STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

STATISTICS FOR DATA SCIENCE

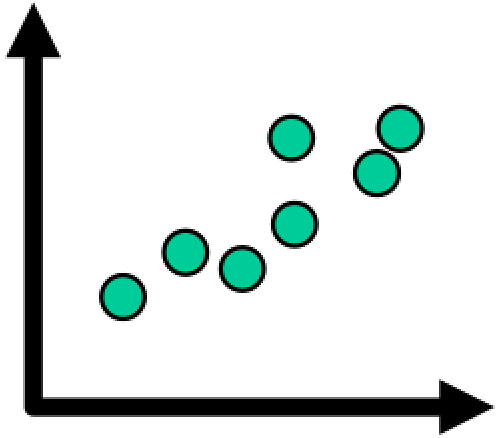
Data Visualization and Interpretation - Scatterplot

D. Uma

Department of Computer Science and Engineering

Scatter plot

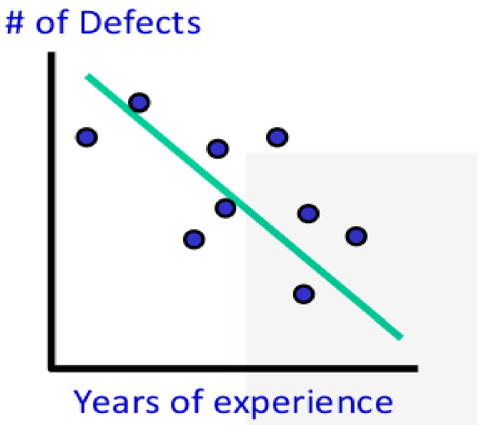
1. A diagram shows whether two variables are correlated
2. Shows pattern in the relationship that cannot be seen by just looking at the data
3. Used as a first step in analyzing correlation between pairs of variables before conducting advanced statistical analyses.
4. Works with both continuous and count data



Data Visualization:Scatter plot

A line manager for example may want to check the relationship
between:

- The number of training hours and employee productivity
- The number of defects and the experience of the staff.
- The equipment downtime and its cost of maintenance.



Data Visualization:Scatter plot

The relationship between

- Driving speed and fuel consumption.

- The number of people working on a shift and the average answer time in a call center.

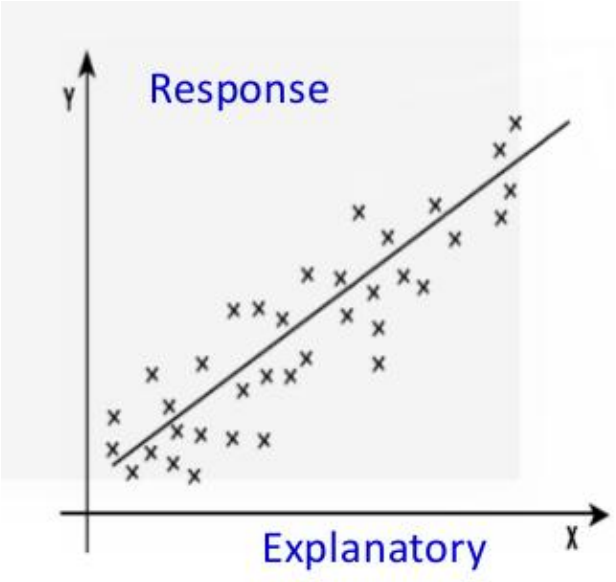
- The number of years of education someone has and the annual income of that person.

Source: www.slideshare.net

Data Visualization:Scatter plot

- When comparing an input with an output variable.
  > The explanatory variable is normally placed on the horizontal axis
  > The response variable is placed on the vertical axis

You may also compare two input or output variables



Source: www.slideshare.net

Data Visualization:Scatter plot

How to Construct a Scatter Plot?

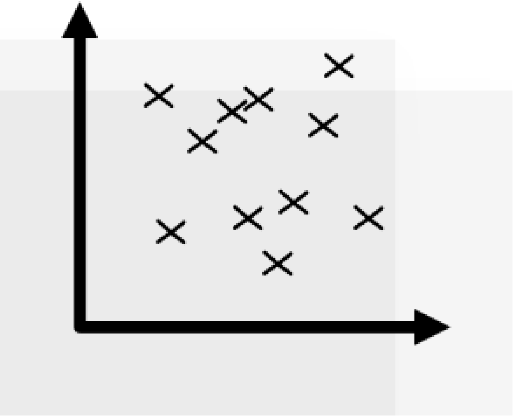
- Collect the two paired sets of data.
- Create a summary table of the data.
- Draw and label the horizontal and vertical axes.
- Plot the data pairs on the diagram by placing a dot at the
- intersection of each data pair.
- Look at how the two variables vary together.

Source: www.slideshare.net

Data Visualization:Scatter plot

Scatter plots can indicate several types of correlation:

No correlation when the data points are scattered randomly without showing any particular pattern.

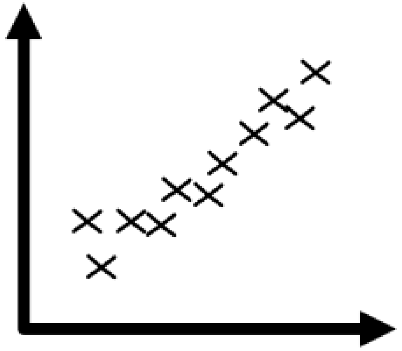


Source: www.slideshare.net

Data Visualization:Scatter plot

Scatter plots can indicate several types of correlation:

A positive correlation occurs when the values of one variable increase as the values of the other also increase



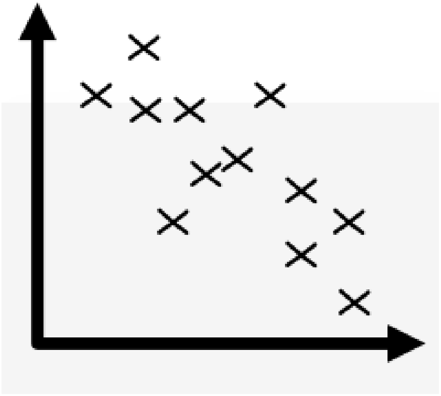
The fitted line slopes from bottom left to top right

Source: www.slideshare.net

Data Visualization:Scatter plot

Scatter plots can indicate several types of correlation:

A negative correlation occurs when the values of one variable increase as the values of the other decrease



