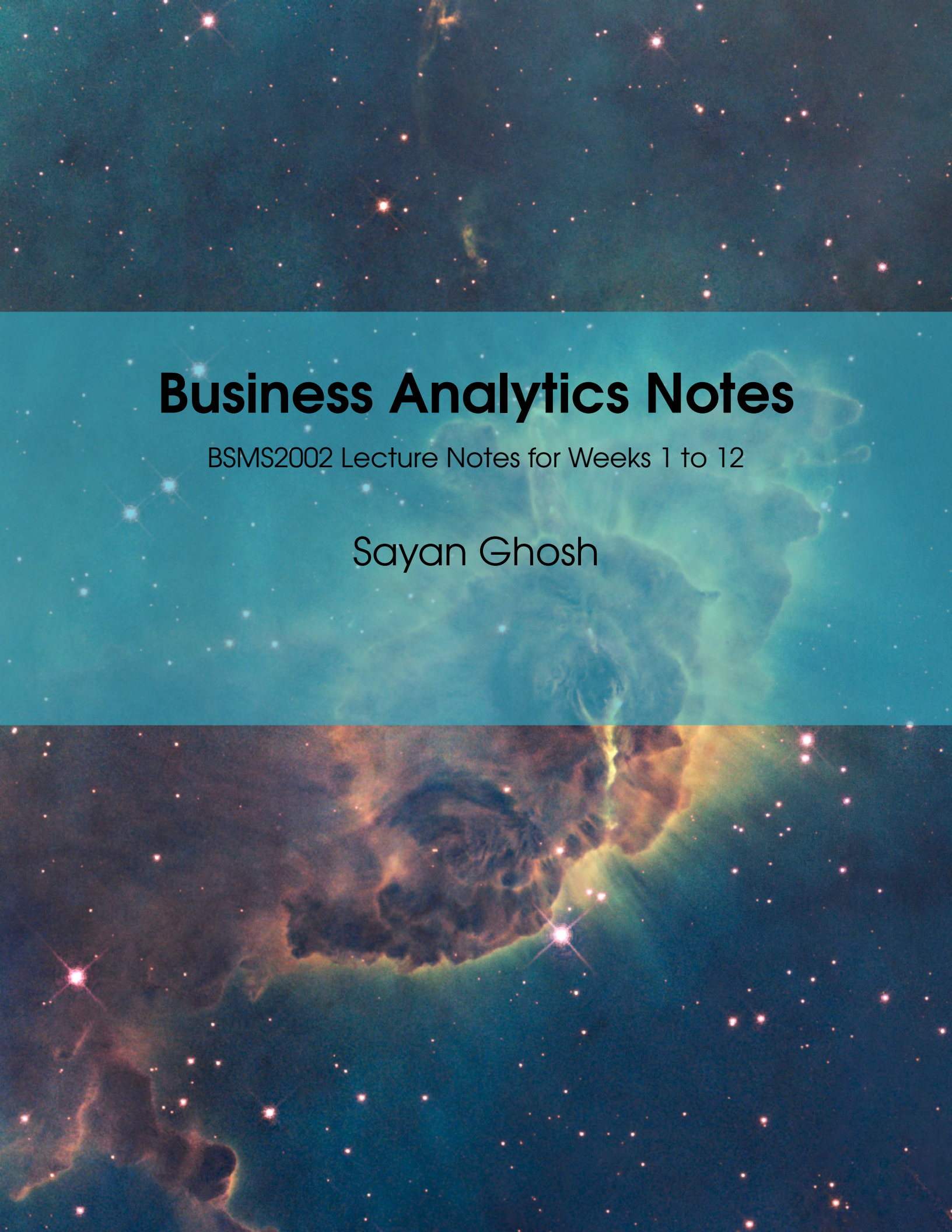


Business Analytics Notes

BSMS2002 Lecture Notes for Weeks 1 to 12

Sayan Ghosh

BUSINESS ANALYTICS (BSMS2002) LECTURE NOTES

HTTPS://GITHUB.COM/SAYAN01/BA-NOTES

This document contains my personal lecture notes of the Business Analytics (BSMS2002) course taught by **Prof. Rahul Marathe R.** at **IIT Madras BS Program.** This is not an official document.

