DATA PREPARATION AND ANALYSIS

UNIT- 5

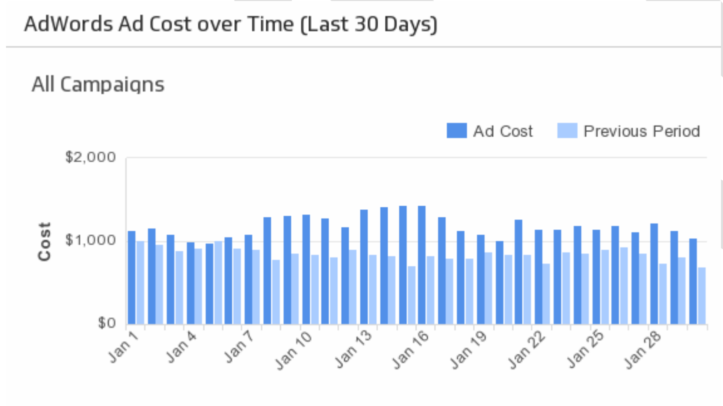
42. List and explain various types of data visualizations. [READ ANY 8]

Answer:

VARIOUS TYPES OF DATA VISUALIZATIONS:

- Bar Chart
- Line Chart
- Scatter plot
- Spark line
- Pie Chart
- Gauge
- Waterfall Chart
- Funnel Chart
- Heat Map
- Histogram
- Box Plot
- Maps
- Tables
- Indicators
- Area Chart
- Radar or Spider Chart
- Tree Map

1. Bar Chart



Bar charts are a popular graph visualization because of how easy you can scan them for quick information. Bar charts organize data into rectangular bars that make it a breeze to compare related data sets.

Use a bar chart for the following reasons:

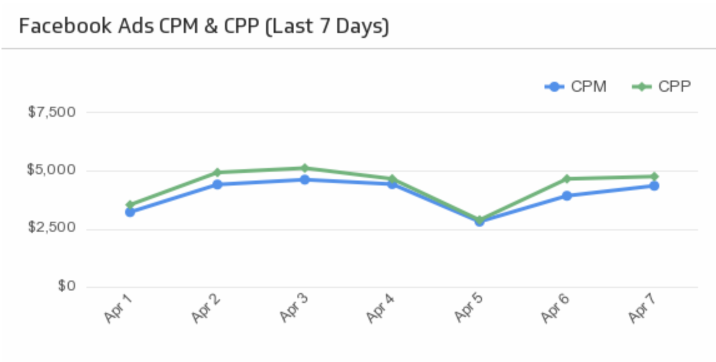
- You want to compare two or more values in the same category
- You want to compare parts of a whole
- You don't have too many groups (less than 10 works best)
- You want to understand how multiple similar data sets relate to each other

Best practices for a bar chart visualization:

If you use a bar chart, here are the key design best practices:

- Use consistent colours and labelling throughout so that you can identify relationships more easily
- Simplify the length of the y-axis labels and don't forget to start from 0 so you can keep your data in order

2. Line Chart



Line charts help to visualize data in a compact and precise format which makes it easy to rapidly scan information in order to understand trends. Line charts are used to show resulting data relative to a continuous variable - most commonly time or money. The proper use of color in this visualization is necessary because different colored lines can make it even easier for users to analyze information.

Use a line chart for the following reasons:

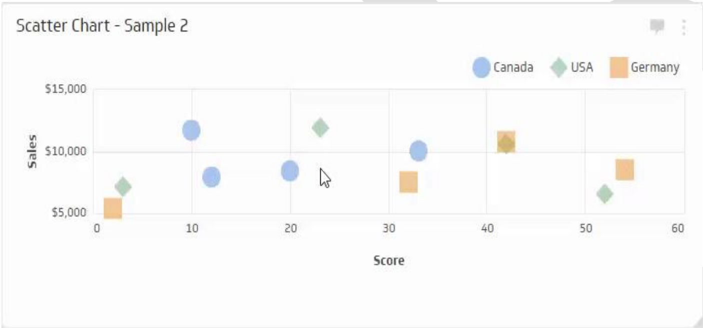
- You want to understand trends, patterns, and fluctuations in your data
- You want to compare different yet related data sets with multiple series
- You want to make projections beyond your data

Best practices for a line chart visualization:

If you use a line chart, here are the key design best practices:

- Along with using a different colour for each category you're comparing, make sure you also use solid lines to keep the line chart clear and concise
- To avoid confusion, try not to compare more than 4 categories in one line chart

3. Scatter plot



Scatter plots are the right data visualizations to use when there are many different data points, and you want to highlight similarities in the data set. This is useful when looking for outliers or for understanding the distribution of your data.

If the data forms a band extending from lower left to upper right, there most likely a positive correlation between the two variables. If the band runs from upper left to lower right, a negative correlation is probable. If it is hard to see a pattern, there is probably no correlation.

Use a scatter plot for the following reasons:

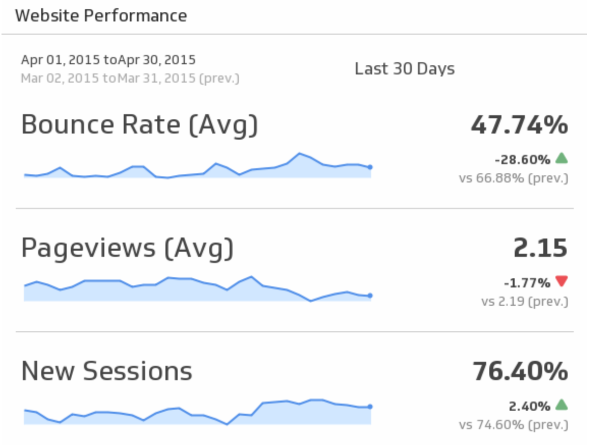
- You want to show the relationship between two variables
- You want a compact data visualization

Best practices for a scatter plot visualization:

If you use a scatter plot, here are the key design best practices:

- Although trend lines are a great way to analyze the data on a scatter plot, ensure you stick to 1 or 2 trend lines to avoid confusion
- Don't forget to start at 0 for the y-axis

4. Spark line:



Spark lines are arguably the best data visualization for showing trends because of how compact they are. They get the job done when it comes to painting a picture for your audience fast. Though, it is important to make sure your audience understands how to read spark lines correctly to optimize their use.

Use a spark line for the following reasons:

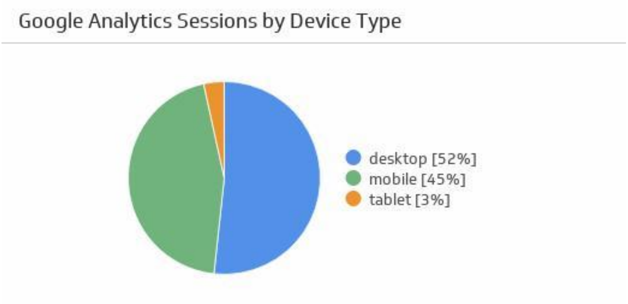
- You can pair it with a metric that has a current status value tracked over a specific time period
- You want to show a specific trend behind a metric

Best practices for a spark line visualization:

If you use a spark line, here are the key design best practices:

- To assist with readability, consider adding indicators on the side that give a better glimpse into the data, like in the example above
- Stick to one colour for your spark lines to keep them consistent on your dashboard

5. Pie Chart



Pie charts are interesting graph visualization. At a high-level, they're easy to read and understand because the parts-of-a-whole relationship is made very obvious. But top data visual experts agree that one of their disadvantages is that the percentage of each section isn't obvious without adding numerical values to each slice of the pie.

Use a pie chart for the following reasons:

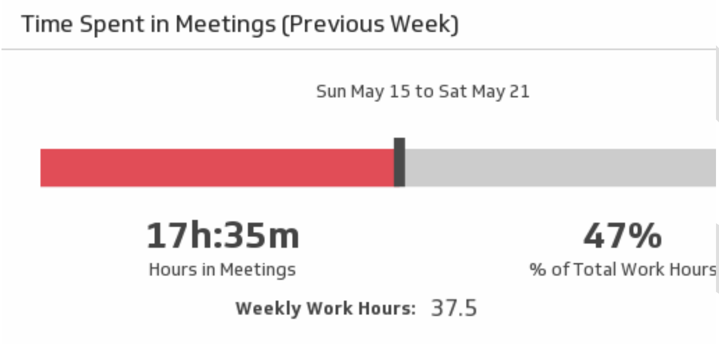
- You want to compare relative values
- You want to compare parts of a whole
- You want to rapidly scan metrics

Best practices for a pie chart visualization

If you use a pie chart, here are the key design best practices:

- Make sure that the pie slices add up to 100%. To make this easier, add the numerical values and percentages to your pie chart
- Order the pieces of your pie according to size
- Use a pie chart if you have only up to 5 categories to compare. If you have too many categories, you won't be able to differentiate between the slices.

