Unit – 1 Introduction to data visualization

Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

Advantages and disadvantages of data visualization

Our eyes are drawn to colors and patterns. We can quickly identify red from blue, and squares from circles. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It's storytelling with a purpose. If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be.

Some other advantages of data visualization include:

Easily sharing information.

Interactively explore opportunities.

Visualize patterns and relationships.

Disadvantages

While there are many advantages, some of the disadvantages may seem less obvious. For example, when viewing a visualization with many different datapoints, it's easy to make an inaccurate assumption. Or sometimes the visualization is just designed wrong so that it's biased or confusing.

Some other disadvantages include:

Biased or inaccurate information.

Correlation doesn't always mean causation.

Core messages can get lost in translation.

Why data visualization is important

The importance of data visualization is simple: it helps people see, interact with, and better understand data. Whether simple or complex, the right visualization can bring everyone on the same page, regardless of their level of expertise.

It’s hard to think of a professional industry that doesn’t benefit from making data more understandable.

Data Preparation

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data. Data preparation can take up to 80% of the time spent on an ML project. Using specialized data preparation tools is important to optimize this process.

Steps in the data preparation process

Data preparation is done in a series of steps. There's some variation in the data preparation steps listed by different data professionals and software vendors, but the process typically involves the following tasks:

Data collection. Relevant data is gathered from operational systems, data warehouses, data lakes and other data sources. During this step, data scientists, members of the BI team, other data professionals and end users who collect data should confirm that it's a good fit for the objectives of the planned analytics applications.

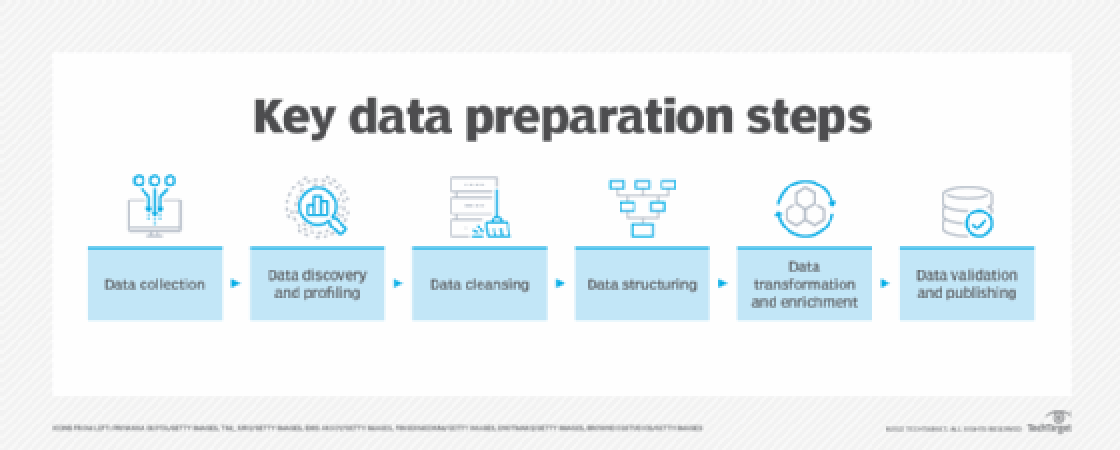
Data discovery and profiling. The next step is to explore the collected data to better understand what it contains and what needs to be done to prepare it for the intended uses. To help with that, data profiling identifies patterns, relationships and other attributes in the data, as well as inconsistencies, anomalies, missing values and other issues so they can be addressed.

Data cleansing. Next, the identified data errors and issues are corrected to create complete and accurate data sets. For example, as part of cleansing data sets, faulty data is removed or fixed, missing values are filled in and inconsistent entries are harmonized.

Data structuring. At this point, the data needs to be modeled and organized to meet the analytics requirements. For example, data stored in comma-separated values (CSV) files or other file formats has to be converted into tables to make it accessible to BI and analytics tools.

Data transformation and enrichment. In addition to being structured, the data typically must be transformed into a unified and usable format. For example, data transformation may involve creating new fields or columns that aggregate values from existing ones. Data enrichment further enhances and optimizes data sets as needed, through measures such as augmenting and adding data.

Data validation and publishing. In this last step, automated routines are run against the data to validate its consistency, completeness and accuracy. The prepared data is then stored in a data warehouse, a data lake or another repository and either used directly by whoever prepared it or made available for other users to access.


