Valores Atípicos

Definición

Un outlier (o valor atípico) es un dato que se encuentra significativamente alejado del resto del conjunto de datos, ya sea en términos de su valor numérico o de su categoría. En datos cuantitativos, los outliers son valores extremadamente altos o bajos en comparación con el resto. En datos cualitativos o categóricos, los outliers son valores que son inusuales, inesperados o raramente observados dentro de su contexto.

Estos pueden surgir por varias razones:

- **Errores de medición o recolección de datos**: Los errores humanos o técnicos durante la recolección o entrada de datos pueden generar valores que no reflejan la realidad.
- **Variabilidad natural en los datos**: Algunos procesos o fenómenos tienen una variabilidad inherente que puede producir valores atípicos ocasionales.
- **Errores experimentales**: Problemas en el diseño del experimento o en la implementación de protocolos pueden resultar en datos inusuales.
- **Distribuciones no normales**: En algunos casos, los datos pueden no seguir una distribución normal, y los valores que parecen atípicos en una distribución normal pueden no ser atípicos en otra distribución.
- **Fraude o manipulación de datos**: En algunos contextos, como auditorías o análisis de fraudes, los outliers pueden ser indicadores de manipulación intencionada de los datos.
- **Eventos raros o excepcionales**: Ocasionalmente, eventos poco comunes pero posibles pueden generar valores atípicos.

Tipos de Valores Atípicos

- **Outliers Univariados**: Valores atípicos que se identifican dentro de una sola variable.
- **Outliers Multivariados**: Valores atípicos que se identifican cuando se consideran varias variables simultáneamente. Un dato puede no parecer un outlier en ninguna variable individualmente, pero en combinación con otras puede ser anómalo.
- **Outliers Categóricos**: Valores categóricos que son inusuales o inesperados dentro del contexto de la/s variable/s.
- **Outliers Contextuales**: Valores que son atípicos en un contexto específico, aunque pueden no serlo en un contexto general.
- **Outliers Colectivos**: Un grupo de puntos de datos que, en conjunto, se desvían del comportamiento esperado del conjunto de datos.

Impacto de Outliers en un Proyecto de Ciencia de Datos

Los valores atípicos pueden tener gran impacto en diversas etapas del proyecto de minería de datos. Este impacto puede variar dependiendo del tipo de outlier, el tipo de modelo utilizado y la naturaleza de los datos. A continuación, se detalla el impacto de los outliers en un proyecto de machine learning según diferentes aspectos:

Distorsión de la Caracterización de Distribuciones

- **Outliers Univariados:** Pueden sesgar medidas descriptivas univariantes y gráficos, dando una falsa impresión de la distribución de los datos de una variable concreta.
- **Outliers Multivariados:** Pueden influir en las medidas de relación, medidas de forma multivariante y demás estadísticas descriptivas cuando se consideran múltiples variables simultáneamente.
- **Outliers Categóricos:** Pueden distorsionar la frecuencia y proporción de las categorías.

Ingeniería de Características

- **Outliers Univariados:** Pueden afectar la creación de nuevas características a partir de transformaciones matemáticas, ya que los valores extremos pueden distorsionar los resultados.
- **Outliers Multivariados:** Pueden complicar la creación de características derivadas de múltiples variables, afectando la identificación de patrones y relaciones.
- **Outliers Categóricos:** La presencia de categorías raras puede complicar la creación de características basadas en combinaciones o interacciones de categorías.

Selección de Características

- **Outliers Univariados y Multivariados:** Los outliers pueden influir en las métricas utilizadas para la selección de características, como la correlación y la importancia de características, llevando a la inclusión o exclusión incorrecta de variables.
- **Outliers Categóricos:** Pueden afectar las técnicas de selección de características categóricas, haciendo que categorías raras parezcan más relevantes de lo que realmente son.

Problemas en el Escalado

- **Outliers Univariados:** Los valores extremos pueden afectar la normalización y el escalado, haciendo que estos sean menos efectivos.
- **Outliers Multivariados:** Pueden complicar la normalización y el escalado en múltiples dimensiones.
- **Outliers Categóricos:** Si se usan técnicas como one-hot encoding, los outliers categóricos pueden aumentar innecesariamente la dimensionalidad.

Desempeño del Modelo

- **Outliers Univariados:** En modelos sensibles a outliers, los outliers pueden influir de manera desproporcionada en los coeficientes o parámetros del modelo.
- **Outliers Multivariados:** Pueden afectar modelos que consideren interacciones entre múltiples variables, como en modelos de clustering.
- **Outliers Colectivos:** Pueden alterar los patrones y relaciones que el modelo debe aprender, especialmente en algoritmos de series temporales o en análisis de datos secuenciales.

