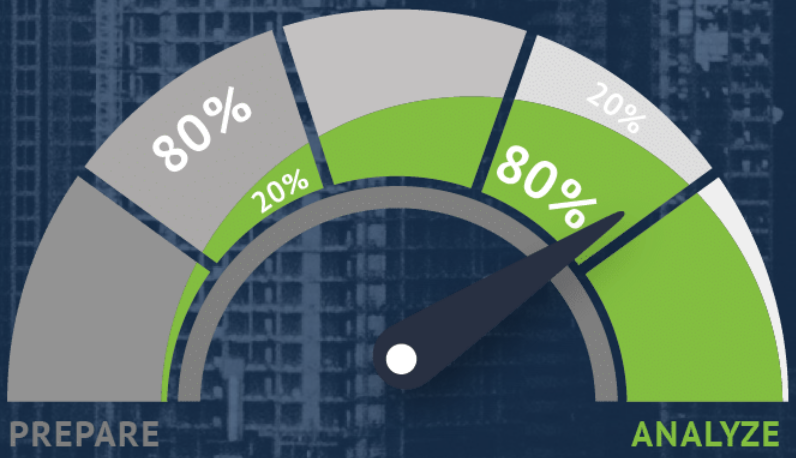
TABLEAU PREP TRAINING HANDBOOK



Written by:
Parker Nokes & Ryan Nokes

Table of Contents

1. INTRODUCTION TO TABLEAU PREP 2

1.1. FEATURES OF TABLEAU PREP 2

2. TABLEAU PREP BASICS 7

2.1. STARTING TABLEAU PREP 7
2.1. CONNECTING TO YOUR DATA 10
2.3. CONNECTING TO MULTIPLE FILES 11
2.4. USING THE UNION STEP 12
2.5. REUSING STEPS IN YOUR FLOW 15
2.6. EXPLORING YOUR DATA 18
2.7. SAVING YOUR WORK 22

3. CLEANING YOUR DATA 23

3.1. GROUPING AND REPLACING 25
3.2. USING A CALCULATED FIELD 28
3.3. CHANGING THE DATA TYPE 32

4. JOINING YOUR DATA 33

4.1. TYPES OF JOINS 33
4.2. ADDING A BRANCH 38
4.3. ADDING A PIVOT 38
4.4 REARRANGING FLOWS 43

5. GENERATING YOUR OUTPUT 43

5.1. PREVIEW YOUR DATA 43
5.2. FINALIZING YOUR FLOW 43
5.3. AGGREGATION 44
5.4. PUBLISHING YOUR OUTPUT 46

6. ADVANCED FEATURES 48

6.1. ADVANCED PIVOTING: ROWS TO COLUMNS PIVOT 48
6.2. PIVOTING ROWS TO COLUMNS WITH CATEGORICAL DATA: 53
6.3. USING THE WILD CARD SEARCH 54
6.4. ADDING DESCRIPTIONS TO FLOWS 57
6.5. IDENTIFYING DATA GRANULARITY AND LEVEL OF DETAIL CALCULATIONS: 59
6.6 USING RANK AND ROW_NUMBER CALCULATIONS 64
6.7 AUTOMATING TABLEAU PREP 70

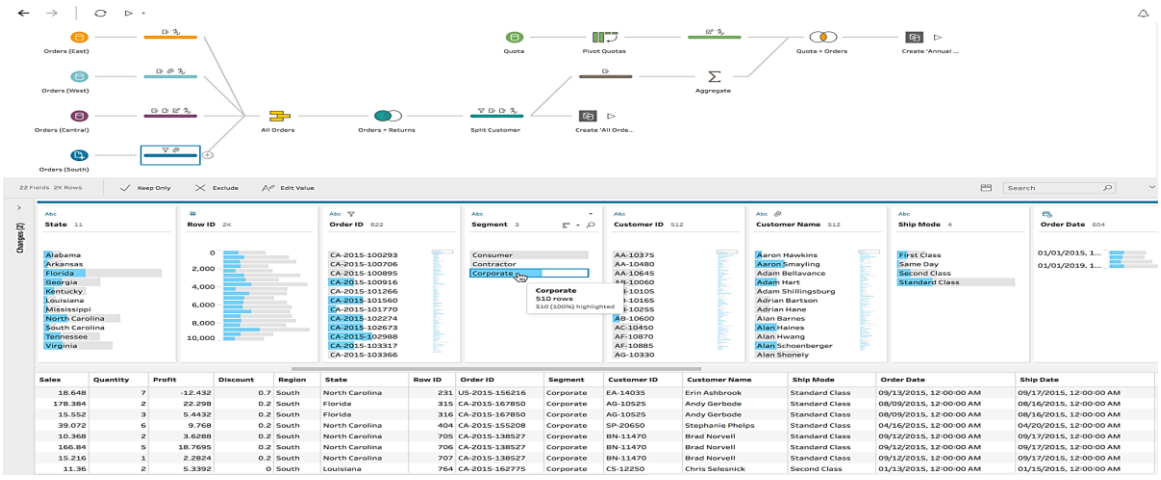
1. INTRODUCTION TO TABLEAU PREP

Tableau prep is a new tool introduced by Tableau to allow you the ability to gather, combine, structure and organize data for analysis within Tableau. Data preparation is a crucial step in the data analysis process to ensure that you use accurate data for your analysis that supports high level decision making.

1.1. Features of Tableau Prep

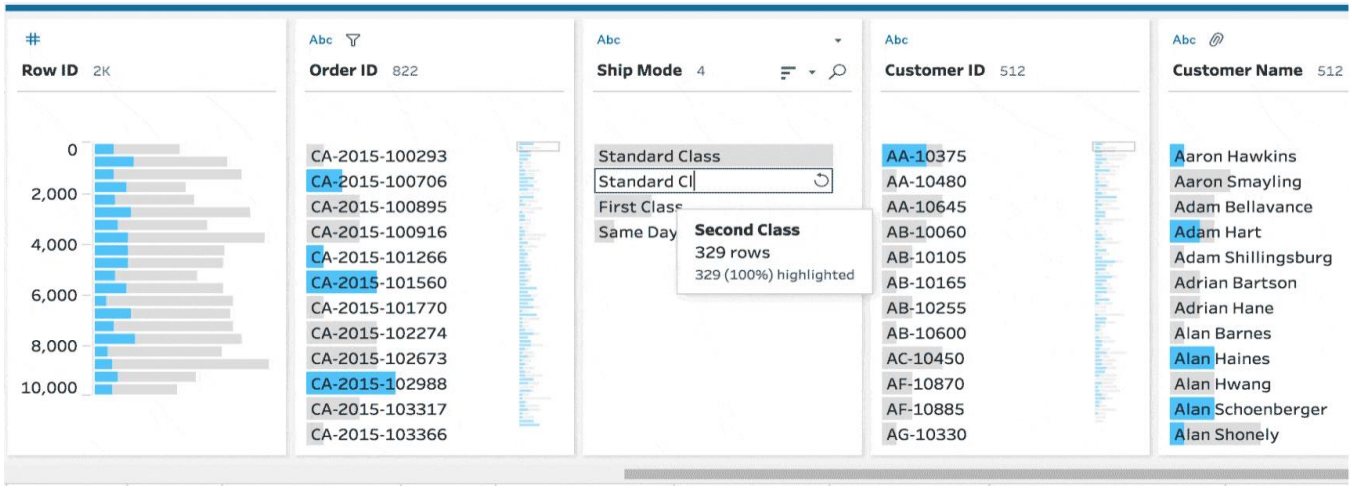
1. Data Overview

Tableau prep provides you with an overview of your row-level data, table columns and the entire preparation process.



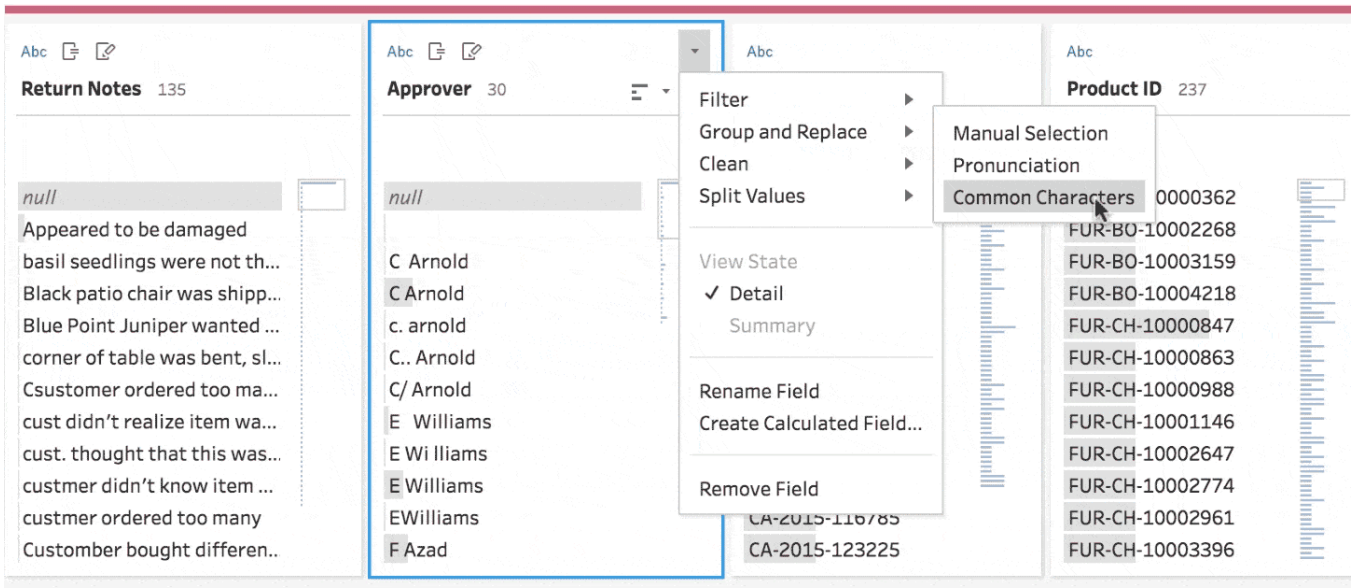
2. Instant results

Tableau prep gives you the ability to see the results of your data preparation process in real-time. This means that you can directly edit a process, join types, data types etc and instantly see the changes in the data.



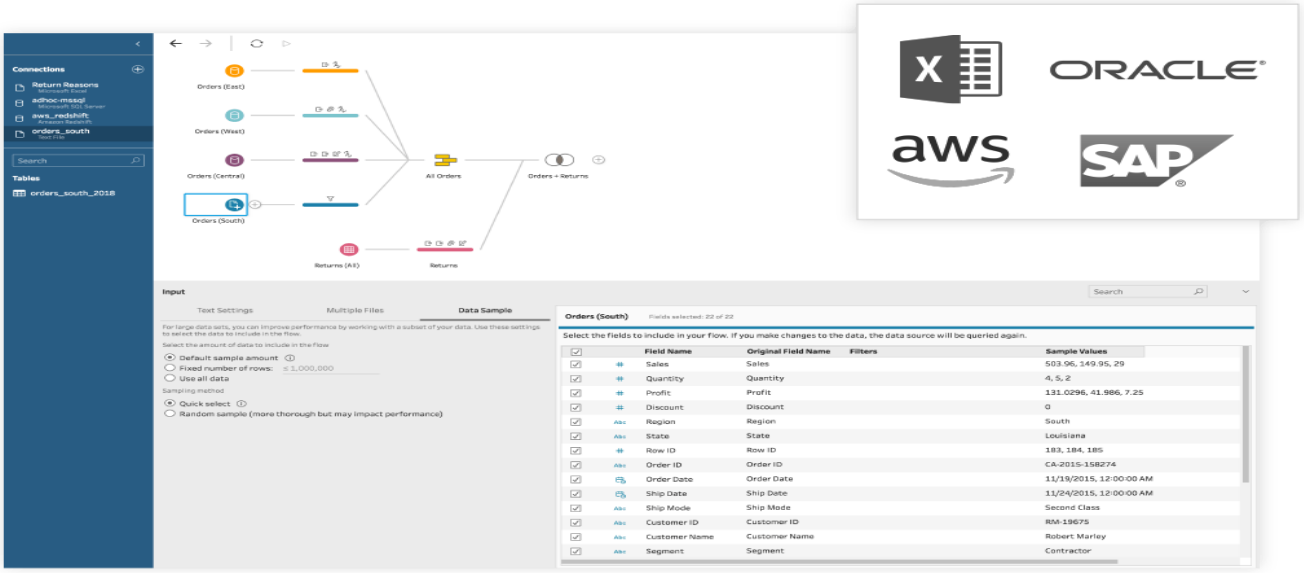
3. Smart features

The Tableau prep tool incorporates certain smart features that helps you to solve common data prep challenges or repetitive tasks to allow you to work faster with your data.



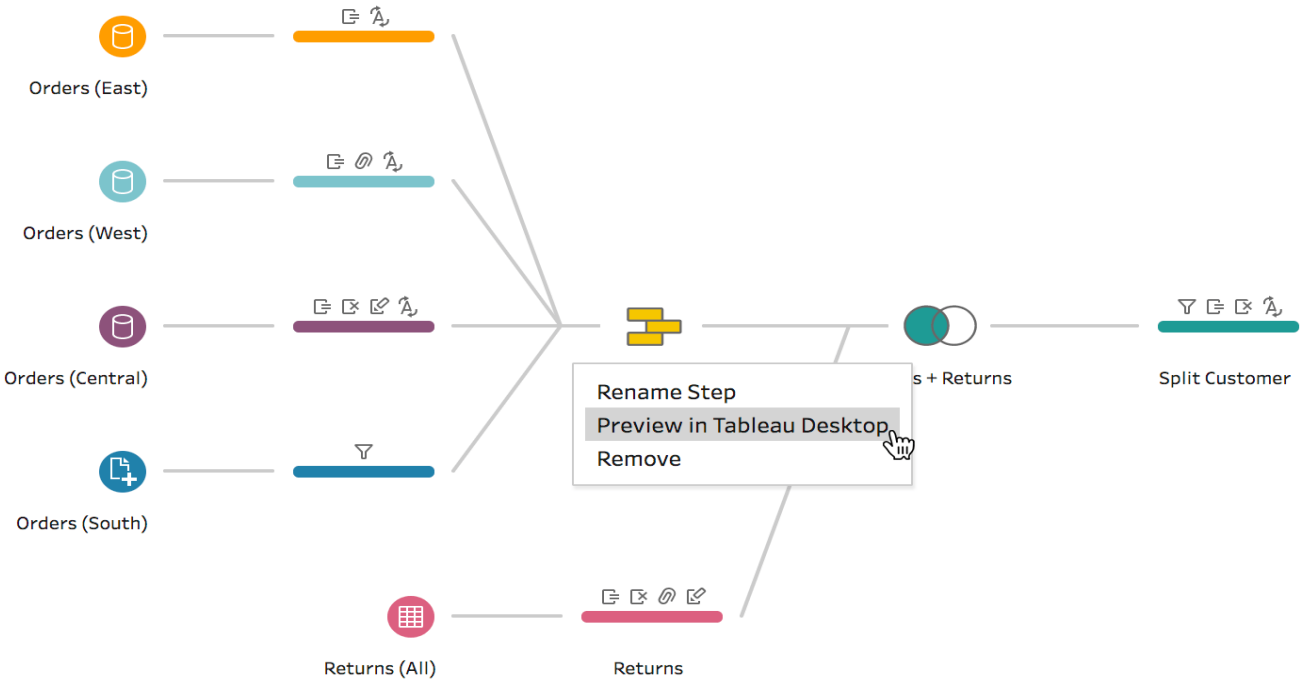
4. Connection to multiple data sources

With Tableau prep you can connect to many different data sources sitting in your database or in a worksheet. It easy to connect to your data just as you would with the Tableau desktop tool.


Connections
Return Reasons
adhoc-mssql
aws_redshift
orders_south
Tables
orders_south_2018
Orders (East)
Orders (West)
Orders (Central)
Orders (South)
All Orders
Orders + Returns
Returns (All)
Returns
ORACLE
aws
SAP
Input
Text Settings
Multiple Files
Data Sample
Default sample amount
Fixed number of rows:
Use all data
Quick select
Random sample (more thorough but may impact performance)
Orders (South)
Select the fields to include in your flow. If you make changes to the data, the data source will be queried again.
Field Name
Original Field Name
Filters
Sample Values
Sales
Quantity
Profit
Discount
Region
State
Row ID
Order ID
Order Date
Ship Date
Ship Mode
Customer ID
Customer Name
Segment

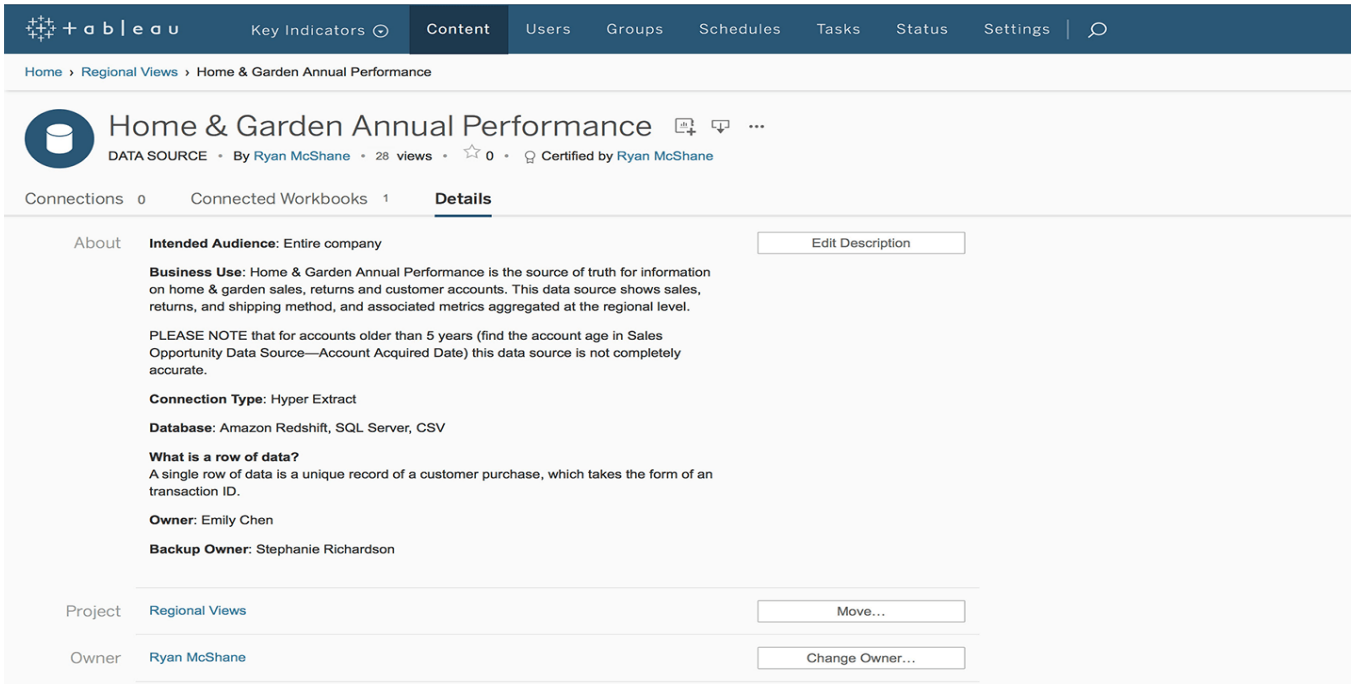
5. Workflows

If you feel like you want to re-use your data prep process with another data source, Tableau prep allows you to easily save workflows, open them with Tableau Desktop or share them with others.



6. Collaborate and Share data securely

With easy integration with Tableau desktop and server, you can share your governed data securely with others in your organization.

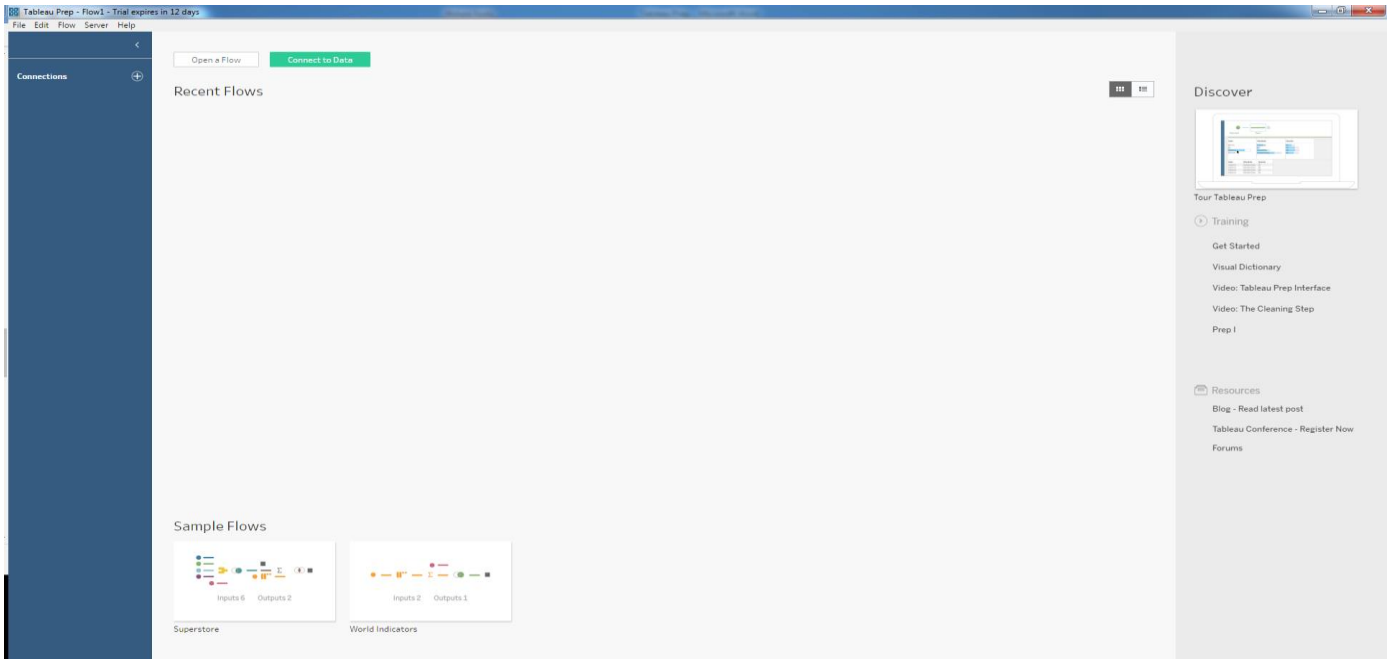

+ableau
Key Indicators
Content
Users
Groups
Schedules
Tasks
Status
Settings
Home › Regional Views › Home & Garden Annual Performance
Home & Garden Annual Performance
DATA SOURCE • By Ryan McShane • 28 views • 0 • Certified by Ryan McShane
Connections 0
Connected Workbooks 1
Details
About
Intended Audience: Entire company
Business Use: Home & Garden Annual Performance is the source of truth for information on home & garden sales, returns and customer accounts. This data source shows sales, returns, and shipping method, and associated metrics aggregated at the regional level.
PLEASE NOTE that for accounts older than 5 years (find the account age in Sales Opportunity Data Source—Account Acquired Date) this data source is not completely accurate.
Connection Type: Hyper Extract
Database: Amazon Redshift, SQL Server, CSV
What is a row of data?
A single row of data is a unique record of a customer purchase, which takes the form of an transaction ID.
Owner: Emily Chen
Backup Owner: Stephanie Richardson
Edit Description
Project
Regional Views
Move...
Owner
Ryan McShane
Change Owner...

2. TABLEAU PREP BASICS

In this chapter, we discover how to connect to your data in Tableau prep and also take a look at the various sections of your Tableau prep start page to discover its features. We will be using the latest version of Tableau prep, version 2019.1.2.

2.1. Starting Tableau Prep

To launch Tableau prep, look for the icon on your desktop or make a search in the list of your installed programs. After you launch Tableau prep, your interface should look like this:



The Tableau Start Page has four main parts; the Menu bar, Connections pane, Open flow pane and Discover pane

- Menu: The start menu is located at the top left side of your Tableau prep start page. It contains drop down menus such as File, Edit, Flow, Server and Help

1. File Menu:

   - ✓ In the file menu a user can open a new workflow using the ‘New’ button
   - ✓ Connect to an existing flow using Open
   - ✓ Save your workflow using Save
   - ✓ Export your workflows using Export Packaged Flow
   - ✓ Exit Tableau prep using Quit

2. Edit

   7. Allows you to perform functions like Undo, Redo, Cut, Copy, Paste and Select

3. Flow

    8. This menu allows you to run any workflow in your window.

4. Server Menu

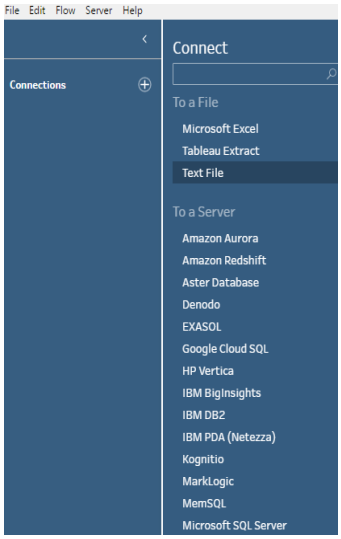
    9. Allows a user to connect to the Tableau Server.

5. Help Menu

    10. This allows you to access to resources such as tutorial videos, sample workflows as well as Tableau support online.

    11. You can also manage your product keys using this menu.

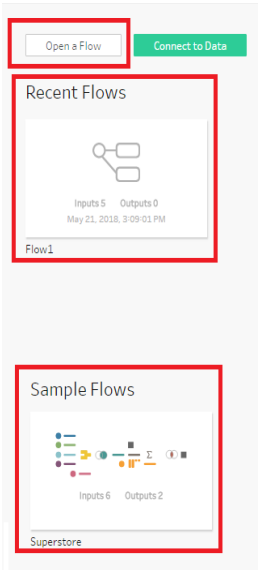
- Connections Pane:



The connect pane is located on the left-hand side of your Tableau prep interface and allows you to connect to different sources such as Excel, Text, Tableau extract or to data sources on servers such as the Microsoft SQL server, Amazon Redshift, Oracle, Teradata, Splunk, Google Cloud SQL etc.

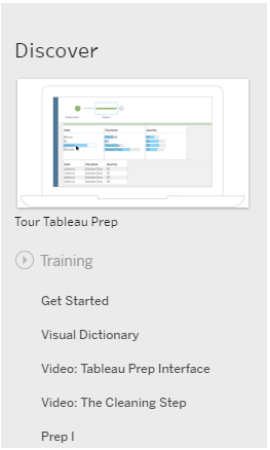
data stored

- Open Flow:



This button allows you to access and load existing workflows stored on your computer. Underneath the open flow button are all recently opened workflows. Tableau also provides quick access to sample workflows at the bottom of your pane.

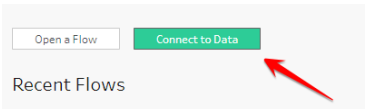
- Discover Pane

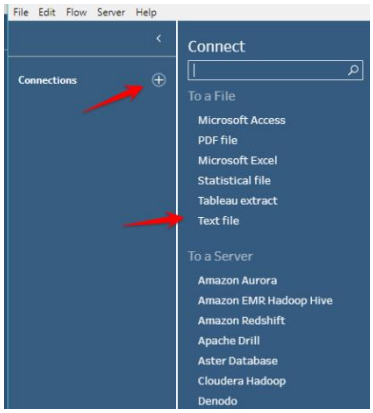


Just like the help menu, this pane on the right-hand side of your start page are resources such as training videos, blogs and Tableau's community forums.

2.1. Connecting to your data

Now that we have explored our start page, let us connect to a data source in Tableau prep. We can do this via the connections icon or by simply clicking on the green button that says 'Connect to data'.

