O'REILLY®

Tableau Prep Up & Running

Self Service Data Preparation for Better Analysis

Early Release
RAW & UNEDITED

Carl Allchin

Tableau Prep

Up and Running

Carl Allchin

Tableau Prep: Up & Running

by Carl Allchin

Printed in the United States of America.

Published by O’Reilly Media, Inc. , 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O’Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (http://oreilly.com). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com .

Editors: Angela Rufino and Michelle Smith

Production Editor: Daniel Elfanbaum

Interior Designer: David Futato

Cover Designer: Karen Montgomery

Illustrator: Rebecca Demarest

July 2021: First Edition

Revision History for the First Edition

- 2020-03-20: First Release

See http://oreilly.com/catalog/errata.csp?isbn=9781492079620 for release details.

978-1-492-07955-2

Chapter 1. How to explain, “Why Self Service Data Prep?”

With every organization swimming in data lakes, repositories, and warehouses, never before have people in organizations had such an enormous opportunity to answer their questions with information rather than just using their experience and gut instinct.

This isn’t that different from where organizations stood a decade ago, or even longer. What has changed is who wants access to that data to answer their questions. No longer is the expectation that a separate function of the business will be responsible for getting that data, now everyone feels they should have access to the data. So what has changed? Self Service Data Visualization. What is about to change to take this to the next level? Self Service Data Preparation.

A Short History of Self Service Data Visualization

More than a decade ago, all things data related were the domain of specialist teams. Data projects were either reporting requests that went to specialist Business Intelligence (BI) teams or Information Technology (IT) teams to set up data infrastructure projects to produce reports from. This was expensive, time consuming, and often resulted in products that were less than ideal for all concerned.

The reason this methodology doesn't work is the iterative nature of BI. Humans are fundamentally intelligent creatures who like to explore, learn, and then ask more questions because they are intrigued. With the traditional IT or BI projects, once the first piece of analysis was delivered, the project was over. However, if one question was answered, others were triggered but as the skills were in different hands to those who have the questions, they simply went unanswered. The business users still tried to cobble together the answers but they were from disparate reports or different levels of aggregation.

This all changed with the rise of Self Service Data Visualization tools like Tableau Desktop. Suddenly with a focus on the user, individuals were able to drag and drop the data fields around the screen to form their own analysis, answer their own questions, and ask their next questions straight away. The previous decade has seen data visualization and analysis become closer to everyone's role, and a significant part of many roles that are now not considered Information Technology roles or part of the data team. The analytical

capacity has come to the business, rather than the business having to go and ask specialists to get the data. This represents a big transformation in how we work and poses a challenge as to what skills people now require.

Accessing the ‘Right Data’

The rise, and entrenchment, of self service data visualization into individuals’ roles has raised other needs and tensions in the analytical cycle. To enable self service, access to data sources has become the next pain point in this cycle. With the right data, optimized for the use in the tools that empower the visual analysis, answers can be found at the speed that the business expert can form the questions. But accessing the ‘right data’ is not that easy. The data assets that have been formed by organizations have been optimized for storage, optimized for tools that now seem to work against the user rather than with them, and are held behind strict security layers to handle greater regulation.

Many data projects are now focused on extracting data from their storage locations. The specialist skills are focused on using data skills to:

- Find data in existing repositories
- Find data in public or third party repositories
- Create feeds of data from previously inaccessible sources / systems

The gap in the process now sits between taking these sources and making them ready for visual analytics.

The Self Service Data Preparation Opportunity

This gap is being challenged by new tools that are allowing the business experts, using self service visual analytics to solve their questions, to access this data. Tableau Prep Builder (Figure 1-1) has brought the same logic that empowered visual analytics to this data preparation step. By using a similar user interface to the one that data visualizers are already accustomed to, Prep Builder has made the transition to self service data preparation a simple one.

