### Tableau Lecture 7: Operations on Dataset

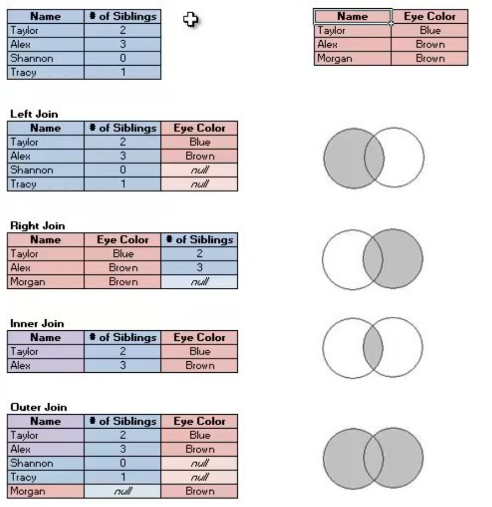
#### **Agenda**

* Joins
  + Building a map visualization
  + Cross-database joins
* Relationships
* Unions
* Blends

### **Joins**

The data that you want to analyze is often made up of a collection of tables that are related by specific fields.

Joining is a method for combining the related data on those common fields. The result of combining data using a join is a table that’s typically extended horizontally by adding fields of data.



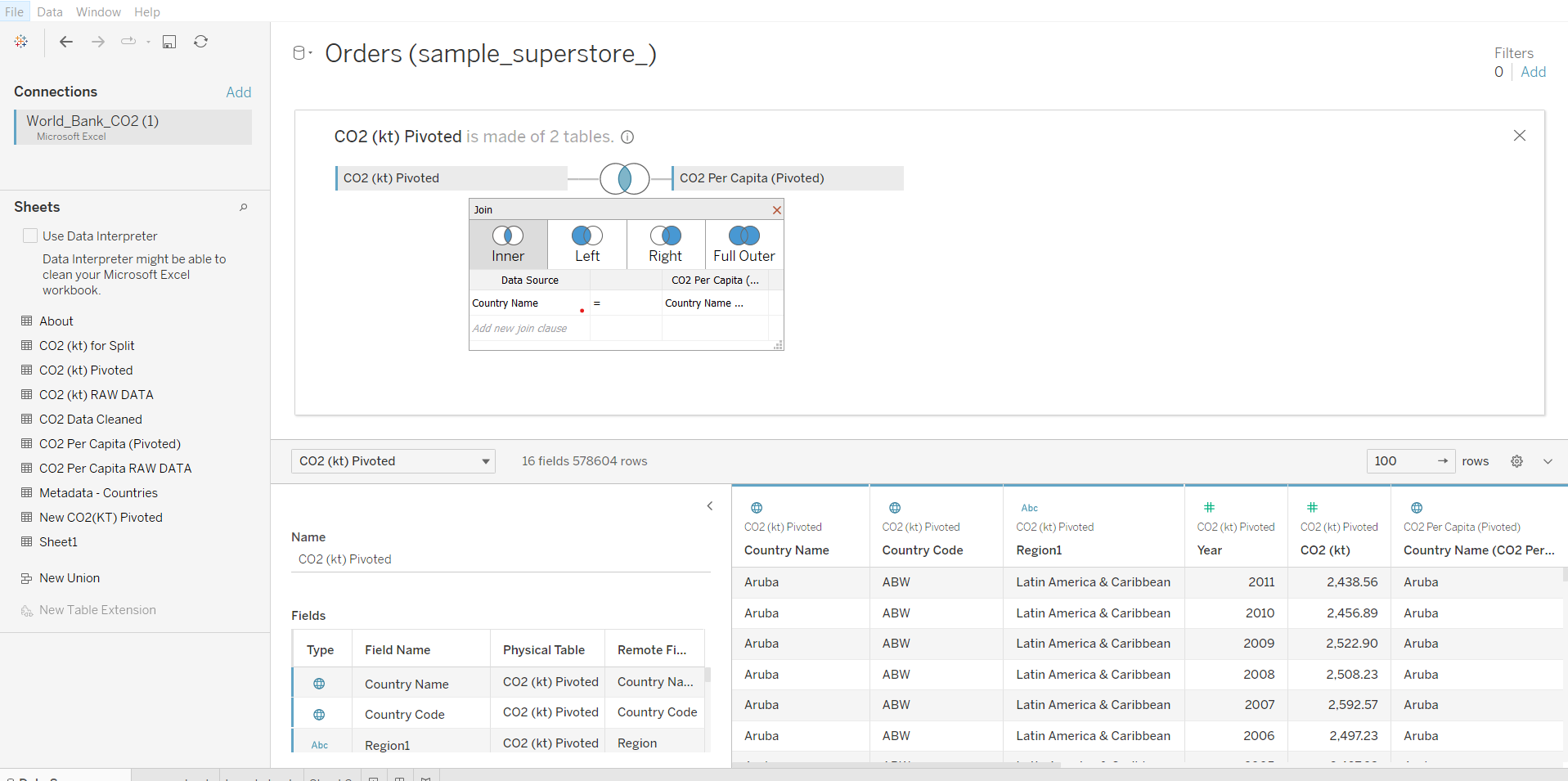
#### **Business problem 1:**

Find per country CO2 KT and per capita emission and filter top 10 countries by CO2 KT.

Dataset : [World\_bank\_CO2.xlsx](https://docs.google.com/spreadsheets/d/1zUMnxp9aPdV5kYWJSoCkVqTEurJvd86d/edit?usp=share_link&ouid=100659516601446935794&rtpof=true&sd=true)

