

finger↑tips

Data Intelligence Solutions

Refreshing Material Part-1

Tableau

What is Data Visualization?

Data Visualization is the process of representing data and information in graphical form. By transforming the written data into charts, and graphics we would be able to identify the trends and patterns better. The goal of data visualization is not to convert data into an image. We tend to understand these visual graphics better and easily, that is why the quarters and percentages are represented as pie charts.

Examples of Data Visualization tools:

- Tableau
- Power Tool
- Infogram

Introduction to Tableau

Tableau is a strong growing data visualization tool. It is a BI tool that helps to interpret the raw data by converting it into a proper visual manner; it may be in the form of a graph, report, chart, pie chart, etc.

The software doesn't require any high-level technical knowledge or programming skills to be operated. The software is very easy for creating visual graphics. The results created can be understood by professionals working at any level.

Tableau Desktop

Tableau Desktop is a data visualization application that will help you to transform any sort of data into graphics within a few minutes. After installation, you can retrieve data from any spreadsheets and present that information in different graphical forms.

Tableau Public

Tableau Public is a free program to facilitate anyone to connect to a spreadsheet or file and create immersive data visualizations for the internet.

With Tableau Public, users can create awesome immersive graphics without the help of programmers or Technical individuals, and publish them easily.

Difference between Dimension and Measure

Dimension	Measure
Dimension is an independent variable	Measure is a dependent variable
The user cannot aggregate dimension	Users can aggregate measure
Dimension is used to compare the data	Measure is numerical value that is used to compare the dimension
Dimensions contain qualitative and categorical data	Measure contains quantitative information

Learn by Doing

Difference between Discrete field and continuous field

Discrete Field	Continuous Field
This field can filter individual data elements	This field can filter through range
Discrete field becomes header in a view	Continuous field becomes axis in view
Brings detail to the view	Brings aggregate to view
Discrete field can have hierarchy	Continuous field cannot have hierarchy

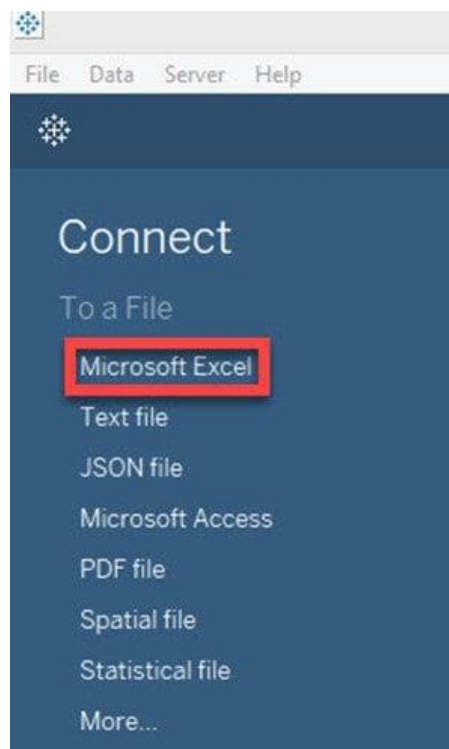
Tableau connection with various data sources

The Tableau tool can be connected to various data sources. It can easily connect to text, excel, PDF files. The tool can connect with servers and web connectors.

Connection to Excel File

Connecting Tableau to Excel is an easy process. You can follow the steps mentioned below:

1. Open the Tableau
2. In the top right corner of the data pane, click on connect the data.
3. Connect window will automatically be opened
4. Select Excel



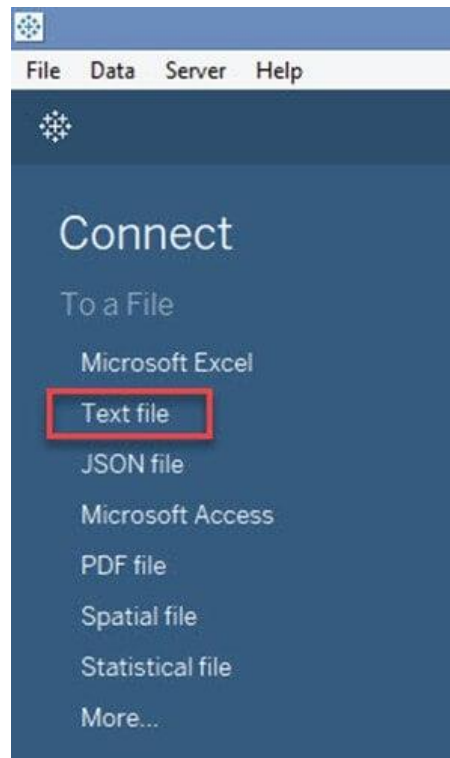
3. Navigate towards the Excel file in your saved location
4. You'll be able to see the Excel file under the connections sections

Even if Tableau detects that your excel data can be optimized then it'll recommend you to use Data Interpreter. This function can help you in cleaning; formatting excels data to ease up your analysis.

Connection to TXT file

Tableau can connect with various text files like .txt, .csv, .tab, .tsv. Follow the step to connect to Text file.

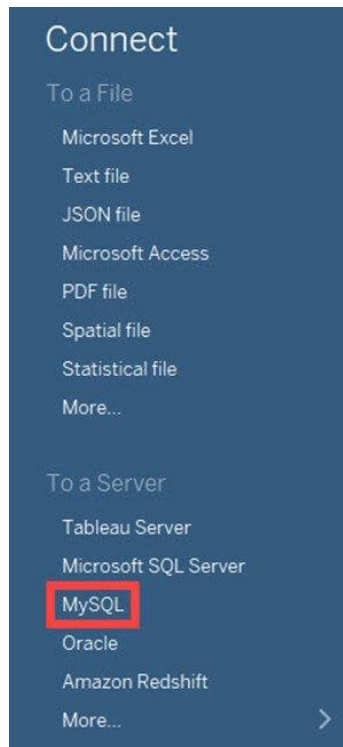
1. Open Tableau, under connect, and click on a Text file.



- 2) Select the file you want to open and click open.
- 3) After the file is opened, click on the sheet tab to start the analysis process.

Connection to Database

1. Under the Data tab, click on the database connection that you wish to connect. For example, if you wish to connect to MySQL, then select the MySQL option.



2) This will open the MySQL connection window.

3) On the screen, you can enter the MySQL server name. Then you're required to enter your username and password. Then finally click on the Sign-in button to connect to the database.

This whole procedure connects the database with Tableau. The user can now select the files from the database and access them in tableau.

Working with Metadata

Just after connecting to the data sources, Tableau captures the metadata details immediately of the sources like columns and their data types. Dimensions, measures, and calculated fields used in the views. Browsing of metadata is possible and changes can be done on its properties according to your specific requirements

Data types in Tableau

String Data type

A string data type is enclosed with either single inverted commas or double inverted commas. For example, "Fingertips", "How are you", etc. We can further divide String Data type into two different types, Char and Varchar.

- **Char String type**

These string types have fixed lengths. When a user enters a value greater than fixed-length then an error would be displayed.

- **Varchar String type**

Learn by Doing

As the name shows, this string type has a variable length. The user can input as many values as required without facing any restrictions.

Numeric Data type

This type of data type has both integer and floating types. The users generally prefer to use integer type over the floating type, as it is easier to round up the decimals after certain limit. It even has a specific function called Round (), which is used for rounding up float values.

Date and Time Data type

Tableau tool supports all sorts of formats for date and time. You can use the format "dd-mm-yy, or mm-dd-yy", or anything else. The data can be in any format like year, month, hour, minute, decade, etc.

Boolean Data type

Boolean Data types are either True or False. These types of Data types are formed at the time of relational calculations. At the times, when the result is unknown, the calculation shows null.

Geographic Data type

Values that are used in maps are geographic data types. The example can be valued like company name, state name, city name, etc.

Cluster or Mixed Data type

When some data set contains values that have a mixture of data types in it then such values can be known as cluster group values or mixed data values, it can be handled manually or you can allow Tableau to operate on it.

What are Joins?

Combining data from multiple places is necessary such as different tables, different data sources to get the desired analysis. Depending upon the structure of the data and the analysis done on it, there are several ways to combine the tables.

Relationships vs Joins

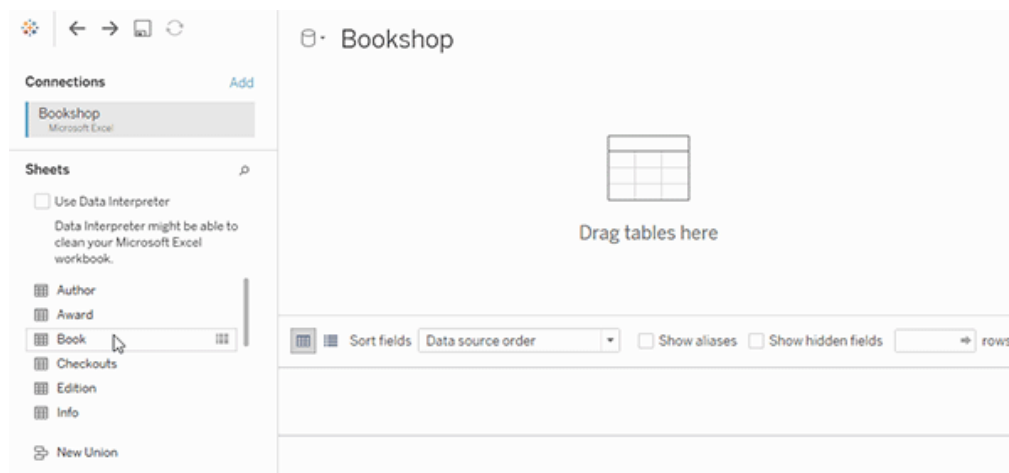
Relationships	Joins
<ul style="list-style-type: none">• They are displayed as flexible noodles between the logical tables.	<ul style="list-style-type: none">• They are displayed with Venn diagram icons between physical tables.
<ul style="list-style-type: none">• They require you to select matching fields between two logical tables.	<ul style="list-style-type: none">• They require you to select join types and join clauses
<ul style="list-style-type: none">• It does not require you to select the join types.	<ul style="list-style-type: none">• In Joins physical tables are merged into a single logical table with a fixed combination of the data.

