

Data Visualization

BY-

ADITYA KADAM [\(TY1G2233657\)](#)

AKASH AUSEKAR [\(TY1G2233629\)](#)

ASMITA KADAM [\(TY1G2233658\)](#)

KALASHRI BORHADE [\(TY1G2233635\)](#)

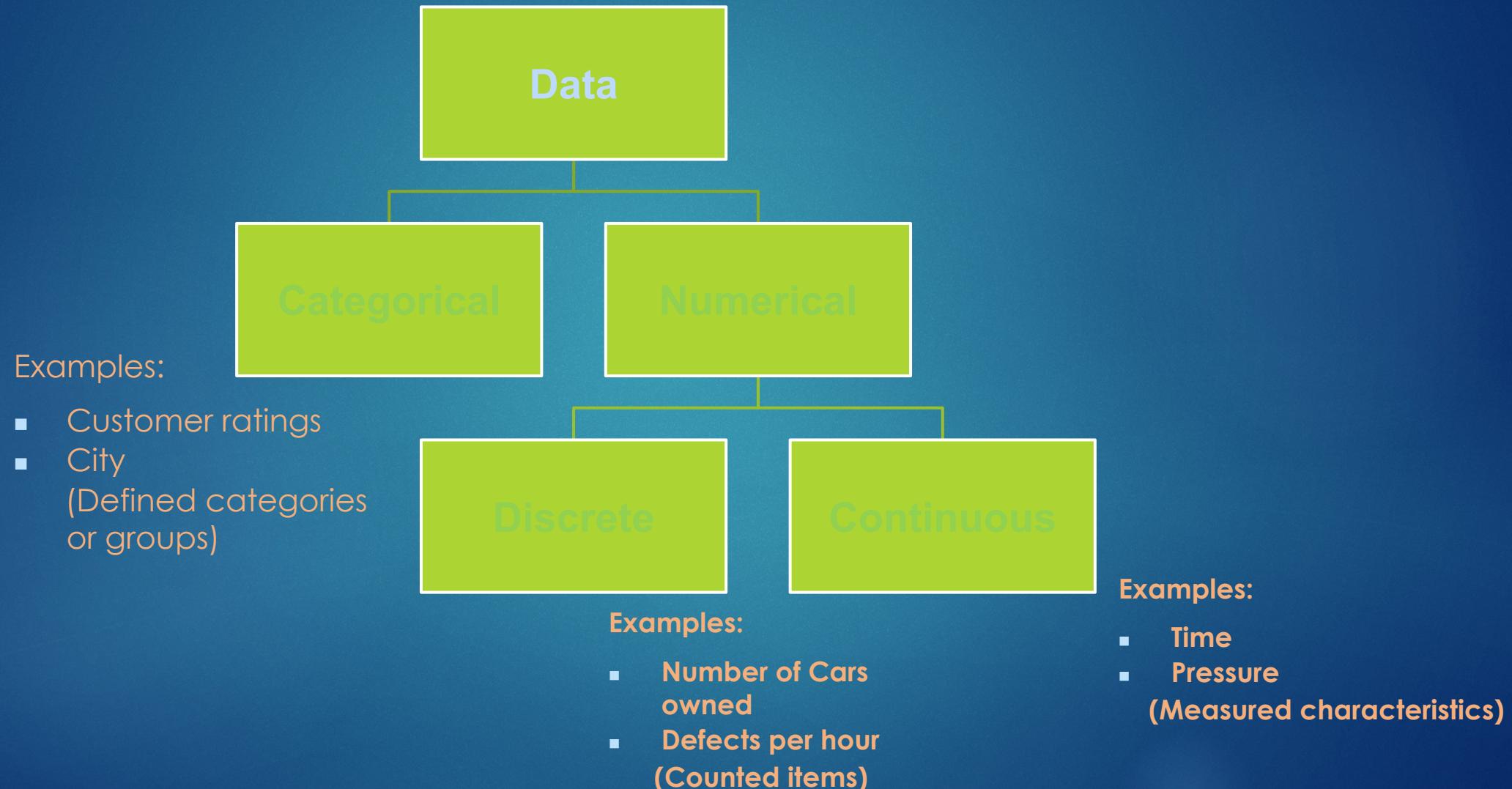
SHANTAPRASAD KAMAT [\(TY1G2233660\)](#)

Guided By- Sanket B
Github ID : bsanketm

Introduction

- ▶ **Data Visualization** is the representation of data in a graphical format. It makes the data easier to understand. Data Visualization can be done using tools like Tableau, Google charts, Data Wrapper, and many more. Excel is a spreadsheet that is used for data organization and data visualization as well. In this article, let's understand Data Visualization in Excel.
- ▶ Excel provides various types of charts like Column charts, Bar charts, Pie charts, Line charts, Area charts, Scatter charts, Surface charts, and much more.

Types of Data



Benefit of visual representation of data

- ▶ Communicate complex information concisely and powerfully
- ▶ 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
- ▶ Describe, explore, summarize a set of numbers
- ▶ Convey a messages about the significance of the data

Principles of Effective Visualization

5

Know purpose

Ensure integrity

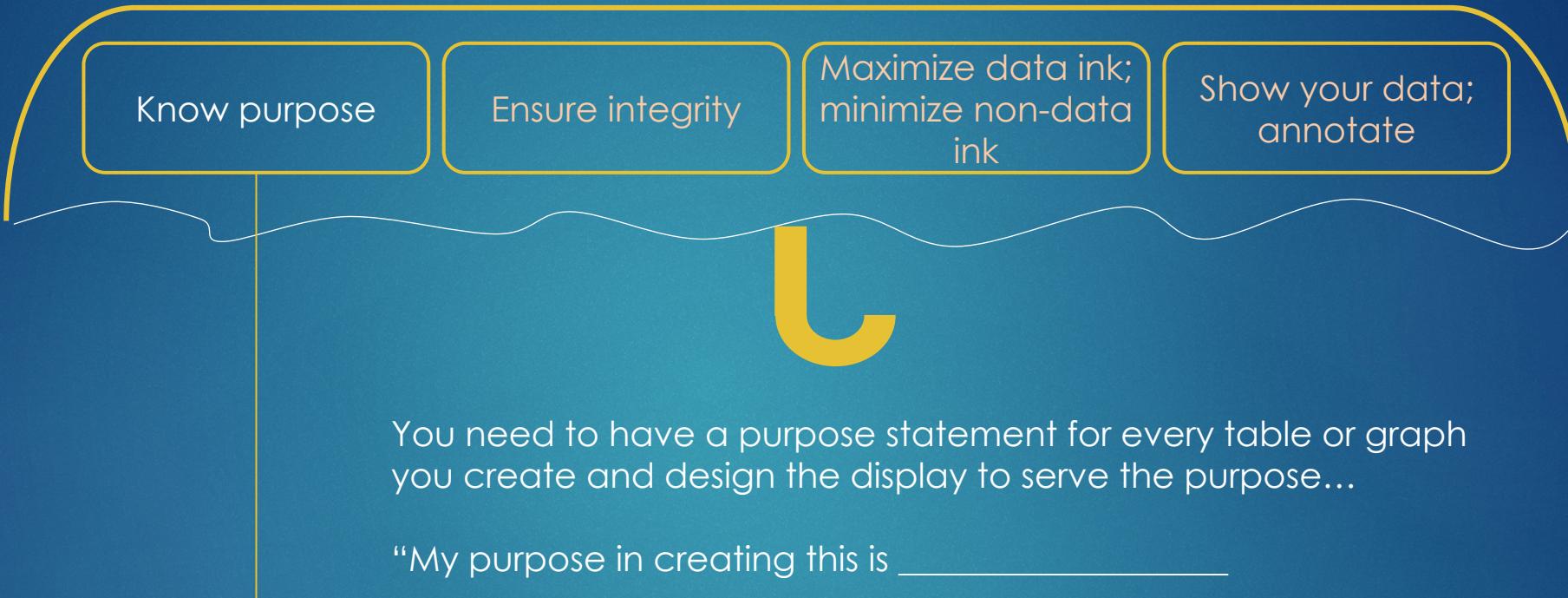
Maximize data ink;
minimize non-data
ink

Show your data;
annotate



Knowing your purpose drives all other decisions

6



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.

Note: a purpose is not necessarily a message.

Ensuring integrity is not just being correct; it also means no distortion

7

Know purpose

Ensure integrity

Maximize data ink;
minimize non-data
ink

Show your data;
annotate

...not only that the information correct, but that it is presented in a way that doesn't distort the truth.



Compare these two earnings-per-share graphs...the numbers are the same, but different scales convey different messages. The graph on the left is deceitful...



