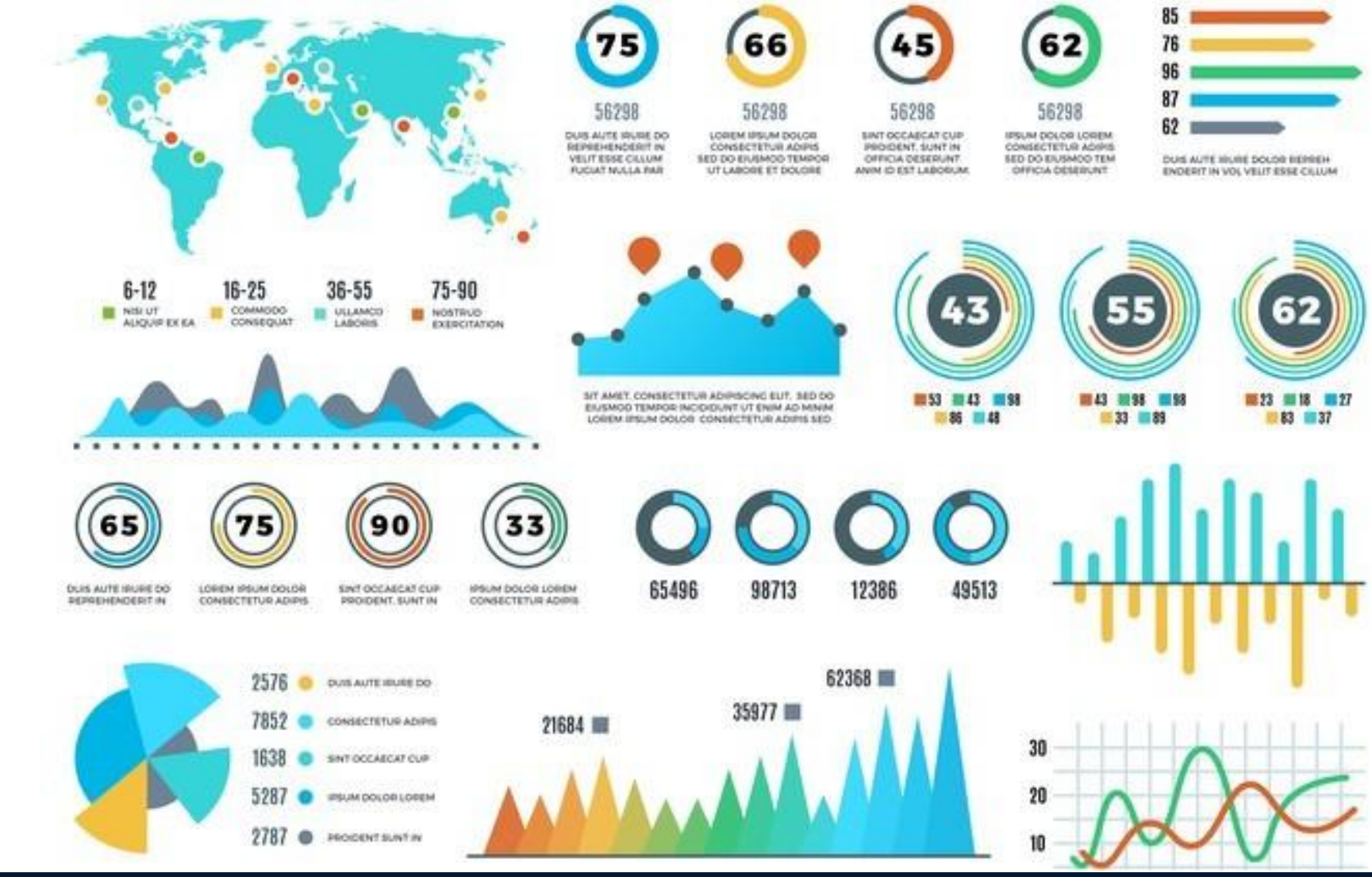
Data Visualization

UNIT-I

MR.G.JASWANTH
ASSISTANT PROFESSOR

- Data visualization
  - What? Why?
  - Benefits
  - Techniques
  - Who uses it?
- Types of Graphs
- Tools
- Techniques in programming
- Best resources

Just check...


75
66
45
62
56298
56298
56298
56298
85
76
96
87
62
6-12
16-25
36-55
75-90
43
55
62
65
75
90
33
65496
98713
12386
49513
2576
7852
1638
5287
2787
21684
35977
62368
30
20
10

Data Visualization

- Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.
- The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.
- The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data Visualization

- Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made.
- Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

Data Visualization

- Data visualization is important for almost every career.
- It can be used by teachers to display student test results, by computer scientists exploring advancements in artificial intelligence (AI) or by executives looking to share information with stakeholders.
- It also plays an important role in big data projects. As businesses accumulated massive collections of data during the early years of the big data trend, they needed a way to get an overview of their data quickly and easily.
- Visualization tools were a natural fit.

Benefits of Data Visualization

- The ability to absorb information quickly, improve insights and make faster decisions;
- An increased understanding of the next steps that must be taken to improve the organization;
- An improved ability to maintain the audience's interest with information they can understand;
- An easy distribution of information that increases the opportunity to share insights with everyone involved;

Benefits of Data Visualization

- Eliminate the need for data scientists since data is more accessible and understandable; and
- An increased ability to act on findings quickly and, therefore, achieve success with greater speed and less mistakes.

Data Visualization Roles

- Showing change over time
- Showing a part-to-whole composition
- Depicting flows and processes
- Looking at how data is distributed
- Comparing values between groups
- Observing relationships between variables
- Looking at geographical data

Change over time

Line chart +Comparisons

Most common chart type for showing change over time. A point is plotted for each time period from left to right; each point's vertical position indicates the feature's value. Points are connected by line segments to emphasize progression across time.

Sparkline +Comparisons

A miniature line chart with little to no labeling, designed to be placed alongside text or in tables. Provides a high-level overview without attracting too much attention. Can also be seen in a sparkbar form, or miniature bar chart (see below).

Change over time

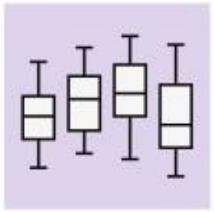
Connected scatter plot ◆ *+Relationships*

Shows change over time across two numeric variables (see scatter plot in *Relationships*). Line segments still connect points across time, but they may not consistently go from left to right like in a line chart.

Bar chart ● *+Distributions* *+Comparisons*

Each time period is associated with a bar; each bar's value is represented in its height above (or below) a zero-baseline. Works best when there aren't too many time periods to show.

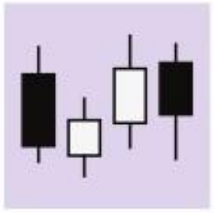
Change over time



Box plot +Distributions +Comparisons

Each time period is associated with a box and whiskers; each set of box and whiskers shows the range of the most common data values. Best when there are multiple recordings for each time period and a distribution of values needs to be plotted.

Tracking change over time is of key interest in the financial domain. One specialist chart developed for this field includes the following:



Candlestick chart

Looks like a box plot, but each box and whiskers encodes different statistics. The box ends indicate opening and closing prices, while color indicates the direction of change.

Part-to-whole composition

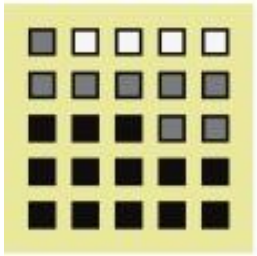
Pie chart

The whole is represented by a filled circle. Parts are proportional slices from that circle, one for each categorical group. Best with five or fewer slices with distinct proportions.

Doughnut chart

A pie chart with a hole in the center. This central area can be used to show a relevant single numeric value. Sometimes used as an aesthetic alternative to a standard progress bar (see stacked bar chart below).

Part-to-whole composition



Waffle chart / grid plot

Squares laid out in a (typically) 10 x 10 grid; each square represents one percent of the whole. Squares are colored based on categorical group size.

Stacked bar chart

A bar chart (see *Change over time* or *Distributions*) where each bar has been divided into multiple sub-bars to show a part-to-whole breakdown. A single stacked bar can be used as an alternative to the pie or doughnut chart; people tend to make more precise judgments of length over area or angle.

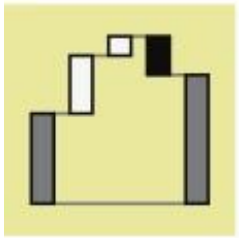
Part-to-whole composition

Stacked area chart ●

A line chart (see *Change over time*) where shaded regions are added under the line to divide the total into sub-group values.

Stream graph ◆

Modified version of the stacked area chart where areas are stacked around a central axis. Highlights relative changes instead of exact values.



Waterfall chart ◆

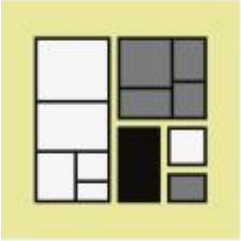
Augments a change over time with a part-to-whole decomposition. Bars on the ends depict values at two time points, and lengths of intermediate floating bars' show the decomposition of the change between points.

Part-to-whole composition

Certain part-to-whole compositions follow a hierarchical form. In these cases, each part can be divided into finer parts on lower levels. Here are a couple of more specialized chart types for visualizing this type of data:

Mosaic plot / Marimekko chart

Can be thought of as a stacked bar divided on both axes. A box is divided on one axis based on one categorical variable, then each sub-box is divided in the other axis based on a second categorical variable.



Treemap

Can be thought of as a more generalized Marimekko plot. Sub-boxes do not need to have a consistent cut direction at a particular hierarchy level, and there can be more than two levels of hierarchy.

Flows and processes

Funnel chart

Seen in business contexts, showing how people encounter a product and eventually become users or customers. One bar is plotted for each stage, whose lengths reflect the number of users. Connecting regions emphasize connections in stages and give the chart type's namesake shape.

Parallel sets chart

Multiple part-to-whole divisions on different dimensions are depicted as parallel stacked bars. Connecting regions show how different subgroups relate to one another between dimensions.

Flows and processes

Sankey diagram ◆

The width of the colored region shows the relative volume at each part of a process. Allows for multiple sources of inputs and outputs to be visualized.

Gantt chart ■

Used for project scheduling, breaking them down into individual tasks. Each task is associated with a bar, providing a timeline for when each task should begin and end.

How data is distributed?

Bar chart +Change over time +Comparisons

Used when a variable is qualitative or takes discrete values. The height of each bar indicates the amount of each categorical group.

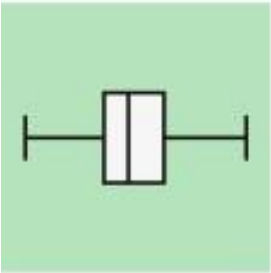
Histogram

Similar to a bar chart, but used when a variable takes continuous numeric values. The variable's numeric range is divided into bins for aggregating counts. Bars are plotted flush against each other to emphasize the variable's continuous nature.

How data is distributed?

Density curve

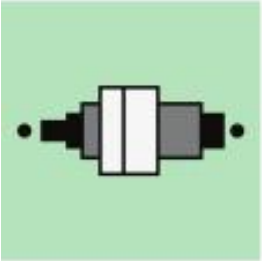
An alternative to the histogram when a variable takes numeric values. Each data point contributes a small amount of local area; the areas are summed across all points to form the full curve.



Box plot +Change over time +Comparisons

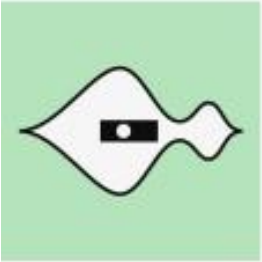
A box and whiskers shows the range of the most common data values. The ends of the box outline the central 50% of the data. More often used to compare distributions between groups rather than as an overall summary.

How data is distributed?



Letter-value plot ◆ *+Comparisons*

Extends the box plot's marking of quartiles with additional boxes that denote eighths, sixteenths, and smaller quantiles. Best when there are lots of data available to make estimates stable.



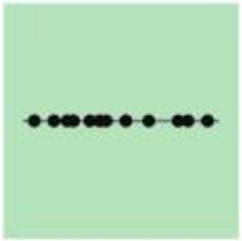
Violin plot ■ *+Comparisons*

Combines a density curve plotted on a center line with a box plot as a statistical summary. More often used to compare distributions between groups rather than as an overall summary.

How data is distributed?

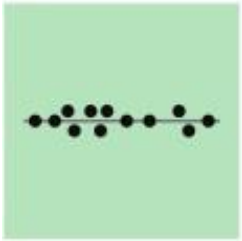
Rug plot

All data points are plotted as tick marks on a straight line with value corresponding precisely with position.



Strip plot

Like a rug plot, but with dots instead of tick marks. Sometimes plotted with points randomly jittered up or down to reduce overlapping.



Swarm plot

Like a strip plot, but deliberate shifting is performed to prevent overlapping. Some horizontal jitter may be needed in order to keep the dot swarm compact.

Comparing values between groups

Bar chart +Change over time +Distributions

Most basic way of comparing numeric values between groups or categories. Each group is assigned a bar; each bar's value is represented in its height above (or below) a zero-baseline.

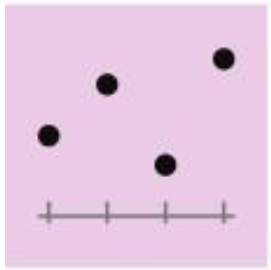
Grouped bar chart +Relationships

Extends a bar chart to compare data across two categorical variables. Each bar corresponds to an intersection of variable levels: categories for one variable are indicated by the bar cluster positions, while the second variable is indicated by bar color or position within each cluster.

Comparing values between groups

Lollipop chart

Replaces the bars of a bar chart with lines and dots. Useful for when there are a lot of groups or categories to plot.



Dot plot

Replaces the bars of a bar chart with just dots. Since value is indicated by position instead of length, the dot plot can be good when a zero baseline is not useful.

Comparing values between groups

Line chart +Change over time

Each line in a line chart shows how values (vertical position) change across time (horizontal). One line is plotted for each group to be compared. Best when there are five or fewer groups to plot.

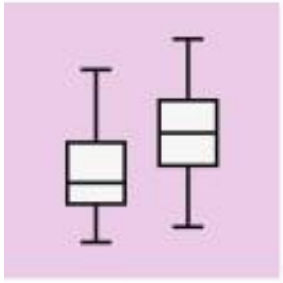
Sparkline +Change over time

Smaller line charts typically with little to no labeling. Designed to show a high-level overview inline with text or tables, but also useful when there are many groups to plot.

Comparing values between groups

Ridgeline

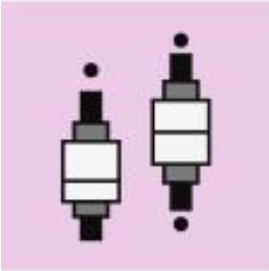
A series of line charts or density curves (see *Distributions*) with partially offset axes used to compare distributions between groups. Best when there are distinct patterns across groups.



Box plot *+Change over time* *+Distributions*

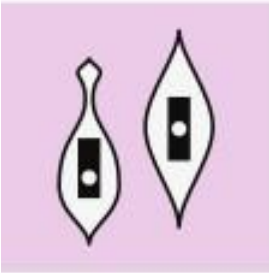
Compares a statistical summary of numeric values between groups. A set of box and whiskers depicting the range of the most common data values (see *Distributions*) is assigned to each group or category.

Comparing values between groups



Letter-value plot ◆ +Distributions

Used in a similar way as the box plot, but a letter-value plot (see *Distributions*) is assigned to each group instead. Best used when there are lots of data in each group so that statistical estimates are stable.



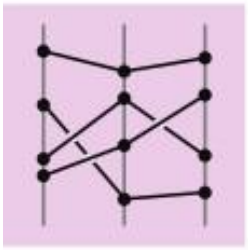
Violin plot ● +Distributions

Compares distributions between groups. A violin assembly of density curve and box plot (see *Distributions*) is assigned to each group or category.

Comparing values between groups

Slope chart

Specialized type of line chart. Two parallel lines indicate different times, with vertical position indicating value. One line segment is drawn between the two times for each data point. Useful for when there are many data points; line slopes provide a quick indicator for direction of change for each one.



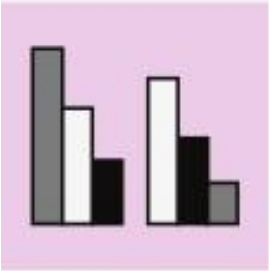
Parallel coordinates plot

Extension of the slope plot for multiple dimensions. Each vertical line now indicates a different variable; each may have its own scale. Useful for observing patterns and relationships in the data. When there are only two variables, a scatter plot (see *Relationships*) is often easier to read.

Comparing values between groups

Bump chart

Modified version of a line chart where vertical position corresponds to rank rather than value. This change allows it to support more categories than a standard line chart.



Grouped bar chart

Normally, grouped bar charts will plot the bars within each group in a consistent order. However, they can instead be sorted by value within each group to emphasize ranking, at the cost of making it more difficult to find each sub-category.

Relationships between variables

Scatter plot

Standard chart type for showing relationships between two numeric variables. Each point's position on the horizontal and vertical axes indicate value on the associated variable.

Bubble chart

Scatter plot with point size dictated by a third numeric variable. Scatter plots can be extended in other ways: point shapes can encode a categorical variable, and color can be used to indicate either categorical or numeric data. It is best to keep a scatter plot to a maximum of three variables to maintain understandability.

Relationships between variables

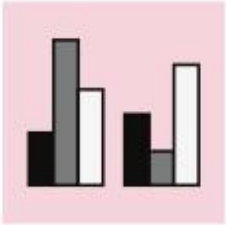
Connected scatter plot ◆

When a third variable represents time, points in a scatter plot can be connected with line segments to show progression in values across time.

Dual-axis bar-line plot ◆

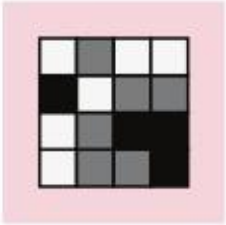
A bar-line plot shares a horizontal axis (typically time) across two chart types: the bar chart and line chart. Useful for when the variables plotted with each chart type are related, but are on different numeric scales.

Relationships between variables



Grouped bar chart +Comparisons

Extension of bar chart (see *Comparisons* or *Distributions*) to two categorical variables. Bar clusters are associated with levels of one variable, while color or position in each cluster indicates levels of the second variable. The length of each bar at the corresponding intersection of levels indicates a value for that group, like data frequency or a summary of a third numeric variable.



Heatmap

Extension of bar charts and histograms (see *Distributions*) to two variables, each of which can be categorical or numeric. Each axis represents groups or bins of values for one of the variables, forming a grid. Cell colors indicate data frequency or a summary of a third variable for each intersection of axis variables.

Relationships between variables

2-d density curve ◆

Extension of density curves (see *Distributions*) to two numeric variables. Colors are mapped to values like in a heatmap, but applied smoothly across the plotted area rather than in discrete bins. Somewhat confusingly, this chart is sometimes also known as a heatmap.

Dendrogram ◆

Specialized chart type to show similarity between data points. The lower the branch connecting two data points is, the more similar they are. Sometimes plotted with an accompanying heatmap to depict the underlying data.

Relationships between variables

Network diagram ◆

Points (nodes or vertices) represent individual entities. Lines (edges) connect entities with a particular relationship. Line thickness may be used to encode value. Vertex positions do not necessarily have any inherent meaning, and may simply be placed just to make connections as clear as possible.

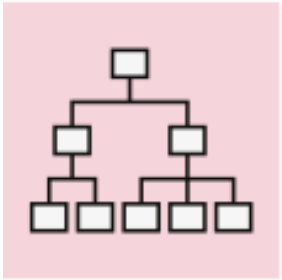
Transit map ◆

Practical application of network diagrams for train and subway systems. Frequently, these take a fair level of abstraction, emphasizing connections between stations rather than their actual geographical locations.

Relationships between variables

Chord diagram ◆

Like a standard network diagram, but vertices are arranged in a circle.



Tree diagram ◆

A network diagram organized to show hierarchical relationships. The direction of each edge corresponds to a relationship between the connected nodes, such as parent-child or senior-junior relationships.

Geographical data

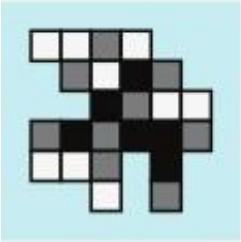
Scatter map

Scatter plot built on top of a geographical map, using geographic coordinates as point positions.

Bubble map

Bubble chart built on top of a geographic map, where point size is an indicator of value. Can also be used to group together points in a scatter map if they are too dense.

Geographical data



2-d histogram ●

Heatmaps can be built on top of geographic areas. Sometimes seen with a hexagon-shaped grid rather than a rectangular grid. May distort the geography on its edges.

Isopleth / contour map ◆

2-d density curve built on top of a geographic map.

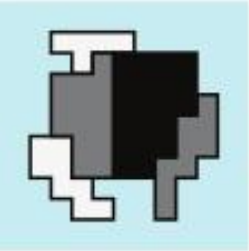
Connection map ◆

Network information and flows built on top of a geographic map.

Geographical data

Choropleth

Similar to a heatmap, but colors are assigned to geopolitical regions rather than an arbitrary grid. Values are often in the form of rates or ratios to avoid distortion due to population density.



Cartogram

Geopolitical regions sized by value. This necessarily requires distortion in shapes and topology.

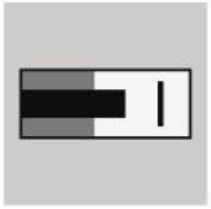
Raw Numbers

Single Value Chart

Show a **raw singular value**

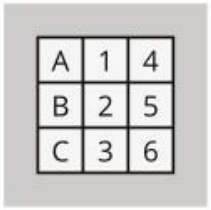
Single Value w/ Indicator

Comparison of a **single value** against a **previous value**



Bullet Chart

Comparison of a **single value** against a **benchmark value**



Table

Show raw values for **multiple data points** on **multiple variables**

Data Visualization Tools

- Tableau
- Infogram
- ChartBlocks
- D3.js
- Google Charts
- Fusion Charts
- Chart.js

Tableau

- Tableau has a variety of options available, including a desktop app, server and hosted online versions, and a free public option.
- There are hundreds of data import options available, from CSV files to Google Ads and Analytics data to Salesforce data.
- Output options include multiple chart formats as well as mapping capability. That means designers can create color-coded maps that showcase geographically important data in a format that's much easier to digest than a table or chart could ever be.

Infogram

- Infogram is a fully-featured drag-and-drop visualization tool that allows even non-designers to create effective visualizations of data for marketing reports, infographics, social media posts, maps, dashboards, and more.
- Finished visualizations can be exported into a number of formats: .PNG, .JPG, .GIF, .PDF, and .HTML. Interactive visualizations are also possible, perfect for embedding into websites or apps.
- Infogram also offers a WordPress plugin that makes embedding visualizations even easier for WordPress users.

ChartBlocks

- ChartBlocks claims that data can be imported from “anywhere” using their API, including from live feeds. While they say that importing data from any source can be done in “just a few clicks,” it’s bound to be more complex than other apps that have automated modules or extensions for specific data sources.
- The app allows for extensive customization of the final visualization created, and the chart building wizard helps users pick exactly the right data for their charts before importing the data.
- Designers can create virtually any kind of chart, and the output is responsive—a big advantage for data visualization designers who want to embed charts into websites that are likely to be viewed on a variety of devices.

D3.JS

- D3.js is a JavaScript library for manipulating documents using data.
- D3.js requires at least some JS knowledge, though there are apps out there that allow non- programming users to utilize the library.
- Those apps include NVD3, which offers reusable charts for D3.js; Plotly's Chart Studio, which also allows designers to create WebGL and other charts; and Ember Charts, which also uses the Ember.js framework.

Google Charts

- Google Charts is a powerful, free data visualization tool that is specifically for creating interactive charts for embedding online.
- It works with dynamic data and the outputs are based purely on HTML5 and SVG, so they work in browsers without the use of additional plugins. Data sources include Google Spreadsheets, Google Fusion Tables, Salesforce, and other SQL databases.
- There are a variety of chart types, including maps, scatter charts, column and bar charts, histograms, area charts, pie charts, treemaps, timelines, gauges, and many others. These charts can be customized completely, via simple CSS editing.

FusionCharts

- FusionCharts is another JavaScript-based option for creating web and mobile dashboards. It includes over 150 chart types and 1,000 map types.
- It can integrate with popular JS frameworks (including React, jQuery, React, Ember, and Angular) as well as with server-side programming languages (including PHP, Java, Django, and Ruby on Rails).
- FusionCharts gives ready-to-use code for all of the chart and map variations, making it easier to embed in websites even for those designers with limited programming knowledge.

Chart.js

- Chart.js is a simple but flexible JavaScript charting library. It's open source, provides a good variety of chart types (eight total), and allows for animation and interaction.
- Chart.js uses HTML5 Canvas for output, so it renders charts well across all modern browsers. Charts created are also responsive, so it's great for creating visualizations that are mobile-friendly.

Visualization using Programming

- Python
  - matplotlib
  - seaborn
  - plotly
  - pylab
- R
  - graphics
  - ggplot2

The importance of Context

The importance of context Before you start down the path of data visualization, there are a couple of questions that you should be able to concisely answer: Who is your audience? What do you need them to know or do? This chapter describes the importance of understanding the situational context, including the audience, communication mechanism, and desired tone. A number of concepts are introduced and illustrated via example to help ensure that context is fully understood. Creating a robust understanding of the situational context reduces iterations down the road and sets you on the path to success when it comes to creating visual content.

Exploratory vs. explanatory analysis

1. **What is Exploratory Analysis?**

   Exploratory analysis is the **first stage**, where you **explore the data** to understand what is happening. You are **finding patterns, relationships, or insights**—even if you don't know what exactly you're looking for

Purpose:

- To *discover* something meaningful.

Real-Time Example 1: E-Commerce Company (Flipkart, Amazon, etc.)

Scenario: The analytics team wants to understand why sales dropped last month.

Exploratory Steps:

- Sales by category
- Sales by city
- Sales by device (mobile / desktop)
- Website traffic
- Delivery delays
- Customer complaints
- Return rate

Exploratory Finding (Pearl Found):

- Sales dropped mainly because **mobile app crashes increased during checkout** after an update.

Real-Time Example 2: Hospital Patient Data

Scenario: A hospital wants to understand factors affecting patient wait time.

Exploratory Steps:

- The data team checks:
- Patients per hour
- Doctor availability
- Emergency vs. routine cases
- Lab processing time
- Equipment availability
- They build many visualizations and test multiple hypotheses.

Exploratory Finding:

- Patients wait time increased due to **a shortage of lab technicians during evening hours.**

Explanatory Analysis (Telling the Story)

2. What is Explanatory Analysis?

Once you find the pearls (important insights),explanatory analysis is how you **communicate** those findings to others.

- Explanatory analysis happens *after* exploratory analysis.
  Now you **communicate the insight clearly** to decision-makers.

Purpose:

- To present the **key insight** with a clear message.

Explanatory Output for E-Commerce Example

Story to Management:

- "Sales dropped by 18% because the mobile app crashed during checkout."

Charts shown:

- Chart 1: Sales trend showing the decline
- Chart 2: Spike in mobile app crash rate
- Chart 3: Drop in successful checkout counts
- Clear. Focused. Actionable.

Short Everyday Example (Students Can Relate)

College Attendance Analysis

Exploratory:

- The teacher explores:
- Week-wise attendance
- Subject-wise attendance
- Online vs offline class attendance
- Attendance vs marks
- Creates 15–20 charts.

Insight Found:

- Attendance is lowest on Mondays because most students come late from their hometown.
- **Explanatory Story to Principal:**
- ✔ Attendance trend by weekday (Monday lowest)
- ✔ Comparison of Monday vs other days
- ✔ Recommendations to improve attendance

Only these final charts matter.

Feature	Exploratory	Explanatory
Purpose	Find insights	Explain insights
Audience	Analyst	Managers, students, teachers
Charts	Many	Few
Style	Messy, trial & error	Clean, simple, story-based
Output	Discover pearls	Show pearls

WHO, WHAT, AND HOW

With Real-Time Industry Examples)

When creating explanatory data visualizations, you must be very clear about:

- **WHO** you are communicating to
- **WHAT** you want them to know or do
- **HOW** data will help you make your point

Below are **simple real-world examples** from different industrie

Real-Time Example 1: E-Commerce Company (Amazon/Flipkart)

Scenario:

- Sales of a popular category have dropped suddenly.

1. WHO (Audience)

- Audience: **Head of Marketing & Product Manager**
- What they care about: revenue, customer experience, sales targets
- They don't want 20 charts — they want the **reason** and **action**

2. WHAT (Message / Desired Action)

- "I want the Marketing Head to understand WHY sales dropped and approve a campaign to recover sales."

3. HOW (Using Data to Support the Message)

- Use 2–3 focused charts:
- Sales trend showing the decline
- Spike in mobile app crashes after an upgrade
- Drop in checkouts on mobile devices

Final Story/Visualization:

"Sales dropped because mobile app crashes increased after the latest update. Fixing the bug can recover 18% of lost revenue."

Real-Time Example 2: Hospital Healthcare Dashboard

Scenario:

- Patient wait time has increased in OPD.

1. WHO

- Audience: **Hospital Administrator**
- They care about: patient satisfaction, operational efficiency

2. WHAT

- "I want the admin to approve hiring 2 additional lab technicians."

3. HOW (Visuals)

- Bar chart: Technician availability per shift
- Line chart: Patient wait time increasing in evening
- Correlation chart: Reduced staff → Increased delays

Story:

- "Wait time increased by 30% due to low staff in evening shifts. Hiring 2 lab technicians will fix the issue."

Real-Time Example 3: College Student Performance Dashboard

Scenario:

- Attendance is decreasing across first-year classes.

1. WHO

- Audience: **Head of Department (HOD)**
- They care about: student discipline, academic performance

2. WHAT

- "I want the HOD to approve a new attendance policy."

3. HOW (Visuals)

- Create visuals that show:
- Attendance trend by day of the week
- Subjects with the lowest attendance
- Reason analysis (many students travel from hometown on Mondays)

Story:

- "Attendance drops sharply on Mondays due to hostel return delays. Moving Monday's first lecture to 10 AM can improve attendance by 22%."

Step	Meaning	Question to Ask	Real-Time Impact
WHO	Your audience	Who makes the decision? What do they care about?	Helps you tailor the message
WHAT	Your message	What should they know or do?	Makes your visualization purposeful
HOW	Using data	Which charts best support the message?	Makes your story convincing

REAL-TIME EXAMPLE (DATA VISUALIZATION): E-COMMERCE SALES DROP

Scenario

Flipkart/Amazon notices that **sales dropped by 18% in the last month**.

Management asks the data team to investigate.

1. THE 3-MINUTE STORY (Real-Time Example)

- "Over the past month, our overall sales dropped by 18%. After exploring the data, we found that most of the decline came from mobile users, especially during the checkout stage.
- When we looked deeper, we discovered that the mobile app crash rate increased significantly after the last software update. Because of these crashes, many customers abandoned their purchases. In fact, checkout completion on mobile dropped from 72% to 51% during the same period.
- We also noticed that desktop users did not experience the same issue, which confirms that the drop is specific to the mobile platform.

- Based on this evidence, it's clear that the mobile app update is causing checkout failures. Fixing the crash issue can recover 15–18% of lost sales. Additionally, pushing a small marketing campaign offering free delivery after the fix could help win back affected customers.
- Therefore, we recommend approving an emergency bug-fix release and a short-term reactivation campaign."
- ✔ Clear
  ✔ Only key points
  ✔ No charts shown (yet)
  ✔ Story is data-driven and concise

2. THE BIG IDEA (1-Sentence Real-Time Example)

"Mobile checkout crashes after the latest app update caused a major sales drop, and fixing this issue immediately is critical to recover lost revenue."

This matches all three Big Idea requirements:

- **Unique point of view** → App crash is the root cause
- **What's at stake** → Major revenue loss
- **Complete sentence** → Stands alone clearly

3. REAL-TIME STORYBOARDING (Using Post-it Style Layout)

(This is how a data visualization analyst would plan the communication BEFORE making any charts.)

