Data Visualization

- Data visualization is the art and practice of gathering, analyzing, and graphically representing empirical information.
- They are sometimes called information graphics, or even just charts and graphs.
- The goal of visualizing data is to tell the story in the data.
- Telling the story is predicated on understanding the data at a very deep level, and gathering insight from comparisons of data points in the numbers

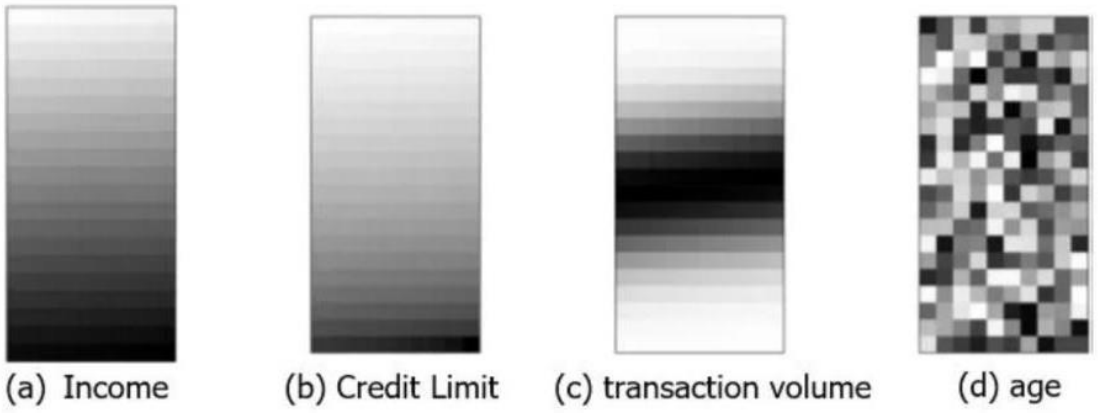
Why data visualization?

- Gain insight into an information space by mapping data onto graphical primitives Provide qualitative overview of large data sets
- Search for patterns, trends, structure, irregularities, and relationships among data.
- Help find interesting regions and suitable parameters for further quantitative analysis.
- Provide a visual proof of computer representations derived.

Categorization of visualization methods

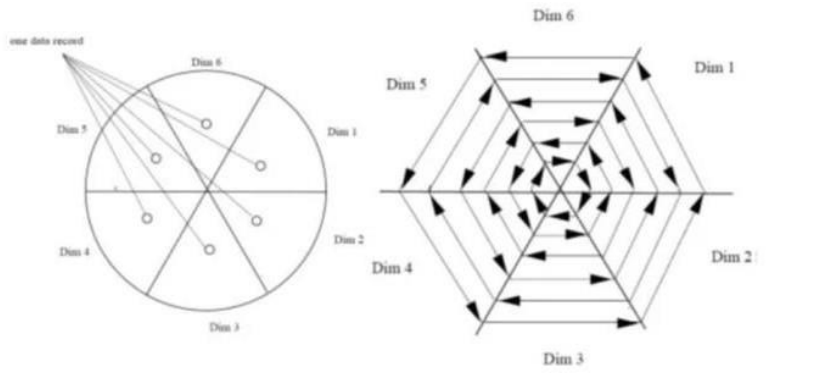
- Pixel-oriented visualization techniques
- Geometric projection visualization techniques
- Icon-based visualization techniques
- Hierarchical visualization techniques
- Visualizing complex data and relations

## Pixel-Oriented Visualization Techniques




*Pixel-Oriented Visualization Techniques*

- For a data set of m dimensions, create m windows on the screen, one for each dimension.
- The m dimension values of a record are mapped to m pixels at the correspondingpositions in the windows.
- The colors of the pixels reflect the corresponding values.




(a) Representing a data record in circle segment (b) Laying out pixels in circle segment

*Laying Out Pixels in Circle Segments*

- To save space and show the connections among multiple dimensions, space filling isoften done in a circle segment.

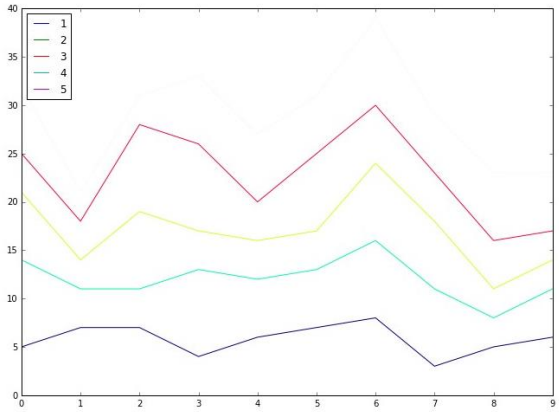
## Geometric Projection Visualization Techniques

Visualization of geometric transformations and projections of the data.Methods

- Direct visualization
- Scatterplot and scatterplot matrices
- Landscapes Projection pursuit technique: Help users find meaningful projections ofmultidimensional data
- Prosection views
- Hyperslice
- Parallel coordinates

Line Plot:

