Verbinden mit PDF-Dateien

Willkommen bei diesem Video zum PDF Connector. Sie können die PDF-Dateien herunterladen und Ihre eigene Version von Tableau verwenden, um dem Kurs zu folgen.

PDF-Dateien können wertvolle Daten für Analysen in Tableau enthalten. Der in Tableau enthaltene PDF Connector ist darauf ausgelegt, die Daten aus Tabellen zu extrahieren. Da in PDF-Dateien keine Metadaten zu den Daten vorhanden sind, kann nach Beendigung der Verbindung und vor dem Durchführen der Analyse noch eine Nachbearbeitung erforderlich sein.

Verbinden mit Datentabellen in PDF-Dateien

Tableau kann in PDF-Dateien enthaltene Kreuztabellen lesen, die idealerweise ungefähr so aussehen – mit Spalten und Zeilen, wobei jede Zeile nur eine Datenzeile enthält. Hier beginnen die Bestandsdaten auf Seite 2. Wir stellen jetzt eine Verbindung zu Tableau her. Wenn wir im Verbindungsbereich die Option „PDF-Datei“ auswählen, wird eine Seitenauswahl eingeblendet. Wir können das ganze Dokument, eine bestimmte Seite oder einen Seitenbereich durchsuchen. Wir werden die Seiten 2 bis 8 betrachten. Und wir ziehen die erste Seite in den Arbeitsbereich. Eine der Zeilen aus der PDF-Tabellenkopfzeile bringt Tableau durcheinander, doch wenn wir den Dateninterpreter aktivieren, erhalten wir die erwarteten Kopfzeilen.

Zusammenführung

Jede Seite wurde als separate Tabelle importiert, doch da diese Tabellen genau gleich strukturiert sind und über dieselben Spaltenüberschriften verfügen, lassen sie sich ganz leicht wieder zusammenführen. Wenn die erste Tabelle bereits im Verknüpfungsbereich vorhanden ist, ziehen wir alle sonst noch gewünschten Seiten in den Zusammenführungsablagebereich unter der ersten Seite. Wir erhalten eine neue Spalte für den Tabellennamen, und wenn wir nach unten scrollen, sehen wir, dass die zusammengeführten Daten von Seite 3 und Seite 2 perfekt ausgerichtet sind.

Bereinigen fehlerhafter Tabellen

Diese PDF-Datei war für den Import genau richtig strukturiert. Doch das wird nicht auf alle PDF-Dateien zutreffen. Generell gilt, dass sich Tableau am besten mit PDF-Dateien verbinden lässt, die eine solche Tabellenstruktur aufweisen, also nur eine einzelne Datenzeile pro Zeile und weder Hierarchien oder verschachtelte Kopfzeilen noch Untertabellen. Doch PDF-Dateien müssen nicht perfekt sein, um importiert werden zu können. Jetzt stellen wir eine Verbindung zu einer Datei her, die etwas komplexer ist als die erste Datei. Die gewünschte Datei befindet sich hier auf Seite 14. (Anmerkung: Tableau betrachtet die absoluten Seitenzahlen, die nicht mit der Seitennummerierung im Dokument übereinstimmen müssen.)

Wir fügen eine weitere Datenquelle hinzu und wählen Seite 14 aus. Mithilfe der Option „PDF-Datei erneut durchsuchen“ im Dropdownmenü „Datenverbindung“ können wir die zu betrachtenden Seiten neu auswählen. Auf dieser Seite gibt es nur eine Tabelle, links hingegen stehen drei Optionen zur Wahl. Tableau hat 3 mögliche Methoden erkannt, um diese Tabelle zu importieren. Wenn wir jede einzeln herausziehen, können wir ihren Inhalt sehen. (Anmerkung: Wenn wir Tabellen von verschiedenen Seiten zusammenführen, wie wir es vorher gemacht haben, muss unbedingt die korrekte Tabellenversion von jeder Seite zusammengeführt werden – und nicht mehrere Versionen derselben Seite.)