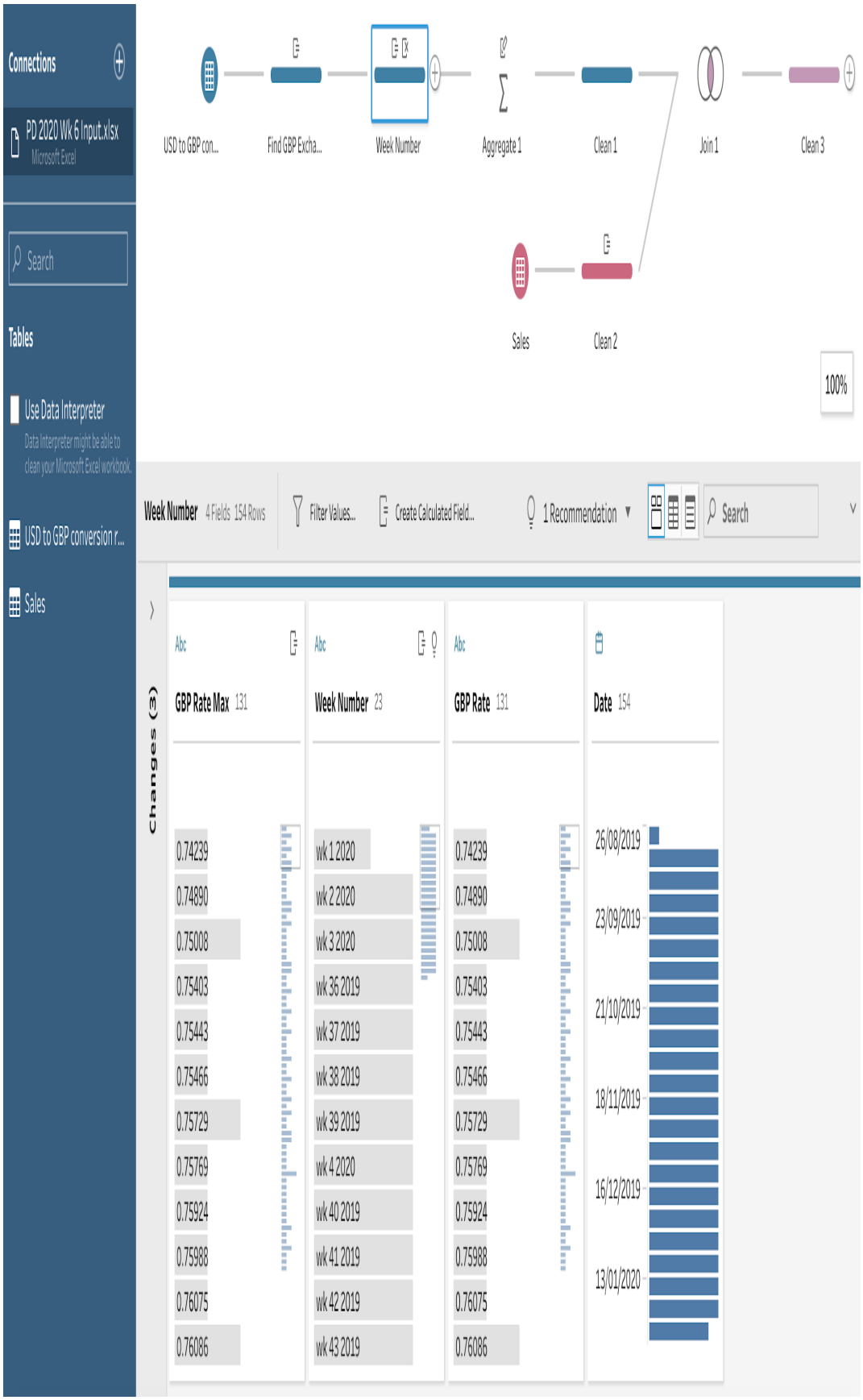

Connections
PD 2020 Wk 6 Input.xlsx
Microsoft Excel
Search
Tables
Use Data Interpreter
Data Interpreter might be able to clean your Microsoft Excel workbook.
USD to GBP conversion r...
Sales
USD to GBP con...
Find GBP Excha...
Week Number
Aggregate 1
Clean 1
Join 1
Clean 3
Sales
Clean 2
100%
Week Number 4 Fields 154 Rows
Filter Values...
Create Calculated Field...
1 Recommendation
Search
Changes (3)
Abc
GBP Rate Max 131
0.74239
0.74890
0.75008
0.75403
0.75443
0.75466
0.75729
0.75769
0.75924
0.75988
0.76075
0.76086
Abc
Week Number 23
wk 1 2020
wk 2 2020
wk 3 2020
wk 36 2019
wk 37 2019
wk 38 2019
wk 39 2019
wk 4 2020
wk 40 2019
wk 41 2019
wk 42 2019
wk 43 2019
Abc
GBP Rate 131
0.74239
0.74890
0.75008
0.75403
0.75443
0.75466
0.75729
0.75769
0.75924
0.75988
0.76075
0.76086
Date 154
26/08/2019
23/09/2019
21/10/2019
18/11/2019
16/12/2019
13/01/2020

Figure 1-1. The Tableau Prep Builder interface

To date my career has taken me through all of the roles in the traditional analytics cycle:

- From a user who went through the pain of waiting for reports to be built for them
- To learning how to build the analysis for themselves
- To wrangle the data sets in databases
- And, eventually to be a trainer of the skills for both data visualization and data preparation

There is still a significant gap between all potential data preparers (we will nickname these individuals as “data preppers”) and those that have the requisite skills needed. The awareness of what to do with the Self Service Data Preparation tools and why they are needed is a significant gap to fill but one that is a worthwhile investment in time.

Tableau Prep Up & Running

This book is designed for exactly this gap. Learning how to utilize the tools needed to tackle the tasks that are acting as the current roadblocks in delivering answers to our questions. The challenges will introduce commonly used techniques to solve these problems away from the pressure of the workplace. Over time, strategies can be formed to prepare the data exactly how you want it whether it comes from files, databases, surveys, pivot tables, messy data, or tangled

text fields. There isn’t a straightforward recipe to follow but through practicing, you’ll soon be able to handle these challenges.

Now you hopefully have a view of why Self Service Data Preparation is required and why learning the techniques we will cover will assist you in your work with data. So how should you approach those challenges and where should you start? This is what will be covered in the next chapter.

Chapter 2. How to Plan Your Prep

> **A NOTE FOR EARLY RELEASE READERS**
>
> With Early Release ebooks, you get books in their earliest form—the author's raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.
>
> This will be the 2nd chapter of the final book.
>
> If you have comments about how we might improve the content and/or examples in this book, or if you notice missing material within this chapter, please reach out to the author at PreppinData@gmail.com.

So you know your data isn't right for the purpose you need it to be; but what do you do about it? The next stage in this process after this realization is key in developing a solution to your challenge but how do you approach this next step when all you see are the issues in front of you?

Our approach to this challenge is a staged approach that should help you plan, define the outcome, and provide a framework of steps to solve your challenges.

To cover this, let's talk through an example data set, but keep it simple (Figure 2-1). Here's some sales data from Chin & Beard Suds Co (a fake retailer that sells soap).

Branch	Product	01 Jan 2019	01 Feb 2019	01 Mar 2019	01 Apr 2019
Wimbledon	Liquid-Soap	173	344	427	470
Wimbledon	Soap-Bar	708	652	377	305
Lewisham_1	Liquid-Soap	276.31	804.22	655.94	789.63
Lewisham_1	Soap-Bar	359.12	647.53	401.26	291.99
Lewisham_2	Liquid-Soap	643	555	686	661
Lewisham_2	Soap-Bar	323	727	810	504

Figure 2-1. Sample data from Chin & Beard Suds Co.

Stage 1 - Know Your Data (KYD)

Without understanding your dataset as it currently stands, you will not be able to deliver the results you need. Sometimes in small datasets, this understanding can be very easy to form. With larger datasets, this process can take longer but it is arguably more important to plan as you can only hold so much information in your own memory. Here's what to look for in datasets (using the example in Figure 2-1):

Columns, rows, and crosstabs - how is the data structured?

> Two columns of dimensions with each other column being a month of values

Headers and fields - are they all there as expected?

> Month headers should ideally be one column with a value alongside

Data Types - what type of data exists in each field?

> Two string values but then each subsequent column is a numeric value

Granularity of Rows - what does each row represent?

> Each row is a different product sold in a store

Nulls - are there any?

None present so can disregard this factor for this example

We often find that a quick sketch of the table will help you think through what is going on in each aspect of your table and apply the questions above. Here's how we'd sketch the data source above (Figure 2-2):

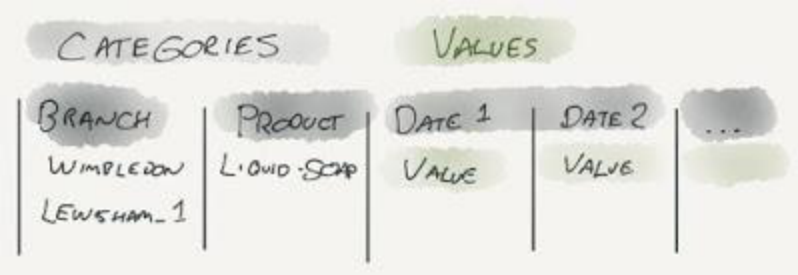


Figure 2-2. Sketch of dataset highlighting Categories and Values

By identifying categorical data and the fields that contain the values you will want for your analysis, you will understand: how complete the data set is, why it isn't ready for analysis yet, and what it might take to prepare the data. Notice, it's not vital that everything is captured. By simplifying the sketch, you start to focus on the core issues rather than drowning in the details.

Stage 2 - The Desired State

As you become more experienced in producing your own data sets, this stage becomes a lot easier and almost second nature. You will be able to look at a data set and know what the desired outcome should be. Whilst you are learning that, this sometimes feels difficult to picture - so sketch it! More on that shortly!

So What Should Your Desired State Be?

For most modern data analysis tools, or visualization software, you will need to structure your data into columns that have a name for that data in the first row. Each subsequent row should be an individual recording / instance. Most tools will require a single input table so all of your data fields you need for your analysis should be in this single table.

The sketch - how to form it

For simple data sets this can be very easy to just doodle out. For wide data sets with lots of columns, this stage can become a bigger task. Start by listing off your categorical data. These are things that you will analyze the numbers by, for example, region, customer, product or course. Each unique combination of these categorical data fields will set the granularity of your Desired State data set. In the example dataset, the combination of Branch, Product, and Date will set the value being recorded.

