FIT3179 Data Visualisation

Tutorial Week 4

Tableau: Multiple Data Sources

Overview

This week's Tableau tutorial consists of two activities.

1. Working with Multiple Data Sources
2. Blending Data Sources

After this studio, you should be able to start designing your data visualisation 1. Please discuss the domain and ideas with your tutor.

1. Working with Multiple Data Sources

So far, you have learned to create a visualisation based on an existing dataset. In some cases, you might need to start designing your visualisation without thinking about whether or not the dataset exists. To bring your design into realisation, you may need to hunt for datasets on the Internet. Stumbling upon dirty datasets which require cleaning is common. You might need to combine pieces of datasets into a single data that can satisfy your visualisation requirements.

This tutorial will teach us the basics of working with multiple data sources. Tableau provides a feature to join multiple data sources together.

For this exercise, we need a fresh new workbook. Save your previous work and create a new workbook by clicking **File -> New**.

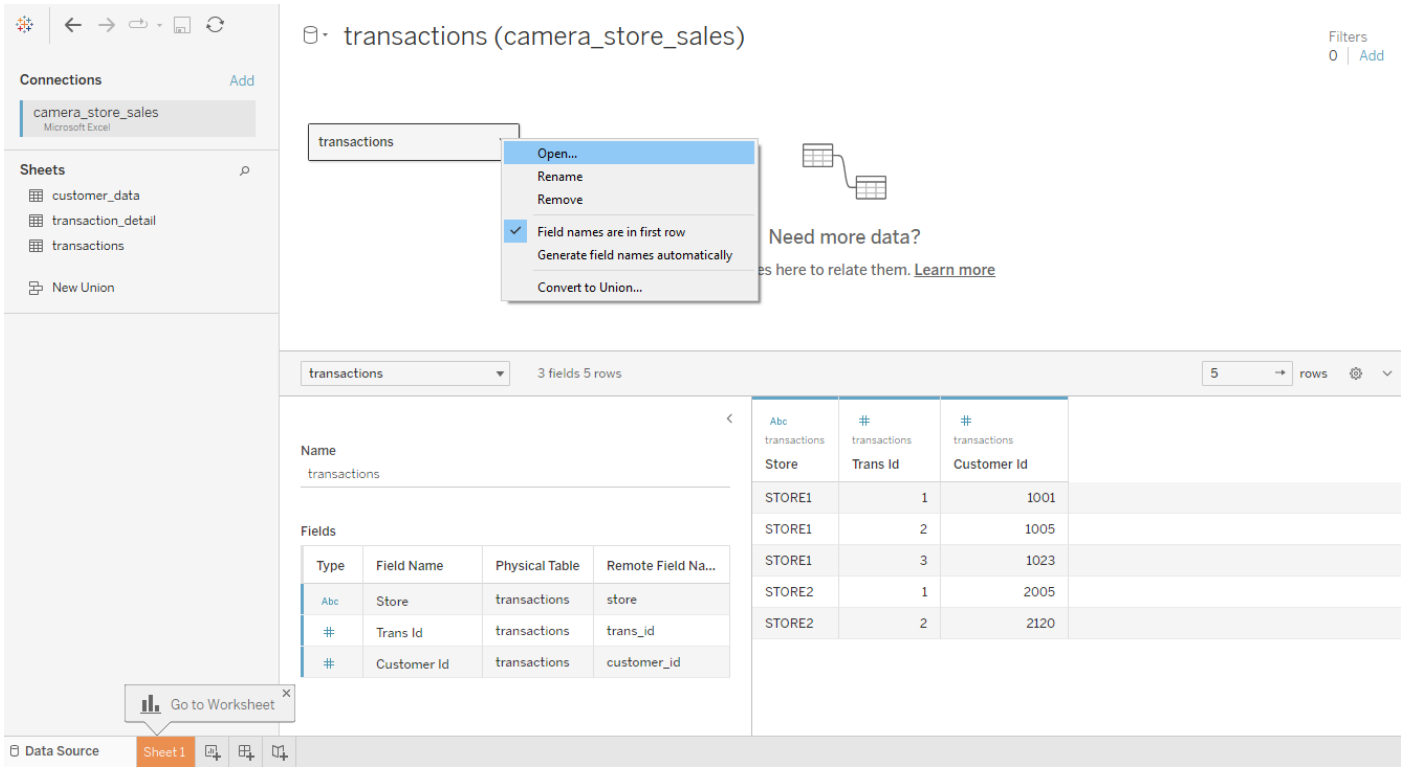
1.1. Joining Datasets via a Single Field

For this activity, we will use two hypothetical datasets to help you recall different join types: **Inner Join**, **Left Join**, **Right Join**, and **Full Outer Join**. Remember your Databases unit?