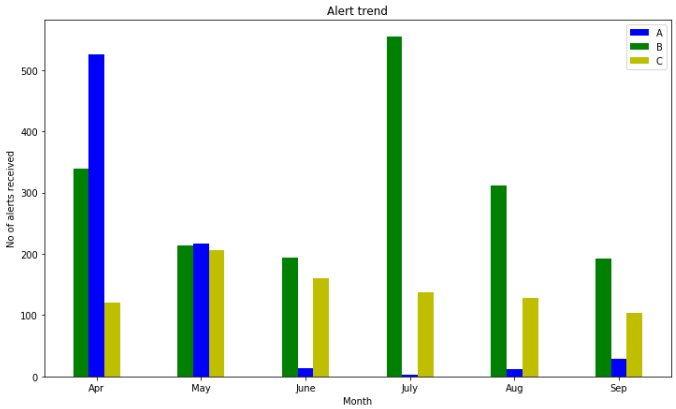
- This is the plot that you can see in the nook and corners of any sort of analysis between 2 variables.




- The line plots are nothing but the values on a series of data points will be connected with straight lines.
- The plot may seem very simple but it has more applications not only in machine learning but in many other areas.
- Used to analyze the performance of a model using the ROC- AUC curve.

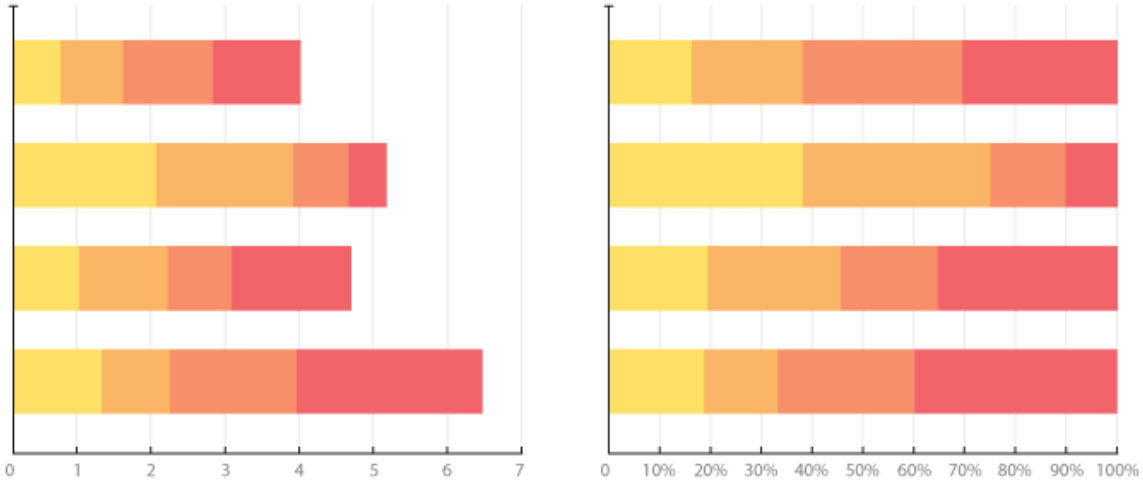
Bar Plot

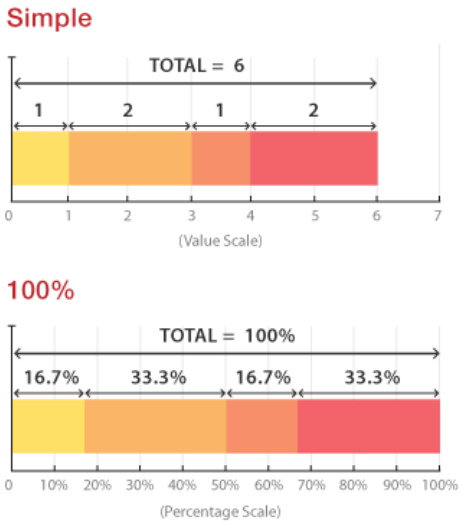
- This is one of the widely used plots, that we would have seen multiple times not just in data analysis, but we use this plot also wherever there is a trend analysis in many fields.
- We can visualize the data in a cool plot and can convey the details straight forward to others.



- This plot may be simple and clear but it's not much frequently used in Data science applications.

Stacked Bar Graph:



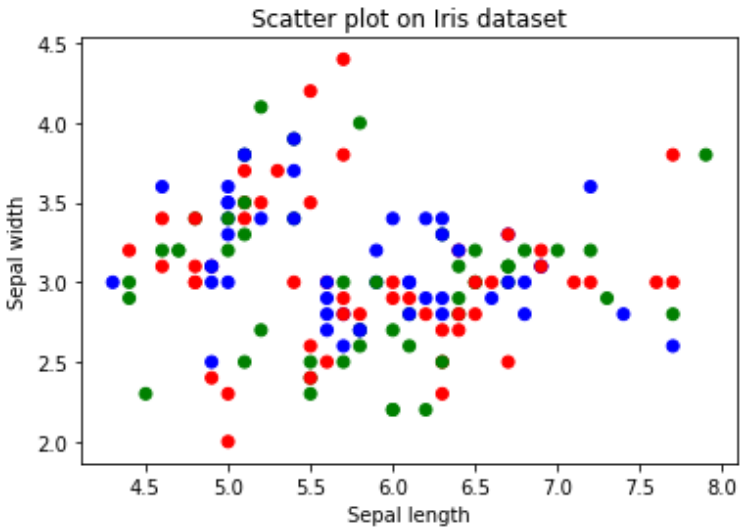




- Unlike a Multi-set Bar Graph which displays their bars side-by-side, Stacked Bar Graphs segment their bars. Stacked Bar Graphs are used to show how a larger category is divided into smaller categories and what the relationship of each part has on the total amount. There are two types of Stacked Bar Graphs:
- 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.