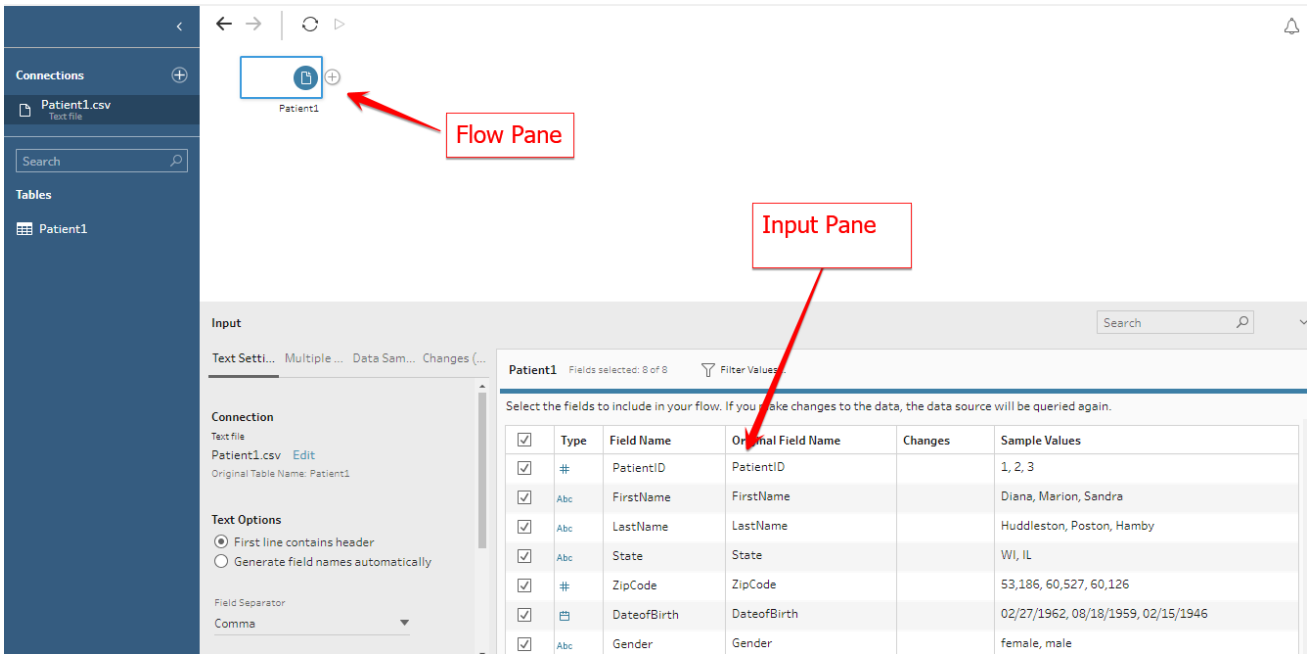




Both of these options will open the connections pane to display the list of data sources we can connect to.

We will be connecting to a flat file from file from our folder.

Click on the text file option. Navigate to the directory where your file is stored and select the **Patient1.csv** file. Your workspace should now display information about your data. The data you're connected to is displayed in your connections pane as well as information about tables in the data source.



1.

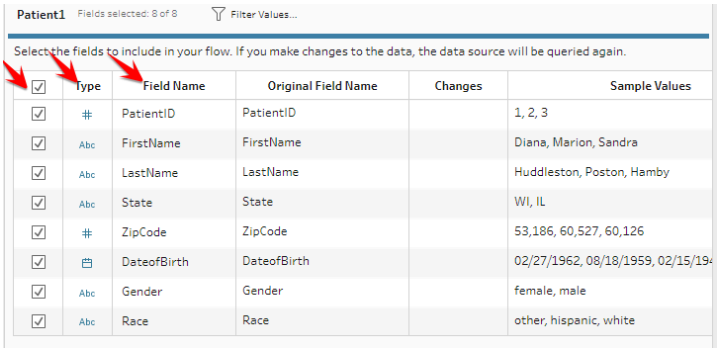
Flow Pane- This canvas allows us to interact with our data and create our workflows. You will also find the table/tables you are building the flows on in this window.

2. Input Pane - This displays a sample of your data as well as information about your data such as data types and field names. We can change data types in this window.

Tableau automatically creates the first input step in your flow with respect to single tables. You can have multiple input steps which may include multiple data sources

Clicking on the input or file in your flow pane gives you a metadata view of your data. This is the first step in your data preparation process. It is useful to take a look at the metadata to find out if there are issues you can fix there. From this meta data view;

- ✓ You can remove fields
- ✓ Rename columns or
- ✓ Change the data types.



Type	Field Name	Original Field Name	Changes	Sample Values
#	PatientID	PatientID		1, 2, 3
Abc	FirstName	FirstName		Diana, Marion, Sandra
Abc	LastName	LastName		Huddleston, Poston, Hamby
Abc	State	State		WI, IL
#	ZipCode	ZipCode		53,186, 60,527, 60,126
	DateofBirth	DateofBirth		02/27/1962, 08/18/1959, 02/15/194
Abc	Gender	Gender		female, male
Abc	Race	Race		other, hispanic, white

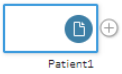
2.3. Connecting to multiple files

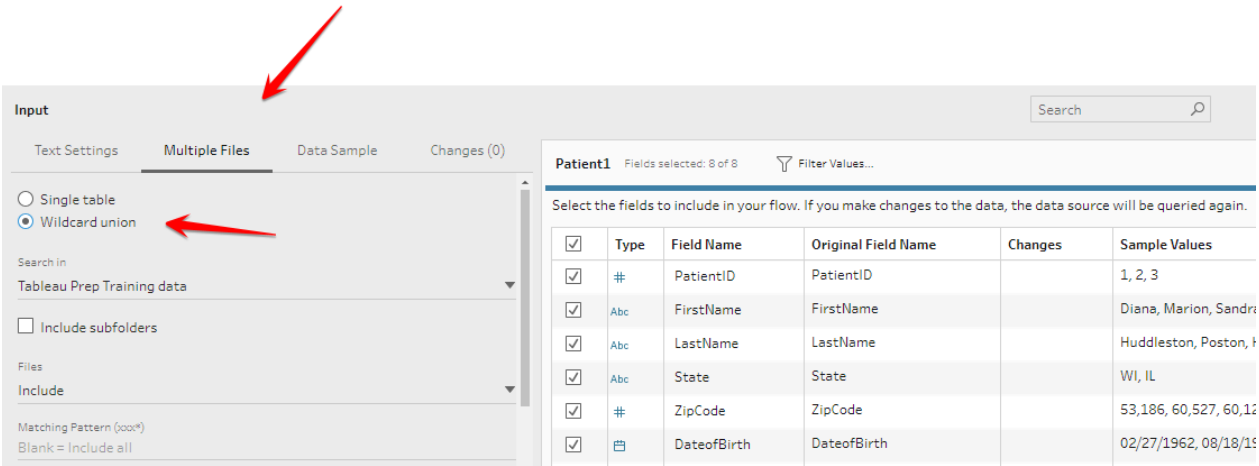
In our folder, you'll notice that there is another patient file, **Patient2.csv.** If we want to join this to the **Patient1.csv** file, we can choose to combine these by adding them individually and using a transformation to create a union or by simply using the wildcard union option. First, we win learn to use the Wildcard union,

1. Click on the Multiple Files tab

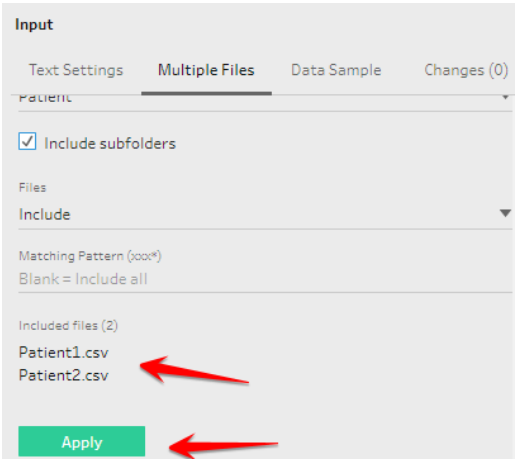
2. Select the option for wildcard union.

If the data you want to union sits in a sub-folder within your main directory, make sure you select that option.





3. You'll notice that the **Patient2.csv** file is populated alongside **Patient1.csv in** Tableau prep because they sit in the same folder. When working with a single data source with similar file names and structure, the wildcard option is a great way to create a union.



4. Hit Apply.

Your data is now combined. To make sure, check the symbol on the data in the Flow pane. It will change to show a plus icon to show that a union has been applied. The graphic on the left below will change to look like the graphic on the right.

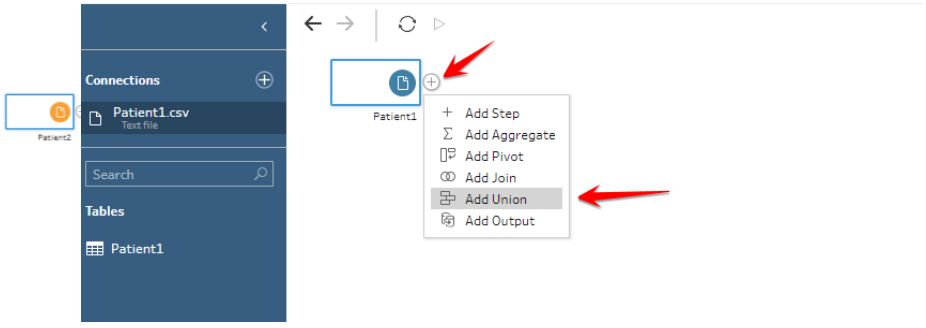
To use the wildcard union, your files must exist in the same folder or directory. Otherwise if you don't see the files, use the search option to locate them by typing a key word and add an asterisk (*) at the end

2.4. Using the Union Step

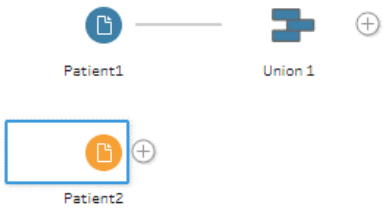
We will now use the second approach to create a union.

To do this connect to the **Patient1.csv** data source as we did before. Use the connect button again and this time around select the **Patient2.csv** data source.

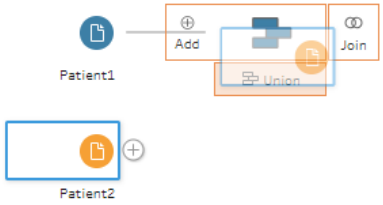
1. Click on the plus icon on the Patient1 file and select **Add Union**.



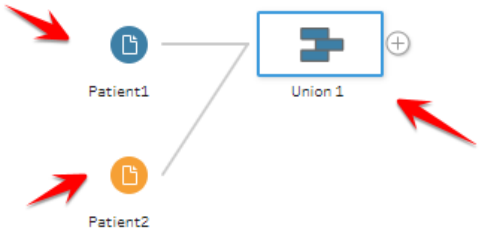
You will notice a Union step will be added.



2. Now drag the **Patient2** on top of the **Union step**. You'll notice three options now, **Add**, **New Union** and **New Join** displayed next to the union step.



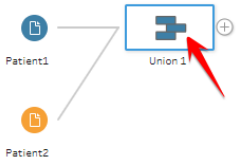
Drop the **Patient2** onto **Add**. You'll notice a connection will be created linking the two files now.

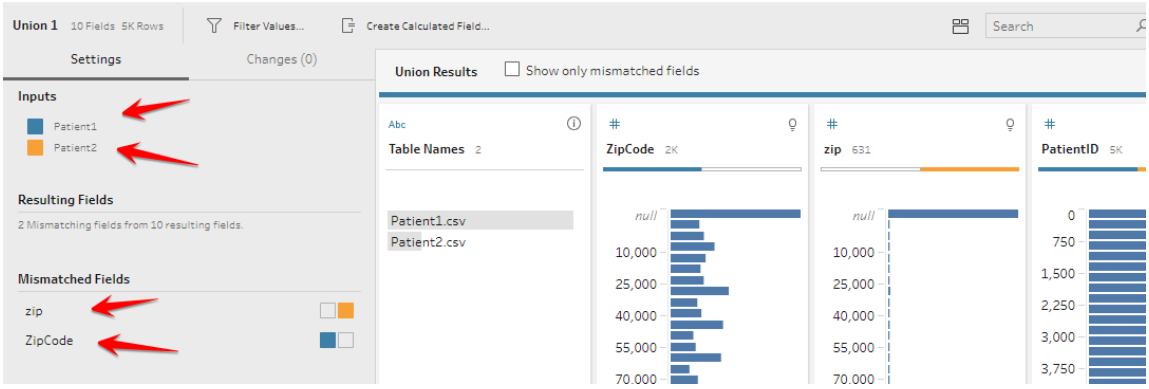


Alternatively, you can simply Drag and hover **Patient2** onto **Patient1** to see the options **Add**, **New Union** and **New Join**.

3. Right-click on the Union step and rename it from Union 1 to **'Union Patients'**.

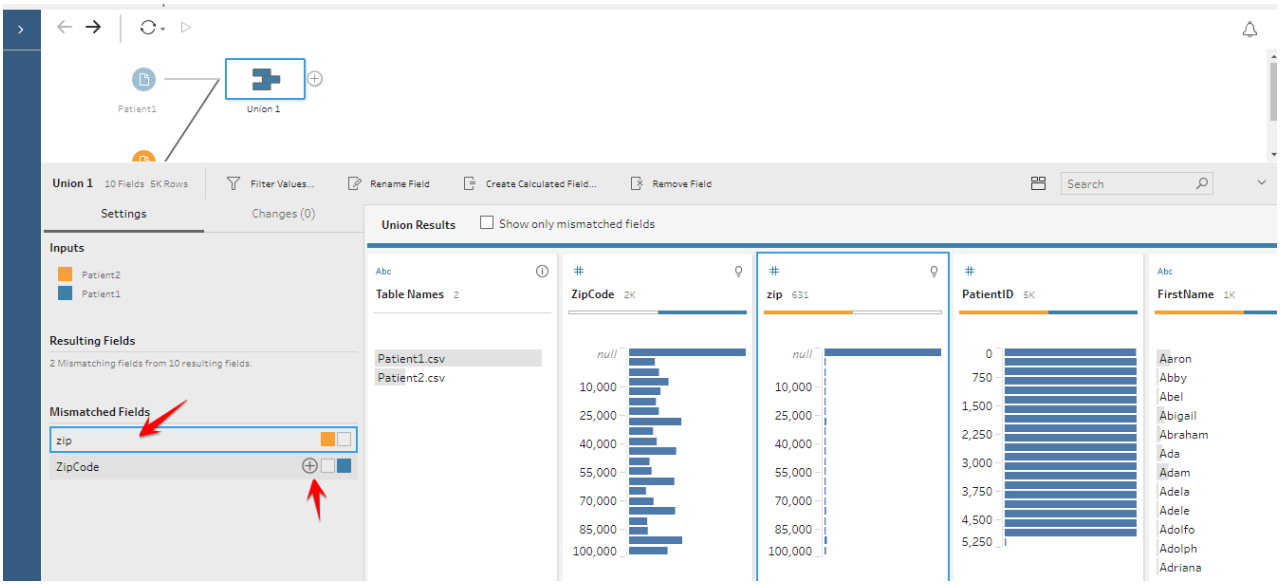
You will notice now that Tableau automatically combines the two files together because the field names match in the various orders. Clicking on the Union step will bring up the Union profile window. In this window, you will notice colors that represent the data files and help you know the makeup of the union. You will most likely see different colors for the files when there is a mismatch between them. In our case Tableau Prep has shown us that the column names **zip** and **ZipCode** are the two mismatched fields in our union.





In your profile window, you can click on the “Show only mismatched fields” to only display the columns that are mismatched.

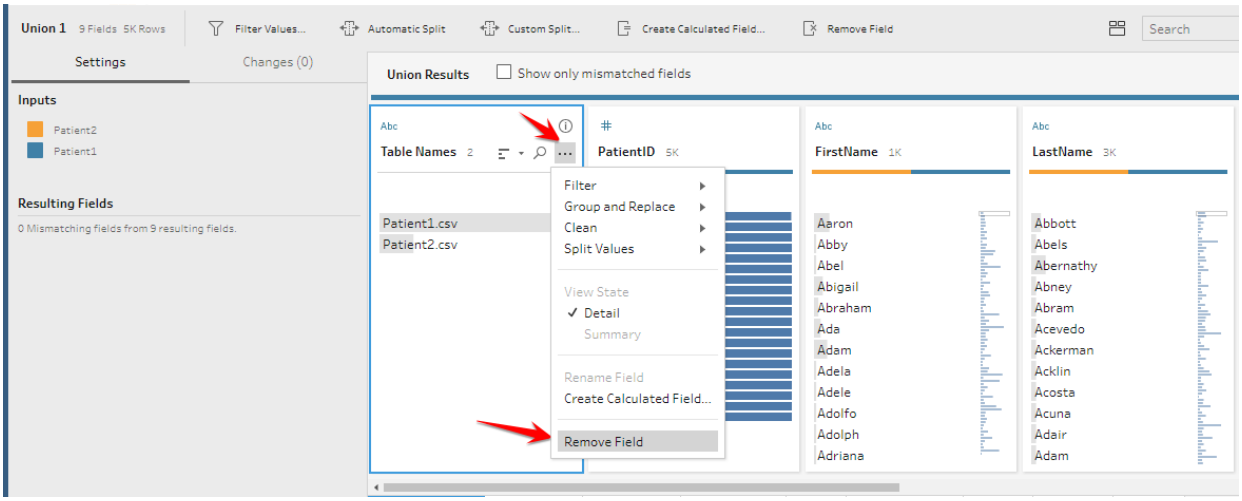
4. To fix this mismatch, click on **zip** then hover on **Zipcode** till you see a plus icon. Click on the plus icon to merge the two fields.



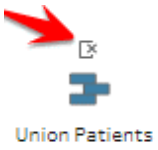
Your union is successful when it shows that there are no mismatched fields.

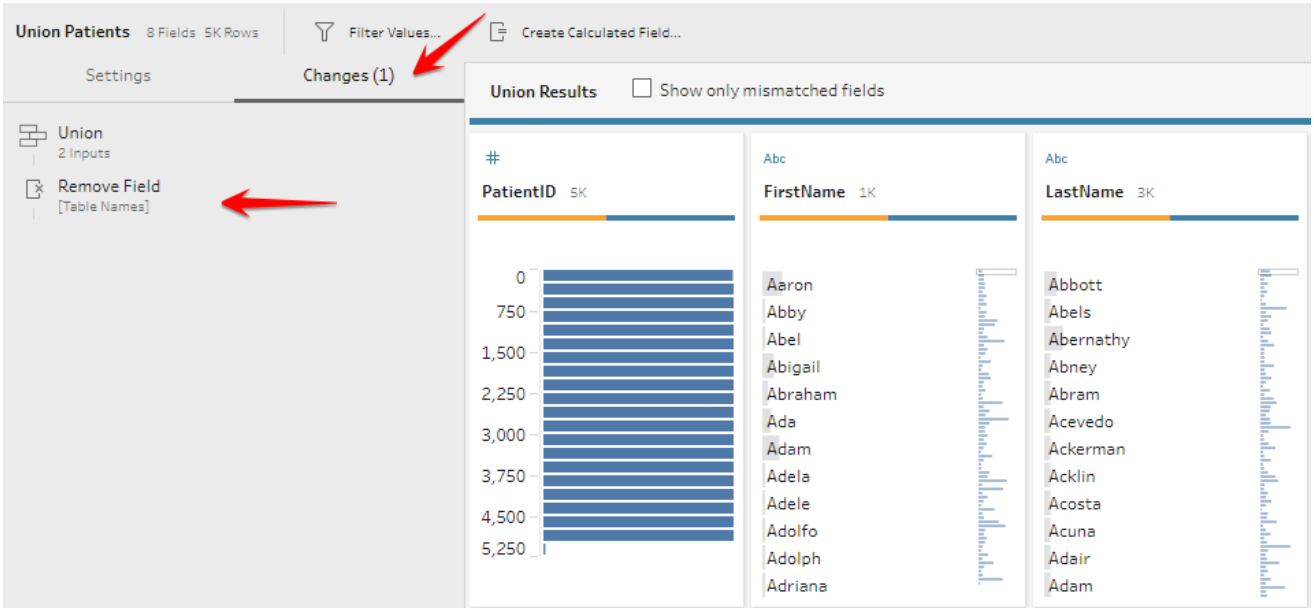
In the Union profile window, you will notice a column named **'Table Names’** which provides you with details of the data sources used in the union. Let’s remove this.

5. Select the column Table Names and then use the dropdown options menu on the top right-hand side to remove the column.



Whenever you make changes at any step in your flow, you will notice an icon appears above the step. If you want to track any changes at any step in your flow, you can click the changes tab in the profile window.



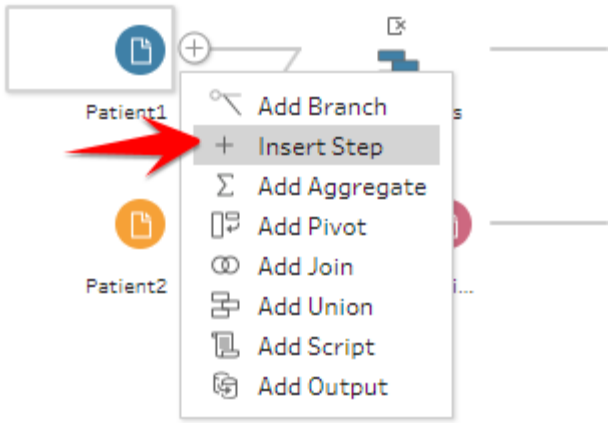


2.5. Reusing Steps in your flow

In step 2.4, we performed a union on **Patient1.csv** and **Patient2.csv**. In both data sets, there exists patient first names and last names. Assuming we want to include a calculation for patient full name in both data sets before performing the union, we will need to insert a step for each data set before including the union. To do this;

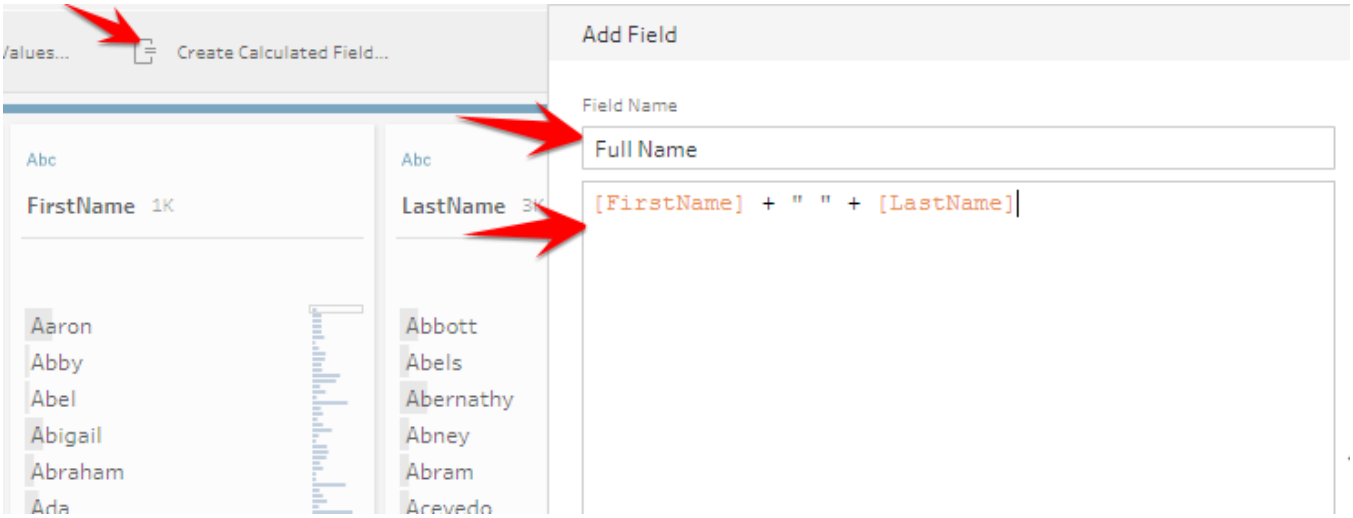
Steps

12. Click the plus icon in front of the Patient1 data set and select **Insert Step**.



13.

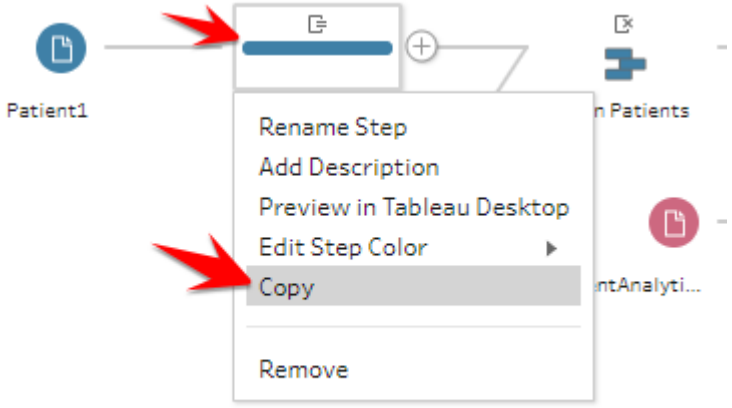
14. In the clean step, click on the **Create Calculated Field** option and in the pop-up window, input the calculation in the screenshot below.



15.

16.

17. Right-click the step you just created for Patient1 and select **Copy**.

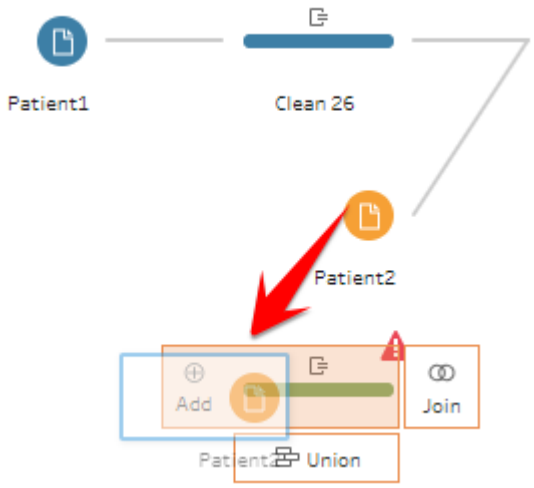


18.

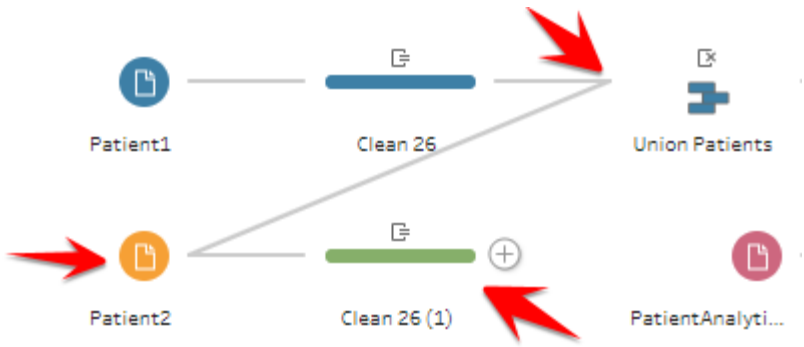
19. Paste the step next to the **Patient2** file.

You will notice there is an error because the new copied step is not attached to the **Patient2** data source yet.

20. Drag the Patient2 and select **Add**, from the hover options.

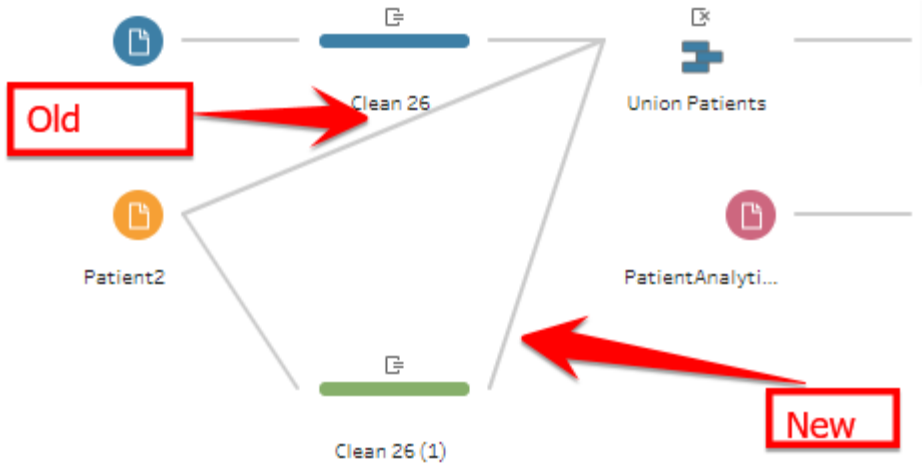
21. 

Your flow should look like the screenshot below. In a situation where you have different column names in both data sources, you may need to edit the transformations in the Changes window.



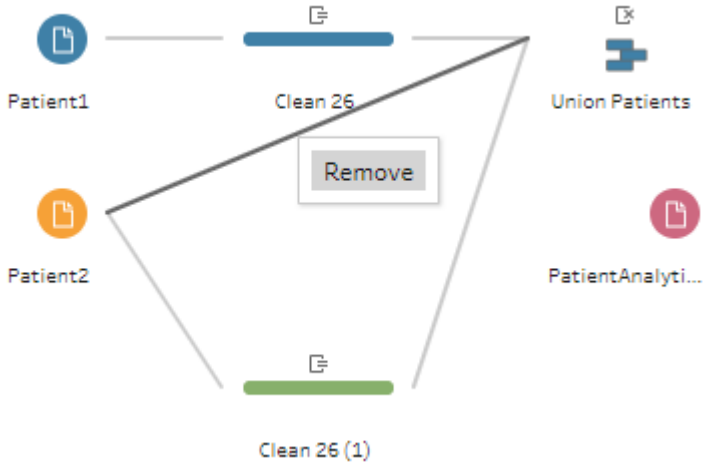
When you include steps in an existing flow, you may need to make some changes to ensure the new step added performs correctly. In the above we notice that the new step we copied is not being unioned to the clean step from **Patient1**. We need to edit this.

22. Drag the clean step from **Patient2** and **Add** it onto the Union step.



23.

24. Right-click on the old flow to the **Union** step and **Remove** it. Your flow should now correctly include the changes you made.