Storyboard: "Why Sales Dropped – Mobile Checkout Crash"

slide **1 – Lead With Big Idea**
"Sales dropped because the mobile app crashed during checkout."

Slide 2 – Overall Sales Trend
Line chart: Monthly sales → Shows 18% drop

Slide 3 – Mobile vs Desktop Comparison
Bar chart: Mobile sales ↓ / Desktop sales stable

Slide 4 – Mobile Crash Rate Timeline
Line chart: Crash rate spiked after update

Slide 5 – Checkout Completion Rate
Funnel visualization:
72% → 51% after update

Slide 6 – Impact on Revenue
Bar chart: Estimated loss = ₹45 crore

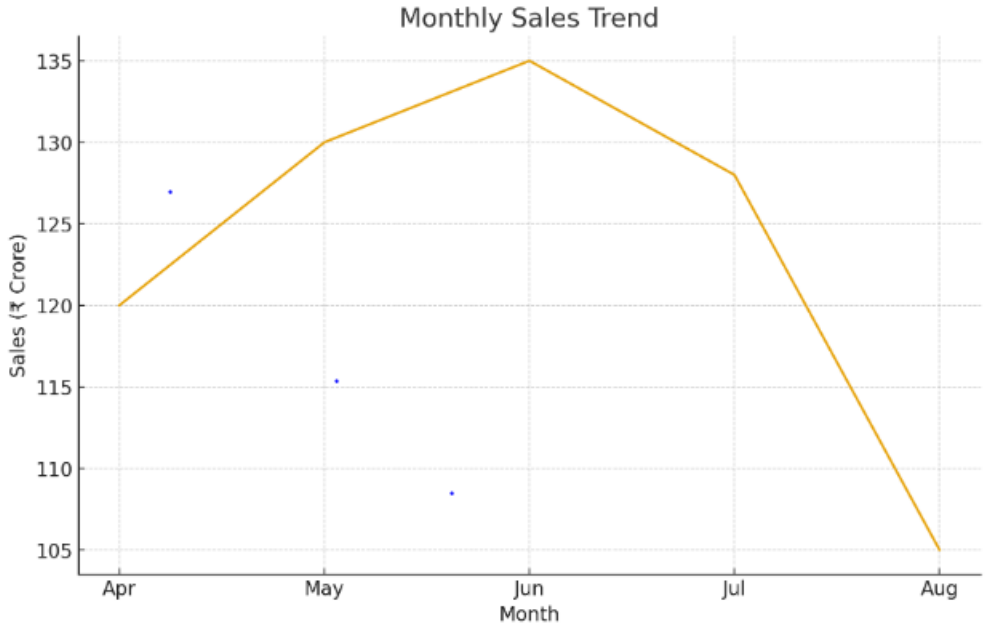
Slide 7 – Recommendation

- Fix crash in next app release
- Offer reactivation coupon
- Monitor checkout success post-release

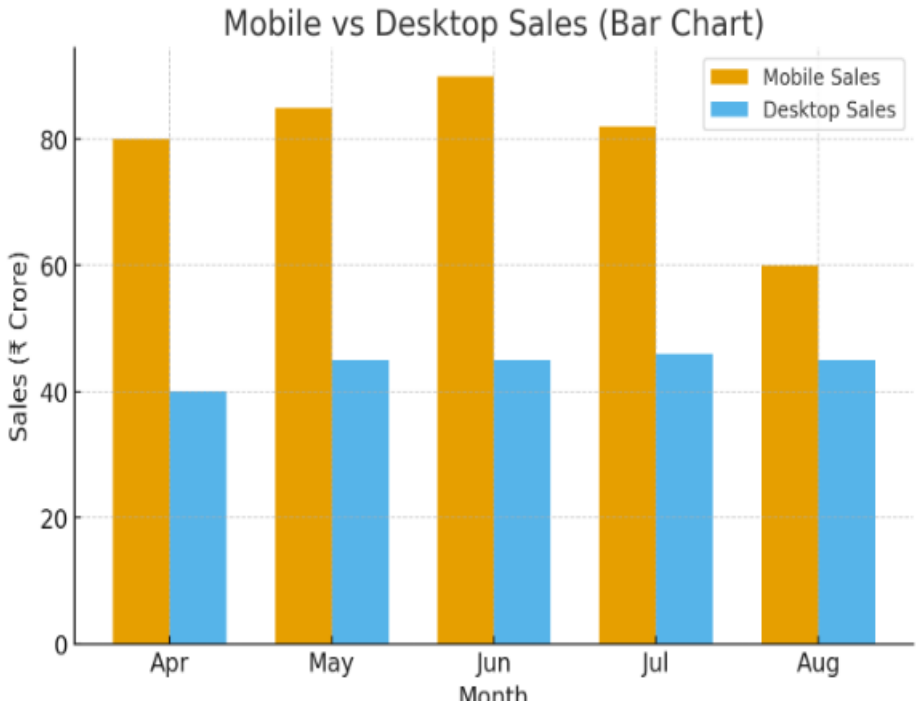
Slide 8 – Big Idea (Reinforced)
"Fixing the crash will recover lost revenue."

✔ Shows how data supports the story
✔ Clear narrative flow
✔ Helps analyst create slides confidently
✔ Ensures audience gets the message quickly

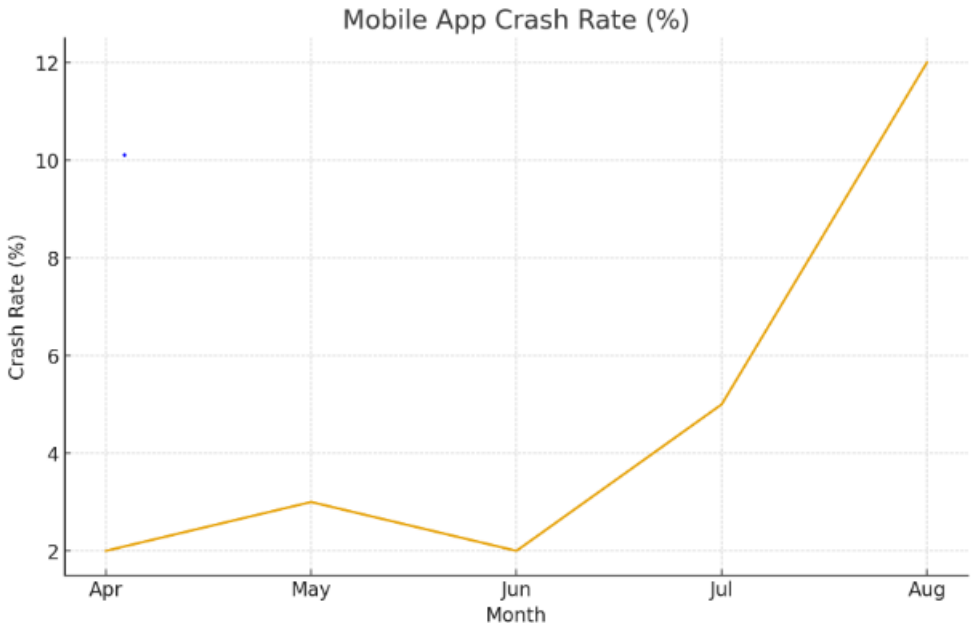
Slide 2 – Overall Sales Trend
Line chart: Monthly sales → Shows 18% drop



Slide 3 – Mobile vs Desktop Comparison
Bar chart: Mobile sales ↓ / Desktop sales stable

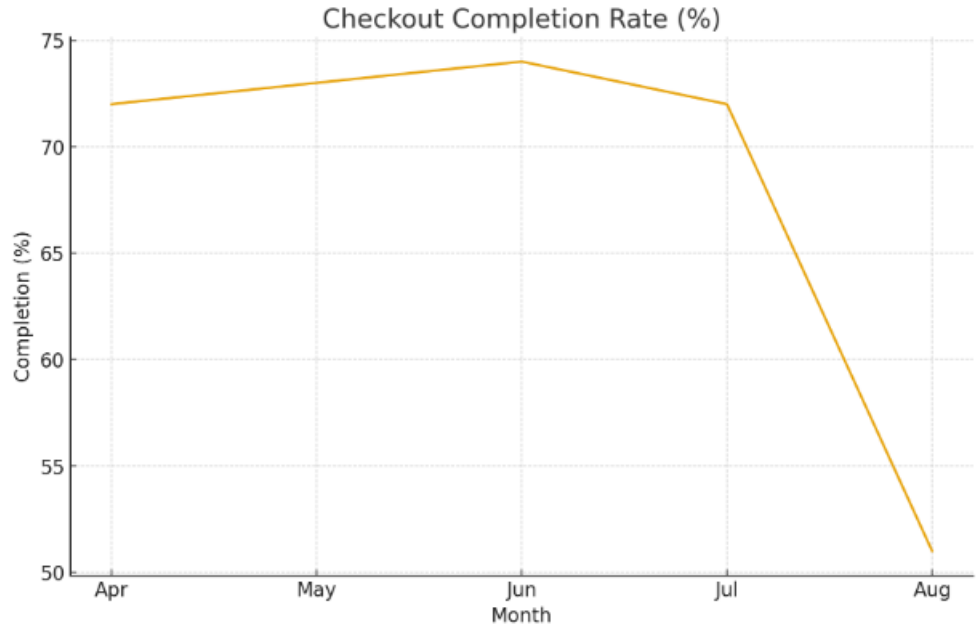


Slide 4 – Mobile Crash Rate Timeline
Line chart: Crash rate spiked after update



Slide 5 – Checkout Completion Rate

Funnel visualization:
72% → 51% after update



CHOOSING AN EFFECTIVE VISUAL

When you only have **one or two important numbers**, you *do not need*

a graph or table. Simple, clear text can communicate the message

better and **faster**.

1.Why use simple text?

- Graphs take space and may confuse the reader when the data is very little.
- A single number can be highlighted easily with big font or bold text.
- Simple text prevents wrong interpretations.

2.When to use simple text

- Use simple text when:
- You have **only 1 or 2 numbers**.
- You want to highlight **one key fact** clearly.
- A graph won't add any extra meaning.

Tables

Tables interact with our verbal system, which means that we read them. When I have a table in front of me, I typically have my index finger out: I'm reading across rows and down columns or I'm comparing values. Tables are great for just that—communicating to a mixed audience whose members will each look for their particular row of interest. If you need to communicate multiple different units of measure, this is also typically easier with a table th

Heavy borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Light borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Minimal borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

FIGURE 2.4 Table borders

HEATMAP

One approach for mixing the detail you can include in a table while also making use of visual cues is via a heatmap. A heatmap is a way to visualize data in tabular format, where in place of (or in addition to) the numbers, you leverage colored cells that convey the relative magnitude of the numbers.

Consider Figure 2.5, which shows some generic data in a table and also a heatmap.

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

LOW-HIGH

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

FIGURE 2.5 Two views of the same data

Graphs

While tables interact with our verbal system, graphs interact with our visual system, which is faster at processing information. This means that a well-designed graph will typically get the information across more quickly than a well-designed table. As I mentioned at the onset of this chapter, there are a plethora of graph types out there. The good news is that a handful of them will meet most of your everyday needs.

The types of graphs I frequently use fall into four categories: points, lines, bars, and area. We will examine these more closely and discuss the subtypes that I find myself using on a regular basis, with specific use cases and examples for each.

Points

Scatterplot

Scatterplots can be useful for showing the relationship between two things, because they allow you to encode data simultaneously on a horizontal *x*-axis and vertical *y*-axis to see whether and what relationship exists. They tend to be more frequently used in scientific fields (and perhaps, because of this, are sometimes viewed as complicated to understand by those less familiar with them). Though infrequent, there are use cases for scatterplots in the business world as well.

For example, let's say that we manage a bus fleet and want to understand the relationship between miles driven and cost per mile. The scatterplot may look something like Figure 2.6.

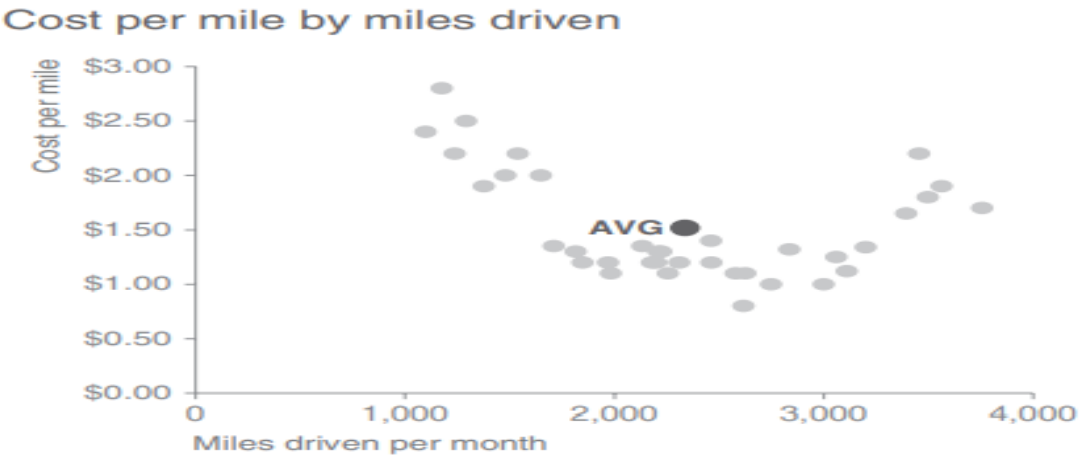


FIGURE 2.6 Scatterplot

Lines

Line graphs are most commonly used to plot continuous data. Because the points are physically connected via the line, it implies a connection between the points that may not make sense for categorical data (a set of data that is sorted or divided into different categories). Often, our continuous data is in some unit of time: days, months, quarters, or years.

Within the line graph category, there are two types of charts that I frequently find myself using: the standard line graph and the slopegraph.

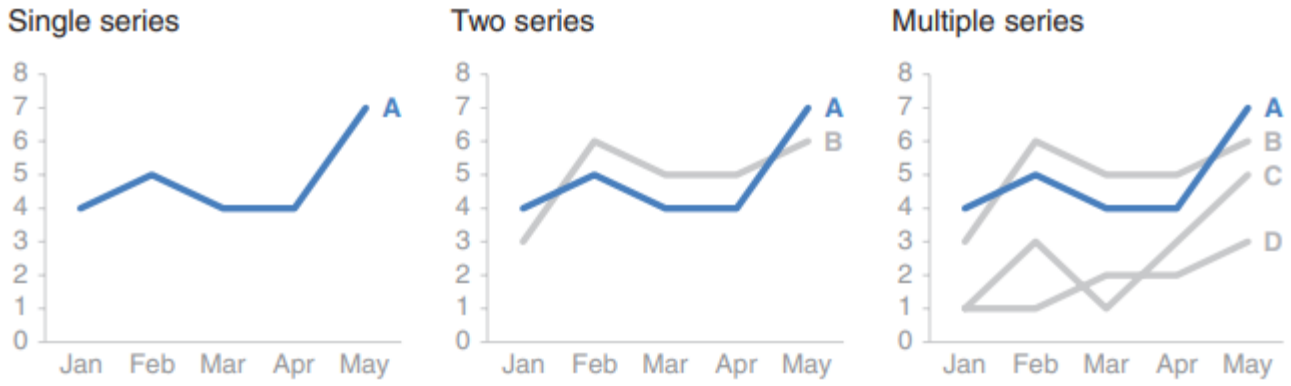


FIGURE 2.8 Line graphs

SLOPEGRAPH

Slopegraphs can be useful when you have two time periods or points of comparison and want to quickly show relative increases and decreases or differences across various categories between the two data points.

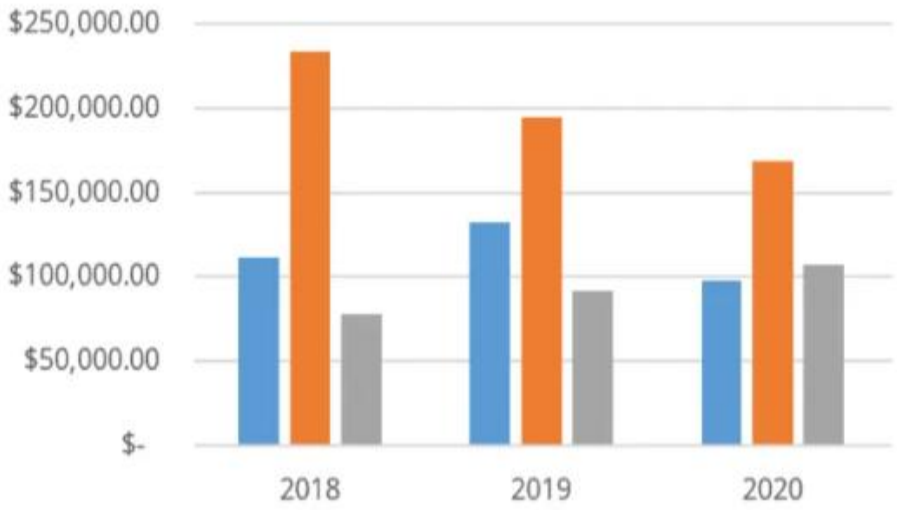
The best way to explain the value of and use case for slopegraphs is through a specific example. Imagine that you are analyzing and communicating data from a recent employee feedback survey. To show the relative change in survey categories from 2014 to 2015, the slopegraph might look something like Figure 2.10.

Slopegraphs pack in a lot of information. In addition to the absolute values (the points), the lines that connect them give you the visual increase or decrease in rate of change (via the slope or direction) without ever having to explain that's what they are doing, or what exactly a "rate of change" is—rather, it's intuitive.

FIGURE 2.10 Slopegraph

Bars

- A **Bar Chart** (or Bar Graph) is a graphical representation of data using **rectangular bars** to compare different categories. The **length or height of each bar** represents the value of that category.



Waterfall chart

The waterfall chart can be used to pull apart the pieces of a stacked bar chart to focus on one at a time, or to show a starting point, increases and decreases, and the resulting ending point.

The best way to illustrate the use case for a waterfall chart is through a specific example. Imagine that you are an HR business partner and want to understand and communicate how employee headcount has changed over the past year for the client group you support.

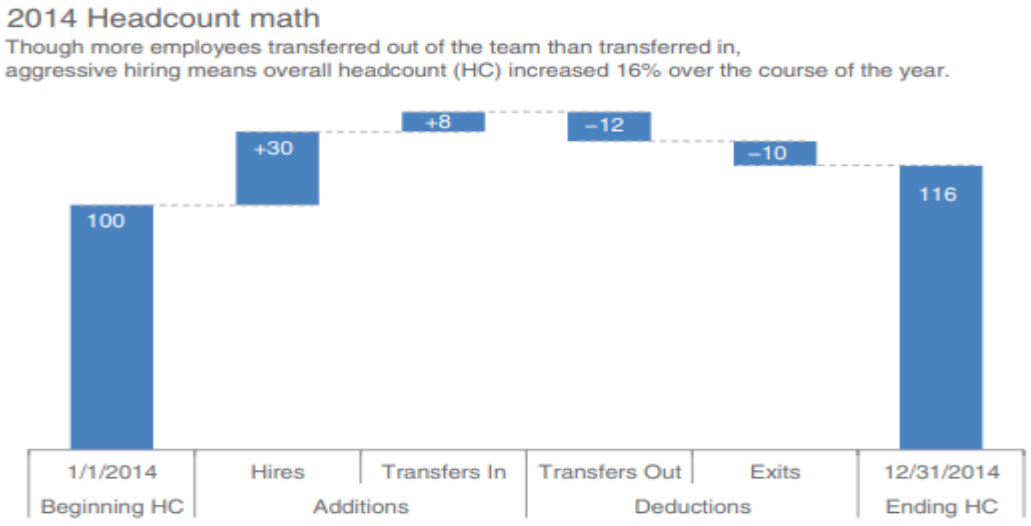


FIGURE 2.17 Waterfall chart

Horizontal bar chart

If I had to pick a single go-to graph for categorical data, it would be the horizontal bar chart, which flips the vertical version on its side. Why? Because it is *extremely easy to read*. The horizontal bar chart is especially useful if your category names are long, as the text is written from left to right, as most audiences read, making your graph legible for your audience. Also, because of the way we typically process information—starting at top left and making z's with our eyes across the screen or page—the structure of the horizontal bar chart is such that our eyes hit the category names before the actual data. This means by the time we get to the data, we already know what it represents (instead of the darting back and forth our eyes do between the data and category names with vertical bar charts).

Like the vertical bar chart, the horizontal bar chart can be single series, two series, or multiple series (Figure 2.18).

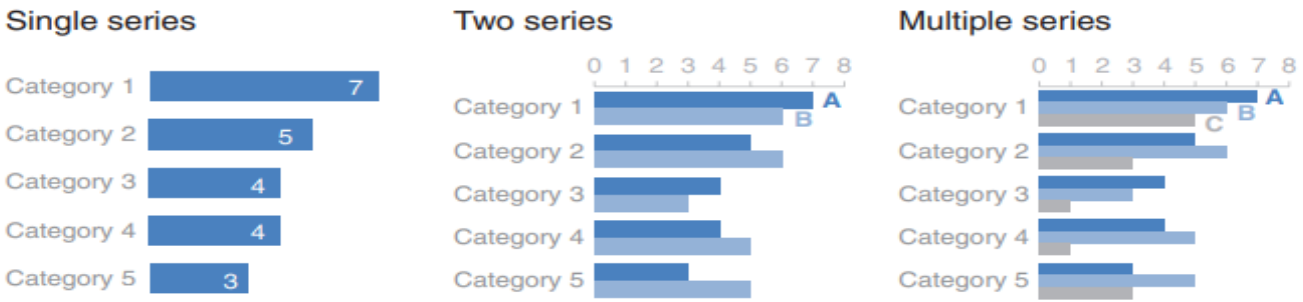


FIGURE 2.18 Horizontal bar charts

Area

I avoid most area graphs. Humans' eyes don't do a great job of attributing quantitative value to two-dimensional space, which can render area graphs harder to read than some of the other types of visual displays we've discussed. For this reason, I typically avoid them, with one exception—when I need to visualize numbers of vastly different magnitudes. The second dimension you get using a square for this (which has both height and width, compared to a bar that has only height *or* width) allows this to be done in a more compact way than possible with a single dimension, as shown in Figure 2.20.

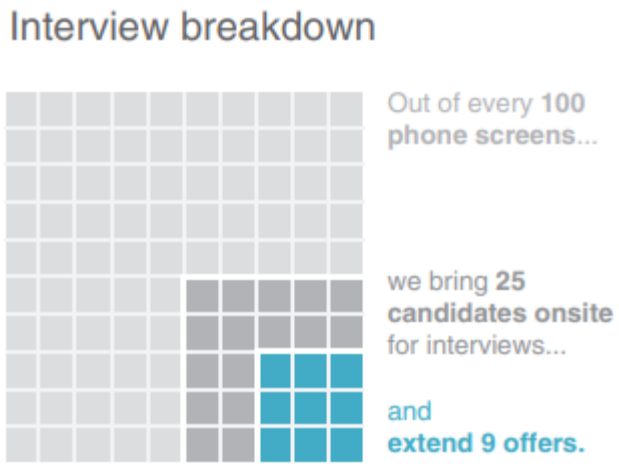


FIGURE 2.20 Square area graph

Pie charts

I have a well-documented disdain for pie charts. In short, they are evil. To understand how I arrived at this conclusion, let's look at an example.

> The pie chart shown in Figure 2.21 (based on a real example) shows market share across four suppliers: A, B, C, and D. If I asked you to make a simple observation—which supplier is the largest based on this visual—what would you say?

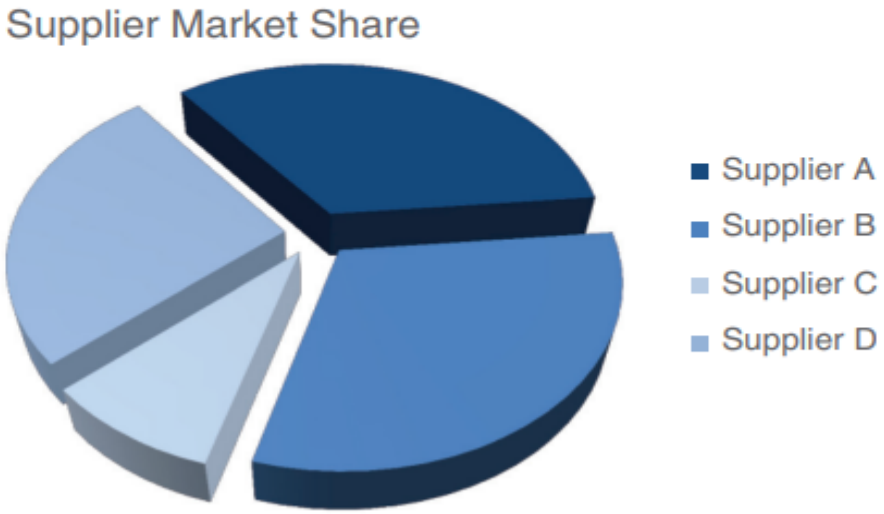


FIGURE 2.21 Pie chart

Donut chart

The donut chart

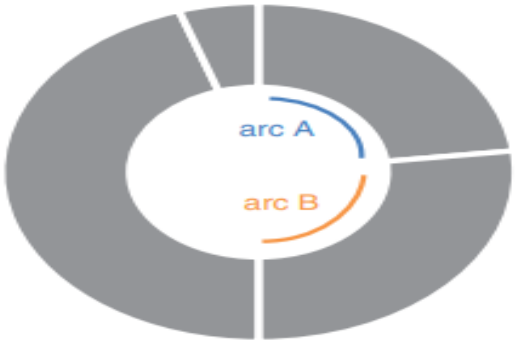


FIGURE 2.24 Donut chart

With pies, we are asking our audience to compare angles and areas. With a donut chart, we are asking our audience to compare one arc length to another arc length (for example, in Figure 2.24, the length of *arc A* compared to *arc B*). How confident do you feel in your eyes' ability to ascribe quantitative value to an arc length?

Not very? That's what I thought. Don't use donut charts.

Never use 3D

One of the golden rules of data visualization goes like this: never use 3D. Repeat after me: never use 3D. The only exception is if you are actually *plotting a third dimension* (and even then, things get really tricky really quickly, so take care when doing this)—and you should never use 3D to plot a single dimension. As we saw in the pie chart example previously, 3D skews our numbers, making them difficult or impossible to interpret or compare.

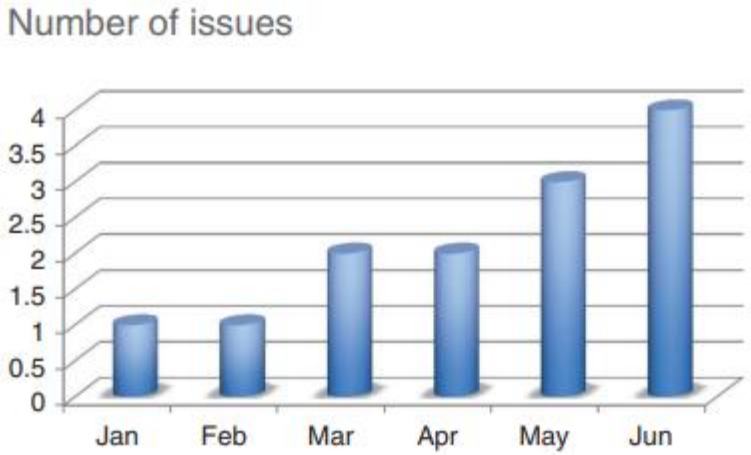


FIGURE 2.25 3D column chart

Secondary y-axis: generally not a good idea

Sometimes it's useful to be able to plot data that is in entirely different units against the same *x*-axis. This often gives rise to the secondary *y*-axis: another vertical axis on the right-hand side of the graph. Consider the example shown in Figure 2.26.

FIGURE 2.26 Secondary *y*-axis

When interpreting Figure 2.26, it takes some time and reading to understand which data should be read against which axis. Because of this, you should avoid the use of a secondary or right-hand *y*-axis. Instead, think about whether one of the following approaches will meet your needs:

1. Don't show the second *y*-axis. Instead, label the data points that belong on this axis directly.
2. Pull the graphs apart vertically and have a separate *y*-axis for each (both along the left) but leverage the same *x*-axis across both.

Reference

- C. B. Jones, *Communicating Data with Tableau: Designing, Developing, and Delivering Your Data to the Masses*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2014.
- W. Playfair, *The Commercial and Political Atlas: Representing, by Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England During the Whole of the Eighteenth Century*. London, UK: J. Debrett, 1786.
- J. Priestley, *A Description of a Chart of Biography*. Warrington, UK: Printed for the author, 1765.
- C. J. Minard, "Carte figurative et approximative des températures pendant l'année 1812" (map of Napoleon's Russian campaign). Paris, France, 1869.

TEXT BOOKS

- Cole Nussbaumer Knaflic, Storytelling with data, Wiley 13th oct 2015.

REFERENCE WEBSITE:

- Data Analysis and Visualization Foundations | Coursera.

UNIT-II

CLUTTER IS YOUR ENEMY!

CLUTTER IS YOUR ENEMY!

Clutter in Data Visualization (Simple Notes + Tables + Examples)

What Is Clutter?

Clutter = Anything in a chart that takes space but does NOT help understanding.

Examples of clutter:

- Extra gridlines
- Unnecessary borders
- Too many colors
- 3D effects
- Shadow effects
- Long data labels
- Decorative icons
- Extra text

Why clutter is bad?

Problem	Why it hurts
Looks complicated	Audience feels overwhelmed
Hard to read	Brain must filter unnecessary elements
Slows understanding	More time to find the important part
Distracts from the message	Focus moves away from insights

Gestalt Principles for Reducing Clutter

These principles help readers quickly understand patterns in your visuals.

We will cover **six principles**:

1. Proximity
2. Similarity
3. Enclosure
4. Closure
5. Continuity
6. Connection

Proximity

Objects that are *close together* are seen as a group.

We tend to think of objects that are physically close together as belonging to part of a group. The proximity principle is demonstrated in Figure 3.1: you naturally see the dots as three distinct groups because of their relative proximity to each other.

FIGURE 3.1 Gestalt principle of proximity

We can leverage this way that people see in table design. In Figure 3.2, simply by virtue of differentiating the spacing between the dots, your eyes are drawn either down the columns in the first case or across the rows in the second case.

FIGURE 3.2 You see columns and rows, simply due to dot spacing

Gestalt Principle of Visual Perception – Proximity

Explanation:

- The principle of proximity states that objects placed close to each other are perceived as belonging to the same group. Our brain naturally groups nearby elements without needing lines, colors, or labels.
- When elements are far apart, we see them as separate groups. When they are close together, we see them as related.

Real-Time (Real-World) Examples of Proximity

1. Students Sitting in a ClassroomStudents sitting close together are seen as one group (same bench or row).
2. Students sitting far apart are seen as different groups.

Similarity

Objects that *look similar* (color, shape, size) are perceived as related.

Objects that are of similar color, shape, size, or orientation are perceived as related or belonging to part of a group. In Figure 3.3, you naturally associate the blue circles together on the left or the grey squares together on the right.

FIGURE 3.3 Gestalt principle of similarity

This can be leveraged in tables to help draw our audience's eyes in the direction we want them to focus. In Figure 3.4, the similarity of color is a cue for our eyes to read across the rows (rather than down the columns). This eliminates the need for additional elements such as borders to help direct our attention.

FIGURE 3.4 You see rows due to similarity of color

Gestalt Principle of Visual Perception – Similarity

- The principle of similarity states that objects that look alike are perceived as belonging to the same group. Similarity can be based on:

Color

Shape

Size

Orientation

- Even if objects are far apart, visual similarity makes our brain group them together.

Real-Time (Real-World) Examples of

1. School or College UniformsStudents wearing the same uniform are perceived as belonging to the same institution.
2. Different uniforms indicate different schools or groups.
3. Traffic Signals and Road SignsWarning signs are usually triangular and red.
4. Information signs are rectangular and blue.
5. Similar color and shape help drivers quickly recognize the type of sign.

Enclosure

Using a box/shaded area groups information together.

We think of objects that are physically enclosed together as belonging to part of a group. It doesn't take a very strong enclosure to do this: light background shading is often enough, as demonstrated in Figure 3.5.