Learn by Doing

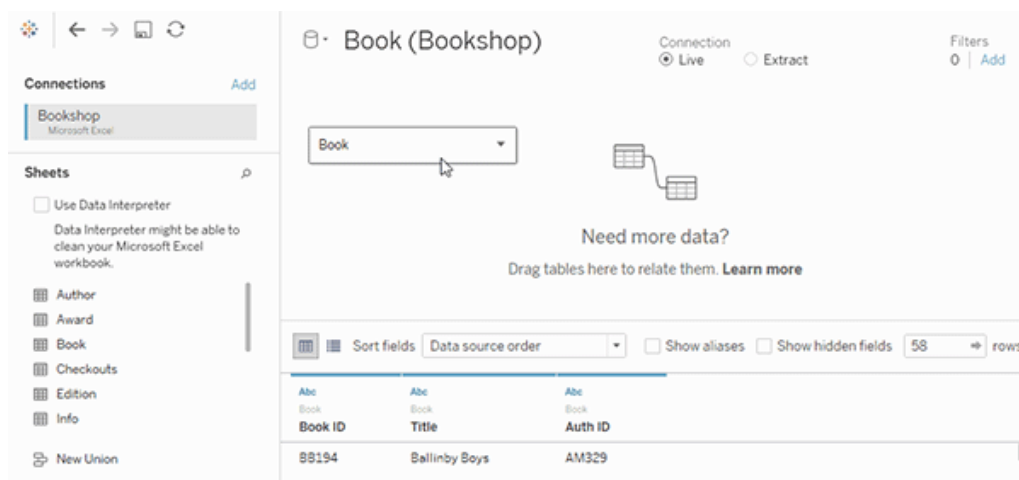
<ul style="list-style-type: none"> It makes all the row and column data from the related tables to get available in the data source. 	<ul style="list-style-type: none"> It May drop unmatched measure values.
<ul style="list-style-type: none"> Maintains the level of the table in the data source and during analysis. 	<ul style="list-style-type: none"> When fields are at different levels of detail, it may duplicate aggregate values.
<ul style="list-style-type: none"> It creates independent domains at multiple levels of detail. Here tables do not get merged together in the data source. 	<ul style="list-style-type: none"> It supports scenarios that require a single table of data, such as extract filters and aggregation.

Creating a Join:

- When you need to create a join, you need to connect all the relevant sources.
- You need to drag the first table to the canvas.

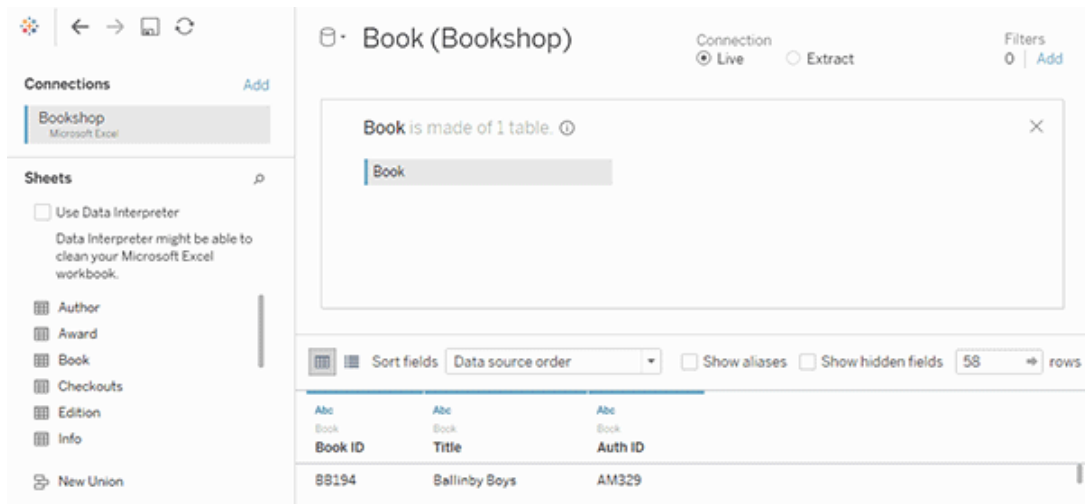


- Now select Open from the menu or double click to the first table to open the join canvas.



- Drag or Double-click to another table to join canvas.

Learn by Doing

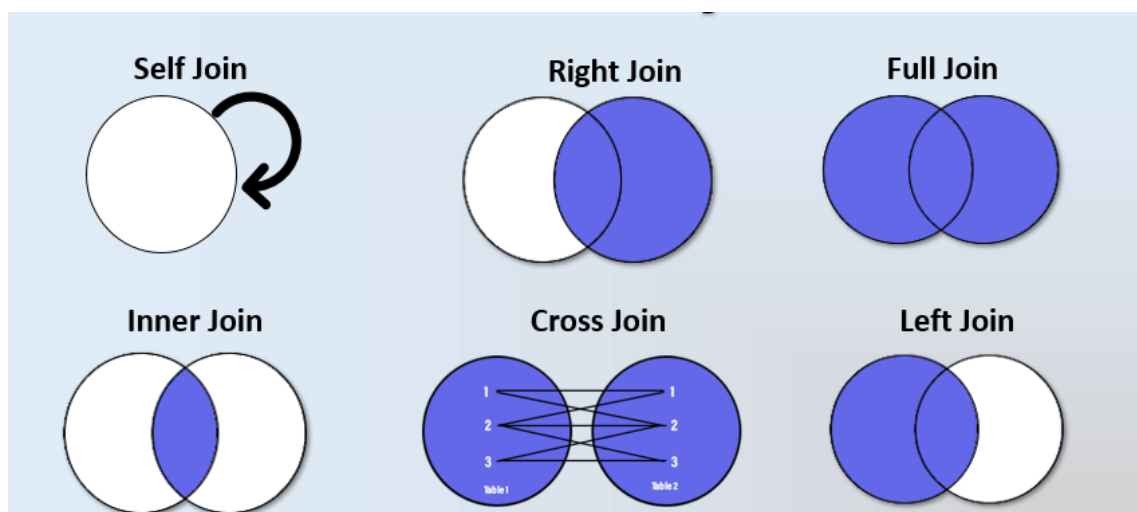


- Directly click the join icon to configure the join.

Types of Join

In Tableau, there are four types of joins named as:

- Inner Join
- Left Join
- Right Join
- Full outer



Union Join

This is another method of joining two or more tables by just appending the rows of the data from one table to another. The tables in which you have used union tables will consist of the same number of fields, and these fields will have matching names and data types.

What are Blends?

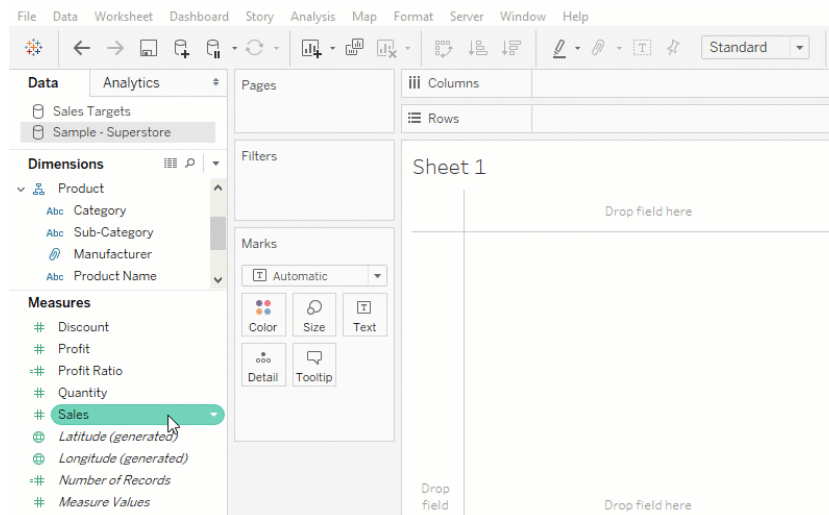
Blending the data is a very powerful and useful tool in Tableau. The user would like to analyze the data that is related to multiple data sources; via Blends, you can analyze that data

Learn by Doing

together in a single view. Blending doesn't combine the data, on the other hand, it queries each data source separately, and finally, the results are displayed together.

Steps for Blending Data

1. Ensure that the Tableau workbook has multiple data.
2. Now, drag a field to view. Remember, from whichever data source the first field is generated that will become the primary data source.



3. Now, go to other data sources and ensure that there is a blend relation with a primary data source.
4. Drag a field from a secondary data source.

Understand Primary and Secondary Data Sources

The basic requirement of blending is it must have a primary data source and at least one secondary data source. The first data source used in the view becomes the primary data source by default. This will restrict the values from the secondary data source. The value that matches with primary data sources appears in the view.

Let's take an example, suppose the primary data source has data on the field as months and has values of January, February and March. Now the view will show Jan, Feb, and March data along with data of secondary data sources. Even, if the secondary data source has values of all the months but it won't be displayed.