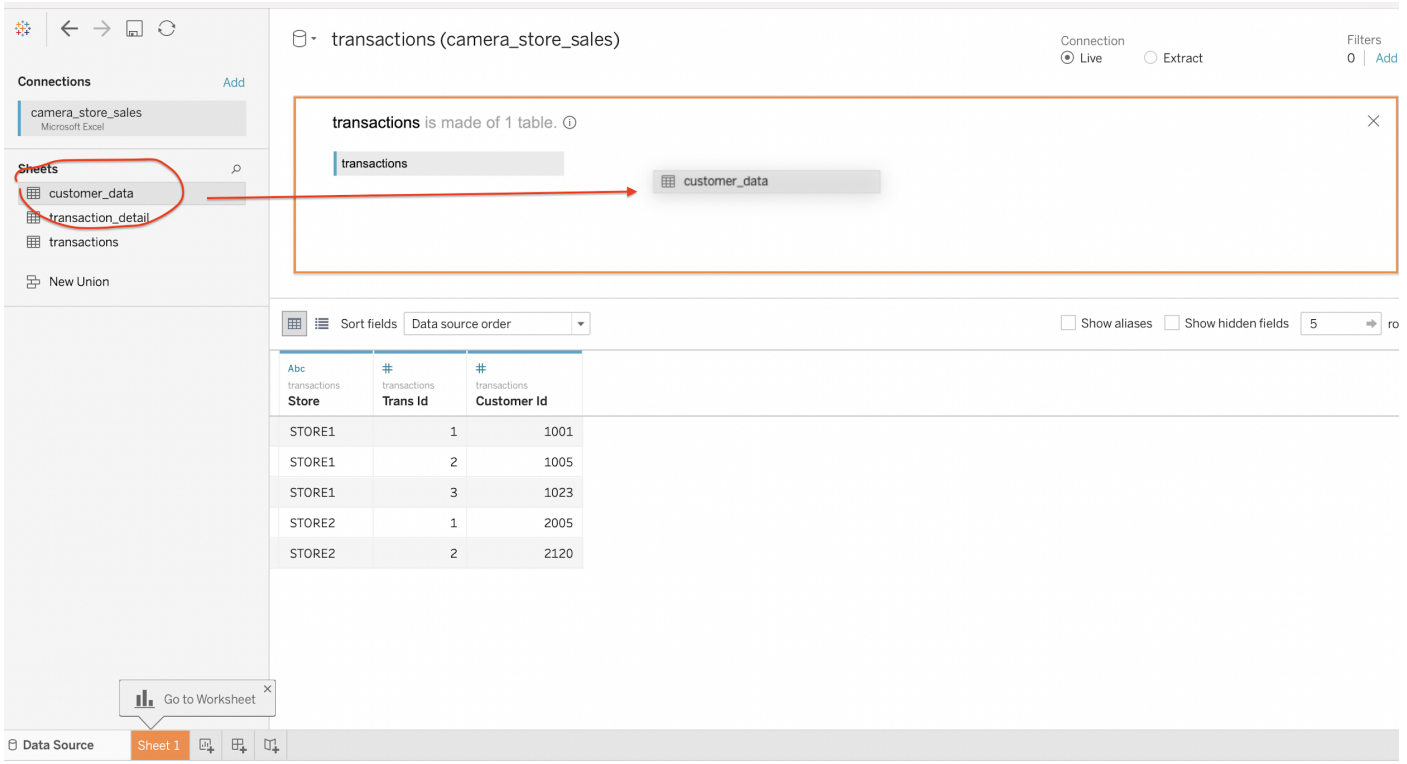
1. Download **camera_store_sales.xls** from the Moodle page.

2. Import the Excel file you just downloaded. The dataset has three sheets: **customer_data**, **transaction_detail**, and **transaction**.

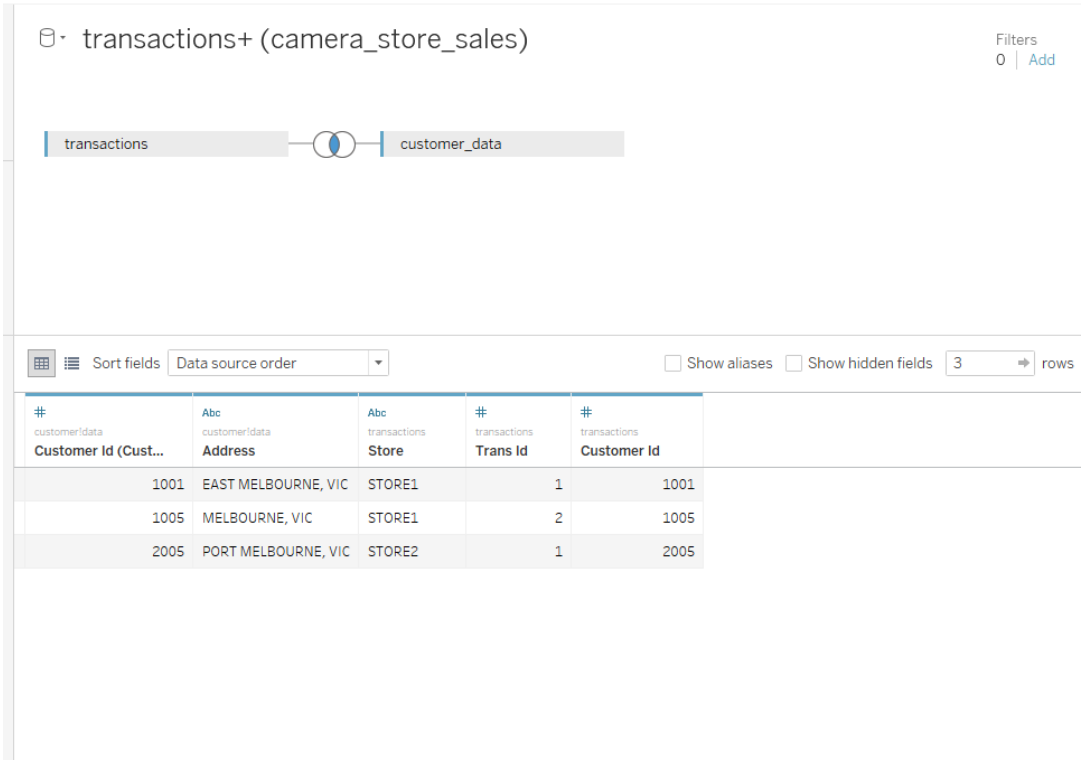
3. Drag and drop the **transactions** sheet. Then right-click on the data box, and select "Open". (You can also double-click on the **transactions** tab to open it)