Key data preparation steps
Data collection
Data discovery and profiling
Data cleansing
Data structuring
Data transformation and enrichment
Data validation and publishing
TechTarget

Data mining

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.

Why is data mining important?

Data mining is a crucial component of successful analytics initiatives in organizations. The information it generates can be used in business intelligence (BI) and advanced analytics applications that involve analysis of historical data, as well as real-time analytics applications that examine streaming data as it's created or collected.

Effective data mining aids in various aspects of planning business strategies and managing operations. That includes customer-facing functions such as marketing, advertising, sales and customer support, plus manufacturing, supply chain management, finance and HR. Data mining supports fraud detection, risk management, cybersecurity planning and many other critical business use cases. It also plays an important role in healthcare, government, scientific research, mathematics, sports and more.

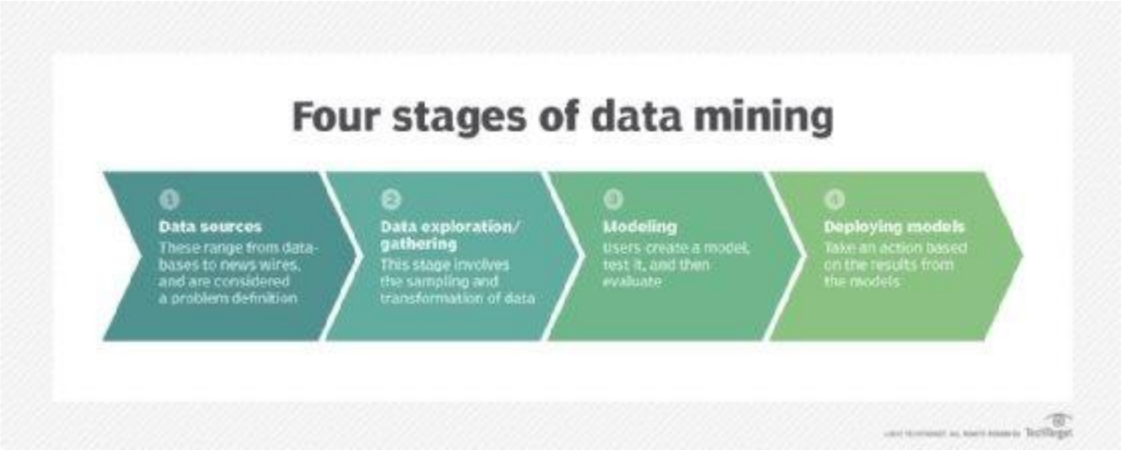
The data mining process can be broken down into these four primary stages:

Data gathering. Relevant data for an analytics application is identified and assembled. The data may be located in different source systems, a data warehouse or a data lake, an increasingly common repository in big data environments that contain a mix of structured and unstructured data. External data sources may also be used. Wherever the data comes from, a data scientist often moves it to a data lake for the remaining steps in the process.

Data preparation. This stage includes a set of steps to get the data ready to be mined. It starts with data exploration, profiling and pre-processing, followed by data cleansing work to fix errors and other data quality issues. Data transformation is also done to make data sets consistent, unless a data scientist is looking to analyze unfiltered raw data for a particular application.

Mining the data. Once the data is prepared, a data scientist chooses the appropriate data mining technique and then implements one or more algorithms to do the mining. In machine learning applications, the algorithms typically must be trained on sample data sets to look for the information being sought before they're run against the full set of data.

Data analysis and interpretation. The data mining results are used to create analytical models that can help drive decision-making and other business actions. The data scientist or another member of a data science team also must communicate the findings to business executives and users, often through data visualization and the use of data storytelling techniques.


Four stages of data mining
1
Data sources
These range from data-bases to news wires, and are considered a problem definition
2
Data exploration/ gathering
This stage involves the sampling and transformation of data
3
Modeling
Users create a model, test it, and then evaluate
4
Deploying models
Take an action based on the results from the models
TechTarget

Types of data mining techniques

Various techniques can be used to mine data for different data science applications. Pattern recognition is a common data mining use case that's enabled by multiple techniques, as is anomaly detection, which aims to identify outlier values in data sets. Popular data mining techniques include the following types:

Association rule mining. In data mining, association rules are if-then statements that identify relationships between data elements. Support and confidence criteria are used to assess the relationships -- support measures how frequently the related elements appear in a data set, while confidence reflects the number of times an if-then statement is accurate.

