Tableau Lecture 1: Intro to Data Visualization, Basic Charts and Operations

Agenda

- Tableau Basics
- Why choose Tableau?
- Tableau Product Suite
- Download & Install Tableau
- TWB vs TWBX file format
- Tableau Public GUI
- Data Fields
- Data Types
- Measures & Dimensions
- Discrete vs Continuous fields
- Rows, Columns, Marks shelf
- Visual Analytics
- Data Connections in Tableau
    - Live connection
    - Extract connection
- Aggregation on Dimensions and Measures
- Basic Charts
    - Line chart
    - Bar chart
    - Pie chart
    - Histogram
- Cleaning and Formatting data for analysis
    - Data Interpreter
    - Pivoting Data
    - Data Source Filter

Dataset: **sample superstore**

Introduction to Data Visualization and Tableau

Dashboard Demonstration

Given the dataset : Netflix Movies & TV shows

Show :-

- Country-wise top 10 directors
- Number of movies/shows per year
- The list of movies and shows
- Ratings count for movies & shows for that country.

Tableau Public Dashboard :
https://public.tableau.com/app/profile/tino3819/viz/NetflixDashboard_16521889408740/Dashboard1

The Basics :

What is Business Intelligence(BI)

- BI combines business analytics, data mining, data visualization, data tools and infrastructure, and best practices to help organizations to make more data-driven decisions.

Why BI?

- It helps companies make better decisions by showing present and historical data within their business context.

What is Data Visualization?

- It is the graphical representation of information and data.
- By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- Humans are good at deriving knowledge from visualizations.

Brief History

- Until the early 21st century, Database, Excel, Access, etc. were used to produce numbers and data.
- The main idea behind Tableau's creation was to make the database industry interactive and comprehensive.
- Tableau is a popular data visualization and business intelligence tool used for reporting and analyzing vast volumes of data.
- Tableau was founded by Pat Hanrahan, Christian Chabot, and Chris Stolte from Stanford University in 2003.

Over the years…

- Tableau has been named a Leader in the Gartner Magic Quadrant for Analytics & Business Intelligence Platforms for the 10th consecutive year.
- Tableau Software has a market capitalization of $14.61 billion and generates $982.95 million in revenue each year.

Why choose Tableau?

- **Pros of Tableau**
  - Quick and interactive visualizations
  - Easy to use for non programmers
  - High performance
  - Mobile friendly
  - Extensive customer resources (Tableau Community)
  - Working with different data sources
  - Easy to upgrade
- **Cons of Tableau**

    - Focuses primarily on visualization and cannot work with uncleaned data. In order to efficiently use Tableau, you need to do proper data cleaning in the underlying database first.
    - Lacks data modeling and data dictionary capabilities for Data Analysts. This means that you've to separately maintain your metrics definitions elsewhere.
    - Lack of version control and collaboration when building data logic and dashboard.

Tableau Product Suite

- Tableau Prep
- Tableau Desktop
- Tableau Server
- Tableau Online
- Tableau Reader
- Tableau Public

References :

- Comparison of Product Suite
- Tableau Pricing
- Understanding License types of Tableau
- Tableau Desktop vs Tableau Public

Download & Install Tableau Public

Pre-read: Tableau Download & Installation Steps

TWB and TWBX file formats

In Tableau, `TWB` and `TWBX` are two different file formats used to save Tableau workbooks.

The choice between these file formats can significantly impact data sharing and collaboration in Tableau.

TWB (Tableau Workbook):

- XML-based text files containing references to data sources and workbook settings.
- Contains workbook layout, visualizations, calculations, and metadata.
- Lightweight and smaller in size compared to TWBX files.
- Used for sharing with collaborators having access to the same data source or when publishing workbooks to Tableau Server or Tableau Online.
- Allows real-time data access for analysis.

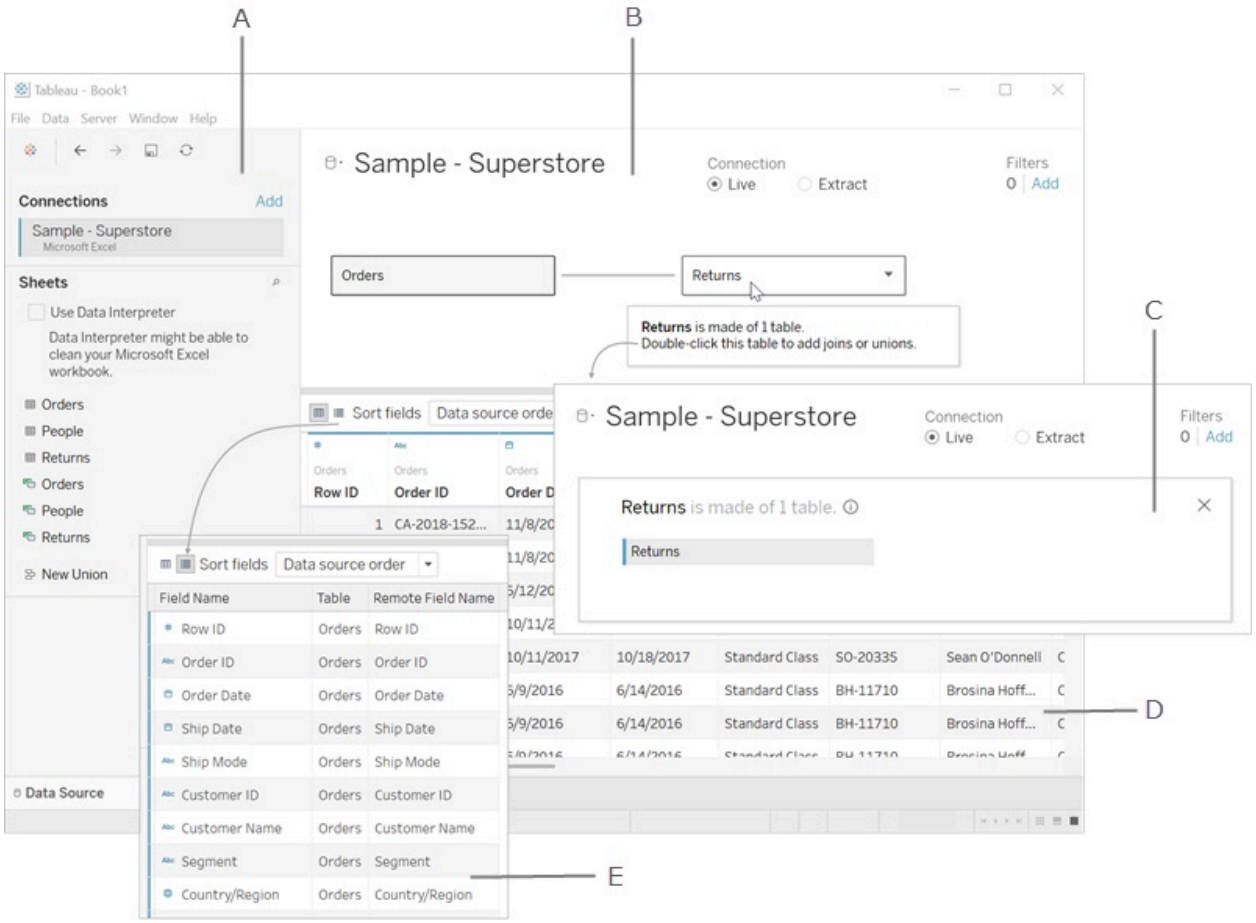
TWBX (Tableau Packaged Workbook):

- Packaged files containing the workbook along with its supporting data and resources.
- Includes workbook, data extracts, images, and custom fonts.
- Larger in size due to packaging of resources.
- Used for sharing with recipients without access to the original data source.
- Enables offline analysis and archiving.

Tableau Public GUI and Basics

Loading dataset in Tableau and Data Source view

Dataset: **link**



A. Left pane - Displays the connected data source and other details about your data.
B. Canvas: logical layer - The canvas opens with the logical layer, where you can create relationships between logical tables.
C. Canvas: physical layer - Double-click a table in the logical layer to go to the physical layer of the canvas, where you can add joins and unions between tables.
D. Data grid - Displays first 1,000 rows of the data contained in the Tableau data source.
E. Metadata grid - Displays the fields in your data source as rows.

Reference:

- https://help.tableau.com/current/pro/desktop/en-us/environment_datasource_page.htm

---

Data Fields

- After you connect to your data and set up the data source with Tableau, the data source connections and fields appear on the left side of the workbook in the Data pane.
- The term **"fields"** refers to **columns**.
- When you connect to a new data source :
    - Each field is automatically assigned a **Data Type** (such as integer, string, date).
    - Tableau assigns each field in the data source as **dimension** or **measure** in the Data pane, depending on the type of data the field contains.
        - Discrete Dimension or Continuous Measure (more common)
        - Continuous Dimension or Discrete Measure (less common)

Reference:

- https://help.tableau.com/current/pro/desktop/en-us/datafields_typesandroles.htm
- https://help.tableau.com/current/pro/desktop/en-us/datafields_understanddatawindow.htm

---

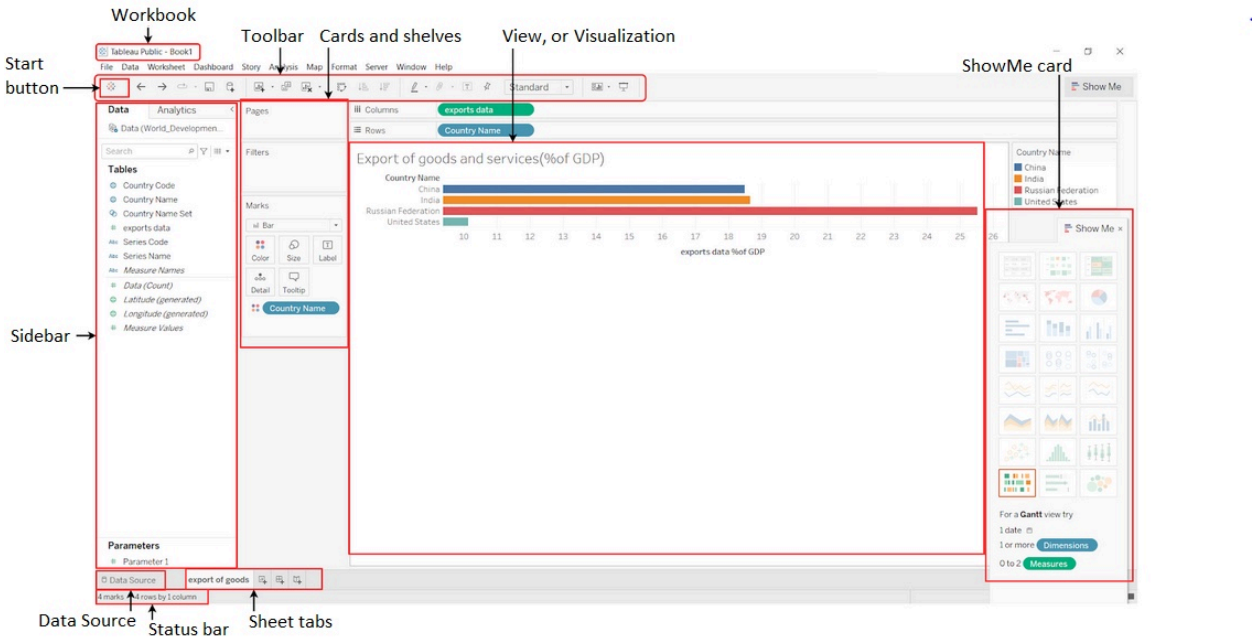
Data Types

- Tableau expresses fields and assigns data types automatically.
- If the data source appoints the data type, Tableau will use that data type.
- If the data source doesn't individually assign a data type, Tableau will assign one

Tableau Data Types

Data Type	Icon
String values (Text)	Abc
Integer values (Numbers)	#
Date values (DD/MM/YYYY or MM/DD/YYYY)	
Date & Time values	
Boolean values (True or False; relational)	T|F
Geographic values (Region, Postal code etc.)	
Cluster group or mixed values	

Tableau View page



Measure & Dimension

1. Measure:

- A measure is a field that is a dependent variable; that is, its value is a function of one or more dimensions.
- Tableau treats any field containing **numeric** (quantitative) information as a measure.

