连接到 PDF

欢迎观看这段关于“PDF 连接器”的视频。您可以下载 PDF，以便在自己的 Tableau 软件副本中跟着操作。PDF 中可能包含可以在 Tableau 中进行分析的宝贵数据。Tableau 中的 PDF 连接器用于帮助用户拉取表中的数据。由于 PDF 缺少关于数据的元数据，您可能需要在连接后进行一些处理，才能开始分析。

连接到 PDF 数据表

Tableau 可以读取 PDF 中的交叉表数据表，这些数据表的理想状态是由列和行组成，每行仅包含一个数据行。在这里，库存数据从第 2 页开始。让我们使用 Tableau 来进行连接。从连接窗格选择 PDF 文件后，我们会看到一个页面选取器。我们可以扫描整个文档、一个特定页面或者某个范围内的页面。我们要查看 2-8 页。我们要将第一页拖至画布。Tableau 没有正确理解 PDF 表标题中的一行内容，但打开数据解释器后，我们就可以获得预期的标题。

并集

每个页面都作为一个单独的表导入，但由于具有相同的结构和相同的列标题，因此可以利用并集轻松地将它们重新放到一起。第一个表已经显示在画布上，我们要将任何需要的其他页面对应的表拖至并集拖放区域，放在第一个表下方。我们得到一个新的表名称列；向下滚动，可以看到来自第 3 页的并集数据与第 2 页完美对齐。

清理有缺陷的表

此 PDF 的结构非常适合进行导入，但并非所有 PDF 都是如此。一般而言，Tableau 与具有以下表结构的 PDF 的连接效果最好：每行包含单个数据行、无分层结构或嵌套标题、无次级表。但我们可以处理不完美的 PDF。我们要连接到另一个文件，这个文件不像上一个那样容易。我们需要的表格在第 14 页。（注意：Tableau 以绝对页码为准，这不一定与文档中的分页方式相同。）

我们要添加另一个数据源，我们选择第 14 页。我们借助“数据连接”下拉菜单中的“重新扫描 PDF”选项，重新选取要查看的页面。该页面只有一个表，但左侧有三个选项。Tableau 检测到可以通过 3 种方法拉取该表格。通过每次显示一种方法，我们可以看到其中包含的内容。（注意：在像前面一样联合多个页面中的表时，如果每个页面有多个版本，我们需要确保联合正确的表版本，而不是相同页面的多个版本。）

我们显示表 1 并使用数据解释器：表 1 似乎信息齐全，但不知为什么，1995-1997 被读取为同一列。如果要使用这个数据版本，我们可以进行清理，方法是进行自定义拆分，在空格处拆分所有列，然后将拆分得到的字段重命名为正确的年份。

但让我们看看表 2 和 3 包含怎样的内容。表 2 看起来像是原始表的底部。而表 3 看起来像是其顶部。我更喜欢表 2 和表 3 的列界定方法，因此我们要使用这两个表。首先，我们对它们进行联合，将表格 2 拖出，放到表格 3 下方。“Inflows”（流入）和列 F1 不匹配。如果同时选定二者并选择“合并不匹配的字段”，我们就可以得到预期的列。让我们将其重命名为“Water Source”（水源）

处理 Null 值

这里有几行 Null 值，这可能是因为“Change in storage”（存储量变化）之类的次级标题被读取为数据行，或者“abstraction from hydroelectricity”（水电汲取）之类的单个数据行被读取为 2 个不同的行。为了删除这些 Null 值，我们将添加数据源筛选器，在右上角单击“添加”。我们要添加筛选器，可以选择包含这些 Null 值的任何列，我使用 F10，单击“确定”，“下一步”，选择“Null”，然后单击“排除”。然后单击“确定”，再次单击“确定”。现在，这些行已经消失，现在只看到我们的数据。

但我要撤销上述操作，因为我们还看到，有几种类型的水源实际上是合计值。我们想对这些水源类型和 Null 值进行一次性筛除。转到“筛选器”>“添加”，我们选择 Water Source（水源）。我们选中在原始 PDF 中是标题或者合计值的任何内容，以及对应值为 Null 行的任何内容。“Hydroelectricity（水电）”出现了两次，一次包含数据，一次包含 Null 值；我现在不动它，以免删除数据。然后我们将依次单击“排除”、“确定”

修复标题和进行透视

现在，我们的行都包含数据，显示 Null 的 hydroelectric 行是一个例外，但我们马上就可以清除它。除了第一列外，我们没有任何标题，但我们可以对原始 PDF 中的值进行交叉引用，确定各列的名称。这些列应该只是年份，从 1995 到 2010。我要加快速度。

我们现在可以透视数据，我们有一个 Year（年份）列，有一个 Million Cubic Meters（百万立方米）列。我们要隐藏表名称列，并将“Year”（年份）的数据类型更改为日期。将“million cubic meters”（百万立方米）的数据类型更改为“整数”；这是最终的列格式，我们可以筛除该列中的所有 Null 值。我们转到“编辑筛选器”>“添加”，选择“Million cubic meters”（百万立方米），并“排除”Null 值。

更改字段成员的别名

水源名称自身还有一些问题。 我们打开该字段的菜单，单击“别名”。我们可以在此更改给定字段的成员别名。我们双击别名，然后键入正确的别名。

- ***Discharge from hydroelectricity*** *generation*（水电发电释放）
- *Groundwater*（地下水）
- *Abstraction for hydroelectricity*（水电汲取）

重新创建组和分层结构

最后，我记得原始 PDF 表有一个结构 — 有水源类别。我们可以在数据窗格构建此结构，我们单击进入工作表 1。首先，“Million Cubic Meters”（百万立方米）实际上是一个度量，因此我们可以将其拖至此处。

随后我们可以创建组。我们将水源拖至行，然后创建第一个组。按住 Control，单击应该属于同一个组的成员：

- *Abstraction for hydroelectricity（水电汲取）*
- *Discharge from hydroelectricity generation（水电发电释放）和*

- *Evapotranspiration（蒸发蒸腾）*
- *To sea and net abstraction（入海和净汲取）*

我们使用工具提示中的别针图标来进行分组。重复并再次分组：“Groundwater”（地下水）、“Ice”（冰）、“Lakes and reservoirs”（湖泊和蓄水池）、“Snow”（雪）、“Soil moisture”（土壤水分）

我们在数据窗格中右键单击这个新字段并编辑该组。首先，我们单击“Precipitation”（降水）并将其分为一组，虽然只有一个成员，我们仍然让它成为单独的组。现在可以对它们进行重命名：*“Inflows”（流入）*、*“Change in Storage”（存储量变化）*、*“Outflows”（流出）*。我们将该分组字段命名为“类别”。通过将原始水源拖放至这个类别字段的上方，我们可以创建一个分层结构，在视图中实现下钻查询。

关于处理 PDF 的提示

连接后的确切清理步骤可能会因具体的 PDF 而异，希望您通过这段视频，了解了一些可以用于清理数据，使其适合分析的工具。我要提醒您，Tableau 在处理包含以下元素的 PDF 时会遇到困难：次级表、标题中的分层结构、应该解读为单行的多行内容。最后请注意，颜色和阴影可能改变数据的解读方式，因为 PDF 必须解析为单元格和数据表。

结语

感谢您观看“PDF 连接器”培训视频。我们邀请您继续观看免费培训视频，进一步了解如何使用 Tableau。