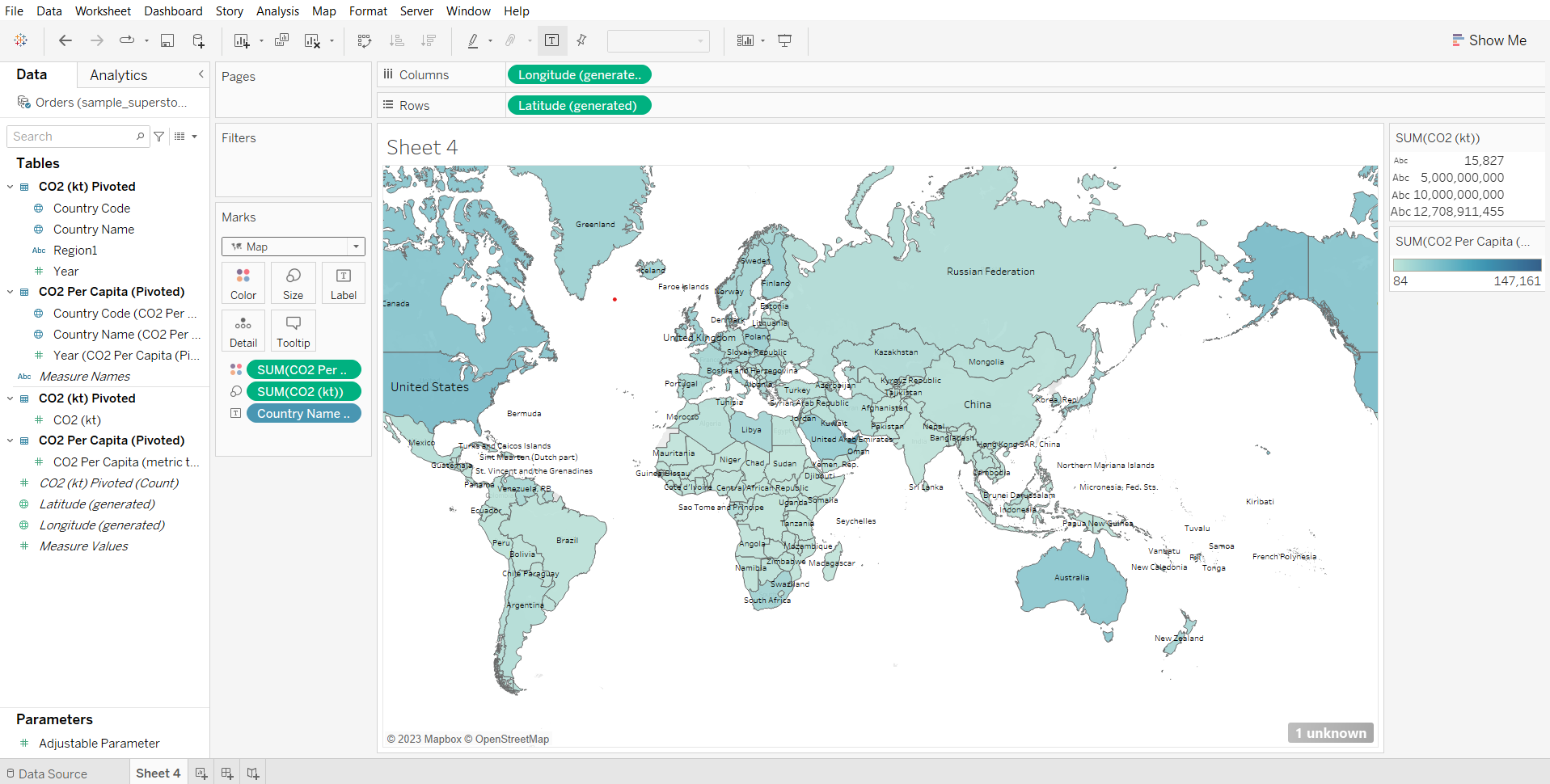
1. Creating Join

* Drag the ‘CO2 (kt) Pivoted’ table to the Canvas
* Double click on the ‘CO2 (kt) Pivoted’ box to get the physical layer.
* Drag ‘CO2 per capita (pivoted)’ to the canvas.
* Click on the join symbol
* Change the key field for both tables to the country name.



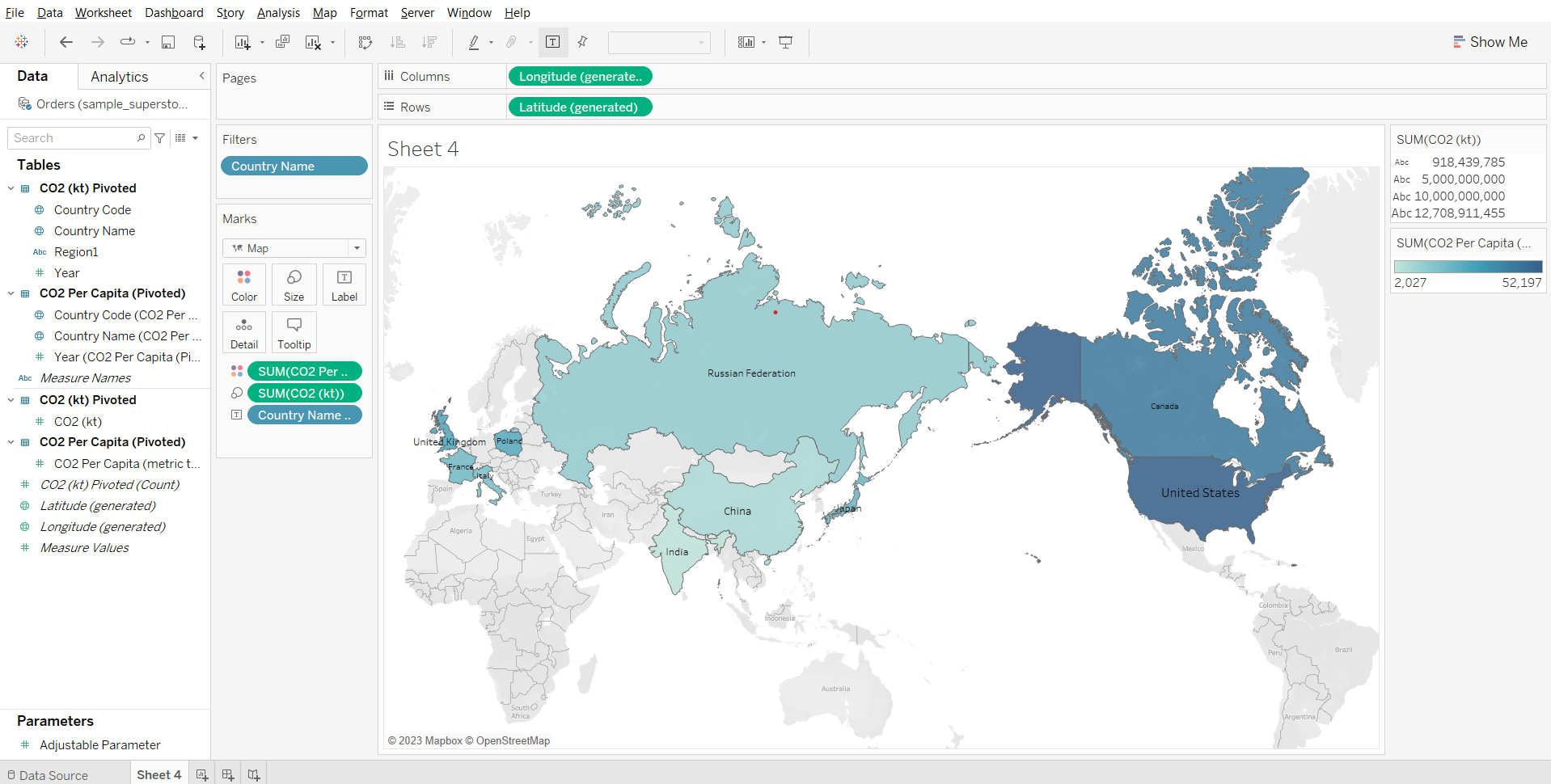
1. Visualization

* Drag Latitude to rows shelf
* Drag longitude to columns shelf
* Drag country name to label in marks card
* Drag CO2 (kt) to size in marks card
* Drag CO2 Per Capita to color in marks card
* In the marks card, change the visualization type to map.