2. Dimension:

- Dimension is a field that can be considered an independent variable.
- By default, Tableau treats any field containing qualitative, **categorical** information as a dimension.

Rule of Thumb :

Generally,

- the measure is the number;
- the dimension is what you “slice and dice” the number by.

Discrete and Continuous fields

- Tableau represents data differently in the view depending on whether the field is discrete (blue), or continuous (green).
- Continuous and Discrete are mathematical terms.
    - Continuous means "forming an unbroken whole, without interruption"
    - Discrete means "individually separate and distinct"

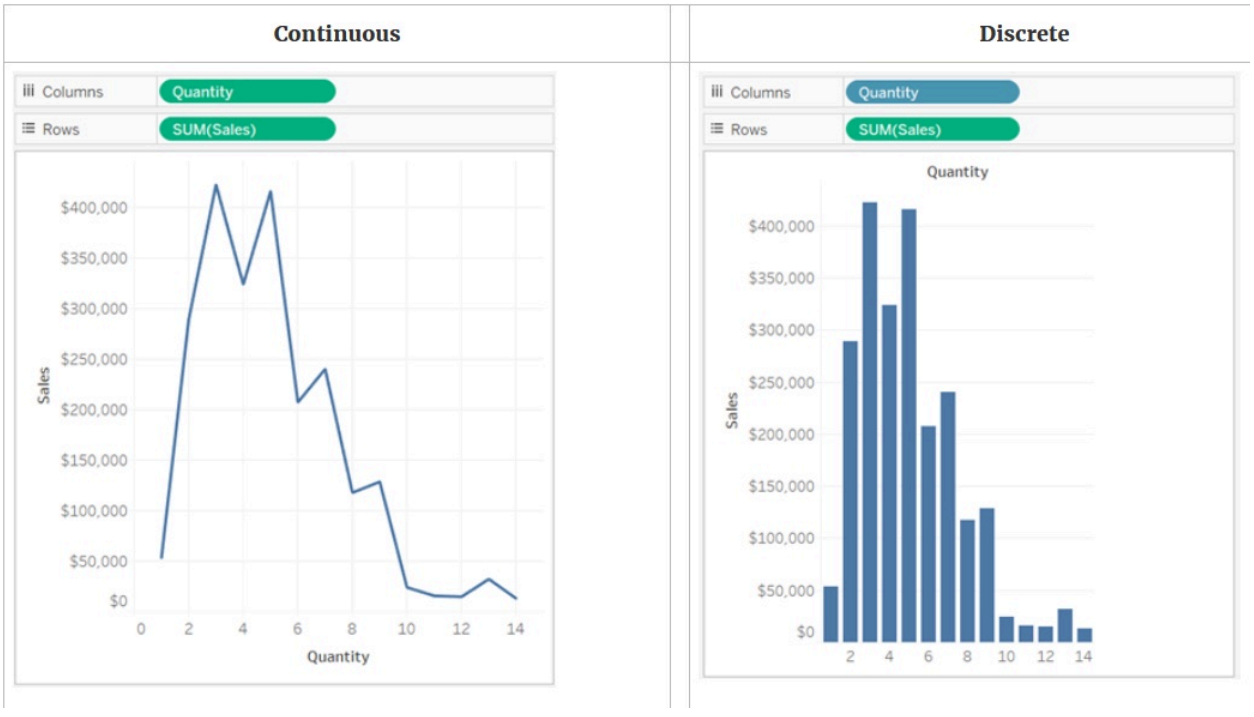
- Green measures SUM(Profit) and dimensions YEAR(Order Date) are continuous. Continuous field values are treated as an infinite range. Generally, continuous fields add axes to the view.
- Blue measures SUM(Profit) and dimensions Product Name are discrete. Discrete values are treated as finite. Generally, discrete fields add headers to the view.

Example :

- Continuous
    - A person's weight, tracked over the past 6 months, has seen a progression from 80 kgs to 85 kgs.
    - Within this range, an infinite number of potential values, such as 81.5, 82.47, 84.999, and so forth, exist.
    - Therefore, the weight of an individual is considered continuous.
- Discrete
    - Consider a scenario where a retailer sold 10 cars in the year 2020 and 15 cars in 2021.
    - In this case, there are only four distinct possible values between these two sales figures: 11, 12, 13, and 14.
    - As a result, the number of cars sold is categorized as discrete.

Rule of Thumb :

- Discrete fields draw headers; Continuous fields draw axes.
- Discrete fields can be sorted; Continuous fields cannot.
- Blue colored field indicates Discrete Field.
- Green colored field indicates Continuous Field.



- In the example on the left (above), because the Quantity field is set to Continuous, it creates a horizontal axis along the bottom of the view.

    - The green background and the axis help you to see that it's a continuous field.
- In the example on the right (above), the Quantity field has been set to Discrete. It creates horizontal headers instead of an axis.
    - The blue background and the horizontal headers help you to see that it's discrete.
- In both examples, the Sales field is set to Continuous.
    - It creates a vertical axis because it is Continuous and it's been added to the Rows shelf.
    - If it was on the Columns shelf, it would create a horizontal axis.
- The green background and aggregation function (in this case, SUM) help to indicate that it's a Measure.
- The absence of an aggregation function in the Quantity field name helps to indicate that it's a Dimension.

Converting Measure to Dimension

- You can convert a field from a measure to a dimension in the current view.
- If you want the change to affect all future uses of the field in the workbook, you can convert a field in the Data pane from a measure to a dimension.

To convert a measure to a dimension in the Data pane, do either of the following :

- Click and drag the field from the measures area in the Data pane and drop it into the dimensions area (above the line).
- Right-click (control-click on a Mac) the measure in the Data pane and select Convert to Dimension.

If you place a field that you converted from a measure to a dimension on a shelf, it now produces headers instead of an axis.

Note: The reason for this will become clearer as the lectures progress.

Measures

- can be placed on the Columns shelf to create the X-axis (horizontal bar chart) and
- on the Rows shelf to create the Y-axis (vertical bar chart).

Dimensions

- can be placed on the Columns shelf to categorize data horizontally across the X-axis (bar charts or histograms) and
- on the Rows shelf to categorize data vertically along the Y-axis (stacked bar charts).

In Tableau, the **"Marks"** card is a key element in determining how your data is visualized.

The Marks card is divided into several shelves. Here are the main elements of the Marks card shelf:

1. **Color**: Determines the color of the marks in your visualization.
2. **Size**: Determines the size of the marks in your visualization.
3. **Label**: Determines the text labels associated with the marks in your visualization.
4. **Detail**: Provides additional level of detail to your visualization for more granular analysis.
5. **Tooltip**: Determines the information displayed when hovering over the marks in your visualization.
6. **Path**: Used in path-based visualizations like line charts to specify the order of data points.
7. **Mark Type**: This dropdown menu allows you to select the type of mark you want to use in your visualization (e.g., bars, lines, circles).

The placement of fields in the Rows and Columns shelves, along with the Marks card settings, defines the structure of your visualization.

Here's a general guideline for placing fields under Marks, Rows, and Columns:

Marks card:

- **Dimensions (Categorical Variables)**:
    - **Color**: Use for categorical variables to distinguish between different categories or groups in your data.
    - **Size**: Use for categorical variables if you want to encode the size of data points based on some measure.

- **Measures (Numeric Variables)**:
    - **Color**: Use for numeric variables if you want to encode data points with a color gradient based on their numeric values.
    - **Size**: Use for numeric variables if you want to encode the size of data points based on their numeric values.
- **Text Variables**:
    - **Label**: Use for text variables if you want to add written information directly on data points in a view.
    - **Tooltip**: Use for text variables if you want to add written information when hovering over data points in a view.

Rows shelf:

- **Dimension (Categorical Variable)**:
    - Drag a dimension to the Rows shelf to create rows in your visualization.
    - Each row represents a unique value of the dimension.
- **Measure (Numeric Variable)**:
    - Drag a measure to the Rows shelf to create rows based on the measure's values.

Columns shelf:

- **Dimension (Categorical Variable)**:
    - Drag a dimension to the Columns shelf to create columns in your visualization.
    - Each column represents a unique value of the dimension.
- **Measure (Numeric Variable)**:
    - Drag a measure to the Columns shelf to create columns based on the measure's values.

Dataset : sample superstore

Business problem 1:

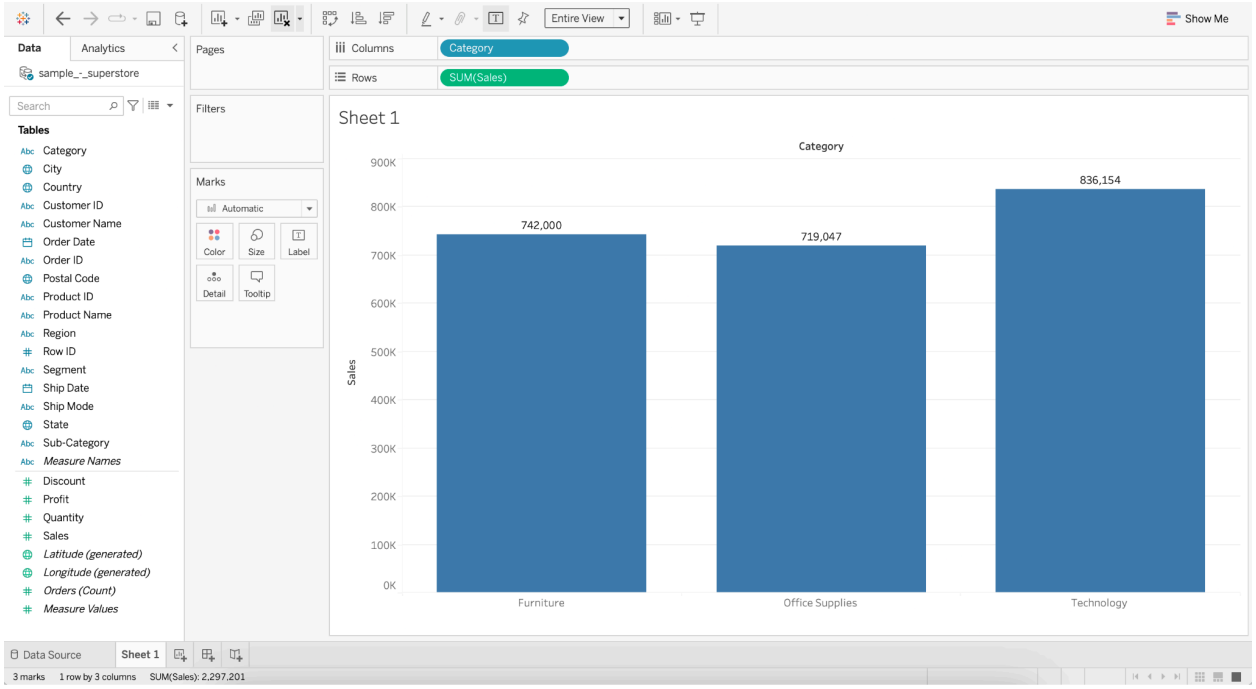
Determine the total sales value of each category.

How to choose what goes to Rows shelf and what goes to Columns shelf?

- **Category** being a **dimension** by which we wish to group the data, goes to **Columns shelf** (X-axis).
- **Sales** being a **measure** that we want to aggregate (SUM), goes to **Rows shelf** (Y-axis).

Steps:

- Use the **Orders** table
- Drag **Category** to Columns
- Drag **Sales** to Rows
- Switch to **Entire View**
- Click on Label and check "Show marks labels"



What is Visual Analytics?

- It is the use of sophisticated tools and processes to analyze datasets using visual representations of the data.