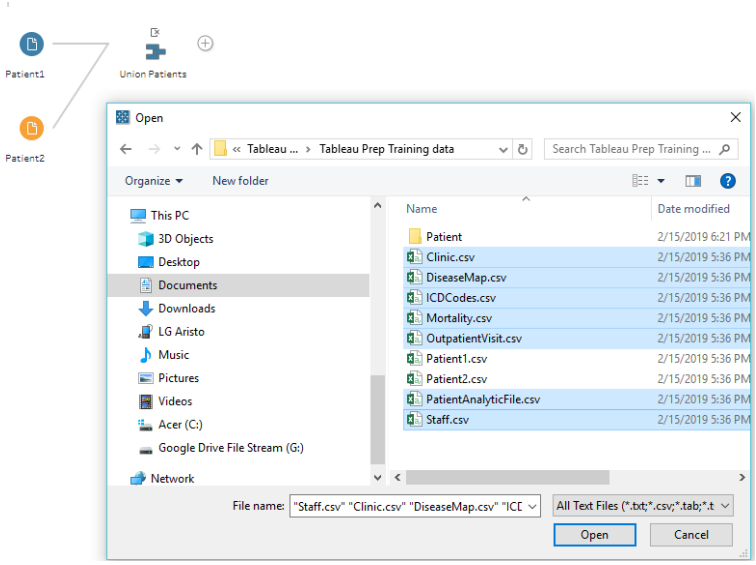
Now you see how easy it is to insert steps into an existing flow and reuse calculations or transformations. This is a useful way to be efficient when preparing your data.

2.6. Exploring your data

Now that we have learned to connect to our data in Tableau prep, let us explore the data to detect the possible problems that may exist. To do this let us bring in all our other data sources into Tableau prep.

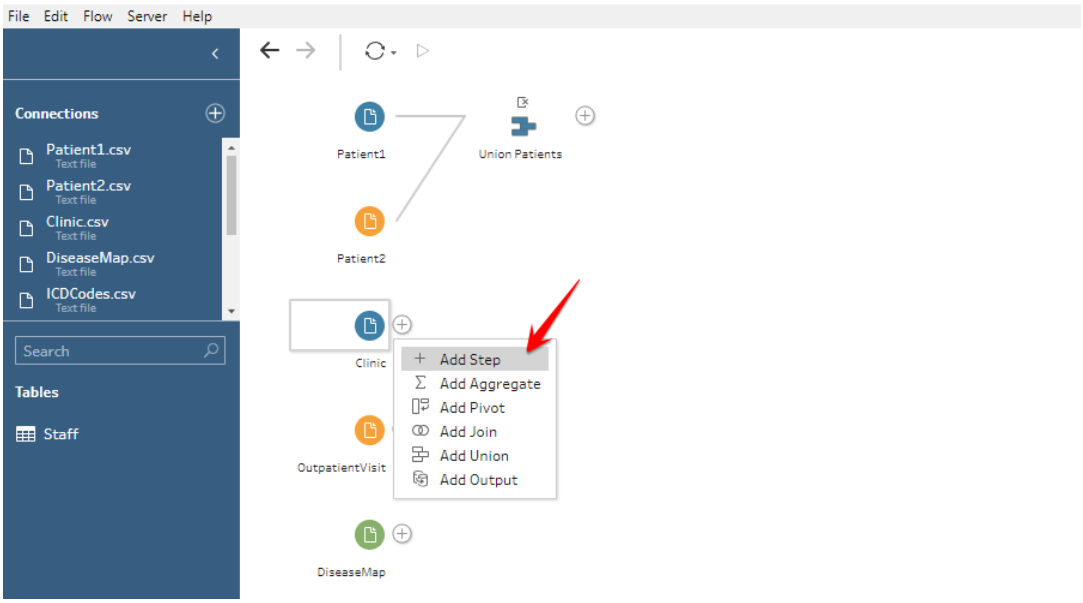
1. In the data connect pane, select text file and navigate to the directory where your data sits.

2. Use CTRL + click (windows) or Command + click (Mac) on your keyboard multi-select the other data sources then select Open.



You will now see all your data sources in the Flow pane. You can also simply drag and drop data directly from your folder to the flow pane of Tableau prep to begin building your flows.

3. Now that we have all the data sources into Prep, click on the plus icon near each data source and select Add Step. This is a useful cleaning step that allows you to see what each file looks like and an easy way to discover errors in your data.



Let's click through the input steps see if we can discover any errors that need to be fixed.

OutpatientVisit

1. There is a code in the ICD_3 column that needs to be filtered out

2. We will also apply a date filter to the data source from Jan 1st 2007 on. Note that there are null VisitDate values. What do we do with these? We can include them or not include them in a filter.

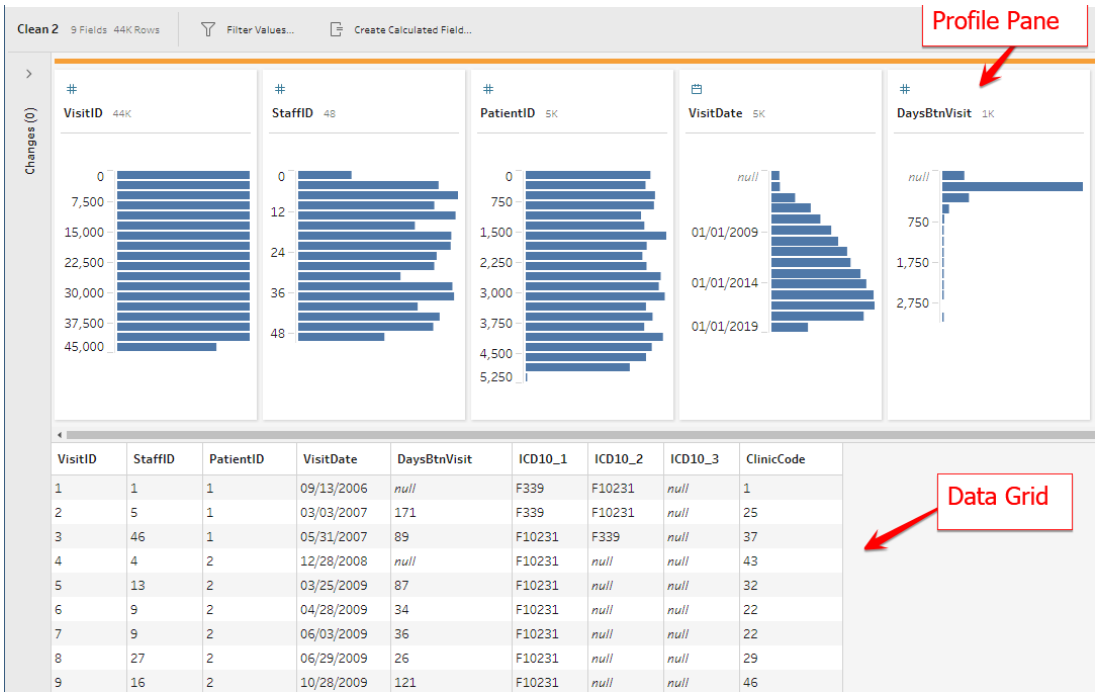
DiseaseMap

1. There is a condition myocardiac_infarction that we will rename to Heart Attack in the Condition Column**.**
2. We have similar fields Liver, LiverMild & LiverSevere in the Condition column we want to group together**.**
3. There are condition names with underscores with spaces that we will replace using a Replace function.

Patient Files

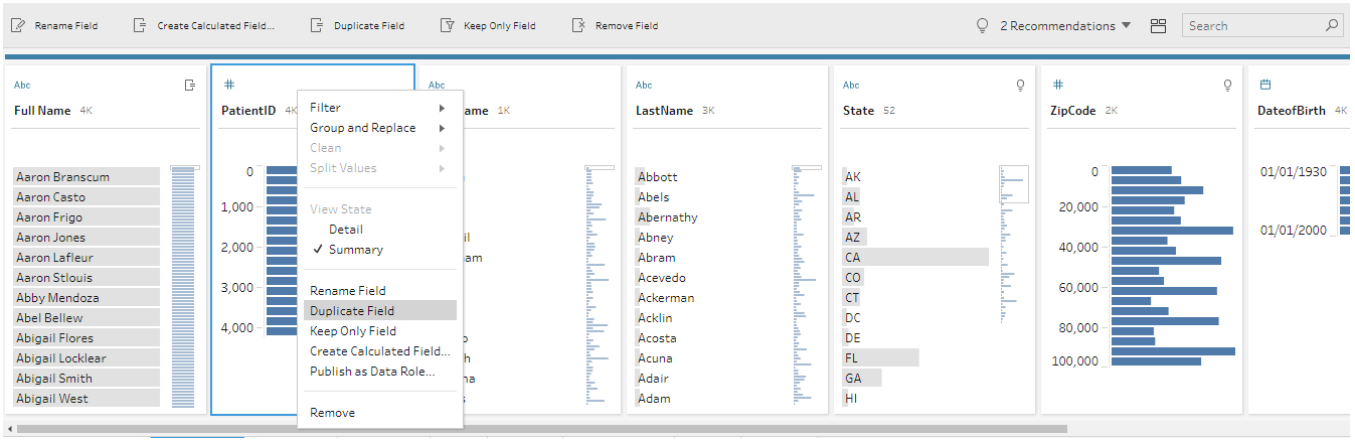
1. We will join Patient to PatientAnalyticFile and Add in Mortality
2. Create a calculation for Full Name**.**
3. We will remove duplicate ID fields.
4. We will clean up the Race column**.**

Did you notice the window that opens up when you Add a step? Two windows, the Profile Pane and Data Grid will be displayed as shown below.

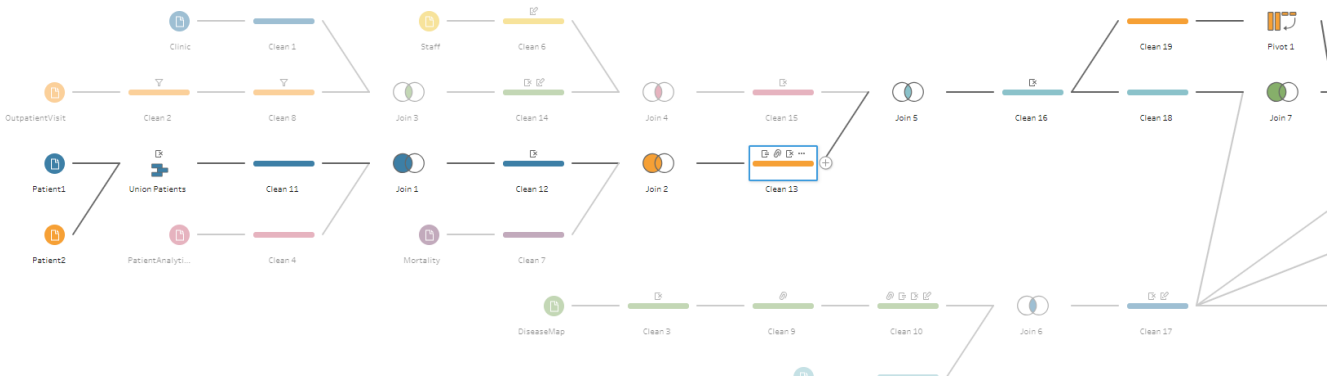


VisitID	StaffID	PatientID	VisitDate	DaysBtnVisit	ICD10_1	ICD10_2	ICD10_3	ClinicCode
1	1	1	09/13/2006	null	F339	F10231	null	1
2	5	1	03/03/2007	171	F339	F10231	null	25
3	46	1	05/31/2007	89	F10231	F339	null	37
4	4	2	12/28/2008	null	F10231	null	null	43
5	13	2	03/25/2009	87	F10231	null	null	32
6	9	2	04/28/2009	34	F10231	null	null	22
7	9	2	06/03/2009	36	F10231	null	null	22
8	27	2	06/29/2009	26	F10231	null	null	29
9	16	2	10/28/2009	121	F10231	null	null	46

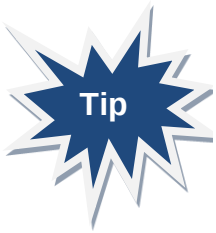
1. Profile Pane- This pane shows you a summary about your data such as number of rows in each subcategory, values of fields as well as information such as outliers or nulls in the data. You can also use the profile pane to make transformations with your data such as duplicating columns, renaming column names, removing columns, creating calculations, filtering values, grouping, publishing data and many more. It can simply be done at any point in your flow when you access a clean step and right-clicking a cloumn to access the various options that exist.



Another useful feature you can use in the data pane is to quickly find out where fields are used in your flow. For example, if we want to discover which parts of the flow, we can find PatientID, we simply need to click on PatientID in the profile pane and the portion of your flow which contains it will be highlighted flow.



2. Data Grid- This gives you information about row level of detail. You can simply click on any information in the profile pane to view the underlying row level details in the data grid.



Right-clicking on a field in the profile pane will bring up the pop-up menu to make changes to that field. By default, Tableau prep names the first Add step as Clean 1. You can double-click this and rename it as you want.

2.7. Saving your work

Two options exist in Tableau prep to save your workflows. You can choose to save as a Tableau workflow file(.tfl) or as a packaged Tableau flow file(.tflx). The only difference between the two options is that the packaged option includes the data used in the workflow. Using the packaged option is important when you want to share your flows with others.

To save, click File, navigate to Save As then select the desired file format.

When you bring in data, only a sample is displayed to maximize performance. If you want to adjust this select Data sample in the input pane and make the necessary changes.

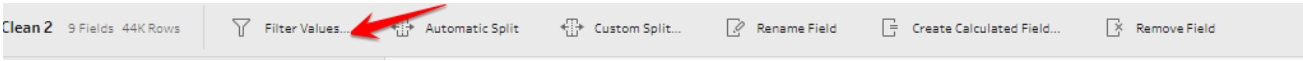
3. CLEANING YOUR DATA

Now that we have explored our data and discovered what changes we want to make, it is now time to go ahead to take the next step and start cleaning the data. As we mentioned before it is good to Add a clean step before performing any data cleaning. Whenever you want to perform data preparation operations such a rename, split, filter and merge etc. you need click the icon next to the input step and ⊕ select **Add step**.

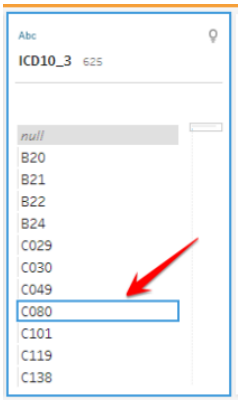
Other steps include

- ✓ Add Branch - allows you to split your flow into different branches
- ✓ Insert Step – allows you to insert a step into an existing flow
- ✓ Add Aggregate- to group and aggregate fields
- ✓ Add Pivot -to swap columns and rows in your data
- ✓ Add Join- to combine two data sources together
- ✓ Add Union - lets you combine two similar data sources
- ✓ Add Output – lets you save your cleaned data as a Tableau Data Extract. We will now start cleaning our data using the OutpatientVisit input.

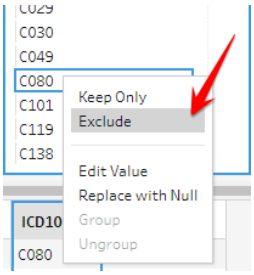
1. Click on the OutpatientVisit clean step.



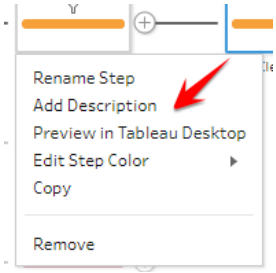
2. In the profile window, locate the column name ICD10_3 and click to select it. You'll notice a new menu displays on top of the Profile pane
3. Find the code C080 in the row data within the ICD110_3 column.



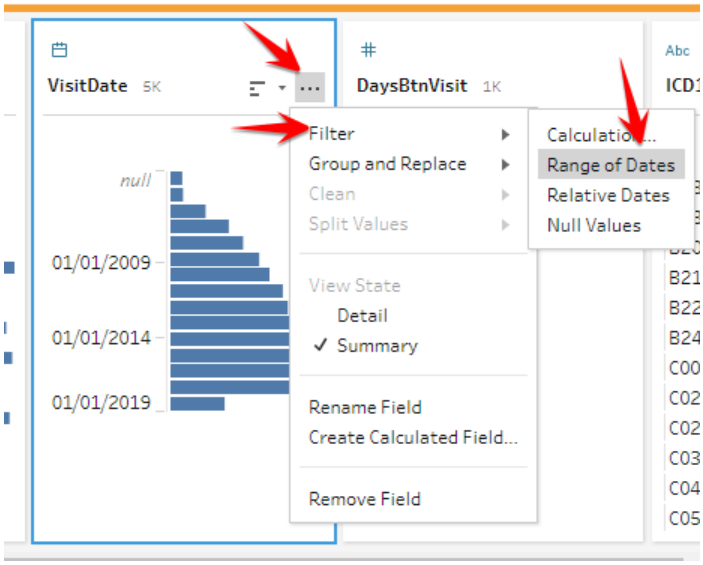
4. Right-click on the code C080 and select Exclude. Alternatively, you can select the code and use the Options at the top of the profile pane to Exclude.



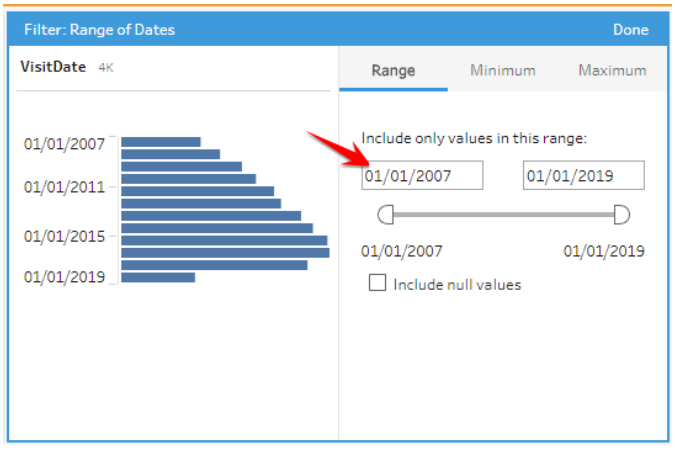
Notice that by default Tableau Prep has a default name for a clean step. You can edit this to allow whoever is viewing your flow to have an idea what kind of changes you made at that step. You can simply right-click the step and input your own custom name. Additionally, you can add a description or color to the step to provide further information. You may also want to preview the data you've just cleaned in Tableau Desktop by using the Right-Click options at the step.



5. Add a description to reflect the changes we just made. Type in "Exclude C080".
6. Add a new clean step (i.e. Add Step)
7. Select the column name VisitDate. On the right-hand side open the drop-down menu and select Filter > Range of Dates

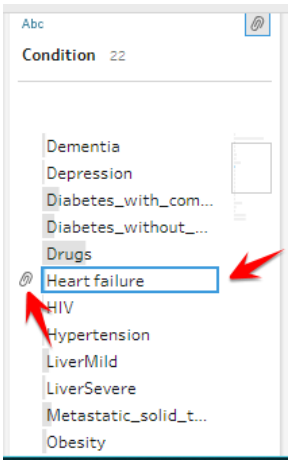


8. Select the date range 01/01/2007 – 01/01/2019 to filter the date



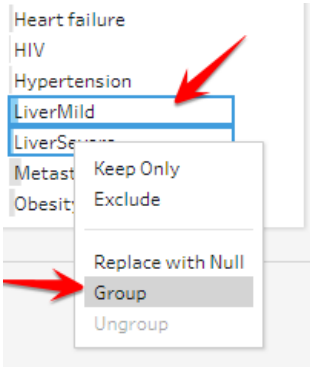
9. Change the default clean step name or Add a description to your step.

3.1. Grouping and Replacing



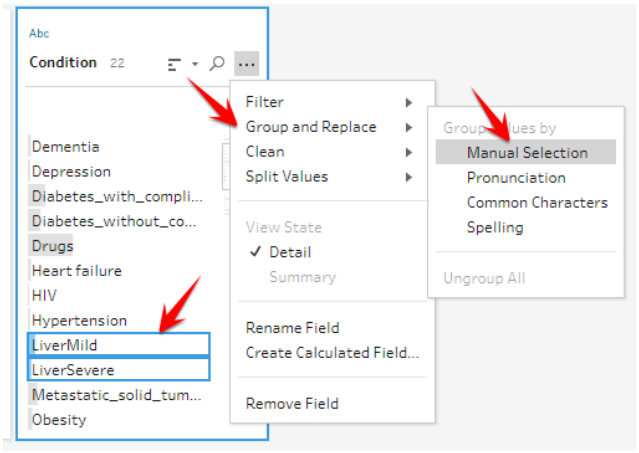
Let us now select the **DiseaseMap** file to fix the issues we identified in our exploration.

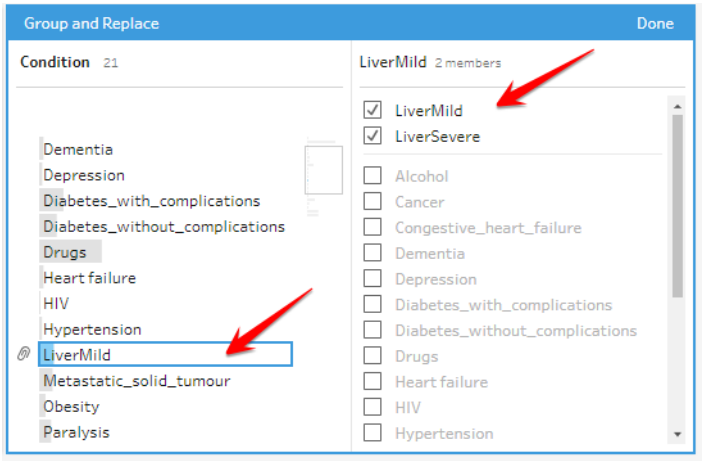
1. First let us remove the DiseaseMapID field**.**
2. Add a clean step. In the Condition double-click on **m**yocardial_infarction**.**
3. Rename this to Heart failure. Notice the paperclip icon next to the renamed field? This lets you know that a grouping has taken place**.**
4. Add a new clean step. Select LiverMild and LiverSevere then Right-click and Select Group.



5. Double-click and rename the field with the paperclip icon to Liver**.**

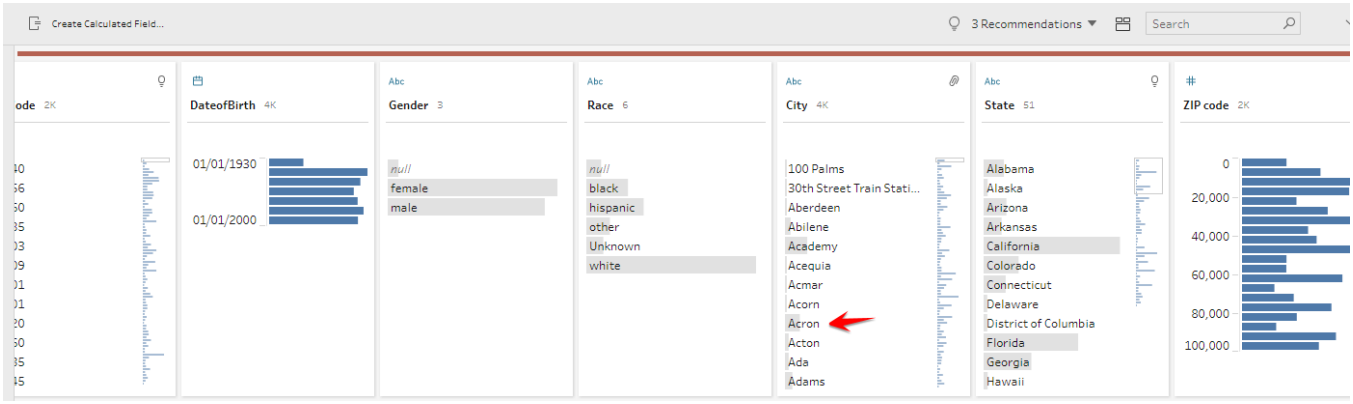
Alternatively, you can also use the options menu to do a grouping. Select Group and Replace > Manual Selection. A new window opens that shows check boxes near the name of the various conditions. Check LiverSevere and LiverMild.





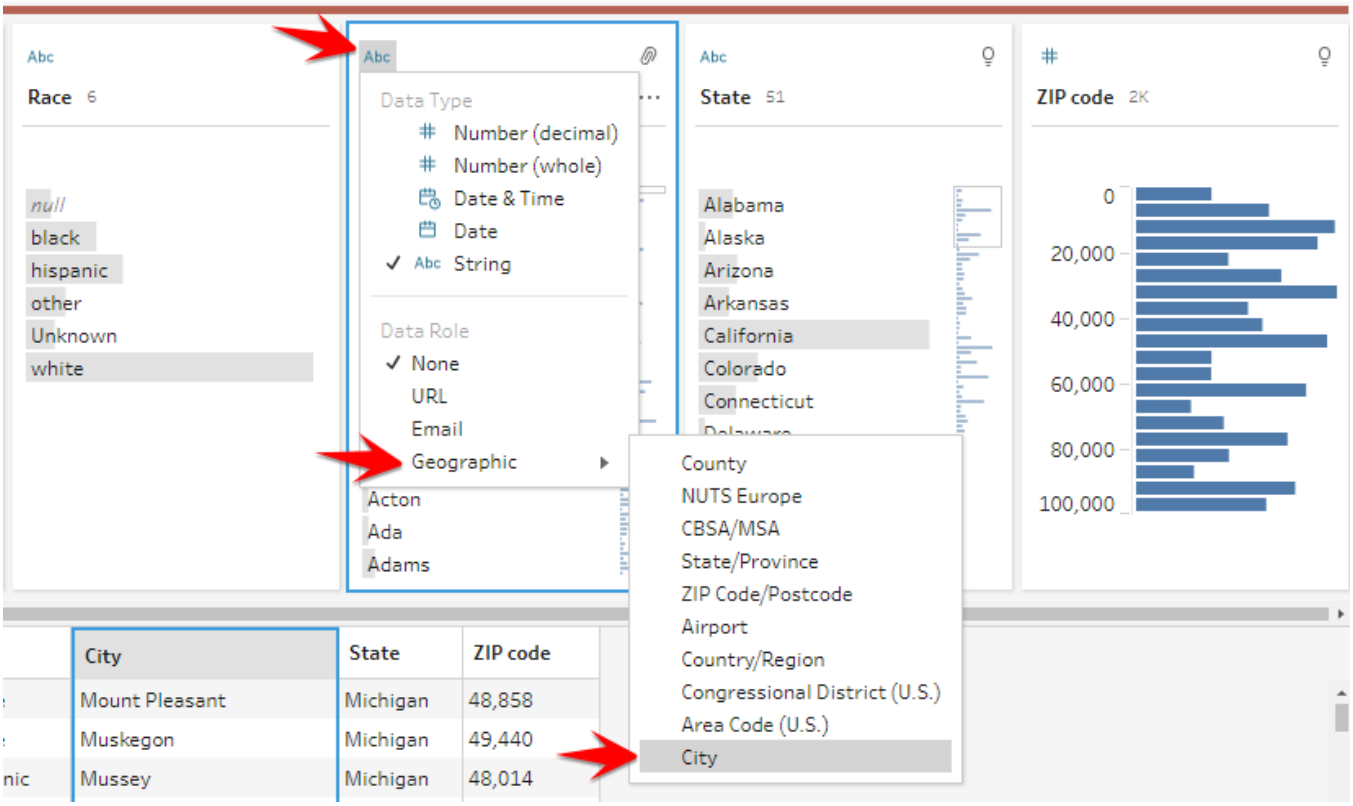
6. Click done. Now double-click and rename the field with the paperclip icon to Liver.

We can also make use Tableau's prep fuzzy matching algorithms to group fields when their values are spelled incorrectly. In the example below, there are values in the City field such as Acron instead of Akron. When you have a situation like this and want to make use of Tableau Prep to automatically fix the spelling errors, you first need to make sure we have converted the City field to a geographic role.

