Se puede sospechar que la frecuencia de ejercicios aumenta la puntuación promedio del examen en los estudiantes? ¿Cuántos puntos de diferencia habría entre un estudiante que no hace ejercicio contra uno que hace ejercicio 6 días a la semana?
- ¿Se puede sospechar que los estudiantes con padres con licenciatura son mejores que los que tienen padres con maestría? ¿Cómo explicarías esto?
- ¿Se puede sospechar que los estudiantes con actividades extracurriculares son mejores que los que no hacen?
- ¿Se puede sospechar que la salud mental de los estudiantes influye en la puntuación promedio de sus exámenes? ¿Cuánta diferencia hay entre un estudiante con salud mental de 1 y uno con salud mental de 10?

Hoja 20 de Ejercicios

Genera un mapa de calor sobre los estudiantes comparando el trabajo de medio tiempo y las actividades extracurriculares, usando la puntuación del examen promedio como objetivo.

```
reporte = students.groupby([
    "part_time_job", "extracurricular_participation"
])["exam_score"].mean().unstack()

print(reporte)

seaborn.heatmap(reporte, cmap="coolwarm")
```

- ¿En qué combinación hay una mayor puntuación promedio del examen (más rojo)?
- ¿En qué combinación hay una menor puntuación promedio del examen (más azul)?

- ¿Cómo explicas que los estudiantes que trabajan de medio tiempo y hacen actividades extracurricules tengan peor puntuación promedio del examen que los estudiantes que no tienen trabajo de medio tiempo y si hacen activades extracurriculares?

Reto de la semana

En este reto tendrás que investigar cómo construir nuevas series de datos a partir de la combinación de otras series de datos existentes.

1. Crea una serie y que sea la suma de la horas de netflix más las horas en redes sociales menos las horas de estudio
2. Crea una serie x que sea la frecuencia de ejercicio entre 7, más las horas de sueño entre 10, más la calificación de salud mental entre 10
3. Crea una serie z que sean las puntuaciones del examen mayores o iguales a 70
4. Dibuja una gráfica *jointplot* usando $x = x$, $y = y$ y $hue = z$ de tipo *hist*
5. Dibuja una gráfica *scatterplot* usando $x = y$, $y = exam_{score}$ y $hue = exam_{score}$
6. Dibuja una gráfica *scatterplot* usando $x = y \cdot (3 - x)^3$, $y = exam_{score}$ y $hue = exam_{score}$

* Para las últimas gráficas tienes que usar `students["exam_score"]` como $exam_{score}$