Secondly, you can now add a column for each of the measures you may want to analyze. At this point you will also want to consider values that might help the analytical tools' performance. Maybe adding a ranking, market share, or running total at this stage will help make your dataset more accessible to novice users but also performant to complete your analysis faster.

Returning to our example, here's what we need to output to complete our analysis #fig_3__sketch_showing_the_structure_of_the_dataset:

Figure 2-3. Sketch showing the structure of the dataset

The Categorical data fields will be Branch, Product, and Date. Note how capturing the data type here is useful too. The output will only have one field for analysis - a simple sales value.

Stage 3 - From Know Your Data to the Desired State

Take your hands away from the keyboard and mouse… for this stage you won't need a computer, just your brain. By looking at the original data and having understood what you want the data to be, you will start to see some of the transitions you will need to take to clean, pivot, join, and aggregate the data as required.

Start by making a list of the transitions you think you will need to make. You may not end up doing them all but you're not building the workflow yet so it doesn't matter. Here are some of the questions we ask at this stage:

Columns:

Too many?

Remove unnecessary fields.

Too few?

Maybe *join* a secondary data set.

Clean Field Names ?

If not, you will need to amend these.

Calculations needed?

If you have all the data fields you need to form the new columns then calculations will be required. By completing your calculations in your *preparation* tool, your analysis tool will require less computing power to focus on forming the data and require less skills for the end user.

Rows:

Too many?

Filter out unnecessary rows. Aggregate the data to be less granular.

Too few?

Pivot columns to create more. Add an additional data set through *unioning* or join a data scaffold.

Clean records? Clumped data? Punctuation where there shouldn't be any?

You will need to handle these different challenges but take note of each change as they will likely be separate data prep steps

Blanks? Should they be there?

Will need to filter them if not or find a way to fill the gaps otherwise.

Multiple data sources:

- Join together to add more columns.
- Union together to add more rows.

Let's return to our example and see how this works for this use case (Figure 2-4):

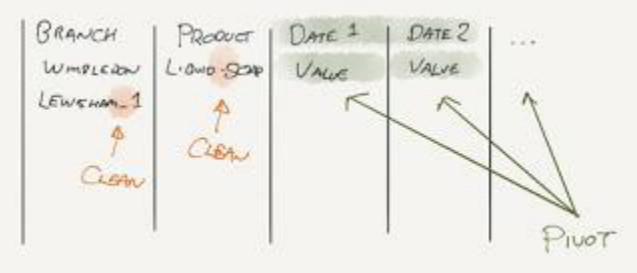


Figure 2-4. Sketch planning the transitions required in the sample dataset

Very quickly we can spot the challenges by taking a few quick notes on my sketch of 'frame' of the data. You might not spot all of the challenges but this stage just gets you thinking about what you have to work with. Here we can see:

- We have two categorical data fields (Branch and Product).
- The rest of the columns we have are all headed by dates and have the sales values inside.

  The Date needs to be pivoted to form our third categorical field. This will change the number of rows of data we have in the data set as we will get an additional row per: month, branch, and product combination.
- The pivoting process will create a column of values.

- Punctuation replacing the space in the Branch and Product fields will need to be removed.

For the rows of data, the analysis requires a row per unique combination of branch (town), product, and month. This means we will need to:

- Aggregate the Lewisham_1 and Lewisham_2 sales together to form the data at the correct granularity. This will also change the number of rows we may have in the Desired State data set compared to the Original Data set.

Here's some other ideas on what other steps we also might need to take:

- Rename fields
- Change data types of fields

Stage 4 - Build It

Ok, you can use the mouse and keyboard now to build out each of those steps. Where do you start? Well, making a basic step-by-step plan would be a good way to go. With Tableau Prep, you can quickly change the order of the transitions or add forgotten ones to go from the original data to the Desired State. You might not get the workflow right the first time but you will be a lot closer for having planned out these steps.

Also, you might not know what tool, transition, or calculation to use to make the change you require, but you will be able to take a step

back to rethink the problem and not get lost from the next steps when you do so.

So let's complete our example. With all the steps captured above, I used Tableau Prep Builder to create this workflow from Inputting the data, to Outputting the file as a csv (Figure 2-5). The first step (icon) in the image below is the 'Input' step where the data is being brought into Prep for processing. The second icon is pivoting multiple columns of data to rows of data instead. This is where all of the different dates are being converted into one column with another column alongside holding the respective value for that date, branch and product.

Figure 2-5. The Tableau Prep Flow from Input to Output

Although in Tableau Prep Builder, the clean step (the third icon) actually contains a lot of detail that is captured in the tools 'Changes pane'. The image below (Figure 2-6) shows not only the steps taken but also what occurred:

- Formed a calculation to change the Product from 'Liquid-Soap' to 'Liquid' and 'Soap-Bar' to 'Bar'.
- Grouping two Lewisham stores together.
- Change the Pivoted field name to Date.
- Change the column now called Date to a date data type.
- Rename the Pivoted Values columns to Sales.

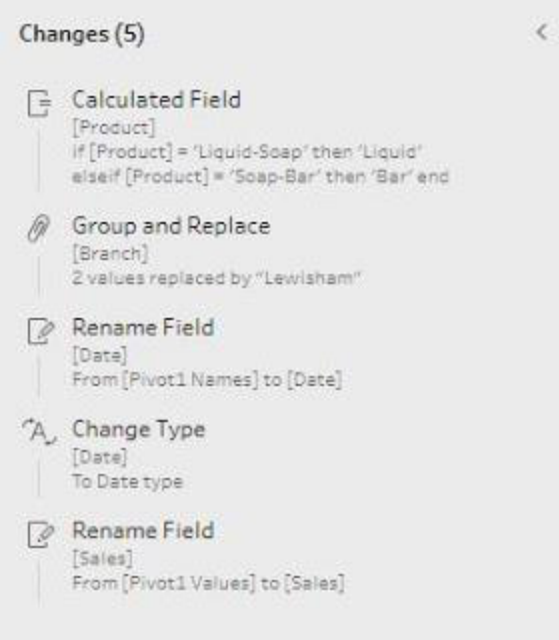


Figure 2-6. The changes pane for the Clean step in Figure 2-5

This leaves us with a nice clean dataset ready for analysis when outputted in the fourth and final step (Figure 2-7):

Branch	Product	Date	Sales
Wimbledon	Liquid	01/03/2019	427
Wimbledon	Bar	01/03/2019	377
Lewisham	Liquid	01/03/2019	655.94
Lewisham	Bar	01/03/2019	401.26
Lewisham	Liquid	01/03/2019	686
Lewisham	Bar	01/03/2019	810
Wimbledon	Liquid	01/04/2019	470
Wimbledon	Bar	01/04/2019	305
Lewisham	Liquid	01/04/2019	789.63
Lewisham	Bar	01/04/2019	291.99
Lewisham	Liquid	01/04/2019	661
Lewisham	Bar	01/04/2019	504
Wimbledon	Liquid	01/02/2019	344
Wimbledon	Bar	01/02/2019	652
Lewisham	Liquid	01/02/2019	804.22
Lewisham	Bar	01/02/2019	647.53
Lewisham	Liquid	01/02/2019	555
Lewisham	Bar	01/02/2019	727
Wimbledon	Liquid	01/01/2019	173
Wimbledon	Bar	01/01/2019	708
Lewisham	Liquid	01/01/2019	276.31
Lewisham	Bar	01/01/2019	355.12
Lewisham	Liquid	01/01/2019	643
Lewisham	Bar	01/01/2019	323

Figure 2-7. The restructured data

Click here to access all the files used if you want to have a go at the exercise yourself.

