

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

4.1 Introduction

4.2 Classification of Visual data analysis techniques

4.3 Data type to be visualized

4.4 Multidimensional Scaling

4.5 Sammon's Mapping

4.6 Interaction Techniques

4.7 Histograms

4.8 Spectral Analysis

4.1 Introduction:

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

“A picture is worth a thousand words”—A popular proverb. Primary Role in the Data Mining Process is during the Data Exploration, and Conditioning phase. Data Exploration and Conditioning required preliminary step before formal analysis i.e. Visual analysis

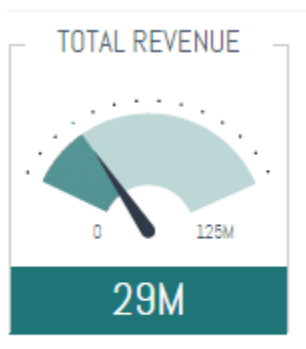
- A free-form data exploration
 - Main idea is to support the data mining goal and subsequent formal analysis
 - Techniques range from basic plots to interactive visualizations
 - Features such filtering, zooming, color and multiple panels
- Usage of Visualization Techniques depends on
- Different data mining tasks such as classification, prediction, clustering etc.
 - Different data mining techniques such as CART, HAC etc.

4.2 Classification of Visual data analysis techniques

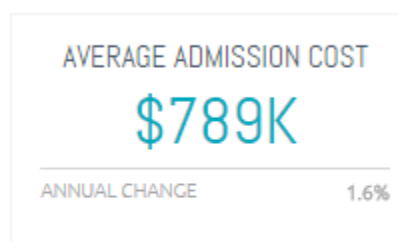
1. Indicator

These are particularly useful when you want to give an instant idea of how well the business is doing on a particular KPI. Incorporating a simple “gauge indicator” visualization shows you immediately whether you’re above or below target, and whether you’re moving in the right direction. This is especially effective if you incorporate color coding, like red or green, or up and down arrows.

Even more straight forward is a Numeric Indicator like the one below right, which gives a simple headline figure and an indication of how this compares to the previous year / quarter / month etc.



Gauge Indicator



Numeric Indicator

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

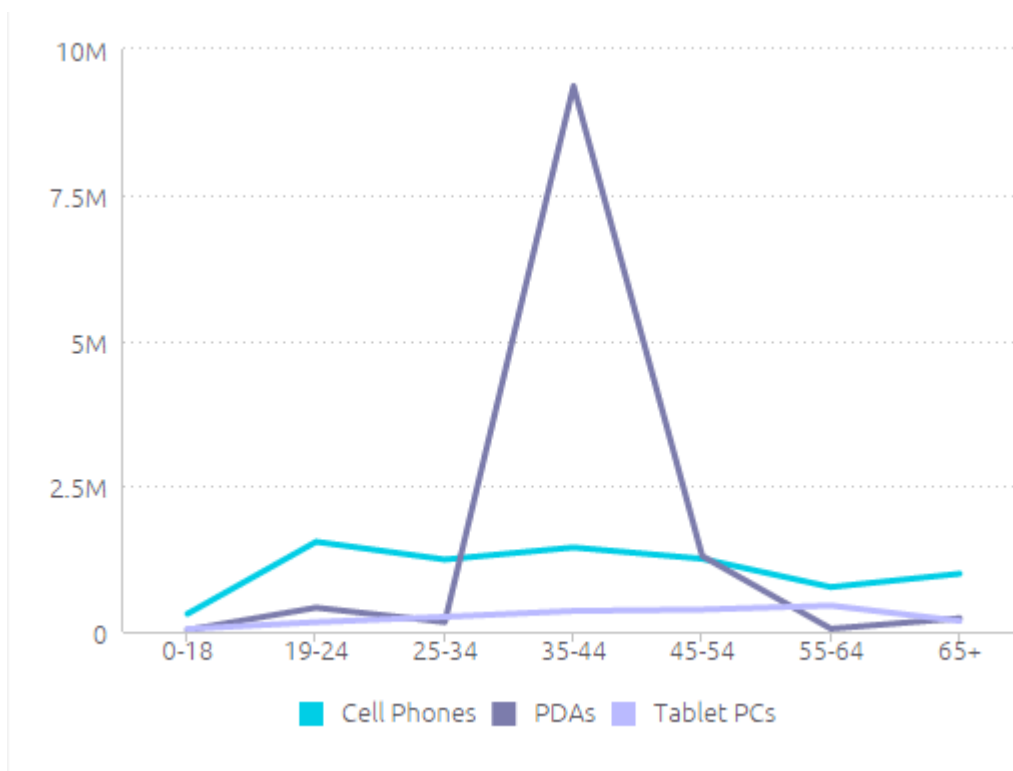
UNIT IV

Subject

Code : SIT1303

2. Line Chart

Line charts are resounding popular for a range of business use cases because they demonstrate an overall trend swiftly and concisely, in a way that's hard to misinterpret. In particular, they're good for depicting trends for different categories over the same period of time, to aid comparison. For example, this graph visualizes sales figures by age group for three different product lines:



Here, you can see at a glance that your biggest customers are 34-45 year old buyers of PDAs, followed by 19-24 year old buyers of cell phones.

3. Bar Chart

Bar charts are great for comparing several different values, especially when some of these are broken into color-coded categories. To illustrate the difference between this and a line graph, let's now take the same information as above and re-visualize it as a bar chart:

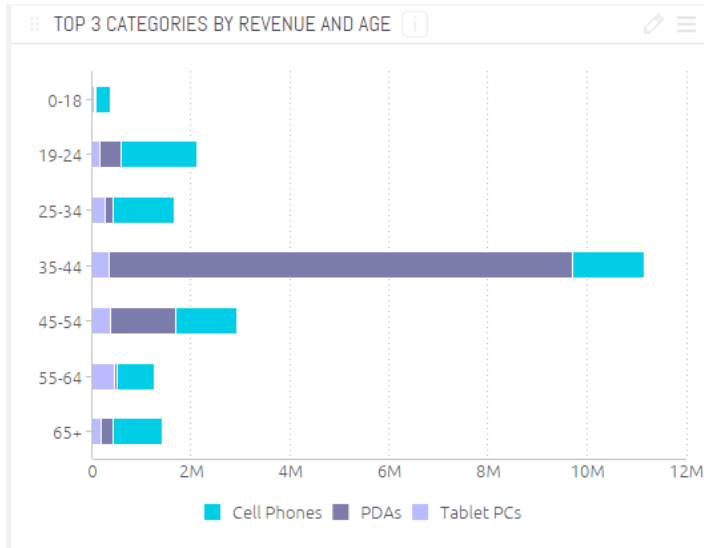
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



While the primary takeaway from the line chart is the huge central spike, representing PDAs bought by 34-45 year olds, here you are encouraged to take in the more granular differences between sales figures for each category within each age group. Since the different product lines are groups by age group, you can also see at a glance which age groups are the most valuable to your business, rather than focussing on the product line.

