PDF への接続

PDF コネクタのトレーニングへようこそ。付属の PDF をダウンロードして、実際に Tableau を操作してみてください。PDF には、Tableau での分析に役立つデータが含まれている可能性があります。Tableau の PDF コネクタは、表からデータを抽出する場合に役立ちます。なお、PDF にはデータに関するメタデータがないため、分析するには接続後にユーザーが作業を行わなければならないこともあります。

PDF のデータ表への接続

Tableau は、PDF からクロス集計のデータ表を読み込むことができます。これが理想的な表です。列と行があり、各行に 1 行のデータが入っています。2 ページ目からは株価データが始まります。Tableau で接続してみましょう。[接続] ペインで [PDF ファイル] を選択すると、ページ選択のダイアログが表示されます。スキャンはドキュメント全体、指定ページ、ページの範囲で行えます。ここでは 2 ～ 8 ページを見てみましょう。そして、最初のページをキャンバスにドラッグします。PDF の表ヘッダーで、ある行が誤って解釈されていますが、データインタープリターをオンにするとヘッダーは正しくなります。

ユニオン

各ページは別々の表になりましたが、どの表も同じ列ヘッダーを持つ同じ構造になっているため、ユニオンで簡単にまとめ直すことができます。最初の表がすでにキャンバスにあるので、その下にあるユニオンのドロップエリアに、まとめたい他のページの表をドラッグします。すると表の名前の新しい列が作成されます。また、下にスクロールすると、3 ページ目からユニオンされたデータは 2 ページ目とぴったり揃っていることがわかります。

不完全な表のクリーンアップ

この PDF はインポートしやすい構造になっていますが、そうではない場合もあります。一般的に、Tableau はこのような表構造を持つ PDF に最もうまく接続できます。各行に単一行のデータがあり、階層も入れ子になったヘッダーもなく、サブテーブルもない構造です。しかし、完璧ではない PDF も利用が可能です。別のファイルに接続してみましょう。最初のものほど簡単にはいかないファイルです。使いたい表はこの 14 ページ目にあります。なお、Tableau は絶対ページ番号で参照するので、ドキュメント内のページ番号と一致するとは限りません。

では別のデータソースを追加し、14 ページ目を選択しましょう。データ接続ドロップダウンの [PDF ファイルを再スキャン] オプションを選択すると、使うページをもう一度指定することができます。このページに表は 1 つしかありませんが、このように左側には選択肢が 3 つ表示されています。これは、Tableau が表の読み込み方に 3 つの可能性があると考えたためです。1 つずつドラッグすると、それぞれの内容を見ることができます。なお、先ほどのように複数のページの表をユニオンするときに、各ページに複数の表がある場合は、同じページの複数の表ではなく適切な表をユニオンしていることを確認してください。

まず、Table 1 をドラッグしてデータインタープリターを使用します。見たところ、Table 1 にはすべての情報があるようですが、なぜか 1995 ～ 1997 年が 1 つの列として読み込まれています。このデータ表を使うのであれば、カスタム分割機能を使い、すべての列をスペースで分割してクリーンアップした後、分割したフィールドの名前をそれぞれの年に変更することもできます。

ですが先に、Table 2 と 3 がどうなっているか見てみましょう。Table 2 は、元の表の下部分のようです。そして Table 3 は上部分のようです。列の分かれ方は Table 2 と 3 の方がいいので、この 2 つを使いましょう。まずユニオンします。Table 2 を Table 3 の下にドラッグしましょう。すると [Inflows] と [F1] が分かれてしまっています。そこで、両方を選択して [一致していないフィールドをマージ] を選択すると、期待通りの列が得られました。名前を「Water Sources」に変えましょう。