Why is it important?

- Users get actionable insights which in turn help organizations make better, data-driven decisions.
- Allows users without data science skills or experience to combine, manipulate, and explore large, dynamic, multi-dimensional, and multi-sourced datasets.

Difference between Visual Analytics and Data Visualization

Data Visualization	Visual Analytics
Graphical/Visual depiction of data to help people better understand the **patterns**, **relationships**, **trends**, and other **meaningful insights** in datasets.	The use of an **analytics program** to perform **advanced analysis** of **complex datasets** allowing users to explore and interact with **dynamic visualizations**.

Tableau Live and Extract Connections

Live and Extracts are two ways you can make the data connection to the Tableau.

Live connection	Extract connection
Live allows you real-time data.	Extract is kind of a batch which needs to be refreshed from time to time to get the updated data.
Whatever changes are done at the data source will be directly available to Tableau Desktop (professional).	Any changes made in the data source won't reflect in the report immediately. It will be reflected when the extract is refreshed.

Note:

- Tableau Public only supports extract connections.

- Tableau Desktop (professional) supports both live and extract connections.

Types of Aggregation on measures and dimensions

- We can add aggregation to dimensions and measures in Tableau.
- Types of aggregation on measures
- We can convert a dimension to a measure and perform aggregation on it.
- Types of aggregation on dimension
- If you want to remove the default aggregation for fields in the view, go to **Analysis** and uncheck the **Aggregate Measures** option.

Business problem 1:

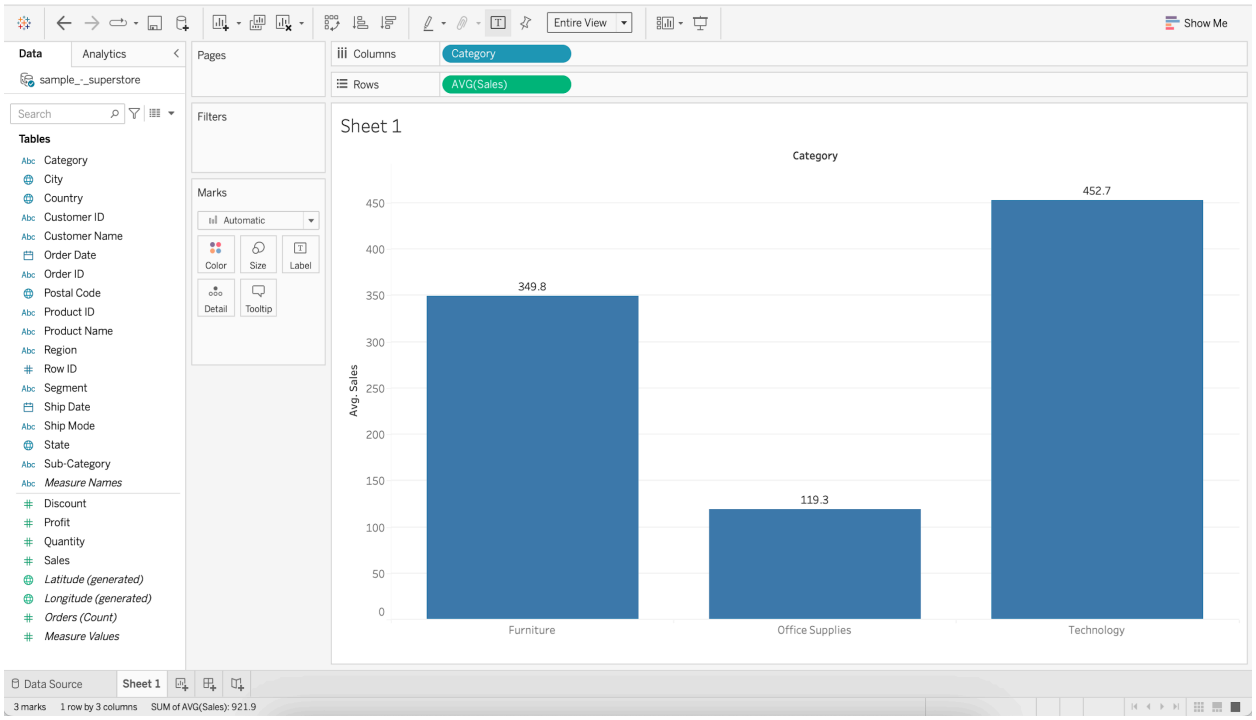
Determine the average sales value of each category.

How to choose what goes to Rows shelf and what goes to Columns shelf?

- **Category** being a **dimension** by which we wish to group the data, goes to **Columns** shelf (X-axis).
- **Sales** being a **measure** that we want to aggregate (AVG), goes to **Rows** shelf (Y-axis).

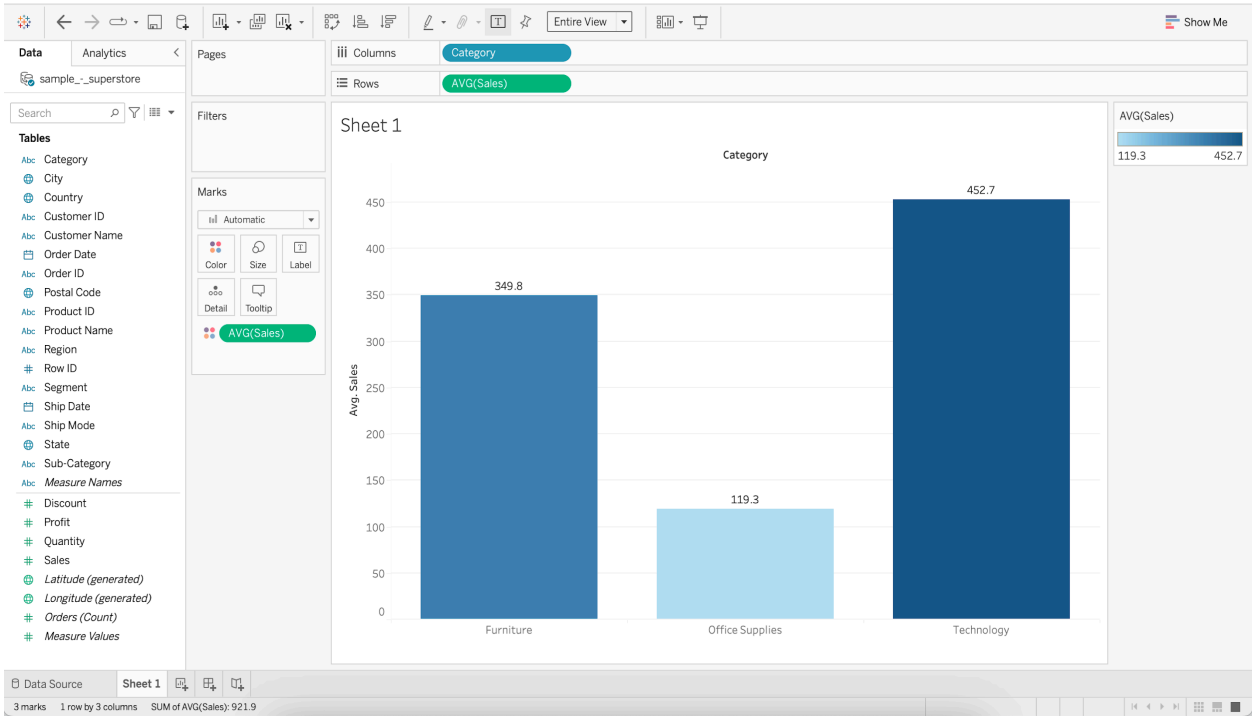
Steps:

- Drag **Category** to Columns
- Drag **Sales** to Rows
- Change aggregation of **Sales** from Sum to Average

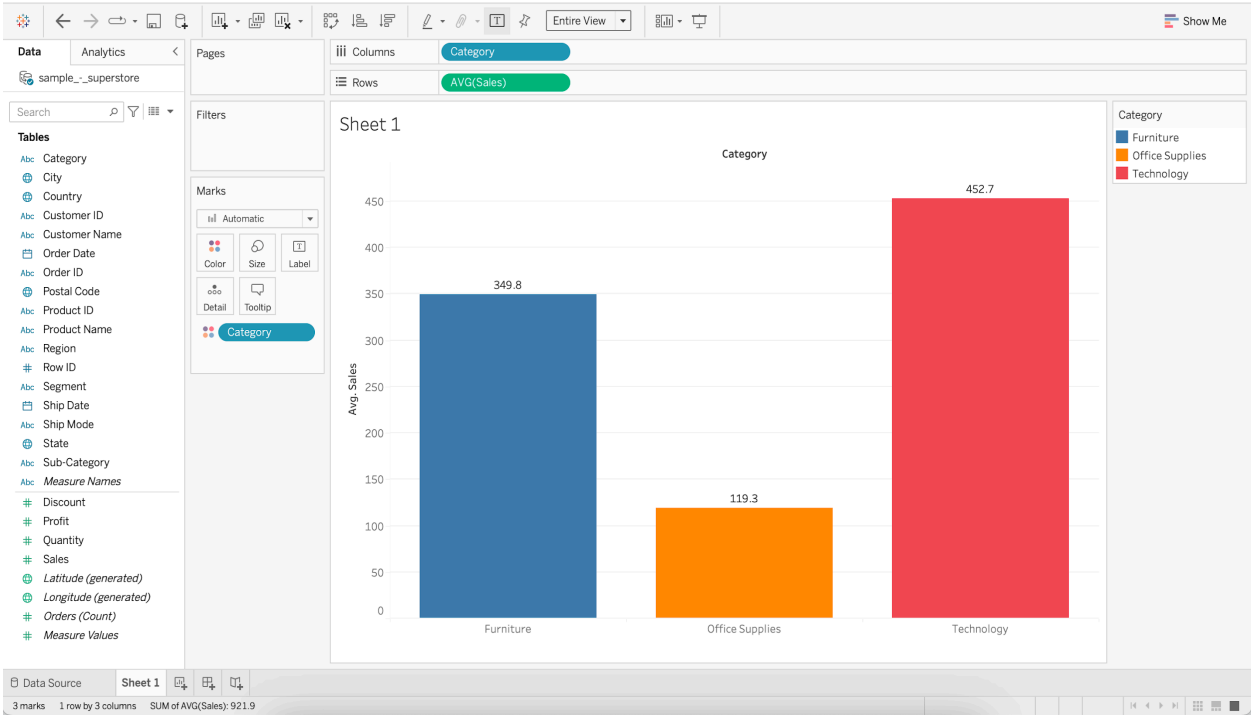


We can demonstrate Color marks shelf -

- If we drag **Sales** (Measure) onto the **Color** marks shelf, it'll assign different shades of a single color to the **AVG of Sales** values.



- If we drag the **Category** (Dimension) to the **Color** marks shelf, we'll get a different color for each category.