- 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.

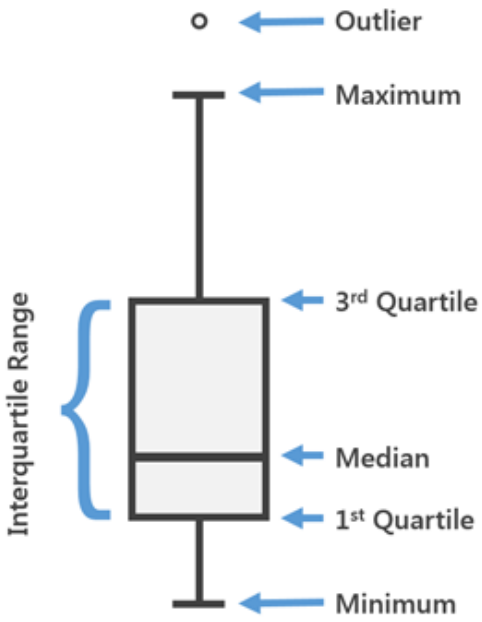
Scatter Plot




- It is one of the most commonly used plots used for visualizing simple data in Machine learning and Data Science.
- This plot describes us as a representation, where each point in the entire dataset is present with respect to any 2 to 3 features(Columns).
- Scatter plots are available in both 2-D as well as in 3-D. The 2-D scatter plot is the common one, where we will primarily try to find the patterns, clusters, and separability of the data.
- The colors are assigned to different data points based on how they were present in the dataset i.e, target column representation.
- We can color the data points as per their class label given in the dataset.

Box and Whisker Plot

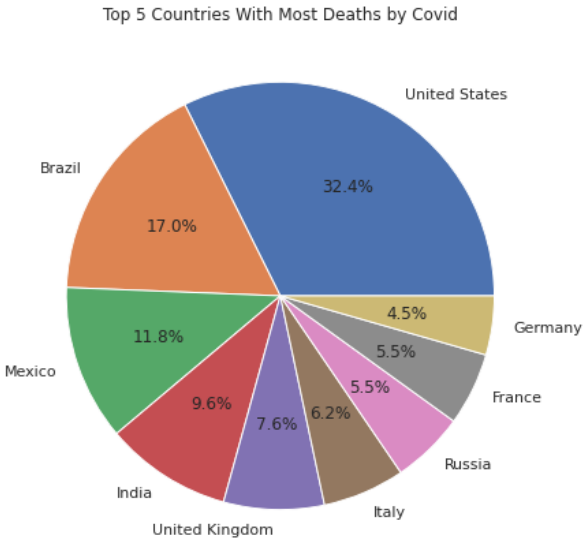
- This plot can be used to obtain more statistical details about the data.
- The straight lines at the maximum and minimum are also called whiskers.
- Points that lie outside the whiskers will be considered as an outlier.
- The box plot also gives us a description of the 25th, 50th,75th quartiles.
- With the help of a box plot, we can also determine the Interquartile range(IQR) where maximum details of the data will be present
- These box plots come under univariate analysis, which means that we are exploring data only with one variable.

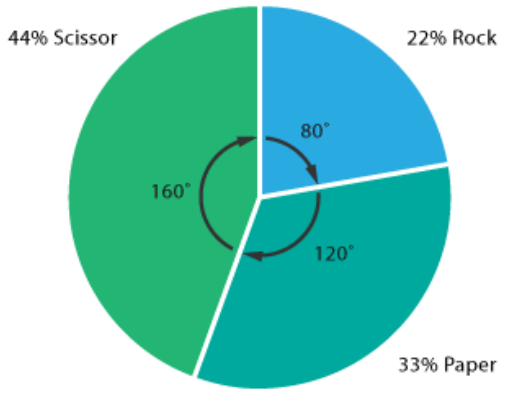



Pie Chart :

A pie chart shows a static number and how categories represent part of a whole the composition of something. A pie chart represents numbers in percentages, and the total sum of all segments needs to equal 100%.

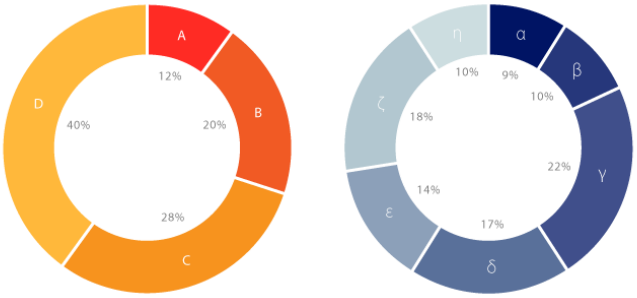
- Extensively used in presentations and offices, Pie Charts help show proportions and percentages between categories, by dividing a circle into proportional segments. Each arc length represents a proportion of each category, while the full circle represents the total sum of all the data, equal to 100%.

