Guía rápida sobre valores ausentes

Cuándo rellenar los valores ausentes y cuándo no

A veces te encontrarás con reglas inflexibles, como "Si falta más del 20% de una variable, elimínala por completo". Pero dudamos en dar reglas inflexibles porque depende mucho del contexto del problema. Como analista de datos, parte de tu trabajo consiste en tener en cuenta los matices a la hora de tomar decisiones. Dicho esto, he aquí algunas directrices:

1. **Nunca rellenes la variable principal de interés, ni nunca utilices la variable principal de interés para rellenar los valores ausentes.** A menudo tu objetivo final es comprender las relaciones entre una variable principal y otras variables. Este es el objetivo, no un paso de preprocesamiento de datos.
2. **Siempre documenta cuándo, dónde, por qué y cómo se rellenaron los valores ausentes.** Siempre tienes que tener una razón justificable para cualquier relleno que hagas. Si te preguntan por qué, debes tener una respuesta.
3. **Los valores rellenados no deberían tener un impacto significativo en tu análisis.** En caso de duda, ejecuta tu análisis dos veces, una con los valores rellenados y otra con los valores ausentes eliminados. Si los resultados son significativamente diferentes, los valores rellenados causan el cambio. Y no quieres que pase esto. Este es un ejemplo de análisis de sensibilidad.

Antes de hablar sobre cómo reemplazar los valores ausentes categóricos, adopta una pose de poder y repite conmigo: "No permitiré que los valores rellenados cambien drásticamente los resultados de mi análisis".

Tratar con valores ausentes

Pues, has determinado que un conjunto de datos tiene valores ausentes. ¿Qué vas a hacer?

1. Informa sobre el problema y averigua si hay alguna forma de obtener los datos completos. Si no la hay, procede al paso 2.

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

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

3. Determina qué tan significativa es su ausencia para el conjunto de datos. ¿Qué porcentaje de los datos representan? En la mayoría de los casos, si no es mucho (digamos, 5-10%, según la situación), puedes eliminarlos.

4. Verifica qué tan significativa es su ausencia para su categoría o columna: llama a los métodos `isnull()` y `count()` e imprime.

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

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

6. Si son **categóricas**:

- Determina si los valores ausentes presentan un patrón, es decir, si su aparición en el conjunto de datos es aleatoria o no. 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:

  - Ausentes completamente al azar (MCAR)
  - Ausentes al azar (MAR)
  - Ausentes no al azar (MNAR)

- Dependiendo del patrón, decide cómo manejarlos:
  - Si los valores ausentes son MCAR o MAR, no hay ningún patrón, por lo que puedes sustituirlos por valores por defecto, es decir, una cadena vacía o una palabra concreta. Utiliza el método `loc[]` y la indexación booleana. El método `fillna()` también puede funcionar, pero no en todos los casos.
  - En el caso de valores MNAR, hay un patrón. Este es el caso más complejo y no nos sumergiremos en detalle en este capítulo.

7. Si son **cuantitativas:**
   - Determina si los 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()`.