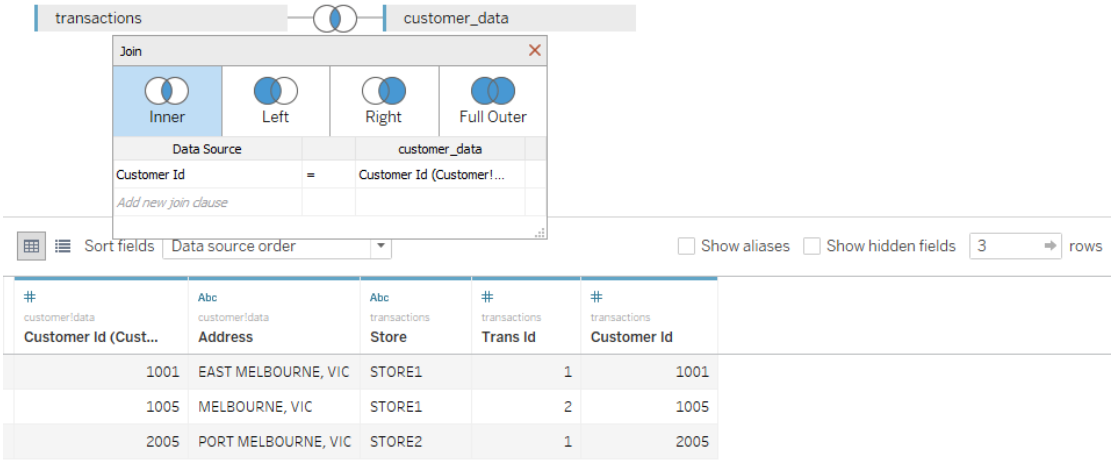
4. Then, drag and drop the **customer_data** sheet to the side of “transactions”.



5. Because both sheets have a common column, **Customer Id**, the **inner join** is automatically performed.



6. If you click on the circle, you can see **Inner Join** is selected. You can also see that the Data Souce (**transactions**) and **customer_data** are linked by the **Customer Id** field. The result of the join is immediately shown in the table below. **Inner Join** will return a set of rows that have common values on both datasets, as shown on the **Venn diagram**. In this case, only customer id **1001**, **1005**, and **2005** are present in the **customer_data** sheet.

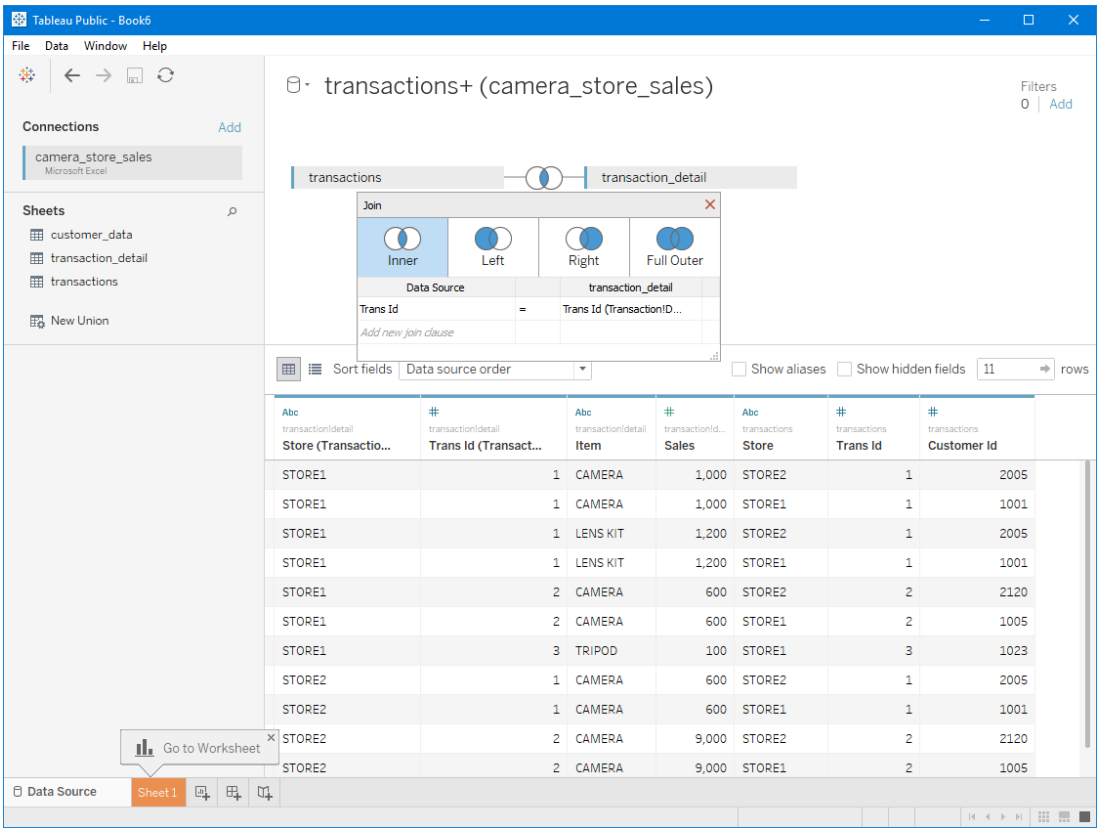


Now, switch to another type of join by clicking the join icon. Inspect the resulting table. Make sure you understand the characteristics of each join!

