Guía rápida para valores ausentes

Cuándo completar valores ausentes... y cuándo no

Algunas veces te encontrarás con reglas estrictas, como "Si falta más del 20% de una variable, borra la variable por completo". Pero dudamos en dar reglas estrictas, ya que mucho depende del contexto del problema. Como analista de datos, parte de tu trabajo es tener en cuenta los matices al momento de tomar decisiones. Dicho esto, he aquí algunas directrices:

1. **Nunca completes la principal variable de interés, y nunca uses la principal variable de interés para completar valores ausentes.** Con frecuencia, tu objetivo final es entender las relaciones entre una variable principal y otras variables. Este es el objetivo, no un paso de preprocesamiento de datos.
2. **Siempre lleva un registro (documenta) sobre cuándo, dónde, por qué y cómo se completaron los valores ausentes.** Siempre ten una razón justificable para cualquier cosa que completes. Si se te pregunta cómo, debes tener una respuesta.
3. **Los valores completados no deberían tener un impacto significativo en tu análisis.** En caso de duda, ejecuta tu análisis dos veces, una con los valores completados y otra con los valores ausentes eliminados. Si los resultados son sustancialmente diferentes, los valores completados están impulsando el cambio. Y no quieres hacer eso. Este es un ejemplo de análisis de sensibilidad.

Antes de hablar sobre cómo remplazar valores categóricos ausentes, toma una postura poderosa y repite después de mí: "No permitiré que los valores completados cambien drásticamente los resultados de mi análisis".

Ocuparse de los valores ausentes

Así que has determinado que un conjunto de datos tiene valores ausentes. ¿Qué haces?

1. Reportar el problema y averiguar si hay una manera de obtener los datos completos. Si no lo hay, continúa con el paso 2.

2. Determina cuántos valores ausentes hay: llama el método `value_counts()` y `print()` .

```
print(file_name['column_name'].value_counts())
```

3. Determina qué tan importante es esa ausencia para el conjunto de datos. ¿Qué proporción de datos representan? En la mayoría de los casos, si no es mucho (digamos, 5-10%, dependiendo de la situación), puedes borrarlos.

4. Comprueba qué tan importante es su ausencia para su categoría o columna: llama a los métodos `isnull()` y `count()` y muéstralo.

```
print(file_name[file_name['row_name'].isnull()].count())
```

5. Determina si los valores ausentes pertenecen a las variables categóricas o cuantitativas.

6. Si son **categóricas:**

- Determina si los valores ausentes muestran o no un patrón; es decir, si su apariencia en el conjunto de datos es o no aleatoria. Si no se puede detectar una correlación con otros valores en las filas en las que aparecen (por ejemplo, en el caso de los encuestados menores de 21 años, una pregunta sobre el alcohol no

tiene respuesta), entonces probablemente sean aleatorios. Existen tres tipos de valores ausentes:

  - Ausencia completamente al azar (MCAR, por sus siglas en inglés).
  - Ausencia al azar (MAR, por sus siglas en inglés).
  - Ausencia no por azar (MNAR, por sus siglas en inglés).

- Dependiendo del patrón, decide cómo manejarlas:
  - Si los valores son MCAR o MAR, no hay un patrón, así que puedes remplazarlos con valores predeterminados: un string vacío o una palabra en particular. Utiliza el método `loc[]` y la indexación booleana. El método `fillna()` también podría funcionar, pero no en todos los casos.
  - Con valores MNAR, hay un patrón. Este es un caso más complejo, y no nos sumergiremos en sus complejidades en este capítulo.

7. Si son **cuantitativas:**
   - Determina si tus datos tienen valores atípicos significativos.
   - Si no hay valores atípicos significativos, calcula la media de tus datos: aplica el método `mean()` a la columna o al conjunto de datos completo.
   - Si tus datos tienen valores atípicos significativos, calcula la mediana de tus datos: aplica el método `median()` a la columna o al conjunto de datos completo.
   - Reemplaza los valores ausentes con la media o la mediana utilizando el método `fillna()`.