Manejo de Datos Faltantes y Outliers - Día 5

Change Status

pending 40 min

Learning Objectives

1. Identificar y cuantificar datos faltantes en datasets usando técnicas sistemáticas
2. Aplicar estrategias apropiadas de imputación según el contexto y tipo de datos
3. Detectar outliers usando métodos estadísticos y visuales

Actividades y Aprendizajes

Aprende todo sobre funciones y módulos en Python con ejemplos prácticos.

Task 1: Detección y Análisis de Datos Faltantes (10 minutos)

Los datos faltantes representan uno de los **desafíos más comunes y peligrosos** en análisis de datos. No son meramente "espacios vacíos", sino indicadores de problemas en la recolección, procesamiento, o calidad de datos que pueden sesgar análisis completos si no se manejan correctamente.

Patrones de Datos Faltantes

Completamente al azar (MCAR): La ausencia de datos no depende de otras variables observadas o no observadas. Ejemplo: errores aleatorios en la medición.

Al azar (MAR): La probabilidad de missing data depende de otras variables observadas. Ejemplo: mujeres menos propensas a reportar ingresos altos.

No al azar (MNAR): La ausencia depende del valor faltante mismo. Ejemplo: personas con ingresos muy altos evitan reportarlos.

Herramientas de Detección en Pandas

isnull() y notnull(): `df.isnull().sum()` cuenta valores faltantes por columna.

info(): `df.info()` muestra tipos de datos y conteo de non-null.

Visualización: Mapas de calor de missing data ayudan a identificar patrones.

Porcentajes: `df.isnull().mean() * 100` muestra porcentaje de datos faltantes.

Task 2: Estrategias de Imputación (10 minutos)

La imputación no es un proceso mecánico, sino una **decisión analítica** que debe considerar el contexto de los datos, el impacto en el análisis, y las características del dataset.

Imputación Simple

Media/Moda/Mediana: `df['columna'].fillna(df['columna'].mean())` - apropiado para datos numéricos sin outliers extremos.

Valor constante: `df.fillna(0)` - útil cuando el cero tiene significado (ej: conteos de eventos).

Forward/Backward fill: `df.fillna(method='ffill')` - apropiado para series temporales donde valores previos son predictores razonables.

Imputación Avanzada

K-Nearest Neighbors: Imputa basado en registros similares en el dataset.

Modelos predictivos: Usa machine learning para predecir valores faltantes basado en otras variables.

Múltiples imputaciones: Crea varias versiones imputadas y combina resultados para manejar incertidumbre.

Consideraciones Éticas y Analíticas

No imputar indiscriminadamente: A veces es mejor analizar datos faltantes como un fenómeno en sí mismo.

Documentar decisiones: Registrar qué método se usó y por qué para reproducibilidad.

Sensibilidad al análisis: Diferentes técnicas de imputación pueden llevar a conclusiones diferentes.

Task 3: Detección y Manejo de Outliers (10 minutos)

Los outliers son valores que **se desvían significativamente** del patrón general de los datos, y su manejo requiere tanto técnicas estadísticas como juicio de dominio.

Métodos Estadísticos de Detección

Z-score: Valores con |z| > 3 se consideran outliers (más de 3 desviaciones estándar).

IQR Method: Valores por debajo de Q1 - 1.5*IQR o por encima de Q3 + 1.5*IQR.

Modified Z-score: Más robusto para datasets pequeños o con distribuciones no normales.

Estrategias de Manejo

Eliminación: `df = df[abs(zscore(df['columna'])) < 3]` - apropiado cuando outliers son errores claros.

Transformación: Aplicar logaritmos o otras transformaciones para reducir impacto de outliers.

Winsorización: Reemplazar outliers extremos con percentiles (ej: 95% para valores muy altos).

Análisis separado: Analizar outliers como grupo separado cuando representan segmentos válidos.

Consideraciones de Dominio

Outliers vs Anomalías: En finanzas, un valor extremo podría ser una anomalía fraudulenta. En medicina, podría ser un caso crítico.

Contexto temporal: Un valor normal en un contexto podría ser outlier en otro.

Distribución de cola larga: En algunos dominios (ingresos, tamaños de ciudades), la distribución natural tiene colas largas.