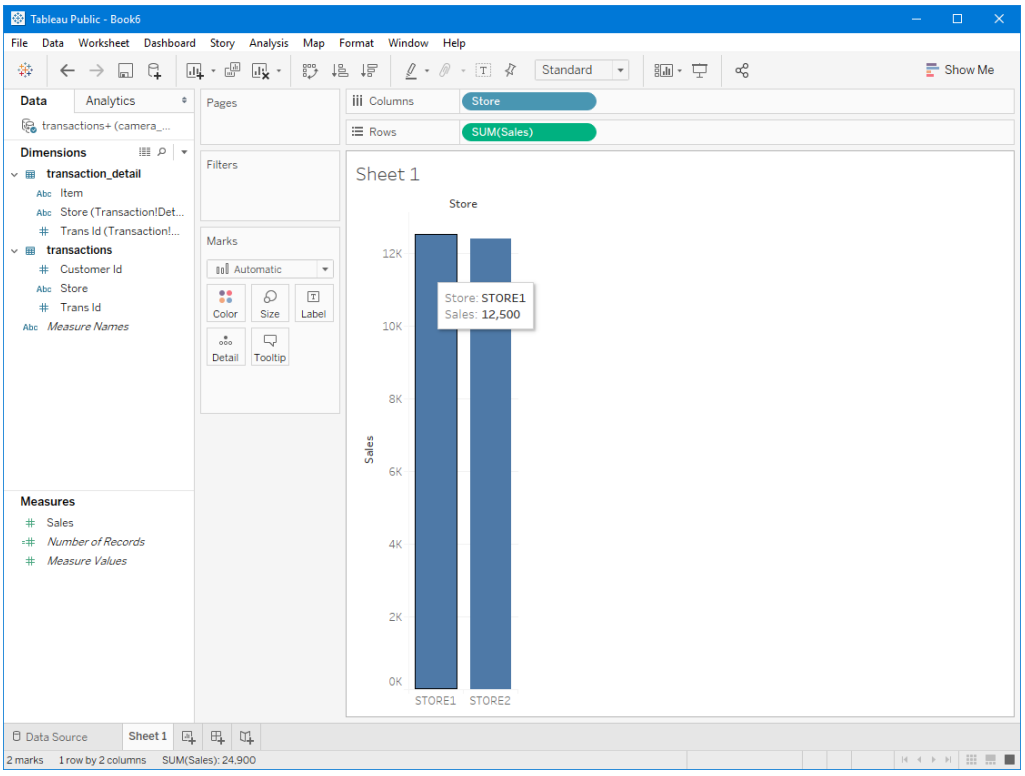
*What are the characteristics of the **Left Join**? **Right Join**? **Full Outer Join**?*

1.2. Joining Datasets via Multiple Fields

1. Once you understand how **Join** works, remove the previously joined dataset. Drop **transactions, click "open"** and add in **transaction_detail**.



2. Now, let's create a new sheet and visualise the sales of each store! Do you still remember how to do this?

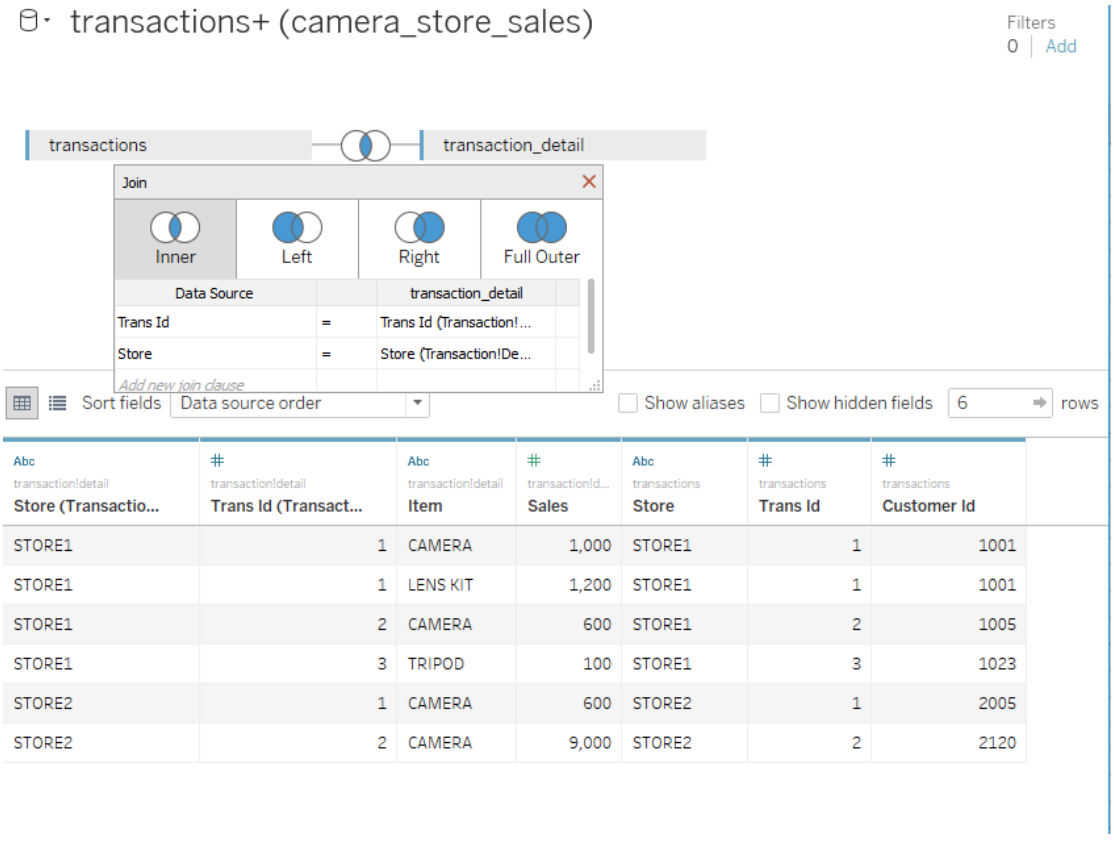


3. The bar chart below shows that the total sales of STORE1 is 12,500. However, if we check the data, STORE1 did not produce that much sales.

store	trans_id	item	sales
STORE1	1	CAMERA	1000
STORE1	1	LENS KIT	1200
STORE1	2	CAMERA	600
STORE1	3	TRIPOD	100
STORE2	1	CAMERA	600
STORE2	2	CAMERA	9000