By Planning your data preparation, you will not just be more focused on the task but also give yourself a solid basis to work from. Some techniques are challenging, especially when you apply them for the first time. Knowing how those techniques are applied and what they should lead to will make your preparation much more likely to be successful. Input and Output datasets can be large and complex, this means the planning may require a significant investment of time before you start making progress on manipulating the data. It would be only normal to want to dive in but the planning effort will save that time in the long term by reducing the risk of going off on tangents or missing key stages.

Exercises

The following exercise was featured on my *Preppin' Data* blog and designed to allow people to practice the techniques discussed in this chapter. In the exercise, you will be able to read about the intention on the exercise and the requirements for the challenge, like they were given to you as a request by someone. The Input and Output datasets are given to allow you to try to meet the challenge set in the exercise. Solutions are found on the blog for you to check what you came up with, but there is no right or wrong solution if you have delivered the output requested.

- **2019 Week 33**

Chapter 3. How to Shape Data

A NOTE FOR EARLY RELEASE READERS

With Early Release ebooks, you get books in their earliest form—the author's raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.

This will be the 3rd chapter of the final book.

If you have comments about how we might improve the content and/or examples in this book, or if you notice missing material within this chapter, please reach out to the author at PreppinData@gmail.com.

As discussed in the previous chapter 'How to...Plan your Prep', the first step of Data Preparation is understanding how the original dataset is structured, and quickly followed by understanding what structure the dataset needs to be in for analysis. This chapter looks at exactly these factors, so when you look at future datasets you can quickly determine what steps you want to take to shape the data for analysis.

What to look at for incoming datasets?

Let's look at a typical input dataset that has been formed within Excel by building a Pivot Table. This example uses Ice Cream sales (Figure 3-1).

Category	Measure	Jan-2019	Feb-2019	Mar-2019	Apr-2019	May-2019	Jun-2019
Mint Choc Chip	Sales	6933	9895	1871	4649	6492	8956
Mint Choc Chip	Profit	965	353	357	469	525	996
Strawberry	Sales	3832	2512	4738	8254	3816	4109
Strawberry	Profit	574	775	715	523	119	949
Vanilla	Sales	5206	1440	6397	2299	1178	7046
Vanilla	Profit	174	921	346	675	968	968

Figure 3-1. Ice Cream Sales and Profit data

When assessing an incoming dataset, it's important to identify both Dimensions of the data as well as Measures. Dimensions are the columns of data that describe the records. For example, the dimensions of the dataset are the Regions a product is sold in or which Category that product belongs to. By Measure, we mean the numeric values of the dataset that are being analysed. Measures might include the number of students in a class at college or the tuition price they are being charged. If the Dimensions are all in individual columns then you can move on to Measures without having to think about any structure changes. The same assessment needs to be made for Measures.

The dataset has been colored (Figure 3-2) to highlight the structure found:

Category	Measure	Jan-2019	Feb-2019	Mar-2019	Apr-2019	May-2019	Jun-2019
Mint Choc Chip	Sales	6933	9895	1871	4649	6492	8956
Mint Choc Chip	Profit	965	353	357	469	525	996
Strawberry	Sales	3832	2512	4738	8254	3816	4109
Strawberry	Profit	574	775	715	523	119	949
Vanilla	Sales	5206	1440	6397	2299	1178	7046
Vanilla	Profit	174	921	346	675	968	968

Figure 3-2. Dataset after being broken into Dimensions and Measures

- Dark Blue - Header for Dimension column
- Light Blue - Dimension value
- Dark Green - Header for Measure column
- Light Green - Measure value

By drawing out the structure of the input dataset, it becomes clear what alterations need to be made once you have an understanding on how data should be structured for analysis.

What Shape is Best for Analysis in Tableau?

When loading data into Tableau Desktop, the software takes the first row of data as the headers for the columns and all subsequent rows as the data points for those headers. Here are the key aspects to consider when structuring data for Tableau:

A single column for each data field

> These will form the data fields that are then dragged & dropped in Tableau.

Is it a Dimension or Measure?

> Tableau will divide all data fields into Dimensions (aspect to split the data up by) and Measures (the data fields to analyse).

A single data type for each data field

> A data field in Tableau (and most other tools) require a single data type. For measures, if they are not numeric, they will not be present in the list of measures.

There are a few aspects that do not matter as well:

Order of columns

Tableau Desktop and Server will absorb a dataset and order the fields shown in the Data Pane in Alphabetical Order. Therefore, there is no need to order the columns.

Order of the rows

You will analyse the rows through the visualizations you build. The charts and graphs can be sorted but the rows of data won't be shown in the order they are absorbed into the tool so therefore, you don't need to worry about the order of the rows in the data source you are forming.

Geographic Roles

Roles that are commonly set in Tableau Desktop to add longitude, latitude, and spatial shapes can not be set and then carried over into Tableau Desktop from Prep. The data field can be set as a String but the Geographic Role will need to be allocated in Desktop. The data fields that you may want to use as Geographic fields should be cleaned but no further actions need to be taken.

In the example above (Figure 3-2), we can see Category is in the correct state. The Header for the Dimension is at the top of the column that contains all the relevant values.

The Measure Header is in the correct location, but is it necessary? The Measures are listed under each individual month. To analyse data over time in Tableau, one column containing all of the different dates would be more preferable and easier to use. Therefore, in this case, the Dates listed as headers at the moment will need to be pivoted.

The Measures that are named in the Measure Column would be much easier to analyse if they were individual columns. Forming one

column for Sales and one for Profit would enable these two columns to be Measures when analysing the data in Tableau.

Changing Dataset Structures in Prep

Of the different types of steps available within Prep, there are three that are key to changing the structure of the input datasets:

Pivot

The Pivot step is the most important when changing the data structure. There are two types of Pivot (Figure 3-3):

Figure 3-3. Column to Rows Pivot and Rows to Columns Pivot step icons

Columns to Rows

> Taking multiple columns and converting them into additional rows of data. The column header is made into a new dimensional column that will contain all other column headers that are involved in the Pivot. The icon for the ‘Columns to Rows’ Pivot is the dark purple icon shown above on the left.

Rows to Columns

> This is the reverse of the Columns to Rows pivot. In this instance, rows of data are converted in additional columns within the dataset. The set-up of the ‘Rows to Columns’ pivot requires a selection of the column that will become the Headers of the new data fields. The set-up also requires a selection of the data field

that will act as the values for the new data fields. In the case that there are multiple values forced into the same cell, a form of aggregation has to be chosen and applied to them.