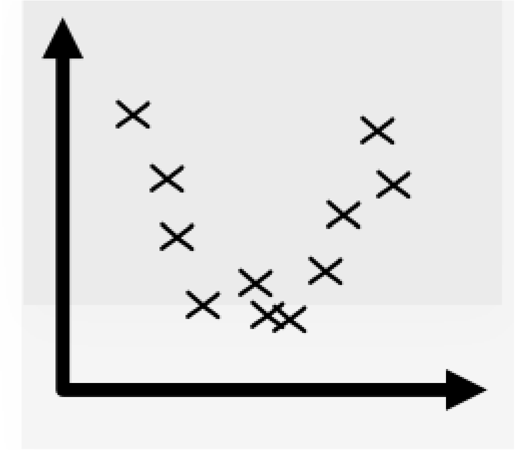
The fitted line slopes from upper left to lower right

Source: www.slideshare.net

Data Visualization:Scatter plot

Scatter plots can indicate several types of correlation:

Scatter plots can also indicate nonlinear relationships between variables

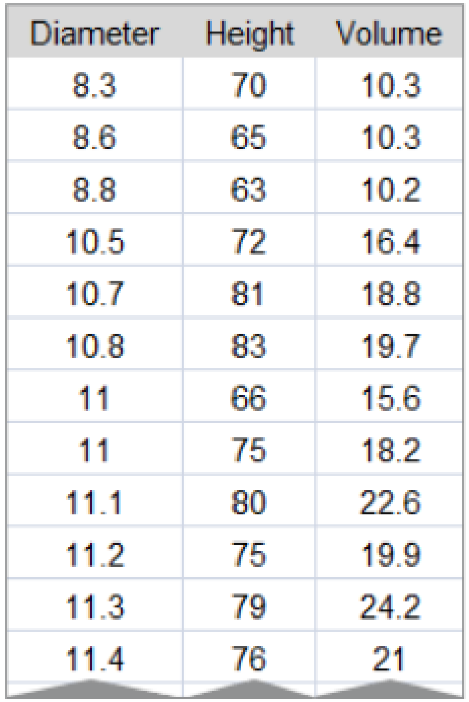


Source: www.slideshare.net

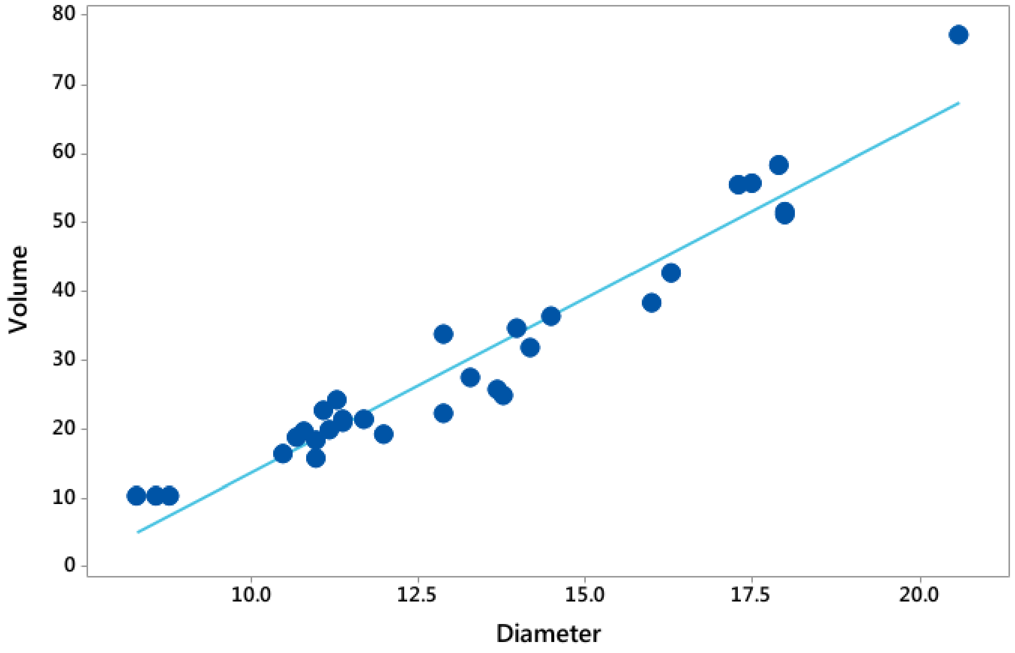
Data Visualization:Scatter plot

Scatter plot

Example



Diameter	Height	Volume
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11	66	15.6
11	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21

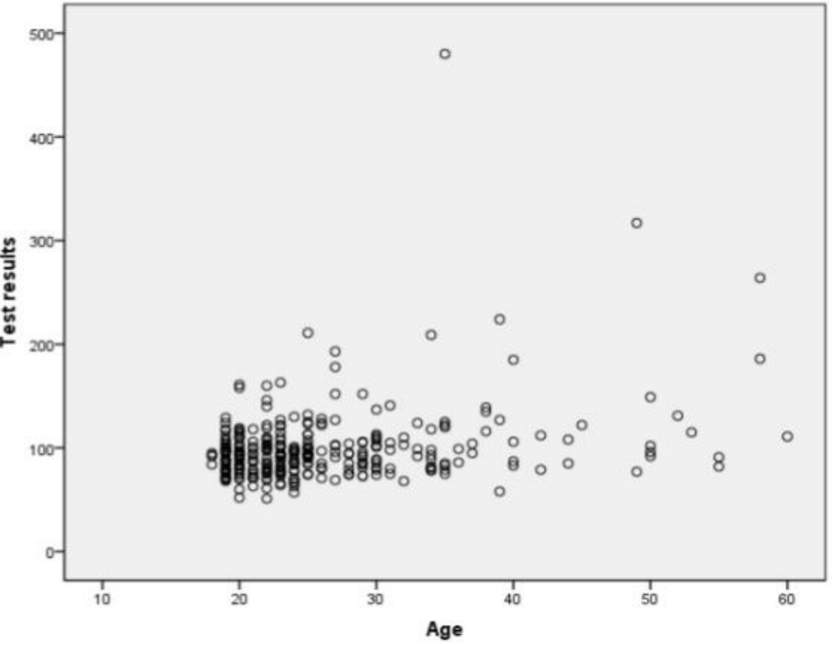


The **volume** and the **diameter** of sample trees in a forest.