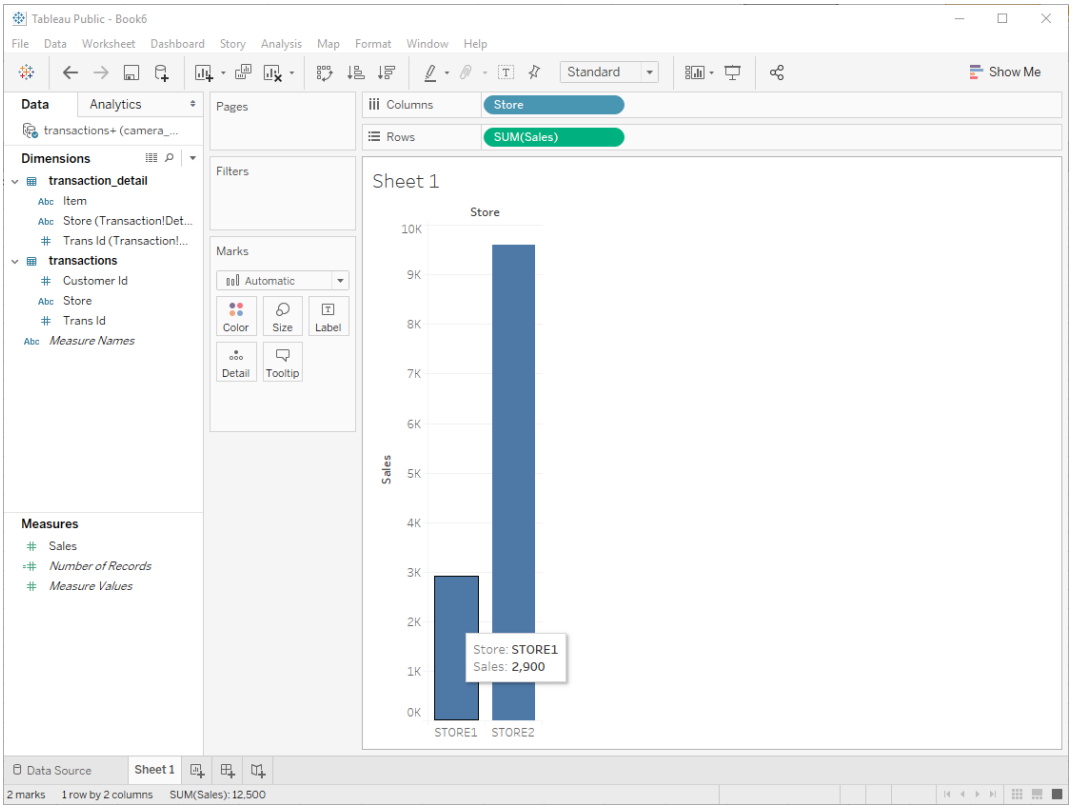
4. Let's go back to the data source and figure out what's wrong with it. Turns out there are many unnecessary repeating rows! This happened because the join condition only considers **trans_id**. In fact, both STORE1 and STORE2 have the same id for different sales. To fix it, we need to add another join condition.

5. Click the **Inner Join** icon and add a new **condition**. In this case, we want to have **store** as another key.

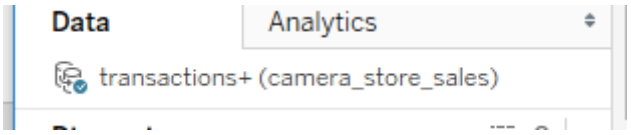


Store (Transactio...	Trans Id (Transact...	Item	Sales	Store	Trans Id	Customer Id
STORE1	1	CAMERA	1,000	STORE1	1	1001
STORE1	1	LENS KIT	1,200	STORE1	1	1001
STORE1	2	CAMERA	600	STORE1	2	1005
STORE1	3	TRIPOD	100	STORE1	3	1023
STORE2	1	CAMERA	600	STORE2	1	2005
STORE2	2	CAMERA	9,000	STORE2	2	2120

6. This looks much better now. If you come back to Sheet1, you can see that the bar chart has been updated.



7. Joining datasets will produce a single data source. If you look at the **Data** tab, you can see there is only one data source named **transaction + (camera_store_sales).** In the next activity, you will use multiple data sources without explicitly joining them together.