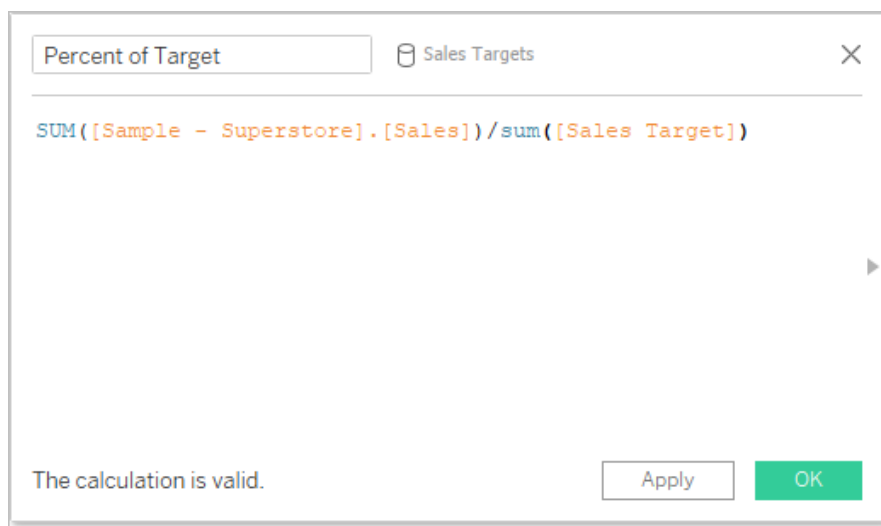
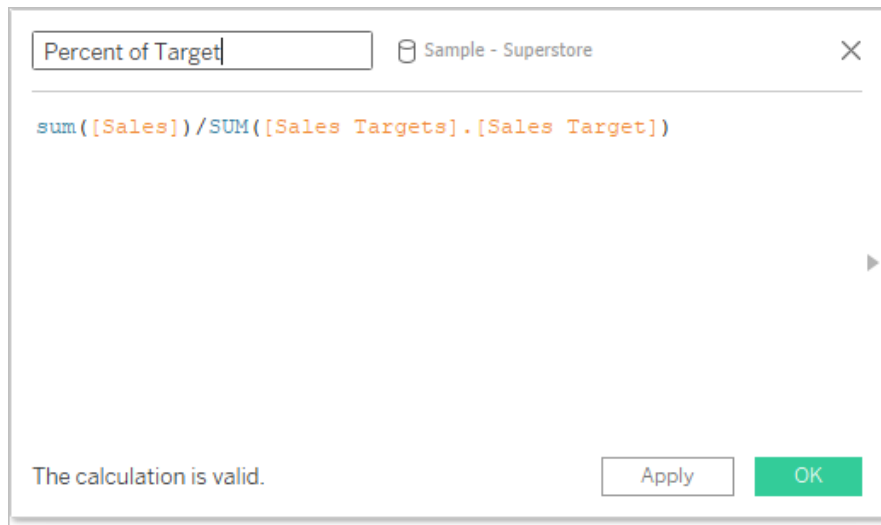
Work across blended data sources

There are certain things that the user should keep in mind while he/she is working across blended data sources. For example, performing calculations with more than one data source can be slightly different and difficult from general calculations.

- **Aggregation**
Any type of field that comes from a different Data source will come with default aggregation "SUM". This default setting can be changed.
- **Dot Notation**

Learn by Doing

While calculation, the field that is from another data source will be referred to as data source using dot notation.



Apart from the calculations part, there are certain limitations in working across blended data sources. The user might not be able to sort the data by field from a secondary data source. Also, the action filter might not work properly with blended data.

Define blend relationships for blending

Tableau requires some common dimensions between different data sources to combine the data. This common dimension required by Tableau is known as the linking field. Active and potential links are identified differently. Like, the active linking field is identified with a link icon and the potential linking field is identified with a broken link icon.

Differences between Joins and Blending

Data Joining	Data Blending
--------------	---------------

Learn by Doing

Data joining is used when the data set is from the same source.	Data blending is used when the data set is from different source.
It can use different types of Joins.	It can use only left join.
Duplication and loss of data is possible.	Duplication and loss of data is not possible.
Data Joining cannot use published sources.	Data joining can use published sources.
Joins data at a row-level	Sends separate query to each dataset, aggregates and then performs blending

Differences between Relationships and Blending

Data Relationship	Data Blending
Data relationship is the data source.	Data blending is the worksheet.
It's not possible to use calculated field as key.	Blends can use calculated field as key.
The context depends on which dimensions are used in the table.	Blends always have a context.
Relationship has a logical layer.	Blend has a logical layer.
Join type is flexible.	Can only use left join.

Marks Card

It is the key element for visual analysis in Tableau, when you drag fields to different properties in the marks card, you can add context and also details to the marks in the view.

Highlighting

It will allow you to call attention to marks of interest by just colouring specific marks and dimming others. Using a variety of tools you can highlight marks.

For example, you can just manually select the marks you want to highlight and can use the legend to select related marks, highlighter to search for marks in context, or create an advanced highlight action.

Sorting

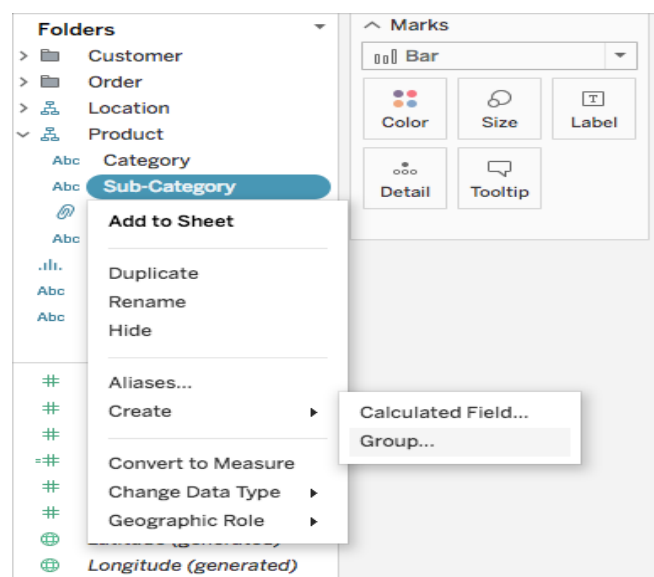
Sorting is an important feature of data analysis. Sorting data of the fields is known as dimensions. While viewing visualization, data can be sorted using single-click options from an axis, header, or field label. In the authoring environment, additional sorting options include sorting manually in the headers and legends, using the toolbar sort icons, or sorting from the sort menu.

Grouping

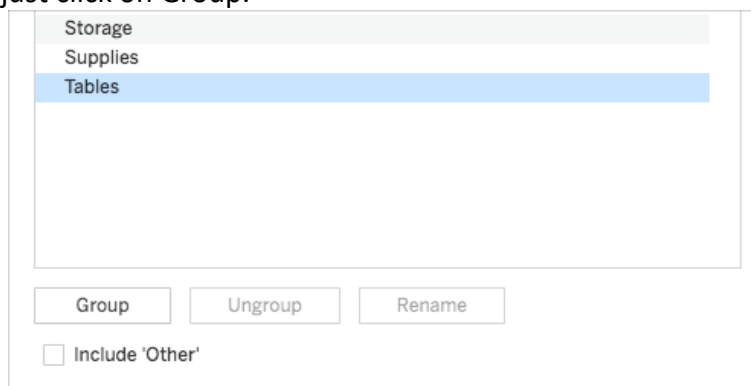
When you group different items that are within a dimension then it can be very useful if you need to color up certain items, sum groups of items together or even follow different item groups. A group can be created to combine the related members in a field.

Creating a Group

Groups can be created in multiple ways, It can be created from a field in the Data pane, or just by selecting the data in the view and then clicking the group icon. Just right-click a field and select Create > Group in the data pane.



When you are in the Create Group dialog box, select the several members you want to group, and then just click on Group.



Sets

Learn by Doing

It can be used to compare and ask questions about a subset of data. They are the custom fields that define a subset of data based on some conditions.

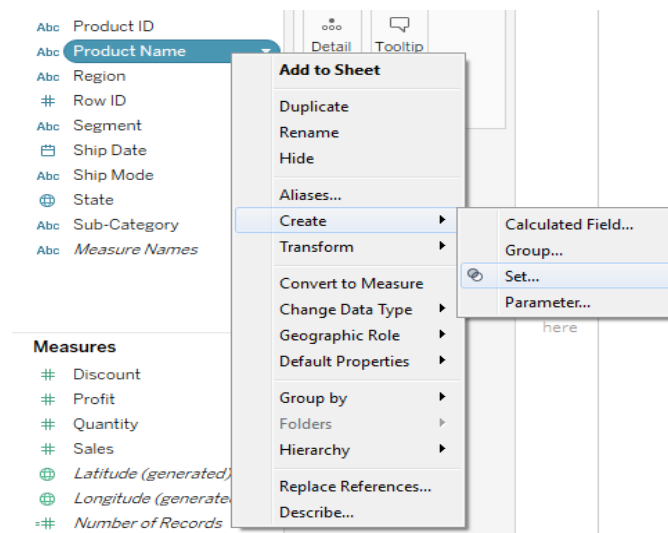
Constant Sets

They are the sets that cannot be changed. If the underlying data changes, the membership of the constant set does not change to reflect these differences, as they are also known as manually created sets.

Computed Sets

In Tableau, there are loose collections of mini-series that are designed to give you an in-depth look into various features of the Tableau software.

They use logic to dynamically update the membership of the set. As this is the key distinction between the constant sets and the computed sets and changes to the data will change the set itself as it re-computes what gets classified as IN the set and as OUT the set.



Bins

They are the containers of equal size that stores data values corresponding to the bin size and also fitting to the bin size.