6. Gauge



Gauges typically only compare two values on a scale: they compare a current value and a target value, which often indicates whether your progress is either good or bad, in the green or in the red.

Use a gauge for the following reason:

- You want to track single metrics that have a clear, in the moment objective

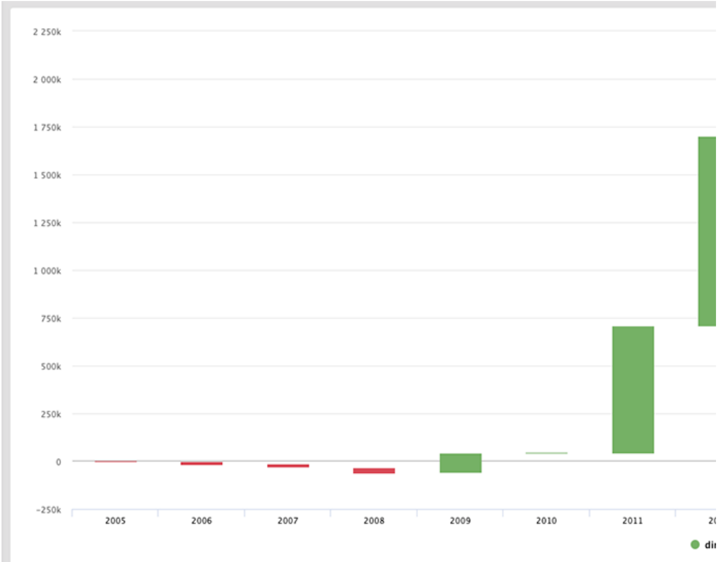
Best practices for a gauge visualization

If you use a gauge, here are the key design best practices:

- Feel free to play around with the size and shape of the gauge. Whether it's an arc, a circle or a line, it'll get the same job done
- Keep the colours consistent with what means "good" or "bad" for you and your numbers

- Use consistent colours and labelling throughout so that you can identify relationships more easily
- Simplify the length of the y-axis labels and don't forget to start from 0 so you can keep your data in order.

7. Waterfall Chart



A waterfall chart is an information visualization that should be used to show how an initial value is affected by intermediate values and resulted in a final value. The values can be either negative or positive.

Use a waterfall chart for the following reason:

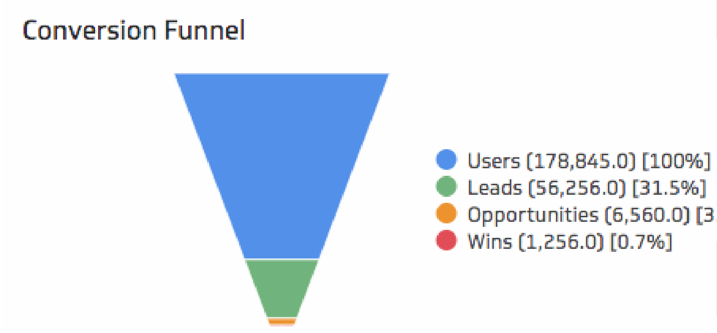
- To reveal the composition or makeup of a number

Best practices for a waterfall chart visualization

If you use a waterfall chart, here are the key design best practices:

- Use contrasting colors to highlight differences in data sets
- Choose warm colors to indicate increases and cool colors to indicate decreases

8. Funnel Chart



A funnel chart is your data visualization of choice if you want to display a series of steps and the completion rate for each step. This can be used to track the sales process, a marketing funnel or the conversion rate across a series of pages or steps. Funnel charts are most often used to represent how something moves through different stages in a process. A funnel chart displays values as progressively decreasing proportions amounting to 100 percent in total.

Use a funnel chart for the following reason:

- To display a series of steps and each step's completion rate

Don't use a funnel chart for the following reason:

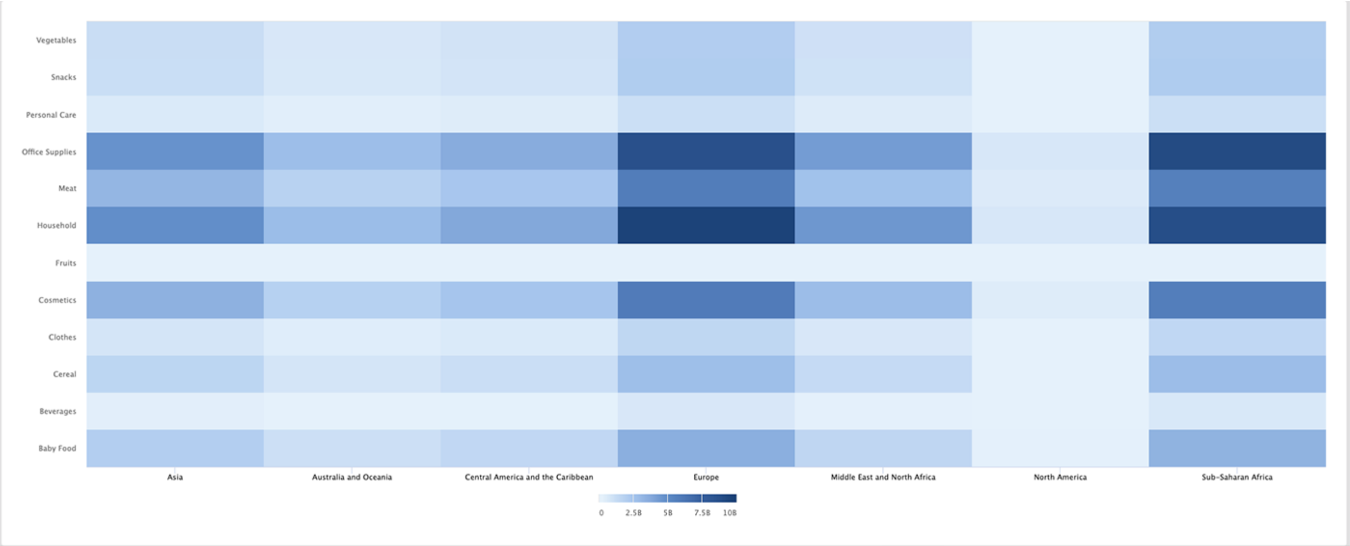
- To visualize individual, unconnected metrics

Best practices for a funnel chart visualization

If you use a funnel chart, here are the key design best practices:

- Scale the size of each section to accurately reflect the size of its data set
- Use contrasting colors or one color in gradating hues, from darkest to lightest as the size of the funnel decreases

9. Heat Map



A heat map or choropleth map is a data visualization that shows the relationship between two measures and provides rating information. The rating information is displayed using varying colors or saturation and can exhibit ratings such as high to low or bad to awesome, and needs improvement to working well.

It can also be a thematic map in which the area inside recognized boundaries is shaded in proportion to the data being represented.

Use a heat map for the following reasons:

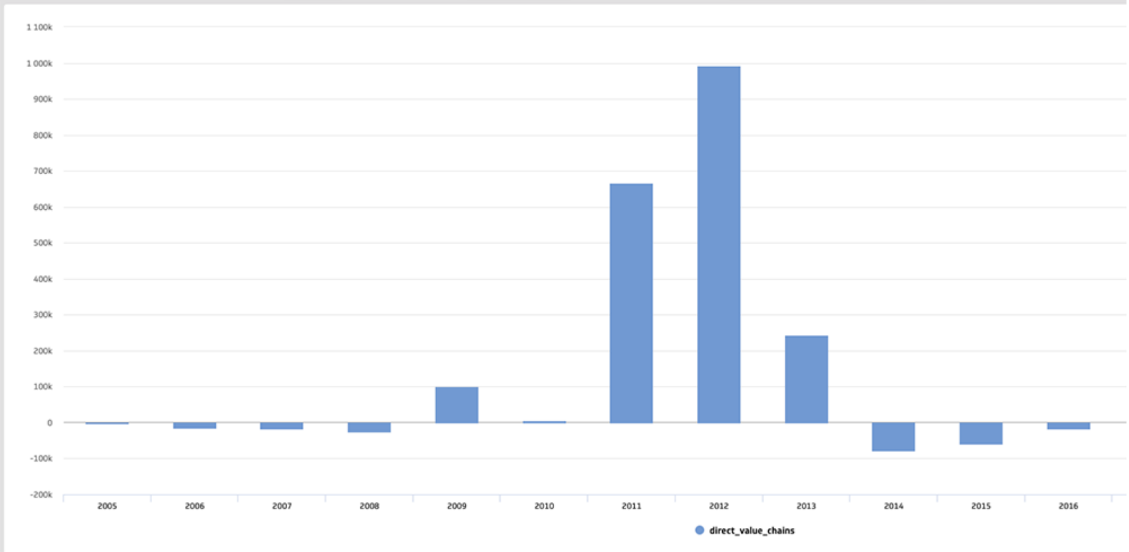
- To show a relationship between two measures
- To illustrate an important detail
- To use a rating system

Best practices for a heat map visualization

If you use a heat map, here are the key design best practices:

- Use a simple map outline to avoid distracting from the data
- Use a single color in varying shades to show changes in data
- Avoid using multiple patterns

10. Histogram



A histogram is a data visualization that shows the distribution of data over a continuous interval or certain time period. It's basically a combination of a vertical bar chart and a line chart. The continuous variable shown on the X-axis is broken into discrete intervals and the number of data you have in that discrete interval determines the height of the bar.