4. Column Chart

Usually it makes sense to use column charts for side-by-side comparison of different values. You can also use them to show change over time, although it makes sense to do this when you want to draw attention to total figures rather than the shape of the trend (which is more effective with a line chart).

For example, the chart below shows total website page views vs sessions on a series of dates. The numbers don't move much from day to day, so a line graph wouldn't reveal anything insightful in terms of trends; rather, the pertinent information here is the concrete number of visitors to the website each day.

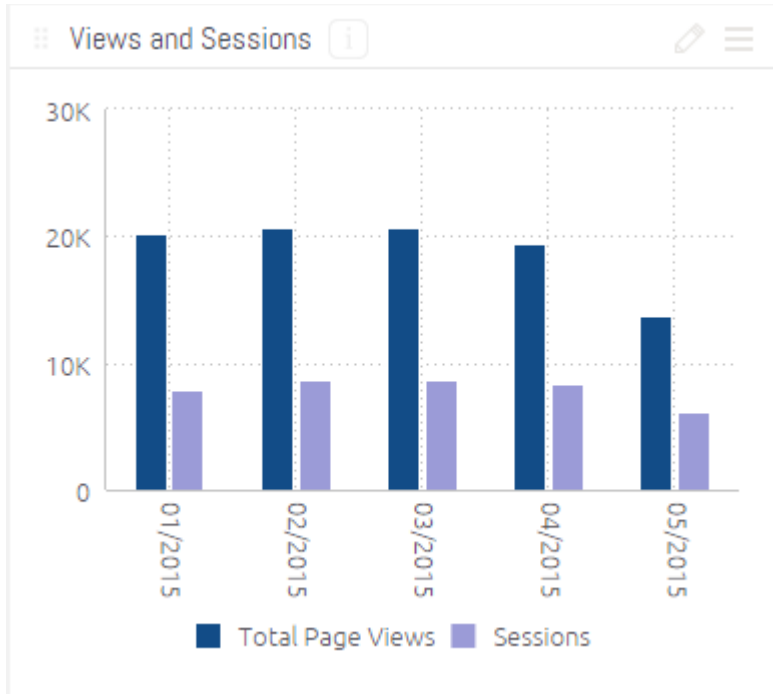
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



If you want to highlight or contrast key figures and an overall trend, you can combine a line *and* column chart, as in the example below.

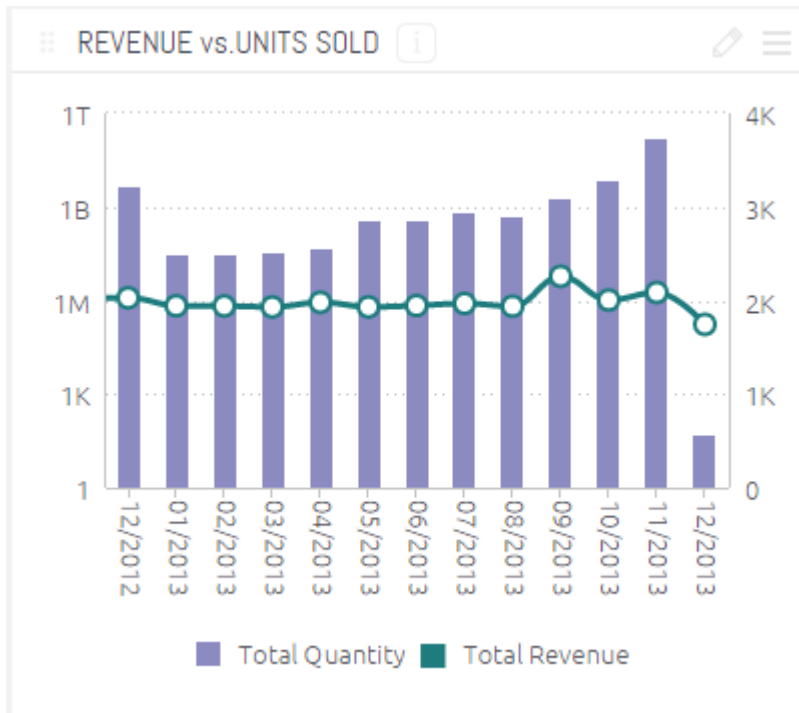
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



As you can see here, the total number of units sold and the total revenue for each month tell a slightly different story; the visualization actually opens up a new line of inquiry into which units are the most profitable, even when fewer are sold – which could prove key in shaping your sales and marketing strategy going forward.

5. Pie Chart

Pie charts are useful for communicating instantaneously what share each value makes up of the whole. They're far more intuitive than simply listing percentages that add up to 100%.

For example, this pie chart illustrates which campaigns bring in the biggest share of total leads. You see at once that AdWords is the most effective source, followed by social media and then webinar signups. An instant insight would illuminate to your marketing team what's working best, helping them to rapidly reassign resources or refocus their efforts to maximise lead generation.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

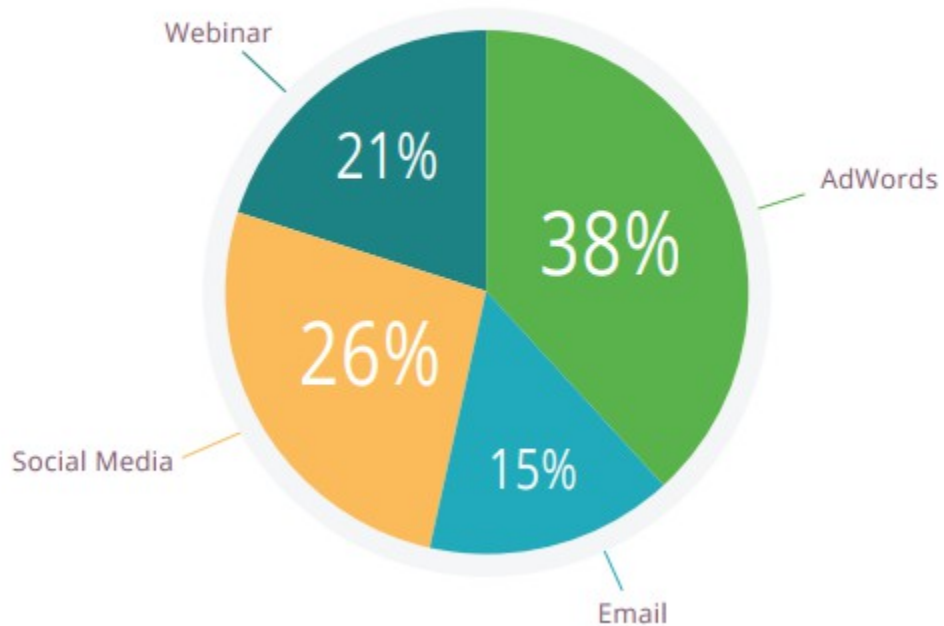
Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

