Introducción a la Ciencia de Datos - 2C 2022

Guía de Trabajos Prácticos Nº 1

Les proponemos comenzar a analizar dos set de datos ubicados en el Campus de la materia a partir de una serie de consignas asociadas a cada dataset.

Parte 1. Trabajo en el laboratorio. Análisis dataset árboles

En el Campus de la materia pueden encontrar varios datasets, en la carpeta `árboles`. Todos provienen de un gran dataset de arbolado lineal de la Ciudad de Buenos Aires que vamos a usar varias veces en la materia y también en Programación 1.

En este caso, elegimos los árboles de algunas especies determinadas que están en avenidas de la ciudad (no están todas, solo las que tienen más de 60 árboles de las especies seleccionadas).

Cada grupo tiene que elegir uno de los datasets y cargarlo en una planilla de Google o en DataWrapper **(ver cápsulas 0 y 1)**.

1. ¿Cuántas filas hay en el dataset que eligieron? ¿Cuántas columnas? ¿Qué representa cada fila (es decir cuáles son las unidades de este dataset? ¿Qué variables son categóricas? ¿Qué variables son numéricas? ¿Son enteros o punto flotante?
2. ¿Cuáles son las especies presentes en el dataset? **Nota**: están en la columna `nombre_científico`. Usar la función UNIQUE en otra pestaña.
3. ¿Cuáles son los valores posibles de las otras variables categóricas? ¿Al hacer esto, encuentran algún problema o error en este dataset?
4. Vamos a realizar un gráfico de dispersión (*scatter plot*) de la altura del árbol (eje vertical) vs. el diámetro a la altura del pecho (eje horizontal). Pero antes de eso, piensen qué tipo de relación esperan encontrar. Hagan un gráfico a mano alzada en un papel.
5. Ahora sí, realicen el gráfico y compárenlo con sus expectativas. ¿Se parece a lo que esperaban? Si la respuesta es no, ¿en qué difiere? ¿por qué puede ser?
6. ¿Qué información te brinda el gráfico? ¿Podrían dar alguna idea de la altura que tienen los árboles para un dado ancho del tronco?
7. Ahora vamos a agregar una variable categórica como el color de los puntos. Usen la especie de los árboles, y discutan en grupo si aparece alguna información nueva o si alguna de las conclusiones del punto anterior cambiaron.
8. Comparen su gráfico con el de otros grupos. Discutan juntos las diferencias y similitudes.
9. ¿Qué otras variables podrían usarse para colorear los puntos, en lugar del nombre científico de los árboles? Prueben repetir los puntos 7 y 8 con alguna de estas variables.

Parte 2. Análisis dataset `healthy_lifestyle_city_2021`

Se analizaron 44 ciudades de todo el mundo para descubrir dónde es más fácil llevar un estilo de vida sano y completo. Desde los niveles de obesidad hasta los índices de contaminación, cada ciudad ha recibido una puntuación en 10 parámetros de vida saludable.

Fuente y detalles del dataset: https://www.lenstore.co.uk/

1. ¿Cuántas variables tiene el dataset? ¿Cuántos datos?
2. A partir de los distintos tipos de datos que contiene cada una de las columnas, responder:
    a. Para las variables categóricas: explicar qué representa la variable y cuáles son las posibles categorías dentro de la misma.
    b. Para las variables numéricas: ¿son continuas o discretas? ¿Qué variables tienen formato *float*? ¿Es necesario en todos los casos?
3. ¿Qué variables se podrían usar para hacer un scatter plot? ¿Cómo se imaginan que va a resultar el gráfico en cada caso? Discutir en grupo. Hagan el esquema en papel y lápiz.
4. 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?
    c. ¿Hay algún punto que sobresalga o que les llame la atención?
5. 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?
6. 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.
7. Sumarle al gráfico anterior la información de alguna variable categórica en forma de color. Pensar cuál y por qué.
8. 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 amigx que esté con ustedes mientras terminan el trabajo, sin explicarles previamente nada del dataset? Si

no es así, ¿qué le cambiarían para que el gráfico comunique mejor de qué se trata? Realicen esos cambios.

b. ¡Pongamos a prueba el punto anterior! Muéstrenle el gráfico a alguna persona, 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!