First release, January 2024

Contents

1 Week 1: Why Visualize Data? 5

1.1 Introduction to Data Visualization 5

1.1.1 Types of Data 5
1.1.2 Categorical Data 5
1.1.3 Numeric Data 6

1.2 Defining the Message 8

1.2.1 The Three Steps of Data Visualization 8
1.2.2 One Data, Many Messages 9

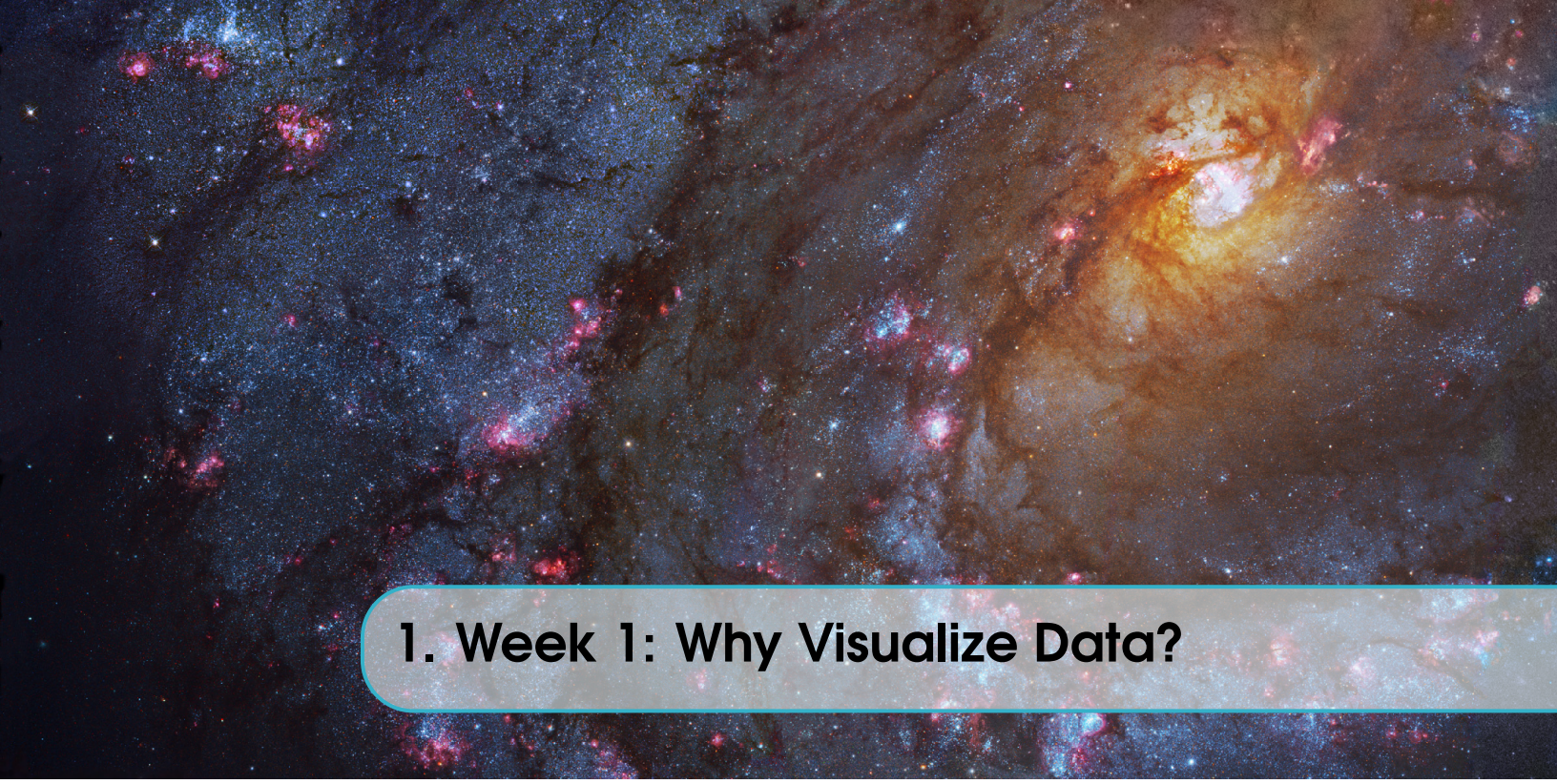
1.3 Choosing the Right Chart Type 9

1.3.1 Chart or Table? 11
1.3.2 Appropriate Chart Type for Different types of Messages 11
1.3.3 Designing the Visualization 11

1.4 Visualizing Data in Dashboards 13

1.5 Support Function 15

1.6 Operations Preserve Convexity of Functions 16



1. Week 1: Why Visualize Data?

1.1 Introduction to Data Visualization

Good data visualization is the key to communicate information clearly and efficiently to users. It is a key step to understand data and make better business decisions.

1.1.1 Types of Data

There are broadly two types of data:

- **Qualitative or Categorical Data:** Qualitative data is descriptive information
  Examples: Colors, textures, smells, tastes, appearance, beauty, etc.
- **Quantitative or Numeric Data:** Quantitative data is numerical information (numbers)
  Examples: Counts, height, weight, area, volume, ratings, etc.

1.1.2 Categorical Data

Definition 1.1.1 — Categorical Data. Categorical data is a type of data that is used to group information with similar characteristics.

Categorical data are labels and the values can belong to only a specific set of categories.

They can be further divided into two types:

- **Nominal Data:** Nominal data is a type of data that is used to label variables without providing any quantitative value or ordering.

- **Ordinal Data:** Ordinal data is a type of data that is used to label variables, but the order of the values matters.

1.1.3 Numeric Data

Definition 1.1.2 — Numeric Data Data. Quantitative or Numeric data is a type of data that is used to measure things.

It can either be **discrete** or **continuous**.

Definition 1.1.3 — Discrete Data. Discrete data is a type of data that can only take certain values.

Definition 1.1.4 — Continuous Data. Continuous data is a type of data that can take any value.