Source: www.slideshare.net

Data Visualization:Scatter plot

Example – An analysis that was conducted for diagnosing the presence of diabetes at a workplace



- The population is generally young (75.8% are below thirty).
- This scatter plot illustrates that there is no obvious relationship between age and glucose levels.
- High glucose levels are found in all ages above twenty, and normal glucose levels are found in higher ages.

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation - Barchart

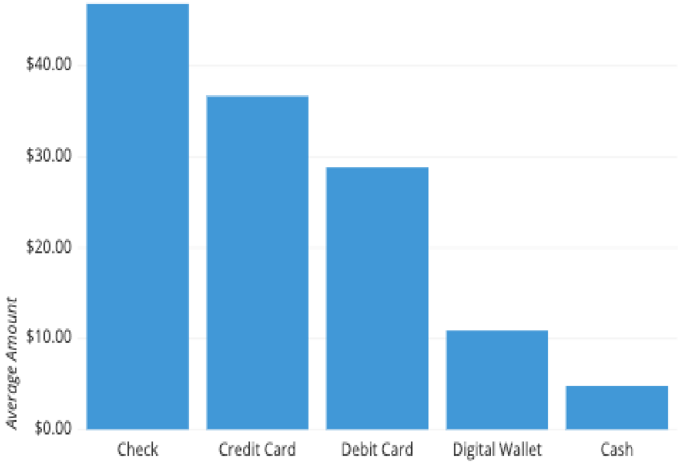
D. Uma

Department of Computer Science and Engineering

Data Visualization:Bar Chart

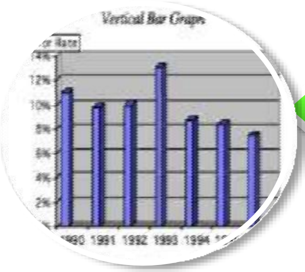
A **bar chart** or **bar graph** is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

A bar graph shows comparisons among discrete categories.

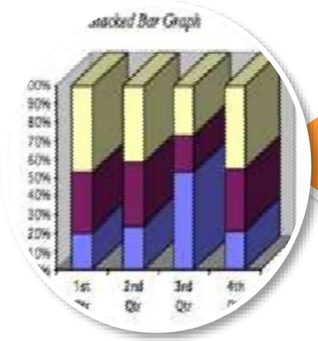


Source: www.google.com

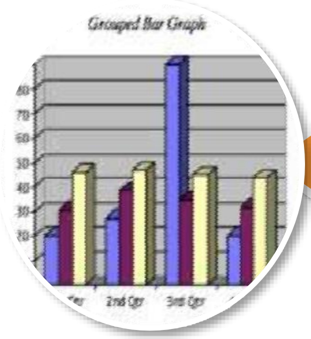
Data Visualization : BAR Chart



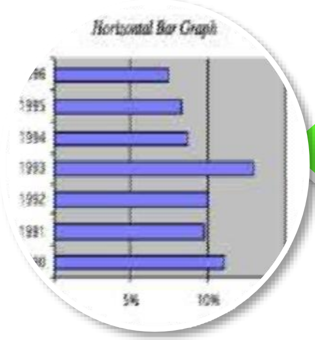
Single(Vertical)