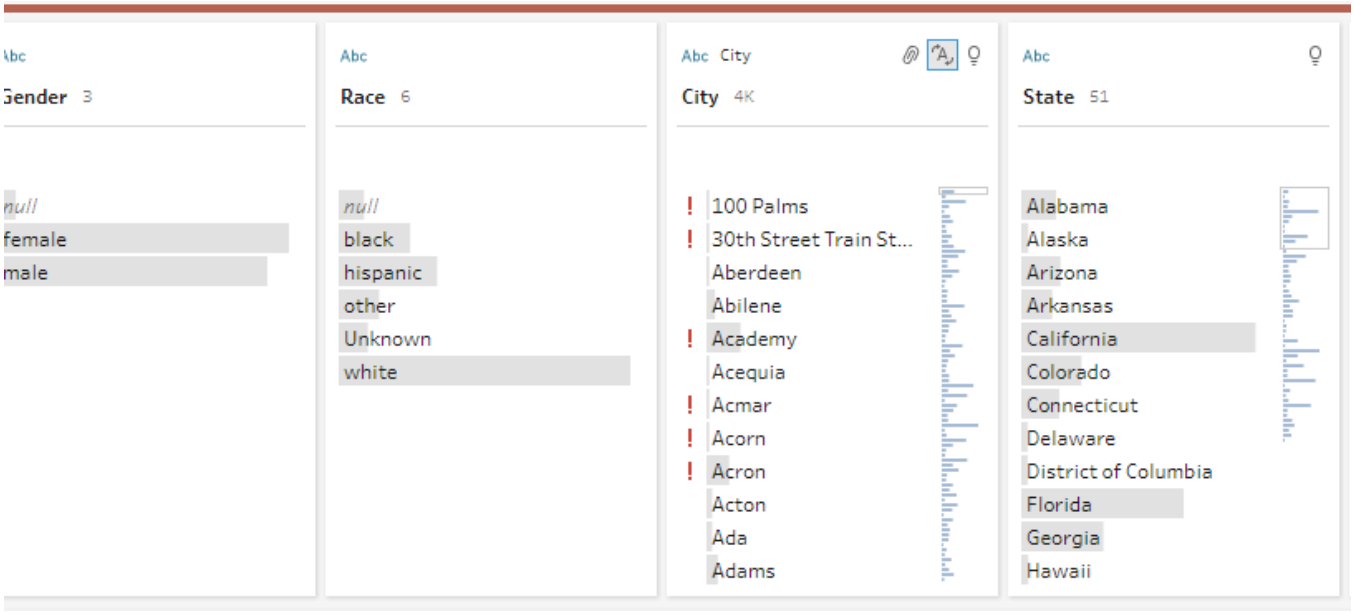


Steps

1. Click on the data type icon on the top of the City column then select Geographic > City

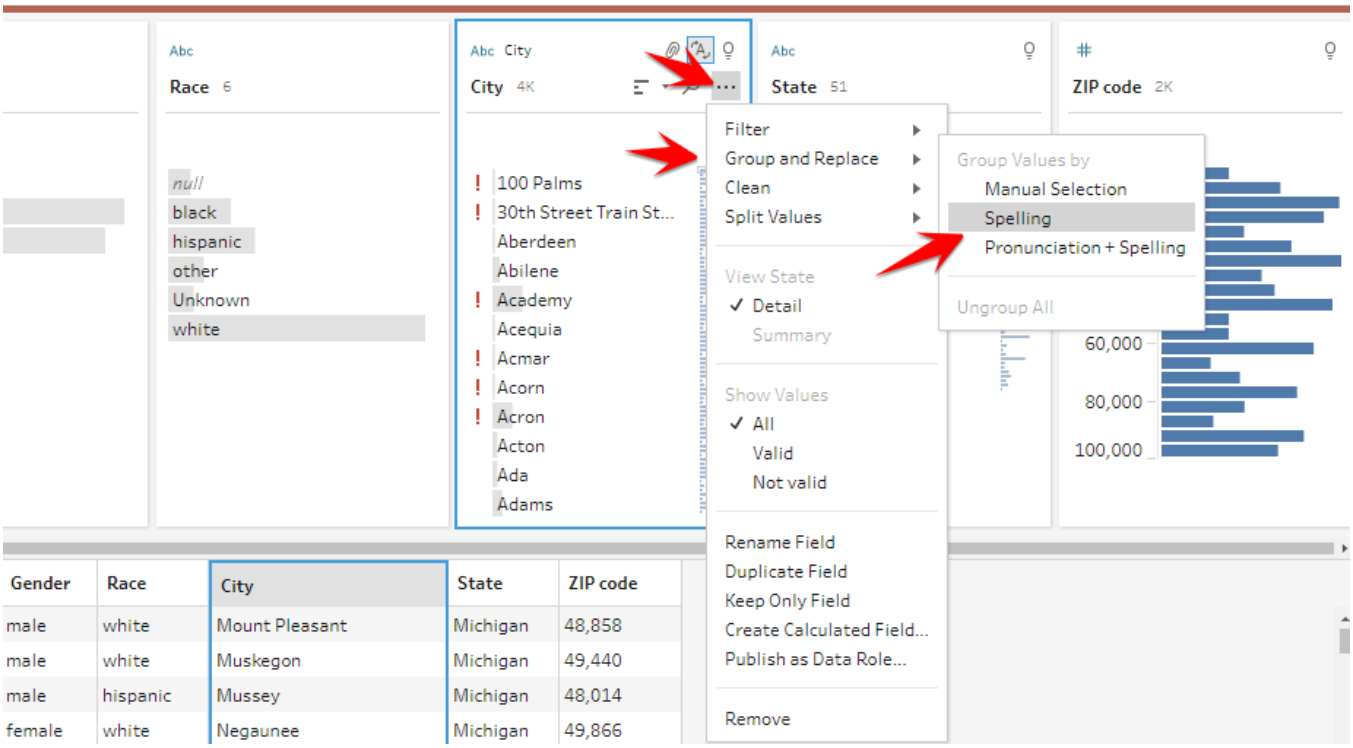
25. 

Notice that after switching the data role, Tableau Prep automatically recognizes the values that are incorrect with a red exclamation mark.

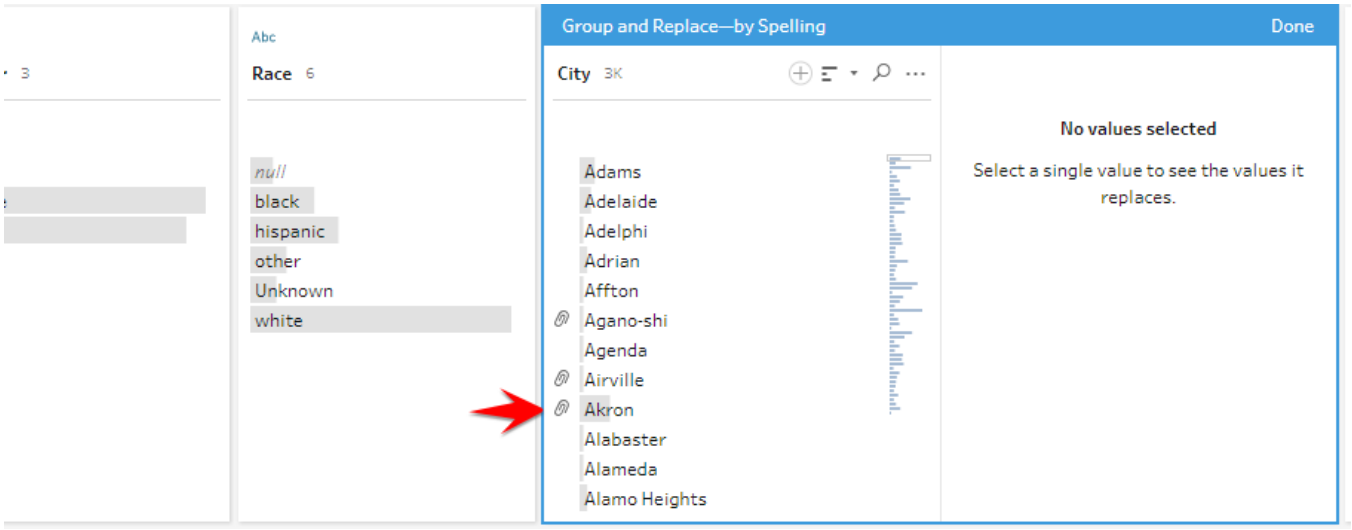


26.

2. Open the Edit menu and select Group and Replace >



3. Hit done. Notice that Akron which was spelled wrongly as Acron is now fixed but this would not always be the case and you may have to use other Group and Replace options to fix the errors.



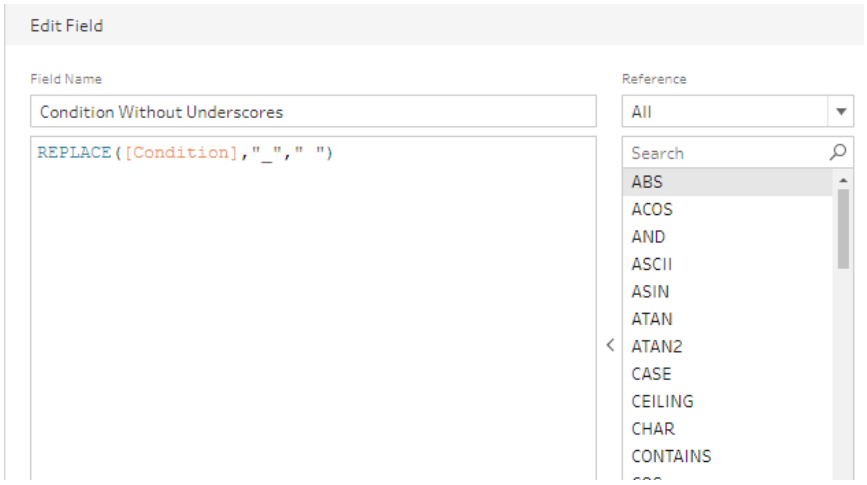
3.2. Using a Calculated Field

In some scenarios you may have to use calculated fields to fix issues with your data as we did previously in chapter 2.5 of this book. Remember that when we explored the **DiseaseMap** file, we realized that Now that we have grouped these files, let us replace the rows in the **Condition** column that have underscores next to them such as "**Diabetes_with_complications**"

To do this:

1. Select the Condition column then **c**lick Create Calculate**d** Field from the toolbar at the top of your Profile pane window**.**

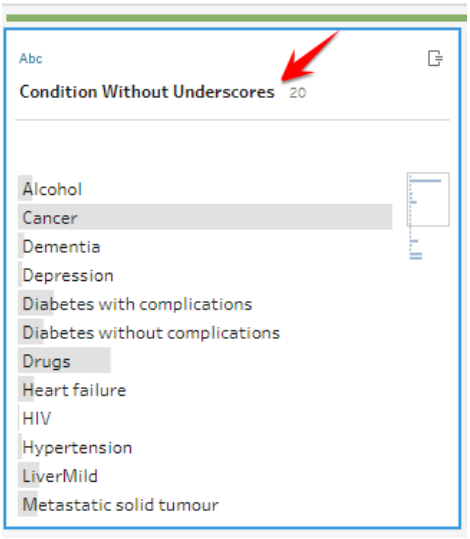
2. In the dialog window that opens, name the calculated field as Condition Without Underscores. For the calculation, input REPLACE([Condition],"_"," ")



3. Click Save.

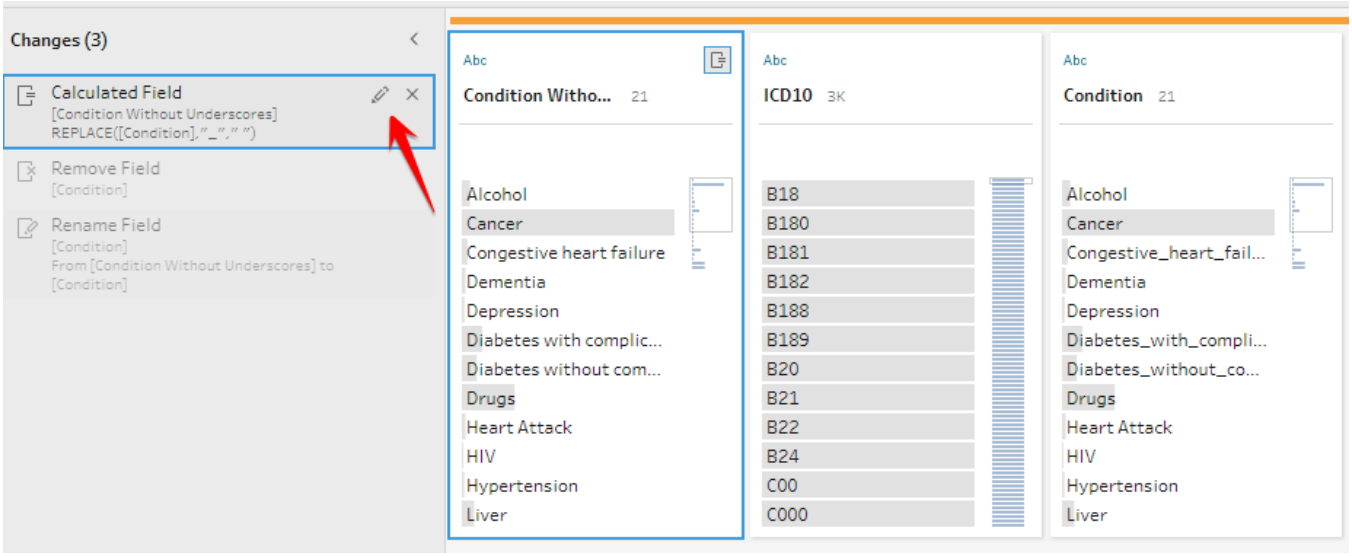
You'll notice that the new field name will be added as a column in your Profile pane.

4. Click on the Condition column now and remove it.
5. Double-click the Column name Condition Without Underscores and rename it as Condition.

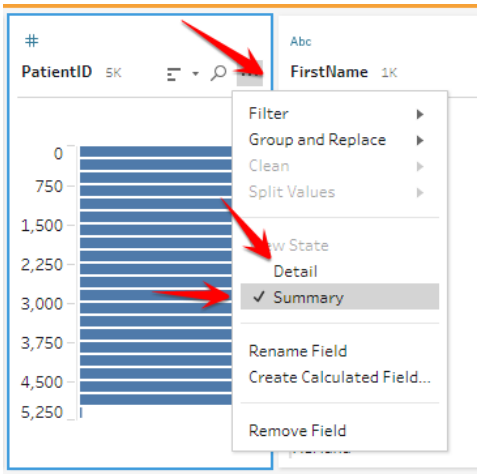


When you make any change to your data in the profile pane, these changes appear on the left-hand side of the profile pane window (see below). This is a good way to track your cleaning process. You can also make any changes in your prep process through this menu.

To edit or remove a step, simply click on the step and hover to the far-right hand-side then select the appropriate action you want. (See image on next page.)

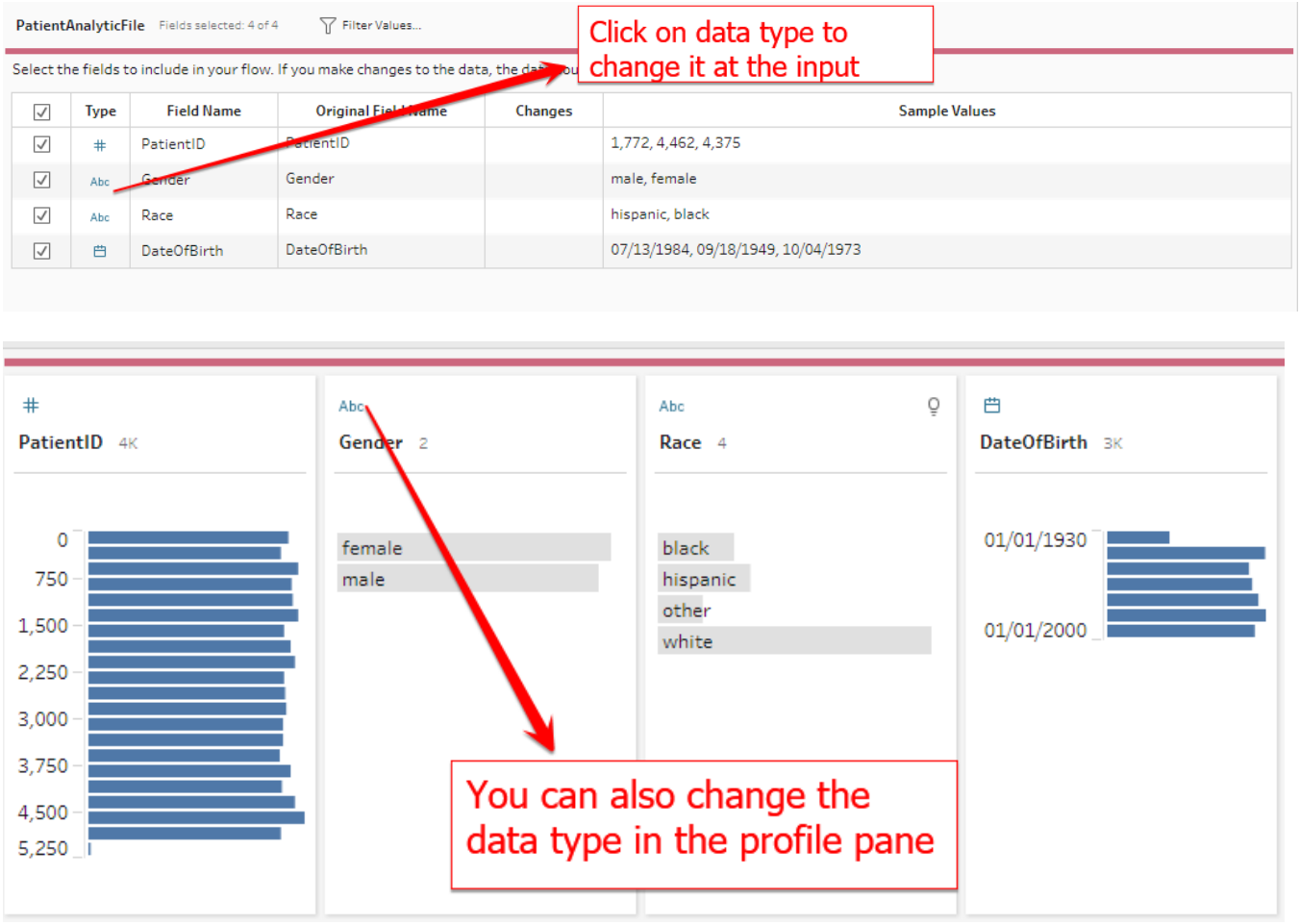


In the profile window, did you notice that some columns show blue bars and others show a text view? You can easily toggle between these views by clicking the options dropdown and selecting either **Summary** or **Detail**. However, they may be times where one of these options may be greyed out. If for example you select a column with string values the **Summary** option would be greyed out.



3.3. Changing the data type

There are two ways to change your data type within Tableau Prep. You can either change the data type at the input stage when you bring in new data or you can add a clean then make this change in the profile window as show below



4. JOINING YOUR DATA

Now that we have explored our data and cleaned it, let us start joining the data into a shape we want. In the next steps, we will now join the **Patient** file to the **PatientAnalyticFile** and then add the **Mortality** file to it. We will achieve this result by using a union step in our flow pane. Before we join our data, let us briefly look at the type of joins we have. Because we have already looked at Unions, we will overlook them in this section.

4.1. Types of Joins

- ✓ Inner Join – When you combine tables with this type of join, the result will only contain values that are a match in both tables based on a common linking field
- ✓ Left Join – When we combine tables with a left join, our result will contain all values from the left table as well as any matched values from the right table. Nulls result when a value in the left table has no corresponding match in the right table.
- ✓ Right join – The results of this join will contain all values from the right table as well as the matching values from the left tables. When values in the right table have no corresponding match in the left, nulls are the result.
- ✓ Full Outer Join – This type of join combines all results in both your right and left table regardless of whether there is a common field between them or matching values.

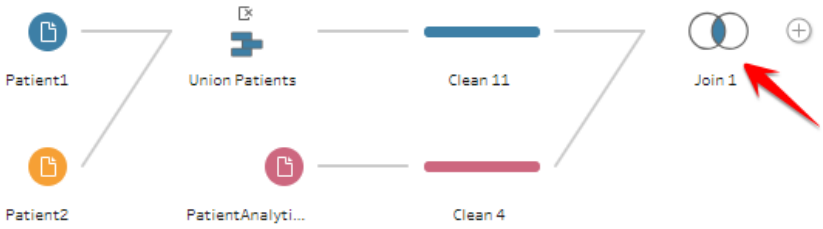
It is important to note that using a Right or Left join depends on the position of files in you flow. To use correctly, have in mind which tables you want to display all values from before deciding whether to use one of these options

Now let's use Tableau Prep to join our data files.

Steps:

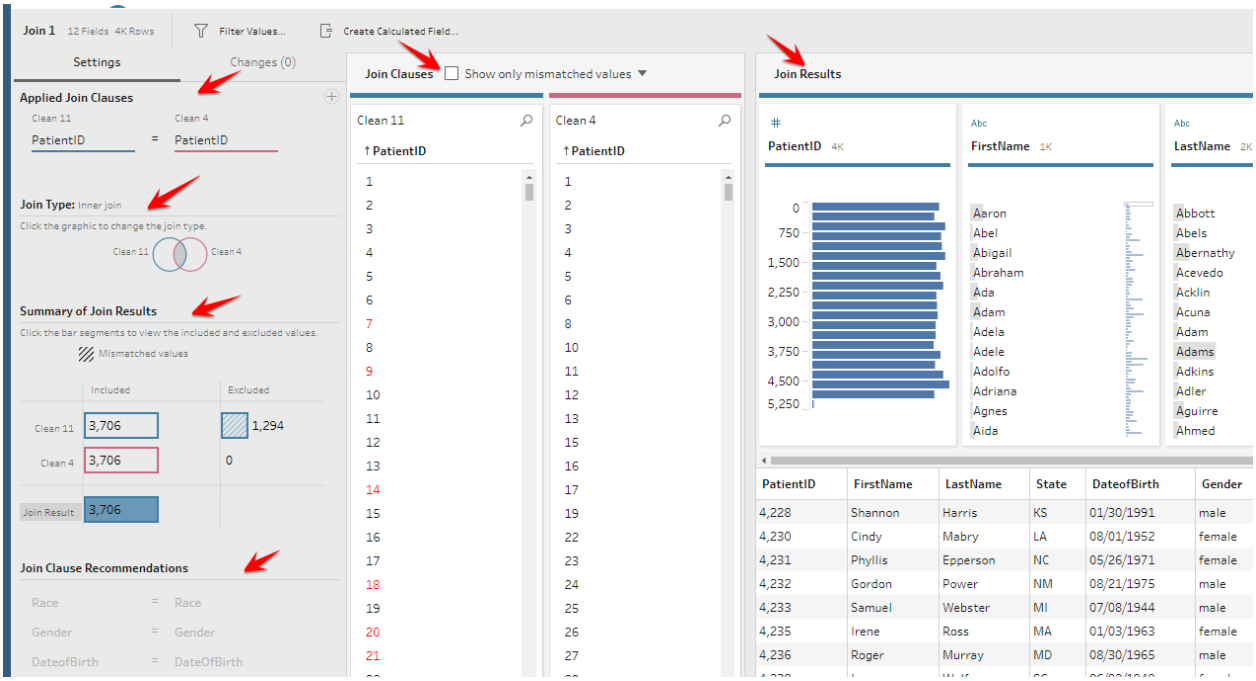
1. Click on the Union Patients step and add another step to it (or the step you combined Patient 1 & 2 together)
2. Click on the PatientAnalyticFile and drag it to the new clean step you created for the Union Patients. Once you hover on top of it, Tableau Prep gives you the option to do a Join or a Union.

3. Drop the PatientAnalyticFile onto the Join. Tableau Prep by default creates an Inner Join but we will change this to a Left Join.



4. Click on the new Join created to bring up the Profile window.

On the left-hand side of the window, you will find information on how Tableau Prep created the join. Let us look at some of the important elements of this window.

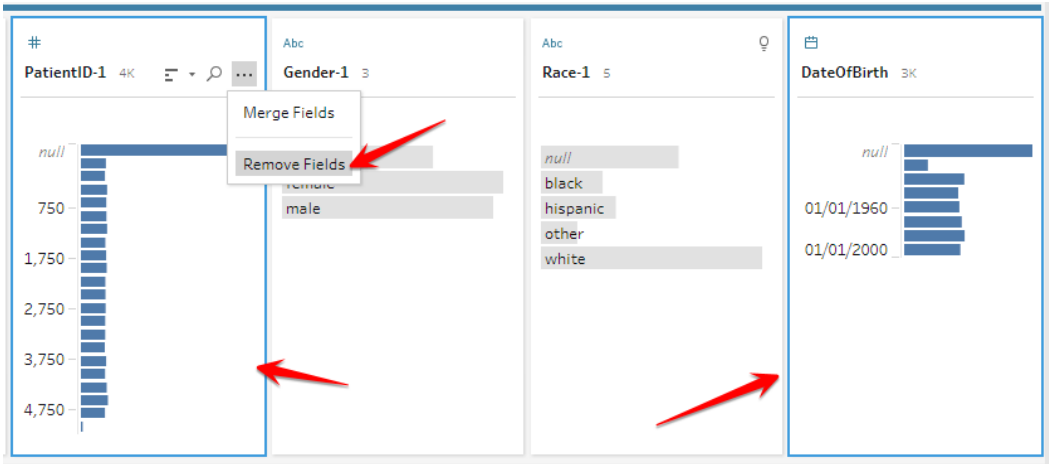


- ✓ **Applied Join Clauses–** Here you will find the common linking field Tableau Prep used to create the join as well as which files or tables they come from. In this case our two files are being joined on the PatientID field. If we want to make changes to the field used to create the link you can do this here.
- ✓ **Join Type –** This is where we make changes to the type of join, we want. As we saw earlier, Tableau Prep is doing an inner join here. We can simply click on the Venn diagram to perform which specific join type we want.
- ✓ **Summary of Join results –** This allows you to see the number rows included in your join as well as what is excluded.
- ✓ **Join Clause Recommendations –** When you want to know which fields match in the files or tables you're joining; this menu gives you that information so you can use those as your linking fields if you want.

5. Now on the Join Type menu click the left-hand side of the Venn diagram to change the join type from an Inner join to a Left Join. Do you notice the difference in your join results?

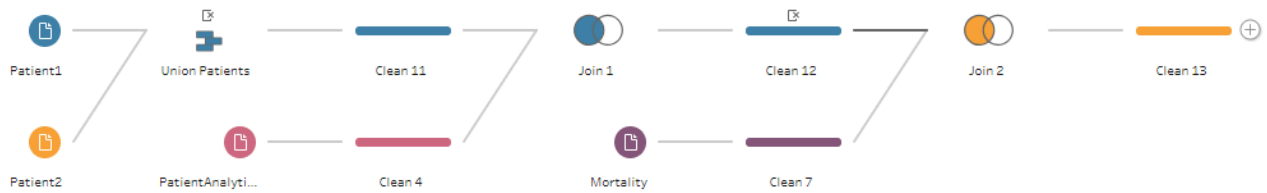
6. Add a New to step to the join. At this step notice that we have some duplication in columns

7. Remove the PatientID-1 and DateOfBirth. You can choose to use CTRL + click (windows) or Command + click (Mac) on your keyboard to select both columns then remove them



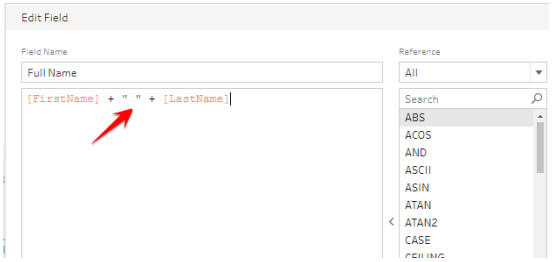
8. Add a new step then drag in the Mortality file and create a Left Join.

9. Add a new step. Your flow should now look like this



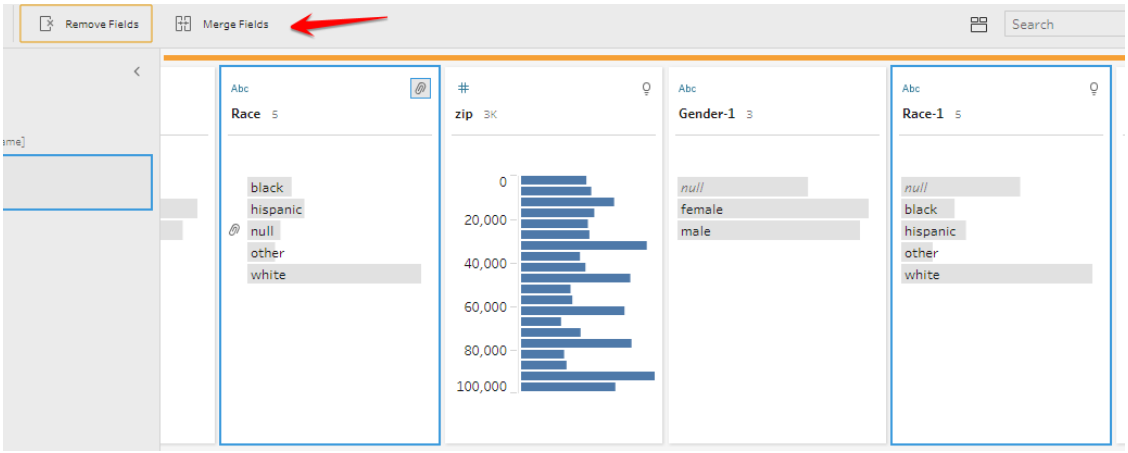
Notice that we have a column name for **FirstName** and **LastName**. Let's create a calculation to merge these together into a **Full Name**.

10. Click Create a calculated field on the top of your profile window and input the calculation below.

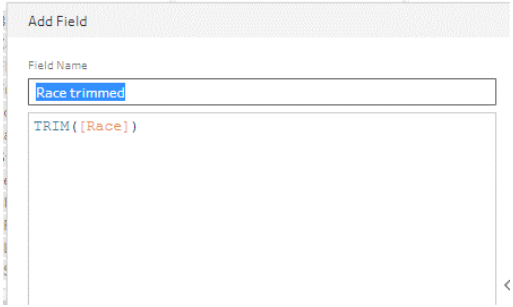


11. Click on the Race column. Notice that we have Two categories null and Unknown. Group these together and make sure the grouped categories are renamed "null".

