

## Introducción a la Ciencia de Datos - 2C 2022

### Guía para la entrega N° 1

Les proponemos analizar por su cuenta el dataset **insurance** que van a poder encontrar en el [campus](#) y el drive de la materia. Este dataset brinda información sobre el precio del seguro médico de varias personas, junto con algunos atributos como:

edad: edad del beneficiario principal

sexo: género del contratante del seguro: mujer, hombre

bmi: índice de masa corporal, que proporciona una comprensión del cuerpo, los pesos que son relativamente altos o bajos en relación con la altura. Idealmente entre 18,5 a 24,9

niños: Número de hijos cubiertos por el seguro médico / Número de personas a cargo

fumador: Fumador sí o no

región: la zona de residencia del beneficiario en los EE.UU., noreste, sureste, suroeste, noroeste.

gastos: Gastos médicos individuales facturados por el seguro de enfermedad

### Análisis dataset **insurance**

1. ¿Cuántas variables tiene el dataset? ¿Cuántos datos?
2. ¿Qué variables se podrían usar para hacer un scatter plot? ¿Cómo se imaginan que va a resultar el gráfico en cada caso? Hagan el esquema en papel y lápiz.
3. Realizar alguno de los gráficos que pensaron en el punto 3. ¿Qué resultados obtuvieron? ¿Resultó como se habían imaginado? ¿Por qué?
  - a. ¿Cómo mejorarían el gráfico?
  - b. ¿Es necesario ajustar los límites de los ejes para poder apreciar mejor la relación?
  - c. ¿Hay algún/algunos puntos que sobresale o que les llame la atención?
4. Volver a realizar el punto 4 para otros pares de variables.
  - a. De todos los gráficos que fueron haciendo ¿Cuál transmite más información? ¿Identifican algún comportamiento llamativo? ¿Cuál es más claro?
5. Elegir uno de los scatter plots realizados en el punto anterior al que le agregaremos una “tercera dimensión” utilizando un *bubble chart* (ver Cápsula 1). Elegir esa variable de tal manera que se pueda aprovechar la nueva información que están sumando al gráfico.
6. Sumarle al gráfico anterior la información de alguna variable categórica en forma de color. Pensar cuál y por qué.
7. Entendamos si nuestra producción está lista:

- a. ¿Qué sería necesario agregar o cambiar para que el gráfico se pueda presentar a alguien que no conozca el dataset? ¿Podrían, por ejemplo, presentárselo a estudiantes de otras materias que estén estudiando en el pasillo, o a algún familiar o amigo que esté con ustedes mientras terminan el trabajo, sin explicarles previamente nada del dataset? Si no es así, ¿qué le cambiarían? Realicen esos cambios.
- b. ¡Pongamos a prueba el punto anterior! Muéstrenle el gráfico a alguna persona, puede ser física o virtualmente. Registren la experiencia: ¿qué preguntas y devoluciones recibieron? ¿Hay alguna forma de resolver esas inquietudes en el mismo gráfico? Si es así, ¡háganlo!

**Subir la producción realizada al campus de la materia en la solapa de “Entrega N° 1” junto con un breve párrafo explicando lo que se ve en el gráfico y la conclusión que obtienen a partir de este análisis.**