Stacked



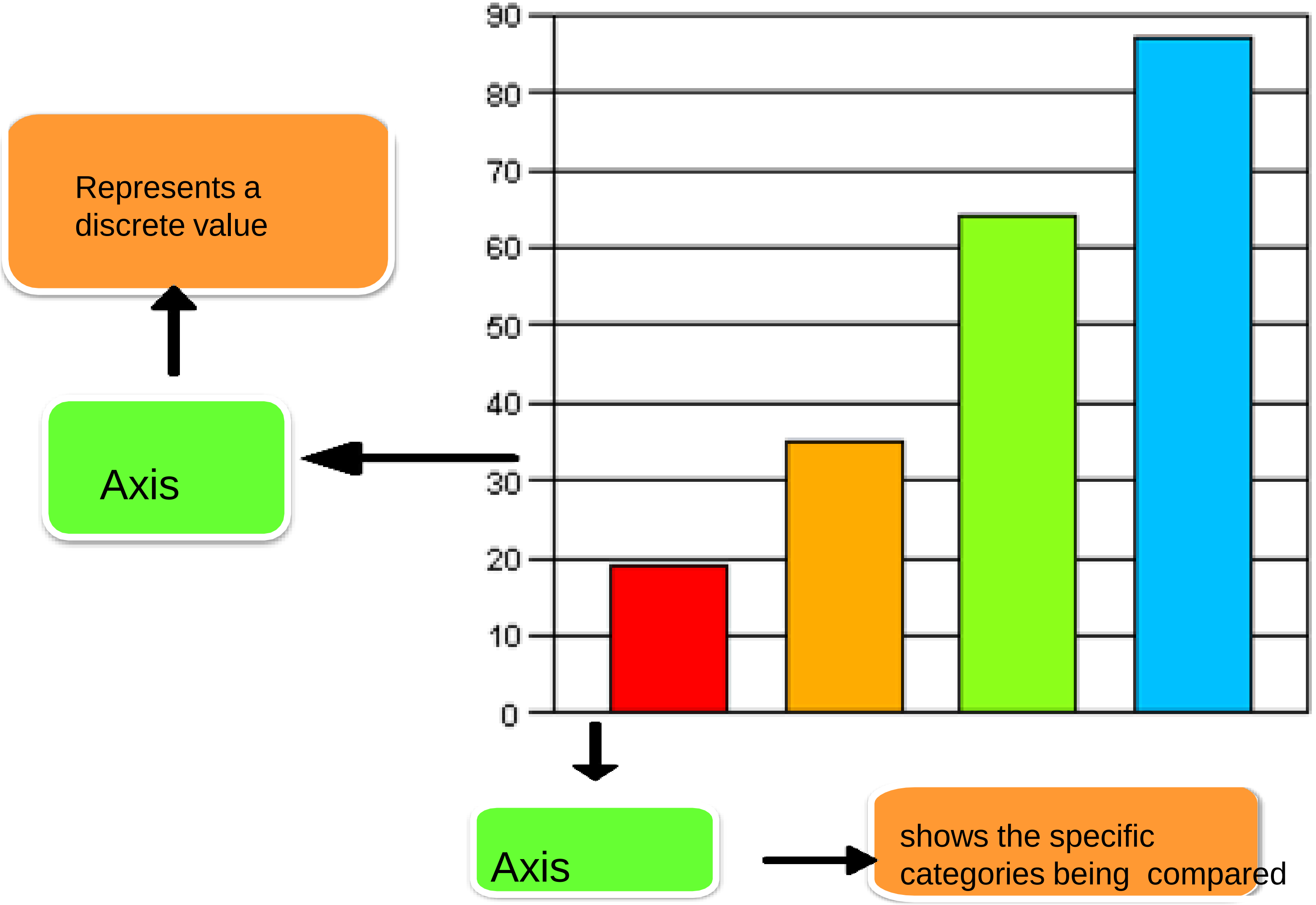
Grouped



Horizantal

www.slideshare.net

Data Visualization:Bar Chart



Represents a discrete value

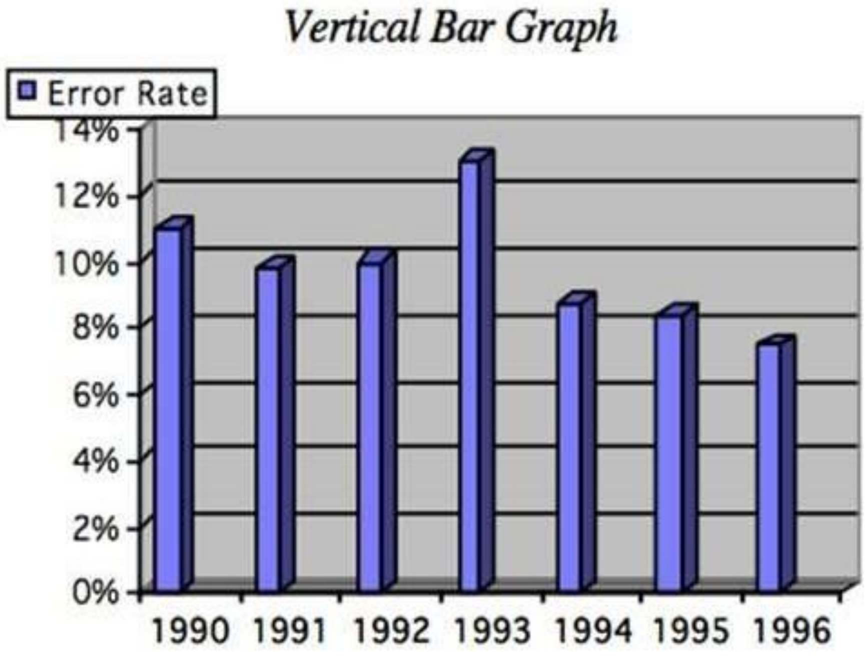
Axis

Axis

shows the specific categories being compared

Single Bar Chart

Single bar graphs are used to convey the discrete value of the item for each category shown on the opposing axis.

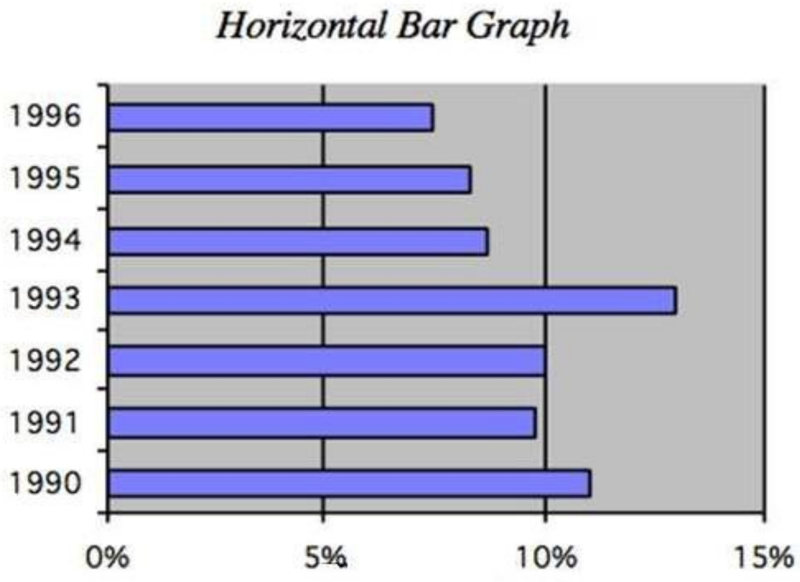


Source: www.slideshare.net

Data Visualization:Bar Chart

Horizontal Bar Chart

It is also possible to draw bar charts so that the bars are horizontal which means that the longer the bar, the larger the category.



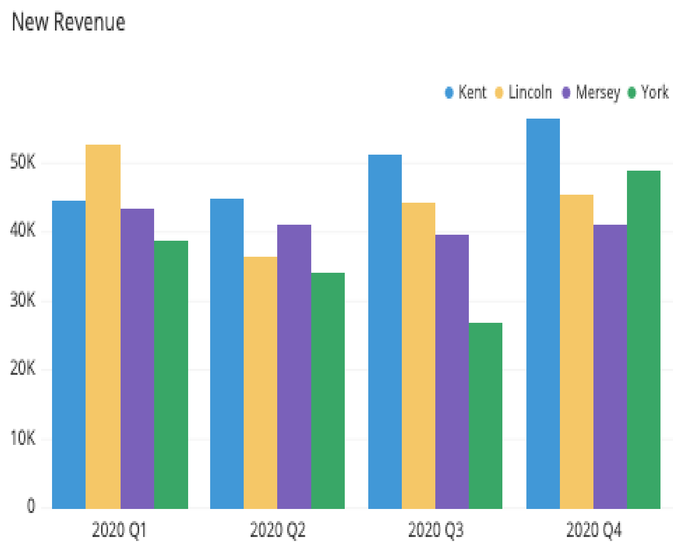
Source: www.slideshare.net

Data Visualization:Bar Chart

Grouped bar chart

Grouped bar charts are Bar charts in which multiple sets of data items are compared, with a single color used to denote a specific series across all sets.

A grouped or clustered bar graph is used to represent discrete values for more than one item that share the same category.



Source: Chartio

Grouped Bar Chart

- Grouped bar charts are a way of showing information about different sub-groups of the main categories.
- But care needs to be taken to ensure that the chart does not contain too much information making it complicated to read and interpret.