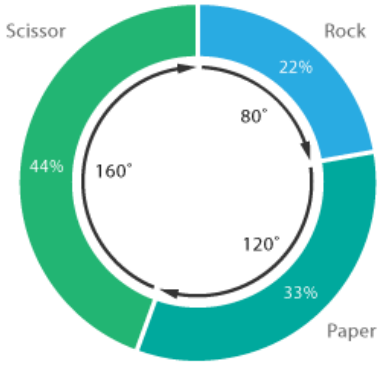



| Data | | | |
|---|---|---|---|
| Rock | Paper | Scissor | TOTAL |
| 2 | 3 | 4 | 9 |
| To calculate percentages | | | |
| 2/9=22% | 3/9=33% | 4/9=44% | 100% |
| Degrees for each "pie slice" | | | |
| (2/9) x 360 = 80° | (3/9) x 360 = 120° | (4/9) x 360 = 160° | 360° |

Donut Chart:



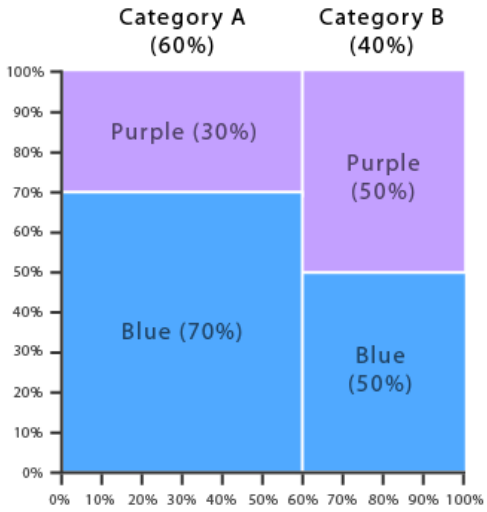



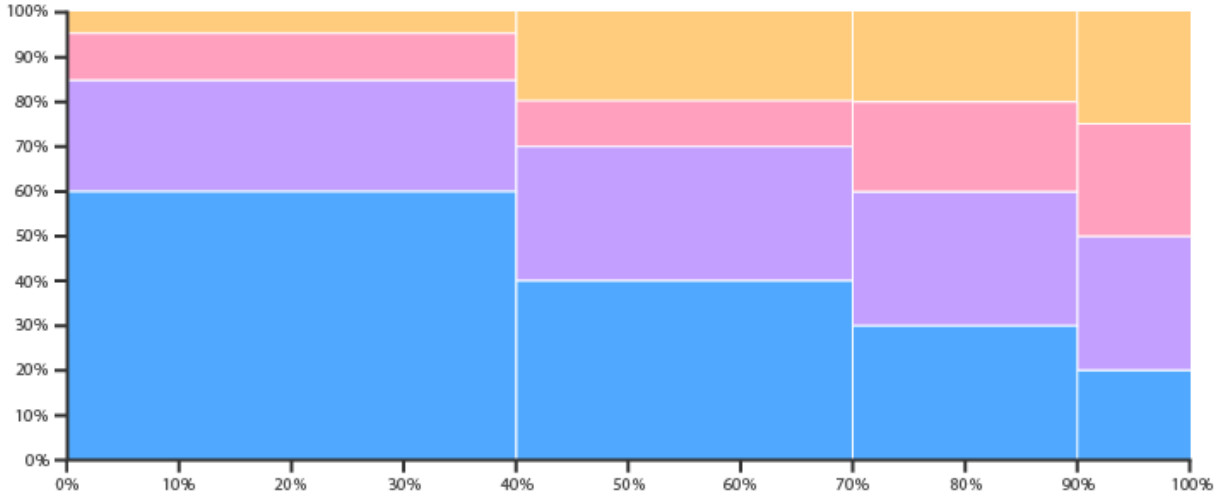

- A donut chart is essentially a Pie Chart with an area of the centre cut out. Pie Charts are sometimes criticised for focusing readers on the proportional areas of the slices to one another and to the chart as a whole. This makes it tricky to see the differences between slices, especially when you try to compare multiple Pie Charts together.

| Data | | | |
|---|---|---|---|
| Rock | Paper | Scissor | TOTAL |
| 2 | 3 | 4 | 9 |
| To calculate percentages | | | |
| 2/9 = **22%** | 3/9 = **33%** | 4/9 = **44%** | 100% |
| Degrees for each "donut slice" | | | |
| (2/9) x 360 = **80°** | (3/9) x 360 = **120°** | (4/9) x 360 = **160°** | 360° |

- A Donut Chart somewhat remedies this problem by de-emphasizing the use of the area. Instead, readers focus more on reading the length of the arcs, rather than comparing the proportions between slices.
- Also, Donut Charts are more space-efficient than Pie Charts because the blank space inside a Donut Chart can be used to display information inside it.