Use the least amount of ink to convey the most amount of information

8

Know purpose

Ensure integrity

Maximize data ink;
minimize non-data
ink

Show your data;
annotate

	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

Table 1 -- Call Totals by Specialty

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

Don't hide data...show it – Use annotations

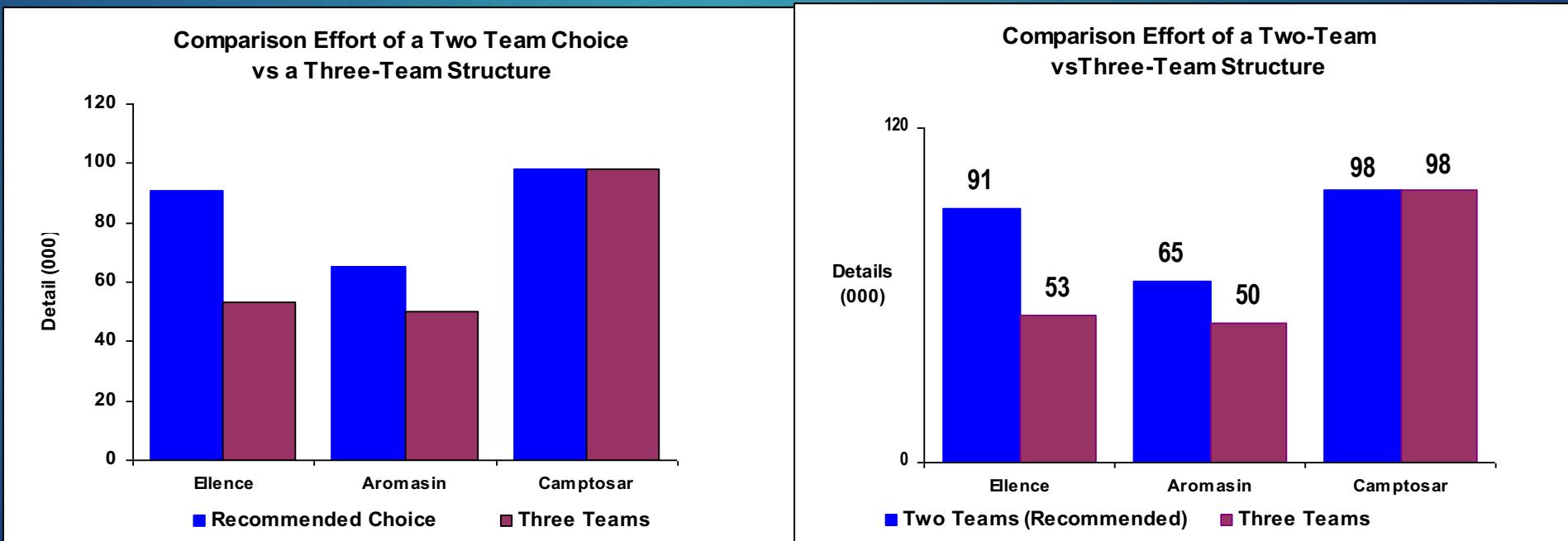
9

Know purpose

Ensure integrity

Maximize data ink;
minimize non-data
ink

Show your data;
annotate



Executing your information display is a three-step process

- What am I trying to communicate?
- What is the message?
- How do I make that message clear at a glance?

Defining Message

Choosing Form

- Should I use text, table, graph or a diagram?
- Or a combination?

- What design principles lead to quick cognitive processing and effective communication of the message?

Creating Design

Defining Message

You've been asked to produce a data display that provides insight into the population of people with AIDS. To the right is some data about the ages of people with AIDS in the U.S.

Given this data, what is the message that you would like to convey?

Majority of AIDS cases are in the 25-44 age groups

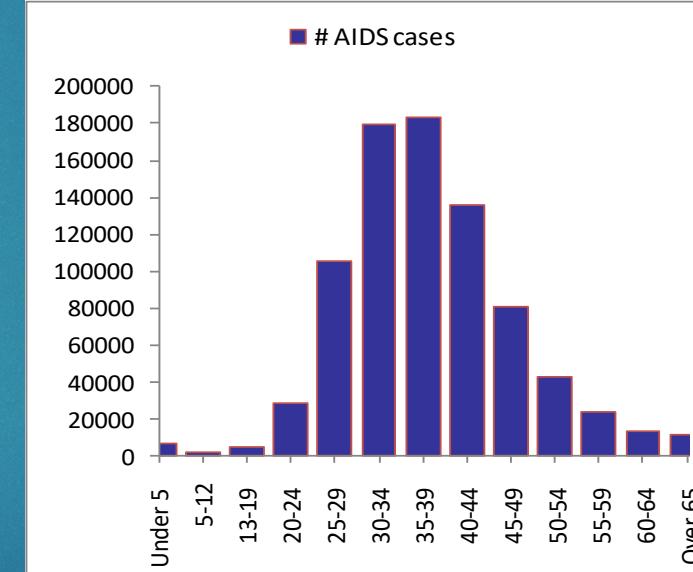
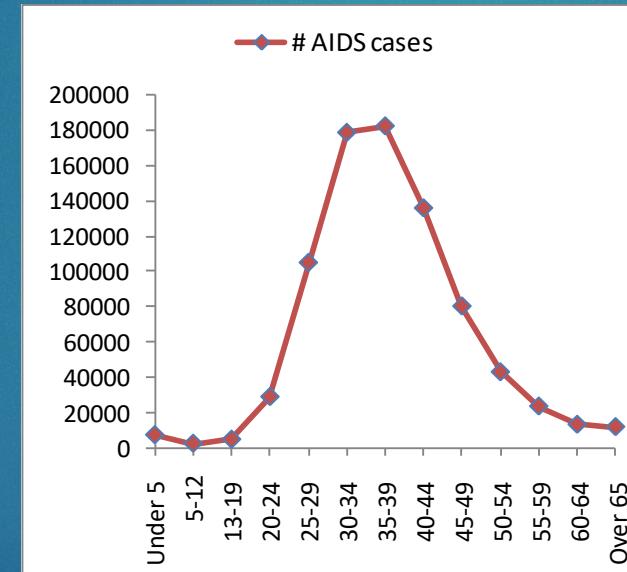
AIDS is most prevalent in the 35-39 age groups

Age Group	# AIDS cases
Under 5	6,975
5-12	2,099
13-19	4,428
20-24	28,665
25-29	10,5060
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

Choosing Form

Below are three ways to display the data to communicate the message from the distribution of AIDS sufferers. What is the most appropriate form?

Age Group	# AIDS cases
Under 5	6,975
5-12	2,099
13-19	4,428
20-24	28,665
25-29	10,5060
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



Choosing Form – Best Practices

13

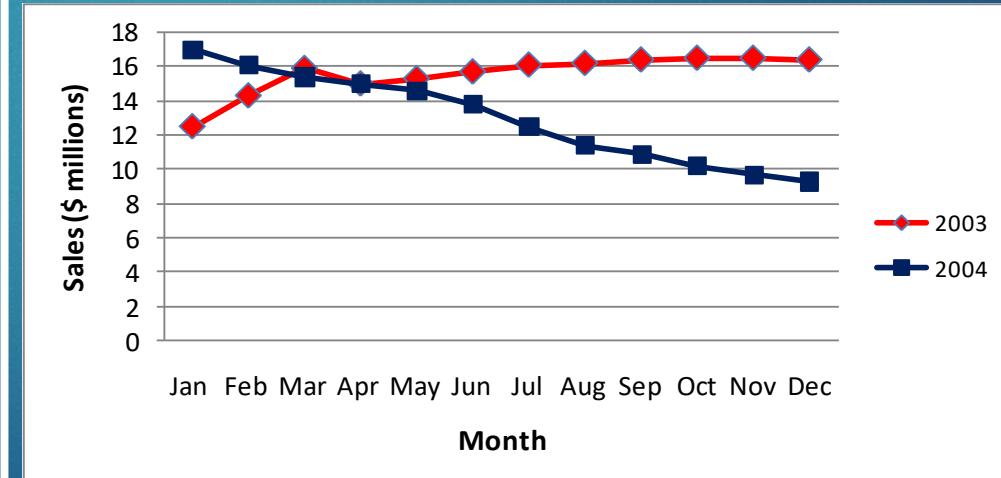
One key decision to make is whether to display the data as a table or a chart

Message: There is no apparent seasonality in sales in 2003 or 2004

Table: Sales by month in 2003, 2004

	2003	2004
Jan	12.5	17.0
Feb	14.3	16.1
Mar	15.9	15.4
Apr	15.0	15.0
May	15.3	14.6
Jun	15.7	13.8
Jul	16.1	12.5
Aug	16.2	11.4
Sep	16.4	10.9
Oct	16.5	10.2
Nov	16.5	9.7
Dec	16.4	9.3

Chart: Sales by month in 2003, 2004



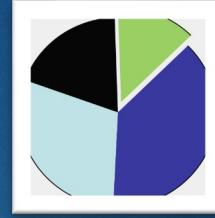
Choose appropriate graph type for message

If your message stresses...

Then choose....

