Conexión a archivos PDF

Bienvenido a este video sobre el conector para PDF. Puede descargar los archivos PDF para seguir la presentación en su propia copia de Tableau.

A veces, los archivos PDF contienen datos valiosos para el análisis en Tableau. El conector para PDF incluido en Tableau está diseñado para ayudar al usuario a extraer datos de tablas. Dado que los archivos de ese tipo carecen de metadatos sobre los datos, es posible que deba hacer algunas tareas después de la conexión para poder realizar su análisis.

Conexión a tablas de datos en archivos PDF

Tableau puede leer tablas de datos en tabulaciones cruzadas de archivos PDF. Lo ideal es que dichas tablas tengan la siguiente apariencia: con columnas y filas, y una única línea de datos en cada fila. Aquí, los datos sobre acciones comienzan en la página 2. Conectémoslos con Tableau. Cuando seleccionamos Archivo PDF en el panel de conexión, aparece un selector de páginas. Podemos examinar todo el documento, una página específica o un intervalo de páginas. Examinaremos las páginas de la 2 a la 8. Y arrastraremos la primera página al lienzo. Una de las líneas del encabezado de tabla del archivo PDF es confusa para Tableau. Si activamos el intérprete de datos, obtenemos los encabezados que esperábamos.

Unión de filas

Cada página apareció como una tabla independiente. Sin embargo, dado que todas tienen la misma estructura y los mismos encabezados de columnas, sus filas se pueden volver a unir con facilidad. Ahora que la primera tabla está en el lienzo, arrastremos las tablas de las demás páginas que nos interesen hasta el área de unión de filas situada debajo de la primera tabla. Obtenemos una nueva columna para el nombre de la tabla. Si nos desplazamos hacia abajo, vemos que los datos unidos de la página 3 se alinean perfectamente con los de la página 2.

Limpieza de tablas imperfectas

Este archivo PDF tenía una buena estructura para su importación, pero no todos los PDF son así. En general, Tableau logra la mejor conexión a los archivos PDF con una fila individual de datos por línea, sin jerarquías ni encabezados anidados y sin subtablas. Sin embargo, no es necesario que esos archivos sean perfectos. Vamos a conectarnos a otro archivo un poco más complejo. La tabla que nos interesa está aquí, en la página 14. (Recuerde: Tableau considera el número de página absoluto, que podría coincidir o no con la paginación del documento).

Agregaremos otra fuente de datos y elegiremos la página 14. La opción “Volver a escanear el archivo PDF…”, en el menú desplegable de la conexión de datos, nos permite volver a elegir las páginas que queremos usar. Solo hay una tabla en esa página, pero aparecen tres opciones a la izquierda. Tableau detecta tres formas posibles de importar esa tabla. Si examinamos una opción a la vez, podemos ver qué contiene cada una. (Recuerde: Al unir tablas de distintas páginas, como hicimos antes, si hay varias versiones de cada página, asegúrese de unir las versiones correctas de las tablas. No una distintas versiones de una misma página).

Examinaremos la tabla 1 y usaremos el intérprete de datos: parece que la tabla 1 contiene toda la información. Sin embargo, por algún motivo, los años entre 1995 y 1997 aparecen en la misma columna. Si deseamos usar esta versión de los datos, podemos limpiarla. Para ello, hagamos una división personalizada separando todas las columnas por espacios. Después, cambiemos el nombre de los campos divididos por los años correspondientes.

Veamos qué ofrecen las tablas 2 y 3. La tabla 2 parece ser la parte inferior de la tabla original. La tabla 3 parece ser la parte superior. Las tablas 2 y 3 tienen una mejor delineación de columnas. Trabajemos con ellas. Primero, unamos esas tablas: arrastremos la tabla 2 debajo de la tabla 3. “Inflows” y la columna F1 no coinciden. Si seleccionamos ambas columnas y elegimos la opción “Fusionar campos no coincidentes”, obtenemos la columna esperada. Cambiemos el nombre por “Water Sources” (Fuentes de agua).