LEAD BREAKDOWN BY CAMPAIGN SOURCE



Note that for a pie chart to be effective, you need to have six categories or fewer. Any more than that and the chart will be too crowded, and the values too indistinct, to garner any insight. Check out this monstrosity, comparing population sizes of US states, as evidence of how a pie chart can communicate very little information at all:

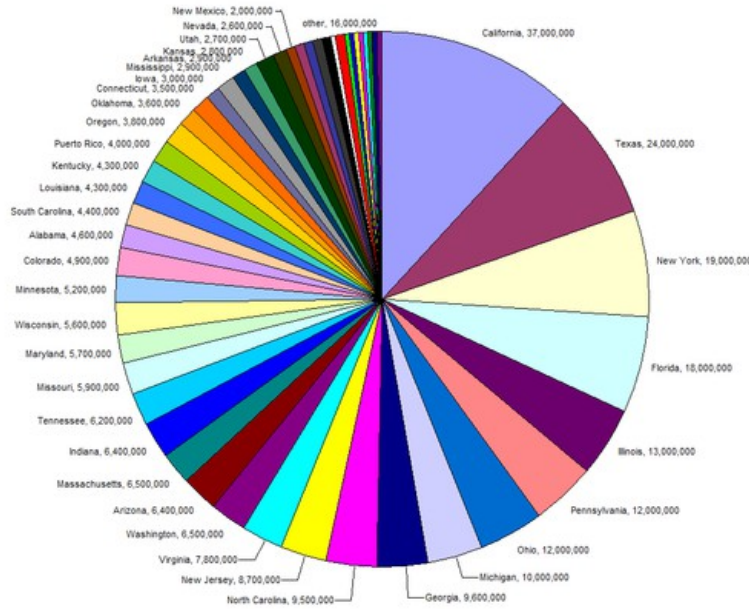
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



Source: European Environment Agency

6. Area Chart

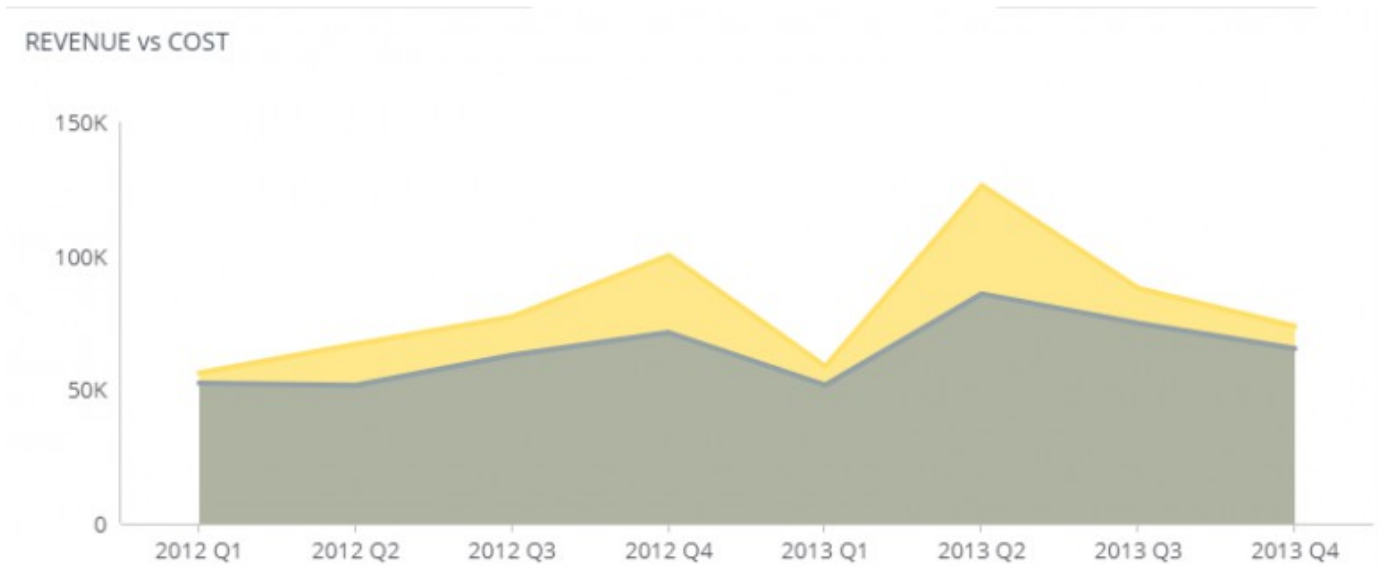
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



Area charts are useful as they give a sense of the overall volume, as well as the proportion of this taken up by each category.

In the example above, you can see how much of one volume (revenue) is overlapped by another volume (cost). This is a great way to impose a reality check on your revenue estimations – you see at once that the yellow sliver of profit is at its thinnest, helping you to assess where cash flow really is tightest, rather than where in the year you’re simply bringing in the most cash.

(Note that layered visualizations like these can start to get confusing when you introduce more than three values into the mix.)

This kind of information can give an instant insight that helps with issues like resource planning, ordering patterns, financial management allocating appropriate storage space, and so on.

7. Pivot Table

Pivot tables aren’t the most beautiful or intuitive ways to visualize data, but they are useful when you want to quickly extract key figures while seeing exact numbers (rather than get a sense of trends), especially if you don’t have access to a self-service BI tool that can automate this for you.