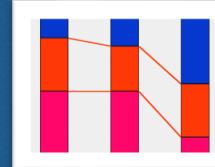
- ▶ Components of one item

- ▶ Pie chart



- ▶ Components of multiple items

- ▶ 100% column / stacked column chart



- ▶ Item comparison

- ▶ Bar chart



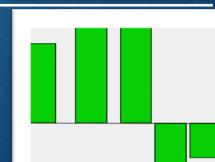
- ▶ Change over time

- ▶ Column / line chart



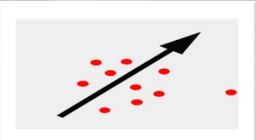
- ▶ Frequency, distribution

- ▶ Histogram

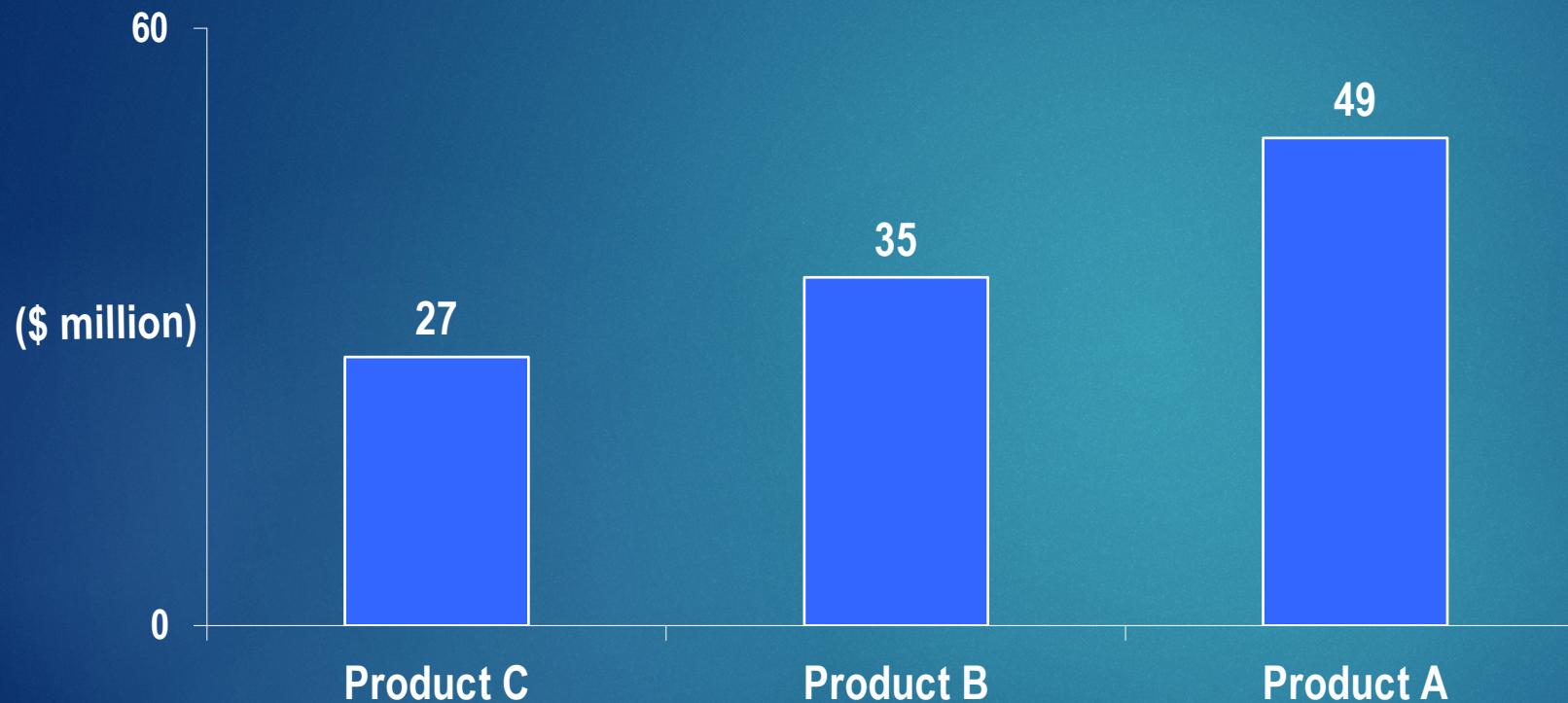


- ▶ Correlation

- ▶ Paired bar, scatter dot

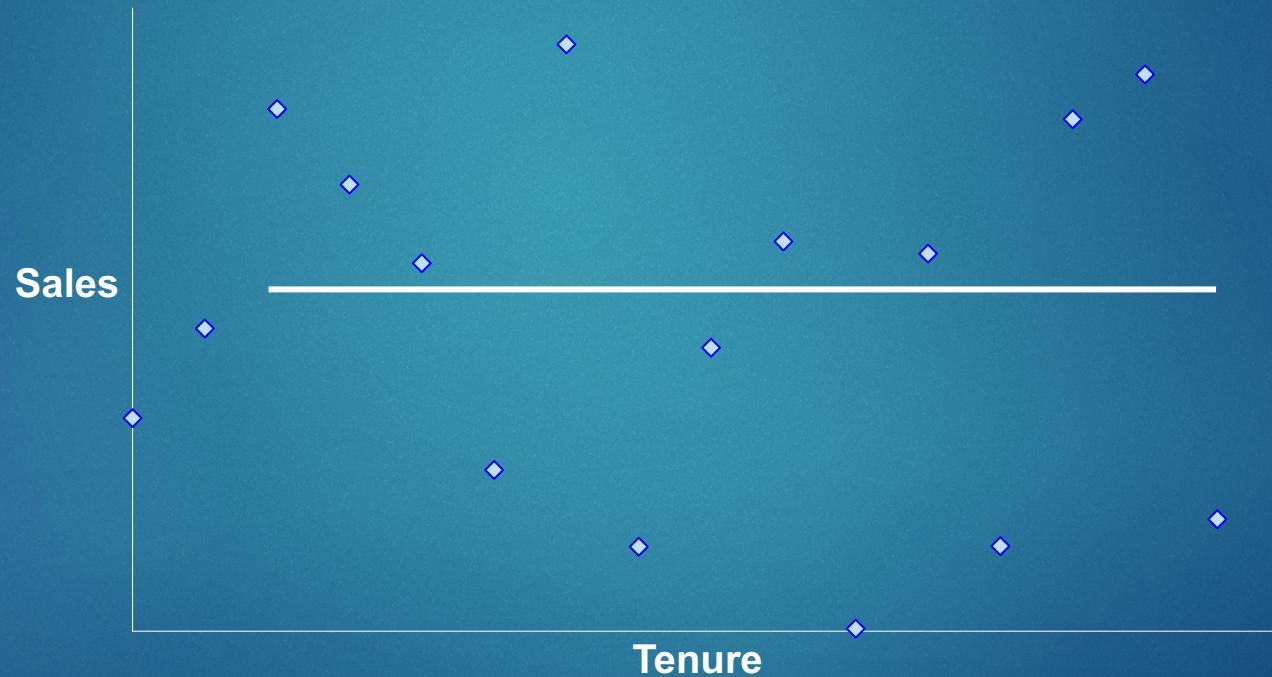


Sales of Product A exceed sales of
B and C



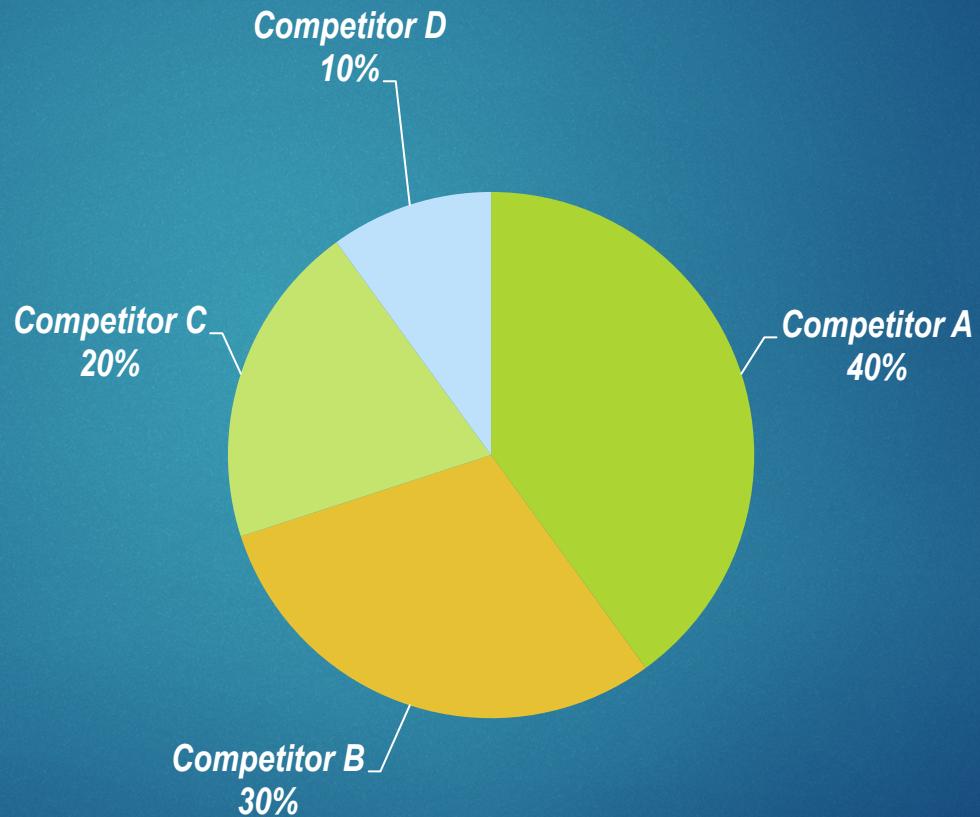
There is no apparent relationship between tenure and sales