Aggregate

Whilst aggregations change the number of rows, the Aggregation step (Figure 3-4) within Prep also can change the structure of the data. The only data fields that continue on in the data flow from an Aggregate step are any fields included as a ‘Group By’ field or an *Aggregation*:

Figure 3-4. Aggregation Step icon

Join

Joins (Figure 3-5) are designed to add additional columns to the original dataset from an additional data source(s). Depending on the Join Type and Join Conditions set, the resulting data set can be very different in terms of the number of data fields, as well as number of rows.

Figure 3-5. the Join Step icon showing an Inner Join

Union

This step (Figure 3-6) can create a different data structure as Unioning mismatched column headers will create a wider dataset. Without merging the mismatched fields, a large number of nulls will be present in any fields that are not contained in both datasets.

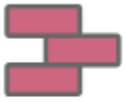


Figure 3-6. the Union step icon

Applying the Techniques to the Example

Let's apply the techniques available in Prep for restructuring data to the Ice Cream example.

Step A: Pivot - Columns to Rows

Creating a single column to contain dates is a key step to take. By taking this step, 'Pivot1 Names' is created as a column to hold the former column headers that are selected within the 'Columns to Rows' step (Figure 3-7):

Category	Measure	Pivot1 Names	Pivot1 Values
Mint Choc Chip	Sales	Jan-2019	6933
Mint Choc Chip	Profit	Jan-2019	965
Strawberry	Sales	Jan-2019	3832
Strawberry	Profit	Jan-2019	574
Vanilla	Sales	Jan-2019	5206
Vanilla	Profit	Jan-2019	174
Mint Choc Chip	Sales	Feb-2019	9895
Mint Choc Chip	Profit	Feb-2019	353
Strawberry	Sales	Feb-2019	2512

Figure 3-7. The result of the Columns to Rows Pivot

Step B: Pivot - Rows to Columns

To create a column for each measure, the ‘Measure’ column needs to be converted into Headers for the new column with the relevant value added underneath for each combination of ‘Category’ and ‘Date’ (Figure 3-8):

Category	Date	Sales	Profit
Mint Choc Chip	Jan-2019	6933	965
Strawberry	Jan-2019	3832	574
Vanilla	Jan-2019	5206	174
Mint Choc Chip	Feb-2019	9895	353
Strawberry	Feb-2019	2512	775
Vanilla	Feb-2019	1440	921
Mint Choc Chip	Mar-2019	1871	357

Figure 3-8. The result of the Rows to Columns Pivot

Restructuring data is a key skill to master and practice as the dataset above is much easier to analyse. However, longer and thinner is not always the intention. A few key elements to aim for are:

- A single data field for each dimension, ideally containing a single data type (like a string or date)
- A single column for a type of date rather than each week / month / year being a separate column
- A column per measure

The closer this data structure is matched, the easier it will be to analyse the data set in question flexibly no matter what the subject of the data is. Following the steps seen above will enable you to reshape the data, but planning what those steps are at the start of the process will make this process much easier.

Chapter 4. How to breakdown Complex Data Preparation Challenges

In previous chapters, techniques for determining the alterations required for preparing a dataset for analysis have been discussed, albeit at a relatively simple level. What about those situations where the challenge isn't straight forward; how do we approach the challenge then? This chapter will cover this exact scenario by taking on one of the most complicated challenges Preppin' Data has covered to date.

Where to begin?

The challenge involves taking NBA results and forming the full conference league tables, including rankings, wins & loses, recent performance, and even winning streaks.

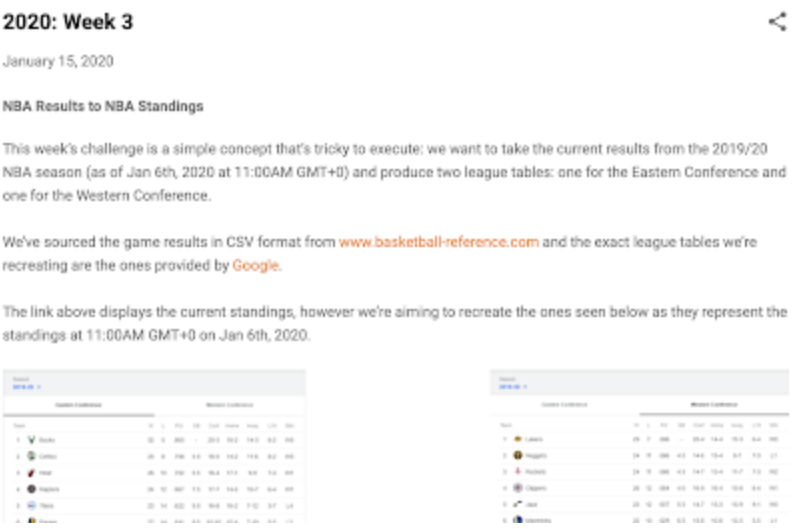


Figure 4-1. Challenge Post for Preppin' Data 2020 Week 3

As covered in Chapter 2, forming an understanding of the input and output will start to allow you to form an overview of the task. Especially on large, complex challenges, planning becomes more important to ensure your are working towards the desired results.

Initial Scoping of the Challenge

Here's my initial scope for the 2020 Week 3 challenge with a few key elements that are reminders from Chapter 2 (Figure 4-2):

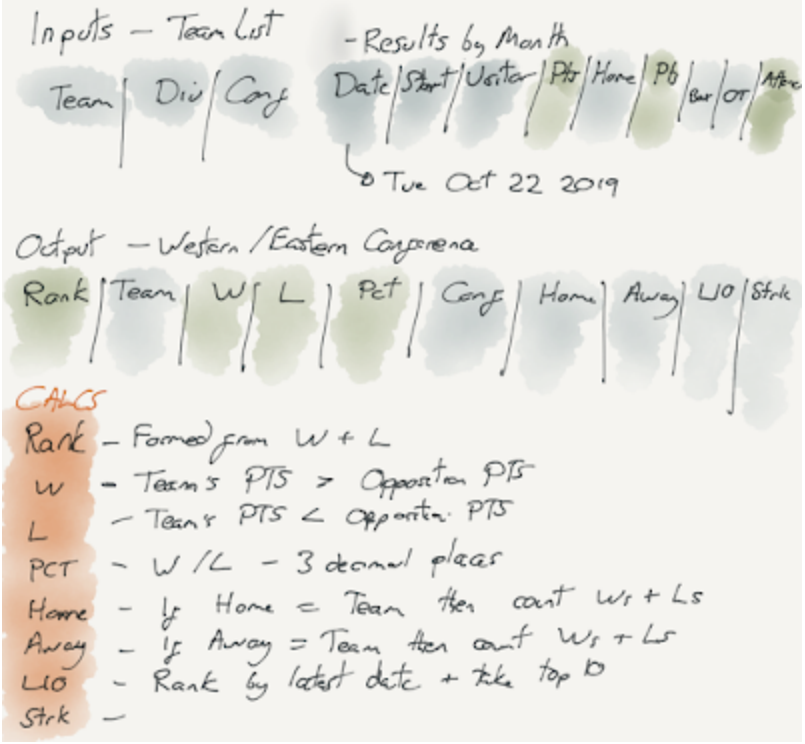


Figure 4-2. Sketched plan to solve the challenge

Map out your inputs - what does each input contain? What dimensions and measures are there in the data? Any data fields that are not in the format you require? What level of granularity does the data exist in?