■ **Example 1.1** The height of a person is a continuous numeric variable. ■

■ **Example 1.2** The number of students in a class is a discrete numeric variable. ■

■ **Example 1.3** The gender of a person is a categorical variable. ■

Benefits of Visualizing Data

- **Communicate complex information:** Visualizations help people understand data quickly and convey a lot of information concisely and powerfully.
- **A Picture is worth a thousand words:** Create a "picture" for reasoning about and analyzing quantitative and conceptual information.
  - Makes cognitive processing easier
  - Provides "content/information rich" view at a glance
  - Directs attention toward the content rather than methodology
- **Focus on the important:** Visualizations help you focus on the important parts of the data and give you a clear idea of what the data is trying to tell you. It lets us identify which factors are more significant than others.

Attributes of Visual Perception

- **Form:**
  - Length
  - Width
  - Orientation
  - Size
  - Shape
  - Curvature
  - Enclosure
  - Spatial grouping
  - Blur
- **Color:**
  - Hue
  - Saturation
  - Brightness
- **Position:**
  - 1D: Position along a common scale
  - 2D: Position in a plane
  - 3D: Position in space
  - Direction of motion

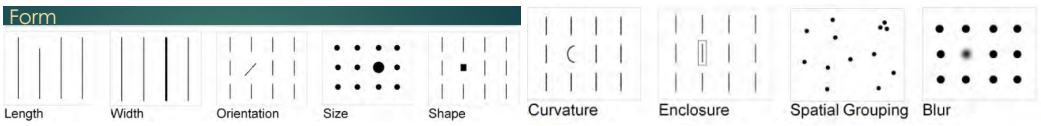
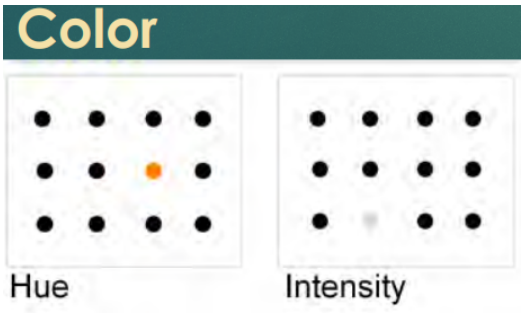
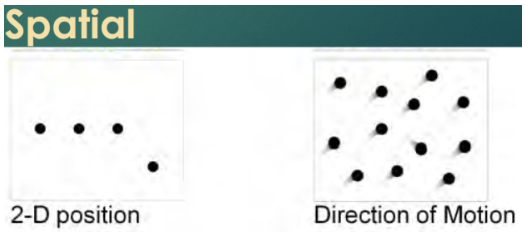


Figure 1.1: Visual Perception: Form



(a) Visual Perception: Color



(b) Visual Perception: Spatial

Figure 1.2: Visual Perception: Color and Spatial

Theorem 1.1.1 — Umbrella Principle. There are four principles of data visualization:

- **Know Purpose**: You need to have a purpose statement for every visualization you create. A purpose is not necessarily a message.
- **Ensure Integrity**: You need to ensure that your visualization is accurate and honest. Not just the data, but also the visualization itself. The presentation should not distort the truth or mislead the audience.
- **Maximize Data Ink; Minimize Non-Data Ink**: Make sure unnecessary elements are removed from the visualization. The goal is to maximize the data-ink ratio.
- **Show Your Data; Annotate**: You need to show your data. You should also annotate your visualization to help the audience understand it.

■ **Example 1.4 Know Purpose:** "My purpose in creating this graph to help the audience see that only a small percentage the patient base are candidates for this specific therapeutic regimen." ■

■ **Example 1.5 Ensure Integrity:** Truncating the y-axis is a common way to mislead the audience. The truncated graph makes the difference between the two groups look much larger than it actually is. ■

Figure 1.3: Baseline Tampering

■ **Example 1.6 Maximize Data Ink; Minimize Non-Data Ink:** The graph should only have lines and labels that are necessary to understand the data. The graph should not have any unnecessary elements like grid lines, unnecessary labels, etc.

■

	CALLS/DAY	DAYS/YEAR	CALLS/YEAR
PCP	8	200	1,600
NEUROLOGY	5	200	1,000
CV	5.5	200	1,100
ONCOLOGY	5	200	1,000
VIROLOGY	5	200	1,000
TRANPLANT	5	200	1,000
P/T DERM	8	200	1,600

(a) Cluttered

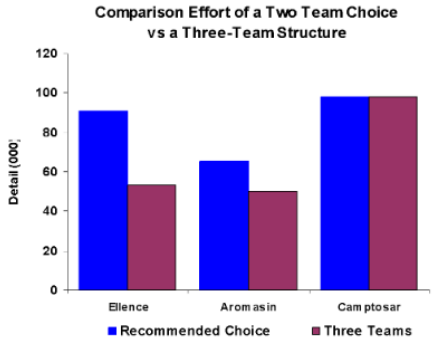
Table 1 -- Call Totals by Specialty

	Calls/Day	*Days/Year*	*Calls/Year*
PCP	8.0	200	1,600
Neurology	5.0	200	1,000
CV	5.5	200	1,100
Oncology	5.0	200	1,000
Virology	5.0	200	1,000
Transplant	5.0	200	1,000
P/T Derm	8.0	200	1,600