Scatter Plot

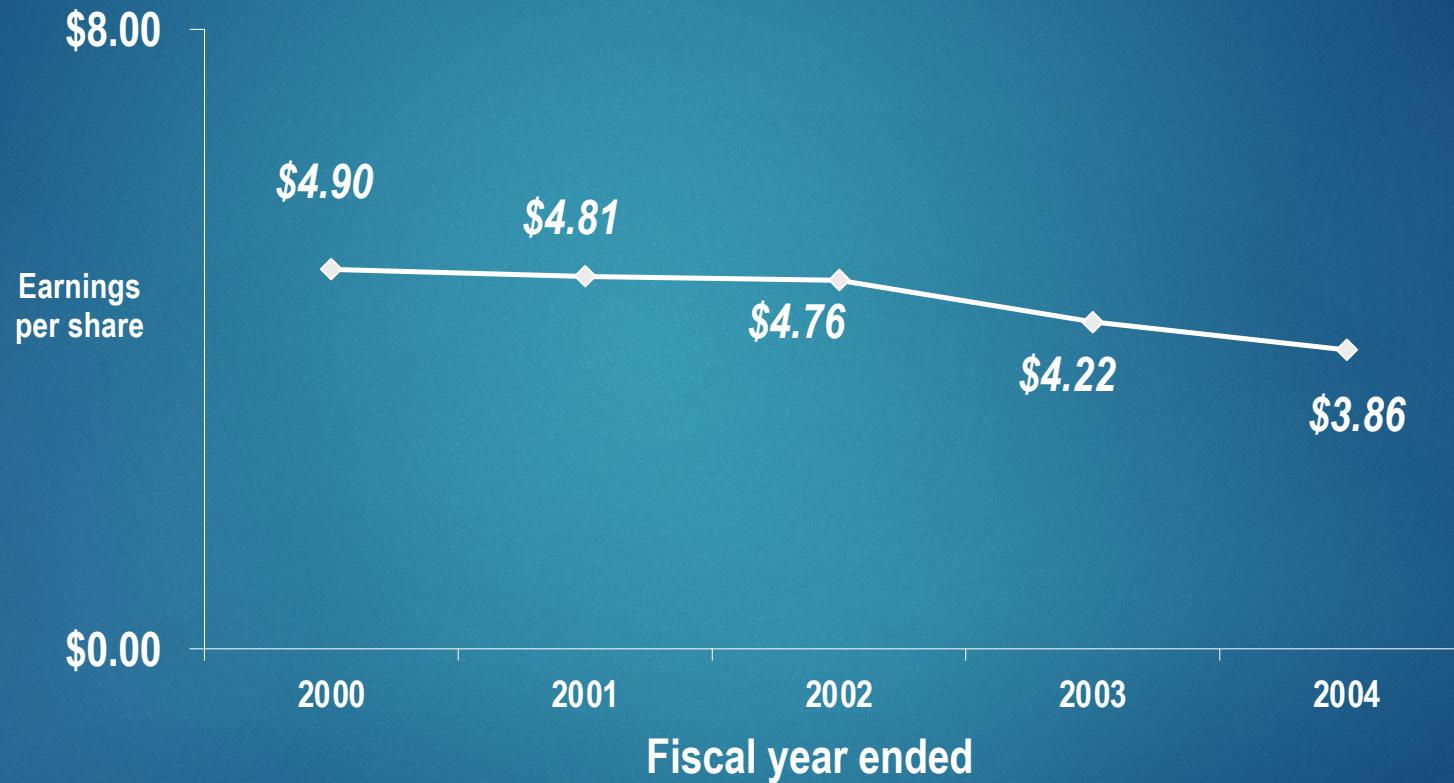


Competitor D has the smallest share of industry sales

Pie Chart

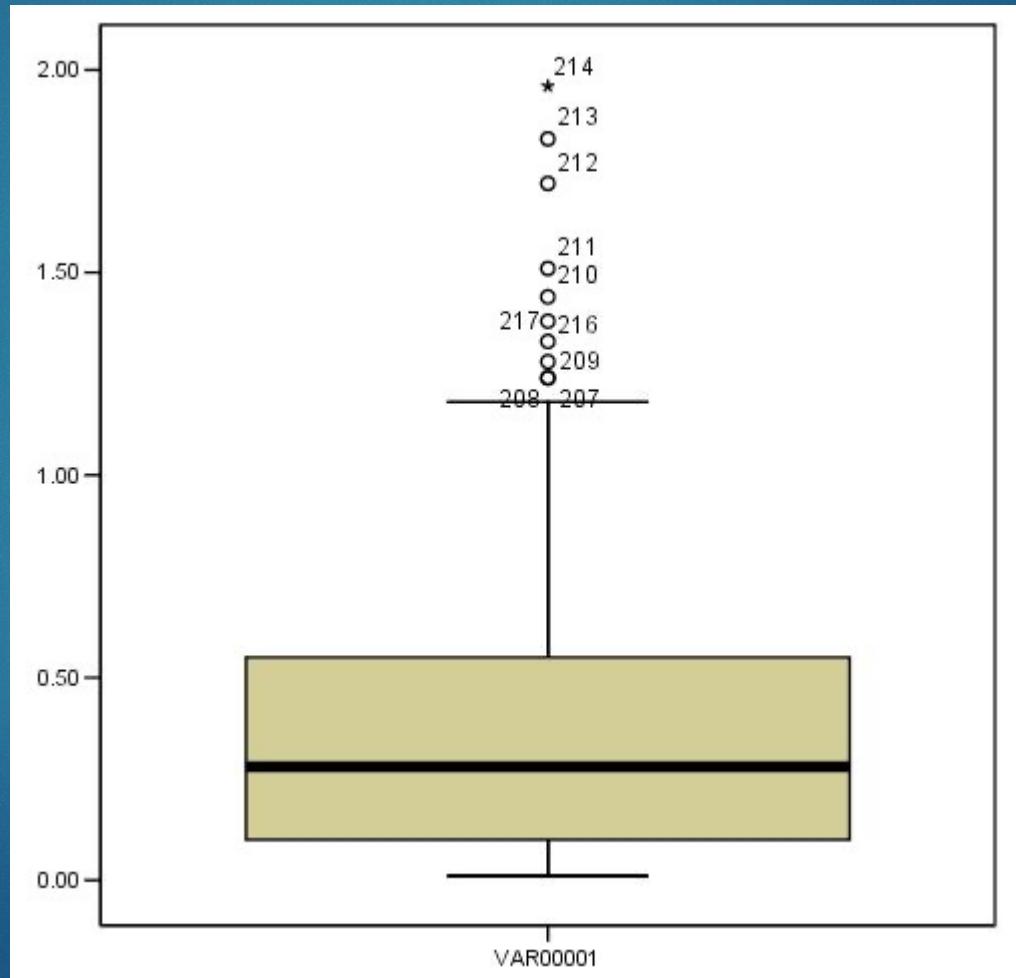


Earnings per share have decreased every year since 2000



There are outliers in the data

Box plot



► Bar Graph:

- A bar graph is a graphical representation of information. It uses bars that extend to different heights to depict value. Bar graphs can be created with vertical bars, horizontal bars, grouped bars, and stacked bars.

► Sparklines:

- A sparkline is a tiny chart in a worksheet cell that provides a visual representation of data. Use sparklines to show trends in a series of values, such as seasonal increases or decreases, Economic cycles, or to highlight maximum and minimum value.

SCATTERPLOT

- ▶ Scatter chart can shows the various type of correlation between variables with certain confidence interval.
- ▶ Scatter plots are used in either of the following situations.
 - When we have paired numerical data
 - When there are multiple values of the dependent variable for a unique value of an independent variable
 - In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems, checking whether two products that appear to be related both occur with the exact cause and so on.
- ▶
- ▶ 1 Select the two variables in the data on which we want to find the relation
- ▶ 2 Click on the insert tab and plot a scatter plot.
- ▶ 3 Then add the trend line and other tools.