In this example, complex patient information is summarized to give you a detailed overview of costs, patient numbers and average days admitted to hospital:


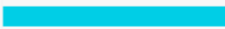








SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

TOP 10 DIAGNOSIS				
DIAGNOSIS	# PATIENTS	AVG COST		AVG DAYS ADMITTED
Bypass	169	\$777,872		5.85
Cardiac Arrest	178	\$777,426		6.24
Chemotherapy	174	\$790,289		6.62
Chronic Headache	173	\$795,728		6.78
Diabetes	191	\$786,282		5.67
Ear infection	175	\$755,058		5.63
EKG	183	\$786,703		6.03
Epilepsy	177	\$785,052		5.99
Hypoglycemia	177	\$777,663		5.85
Radiotherapy	196	\$776,702		6.15

8. Scatter Chart

These represent categories by circle color and the volume of the data by circle size; they're used to visualize the distribution of, and relationship between, two variables.

For example, the chart below visualizes each product line by the number of units sold and the revenue this brings in, representing the value in physical size. It also breaks this down by gender (hovering over the circles would reveal the name of the product in the original).

In this scenario, you would determine that your most frequent (and profitable) clients are currently men – which could, for example, lead you either to focus more marketing effort on male shoppers, or to seek out more effective ways of engaging female customers, depending on your business priorities.

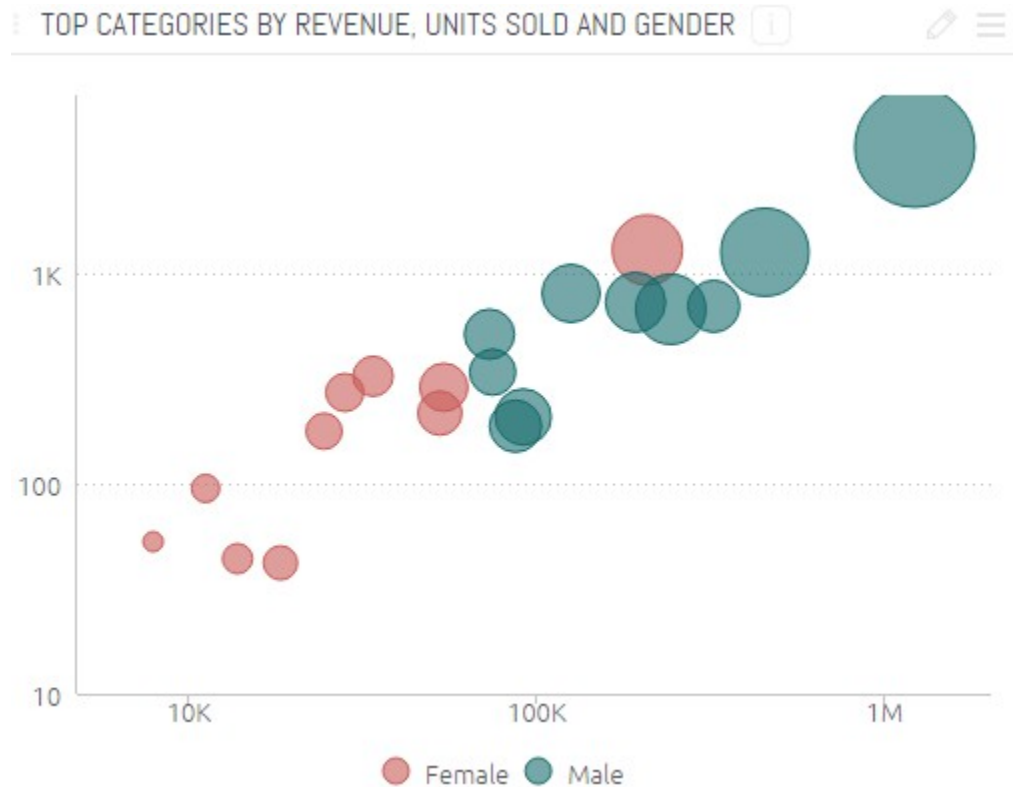
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



9. Bubble Chart

Similar to Scatter Charts, Bubble Charts depict the weight of values by circle circumference size. However, they differ in that they pack many different values into one small space and only represent a single measurement per category. They are useful when you want to demonstrate how a handful of categories are highly significant compared to a sea of insignificant ones.

For example, take this bubble chart based on [this research by the New York Times](#), which breaks down how the US government's \$3.7 trillion in "welfare" is actually spent:

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

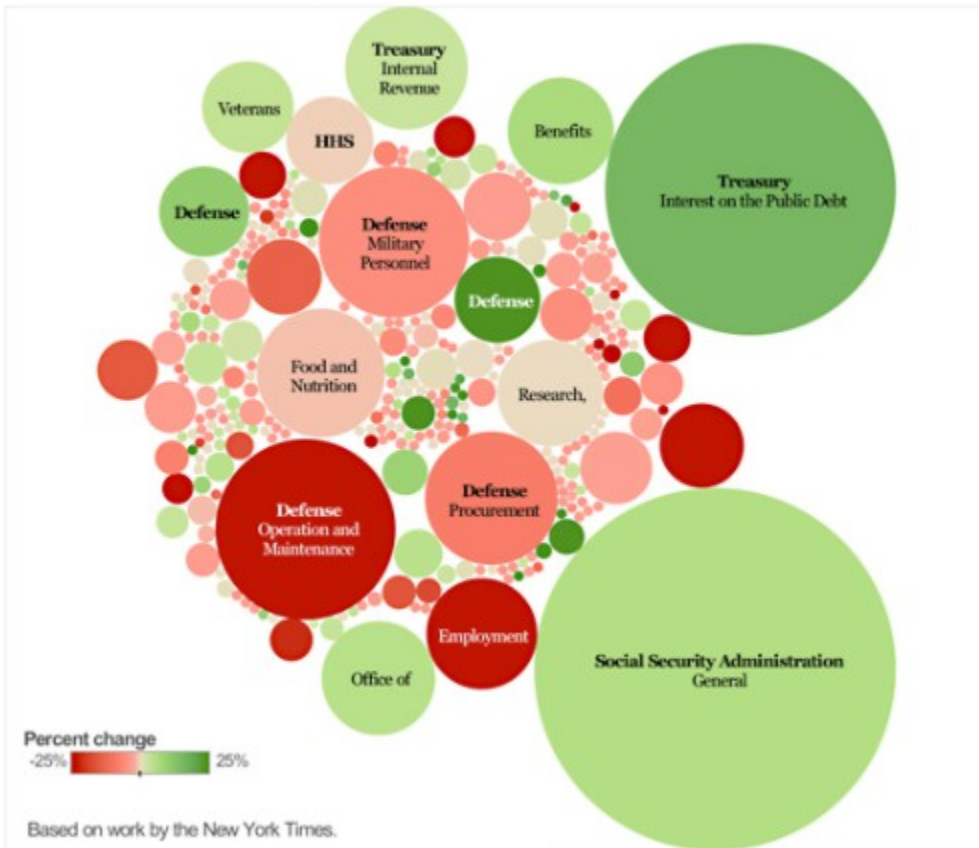
Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

How \$3.7 Trillion is Spent



Source: The

New York Times

You see immediately here that the percentage most people think of as welfare (i.e. Benefits) is dwarfed in comparison to admin costs, defense-connected spending and interest, while most outgoings wrapped into this category are so tiny they are barely visible.