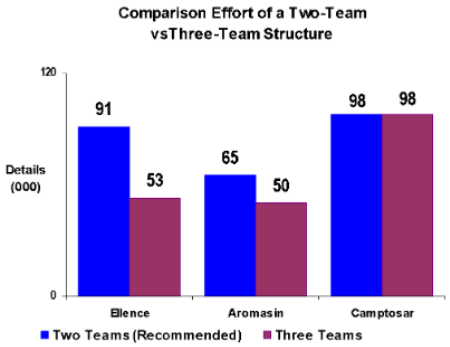
(b) Clean

Figure 1.4: Maximize Data Ink; Minimize Non-Data Ink

■ **Example 1.7 Show Your Data; Annotate:** You should always show your data. You should also annotate your visualization to help the audience understand it.



(a) No Annotations



(b) Annotations

■

1.2 Defining the Message

Definition 1.2.1 — Message. The message is the main point that you want to convey to the audience through your visualization.

1.2.1 The Three Steps of Data Visualization

1. **Defining the Message:** The first step is to define the message that you want to convey to the audience.
2. **Choosing the Right Chart Type:** The second step is to choose the right chart type to convey the message. "Should I use a bar chart or a line chart? or maybe a pie chart?"
3. **Designing the Visualization:** The third step is to design the visualization. "What colors should I use? Should I use a 3D chart or a 2D chart?" This step should ensure quick cognitive processing of the visualization.

1.2.2 One Data, Many Messages

A single data set can have many messages. The message that you want to convey depends on the audience and the context.

■ **Example 1.8** AIDS Data

You have been given the data of AIDS patients in the US. You can use this data to convey many messages:

Age Group	Number of AIDS Cases
Under 5	6,975
5-12	2,099
13-19	4,428
20-24	28,665
25-29	105,060
30-34	179,164
35-39	182,857
40-44	136,145
45-49	80,242
50-54	42,780
55-59	23,280
60-64	12,898
Over 65	11,555
Total	**816,148**

Table 1.1: AIDS Data

- **Message 1:** The age range of 25-44 has the highest number of AIDS patients.
- **Message 2:** Even kids below the age of 5 have AIDS.

1.3 Choosing the Right Chart Type

There are many chart types available to visualize data. Choosing the right chart type is very important. The chart type should be able to convey the message that you want to convey to the audience.

■ **Example 1.9** For the above AIDS data 9, to show the distribution of data among the age groups, you can use a pie chart or a bar chart. However, a line chart would not be a good choice.

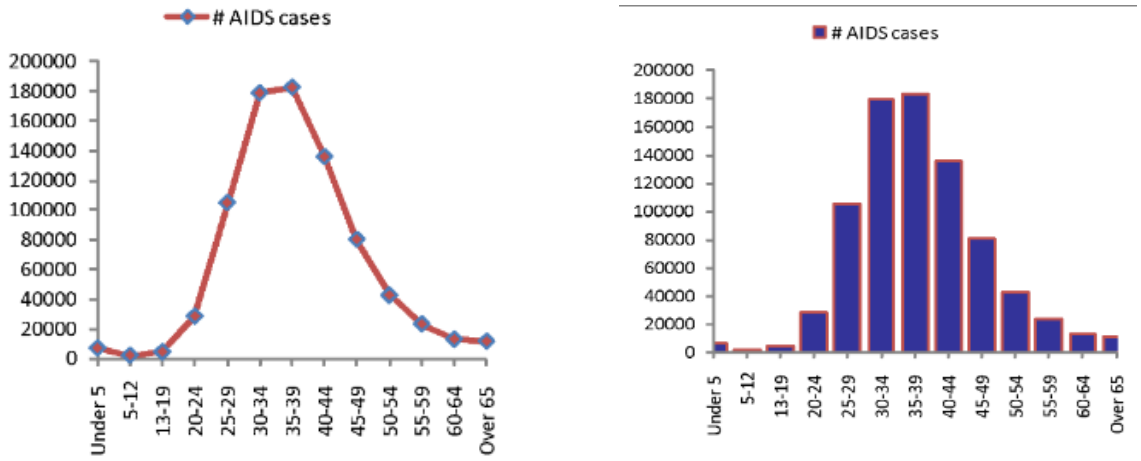
Pie Chart: Use this if number of categories is small (less than 5).

Bar Chart: Use this if number of categories is large (more than 5).

■

■ **Example 1.10** **Seasonality in Sales Data**: Use a line chart to show seasonality in sales data or other time series data.