Histograms give an estimate as to where values are concentrated, what the extremes are and whether there are any gaps or unusual values throughout your data set.

Use a histogram for the following reason:

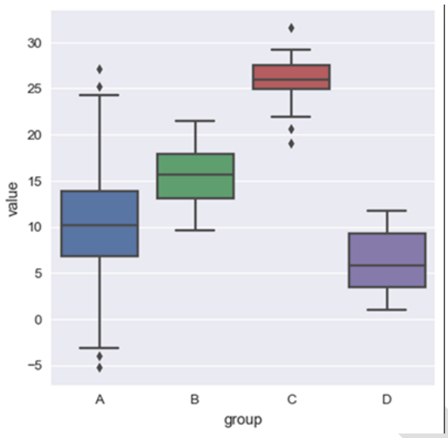
- To make comparisons in data sets over an interval or time
- To show a distribution of data

Best practices for a histogram visualization

If you use a histogram, here are the key design best practices:

- Avoid bars that are too wide that can hide important details or too narrow that can cause a lot of noise
- Use equal round numbers to create bar sizes
- Use consistent colours and labeling throughout so that you can identify relationships more easily

11. Box Plot



(Source: Python Graph Gallery)

A box plot, or box and whisker diagram, is a visual representation of displaying a distribution of data, usually across groups, based on a five number summary: the minimum, first quartile, the median (second quartile), third quartile, and the maximum.

The simplest of box plots display the full range of variation from minimum to maximum, the likely range of variation, and a typical value. A box plot will also show the outliers.

Use a box plot for the following reasons:

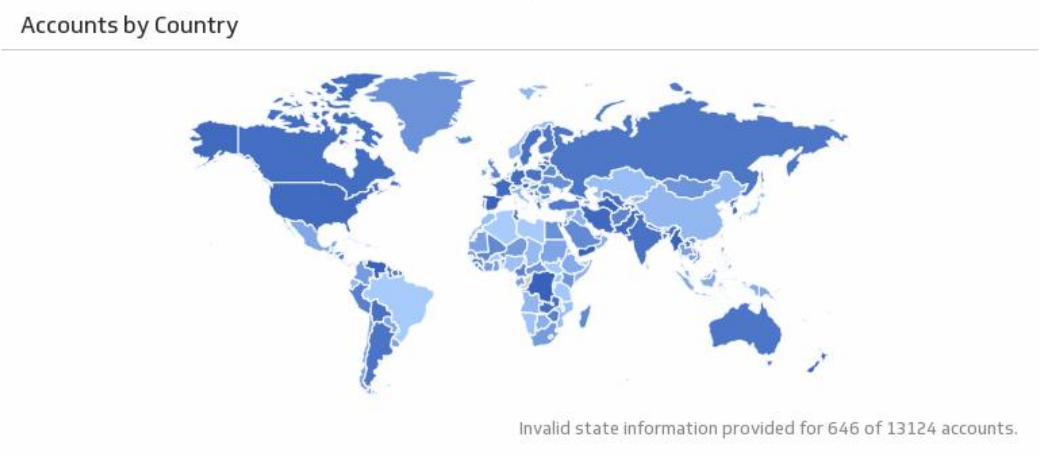
- To display or compare a distribution of data
- To identify the minimum, maximum and median of data

Best practices for a box plot visualization

If you use a box plot, here are the key design best practices:

- Ensure font sizes for labels and legends are big enough and line widths are thick enough to understand the findings easily
- If plotting multiple datasets, use different symbols, line styles or colour to differentiate each
- Always remove unnecessary clutter from the plots

12. Maps

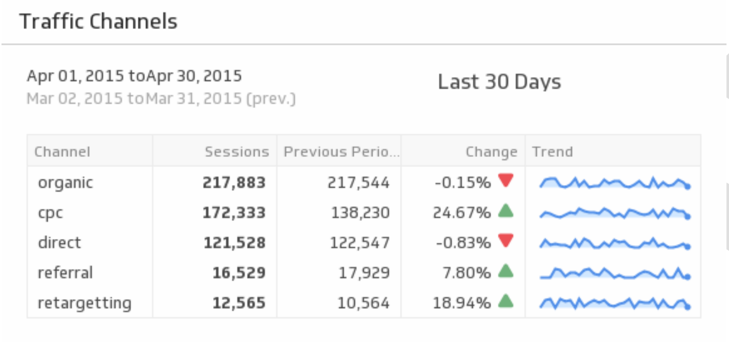

Accounts by Country
Invalid state information provided for 646 of 13124 accounts.

Maps are an amazing visualization to add to your dashboard if organizing data geographically tells an important story for your business. For example, if your dashboard is looking looking at monthly sales, it could be extremely useful to see the geographic locations of your customers.

Above, you'll find a map visualization that integrates with Salesforce to measure accounts by country. Keep in mind that if your dashboard is looking at daily sales, this visualization may provide less value to your day-to-day discussions.

Use a map for the following reason:

13. Tables



Traffic Channels

Apr 01, 2015 toApr 30, 2015
Mar 02, 2015 toMar 31, 2015 (prev.)

Last 30 Days

Channel	Sessions	Previous Perio...	Change	Trend
organic	**217,883**	217,544	-0.15% ▼	
cpc	**172,333**	138,230	24.67% ▲	
direct	**121,528**	122,547	-0.83% ▼	
referral	**16,529**	17,929	7.80% ▲	
retargetting	**12,565**	10,564	18.94% ▲	

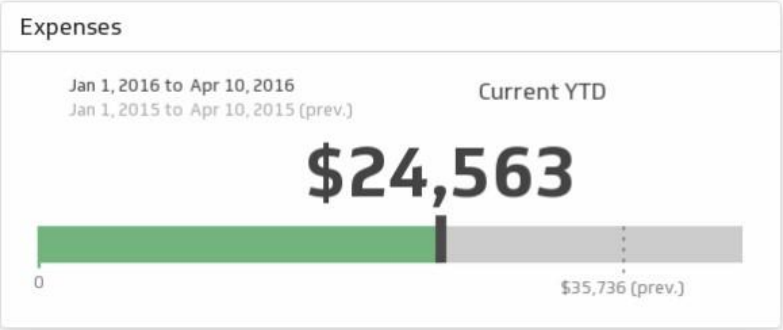
Tables can display both data points and graphics, such as bullet charts, icons, and spark lines. This visualization type also organizes your data into columns and rows, which is great for reporting.

Above is an example of how to bring in your Google Analytics data into a table, so that you can see all the information you need in one place.

One thing to keep in mind is that tables can sometimes be overwhelming if you have a dashboard with many metrics that you want to display. It's important to find a happy medium between large amounts of data (confusing) and too little data (waste of dashboard space).

Use a table for the following reasons:

14. Indicators:



Expenses

Jan 1, 2016 to Apr 10, 2016
Jan 1, 2015 to Apr 10, 2015 (prev.)