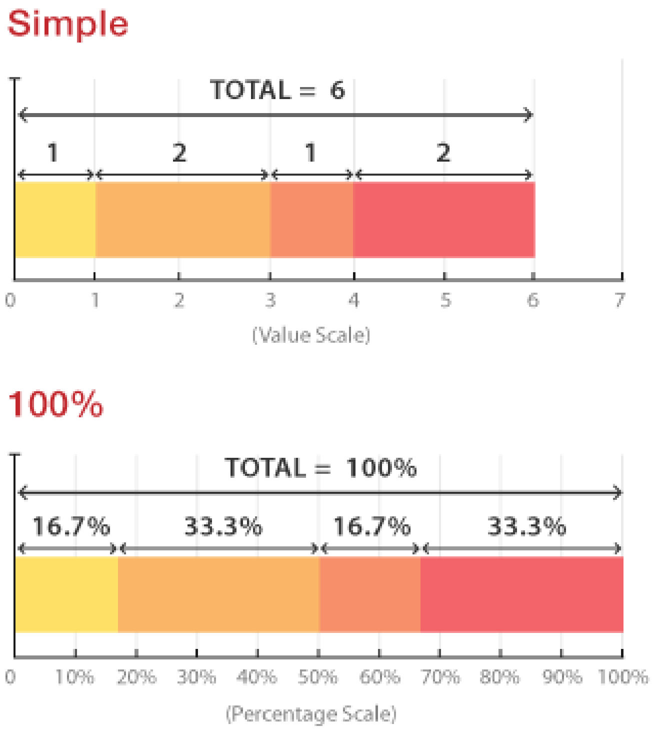
Source: www.slideshare.net

Stacked Bar Chart

- Simple Stacked Bar Graphs place each value for the segment after the previous one.

- The total value of the bar is all the segment values added together.

- Ideal for comparing the total amounts across each group/segmented bar.

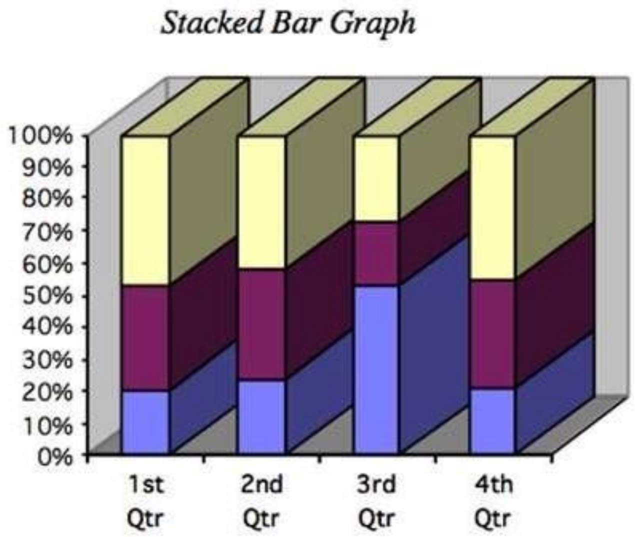


Source:datavizcatalogue.com

Stacked Bar Chart

100% Stack Bar Graphs show the percentage-of-the-whole of each group and are plotted by the percentage of each value to the total amount in each group.
This makes it easier to see the relative differences between quantities in each group.

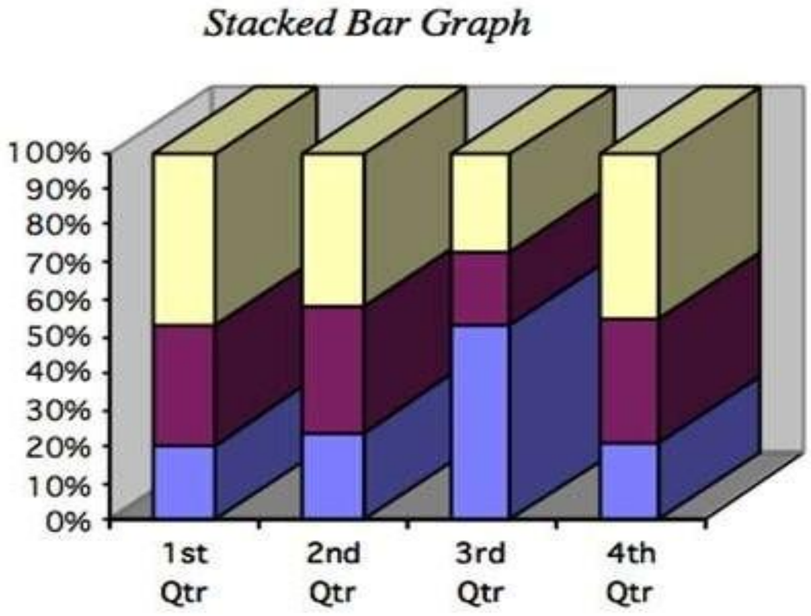
One major flaw of Stacked Bar Graphs is that they become harder to read the more segments each bar has. Also comparing each segment to each other is difficult, as they're not aligned on a common baseline.



Source: www.slideshare.net

Stacked Bar Chart

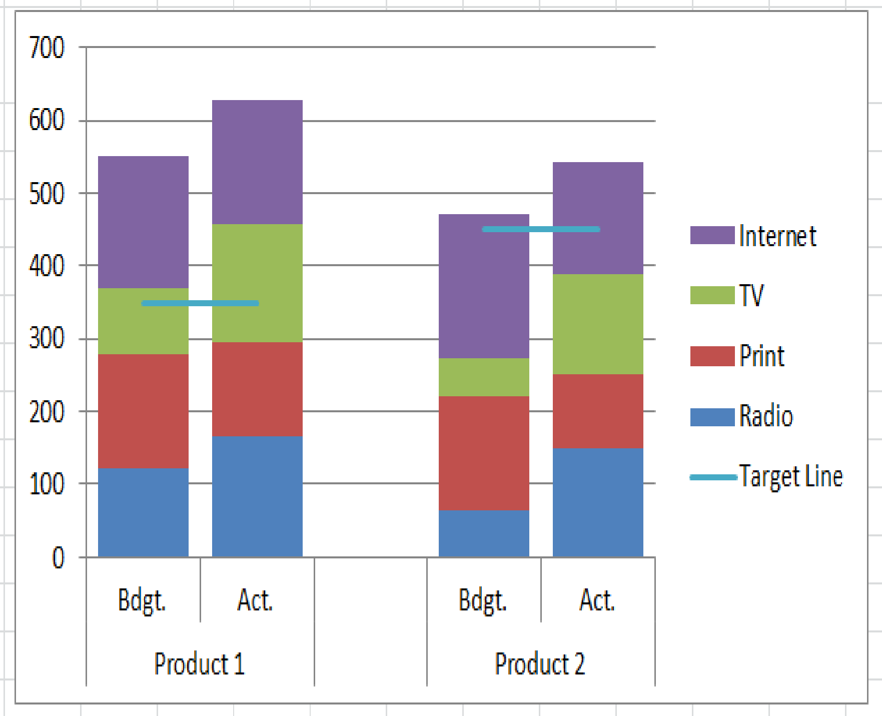
Some bar graphs have the bar divided into subparts that represent the discrete value for items that represent a portion of a whole group.