Types

❖ Types of correlation

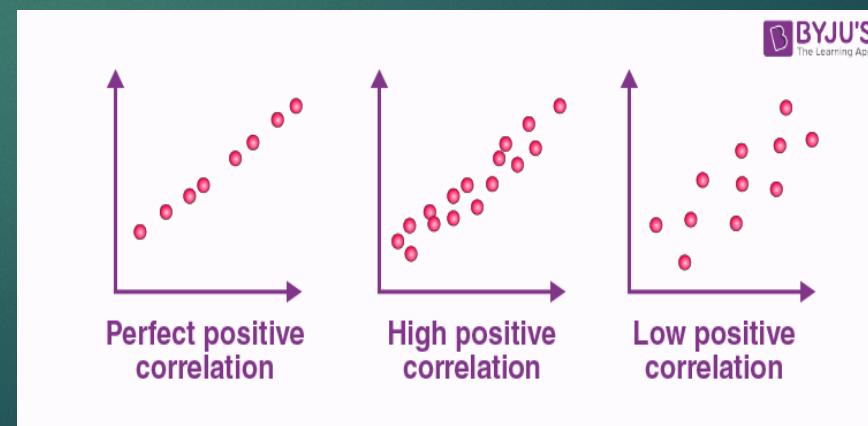
► The scatter plot explains the correlation between two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

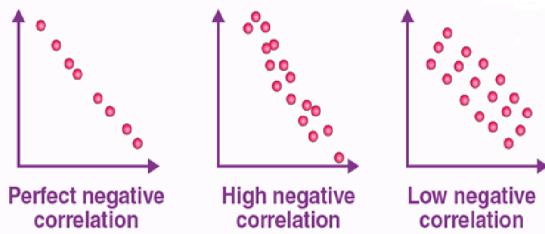
1. Positive Correlation
2. Negative Correlation
3. No Correlation

► Positive Correlation

► When the points in the graph are rising, moving from left to right, then the scatter plot shows a positive correlation. It means the values of one variable are increasing with respect to another. Now positive correlation can further be classified into three categories:

- **Perfect Positive** – Which represents a perfectly straight line
- **High Positive** – All points are nearby
- **Low Positive** – When all the points are scattered





Negative Correlation

When the points in the scatter graph fall while moving left to right, then it is called a negative correlation. It means the values of one variable are decreasing with respect to another. These are also of three types:

- **Perfect Negative** – Which form almost a straight line
- **High Negative** – When points are near to one another
- **Low Negative** – When points are in scattered form

Advantages

1. It helps to obtain the relationship between two or more variables.
2. A scatter plot is the best method to show the non-linear pattern.
3. Anyone can easily determine the minimum and maximum value of the range of data flow.
4. Perception and reading are so clear.
5. Plotting the graph is moderately basic.

Disadvantages

1. A scatter plot chart does not help determine the variables' precise relationship.
2. A scatter plot chart is not a quantitative measure, which helps to get the result numerically.
3. It only gives an estimated idea of the relationship.

Line Chart

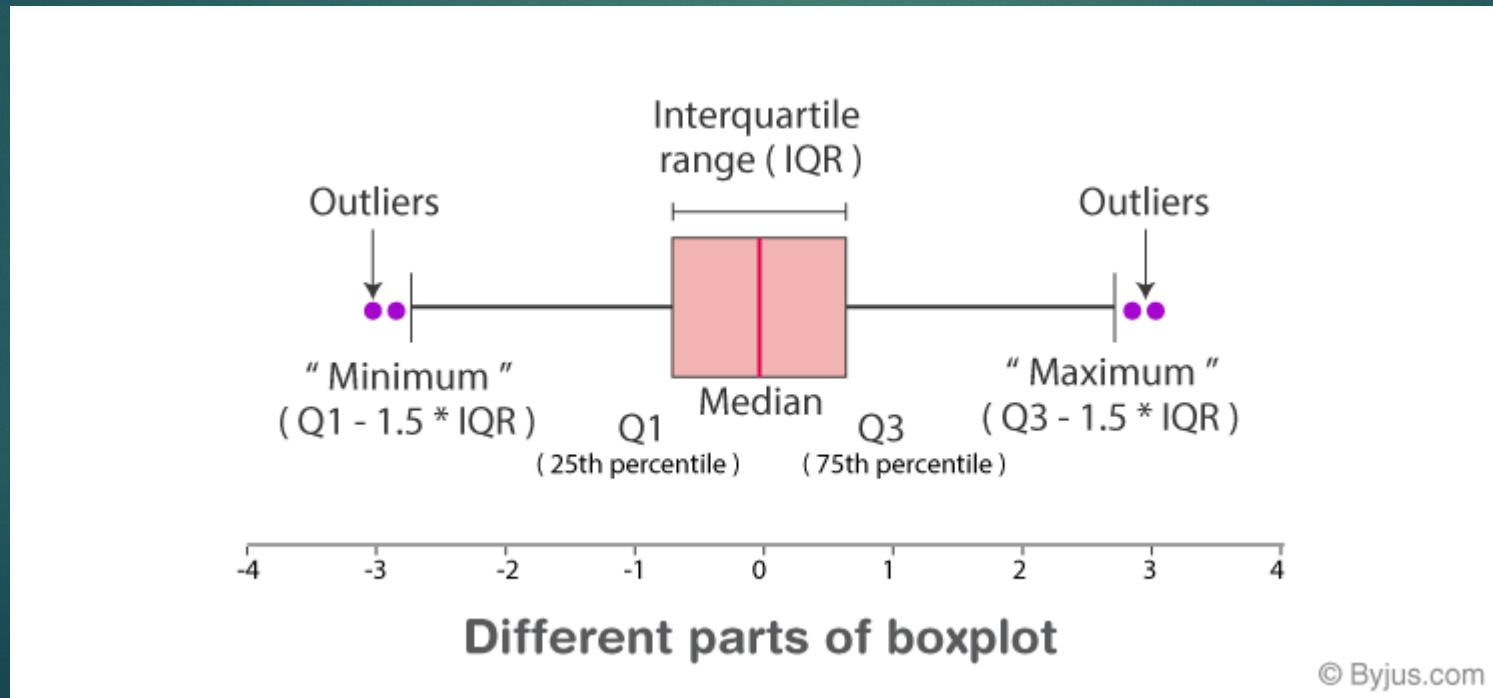
- The Line Chart is one of many different chart types that can be used for visualizing data.
- A Line Chart, also referred to as Line Graph or Line Plot , Connects a series of data points using a line.
- When to use line charts? When changes are minor, it is better to use Line Charts rather than Bar Charts. Line charts can also compare changes over the same period for more than one group.
- A Line Chart is used to show the change in information over time. The horizontal axis is usually a time scale; for eg. Minutes, hours, days, months, or years.
- What type of data is used in a Line Chart? Continuous data is appropriate for Line Chart. For eg. Blood Pressure, Age, etc.
- How to analyze a Line Graph? The changing slope of the line segments emphasizes changes, trends, and patterns. For a single series of data, assess the changes in the line to identify trends and patterns. When you have multiple metrics, compare their lines to determine whether they have the same trend and patterns.

Bubble Chart

- The Bubble Chart is an extension of scatterplot.
- A Bubble Chart is used to visualize a data set with two to four dimensions. The first two dimensions are visualized as coordinates, the third as color and the fourth as size.
- Each bubble in a chart represents a single data point.
- Bubble Charts, also known as bubble plots or bubble graphs, are used when data needs a third dimension to provide richer information to viewers. A bubble plot is a relational chart designed to compare three variables.
- A Bubble Chart is created from a data table with three columns. Two columns will correspond with the horizontal and vertical positions of each point, while the third will indicate each point's size. One point will be plotted for each row in the table.

Box and Whiskers Plot

- ▶ A simple way of representing statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles, usually with a vertical line inside to indicate the median value. The lower and upper quartiles are shown as horizontal lines either side of the rectangle.



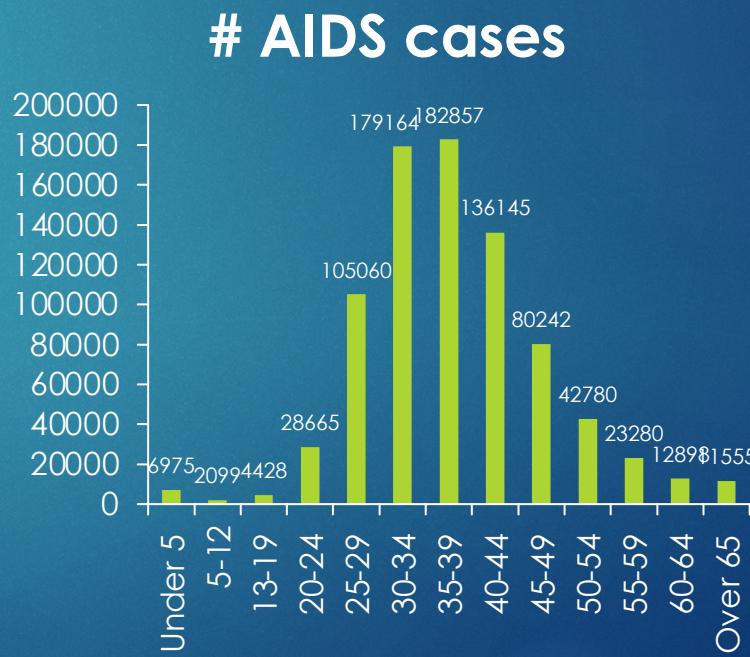
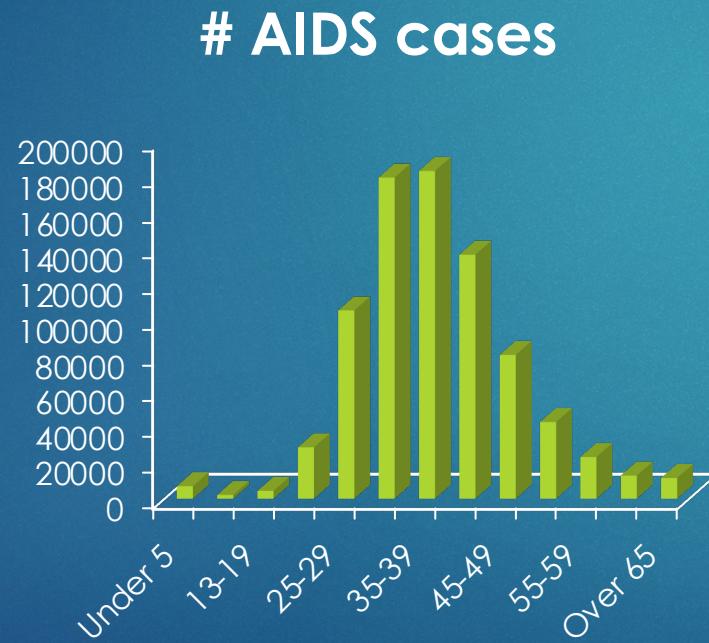
WATERFALL CHART

