第2章
小売店のデータでデータ加工を行う10本ノック

本章では、小売店の売上履歴と顧客台帳データを用いて、データの分析や予測を行うための重要なノウハウである「データの加工」を習得します。小売店のデータは、ECサイトのシステムによって管理されたデータと違って、人間の手を介在します。このため、日付などの入力ミスや、データの抜け漏れ等、人間ならではの「間違い」を多く含みます。そうしたデータを、ECサイトの場合と同じ方法で読み込もうとすると、おのずと誤った結果を導き出してしまったり、そもそも処理することができなかったりと、さまざまな問題が発生します。そうした「汚い」データを扱う練習問題として、小売店のデータは適していると言えます。

小売店以外のビジネス現場でも、Excel等による手入力の作業は少なくありません。手入力で作成されたデータは機械的な入力チェックなどが行われていないため、データはおのずと「汚く」なってしまい、そのままではデータ分析に活用できないのです。人間からすると大差ないように見えてしまう「半角」と「全角」の違いなども、データ分析においては誤作動を引き起こす「汚い」データとなってしまいます。その他にも、異なる部署から集められたデータを扱おうとする場合には、それぞれ独自のシステムでデータが管理されており、それらを統一的に扱うのは容易ではありません。実際のビジネス現場でデータ分析を行おうとすると、さまざまな「汚い」データに直面し、一筋縄ではいかないケースに戸惑う事も多いのです。

本章を通し、より現場に近い「汚い」データを処理する経験を積むことで、ビジネス現場ならではの種々雑多な状況に対処できる力を身につけましょう。それでは、前章と同様に、「顧客の声」と「前提条件」を確認したうえで、データの読み込みを行っていきましょう。

ノック11：データを読み込んでみよう
ノック12：データの揺れを見てみよう
ノック13：データに揺れがあるまま集計しよう
ノック14：商品名の揺れを補正しよう
ノック15：金額欠損値の補完をしよう
ノック16：顧客名の揺れを補正しよう
ノック17：日付の揺れを補正しよう
ノック18：顧客名をキーに2つのデータを結合（ジョイン）しよう
ノック19：クレンジングしたデータをダンプしよう
ノック20：データを集計しよう