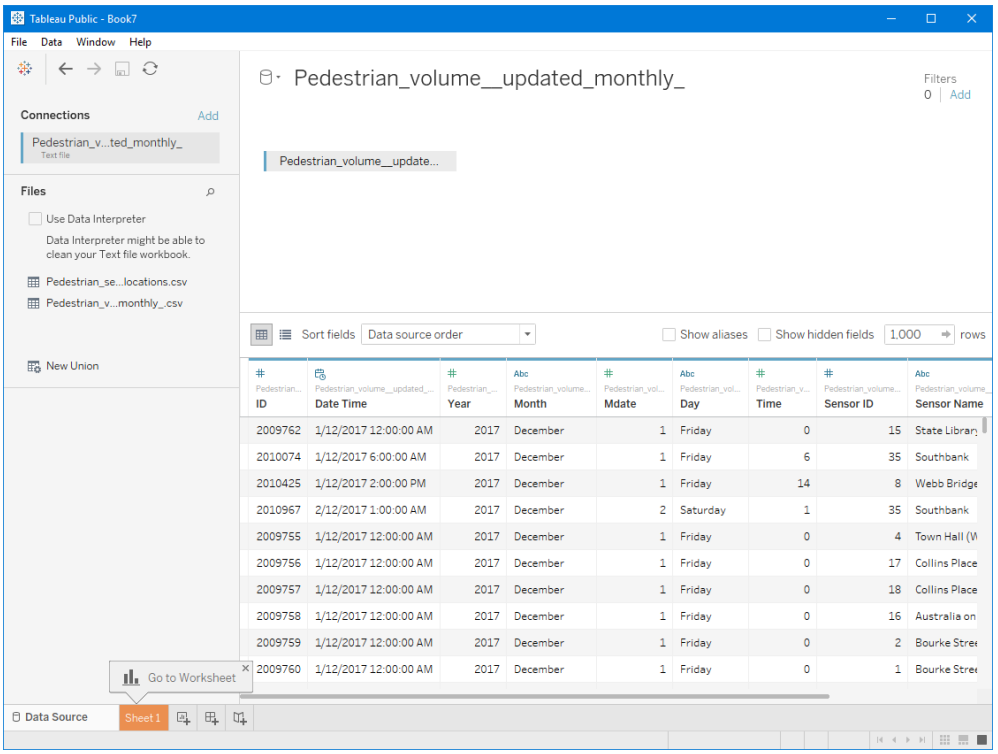
2. Blending Data Sources

Blending means joining datasets on the fly. Instead of creating a single table that contains all the data we need, we can create links between data sources on demand. For this activity, we will use a different dataset.

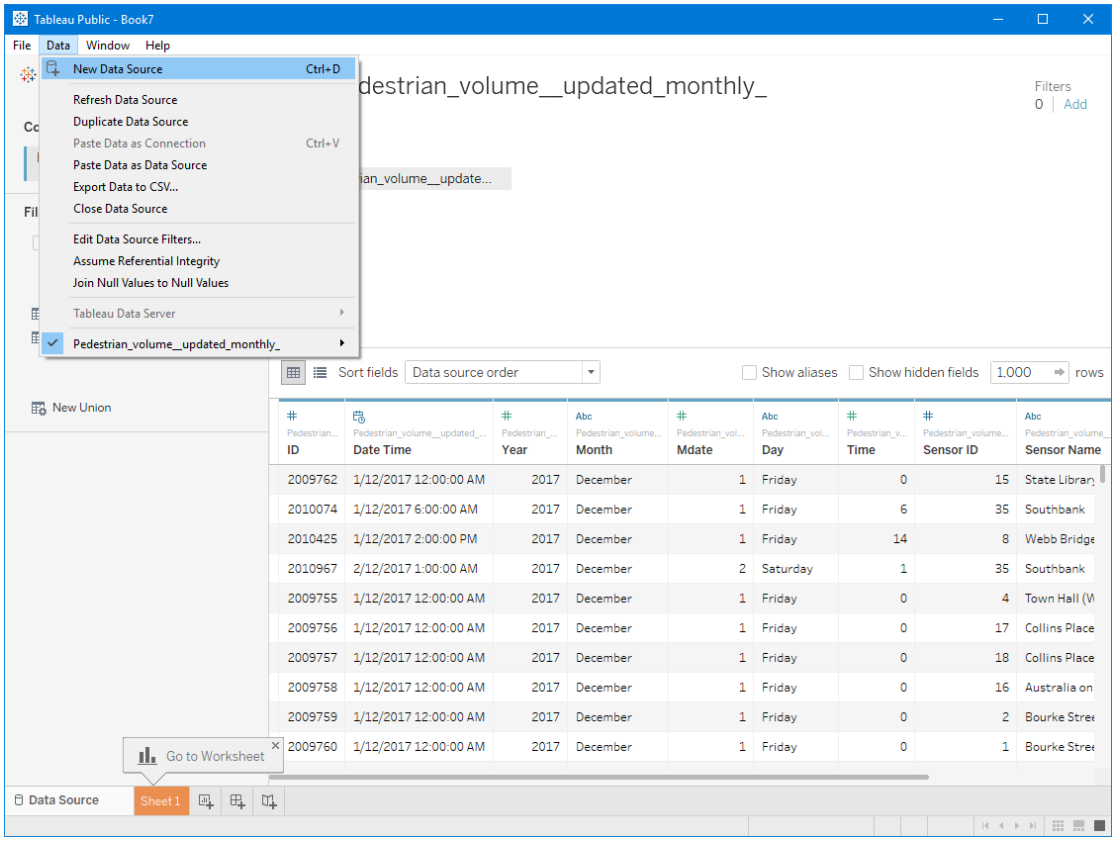
Save the previous workbook and create a new one.

1. Download **Pedestrian_sensor_locations.csv** and **Pedestrian_volume__updated_monthly_.csv**

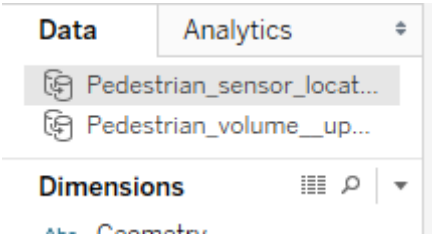
2. Open Tableau and import **Pedestrian_volume_update_monthly.csv**



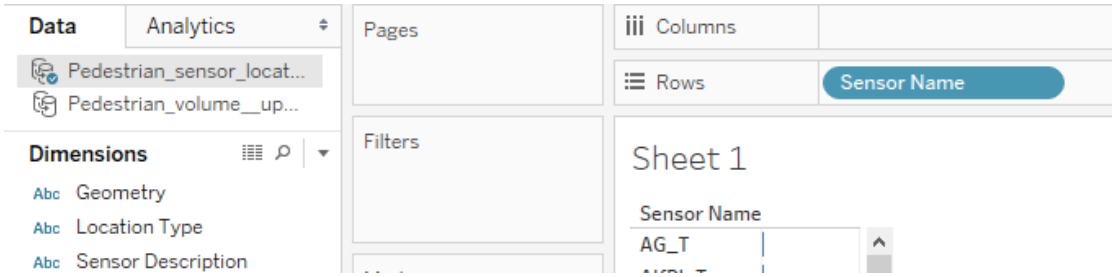
3. Add a new data source and select **Pedestrian_sensor_locations.csv**