Classification. This approach assigns the elements in data sets to different categories defined as part of the data mining process. Decision trees, Naive Bayes classifiers, k-nearest neighbor and logistic regression are some examples of classification methods.

Clustering. In this case, data elements that share particular characteristics are grouped together into clusters as part of data mining applications. Examples include k-means clustering, hierarchical clustering and Gaussian mixture models.

Regression. This is another way to find relationships in data sets, by calculating predicted data values based on a set of variables. Linear regression and multivariate regression are examples. Decision trees and some other classification methods can be used to do regressions, too.

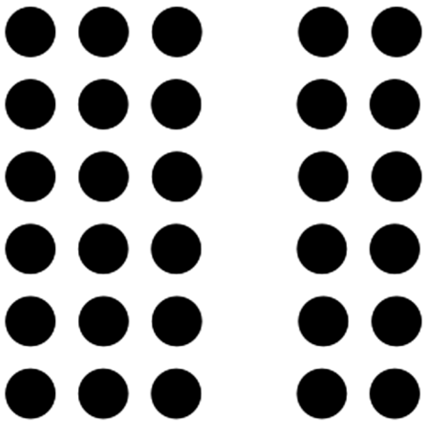
Sequence and path analysis. Data can also be mined to look for patterns in which a particular set of events or values leads to later ones.

Neural networks. A neural network is a set of algorithms that simulates the activity of the human brain. Neural networks are particularly useful in complex pattern recognition applications involving deep learning, a more advanced offshoot of machine learning.

Principles Perception

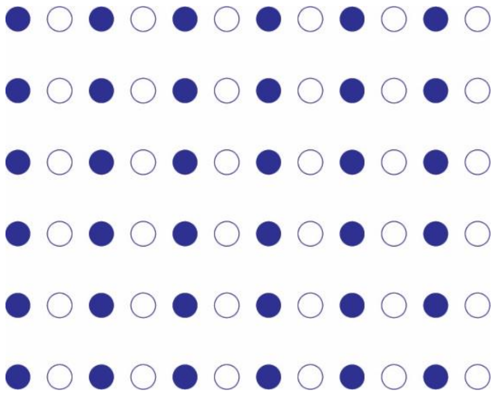
Gestalt principles are the different ways individuals group stimuli together in order to make a whole that makes sense to them. These principles are divided up into five categories: proximity, similarity, continuity, connectedness, and closure.

Proximity



The Gestalt principle of proximity is when an individual perceives several objects that are close together as belonging together. An example of this is in the picture above. In the picture the dots are all the same color, size, and shape. The only reason that we perceive two different blocks of dots is because of their position, and how close they are to each other. If these dots were to be miles and miles apart, then we would not perceive them as being a group.

Similarity



If proximity is due to position, then the Gestalt principle of similarity is how we piece information together by how similar objects are. For example, if there were five dogs of all different breeds and five cats of different breeds, then we would group them as cats and dogs. Here, positions do not matter, because we are looking into how similar the objects are to each other.

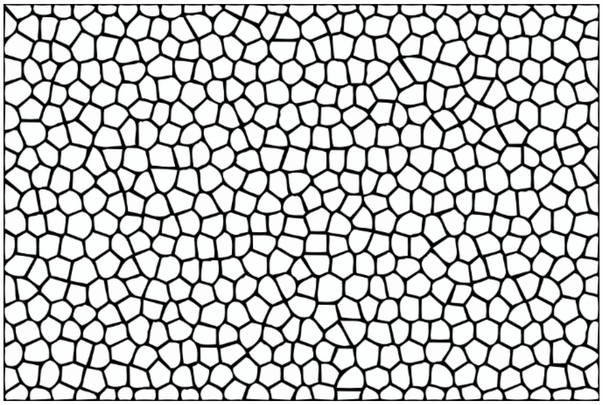
Another example is the picture above. When looking at these dots one would say that there are two groups. There are white dots and there are blue dots. We perceive these two groups as such, because they have the same shape. The only difference we see is in similarity, or in this case, the color. If all of the dots were blue, then we would say that there was one group of blue dots.

Continuity

The third Gestalt principle is continuity. Continuity is that our brains tend to see objects as continuous or smooth rather than disjointed or discontinuous. A great example of this phenomenon is a movie. Movies are just millions of pictures put together and flipped through at a fast rate. Your brain brings all of these pictures, these disjointed pictures, together into one cohesive, smooth unit.