Business problem 2:

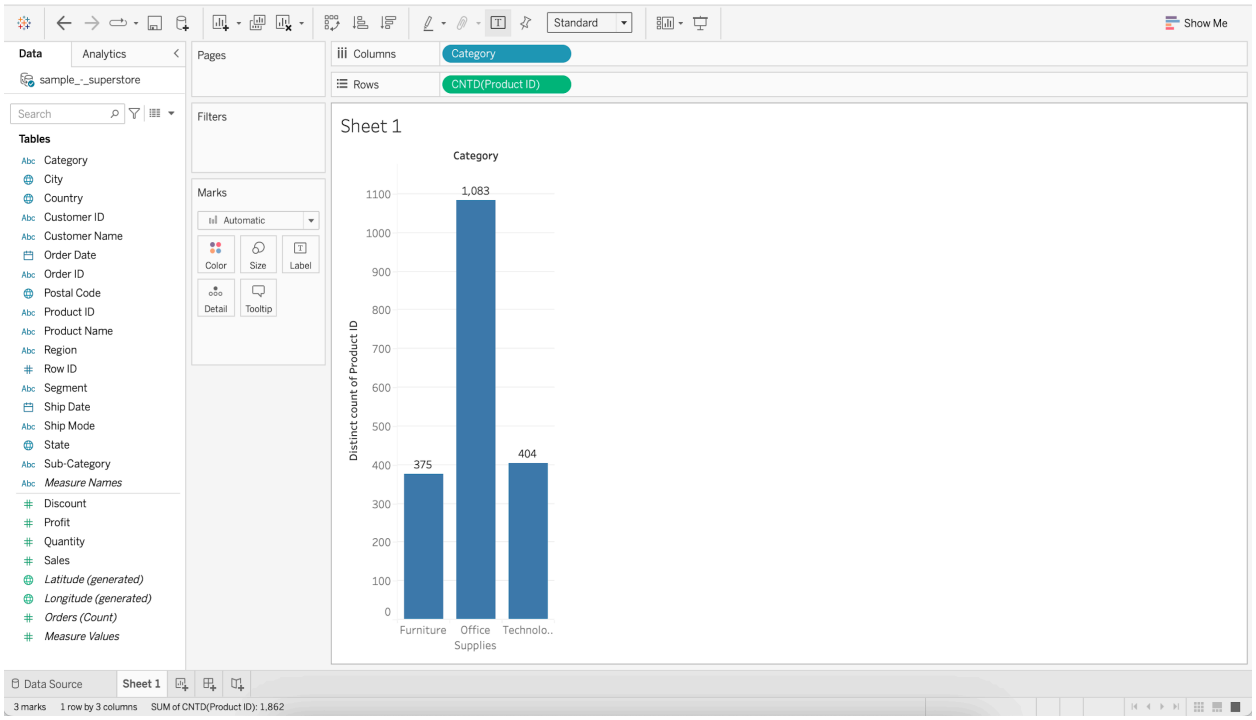
Determine total number of products within each category.

How to choose what goes to Rows shelf and what goes to Columns shelf?

- **Category** being a **dimension** by which we wish to group the data, goes to **Columns** shelf (X-axis).
- **Product ID** is a **dimension** but it still goes to **Rows** shelf (Y-axis) since we plan to use it as a **measure** and count (CNTD) the unique Product IDs.

Steps:

- Drag **Category** to Columns
- Drag **Product ID** to Rows
- Change aggregation of **Product ID** to Count Distinct



Line chart

They connect individual data points in a view. They provide a simple way to visualize a sequence of values and are useful when you want to see trends over time.

Discrete line chart

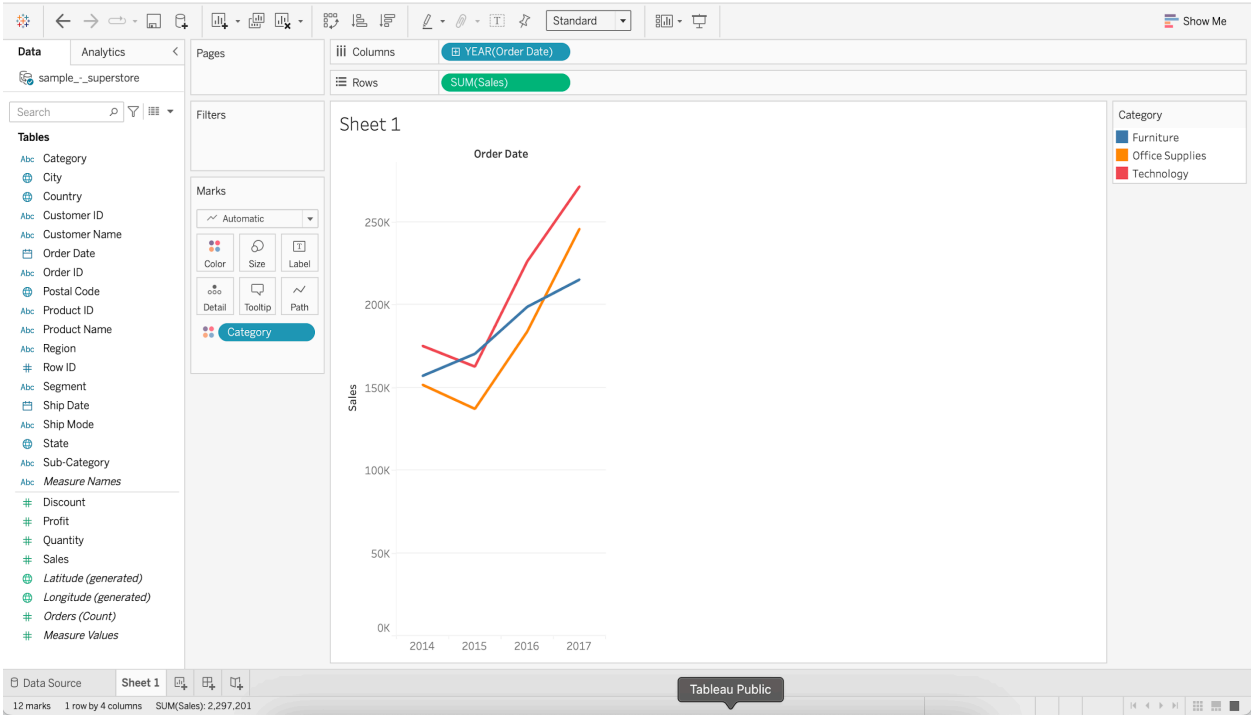
Business problem 3 : Find the product category that has the highest sales throughout most of the years.

How to choose what goes to Rows shelf and what goes to Columns shelf?

- Year of **Order Date** being a **dimension** by which we wish to group the data, goes to **Columns** shelf (X-axis).
- **Sales** being a **measure** that we want to aggregate (SUM), goes to **Rows** shelf (Y-axis).
- **Category** being a **dimension**, if put into the **Color** marks shelf, we'll get a different color for each category.

Steps:

- Drag **Order Date** to Columns
- Drag **Sales** to Rows
- Add **Category** to Color



Continuous line chart

Business problem 4 : Find the year and month that had the highest and lowest sales.

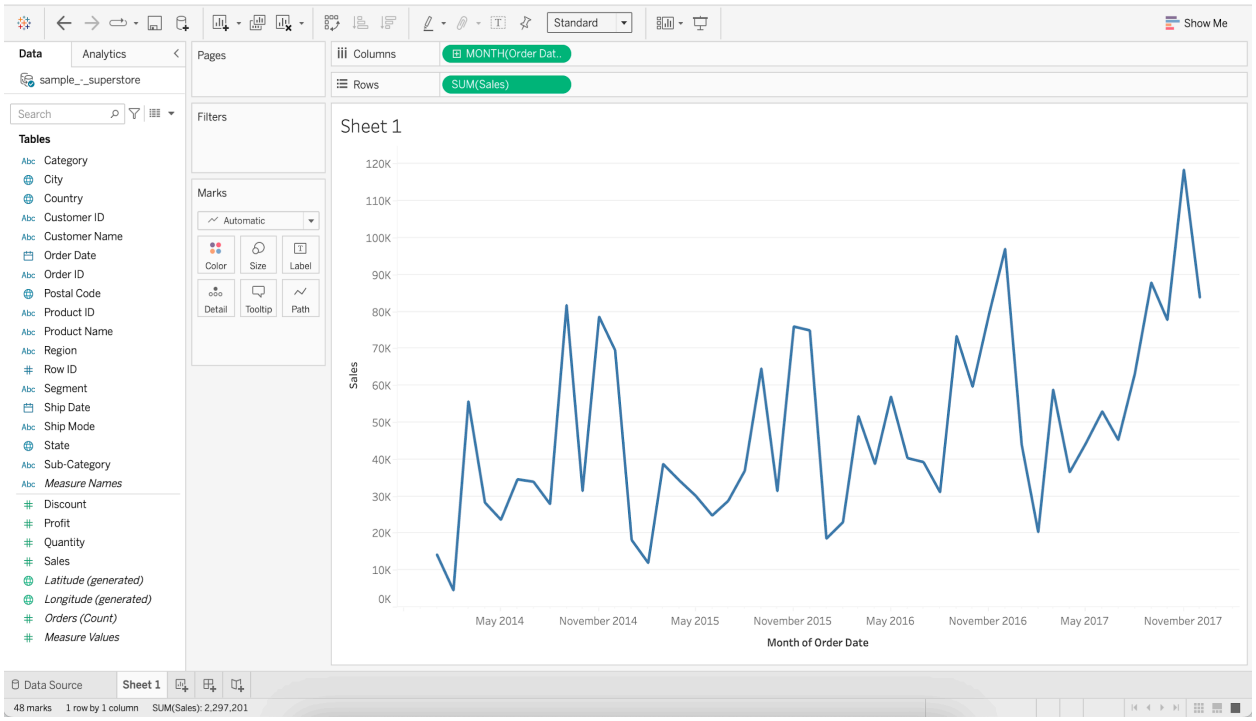
How to choose what goes to Rows shelf and what goes to Columns shelf?

- Month of **Order Date** being a **dimension** by which we wish to group the data, goes to **Columns** shelf (X-axis).
- **Sales** being a **measure** that we want to aggregate (SUM), goes to **Rows** shelf (Y-axis).

Steps:

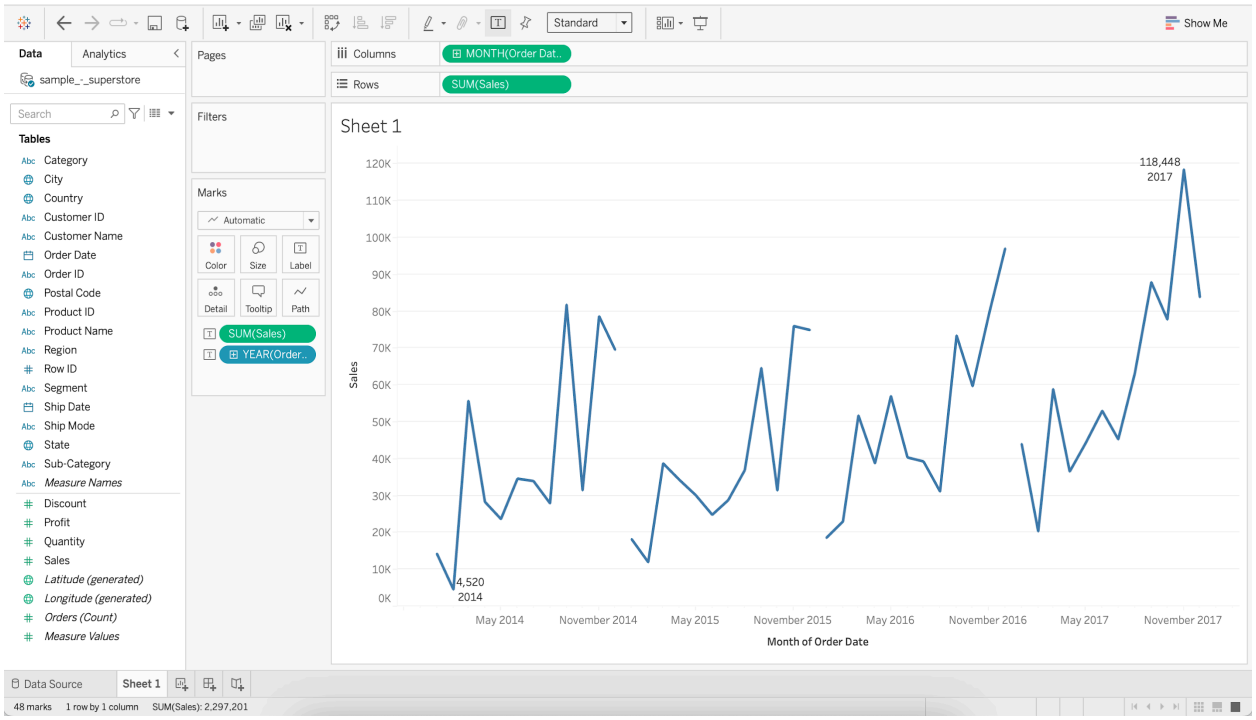
- Drag **Order Date** to Columns
- Drag **Sales** to Rows

- Switch from YEAR (E.g. 2015) to MONTH (E.g. May 2015)



Annotating the Line chart with min and max Sales -

- Add **Sales** to Label
- Add **Order Date** to Label
- Click on Label and choose "Marks to Label" as Min/Max



- Putting **Sales** (Measure) on the **Label** marks shelf will add written information about **SUM of Sales** to the chart.
- Putting **Order Date** (Dimension) on the **Label** marks shelf will add written information about **YEAR of Order Date** to the chart.

