FEB 25

SPRINT 1 - MEETING 8

Exploración de datos

Grabación

Presentación

El conocimiento está hecho de preguntas

> *"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." – John Tukey* **Estadístico.***

En las últimas bitácoras fuimos armándonos de conceptos y herramientas que nos permiten dar el primer paso que suele realizar un/a Data Scientist: el **análisis exploratorio de datos.** Si bien ya lo hemos mencionado y hemos visto algunas conceptos fundamentales, es momento de darle una mirada integradora.

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) lidia con la siguiente pregunta: ¿qué aspecto tienen nuestros datos?

Responderla es una etapa fundamental de nuestro flujo de trabajo como Data Scientists y utilizaremos las visualizaciones y la estadística predictiva para hacerlo. A veces un EDA se convierte en el único objetivo ya que constituye un entregable en sí mismo.

Si bien hay que dejar que "los datos nos guíen"... ¿qué preguntas y curiosidades nos despierta el dataset a medida que lo exploramos?.

Ciertos pasos y cuidados suelen ser comunes en cualquier análisis. Veámoslos.

Primeros pasos con un Dataset

Antes de continuar, te pedimos que pienses, en tu opinión y a partir de lo que vimos en encuentros y bitácoras anteriores, cuáles deberían ser los primeros pasos con un dataset. Animate y escribelos.

Luego, a medida que vayas leyendo, presta atención a si puedes recordar qué herramientas vimos (librerías, comandos de Python, etc.) que te pueden ayudar a realizar cada paso.

En primer lugar, es importante que respondas algunas preguntas:

- **¿Quién construyó este conjunto de datos, cuándo, por qué y cómo?** Es raro que un conjunto de datos contenga los datos de toda una **población,** sino que en general suele consistir en una **muestra**. Se espera que la muestra sea representativa de la población, aunque ese no siempre es el caso. Saber quién, cómo y cuándo construyó el dataset nos ayuda a pensar en qué sesgos puede tener y cuán importantes son. Además, nos indica a quién consultar si necesitamos saber más sobre nuestros datos. ¡Muchas veces quienes arman los conjuntos de datos están muy bien predispuestos a responder nuestras preguntas!
- ¿Cuál es el **tamaño** de nuestro conjunto de datos? ¿Cuántas **instancias** (registros) y **columnas** tiene?
- ¿Qué significa cada columna? ¿Cuáles tienen **valores** numéricos y cuáles categóricos? ¿Qué valores pueden tomar? ¿En qué **unidades** están las mediciones? ¿Hay valores **faltantes?**

Es probable que el primer conjunto de preguntas ("¿Quién construyó este conjunto de datos, cuándo, por qué y cómo) podamos responderlo recién después de haber explorado un poco nuestro dataset. Es importante tener presente que pueden tener sesgos y que las conclusiones que saquemos en un EDA son sobre nuestro dataset, no sobre la población que representan. Para afirmar que una conclusión a la que llegamos puede ser aplicada a la población, debemos estar seguros que el dataset es representativo de toda la población o que los sesgos fueron correctamente considerados y corregidos.

Una vez concluído este punto, se suele seguir con un resumen de estadísticas básicas de cada columna, calculando: cantidad de elementos, máximo, mínimo, media, desviación estándar, cuartiles, etc.

Seguimos con la parte más importante: las visualizaciones. Aquellas que no suelen faltar son:

- **Histogramas:** ¿Cuál es la forma general de la distribución de cada variable? ¿Existen valores atípicos? Si nuestro dataset tiene muchas variables (columnas), podemos elegir las que consideremos más importantes o las que nos generen más intriga.
- **Diagramas de dispersión de a pares:** ¿cómo se relacionan las variables entre sí? Al igual que con el histograma, dependerá de nuestra valoración.
- **Mapa de calor de correlaciones:** la correlación nos ayuda a ver qué variables están relacionadas entre sí. Si bien con una una matriz de coeficientes de correlación es suficiente, darle un toque visual con un mapa de calor hace que la información se interprete más fácilmente y mejor. Estar atentos/as a qué variables están correlacionadas entre sí nos ayudará a construir mejores modelos.