1. Applying filter

* Drag country name to filters card
* Go to the ‘Top’ tab, choose the CO2 (kt) field, click OK.



**Geographic data type:**

#### **Format geographic data in Tableau**

* Depending on the type of map you want to create, you must assign certain data types, data roles, and geographic roles to your fields (or columns).
* For example, in most cases, your latitude and longitude fields should have a data type of number (decimal), a data role of measure, and be assigned the Latitude and Longitude geographic roles.
* All other geographic fields should have a data type of string, a data role of dimension, and be assigned the appropriate geographic roles.

#### **Change the data type of a column**

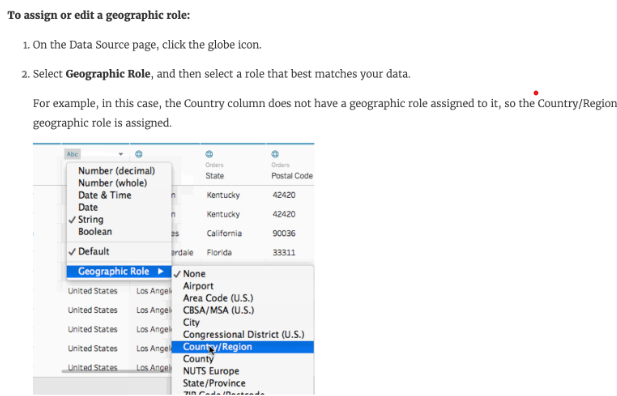
* Tableau might incorrectly assign a Postal Code column a data type of Number (whole). To create map views, your Postal Code data must have a data type of String.

### **Assign geographic roles to your geographic data:**

#### **How does tableau know where to plot on a map?**

When you assign the correct geographic role to a field, Tableau assigns latitude and longitude values to each location in that field by finding a match that is already built into the installed geocoding database. This is how Tableau knows where to plot your locations on the map.

**Geographic roles** are sometimes **automatically assigned** to your data. You can tell a geographic role has been assigned to your data because the column includes a **globe icon**.  
If a **geographic role** is not automatically assigned, you can **manually assign one to your field**.

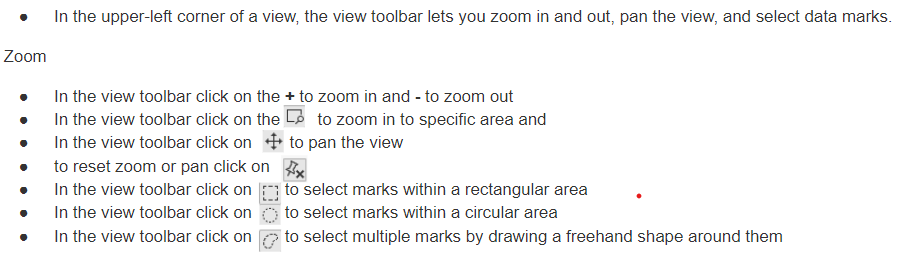


### **Importance of maps:**

* You have some location data in your data source
* You think that it would enhance getting better insight from data compared to other visualization
* It makes representing data in much more beautiful manner
* To answer spatial questions like
  + What is state wise sales?
  + Which states have the highest profit?
  + In which state we had the highest sales for a particular product?

### **Few things to keep in mind when creating a map view in tableau**

* If your data source has **latitude and longitude fields** then they should have a **data type of numeric(decimal)** and a **data role of measure** should be assigned to them.
* And other **geographical fields like country, city, state** should be assigned **string data type** and a **data role of dimension.**
* Tableau will automatically assign geographic roles based on the field name and values present in the data.
* **If our data source does not contain latitude and longitude values we can still create maps using the geographical fields available in the data source.**