While bubble charts like these are often used to make a stark political point, you can also use this to great effect in your business to demonstrate things like misplaced priorities, actual comparative costs and values, or to highlight areas of highest spending when looking to streamline activities and cut costs.

10. Treemap

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

Treemaps are useful for displaying hierarchies and comparative value between categories and subcategories, as well as allowing you to retain detail while projecting an instant sense of which areas are most important overall.

You achieve this by nesting color-coded rectangles inside each other, weighted to reflect their share of the whole. This treemap depicts the value of different marketing channels, which are then broken down by country. You see at a glance that AdWords is your most successful channel, but that the US is your most valuable destination, across all channels.



11. Polar Chart

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

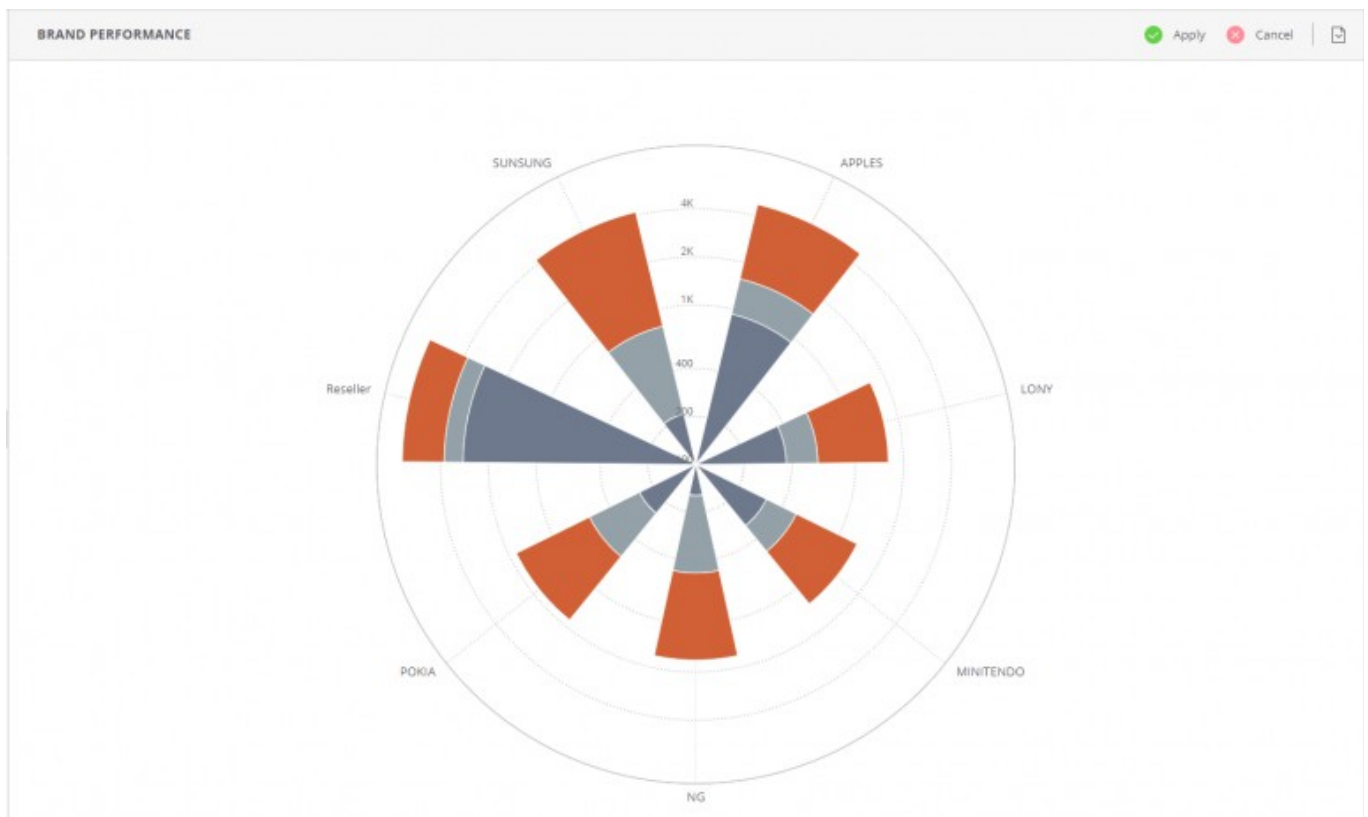
UNIT IV

Subject

Code : SIT1303

A polar chart (or polar area diagram) is a type of pie chart. However, instead of depicting each value's share of the whole by the size of the angle, all the sectors have equal angles, and the value is shown by the how far it reaches from the center of the circle.

The example below is from a sales dashboard depicting sales of multiple brands. Each segment represents a brand name, while red represents new products, light gray represents refurbished products, and dark grey means "unspecified".



12. Area Map/Scatter Map

These kinds of data visualizations allow you to see immediately which geographical locations are most significant to your business. Data is visualized as points of color on a map; values are represented by circle size.

For example, the map below depicts website visitors by location, while the color indicates the percentage of conversions (the brighter the green, the higher the conversion rate).