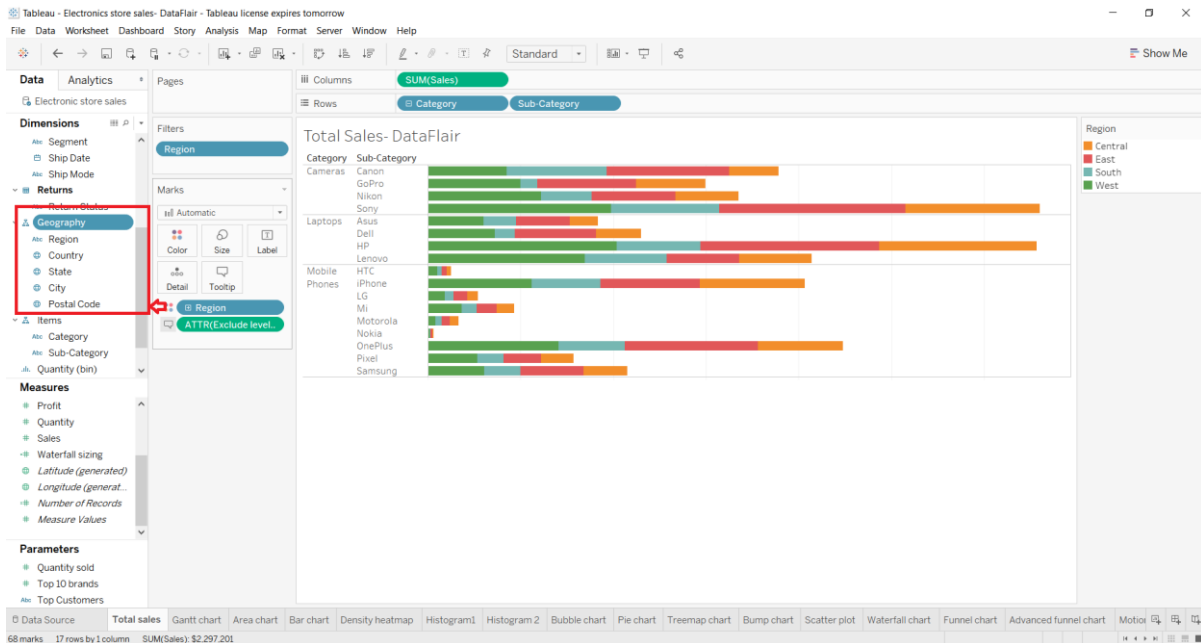
Bins group a set of data into the groups of equal intervals or size making it a systematic distribution of the data. In Tableau, data from any discrete field can be taken so that bins can be created.

Hierarchies

When the data source is connected, Tableau will automatically separate the data fields into hierarchies so that you can easily break them down.

For Example, Suppose if you have a set of fields named Region, State, and Country, you can directly create a hierarchy by using these fields so that you can quickly drill down the between levels.

Learn by Doing



Filters and its type

Filters are a great way to manage large data sets. The tool can remove irrelevant data and reduce the size of data for faster processing and analysis of data.

There are five different types of Filters in Tableau that are mentioned below:

1. Extract Filter

This type of data is used to extract data from different sources. Using an extract filter reduces the tableau queries in the data.

2. Data Source Filter

This filter is used to restrict certain sensitive data from viewers. But, viewers have certain rights of access to view the data. One important thing to remember is that the data source filter and extract filter are not at all linked together.

3. Context Filter

This filter helps to create data sets by implying relevant presets for compilation. The context filter adds actionable context to the data analysis process.

4. Dimension Filter

These filters are applied to the dimension field. It includes filtering based on a certain category of text or numerical data.

5. Measure Filter

This filter is applied on the Measure field. A measure field has quantitative data and thus filtering is based on the calculation part.

6. Visual Filter

While converting the data into visual form, you have to think not only as an analyst but also as a designer and as an end reader of data.

7. Interactive Filter

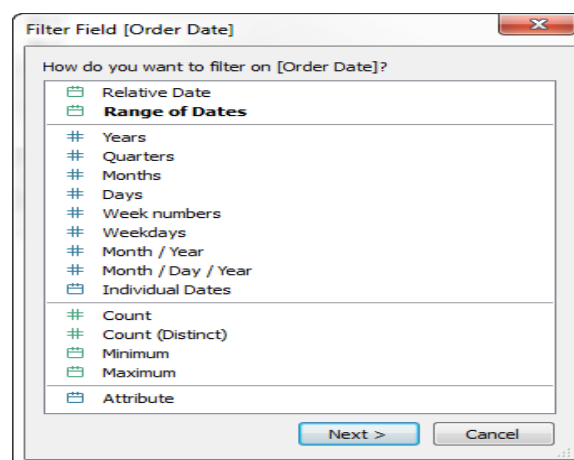
Learn by Doing

You can add context and make your data interactive using certain actions. Users are going to use your data and interact with it. You need to make sure that your data is worth for the users to interact with. Below mentioned are different types of actions:

- Highlight
- Filter
- URL
- Parameter
- Set values

8. Date filter

Under this Filter, the user can select to filter on a relative date, filter between a range of date, discrete date, or individual date. The moment you drag a date filter, this type of box will be discovered:



Charts in Tableau

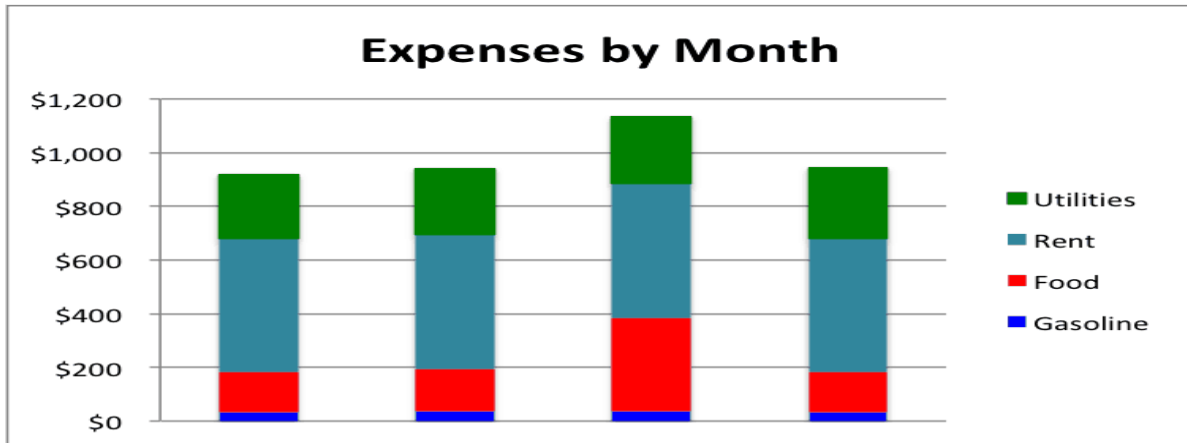
Let's see different forms of charts that we can make in Tableau.

Bar Charts

Bar Charts are simple and can be used your data can be divided into various categories. With the help of bar charts, you can identify trends, compare high and low values, and compare historical data in just one glance.

Stacked Bar Chart

A stacked Bar Chart is a simple bar chart with further segmented bars. The bars in the chart are then categorized further. The bars are internally divided to provide more advanced details.



Line Charts

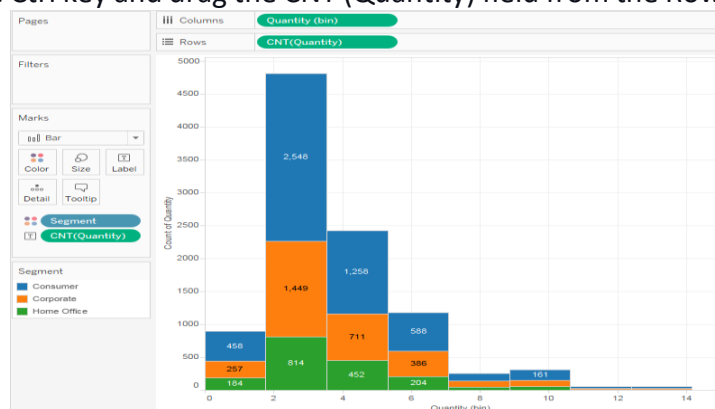
Line charts show the information via data points that are connected by line segments. The result drawn from this chart is easy to understand and visualize.

Histogram

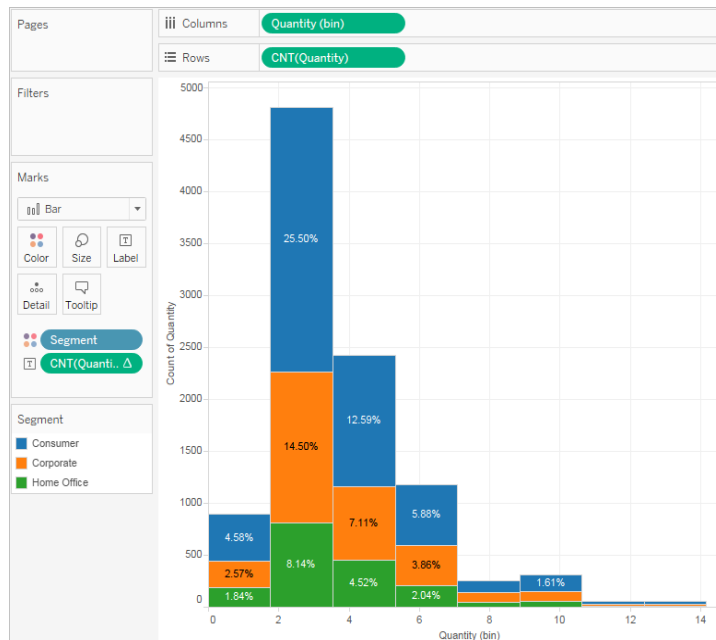
A histogram represents the frequencies of values of a variable bucketed into ranges. The histogram is similar to a bar chart but it groups the values into continuous ranges. Each bar in the histogram represents the height of the number of values present in that range.

A step-by-step process of creating Histogram

1. Under Toolbar, click on show me and then click on histogram chart.
2. Now, drag a segment to color
3. Hold down the Ctrl key and drag the CNT (Quantity) field from the Rows shelf to Label.