Convergencia de Algoritmos

- **Outliers Univariados y Multivariados:** Los algoritmos basados en gradientes pueden tener dificultades para converger si hay outliers, ya que los gradientes pueden ser excesivamente grandes.

Sobreajuste

- **Outliers Univariados y Multivariados:** Los modelos pueden sobreajustarse a los outliers, especialmente en conjuntos de datos pequeños, donde los outliers pueden tener una influencia significativa en el entrenamiento.
- **Outliers Colectivos:** Pueden causar que el modelo capture patrones no representativos de la población general.

Métricas de Evaluación

- **Outliers Univariados:** Las métricas de evaluación pueden ser altamente sensibles a los outliers.
- **Outliers Multivariados:** Pueden afectar las métricas que consideren múltiples dimensiones de error.
- **Outliers Colectivos:** Pueden influir en la evaluación de series temporales y análisis secuenciales.

Validación Cruzada

- **Outliers Univariados y Multivariados:** Pueden causar variabilidad en los resultados de la validación cruzada, llevando a una estimación inestable del desempeño del modelo.

Confusiones en la Interpretación

- **Outliers Univariados:** Pueden llevar a interpretaciones incorrectas de los patrones en los datos.
- **Outliers Multivariados:** Pueden complicar la interpretación de las relaciones entre variables en modelos multivariados.
- **Outliers Categóricos:** Pueden afectar la claridad de las interpretaciones categóricas, introduciendo categorías raras que confunden el análisis.

Mantenimiento del Modelo

- **Todos los tipos de Outliers:** La necesidad de manejar outliers en datos en tiempo real puede complicar el mantenimiento y la actualización del modelo, requiriendo técnicas adicionales para la detección y el tratamiento de outliers en producción.

Identificación de Valores Atípicos Univariantes

Método Z-Score

Dada una variable cuantitativa $X = \{x_1, x_2, \cdots, x_n\}$ el Z-score es una medida estadística que describe la relación de los valores x_i de la variable con respecto a su media $\bar{x}$. Se calcula restando la media de la variable y dividiendo por su desviación estándar s_X cada valor x_i.

Cálculo del Z-Score

El Z-Score se calcula como:

$$Z = \frac{X - \bar{x}}{s_X}$$

$$z_i = \frac{x_i - \bar{x}}{s_X}$$

Donde:

- X son todos los valores de la variable de estudio.
- x_i es un valor concreto de X.
- $\bar{x}$ es la media de la variable X.
- s_X es la desviación estandar de la variable X

Identificación de Outliers usando el Z-Score

- Se define un umbral u arbitrario, usualmente 3, el cual es interpretado como las desviaciones típicas con respecto a la media.
- Se comparan los valores z_i contra el umbral definido.
- Los valores que cumplan $|z_i| > u$ son candidatos a ser considerados valores atípicos.

Litaciones del método Z-Score

- **Suposición de Normalidad**: El método Z-Score aplicado a una variable X supone que la variable X sigue aproximadamente una distribución normal.
- **Umbral Arbitrario**: El umbral típico $u = 3$ puede no ser adecuado para todas las variables o contextos.
- **Sensibilidad a la media y desviación estándar**: En variables con pocas observaciones o con outliers altamente significativos la media y la desviación estándar pueden no ser representativas del verdadero comportamiento de la población y, el Z-Score, al depender de estas, puede clasificar erróneamente datos como outliers o viceversa.

Método IQR

Dada una variable cuantitativa $X = \{x_1, x_2, \cdots, x_n\}$, el método IQR es una medida estadística que describe la relación de los valores x_i con respecto a sus cuartiles $Q_1 = P_{0.25}$ y $Q_3 = P_{0.75}$.

Cálculo del IQR

- **Definición de Percentiles (P_k)**: Es el valor por debajo del cual se encuentran el $k\%$ de los valores de la variable X cuando están ordenados en forma ascendente. Se calcula como:

  $$P_k = x_{(\frac{k(n+1)}{100})}$$

  Donde:

  - $x_{(i)}$ es el valor x en la posición i dentro de la lista ordenada de los valores de X.
  - n es el número de observaciones de la variable X.
  - k la posición dentro de los valores ordenados.
- **Definición de IQR**: Es una medida estadística que mide la amplitud de una distribución entre el primer cuartil ($Q_1 = P_{0.25}$) y el tercer cuartil ($Q_3 = P_{0.75}$) capturando la dispersión central del 50% de los valores que componen la distribución. Se calcula como:

  $$IQR = Q_3 - Q_1 = P_{0.75} - P_{0.25}$$

  Donde:

  - $Q_3 = P_{0.75}$: es el valor que por debajo del cual se encuentran el 75% de los valores ordenados de la variable X.
  - $Q_1 = P_{0.25}$: es el valor que por debajo del cual se encuentran el 25% de los valores ordenados de la variable X.