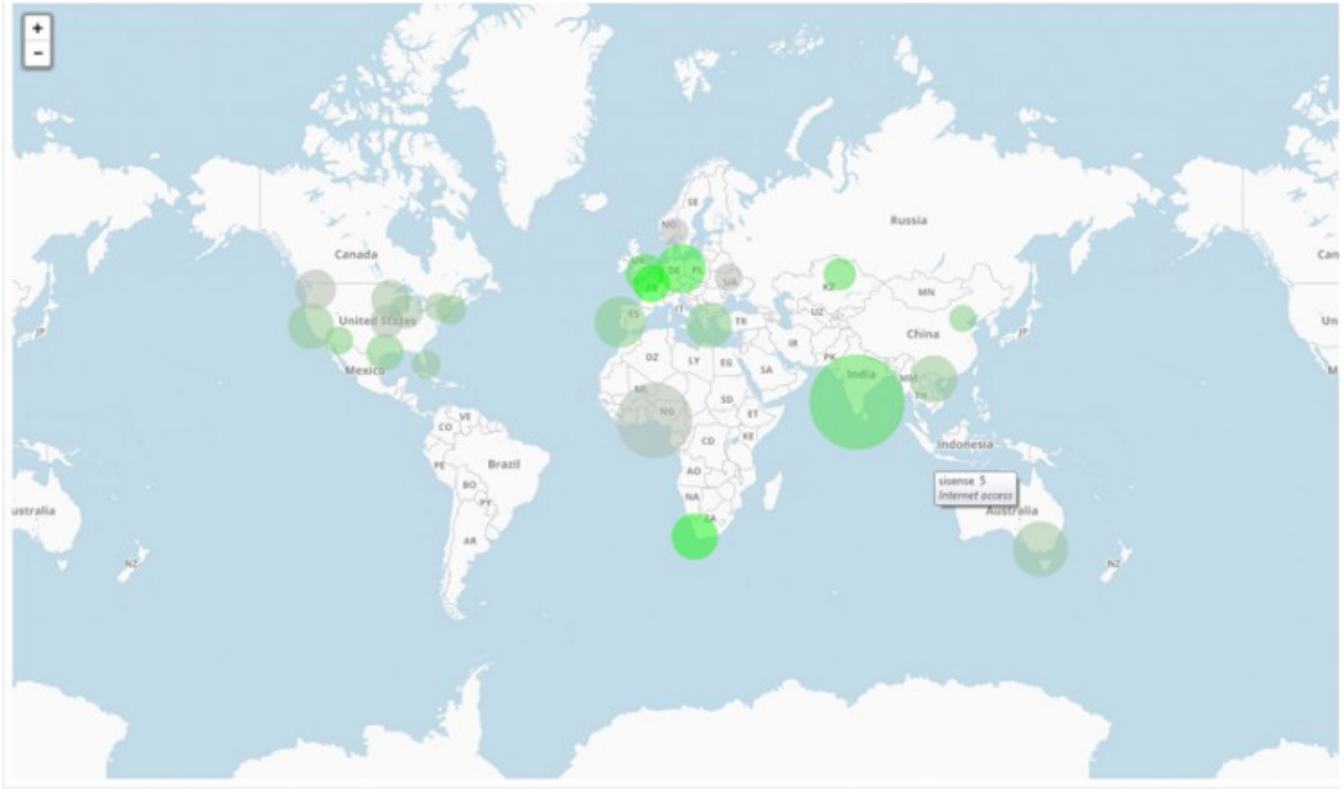
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



This kind of representation is incredibly useful as it gives two vital pieces of information at a glance: where in the world *most* of your visitors are from, compared to where in the world your most *valuable* visitors are from. Insights like these can show up weaknesses in a marketing strategy in seconds.

13. Funnel Chart

This is a very specific type of visualization that depicts the decreasing values as customers move through the sales funnel. The beauty of it is that it brings to life your conversion rates at each step, so you can see quickly where you are losing people in the process. The funnel chart below shows the number of people at each demand stage, from initial website visit, through every touchpoint until a final sale:

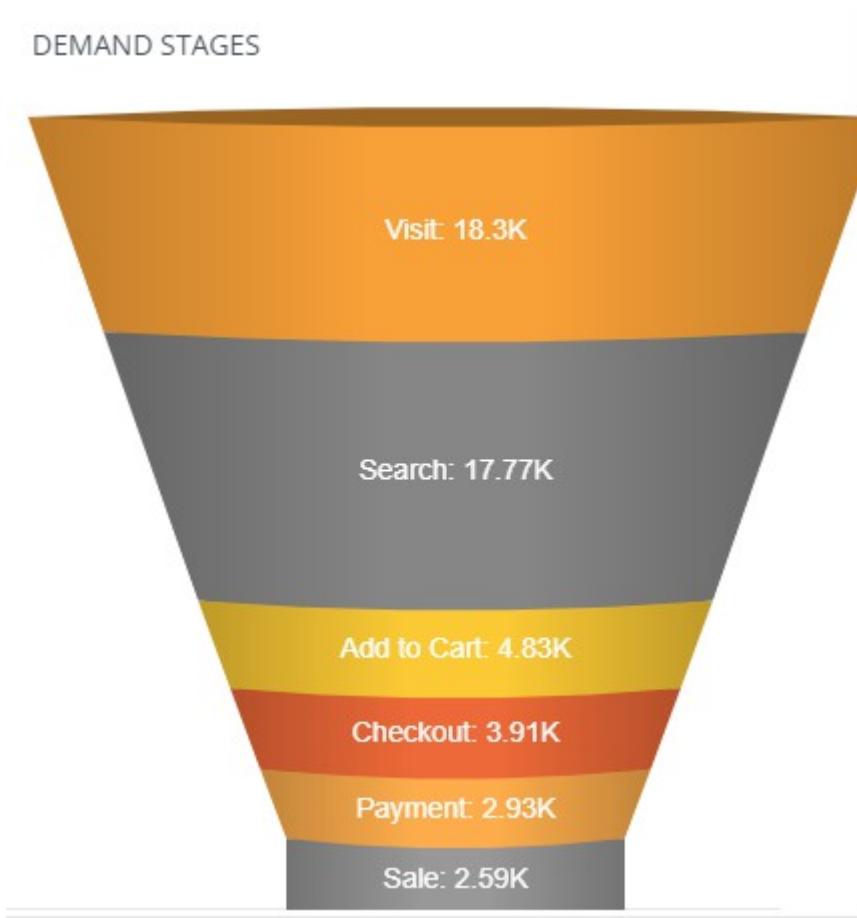
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



14. Fisheye/Cartesian Distortion

Finally, this isn't a data visualization style per se, but rather a useful addition that allows you to zoom into the details in a more complex visualization, like a force-directed graph or bubble chart. As you move your cursor over a graph, the area you're looking expands in fisheye view, allowing you to dip in and out to see more granular details as needed.

Whichever type of data visualization you opt for, remember that, to make it accurate and effective, the software you use must be able to interact effectively with your data.