Bar chart

It is a chart with rectangular bars where lengths and heights are proportional to the value that they represent.

Horizontal bar chart

It is a bar graph that represents data horizontally, dimensions are present on the vertical axis and data values on the horizontal axis.

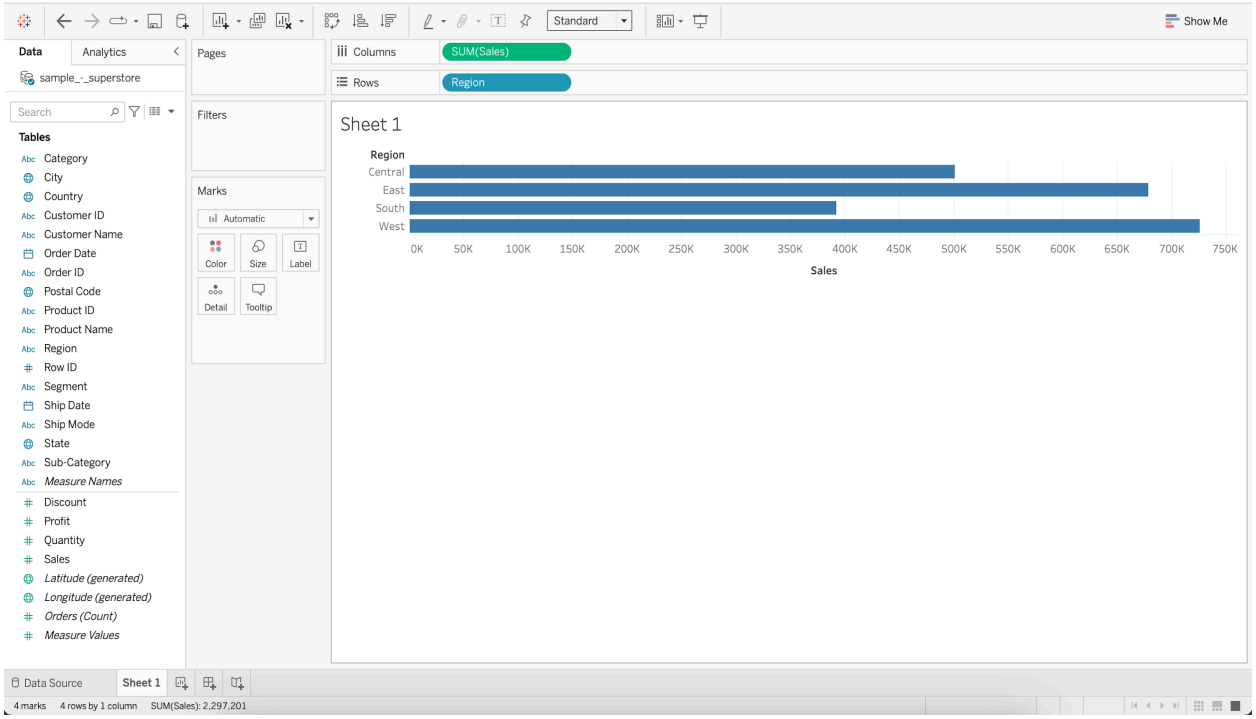
Business problem 5 : Find sales and profit by region for the year 2017.

How to choose what goes to Rows shelf and what goes to Columns shelf?

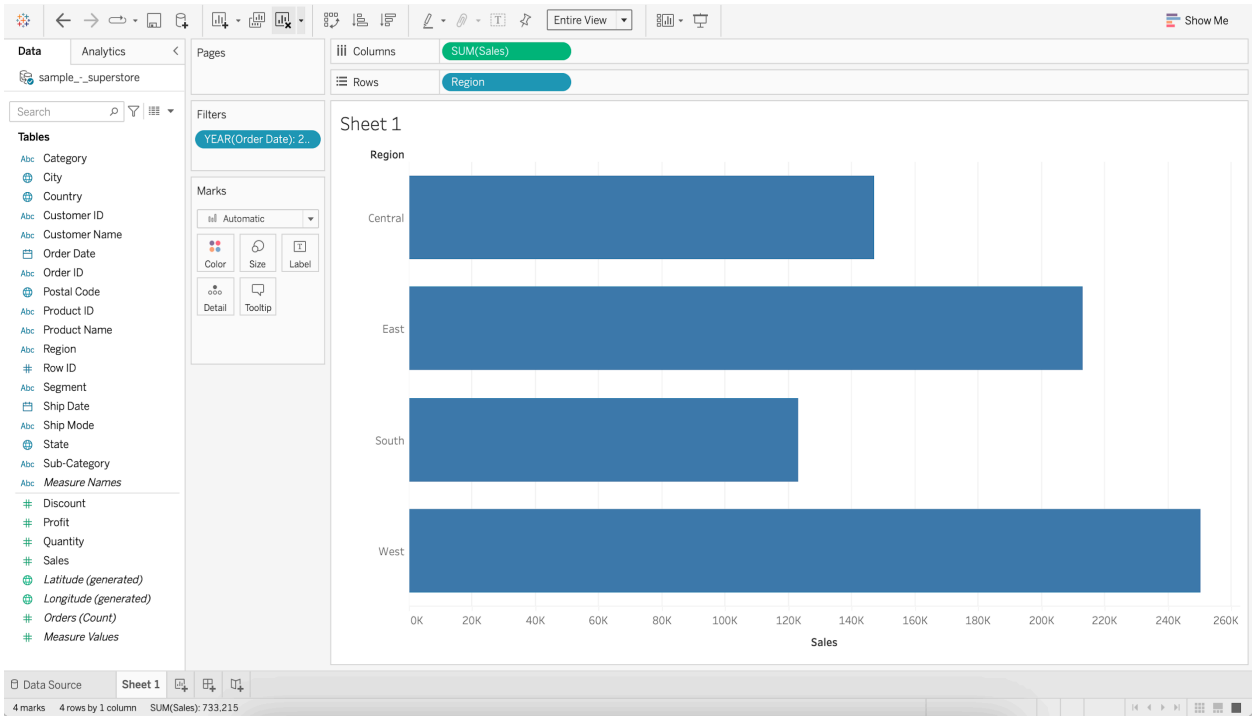
- Usually, **Sales** (Measure) would go to Rows shelf and **Region** (Dimension) would go to Columns shelf.
- But since we're creating a **Horizontal bar chart** (instead of a Vertical bar chart),
    - **Sales** (Measure) goes to **Columns** shelf.
    - **Region** (Dimension) goes to **Rows** shelf.

Steps:

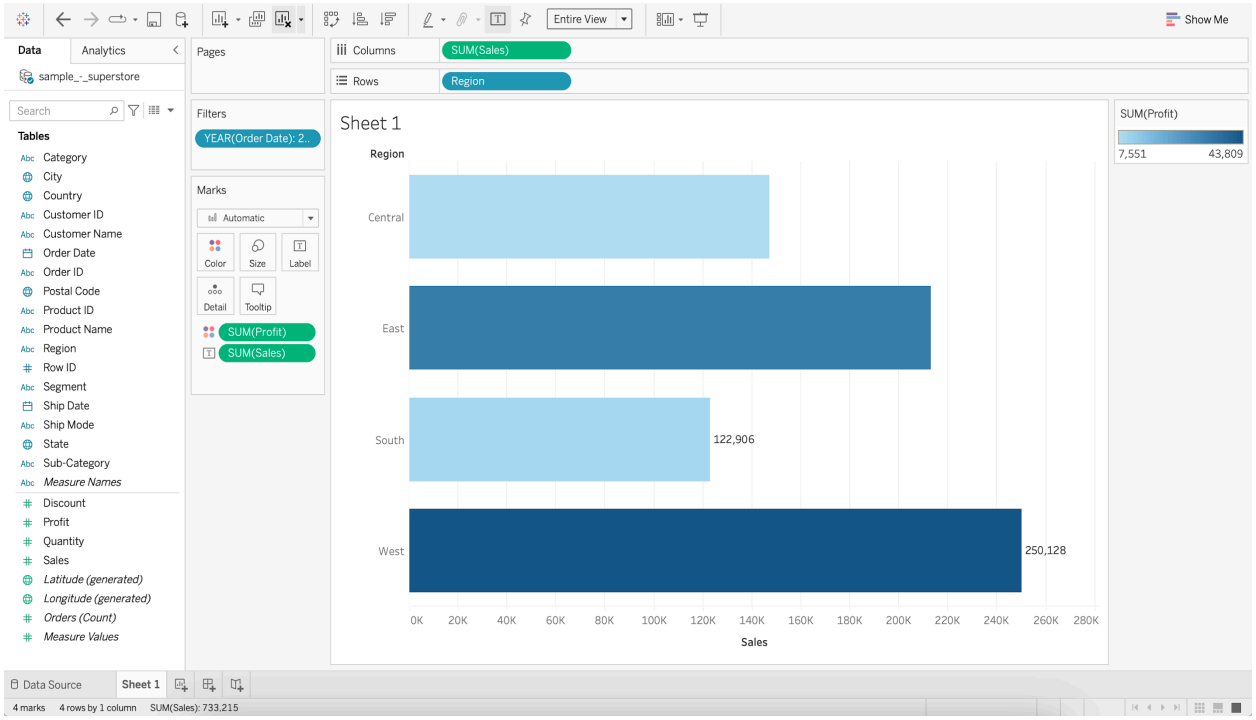
- Drag **Sales** to Columns
- Drag **Region** to Rows



- Put **Order Date** into Filters
- Select Years and press next
- Select 2017 and press OK



- Add **Profit** to Color
- Add **Sales** to Label



- Adding **Profit** (Measure) to the **Color** marks shelf assigns different shades of a single color to the **SUM of Profit** values.

- Adding **Sales** (Measure) to the **Label** marks shelf will add written information to the chart about the **SUM of Sales** values.