4. Right-click the CNT (Quantity) field on the Marks card and select Quick Table Calculation > Percent of Total.



5. In the Table Calculation dialog box, change the value of the Compute Using field to Cell.

The Table Calculation dialog box is shown with the following settings:

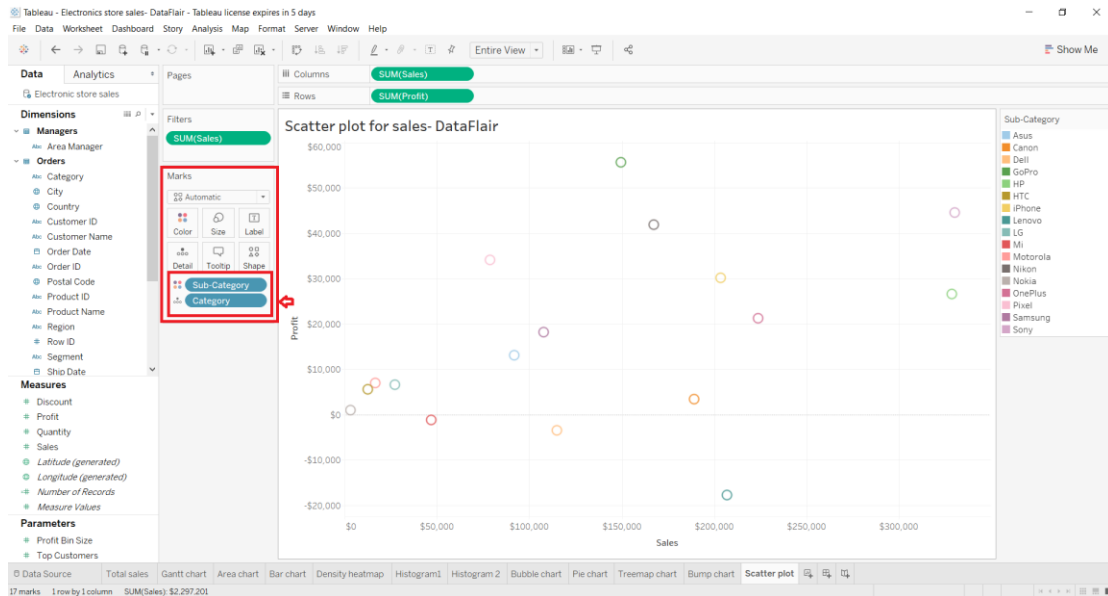
- Calculation Type:** Percent of Total
- At the level:** (empty)
- Compute total across all pages:** ☐
- Compute Using:** Cell (selected)
- Segment:** ☒
- Quantity (bin):** ☐

Scatter plot

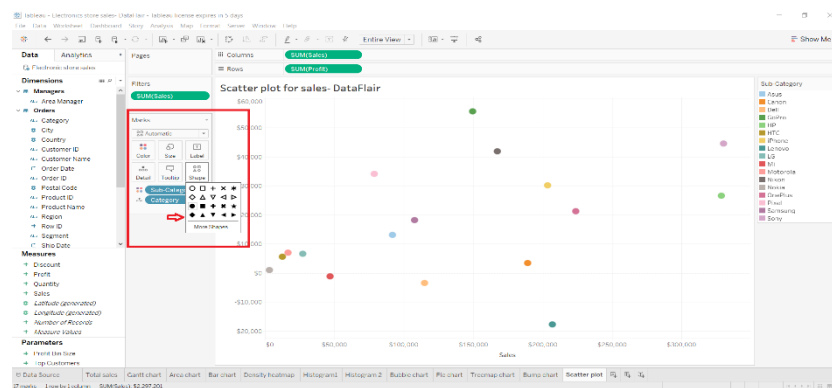
Scatter plot also known as scatter chart or scatter graph, uses dots to represent value for two different variables. The place of each dot depicts value on the horizontal and vertical axis. This type of visualization can be used to study the relationship between two different variables.

Process of creating a scatter Plot in Tableau:

1. Select the Measure
2. Drag Measure to the Rows Section
3. Select Two Dimension Fields

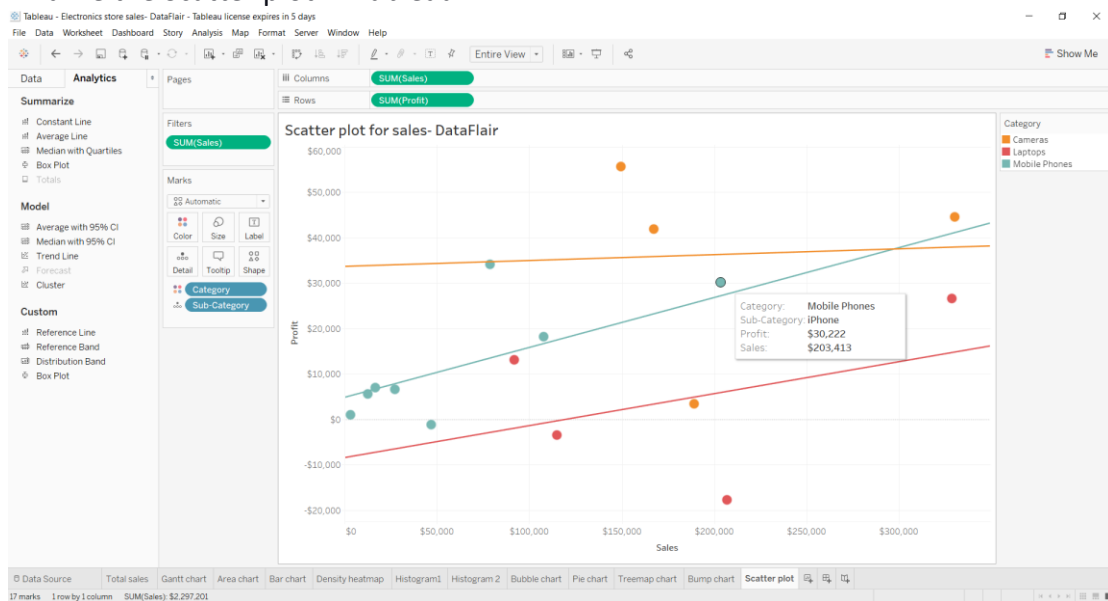


4. Customize Scatter Plot in Tableau



5. Increase Detail of Scatter Plot

6. Finalize the Scatter plot in Tableau



Learn by Doing

Dual Axis Chart

In the dual-axis, two axes are layered and displayed on top of each other. This type of chart allows you to compare multiple measures together. These types of charts are also known as combo charts.

Maps

If the user wants to analyze the data information in a geographical format, then he/she can plot the data on a map in Tableau. It is appropriate to use maps when you have a spatial question in your mind. A spatial question is something related to size, position, or area.

Tableau provides different types of options in maps:

- Proportional symbol maps
- Choropleth maps
- Point distribution maps
- Heat maps
- Flow maps
- Spider maps

Bullet chart

A bullet chart is an advanced version of a Bar chart where we can compare two different variables on a single bar only. In the bullet chart, the main primary variable is shown by the dark color bar and the second variable is displayed in the light color bar.

Bar-in-bar chart

The bar-in-bar chart is two bars on a single chart, interlocking each other. It is used when there's a requirement to study two variables simultaneously for comparison purposes.

Level of Detail

It allows you to compute the values at the data source level and the visualization level. LOD expressions are used to run queries that are complex and involve many dimensions at the data source level instead of bringing all the data to the Tableau interface.

Different Types of LOD expressions

FIXED

It computes a value using a specified dimension, which is without reference to the dimension in the view.

INCLUDE

It computes the values using the specified dimensions in addition to whatever dimensions are in the view.

INCLUDE level of detail expressions are useful when you wanted to calculate at a fine level of detail in the database and then re-aggregate and show the level of details in your view. Fields based on INCLUDE level of detail expressions will change when you add or remove dimensions from the view.

Learn by Doing

EXCLUDE

It declares dimensions to omit from the view level of details. They are useful for the 'percent of total' or 'difference from overall average' scenarios and can be comparable to such features as Totals and Reference Lines.

Expressions Syntax

Level of detail expression has this structure:

```
{[FIXED | INCLUDE | EXCLUDE] <dimensions declaration > : < aggregate expression > }
```

Aggregation and replication with LOD expressions

FIXED-Aggregated

Results

If the LOD expression's results are more granular, the values from the LOD Expression are aggregated to create the view.

- LOD of the view will be simply segmented, but LOD expression will be fixed at a more granular level of both segment and category.
- LOD Expression's values for each category are aggregated into a single value per segment which is displayed in the view as the result.
- If we wrap a level of detail expression in aggregation when we create it, Tableau will use the aggregation specified rather than choosing one when that expression is placed on a shelf.

FIXED-Replicated Results

- If the dimension declaration is more aggregated than the view, the values from the LOD Expression are replicated to create the view.
- The LOD of the view is both Category and segment, But LOD Expression will be fixed at a less granular level of just segment.

The LOD expression's values for each segment will be replicated for each category within a segment to be displayed in the view.

Dashboards

It is a collection of several views, which will let you compare a variety of data simultaneously.

Building a Dashboard

When you've created one or more sheets you will be able to combine them in a dashboard and can add interactivity, and much more.

You can create a dashboard in the same way you create a new worksheet. On the left from the Sheets, drag views to your dashboard at the right.

Adding interactivity

Interactivity can be added to dashboards to enhance the data insights.

Learn by Doing

Adding dashboard objects and setting their options

You can add dashboard objects, in addition to the sheets that will add visual appeal and interactivity.

- Horizontal and Vertical objects provide layout containers that let you group related objects together and then fine-tune how your dashboard resizes when users interact with them.
- Text objects will provide headers, explanations, and other information.
- Image objects add to the visual flavour of a dashboard, you can link them to a specific target URL.
- Web Page objects display target pages in the context of your dashboard.
- Blank objects will help you adjust the spacing between dashboard items.
- Navigation objects will let your audience navigate from one dashboard to another, or maybe to other sheets or stories. Text or image can be displayed to indicate the button's destination to your users, specific custom.
- Download objects will let your audience quickly create a PDF file, PowerPoint Slide, or PNG image of an entire dashboard, or a crosstab of selected sheets. Formatting options are similar to Navigation objects. Crosstab download is possible only after publishing to Tableau Online or Tableau Server.
- Extension objects will let you add unique features to the dashboards or integrate them with the applications outside Tableau.
- Ask Data objects will let users enter conversational queries which are specific for data source fields, which are optimized for specific audiences such as sales, marketing, and support staff.

Adding an object

Objects section which is at left, drag an item to the dashboard on the right.



Copying objects

Copy and paste objects that are either within the current dashboard, or from the dashboards in the other sheets and files. Objects can also be copied between the tableau desktop and even in your web browser.