Source: www.slideshare.net

Stacked Bar Charts

Stacked bar charts are similar to grouped bar charts in that they are used to display information about the sub-groups that make up the different categories.



Stacked bar charts can also be used to show the percentage contribution different sub- groups contribute to each separate category.

Source: www.google.com

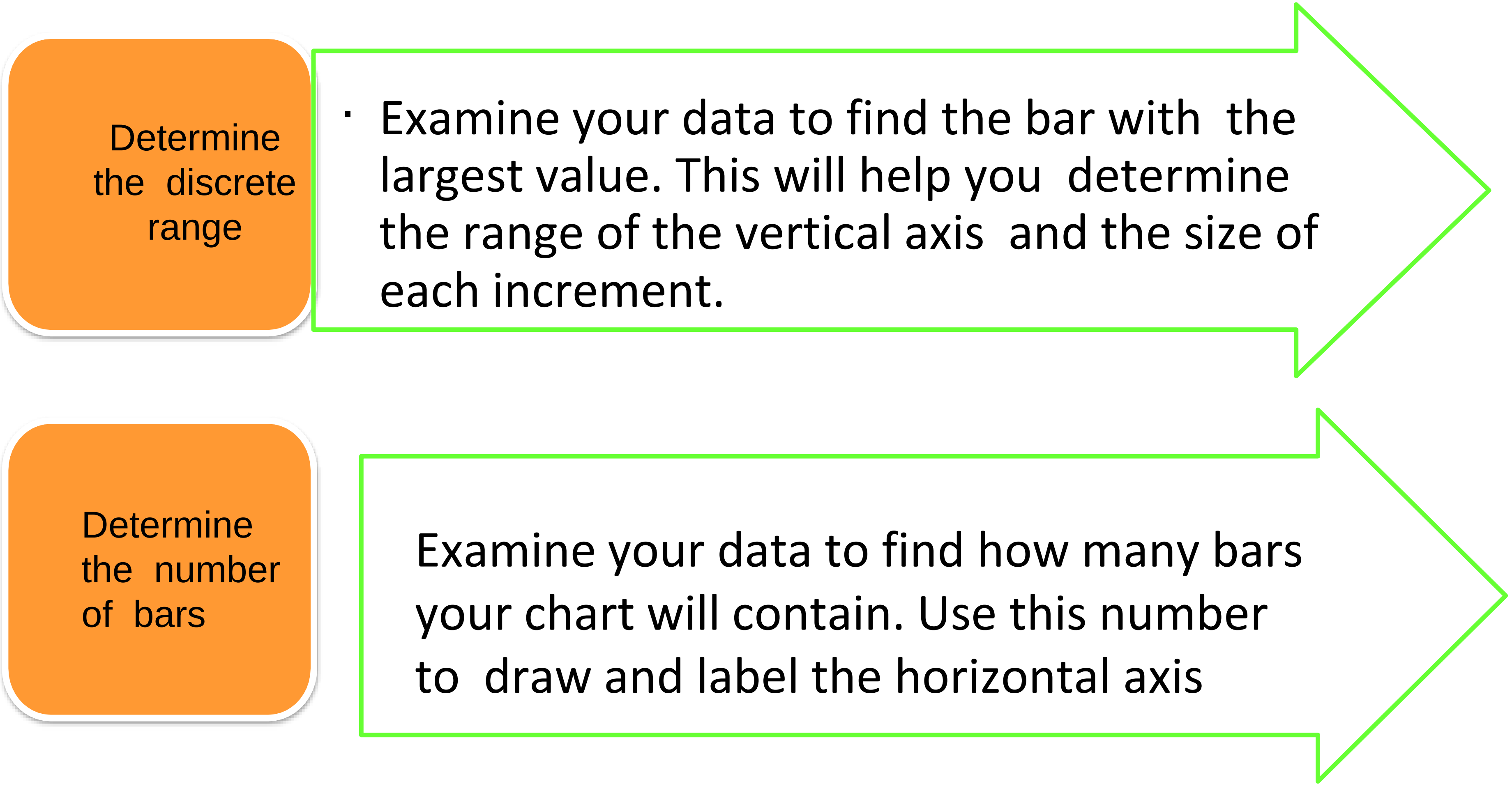
Uses of Bar Chart

Useful for comparing classes or groups of data.

In bar charts, a class or group can have a single category of data, or they can be broken down further into multiple categories for greater depth of analysis.