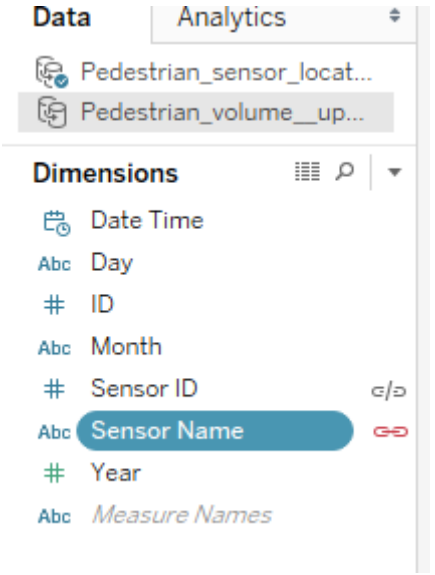
4. Go to **Sheet 1**. You will see two data sources on the **Data** tab. Both data sources have the same icon because we have not yet created any visualisation.



5. The main data source is determined by the first field that you put into the visualisation. To illustrate this concept, let's drop **Sensor Name** from the first data source.

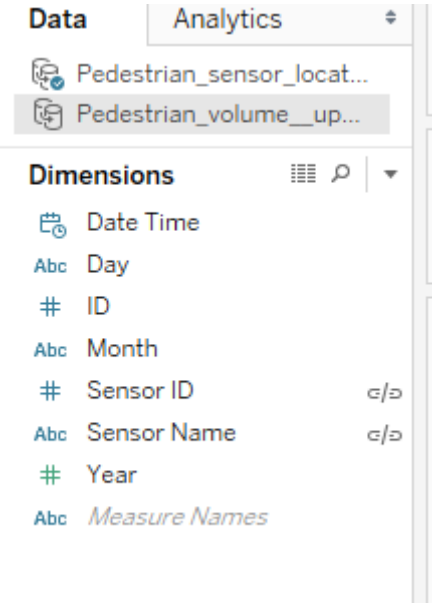


6. As you can see, the icon of the first data source has changed. Now, click on the second dataset. You will notice the link icons on both **Sensor ID** and **Sensor Name** dimensions. In this case, the **Sensor Name** link is active because Tableau automatically detected the same field name from the main data source (the first one).



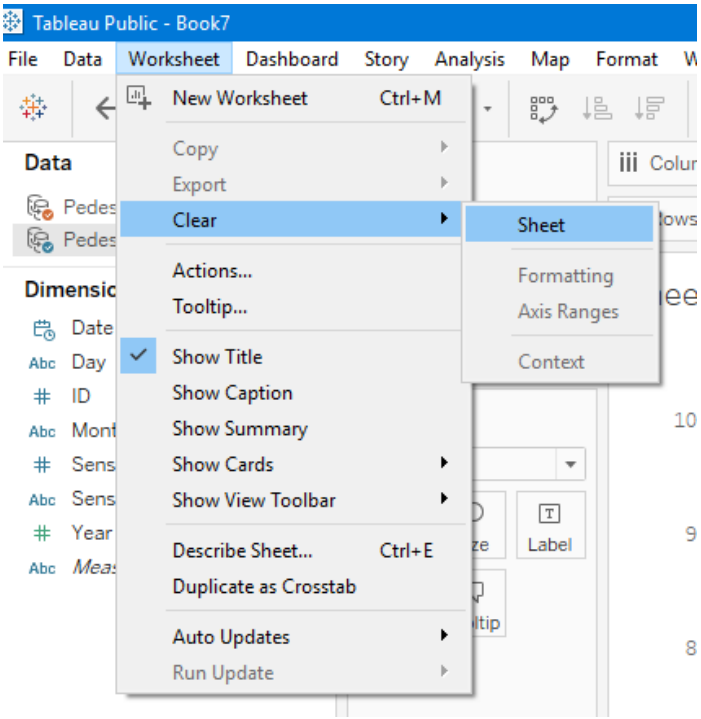
7. Now, remove **Sensor Name** field from the **Rows** and add the **Location Type** from the first data source.

8. After that, click the second data source. The **Sensor Name** link is no longer active because Tableau cannot determine to which field **Location Type** is connected.