You will not be able to copy when:

- Sheets are in a dashboard
- Items that rely on a specific sheet, such as filters, parameters, and legends
- In Layout containers, you can't copy inside them, like a sheet or filter
- Objects on a device layout

Learn by Doing

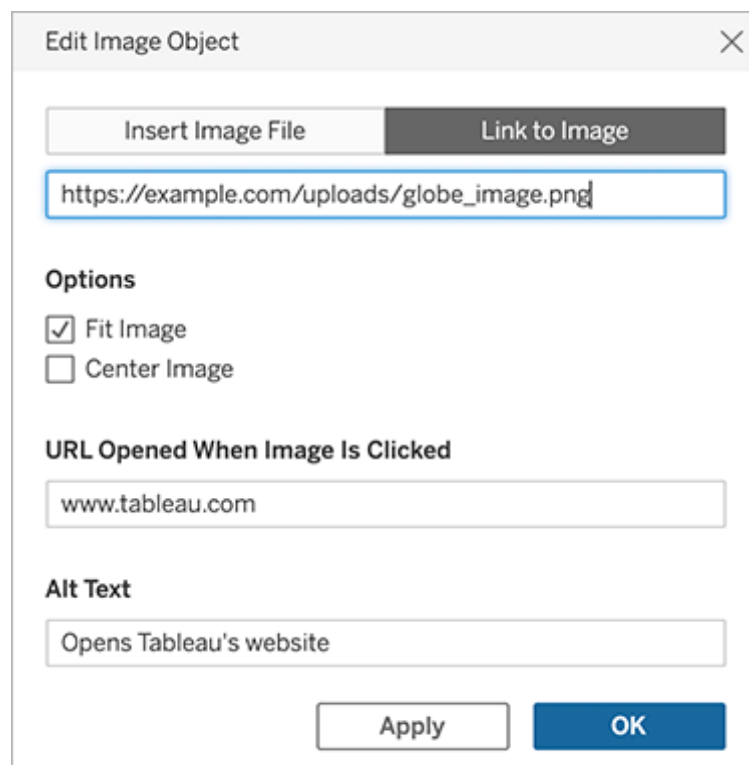
- Dashboard titles

Detailed options for image objects

When you are with the Image object, you can either insert an image file into the dashboard or link to images posted on the web. In either case, you can specify a URL the image will open up when clicked, which will add interactivity to your dashboard.

URLs for the web-based images will require the HTTPS prefix so that security can be improved. For image URLs with the other prefixes, use the Web page object.

- From the object section which is at the left, just drag an image object to your dashboard at right. Or, on an existing Image object in a dashboard, click the pop-up menu in the upper corner, and choose Edit Image.
- Now click either to the Insert Image file so that an image file can be embedded into the workbook or link to Link to Image to link to a web-based image.
- You can consider linking to a web-based image when:
 - The image is very large and the dashboard audience will be going to view it in a browser
 - The image is an animated GIF file.



- When you're inserting an image, just click Choose to select the file and if you're linking to an image, enter its web URL.
- Set remaining image fitting, URL linking, and alt text options.

Designing dashboards for devices

Learn by Doing

Dashboards can include layouts for different types of devices that span a wide range of screen sizes. Later on, you can publish these layouts to Tableau Server or Tableau Online, when someone views your dashboard they experience a design optimized for their phone, tablet, or even desktop.

How is the Default dashboard related to device layouts?

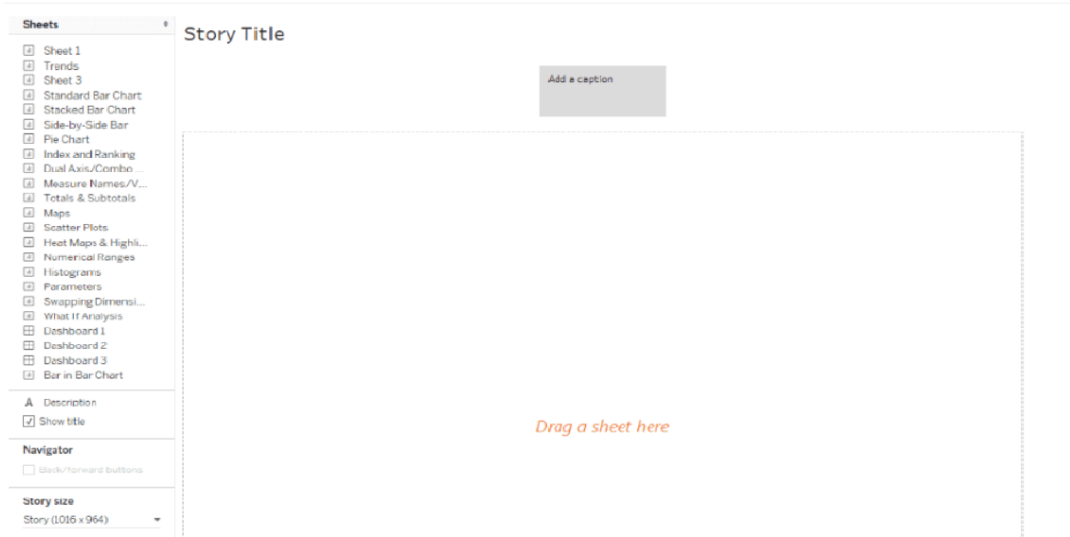
They appear on the Dashboard tab, under Default. Initially, each of the device layouts contains every item in the Default dashboard and it also derives its size and layout from Default as well. When you think of the Default dashboard as the parent, and the device layouts such as desktop, tablet, and phone as its children. Any view, filter, action, legend, or parameter that you want to add to a device layout must first exist in the default dashboard.

Story Points

In Tableau story points are the very powerful feature of Tableau. Tables of numbers can be generally seen as Excel, with a chart or two where we need to create and maintain those, and sometimes they aren't always done well, and excel can encode one or maybe two values at the most but a tableau chart can encode several around 6+ variables. Tableau story points with actions can solve your problems.

You can create a dashboard using Tableau story Points

- You have to think and structure your analysis like a story. It should have a start point and a series of steps that will take your story to a further level with a key insight you want your audience to take away.
- Now you can use Tableau story points to fall into the line.
- You have to click on the icon at the bottom named "Create Story". Which a new tab as "Story 1" gets created.



- Now you need to add a title to the tableau story that will apply to your entire dashboard.
- Here you can drag your charts, dashboard, text box, image, or webpage, onto the dashboard canvas.
- Title and captions on your charts and images should make sense to your end-user.

Learn by Doing

- To do that double click in the “Add a caption” box which is at the top and you need to add a descriptive sentence that helps build the action.
- Now click enter. A new tableau story point will appear to you.
- Later on, you can repeat these steps to build your entire story
- Now to format your Tableau story points or even title, click on the story menu linked at the top and click on format.

Sharing options in Tableau

In Tableau, the users can share projects, collections, data sources, and much more directly with the users or by copying a link to the content.

Sharing directly with other Users

Follow the below-mentioned steps to share in Tableau directly with other users.

- Open the content you want to share and open the actions option.
- Under that select share from the menu.
- Under the message option, add the additional note if you want to add any.
- Now, click the share button.

When you share content directly with the users, they receive the notification. The content is then added to their shared with me page.






Copying a link to share

You can generate the link for a piece of content and copy that to share with other users.

- Open the content you want to share and open the actions option.
- Under that select share from the menu.
- Click on the copy link button.
- Now, paste the link into an application to share it with other users.

Presenting your reports

The presentation mode is used when the user wants to share the findings using this mode. When the user uses this view, Tableau hides the toolbar and menu option and displays only the view. There are different controls available in the presentation mode like:

	Show Filmstrip - shows the sheets as thumbnails at the bottom of the workspace.
	Show Tabs - shows the sheet tabs at the bottom of the workspace.
	Previous/Next Sheet - advances forward or backward through the sheets in a workbook.
	Enter/Exit Full Screen - switches between expanding the workbook to fill the entire screen and showing it in a window.
	Exit Presentation Mode - returns the workbook to showing the entire workspace including the menus, toolbar, and the Data pane.

Printing your reports

You can print your reports of Tableau to a printer or as PDF. But, before hitting the print option you're required to do a Page Setup.

Export Data from Tableau

After you join the tables and data from more than one connection, you might want to share your Tableau Dashboards. The user can export data from Tableau in the following ways:

- **Export your data to .csv file**
.csv is the most basic, simple, and used format for the data. You can export the data in two ways:
On the main data source page, select Data> Export Data to CSV
On the sheet tab, drag a field to the Columns or Rows shelf, click the View Data icon in the Data pane, and click the export all button.
- **Export the Data Source**
When you finish connecting to your data, the user can export and save your data source as Tableau Data Source. The data will be stored as a .tds file. This will create a shortcut for your program.
- **Export Data used in the view**
The user also has the option to export data that is in the view without exporting whole data. The fields that are exported come from the fields on the shelves of the sheet. However, fields that function as external filters, in other words, the fields that appear only on the filters shelf, are not included in the export.

SQL

What is Database-

A database is a systematic way of collecting data. It supports the manipulation of data and electronic storage. Databases make the management of data easy and less complex.

We will be able to organize the data into tables, rows, columns, and indexes which make it easier to find relevant information.

The main purpose of the database is to operate a large amount of information by storing and managing data.

Generally, modern databases are managed by database management systems (DBMS).

Types of Database Management Systems-

Database Management systems can be of several types. Here is a list of some common database management systems.

- Hierarchical databases
- Network databases
- Relational databases
- Object-oriented databases
- Graph databases
- Centralized database
- Distributed database
- NoSQL databases

Hierarchical databases

Here is a hierarchical database management system, data is stored in a node which is like a parent-child relationship. Besides actual data, records also contain information about the relationships of groups of parent/child relationships.

In Hierarchical databases data is organized into a tree-like structure, the data is stored in such a way that each field contains only one value, and records are linked to each other via links to a parent-child relationship. Each of the child records can only have one parent whereas a parent can have multiple children, if we want to retrieve a field's data, we need to traverse through each tree until the record is found.

Network Databases

It uses a network structure to create a relationship between entities. They are mainly used on large digital computers. They are hierarchical databases, but unlike hierarchical databases

Learn by Doing

where one node can have a single parent only, here in a network node multiple entities can have a relationship, it looks like a record of interconnected network.

Here children are known as members and parents are known as occupiers. The main difference between each child or member is that it can have more than one parent.