■



(a) Line Chart

(b) Histogram

Figure 1.6: Distribution data: Histogram is a better choice if discrete groups

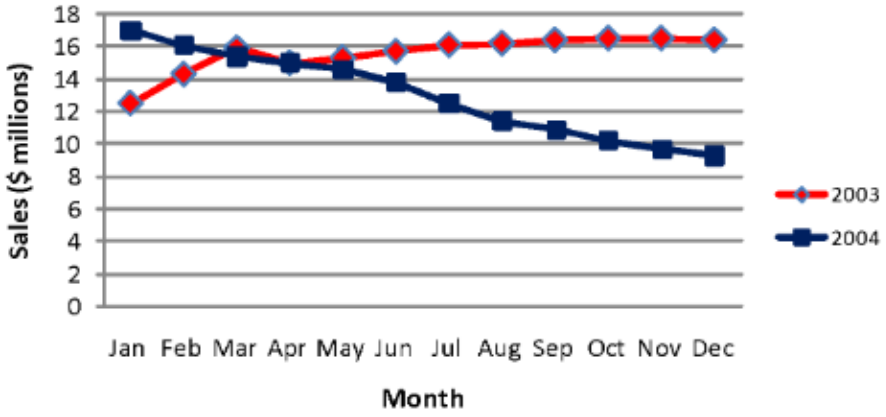


Figure 1.7: Seasonality in Sales Data

1.3.1 Chart or Table?

Sometimes a table is better than a chart. If the data is small and the audience needs to see the exact values, then a table is better than a chart.

This is also the case if there is no pattern in the data to be shown.

1.3.2 Appropriate Chart Type for Different types of Messages

We can also use a simple heuristic to choose the right chart type based on the message that we want to convey.

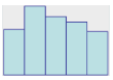
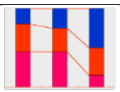
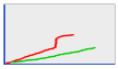
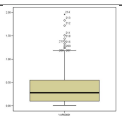
Message	Chart Type	Example
Item Comparison	Bar Chart	
Components of one item	Pie Chart	
Components of multiple item	Stacked Bar Chart	
Distribution	Histogram	
Relationship	Scatter Plot	
Location	Map	
Change over Time	Line Chart	
Outliers	Box Plot	

Table 1.2: Appropriate Chart Type for Different types of Messages

1.3.3 Designing the Visualization

Designing is a very important step in data visualization. The design should ensure quick cognitive processing of the visualization. The design should also be aesthetically pleasing.

A few best practices for designing visualizations are:

- **Avoid 3D Charts**: 3D charts are hard to read and understand. They also distort the data by violating the **Area Principle**.
- **Avoid Legends**: Legends are hard to read and understand. It is better to label the data directly.
- **Maintain High Contrast**: The colors used in the visualization should have high contrast. This makes the visualization easy to read.

- **Annotate Data to highlight important points**: Annotations help the audience understand the visualization. They also help the audience focus on the important points.

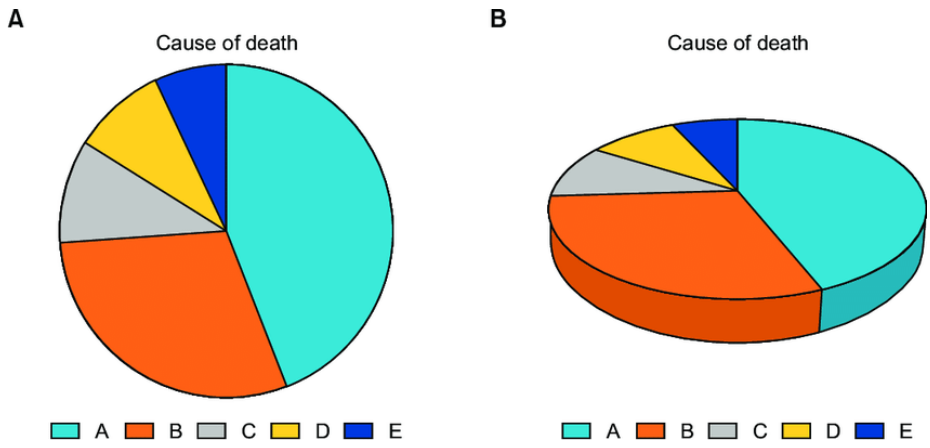
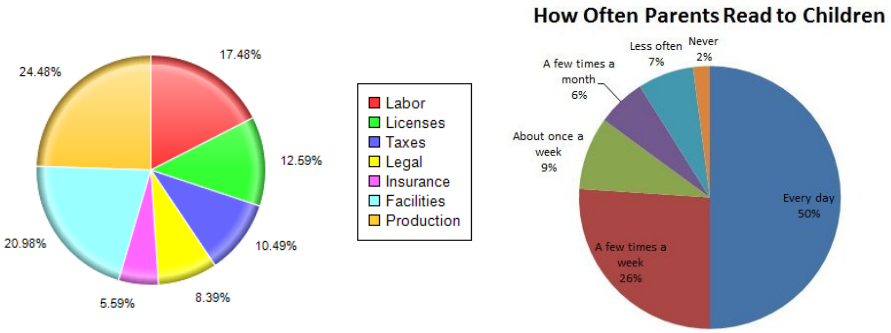