Source: www.slideshare.net

Data Visualization:Bar Chart



Determine the discrete range

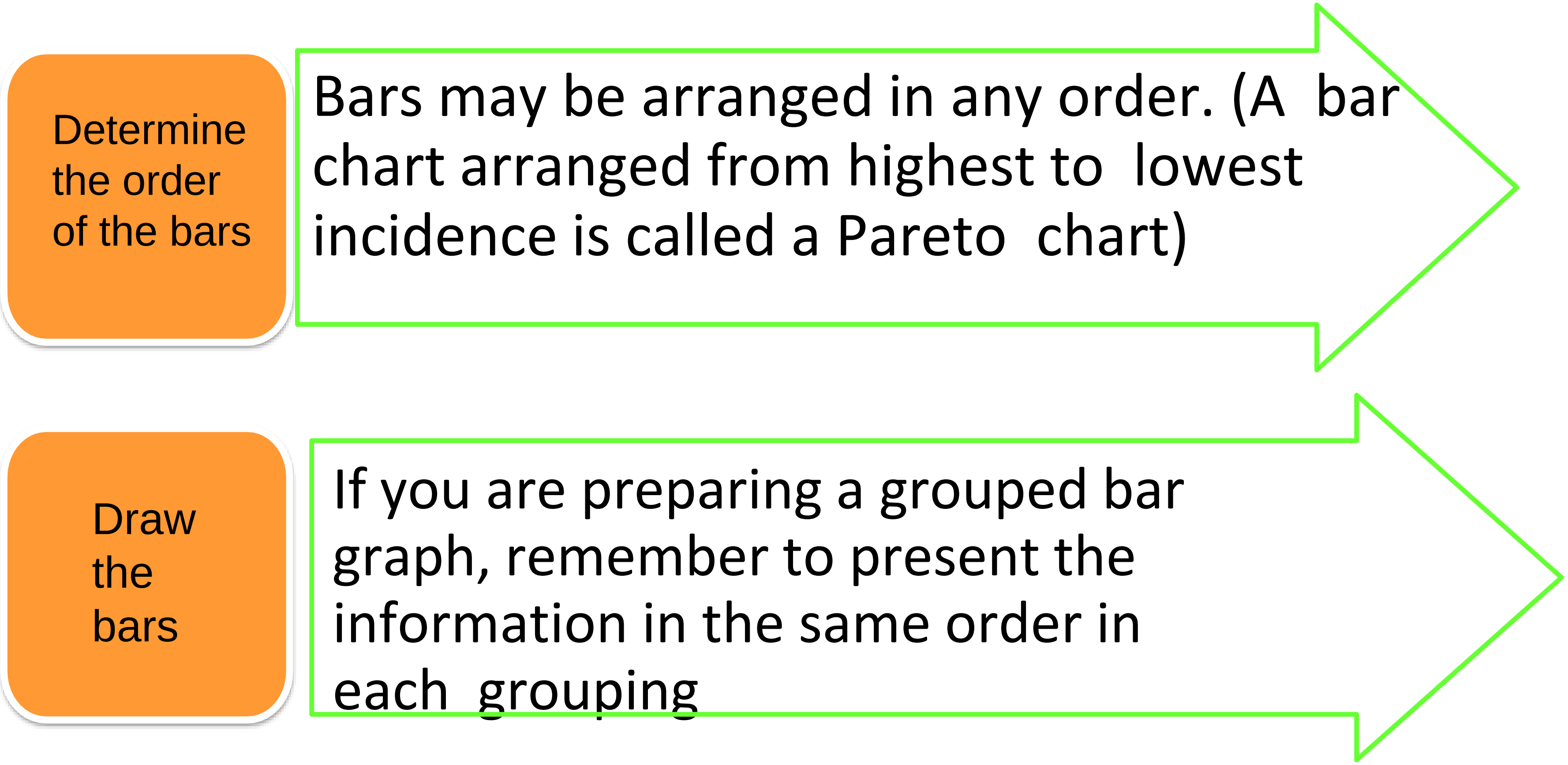
- Examine your data to find the bar with the largest value. This will help you determine the range of the vertical axis and the size of each increment.

Determine the number of bars

Examine your data to find how many bars your chart will contain. Use this number to draw and label the horizontal axis

Source: www.slideshare.net

Data Visualization:Bar Chart



Determine the order of the bars

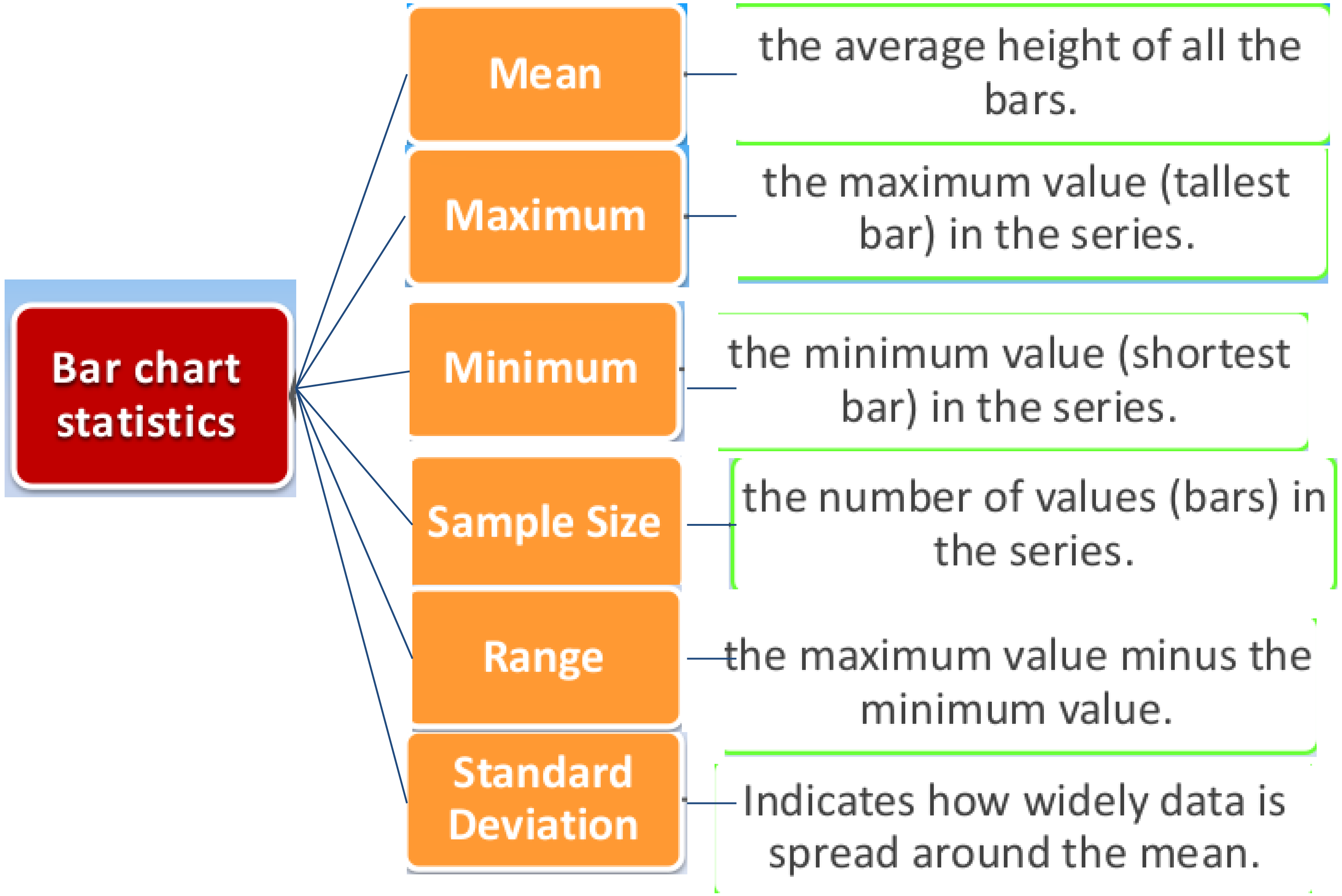
Bars may be arranged in any order. (A bar chart arranged from highest to lowest incidence is called a Pareto chart)