12. Select the Race and Race-1 columns. On the top of the profile window select Merge Fields



13. Use a calculated field to remove the leading spaces in the merged Race column with the calculation below**.**



Alternatively, we can fix the extra spaces by Right clicking the Race column and selecting Clean > Trim spaces

14. Remove the Race column and rename Race trimmed to Race

15. Remove the PatiendID-1 column

16. Merge the Gender and Gender-1 column.

This completes our flow for all the Patient data we want, i.e. Patient1, Patient2, PatientAnalyticFile and Mortality. The next stage is to complete our Patient Visit data. We have already cleaned up the Outpatient file. We are going to add Clinic data and Staff data to this and then clean up the join results. After this we will join the Patient data to the Visit data.

Steps

Drag the Clinic clean step onto the final step of the Outpatient visit file we cleaned earlier to create a join. Leave this an inner join.

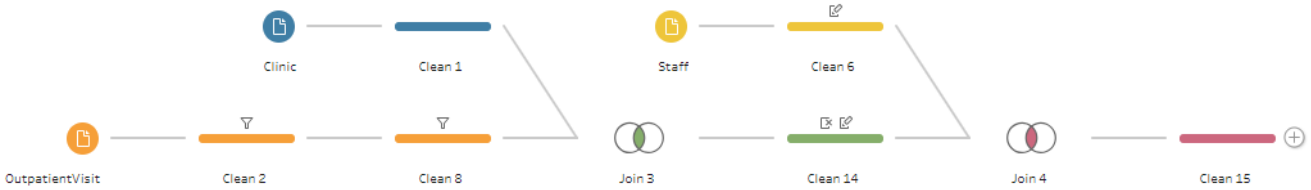
27. Add a clean step. Remove ClinicCode-1 and rename ClinicCode as ClinicID

28. Click on the clean step for the Staff file and Rename the following**:**

    FirstName > StaffFirstName, LastName > StaffLastName, HireDate > StaffHireDate, HourlyRate > StaffHourlyRate, Salary > StaffSalary, PayType > StaffPayTye

29. Now drag the Staff file onto the clean step in 2 and create an Inner join.

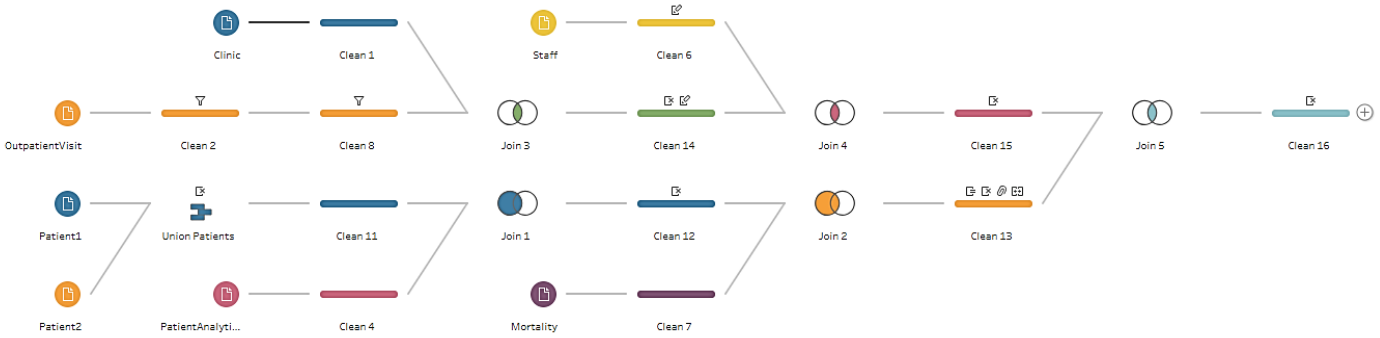
30. Add a clean step. Your flow should now look like this.



31. Remove StaffID-1 from the clean step

32. Now drag this flow above to the Patient flow to create an Inner join**.**

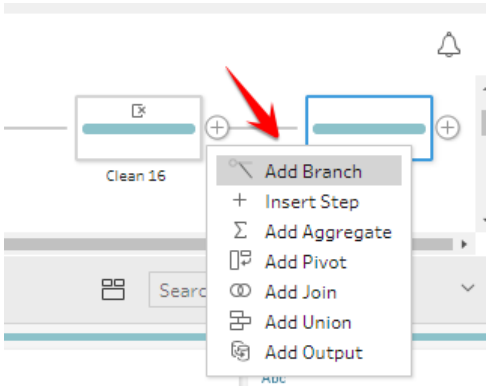
33. Add a clean step and remove the PatientID-1 and PatientID-2 duplicates. Your flow should now look like this.

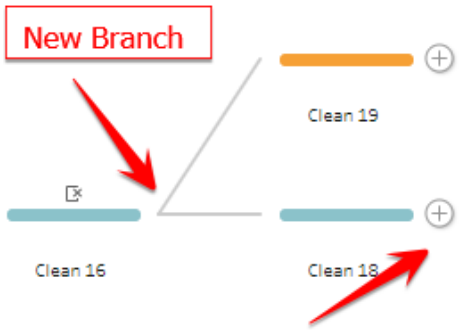


4.2. Adding a Branch

We're now going to pivot the **ICDCode** column in our flow above so that it can be added to a join between the **DiseaseMap** file and the **ICD codes**. This is due to the fact that we want to generate two separate outputs. One output that focuses on our **Visits** and another that focuses on our **Conditions**. To do this we will need to Add a **Branch** to our flow to perform the pivot.

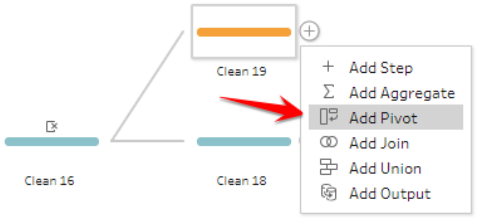
34. Add another clean step. Now click the plus icon next to the previous step and select Add Branch.

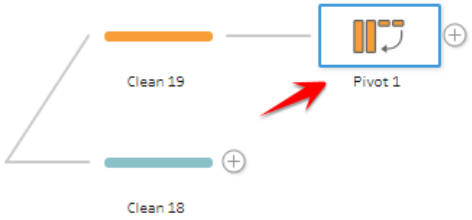




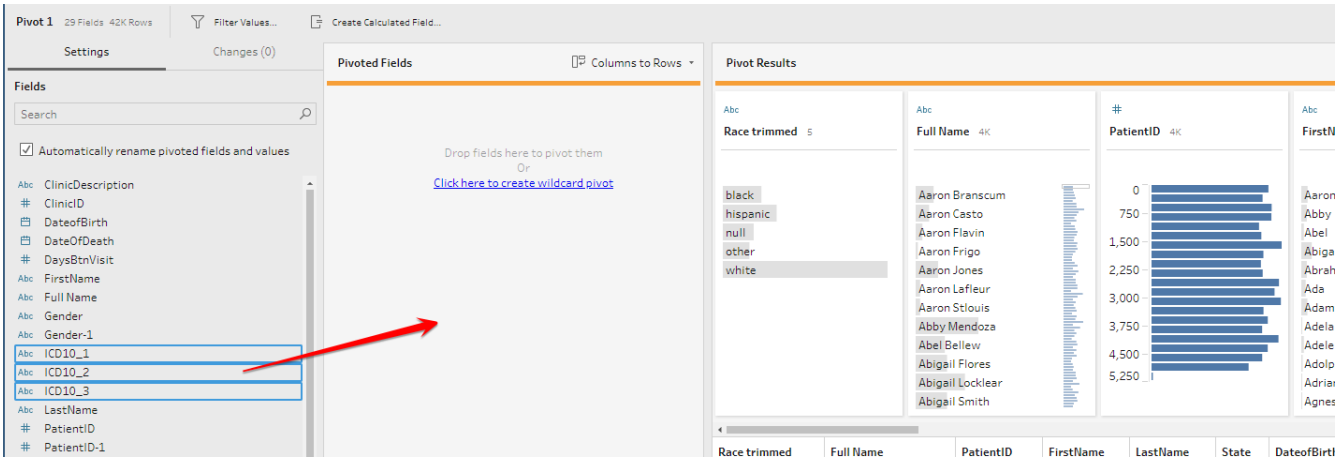
4.3. Adding a Pivot

1. Now click the plus sign on the clean step for the new branch and select Add Pivot



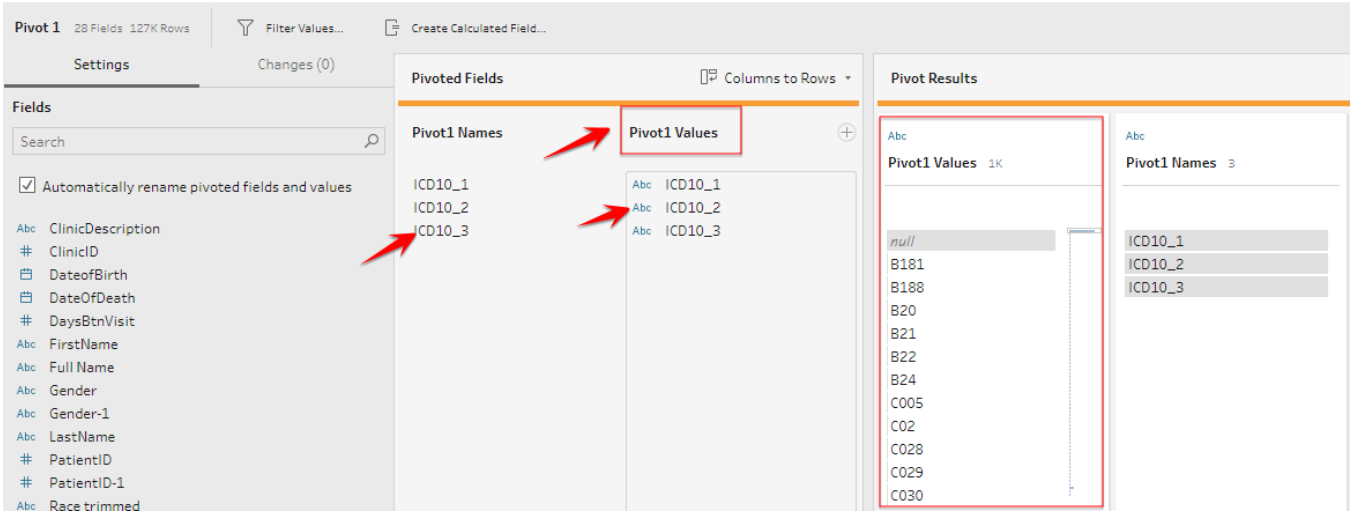


After you select the **Pivot** step a new profile window displays that requires us to make some selections.



1. Select ICD10_1, ICD10_2, ICD10_3 from the left-hand side of the pane and drop it into the Pivoted Fields column

Now all these 3 columns we selected would be merged into one column.

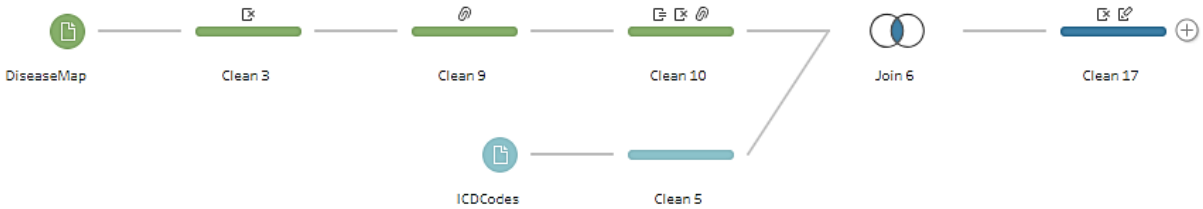


2. Double-click on Pivot Values and rename it to ICD10

Now we're going to join our clean **DiseaseMap** file to **ICDCodes**.

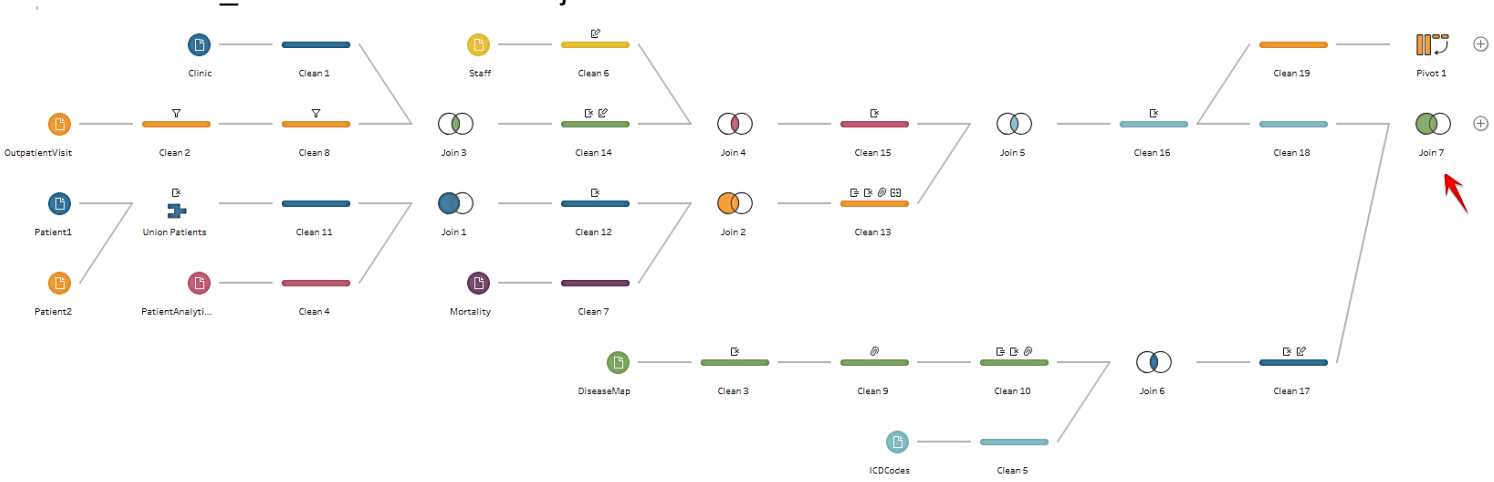
Steps

1. Click on the ICDCodes clean step and drag it onto the DiseaseMap clean step to create an Inner join
2. Add a clean step then remove the duplicate field ICD10-1.
3. Rename ICD10Descr to Condition Description



Our next step would be to join the two flows we have created together.

1. Drag the last clean step in the DiseaseMap + ICDCodes flow above and drop it onto the last clean step of Patient+ Visit data flow to create a join.
2. Change the join from an Inner to a Left join. By default, Tableau prep will use the clause ICD_1=ICD10 to create the join. We will leave it as is.

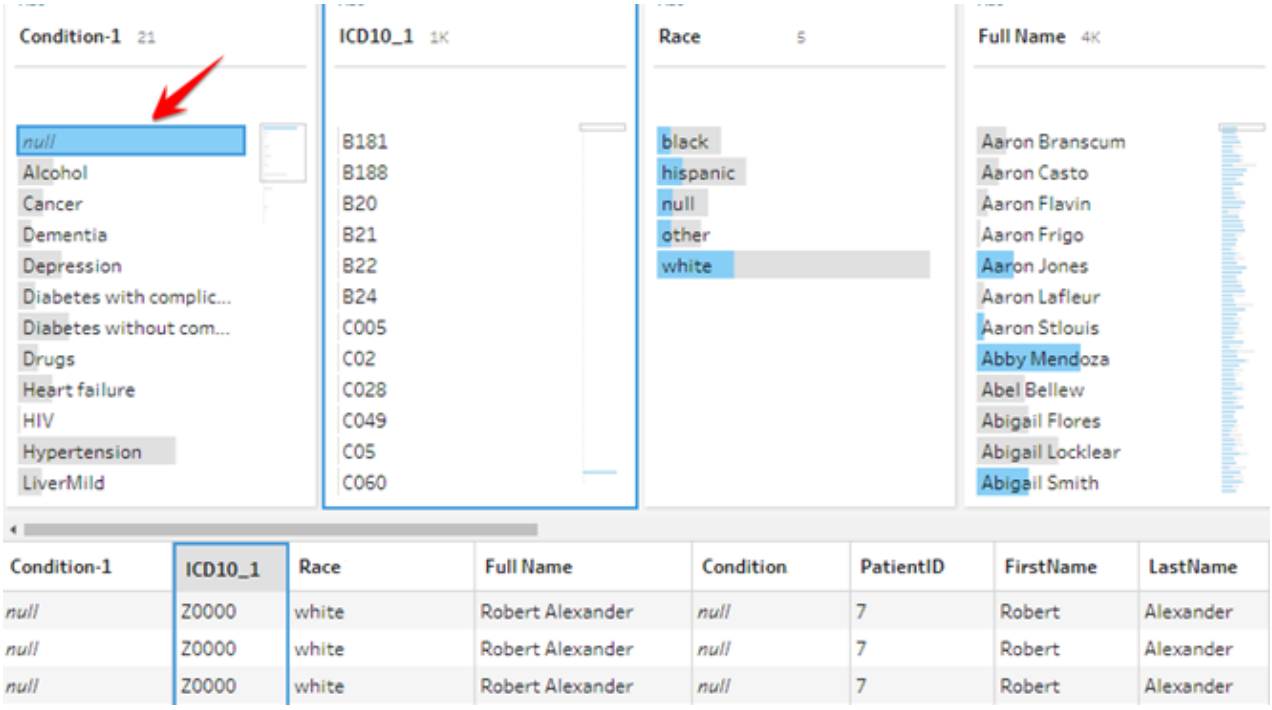


3. Add a clean step

Note that we're not using the new Branch we created yet.

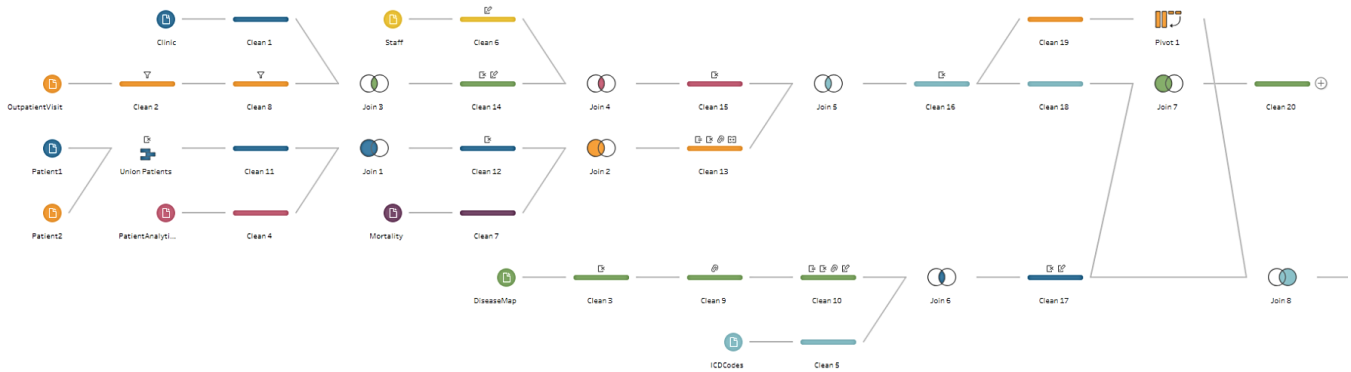
4. Click on the Pivoted ICDcodes step and join this with the DiseaseMap + ICDCodes flow.
5. Because we want to bring all values from the DiseaseMap + ICDCodes flow, we will create a Right-join here.
6. Add a Step

When you explore the data, you will discover that when the Condition column is null, the code is **Z0000**. This code simply means that no abnormal condition was found in a patient or that they are healthy.



Condition-1	ICD10_1	Race	Full Name	Condition	PatientID	FirstName	LastName
null	Z0000	white	Robert Alexander	*null*	7	Robert	Alexander
null	Z0000	white	Robert Alexander	*null*	7	Robert	Alexander
null	Z0000	white	Robert Alexander	*null*	7	Robert	Alexander

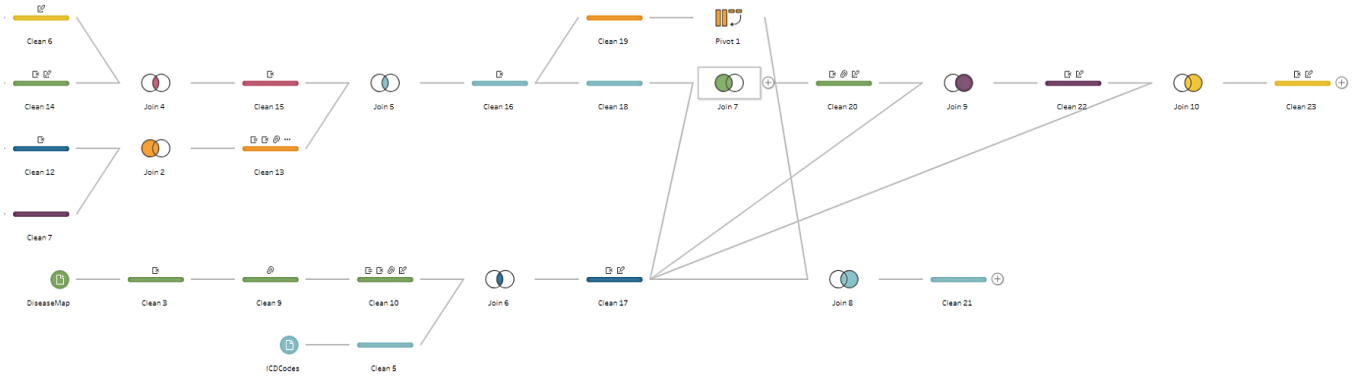
7. Rename Condition to Condition1 and Condition Description to Condition Description2.
8. Group null fields and rename them as “Healthy”



9. Our flows should now look like this. Note that no two flows may look exactly alike in shape. Now click on the clean step in the Patient + Visit flow path. Let us explore our data at this step.

What does the Nulls in the Condition column represent? Let us join this back to the **DiseaseMap + ICDCode** flow path to find more insight about these Nulls.

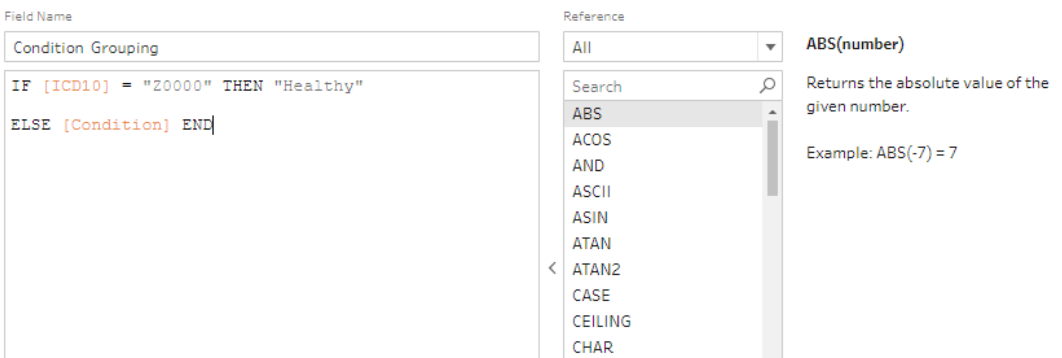
1. Change the default join clause to ICD10=ICD10_2 and create a Right join. Remember that this could be the other way around and you would be creating a Left join depending how your files are structured in the flow process.
2. Add a clean step and rename Condition to Condition2 as well as Condition Description to Condition Description2
3. Remove the ICD10 column
4. Drag the step you have just cleaned and drag it to the DiseaseMap + ICDCode flow path. This time use the clause ICD10=ICD10_3 and create a Right-join.
5. Rename Condition to Condition3 and Condition Description to Condition Description3
6. Remove the ICD10 column



7. Now that this part of the flow is complete, let us focus on the DiseaseMap +ICDCodes path and complete that flow. Since we've also joined this flow to the main flow through the Pivot branch we created, we need to fix the Z0000 issue here as well. This time we will use a calculated field instead of the Grouping option.

Steps

1. Click the clean step. Select the Create Calculated Field option at the top of the profile window and input the calculation below:



2. Remove the Condition column.
3. Rename Condition Grouping to Condition
4. Filter out the Nulls in the Condition column by right-clicking the Null value and selecting "Exclude."

Now that we have finished preparing our flows, let us learn to Generate Outputs

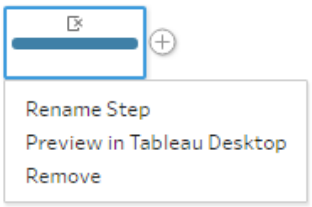
4.4 Rearranging Flows

Oftentimes when building complex flows, your diagram might look messy and it may be difficult to track the flow. In Tableau Prep, it is easy achieve this feature by simply dragging and dropping data sources or moving objects around. Doing this improves the overall look of your data prep diagram and makes it easy for the next person to track what you have done.

5. GENERATING YOUR OUTPUT

5.1. Preview your data

Now that we have completed our data cleaning process, it is time to generate our final output. Generating an output completes your flow process and you cannot add any step or clean your data further after this step. Before you generate your output, it is advisable to preview your cleaned data to ensure that everything is the way you want before you generate your extract. This is an important step because depending on the size of your data sources or the complexity of your workflow, running the flow might take a long time to complete. You wouldn't want to find out that there are errors to your output after waiting hours for it! To preview our data tight-click on the **Final clean** step and select preview. Tableau Prep would run your flow and automatically open your cleaned data in Tableau Desktop.



5.2. Finalizing your flow

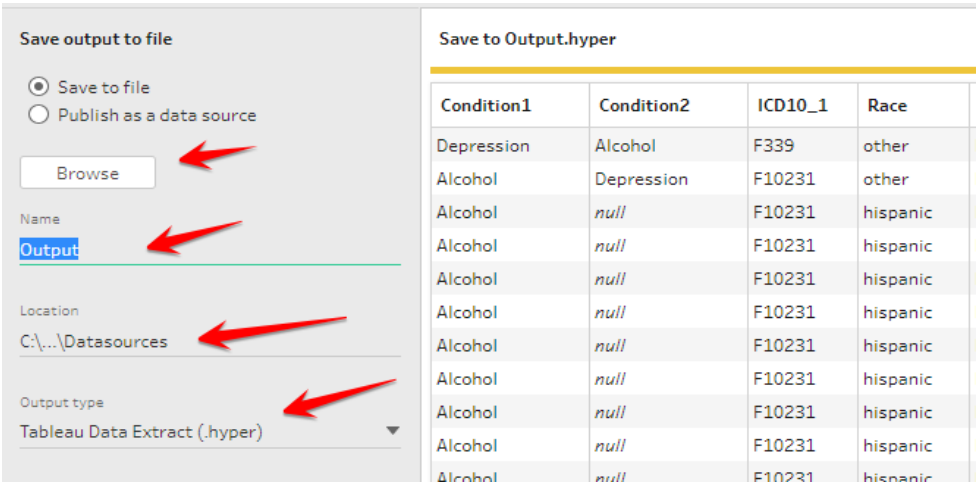
Now that we are sure the data looks exactly as we want, let us generate an output

Steps

1. This way we can explore our data before we generate an output. Now that we are satisfied that everything is correct, let's generate our Tableau extract.
2. Click on the final step in the Patient + Visit data flow icon
3. Click the plus icon and select Add Output.
4. Rename the Output step to Main Output

You will notice several options on the left-hand side of your output pane. We can use the **Browse** option to change the default directory also change the default extract name. By default, the output goes to My Tableau Prep repository in your documents folder.

5. Click on Browse and in the dialog, box change the name from Output to Main Healthcare Output. You also change the directory for your data source here.



We can also choose to change the output type from the default Tableau Data Extract(hyper) to a Tableau Data Extract (. hyper) to Tableau Data Extract(.tde) or a Comma Separated Values (.csv). We will leave it as the default hyper extension.

6. At this point you can click Run workflow from the Output pane or use the 'play' icon from your Output step to generate your output.

You can choose to complete all your flows then run them all together to generate the various outputs. Do not run the flow for now. Let us finalize the other workflows first.

7. Click the final clean step for Disease Map + ICDCodes flow path and add an Output step.
8. Rename your step Pivoted Healthcare Output

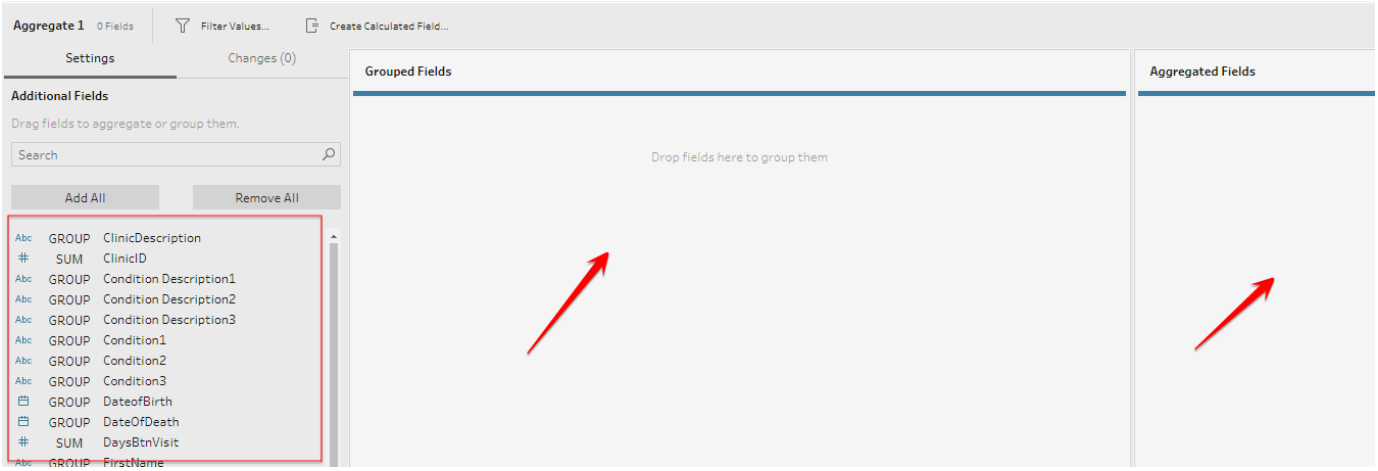
5.3. Aggregation

Now let's assume we want to generate an output that allows us to find the number of visits by month. We can generate this by aggregating our data at the month level of detail.