****

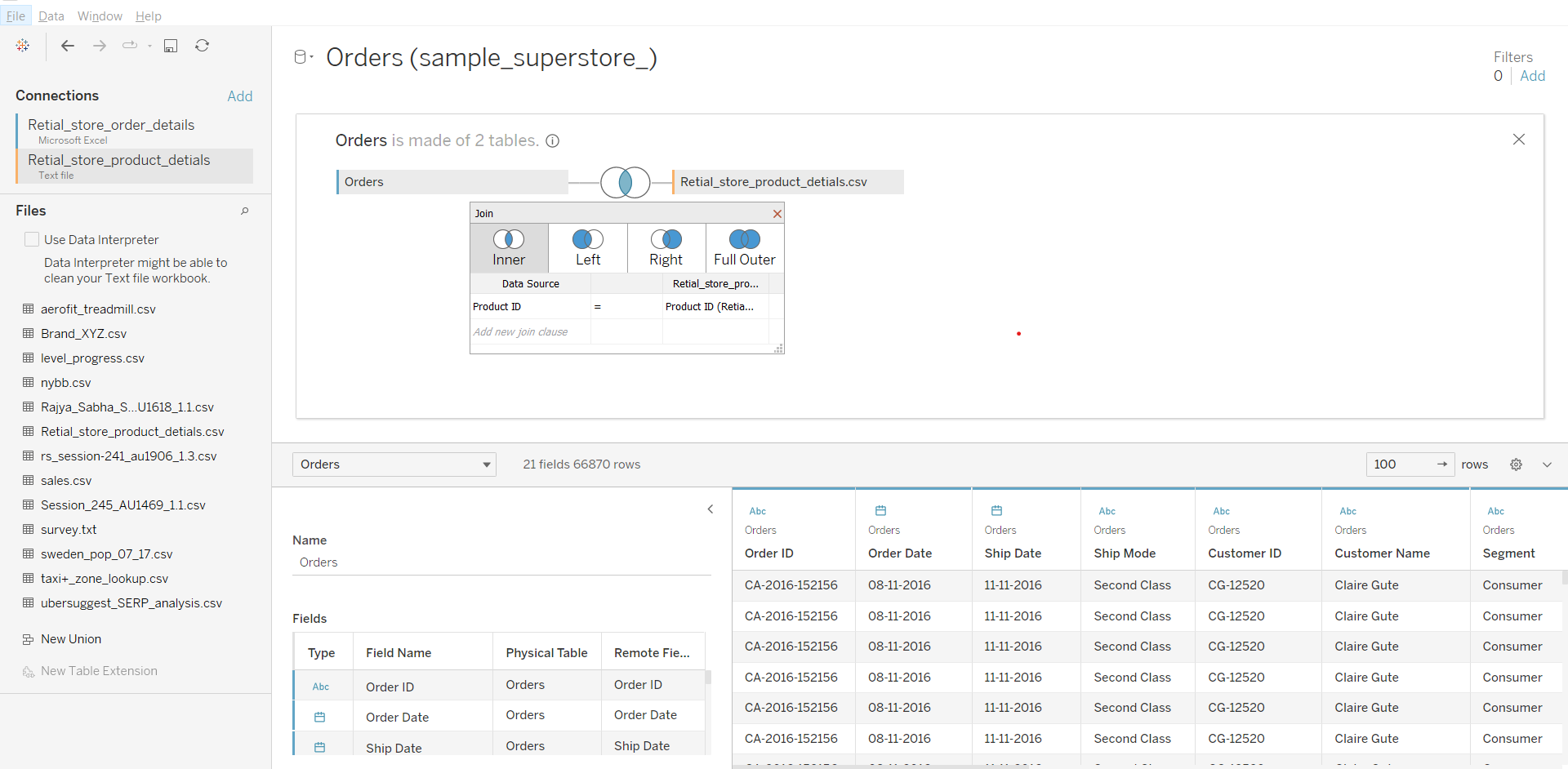
**Cross-database Joins**

* In Tableau, a cross-database join is when you join two or more tables from different databases using a common field. This creates a single table in Tableau that can be used to create visualizations.
* Cross-database joins require a multi-connection data source - that is, you create a new connection to each database before you join the tables.
* We’ll try to join tables from a CSV file and an Excel file.
  + Dataset:
    - [Retail\_store\_order\_details.xlsx](https://docs.google.com/spreadsheets/d/1Sa5n50BOhbW0O3YDXTnL51IZj9vSj4Lk/edit?usp=share_link&ouid=100659516601446935794&rtpof=true&sd=true)
    - [Retail\_store\_product\_details.csv](https://drive.google.com/file/d/1V69qfTOTadbBfK6BOq8hJAFPvMjDJCXY/view?usp=share_link)

#### **Sample problem:**

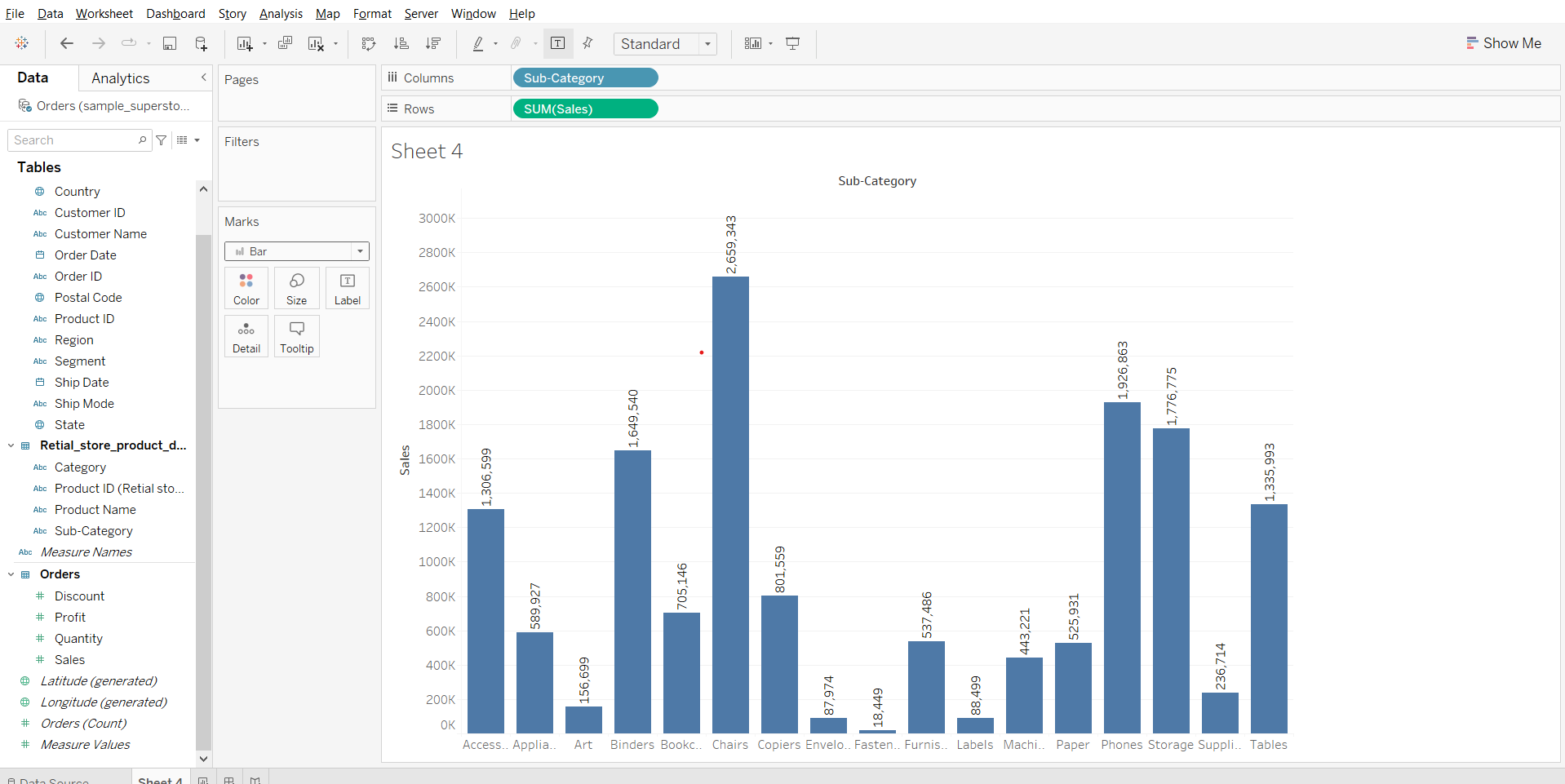
Identify subcategories that made the highest and lowest sales.

1. Loading the two different data sources and adding an inner join between the two tables  
   Note: one is the orders table and the other is the products table.



1. Solving the business problem:

* Go to a worksheet
* Drag Sub-Category to Columns
* Drag Sales to Rows



**Relationships**

Relationships are a dynamic, flexible way to combine data from multiple tables for analysis.