Marimekko Chart:



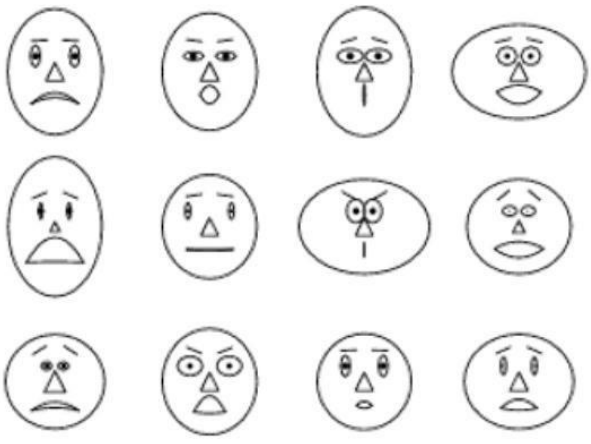

Also known as a *Mosaic Plot*.

- Marimekko Charts are used to visualise categorical data over a pair of variables. In a Marimekko Chart, both axes are variable with a percentage scale, that determines both the width and height of each segment. So Marimekko Charts work as a kind of two-way 100% Stacked Bar Graph. This makes it possible to detect relationships between categories and their subcategories via the two axes.
- The main flaws of Marimekko Charts are that they can be hard to read, especially when there are many segments. Also, it's hard to accurately make comparisons between each segment, as they are not all arranged next to each other along a common baseline. Therefore, Marimekko Charts are better suited for giving a more general overview of the data.

## Icon-Based Visualization Techniques

- It uses small icons to represent multidimensional data values
- Visualization of the data values as features of icons
- Typical visualization methods
    - Chernoff Faces
    - Stick Figures

Chernoff Faces



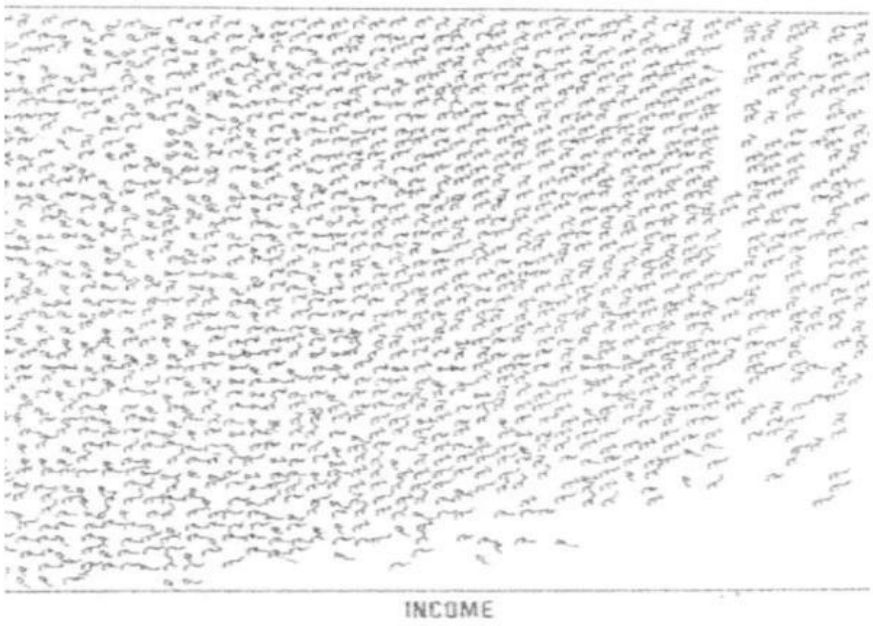
Chernoff Faces

A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.

- The figure shows faces produced using 10 characteristics–head eccentricity,

eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening. Each assigned one of 10 possible values.