Relational Databases

Here is a relational database management system (RDBMS), the relationship between the data is relational and data here is stored in the form of tables which consist of columns and rows. Each column of a table represents an attribute and each row in a table represents a record. Each field in a table represents a data value.

Structured Query Language (SQL) is the language used to query RDBMS, which includes inserting, updating, deleting, and searching records.

It works on each table that has a key field that uniquely indicates each row, later on, these fields can be used to connect one table of data to another. Relational databases are more widely used databases. Some examples can be Oracle, SQL Server, MySQL, SQLite, and IBM DB2.

Object-Oriented Model

It provides full-featured database programming capabilities while containing native language compatibility which adds the database functionality to object programming languages and creates a more manageable codebase. It uses small, recyclables separated from software called objects, where objects themselves are stored in the object-oriented database.

Here each object contains two elements:

- It can be data (e.g., sound, video, text, or graphics).
- Instructions, or software programs called methods

Graph Databases

These Databases are NoSQL databases as they use a graph structure for semantic queries. The data in this database is stored in the form of nodes, edges, and properties. In a graph database, a Node represents an entity or instance such as a customer, person, or car where a node is equivalent to a record in a relational database system. An Edge in a graph represents a relationship that connects nodes.

Azure Cosmos DB, SAP HANA, Oracle Spatial, and Graph are some popular graph databases. It is also supported by some RDBMS like Oracle and SQL servers.

Centralized Database

It is a type of database that is stored, located as well as maintained at a single location only. This type of database is modified and managed from that location itself. This location is thus mainly any database system or a centralized computer system. The centralized location is accessed via an internet connection (LAN, WAN, etc). This centralized database is mainly used by institutions or organizations.

Distributed Database

It is a type of database that consists of multiple databases that are connected and are spread across different physical locations. The data that is stored in various physical locations can thus be managed independently of other physical locations. The communication between databases at different physical locations is thus done by a computer network.

NoSQL Databases

These databases do not use SQL as their primary data access language. A graph database, network database, object database, and document database are some common NoSQL databases. This database does not have predefined schemas, which makes it the best. It also allows developers to make changes without affecting applications. It has five major categories as Column, Document, Graph, Key-value, and Object databases.

What is a Relational Database Management system-

A relational Database Management System (RDBMS) is an information management system that is oriented on a data model. All the information here is properly stored as tables. Generally, it arranges information into allied rows and columns. Some examples of RDBMS are SQL Server, MySQL, SQLite, Oracle, and MariaDB.

Features of RDBMS:

- Here information can be saved in the tables.
- Several users can access it together which can be managed by a single user.
- Here data are always saved in rows and columns.
- Indexes are used to get the information.

Data Types in SQL-

They are used to represent the nature of different types of data that can be stored in the database table. Let's take an example, Suppose there is a table that has a particular column, if we want to store data that is of string type in it then we have to declare a string data type of this column.

Data types can be mainly classified into three categories for ever databases.

- String Data types
- Numeric Datatypes
- Date and time Data types

What are SQL Operators?

Every database administrator and a user uses SQL queries for accessing and manipulating the data of database tables and views.

This manipulation and retrieving of the data are performed with the help of reserved words and characters, which are generally used to perform some of the operations such as arithmetic operations, logical operations, comparison operations, compound operations, etc.

In SQL, reserved words and characters are known as operators, they are generally used with a WHERE clause in an SQL query. Here In SQL, an operator can be either a unary or binary operator. The unary operator uses only one type of operand while performing the unary operation, whereas the binary operator uses two types of operands for performing the binary operations.

Types of Operator-

SQL operators can be of the following categories:

- SQL Arithmetic Operators
- SQL Comparison Operators
- SQL Logical Operators
- SQL Compound Operators
- SQL Unary Operators

Some important Comparison Operators Performed on the SQL data:

Operator	Description	Example
=	It checks if the values of two operands are equal or not, if yes then the condition becomes true.	Here (a = b) is not true.
!=	It checks if the values of two operands are equal or not, if values are not equal then the condition becomes true.	Here (a != b) is true.
<>	It checks if the values of two operands are equal or not, if values are not equal then the condition becomes true.	Here (a <> b) is true.
>	It checks if the value of the left operand is greater than the value of the right operand, if yes then the condition becomes true.	Here (a > b) is not true.

Some Important SQL Logical Operators:

Operators	Description of the operators
ALL	It is used to compare a value to all other values in the value set.

Learn by Doing

AND	This operator allows the existence of multiple conditions in an SQL statement's where clause.
ANY	It is used to compare a value to any applicable value in the list as per the condition.
BETWEEN	It is used to search for the values which lie within a set of values, given the minimum value and the maximum value.

Storage Engines

Storage engines (underlying software components) are MySQL components, that can handle the SQL operations for different table types to store and manage information in a database. InnoDB is mostly used general-purpose storage engine, by using storage engines you can easily interact with a file at an OS level so that data can be stored in it.

Keys in SQL-

It is just a single attribute which is a column that can uniquely identify a row.

Some types of Keys are:

- **Primary key-**
The primary key helps to identify every record that is present in the table uniquely. We can have only one primary key in a table while there can be multiple unique keys.
- **Super Key-**
It is a set of attributes that can be one or more than one that collectively identifies an entity set.
- **Candidate key-**
It is a minimal super key. An entity set can have more than one candidate key.
- **Foreign key-** It is used to define a relationship between two tables.
- **Alternate key-** It is a table that has more than one candidate key, and after choosing the primary key from those candidate keys, the rest of the candidate keys are known as alternate keys of that table.
- **Composite Key-** They are the keys with more than one column.

DDL and DML Statement-

Some of the Major SQL statements are:

- **Data Definition Language (DDL) Statements:**

It consists of some SQL commands which can be used to define the database schema, as it simply deals with the description of the database schema and is also used to create and modify the structure of the database objects. DDL is a set of SQL commands used mostly by advanced users.

Some DDL Commands:

- **CREATE:** It is used to create the database or objects (which can be a table, index, function, views, store procedure, and triggers).
- **DROP:** It is used to delete objects from the database.
- **ALTER:** It is used to alter the structure of the database.
- **TRUNCATE:** It is used to remove all records from a table, including all spaces allocated for the records are removed.
- **COMMENT:** It is used to add comments to the data dictionary.
- **RENAME:** It is used to rename an object existing in the database.

CREATE, ALTER, and DROP commands require exclusive access to the specified object. An ALTER TABLE statement fails if another user has an open transaction on the specified table.

Some more DDL Commands like GRANT, REVOKE, ANALYZE, AUDIT, and COMMENT commands do not require exclusive access to the specified object.

- **Data Manipulation Language (DML) Statements:**

It deals with the manipulation of data that is present in the database belonging to Data Manipulation Language and includes most of the SQL statements. It is the component of the SQL statement which controls access to the data and the database.

Some DML commands:

- **INSERT:** It is used to insert data into a table.
- **UPDATE:** It is used to update existing data within a table.
- **DELETE:** It is used to delete records from a database table.
- **LOCK:** Table control concurrency.
- **CALL:** Call a PL/SQL or JAVA subprogram.

EXPLAIN PLAN: It describes the access path to data.

Functions in SQL

They are the methods where data operations are performed. SQL has many in-built functions used to perform string concatenations, mathematical calculations etc.

Learn by Doing

SQL functions are categorized into the following two categories:

Aggregate Functions

Scalar Functions

Aggregate SQL Functions

The Aggregate Functions in SQL perform calculations on a group of values and then return a single value.

Scalar SQL Functions

The Scalar Functions in SQL are used to return a single value from the given input value.

SQL Joins

SQL joins are used when we need to combine records from two or more two different tables in a database, In other words, it helps us to retrieve data from two or more database tables, where these two tables are related to each other using primary and foreign keys.

Types of joins available in SQL

There are many different types of joins available in SQL, to understand them let's prepare sample data.

- **Multiple Join-**

This Join is used to perform multiple joins in a query statement that can retrieve the data by combining the records of more than one table. Whenever we perform joining of more than one join in a single query statement, then we are making use of multiple joins.

Inner join-

The INNER JOIN keyword selects all rows from both tables as long as the condition is satisfied. This keyword will create the result-set by combining all rows from both the tables where the condition satisfies

i.e value of the common field will be the same.

Left join-

Left join returns all the rows of the table on the left side of the join and matching rows for the table on the right side of the join. For the rows for which there is no matching row on the right side, the result-set will contain null. LEFT JOIN is also known as LEFT OUTER JOIN.

Right join-

RIGHT JOIN is similar to LEFT JOIN. This join returns all the rows of the table on the right side of the join and matching rows for the table on the left side of the join. For the rows for which

Learn by Doing

there is no matching row on the left side, the result-set will contain null. RIGHT JOIN is also known as RIGHT OUTER JOIN.

Full Join-

FULL JOIN creates the result-set by combining results of both LEFT JOIN and RIGHT JOIN. The result-set will contain all the rows from both tables. For the rows for which there is no matching, the result-set will contain NULL values.

Subqueries in SQL-

It is a query within another SQL query and which is embedded within the WHERE clause.

It is used mostly when we need to return data that will be used in the main query as a condition to further restrict the data to be retrieved.

They can be easily used with the SELECT, INSERT, UPDATE, and DELETE statements along with the operators like =, <, >, >=, <=, IN, BETWEEN, etc.

Some rules that need to be followed-

- It must be always enclosed within parentheses.
- It can have only one column in the SELECT clause, whereas multiple columns are in the main query for the subquery to compare its selected columns.
- An ORDER BY command cannot be used in a subquery, the main query can use an ORDER BY. Also, the GROUP BY command can be used to perform the same function as the ORDER BY in a subquery.
- Subqueries that return more than one row can only be used with multiple value operators such as the IN operator.
- The SELECT list cannot include any references to values that evaluate a BLOB(Binary large objects)
- , ARRAY, CLOB, or NCLOB.
- It cannot be immediately enclosed in a function.
- BETWEEN operator cannot be used with a subquery.