Current YTD

$24,563

- Geography is an important part of your data story

Best practices for a map visualization

If you use a map visualization, here are the key design best practices:

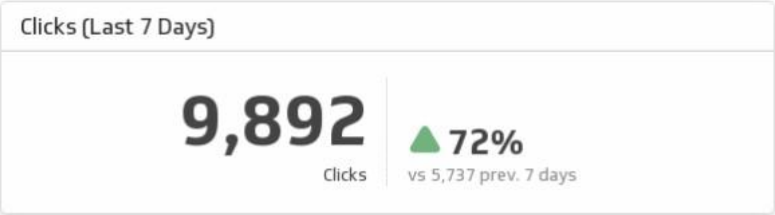
- Avoid using multiple colours and patterns on your map. Use varying shades of the same colour instead
- Make sure to include a legend with your map, so that everyone understands what the data means

- You want to display two-dimensional data sets that can be organized categorically
- You can drill-down to break up large data sets with a natural drill-down path

Best practices for a table visualization

If you use a table, here are the key design best practices:

- Be mindful of the order of the data. Make sure that labels, categories and numbers come first then move on to the graphics
- Try not to have more than 10 different rows in your table to avoid clutter



Clicks (Last 7 Days)

9,892

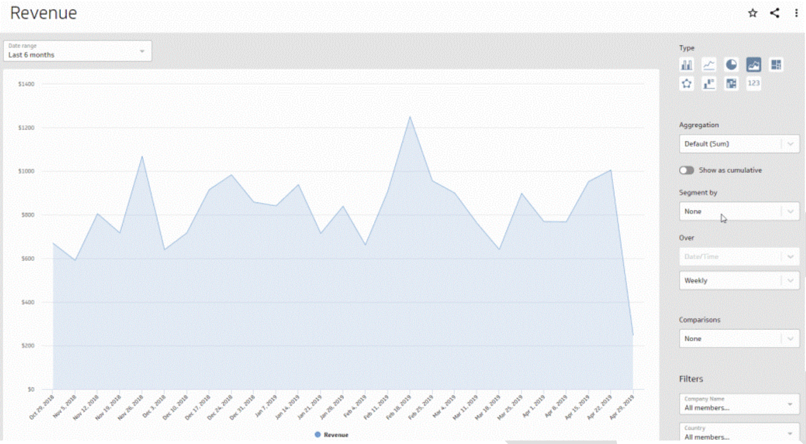
Clicks

▲ 72%

vs 5,737 prev. 7 days

Indicators are useful for an at a glance view of a metric you need to keep track of. An indicator is simply a number showing the current value of whichever performance metric you're tracking. To make it more useful, add a comparison to the previous time period to show whether your metric is tracking up or down.

15. Area Chart



Revenue

Date range
Last 6 months

Type

Aggregation
Default (Sum)

Show as cumulative

Segment by
None

Over
Date/Time
Weekly

Comparisons
None

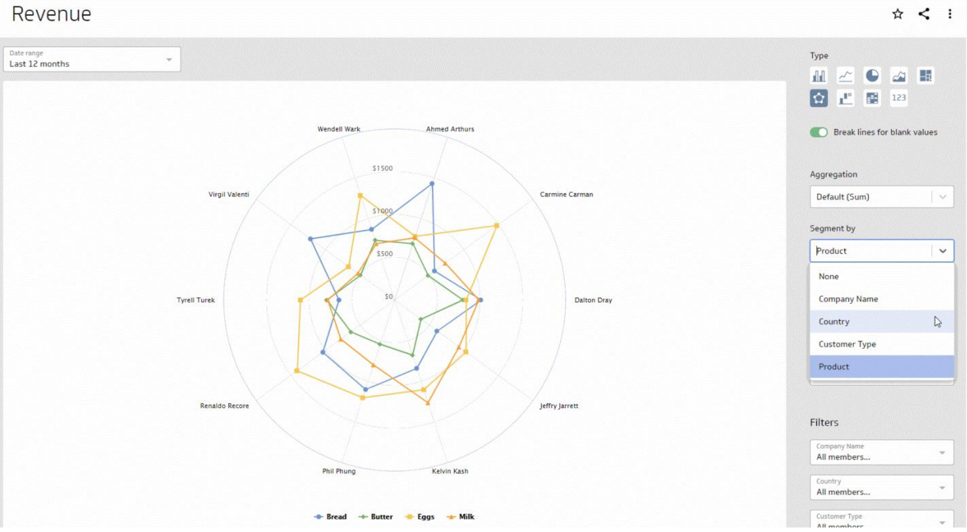
Filters

Company Name
All members...

Country
All members...

An area chart is very similar to a line graph but may do a better job at highlighting the relative differences between items. Use an area chart when you want to see how different items stack up or contribute to the whole.

16. Radar or Spider Chart



Revenue

Date range
Last 12 months

Type

Break lines for blank values

Aggregation
Default (Sum)

Segment by
Product

Filters

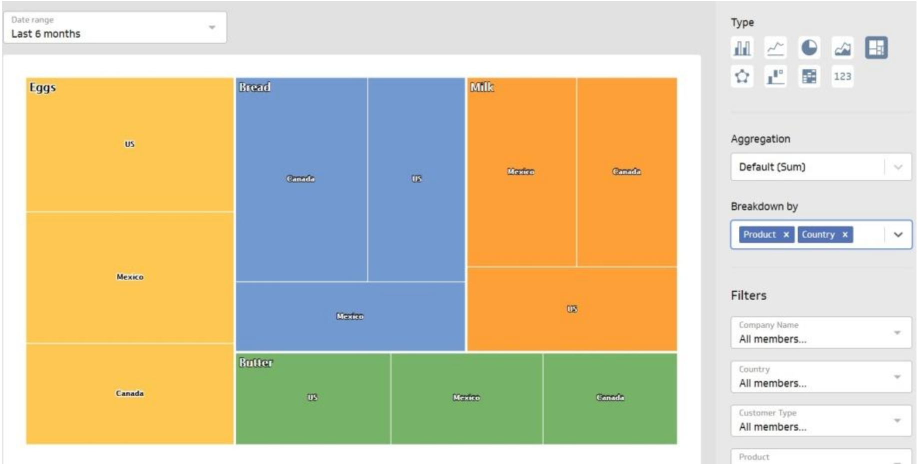
Company Name
All members...

Country
All members...

Customer Type
All members...

A radar chart is useful for understanding the relative differences between items in your data. Radar charts make it easy to compare multiple items and see if there are differences that may be worth further investigation.

17. Tree map



Date range
Last 6 months

Aggregation
Default (Sum)

Breakdown by
Product × Country ×

Filters

Company Name
All members...

Country
All members...

Customer Type
All members...

Product

A treemap is a visual tool that can be used to break down the relationships between multiple variables in your data. They can be used strictly as a presentation vehicle to show how your products roll up into different categories, for example. A treemap can be broken down into 2-3 different layers to show the hierarchical relationship between items.

43. Illustrate how data visualizations are implemented in designing statistical analysis.

44. How can you design the problem and data to be presented into an information visualization?

Answer:

Design plays a key role:

Designing data visualization is not just about the visuals, but why those visuals matter in the data analysis process and how they can be of actual use for the user. We work on designing for iterative data exploration, a guided experience that helps the business user get to their business answers as quickly as possible, and a flexible work flow that supports analytics experts and novices alike. Design work in this field can have powerful implications for data users and effect on how businesses operate.

5 Steps to Designing Information Visualization: The overall process is simple and once you‘ve reviewed the process; it should feel like common sense:

1. Define the problem

2. Define the data to be represented

3. Define the dimensions required to represent the data

4. Define the structures of the data

5. Define the interaction required from the visualization

1. Define the Problem

- As with any user experience work; the first step is to define the problem that your information visualization will solve. This will usually require some user research to answer the questions; —what does my user need from this? and —how will they work with it? You may be trying to explain something to a user or you may be trying to enable them to make new connections or observations; it might even be that the user is trying to prove a theory.
- You should also take into account any specific factors that are unique to the user base during this research. What is their level of education or ability with data handling? What kind of experience do they have with the data in the past? This will guide the level of complexity of the output and clarify the overall needs of the user.

2. Define the Data to be represented:

There are three main types of data that can be represented through information visualization and the way that they are mapped can vary dramatically – so it pays to have it clear in your mind before you start designing, what data will you use?

1. Quantitative data – this data which is numerical.

2. Ordinal data – data which doesn‘t have numbers but does have an intrinsic form of order. (Think days of the week, for example.)

3. Categorical data – data which has neither numbers nor intrinsic order. (Such as business names or place names).

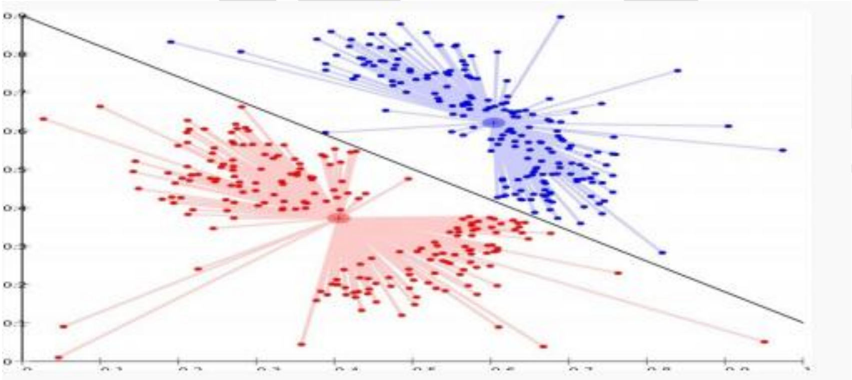
3. Define the Dimensions Required to Represent the Data:

- The number of dimensions or attributes of a data set must be considered carefully as it will determine, to a great extent, the possible information visualizations that can be used to represent the data. The more dimensions that are represented in the data – the more confusing it can be to comprehend the information visualization. Thus it‘s worth noting that the data with large numbers of dimensions may well benefit from using a highly interactive representation rather than a static one.
- Dimensions can be either dependent or independent of each other. It is the dependent dimensions which vary and which we would expect to need to analyze with respect to the independent dimensions. There are four types of analysis which can be conducted based on the number of dependent dimensions to be studied:

1. Univariate analysis – where a single dependent variable is studied against independent variables

2. Bivariate analysis – where two dependent variables are studied against independent variables 3. Trivariate analysis – where three dependent variables are studied against independent variables.