Stick Figure

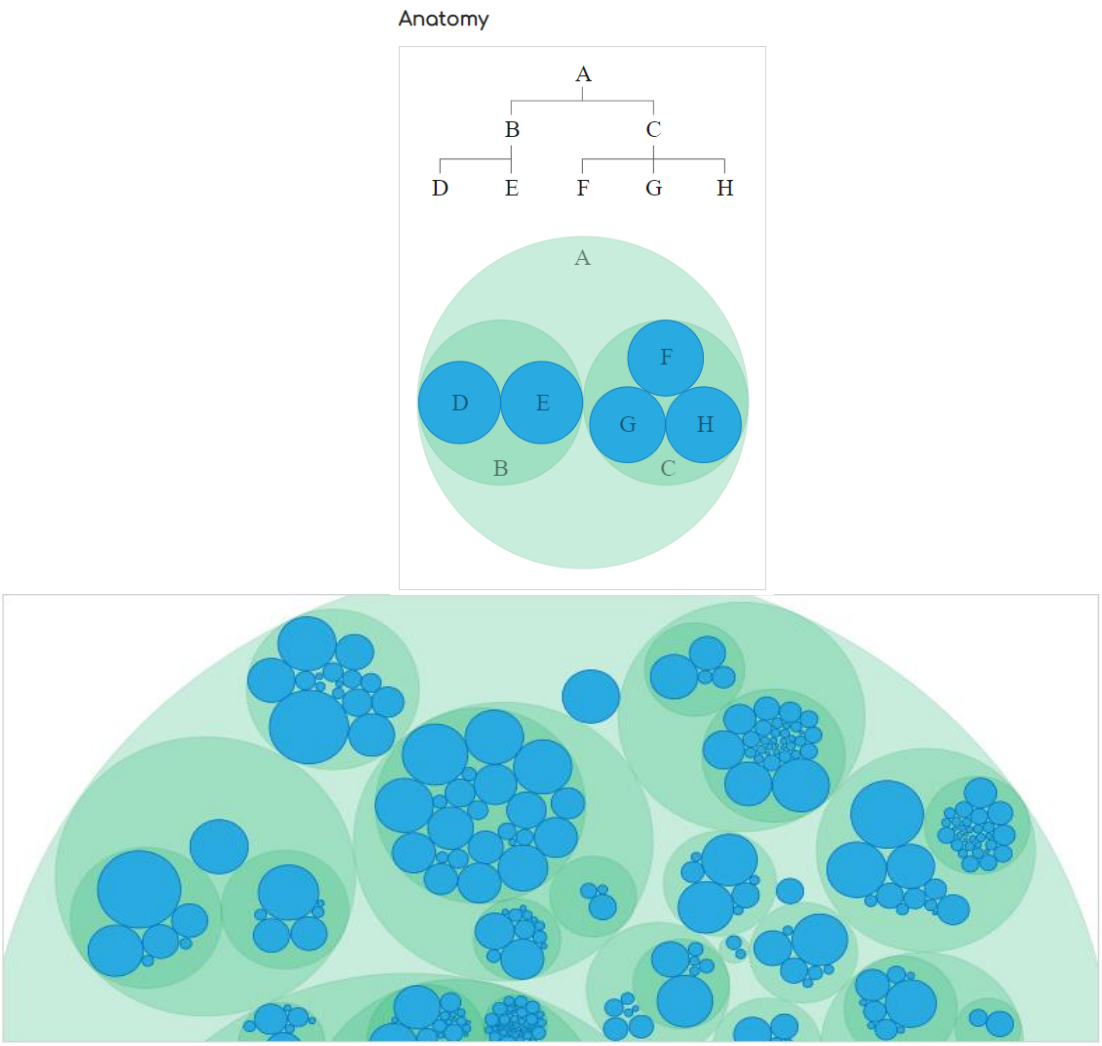



*Stick Figure*

- A census data figure showing age, income, gender, education
- A 5-piece stick figure (1 body and 4 limbs w. different angle/length)
- Age, income are indicated by position of the figure.
- Gender, education are indicated by angle/length.
- Visualization can show a texture pattern.
- 2 dimensions are mapped to the display axes and the remaining dimensions are mapped to the angle and/ or length of the limbs.