> In this case, only the ‘Date’ field seems to be something that may need cleaning.

Map your outputs - How many output files will be required? What format will you need certain fields to be? Think about the granularity of the data required.

> In this case each time will have a single row within the output. This means there will be a lot of aggregation to take the game results to the level required in the output.

Understand the gaps - what fields do not exist within the data and how are they to be formed?

> This will create a list of fields that need to be created through either pivoting, joins or calculations.

At this stage, it isn't imperative that you solve all the issues that stand in the way from the input data to the output dataset that will be used for analysis. As discussed in Chapter 2, you might not spot all of the challenges in the dataset at this stage, but by working with the data, those challenges will emerge.

Logical Steps

Breaking the challenge down to individual chunks makes the overall challenge a lot easier to work out how best to solve it. Without doing so, the overall challenge may be insurmountable. The formation of calculations you know that you need is a simple step to take as they will help:

- Guide what steps you need to take to be able to find a solution
- Give an order in which to do things

Let's take 'Wins' as an example in this challenge. As you have individual game results, you need to determine who the winner is. This is easier said than done as for each 'win' there is a loss as well. Therefore, two rows are required for each game, one to record the winner, one the loser. To ensure I capture every team's games, I first use the 'Team List' data source and join all game results on to that twice, once matching the home team and again to match the away team to the original team list team name (Figure 4-3).

Figure 4-3. Join set-up to create a set of results per team for home and away games

Using two calculations then allows me to form that team's points (Figure 4-4) and the opposition's points (Figure 4-5) for the home games:

Figure 4-4. Calculation forming Own Score by returning the home team's points value

Figure 4-5. Calculation forming the Opposition's Score by returning the away team's points value

These calculations can then be repeated but this time testing whether the team list team name is the away team. Figure 4-7 is how to calculate the Away Team's score where the focus team plays away. A second calculation can form the oppositions score as per Figure 4-5:

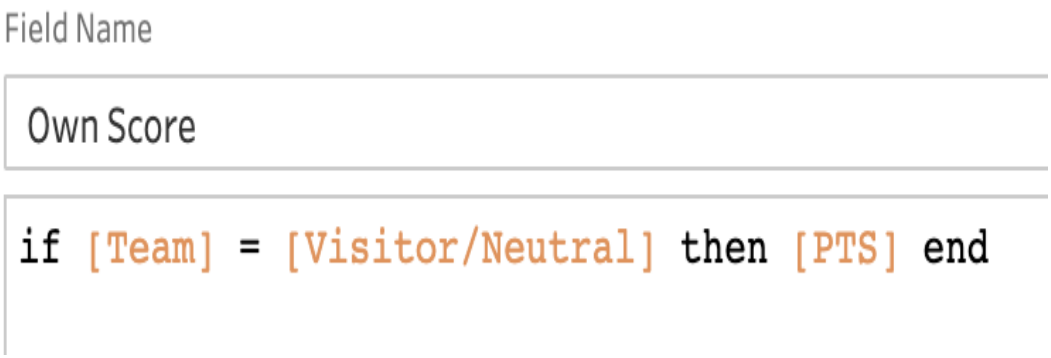


Figure 4-6. Calculation forming the Away team's own score from the second Join step in Figure 4-3

These calculations can then be assessed for who won or lost the game (Figure 4-7). The same approach can then be taken for games in which the team played away and the results Unioned together.

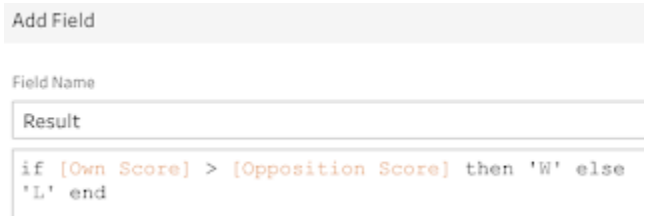


Figure 4-7. Calculation forming the Result of each game

The two flows can then be Unioned together to create one large dataset to record each team's full set of games in the season (Figure 4-8):

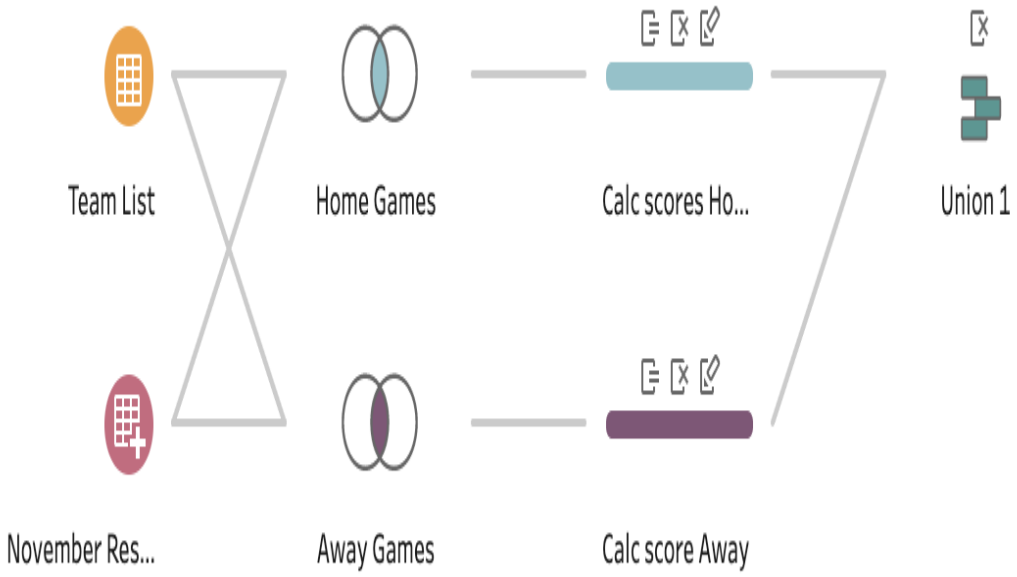


Figure 4-8. First stage of Preppin' Data 2020 Week 3 challenge

Making Changes

This approach can then be repeated to ensure you are tackling each of those calculations in turn. As you determine a solution for each sub-challenge, you may need to change the ordering of the steps, or copy and paste entire sections. This is easily done. By right-clicking on the linking line between two steps, you can delete it and then drag the step from the 'pre-step' to the 'Add' part of the step you wish to link it to in your flow (Figure 4-9)

Figure 4-9. Reconnecting the steps

Be Ready to Iterate

Often only by working with the data will the solution come to fruition as otherwise it might be difficult to imagine how exactly the data will behave during the transformational steps you are making.

For example, when forming the W, L, Home, and Away columns, I knew they would involve similar calculations but wasn't sure which order I would handle them in. These columns represent the total wins and losses as well as the record of wins and losses achieved in home and away games. Here is the flow of how I approached this task (Figure 4-10):

Figure 4-10. Flow to form Wins, Loses for Home and Away Games

To work through this sub-challenge, the primary data point to form for each team was whether the team won or lost a game. I already had captured whether the team at the focus of the game had scored more points than the opposition. If they had scored more, I had formed a column of 'W', otherwise I returned a column of 'L'. This logic was correct but wasn't ideal for aggregating to create all the relevant

totals of Wins and Losses. Therefore, I pivoted this column to create a single column of both wins and losses and to make the next counts simple, I had created a simple calculation of 1 to add in to each win or loss column depending on the result (Figure 4-11).