Your data visualization software should be able to handle whichever data sources you throw at it. You must be able to clean and prepare your data properly. You should be able to incorporate a [powerful external visualization tool like D3](#) to enhance your results. Without a powerful,

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

flexible platform, you could end up creating a beautiful structure... but on very shaky foundations.

4.3 Data type to be visualized

Seven data types

- 1D linear
- 2D map
- 3D world
- Multidimensional
- Temporal
- Tree
- network

– First four are dimensional, last three are structural

Tasks:

- Overview
- Zoom
- Filter
- details-on-demand
- relate

- history,

- extract

1D linear data

- Items which can be organized sequentially e.g. text document, list of names
- Design issues:
 - Colours, sizes, layout
 - Scrolling, selection methods
- Example user tasks: check which items have some required attribute

2D map data

- Items make up some part of the 2D area– Not necessarily rectangular, e.g. Lake on Google Map
- e.g. Geographic map, floor plans
- Example user tasks: finding items, finding path between items

3D world data

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

- Items with complex relationships with other items– e.g. Volume, temperature, density
- e.g. Medical imaging, architectural drawing, scientific simulations
- Design issues: position, orientation and navigation for viewing 3D application
- Example user tasks: temperature, density

Multidimensional data

- Items with n attributes in n-dimensional space
- Relational database contents can be treated this way
- Interface may allow user to view 2 dimensions at a time

Temporal data

- Very close idea to 1D sequential data, but warrant a distinct data type in the taxonomy as temporal data is so common
- e.g. Stock market data, weather
- Items have a beginning and end time, may overlap in time
- Example user tasks: finding events during a time period, searching for periodical behaviour

Tree data

- Non-root items have a link to a parent item
- Items, links can have multiple attributes
- e.g. Windows file explorer
- Example user tasks: how many items are children of a node, how deep or shallow is the graph

Network Data

- Items linked to arbitrary number of other items
- Example user task: shortest path, least costly path
- How to visualize, layout the network?

Seven Basic Tasks

Overview task

- Give user overview of entire set of items
- e.g. Zoom out, Field-of-view box, Fish eye strategy
- If complete overview is impossible, is there an effective overview strategy?

Zoom task

- Allow users to focus in on or enlarge items of interest
- May allow users to control the “zoom factor”
- Extra importance if small display is a possibility– e.g. Google maps on smartphone

Filter task

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

- Allows users to remove items eg:Dynamic queries applied to data
- Highlighting desired items

Details-on-demand task

- Selecting item(s) and allowing users to get more details
- e.g. Click on an item, new window
- e.g. Searching for a book on Amazon, can click on different editions and get more details

Seven Basic Tasks

Relate task

- Relating items within a set
- How to show relationships?
 - Proximity
 - Containment
 - Connected lines
 - Colour-coding

History task

- History of tasks which can be undone, replayed,refined
- Much work is a “process”, allowing for refinement, steps, important e.g. Retrieval of past searches

Extract task

- Extraction of items– Based on query parameters
- Allow user to “save”, publish, examine extracted items

4.4 Multidimensional Scaling

Multidimensional scaling (*MDS*) can be considered to be an alternative to factor analysis. In general, the goal of the analysis is to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities (distances) between the investigated objects. In factor analysis, the similarities between objects (e.g., variables) are expressed in the

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

correlation matrix. With MDS, you can analyze any kind of similarity or dissimilarity matrix, in addition to correlation matrices.

Logic of MDS

The following simple example may demonstrate the logic of an MDS analysis. Suppose we take a matrix of distances between major US cities from a map. We then analyze this matrix, specifying that we want to reproduce the distances based on two dimensions. As a result of the MDS analysis, we would most likely obtain a two-dimensional representation of the locations of the cities, that is, we would basically obtain a two-dimensional map.

In general then, MDS attempts to arrange "objects" (major cities in this example) in a space with a particular number of dimensions (two-dimensional in this example) so as to reproduce the observed distances. As a result, we can "explain" the distances in terms of underlying dimensions; in our example, we could explain the distances in terms of the two geographical dimensions: north/south and east/west.

Orientation of axes. As in factor analysis, the actual orientation of axes in the final solution is arbitrary. To return to our example, we could rotate the map in any way we want, the distances between cities remain the same. Thus, the final orientation of axes in the plane or space is mostly the result of a subjective decision by the researcher, who will choose an orientation that can be most easily explained. To return to our example, we could have chosen an orientation of axes other than north/south and east/west; however, that orientation is most convenient because it "makes the most sense" (i.e., it is easily interpretable).

Computational Approach

MDS is not so much an exact procedure as rather a way to "rearrange" objects in an efficient manner, so as to arrive at a configuration that best approximates the observed distances. It actually moves objects around in the space defined by the requested number of dimensions, and checks how well the distances between objects can be reproduced by the new configuration. In

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

more technical terms, it uses a function minimization [algorithm](#) that evaluates different configurations with the goal of maximizing the goodness-of-fit (or minimizing "lack of fit").

Measures of goodness-of-fit: Stress. The most common measure that is used to evaluate how well (or poorly) a particular configuration reproduces the observed distance matrix is the stress measure. The raw stress value *Phi* of a configuration is defined by:

$$\text{Phi} = \sum_{i,j} \delta_{ij} [d_{ij} - f(\delta_{ij})]^2$$

In this formula, d_{ij} stands for the reproduced distances, given the respective number of dimensions, and δ_{ij} (delta_{ij}) stands for the input data (i.e., observed distances). The expression $f(\delta_{ij})$ indicates a nonmetric, monotone transformation of the observed input data (distances). Thus, it will attempt to reproduce the general rank-ordering of distances between the objects in the analysis.

There are several similar related measures that are commonly used; however, most of them amount to the computation of the sum of squared deviations of observed distances (or some monotone transformation of those distances) from the reproduced distances. Thus, the smaller the stress value, the better is the fit of the reproduced distance matrix to the observed distance matrix.

Shepard diagram. You can plot the reproduced distances for a particular number of dimensions against the observed input data (distances). This scatterplot is referred to as a *Shepard* diagram. This plot shows the reproduced distances plotted on the vertical (*Y*) axis versus the original similarities plotted on the horizontal (*X*) axis (hence, the generally negative slope). This plot also shows a step-function. This line represents the so-called *D-hat* values, that is, the result of the monotone transformation $f(\delta_{ij})$ of the input data. If all reproduced distances fall onto the step-line, then the rank-ordering of distances (or similarities) would be perfectly reproduced by the respective solution (dimensional model). Deviations from the step-line indicate lack of fit.

How Many Dimensions to Specify?