Another great example of continuity is music. Music is individual notes that are strung together. Our brains bring those notes together into one smooth unit through continuity.

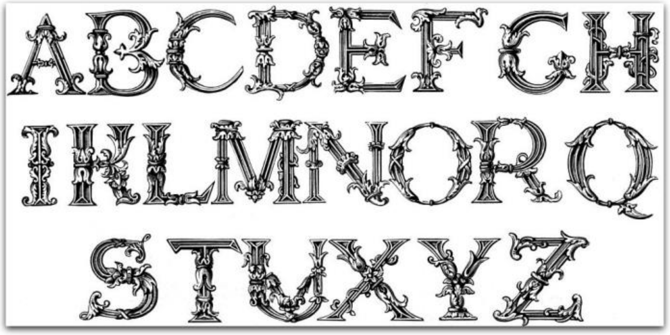
Connectedness



Connectedness is the fourth principle of the Gestalt principles. Connectedness is when we see connections in disjointed objects. One example of this is when you can see the image that will be made on a connect the dots picture before you connect the dots. For example, when people find constellations in the sky they see a picture made up of dots.

Another example of connectedness is a mosaic. A mosaic is made up of tiny broken pieces of glass or tile that are all put together in a collage to make a new, unified whole.

Closure



Closure is the final Gestalt principle. Closure is when individuals fill in the blanks. This means that the brain sees the big picture even when an element of that picture may be missing. An example of this is in the alphabet.

You may not have noticed that some letters were missing (see image above), because your brain knows what that sequence of letters is supposed to look like, and you perceived that the missing letters were there.

Another example of the Gestalt principle of closure is seen in the picture to the left. While these roughly drawn shapes are not finished, we can perceive that these shapes are a circle and a rectangle. Our ability to see closure with almost finished objects fills in the missing information.

Color Choices in Data Visualizations

Data visualizations are a useful way to present data points and insights. Using visual aids, charts and diagrams allow viewers to interpret information, make connections and remember keynotes that visualized data reports represent. Along with the physical structure of data in a visualization, the color choices and contrasts enhance its effectiveness. Understanding the benefits of color choice can help you in developing informational and engaging data visualizations. In this article, we discuss the importance of color in data visualization and how to choose colors.

Consider these steps when choosing colors for data visualizations:

1. Analyze collected data insights

Before choosing the colors for your data visualization, analyze the collected data insights to determine the necessity of using a variety of colors. For data with two main values or categories, it may not be necessary to create a color palette to represent your data. If you have three or more values, choosing colors can be more beneficial.

2. Evaluate other visualizations

Review other data visualizations presented within the company to identify color palettes, themes and specific color associations. These other visualizations allow you to assess established color correlations with certain values or categories relevant to the new data visualization. Keep these color trends consistent by maintaining previously assigned colors when creating your visualization.

3. Limit color variations

Having too many colors can make the data visualization more complex and harder to understand, so it's important to stay within six or fewer different colors so that the color values remain distinct from one another. Choosing colors that are right next to each other on the color wheel can make it hard to distinguish. When looking at a color wheel with 12 colors, using six of those colors allows you to choose colors that aren't directly next to each other.

4. Group together similar data

If you have more than six values or categories to present in your data visualization, consider grouping similar values together to use one color for them. When grouping these values together, you can also use different shades of the color to demonstrate its slight variation from the main color value. The dominant value can maintain the original color, while relevant data points can vary with lighter or darker shades of the color.

5. Consider the type of color palette

With your data visualization application software or tool, you can find different pre-made or recommended color palettes that you can use. These palettes often consider color theories and shade variations to allow you to choose a palette and adapt it to suit your specific visualization needs. Depending on the type of data and report you want to present, there are different palette categories you can implement, such as:

Diverging: A diverging color palette is useful when representing three main values, with one category as the median or average, while the other two diverge from it. This uses a neutral color in the middle of the palette for the median and two opposite colors with varying shades to demonstrate their divergent values.

Qualitative: A qualitative color palette can best represent values of distinct categories using different colors. These palettes work for line graphs, bar graphs and pie charts.

Sequential: A sequential color palette uses one color with multiple shade variations that can vary from an almost black shade to an almost white tone. This type of color palette may suit values that represent a single category of data.