Wir ziehen Tabelle 1 heraus und verwenden den Dateninterpreter: Es scheint, als würde Tabelle 1 alle unsere Daten enthalten, doch aus irgendeinem Grund werden die Jahre 1995 bis 1997 als eine einzige Spalte gelesen. Wenn wir diese Version der Daten verwenden möchten, könnten wir dieses Problem bereinigen, indem wird die Option

„Benutzerdefiniertes Teilen“ verwenden und alle Spalten an Leerzeichen teilen und die geteilten Felder dann in die passenden Jahreszahlen umbenennen.

Aber sehen wir uns jetzt an, wie die Tabellen 2 und 3 aussehen. Tabelle 2 sieht wie der untere Teil der Originaltabelle aus. Und Tabelle 3 sieht wie der obere Teil aus. Mir gefällt die Spaltentrennung in den Tabellen 2 und 3 besser, daher werden wir damit arbeiten. Zuerst werde ich die Tabellen zusammenführen, indem ich Tabelle 2 unter Tabelle 3 ziehe. Es besteht eine Diskrepanz zwischen „Inflows“ (Zuflüssen) und F1. Wenn wir beide Felder auswählen und die Option „Nicht übereinstimmende Felder zusammenführen“ verwenden, erhalten wir die erwartete Spalte. Wir können dieses Feld in „Water Source“ (Wasserquelle) umbenennen.

Umgang mit Null-Werten

Es gibt mehrere Zeilen mit Null-Werten. Das liegt entweder daran, dass ein Untertitel wie „Change in storage“ (Änderung im Speicher) als Datenzeile gelesen wurde, oder daran, dass eine einzelne Datenzeile wie zum Beispiel „Abstraction from hydroelectricity“ (Entnahme aus Wasserkraftanlage) als 2 verschiedene Zeilen gelesen wurde. Um diese Null-Werte zu entfernen, fügen wir einen Datenquellenfilter aus der oberen rechten Ecke hinzu. Klicken Sie auf „Hinzufügen“. Wir fügen einen Filter hinzu. Dazu wählen wir irgendeine der Spalten aus, die diese Null-Werte enthält. Ich werde F10 verwenden, auf „OK“ und auf „Weiter“ klicken und dann die Optionen „Null“ und „Ausschließen“ auswählen. Klicken Sie dann auf „OK“ und erneut auf „OK“. Jetzt sind diese Zeilen verschwunden, und es sind nur noch unsere Daten enthalten.

Doch das werde ich wieder rückgängig machen, weil wir auch sehen, dass es unterschiedliche Typen von Wasserquellen gibt, bei denen es sich eigentlich um Gesamtwerte handelt. Wir möchten diese Gesamtwerte und die Null-Werte in einem Durchgang entfernen. Wir gehen zu „Filter“ > „Hinzufügen“ und wählen die Option „Water Source“ (Wasserquelle). Wir aktivieren alle Optionen, die eine Kopfzeile oder einen Gesamtwert aus der ursprünglichen PDF-Datei bzw. eine Null-Zeile für die Werte darstellen. „Hydroelectricity“ (Wasserkraft) ist zweimal vorhanden, einmal mit Daten und einmal mit Null-Werten, was wir zunächst so belassen, um die Daten nicht zu entfernen. Dann klicken wir auf „Ausschließen“ und auf „OK“.

Reparieren von Kopfzeilen und Pivot-Funktion

Jetzt sind die Zeilen mit Daten und eine Zeile mit Null-Werten für „Hydroelectric“ (Wasserkraft) vorhanden, doch das können wir gleich bereinigen. Abgesehen von dieser ersten Zeile gibt es keine Kopfzeilen, aber wir können die Werte zur ursprünglichen PDF-Datei querverweisen, um zu erkennen, was die einzelnen Werte darstellen. Die Spalten sollten nur die Jahre von 1995 bis 2010 darstellen. Ich beschleunige das mal.

Jetzt können wir unsere Daten so drehen, dass eine Spalte für „Year“ (Jahr) und eine Spalte für „Million Cubic Meters“ (Millionen Kubikmeter) vorhanden ist. Wir blenden die Spalte mit dem Tabellennamen aus und ändern den Datentyp für „Year“ (Jahr) in ein Datum. Ändern Sie den Datentyp für „Million Cubic Meters“ (Millionen Kubikmeter) in Ganzzahlen. Da die Daten jetzt im endgültigen Spaltenformat vorliegen, können wir jegliche Null-Werte aus dieser Spalte hier herausfiltern. Wir gehen zu „Filter bearbeiten“ > „Hinzufügen“, wählen die Option „Million Cubic Meters“ (Millionen Kubikmeter) und dann „Null“ > „Ausschließen“.