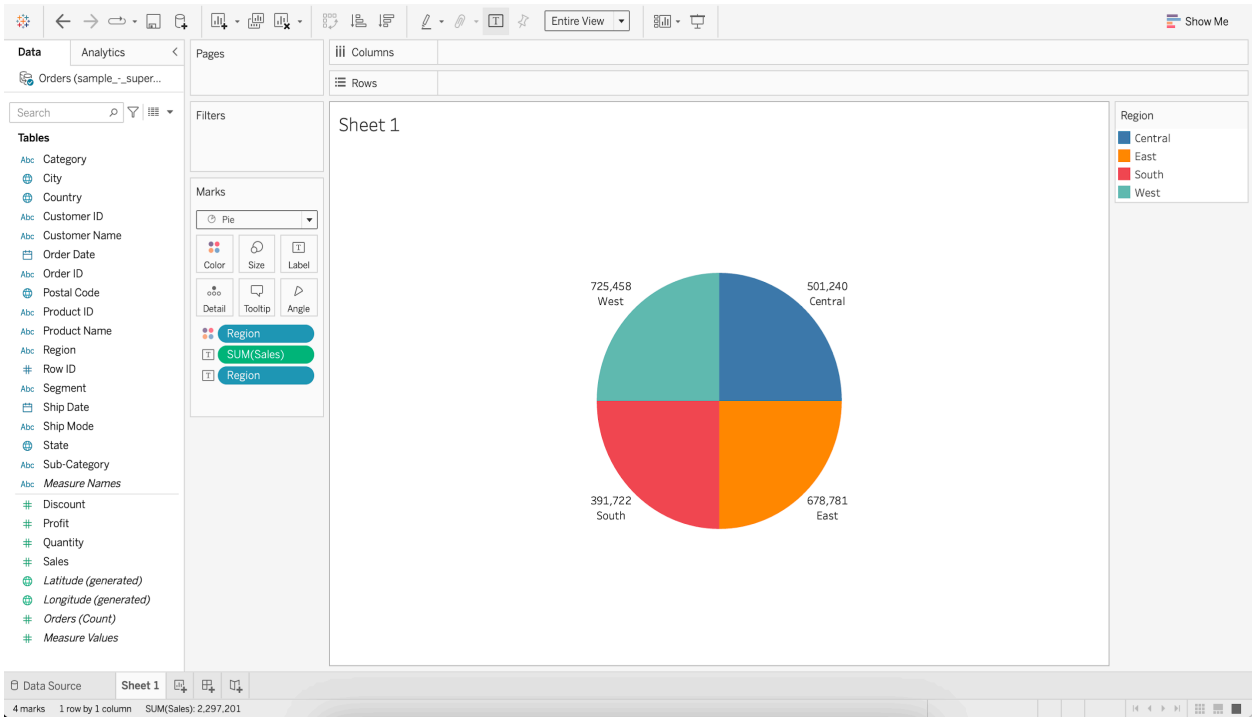
Pie chart

It represents data as a slice of a circle with different sizes and colors.

Business problem 6: Show relative percentage of sales and profit by region for the year 2017.

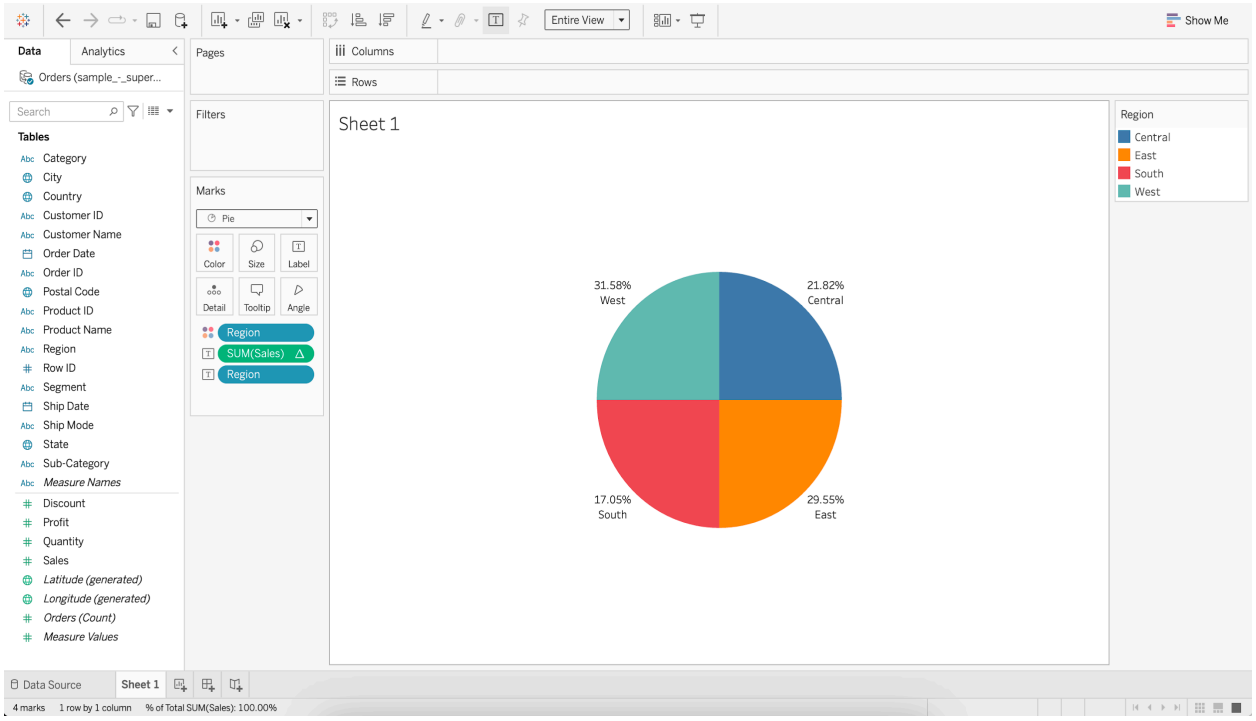
Steps:

- Drag **Sales** to the Sheet
- Change Marks from Automatic to Pie
- Add **Region** to Color
- Add **Region** to Label
- Switch to Entire View

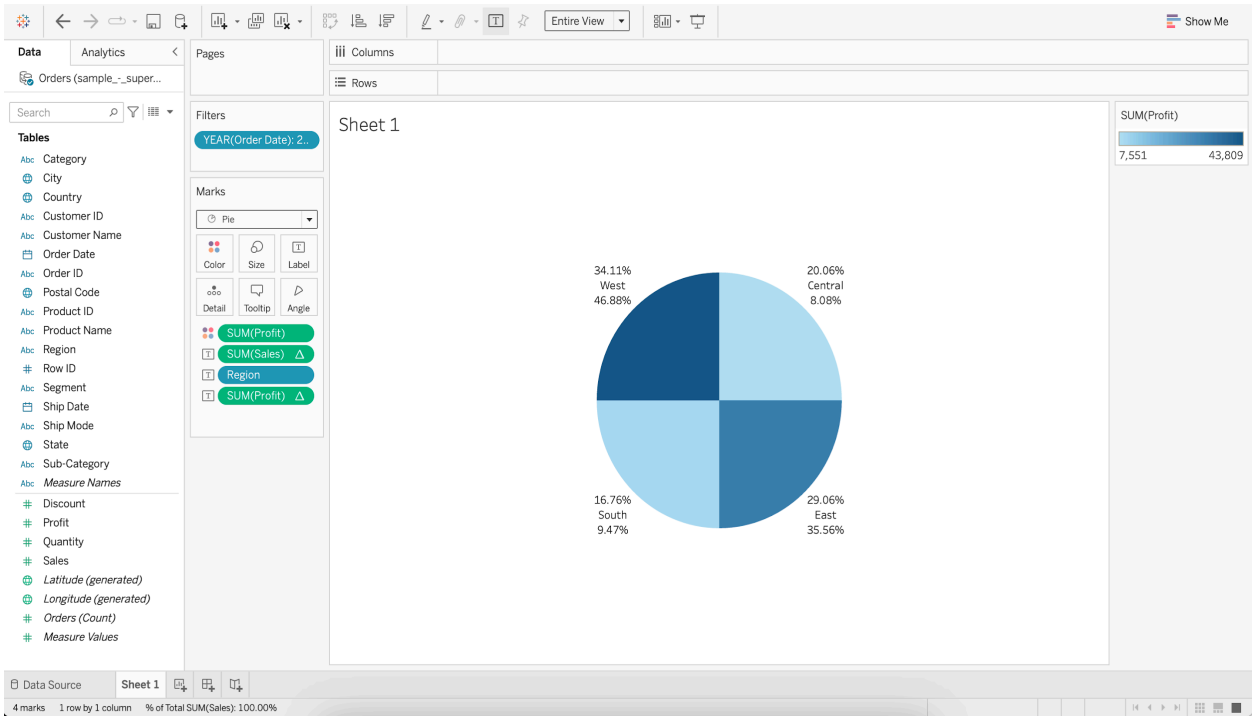


- If we drag **Region** (Dimension) to the **Color** marks shelf, we'll get a different color for each region.

- Putting **Region** (Dimension) on the **Label** marks shelf will add written information about region to the chart.
- Adding **Sales** (Measure) to the **Label** marks shelf will add written information to the chart about the **SUM of Sales** values.
- For SUM(Sales), go to Quick Table Calculation
- Select Percent of Total

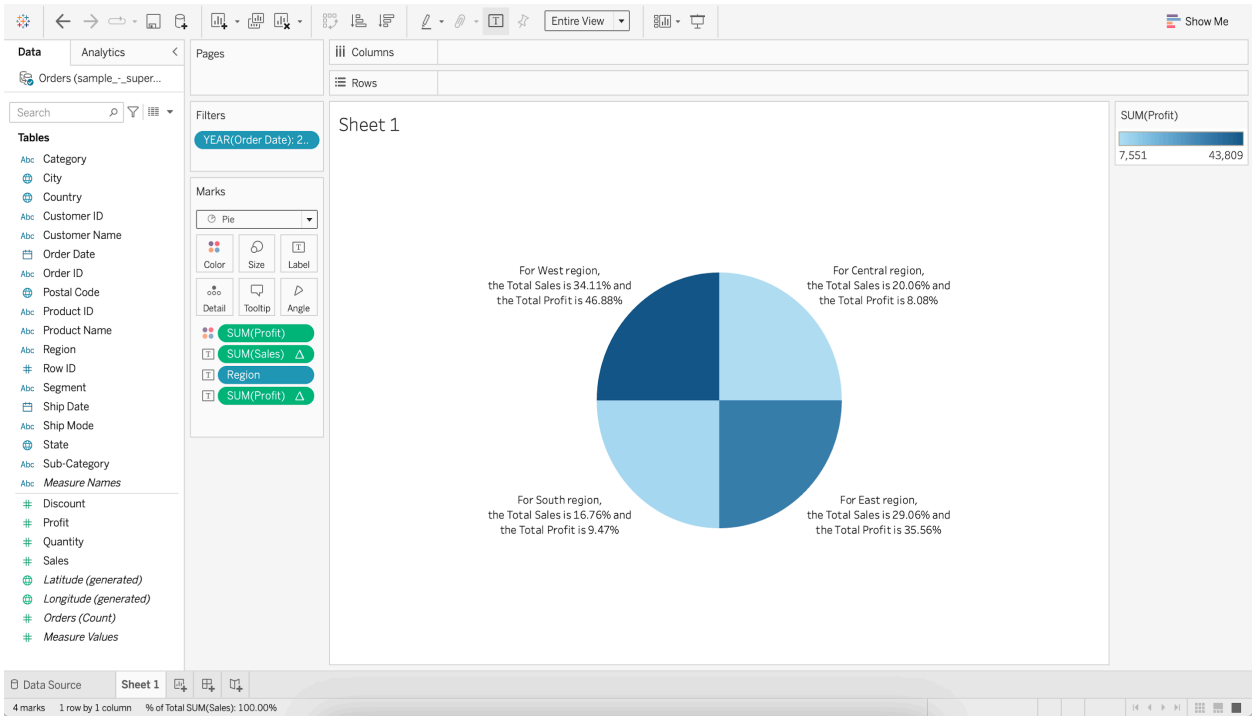


- Put **Order Date** into Filters
- Select Years and press next
- Select 2017 and press OK
- Add **Profit** to Color
- Add **Profit** to Label
- For SUM(Profit), go to Quick Table Calculation
- Select Percent of Total



Annotating the Pie chart -

- Go to Label and edit the Text
- Enter
    - For < Region > region,
    - the Total Sales is <% of Total SUM(Sales)> and
    - the Total Profit is <% of Total SUM(Profit)>.



Histogram

It is a chart that displays the shape of a distribution. A histogram looks like a bar chart but groups values for a continuous measure into ranges, or bins.

Business problem 7: Find the distribution of the number of quantities sold based on the quantity bins.