FIGURE 3.5 Gestalt principle of enclosure

One way we can leverage the enclosure principle is to draw a visual distinction within our data, as is done in the graph in Figure 3.6.

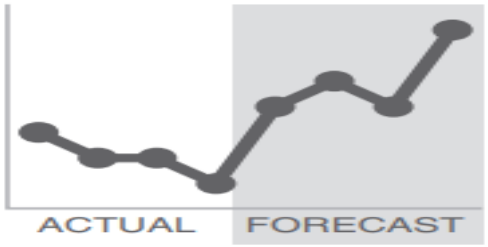


FIGURE 3.6 The shaded area separates the forecast from actual data

Gestalt Principle of Visual Perception – Enclosure

- The principle of enclosure states that objects enclosed within a common boundary are perceived as belonging to the same group. The enclosure does not need to be strong—even light background shading, a box, or an outline is enough for our brain to group elements together.

Real-Time (Real-World) Examples of Enclosure

1. 1.Forms on WebsitesLogin details (Username and Password) are placed inside a box or shaded area.
2. Users immediately understand these fields belong together.
3. 2. Mobile App CardsInformation shown inside a card layout is perceived as one unit.
4. Example: Contact cards, news cards, product cards.

Closure

Our brain fills in missing shapes automatically.

The closure concept says that people like things to be simple and to fit in the constructs that are already in our heads. Because of this, people tend to perceive a set of individual elements as a single, recognizable shape when they can—when parts of a whole are missing, our eyes fill in the gap. For example, the elements in Figure 3.7 will tend to be perceived as a circle first and only after that as individual elements.

FIGURE 3.7 Gestalt principle of closure

It is common for graphing applications (for example, Excel) to have default settings that include elements like chart borders and background shading. The closure principle tells us that these are unnecessary—we can remove them and our graph still appears as a cohesive entity. Bonus: when we take away those unnecessary elements, our data stands out more, as shown in Figure 3.8.

FIGURE 3.8 The graph still appears complete without the border and background shading

Gestalt Principle of Visual Perception – Closure

- The principle of closure states that people naturally perceive incomplete shapes as complete. When part of a shape is missing, the human brain automatically fills in the gaps to create a familiar and simple form.
- We prefer simplicity and recognizable patterns, so we see the whole object first before noticing the missing parts.

Real-Time (Real-World) Examples of Closure

1. Company LogosLogos like WWF (panda) or Olympic rings are not fully closed.
2. Our brain completes the shapes automatically.
3. 2. Dashed Road LinesBroken white lines on highways are perceived as a continuous line.
4. Even though parts are missing, drivers see a complete lane.

Continuity

Our eyes follow smooth paths.

Example:

In line charts, a single clean line is enough.

The principle of continuity is similar to closure: when looking at objects, our eyes seek the smoothest path and naturally create continuity in what we see even where it may not explicitly exist. By way of example, in Figure 3.9, if I take the objects (1) and pull them apart, most people will expect to see what is shown next (2), whereas it could as easily be what is shown after that (3).

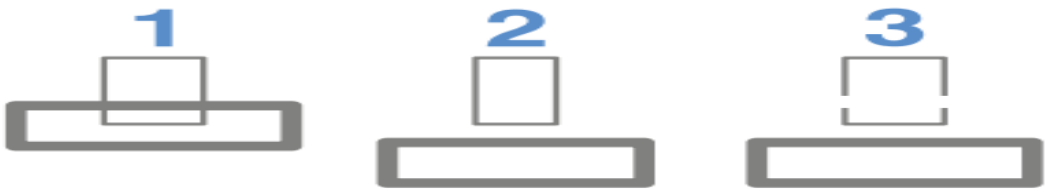


FIGURE 3.9 Gestalt principle of continuity

In the application of this principle, I've removed the vertical *y*-axis line from the graph in Figure 3.10 altogether. Your eyes actually still see that the bars are lined up at the same point because of the consistent white space (the smoothest path) between the labels on the left and the data on the right. As we saw with the closure principle in application, stripping away unnecessary elements allows our data to stand out more.

FIGURE 3.10 Graph with *y*-axis line removed

Gestalt Principle of Visual Perception – Continuity

- The principle of continuity states that our eyes naturally follow the smoothest and most continuous path when viewing visual elements. Even if a line or path is not fully drawn, the brain assumes continuity rather than sudden breaks or changes.
- This principle is similar to closure, but continuity focuses on direction, alignment, and flow.

Real-Time (Real-World) Examples of Continuity

1. Roads and Railway TracksRoads and tracks appear continuous even when partially hidden by trees or bridges.
2. Our brain expects the path to continue smoothly.
3. 2. Line GraphsPoints connected by lines are seen as a continuous trend.
4. Even without axis lines, data appears aligned due to consistent spacing

Connection

When items are visually connected, the brain groups them.

The final Gestalt principle we'll focus on is connection. We tend to think of objects that are physically connected as part of a group. The connective property typically has a stronger associative value than similar color, size, or shape. Note when looking at Figure 3.11, your eyes probably pair the shapes connected by lines (rather than similar color, size, or shape): that's the connection principle in action. The connective property *isn't* typically stronger than enclosure, but you can impact this relationship through thickness and darkness of lines to create the desired visual hierarchy (we'll talk more about visual hierarchy when we discuss preattentive attributes in Chapter 4).

FIGURE 3.11 Gestalt principle of connection

One way that we frequently leverage the connection principle is in line graphs, to help our eyes see order in the data, as shown in Figure 3.12.

FIGURE 3.12 Lines connect the dots

As you have learned from this brief overview, the Gestalt principles help us understand how people see, which we can use to identify unnecessary elements and ease the processing of our visual communications. We aren't done with them yet. At the end of this chapter, we'll discuss how we can apply some of these principles to a real-world example.

Gestalt Principle of Visual Perception – Connection

- The principle of connection states that objects that are physically connected by lines, curves, or paths are perceived as belonging to the same group. This connection is often stronger than similarity in color, shape, or size.
- In Figure 3.11, even if shapes differ in color or size, our eyes first group the ones that are joined by lines. That is the connection principle at work.Although enclosure is usually stronger than connection, designers can make connections more prominent by using thicker or darker lines, creating a clear visual hierarchy.

Real-Time (Real-World) Examples of Connection

1. Line GraphsData points connected by lines are seen as part of the same data series.
2. The connection helps us easily see trends and order.
3. 2. FlowchartsSteps connected with arrows show the sequence of actions.
4. The connection explains how one step leads to another.

Lack of Visual Order in Data Visualization

Meaning

A chart lacks *visual order* when:

- Information is scattered randomly
- Graph elements are not aligned
- There is no clear hierarchy
- Viewer's eyes don't know where to look first

This creates **cognitive load** and makes the message harder to understand.

Why It Happens

- Inconsistent spacing
- Misaligned text, bars, labels
- Too many fonts/colors
- No grouping or logical structure
- Irregular scale or axis intervals

Alignment

Meaning

- Alignment means placing elements so that:
- Text lines up
- Bars or points share the same starting baseline
- Titles, labels, and legends are visually connected
- Tables and charts follow a clean grid

Good alignment gives the viewer a **smooth visual path**.

Why Alignment Matters

- Reduces eye movement
- Makes comparison effortless
- Creates professional, clean visuals
- Improves perceived reliability of data

Non-Strategic Use of Contrast — Explained Simply

What is Contrast?

Contrast is when you make one object stand out by making it different from others — using:

- Color
- Size
- Bold text
- Shape
- Thickness
- Highlighting

Decluttering: Step-by-Step

What is "Decluttering: Step-by-Step" in Data Visualization?

1. Decluttering means removing all unnecessary visual elements from a chart so the audience can clearly understand the main message.
2. The step-by-step approach gives you a systematic method to clean a messy chart, reduce cognitive load, and highlight the insight that truly matters.

1.Remove Non-Essential Chartjunk

These include:

- Heavy gridlines
- Background shading
- Patterns
- Borders
- Unnecessary icons
- Drop shadows
- 3D effects

- **Example**

 Bad line chart → Thick gridlines + 3D + background color

✔ Good chart → Clean white background + faint gridlines or none

2.Reduce Text Overload

Check for:

- Long axis titles
- Repeated labels
- Long sentence-like legends
- Redundant decimals

Example

- "Revenue in Indian Rupees (INR) for the period 2015–2024“

✔ “Revenue (₹ Cr)”

Remove:

- 2 decimal places → keep 1 or 0
- Legend if labels can go directly on chart

3.Group Related Information

Apply **Gestalt principles**:

- Proximity
- Similarity
- Enclosure
- Alignment
- Connection

Example

Put:

- Title → top
- Chart → center
- Key message → near the data point it refers to

4.Apply Consistent Formatting

Fix:

- Colors
- Font sizes
- Bar or line thickness
- Label alignment
- Spacing

Example

 Every bar = different color

✔ All bars = one color + one highlight color

This creates visual order.

5.Use Contrast Strategically (Only Where Needed)

This is the most important step.

Highlight:

- One bar
- One line
- One value
- One region
- One trend

Example

- ✓ All bars = grey
- ✓ Important bar = blue
- ✓ Message becomes instantly clear.

6.Add a Clear Message (Your "Big Idea")

A clean chart still fails **if it doesn't communicate the point**.

Add:

- Short title with conclusion
- Optional annotation near the key data point

Example

Title:
"Mobile sales dropped sharply in August."

- ✓ Annotation:
  "↓ 27% compared to July"
- ✓ This makes your visual **explanatory**, not exploratory.

Focus Your Audience's Attention — Student Notes

When we create a chart, slide, or dashboard, our audience does **not** see everything at once. Their brain picks up certain things faster than others.

If we understand **how people see**, we can **guide their attention** exactly where we want.

1. You Don't See With Your Eyes — You See With Your Brain

How seeing works (simple explanation):

1. Light hits an object
2. Eyes capture it
3. Brain interprets it — this is actual "seeing"
4. Your brain decides *what to notice first*, not your eyes.

Quick Lesson on Memory (Very Important for Data Visualization

When people look at a visual, three types of memory work together:

A. Iconic Memory (0.1 seconds — automatic)

- Works instantly
- Helps detect differences in what we see
- Reacts to things like **color, size, boldness, movement, etc.**

Example:

You see a red dot among 50 grey dots — you notice the red dot instantly **without thinking**.

This is iconic memory picking up **preattentive attributes**.

B. Short-Term Memory (very limited: 4 items only!)

- Can hold only **4 chunks** of visual information
- If you show too many colors / lines / legends, people get confused

Example of overload:

A line chart with **10 colored lines** + legend → audience works too hard → message lost.

Fix:

Label lines directly instead of using a legend → reduces cognitive load.

C. Long-Term Memory (stored for life)

- Helps with **pattern recognition**
- Connections form faster when visuals + words are combined

Example:

Seeing the Eiffel Tower immediately reminds you of "Paris".

Using visuals helps your message stick longer.

3.Preattentive Attributes — The Most Powerful Tool in Visual Design

- Preattentive attributes are things the brain notices **automatically** in less than a second.

- Common Preattentive Attributes

Type	Examples
Color	Red among grey
Size	One big circle among small circles
Position	Higher/lower on the page
Shape	Triangle among circles
Intensity	Dark vs light
Orientation	Tilted line among vertical lines
Enclosure	Highlighted with a box
Added marks	Underlining, symbols

Preattentive Attributes in TEXT — Real-Time Examples

1. Preattentive Attributes in TEXT — Real-Time Examples

Example 1: Customer Feedback Summary

Context: You want management to quickly understand customer compliments.

Before (no preattentive attributes — hard to scan):

- Great products: "These products are clearly the best in class."
- Replacement parts are shipped when needed: "You sent me gaskets without me having to ask."
- Problems resolved promptly: "Bev in billing fixed my issue fast."
- Customer service exceeds expectations: "Account manager called after hours."

After (using bold/color hierarchy):

What Customers Love About Us

Great products – "Best in class!"

Fast replacement parts – "Gaskets arrived before I even asked."

Quick issue resolution – "Billing problem solved immediately."

Outstanding customer service – "Called even after business hours."

Why this works:
Bold headings act as **visual anchors**, letting the reader scan instantly.

Storytelling with Iterative Visuals — Real-Time Examples

Example 2: Repeating Same Graph with Different Emphasis

Scenario: Manufacturing Defects Analysis

1.Slide 1 → Show all defects (grey)

Just to familiarize the audience.

2.Slide 2 → Highlight only "Overheating" in blue

Message: This is the biggest defect.

3.Slide 3 → Add text bubbles explaining why overheating happens

Message: Root causes.

- This stepwise reveal keeps audience focused and helps the story flow.

Lessons in Storytelling

1. Why Storytelling Matters (Real-Time Example)

Example: Company Sales Report

If a manager says:

- "Sales dropped by 8% last quarter."
- It gives information, but **no urgency**.
- If the same data is told as a story:

"Last quarter, our sales dropped 8% for the first time in three years. If this continues, we will miss our annual target by ₹12 crores."

- **Impact:**
  Story creates **context, emotion, and urgency**.

2. Red Riding Hood → Business Parallel

Story ElementBusiness / Data ExampleRed Riding HoodBusiness ManagerGrandma sickFalling performanceWolfMarket competitionWoodsmanData-driven solutionHappy endingBusiness recovery

- **Lesson:**
  Every data story has:
- A problem
- A conflict
- A solution

Three-Act Structure

3. Three-Act Structure with Real-Time Example

Example: Customer Churn Analysis

Act 1 – Beginning (Setup)

- "Over the last 6 months, customer churn has increased from 4% to 7%."
- (Problem introduced)

Act 2 – Middle (Conflict)

- "Data shows churn is highest among customers who wait more than 3 days for support responses."
- (Tension builds)

Act 3 – End (Resolution)

- "By reducing response time to under 24 hours, we can lower churn back to 4%."
- (Call to action)

4. Storytelling vs Bullet Points (Corporate Example)

Bullet-Point Approach

- Customer satisfaction: 68%
- Response time: 72 hours
- Complaints increased

Storytelling Approach

- "As response time increased to 72 hours, customer satisfaction dropped to 68%, leading to a rise in complaints."

 Why it works:
Connects cause → effect → consequence.

5. Robert McKee's Balance → Imbalance (Example)

Example: E-commerce Website

Balance: Website conversion rate = 4.5%

Imbalance: Conversion dropped to 2.9% after website redesign

Conflict: Customers abandon cart on payment page

Resolution: Simplify checkout → restore conversions

 Story explains change, not just numbers.

6. Kurt Vonnegut's Rules in Practice

Example: Dashboard Design

 Bad:

"This visualization demonstrates multi-dimensional trends across quarterly metrics…"

 Good:

- "Sales fell sharply in Q3 due to delayed deliveries."

 Simple, clear, human language wins.

7. Constructing a Data Story (End-to-End Example)

Scenario: Low Employee Productivity

Beginning (Context)

- "Employee productivity has fallen 10% in the last year."

Middle (Evidence)

- Absenteeism increased by 15%
- Overtime costs up by ₹40 lakhs
- Teams with flexible hours perform better

End (Action)

- "Introduce flexible work hours starting next quarter."

8. Audience as the Main Character (Example)

Example: Marketing Team Presentation

- Wrong focus:

"We built a churn model using Python."

- Right focus:
  - "This model helps **you** identify high-risk customers before they leave."
- **Audience = Hero of the story**

9. Bing, Bang, Bongo (Real-Time Example)

Example: Quarterly Business Review

Bing (Executive Summary):

- "Sales declined, costs increased, and we need pricing changes."

Bang (Details):

- Sales data by region
- Cost breakdown
- Pricing comparison with competitors

Bongo (Conclusion):

- "To recover profits, we recommend a 5% price adjustment."

10. Reverse Storyboarding (Example)

After creating a dashboard:

- Slide 1: Problem overview
- Slide 2: Root cause
- Slide 3: Financial impact
- Slide 4: Recommendation

 If flow feels confusing → reorder slides.

11. Fresh Perspective (Real-World Example)

- Before sending a report to leadership:
- Give it to a colleague from another department

Ask:

- "What is the main message?"
- "What action is expected?"

 If they struggle → story is unclear.

Reference

- C. B. Jones, *Communicating Data with Tableau: Designing, Developing, and Delivering Your Data to the Masses*, 1st ed. Sebastopol, CA, USA: O
- R. McKee, *Story: Substance, Structure, Style, and the Principles of Screenwriting*. New York, NY, USA: HarperCollins, 1997.'Reilly Media, 2014, ch. 3
- K. Vonnegut, "Shapes of Stories," lecture recorded at University of Visual and Performing Arts, Cleveland, OH, USA, 2000. [Online]. Available: Various YouTube archives.-4.

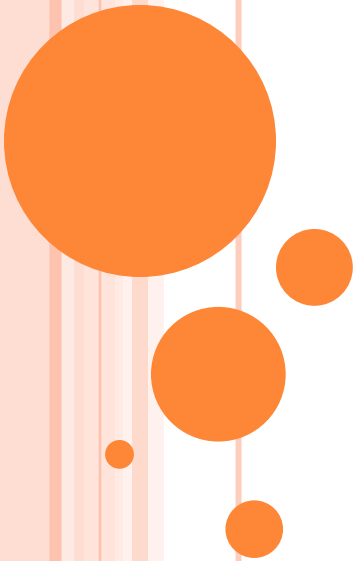
TEXT BOOKS

- Cole Nussbaumer Knaflic, Storytelling with data, Wiley 13th oct 2015.

REFERENCE WEBSITE:

- Data Analysis and Visualization Foundations | Coursera.

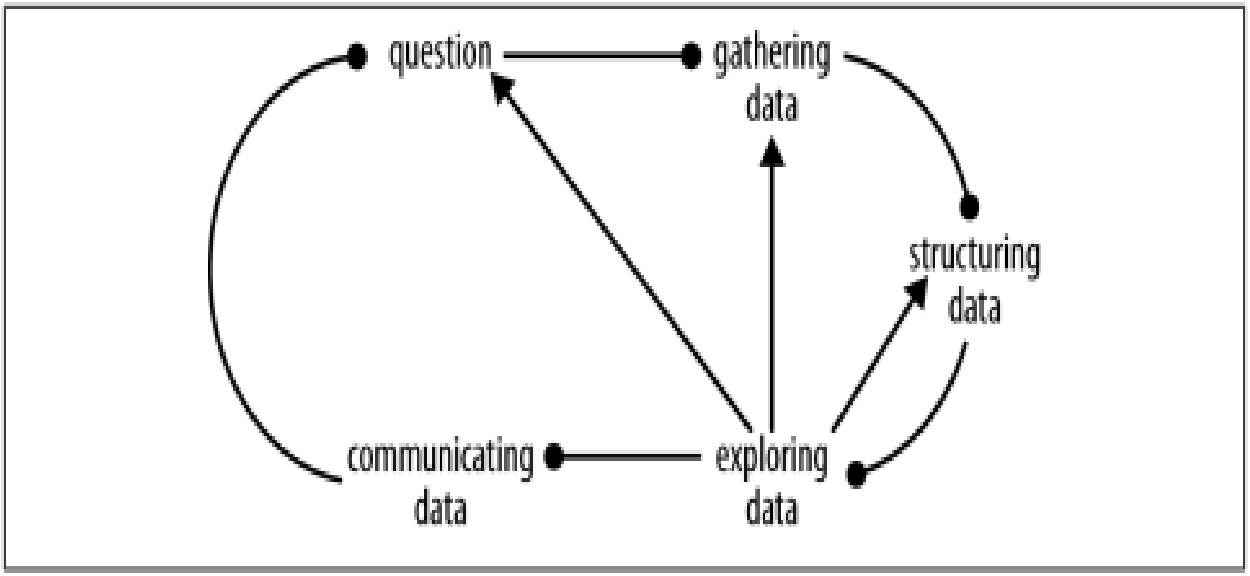
UNIT-III



Communicating data

- Communicating data through visualization involves transforming complex data into simple graphical representations like charts, graphs, and maps to make insights accessible, identify patterns, and tell a compelling story

A Step in the Process



- **Step-by-step explanation of the diagram**

Question

- This is the starting point.
- You clearly define *what you want to know* or *what problem you want to solve*.
- Example: *Why are sales declining in the last quarter?*

Gathering Data

- Data relevant to the question is collected.
- Sources may include databases, surveys, sensors, logs, Excel files, APIs, etc.
- If data is insufficient, you may return to refine the question.

Structuring Data

- Raw data is cleaned and organized.
- Activities include removing duplicates, handling missing values, formatting columns, and creating tables.
- This step prepares data for proper analysis.

Exploring Data

- Initial analysis to understand patterns, trends, and outliers.
- Uses descriptive statistics and basic visualizations (bar charts, line graphs).
- Insights gained here may send you back to:
  - **Gather more data**
  - **Restructure data**
  - **Refine the question**

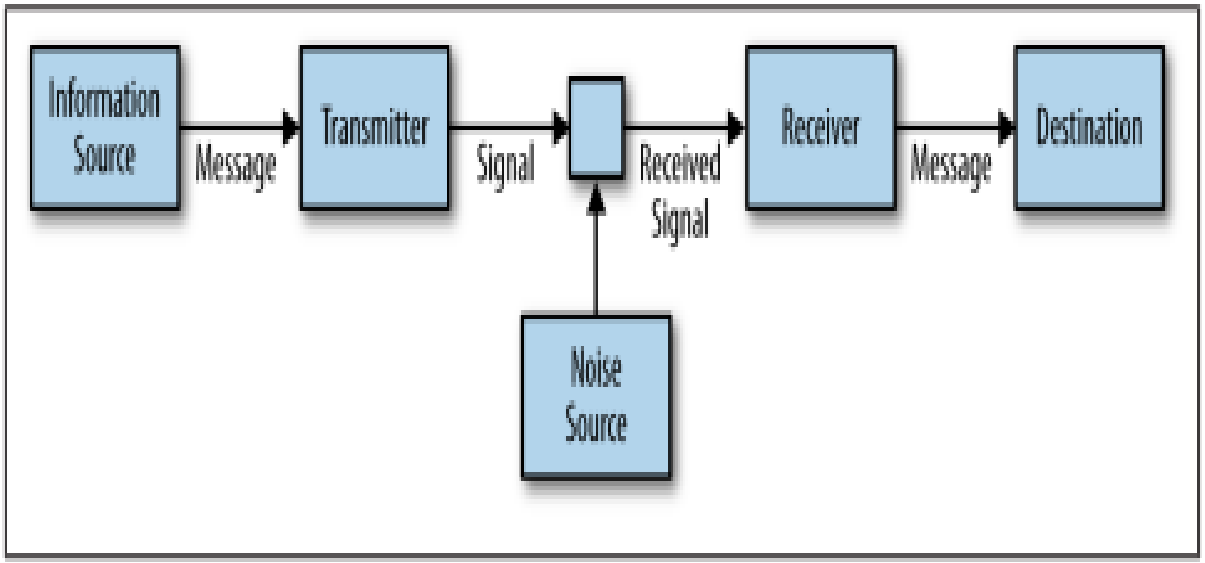
Communicating Data

- Final insights are presented using dashboards, charts, reports, or stories.
- Focus is on clarity and decision-making, not just numbers.
- Example: presenting findings to management using Tableau or Power BI.

Feedback Loop (Iteration)

- Communication often leads to **new questions**.
- The process loops back to the **question** stage, starting the cycle again.

A Model of Communication



- **Step-by-step explanation of the diagram**

Information Source

- Generates the original information or data.
- Example: a person speaking, a computer generating data, a sensor collecting readings.

Message

- The information produced by the source in a usable form (text, voice, image, data).

Transmitter

- Converts the message into a **signal** suitable for transmission.
- This may involve encoding, modulation, or signal processing.
- Example: a mobile phone converting voice into electrical signals.

Signal

- The encoded form of the message that travels through the communication channel.
- Can be electrical, optical, or electromagnetic.

Channel

- The medium through which the signal travels.
- Example: air, optical fiber, copper wire.
- *(Implied in the diagram between transmitter and receiver)*

Noise Source

- Any unwanted disturbance that affects the signal during transmission.
- Examples: electromagnetic interference, thermal noise, static.
- Noise can distort or corrupt the signal.

Received Signal

- The signal as it arrives at the receiver after being affected by noise.

Receiver

- Processes the received signal to recover the original message.
- Includes demodulation, decoding, and error correction.

Message (Output)

- The recovered information after reception.

Destination

- The final user or system that receives the message.
- Example: a listener, a computer, a display screen.

Six Principles of Communicating Data

In order to address these three types of communication problems, I'd like to propose six principles to consider when communicating data. They are numbered in the general order that they transpire, though it's fully recognized that this process is highly iterative and rarely pro- ceeds in a straight line. Communicating is a creative process—one that involves crafting and refining a message—and as such it will neces- sarily involve many loops:

1. Know your goal
2. Use the right data
3. Select suitable visualizations
4. Design for aesthetics
5. Choose an effective medium and channel
6. Check the results

Principle #1: Know Your Goal

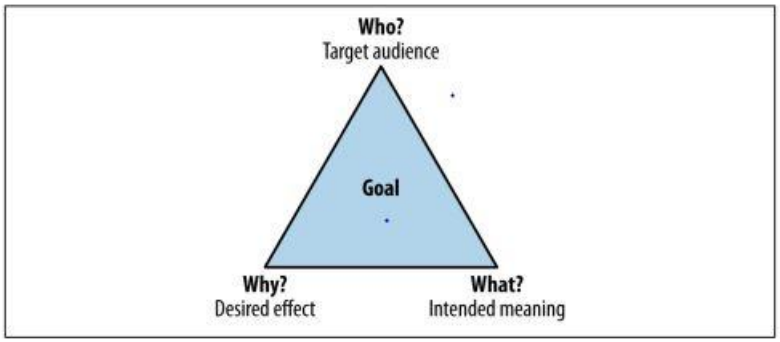


Figure 1-4. Elements of the goal

- Who are you trying to communicate with? (target audience)
- What do you want them to know? (intended meaning)
- Why? What do you want them to do about it? (desired effect)

EXAMPLES:: **Student Performance Report**

Goal: Help teachers identify **students who need support.**

The visualization should focus on:

- Low-scoring students
- Subjects with declining performance
- Attendance vs marks

If the goal is not defined, you may end up showing:

- Average class marks
- Gender distribution
- Highest scores

These don’t help the teacher take action.

Principle #2: Use the Right Data

Choosing the right data is *as important* as choosing the right chart.

Good visuals fail when the data behind them is:

- too little
- too much
- misleading
- statistically weak
- or ethically incorrect

Your goal is to select **only the data that strengthens the message**—nothing more, nothing less.

Example : Hospital Using Wrong Average Waiting Time

- A hospital proudly displayed:
- "Average patient waiting time = 12 minutes."

But the data hid the truth:

- Some patients waited **1 minute**
- Others waited **2 hours**

The **median** was the better metric (35 minutes).
Once shown the correct data, the hospital improved staffing.

- **Right Data Principle**
  Correct statistical measure matters more than flashy visuals.

Principle #3: Select Suitable Visualizations

- ✓ Below are **practical scenarios** showing how choosing the right visual encoding helps convey the intended meaning clearly.
- ✓ Once you've identified the data that you'll need to make your point, the next step is deciding how to encode the message. Encoding the data means converting the data values themselves into abstract graph- ical representations, like size or color or shape.

Example 1: E-commerce company wants to compare product ratings

Scenario

Flipkart wants to analyze how many products fall into each star rating (1-star to 5-star).

Data type

- Rating level → **Ordinal**
- Count → **Quantitative**

Suitable Encoding

- **Ordered bars** using **length**

Best Visualization

- **Ordered bar chart**

Why?

- Shows the natural order (1 < 2 < 3 < 4 < 5) clearly.

Example 2: Comparing Marks Across Subjects

Scenario

- You want to compare your marks in Math, Science, English, Social, and Hindi.

Data

- Subject → **Nominal**
- Marks → **Quantitative**

Best Visualization

- **Bar chart**

Why?

- Length of the bar helps you quickly see your strongest and weakest subjects.

Example:

Math 95, Science 88, English 78…

- What are the most effective types of visualizations for your data type? Once you've identified what data type or types you will need to get your point across, you need to decide what variables you will use to encode the data (see Figure 1-6).

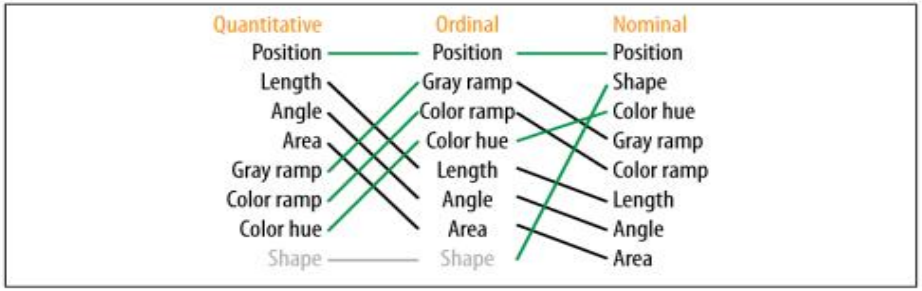


Figure 1-6. Effectiveness of data encoding

- A few points are immediately obvious:
- Position is the most effective form of encoding for all data types.
- Length, angle, and area decrease in effectiveness from quantitative to ordinal to nominal.
- Color hue increases in effectiveness from quantitative to ordinal to nominal.

Principle #4: Design for Aesthetics

Below are **clear, real-time, practical examples** that explain why **aesthetics matter in data visualization** and how poor aesthetic choices negatively affect communication.

These examples are easy to understand and come from **real student, business, and daily life scenarios**.

1. Poor Color Schemes

Real-Time Example:

- A school principal shows attendance data using colors like **red and green**.
  But some students (or teachers) are color-blind and cannot distinguish these colors.

Why This Is a Problem:

- Data becomes unreadable for color-blind viewers.
- Color carries meaning, so wrong or low-contrast colors confuse people.

Better Approach:

- Use high-contrast palettes: blue, orange, gray.
  Use color only where it adds meaning.

Why Aesthetics Actually Matter (In Simple Terms)

- Aesthetics:
  - ✔ Makes people pay attention
  - ✔ Helps memory
  - ✔ Enhances clarity
  - ✔ Makes insights easier to understand
  - ✔ Supports professionalism
  - ✔ Ensures data doesn't just inform but also **engages**
- Aesthetic errors:
  - ✘ Distract
  - ✘ Confuse
  - ✘ Overwhelm
  - ✘ Slow down comprehension
  - ✘ Look unprofessional

The goal is **not decoration**, but **clear, beautiful communication**.