Draw the bars

If you are preparing a grouped bar graph, remember to present the information in the same order in each grouping

Source: www.slideshare.net

Data Visualization:Bar Chart



Bar chart statistics

- Mean: the average height of all the bars.
- Maximum: the maximum value (tallest bar) in the series.
- Minimum: the minimum value (shortest bar) in the series.
- Sample Size: the number of values (bars) in the series.
- Range: the maximum value minus the minimum value.
- Standard Deviation: Indicates how widely data is spread around the mean.

Source: www.slideshare.net

Difference between Bar and Histogram

Bar

Type of Data

In bar graphs are usually used to display "**categorical data**", that is data that fits into categories.

Histogram

Type of Data

Used to present "**continuous data**", that is data that represents measured quantity where, at least in theory, the numbers can take on any value in a certain range.

STATISTICS FOR DATA SCIENCE

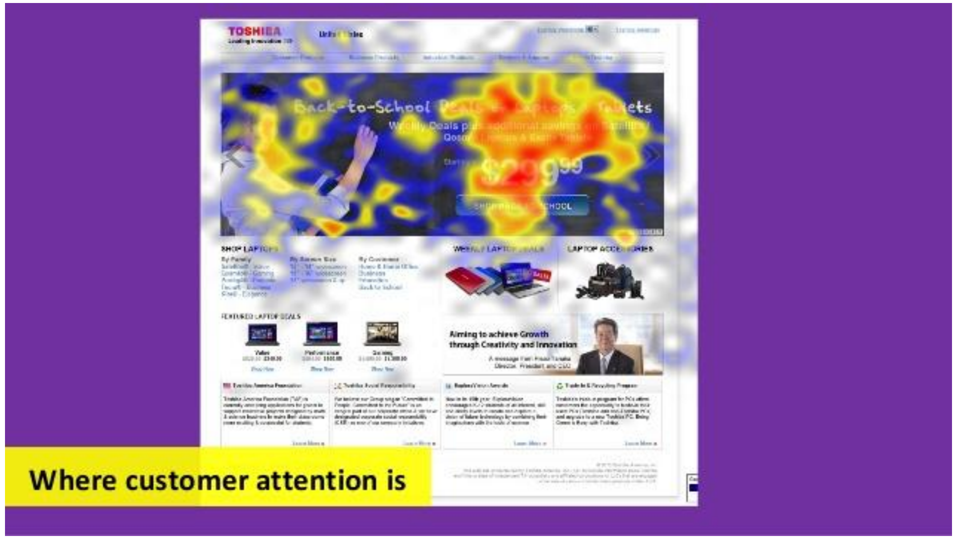
Data Visualization and Interpretation – Heat Map

D. Uma

Department of Computer Science and Engineering

Data Visualization: Heat Map

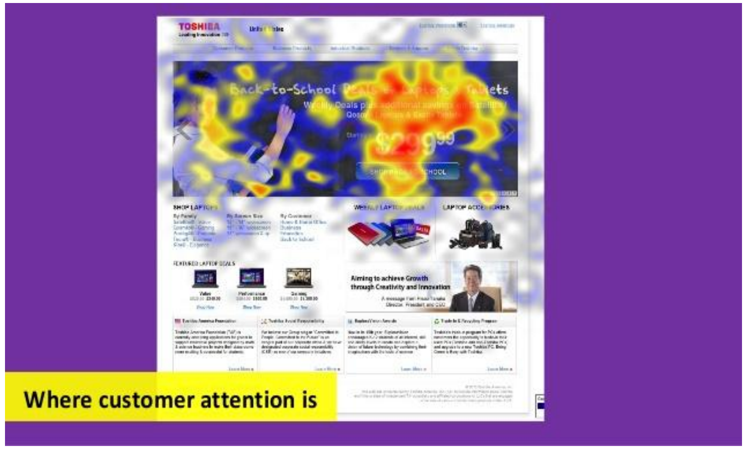
A **heat map** (or **heatmap**) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

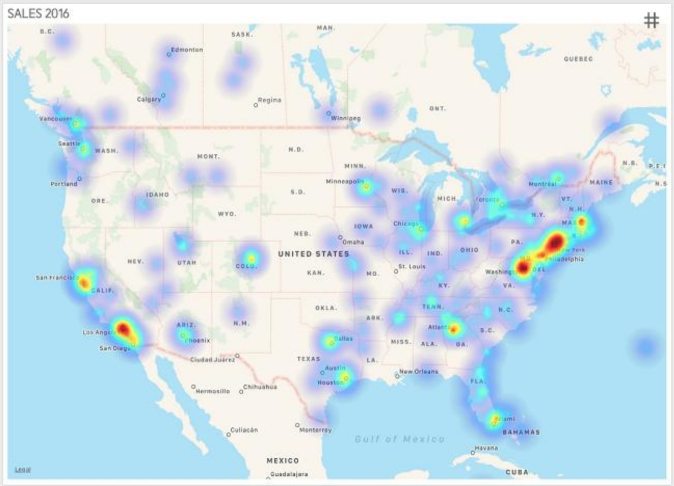

TOSHIBA
Back-to-School Deals on Laptops & Tablets
SHOP LAPTOPS
WEEKLY LAPTOP DEALS
LAPTOP ACCESSORIES
FEATURED LAPTOP DEALS
Aiming to achieve Growth through Creativity and Innovation
Where customer attention is

Data Visualization: Heat Map

A heat map is data analysis software that uses color the way a bar graph uses height and width: as a data visualization tool.

If you're looking at a web page and you want to know which areas get the most attention, a heat map shows you in a visual way that's easy to assimilate and make decisions from.





Source: www.infragistics.com

THANK YOU

D. Uma

Department of Computer Science

umaprabha@pes.edu