6. Assign colors to values

When creating a data visualization using colors to represent values, create labels or a dedicated legend that defines what color represents what value or category. For visualizations like graphs or plots that have an x-axis and y-axis, it's important to remember to label what those axes represent and to note the value increments for each line. Maintain company terminology throughout the labeling process and consider using company-established colors or a color theme that incorporates company colors.

7. Highlight important data points

Consider using color to highlight data points in your visualization. You can use muted or unsaturated colors to minimize the visual impact of other values while using saturated colors to create a distinct separation that highlights a few areas. Depending on the purposes of your visualization, you may choose to avoid using color on any value other than the desired points you want to highlight.

8. Determine the background-color

The background of your visualization can affect the perception of colors when there's a lack of contrast between the background and data points. Aim for a neutral background such as black or white. If you're using a wide range of color shades or tones that may look similar to black or white, choose a color with a different undertone to add contrast.

9. Use online resources and tools

You can find online color palette resources through a browser search to identify a tool that can help you achieve your desired visualization results. These tools can allow you to input specific colors you want and it produces a color palette that accommodates those colors while remaining distinguishable. Other tools may suggest different visualization types that may best suit the data you want to represent.

Data Visualization Design

Data visualization can be a powerful tool. Only, however, when done correctly. As we’ve mentioned, a poorly designed visualization can end up doing more harm than good. So, it’s important to make sure that your data visualizations are effective.

When designing your dashboards and visualizations, there are certain principles or tips you should keep in mind to achieve this efficacy. These will enhance the value and effectiveness of your visualisations. They are:

Know your audience and your objective.

Choose the right types of visualizations.

Make them organized, consistent and intuitive.

Give context.

Less is more.

Use colors wisely.

1. Know Your Audience and Your Objective

Before choosing your datavis design, it’s essential that you know what you want to achieve from your visualizations, and who will be viewing them.

This is essential because if you design based on what you want to communicate to your end-end-viewer, it’s more likely that they will easily be able to grasp that information.

Your job is to make it easy for your viewer to make the business decisions they need to based on the data you are displaying for them. So, you will need to ask yourself what question they are trying to answer with this data and work from there.

You will also need to assess how familiar they are with the information you are presenting. And, you should keep in mind their abilities to read different kinds of graphs and charts. From there you can decide how simple or complex your visualization can be, and whether you need to add any explanatory notes.

2. Choose the Right Types of Visualizations

In order to choose the right kinds of visualizations for you and your stakeholders, you have to know a little about the different kinds and what purpose they serve. Let’s take a look at a few of the most popular options:

Bar graphs - Most people are familiar with bar charts which show bars plotted along axes and are used to compare different factors or categories. They are great for making comparisons.

Tables - Another old favorite, tables are made up of rows and columns and are good for showing a lot of information in a tidy, easy-to-read way.

Line Charts - These involve points that are plotted along axes and are good for tracking trends and changes over time.

Scatterplots - These show different variables plotted alongs axes with dots. The dots form patterns which allow the viewer to draw their conclusions. They provide a good way to show non-linear patterns.

Pie Charts - With these you can assign different variables or different quantities to portions of the circle (or pie) to then compare those variables. They are a simple, easy-to-understand chart.

Infographics - these are illustrations that offer an easy way to view a lot of information. When done well they are aesthetically pleasing and clear.

Word clouds - Essentially a visual representation of words, a word cloud will highlight and show words that come up in data with higher frequency. These are great for keyword data.

Maps - Maps are another familiar visualization and are great for showing data related to geographical regions or locations.

To get more information on the different types of visualizations, take a look at our data visualization types post.

3. Make Your Visualizations Organized, Consistent and Intuitive

The whole point of data visualization is that the viewer will understand the data better than if it were in its raw form. It makes sense then, that the visualizations need to be intuitive and well organized.

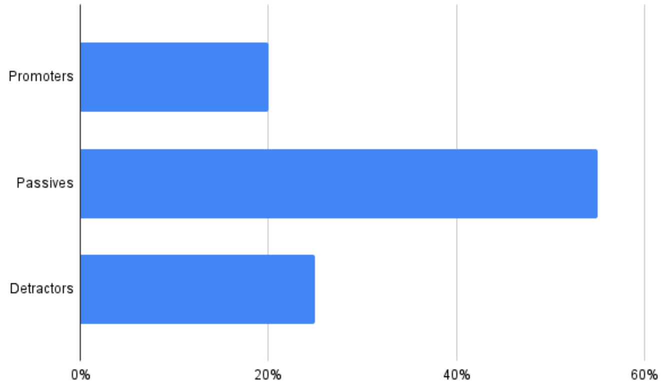
When you are designing you are shooting for clarity above all else.