EXAMPLE

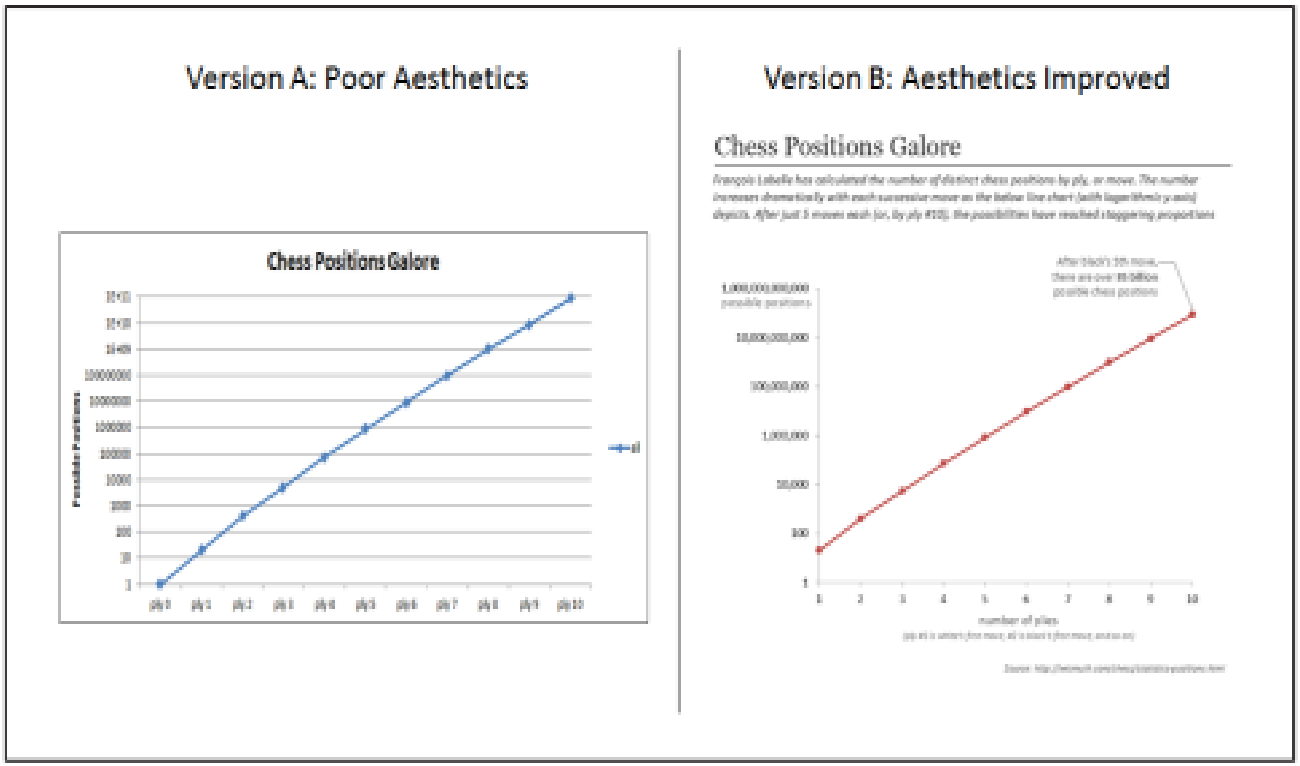


Figure 1-7. Two versions of the same line plot

There are a number of aesthetic elements of every data visualization, and a handful of common mistakes people make when creating them:

- Poor color schemes
- Distracting fonts
- Many different fonts
- Sloppy alignment
- Vertical or angled labels
- Dark background colors
- Thick borders or grid lines
- Useless images and clip art
- Lazily accepting most software defaults

Principle #5: Choose an Effective Medium and Channel

What form the message takes (medium) and how it gets delivered to the audience (channel) are critical elements of any data communication effort. Care needs to be taken in selecting the "how," the "when," and the "where" to improve the chances that your audience is reached and your goals are met.

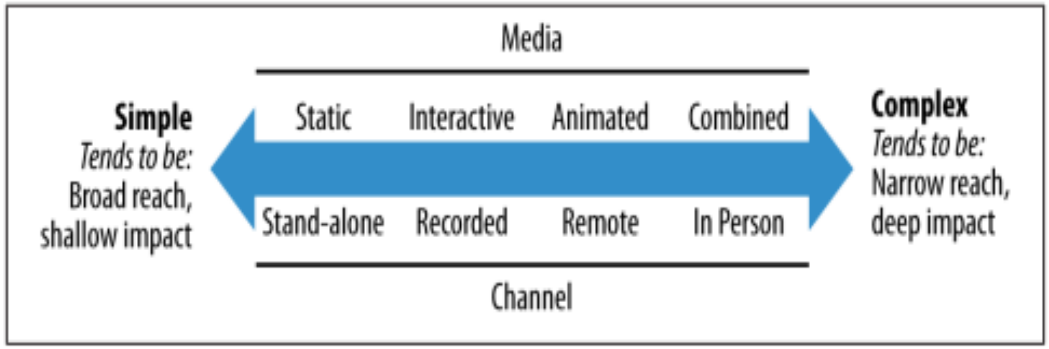


Figure 1-9. A spectrum of data communication types

Why Medium & Channel Matter

- The 'how', 'when', and 'where' decide communication success.
- Different audiences need different formats.
- Goal: Maximize clarity, attention, and impact.

1.Standalone Graphic vs. Narrated Graphic

Real-Time Example: Annual School Results

- A school wants to share **annual exam performance** with parents.
- **Standalone Graphic (Static PDF or Image)**
- They send a **PDF chart** showing average marks in each subject.

Good for:

✔ Quick reading
✔ Parents who check WhatsApp messages
✔ No explanation needed

Narrated Graphic (Teacher Video Explanation)

- The teacher records a **3-minute video** explaining why science scores increased, why English dropped, etc.

Good for:

✔ Context and storytelling
✔ Parents who need clarity
✔ Explaining causes behind the numbers

2. Static vs. Interactive Visualizations

Real-Time Example: College Attendance Dashboard

A college tracks attendance daily.

Static Visualization (Email Screenshot)

- The dean sends a **static chart** by email showing:
- Overall attendance trend
- Absent count

Good for:

✔ Quick overview
✔ No interaction required

Interactive Visualization (Power BI / Tableau Dashboard)

- Students and faculty use an **interactive dashboard** to:
- Filter by department
- Filter by date
- Check individual attendance

Good for:

✔ Exploration
✔ Department-specific insights
✔ What-if analysis

3. Animated Visualizations

Real-Time Example: Startup Growth Over 5 Years

- A startup wants to show investors how users increased from 2018–2024.

Animated Chart (GIF or Video Chart)

- A moving line chart shows user growth year-by-year.

Good for:

✔ Showing progression over time
✔ Making investors emotionally engaged
✔ Highlighting momentum

Why it works:

Animation helps show **change**, not static points.

Principle #6: Check the Results

It is a good habit in general to incorporate into your efforts feedback loops and checkpoints that help you gauge whether you've achieved your intended results or not. This allows for course correction in the case of woefully unmet goals, or fine-tuning in the case of slight miscues.

There are a few questions to ask when you check the results. We'll call this the "RUI":

Reach

Did the audience even receive your message at all? Who did and who didn't?

Understanding

Did the audience interpret the data message in the way you intended?

Impact

Did the audience react in the way you wanted them to react? Asking these questions will help you hone your message and com- municate data better, and it also will show an appropriate degree of respect to your audience.

1. Reach — Did the audience receive your message?

Example 1: Sales Dashboard Email

- You send a monthly **Sales Performance Dashboard** via email to 50 sales managers.

What you check:

- Email open rate: Only **32 out of 50** opened it.
- Dashboard clicks: Only **18** interacted with it.

Insight:

You *reached* only ~60% of the intended audience → many sales managers did not even see the dashboard.

Action:

Send reminders, repost on MS Teams, and present a short summary in the weekly sales meeting.

1. What is Tableau?

Tableau is a powerful data visualization and analytics tool used to turn raw data into clear, meaningful visual insights.

Think of Tableau as a tool that helps you:

- See your data visually
- Find patterns and trends
- Build dashboards
- Share insights with others

It is widely used in **business, finance, marketing, healthcare, supply chain, and data science**.

2. Why Do We Use Tableau?

✔ **Easy to Use**

- Drag and drop — no coding needed.

✔ **Connects to Any Data**

- Excel, SQL, CSV, cloud databases, Google Sheets, etc.

✔ **Fast Visualization**

- Instant charts, maps, dashboards.

✔ **Interactive Dashboards**

- Filters, parameters, actions → powerful for decision-making.

✔ **Industry Standard**

- Used by companies like Amazon, Deloitte, and Netflix.

Connecting to Data in Tableau

Tableau can connect to almost every kind of data, such as:

✔ **Files**

- Excel (.xlsx)
- CSV (.csv)
- PDF
- JSON
- Spatial files
- Text files

✔ **Databases**

- MySQL
- SQL Server
- Oracle
- PostgreSQL
- Snowflake
- Google BigQuery

✔ **Cloud Sources**

- Google Sheets
- AWS Redshift
- Salesforce
- Dropbox

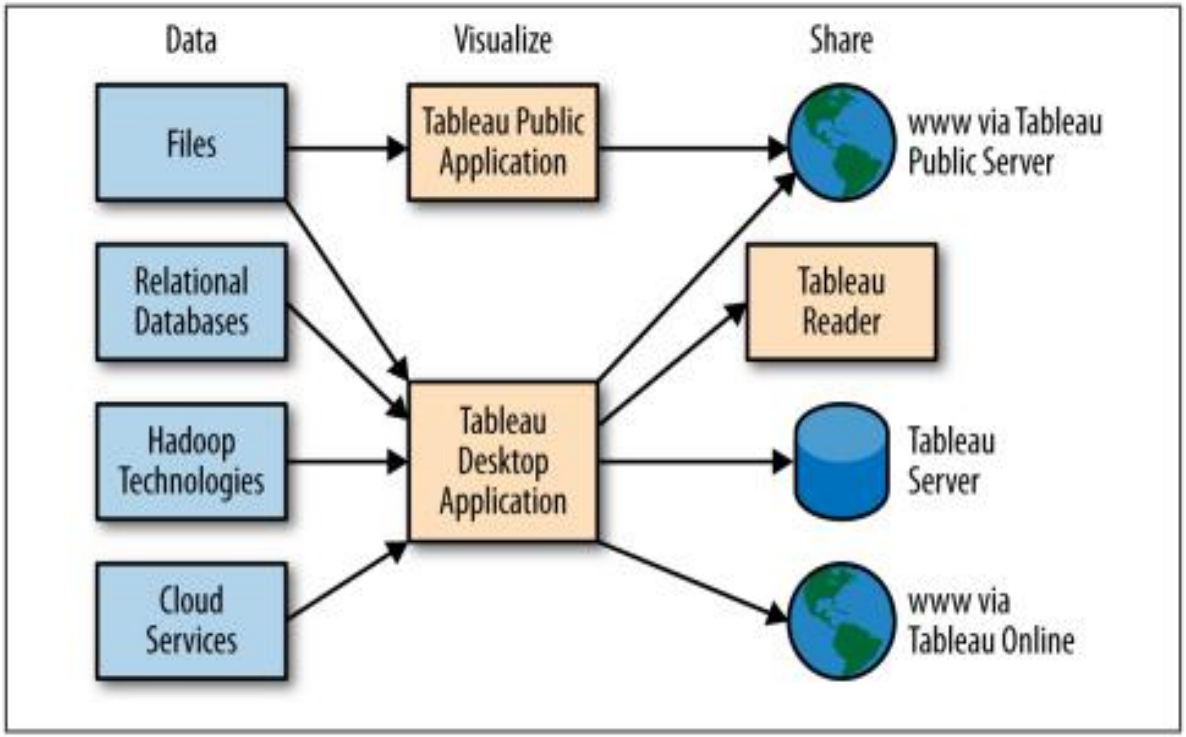


Figure 2-1. Tableau product diagram

⭐ 5. Common Charts in Tableau (with Examples)

Chart Type	Purpose	Example
Bar Chart	Compare categories	Sales by region
Line Chart	Show trends over time	Monthly revenue trend
Pie Chart	Show parts of a whole	Market share contribution
Maps	Show location-based data	Sales by country/state
Scatter Plot	Show relationships	Advertising spend vs Sales
Heat Map	Highlight patterns	Profit by category & region

6. Real-Time Examples Students Can Relate To

Example 1: College Attendance Dashboard

- Visualizes student attendance by subject, month, and department.

Example 2: E-commerce Sales Dashboard

- Shows:
- Daily orders
- Top-selling products
- Profit by category
- Students see real industry scenarios easily.

Example 3: Social Media Performance

- Visualizes:
- Likes
- Shares
- Engagement rate
- Growth over time
- Useful for marketing students.

Example 4: Finance Portfolio Dashboard

- Shows:
- Stock returns
- Risk levels
- Asset allocation
- Helps students learning finance/IB.

9. Advantages of Tableau

- ✔ **Very fast performance**
- ✔ **Drag-and-drop simplicity**
- ✔ **Easy to create dashboards**
- ✔ **Supports big datasets**
- ✔ **Beautiful, professional visuals**
- ✔ **Interactive & shareable**

10. Limitations of Tableau (Simple to explain)

- Cannot clean data deeply (need Tableau Prep or Excel/Python).
- Expensive (Desktop version).
- Limited advanced statistical modeling (compared to Python).

How Much and How Many

When we communicate data, what we are really doing is **making comparisons**. Humans naturally compare things to understand the world—bigger vs smaller, more vs less, higher vs lower. In data visualization, almost every comparison comes down to answering **one of two basic questions**:

- **How much?**
- **How many?**
- Understanding the difference between these two helps us choose the *right chart* and *right method* to present data clearly and honestly.

What Does "How Much" Mean?

- "How much" is used for **measured quantities**—things that cannot be counted one by one.

Examples:

- How much revenue did we earn?
- How much garbage was collected?
- How much time did a flight take?

These values are usually totals, averages, or sums. For example, garbage collected in tons or revenue in dollars.

Key idea:

- Data is often **already aggregated** (summed or averaged).
- Charts like **bar charts, dot charts, and tables** work well.
- **Bar length or position** helps people quickly see differences.

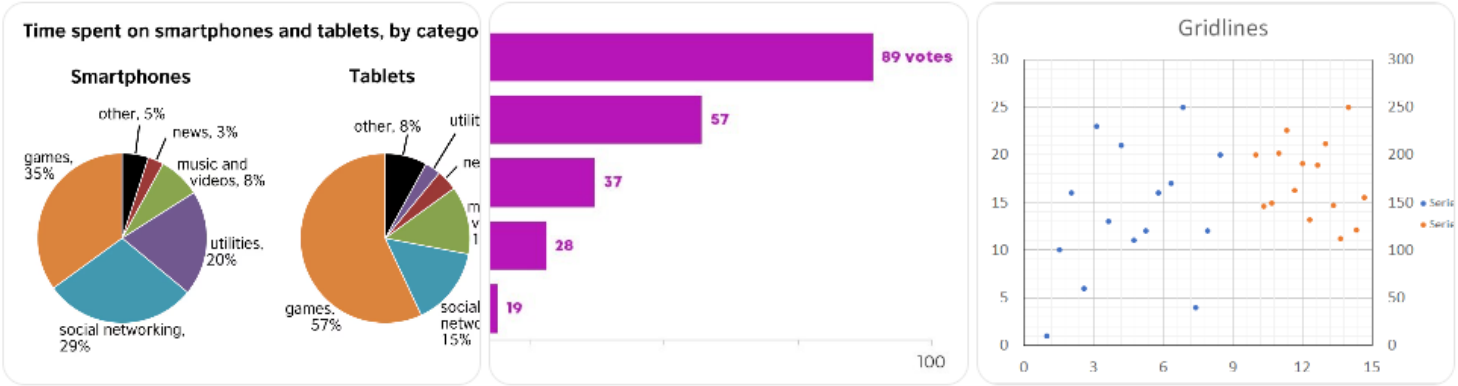
Why bar and dot charts work best:

- Our eyes are very good at comparing **length and position**.
- A horizontal bar chart makes it easy to see:
  - Who has the most
  - Who has the least
  - How large the differences are

Tables are best when:

- Exact numbers matter
- Precision is more important than visual comparison

Choosing the Right Chart Matters



Not all charts communicate "how much" equally well.

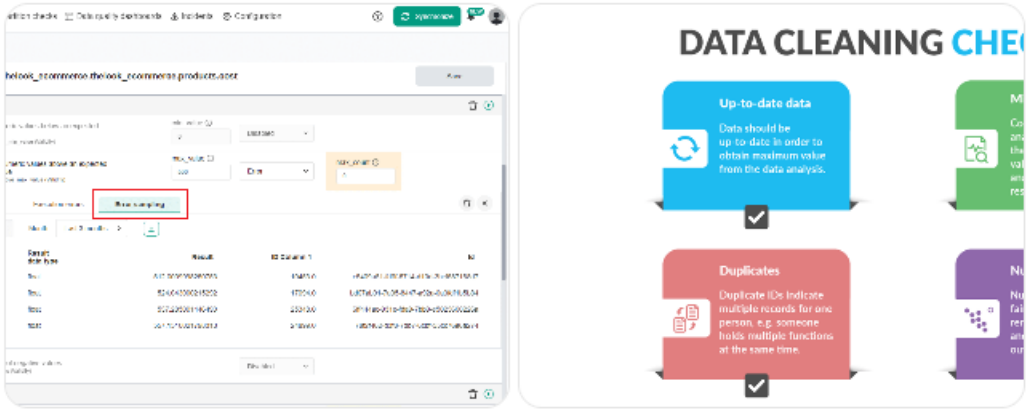
- **Bar charts** → Best for clear comparison
- **Dot charts** → Even more precise, use position instead of length
- **Pie charts** → Poor choice for comparison (hard to judge angles)
- **Tables** → Best for accuracy, not quick insight

The key rule is:

> **Choose the chart based on what the audience needs to do with the data.**

- **What Does "How Many" Mean?**
- "How many" is used for **countable items**—things that exist as individual records.
- Examples:
- How many customers?
- How many complaints?
- How many incidents or events?
- Unlike "how much," this data often starts at the **row level**, where each row represents one event.
- **Two common ways to count:**
- **Number of Records** – counts rows automatically
- **COUNT of a field** – counts unique IDs or categories
- Both methods answer the same question:
- "How many times did this happen?"

Using Counts to Discover Problems



Counting data can also help **detect errors**:

- Misspellings
- Duplicate entries
- Incorrect categories

For example, if a city name appears only once but looks wrong, it signals a data quality issue.

Histograms: "How Many of How Much?"

- A **histogram** answers a combined question:
- "How many records fall into each range of values?"
- Instead of categories like cities or departments, histograms use **numeric ranges** (called bins).

Example:

How many districts collected:

- 1,000–1,500 tons?
- 1,500–2,000 tons?

Important ideas:

- Bin size affects readability
- Too many bins → cluttered
- Too few bins → oversimplified
- Missing ranges should be shown as zero, not ignored
- A good histogram balances clarity and detail—*not too many, not too few*.

Key Takeaways (In Simple Terms)

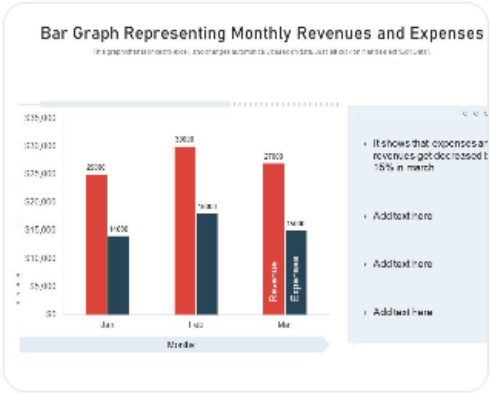
- All data communication starts with **comparison**
- Every comparison answers:
  - **How much** (measured values)
  - **How many** (counts of things)

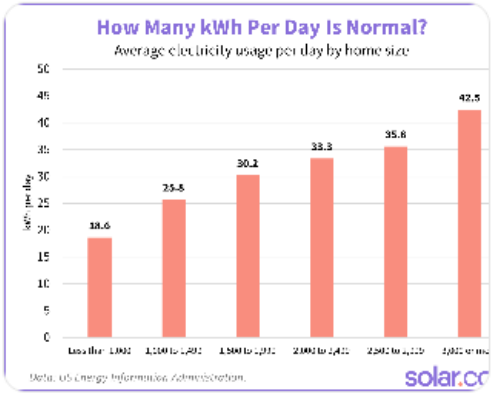
The **chart choice depends on the task**

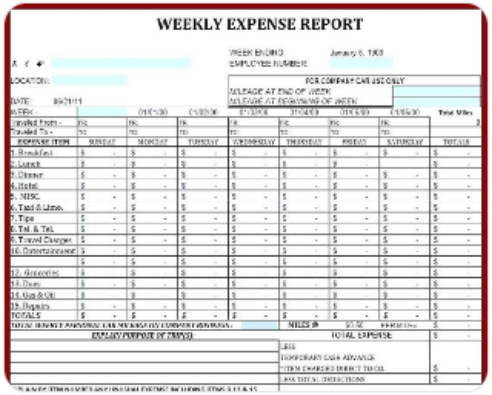
- Bars and dots for comparison
- Tables for precision
- Histograms for distribution

Good visuals build **trust** by being clear, accurate, and honest

1. Real-Time Examples of "How Much" (Measured Data)







Example 1: Company Sales Revenue

Question:

- *How much revenue did each product generate last month?*

Real-life situation:

A retail manager wants to compare revenue from:

- Mobile Phones
- Laptops
- Accessories

Data type:

- Revenue in ₹ or $ (measured, not counted)

Best chart:

- Horizontal bar chart or dot chart

Why:

Revenue is a **continuous value**, and comparing bar lengths makes differences obvious.

Example 2: Customer Complaints

Question:

- *How many complaints were received from each city?*

Real-life situation:

A telecom company counts complaints from:

- Chennai
- Bengaluru
- Hyderabad

Data type:

- Number of complaints (events)

Why "how many":

Each complaint is a **separate record**.

Summary

How much and how many are foundational types of comparisons that we make every day. Whenever data is communicated, one of these two comparisons (or both) is being made. Knowing your options, and evaluating which will work best for the task at hand, is an important first step in encoding the message. We've seen how to work with Tableau when communicating the most basic of comparisons: those in which only one variable is involved. Now, let's move on to explore how to communicate data when more than one comparison is being made.

Ratios and Rates

Introduction

When we compare data using **absolute values** (total numbers), the comparison may be misleading because different regions have different populations, sizes, and activities.

To make **fair and meaningful comparisons**, we use **normalized measures** such as **ratios, rates, proportions, and percentages**.

These measures help answer questions like:

- How much **per person**?
- How much **relative to something else**?

Normalized Comparisons

Normalized comparisons put data on a **common scale**, allowing accurate "apples-to-apples" comparisons.

The four common types are:

- Ratio
- Rate
- Proportion
- Percentage

1. Ratio

Definition:

A **ratio** is a comparison of two quantities with the **same unit**, expressed as a quotient.

Formula:

Ratio = Quantity A / Quantity B

Example:

- Manhattan produced **0.264 tons of recyclables for every 1 ton of refuse**.
- Recycle to Refuse Ratio = 0.264

Key Points:

- Units are the same (tons / tons)
- Can be written as x:y, x/y, or a decimal

2. Rate

Definition:

A **rate** is a ratio where the numerator and denominator have **different units**.

Formula:

- Rate = Quantity / People or Time or Area

Examples:

- Population density = people per square mile
- Trash production = pounds per person
- Crime rate = crimes per year

Key Points:

- Units are different
- Rates are often predictive
- Frequently used in real-world analysis

3. Proportion

Definition:

A **proportion** is a ratio where the numerator is a **part of the total**.

Range:

- 0 to 1

Example:

- 0.169 of NYC's population lives in the Bronx

4. Percentage

Definition:

A **percentage** is a proportion multiplied by 100.

Range:

- 0% to 100% (can exceed 100%)

Example:

- 16.9% of NYC residents live in the Bronx
- Sales increased by 150%

- **Creating Ratios in Tableau**
- **Step 1: Create a Calculated Field**
- Tableau allows creating new fields using **Calculated Fields**.
- Example:
- RecyclableTonsCollected = SUM([PaperTonsCollected]) + SUM([MGPTonsCollected])

Step 2: Create the Ratio

- Recycle to Refuse Ratio = SUM([RecyclableTonsCollected]) / SUM([RefuseTonsCollected])
- **Important Note:**
- Always use **SUM()** for both numerator and denominator
- This avoids misleading results when aggregating data

Visualizing Ratios

- **Bar Chart**
- Rows → Borough
- Columns → Recycle to Refuse Ratio
- Sort in descending order
- **Purpose:**
- Compare recycling performance across boroughs

Community District Analysis

- To avoid incorrect aggregation:
- Compare **community districts within each borough**
- Use grids or combined dimensions

Highlight Tables

Purpose:

- Quickly identify highest and lowest values
- Color intensity shows magnitude

Advantage:

- Easy pattern recognition

Limitation:

- Difficult to rank many values precisely
- **Ranking Data in Tableau**

Method 1: INDEX() Function

- Rank = INDEX()
- Compute using combined dimension
- Convert to **Discrete**

Method 2: Rank Table Calculation

- Quick Table Calculation → Rank
- Better handling of ties

Rates Using Multiple Data Sources

Example:

- **Trash produced per person**
- **Data Required:**
- Trash collected (DSNY data)
- Population (Census data)

Data Blending:

- Link data using common fields:
  - Borough
  - CommunityDistrict

Creating a Rate in Tableau

- Refuse per Person (lbs) = SUM([RefuseTonsCollected]) * 2000 / SUM([Population])

Why Pounds?

- Easier to understand per person than tons

Interpreting Results Carefully

- Data shows **what**, not always **why**
- High rates may be due to:
  - Industries
  - Commercial waste
  - Collection methods

- Avoid jumping to conclusions without context.

Summary

- Ratios and rates create fair comparisons
- Ratios use same units; rates use different units
- Tableau Calculated Fields help create normalized measures
- Highlight tables and bar charts aid comparison
- Ranking improves readability
- Data blending allows advanced analysis
- Always interpret results responsibly

Reference

- C. B. Jones, *Communicating Data with Tableau: Designing, Developing, and Delivering Your Data to the Masses*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2014.
- C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL, USA: Univ. of Illinois Press, 1949.

TEXT BOOKS

- Ben Jones, Communicating Data with Tableau, O‘Reilly June 2014

REFERENCE WEBSITE:

- Data Visualization | Coursera.

UNIT-IV

Proportions and Percentages

In data visualization, proportions and percentages help us explain **relationships**, not just raw numbers. Instead of saying *how much* something happened, they tell us **how important it is relative to the whole**. This is crucial because audiences usually care more about *share*, *contribution*, or *change* than absolute values.

Proportions are ratios between 0 and 1, while percentages express the same idea out of 100. Both are powerful because they allow fair comparisons—even when totals differ.

Common Ways Proportions Are Used

In practice, proportions usually answer one of three questions:

- **Part-to-whole** – How much does one part contribute to the total?
- **Current-to-historical** – How does today compare with the past?
- **Actual-to-target** – How close are we to a goal?

This example focuses on **part-to-whole**, which is one of the most common and useful comparisons.

Why Filtering Matters Before Showing Proportions

When visualizing proportions (like batting averages or home-run shares), **context matters**. Including players with very few opportunities (for example, only two at bats) can distort the story. A 0.500 batting average looks impressive—until you realize it's based on almost no data.

- From a data visualization standpoint:
- Filters improve **fairness**
- They reduce **noise**
- They prevent **misleading conclusions**

Interactive filters (like Tableau's Quick Filters) are especially effective because they let users explore the data themselves and see how results change.

Part-to-Whole

Choosing the Right Chart for Part-to-Whole

Not all charts communicate proportions equally well:

- **Pie charts** struggle when there are many categories. Small differences are hard to see, and middle slices blur together.
- **Stacked bars** show totals but make comparisons across segments difficult.
- **Bar charts with percent labels** work better but can still feel heavy.
- The **dot chart** stands out as the most effective option:
- It clearly shows **rank**
- It allows easy comparison of both **relative (percentage)** and **absolute (count)** values
- It avoids clutter and visual distortion
- From a data visualization perspective, dot charts are often the best choice for part-to-whole comparisons with many categories.

Part-to-Whole (Most common real-time use)

🎓 Example: Student Marks Distribution

Out of **500 total marks**:

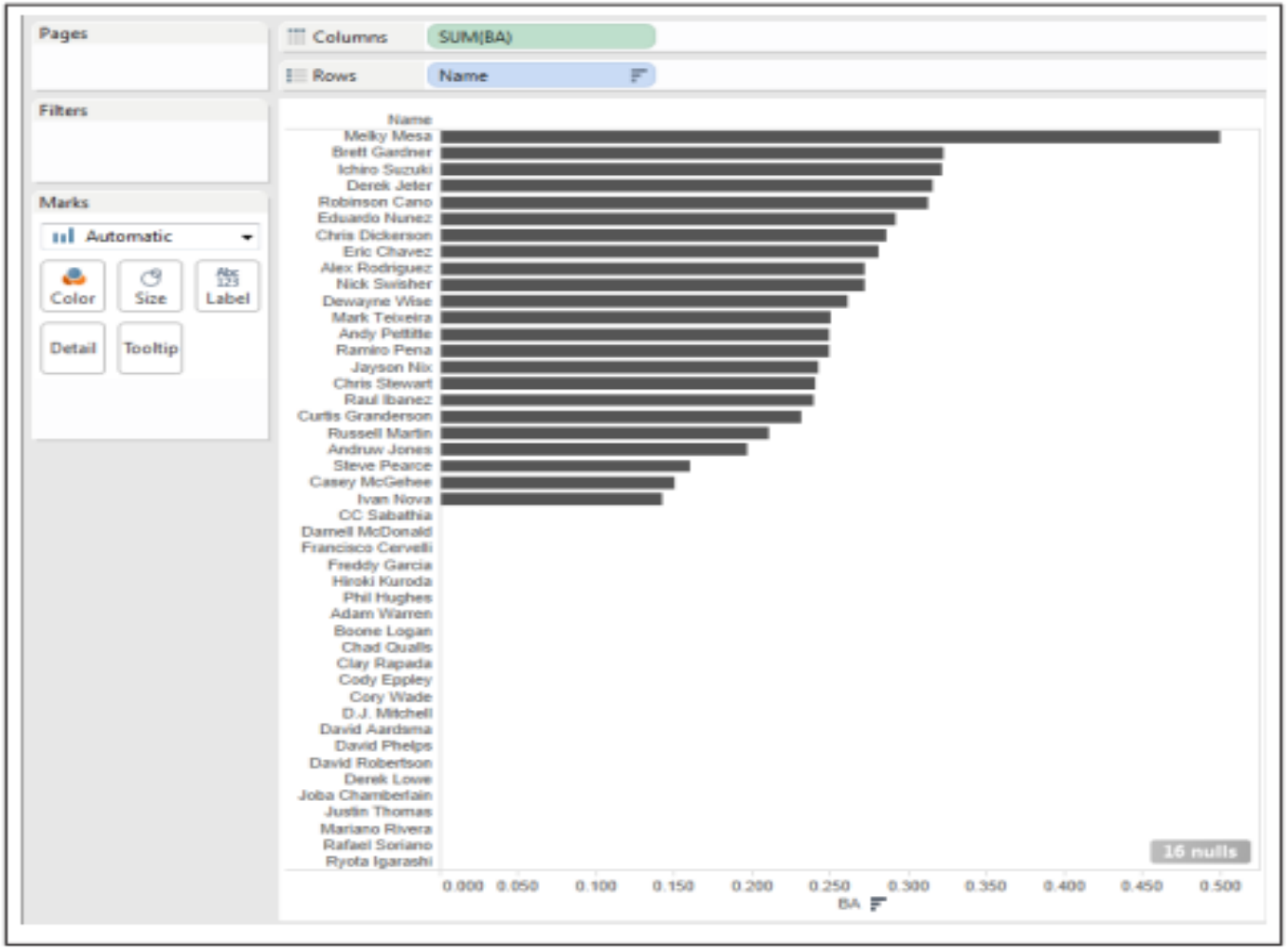
Subject	Marks	Percentage
Math	150	30%
Science	125	25%
English	100	20%
Social Studies	75	15%
Computer Science	50	10%

Visualization choice:

- **Stacked bar** → shows total score
- **Dot chart** → makes subject strengths obvious

✅ **Insight:**

Teachers and students can quickly identify **strong and weak areas**.


Pages
Columns
SUM(BA)
Rows
Name
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Name
Melky Mesa
Brett Gardner
Ichiro Suzuki
Derek Jeter
Robinson Cano
Eduardo Nunez
Chris Dickerson
Eric Chavez
Alex Rodriguez
Nick Swisher
Dewayne Wise
Mark Teixeira
Andy Pettitte
Ramiro Pena
Jayson Nix
Chris Stewart
Raul Ibanez
Curtis Granderson
Russell Martin
Andruw Jones
Steve Pearce
Casey McGehee
Ivan Nova
CC Sabathia
Darnell McDonald
Francisco Cervelli
Freddy Garcia
Hiroki Kuroda
Phil Hughes
Adam Warren
Boone Logan
Chad Qualls
Clay Rapada
Cody Eppley
Cory Wade
D.J. Mitchell
David Aardsma
David Phelps
David Robertson
Derek Lowe
Joba Chamberlain
Justin Thomas
Mariano Rivera
Rafael Soriano
Ryota Igarashi
16 nulls
0.000
0.050
0.100
0.150
0.200
0.250
0.300
0.350
0.400
0.450
0.500
BA

Proportions as Waterfall Charts

A **waterfall chart** is a powerful way to show **how individual parts build up to a total**. Instead of stacking everything in one bar, the values are arranged step-by-step so the viewer can see **each contribution clearly**.