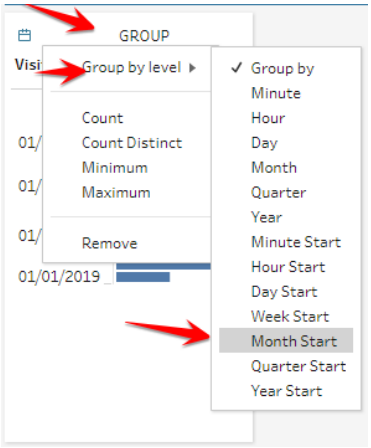
Steps

1. First click on the plus sign on the final clean step for our Main Output and Add a Branch.
2. Add a clean step.

3. Add an Aggregate Step. You should see an Aggregate profile window as shown below.

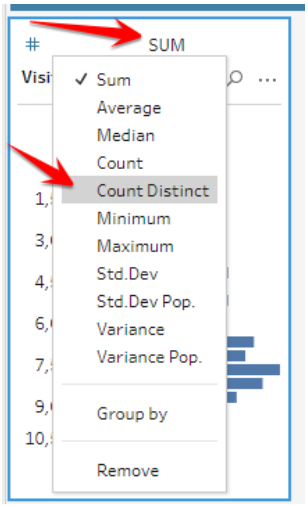


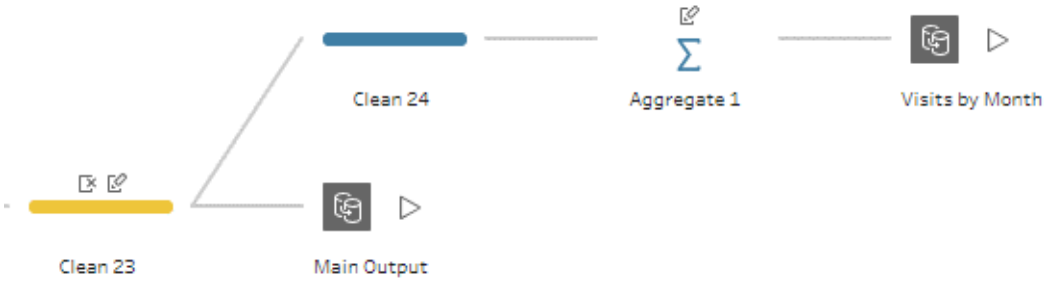
The left hand-side of the window are the fields that we will select to perform the aggregation. We will group **VisitDate** at the month level and perform a count distinct calculation on the **VisitID** to find the visits per month.



4. From the left-hand pane, drag GROUP VisitDate to the Grouped Fields window.
5. Click on GROUP and in the drop-down menu select Group level > Month Start

6. Now select SUM VisitID from the from the left-hand pane and place it on the Aggregated Fields column. By default, the aggregation would be a SUM. Because we want to find the number of visits, we have to perform a count of distinct VisitID's.
7. Click on SUM and in the drop-down menu select Count Distinct
8. Double-click on VisitID and rename it to Count of Visits
9. Add an Output step and name it Visits by Month. Your output should look like this:

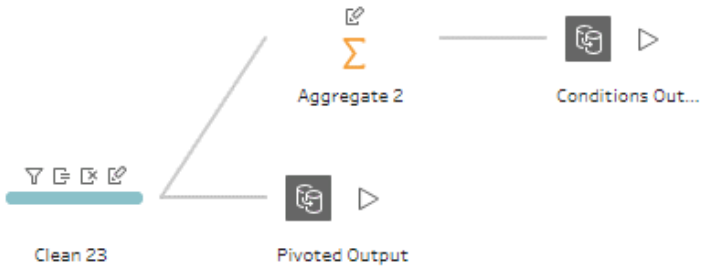




Now, let us assume we want to find the number of Visits by Condition by Year from the DiseaseMap + ICDCodes flow. We need to aggregate data here.

Steps

1. First add a Branch to the final clean step.
2. Add a clean step.
3. Add an Aggregate Step.
4. Drag Condition, Gender and VisitDate to the Group Fields column.
5. Change the VisitDate level of detail to Year.
6. Drag VisitID to the Aggregated Fields column. Change the aggregation to Count Distinct and rename VisitID to Count of Visits.
7. Add an Output Step. Your final output should look like this:



8. You can now run all your flows to generate your output by clicking on Flow > Run All from the Tableau prep menu. You can also use the play icon at the top of your Tableau Prep flow

5.4. Publishing your Output

In Tableau Prep, if you do not want to save your output onto your local machine, you may choose to save your output to the Tableau Server. To do this,

1. Add an Output step to your flow.
2. On the left-hand side of your Output profile pane, select Publish as a data source
3. Enter your Tableau Server credentials and publish your data.

Now your data had been cleaned saved to your computer or published on the server. You are ready to conduct your analysis.

6. ADVANCED FEATURES

We have now built out an entire complex Tableau Prep output joining in multiple data sources, data exploration, cleaning and grouping data, filtering, aggregating data, and then published our data source. The following section is going to cover more advanced functionality of Tableau Prep as well as tricks and workarounds discovered by the community. We will also cover how to automate your Tableau Prep flows so that you are always working with fresh data.

6.1. Advanced Pivoting: Rows to Columns Pivot

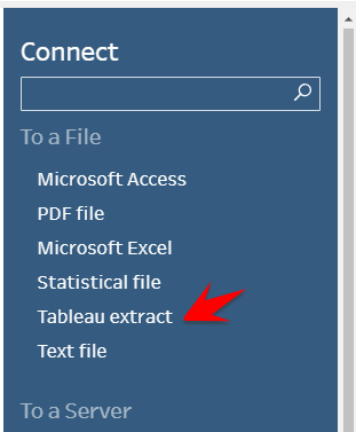
In a previous section, you learned how to transpose values in columns into rows making your data narrower and longer. As a more advanced pivot, we are going to take values in individual rows and pivot them as column values accomplishing the inverse from pivoting columns to rows – making the data wider and potentially shorter depending on what we plan to do after we pivot.

Why this is useful: We can now do filtering on specific values in Tableau. A prominent example would be analyzing survey data. Say your first few questions were demographic information about that customer. We can now translate that information into columns to show in every row across our dataset allowing us to filter by gender, race, income or whatever fields we collected. This demographic info could also be used across our organization for other analytical purposes. Another excellent example for this is for aggregations of specific values in a column. This is the example we are going to tackle next: showing an aggregated number of visits for each type of our clinics at the patient level.

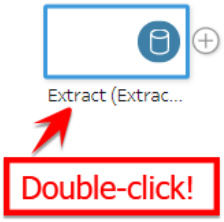
Steps:

35. Ensure that everything we have done up to this point is completed and you have run the main output (not the pivoted one) saving it somewhere that you can find it.
36. Open up a new Tableau Prep file.

37. Click connect to data, select Tableau extract, navigate to the file, and open it up.



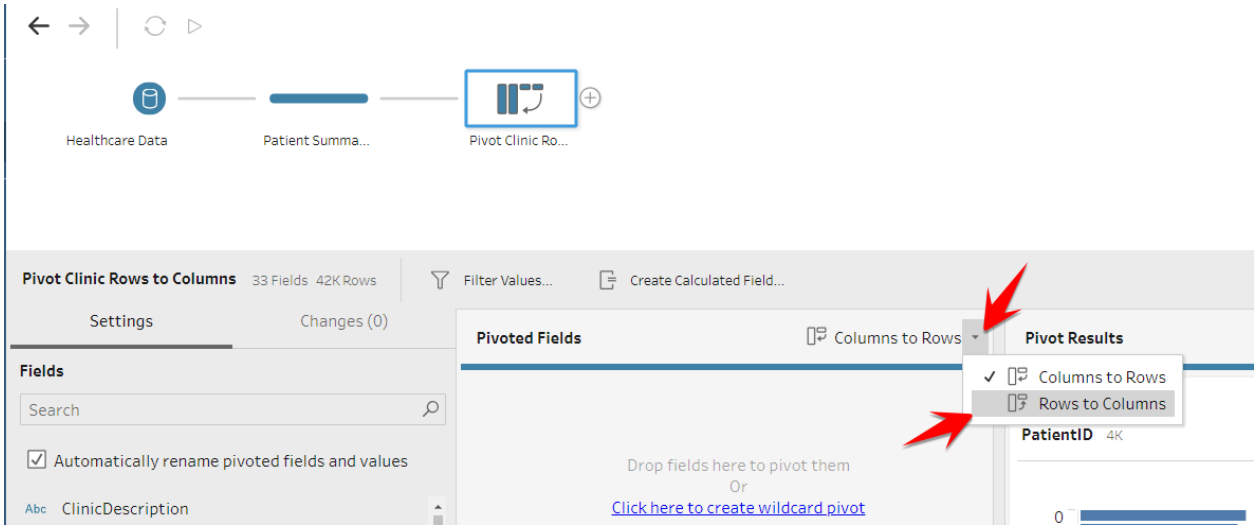
38. You will notice in the beginning of this manual the CSV files were all named in the main window. For whatever reason, when a Tableau Extract is pulled in it will give it a name of Extract (Extract.Extract). This isn't very user friendly and can become a mess if you are working with multiple Tableau Extracts and they are all named Extract. Double-click on the name beneath the icon and rename it to Healthcare Data.



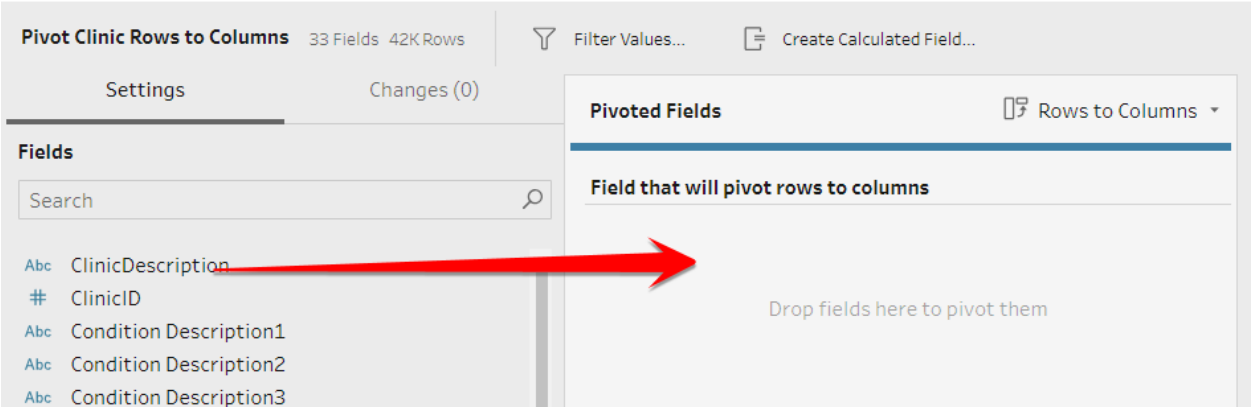
39. Create a new clean step. We are not going to do anything in this clean step, but it helps us compartmentalize what is happening as will make another branch of the main output in a later section. Call it Patient Summary.

40. Click the + icon and select Add Pivot. We did this step when we pivoted columns to rows earlier. Name it Pivot Clinic Rows to Columns.

41. In the middle column Pivoted Fields, you will see a dropdown menu that currently is Columns Rows. Select the dropdown and switch it to Rows to Columns.

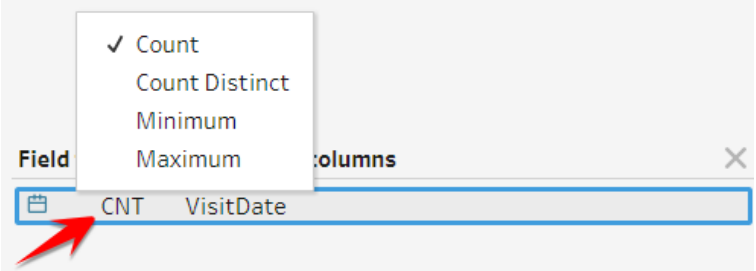


42. The middle column will change slightly. Let's see it in action. Find ClinicDescription and drag it onto the top pane in Pivoted Fields.



43. Tableau Prep will give you an angry red icon in the top-right hand corner because it now needs to know what to aggregate the fields by. Find VisitDate and drag that into the lower pane.

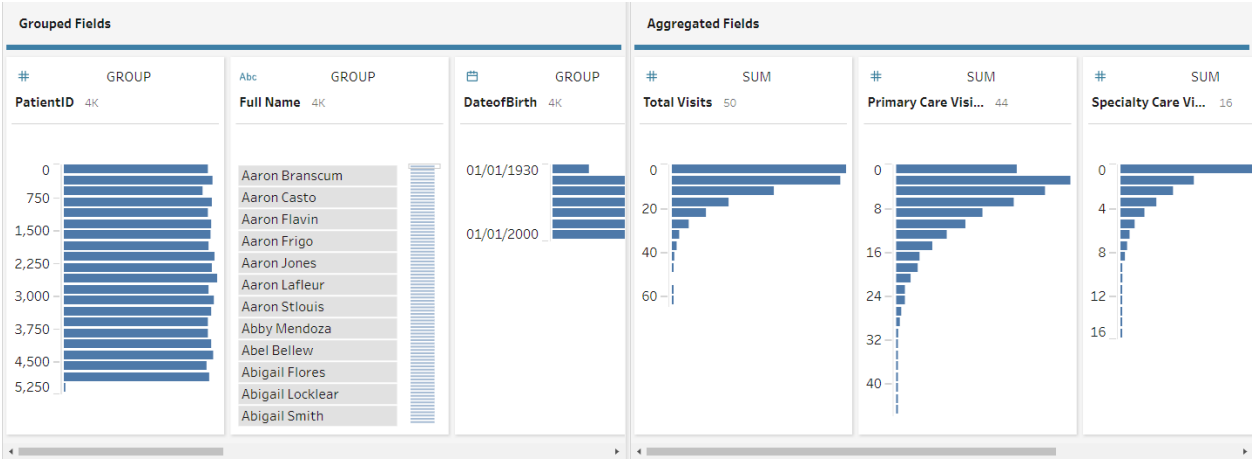
44. Note that you can change the data type and how it is aggregated selecting either icon as shown:



45. You will see that it loaded three distinct values of Primary Care, Specialty Care, and Emergency Department which translates to three additional columns.
46. What happened here? Each visit has a type of clinic associated with it. Some visits are to primary care, others are specialty visits, and others are to emergency room/emergency visits. We want to get a summary of all visits at the patient level instead of each individual visit. In Tableau Prep, we tell it to take each of those clinic types and put them onto columns. We use a CNT or CNTD of VisitDate to give us a 1 if that was the type or a 0 if it wasn't.
47. Before we move on, create a calculated field that will denote Total Visits. Select Create a Calculated Field. Add up Primary Care, Specialty Care, and Emergency Department. In its current state, it isn't very useful to us, but will be in the next step.

48. If you browse through the data exploration panes you will see that we are still at the visit level. Let's put to use our skills in aggregation to get this data at the patient level. Click Add Aggregate and call it Aggregate to Patient.
49. In the Grouped Fields section add:
50. PatientID
51. Full Name
52. Date of Birth

53. Date of Death
54. Gender
55. Race
56. State
57. Zipcode
58. In the Aggregated Fields section:
59. Total Visits (that we created in step 13)
60. Primary Care
61. Specialty Care
62. Emergency Department
63. By default, the aggregations should all be SUMs, but ensure that each field is a SUM. This step summed up all the 0s and 1s and our data is at the patient level now along with other relevant patient data.
64. I renamed the original three types of care to have Visits at the end of their field name to be more descriptive. It should look something like this:

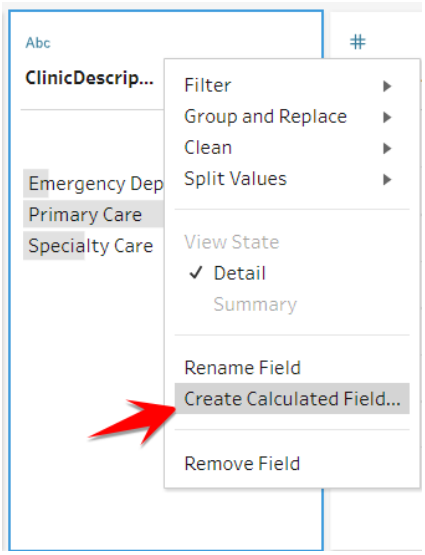


65. Feel free to explore the data now. What jumps out at you? Some people have *a lot* of visits to the hospital.
66. Create an output step and run it saving to somewhere accessible.

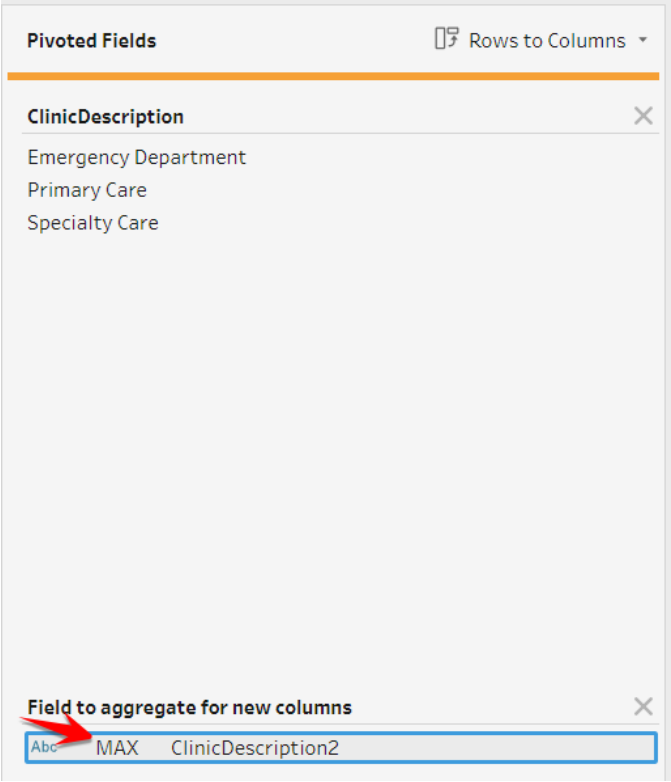
6.2. Pivoting Rows to Columns with Categorical Data:

Before starting the how-to tutorials, we saw an example of how to transform demographic data (that was answered in a survey question) into a column form. This is text data as opposed to numeric data that we just finished pivoting. If you noticed our original field disappeared once we put it in the pivoting section. Here is a workaround to get the text data to show up. Note that I can demonstrate the concept on this data, but it would not make practical sense for this dataset (should be fine for your data though!).

2. In our first clean step, create a calculated field that duplicates ClinicDescription. You can do this by right-clicking the field and then selecting Create Calculated Field. Name it something like ClinicDescription2.



3. In your pivot step, you put ClinicDescription in the first Pivoted Fields section and then put ClinicDescription2 in the bottom half of Pivoted Fields. Set ClinicDescription2 to MAX. MAX is a trick that you can use in Tableau Prep or Tableau Desktop to have individual text values appear.



4. Instead of 1s and 0s, you will now see the clinic type or null. You will need to do an aggregate step to get it at the level of granularity that you want. Change the aggregate from a SUM to a MAX. Aggregates will replace the null values with the text value. In my survey example, you will now see their demographic information in every column across each individual's responses.

6.3. Using the Wild Card Search

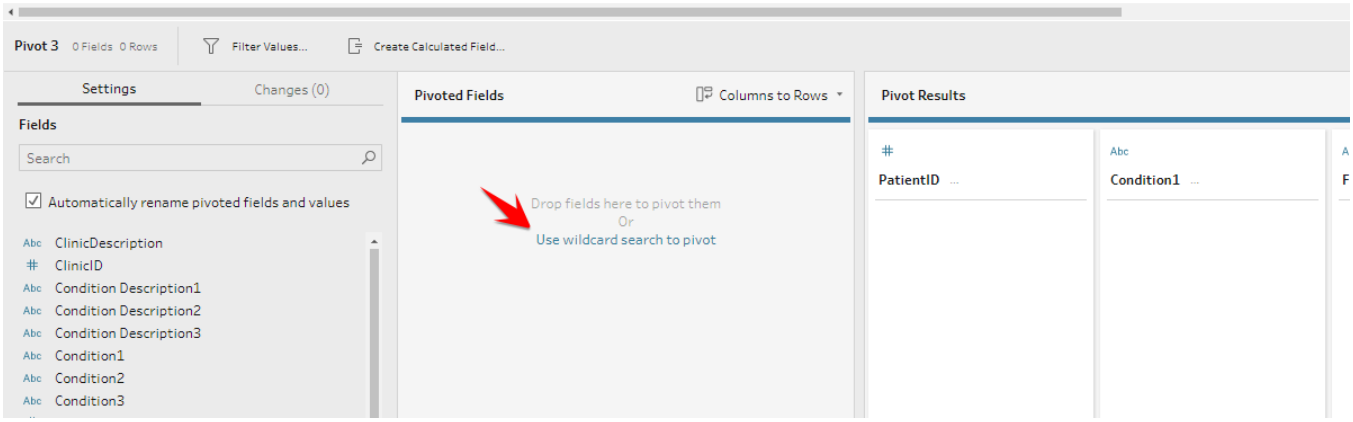
The wild-card search enables you to pivot columns to rows instantly when you're working with a large data set that changes a lot over time. The wildcard search option pivots your data based on a wild-card pattern match so that when new fields are added or removed from your table, Tableau Prep recognizes the change to your schema and updates your pivot results when you run your flow.

Steps:

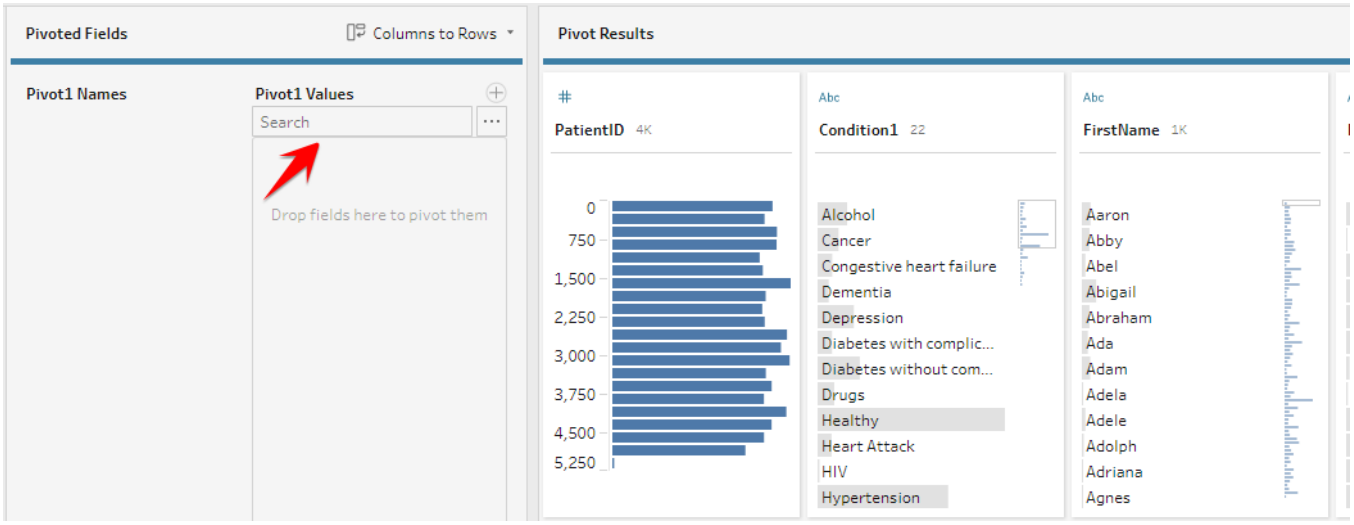
1. Connect to the Healthcare data source from the connections pane

2. Click on the plus icon and select **Add Pivot**

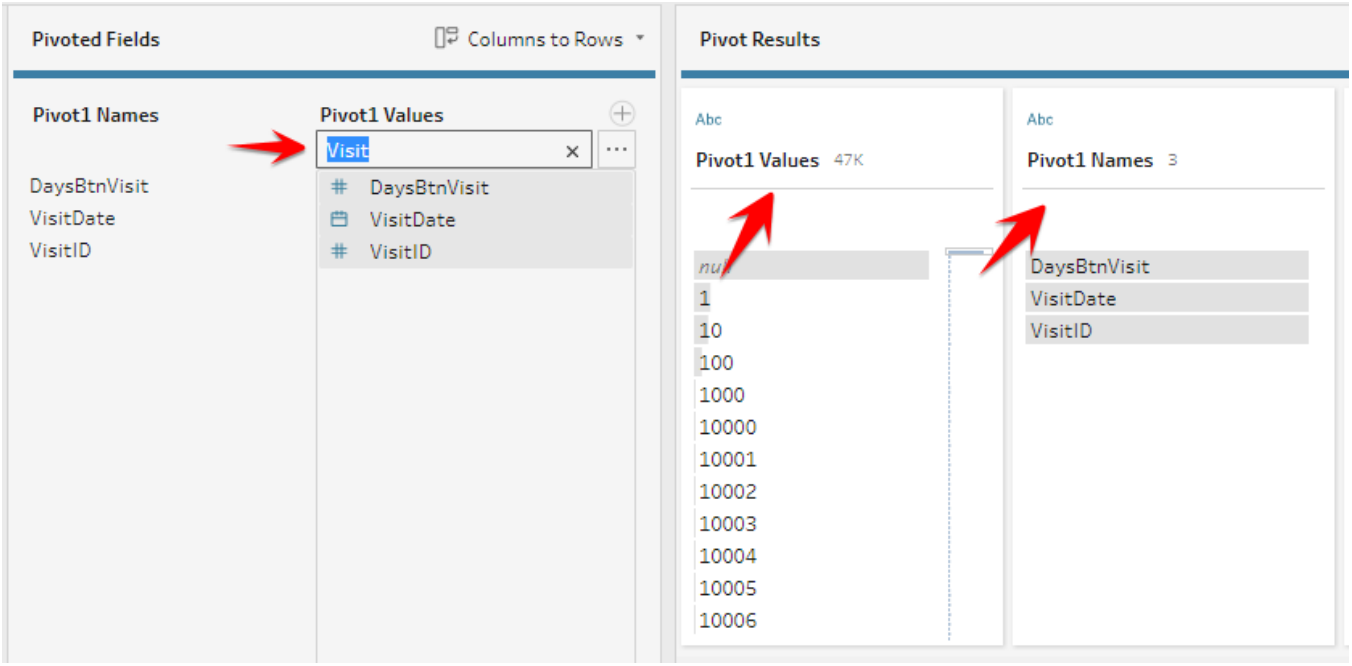
3. In the Pivoted fields pane that pops up, select **Use wildcard search to pivot**



4. In the Pivot fields search window, enter a value or partial value of the field you want to use to search for. Type "**Visit**" into the search window.

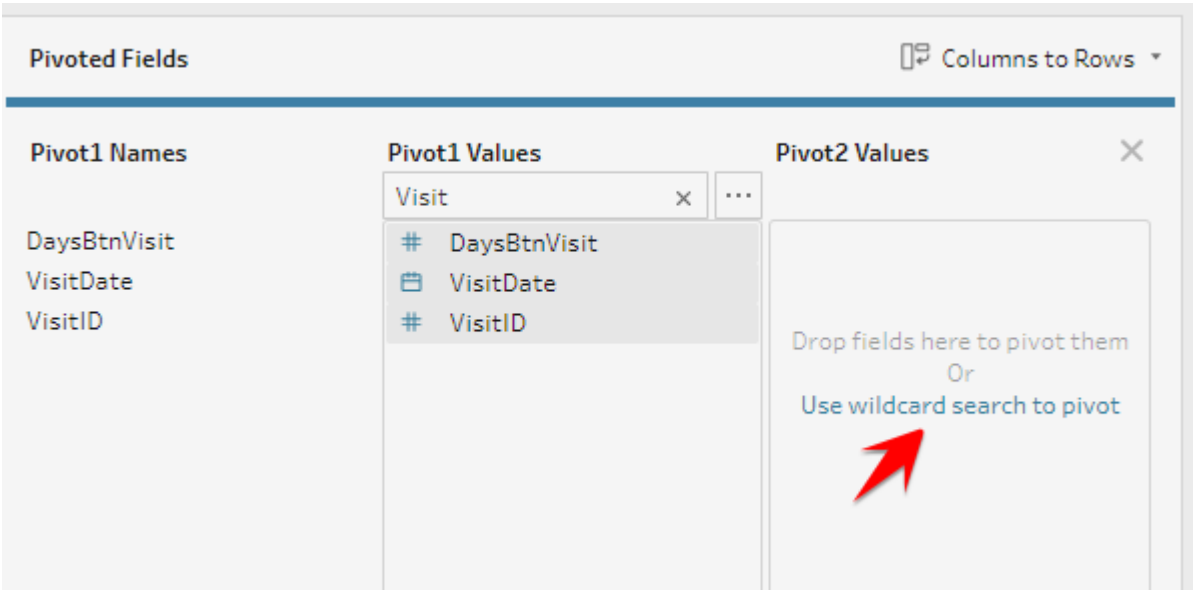


Notice in the screenshot below how Tableau performs the pivot using the keyword we put into the search window.