Identificación de Outliers usando el método IQR

- Se define una constante arbitraria c, usualmente 1.5.
- Se define un límite superior como $ls = Q_3 + c \cdot IQR$.
- Se define un límite inferior como $li = Q_1 - c \cdot IQR$.
- Cualquier valor que x_i que cumpla $x_i > ls$ o $x_i < li$ es candidato a ser considerado valor atípico.

Limitaciones del método IQR

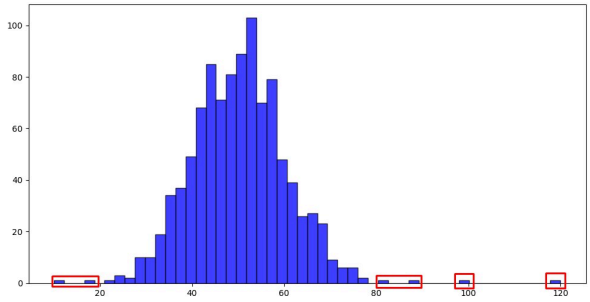
- **Interpretación Limitada en Distribuciones Multimodales**: En distribuciones multimodales (distribuciones con múltiples picos), el IQR puede no reflejar adecuadamente la dispersión y puede fallar en la identificación de outliers que se encuentran en diferentes modos de la distribución.
- **Límite Arbitrario**: La constante multiplicadora c utilizada para determinar los límites superior e inferior es arbitraria y puede no ser adecuada para todos los conjuntos de datos o contextos.
- **Interpretación Limitada en Distribuciones con Colas Pesadas**: El IQR puede no capturar adecuadamente estos outliers porque las colas pesadas pueden extenderse más allá de los límites.

Detección de Outliers Mediante Histogramas

Un histograma es una representación gráfica de la distribución de una variable de cualquier tipo X. Consiste en una serie de barras rectangulares adyacentes, donde cada barra representa una clase o intervalo de valores (Llamados bins). La altura de cada barra es proporcional a la frecuencia de los valores dentro de ese intervalo. Es particularmente útil para identificar la forma de la distribución, la dispersión y la presencia de valores atípicos.

Identificación de Outliers usando un Histograma

Cualquier valor o rango cuya barra esta aislada del centro de masas de la distribución es candidato a ser considerado outlier.



Limitaciones de la Identificación de Outliers usando Histogramas

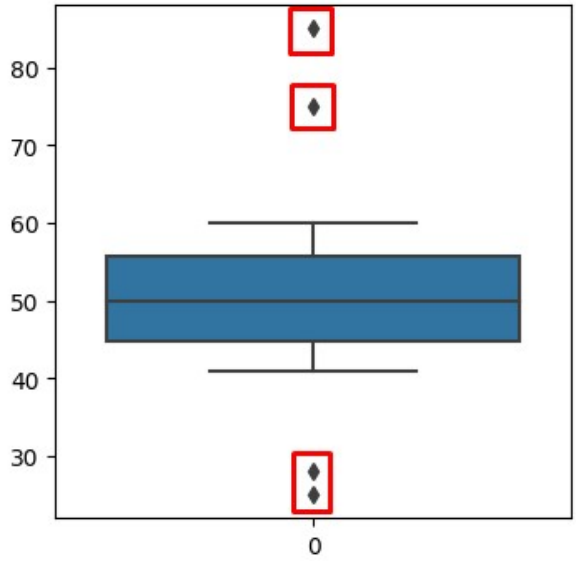
- **Dependencia del Número de Bins**: La elección del número de bins (segmentos) en un histograma puede afectar significativamente la visualización de los datos. Un número muy alto de bins puede hacer que el histograma sea ruidoso y difícil de interpretar, mientras que un número muy bajo puede ocultar la presencia de outliers al agrupar demasiados datos en cada bin.
- **Subjetividad en la Interpretación**: La identificación de outliers a partir de histogramas es en gran medida visual y subjetiva.
- **Interpretación Limitada en Distribuciones con Colas Pesadas**: En distribuciones con colas pesadas, los valores extremos pueden ser parte de la distribución esperada y no outliers verdaderos.
- **Interpretación Limitada en Distribuciones Complejas**: En distribuciones multimodales o muy asimétricas, los histogramas pueden no representar adecuadamente la distribución de los datos.
- **Sensibilidad a la Escala**: Los histogramas pueden ser sensibles a la escala de los datos y pueden necesitar una transformación previa para visualizar correctamente los outliers.