Real-Time Example: Monthly Company Revenue

Assume a company's **total revenue = ₹10 crore**, coming from different departments:

Department	Revenue (₹ crore)
Online Sales	3.5
Retail Sales	2.8
Corporate Clients	2.0
Exports	1.2
Others	0.5

Why a waterfall chart works here:

- Each bar **starts where the previous one ends**
- The final bar shows the **grand total**
- Viewers clearly see **how each department adds to total revenue**

Business insight:

Management can instantly identify which departments contribute most to overall revenue.

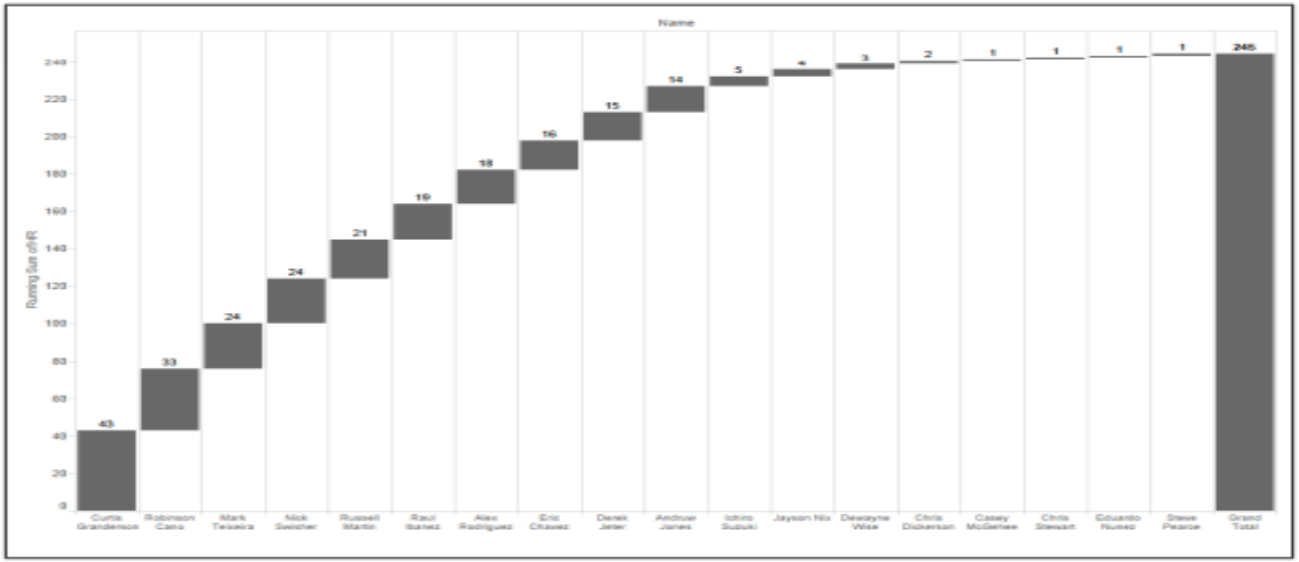


Figure 5-12. Home run data as a waterfall chart

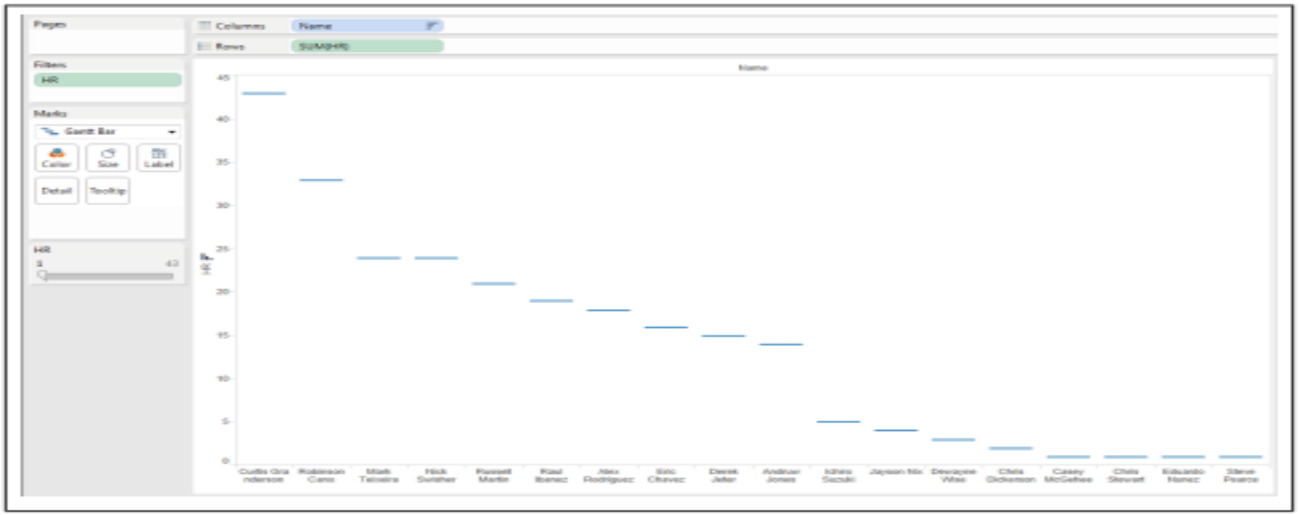


Figure 5-13. Waterfall chart in progress

Why Gantt-Based Waterfall Charts Are Useful

- Using Gantt-style bars creates a **staircase effect**, which:
- Visually reinforces the **part-to-whole relationship**
- Makes cumulative growth intuitive
- Helps explain complex totals to non-technical audiences
- This style is often used in:
- Financial reports
- Profit & loss breakdowns
- Cost analysis dashboards

Current-to-Historical Comparison (Real-Time Example)

📈 Example: Year-on-Year Sales Growth

Year	Sales (₹ crore)
2023	120
2024	138

Percentage change:

$$\frac{138 - 120}{120} \times 100 = 15\%$$

Best visualization:

- **Bullet graph**
- **Bar + reference line**
- **Line chart with labels**

📊 Insight:

Instead of just saying "sales increased," we can say **sa[↓] grew by 15% compared to last year**, which is much more meaningful.

Bullet Graph Explained with Real-Time Use

🎯 Example: Comparing Performance Across Branches

Suppose we compare **2024 sales** (current) against **2023 sales** (historical):

Branch	2023 Sales	2024 Sales
Chennai	100	115
Bangalore	120	110
Hyderabad	90	95

How to read a bullet graph:

- **Bar** → current year performance (2024)
- **Vertical line** → last year's performance (2023)
- **Bands** → performance ranges (60%, 80%, 100%)

📊 Interpretation:

- Chennai exceeded last year
- Bangalore underperformed
- Hyderabad improved slightly

Reference Lines (Explained Simply)

Reference lines are visual guides added to charts to help viewers quickly compare actual values against a benchmark such as an average, previous year value, target, or limit. Instead of forcing the audience to calculate differences mentally, reference lines make comparisons **instant and intuitive**.

In Tableau, reference lines and bands can be added to **any chart**, not just bullet graphs. Bullet charts simply come with them pre-built.

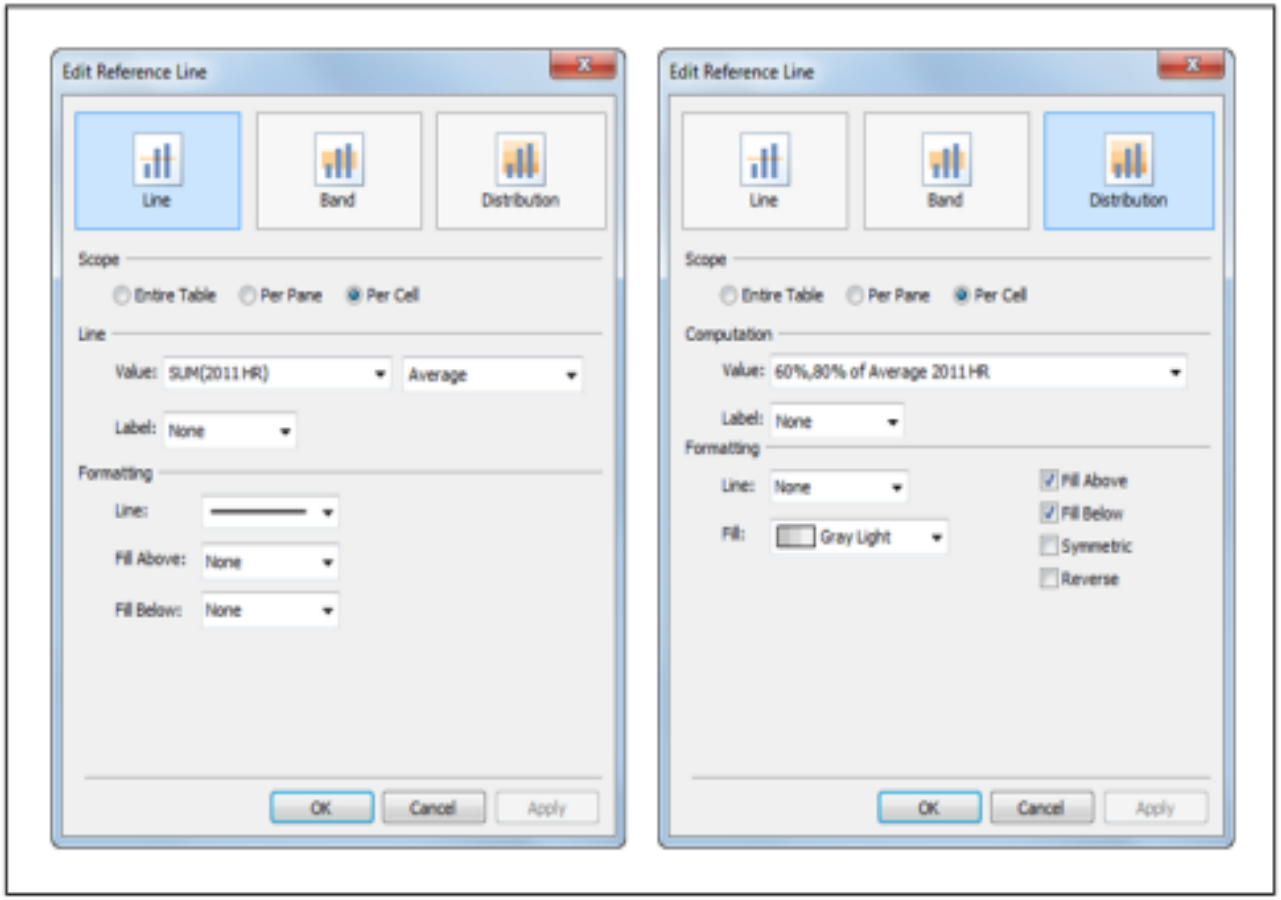


Figure 5-18. Two reference lines automatically created along with the bullet graph

Actual-to-Target Comparison (Key Business Use Case)

- Actual-to-target compares **what has happened** with **what should have happened**.
- Unlike part-to-whole, this comparison **can exceed 100%**.

Real-Time Example 4: Sales Target Achievement

Sales Rep	Target (₹ Lakh)	Actual (₹ Lakh)	% Achieved
Rep A	50	55	110%
Rep B	50	45	90%
Rep C	50	40	80%

Best visualization:

- **Bullet chart**
- **Progress bar**
- **Bar chart with target line**

📊 **Insight:**

- Rep A exceeded target
- Rep B slightly missed target
- Rep C needs improvement

Mean and Median

Mean and Median – Explained for Real Life

When we present data, we almost always include an **average** or a **median**. We do this because humans are not good at understanding long lists of numbers. Measures of central tendency help us summarize large datasets into a single value that represents what is *typical*.

However, choosing the **wrong measure** can seriously mislead the audience. That's why understanding **mean vs median** is essential in data visualization.

Mean, Median, and Mode (In Simple Words)

- **Mean (Average):**
  Add all values and divide by the number of values.
  ➜ Very sensitive to extreme values (outliers).
- **Median:**
  The middle value when data is sorted.
  ➜ More stable and often more realistic.
- **Mode:**
  The most frequently occurring value.
  ➜ Useful in limited scenarios (like product sizes or grades).

Normal Distribution

- Many natural measurements follow a normal distribution:
- Human height
- Test scores
- Manufacturing tolerances
- In such cases:
- Mean = Median = Mode
- So using the **average** is safe and meaningful.

Real-Time Example 2: Employee Salaries (Non-Normal Data)

Scenario:

Monthly salaries (₹):

> 25,000, 28,000, 30,000, 32,000, 35,000, 40,000, 2,00,000

- Mean ≈ ₹55,700
- Median = ₹32,000

Key Question:

Which number represents a *typical employee salary*?

📊 **Answer:** Median.

Why?

- One highly paid manager inflates the mean
- Most employees earn close to ₹32,000

🚨 **Problem with mean:**

More than 80% of employees earn **less than the average**.

What is Normal Distribution?

- A normal distribution, also known as a Gaussian distribution, is a fundamental probability distribution in statistics characterized by its symmetrical, bell-shaped curve. Here are its key features:
- **Symmetry**: The distribution is perfectly symmetrical around its mean, meaning the left side is a mirror image of the right side.
- **Bell-Shaped Curve**: The majority of data points cluster around the mean, with the frequency of observations tapering off as you move away from the mean. This creates the characteristic bell shape.
- **Mean, Median, and Mode**: In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.
- **Standard Deviation**: The spread of the data is determined by the standard deviation. Approximately 68% of the data falls within one standard deviation of the mean, 95% falls within two standard deviations, and 99.7% falls within three standard deviations. This is known as the empirical rule or the 68-95-99.7 rule.
- **Z-Scores**: Z-scores, which measure the number of standard deviations a data point is from the mean, can be used to compare data points from different normal distributions.

Example of Normal Distribution

- An example of a normal distribution can be found in the distribution of human heights. If you measure the heights of a large number of people, the data tends to cluster around a central value (the mean height) with a symmetric distribution on either side. Most people have heights that are close to the average, and as you move further from the average, fewer people have those heights. This creates the characteristic bell-shaped curve of a normal distribution.
- **IQ Scores**: IQ scores are designed to follow a normal distribution with a mean of 100 and a standard deviation of 15. This means that approximately 68% of the population has an IQ between 85 and 115, and 95% have an IQ between 70 and 130.

Box Plots

Box Plots – Best Tool for Central Tendency

A **box plot** shows:

- Minimum
- Maximum
- Median
- Interquartile range
- Outliers
- Mean (optional)

Real-Time Example: Department-wise Salaries

Using box plots:

- HR salaries are tightly grouped
- IT salaries show wide variation
- Sales team has extreme outliers (bonuses)

📊 **Why box plots are powerful:**

- You see spread + center together
- You immediately spot skewness and outliers ↓

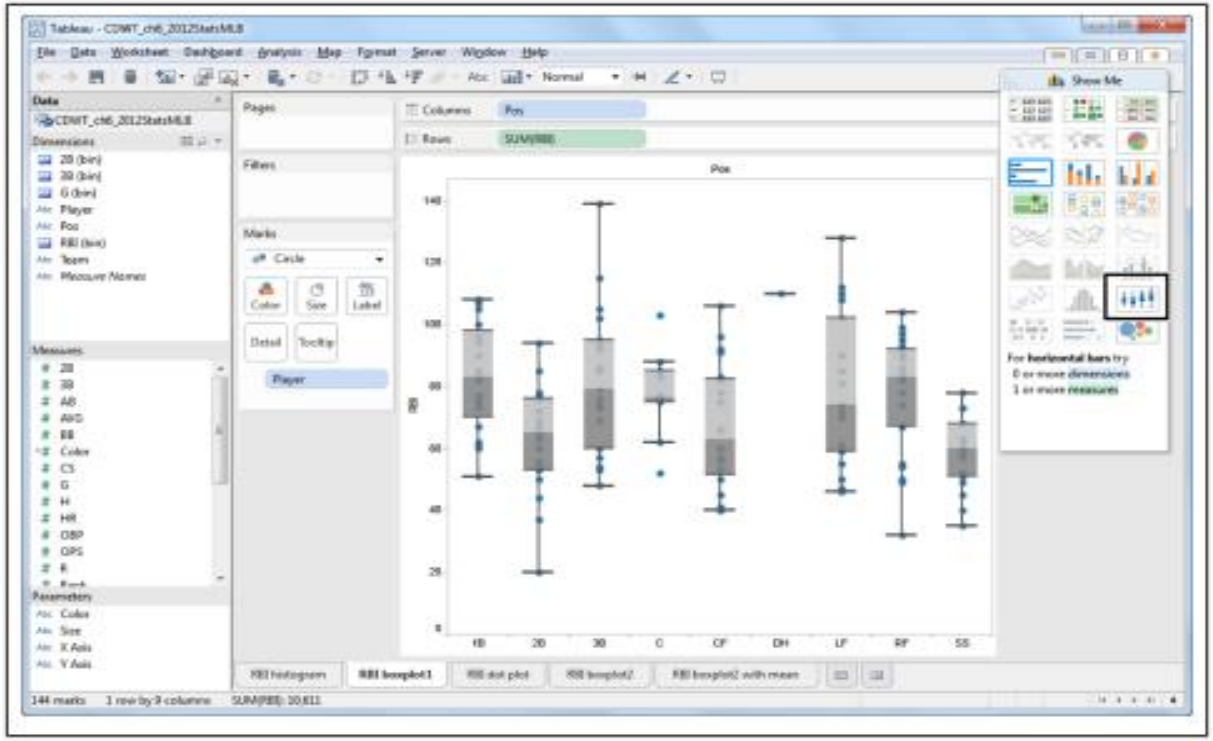


Figure 6-4. The box-and-whisker plot

What is Non-Normal Distribution?

A non-normal distribution is any statistical distribution that does not conform to the bell-shaped, symmetrical pattern of a normal distribution. Non-normal distributions can exhibit various shapes and characteristics, making them more complex to analyze. Here are some common types of non-normal distributions:

- **Skewed Distributions**:
- **Positively Skewed (Right Skewed)**: In a positively skewed distribution, the tail on the right side of the distribution is longer or fatter than the left side. An example is income distribution, where a small number of people earn significantly more than the majority.
- **Negatively Skewed (Left Skewed)**: In a negatively skewed distribution, the tail on the left side is longer or fatter than the right side. An example is the age at retirement, where most people retire around a certain age, but a few retire much earlier.
- **Bimodal and Multimodal Distributions**: These distributions have two or more peaks. An example is the distribution of test scores in a class where there are two distinct groups of students, such as high achievers and low achievers.
- **Uniform Distribution**: This distribution has no peaks and is flat, meaning each outcome is equally likely. An example is the roll of a fair die, where each number from 1 to 6 has an equal probability of occurring.
- **Heavy-Tailed Distributions**: These distributions have tails that are not exponentially bounded. Examples include the Cauchy distribution and financial returns, where extreme values (outliers) are more likely than in a normal distribution.
- **Exponential Distribution**: This distribution describes the time between events in a Poisson process. An example is the time between arrivals of customers at a store.

Examples of Non-Normal Distribution

Some examples of non-normal distributions are:

- **Income Distribution**:
  - **Positively Skewed Distribution**: Income distributions in most countries are positively skewed, with a long right tail. This means that a small number of people earn much higher incomes than the majority, creating a skew to the right.
    - **Example**: The distribution of household incomes in the United States is positively skewed, with the majority of households earning below the mean income and a few households earning significantly above it.
- **Age at Retirement**:
  - **Negatively Skewed Distribution**: The age at which people retire is often negatively skewed, with more people retiring at the common retirement age (e.g., 65) and fewer people retiring much earlier.
    - **Example**: In many countries, the age at retirement shows a peak around the typical retirement age, with a longer tail towards younger ages where fewer people retire early due to various reasons.

Sensitivity to Outliers: A Hypothetical Scenario

- Now imagine adding **one superstar player** earning **$20 million** (like Messi or Ronaldo).
- **What happens?**
- **Mean salary** jumps dramatically
  → From about **$164K to nearly $200K**
- **Median salary** increases by **less than $100**
- **Why?**
- Mean is **highly sensitive to outliers**
- Median is **robust and stable**
- This shows why the median is preferred when describing **"typical" values** in skewed distributions.

	2012 guaranteed compensation	If a player earning $20M joined MLS	% earning less than
Mean	$ 163,934	$ 199,999	84%
Standard Deviation	$ 423,846	$ 945,895	N/A
1st Quartile	$ 44,100	$ 44,100	25%
Median	$ 81,250	$ 81,332	50%
3rd Quartile	$ 160,500	$ 160,875	75%
Mode	$ 44,000	$ 44,000	10%

Figure 6-12. Hypothetical statistics resulting from the addition of a $20M player

Variation and Uncertainty

- **Variation and Uncertainty – Explained in Simple Words**
- **1 Why This Principle Matters**
- When we see data, we often become overconfident.
- Example:
- "Average sales increased."
- "Customer satisfaction is 82%."
- "Profit margin is 15%."
- But the real question is:
- **How much do the values vary?**
  How sure are we about this result?
- That's where **variation** and **uncertainty** come in.

PART 1: Variation

What is Variation?

Variation means:

How different individual values are from each other.

- If all values are close together → low variation
  If values are very spread out → high variation

Simple Example

- Imagine two companies:

Company A

Monthly profits (in lakhs): 10, 11, 9, 10, 10

- Very consistent → Low variation

Company B

Monthly profits: 2, 25, 3, 40, 5

- Very unstable → High variation
- Even if the average profit is similar,
  Company B is **much riskier**.

Important Idea from the Chapter

- Just showing **average** is misleading.
- If we only show:
- "Average strikeouts per year"
- We miss:
- Which teams performed badly
- Which teams performed extremely well
- Whether there were outliers
- □ Showing variation gives a **more honest picture**.

Real-Time Business Example (Variation)

Example: Manufacturing Company – Defect Rate

A factory produces 10,000 mobile phones daily.

- Management sees:
- Average defect rate = 2%
- They think:
  "Everything is fine."
- But when variation is shown using a control chart:
- Some days = 1%
- Some days = 6%
- Some days = 0.5%
- Now we understand:
  - The process is unstable.

Control Chart (Real-World Application)

- A control chart shows:
- Average line
- Upper Control Limit (UCL)
- Lower Control Limit (LCL)
- Outliers
- Trends

- Shifts
- If values cross UCL or LCL → Something unusual happened.

Example:

- Machine malfunction
- Supplier issue
- Worker error

This is heavily used in:

- Manufacturing
- Six Sigma
- Quality control
- Banking fraud detection
- Stock market monitoring

Control Chart (Real-World Application)

- A control chart shows:
- Average line
- Upper Control Limit (UCL)
- Lower Control Limit (LCL)
- Outliers
- Trends
- Shifts
- If values cross UCL or LCL → Something unusual happened.
- Example:
- Machine malfunction
- Supplier issue
- Worker error
- This is heavily used in:
- Manufacturing
- Six Sigma
- Quality control
- Banking fraud detection
- Stock market monitoring

PART 2: Uncertainty

What is Uncertainty?

Uncertainty happens when:

We use a sample to make conclusions about the whole population.

- Example:
  You survey 100 customers out of 10,000.
- You find:
  70% are satisfied.
- But can we confidently say:
  "70% of all customers are satisfied"?
- Not exactly.

Because:

- It's just a sample
- Sample may not represent everyone

Confidence Intervals (Very Important Concept)

- Instead of saying:
- Satisfaction = 70%
- We say:
- Satisfaction = 70% ± 5% (95% confidence)
- This means:
  We are 95% sure that real satisfaction lies between 65% and 75%.
- That is honest communication.

Real-Time Example (Uncertainty)

Example: Investment Banking / Market Research

- Imagine a company wants to launch a new product.
- They survey 200 people.
- 120 say they will buy.
- Sample result:
  60% purchase intention
- But total target market = 1,00,000 people.
- Instead of saying:
  "60% of people will buy"
- We should say:
  "Estimated demand is 60% with a 95% confidence interval of 53% to 67%."
- Now management understands:
  There is uncertainty.

Why This Is Very Important in Finance

- In economics and finance:
- Stock returns are not normally distributed
- Markets have extreme outliers
- Financial crises are unpredictable
- Using simple standard deviation blindly can be dangerous.
- This is what:
- Benoit Mandelbrot
- Nicholas Nassim Taleb
- warned about.
- Financial markets have:
  - Fat tails
  - Extreme events
  - Non-normal distributions
- So we must choose the right statistical tools.

Key Lessons from This Chapter

- Never show only averages.
- Always show variation if it matters.
- Use control charts for processes over time.
- Use confidence intervals when dealing with samples.
- Be honest about what you don't know.
- Do not overstate conclusions.

Simple Exam-Ready Summary

Variation refers to how much data points differ from one another, while uncertainty refers to the lack of confidence when making conclusions about a population using sample data. When communicating data, it is important to show variation through tools like box plots or control charts and to show uncertainty through confidence intervals to avoid misleading the audience.

References

- C. B. Jones, *Communicating Data with Tableau: Designing, Developing, and Delivering Your Data to the Masses*, 1st ed. Sebastopol, CA, USA: O
- B. B. Mandelbrot and R. L. Hudson, *The (Mis)Behavior of Markets: A Fractal View of Risk, Ruin, and Reward*. New York, NY, USA: Basic Books, 2004.
  N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable*, 2nd ed. New York, NY, USA: Random House, 2010 'Reilly Media, 2014

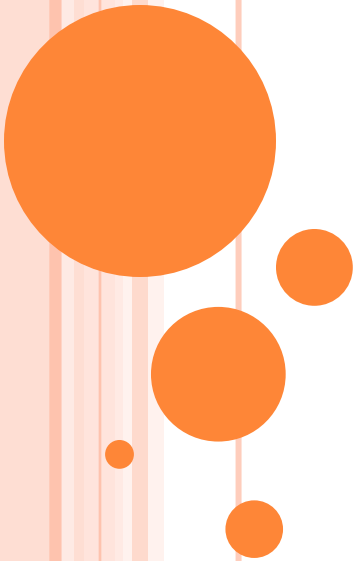
TEXT BOOKS

- Ben Jones, Communicating Data with Tableau, O‘Reilly June 2014

REFERENCE WEBSITE:

- Data Visualization | Coursera.

UNIT-V



Multiple Quantities in Data Visualization

- So far, single-variable charts (like bar charts or histograms) help us understand **one variable at a time**.
- But real-world insights often come from analyzing **relationships between multiple variables**.
- When we analyze **multiple quantities**, we study the relationship between two or more variables in the same view. In Tableau, the most powerful chart for this is the **Scatterplot**.

Why Multiple Quantities Matter

Single-variable charts (bar charts, pie charts) show **one measure at a time**.

But business and analytics decisions require relationships like:

- Sales vs Profit
- Marketing Spend vs Revenue
- GDP vs Inflation
- Goals vs Assists

Studying these relationships helps identify:

- Patterns
- Trends
- Outliers
- Correlations

- Important: **Correlation does not imply causation**
  Just because two variables move together doesn't mean one causes the other.

Scatterplots

A **scatterplot** is one of the most powerful tools for visualizing relationships between two quantitative variables.

- It:
- Uses a two-dimensional plane
- Plots one variable on the X-axis
- Plots another on the Y-axis
- Shows each data point as a dot
- Scatterplots help us:
- Identify relationships
- Detect clusters
- Spot outliers
- Compare many values at once

Example: Hockey Career Statistics

- The example discussed a scatterplot of top 100 career point leaders in professional hockey.
- **Variables Used:**
- Goals (G) → X-axis
- Assists (A) → Y-axis
- Points (P = G + A) → Size of circle
- Games Played (GP) → Color of circle
- This allows comparison across **four dimensions in one chart**:
- Position (Goals)
- Position (Assists)
- Size (Total Points)
- Color (Games Played)

Identifying “Who Is Who” in Scatterplots

- When many points are displayed, identifying them becomes difficult.
- There are three solutions:

1. Labels

- Add names directly next to points
- Good for highlighting a few key data points
- Too many labels create clutter

2. Tooltips

- Appear when hovering over a point
- Show detailed information
- Keeps the chart clean
- Interactive and user-friendly

3. Annotations

- Manually highlight specific points
- Best for emphasizing key outliers
- Useful for storytelling

Making the Scatterplot Exploratory (Using Quick Filters in Tableau)

- So far, the scatterplot helped us see how extraordinary **Wayne Gretzky** was.
- But what if we don't just want to *show* something? What if we want users to *explore* the data themselves?
- That's where **Quick Filters** come in.

Why Add Quick Filters?

Quick Filters make a static chart interactive.

Instead of showing just one story, they allow users to:

- Filter by player position
- Adjust career penalty minutes (PIM)
- Adjust plus-minus (+/–)
- Discover their own insights

This transforms the chart from a presentation tool into an exploration tool.

Figure 8-9. Quick Filters turn a scatterplot into an exploratory interactive

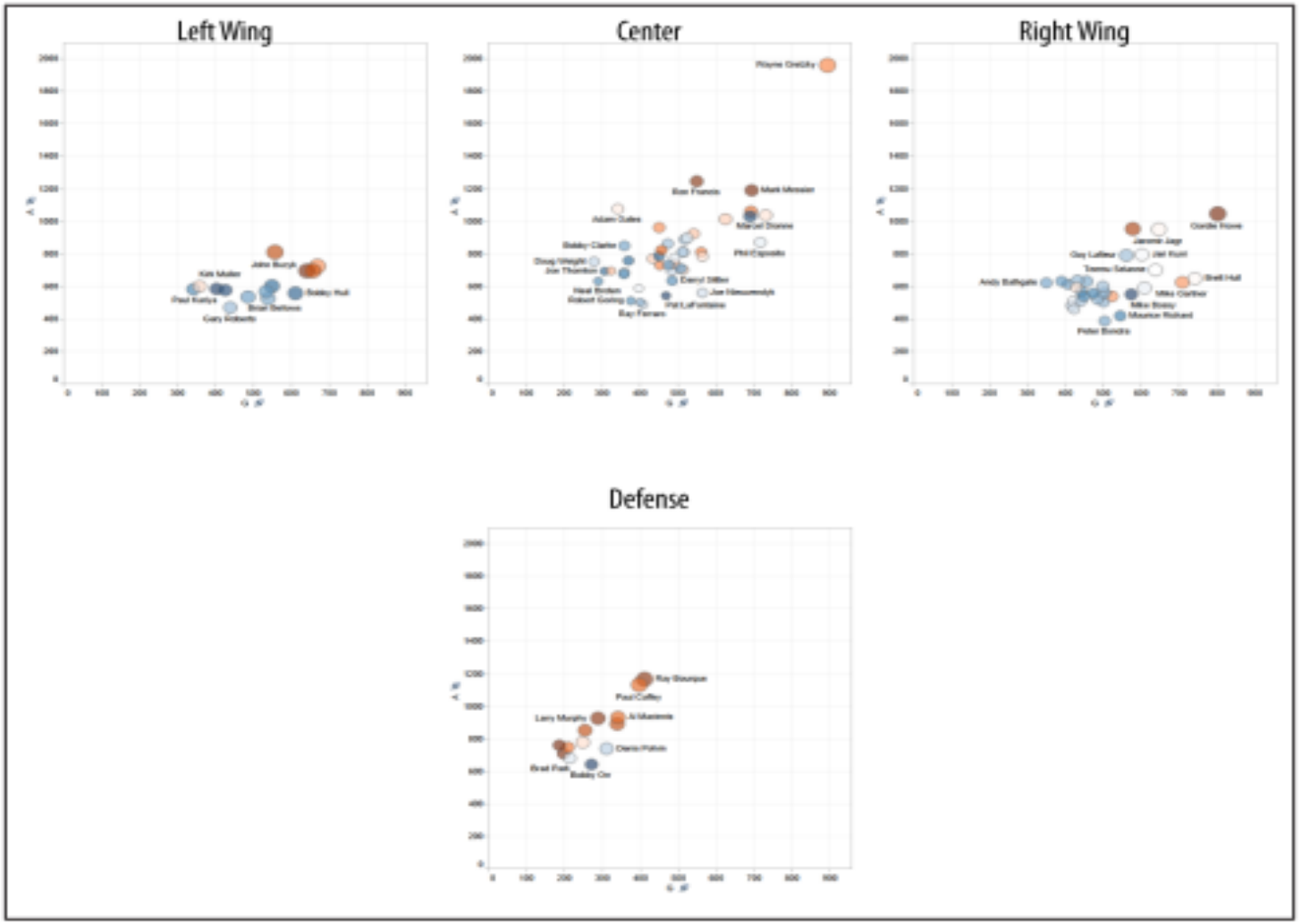


Figure 8-10. Viewing the scatterplot one position at a time

Example: Adding Filters in Tableau

- We add filters for:
- **Position (Pos)**
- **Plus/Minus (+/–)**
- **Penalty Minutes (PIM)**

Steps:

- Right-click each field (Pos, +/–, PIM)
- Select **Show Quick Filter**
- Tableau adds:
  - Drop-down list for Position
  - Range sliders for +/– and PIM
- Now users can:
- Select only defensemen
- See players with at least +100 plus-minus
- Filter by low penalty minutes

What Patterns Appear?