- ▶ A waterfall chart shows a running total as values are added or subtracted. It's useful for understanding how an initial value (for example, net income) is affected by a series of positive and negative values.
- ▶ The columns are color coded so you can quickly tell positive from negative numbers. The initial and the final value columns often start on the horizontal axis, while the intermediate values are floating columns. Because of this "look", waterfall charts are also called bridge charts.



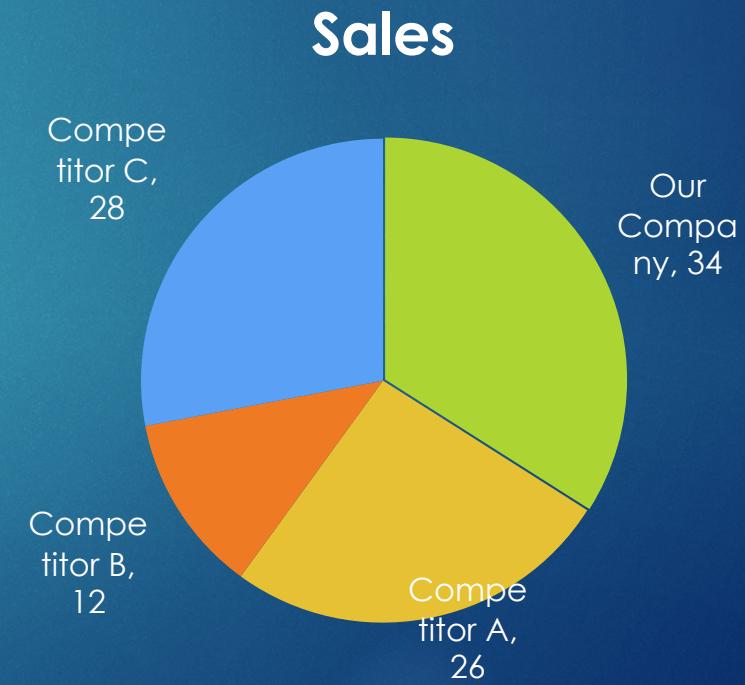
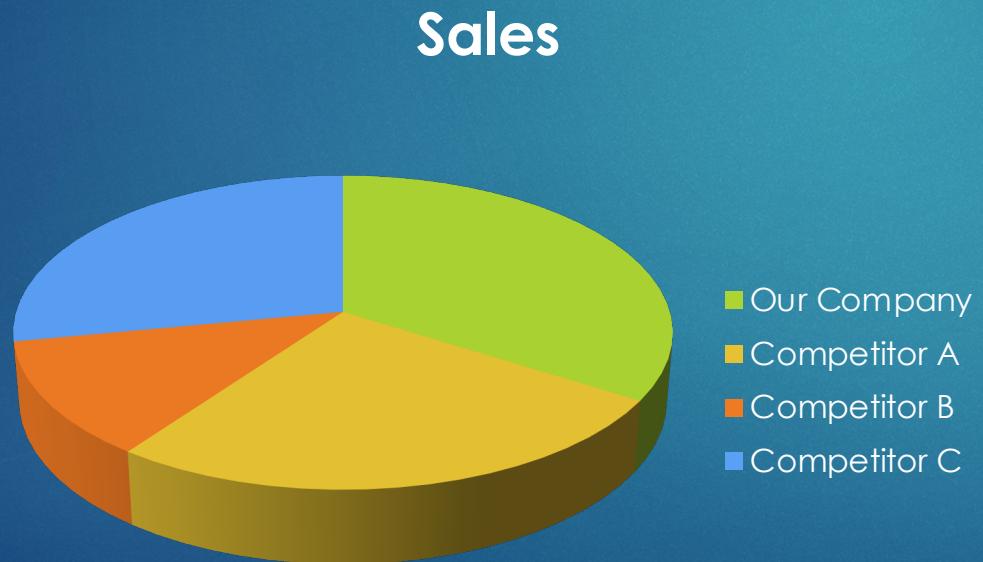
Creating Design

Below are two designs of the column data representation of the AIDS data. Which design leads to a chart that is easier to understand?



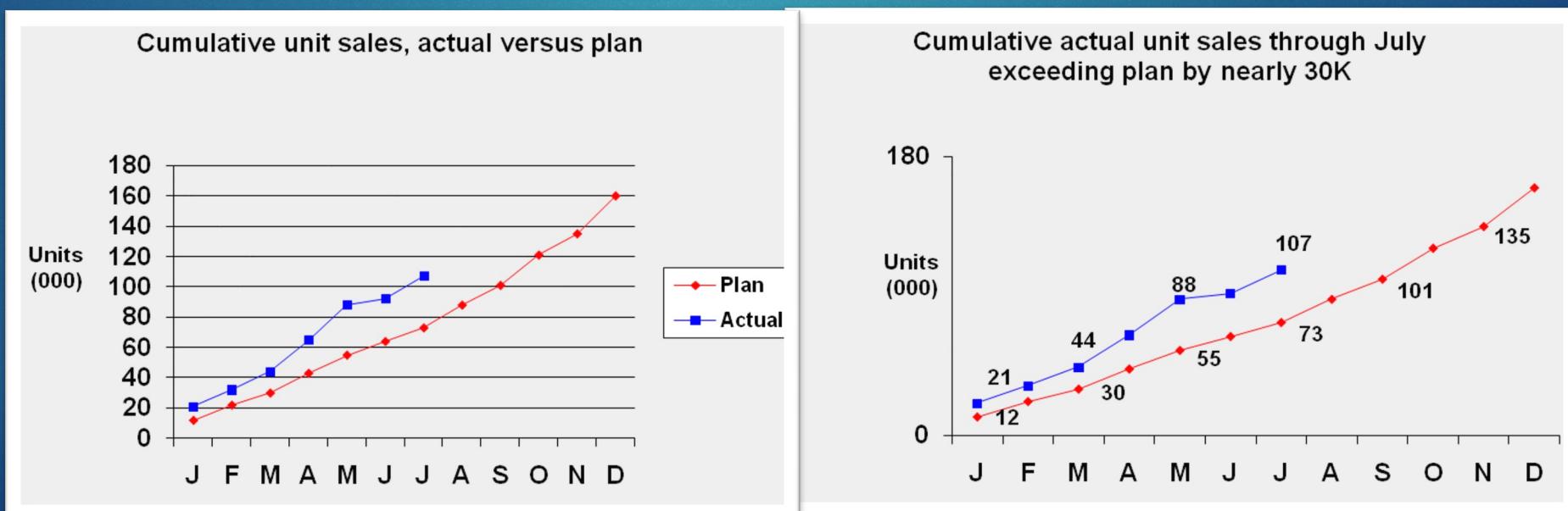
Creating Design – Best Practices

- Avoid 3-D effects
- Avoid legends; consider using labels
- Avoid contrasting borders around objects
- Use annotations to highlight key data changes or to focus on specific data points