NULL 値の扱い方

NULL の行がいくつかあります。これは、「Change in storage」などのサブヘッダーがデータ行として読み込まれたか、「Abstraction from hydroelectricity」などの単一行が 2 つの行に分割されたためです。NULL をなくすには、データソースフィルターを追加します。それには右上の [追加] をクリックして、フィルターを追加してください。NULL があるどの列でも選択できますが、ここでは [F10] を選択して [OK] をクリックし、次に [NULL] と [除外] を選択します。[OK] をクリックし、もう一度 [OK] をクリックします。これで NULL の行がなくなりデータだけになりました。

ですが、いったん元に戻します。水源の種類の中に実際は合計値を示しているものもあるためです。では、合計値と NULL をすべて一度に除外してみましょう。[フィルター] > [追加] をクリックして、[Water Sources] を選択します。次に、元の PDF でヘッダーか合計値になっているものを選択します。まず、値が NULL になっている行です。[hydroelectricity] は、データのあるものと NULL のものの 2 つがあり、データは削除したくないのでここでは手をつけません。[除外] をクリックして [OK] をクリックします。

ヘッダーの修正とピボット

これでほとんどがデータの行になりました。[hydroelectricity] の NULL の行が 1 つ残っていますが、後でクリーンアップします。最初の列以外はヘッダーがありませんが、元の PDF で値を見れば、どれが何かを調べることができます。列は 1995 年から 2010 年までのはずです。細かいところは割愛します。

ここで、データのピボットを行います。すると、年の列と 100 万立方メートルの列ができました。[Table Name] の列は非表示にし、[Year] のデータ型を [日付] に変更します。[Million Cubic Meters] のデータ型は [数値 (整数)] にします。これで列の体裁が整ったので、この列から NULL を除外することができます。[フィルターの編集] > [追加] で [Million Cubic Meters] を選択し、[NULL] と [除外] を選択します。

フィールドのメンバーに別名を再び付ける

水源の名前自体にはまだ問題があります。 そこで、このフィールドのメニューを表示し、[別名] を選択します。ここでは、どのフィールドのメンバーでも別名を付け直すことができます。別名をダブルクリックして、正しい名前を入力してください。

- **Discharge from hydroelectricity** generation
- Groundwater
- **Abstraction for** hydroelectricity

グループと階層を再び作成する

最後に、元の PDF 内の表には構造がありました。水源のカテゴリーです。この構造は [データ] ペインで作成できるので、[シート 1] をクリックします。まず、[Million Cubic Meters] は実際にはメジャーなので、そこにドラッグします。

では、グループを再作成しましょう。[Water Sources] を [行] にドラッグして最初のグループを作成します。Ctrl キーを押しながら、同じグループに入れるメンバーをクリックします。

- Abstraction for hydroelectricity
- Discharge from hydroelectricity generation
- Evapotranspiration
- To sea and net abstraction

ツールヒントのクリップアイコンをクリックすると、グループが作成されます。同じ操作で、もう一度グループを作成します。[Groundwater]、[Ice]、[Lakes and reservoirs]、[Snow]、[Soil moisture] です。

[データ] ペインでこの新しいフィールドを右クリックして、グループを編集しましょう。まず、[Precipitation] を選択して [グループ] をクリックし、他にメンバーのないそれだけのグループを作成します。次にそれぞれの名前を、「Inflows」、

「Change in storage」、「Outflows」に変更します。そして、グループ化したフィールドの名前を「Categories」にします。元の [Water Sources] をこの新しい [Categories] フィールドの上にドラッグすると、階層が作成され、ビューでドリルダウンできるようになります。

PDF 操作のヒント

PDF に接続した後のクリーンアップの詳しい手順は PDF ごとに異なりますが、データをクリーンアップして分析の準備を整えるためのツールは、このビデオでおわかりになったと思います。なお Tableau は、一部の PDF からはデータをうまく読み込めません。PDF に、サブテーブル、ヘッダー内の階層、単一行として認識されるべき複数の行のコンテンツがある場合です。最後になりますが、色や網掛けによっては、PDF からセルとデータ表への変換方法が変わり、そのためデータの認識のされ方も変わることがあります。

最後に

PDF コネクタのトレーニングビデオを視聴いただき、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。