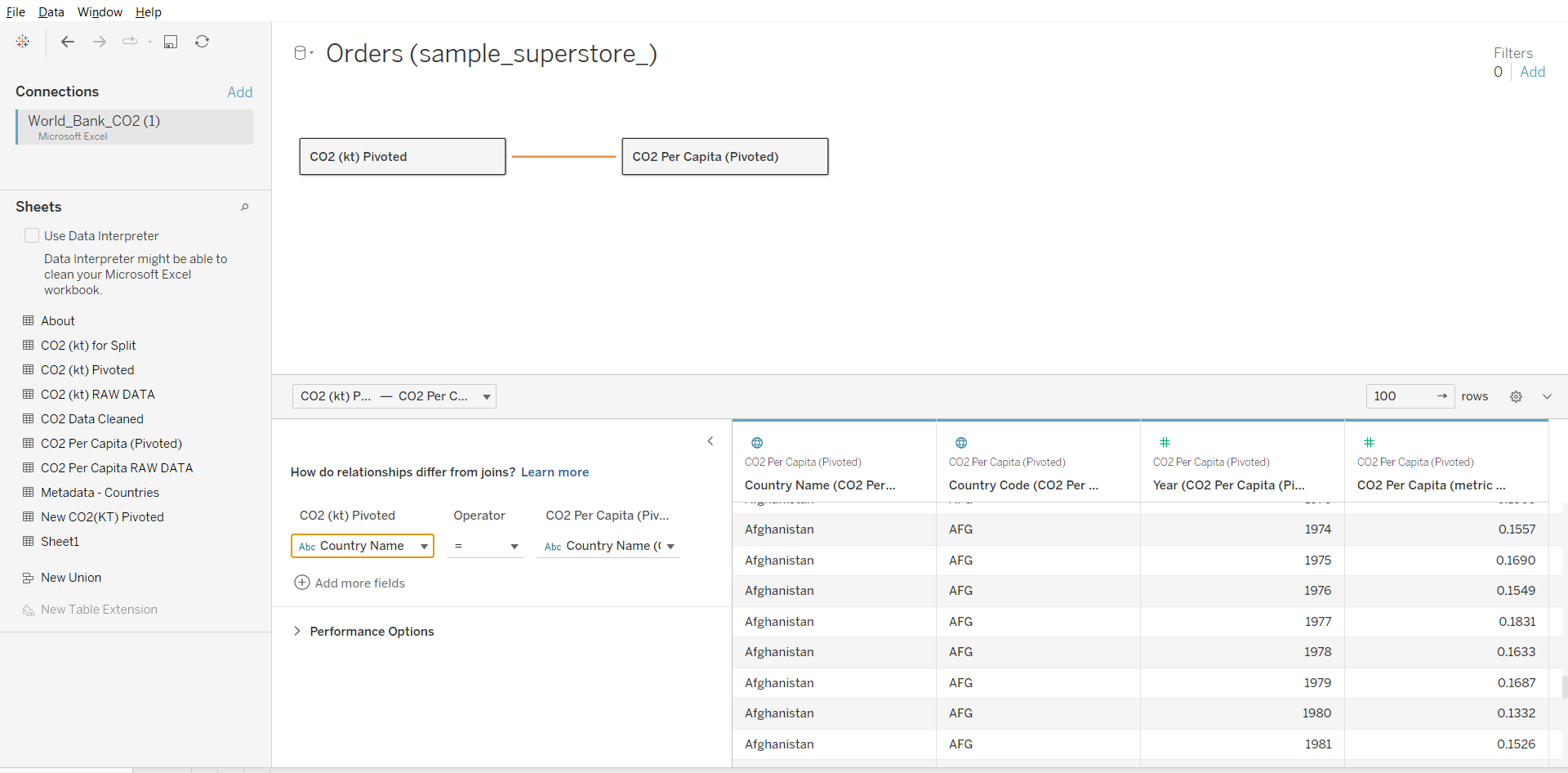
A relationship describes how two tables relate to each other, based on common fields, but does not merge the tables together.  
When a relationship is created between tables, the tables remain separate, maintaining their individual level of detail and domains.

Think of a relationship as a contract between two tables. When you are building a viz with fields from these tables, Tableau brings in data from these tables using that contract to build a query with the appropriate joins.

[Advantages of Relationships](https://help.tableau.com/current/pro/desktop/en-us/relate_tables.htm#:~:text=Relationships%20provide%20several%20advantages%20over%20using%20joins%20for%20multi%2Dtable%20data%3A)[Relationships vs Joins](https://help.tableau.com/current/server/en-us/datasource_relationships_learnmorepage.htm)

**Note:** We’ll be using the same example as we saw for Joins but instead of Joins here we’ll be building relationships for understanding the topic.

* Add the World Bank CO2 dataset
* Drag CO2 (kt) Pivoted on the canvas
* Drag CO2 Per Capita Pivoted on the canvas
* In the clause, set the field for both tables to the country name.