Detección de Outliers mediante Diagrama de Caja y Bigotes (Boxplot)

Un boxplot, también conocido como diagrama de caja y bigotes, es una representación gráfica que muestra la distribución de variable cuantitativa X basado en un resumen de cinco métricas: mínimo, primer cuartil ($Q1$, $P_{0.25}$), mediana ($Q2$, $P_{0.5}$), tercer cuartil ($Q3$, $P_{0.75}$) y máximo. Es particularmente útil para identificar la forma de la distribución, la dispersión y la presencia de valores atípicos.

Identificación de Outliers usando un Boxplot

Los valores atípicos se pueden detectar como puntos más allá de los bigotes. Estos valores atípicos son calculados usando el método IQR.



Limitaciones de la Identificación de Outliers usando Boxplots

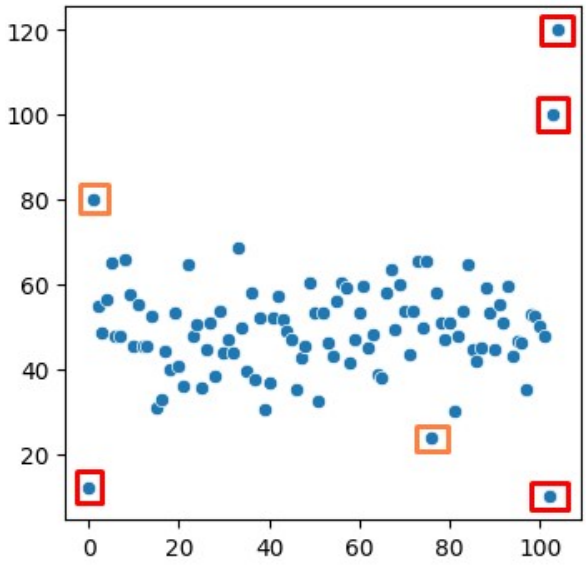
Los Diagramas de Cajas y Bigotes usan el método IQR para marcar los valores atípicos, por tanto, las limitaciones de estos son las mismas que las del método IQR descritas en puntos anteriores.

Detección de Outliers mediante un Gráfico de Dispersión (Scatterplot)

Los scatterplots univariantes, donde una variable cuantitativa X se grafica contra su índice, son útiles para visualizar la dispersión de los datos a lo largo de un eje temporal o secuencial y para identificar outliers.

Identificación de Outliers usando un Scatterplot

Cualquier valor que se puede señalar como suficientemente alejado del centro de masa de los demas valores es candidato a ser considerado valor atípico.



Limitaciones de la Identificación de Outliers usando Scatterplots

- **Subjetividad de la Interpretación**: La interpretación de lo que se considera un outlier puede variar dependiendo del analista y del contexto.
- **No Cuantitativo**: Los scatterplots son visuales y no proporcionan un criterio cuantitativo específico para la identificación de outliers.
- **Menos Eficaz para Datos No Secuenciales**: Son menos efectivos para datos que no tienen un orden temporal o secuencial claro.

Detección de Outliers Mediante Tabla de Frecuencias

Una tabla de frecuencias es una forma de resumir datos categóricos o discretos, mostrando la frecuencia de cada valor o categoría de cualquier tipo X, usualmente categórica. Puede incluir frecuencias absolutas, relativas , acumuladas y acumuladas relativas.

Calculo de Frecuencias

- **Frecuencia absoluta(f_i)**: La frecuencia absoluta de un valor x_i sobre una variable X es el número de veces que la variable X toma el valor x_i en los diferentes registros de un conjunto de datos. Para un valor específico $x_i \in X$, la frecuencia absoluta f_i, se calcula contando cuántas veces aparece x_i en los diferentes registros de un conjunto de datos.
- **Frecuencia Relativa(p_i)**: La frecuencia relativa de un valor x_i sobre una variable X es la proporción de veces que la variable X toma el valor x_i respecto el total de registros de un conjunto de datos. Para un valor específico $x_i \in X$, la frecuencia relativa p_i se calcula como:

$$p_i = \frac{f_i}{n}$$

Donde:

- n el número total de observaciones o registros de una variable X.
- **Frecuencia Absoluta Acumulada(F_i)**: La frecuencia absoluta acumulada de un valor x_i sobre una variable X es la suma acumulada de todas las frecuencias absolutas hasta ese valor. Para un valor específico $x_i \in X$, la frecuencia absoluta acumulada F_i se calcula como:

$$F_i = \sum_{j=1}^{i} f_j$$

- **Frecuencia Relativa Acumulada(P_i)**: La frecuencia relativa acumulada de un valor x_i sobre una variable X es la suma acumulada de todas las frecuencias relativas hasta ese valor. Para un valor específico $x_i \in X$, la frecuencia relativa acumulada P_i se calcula como:

$$P_i = \sum_{j=1}^{i} p_j$$

Identificación de Outliers usando una Tabla de Frecuencias

- Se selecciona un umbral arbitrario u, se suele utilizar un porcentaje pequeño de la muestra total como umbral.
- Cualquier categoría la cual cumpla $-u > x_i$ o $x_i > u$ es candidata de ser considerada outlier.