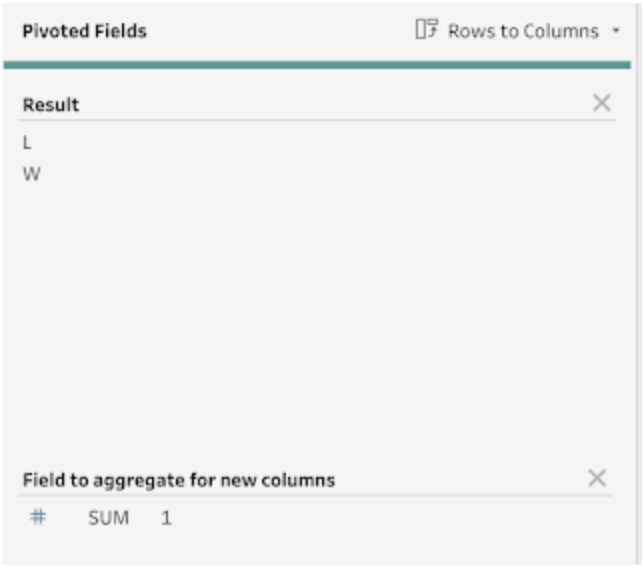


Figure 4-11. Pivot to make Win / Loss counts

Knowing that the next steps would require an Aggregation step, I began to recognise the process would need to be split at this point into two streams. The Aggregation step only returns the aggregated values and the 'Group By' data fields (to understand the Aggregation step in Prep read this 'How to... Aggregate' chapter). Even with very careful planning, it would be unlikely that I would have thought about taking the step to add an additional branch to the flow at this point. Prep is a fantastic tool for being agile once you prove that your logic has differed from your initial thoughts. As I needed to aggregate at an overall level as well as splitting out Home and Away records, I needed two separate Aggregate steps (Figure 4-12).

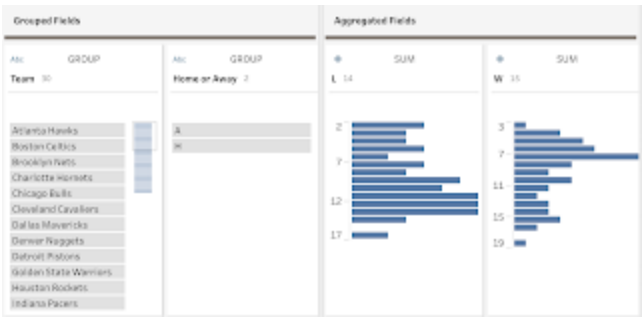


Figure 4-12. Aggregating Wins and Losses columns

These two flows were eventually joined back together to form one overall view of the team's records at the end of the flow shown above.

In summary, the power of being able to iterate your approach is key in data preparation as often there are cleaning issues you didn't spot in your initial approach or your attempted solution didn't work as intended. Enjoying the challenge of data preparation comes down to being able to focus on problem-solving. This is much easier if you break the challenges down into more manageable chunks.

Exercises

The following exercises were featured on my *Preppin' Data* blog and designed to allow people to practice the techniques discussed in this chapter. In the exercises, you will be able to read about the intention on the exercise and the requirements for the challenge, like they were given to you as a request by someone. The Input and Output datasets are given to allow you to try to meet the challenge set in the exercise. Solutions are found on the blog for you to check what you came up with, but there is no right or wrong solution if you have delivered the output requested.

- 2019 Week 17
- 2020 Week 3

Chapter 5. How to Not Need Data Prep At All

A NOTE FOR EARLY RELEASE READERS

With Early Release ebooks, you get books in their earliest form—the author's raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.

This will be the 5th chapter of the final book.

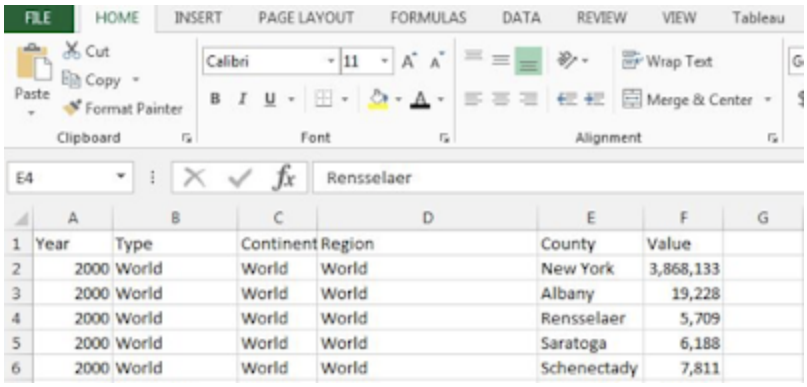
If you have comments about how we might improve the content and/or examples in this book, or if you notice missing material within this chapter, please reach out to the author at PreppinData@gmail.com.

Sometimes we all overthink problems and this happens a lot in Data Preparation. Once you have some strong skills, it can be easy to overthink the issue and take additional steps you simply don't need. This chapter will cover when the data visualization tool can be used to complete the data preparation process and what situations you may want to assess to determine if that is the correct choice.

History of Data Preparation in Tableau

Back in 2010, when trying Tableau Public for the first time in version 5.2 and becoming a heavy user of Desktop version 7 onwards, completing your data preparation in Tableau was tough unless you were doing very simple use cases. Tableau Prep was just a glint in the eye of the Tableau Developers at that point.

Common data prep tasks had to happen outside of Tableau Desktop for a number of reasons, like performance (slower processing speeds), but it was the lack of functionality that caused the most significant issues. There was often the need to use external tools to complete tasks. The Tableau Excel Add-In Extension (see the top right of Figure 5-1) was the principle way to deal with pivoting data fields to change columns into rows of data. This was very useful for dealing with survey data and where dates were held in separate columns (i.e. not Tableau friendly).



	A	B	C	D	E	F	G
1	Year	Type	Continent	Region	County	Value	
2	2000	World	World	World	New York	3,868,133	
3	2000	World	World	World	Albany	19,228	
4	2000	World	World	World	Rensselaer	5,709	
5	2000	World	World	World	Saratoga	6,188	
6	2000	World	World	World	Schenectady	7,811	

Figure 5-1. The Tableau Excel Add-In in Excel's Menu

Complex joins, removing data for better performance, and unions were not part of the data connection window in Desktop as it is now. This meant that to use data sources, the preparation would have to happen in the database the data originated in.

As Tableau's development aim has always been to keep the user in the flow of their analysis, they have built more preparation functions into the core tool, Tableau Desktop. The challenge with this was that all the data preparation happened within the Data Connection window and space was becoming limited. The complexity of the challenge could not be kept within Tableau's usual simple User Interface until

version 2018 changed that approach by spinning out Tableau Prep Builder as a separate tool.