- When filtering by position:
- **Centers** often have very high total points (like Gretzky)
- **Defensemen** cluster differently:
  - More assists than goals
  - Played more games
- Fewer left wingers appear in the top 100
- Instead of telling users this, we let them discover it.
- That's the power of exploratory visualization.

Adding Background Images (Aesthetic Enhancement)

The author added a hockey image behind the scatter plot.

Steps:

Go to **Map → Background Images**

Add image

Lock aspect ratio

Always show entire image

- Important:

Background images should be used carefully.

Too much decoration can:

Distract from data

Reduce clarity

Data visualization should prioritize clarity over decoration.

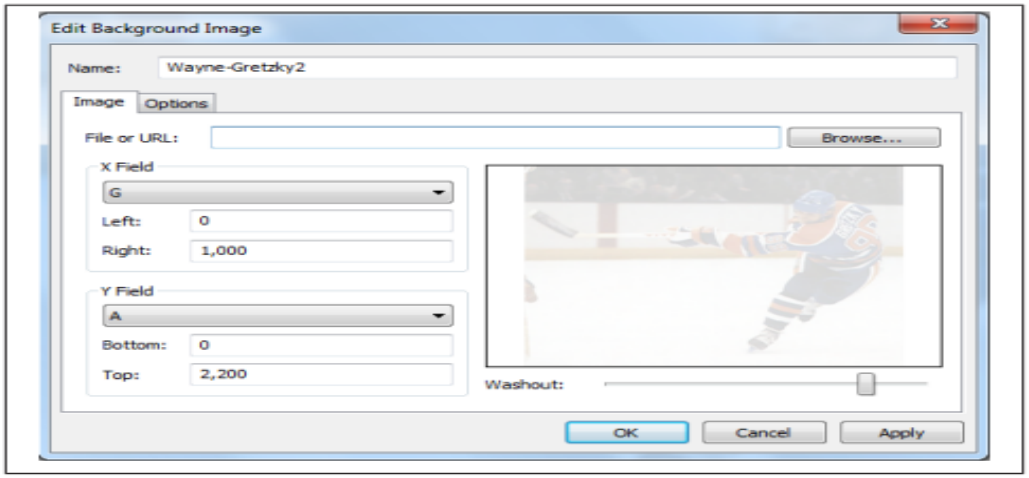


Figure 8-11. Adding a background image (photo by B. Bennett/Getty

Stacked Bars

Stacked Bar Chart for Per-Game Rates

Next, the analysis moves from totals to **rates**.

A rate is:

A ratio where numerator and denominator have different units.

- Example:
- Goals per Game
- Assists per Game
- Points per Game

Creating Calculated Fields

Since per-game stats weren't in the data:

Create calculated fields:

- Goals per Game = Goals / Games Played
- Assists per Game = Assists / Games Played
- Points per Game = Points / Games Played

Now we compare scoring efficiency instead of total career stats.

Stacked Bar Chart

Steps:

1. Player → Rows
2. Measure Values → Columns
3. Measure Names → Color
4. Keep only Goals per Game & Assists per Game
5. Sort by Points per Game (Descending)

What Do We See?

- **Wayne Gretzky** and **Mario Lemieux** stand clearly above others.
- Lemieux had a higher goals-per-game rate.
- Mike Bossy scored goals at an even higher rate than both.

 Limitation of Stacked Bars

Stacked bars are good for:

- Comparing totals
- Comparing first segment (baseline)

But they are bad for:

- Comparing middle/top segments (no common baseline)
- If comparing goals vs assists directly is more important,
  - Use a **Dual Dot Chart** instead.

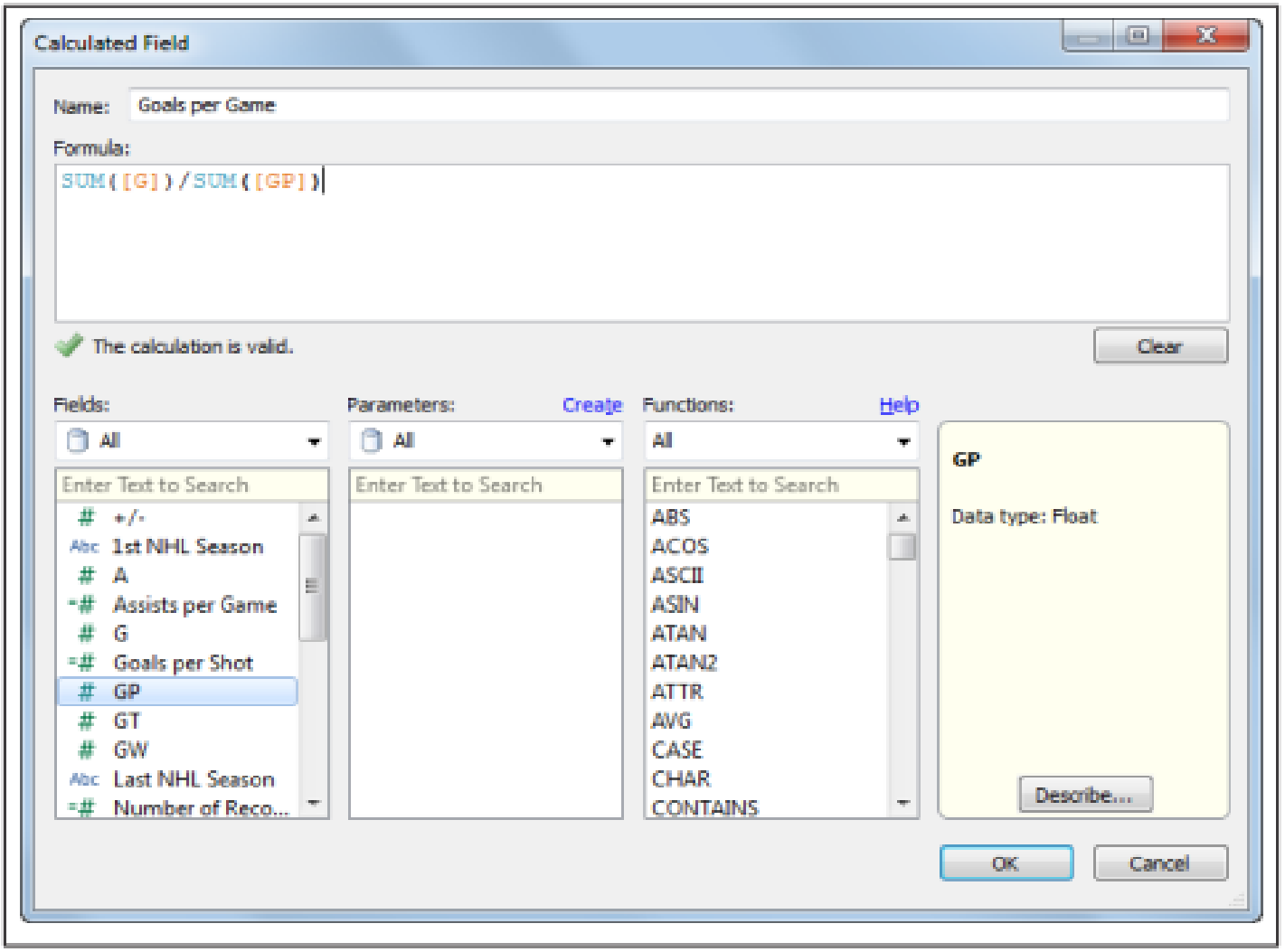


Figure 8-12. Creating per game rates with Calculated Fields

1. Drag Player to the Rows shelf, creating an alphabetical list of player names.
2. Drag Measure Values from the Measures area to the Columns shelf, creating bars that are the sum of all Measures for each player.
3. Drag Measure Names from the Dimensions area to the Colors shelf, breaking up the bars by color.
4. Drag everything except AGG(Assists per Game) and AGG(Goals per Game) out of the Measure Values area below the Marks card, which leaves only two corresponding bars for each player.
5. Change the colors by clicking on the down arrow to the right of the Measure Names area header and clicking Edit Colors.
6. Click the down arrow in the blue Player pill on the Rows shelf and select Sort, choosing a descending sort order by the Points per Game field.

Regression and Trend Lines

Now the analysis studies the relationship between:

- Shots (Independent Variable → X-axis)
- Goals (Dependent Variable → Y-axis)

Why?

Because shots can lead to goals.

Adding Trend Lines in Tableau

Steps:

1. Create scatterplot (Shots vs Goals)
2. Right-click → Trend Lines → Show Trend Lines

You can:

- Show separate trend lines per position
- Or show one overall trend line

Understanding the Trend Line

The trend line shows:

- A positive slope → More shots = More goals
- Strong correlation
- Statistical equation
- p-value (very small → relationship is significant)

R-Squared (R^2)

- R^2 tells us:

How well the model explains the variation in data.

If R^2 is high:

- Strong fit

If R^2 is low:

- Weak fit

When forcing the line through zero:

- R^2 becomes artificially high
- Removing the constraint gives more realistic value

Interpreting the Scatterplot with Trend Line

Points above the line:

- Higher shooting efficiency
- Fewer shots needed per goal

Points below the line:

- Lower efficiency
- More shots needed per goal

The farther a point is from the line:

- The more unusual that player is compared to average trend

Overall Summary

This chapter teaches three big ideas:

1. Turn static charts into exploratory dashboards using Quick Filters.
2. Use calculated fields to create rates for deeper analysis.
3. Use regression and trend lines to measure strength of relationships.

- It shows how visualization moves from:
- Showing totals
  → To showing efficiency
  → To measuring statistical relationships

Quadrant Chart

The Quadrant Chart (Explained Clearly)

A **Quadrant Chart** is a scatterplot divided into **four sections** using two reference lines:

- One **vertical line**
- One **horizontal line**

These lines split the chart into four parts (quadrants), helping us compare performance relative to averages.

Instead of just seeing correlation, we now see **categorization**.

Example: Shots vs Goals (Hockey Players)

Suppose we already created a scatterplot:

- X-axis → Shots
- Y-axis → Goals
- Each dot → One player

Now we divide the chart into four quadrants using the **mean (average)** values.

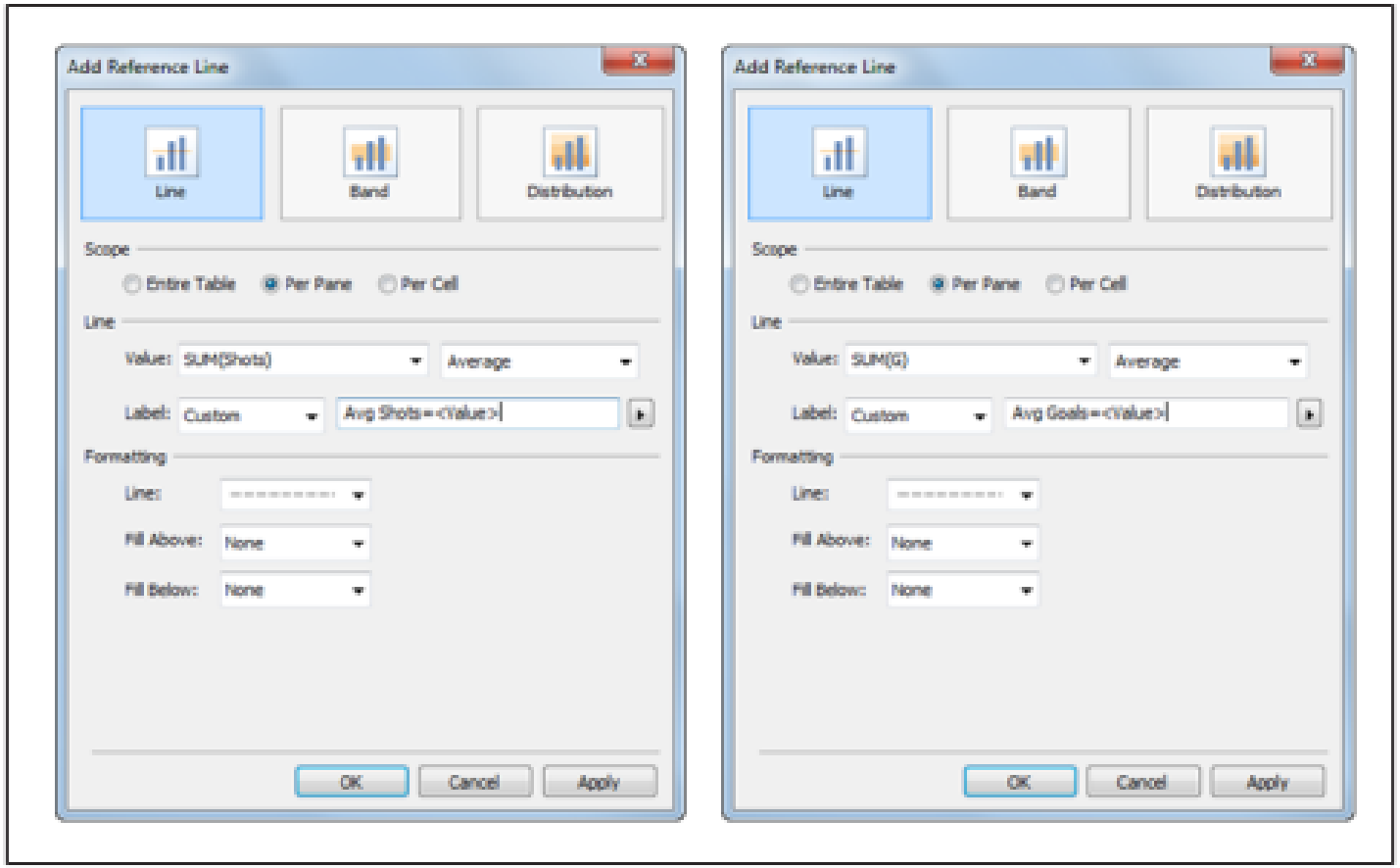


Figure 8-20. Adding vertical (left box) and horizontal (right box) reference lines

Changes Over Time

Changes Over Time (Explained Clearly)

When communicating data, **time is one of the most important dimensions**.

Almost every business metric includes time:

- Monthly unemployment rate
- Quarterly earnings per share
- Daily stock prices
- Annual revenue growth

If time is not shown, the data often loses meaning.

Earlier, we analyzed totals (like total career goals).

But now we ask a deeper question:

How do values change over time?

Why Time Analysis Is Important

- Studying time helps answer:
- Is performance improving or declining?
- Are there upward or downward trends?
- Are there seasonal patterns?
- Is variation meaningful or just random noise?
- Did a specific event cause a shift?
- Time unlocks trend insights that totals cannot show.

Origin of Time Charts

- The first line chart was created in 1786 by:

William Playfair

- In his book:

The Commercial and Political Atlas

He showed imports and exports over time using two lines.

This introduced the world to:

- Time on the X-axis
- Quantity on the Y-axis
- Left-to-right progression

Even today, we follow this same structure.

The Line Chart (Most Common Time Chart)

- The most effective way to show change over time is a **line chart**.

Standard Structure:

- X-axis → Time (left to right)
- Y-axis → Measure (Sales, Profit, Price, etc.)

This layout feels natural and easy to understand.

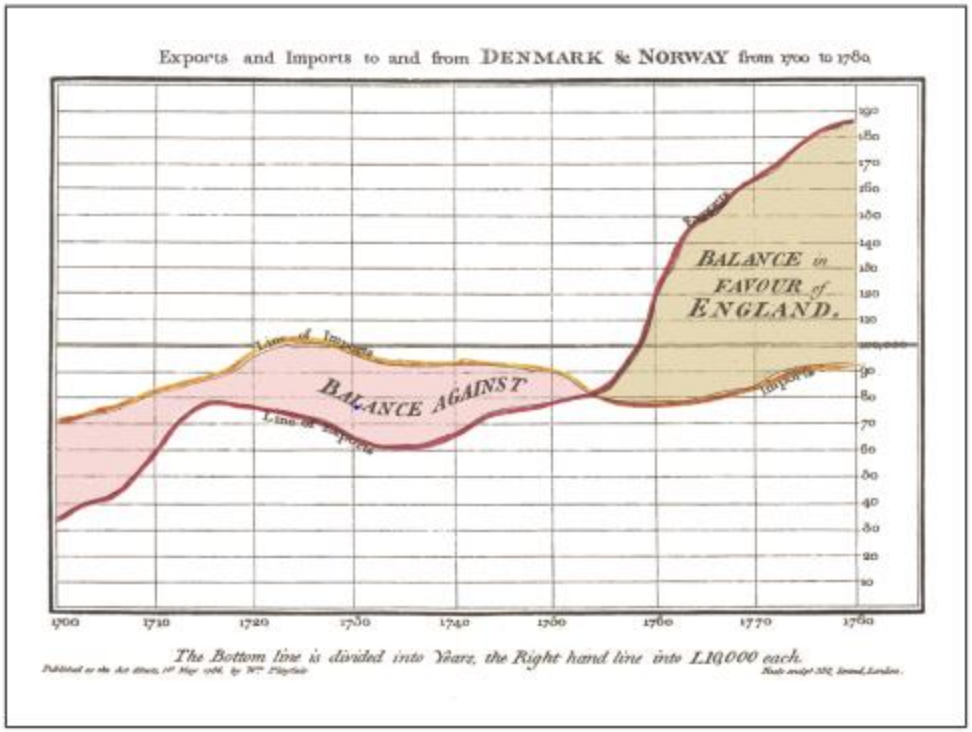


Figure 9-1. The first line chart, by William Playfair in 1786

The Line Chart

Line Chart Example: Strikeouts Over Time (Baseball Data)

We are working with a **pre-aggregated dataset**, meaning the data is already summarized.

- The dataset contains:

1. **Year**
2. **Average Strikeouts per Game (SO)**
3. **Number of Teams (Tms)**
4. **Total Number of Pitchers (#Pitch)**

Our goal:

- Understand how strikeouts per game changed from 1871 to present.

- **Creating the Line Chart in Tableau**

Step 1:

- Drag **SO (Strikeouts per Game)** → Rows shelf

Step 2:

- Right-click drag **Year** → Columns shelf
  Select **Year (continuous)**
- This creates a **line chart** where:
- X-axis = Year (Time)
- Y-axis = Average Strikeouts per Game

What Does the Chart Show?

The pattern shows:

✔ A long-term upward trend

✔ Strikeouts per game have increased over 150+ years

- At first glance, it looks dramatic.
- But when we add a **Trend Line**:
- Right-click chart → Trend Lines → Show Trend Lines
- We discover something interesting:

 The increase is only **0.033 strikeouts per game per year**

That means:

- Every 10 years → increase of 0.33 strikeouts per game
- Roughly 1 extra strikeout every 3 games per decade
- So the trend is steady, but slow.

 Observing Deviations from Trend

- The line does not increase smoothly.
- There are noticeable fluctuations:

Before 1920:

- High variability
- Strike zone rules were changing
- Batters could request pitch height (before 1887)
- Baseball rules were unstable
- This caused unpredictable strikeout patterns.

- **Adding a Vertical Reference Line (1969)**
- To highlight this event:
- Right-click chart
- Select **Add Reference Line**
- Enter value = 1969
- Apply to entire table
- Now viewers can clearly see:
- Before 1969
- After 1969
- This helps tell the story visually.

Open the data set and worksheet.

1. Left-click the Measure SO, and drag and drop it onto the Rows shelf.
2. Right-click on the Measure Year, and drag it to the Columns shelf. When you release the right mouse button, select the top option, Year.

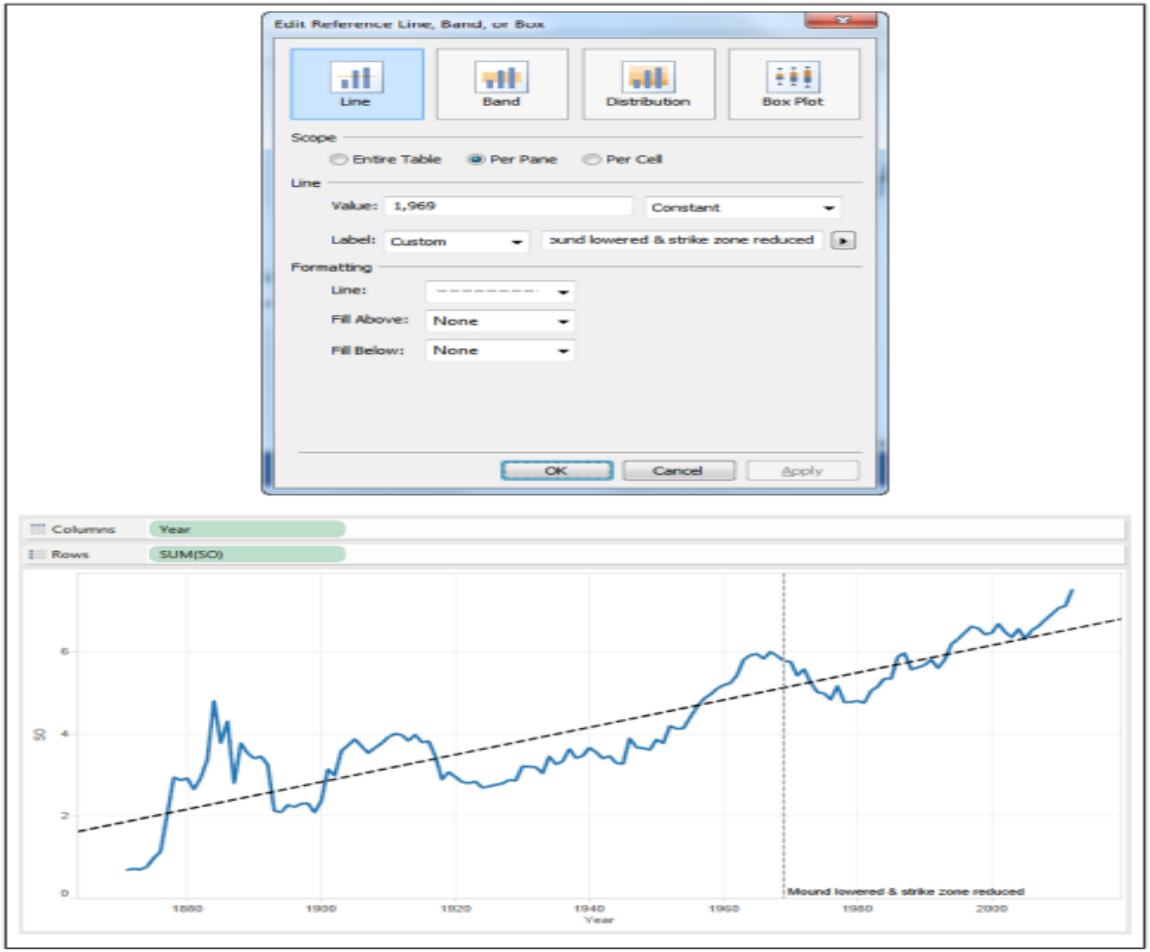


Figure 9-5. Updated line chart with added reference line

The Dual-Axis Line Chart

A dual-axis line chart, or combo chart, plots two distinct data series on a single graph using a shared X-axis (usually time) and two independent vertical Y-axes. This visualization enables direct comparison of metrics with different units or scales, such as revenue and profit percentage, to identify correlations, trends, and relationships.

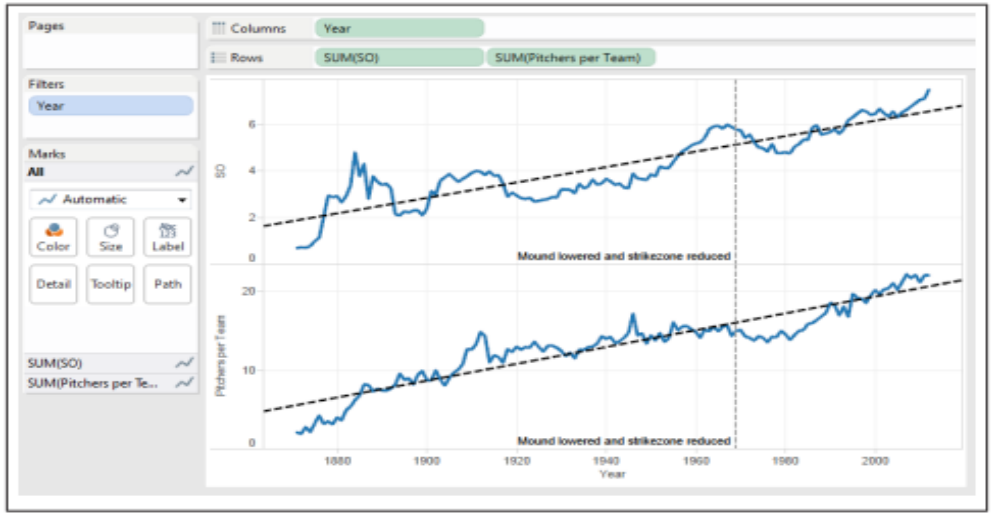


Figure 9-6. Two line charts, shown one above the other

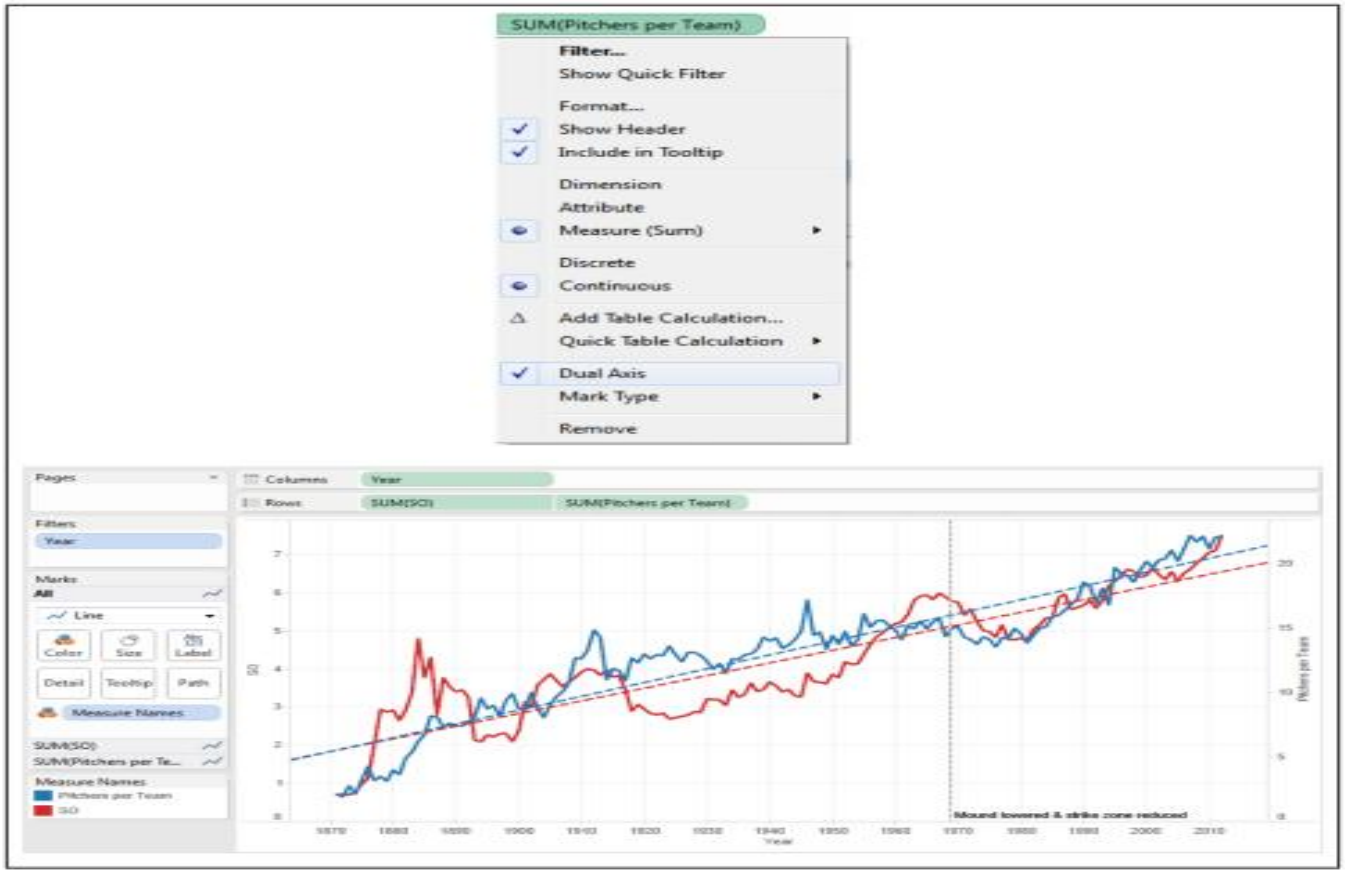


Figure 9-7. The dual-axis line chart

Notice that a number of elements have changed:

a. • First, and most obviously, the lines have been placed together and Tableau has automatically colored one chart (and its trend line) red to differentiate it from the other.

b. • Second, notice that the second y-axis (Pitchers per Team) has been placed on the righthand side of the chart, and that it has a different scale than the lefthand y-axis (SO).

c. • Third, notice that the green pills in the Rows shelf have changed shape: their adjacent edges are flat instead of round.

d. • Finally, notice that the Marks card area to the left of the lines now has three different sections: All, SUM(SO), and SUM(Pitchers per Team). Click on these headers to see how the Marks card changes to allow us to control each line separately, or both together.

The Connected Scatterplot

The Connected Scatterplot in Tableau

A connected scatterplot is an innovative way to show **how two variables change over time**, without putting time on the x-axis.

Instead of:

- X-axis = Year
- Y-axis = Strikeouts

We use:

- X-axis = Pitchers per Team
- Y-axis = Strikeouts per Game

Time is shown by connecting points in chronological order.

Step 1: Create a Basic Scatterplot

- In Tableau:
- Drag **Pitchers per Team** → Columns
- Drag **SO (Strikeouts per Game)** → Rows
- Marks type → Circle
- You now have a scatterplot.

Step 2: Add a Trend Line

- Right-click in chart
- Select **Trend Lines → Show Trend Lines**
- **Interpretation:**
- The trend line shows:
- **$R^2 = 0.71$**
- This means **71% of variation in strikeouts is explained by pitchers per team**.
- That is a strong positive relationship.

Step 3: Add Time Information

- **Option 1: Add Year to Color & Label**
- Drag **Year** → Color
- Drag **Year** → Label
- Now:
- Early years appear lower-left
- Recent years appear upper-right
- This gives some idea of progression.

Step 4: Create Connected Scatterplot

- To show time clearly:
- Change **Marks type → Line**
- Drag **Year** → Path
- Now Tableau connects each year in chronological order.
- The dots now move like a "journey" through time.

Step 5: Filter to 1981–Present

- Because too many points make it messy:
- Drag **Year** → Filters
- Select range → From 1981 to Present
- Now the pattern becomes clear.

Key Insight (Very Important)

- Between **2007 and 2012**:
- Strikeouts per game increased
- Pitchers per team stayed almost constant
- This is important because:
- It shows that strikeouts increased even without adding more pitchers.
- So something else caused the increase:
- Higher pitch velocity
- Better analytics
- More aggressive hitting styles
- Advanced training methods
- This is something we could NOT easily see in a normal time series chart.

