Datos Ausentes Método de sustitución	Condiciones	Ventajas	Desventajas	Comentarios
Sustitución por la Media / Mediana	Los valores ausentes faltan de forma completamente aleatoria	• Fácil de implementar • Rápida forma de construir conjuntos de datos	• Distorsiona la varianza original • Distorsiona la covarianza / correlación con otras variables en el conjunto de datos	Sustitución por la Media / Mediana + adición de una variable para indicar 'ausencia' es bastante común en las competencias de ciencia de datos. La media debe ser usada cuando la variable está normalmente distribuida y la mediana para el resto de los casos Sin embargo, la mayoría de las personas sustituyen por la media independientemente de la distribución de la variable
Sustitución por una muestra aleatoria	Los valores ausentes faltan de forma completamente aleatoria	• Fácil de implementar • Rápida forma de construir conjuntos de datos • Preserva la varianza de la variable original	• Aleatoriedad • Distorsiona la covarianza / correlación con otras variables en el conjunto de datos	La sustitución por una muestra aleatoria consiste en tomar una muestra aleatoria de la variable donde los valores están disponibles y usarlos para imputar los valores faltantes. No es tan frecuente en las competencias de datos, pero es usada en la industria. Se necesita controlar la aleatoriedad cuando se están evaluando 'scoring' nuevos clientes, ya que clientes con las mismas condiciones deben recibir el mismo tratamiento.
Añadir indicador de ausencia	Datos ausentes tienen poder predictivo	• Fácil de implementar • Captura la importancia de la 'ausencia' de los datos si existe alguna.	• Incrementa el espacio/número de variables • Puede generar múltiples indicadores de ausencia que son similares o altamente correlacionados.	Imputación por la Media / Mediana / Moda + adición de una variable para indicar 'ausencia' es bastante común en las competencias de ciencia de datos y en las organizaciones
Imputación por valores al final de la distribución	Los valores no están ausentes de forma aleatoria	• Fácil de implementar • Captura la importancia de la 'ausencia' de los datos si existe alguna	• Distorsiona la distribución original de la variable • Si la 'ausencia' no es importante, puede enmascarar el poder predictivo de la variable original. • Si el número de datos ausentes/NA es alto, enmascara los valores extremos de la distribución • Si el número de datos ausentes/NA es bajo, los valores ausentes reemplazados pueden ser considerados outliers. Estos a su vez pueden ser pre-procesados en un paso más adelante de la ingeniería de variables	Usado en compañías, que no quieren atribuir a los valore faltantes las ocurrencias más comunes de la variable (media / mediana). El razonamiento es que si el valor faltante está ausente por una razón en particular, los NA no deben ser reemplazados por la media, ya que los hace parecer como la mayoría de las observaciones. Por el contrario, en este método, los NA están marcados como 'diferentes' ya que se les asigna un valor al final de la distribución, donde las observaciones están poco representadas en la población,.
Imputación por un valor arbitrario	Los valores no están ausentes de forma aleatoria	• Fácil de implementar • Captura la importancia de la 'ausencia' de los datos si existe alguna	• Distorsiona la distribución original de la variable • Si la 'ausencia' no es importante, puede enmascarar el poder predictivo de la variable original. • Difícil de decidir cuál es el valor arbitrario que debemos usar. • Si el valor está fuera de la distribución original puede enmascarar o crear outliers	Cuando las variables son capturadas por organizaciones externas, como agencias crediticias, ellos colocan números arbitrarios en los datos ausentes, para indicar la 'ausencia' de los datos. No es una práctica común en las competencias de datos, pero sí es común en organizaciones que generan datos. Números arbitrarios comunes son 9999, -9999.
Imputación por la categoría más frecuente	Valores están ausentes de forma aleatoria	• Fácil de implementar • Rápida forma de construir conjuntos de datos	• Distorsiona la relación de la etiqueta más frecuente con las otras variables en el conjunto de datos • Puede llevar a una sobre-representación de la etiqueta más frecuente si hay un número alto de valores ausentes	Este es el equivalente a la imputación por la moda, en este caso, usado solamente para variables categóricas. (Sustitución por la moda no es normalmente usado por las variables numéricas).
En variables categóricas: sustitución NA como una categoría adicional	Ninguna	• Fácil de implementar • Captura la importancia de la 'ausencia' de los datos si existe alguna	• Si el número NA es bajo, crea una categoría adicional que puede conllevar a una etiqueta adicional para valores raros o poco comunes	Método popular, ya que trata los valores ausentes como una categoría independiente, sin imponer condiciones en su significado.