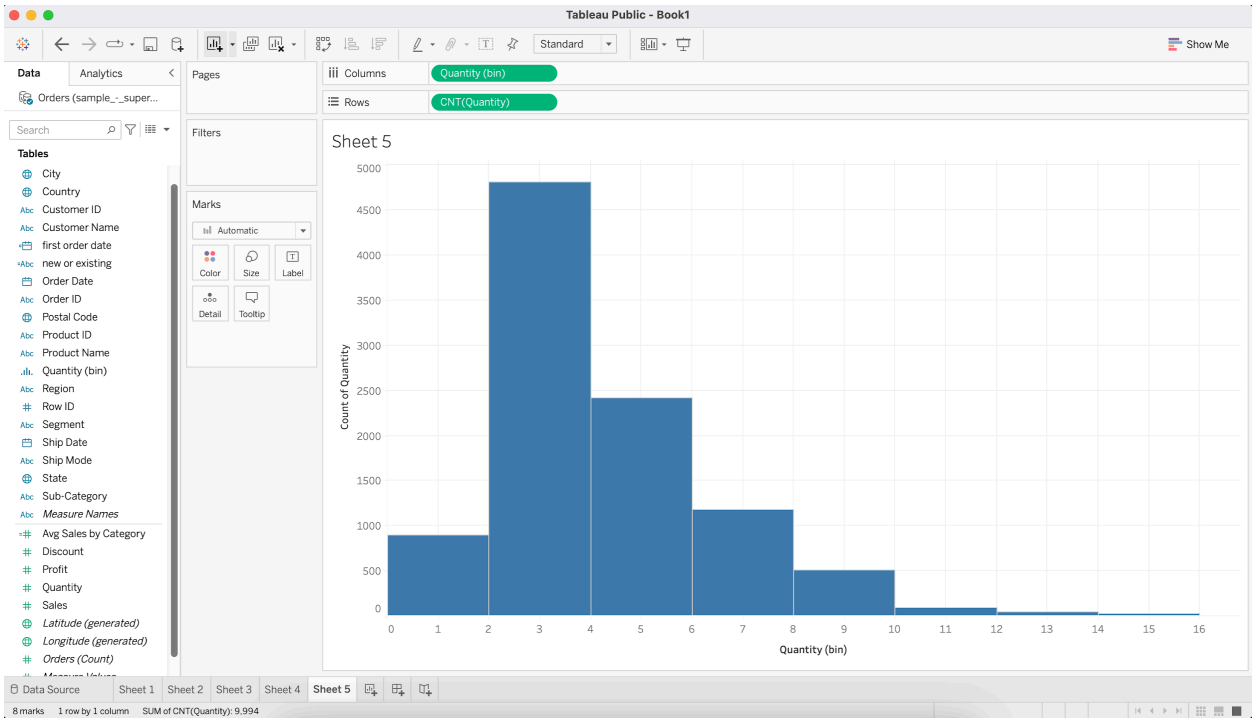
How to choose what goes to Rows shelf and what goes to Columns shelf?

- **Quantity (bin)** being a **dimension** by which we wish to group the data, goes to **Columns** shelf (X-axis).
- **Quantity** being a **measure** that we want to aggregate (CNT), goes to **Rows** shelf (Y-axis).

Steps:

- For **Quantity**, go to Create and choose Bins
- Change the Size of bins to 2

- Drag **Quantity (bin)** to Columns
- Drag **Quantity** to Rows
- Change aggregation of **Quantity** to Count
- Set **Quantity (bin)** to Continuous



Cleaning and Formatting data for analysis

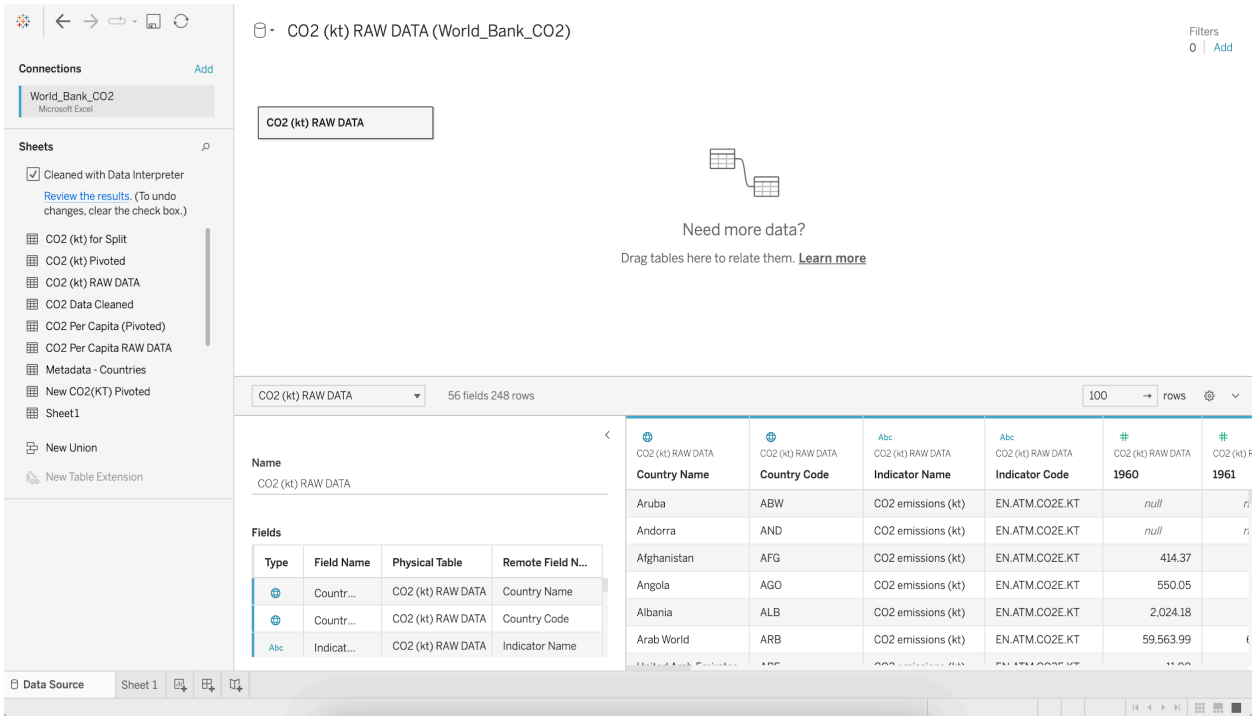
Dataset : World_Bank_CO2.xlsx [Use CO2 (kt) RAW DATA table]

Business problem 8:

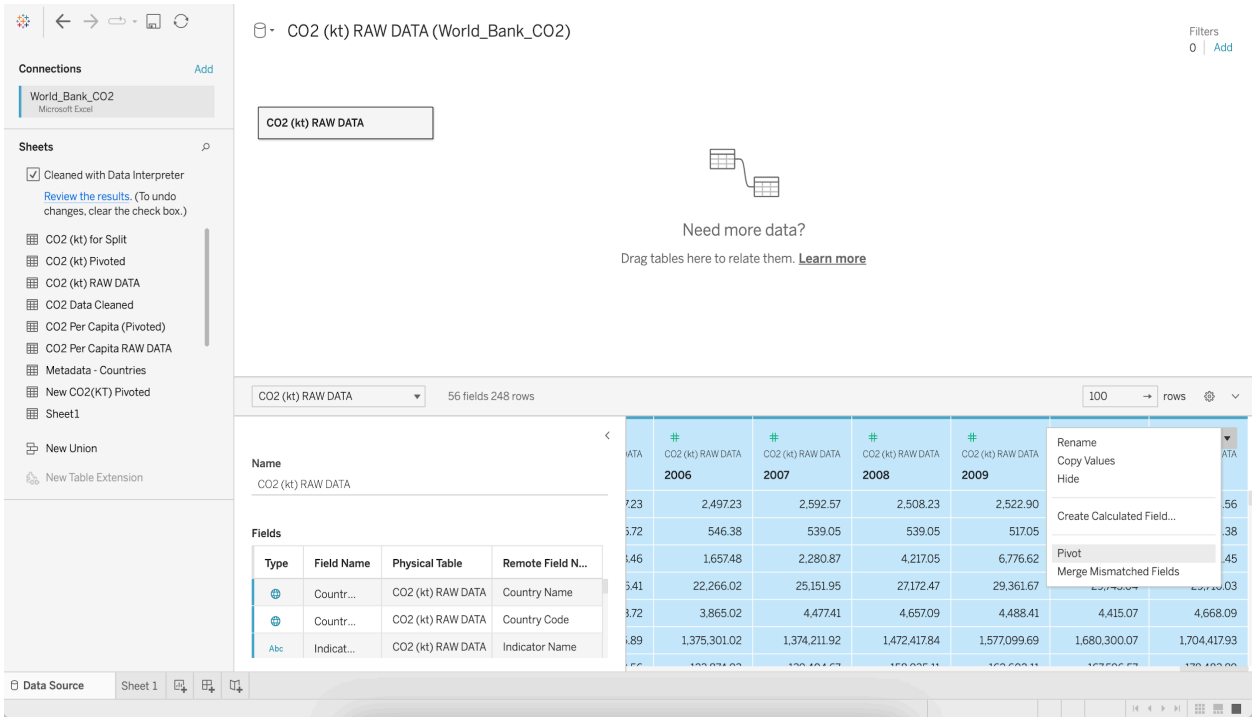
Find the CO2 emissions per country per year for those countries which have CO2 emissions above 2000 kt.

Steps:

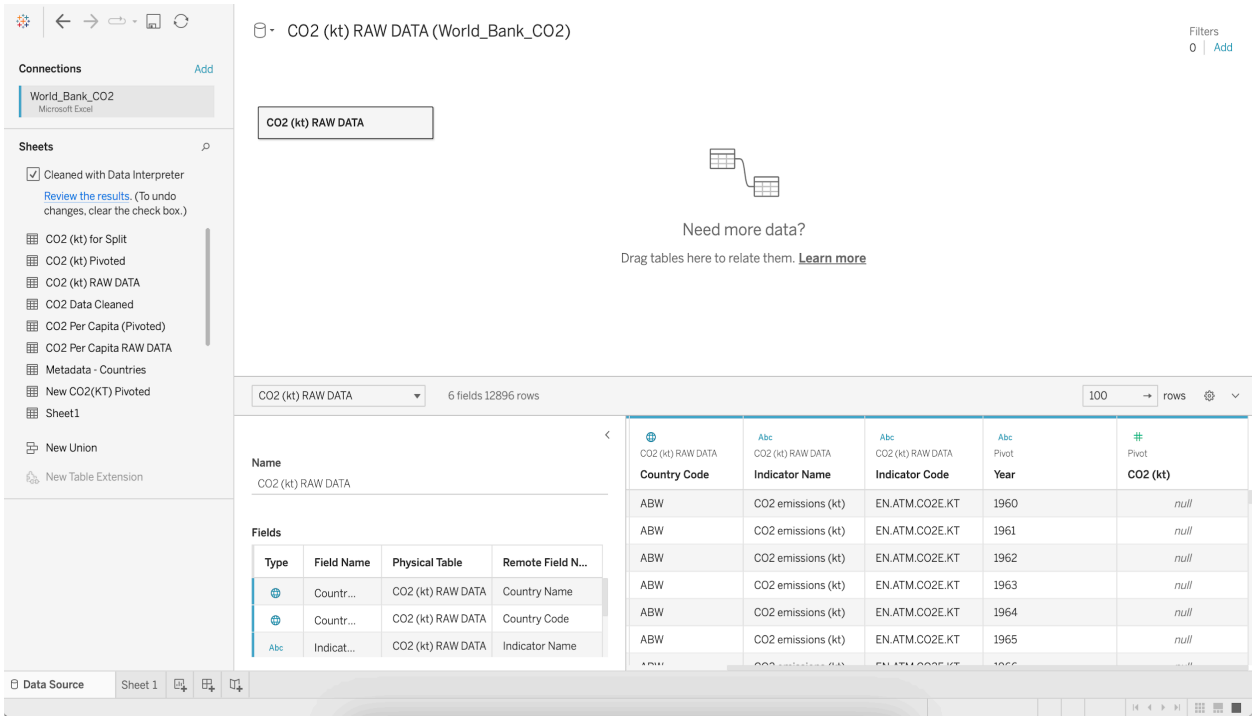
- Load the **World_Bank_CO2.xlsx** dataset
- Drag the **CO2 (kt) RAW DATA** table to the Canvas
- Check the **Use Data Interpreter** option