You don’t want the viewer to have to work hard to understand what they are seeing. Make sure your data is set out in a logical format. This could be alphabetically, by value or another criteria, depending on your data.

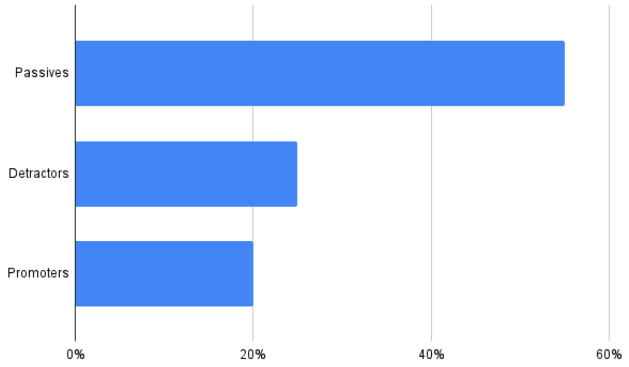
You should also take into consideration the hierarchy of data, placing different elements in certain places to attract more attention, and make sure that you have white space in your design.

These considerations should also be taken with regards to labels, fonts and colors (we’ll go into more depth on colors shortly). All of these elements are useful only when they are error-free, help the viewer understand the data, and are not distracting.

Your data should follow a natural order. To give you an example, instead of doing this:



You should do this:



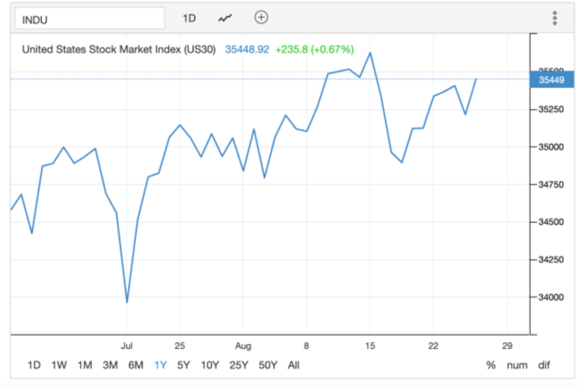
4. Give Context

To help your viewer quickly understand what they are seeing, it helps to provide previous data as context. Data rarely exists in a vacuum so without context your results might actually be misleading.

In order to provide this context, you can add a benchmark or a zero baseline. You could also add short explanatory notes (emphasis on short).

Either way, by comparing it against existing data or insights, the viewer can easily tell how what they are seeing relates to what they already know.

The scale you chose also matters. In the below example of the US stock market, it’s important for small variances to be visible to the viewer. So the scale must be adapted according to this requirement:

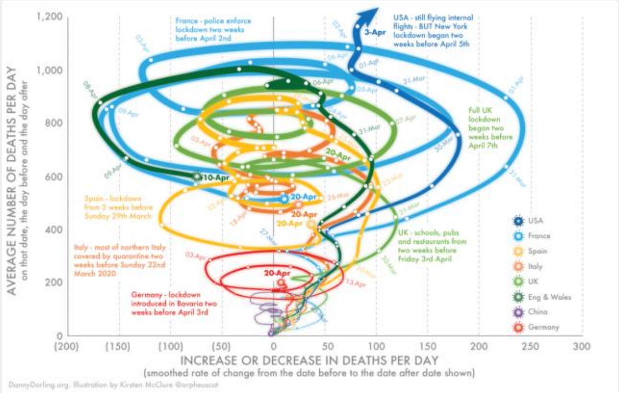

INDU
1D
United States Stock Market Index (US30) 35448.92 +235.8 (+0.67%)
35449
35250
35000
34750
34500
34250
34000
Jul
25
Aug
8
15
22
29
1D 1W 1M 3M 6M 1Y 5Y 10Y 25Y 50Y All
% num dif

5. Less Is More

Following on from point number 3 regarding clarity and consistency, what you leave in, or more importantly, leave out, matters. Take out anything that doesn’t add value or that detracts the viewers attention.

As you move along the design process, continually refer to your original objective. Question whether the elements you add to your visualization are getting you closer to answering that objective. If they are not, remove them.