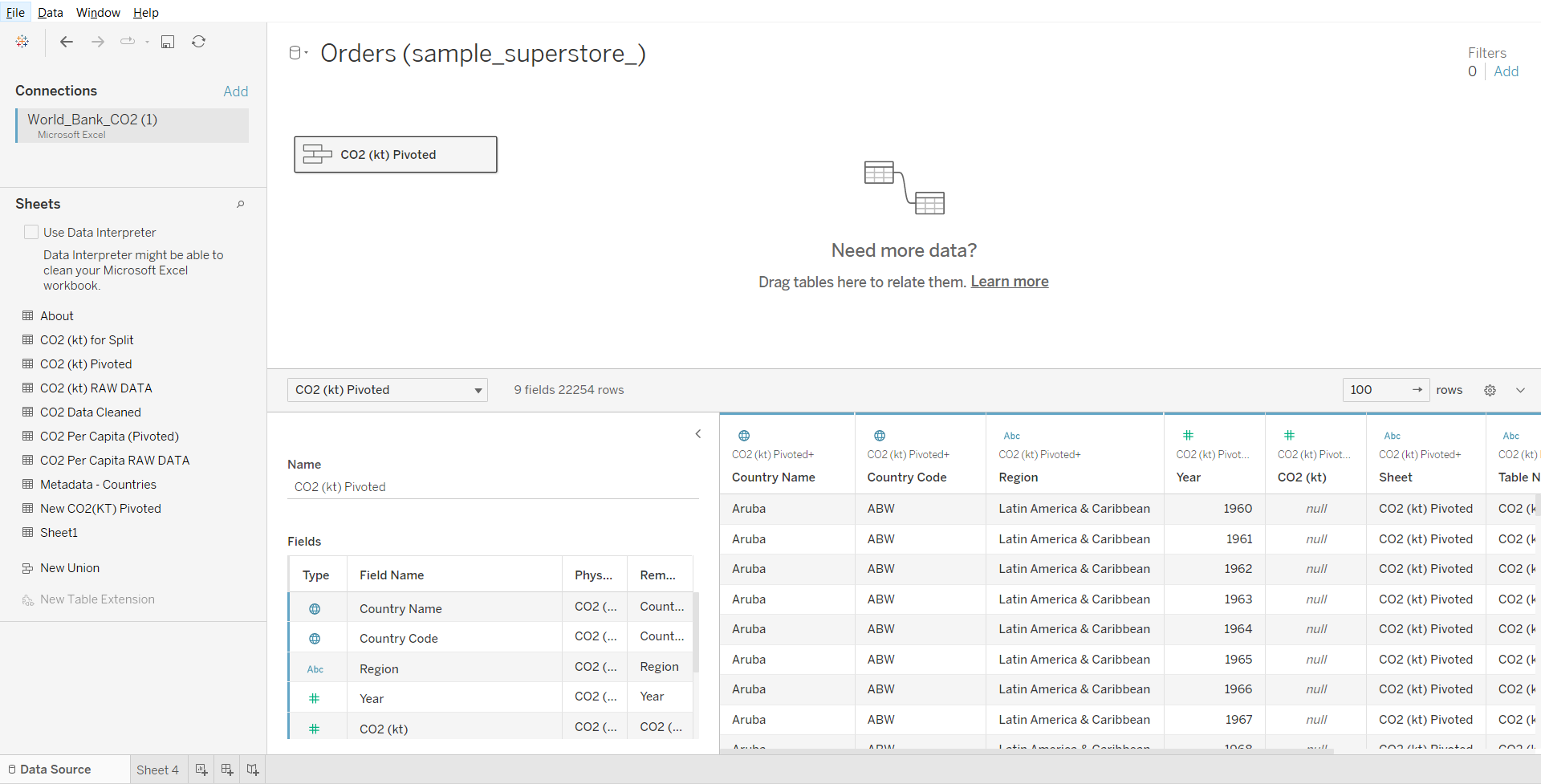
**Union**

Though union is not a type of join, union is another method for combining two or more tables by appending rows of data from one table to another.  
Make sure the tables you union have the same number of fields, the same field names, and the fields are the same data type.

Data: [World\_bank\_CO2.xlsx](https://docs.google.com/spreadsheets/d/1zUMnxp9aPdV5kYWJSoCkVqTEurJvd86d/edit?usp=share_link&ouid=100659516601446935794&rtpof=true&sd=true)

**Perform Union on table CO2 (kt) Pivoted and New CO2 (kt) Pivoted**

* Drag CO2 (kt) Pivoted to the canvas
* Double click ‘CO2 (kt) Pivoted’ to open the physical layer.
* Drag New CO2 (kt) Pivoted just below ‘CO2 (kt) Pivoted’ on the canvas.
* Close the physical layer.



**Blending**

Data blending is a method for combining data from multiple sources. Data blending brings in additional information from a secondary data source and displays it with data from the primary data source directly in the view.

**Example:** Consider the Sales data is present in a relational database and Sales Target data in an Excel spreadsheet. Now, to compare actual sales to target sales, you can blend the data based on common dimensions to get access to the Sales Target measure. The two sources involved in data blending are referred to as primary and secondary data sources.

**Note:** A left join is created between the primary data source and the secondary data source.

#### 

#### **Business problem 2 :**

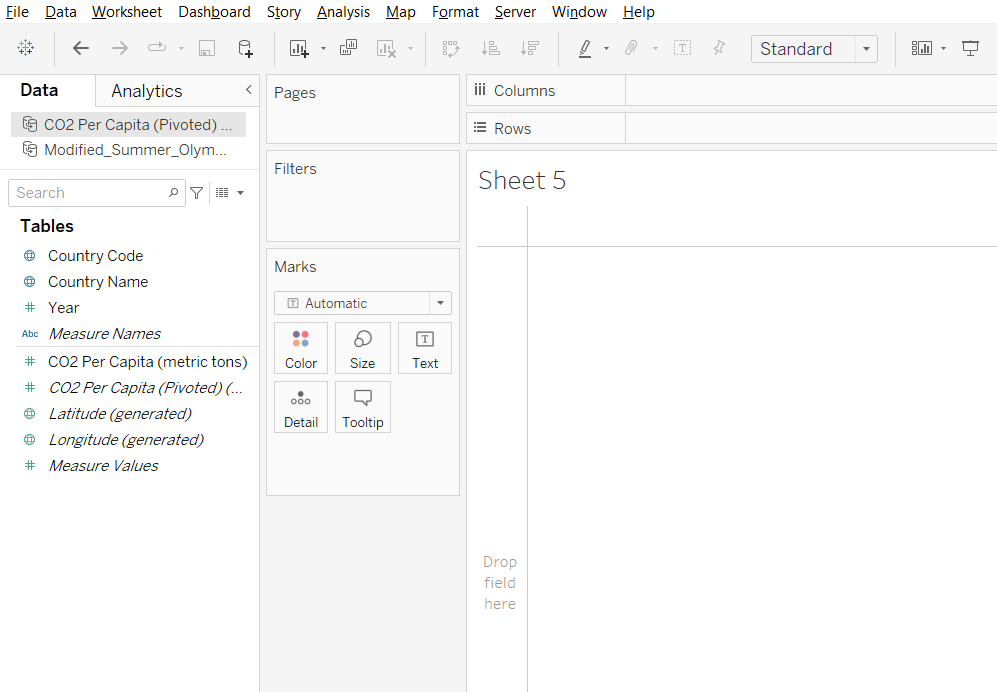
For each country, find the total number of olympic medals won and CO2 per capita (metric tons) emission till 2008.

Dataset : [Modified\_Summer\_Olympic\_medallists\_1896-2008.xlsx](https://docs.google.com/spreadsheets/d/1nD7DYtJlsnbKQEAvHfY13xzvBsuL3pM1/edit?usp=share_link&ouid=100659516601446935794&rtpof=true&sd=true) & [World\_bank\_CO2.xlsx](https://docs.google.com/spreadsheets/d/1zUMnxp9aPdV5kYWJSoCkVqTEurJvd86d/edit?usp=share_link&ouid=100659516601446935794&rtpof=true&sd=true)

Use table Team event fixed from Modified\_Summer\_Olympic\_medallists\_1896-2008  
Use table CO2 per capita (Pivoted) from World\_bank\_CO2 dataset

