

ブレンドと結合の使い分け

「データのブレンドとクロスデータベース結合の比較」のトレーニングによるこそ。

付属の練習ワークブックをダウンロードして、実際に Tableau を操作してみてください。

簡単に言うと、クロスデータベース結合では、2 つのデータソースを行レベルで統合して 1 つのデータソースを作成します。ブレンドでは、各データソースから別々にクエリ結果を取得して、ビューで集計します。クロスデータベース結合とデータブレンディングの詳細については、それぞれのトレーニングビデオをご覧ください。

1 対多のリレーションシップ

まず、クロスデータベース結合のトレーニングの [Product 2016] と [Sales 2016] のデータソースのブレンドを設定するところから始めます。市場ごとのカテゴリ別売上を簡単に分析してみましょう。

- **Product** のデータソースから [カテゴリ] を [行] にドラッグします。
- 次に、**Sales** のデータソースに切り替えます。
 - [製品 ID] のリンクをクリックして、ブレンドのリレーションシップを設定し、
 - [売上] を [列] にドラッグします。
- これでデータがブレンドされ、カテゴリ別の全体的な売上がわかります。
- ここで、[市場] を [色] にドラッグするとどうなるでしょうか。棒の大部分がアスタリスクの灰色になりました。なぜでしょうか。

ここでは [製品 ID] でブレンドしましたが、このデータは市場に対して 1 対多のリレーションシップを持っています。淡い青緑の棒は、米国・カナダの市場のみで販売されている製品 ID を示しています。これに対して、灰色の部分はすべて、複数の市場で販売されている製品であり、集計されてアスタリスクで表示されています。ブレンドでは、データソースごとの結果が別々に集計され、ビューでまとめて表示されるので、1 対多のリレーションシップがある場合、簡単にデータを視覚化することはできません。

このような場合は、ブレンドよりもクロスデータベース結合が適しています。このデータソースは、製品 ID で行レベルで結合されています。最初のビューから作り直してみましょう。

- [カテゴリ] を [行] に、
- [売上] を [列] に、
- [市場] を [色] にドラッグします。

これで期待通りの結果が得られました。カテゴリ別の売上が市場ごとに表示されています。データを結合すると、売上情報に製品情報が補足されただけで、[売上] の元のレコード数は変わりませんでした。

結合による行数の増加

では、いつでもクロスデータベース結合を使えばよいのでしょうか。ここで、別のデータセットに切り替えてみましょう。今度はデータブレンディングのトレーニングビデオのオフィスシティとコーヒーチェーンを使用します。この 2 つは、同じ親会社が所有する小売りチェーンです。オフィスシティは米国の全 50 州で、コーヒーチェーンは 20 州で営業しています。今回は、両方の組織の合算売上を 50 州すべてについて見てみたいとしましょう。

ここには、共通フィールドである [州] で結合したデータソースがあります。

- オフィスシティのデータには 50 州すべてが含まれているので、このデータソース内の [オフィスシティ] の接続から [州] をドラッグします。
- [加算された売上] という計算を右クリックして編集すると、オフィスシティの売上とコーヒーチェーンの売上が加算されていることを確認できます。
 - Zero Null 関数が使われているので、コーヒーチェーンの売上がない州の NULL は 0 として処理されています。
- [加算された売上] を [列] にドラッグします。
- ご覧ください。億単位の売上がある州があるとすれば、ビジネスは大繁盛ですね。しかし、たまたまそれは事実でない知っているのもっと掘り下げてみましょう。

このダッシュボードには、4 つの州のオフィスシティの売上と、同じ 4 つの州のコーヒーチェーンの売上が表示されています。合算売上は、中央の数式にあるように、両方の値を加算したものになるはずですが。

- ブレンドした (黄色の) 合算売上は正しい数字になっていますが、結合した (紫の) データソースでは、値が膨れ上がっています。

レコード数をオフィスシティ、コーヒーチェーン、結合したデータソースで比較してみると、明らかに増大しています。

[州] で結合したことにより、コーヒーチェーンの製品と日付、オフィスシティの行 ID など、一意の情報と州を個々に組み合わせて行が作成されています。このため、同じ売上の値が複数回カウントされる結果となっています。

これに対して、ブレンドでは正しい値が出ています。[州] について 1 つのデータソースから売上の値を取り出し、もう 1 つのデータソースからその州の売上の値を取り出して、それらを加算しているからです。今回の場合は、これが望ましい動作です。

ご覧のように、どのフィールドで結合やブレンドを行うかが、分析に大きな影響を及ぼします。正しいフィールドを選択すること、そして、ブレンドや結合がどのように行われるかを理解していることが、どちらを使うべきかを判断するうえで非常に重要です。

それぞれのメリット

クロスデータベース結合を行ってデータソースを統合すると、さまざまなメリットがあります。

- 結合は行レベルで行われるので、データブレンディングのように 1 対多の関係による制限がなく、アスタリスクが表示されることはありません。
- データのブレンドとは異なり、単一のデータソースを抽出して保存し、パブリッシュして、簡単に他のユーザーと共有できます。他のデータソースと同じように扱うことができます。

ただし、結合するフィールドの粒度が適切でないと、結合によってデータセットのサイズが不自然に膨れ上がり、誤った情報が引き出されるおそれがあります。

データのブレンドでは、データソースごとに別々にクエリを送り、ビューで必要な粒度にそれらを集計することで、分析に必要な結果を得ることができます。

まとめ

ブレンドとクロスデータベース結合によるデータソースの結合のトレーニングビデオをご視聴いただき、ありがとうございます。Tableau の使用方法について、引き続き無料のトレーニングビデオをご覧ください。