But what should you still attempt in Tableau Desktop compared to using Tableau Prep?

Simple Joins

Joins have been a feature of data connections in Desktop for years (Figure 5-2). Yes, Tableau Prep can of course handle these but by keeping the join in Desktop, you can flexibly change the join type and conditions whenever you want. With the addition of Join Calculations, most situations where you need additional columns from additional data sources can, and should, be handled in Desktop.

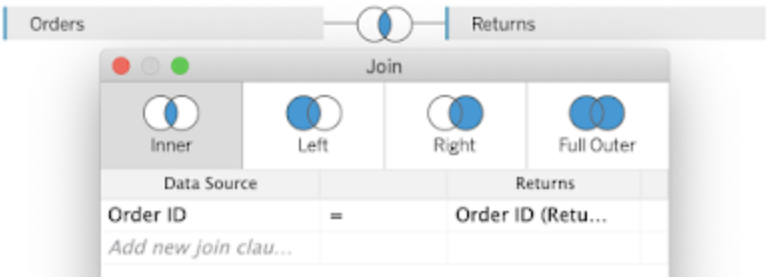


Figure 5-2. Join set-up in Tableau Desktop

WHEN TO MOVE SIMPLE JOINS TO PREP?

When making multiple, different joins, things can get confusing in Desktop. With Prep's profile pane, it's much easier to see when join conditions have gone wrong or created a lot of nulls that you were not expecting. Also, look to use Prep when you need to change the level of aggregation of one of your data sources before joining the data sets together.

Unions

Like Joins, when Unions were added to the Tableau Desktop Connection Window (Figure 5-3), the number of times data prep had to be completed outside of Tableau Desktop reduced dramatically. Unions in Desktop are quite flexible. So much so, the options for Unions within Prep are very similar.

Figure 5-3. Union Iconography in Tableau Desktop

A basic Union can be added in Desktop by dragging the additional dataset underneath the original connection (Figure 5-4).

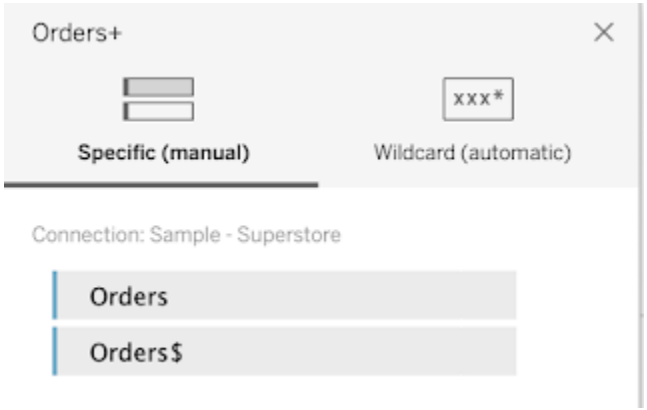


Figure 5-4. Union set-up in Tableau Desktop

Once the union is formed, there is the option to edit the Union or add additional data sources to that union. More complex unions can be completed through using Wildcard Unions (Figure 5-5). There's more explanation on Wildcard Unions in this How To... Union chapter.

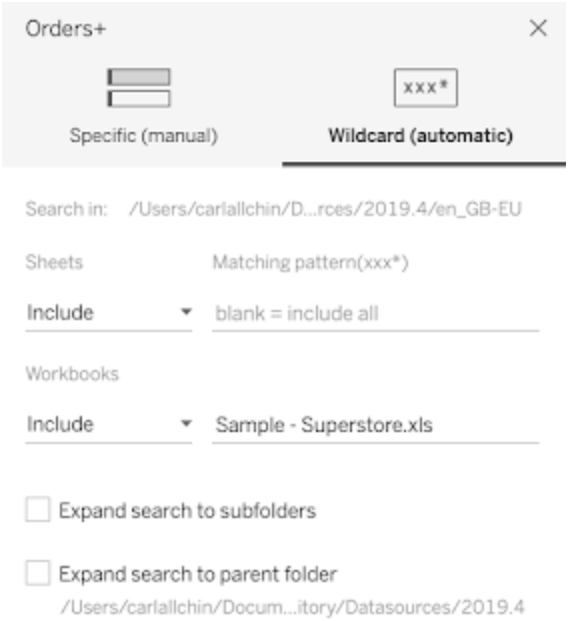


Figure 5-5. Tableau Prep Wildcard Union configuration pane

Similar to Prep, the Union can be wildcarded at the Sheet or Workbook level when connecting to Excel sheets. Different data types have different options at this point so planning what data you need and where it is is key. Being able to explore the effects of the unioned data straight away in Desktop is a reason to start in the tool unless you need to do other data preparation steps.

WHEN TO MOVE UNIONS TO PREP?

When unioning data with different structures (column names), the resulting data set from the union will lead to a lot of ‘null’ values. Being able to deal with those nulls before progressing to your analysis can be complex so the calculations to clean this up are important.

Single Pivots

When pivoting was added to Desktop (Figure 5-6), the need to use external tools like the Excel Add-In for Tableau was considerably reduced. Survey data has long been a considerable challenge in Desktop but being able to pivot a column, or set of columns, meant that external tools were no longer the only option to prepare and shape the data set.

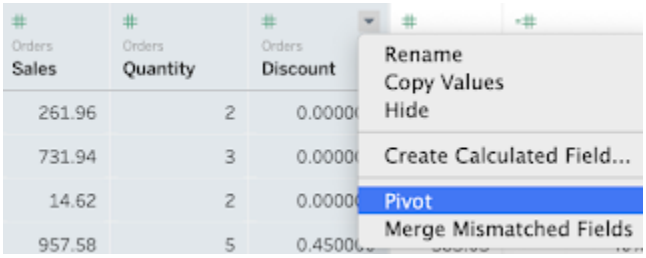


Figure 5-6. Pivoting selected Columns in Desktop

The resulting data field names are far from useful but are easy to adjust (Figure 5-7).

Abc Pivot Pivot Field Names	# Pivot Pivot Field Values
Discount	0.000000
Discount	0.000000
Discount	0.000000

Figure 5-7. Data Field Names resulting from the Pivot in Desktop

WHEN TO MOVE SINGLE PIVOTS TO PREP?

The limitation of pivoting within Desktop is that you can only pivot once. Many data sets are much more complicated than this, especially that pesky survey data.

Review/Handover

Tableau Prep is a piece of software that allows a lot of traceability. By this we mean it is largely self-documenting as you build the data preparation flow. The use of iconography, the Changes Pane, and ability to step through the changes one-by-one allows anyone to retrace the journey from input to output and why the stages are likely to exist. With renamed steps and added descriptions, the work is a lot easier to understand than a laboriously written out project handover Word document that many of us have had to consume over the years.

WHEN TO MOVE REVIEW/HANDOVER TO PREP?

Do the data preparation in Prep and save yourself the time in writing the documentation for handover separately.

Closing Summary

All these steps in isolation have been possible within Desktop but the combination of these techniques and others result in the need for Tableau Prep. Also handing over complex workbooks with multiple stages of data preparation in Desktop is far from ideal for the recipient so again, passing this work to Prep is useful outside of just handling the more complex functionality that will feature in additional ‘How to...’s.