4. Multivariate analysis – where more than three dependent variables are studied against independent variables.

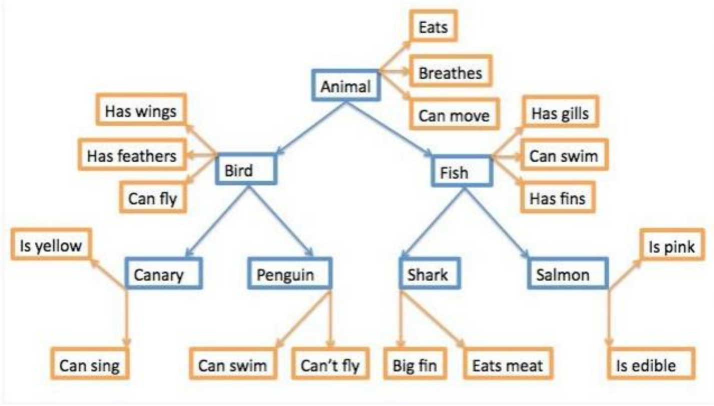


An image of multi-variate analysis where relationships between data points are numerous and dependant.

4. Define the Structures of the Data:

 This is all about examining how the data sets will relate to each other, common relationship structures include: Linear relationships – where data can be shown in linear formats such as tables, vectors, etc.

- Temporal relationships – where data changes over the passage of time
- Spatial relationships – data that relates to the real world (such as map data or an office floor plan) this is sometimes also known as a geographical relationship
- Hierarchical relationships – data that relates to positions in a defined hierarchy (from an office management structure to a simple flowchart)
- Networked relationships – where the data relates to other entities within the same data



An example of a hierarchical network model is shown above.

5. Define the Interaction Required from the Visualization:

The final part of the design process requires that you understand the level of interaction required from the information visualization by the user. There are three categories of interaction:

1. Static models – these models are presented —as is|| such as maps in a Road Atlas that you keep in a car. They cannot be modified by the user.

2. Transformable models – these models enable the user to transform or modify data. They may allow the user to vary parameters for analysis or choose a different form of visual mapping for the data set.

3. Manipulable models – these models give the user control over the generation of views. For example; they may allow a user to zoom in or zoom out on a model or to rotate 3-dimensional models in space for viewing from other angles. It‘s worth noting that you can combine transformable and manipulable models to create the highest level of interaction in information visualization.

~~~~~~

48. Compare and contrast the hierarchy and network type of visualizations.

Answer:

Hierarchy Visualizations:

Data visualizations that belong in the hierarchical category are those that order groups within larger groups. Hierarchical visualizations are best suited if you’re looking to display clusters of information, especially if they flow from a single origin point.

The downside to these graphs is that they tend to be more complex and difficult to read, which is why the tree diagram is used most often. It is the simplest to follow due to its linear path.

Examples of hierarchical data visualizations include:

- Tree diagrams
- Ring charts
- Sunburst diagrams

Network Visualizations:

Datasets connect deeply with other datasets. Network data visualizations show how they relate to one another within a network. In other words, demonstrating relationships between datasets without wordy explanations.

Examples of network data visualizations include:
~~~~~~

- Matrix charts
- Node-link diagrams
- Word clouds
- Alluvial diagrams

49. Write about Interactive data visualization.

Answer:

Interactive visualization or interactivevisualisation isa branch of graphic visualization in computerscience that involves studying how humans interact with computers to create graphic illustrations of information and how this process can be made more efficient.

For a visualization to be considered interactive it must satisfy two criteria:

- **Human input**: control of some aspect of the visual representation of information, or of the information being represented, must be available to a human, and
- **Response time**: changes made by the human must be incorporated into the visualization in a timely manner. In general, interactive visualization is considered a soft real-time task.
- One particular type of interactive visualization is virtual reality (VR), where the visual representation of information is presented using an immersive display device such as a stereo projector (see stereoscopy). VR is also characterized by the use of a spatial metaphor, where some aspect of the information is represented in three dimensions so that humans can explore the information as if it were present (where instead it was remote), sized appropriately (where instead it was on a much smaller or larger scale than humans can sense directly), or had shape (where instead it might be completely abstract).
- Another type of interactive visualization is collaborative visualization, in which multiple people interact with the same computer visualization to communicate their ideas to each other or to explore information cooperatively. Frequently, collaborative visualization is used when people are physically separated. Using several networked computers, the same visualization can be presented to each person simultaneously. The people then make annotations to the visualization as well as communicate via audio(i.e., telephone), video (i.e., a video-conference), or text (i.e., IRC) messages.
- The Programmer's Hierarchical Interactive Graphics System (PHIGS) was one of the first programmatic efforts at interactive visualization and provided an enumeration of the types of input humans provide. People can:
  - *Pick* some part of an existing visual representation;
  - *Locate* a point of interest (which may not have an existing representation);
  - *Stroke* a path;
  - *Choose* an option from a list of options;
  - *Valuate* by inputting a number; and

- *Write* by inputting text.

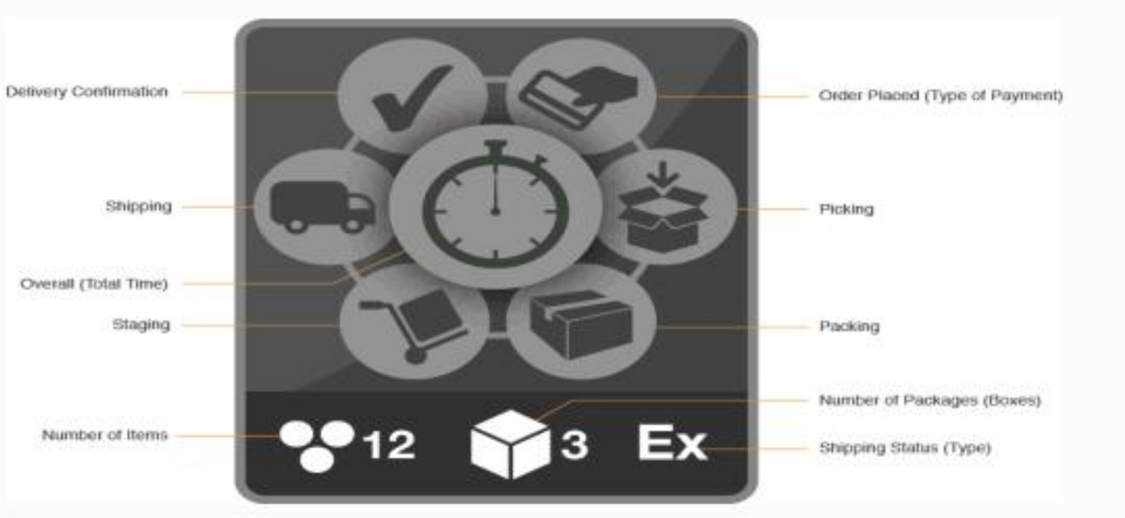
The following are five key properties of Interactive Visualization:

1. The Novice User. Even novices must be able to examine data and find patterns, distributions, correlations, and/or anomalies. They must be able to build and use tools thatenable faster decisions based on real-time information. As the National Research Council of the National Academies of Sciences states, even –naïve users‖ should be able to –carry out massive data analysis without a full understanding of systems and statistical uses.‖ *Frontiers in Massive Data Analysis*(National Academy of Sciences 2013). And while data scientists play an indispensable role in today‘s corporation, business line executives should not have to rely on them to run analytics and make the inferences that are the basis for decisions. As McKinsey puts it, –sophisticated analytics solutions . . . must be embedded in frontline tools so simple and engaging that managers and frontline employees will be eager to use them daily.