Steps to connected to scatterplot

- I prefer to format connected scatterplots to look like the view shown in Figure 9-13. To do this, the following steps are required:

1. Drag a second Pitchers per Team pill onto the Columns shelf to the right of the first and make it a dual-axis plot. 160 | Chapter 9: Changes Over Time www.it-ebooks.info
2. Right-click in the top x-axis and select Synchronize Axis, then un- check “Show Header.” The top x-axis should disappear.
3. Take Measure Names off of the Color shelf of “All” by dragging and dropping it back into the Measures area.
4. Change Mark type for Pitchers per Team (2) from Line to Circle, and take SUM(Year) off of the Labels shelf.
5. In the Marks panel for Pitchers per Team (2), change the Circle color from blue to white and give each circle a gray border.
6. In the Marks panel for Pitchers per Team, change the color of the line to gray and reduce the size.

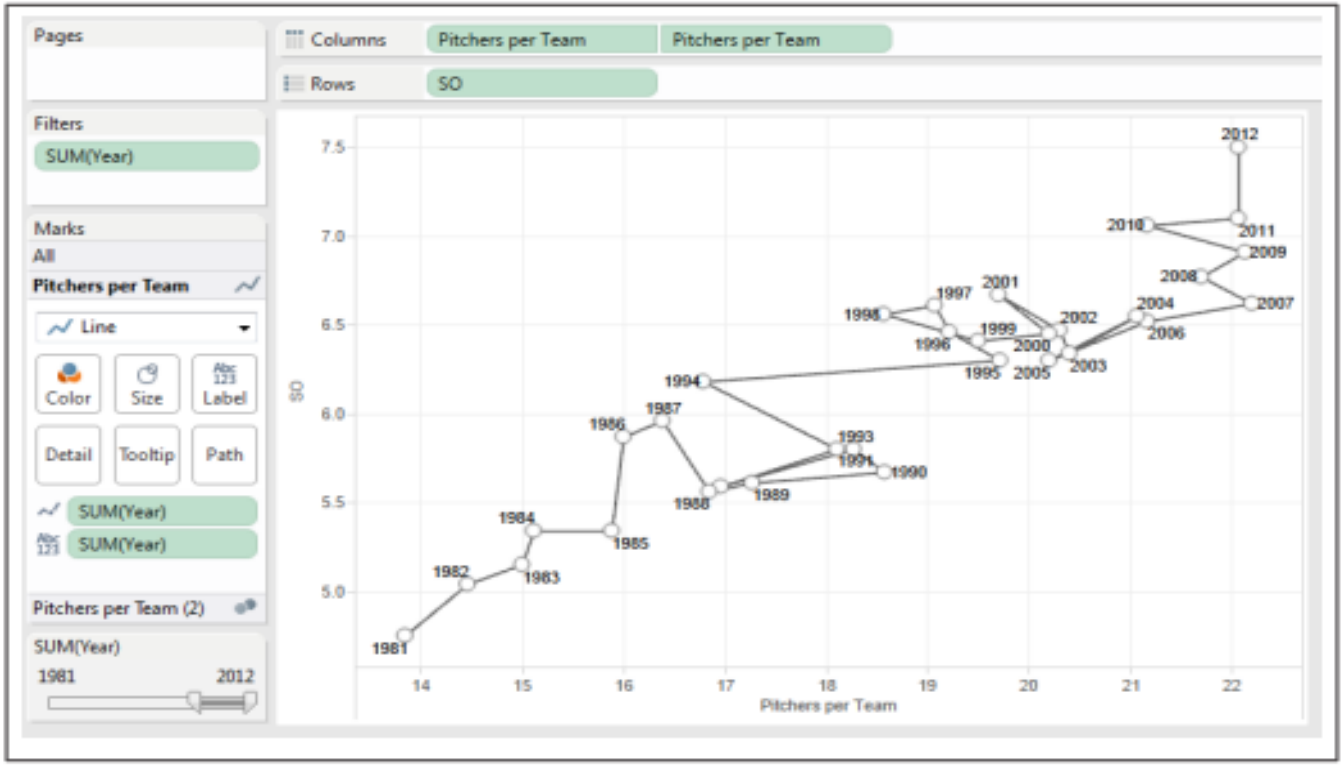


Figure 9-13. The formatted connected scatterplot

The Date Field Type and Seasonality

Date Field Type and Seasonality in Tableau

In this example, we analyze **New York City rat sightings data** to understand:

- When do most rat sightings occur?
- Is the trend increasing or decreasing?
- Is there seasonality?
- Can we forecast future sightings?

The dataset contains a field:

- **Created Date** (format: mm/dd/yyyy)
- Each row represents **one reported rat sighting**.

1.Creating a Basic Time Series

- **Steps in Tableau:**

1. Drag **Created Date** → Columns
2. Drag **Number of Records** → Rows

- Tableau automatically creates a **Yearly Line Chart**.
- **What Happens?**
- Tableau treats the date as **Year(Created Date)**
- The pill is **blue**, meaning it is **Discrete**
- Discrete = categorical grouping (separate years)
- This produces an **annual timeline**.

2.Drill Down Using + and – (Date Hierarchy)

Next to Year(Created Date), you see a **+ symbol**.

Click:

- **+ → Breaks into Quarter**
- Click + again → Breaks into Month
- Click + again → Breaks into Day
- Click – → Aggregates back up

This is called **date hierarchy drill-down**.

It allows us to increase or decrease aggregation.

3.Discrete vs Continuous Dates

- By default:
- Year(Created Date) is **Discrete (Blue)**

To change:

1. Click dropdown on Year(Created Date)
2. Choose options below the divider (Year, Month, Week, Day)

- Now the pill turns **Green** → meaning **Continuous**

4.Identifying Seasonality

When viewed by **Month (Continuous)**:

We notice:

1. Summer → High rat sightings
2. Winter → Low rat sightings

This repeating yearly pattern is called:

Seasonality

- Seasonality = regular, repeating patterns within a time period.

5.Comparing Boroughs

- To check if seasonality differs by location:
- **Step:**

Drag **Borough** → Color shelf

- Now:
- Each borough has its own line
- All boroughs show similar seasonal peaks in summer
- Seasonality holds true across locations

6.Cleaning Data (Outliers)

There was:

- One sighting labeled “Unspecified”
- Appeared as a single brown dot

We can:

- Click the mark
- Choose **Exclude**
- Data cleaning improves clarity.

7.Adding Forecast in Tableau

- Tableau can predict future values.

Steps:

- Right-click in graph area
- Select **Forecast → Show Forecast**

Now:

- Future months appear in lighter shade
- Forecast is based on historical trend and seasonality

Forecast Options

Right-click → **Forecast → Forecast Options**

You can modify:

- Forecast length
- Seasonality inclusion
- Model type

To see details:

- Right-click → **Describe Forecast**

What the Forecast Shows

- Because rat sightings show:
- Clear seasonality
- Slight upward trend
- Tableau predicts:
- Continued seasonal peaks in summer
- Continued winter drops

The Timeline

The timeline was first introduced in 1765 by
Joseph Priestley
in his work *Chart of Biography*.

In that early timeline:

- Each bar represented the lifespan of a historical figure.
- Time moved from left to right.
- It visually showed overlap between lives.

Today, timelines are widely used in:

- Project management
- Historical analysis
- Event tracking
- In Tableau, we create timelines using the **Gantt Bar chart**.

Example: U.S. Presidents Timeline

We use milestone data for presidents of the
United States of America.

The dataset includes:

- President
- Born
- Died
- Took Office
- Left Office
- Party

Step 1: Create Calculated Fields

- Because:
- Some presidents are still alive
- The current president hasn't left office

We must handle NULL values.

Calculated Field 1: Life Span

IF ISNULL([Died])
THEN TODAY() - [Born]
ELSE [Died] - [Born]
END

Calculated Field 2: Time in Office

IF ISNULL([Left Office])
THEN TODAY() - [Took Office]
ELSE [Left Office] - [Took Office]
END

- These formulas:
- Calculate duration correctly
- Replace missing dates with today's date

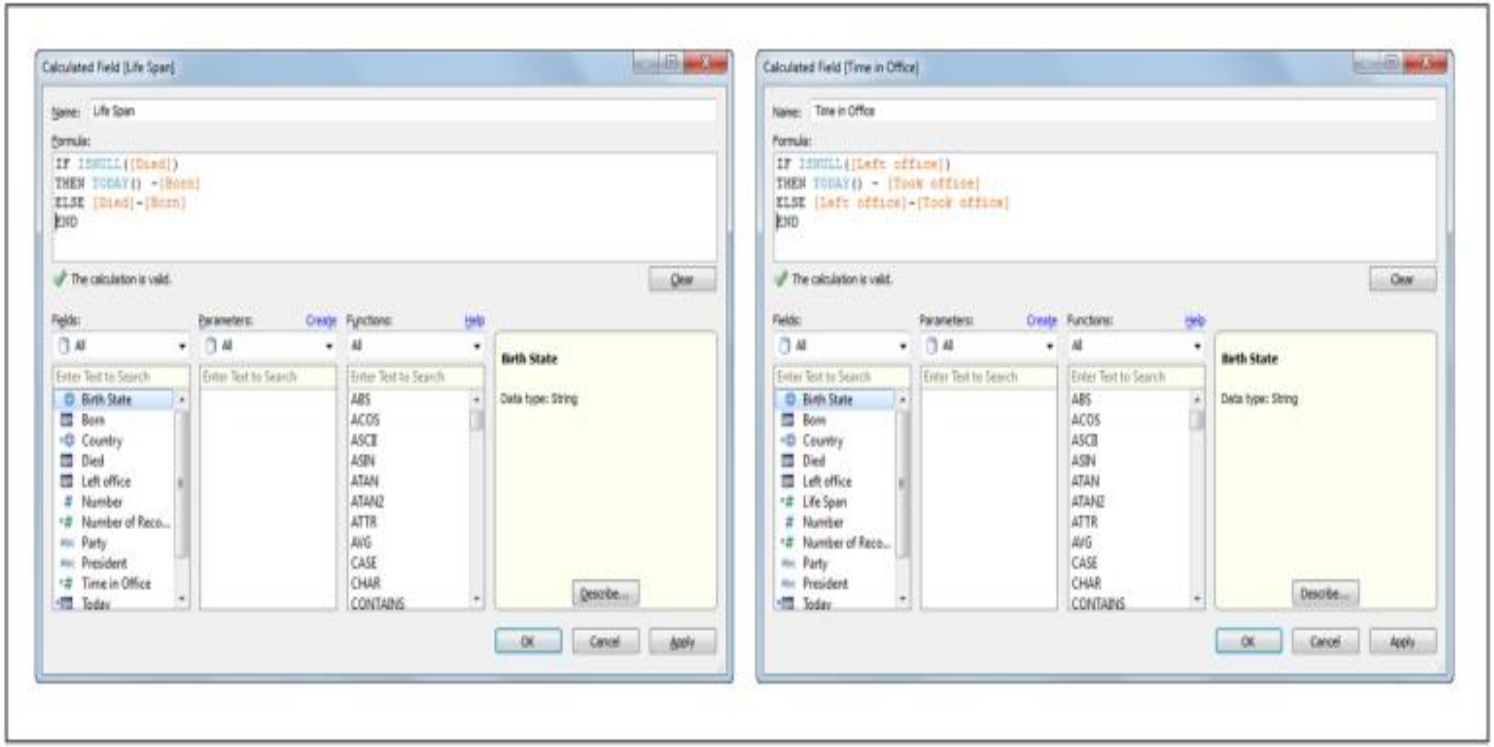


Figure 9-21. Calculating life span and time in office

Step 2: Create Basic Gantt Timeline

Steps in Tableau:

1. Drag **Took Office** → Columns
    1. Change to **Day (Continuous)** (Green pill)
2. Change Marks type → **Gantt Bar**
3. Drag **Time in Office** → Size shelf
4. Drag **Party** → Color shelf

You now see:

- Horizontal bars
- Each bar = duration in office
- Colored by political party

Step 3: Break Out by President

Drag **President** → Rows

This lists presidents alphabetically.

To sort chronologically:

- Click President pill → Sort
- Sort by **Took Office (Ascending)**

Now the timeline shows presidents in historical order.

What This Timeline Shows

- Length of each presidency
- Party transitions
- Overlaps between administrations
- Short terms (e.g., deaths in office)

It becomes easy to visually spot:

- Presidents who died in office
- Long vs short administrations
- Political dominance periods

- **Step 4: Add Life Span (Dual Axis Timeline)**

Now we enhance the visualization by adding lifespan.

Steps:

1. Drag **Born** → Left of Took Office on Columns
2. Change to **Day (Continuous)**
3. In Marks for Born:
    1. Replace Time in Office with **Life Span**
    2. Remove Party from Color
4. Right-click Took Office → **Dual Axis**
5. Synchronize Axis
6. Hide top header

The Slopegraph

A slopegraph is used to show **change between two points in time**.

Step 1:

- Get the Data League tables for each season up to a chosen game number are avail- able online. The teams' results up to game 15 for both the 2012/2013 and 2013/2014 seasons were copied and pasted into an Excel spread- sheet, with an added column for Year, as shown in Figure 9-26.

	A	B	C	D	E	F	G	H	I	J	K	L
1	POS	Year	LP	CLUB	P	W	D	L	GF	GA	GD	PTS
2	1	2013	-1	Arsenal	15	11	2	2	30	11	19	35
3	2	2013	-2	Liverpool	15	9	3	3	34	18	16	30
4	3	2013	-3	Chelsea	15	9	3	3	30	17	13	30
5	4	2013	-4	Manchester City	15	9	2	4	41	15	26	29
6	5	2013	-5	Everton	15	7	7	1	23	14	9	28
7	6	2013	-6	Tottenham	15	8	3	4	15	16	-1	27
8	7	2013	-7	Newcastle	15	8	2	5	20	21	-1	26
9	8	2013	-8	Southampton	15	6	5	4	19	14	5	23
10	9	2013	-9	Manchester	15	6	4	5	22	19	3	22
11	10	2013	-11	Swansea City	15	5	4	6	21	20	1	19
12	11	2013	-10	Aston Villa	15	5	4	6	16	18	-2	19
13	12	2013	-13	Hull City	15	5	3	7	13	19	-6	18
14	13	2013	-12	Stoke City	15	4	5	6	15	20	-5	17
15	14	2013	-14	Norwich City	15	5	2	8	14	28	-14	17
16	15	2013	-15	West Bromwich	15	3	6	6	17	21	-4	15
17	16	2013	-16	Cardiff City	15	3	5	7	11	22	-11	14
18	17	2013	-17	West Ham	15	3	4	8	13	19	-6	13
19	18	2013	-18	Fulham	15	4	1	10	14	26	-12	13
20	19	2013	-19	Crystal Palace	15	4	1	10	10	22	-12	13
21	20	2013	-20	Sunderland	15	2	2	11	12	30	-18	8
22	1	2012	-1	Manchester	15	12	0	3	37	21	16	36
23	2	2012	-2	Manchester City	15	9	6	0	28	11	17	33
24	3	2012	-3	Chelsea	15	7	5	3	25	16	9	26
25	4	2012	-4	Tottenham	15	8	2	5	28	23	5	26
26	5	2012	-5	West Bromwich	15	8	2	5	24	19	5	26
27	6	2012	-6	Everton	15	5	8	2	25	19	6	23
28	7	2012	-7	Swansea City	15	6	5	4	23	17	6	23

Figure 9-26. The first 15 game results in Excel

If you read Andy's blog post, you'll notice that this spreadsheet is structured differently than Andy's. He had one column for 2012/2013 results and another column for 2013/2014 results. I've structured the spreadsheet in this way so that I can use `Year` as a Measure in Tableau.

step 2: Connect Tableau This is a very straightforward step: Open Tableau, click Connect to Data, and find your results spreadsheet.

Step 3: Create a Parameter and Matching Calculated Field Before creating the slopegraph, let's make a Parameter that will allow users to choose which stat to chart. This is a technique we'll explore more in later chapters as well. Right-click anywhere in the Dimensions or Measures panel to the left, and select Create New Parameter. Fill out the dialog box as shown in Figure 9-27

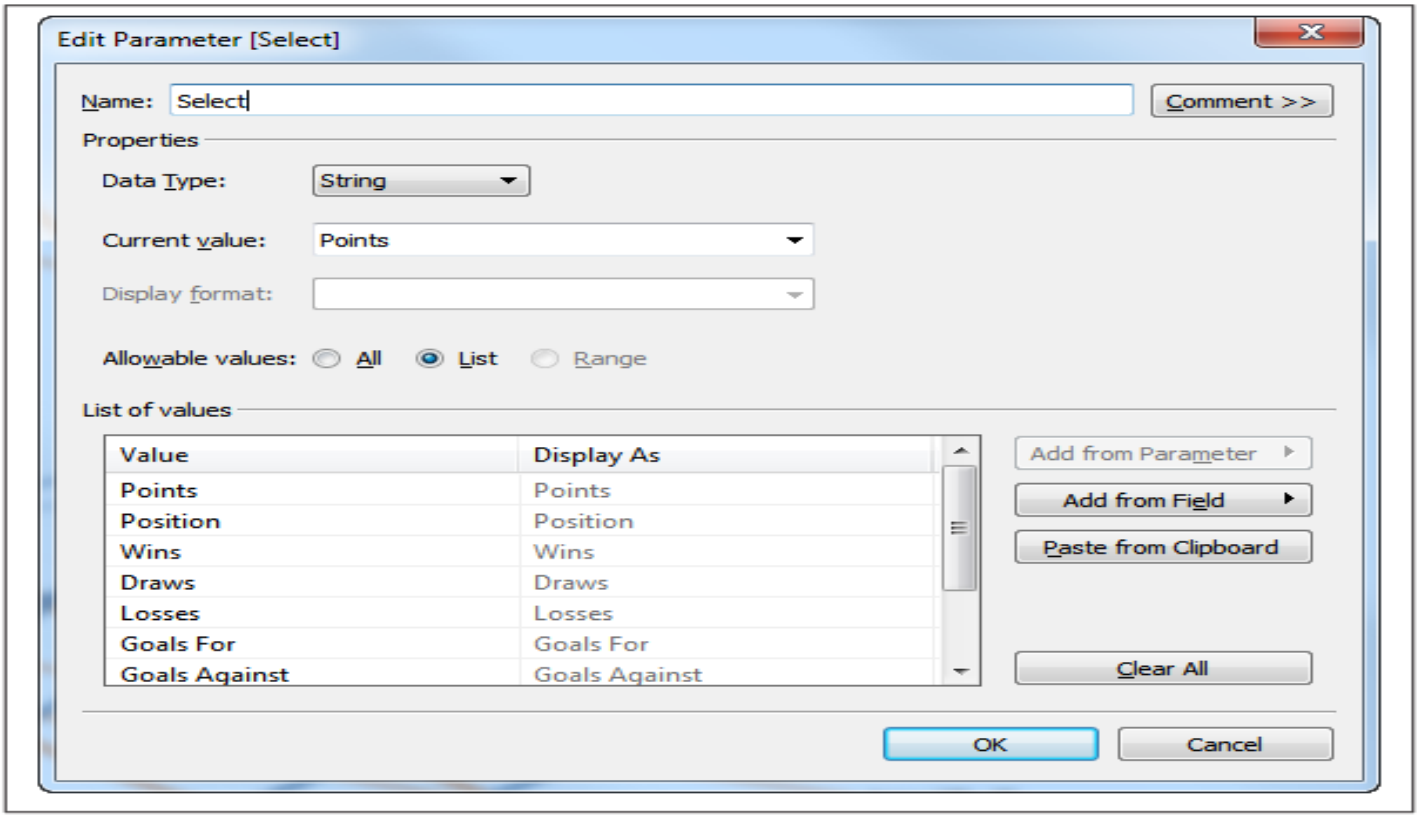


Figure 9-27. Creating a Parameter in Tableau

Click *OK*, and then right-click the newly created Parameter in the area to the bottom left and select *Show Parameter Control*. You'll see a drop-down select appear in the upper right. You can use this to change the value of the parameter.

We now need to create a calculated field to link to the different team stats based on the user's choice. Right-click on the parameter, select *Create Calculated Field*, and fill out the dialog box as shown in

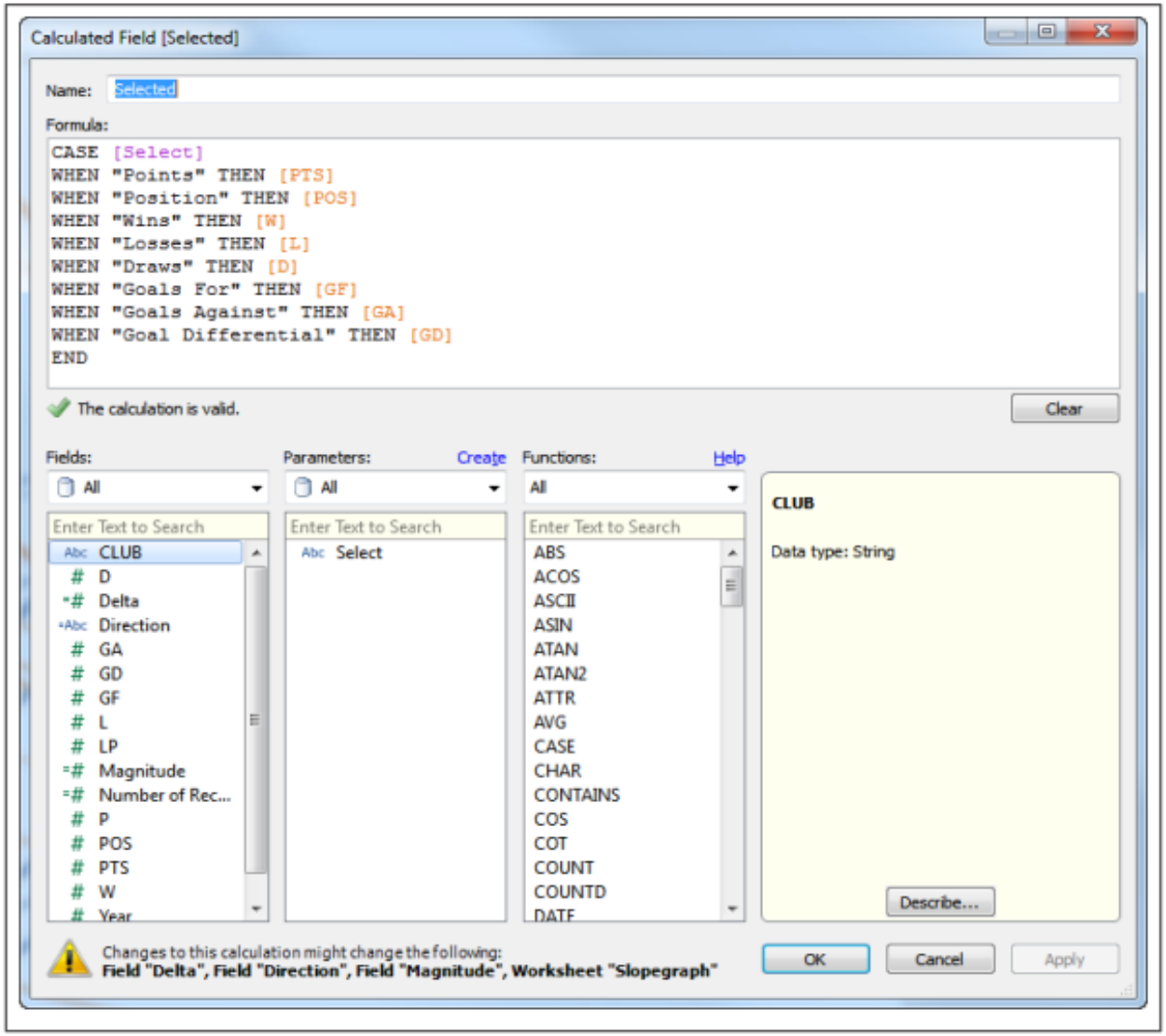


Figure 9-28. Mapping a Calculated Field to a Parameter

Step 4: Create the Basic Slopegraph Now that we have this Selected data field mapped to the Parameter, we can use it to create our basic slopegraph as follows:

- 1. Drag Year to the Columns shelf, and change it to discrete (blue pill) by clicking the down arrow and selecting Discrete.
- 2. Drag the Selected calculated field to the Rows shelf.
- 3. Change the Marks type from Automatic to Line.
- 4. Drag the CLUB Dimension to the Detail card and resize the view (making it wider).
- 5. Drag another instance of the CLUB Dimension to the Label card, and then click on Label and select Line Ends in the "Marks to Label" area.
- Step 4, numbers 1–5 are shown in Figure 9-29.

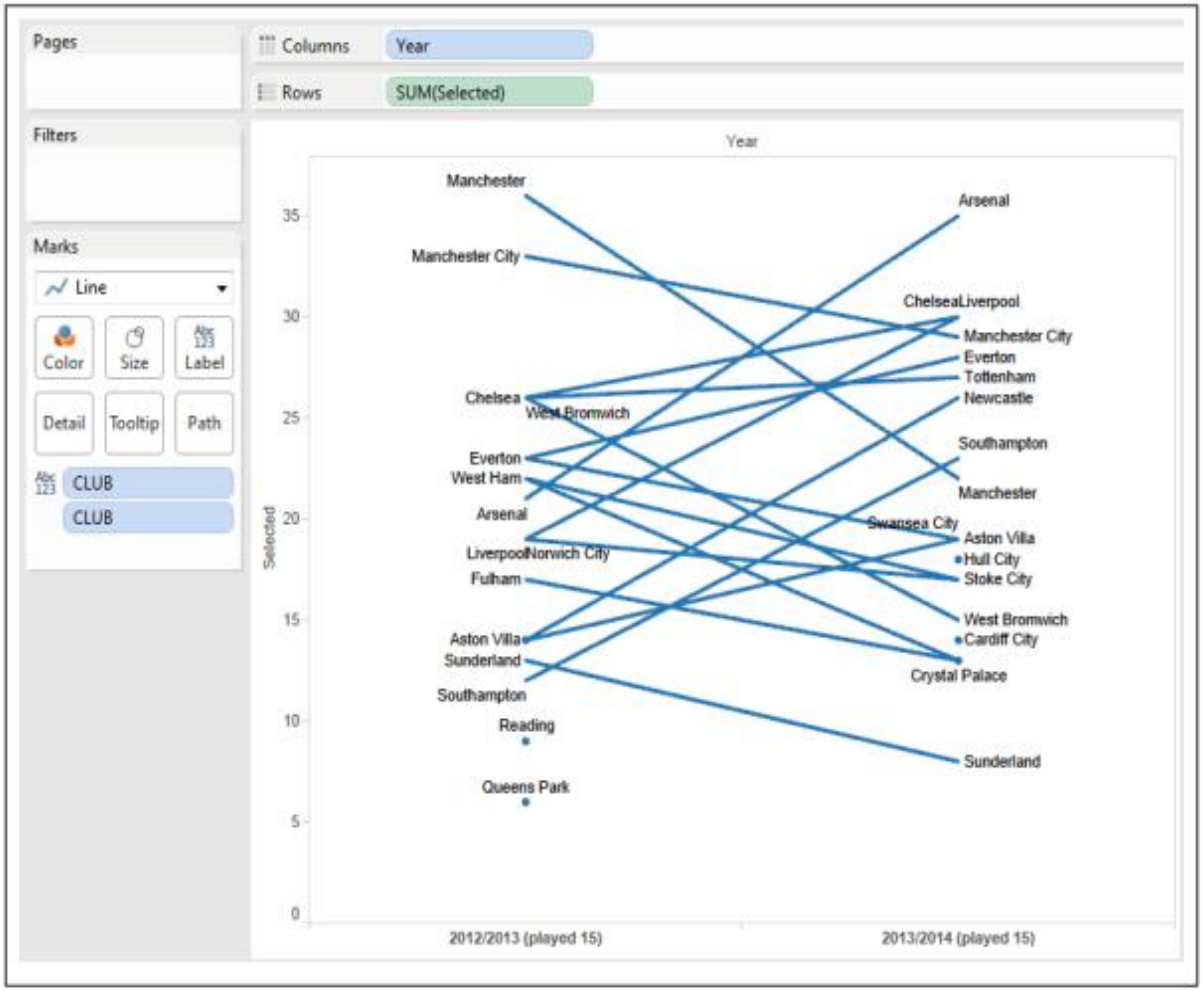


Figure 9-29. Creating a basic slopegraph in Tableau

Step 5: Add Line Coloring and Thickness In order to make the lines one color for increasing values and another color for decreasing values, and to change their thickness based on the magnitude of the change, we'll need to create three more calculated fields:

Delta

- The first calculated field computes the change in value of the se- lected statistic from one year to the next, as shown in Figure 9-30.

Direction

- The second calculated field gives one string for values that got better and another for values that got worse. This will be useful for coloring the lines. See Figure 9-31.

Magnitude

- This final calculated field yields the absolute value of the change, or the magnitude. This will be helpful for making lines thicker or thinner based on the magnitude of the change. See Figure 9-32.