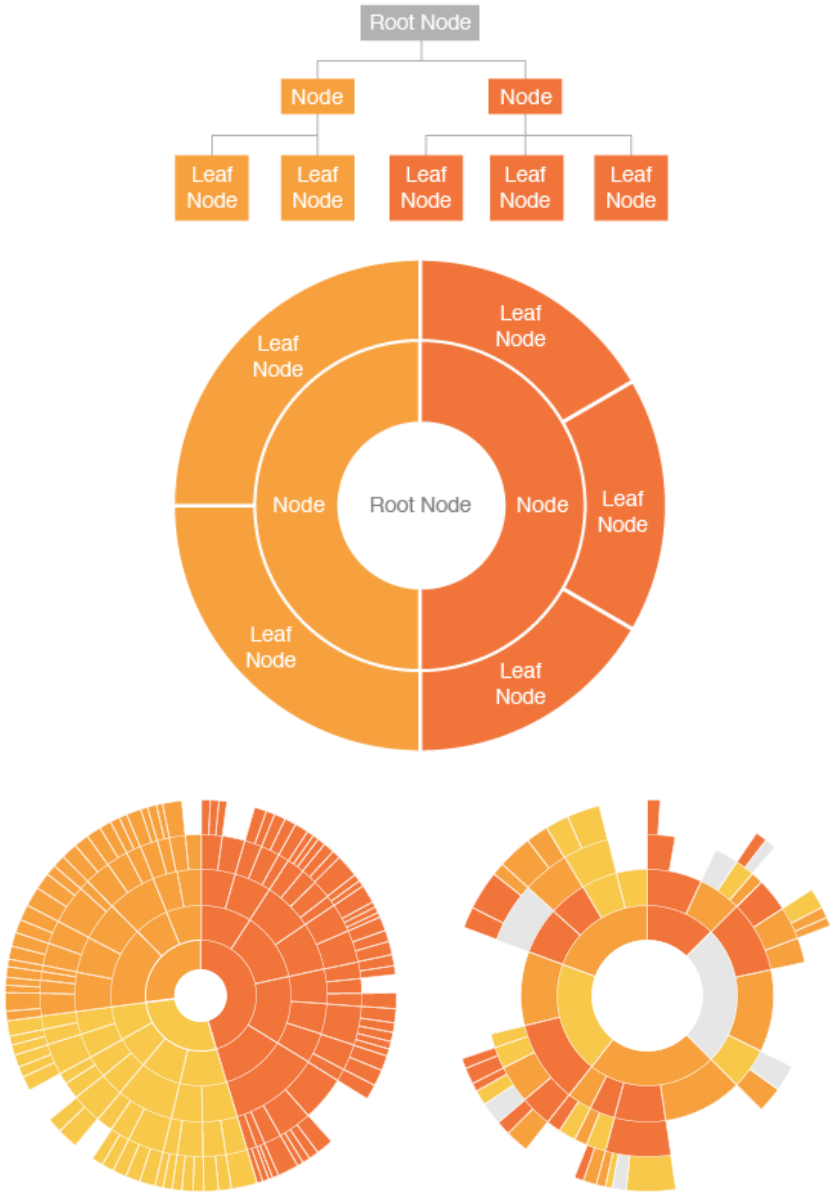
## Hierarchical Visualization

### Circle Packing




- Circle Packing is a variation of a Treemap that uses circles instead of rectangles. Containment within each circle represents a level in the hierarchy: each branch of the tree is represented as a circle and its sub-branches are represented as circles inside of it. The area of each circle can also be used to represent an additional arbitrary value, such as quantity or file size. Colour may also be used to assign categories or to represent another variable via different shades.
- As beautiful as Circle Packing appears, it's not as space-efficient as a Treemap, as there's a lot of empty space within the circles. Despite this, Circle Packing actually reveals hierarchal structure better than a Treemap.

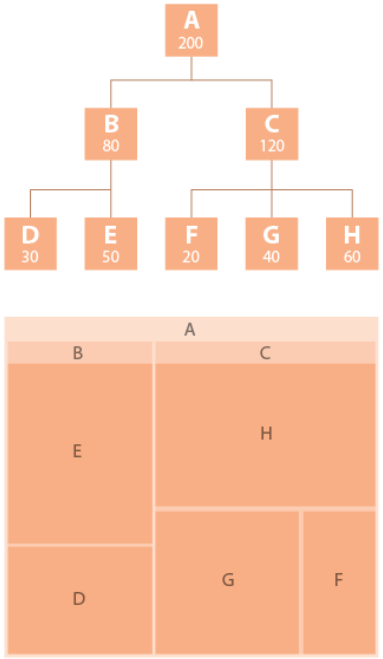
Sunburst Diagram

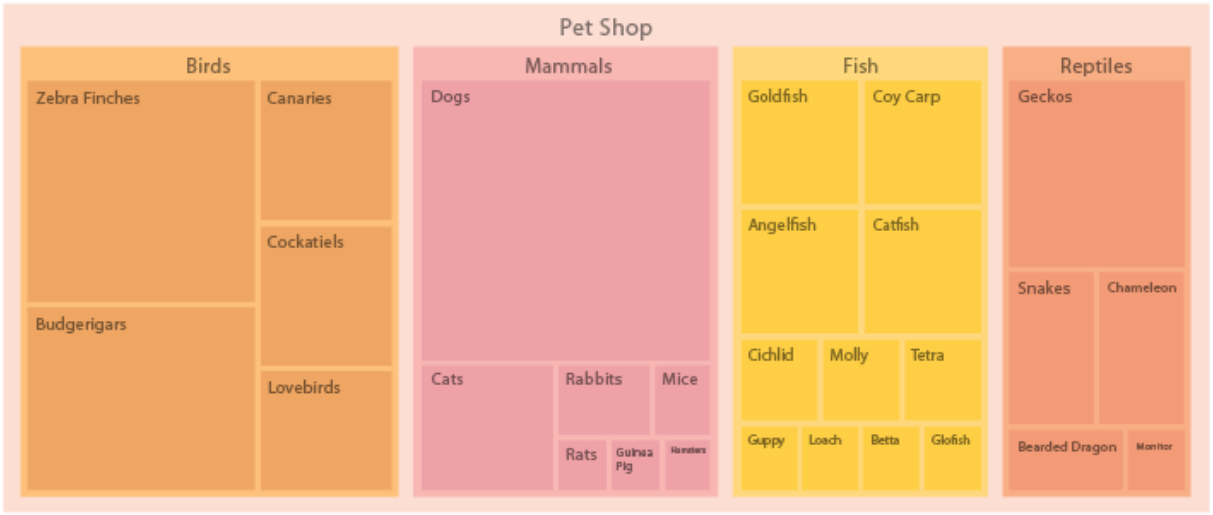



- As known as a *Sunburst Chart, Ring Chart, Multi-level Pie Chart, Belt Chart, Radial Treemap*.
- This type of visualisation shows hierarchy through a series of rings, that are sliced for each category node. Each ring corresponds to a level in the hierarchy, with the central circle representing the root node and the hierarchy moving outwards from it.

- Rings are sliced up and divided based on their hierarchical relationship to the parent slice. The angle of each slice is either divided equally under its parent node or can be made proportional to a value.
- Colour can be used to highlight hierarchal groupings or specific categories.