Neubenennen der Elemente eines Feldes

Mit den Namen der Wasserquellen selbst gibt es noch immer einige Probleme. Wir öffnen das Menü für dieses Feld und klicken auf „Aliasse“. Hier können wir die Elemente eines bestimmten Feldes neu benennen. Wir doppelklicken auf das Alias und geben den gewünschten neuen Namen ein.

- ***Discharge from hydroelectricity generation*** *(Abfluss aus Wasserkraftanlage).*
- *„Groundwater“ (Grundwasser).*
- ***Abstraction for hydroelectricity*** *(Entnahme für Wasserkraft).*

Wiederherstellen von Gruppierungen und Hierarchien

Wenn ich mich recht entsinne, war in der ursprünglichen PDF-Tabelle eine Struktur vorhanden – es gab Kategorien von Wasserquellen. Wir können diese Struktur im Datenbereich nachbilden, deshalb klicken wir uns jetzt zu Blatt 1 durch. Zunächst einmal ist die Option „Million Cubic Meters“ (Millionen Kubikmeter) eigentlich eine Kennzahl, deshalb werden wir sie nach hier unten ziehen.

Dann können wir unsere Gruppen wiederherstellen. Wir ziehen die Wasserquellen in „Zeilen“ und erstellen daraus die erste Gruppe. Klicken Sie bei gedrückter Control-/Strg-Taste auf die Elemente, die zur selben Gruppe gehören sollen:

- *Abstraction for hydroelectricity (Entnahme für Wasserkraft).*
- *Discharge from hydroelectricity generation (Abfluss aus Wasserkraftanlage).*
- *Evapotranspiration*
- *To sea and net abstraction (Ins Meer und Entnahme aus dem Netz)*

Wir verwenden das Büroklammersymbol in der QuickInfo, um die Gruppierung vorzunehmen. Wir wiederholen diesen Schritt, um weitere Elemente zu gruppieren: *„Groundwater“ (Grundwasser), „Ice“ (Eis), „Lakes and reservoirs“ (Seen und Stauseen), „Snow“ (Schnee) und „Soil moisture“ (Bodenfeuchte).*

Jetzt klicken wir mit der rechten Maustaste auf dieses neue Feld und bearbeiten die Gruppe. Als Erstes klicken wir auf „Precipitation“ (Niederschlag) und machen daraus eine eigene Gruppe, obwohl es nur ein einzelnes Element ist. Jetzt können wir die Gruppen umbenennen: *„Inflows“ (Zuflüsse), „Change in Storage“ (Änderung des Speichers), „Outflows“ (Abflüsse).* Und wir nennen das neu gruppierte Feld „Kategorien“. Wenn wir die ursprünglichen „Water Sources“ (Wasserquellen) über dieses Kategorienfeld ziehen, können wir eine Hierarchie erstellen und ein Drilldown in der Ansicht ermöglichen.

Tipps zum Arbeiten mit PDF-Dateien

Die genauen Schritte zur Bereinigung einer PDF-Datei nach dem Herstellen der Verbindung können variieren, doch dieses Video hat Ihnen hoffentlich einige Tools gezeigt, mit denen Sie die Daten so bereinigen können, dass sie für eine Analyse bereit sind. Zur Erinnerung: Tableau bereiten PDF-Dateien Probleme, die Untertabellen, Hierarchien in Kopfzeilen und mehrere Inhaltszeilen enthalten, die als eine einzelne Zeile interpretiert werden sollten. Beachten Sie zum Abschluss auch noch, dass Farben und Schattierungen die Interpretation der Daten verändern können, wegen der Art und Weise, wie PDF-Dateien nach Datenzellen und -tabellen analysiert werden müssen.

Fazit

Vielen Dank, dass Sie sich dieses Schulungsvideo über den PDF Connector angesehen haben. Sehen Sie sich auch unsere anderen kostenfreien Schulungsvideos zur Nutzung von Tableau an.