Otras prácticas útiles son:

- **Separar por clases:** podemos repetir los mismos análisis y gráficos, pero ahora agrupando por grupos de instancias que pertenezcan a una misma clase o compartan el valor de una variable categórica.
- **Buscar registros familiares o interpretables:** muchas veces, es muy útil familiarizarse con algunos registros. De esta forma, siempre tendrás algunos ejemplos en mente que te ayudarán a "llevar a tierra" algunas características de tu dataset. Además, puedes seguir su evolución a lo largo del análisis. Los registros generalmente están asociados con una persona, lugar o cosa de la que ya tenemos algún conocimiento, por lo que podemos ponerlo en contexto y evaluar la consistencia de los datos que tiene. Algunos registros son particularmente interesantes, como aquellos que toman valores máximos o mínimos de algún campo importante.
- **Submuestreo:** si bien nuestras computadoras nos permiten trabajar con grandes conjuntos de datos, puede ocurrir que esto no sea siempre conveniente. Puede ser que queramos hacer una iteración rápida de un análisis, haciendo algunas pruebas rápidas, hasta obtener una forma más definida de lo que realmente queremos hacer. O tal vez nuestro dataset es demasiado grande para nuestra comprensión y preferimos empezar por algo más pequeño. En cualquier caso, siempre podemos considerar submuestrear nuestro dataset (preferiblemente al azar) y trabajar sobre ese subconjunto.

Por último, siempre deja que el dataset te guíe en las preguntas que quieres responder. Estos pasos iniciales te ayudarán a salir de la página en blanco, y ya verás cómo automáticamente empezarán a surgir preguntas para responder: un buen análisis exploratorio de datos cuenta una historia.

Algunas buenas prácticas para graficar. *Como mencionamos, visualizar es una parte fundamental de este trabajo. Por eso, aprovechamos a dejar algunas recomendaciones útiles para mejorar nuestros gráficos.* *Nunca te olvides de:**

- *Ponerle etiquetas a los ejes, junto con sus unidades si corresponde. Por ejemplo, si el eje x representa edades, es importante aclarar si corresponde a años (para el general de la población) o días o semanas (muy importante si estamos trabajando con datos sobre bebés).*
- *Ajustar la escala de forma tal de aprovechar mejor el espacio de la figura. Si es necesario, también podemos cambiar su tamaño.*
- *Si nuestro gráfico tiene más de un curva o puntos de distinto color o trazo, es importante que en la leyenda del gráfico indiquemos qué representa cada una.*

Realizar gráficos, además del aspecto técnico, requiere también **prestar atención a su diseño.** Una buena forma de adquirir un criterio estético es mirando buenas y malas visualizaciones. Decidiendo en qué acertaron y en qué fallaron. La práctica, obviamente, es fundamental. Leer consejos sobre visualizaciones, también te ayudará. No nos cansaremos de repetirlo: visualizar datos y resultados correctamente es fundamental en el trabajo del/la Data Scientist: si logramos visualizaciones efectivas, nuestros gráficos hablarán por nosotros y contarán historias no solo poderosas sino verdaderas.

¡Aprende mirando!

Si bien ya vienes adquiriendo las herramientas para el Análisis Exploratorio de Datos, todavía no las pusimos en práctica todas juntas. En el siguiente video empezamos un EDA sobre un dataset que puedes descargar aquí. El objetivo es que veas cómo es un comienzo realista de un EDA, ya que muchas veces en la teoría parece más lineal y ordenado de lo que realmente es. Como el nombre lo indica, ¡es una etapa de exploración!

Capítulo 2: Análisis Exploratorio de Datos (EDA)

¡Prepárate para el próximo encuentro!

Profundiza
Te invitamos a conocer más sobre el tema de esta bitácora.

Comunidad
Sumérgete en la red donde los profesionales de DS interactúan.

Lista de recursos

- Profundiza
- Comunidad