2. Driving Processes. The solution must allow the user to establish KPIs that provide the rules that drive processes. These must be displayed visually—for example, by color—in real time based on defined thresholds. Likes its architecture, Interactive Visualization is a means to an end – to stimulate informed action. Thus, for example, when a fire engulfs the third floor of a company‘s office space, triggers are set off that alert proper actors such as the municipal fire department. Interactive Visualization displays the department‘s efforts through phases of the process—discovery, initial actions, mitigation, stabilization, and recovery. As each phase is completed, analytics-based data is represented in real time in green, thereby ending with the most intuitive color cycle we know: red (danger; take action); yellow (pause; remediation is underway); and green (problem solved; situation clear). With icons changing color based on pre-defined thresholds (rules) run against multiple data streams by an analytics engine, data can be understood equally by management and analysts with no need for technical translation. It is important that the status of a process (fire), a person (fireman), or a physical asset (fire truck) must be depicted visually either independent of one another or as they correlate. Each representation must be both simple and highly granular, allowing a user to understand huge amounts of data with little or no training.

3. Data Must Tell A Story. An intuitive, visual workplace that it easy to master is based on easily digestible interactive patterns. Data must tell a story that instantly relates the performance of a business and its assets. Almost every Interactive Visualization narrative takes place across multiple layers. Users must thus be able to select data elements and filters, and then highlight and modify options to change data perspectives – from high-tech overviews down to the most granular detail. For example, one might overlay data on top of maps, diagrams such as building schematics, or even atop steps underway in a process. A story will emerge that places data in context and in real time as needed. Visual Cue is one company that has taken telling business process stories to a new level. An example of one of their active business tiles is shown here. According to CEO Kerry Gilger, —users want a complicated story made simple so that they act on it. The story needs to unfold simply, in real time, and in intuitive diagrams that

can prompt immediate.



4. Data Correlation. The user should immediately know not only of hot spots that require attention, but also effortlessly find trends based on the dynamic relationship between multiple data streams and the data derived from them by means of predictive analytics.

5. Prescriptions: —What should happen next?‖ According to Gartner, analytics evolves through four phases. The third and most discussed phase in today‘s market is predictive analytics – the application of rules and algorithms against data streams in order to yield actionable intelligence. This answers the question: —What is going to happen?‖ World-class Interactive Visualization and underlying analytics capabilities surpass that standard by offering prescriptive analytics(—What should happen next?‖) to drive real-time asset behaviour modification. This is the pinnacle of Gartner‘s evolution of analytics. It is closely linked to the need to drive processes discussed above. Recommendations may range from (i) re-routing a cargo ship to a different port based on the ratio of fuel loss to cargo weight, to (ii) suggesting additional training for underperforming members of a call center.

~~~~~~

50. Explain different types of correlations.

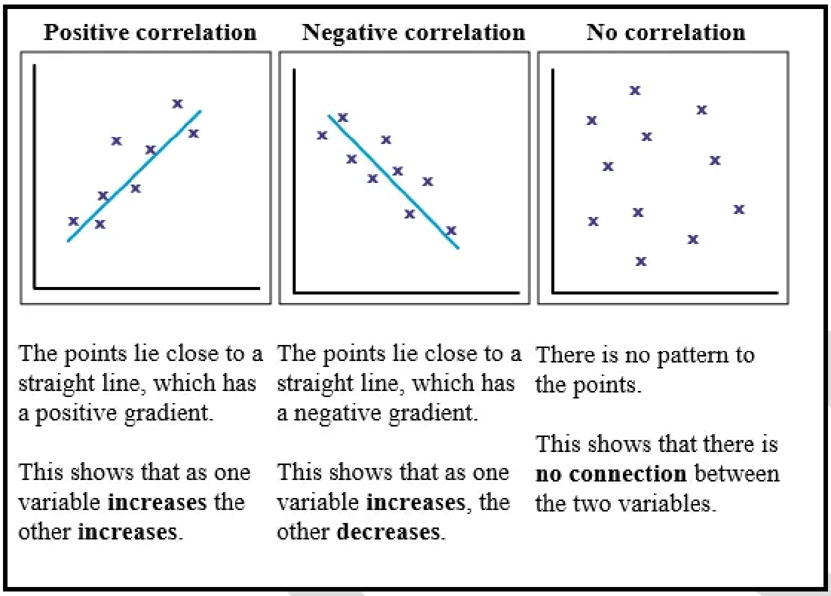
Answer:

Correlation:

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a co relational study: a positive correlation, a negative correlation, and no correlation.

- A **positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be height and weight. Taller people tend to be heavier.
- A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
~~~~~~

- A **zero correlation** exists when there is no relationship between two variables. For example there is no relationship between the amount of tea drunk and level of intelligence.



The points lie close to a straight line, which has a positive gradient.

This shows that as one variable **increases** the other **increases**.

The points lie close to a straight line, which has a negative gradient.

This shows that as one variable **increases**, the other **decreases**.

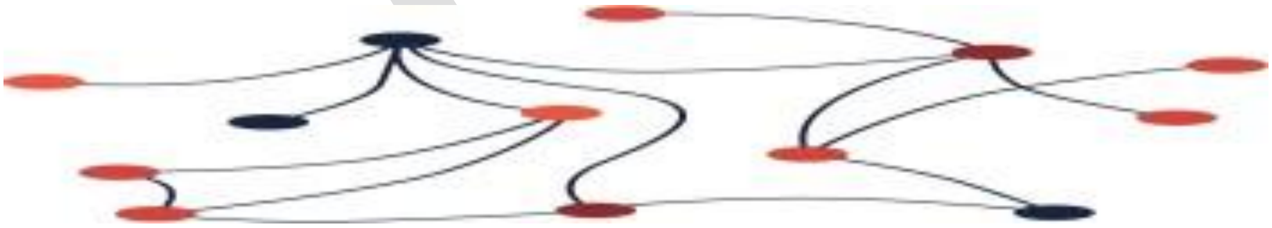
There is no pattern to the points.

This shows that there is **no connection** between the two variables.

51. Write short notes on Network data visualization.

Network Visualization:

- Network Visualization (also called Network Graph) is often used to visualize complex relationships between huge amounts of elements. This type of visualization illuminates relationships between entities. Entities are displayed as round nodes and lines show the relationships between them.
- Network visualization displays undirected and directed graph structures.
- This type of visualization illuminates relationships between entities. Entities are displayed as round nodes and lines show the relationships between them. The vivid display of network nodes can highlight non-trivial data discrepancies that may be otherwise be overlooked.



52. Write a short note on Hierarchical data visualization.

Answer:

Hierarchical Data Visualization:

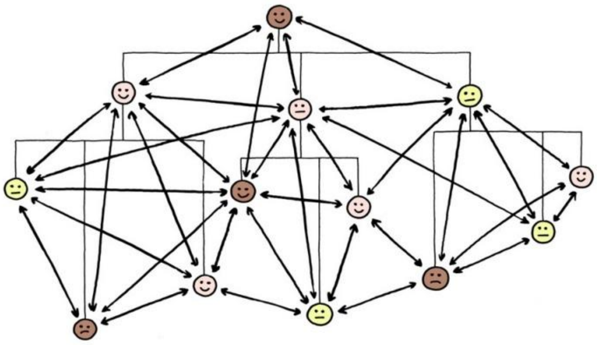
Hierarchical data, a very special type of network data, is represented by hierarchical data visualization.

Typically, the principle of connection represents network data. However, hierarchical data relies on the principle of containment.

A classic example of hierarchical data visualization is the file and folder system found on your computer. You have a folder that contains more folders.

Other common types of hierarchical data visualization forms are tree diagrams, cone tree diagrams, botanical tree diagrams and tree map diagrams.

How to Show Hierarchical Data with Information Visualization

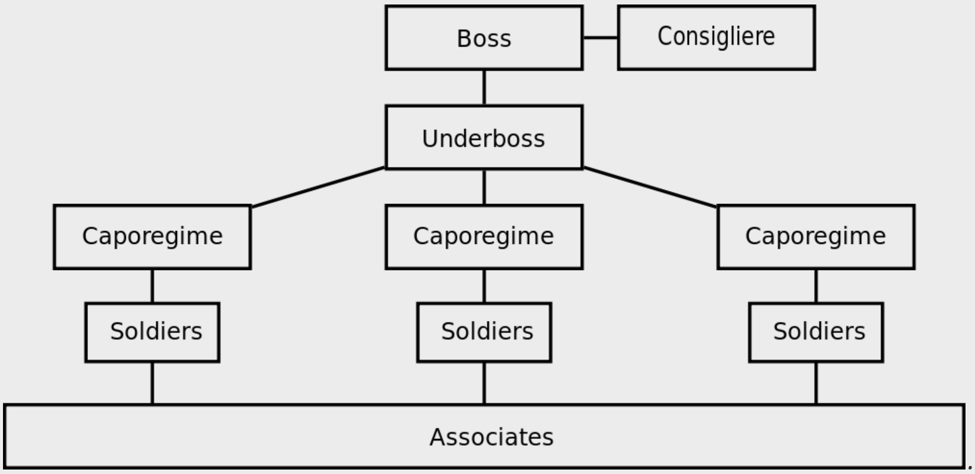


Hierarchical data is essentially a specialized form of network data – in that while entities within the dataset do not have dependent relationships; they are all related to each other by the principle of containment. They, unlike standard data networks, do not use the principle of connection.