Charts like the following are a good example of how more detail can lead to confusion rather than clarity. There is too much for the eye to focus on and too many threads to follow:



6. Use Color Wisely

Color is both powerful and influential in data visualization. It can capture your audience's attention faster and provide strong visual queues. It can also be confusing and distracting, depending on how you use it.

Some good principles to follow with regards to color include:

- Stay consistent with your use of colors. Do not interchange them in the same visualization.
- Don't use too many different colors as this can be distracting.
- Choose high contrast color schemes over lighter colors. This makes it easier for people to read your visualizations.
- Respect existing color associations. For example, don't try to use red for positive and green for negative.

When choosing colors it's best to be as inclusive as possible. Not everyone views color the same, for instance those with color blindness to do. When designing your visualizations, there are several tools that can help you plan for this. Adobe, for instance, provides color blindness filters that allow you to see how your work as a person with color blindness would.

Text Data Visualization

Text visualization is the technique of using graphs, charts, or word clouds to showcase written data in a visual manner. This provides quick insight into the most relevant keywords in a text, summarizes content, and reveals trends and patterns across documents.

Companies use text visualization to:

Summarize large amounts of text. Automatically highlight key terms in a series of texts, and categorize text by topic, sentiment, and more, saving hours of reading time. How long would it take you to read 500 online reviews? With a word cloud or data visualization dashboard, you can understand text data at a glance.

Make text data easy to understand. The human brain loves visual data. In fact, we are able to process images much faster than text. Text visualization is an effective way of simplifying complex data and communicating ideas and concepts to team managers.

Find insights in qualitative data. Customer feedback holds a trove of insights. Through text visualization, you can get an overview of the features, products, and topics that are most important to your customers. Learn what their pain points are and what you're doing right.

Discover hidden trends and patterns. Analyze and visualize insights over time to detect fluctuations, and quickly find the root cause.

Interactive Animated Visualization

There are four key steps in creating a world-class interactive data visualization.

Data integration. Collect raw data and turn it into clean, analytics-ready information by performing data replication, ingestion and transformation. Then store it in a data lake or data warehouse.

Goal definition. Define the business objective you're trying to achieve and the data insights you seek. For example, are you trying to optimize a production process or track the ROI of your marketing efforts.

Visualization design. Design begins with selecting KPIs and types of graphs, charts, and maps that best tell your story. Keeping your visualizations clean and simple will help users understand and work with the data.

Collaboration and sharing. Allow all approved users to explore the data freely to uncover their own insights. Your software should allow users to embed your visualizations in other applications and to engage with them on their mobile devices.

Temporal Data Visualizations

Temporal Visualizations (or timelines) are similar to one-dimensional linear visualizations. Because timelines are widely used and vital enough for medical records, project management, and historical presentations, they are considered a separate data type. Temporal data is characterized by items that have a start and finish time, and items may overlap each other. Timeline visualizations usually include all events before, after, or during some time period or moment.

Examples of Temporal Visualizations include: timelines, Gantt charts, stream graphs, arc diagrams/thread arcs, tree rings/concentric circle graphs, time series charts/graphs, and alluvial diagrams

Tools for Temporal Visualizations

d3.js

From the developers of Provotis, d3.js is a small, free JavaScript library for manipulating documents based on data. Can produce choropleth, motion chart, hib plot, and fisheye distortion visualizations.

Time Flow

Analyze temporal data with five different displays: Timeline, Calendar, Bar Chart, Table, and List views.

Simile Timeline

Timeline is a web widget for visualizing temporal data. You can make interactive, detailed, timelines.

Simile Time plot

Time plot is a DHTML-based AJAX widget for plotting time series and laying time-based events over them.

Provotis

A graphical approach to visualization, Provotis composes custom views of data with simple marks such as bars and dots and defines marks through dynamic properties that encode data. Protovis is mostly declarative and designed to be learned by example. It is no longer under active development.

Exploratory data analysis (EDA)

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

Exploratory data analysis

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

Exploratory data analysis tools

Specific statistical functions and techniques you can perform with EDA tools include:

- Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at.
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
- K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- Predictive models, such as linear regression, use statistics and data to predict outcomes.

Types of exploratory data analysis

There are four primary types of EDA:

- **Univariate non-graphical.** This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it's a single variable, it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
- **Univariate graphical.** Non-graphical methods don't provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:
    - Stem-and-leaf plots, which show all data values and the shape of the distribution.
    - Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
    - Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
- **Multivariate nongraphical:** Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
- **Multivariate graphical:** Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.