Figure 1.8: Area Principle is violated in 3D Charts; avoid them



(a) Pie Chart with Legend

(b) Pie Chart with Labels

Figure 1.9: Pie with Labels is easier to read than Pie with Legend

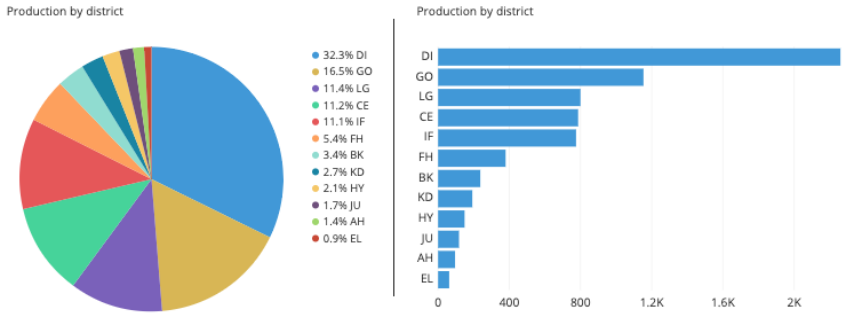


Figure 1.10: Bar Charts are better than Pie Charts if the number of categories is large

1.4 Visualizing Data in Dashboards

Definition 1.4.1 — Dashboard. A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen, so the information can be monitored at a glance.

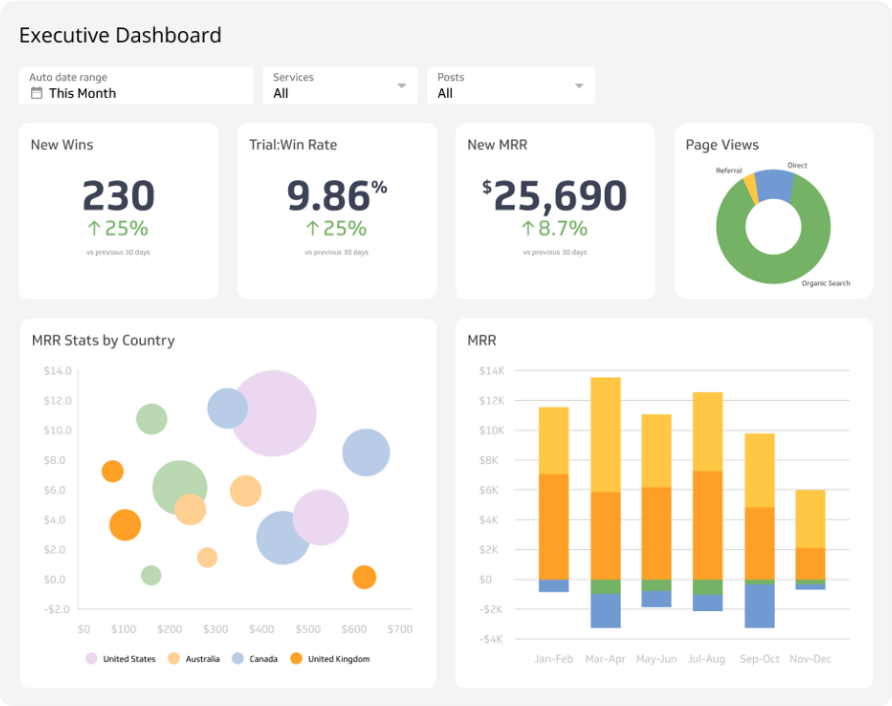


Figure 1.11: Example Dashboard

R f is convex $\Rightarrow -f$ is concave.

Level set: $S_\alpha f = \{x | f(x) \leq \alpha\}$.
$S_\alpha f$ is convex $\Leftrightarrow f$ is convex.

Definition 1.4.2 — Strongly convex. $f : C \to \mathbb{R}$ is strongly convex with modules μ if $\forall x, y \in C$, $\forall \alpha \in (0,1)$, $f(ax+(1-a)x) \leq af(x)+(1-a)f(x)-\frac{1}{2\mu}\alpha(1-\alpha)\|x-y\|^2$.

R

- f is 2nd-differentiable, f ix convex $\iff \nabla^2 f(x) \rangle 0$.
- f is strongly convex $\iff \nabla^2 f(x) \rangle \mu I \iff x \geq \mu$

Definition 1.4.3 — 2. $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is convexif $x, y \in \mathbb{R}, \alpha \in (0,1), f(ax+(1-a)x) \leq af(x)+(1-a)f(x)$.

The effective domain of f is $domf = \{x | f(x) \langle +\infty\}$

■ **Example 1.11 — ludcator function.** $\delta_c(x) = \begin{cases} 0 & x \in C \\ +\infty & elsewhere \end{cases}$.

$dom\delta_c(x) = C$ ■

Definition 1.4.4 — Epigraph. The epigraph of f is $epif = \{(x,\alpha) | f(x) \leq \alpha\}$

The graph of $epif$ is $\{(x, f(x) | x \in domf\}$.

Definition 1.4.5 — III. A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is

Theorem 1.4.1 $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is convex $\iff \forall x, y \in \mathbb{R}^n, \alpha \in (0,1), f(ax+(1-a)x) \leq af(x)+(1-a)f(x)$.