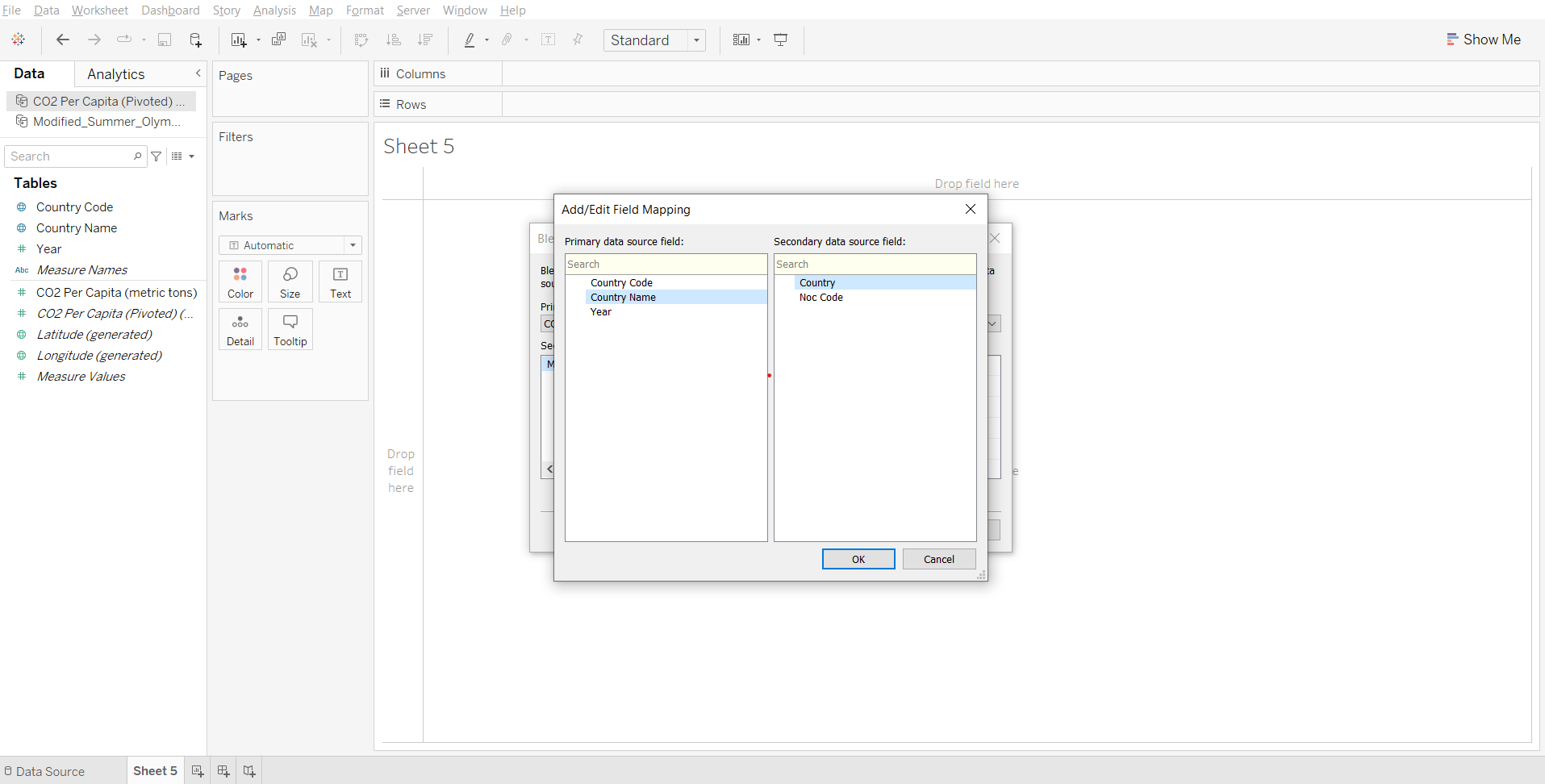
1. Loading 2 Data sources in a Single workbook

* Since world bank CO2 dataset is already connected, drag CO2 (kt) Pivoted to the canvas
* Go to worksheet
* In the Data menu, click new data source and select the modified summer olympics excel file.
* Go to Data Source and from modified summer olympics file, drag the team events fixed table to the canvas.



1. Edit Blend Relationship

* Click on CO2 Per Capita in Data panel
* Go to the Data menu, click Edit Blend Relationships…
* Select Custom and click on Add…
* Choose the country name from the Primary Data Source field, click ok on both dialog boxes.

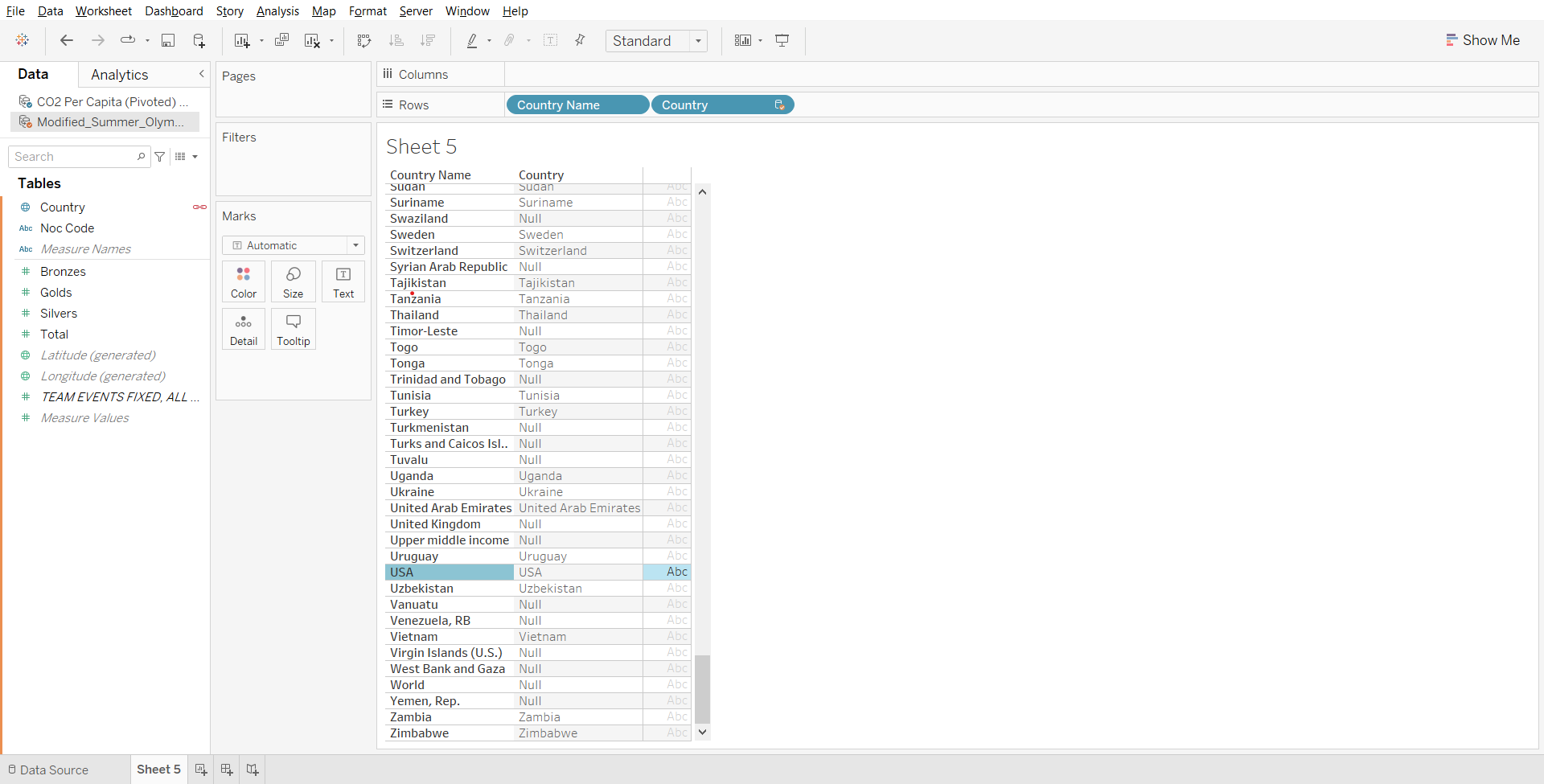


1. Edit Alias

* If you drag the country name from CO2 data to the rows shelf, an orange link appears next to ‘country’ in Olympics data, saying - ‘Stop using country as a linking field’.