A hierarchy begins with a root entity. This might be the CEO of a company, the name of a book, the title of a folder, etc. and then the root entity has at least one “child node” and every further child node has zero or more children.

An entity which comes below another is a child node to the entity above. Similarly, an entity which comes above another is a parent node to the node below.

Hierarchical data is shown in tree graphs; so called because of their similarity to a tree’s structure (though a tree which has been turned upside down so that the root is at the top and the branches form below it).



Above we see a simple tree diagram for the structure of a mafia family. The root entity is the boss of the family and the underboss is the first child entity. This is a very basic hierarchical relationship and it is possible to map much more complex hierarchies using information visualization techniques.

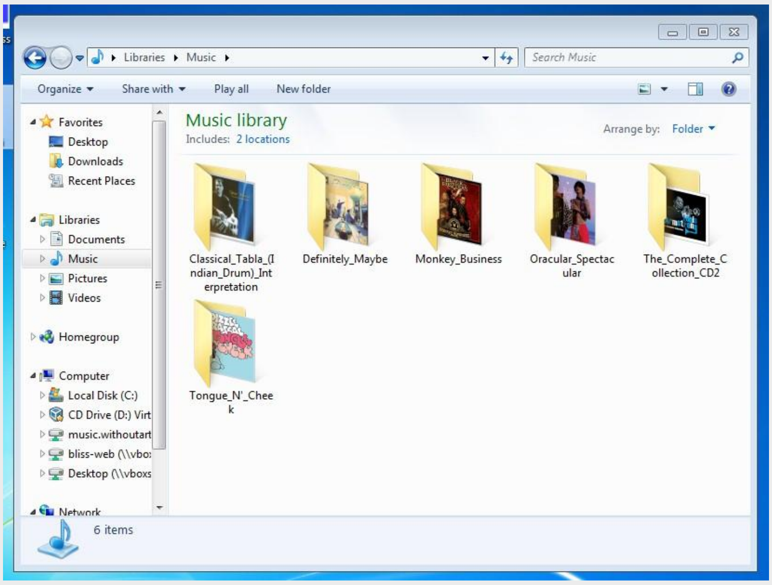
This diagram also includes a “sibling node” in the form of a consigliere who is not the boss of the organization but whose authority is equal to that of the boss.

The term “tree diagram” was coined by Noam Chomsky in his 1965 work; Aspects of the Theory of Syntax.

1. The File and Folder System

One of the most common hierarchies, which many of us deal with daily, is the computer file system. There is a root directory which then has a selection of child folders, which in turn have child folders, and in some or all of these folders there are files to be found.

The file tree is normally rendered in a visual format by the operating system. To provide a certain level of familiarity it uses images of the classic paper folder to connect the viewer with the property of the abstract file system used on the disc drive. This tree is interactive in nature and clicking on any given folder enables the user to determine what is inside that folder. However, it is also possible to use the command shell of an operating system to present this information textually too.



The Windows file system above is essentially a hierarchical tree and one with which many of us are already incredibly familiar with.

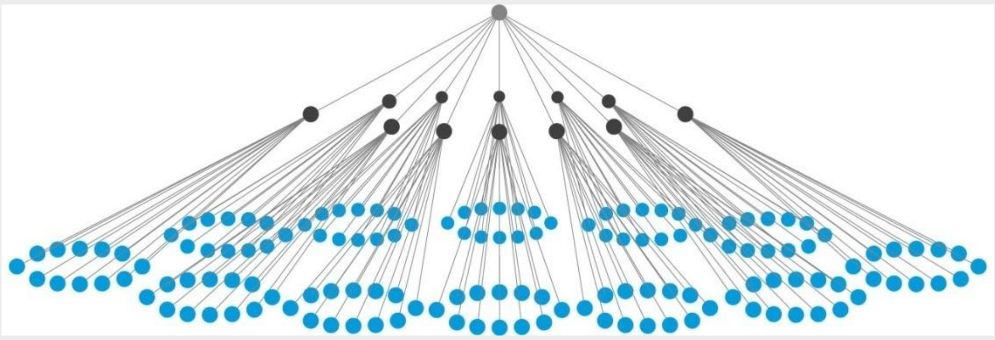
<u>2. Cone Tree Diagram:</u>

A cone tree is a 3-D hierarchy model which was developed at Xerox PARC in the 1990s. It was designed to enable the representation of hierarchies with large multiples of nodes. The 3D means that the physical limitations of displaying complexity on a flat screen can, to some extent, be overcome.

It works by beginning with a root node and then arranging all the child nodes of that root equidistant from the parent. This forms a cone with some transparency. The process is then performed again and again for each set of child nodes and the diameter of the cone is reduced at each level of the hierarchy.

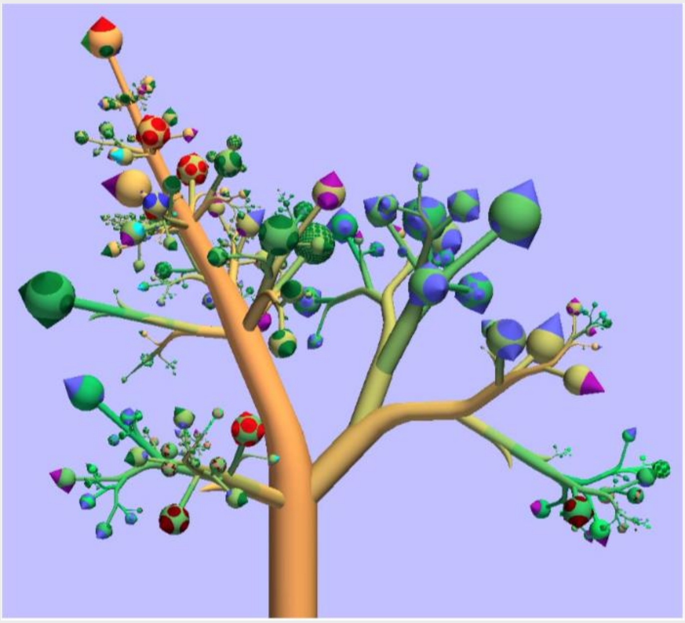
In general cone tree diagrams are generated using software which enables interactivity and a useful property of these diagrams is the ability to rotate them so that a particular child is occluded (hidden) by another child.

You can also get an idea of the numbers of child entities on any parent simply by observing the density of shading on any given cone. (The edges used to define the cone are properties of the children).



3. The Botanical Tree Diagram

The botanical tree diagram was invented by researchers at Eindhoven University of Technology in the Netherlands. They had noticed that the limitation of tree diagrams was that they can quickly become too complex to be functional and then they noticed that real trees had leaves. They also saw that it didn't matter how many leaves and branches there were on a tree – these were always distinct entities visually. So they extended the concept of tree diagrams by adding leaves and branches.

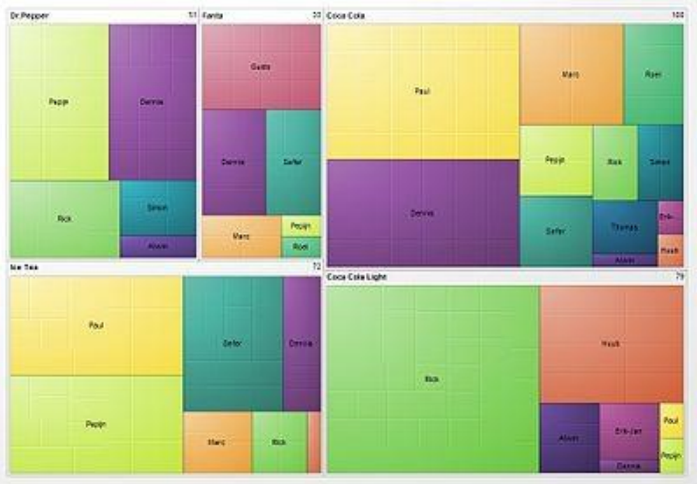


4. The Treemap Diagram

The treemap was invented Ben Shneiderman of the University of Maryland in 1990. It represents hierarchies by using all the available space and in the form of nested rectangles.

The rectangles can be defined in proportion to the "space" that they take up within the data set. These information visualizations can be very useful for comparing nodes and see patterns within them.