Categoría	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Acumulada Relativa
A	10	0.20	10	0.20
B	15	0.30	25	0.50
C	5	0.10	30	0.60
D	20	0.40	50	1.00

Limitaciones de la Identificación de Outliers usando Tabla de Frecuencias

- **Subjetividad en el Umbral**: La elección del umbral de frecuencia mínima es subjetiva y puede variar dependiendo del analista y del contexto.

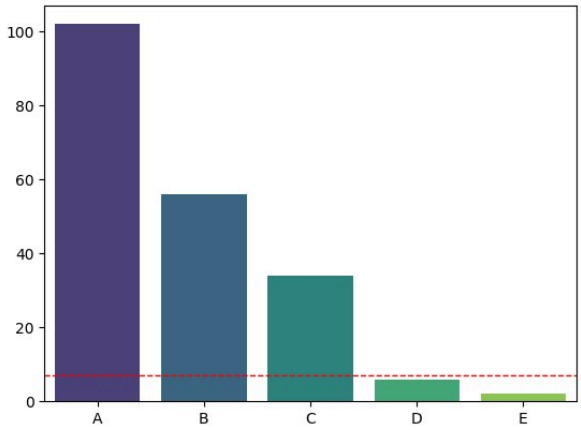
- **No Considera la Variabilidad**: No considera la variabilidad intrínseca de las categorías, lo que puede llevar a la identificación de outliers que son legítimos en ciertos contextos.
- **Menos Eficaz para Datos Muy Desbalanceados**: En conjuntos de datos extremadamente desbalanceados, este método puede identificar muchas categorías como outliers.

Detección de Outliers mediante un Gráfico de Barras (Barplot)

Un gráfico de barras es una representación gráfica que utiliza barras rectangulares para mostrar la frecuencia, cantidad o proporción de diferentes categorías o valores de una variable de cualquier tipo X, usualmente categórica. Las barras pueden orientarse vertical u horizontalmente y su altura (o longitud) es proporcional al valor que representan.

Identificación de Outliers usando un Gráfico de Barras

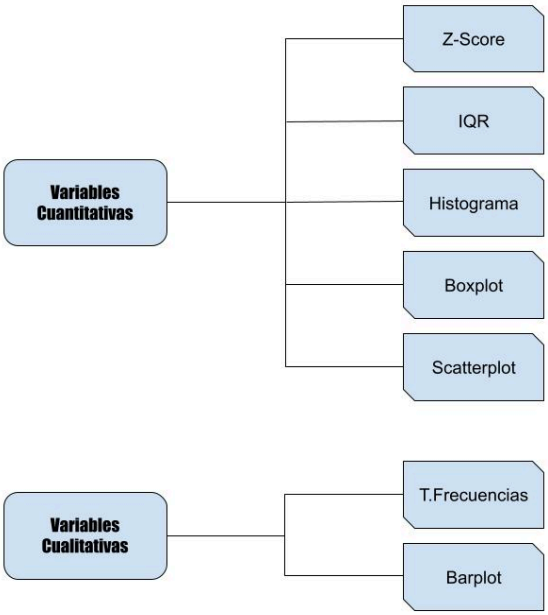
- Se selecciona un umbral arbitrario u, se suele utilizar un porcentaje pequeño de la muestra total como umbral.
- Se dibuja el umbral como linea horizontal en el gráfico de barras
- Cualquier categoría por debajo del umbral u es candidata a ser considerada outlier.



Limitaciones de la Identificación de Outliers usando Barplots

Los gráficos de barras y las tablas de frecuencias usan el mismo umbral para clasificar las categorías de la variable X como valores atípicos, por tanto, sus limitaciones son las mismas.

Resumen de Identificación de Valores Atípicos Univariantes



Identificación de Valores Atípicos Multivariantes

Distancia de Mahalanobis

Dadas p variables cuantitativas $X_1, X_2, \cdots, X_p$, la distancia de Mahalanobis para una observación individual i, $\mathbf{x_i} = (x_1, x_2, \cdots, x_p)$ con respecto a la media multivariada del conjunto de datos se define como:

$$D_M(\mathbf{x_i}) = \sqrt{(\mathbf{x_i} - \bar{\mathbf{x}})^\top \mathbf{S}^{-1} (\mathbf{x_i} - \bar{\mathbf{x}})}$$

Donde:

- $\mathbf{x_i}$ es el vector de datos de la observación i, $\mathbf{x_i} = (x_1, x_2, \cdots, x_p)$.
- $\bar{\mathbf{x}}$ es el vector de medias de las variables, $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)$.
- $\mathbf{S}$ es la matriz de covarianza de las variables $X_1, X_2, \cdots, X_p$.
- $\mathbf{S}^{-1}$ es la inversa de la matriz de covarianza $\mathbf{S}$.
- $(\mathbf{x_i} - \bar{\mathbf{x}})^\top$ es el vector transpuesto de las diferencias entre el punto compuesto por la observación i y las medias.