This is because the name of some countries (like the USA) is written differently in the two files. We can set an alias of a country if needed using the following steps -

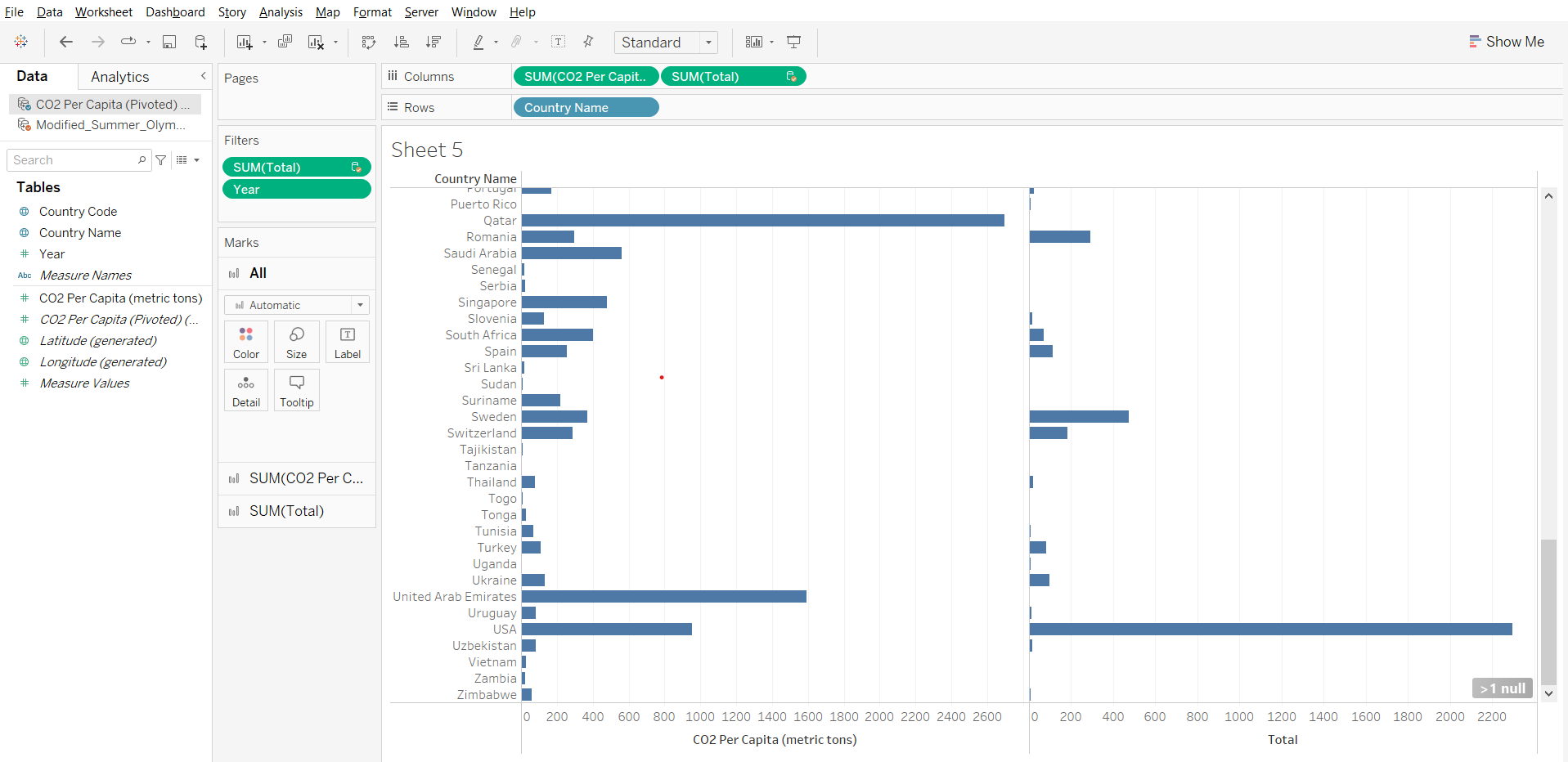
* In CO2 data, from the country name dropdown in the data panel, click Aliases…
* Set the alias for the United States as the USA.
* Drag country name from CO2 data and country from Olympics data to rows shelf.
* Observe that the entries for ‘USA’ are next to each other.



**Note:** If the same error pops up when bringing country name from one table just click on the orange link that appears

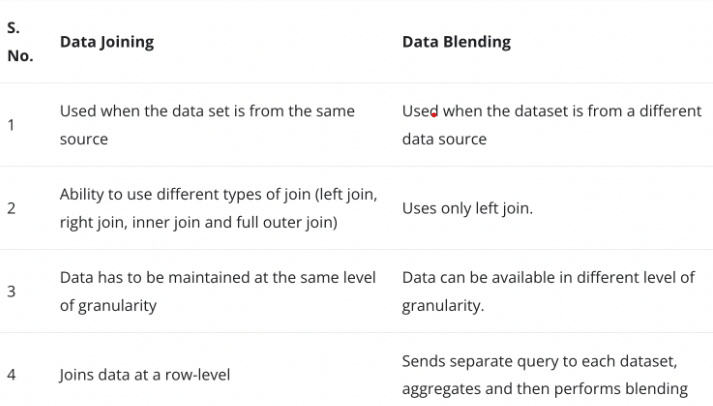
1. Solving the business problem:

* Drop country name to rows
* Drop CO2 Per Capita to columns
* Drop Total from Olympics data to columns
* Click on ‘>149 nulls’ in bottom right corner of the sheet and select Filter data
* Drag Year in CO2 data to filters and set maximum value to 2008.

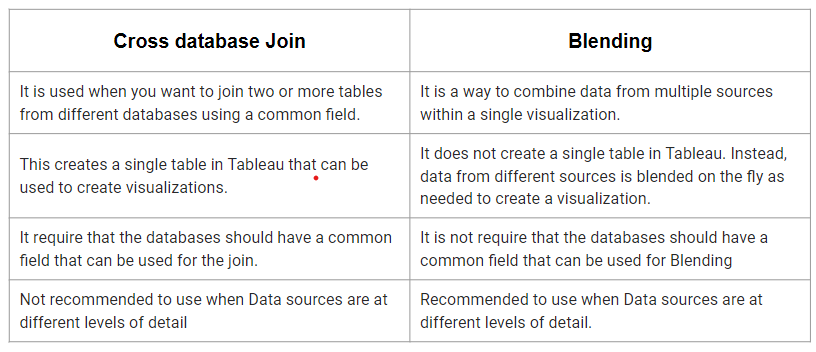


### 

### **Joins vs Blend**

****

### **Cross-database Join vs Blending**

****