The math involved to create a treemap is quite complex but the good news is that you don't have to do that math; there are plenty of software packages available that can do this for you.



Above is a treemap of market share for different soft drinks. Comparing Coke, Coke Light, Ice Tea, Fanta and Dr. Pepper.

When Al Shalloway, the founder and CEO of Net Objectives, said; "Visualizations act as a campfire around which we gather to tell stories." we wonder if he had envisioned such complex stories?

Representing hierarchical information is quite straightforward; in the majority of cases the information designer will either use a tree diagram (or a variant of a tree diagram) or a treemap to display the data in an efficient format for the user.

~~~~~~

53. Implement how the functionality of data visualization to interactive data analysis?

Data Visualization:

- Visualize distribution of a variable and see patterns with line charts, histograms and trend lines.
- See relationships between variables with scatter plots (E.g., relationship between calls and number of sales) and heat maps.
- Classify variables (e.g., percent of gross revenue by salesperson) with bar charts.

How Does Interactive Data Analysis Work?

Interactive data analysis works mostly in a loop fashion.

- Defining the problem. Every projects starts with a problem statement. What problem are you trying to solve? *What is your ultimate goal? How is increased understanding derived from data analysis going to bring you closer to your goal?*
- Generating questions: Typically, the problem needs to be translated (implicitly or, better, explicitly) into a number of data analysis questions.
- Gathering, transforming and familiarizing with the data. Some projects have data available, whereas some others require some degree of data search or generation. In any case, all projects require the analyst to familiarize with the content and its meaning and perform multiple transformations, both to familiarize with the data (e.g., often slicing, dicing and aggregating the data) and to prepare it for the analysis one is planning to perform.
- Creating models out of data. Using statistical modeling and machine learning methods can be useful when the question asked can be answered more easily by building a model. Models can be extremely powerful tools for exploration and hypothesis generation.
- Visualizing data and models: This is where the results obtained from data transformation and querying (or from some model) are turned into something our eyes can understand.
~~~~~~

- Interpreting the results. Once the results have been generated and represented in some visual format, they need to be interpreted by someone. This is a crucial step and an often overlooked one. Behind the screen there is a human who needs to understand what all those colored dots and numbers mean.
- Generating inferences and more questions. All of these steps ultimately lead to creating some new knowledge and, most of the time, generating additional questions or hypotheses.

~~~~~~

54. Explain about data correlation and why is it important to identify correlations?

Correlation:

Correlation is a statistical measure. Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

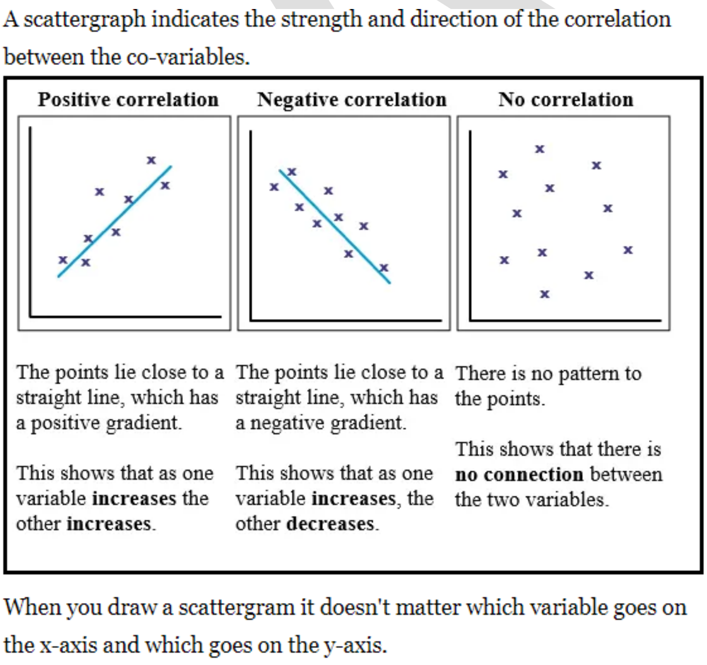
Two features (variables) can be positively correlated with each other. It means that when the value of one variable increases then the value of the other variable(s) also increases.

Types of correlation:

- ✓ A **positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be height and weight. Taller people tend to be heavier.
- ✓ A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
- ✓ A **zero correlation** exists when there is no relationship between two variables. For example there is no relationship between the amount of tea drunk and level of intelligence.

Scattergrams

A correlation can be expressed visually. This is done by drawing a scattergram (also known as a scatterplot, scatter graph, scatter chart, or scatter diagram).



A scattergraph indicates the strength and direction of the correlation between the co-variables.

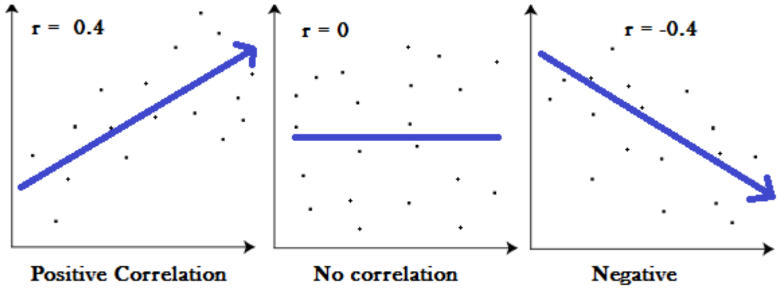
Positive correlation	Negative correlation	No correlation
The points lie close to a straight line, which has a positive gradient. This shows that as one variable **increases** the other **increases**.	The points lie close to a straight line, which has a negative gradient. This shows that as one variable **increases**, the other **decreases**.	There is no pattern to the points. This shows that there is **no connection** between the two variables.

When you draw a scattergram it doesn't matter which variable goes on the x-axis and which goes on the y-axis.

The Correlation Coefficient

A correlation coefficient is a way to put a value to the relationship. Correlation coefficients have a value of between -1 and 1. A “0” means there is **no relationship** between the variables at all, while -1 or 1 means that there is a **perfect**
~~~~~~

negative or positive correlation (negative or positive correlation here refers to the type of graph the relationship will produce).



Importance of Correlation:

- Correlation helps us in determining the degree of relationship between variables. It enables us to make our decision for the future course of actions.
- Correlation analysis helps us in understanding the nature and degree of relationship which can be used for future planning and forecasting.
- Forecasting without any prior correlation analysis may prove to be defective, less reliable and more uncertain. If it is based upon the result of correlation analysis, it will be more reliable.

Identifying correlations can help marketers:

- Make SEO/PPC campaign decisions
- Test those campaign decisions
- Determine if those relationships are yielding profitable results.

---

55. Give a brief review of general aspects of interaction on data visualization

Visualization has become a valuable means for data exploration and analysis. Interactive visualization combines expressive graphical representations and effective user interaction.

Interactive data visualization refers to the use of modern data analysis software that enables users to directly manipulate and explore graphical representations of data.

Some common data visualization interactions that will help users explore their data visualizations include:

- ✓ **Brushing**: Brushing is an interaction in which the mouse controls a paintbrush that directly changes the color of a plot, either by drawing an outline around points or by using the brush itself as a pointer.
- ✓ **Painting**: Painting refers to the use of persistent brushing, followed by subsequent operations such as touring to compare the groups.
- ✓ **Identification**: Identification, also known as label brushing or mouseover, refers to the automatic appearance of an identifying label when the cursor hovers over a particular plot element.
- ✓ **Scaling**: Scaling can be used to change a plot's aspect ratio, revealing different data features. Scaling is also commonly used to zoom in on dense regions of a scatter plot.
- ✓ **Linking**: Linking connects selected elements on different plots. One-to-one linking entails the projection of data on two different plots, in which a point in one plot corresponds to exactly one point in the other. Brushing an area in one plot will brush all cases in the corresponding category on another plot.

Some major benefits of interactive data visualizations include:

- ✓ **Identify Trends Faster** - Direct manipulation of analyzed data via imagery makes it easy to understand and act on valuable information.
- ✓ **Identify Relationships More Effectively** - Enables users to identify otherwise overlooked cause-and-effect relationships throughout definable timeframes. This is especially useful in identifying how daily operations affect an organization's goals.

- ✓ **Useful Data Storytelling** - A visual data story in which users can zoom in and out, highlight relevant information, filter, and change the parameters promotes better understanding of the data by presenting multiple viewpoints of the data.
- ✓ **Simplify Complex Data** - Incorporating filtering and zooming controls can help untangle and make these messes of data more manageable, and can help users gain better insights.

A static data visualization is one that does not incorporate any interaction capabilities and does not change with time, such as an infographic focused on a specific data story from a single viewpoint.

Explain how graphs are used in network visualization