Cómo trabajar con los valores nulos

Hay varias filas con valores nulos. Eso sucede porque algún encabezado secundario, como “Change in storage”, se leyó como una fila de datos, o bien, porque alguna fila individual, como “abstraction from hydroelectricity”, se leyó como dos filas distintas. Para deshacernos de estos valores nulos, agreguemos un filtro de fuentes de datos. En la esquina superior derecha, hagamos clic en Añadir. Agregaremos un filtro. Podemos elegir cualquier columna que tenga valores nulos. Seleccionemos F10 y hagamos clic en Aceptar. Después, seleccionemos Nulo y luego, Excluir. A continuación, hagamos clic en Aceptar dos veces seguidas. Ya no vemos esas filas, solo quedan los datos.

Vamos a deshacerlo, ya que también vemos que hay varios tipos de fuentes de agua que son totales. Vamos a filtrar los valores totales y los nulos a la vez. Vayamos a Filtros, hagamos clic en Añadir y seleccionemos Water Source. Marquemos todos los encabezados y los totales del PDF original, y todo lo que corresponda a una fila de valores nulos. “Hydroelectricity” aparece dos veces, una con datos y otra con valores nulos. Por ahora, lo dejaremos para no eliminar datos. Después, hagamos clic en Excluir y en Aceptar.

Corrección de encabezados y dinamización

Como resultado, quedan las filas con datos y una fila de valores nulos correspondientes a “hydroelectric”. Podemos limpiarla en un momento. No tenemos encabezados, excepto en esta primera columna, pero podemos hacer una referencia cruzada de los valores con el archivo PDF original para saber a qué corresponde cada valor. Las columnas deberían ser los años de 1995 a 2010. Lo haremos rápidamente.

Ahora, podemos dinamizar nuestros datos con una columna de años (Year) y otra de millones de metros cúbicos (Million Cubic Meters). Ocultaremos la columna con el nombre de la tabla y cambiaremos el tipo de datos a Year para indicar una fecha. Cambiemos el tipo de datos de Million Cubic Meters por números enteros y, ahora que tenemos el formato final de las columnas, podemos filtrar los valores nulos de esta columna aquí. Vayamos a Editar filtros, hagamos clic en Añadir, seleccionemos Million Cubic Meters y excluimos los valores nulos.

Identificación de los elementos de un campo

Aún hay problemas con los nombres de las fuentes de agua. Abramos el menú de este campo y hagamos clic en Alias. Aquí podemos cambiar los alias de los elementos de un campo. Haremos doble clic sobre el alias y escribiremos el nombre que debería tener.

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

Recreación de grupos y jerarquías

Por último, recuerdo que la tabla original del archivo PDF tenía una estructura que incluía categorías de fuentes de agua. Podemos crear esa estructura en el panel de datos. Hagamos clic en Hoja 1. Antes que nada, Million Cubic Meters es una medida, así que la arrastramos hasta aquí abajo.

Después, podremos volver a crear los grupos. Llevaremos las fuentes de agua a las filas y crearemos el primer grupo. Presionemos la tecla Control y hagamos clic en los miembros que deben pertenecer al mismo grupo:

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

Usaremos el icono del clip en la descripción emergente para realizar la agrupación. Vamos a repetir y volver a agrupar: *Groundwater, Ice, Lakes and reservoirs, Snow, Soil moisture*

Hagamos clic con el botón secundario en este nuevo campo en el panel de datos y editemos el grupo. Primero, hagamos clic en Precipitation y realicemos la agrupación para darle su propio grupo, aunque sea un único elemento. Ahora, podemos cambiar los nombres de los grupos. *Inflows, Change in Storage, Outflows.* Le daremos el nombre Categories (categorías) al campo agrupado. Si arrastramos el campo de fuentes de agua original y lo soltamos sobre este nuevo campo de categorías, podemos crear una jerarquía y habilitar la exploración de jerarquías en la vista.

Consejos para trabajar con archivos PDF

Los pasos exactos para limpiar un archivo PDF después de conectarse a él pueden variar. Aun así, esperamos que, en este video, haya descubierto algunas herramientas útiles para limpiar sus datos y prepararlos para el análisis. Recuerde que Tableau tendrá problemas con los archivos PDF que contengan subtablas, jerarquías en los encabezados y varias filas de contenido que deberían interpretarse como una fila individual. Por último, tenga en cuenta que los colores y las sombras pueden influir sobre la interpretación de los datos, ya que los archivos PDF deben dividirse en celdas y tablas de datos.

Conclusión

Gracias por ver este video de capacitación sobre el conector para PDF. Lo invitamos a continuar viendo los videos gratuitos de capacitación y obtener más información acerca de Tableau.