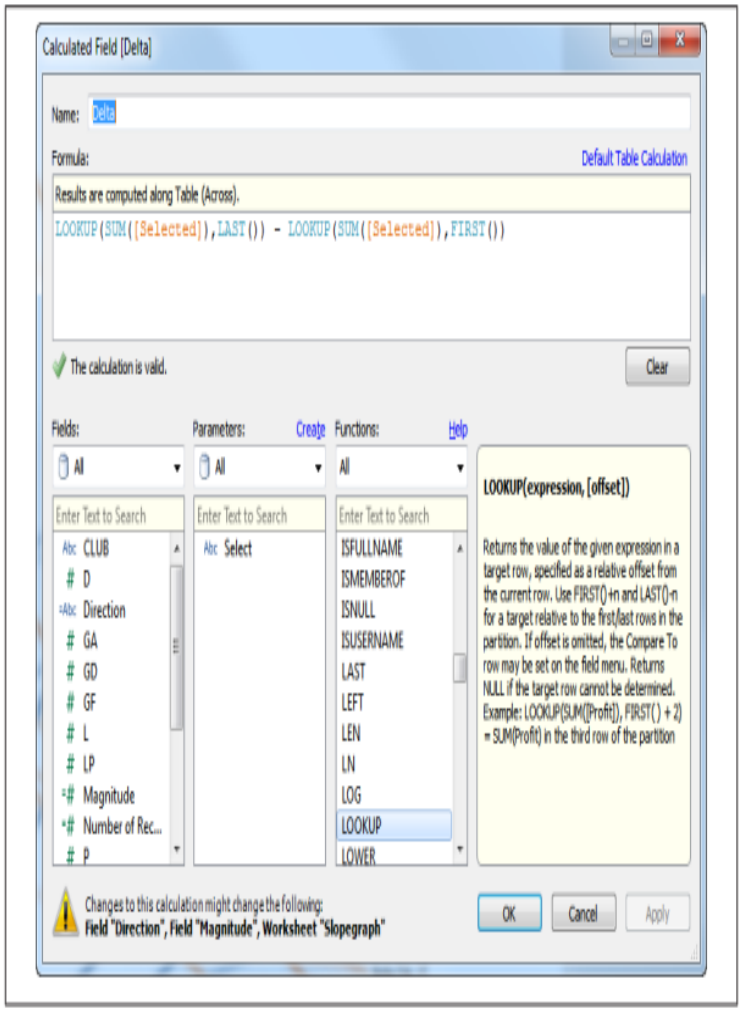


Figure 9-30. Creating a Calculated Field to compute the change in values

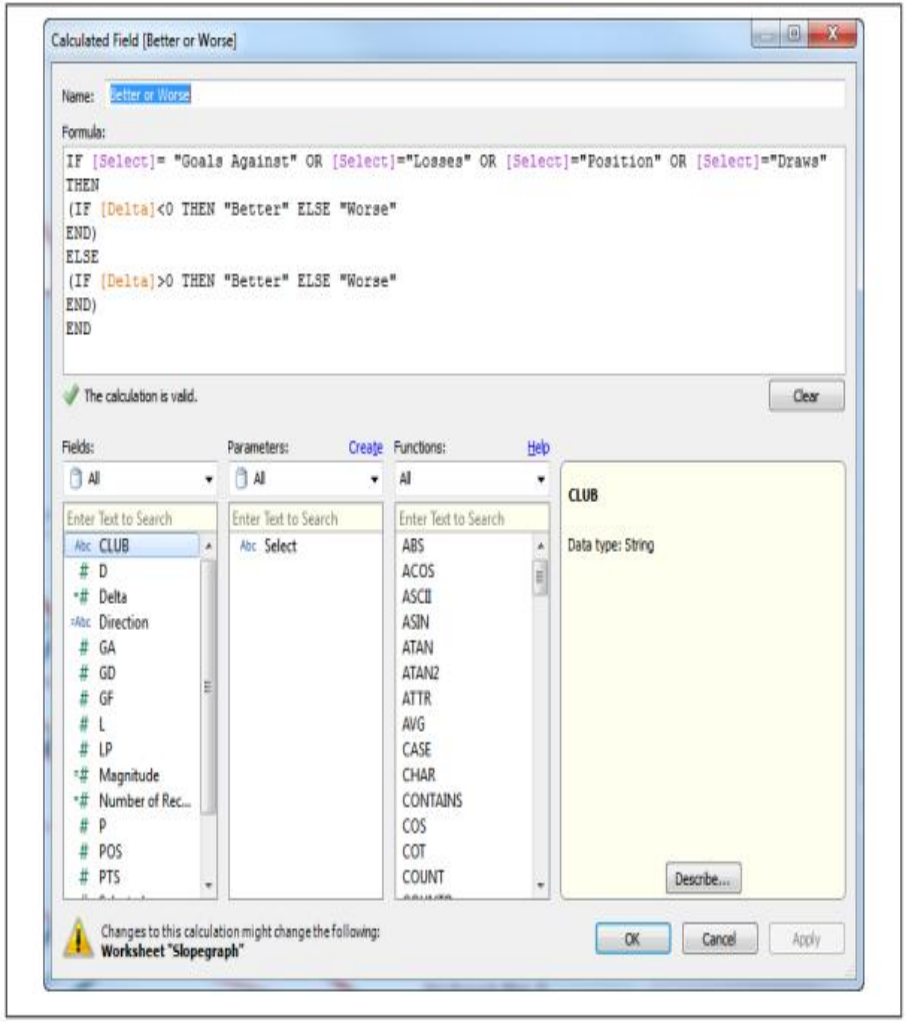


Figure 9-31. Creating a Calculated Field to categorize the direction of change

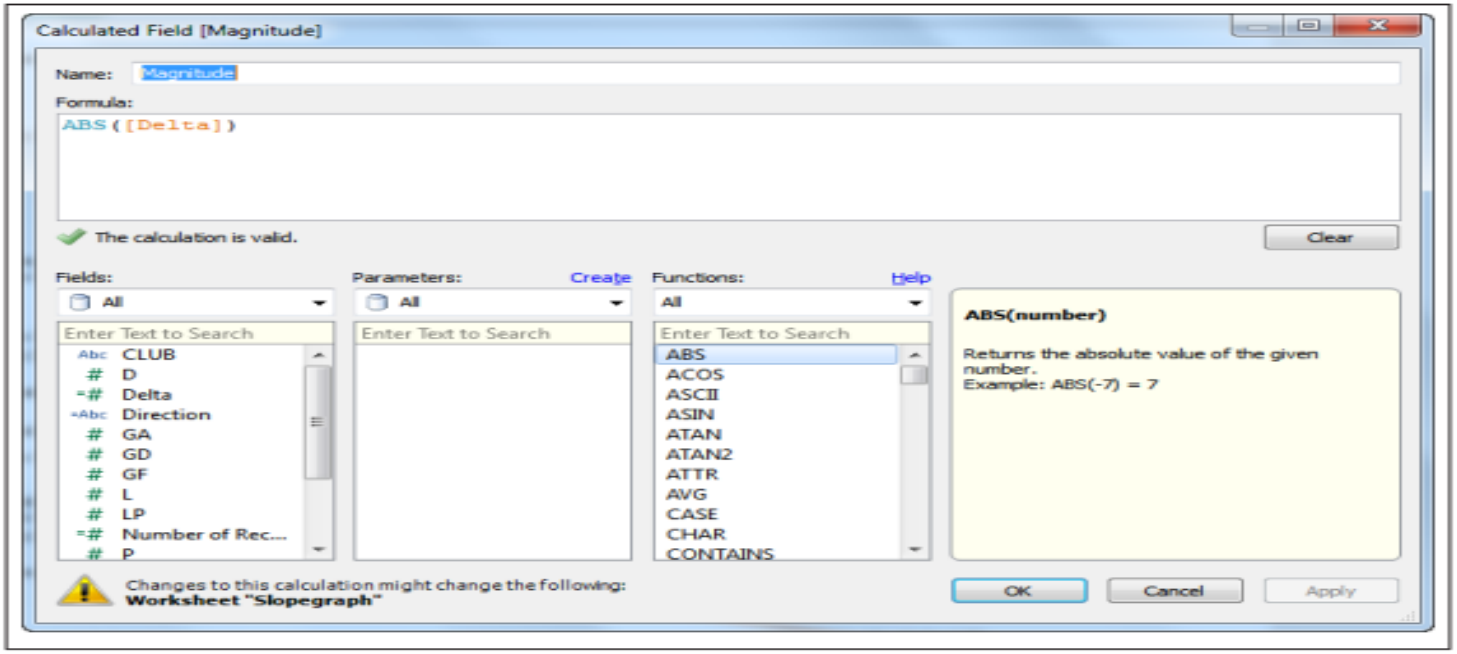


Figure 9-32. A Calculated Field that determines the absolute value, or magnitude, of the change

Now that these fields are created, let's do the following to complete the slopegraph itself:

1. Drag `Direction` to Color.
2. Drag `Magnitude` to Size.
3. Drag `Selected` to Label and change the label so that the Club name and the value are in line, with a comma separating them.
4. Filter out the Clubs that were either promoted or relegated after the 2012/2013 season.
5. Clean up the fonts (change them all to Gill Sans MT).

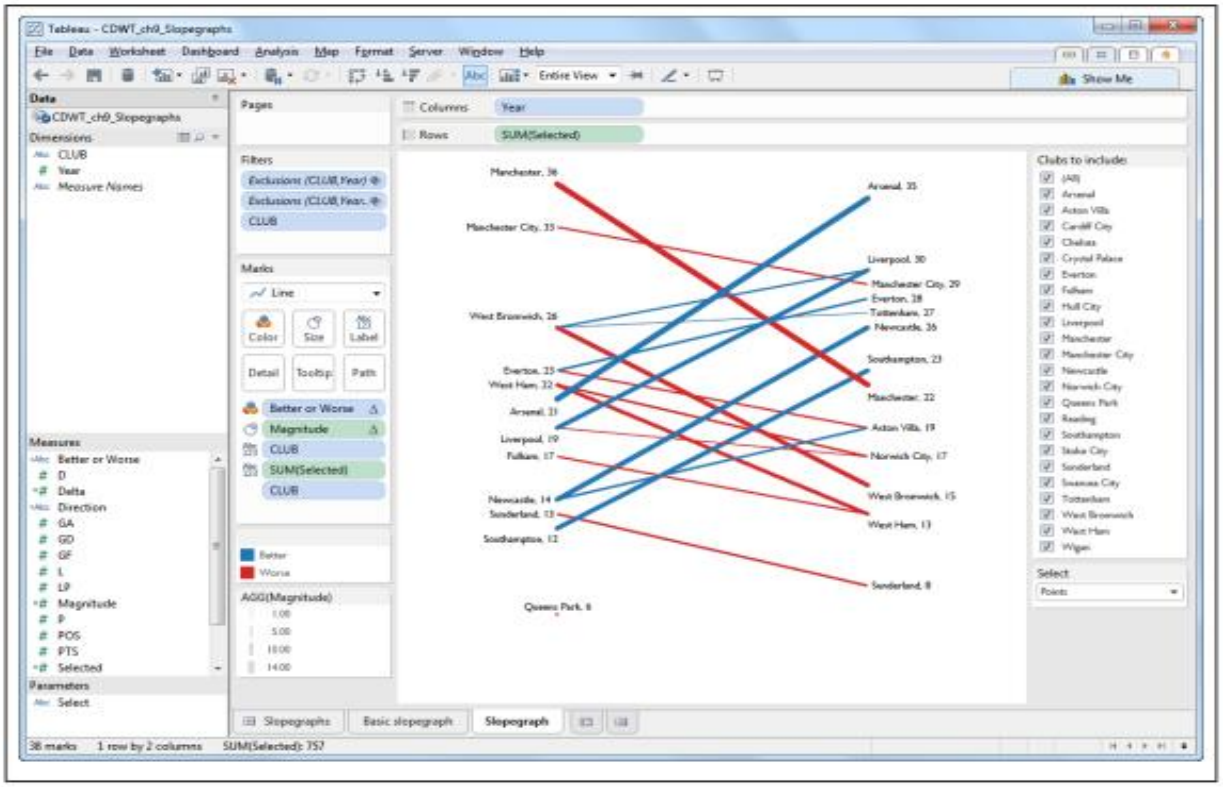


Figure 9-33. An updated slopegraph

I've also formatted the tooltips to yield a nice result when mousing over any of the line ends, and I've hidden Marks that were placed in awkward positions on the slopegraph that I couldn't adjust.

Step 6: Design the Dashboard We'll cover dashboards much more starting in Chapter 12, but let's see how a finished dashboard could look using this slopegraph example. After creating the slopegraph, we can add it to a new dashboard, add the parameter control and a drop-down filter for Clubs as floating dashboard objects, and then add a title and data source/reference in- formation at the bottom. With this view, we can do a whole lot more than find out what's behind Manchester's performance; we can also notice other big changes, such as Liverpool's suddenly prolific offense (select "Goals For"), or South- ampton's dramatic improvement in defense ("Goals Against" drops from 32 in 2012/2013 to only 14 in 2013/2014). See Figure 9-34.

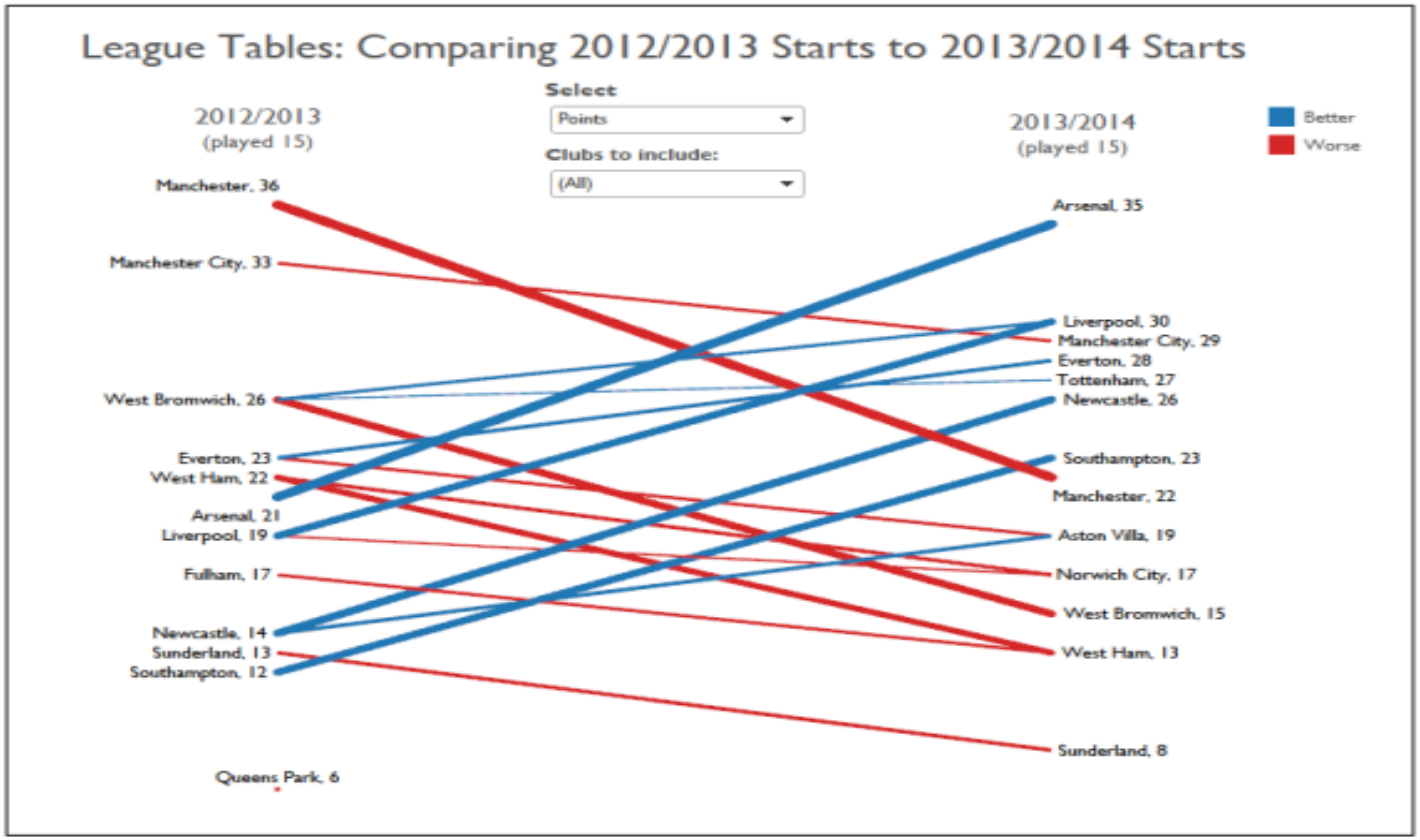


Figure 9-34. A slopegraph comparing the points earned by clubs through 15 games

This is the value of the slopegraph. It allows us to make a whole host of point-to-point comparisons, and the largest magnitude changes literally jump to the surface.

Maps and Location

Maps in Data Visualization

- Maps are powerful because they show **where things are located in relation to each other**.
- When we hear "map," we usually think of a world map. But historically, the earliest known maps were celestial maps — such as the star patterns found in the **Lascaux cave paintings** in France.

What makes something a map?

- At least one visual encoding represents **physical location**.

Types of Maps

Maps are not limited to geography. Examples include:

- One special map
- Circle Maps
- Heat maps (e.g., baseball pitch locations)
- Weather maps
- Election maps

What they all share:

- Position (x, y) encodes real-world location.

One Special Map

One Special Map – Minard's Masterpiece

- One of the most famous maps in history was created by
  Charles Joseph Minard.
- It shows the disastrous 1812 Russian campaign of Napoleon Bonaparte.
- This map is often called a masterpiece of data visualization.
- **What the Map Represents**
- Minard's map shows the march of Napoleon's army:
- From France → To Moscow (Advance)
- From Moscow → Back to France (Retreat)
- It visually encodes multiple variables at once.

How the Data Is Encoded

1.Position (Geography)

- The path of the army is drawn over a map using geographic coordinates.

2.Army Size (Quantity)

- The **thickness of the band** represents the number of soldiers.
- Thick band → Large army
- Thin band → Fewer soldiers
- At the start:
- About 422,000 soldiers
- At the end:
- Only around 10,000 survived
- The narrowing band visually shows massive loss of life.

3.Direction

- Light-colored band → Advance to Moscow
- Black band → Retreat from Moscow
- This clearly separates forward and return journeys.

4. Temperature (Time Series)

- At the bottom of the map:
- A line chart shows temperatures during the retreat.
- Temperatures dropped below -30°C.
- This explains why the army suffered such heavy losses.

- **Connection to Modern Tools (Like Tableau)**
- Today, with tools like Tableau, we can create similar positional visualizations using:

1.Circle Maps (Symbol Maps)

- Location = Position
- Size of circle = Value

2. Filled Maps (Choropleths)

- Regions shaded by intensity
- Color represents magnitude

3.Combined Maps

- Size + Color + Location together
- Similar to how Minard combined encodings
- Later, we can also create:
- Flow maps
- Shape maps
- Custom background image maps

Circle Maps

What Is a Map in Data Visualization?

A map is a graphical representation of a real-world place.

- It shows:
- Roads
- Borders
- Rivers
- Cities
- Regions

The spacing of objects on the map reflects their real-world geographic positions.

Because maps encode **physical location**, they are powerful tools for visualizing data tied to place.

What Is a Circle Map (Bubble Map)?

A circle map places a **circle (symbol)** at a geographic location.

The circle's:

1. **Size** → Represents magnitude (e.g., population)
2. **Color** → Represents category or intensity
3. **Position** → Represents geographic location

It is sometimes called a **Bubble Map**.

Why Use Circles?

- We use circles because:
- The center clearly represents a single point
- No corners that suggest a different location
- Visually simple and intuitive
- Easy to compare sizes
- The viewer naturally focuses on the center of the circle.

Three Ways to Create a Circle Map in Tableau

- Tableau automatically recognizes geographic fields (marked with a globe icon).

Method 1: Double-Click Geographic Field

- Double-click on a geographic field (e.g., Borough or County)
- Tableau automatically:
  - Generates Latitude and Longitude
  - Creates a symbol map
  - Places blue circles at correct locations
- This is the fastest method.

Method 2: Use "Show Me"

- Click the geographic field
- Open **Show Me** panel
- Select **Symbol Map**
- Tableau builds the circle map automatically.

Method 3: Manual Drag Method

- Drag **Latitude (Generated)** → Rows
- Drag **Longitude (Generated)** → Columns
- Drag **County/Borough** → Detail
- Now Tableau places a circle at the geographic center of each region.

Second Encoding

- **Adding a Second Encoding in Tableau (Circle Map Example)**
- In Tableau, **adding a second encoding** means using another visual element (like color) to represent an additional measure in the same chart. This makes the visualization richer and more informative.
- In your example:
- **Circle Size → Population**
- **Color Intensity → Population Density**
- This is called **dual encoding** (two visual attributes representing two different measures).
- **Step-by-Step in Tableau**
- **Step 1: Create the Base Circle Map**
- Drag **County** (or Borough) to the view.
- Ensure geographic role is assigned (County → Geographic Role → County).
- Tableau generates a map automatically.
- Change Marks type to **Circle**.
- Drag **Population → Size** shelf.
- □ Result: Larger circles = Higher population.

Step 2: Add the Second Encoding (Color)

- Drag **Population Density** (calculated field) → **Color** shelf.
- Tableau automatically applies a color gradient.
- Edit Colors:
  - Choose a **sequential color palette**
  - Use darker shade for higher values
- □ Result:
- Darker circles = Higher density
- Lighter circles = Lower density

When Marks Multiply

- **Building the Global Circle Map**
- **Dataset:**
- **Sample – World Bank Indicators (Excel)**
  Measure: *Internet users (per 100 people)*
  Time: 2000–2010

Step 1: Create the Map

- Double-click **Country / Region** → Tableau creates a world map.
- Drag **B: Internet users (per 100) → Size**
- Drag **Region → Color**
- Now each country becomes one circle (mark).
- **Important Concept: Why Did Tableau Create AVG(...)?**
- When you dragged the measure to Size, Tableau created:
- AVG(B: Internet users (per 100))
- Why?
- Because:
- Each country has **multiple records** (one for each year).
- Tableau must aggregate multiple values into one number per country.
- By default, Tableau uses **Average (AVG)**.
- You could also change aggregation to:
- SUM
- MIN
- MAX
- COUNT
- But average across 2000–2010 doesn't show growth clearly.

Step 2: Add a Year Filter (Very Important)

- Since Internet usage changes every year, showing an average hides growth trends.

Add Year Filter:

- Right-click **Date (year) → Show Quick Filter**
- Change filter type to:
  - Single Value (Slider)
    OR
  - Single Value (Dropdown)
- Why use single value?
  Because users usually want:
  ✔ 2000
  ✔ 2005
  ✔ 2010
- Not random combinations like 2000 + 2003 + 2009.

Filled Maps

Filled Maps (Choropleth Maps) in Tableau

- Filled maps are ideal when you want to show **rate, ratio, or density** across geographic regions.
- Instead of circles, Tableau fills each country (or state) shape with color intensity based on a measure.
- This is commonly seen in election maps or public health dashboards.

Why Filled Maps Work Well for Rates

- In your example, we are using:
- **Internet users (per 100 people)**
- This is a **rate (density)**, not a total quantity

Example: Iceland vs India (2010)

- Iceland → Very high Internet adoption rate
- India → Lower rate
- But…
- India has **far more total users** because its population is huge.
- So:
- Iceland → High rate, small total
- India → Lower rate, massive total
- This is why circle size (mass) can mislead when the metric is actually a rate.
- Filled maps solve this problem.

Country / Region	B: Internet users (per 100)	P: Population (count)	B: Internet users (calculated)
Iceland	96	0.3M	305,319
India	8	1,224.6M	97,969,146

Figure 10-10. Comparing Internet usage in Iceland and India, 2010

The World Bank data set includes the first three columns, and the fourth was made with a simple calculated field, as shown in Figure 10-11.

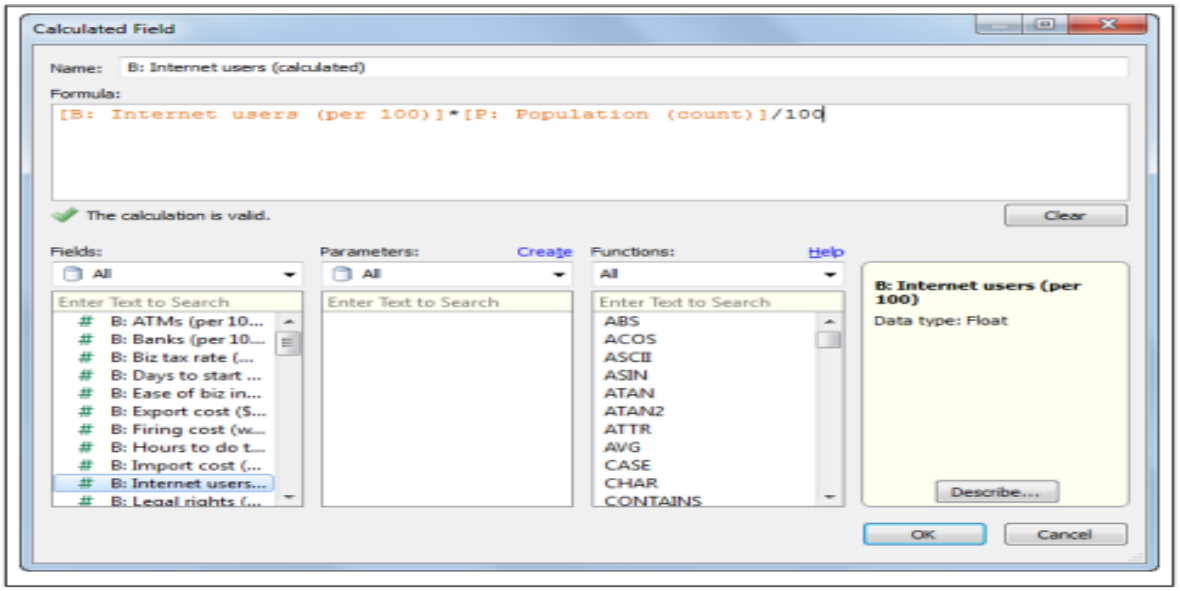


Figure 10-11. Calculating the absolute number of Internet users

How to Create a Filled Map in Tableau

Step 1: Duplicate Your Circle Map Sheet

- Right-click sheet → Duplicate
- Rename it "Internet Usage – Filled Map"

Step 2: Convert to Filled Map

Click **Show Me**

Select **Filled Map**

Tableau:

✔ Removes circles
✔ Colors each country shape
✔ Applies a color gradient

Sequential Color Palette (Best for Rates)

- Tableau defaults to a **Sequential** palette:
- Light Green → Low rate
- Dark Green → High rate
- This is effective because:
- One hue
- Increasing darkness = increasing value
- Clear interpretation
- We can now see:
- North America → High adoption
- Europe → High adoption
- Africa → Low adoption
- South America / Asia → Mid-range

- **Important Step: Fix the Scale**
- Tableau default:
- Min = 0
- Max = 96 (dataset maximum)
- Midpoint ≈ 48 (not meaningful)
- Better approach:
- Set Minimum = 0
- Set Maximum = 100
- Set Midpoint = 50
- Now white = 50 users per 100 (clear meaning).
- Now you can easily identify countries around 50%.
- Examples:
- Italy
- Portugal
- Morocco
- Uruguay
- **New Problem: Missing Data**
- Example:
- North Korea
- It appears near the midpoint color.
- But actually:
- It has **missing data**
- Tableau simply doesn't fill it
- It blends with background color
- Another example:
- Western Sahara
- No data → Looks similar to midpoint values.
- This is a major risk in filled maps:
- Missing values may look like meaningful values

Solutions to Missing Data Issue

- **Option 1: Change Diverging Palette**
- Use Orange–Blue instead of Orange–White–Blue.
- **Option 2: Change Background Map Style**
- Menu → Map → Map Options
  Change Background from **Gray → Dark**
- This makes missing areas more obvious.
- **Important: Grayscale Consideration**
- If someone prints your dashboard in black & white:
- Diverging palette becomes confusing.
- High and low values may both appear dark.
- Sequential palette works better in grayscale because:
- Darker = Always higher
- Lighter = Always lower
- Rule:
- If grayscale viewing is possible → Use Sequential.

Circle Map vs Filled Map

Circle Map	Filled Map
Good for totals (mass)	Good for rates (density)
Shows magnitude via size	Shows concentration via color
Can overlap	No overlapping
Better for comparing totals	Better for comparing intensity

Dual-Encoded Maps

- **Dual-Encoded Maps in Tableau**
- So far, each map showed **one variable**:
- Internet usage rate (per 100 people)
- Now we want to show **two variables at once**:
- 1.**Rate** → Internet users per 100 (density)
  2.**Total** → Absolute number of Internet users (mass)
- This is called **dual encoding** — using two visual properties to represent two different measures.

Step 1: Create the Absolute Number of Users

- The dataset includes:
- Internet users (per 100)
- Population
- So we create a calculated field:

```
Internet Users (Absolute) =
[Internet users (per 100)] / 100 * [Population]
```

Now we have:

- Density (rate)
- Mass (total users)

Method 1: Dual-Axis Map

- Start from the **Filled Map** (rate shown by color).

Steps:

1. Drag another **Latitude (generated)** to Rows.
2. Click dropdown → **Dual Axis**
3. In the second Marks card:
    1. Change Mark type → **Shape**
4. Remove rate from color
5. Drag **Internet Users (Absolute)** → **Size**
6. Make circles darker gray
7. Increase circle size

What This Shows

- **Country color** → Rate (per 100)
- **Circle size** → Total users

Example Insights (2010):

- China
  - Light color (moderate rate)
  - Largest circle (most users)
- Japan vs India
  - Similar circle sizes (similar totals)
  - Japan darker (higher adoption rate)
- Scandinavia
  - Darkest colors (very high rate)
  - Small circles (small population)
- **Problem with Dual-Axis Map**
- Circles overlay country shapes
  Visual clutter
  Occlusion (one mark hides another)
  Slightly intimidating for first-time viewers
- It gives rich insight — but requires effort to interpret.

A Dual-Axis Map

Let's start with the filled map we created (shown in Figure 10-14). We'll take the following steps to create a dual-axis map:

1. Drag another `Latitude (generated)` pill to the Rows shelf to the right of the first one.
2. Click the small down arrow in the second `Latitude (generated)` pill and select *Dual Axis* (as shown in Figure 10-20).
3. In the Marks card, open the panel for `Latitude (generated) (2)` and change the Mark type from *Automatic* (or Filled Map) to *Shape*.
4. Still in the Marks card area for `Latitude (generated) (2)`, remove `AVG(B: Internet users (per 100)` from the color shelf and add `B: Internet users (calculated)` to the Size shelf.

5. Use the Color and Size cards to change the color of the rings to a darker gray and increase the size of all of the rings to produce the final dual-axis map shown in Figure 10-21.

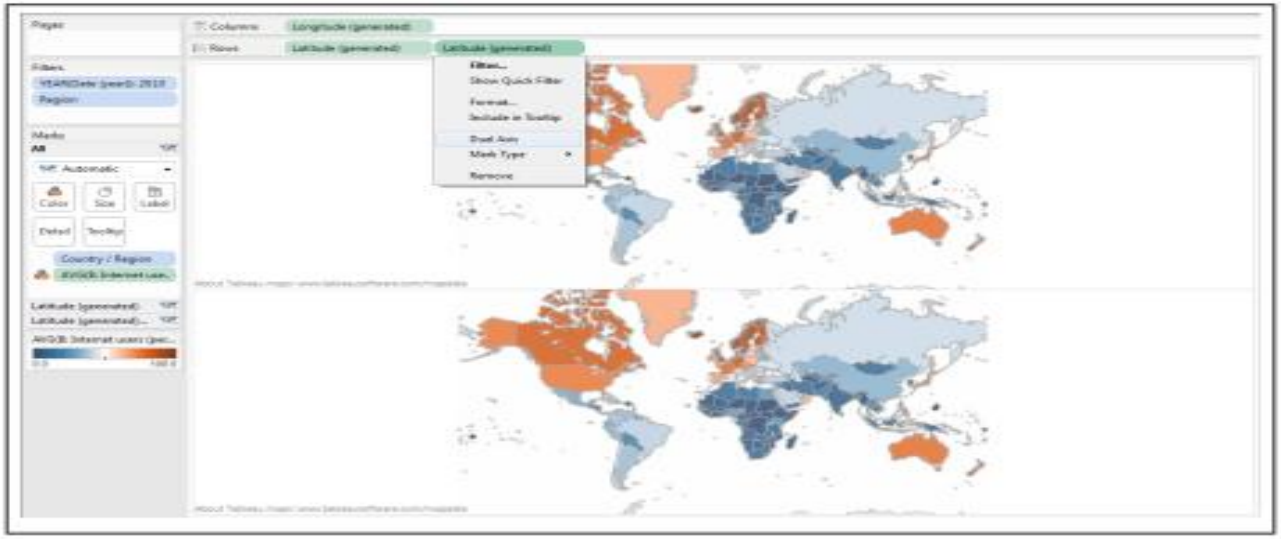


Figure 10-20. Create a dual-axis map

Figure 10-21. A dual-axis map showing both rate (color) and number (circle sizes) of Internet users

Dual-Encoded Circle Map

Method 2: Dual-Encoded Circle Map (Cleaner Option)

- Instead of overlaying shapes, use circle properties:
- **Size**
- **Color**
- **Steps:**

1. Duplicate the original Circle Map.
2. Remove **Region** from Color.
3. Change Marks → Circle.
4. Move:
    1. Rate → Color
    2. Absolute users → Size
5. Set Color palette:
    1. Orange–White–Blue Diverging
    2. Start: 0
    3. End: 100
6. Increase circle size.

What This Shows

- Large circle = More total users
- Orange = High adoption rate
- Blue = Low adoption rate
- Now comparisons are clearer and less cluttered.

Comparing the Two Methods

Feature	Dual-Axis Map	Dual-Encoded Circle Map
Visual Complexity	Higher	Lower
Occlusion	Yes	Minimal
Learning Curve	Steeper	Easier
Cleanliness	More layered	More readable

Most viewers find the **dual-encoded circle map easier to interpret**.

References

- W. Playfair, *The Commercial and Political Atlas: Representing, by Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England During the Whole of the Eighteenth Century*. London, UK: J. Debrett, 1786
- "Communicating Data with Tableau" by Ben Jones, covering scatterplots (Ch. 8), time series, and maps (Ch. 9).
- C. B. Jones, *Communicating Data with Tableau: Designing, Developing, and Delivering Your Data to the Masses*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2014.
- J. Priestley, *A Description of a Chart of Biography*. Warrington, UK: Printed for the author, 1765.
- C. J. Minard, "Carte figurative et approximative des températures pendant l'année 1812" (map of Napoleon's Russian campaign). Paris, France, 1869 (reprinted in *Cosmos*, vol. 18, pp. 5-11, 1861).

TEXT BOOKS

- Ben Jones, Communicating Data with Tableau, O‘Reilly June 2014

REFERENCE WEBSITE:

- Data Visualization | Coursera.