Proof. $\Rightarrow$ take $x, y \in domf$, $(x, f(x)) \in epif, (y, f(y)) \in epif$. ■

■ **Example 1.12 — Distance.** Distance to a convexset $d_c(x) = \inf\{\|z-x\| | z \in C\}$. Take any two sequence $\{y_k\} and \{\bar{y}_k\} \subset C$ s.t. $\|y_k - x\| \to d_c(x)$, $\|\bar{y}_k - \bar{x}\| \to d_c(\bar{x})$. $z_k = \alpha y_k + (1-\alpha)\bar{y}_k$.

$$\begin{aligned} d_c(\alpha x+(1-\alpha)\bar{x}) &\leq \|z_k - \alpha x - (1-\alpha)\bar{x}\| \\ &= \|\alpha(y_k - x) + (1-\alpha)(\bar{y}_k - \bar{x})\| \\ &\leq \alpha\|y_k - x\| + (1-\alpha)\|\bar{y}_k - \bar{x}\| \end{aligned}$$

Take $k \to \infty$, $d_c(\alpha x+(1-\alpha)\bar{x}) \leq \alpha d(x)+(1-\alpha)d(\bar{x})$ ■

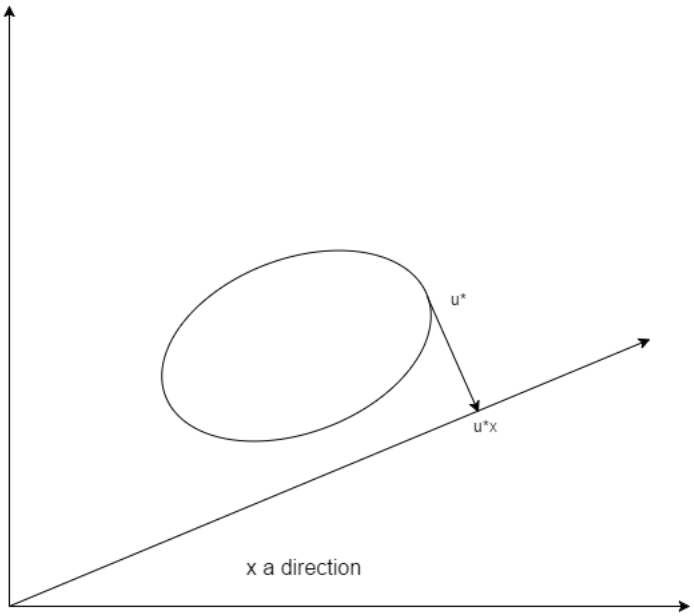
■ **Example 1.13 — Eigenvalues.** Let $X \in S^n := \{n \times nsymmetricmatrix\}$. $\lambda_1(x) \geq \lambda_2(X) \geq \ldots \geq \lambda_n(x)$.
$f_k(x) = \sum_1^n \lambda_i(x)$.
Equivalent characterization

$$f_k(x) = \max\{\sum_i v_i^T X v_i | v_i \perp v_j, i \neq j\}$$
$$= \max\{tr(V^T XV | V^T V = I_k\}$$
$$\max\{tr(VV^T X)\} \text{by circularity}$$

Note $\langle A,B\rangle = tr(A,B)$ is true for symmetric matrix.
$\langle A,A\rangle = |A|_F^2 = \sum_i A_{ii}^2$ ∎

1.5 Support Function

Take a set $C \in \mathbb{R}^n$, not necessarily convex.The support function is $\sigma_C = \mathbb{R}^n \to \bar{\mathbb{R}}$. $\sigma_C(x) = \sum\{\langle x,u\rangle | u \in C\}$.



Fact 1.5.1 The support function binds the supporting hyper-plane.