Treemap:



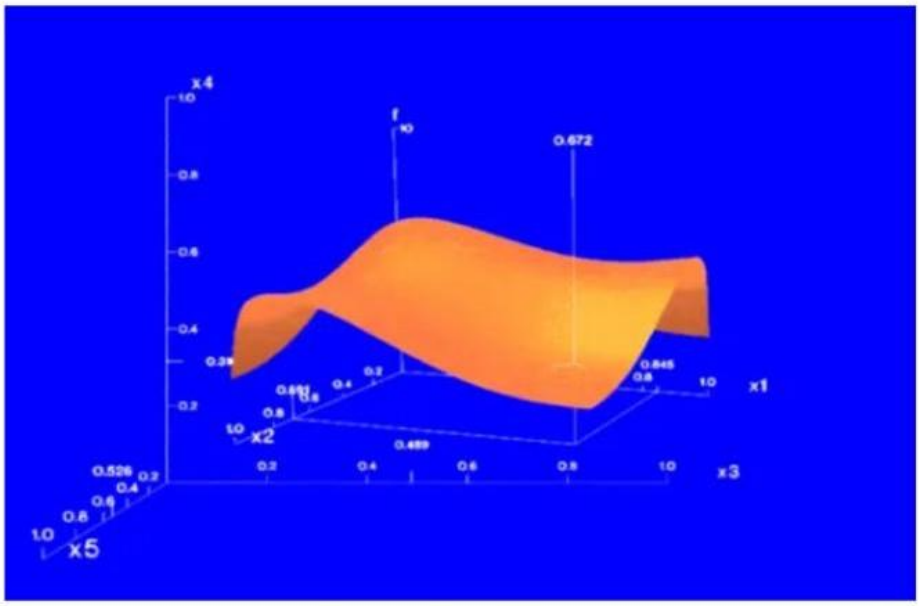




- Treemaps are an alternative way of visualising the hierarchical structure of

a Tree Diagram while also displaying quantities for each category via area size. Each category is assigned a rectangle area with their subcategory rectangles nested inside of it.

- When a quantity is assigned to a category, its area size is displayed in proportion to that quantity and to the other quantities within the same parent category in a part-to-whole relationship. Also, the area size of the parent category is the total of its subcategories. If no quantity is assigned to a subcategory, then it's area is divided equally amongst the other subcategories within its parent category.
- The way rectangles are divided and ordered into sub-rectangles is dependent on the tiling algorithm used. Many tiling algorithms have been developed, but the "squarified algorithm" which keeps each rectangle as square as possible is the one commonly used.
- Ben Shneiderman originally developed Treemaps as a way of visualising a vast file directory on a computer, without taking up too much space on the screen. This makes Treemaps a more compact and space-efficient option for displaying hierarchies, that gives a quick overview of the structure. Treemaps are also great at comparing the proportions between categories via their area size.
- The downside to a Treemap is that it doesn't show the hierarchal levels as clearly as other charts that visualise hierarchal data (such as a Tree Diagram or Sunburst Diagram).

## Visualizing Complex Data and Relations




*Hierarchical Visualization*

- For a large data set of high dimensionality, it would be difficult to visualize alldimensions at the same time.
- Hierarchical visualization techniques partition all dimensions into subsets (i.e., subspaces).
- The subspaces are visualized in a hierarchical manner
- "Worlds-within-Worlds," also known as n-Vision, is a representative hierarchicalvisualization method.
- To visualize a 6-D data set, where the dimensions are F,X1,X2,X3,X4,X5.
- We want to observe how F changes w.r.t. other dimensions. We can fix X3,X4,X5dimensions to selected values and visualize changes to F w.r.t. X1, X2
- Most visualization techniques were mainly for numeric data.
- Recently, more and more non-numeric data, such as text and social networks, havebecome available.
- Many people on the Web tag various objects such as pictures, blog entries, and productreviews.
- A tag cloud is a visualization of statistics of user-generated tags.

- Often, in a tag cloud, tags are listed alphabetically or in a user-preferred order.
- The importance of a tag is indicated by font size or color.


*Visualizing Complex Data and Relations*

Word Cloud:


Also known as a *Tag Cloud*.

- A visualisation method that displays how frequently words appear in a given body of text, by making the size of each word proportional to its frequency. All the words are then arranged in a cluster or cloud of words. Alternatively, the words can also be arranged in any format: horizontal lines, columns or within a shape.
- Word Clouds can also be used to display words that have meta-data assigned to them. For example, in a Word Cloud with all the World's country's names, the population could be assigned to each name to determine its size.

- Colour used on Word Clouds is usually meaningless and is primarily aesthetic, but it can be used to categorise words or to display another data variable.
- Typically, Word Clouds are used on websites or blogs to depict keyword or tag usage. Word Clouds can also be used to compare two different bodies of text together.
- Although being simple and easy to understand, Word Clouds have some major flaws:
- Long words are emphasised over short words.
- Words whose letters contain many ascenders and descenders may receive more attention.
- They're not great for analytical accuracy, so used more for aesthetic reasons instead.

*Source: https://datavizcatalogue.com/index.html*