We can choose to search for more fields to add to the pivot. To do this

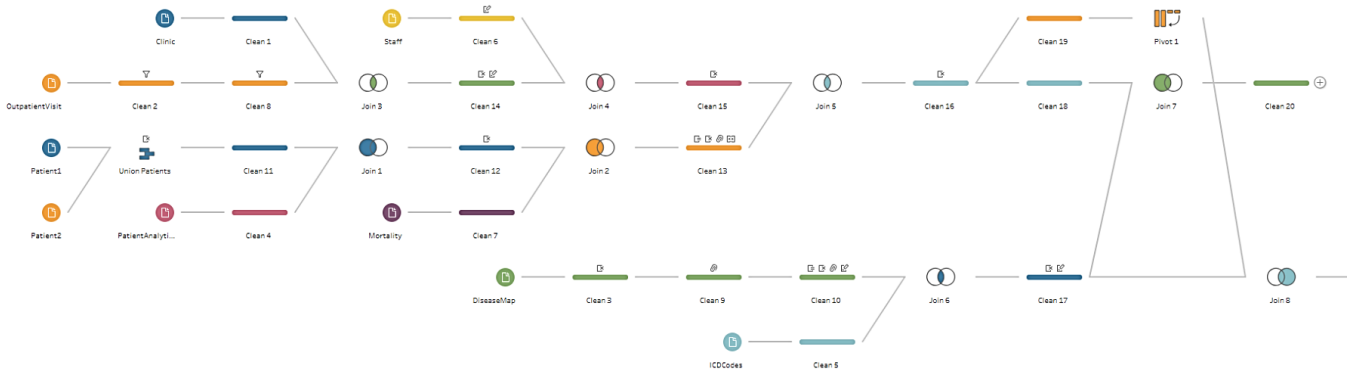
5. Click the plus icon next to the **Pivot Values**. You will see a new window for **Pivot Values 2**. Type in the field you want to include to your pivot



In some situations, Tableau Prep may fail to recognize the field name you type into the search window. If this happens, modify your keyword for more accurate results. This approach to pivoting is very dynamic because any new data added to your data source that matches your wild card search pattern is automatically detected and included in your pivot results when you run your flow.

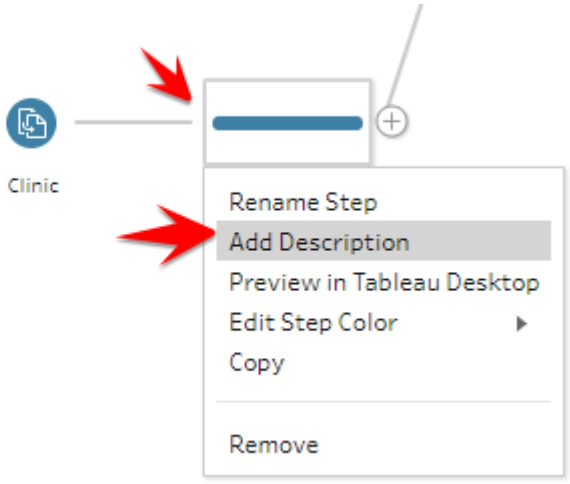
6.4. Adding Descriptions to Flows

Adding descriptions to your Tableau Prep flows is a useful way in documenting your process to serve as a walkthrough for whoever wants to reuse your flows. Imagine looking at the screenshot below in an attempt to figure out a complex calculation that was used to duplicate it in your work. This would be tedious and take a lot of time to figure out. Use descriptions can help you to spend less time for you and your organization in its data preparation process especially when your prep process is often reused.

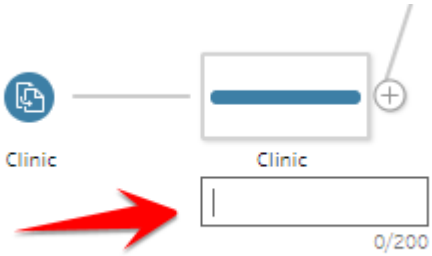


Steps

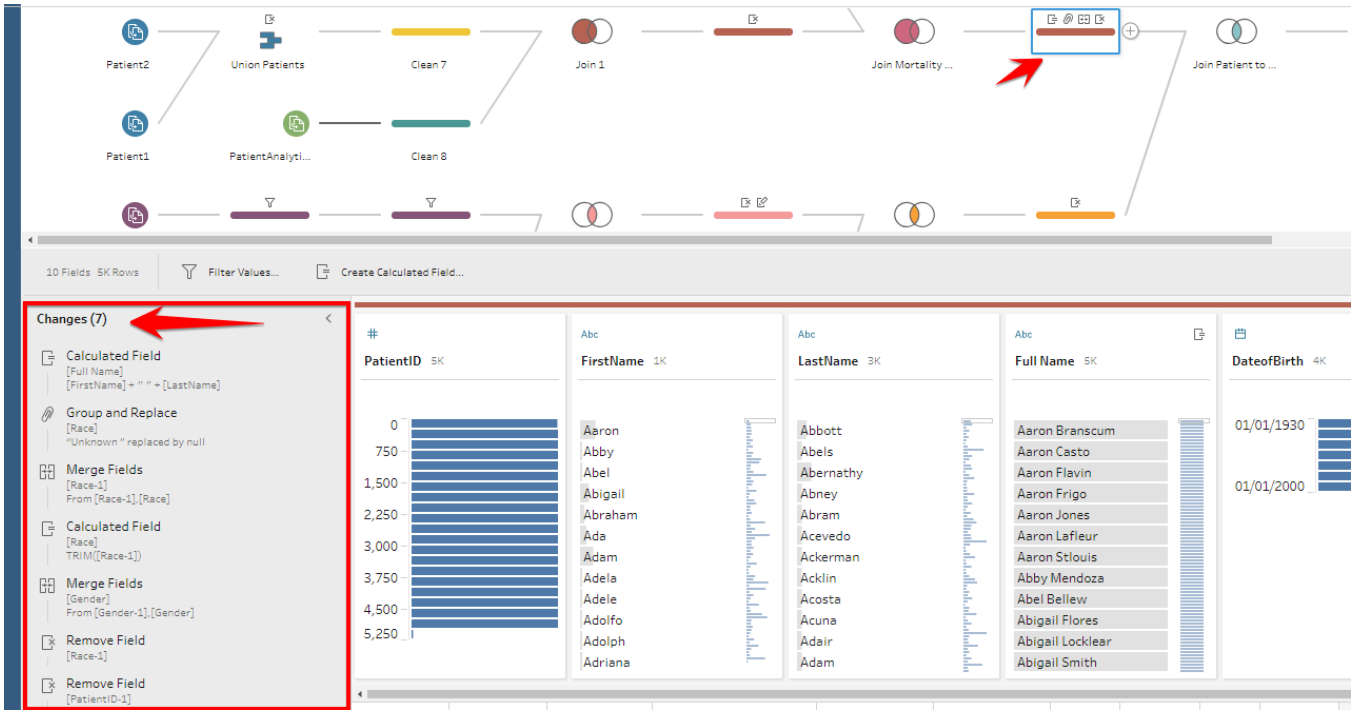
1. To create a description in your flow, right-click any step in your flow and select Add Description



2. In the text box that opens, type in your description for that particular step you are on in the flow. This could be anything regarding the changes you have made to the flow.

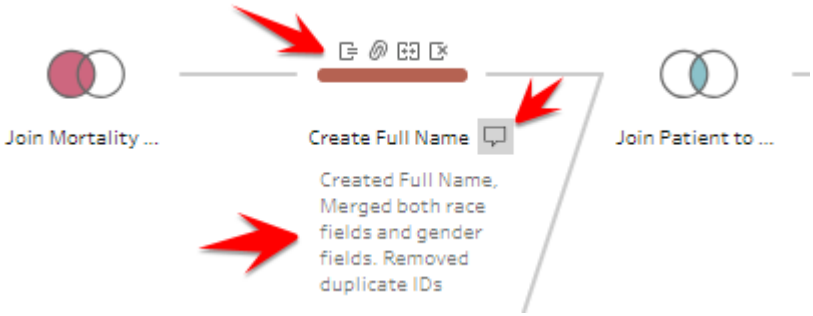


In the screenshot below, we will attempt to write a short description of all the steps we have taken at this clean step

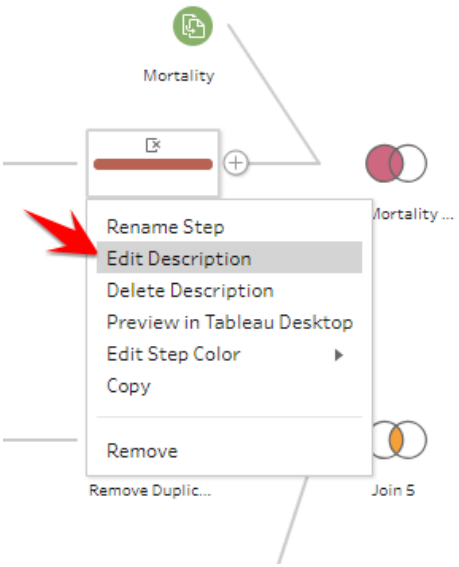


3. Notice at this stage of our flows we have made 7 different transformations or changes to the data. Now type in a description that summarizes these transformations.

Your summary would like the screenshot below



4. To edit your description, right-click on the step and select **Edit Description**



6.5. Identifying Data Granularity and Level of Detail Calculations:

Sometimes you want to compare a value or see a record in its current level of granularity compared to something at a higher level of granularity. This is the equivalent of using fixed level of detail calculations in Tableau Desktop, which Tableau Prep also supports.

Why this is useful: Say you want to compare a sub-category's total sales to the parent category's total sales at the row level, you could use a fixed level of detail calculation to sum up the parent category's sales and append it to the sub-category row. Another possibility is that your HR department administers company compliance training with videos and quizzes. Each quiz requires a certain number of questions to be passed, otherwise it requires you to continue to retake it until you have a passing grade. This would help in finding a user's most recent passing attempt for a summary report to HR. Or, in our case, we are going to find a patient's first and last visit date and then add those as columns in our patient summary data that we created in the previous section, to give a more complete image of each patient.

Understanding Level of Detail Calculations:

Level of Detail calculations have existed in Tableau Desktop for several years, but were added into the May 2020 release of Tableau Prep. In Tableau Desktop, the formula names included

FIXED, INCLUDE, and EXCLUDE. Tableau Prep only has FIXED, as INCLUDE and EXCLUDE are dependent on filters in the view of what is being built in Tableau Desktop.

Level of Detail calculations as a concept can be confusing for many, so let's do a brief review. (A deeper dive can be found here: https://www.tableau.com/learn/whitepapers/understanding-lod-expressions)

FIXED follows a format of { FIXED DATA_PARTITION : AGGREGATED_DATA }. If you are familiar with sub-queries in SQL, these can be similar. The DATA_PARTITION part of the calculation defines a section of data that will be examined when performing an aggregation. If we don't specify a partition it will look at all data, otherwise it will calculate the aggregation for each partition. AGGREGATED_DATA is how we aggregate the data up. There must be an aggregation specified, whether that is a SUM, AVG, MAX, MIN, etc. Looking at an example, say we had simplified patient visit data as follows:

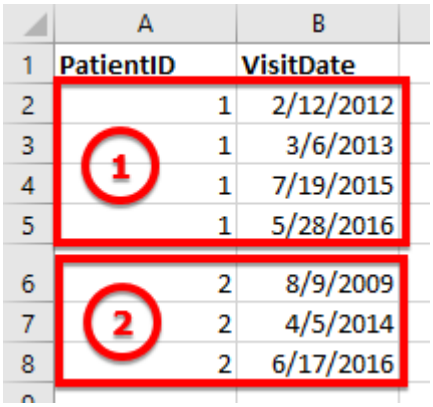
	A	B
1	**PatientID**	**VisitDate**
2	1	2/12/2012
3	1	3/6/2013
4	1	7/19/2015
5	1	5/28/2016
6	2	8/9/2009
7	2	4/5/2014
8	2	6/17/2016
9		

Two patients with multiple visits a piece. Our task is to find the first visit date for each patient. If we constructed a FIXED calculation that looks like this:

This calculation looks across all patients and finds the minimum visit date. The resulting data would look like:

	A	B	C
1	PatientID	VisitDate	First Visit Date
2	1	2/12/2012	8/9/2009
3	1	3/6/2013	8/9/2009
4	1	7/19/2015	8/9/2009
5	1	5/28/2016	8/9/2009
6	2	8/9/2009	8/9/2009
7	2	4/5/2014	8/9/2009
8	2	6/17/2016	8/9/2009
9			

We are on track, but not quite there. If we add in Patient ID into the DATA_PARTITION part of the calculation, now we will be finding the minimum date within each patient partition. Our partitions would look like:

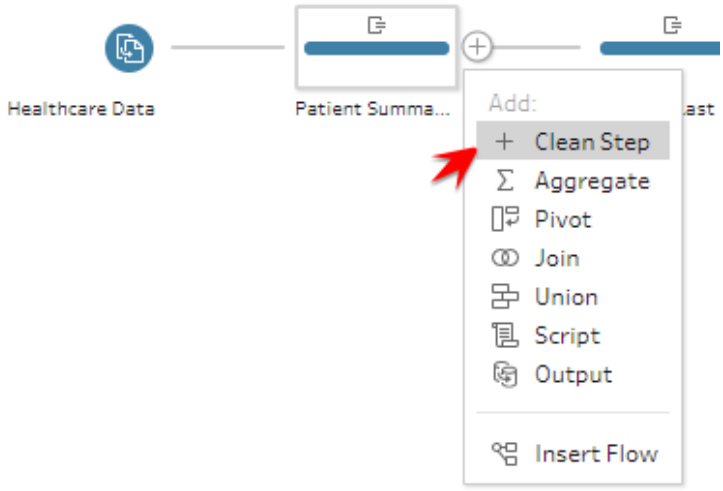


All visits for a given patient will make up the partition. Now our resulting data looks correct:

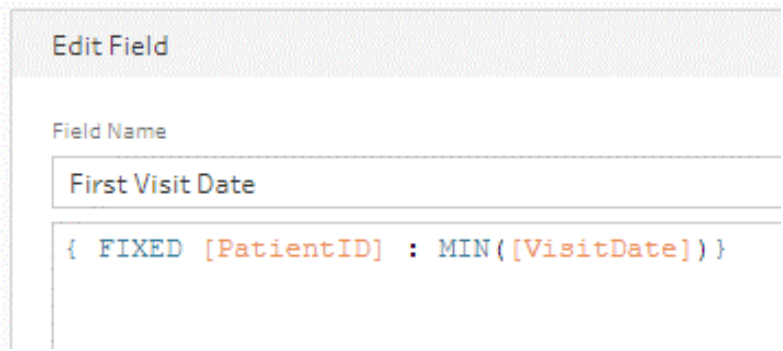
	A	B	C
1	PatientID	VisitDate	First Visit Date
2	1	2/12/2012	2/12/2012
3	1	3/6/2013	2/12/2012
4	1	7/19/2015	2/12/2012
5	1	5/28/2016	2/12/2012
6	2	8/9/2009	8/9/2009
7	2	4/5/2014	8/9/2009
8	2	6/17/2016	8/9/2009

Steps:

1. This builds off the file that was used in section 6.1. At the first step where we brought the file in, click the + after the first clean step called Patient Summary to insert a new clean step.

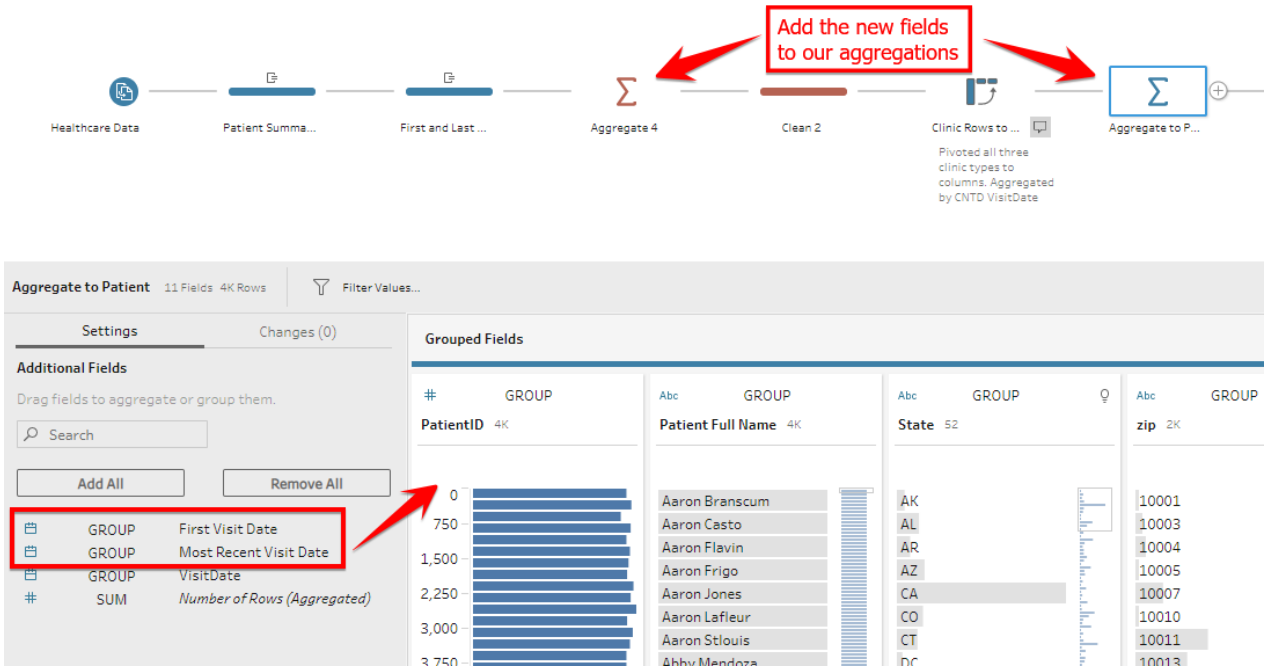


2. Create a new calculation called First Visit Date (Notice that the PatientID column is our partition as previously described):



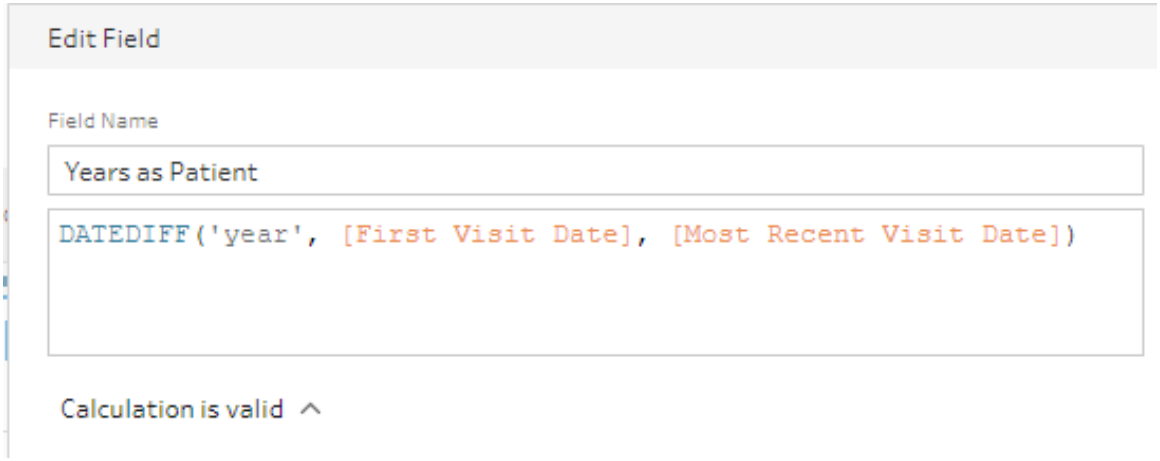
3. An additional calculation following the same format, but this time calling it Most Recent Visit Date and using a MAX instead of a MIN.

4. The new fields that were created need to be added into the aggregations that we created in section 6.1. Add both of those fields into the Grouped Fields section of the aggregation steps.



5. Explore the data for a second at the last node in our flow. We have total visits along with the type of visit broken out by patient. We also now have their first and last visit dates with us. What do you see? Have some patients been with us forever with a high volume of visits? Are there other more recent patients that also have a high number of visits?

6. Let's add one additional item for more insight. We are going to use the DATEDIFF calculation which calculates the difference between two date periods. It can calculate years, months, days, down to the second. In our case, we are going to calculate the number of years a patient has been seeing us.



7. DATEDIFF has three values that get passed to it. The first is what date period you want to calculate the difference in. The value goes in quotes. The second value is beginning date and the last passed value is the most recent date.

8. How are our years as patient distributed now? Which year groups have the most patients in them?

6.6 Using RANK and ROW_NUMBER Calculations

Rank assigns a rank to each row based on an aggregation (SUM, AVG, etc.) while the row_number formula assigns a sequential row number. At a basic level, a rank can assign a rank value to the largest value down to the lowest value (or vice versa) which is useful by itself. New possibilities beyond this emerge to augment your existing data. We can rank across all of our data, but we can also rank within a partition opening up interesting information that can be added to our data. These calculations act similarly to Table Calculations within Tableau Desktop.

Why this is useful:

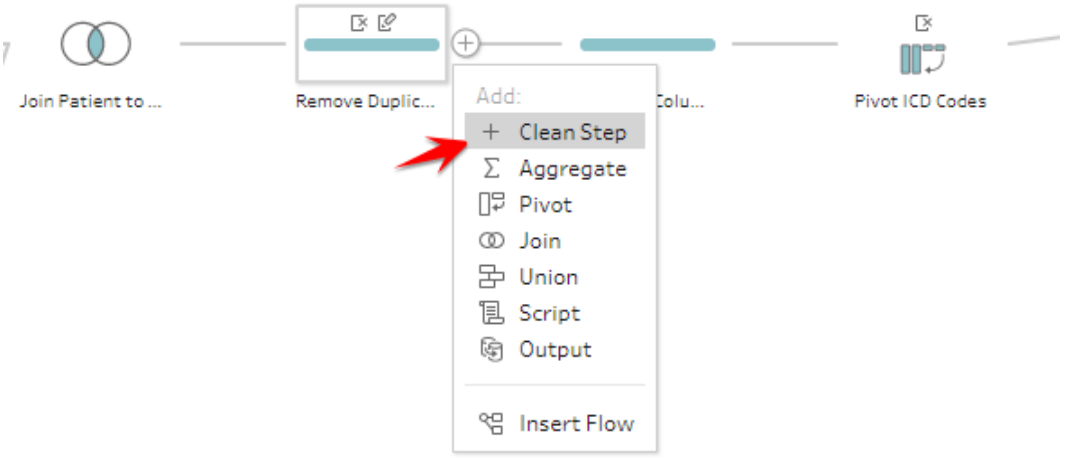
Say you are in the retail or ecommerce space. Each week or month you track your most popular products with their ranked value of total sales. In addition, you want to track how the rank changes over time between each period (i.e. Product X went up 2 ranks, Product Y went down 10 ranks, etc.). Tableau Prep can perform both of these operations.

Duplicate data is often a problem that data analysts, engineers, and scientists face in their data wrangling tasks. Removing duplicate data becomes easier in Tableau Prep by using row numbers along a partition of the primary key (or whatever column is being duplicated) and removing any values besides the first one.

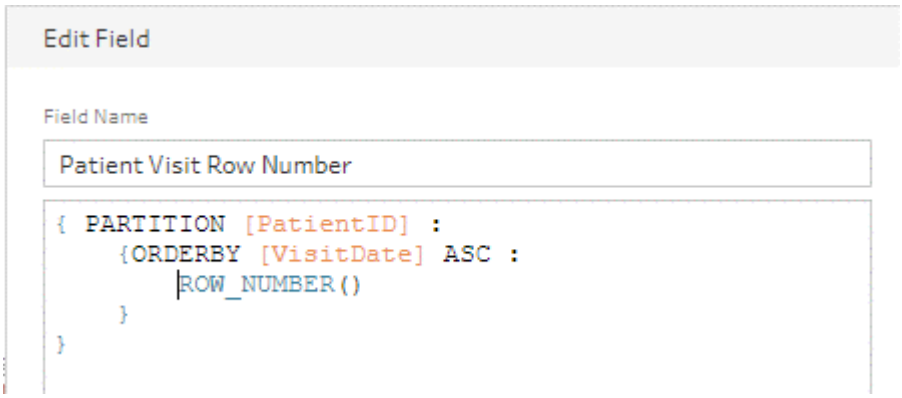
Perhaps you are a Healthcare Administrator and you are wanting to track the average days between visits to your medical facilities by your patients. Or, you desire to run patient medical history through a predictive model to predict the relationship between days between visits and their medical conditions to help in identifying at-risk patients. We will be calculating the number of days between visits for our patients in this example.

Steps:

1. We are going to return to our original file, before we worked on FIXED calculations and row to column pivots.
2. Right before the point of pivoting our ICD10 codes to rows, we are going to add some steps.

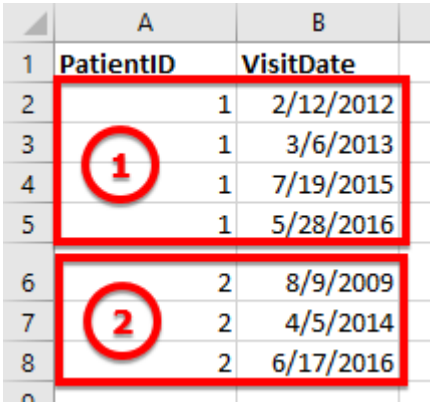


3. Create a calculation as follows:



4. There are four parts to this calculation

PARTION: Partition defines what our window will be. Using our example from the Level of Detail calculations, all visits for a given patient will be our window or partition. In the following image, Patient 1 has 4 visits and Patient 2 has 3 visits.



ORDERBY: This defines how the rank or row number will be assigned. In our example, we are ordering by each VisitDate. If it is was retail data, it might be ordered by sales. If I was looking at HR data and how long each employee was with the company, I might order this by the length with company.

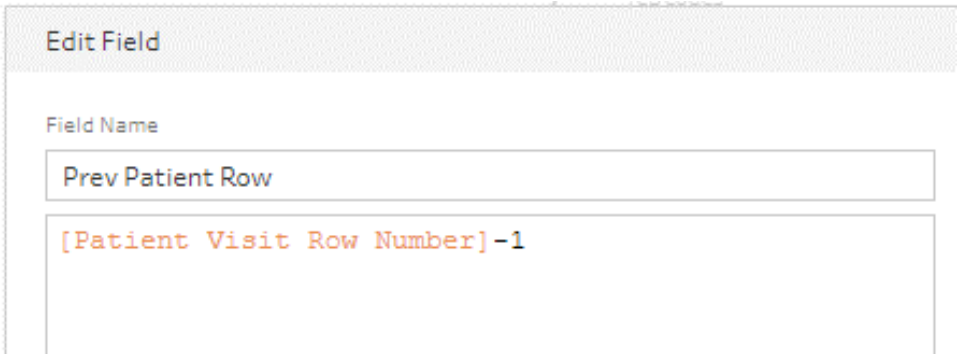
ORDER TYPE: Two options of either **"ASC"** or **"DESC"** which orders the data in ascending or descending order. We choose to order ascendingly, as I want the first dates

to be ranked first for other calculations that we will add in the next steps.
RANK TYPE: There are 5 options to choose from here, **RANK(), RANK_DENSE(), RANK_MODIFIED, RANK_PERCENTILE(),** and **ROW_NUMBER().** (More information about these can be found here: https://help.tableau.com/current/prep/en-us/prep_calculations.htm#supported-analytic-functions). In a nutshell, Rank and Row Number do similar things. If you only need sequential numbering, use ROW_NUMBER. If you need competition ranking, use RANK.