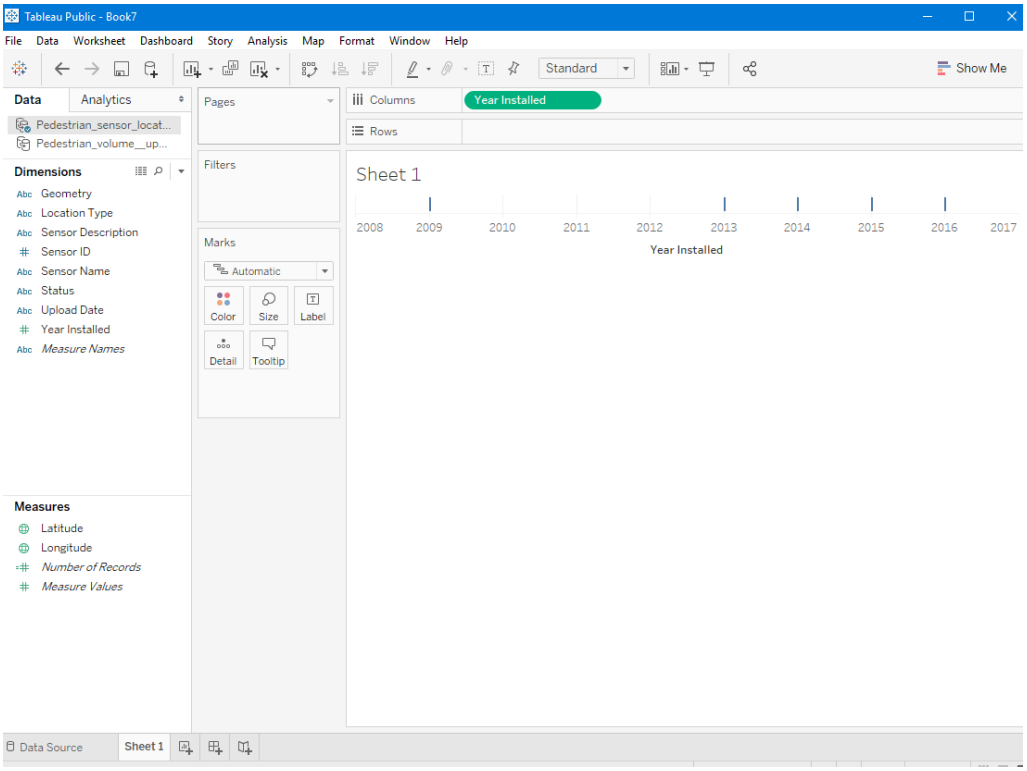


9. If you want to make the second data source the main data source, **clear** your **Sheet** and drop a field from the second data source first.

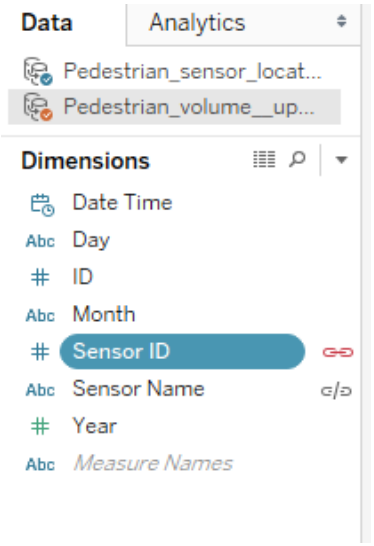
10. Let's create a simple bar chart using these two data sources. First, clear your Sheet by clicking **Worksheet Clear Sheet**



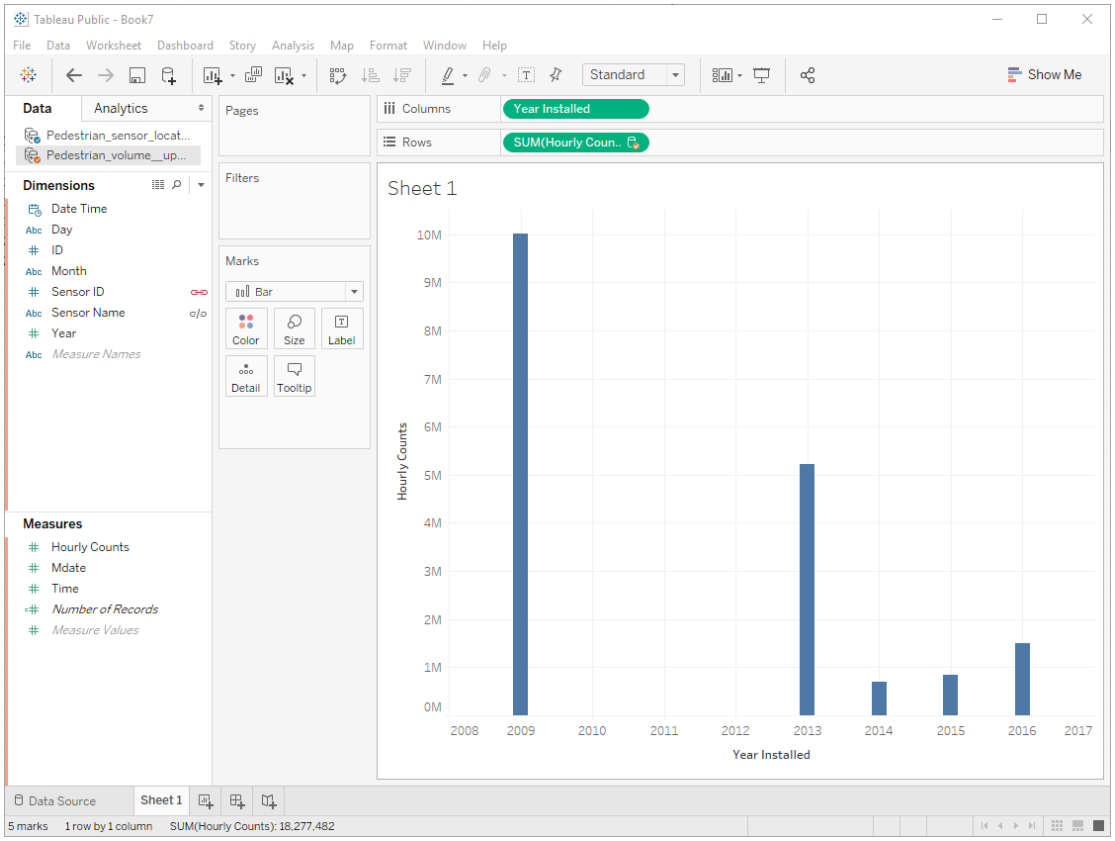
11. Suppose we want to know the **number of pedestrians** counted by the sensor on the first year of instalment. Drop the **Year Installed** dimension from the first data source to Columns.



12. Click the second data source and activate the **Sensor ID** link.



13. Drop the **Hourly Counts** measure to Rows and set the **Marks** to **Bar**.



14. Here you go. Now you have a simple bar chart showing the **total pedestrian count** on the first year the sensor was installed! This visualisation does not tell you much; however, the focus of this activity is to introduce Data Source Joining and Blending so that you can use it later in your fascinating visualisation project.

Look at the Data Sources tab; can you recognise which one is the main source based on the icon?