Stored Procedures-

This code in SQL can be stored for later use and can be used many times. In that case,

Learn by Doing

whenever you need to execute a query you can just simply call the stored procedures to perform the task, you can even pass parameters to an already stored procedure, as the stored procedure will act based on the parameters' values.

Types of Stored Procedures-

They are of two types mainly:

- **User-defined Stored Procedures:**

They provide one or more SQL statements for selecting, updating, or removing data from database tables. They are specified by the user who accepts input parameters and returns output parameters. DDL and DML commands are used together in a user-defined procedure.

- **System Stored Procedures:**

When SQL Server is installed, It creates system procedures. The system stored procedures prevent the administrator from querying or modifying the system and database catalog tables directly. Developers often ignore system stored procedures.

Filter-

When you want to refine your query by running your aggregations against a set of limited values in a column then you can use the **FILTER** keyword. As the filter clause will extend aggregate functions such as sum, avg, count, etc. By using an additional WHERE clause. The result of the aggregate will be built up from only the rows which satisfy the additional WHERE clause too.

STATISTICS

What is Statistics?

Statistics is a branch of Mathematics dealing with Data Collection, Organization, Analysis, Interpretation, and Presentation.

As defined by the American Statistical Association (ASA)- is the science of learning from data and measuring, controlling, and communicating, uncertainty. Statistics is the art of learning from data. In other words, we can say that it is concerned with the collection of data, subsequent description, and their analysis, which leads it to draw of conclusion, Moreover, in statistics, we will be going to study a large collection of people or objects.

Types of Statistics Descriptive Statistics

Descriptive Statistics, as the name implies means describe the description of the data. In other words, we can say that it summarizes a given data set, which can be either a representation of the entire population or a sample of a population. In descriptive statistics, there is no uncertainty in the statistics describing the data that we have collected.

Inferential statistics

In Inferential statistics, we are working on the samples, because it is too difficult or we can say expensive to collect data from the whole population that we are interested in.

While descriptive statistics can only summarize a sample's characteristics, inferential statistics use the sample to make reasonable guesses about the larger population.

What is Probability?

In the most literal sense, the probability is the likelihood of the occurrence of an event.

Probability of an event= (Number of favorable outcomes)/ (Total Number of Possible Outcomes)

$$P(A) = n(E)/n(S)$$

Types of Probability Distribution

Uniform distribution is fairly simple. Every value has a change of incidence that is equal. The distribution is thus made up of random values with no trends in them.

Normal Distribution-The "Bell Curve" is a Normal Distribution and some data that follows it closely, but not perfectly (which is usual). It is often called a "Bell Curve" because it looks like a bell.

Learn by Doing

Binomial Distribution is a type of distribution that has two possible outcomes (the prefix “bi” means two, or twice). For example, a coin toss has only two possible outcomes

Variable

A variable is a characteristic of a unit that is being observed and can assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned.

Some types of variables in the field of Data Science are listed below:

- Numerical
- Categorical

Numerical: This category of variables is the variable that deals with numbers only. This can now be divided into 2 subcategories.

- Discrete: A discrete variable is a numeric variable. Observations can be taken as a value that is based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction that lies between one value and the next closest value.
- Continuous: A Continuous Variable is a numerical Variable that deals with quantities such as continuous quantities or fractional quantities.
 - o Height
 - o Age
 - o Temperature.

Categorical: Categorical variables have values that describe a ‘quality’ or ‘characteristic’ of a data unit, like ‘what type’ or which category. Categorical variables can be put into categories.

Some terminologies in Probability

- **Experiment:** It is an activity whose outcomes are not known, is called an experiment. Every experiment has a few favorable outcomes and a few unfavorable outcomes.
- **Event:** An event is a trial with a clearly defined outcome. An example of an event can be getting a tail while tossing a coin which is an event.
- **Random Event:** An event that cannot be predicted easily is a random event. For these events, the probability value is very less. An example of a Random event can be seeing a shooting star as a random event.
- **Trial:** The number of numerous attempts in the process while experimenting is called trials, or we can say that any particular act of a random experiment is called a trial. An example of a trial can be the tossing of a coin.
- **Outcome:** The result of a trial can be termed an outcome. An example of an outcome can be a footballer, either he will hit the goal or miss the goal.
- **Mutually Exclusive Events:** When the happening of one event prevents the happening of another event then they are known as mutually exclusive events, or we can say that two events are mutually exclusive if they cannot occur at the same time.

Learn by Doing

Sample

A sample is a subset of the population, to study the larger population we select a sample. In sampling, we select a portion of a larger population and study that portion to gain information about the population.

Methods of Sampling – Probability & Non-Probability Sampling

Probability Sampling

Probability sampling represents a group of sampling techniques that help researchers to select units from a population that they are interested in studying. Types of Probability Sampling.

- **Simple Random Sampling:** With simple random sampling, there is an equal chance (probability) that each of the units could be selected for inclusion in our sample.
- **Systematic Sampling:** Every member of the population here is listed with a number, but instead of randomly generating the numbers, here individuals are chosen at regular intervals.
- **Stratified Sampling:** With the stratified random sample, there is an equal chance (probability) of selecting each unit from within a particular stratum (group) of the population when creating the sample.
- **Clustered Sampling:** It is a method where we divide the entire population into sections or clusters that represent a population. This method is good for dealing with large and dispersed populations

Non-Probability Sampling

Non-probability sampling represents a group of sampling techniques that help researchers to select units from a population that they are interested in studying. Its Types are:

- **Quota Sampling:** With proportional quota sampling, the aim is to end up with a sample where the strata (groups) being studied (e.g., males vs. females students) are proportional to the population being studied.
- **Convenience Sampling:** A convenience sample is simply one where the units that are selected for inclusion in the sample are the easiest to access.
- **Snowball Sampling:** Snowball sampling is particularly appropriate when the population you are interested in is hidden and/or hard to reach. It can be used to recruit participants via other participants.
- **Judgement Sampling:** Also known as selective, or subjective, sampling, this technique relies on the judgment of the researcher when choosing who to ask to participate

Learn by Doing

Range

A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. It is based on two extreme observations. Hence, gets affected by fluctuations.

- If X max and X min are the two extreme observations then

$$\text{Range} = X \text{ max} - X \text{ min}$$

Quartile

The quartiles will divide the data set into quarters. The first quartile, (Q1) will be the middle number between the smallest number and the median of the data. The second quartile, (Q2) will be the median of the data set. The third quartile, (Q3) will be the middle number between the median and the largest number.

A quartile divides a sorted data set into 4 equal parts so that each part represents $\frac{1}{4}$ of the data set.

Variance

Variance is the average squared deviation from the mean of a set of data. Variance is generally used to find the standard deviation.

Measures of Central Tendency

- **Mean:** It is simply the average of all the data (salary) values. Add all the numbers then divide by the number of numbers.
- **Median:** It is the value in the middle when the data items are arranged in ascending order. It needs the arrangement of a series in ascending or descending order
- **Mode:** It is the most frequently occurring value in a series of data in case of no repeating values, there would be no mode.

T-Test

To evaluate whether there is a significant difference between the means of two groups that may be related in some ways, a t-test is a sort of inferential statistic that is utilized. It is typically employed when data sets, such as the one representing the results of tossing a coin 100 times, would follow a normal distribution and might contain unidentified variances.

Types of T-Test

There are three types of T-tests.

Learn by Doing

- One sample T-test
- Independent Two-sample T-test
- Paired sample t-test

Chi-Square

The Chi-Square statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi-Square test is that no relationship exists between the categorical variables in the population; they are independent. It is used to determine whether there is a statistically significant association between the categorical variables. Thus, it finds out if the relationship exists between any two business parameters that are of categorical data type

Anova

It stands for Analysis of Variance. It is used to determine whether there is a statistically significant difference among more than two group means.

Example Anova-

We could use the One-way Anova test to determine if out of three or more rivers, at least two of them differ significantly from each other in terms of pH, TDS, etc.

We could determine if at least two regions differ significantly in terms of average sales of a particular product category

Skewness

It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution.

It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0

Hypothesis

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is an assumption that we make about the population parameter.

Critical Value

We need to consider the following two facts. One significance level is the probability of rejecting a correct null hypothesis.

The sampling distribution for a test statistic assumes that the null hypothesis is correct.

P-value

A statistical hypothesis test may return a value called p or the p-value. This is a quantity that we can use to interpret or quantify the result of the test and either reject or fail to reject the null hypothesis. This is done by comparing the p-value to a threshold value chosen beforehand called the significance level. The significance level is often referred to by the Greek lower-case letter alpha.

Degree of Freedom in Stats

In statistics, the degree of freedom is the number of values that can be involved in a calculation and has the freedom to vary. It can be computed to some of the statistical validity tests like t s, chi-square tests, and the more elaborated f-tests. The Formula is $df = N - 1$.

Skills of Data Scientist

There are two different types of skills required by a data scientist and they're technical and non-technical skills. Various technical skills are machine learning, deep learning, data visualization, data wrangling, etc. And, various non-technical skills required are strong communication skills, data intuition, and strong business acumen.

Machine Learning

Machine learning is an interesting branch of artificial intelligence. Machine learning helps us in accessing data in new ways. For example, Facebook recommends you the ads for products that you searched on other platforms. This amazing technology helps the machines access data from the systems and performs smart tasks.

Data visualization

Data visualization is also known as information visualization. It is the process of translating data and information into a visual context. The visual context involves a chart, bar, graph, etc. This step in the process shows that the information is collected and processed also. The visualized information allows the user to conclude.

Data wrangling

It is the process of cleaning, organizing, and transforming raw data into the format that is desired. The exact method of data wrangling varies as it depends on the project and type of data. Data wrangling helps in making the raw data useful. Accurately wrangled data help and guarantee that correct data is entered.

Learn by Doing

Communication skills

Data scientists extract, understand and analyze data. However, to be successful in the role and to benefit the organization, you must be able to successfully communicate the results with your team members who are not from the same background as you.

Computer programming

As a data scientist, you need to know the various programming language. Language like Python, C/C++, SQL, and Java is the most common coding language required in a data science role.