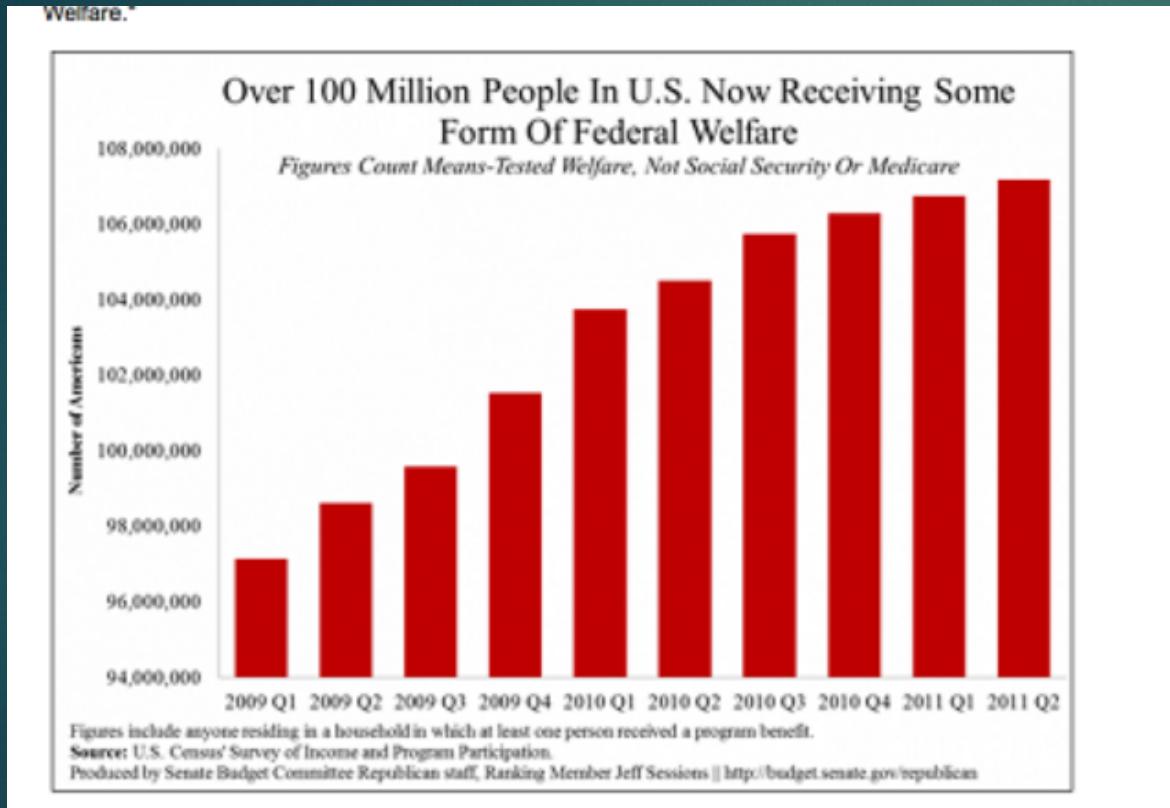


Graphs: Best practices lead to uncluttered look

- Design graph to support message; consider using talking head here, too
- Use minimal grid lines, ideally none
- Use thin lines, thin axes, thin bars, thin arrows to show trends
- Display subtle but visible data point marks—only enough to show trend
- Use minimal tick marks to display scale: usually just min/max on Y axis
- Label axes; label graphic items
- Opt for value labels wherever possible; delete some to avoid clutter

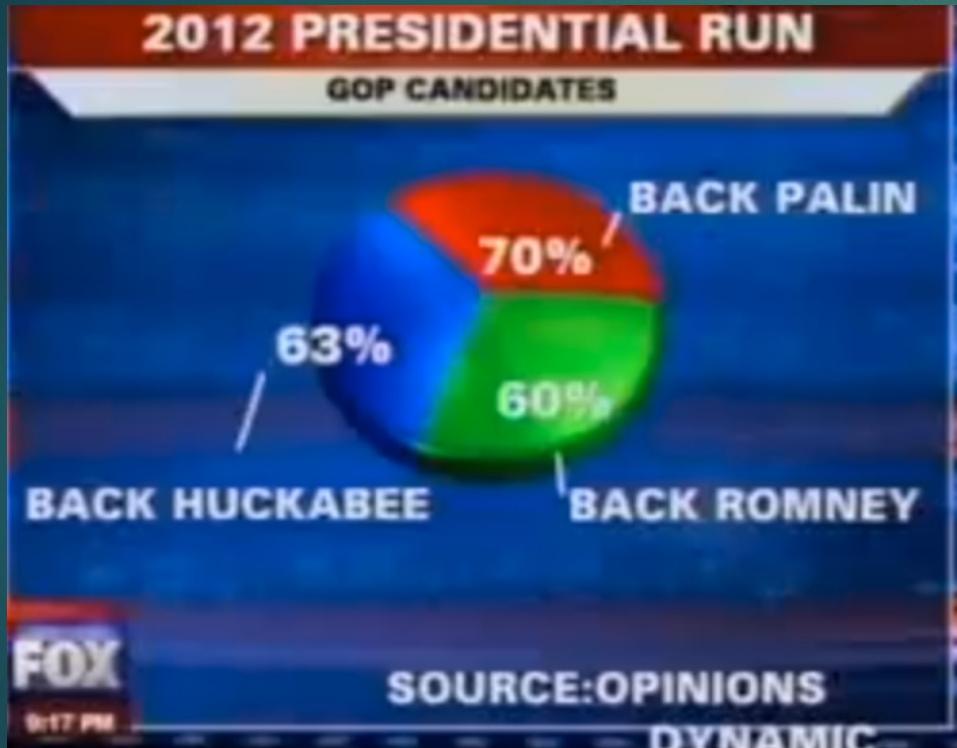


Misleading axes



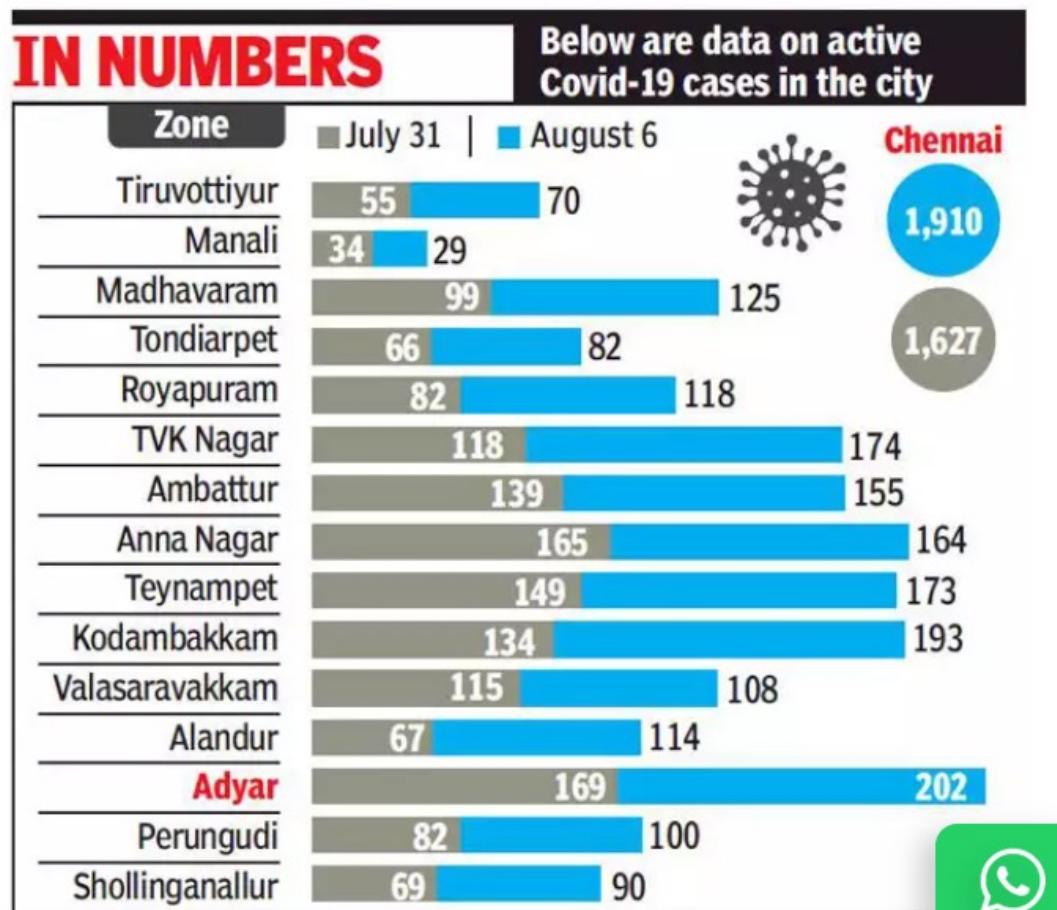
- ▶ It appears that the welfare problem is spiraling out of control.
- ▶ But note where the y-axis starts!