Identificación de Outliers usando la Distancia de Mahalanobis

- Se calcula la distancia de Mahalanobis para cada observación i.
- Se define un umbral u, usualmente el valor crítico de la distribución chi cuadrado con p grados de libertad, χ^2_p
- Cualquier observación o punto $\mathbf{x_i}$ que cumpla $\mathbf{x_i} > u$ es candidata a ser considerada valor atípico.

Limitaciones al uso de la Distancia de Mahalanobis Para la Detección de Valores Atípicos

- **Asume Distribución Normal Multivariada**: La distancia de Mahalanobis se basa en la media y la matriz de covarianza, asumiendo que los datos siguen una distribución normal multivariada.
- **Asume no Multicolinealidad**: La inversión de la matriz de covarianza, necesaria para el cálculo de la distancia de Mahalanobis, puede ser singular o casi singular si hay multicolinealidad en los datos lo que imposibilitaría o perjudicaría el cálculo de S^{-1}.
- **Sensibilidad a Outliers**: La media y la matriz de covarianza son necesarias para el cálculo de la distancia de Mahalanobis y estas son sensibles a outliers y la presencia de estos puede resultar en medidas inexactas. Esto se puede solucionar usando alternativas a la media y la varianza robustas.
- **Asume Linealidad**: La distancia de Mahalanobis se basa en una relación lineal entre las variables. Si hay relaciones no lineales entre las variables, la distancia de Mahalanobis puede no capturar adecuadamente las distancias verdaderas, llevando a identificaciones incorrectas de outliers.
- **Interpretación en Altas Dimensiones**: En altas dimensiones, el exceso de variables, puede hacer difícil distinguir entre puntos normales y outliers.
- **Computacionalmente Costoso**: El cálculo de la distancia de Mahalanobis, especialmente la inversión de la matriz de covarianza, puede ser computacionalmente costoso para conjuntos de datos grandes.
- **Umbral Arbitrario**: La elección del umbral u, al ser arbitrario, no tiene porque ser adecuado para todos los conjuntos de datos.

Método IQR Generalizado

Para cualquier conjunto de p variables cuantitativas $X_1, X_2, \cdots, X_p$, Se puede generalizar el método IQR univariante añadiendo un condicional g el cual determine el número de variables máximas aceptadas por observación que cumplan con el método IQR univariante.

Identificación de Outliers usando el método IQR generalizado

- Se define una constante arbitraria c, usualmente 1.5.
- Se define un condicional g, usualmente 1.
- Se define un límite superior como $ls = Q_3 + c \cdot IQR$.
- Se define un límite inferior como $li = Q_1 - c \cdot IQR$.
- Cualquier valor que $x_{ij} \in \mathbf{x_i}$, siendo $\mathbf{x_i}$ la observación i, que cumpla $x_{ij} > ls$ o $x_{ij} < li$ será contabilizada en una variable $cont_i$.
- Cualquier observación $\mathbf{x_i}$ con un contador asociado $cont_i$ cumple que $cont_i > g$ es candidata a ser considerada outlier.

Limitaciones del método IQR Generalizado

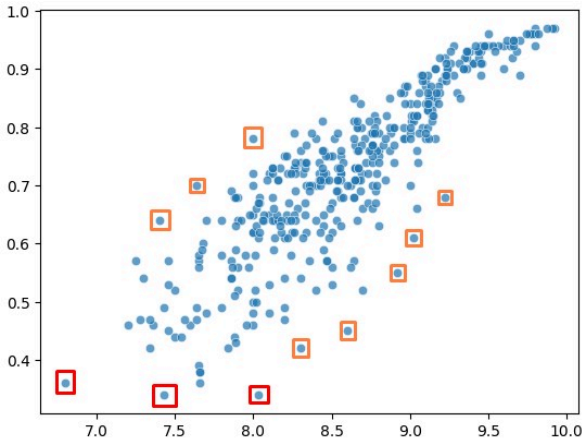
Las limitaciones del método IQR Generalizado son las mismas que las del método IQR añadiendo otro nivel de subjetividad introducido por la elección arbitraria del condicional g.