Supporting functions are

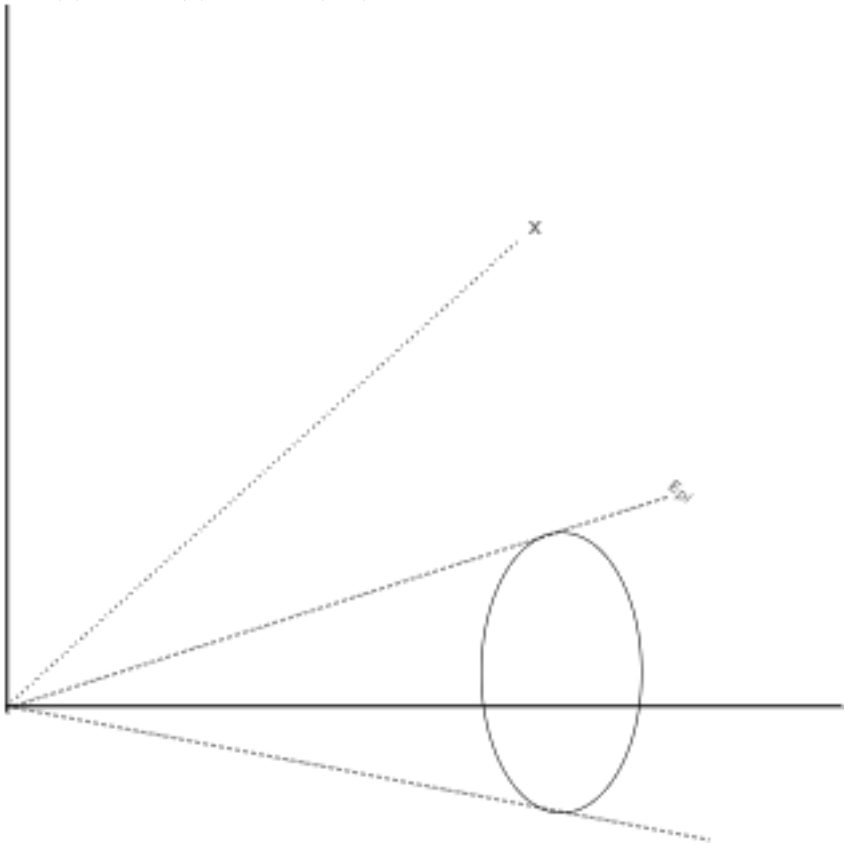
- Positively homogeneous
  $\sigma_C(\alpha x) = \alpha\sigma_C(x) \forall \alpha \rangle 0$
  $\sigma_C(\alpha x) = \sup_{u\in C}\langle \alpha x, u\rangle = \alpha \sup_{u\in C}\langle x,u\rangle = \alpha\sigma_C(x)$
- Sub-linear(a special case of convex, linear combination holds $\forall\alpha$.
  $\sigma_C(\alpha x + (1-\alpha)y) = \sup_{u\in C}\langle \alpha x + (1-\alpha)y, u\rangle \leq \alpha \sup_{u\in C}\langle x,u\rangle + (1-\alpha)\sup_{u\in C}\langle y,u\rangle$

∎ **Example 1.14 — L2-norm.** $\|x\| = \sup_{u\in C}\{\langle x,u\rangle, u \in \mathbb{R}^n\}$.
$\|x\|_p = \sup\{\langle x,u\rangle, u \in B_q\}$ where $\frac{1}{p} + \frac{1}{q} = 1$. $B_q = \{\|x\|_q \leq 1\}$.
The norm is

- Positive homogeneous
- sub-linear
- If $0 \in C$, σ_C is non-negative.
- If C is central-symmetric, $\sigma_C(0) = 0$ and $\sigma_C(x) = \sigma_C(-x)$

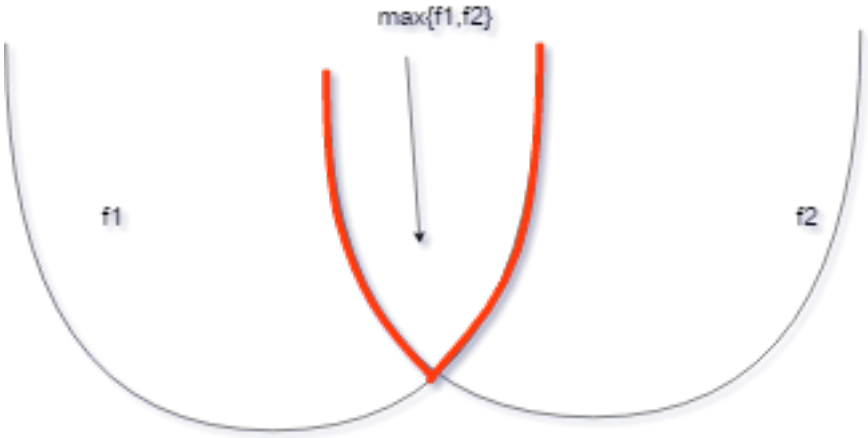
∎

Fact 1.5.2 — Epigraph of a support function. $epi\sigma_C = \{(x,t)|\sigma_C(x) \leq t\}$. Suppose $(x,t) \in epi\sigma_C$. Take any $\alpha > 0$. $\alpha(x,t) = (\alpha x, \alpha t)$.
$\alpha\sigma_C(x) = \alpha\sigma_C(x) \leq \alpha t$. $\alpha(x,c) \in epi\sigma_C$



1.6 Operations Preserve Convexity of Functions

- Positive affine transformation
  $f_1, f_2, \ldots, f_k \in cvx\mathbb{R}^n$.
  $f = \alpha_1 f_1 + \alpha_2 f_2 + \ldots + \alpha_k f_k$
- Supremum of functions. Let $\{f_i\}_{i\in I}$ be arbitrary family of functions. If $\exists x \sup_{j\in J} f_j(x) < \infty \Leftrightarrow$ $f(x) = \sup_{j\in J} f_j(x)$



- Composition with linear map.
  $f \in cvx\mathbb{R}^n$, $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map. $f \circ A(x) = f(Ax) \in cvx\mathbb{R}^n$

$$\begin{aligned} f \circ A(x) &= f(A(\alpha x + (1-\alpha)y)) \\ &= f(A\alpha x + (1-\alpha)Ay) \\ &\leq \alpha f(Ax) + (a-\alpha) f(Ay) \end{aligned}$$