



Big Data Real-Time Analytics Com Python e Spark 3.0

Big Data Real-Time Analytics Com Python e Spark Versão 3.0

Valores Outliers e Z-score

Outliers são pontos de dados que estão longe de outros pontos de dados. Em outras palavras, são valores incomuns em um conjunto de dados. Os valores atípicos são problemáticos para muitas análises estatísticas porque podem fazer com que os testes percam descobertas significativas ou distorçam os resultados reais.

Infelizmente, não existem regras estatísticas estritas para identificar definitivamente os outliers. Encontrar outliers depende do conhecimento da área de assunto e da compreensão do processo de coleta de dados. Embora não haja uma definição matemática sólida, existem diretrizes e testes estatísticos que você pode usar para encontrar candidatos discrepantes.

Outliers são um conceito simples - são valores notavelmente diferentes de outros pontos de dados (normalmente distantes da média dos dados) e podem causar problemas em procedimentos estatísticos.

Usando Z-scores Para Detectar Valores Outliers

Z-scores são o número de desvios padrão acima e abaixo da média. Por exemplo, um escore Z de 2 indica que uma observação está dois desvios padrão acima da média, enquanto um escore Z de -2 significa que está dois desvios padrão abaixo da média. Um Z-score de zero representa um valor que é igual à média.

Para calcular o Z-score de uma observação (X), subtraia a média (μ) e divida pelo desvio padrão (σ). Matematicamente, a fórmula para esse processo é a seguinte:

$$Z = \frac{X - \mu}{\sigma}$$

Quanto mais longe o Z-score de uma observação estiver de zero, mais incomum essa observação será. Um valor de corte padrão para encontrar valores discrepantes são escores Z de +/- 3 ou acima.

Esta não é a única forma de identificar valores outliers em uma variável, mas é um método bastante efetivo. Usaremos no Estudo de Caso deste capítulo.