Wrong data



- ▶ The weights of the categories do not add up

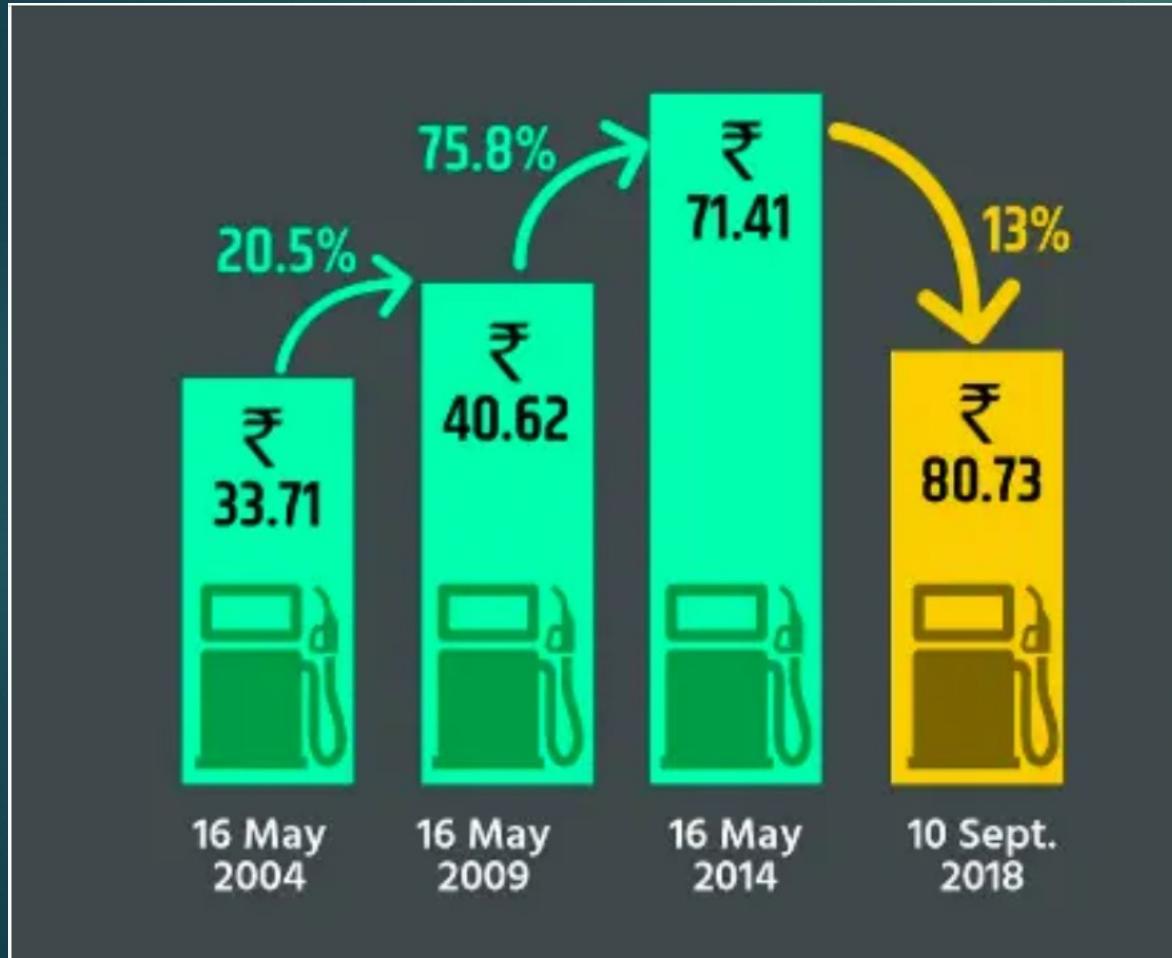
Inappropriate chart type



- ▶ It is difficult to compare the two datasets using the stacked bar chart.
- ▶ Bar chart with two series of columns would be better.

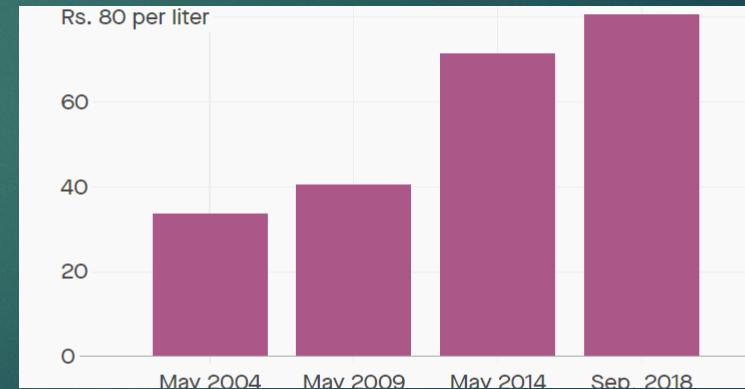
Incorrect bar chart

%increase in petrol prices



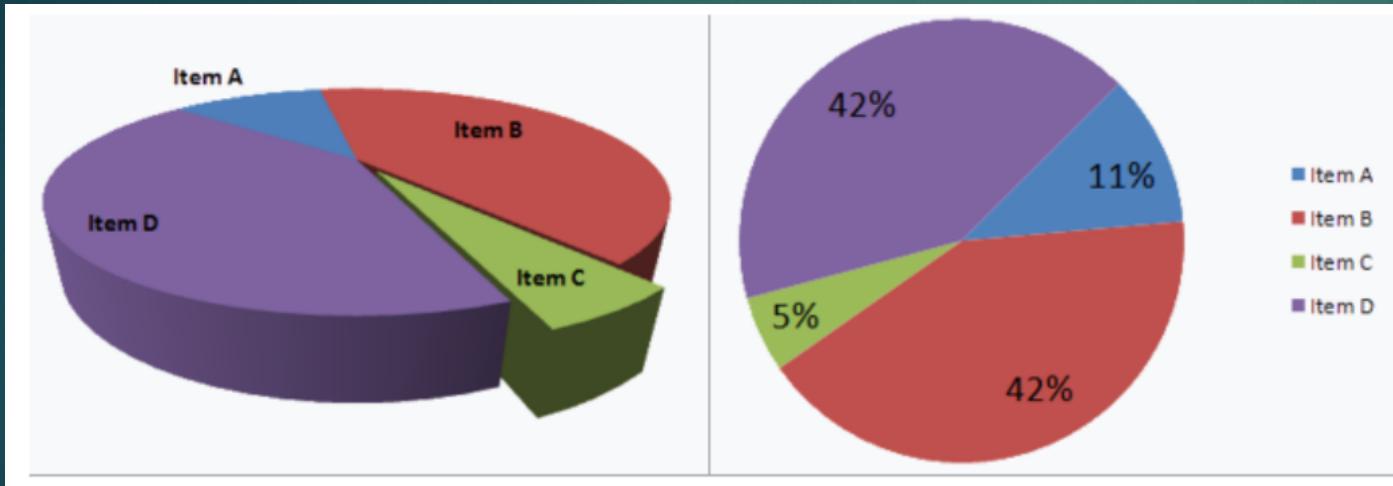
- ▶ An arrow points downward to Sept data point, however there is a positive change to the price.
- ▶ However, the %increase has reduced for Sept.

Accurate chart might be



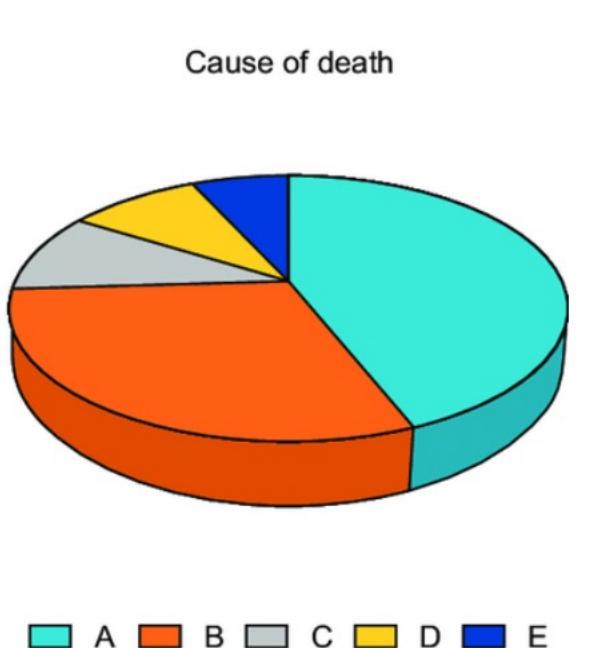
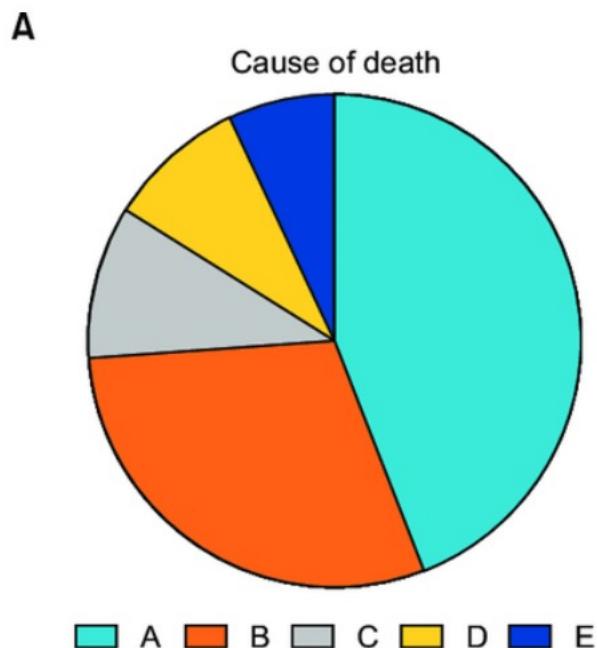
<https://qz.com/india/1388030/indiias-bjp-peddled-a-prime-example-of-chartjunk-to-its-millions-of-followers/>

Misleading 3D Pie Chart



- ▶ The slices closer to you will appear larger than those further away.
- ▶ Compare Item A and Item C in both the graph

Misleading 3D Pie Chart..



▶ A and B look quite similar in the 3D graph!

In, J., & Lee, S. (2017). Statistical data presentation. *Korean journal of anesthesiology*, 70(3), 267.

Thank you