Detección de Outliers mediante un Gráfico de Dispersión Bivariante (Scatterplot)

Un gráfico de dispersión, también conocido como diagrama de dispersión o scatter plot, es una representación gráfica que utiliza coordenadas cartesianas para mostrar valores de dos variables cuantitativas X_1, X_2. Cada punto en el gráfico representa un par de valores de estas variables. Esta graficación es particularmente útil ya que permite observar la relación entre las dos variables y detectar puntos que se desvían significativamente del patrón general.

Identificación de Outliers usando un Scatterplot

Cualquier valor que se puede señalar como suficientemente alejado del centro de masa de los demas valores es candidato a ser considerado valor atípico.



Limitaciones de la Identificación de Outliers usando Scatterplots

- **Subjetividad de la Interpretación**: La interpretación de lo que se considera un outlier puede variar dependiendo del analista y del contexto.
- **No Cuantitativo**: Los scatterplots son visuales y no proporcionan un criterio cuantitativo específico para la identificación de outliers.

¿Cuándo Tratar Outliers?

El tratamiento de outliers es una decisión crítica que puede afectar significativamente el objetivo y rendimiento del análisis. A continuación se presentan algunas consideraciones sobre cuándo tratar estos valores atípicos:

- **Errores de Medición o Entrada de los Datos**: Si los outliers son claramente el resultado de errores, deben ser corregidos o eliminados.
- **Tipos de Modelo o Análisis**: Algunos modelos o algoritmos son más robustos que otros, esto es importante tenerlo en cuenta, pues si, el modelo o el análisis a realizar sobre un conjunto de datos puede no ser neceseario tratar los valores atípicos.
- **Dependencia de la Distribución**: Si los datos siguen una distribución con colas largas, los valores extremos pueden no ser verdaderos outliers por lo que no sería necesario tratarlos.
- **Objetivo del Análisis o Contexto**: En algunas tareas, como la detección de anomalías, los outliers son el foco del análisis por lo que no deben ser modificados o eliminados.

Sensibilidad a Outliers en Modelos de Machine Learning

Modelo	Sensibilidad a Outliers	Razón
Regresión Lineal	Alta	Los outliers pueden influir significativamente en la línea de mejor ajuste.
Regresión Logística	Alta	Los outliers en las características pueden distorsionar la predicción de probabilidades.
K-Nearest Neighbors	Media	Los outliers pueden afectar las distancias y, por ende, las predicciones de los vecinos más cercanos, pero el impacto puede ser mitigado ajustando K y la métrica de distancia.
SVM (Kernel Lineal)	Alta	Los outliers pueden afectar la posición del hiperplano de separación.
Árboles de Decisión	Baja	Los árboles de decisión son menos sensibles a los outliers porque dividen el espacio de las características de forma recursiva.
Bosques Aleatorios	Baja	La agregación de múltiples árboles de decisión ayuda a mitigar el impacto de los outliers.
Gradient Boosting	Media	Aunque cada árbol individual es sensible, la técnica de boosting puede reducir el impacto global de los outliers mediante la ponderación de errores.
Regresión Robusta	Baja	Diseñada específicamente para ser menos sensible a los outliers, utilizando métodos como la regresión de Huber o la regresión de M-estimación.
K-Means Clustering	Alta	Los outliers pueden influir en la posición de los centroides.
DBSCAN	Baja	Este algoritmo de clustering es capaz de identificar y manejar outliers naturalmente.
Redes Neuronales	Media	La sensibilidad puede ser mitigada mediante técnicas como la normalización y la regularización, pero los outliers aún pueden afectar el entrenamiento.
PCA	Alta	Los outliers pueden influir en la dirección de los componentes principales, distorsionando la representación de los datos.
Isolation Forest	Baja	Diseñado específicamente para detectar outliers mediante el aislamiento de observaciones.

Métodos para Tratar valores Atípicos.

Aunque cada modelo, contexto o análisis específico tienen su forma particular de tratar los valores atípicos, se intentará dar, de manera generalizada un enfoque para tratar estos valores.