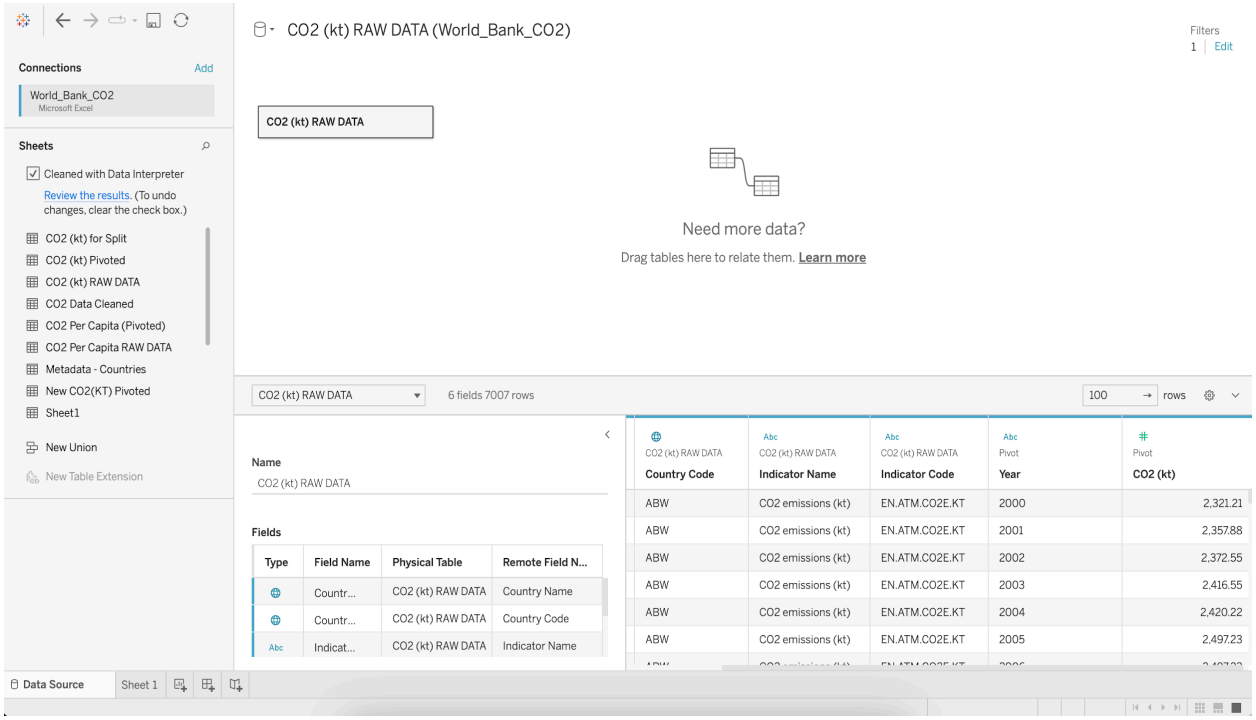
- Select the columns **1960** to **2011** and **Pivot**



- Rename the PIVOT FIELD NAMES column as **Year**
- Rename the PIVOT FIELD VALUES column as **CO2 (kt)**



- Go the **Filters** in the upper right corner and click Add
- Again click Add and Select a field as **CO2 (kt)**
- Select the minimum Range of values as **2000**



Data Interpreter

- When you track data in Excel spreadsheets, you create them with the human interface in mind.
- To make your spreadsheets easy to read, you might include things like titles, stacked headers, notes, maybe empty rows and columns to add white space, and you probably have multiple tabs of data too.
- When you want to analyze this data in Tableau, these aesthetically pleasing attributes make it very difficult for Tableau to interpret your data.

That's where a Data Interpreter can help.

What does a Data Interpreter do?

- It can give you a head start when cleaning your data by detecting things like titles, notes, footers, empty cells, etc.
- And bypassing them to identify the actual fields and values in your data set.

Note:

- Data Interpreter is only available for Microsoft Excel, Text (.csv) files, PDF files and Google Sheets.
- For Excel, your data must be in the .xls or .xlsx format.

For more information:

https://help.tableau.com/current/pro/desktop/en-us/data_interpreter.htm

Pivoting Data

- Sometimes analyzing data from a spreadsheet or crosstab format can be difficult in Tableau.
- Tableau prefers data to be **"tall"** instead of **"wide"**, which means that you often have to pivot your data from columns to rows so that Tableau can evaluate it properly.

Note: Pivoting from rows to columns is not available in Tableau Public.

For more information: https://help.tableau.com/current/prep/en-us/prep_pivot.htm

Data Source Filter

- The data source filter is used to filter the data in data source proportion.
- It restricts the files present in the dataset.
- A data source filter works on both Live connection and Extract connection.

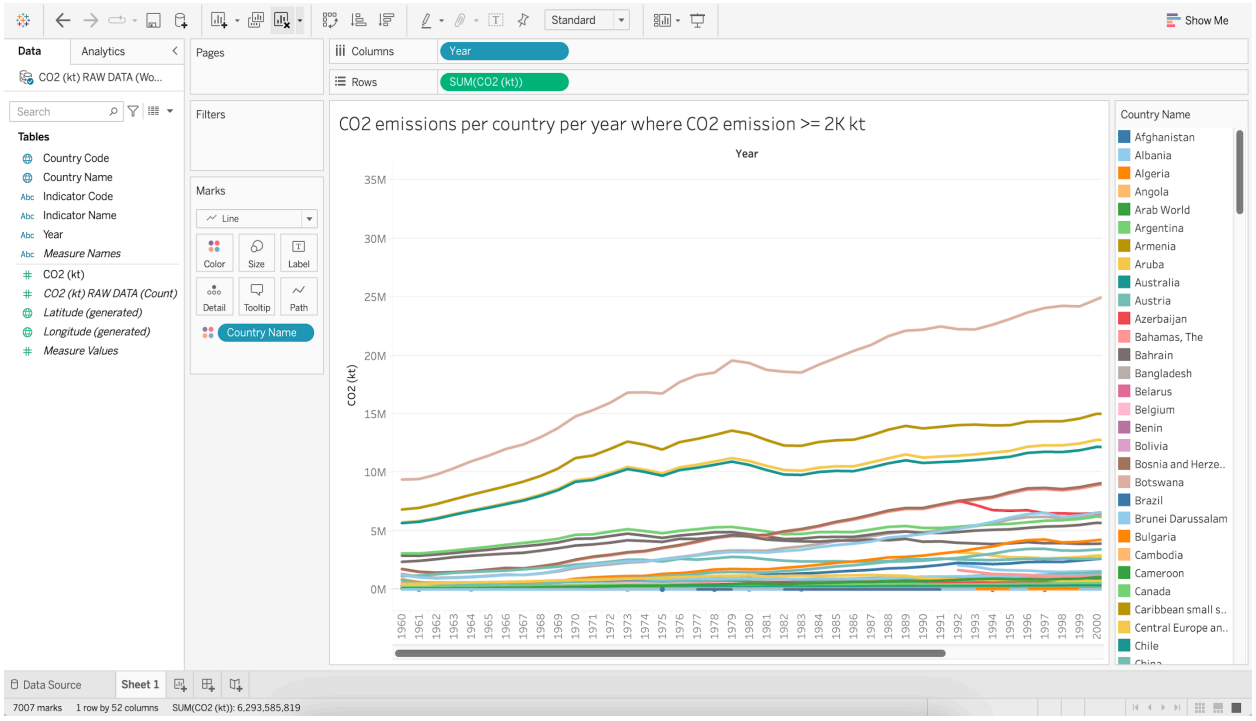
We can use the **Filter shelf** to filter data in the view.

Business problem 9:

Find the CO2 emissions per country per year where CO2 emission >= 2K kt.

Steps:

- Drag **Year** to Columns
- Drag **CO2 (kt)** to Rows
- Change Marks from Automatic to Line
- Add **Country Name** to Color
- Edit Title as "CO2 emissions per country per year where CO2 emission >= 2K kt"



- **Year** being a **dimension** by which we wish to group the data, goes to **Columns** shelf (X-axis).

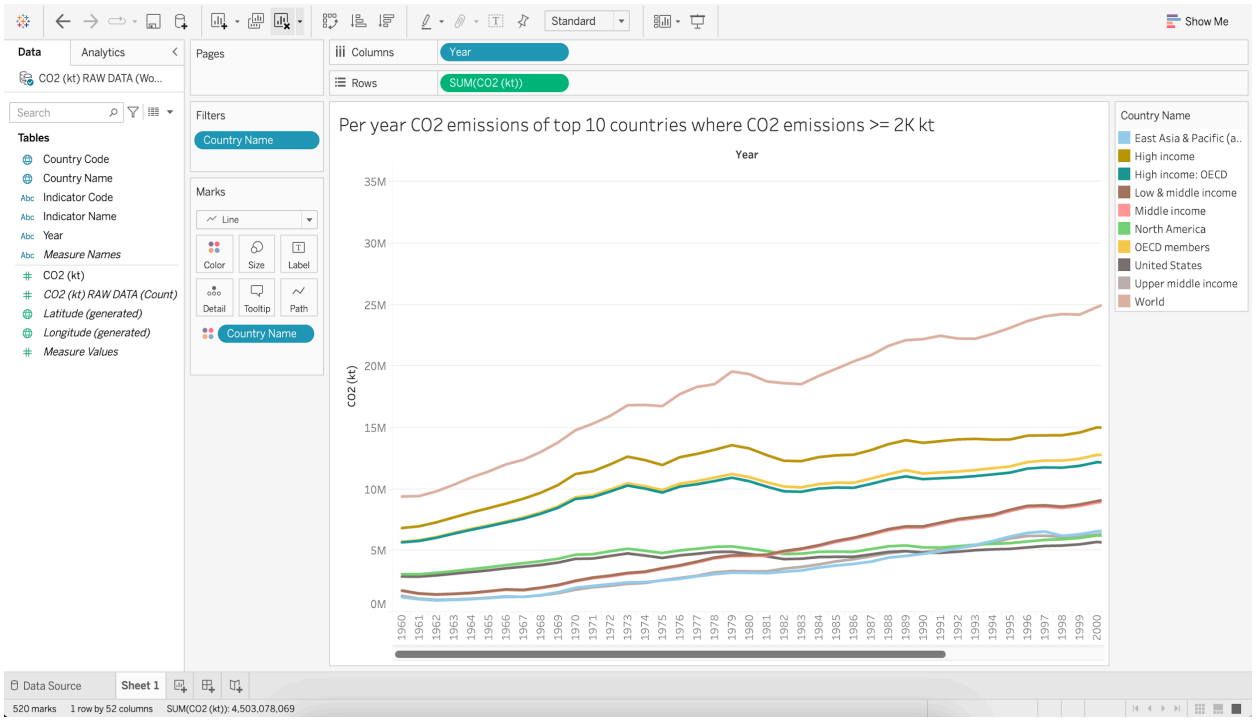
- **CO2 (kt)** being a **measure** that we want to aggregate (SUM), goes to **Rows** shelf (Y-axis).
- **Country Name** being a **dimension**, if put into the **Color** marks shelf, we'll get a different color for each country.

Business problem 10:

Find the per year CO2 emissions of top 10 countries where CO2 emissions >= 2K kt.

Steps:

- Put **Country Name** into Filters
- Go to Top and Select **Top 10** Fields by **Sum** of **CO2 (kt)**
- Edit Title as "Per year CO2 emissions of top 10 countries where CO2 emissions >= 2K kt"



Creating & Uploading Dashboard

- Re-iteration of the point - create a dashboard and upload it online.

- We need to show a basic dashboard based on the worksheets created above (Business problem 1 to 7).
- We will also show the steps of logging into the Tableau Public online account and publishing the dashboard.