5. Our calculation is giving a row number to each patient's visit ordered by the visit date. Using the same example, this would be our resulting data:

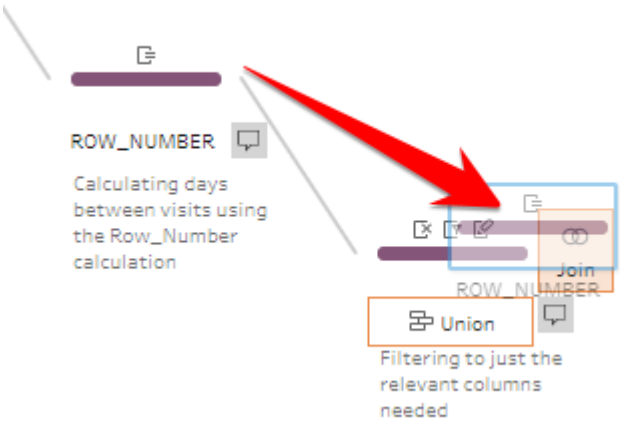
PatientID	VisitDate	Patient Visit Row Number
1	2/12/2012	1
1	3/6/2013	2
1	7/19/2015	3
1	5/28/2016	4
2	8/9/2009	1
2	4/5/2014	2
2	6/17/2016	3

6. Notice that the ranking or numbering restarts with each patient.
7. Call this clean step "Row Numbering"
8. Create an additional calculation with the following formula:

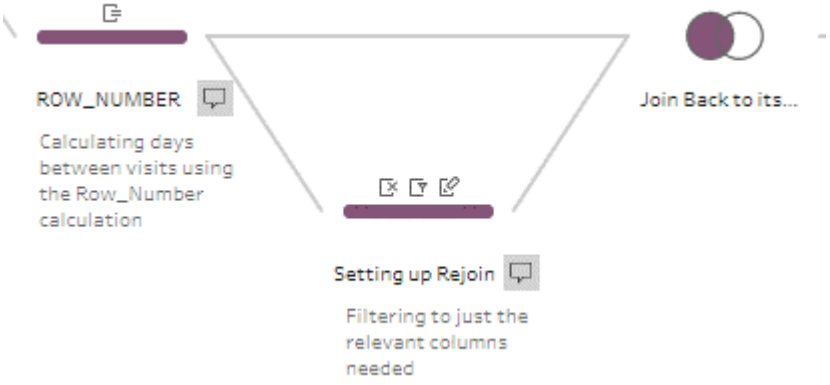


9. This new column will be used to join the previous row's visit date to itself to calculate the days between the previous visit.
10. Create a new clean step after the one we just created and call it "Setting up Rejoin".
11. Keep only the columns "Patient Visit Row Number", "PatientID", "VisitDate". Rename VisitDate to "Previous Visit Date".

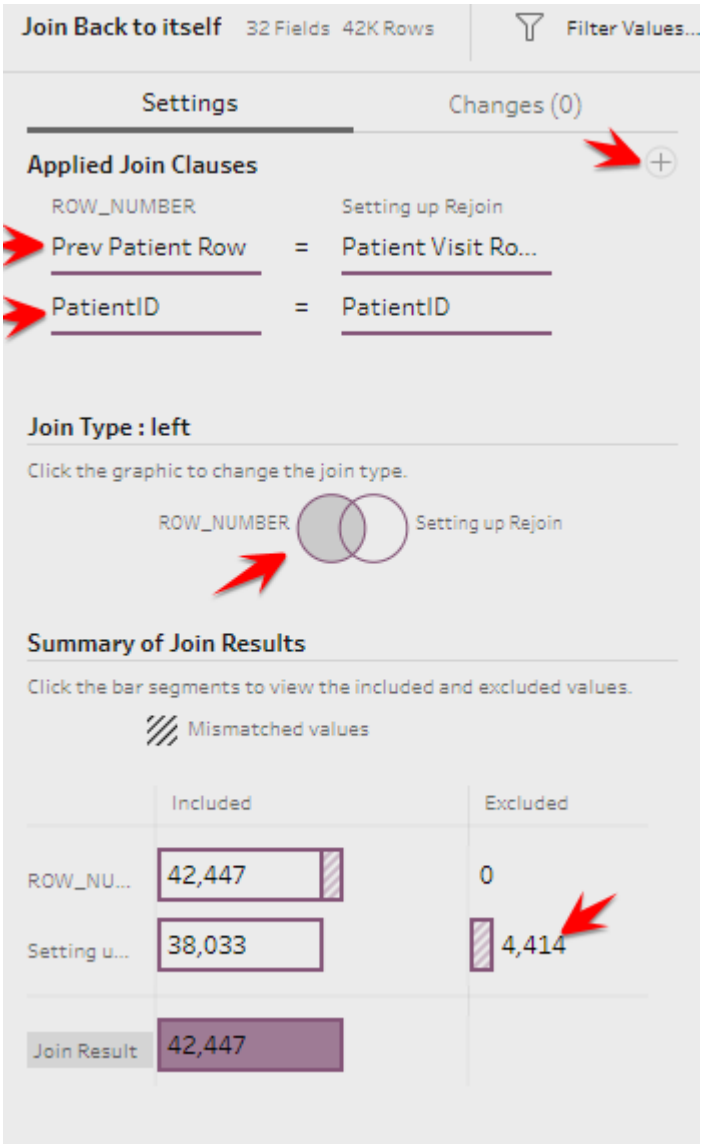
12. Drag the Row Numbering Node onto this new clean step that we created to create a join. This will be joining the data to itself so you will see a triangle once done (you may need to readjust the nodes after doing this).



13. Resulting join to itself:



14. Two join clauses will be needed for this. You will join Prev Patient Row from the Row Numbering node to Patient Visit Row Number in the Setting up Rejoin node. Along with PatientID to PatientID to make it line up to the right person. You can do this by clicking the plus icon in the top right of the join menu.

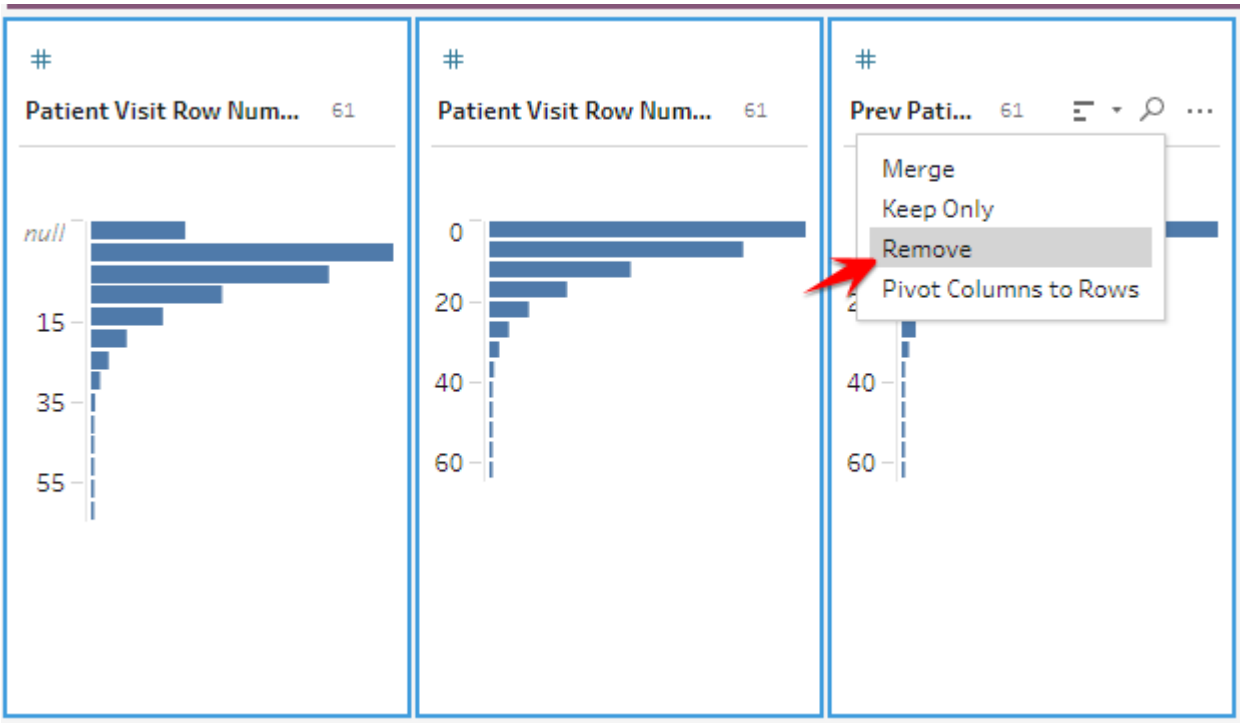


15. Make it a left join on the Row Numbering side.
16. You will also notice that there are values excluded, which is right. (This would be all of the "0" row numbers after we took the rank minus one. Those don't make sense to keep as they don't exist. In other words, the first visit won't have any visits before it.)
17. What happened here and why does this work? We numbered a patient's visits sequentially. A value of 1 is their first visit with us ever, 2 is their second visit, 3 is third, and so on. An additional column was added of the current row number with one subtracted. 0,1,2,3 and so on. Next, the original data was left joined to the branch of the previous visit matched to the current. This pulls in the previous visit date at the row level to calculate the number of days between them. For example, somebody visits the medical clinic on January 1st which is visit 1. 6 months later, they visit again on July 1st which is visit 2. Visit 2 gets joined to Visit 1 because we know that Visit 1 preceded visit 2. Now that this data is joined in at the row level, the number of days since their previous visit can be calculated.

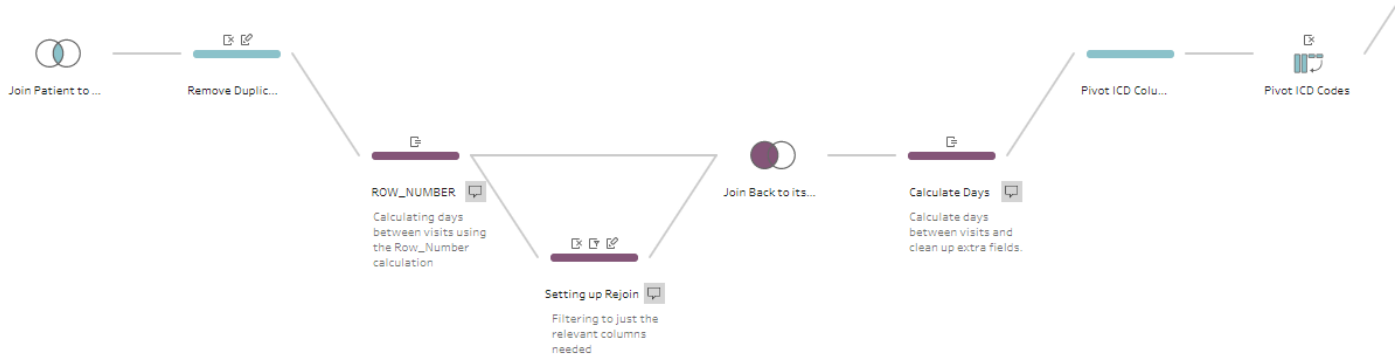
18. Add a new clean step and add in the following column:

19. The ISNULL clause is for first visits to our facility. It comes in as a 0 despite Previous Visit Dates being null for first visits so this ensures that it is null.

20. Finally, remove all the row number columns as they are no longer needed.



21. Re-attach our data flow into the pivoting at the end. You may need to delete some existing connection lines to get this attached in. I colored my line in purple and indented it below to show that this was added later.



6.7 Automating Tableau Prep

What are your awesome data preparation skills worth if every day you walk into the office and you have a slew of requests to run flows previously created to update data? Luckily, we have two options to automate this process, so you don't have to think about it (or be bothered by it).

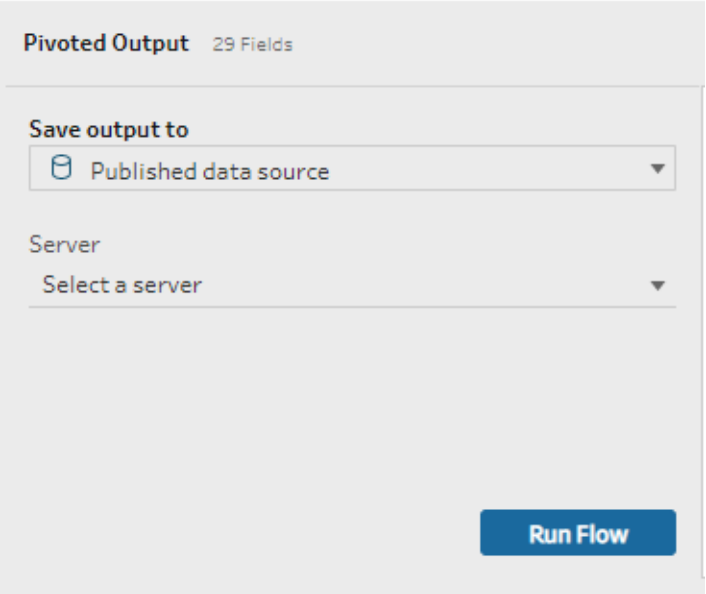
Option 1: Tableau Server/Online

Of the two options, the best option is to publish your flow onto Tableau Server or Tableau Online. Once published, it can be set to refresh on a recurring schedule. (From here on out, Tableau Server will refer to both Tableau Server and Tableau Online as the instructions are the same.) Unfortunately, this is an add-on service that is charged per user, per month for all users registered to Tableau Server (including users who only consume content with "Viewer" license types). This service is often cost-prohibitive for businesses but is easy and convenient to use if your company does have it.

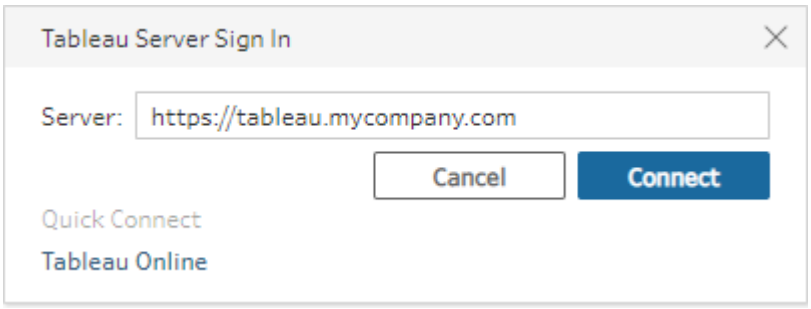
Your database credentials will be published up to Tableau Server (encrypted!) along with the flow steps so that your data will always stay fresh as it pulls directly from the database. Obviously, if you published a flat excel or csv file that file will get published up to the server, but it does not have a way to automatically refresh. Note that Tableau Server (not Online, in this case) can connect to and publish files hosted on a shared network drive as long as the Tableau Server has access to the network drive.

Steps:

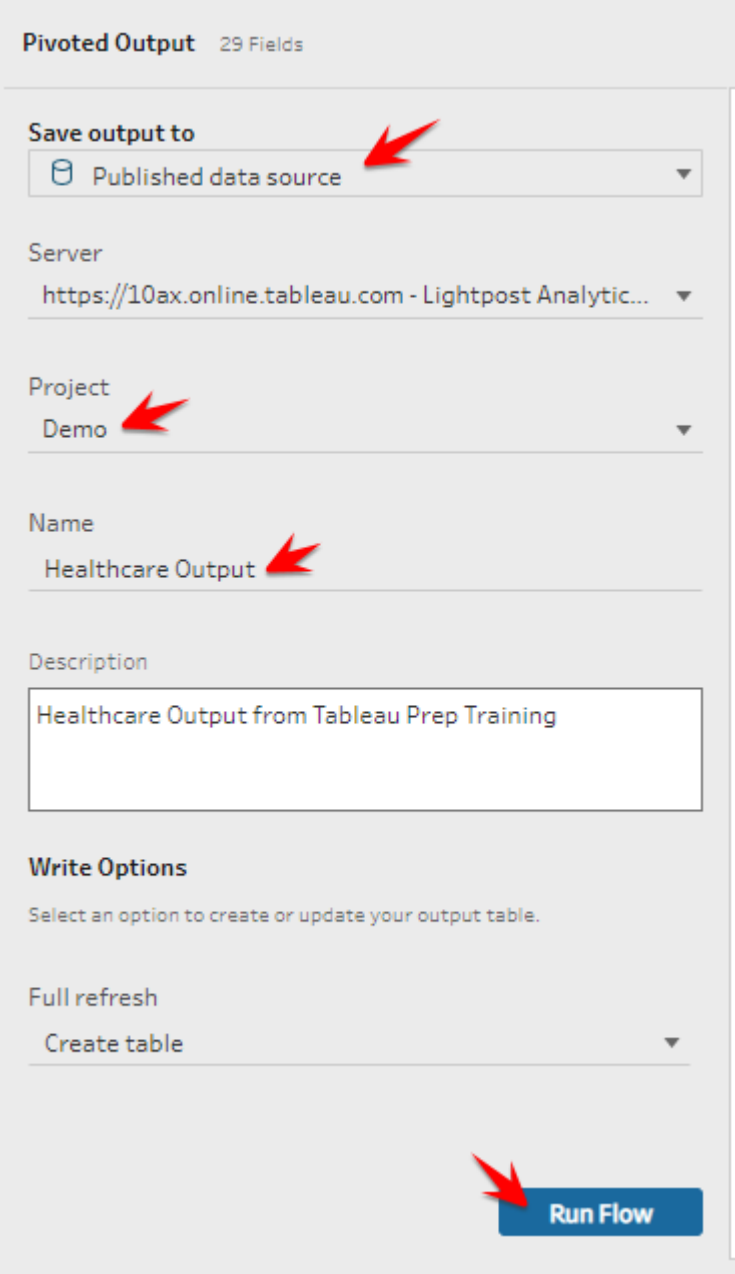
1. Ensure that your Tableau Prep Output step is set to publish as a datasource to Tableau Server instead of a flat file to your computer (Unless you have set it to publish to a network drive that has already been configured to be a safe path for Tableau Server to access). Skip to step 7 if you already have this configured.
2. If it is not, create an output step. If you are not logged into Tableau Server, the menu on the bottom left will appear like this:



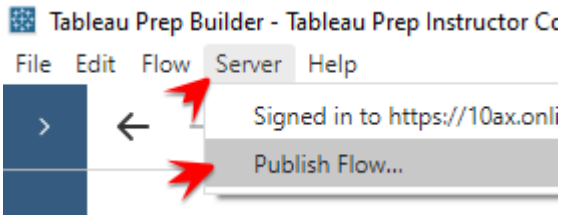
3. Select the dropdown on “Select a server” and choose “Sign in”.
4. A new menu appears. If your company is using Tableau Server then you will enter the URL for that in the box. This might be a URL like “https://tableau.mycompany.com” or it might be an IP address like 10.25.100.1. If your company is using Tableau Online then select “Tableau Online” under quick connect.



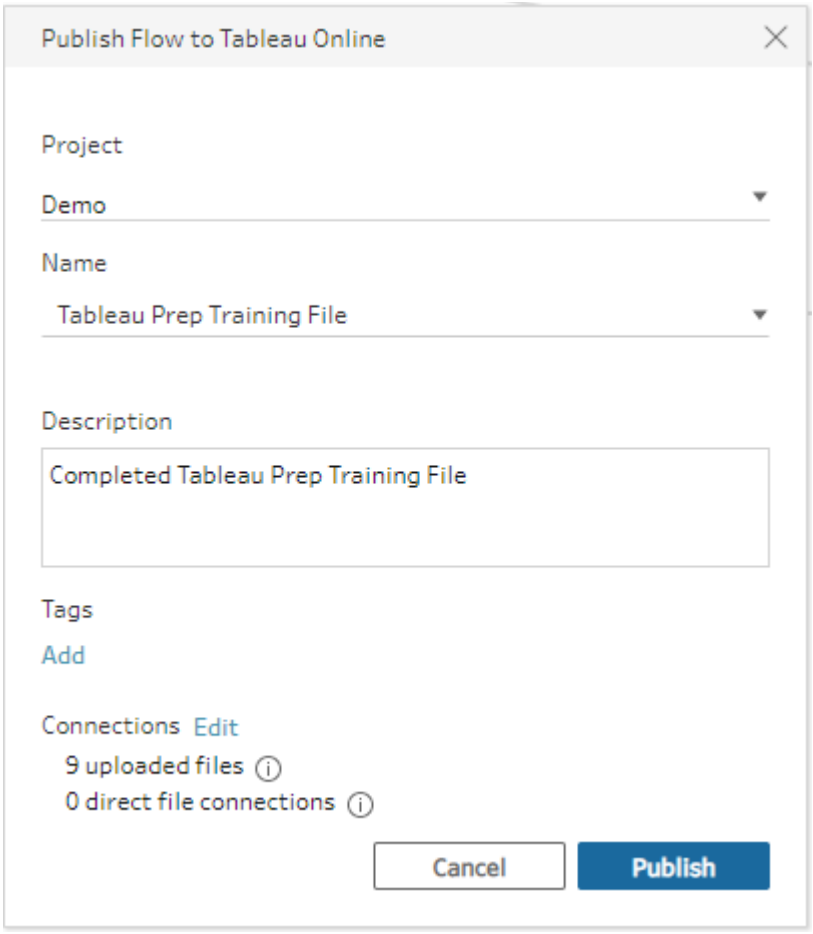
5. Enter your Tableau Server/Online credentials in the new box that appears.
6. Once connected, choose a Project/Folder, give it a name, a description if desired, and click Run Flow to ensure that it runs correctly on your computer.



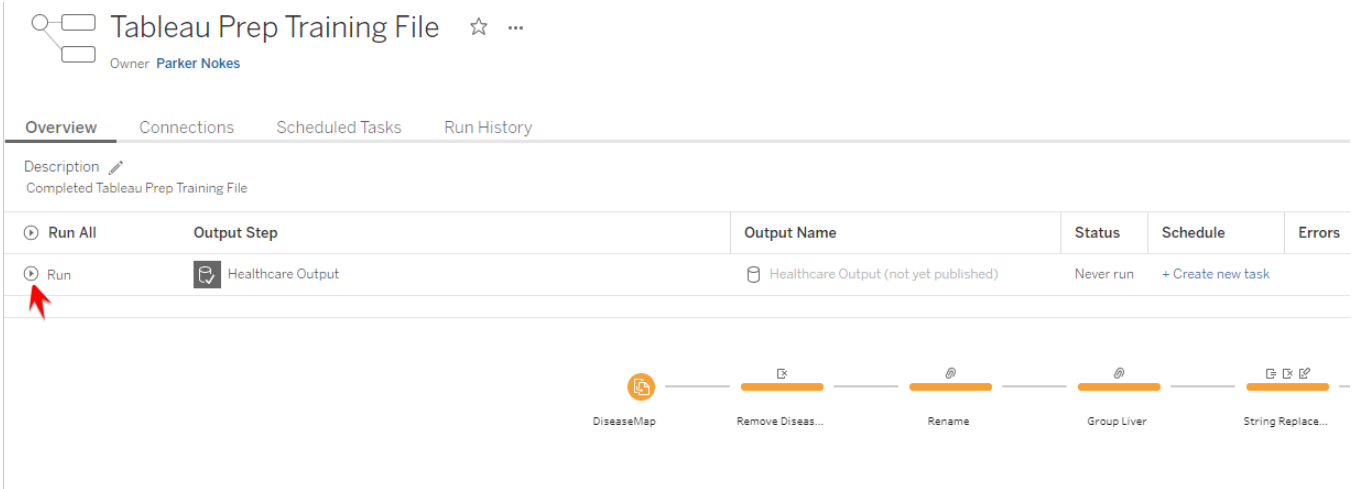
7. To publish the flow, in the top left dropdown menus, hover over the “Server” section and choose “Publish Flow”. On Mac, this will be in the top bar across your screen.



8. A similar menu that we saw in step 6 appears, we need to put our published flow into a Project/Folder, give it a name, and a description if desired.

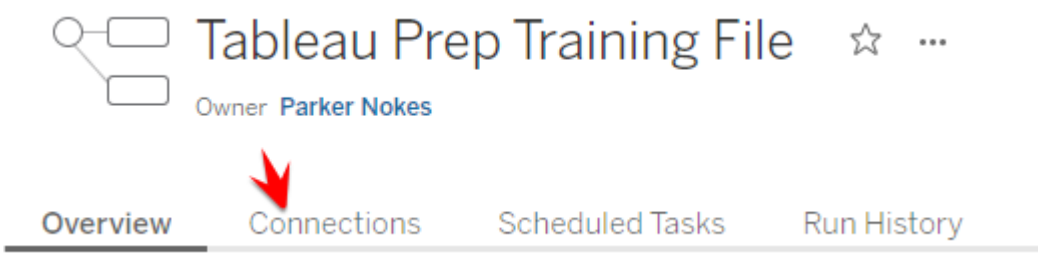


9. Click “Publish”!
10. Your browser will open up a new tab with the published Flow File. Log in if needed.

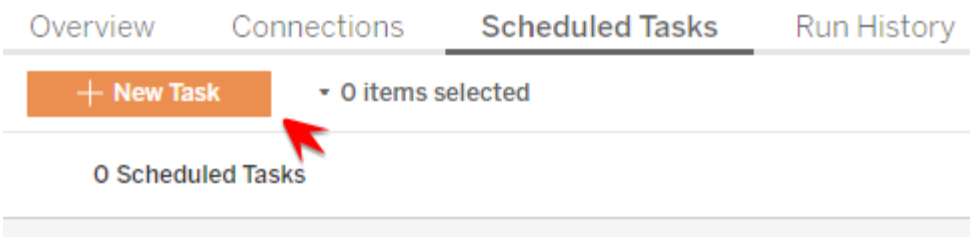


11. In the bottom pane, your flow steps will be shown. Above that, each of the outputs of your file will be shown with a run button. Go ahead and click Run to ensure that it is working once connected to the server. If you are using a database connection and it doesn’t work now on the Server, but did work on your computer, you may need to talk with your IT administrators to ensure that Tableau Server can communicate with your databases.

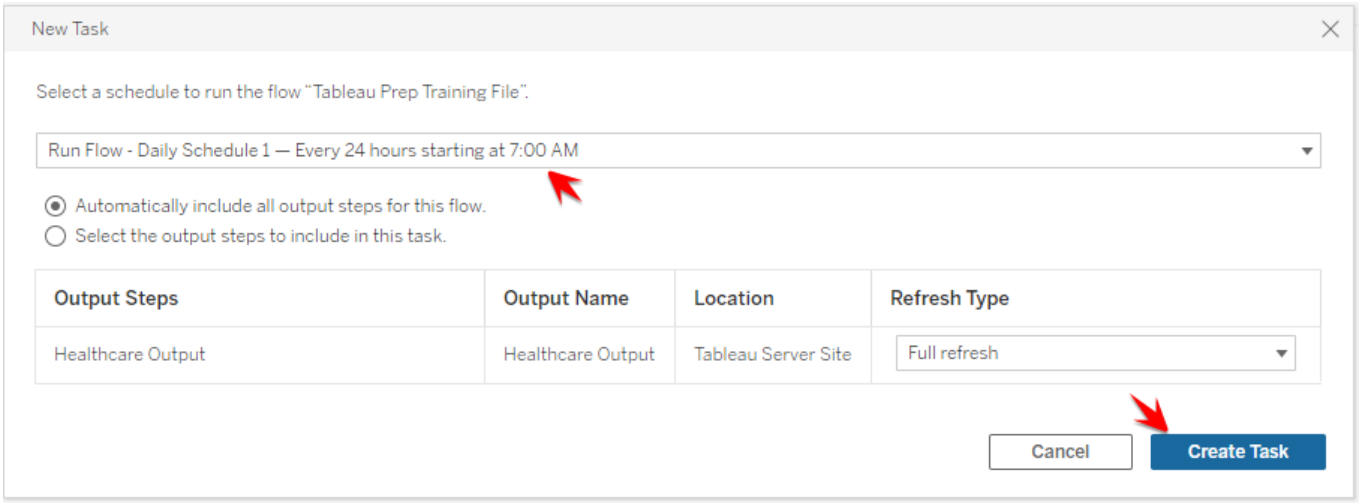
12. Click “Connections” in the top menu.



13. This area shows what files or databases you are connected to. You can edit those connections here.
14. Click on Scheduled Tasks.
15. Click on “New Task” in the top left.



16. This menu allows you to schedule when flows happen. By default, there are usually several run schedules available that fit most needs, but if you have special use cases that aren’t listed, talk to your Tableau Server Administrator to get new schedules created.
17. Select a schedule. If you want all outputs to run, leave the default. If you want only specific outputs to run (maybe one output is run daily while another one is only run weekly), those can be specified.



18. You should be all set for your data to be refreshing automatically. If it fails, it will send an email to you which may need to be investigated, but if everything is set up correctly, there shouldn't be any future issues.
19. You can now open Tableau Desktop and connect to a Tableau Server Datasource. Opening the Tableau file with the server datasource will automatically pull in the new data that your Tableau Prep flow outputs to.

Option 2: Using Tableau Prep's Command Line Tool on a Schedule

Tableau Prep has a command line tool that allows you to run previously created flows. Using the command line sounds scary to many, but their tool is simple to use. Note that you will need initial administrator approval (and their credentials) to get it scheduled, but once approved it should run without future approval. The steps to use this won't be detailed here, but Tableau has great documentation on it: https://help.tableau.com/current/prep/en-us/prep_run_commandline.htm

To get it fully automated, you will need to create a .bat file for windows. This is a file that runs command line commands. The same commands that you use to run it from the command line will be copied into a text file and saved as a .bat file. Then you can use Windows Task Scheduler to run the .bat file. Google "How to run .bat file in Windows Task Scheduler" for more specific instructions.

The Mac equivalent is a bash script that runs in the Automate program. Google "How to automate bash script on Mac" for instructions.

You have built some pretty complex flows in Tableau now! All of these things should give you enough skills to be dangerous for any data cleaning and preparation project thrown your way. Tableau Prep is a constantly evolving tool with new features added almost every month. You are going to want to follow those changes and other features that are added that will make your job easier.

For more information on using Tableau Prep, check out Tableau's training videos at https://www.tableau.com/learn/training