- **Eliminación de Outliers**
  - **Descripción**: Eliminar las observaciones que son consideradas outliers.
  - **Cuándo usarlo**: Cuando los outliers son errores evidentes o irrelevantes para el análisis.
  - **Ventajas**:
    - Sencillo y fácil de implementar.
    - Mejora la interpretación de los resultados.
  - **Desventajas**:
    - Puede llevar a la pérdida de información valiosa.
    - No siempre es adecuado si los outliers son parte intrínseca de los datos.
- **No Hacer Nada**
  - **Descripción**: Dejar los outliers en el conjunto de datos sin realizar ningún tratamiento.
  - **Cuándo usarlo**: Cuando se tiene certeza de que los outliers son representativos de la variabilidad natural de los datos y no se desea alterar el conjunto de datos.
  - **Ventajas**:
    - Mantiene la integridad del conjunto de datos original.
    - No se pierde información.
  - **Desventajas**:
    - Los outliers pueden distorsionar los resultados del análisis.
    - Puede dificultar la interpretación de los resultados.
- **Transformación de Datos**
  - **Descripción**: Aplicar transformaciones matemáticas para reducir el impacto de los outliers.
  - **Cuándo usarlo**: Cuando los datos tienen distribuciones sesgadas o colas largas.
  - **Ventajas**:
    - Reduce la influencia de los outliers.
    - Puede hacer que los datos se ajusten mejor a las suposiciones del modelo.
  - **Desventajas**:
    - Puede ser difícil interpretar los datos transformados.
    - No elimina completamente el problema de los outliers.
- **Imputación de Valores**
  - **Descripción**: Reemplazar los outliers con valores más consistentes con el resto del conjunto de datos.
  - **Cuándo usarlo**: Cuando se quiere mantener el tamaño del conjunto de datos sin eliminar observaciones.
  - **Ventajas**:
    - Mantiene el tamaño del conjunto de datos.
    - Evita la pérdida de información.
  - **Desventajas**:
    - Puede introducir sesgos si no se hace correctamente.
    - No siempre es fácil decidir qué valor utilizar para la imputación.
- **Métodos Robustos**
  - **Descripción**: Utilizar algoritmos y técnicas que son inherentemente menos sensibles a los outliers.
  - **Cuándo usarlo**: Al usar modelos que pueden manejar outliers de manera efectiva.
  - **Ventajas**:
    - Evita la necesidad de eliminar o transformar los datos.
    - Mantiene la integridad del conjunto de datos original.

    - **Desventajas**:
        - Puede ser computacionalmente intensivo.
        - No siempre está claro si el modelo robusto está interpretando los outliers de manera adecuada.
- **Análisis Separado**
    - **Descripción**: Analizar los outliers por separado para entender mejor su naturaleza antes de decidir cómo tratarlos.
    - **Cuándo usarlo**: Cuando se sospecha que los outliers pueden contener información valiosa.
    - **Ventajas**:
        - Permite un entendimiento profundo de los outliers.
        - Puede revelar patrones o problemas subyacentes en los datos.
    - **Desventajas**:
        - Puede ser laborioso y requerir mucho tiempo.
        - No siempre resulta en un claro curso de acción para el tratamiento de los outliers.

Tipos de Imputación

- **Imputación por la Media**
    - **Descripción**: Reemplazar los valores faltantes o atípicos con la media de la columna.
    - **Cuándo usarlo**: Cuando los datos siguen una distribución aproximadamente normal y no hay muchos valores faltantes.
    - **Ventajas**:
        - Simple y rápido de implementar.
        - Mantiene la media de la columna.
    - **Desventajas**:
        - Reduce la variabilidad de los datos.
        - No es adecuado para datos con distribuciones sesgadas.
- **Imputación por la Mediana**
    - **Descripción**: Reemplazar los valores faltantes o atípicos con la mediana de la columna.
    - **Cuándo usarlo**: Cuando los datos tienen outliers o distribuciones sesgadas.
    - **Ventajas**:
        - Robusto a outliers.
        - Mantiene la tendencia central sin distorsión.
    - **Desventajas**:
        - Puede no reflejar adecuadamente la variabilidad de los datos.
- **Imputación por la Moda**
    - **Descripción**: Reemplazar los valores faltantes o atípicos con el valor más frecuente de la columna.
    - **Cuándo usarlo**: Para datos categóricos o discretos.
    - **Ventajas**:
        - Fácil de implementar.
        - Mantiene la moda de la columna.
    - **Desventajas**:
        - No es útil para datos continuos.
        - Puede introducir sesgos si la moda no es representativa.
- **Imputación Basada en Modelos**
    - **Descripción**: Utilizar modelos predictivos para estimar los valores faltantes o atípicos.
    - **Cuándo usarlo**: Cuando se dispone de suficientes datos y recursos computacionales.

- **Ventajas**:
  - Puede capturar relaciones complejas entre variables.
  - Generalmente proporciona estimaciones más precisas.
- **Desventajas**:
  - Más complejo y computacionalmente intensivo.
  - Requiere un modelo bien ajustado.

- **Imputación Estocástica**
  - **Descripción**: Reemplazar los valores faltantes con valores generados aleatoriamente dentro de una distribución estimada de los datos.
  - **Cuándo usarlo**: Cuando se quiere mantener la variabilidad de los datos.
  - **Ventajas**:
    - Mantiene la variabilidad y distribución de los datos originales.
  - **Desventajas**:
    - Puede introducir variabilidad adicional no deseada.
    - Complejo de implementar.