If you are familiar with factor analysis, you will be quite aware of this issue. If you are not familiar with factor analysis, you may want to read the [Factor Analysis](#) section in the manual;

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

however, this is not necessary in order to understand the following discussion. In general, the more dimensions we use in order to reproduce the distance matrix, the better is the fit of the reproduced matrix to the observed matrix (i.e., the smaller is the stress). In fact, if we use as many dimensions as there are variables, then we can perfectly reproduce the observed distance matrix. Of course, our goal is to *reduce* the observed complexity of nature, that is, to explain the distance matrix in terms of fewer underlying dimensions. To return to the example of distances between cities, once we have a two-dimensional map it is much easier to visualize the location of and navigate between cities, as compared to relying on the distance matrix only.

Sources of misfit. Let's consider for a moment why fewer factors may produce a worse representation of a distance matrix than would more factors. Imagine the three cities *A*, *B*, and *C*, and the three cities *D*, *E*, and *F*; shown below are their distances from each other.

	A	B	C		D	E	F	
A	0				D	0		
B	90	0			E	90	0	
C	90	90	0		F	180	90	0

In the first matrix, all cities are exactly 90 miles apart from each other; in the second matrix, cities *D* and *F* are 180 miles apart. Now, can we arrange the three cities (objects) on one dimension (line)? Indeed, we can arrange cities *D*, *E*, and *F* on one dimension:

D---90 miles---E---90 miles---F

D is 90 miles away from *E*, and *E* is 90 miles away from *F*; thus, *D* is 90+90=180 miles away from *F*. If you try to do the same thing with cities *A*, *B*, and *C* you will see that there is no way to arrange the three cities on one line so that the distances can be reproduced. However, we can arrange those cities in two dimensions, in the shape of a triangle:

A
90 miles 90 miles
B 90 miles **C**

Arranging the three cities in this manner, we can perfectly reproduce the distances between them. Without going into much detail, this small example illustrates how a particular distance matrix implies a particular number of dimensions. Of course, "real" data are never this "clean," and

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

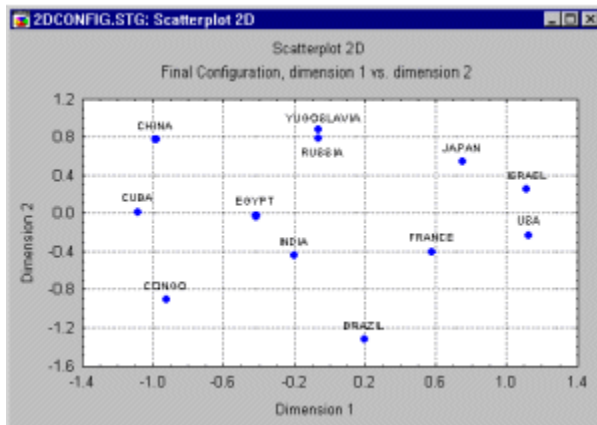
Code : SIT1303

contain a lot of noise, that is, random variability that contributes to the differences between the reproduced and observed matrix.

Interpretability of configuration. A second criterion for deciding how many dimensions to interpret is the clarity of the final configuration. Sometimes, as in our example of distances between cities, the resultant dimensions are easily interpreted. At other times, the points in the plot form a sort of "random cloud," and there is no straightforward and easy way to interpret the dimensions. In the latter case, you should try to include more or fewer dimensions and examine the resultant final configurations. Often, more interpretable solutions emerge. However, if the data points in the plot do not follow any pattern, and if the stress plot does not show any clear "elbow," then the data are most likely random "noise."

Interpreting the Dimensions

The interpretation of dimensions usually represents the final step of the analysis. As mentioned earlier, the actual orientations of the axes from the MDS analysis are arbitrary, and can be rotated in any direction. A first step is to produce scatterplots of the objects in the different two-dimensional planes.



Three-dimensional solutions can also be illustrated graphically, however, their interpretation is somewhat more complex.

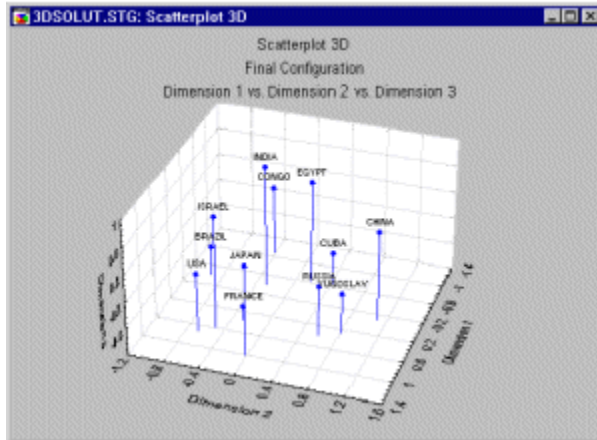
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303



In addition to "meaningful dimensions," you should also look for clusters of points or particular patterns and configurations (such as circles, manifolds, etc.). For a detailed discussion of how to interpret final configurations, see Borg and Lingoes (1987), Borg and Shye (in press), or Guttman (1968).

Applications

The "beauty" of MDS is that we can analyze any kind of distance or similarity matrix. These similarities can represent people's ratings of similarities between objects, the percent agreement between judges, the number of times a subjects fails to discriminate between stimuli, etc. For example, MDS methods used to be very popular in psychological research on person perception where similarities between trait descriptors were analyzed to uncover the underlying dimensionality of people's perceptions of traits (see, for example Rosenberg, 1977). They are also very popular in marketing research, in order to detect the number and nature of dimensions underlying the perceptions of different brands or products & Carmone, 1970).

In general, MDS methods allow the researcher to ask relatively unobtrusive questions ("how similar is brand A to brand B") and to derive from those questions underlying dimensions without the respondents ever knowing what is the researcher's real interest.

MDS and Factor Analysis

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

Even though there are similarities in the type of research questions to which these two procedures can be applied, MDS and factor analysis are fundamentally different methods. Factor analysis requires that the underlying data are distributed as multivariate normal, and that the relationships are linear. MDS imposes no such restrictions. As long as the rank-ordering of distances (or similarities) in the matrix is meaningful, MDS can be used. In terms of resultant differences, factor analysis tends to extract more factors (dimensions) than MDS; as a result, MDS often yields more readily, interpretable solutions. Most importantly, however, MDS can be applied to any kind of distances or similarities, while factor analysis requires us to first compute a correlation matrix. MDS can be based on subjects' direct assessment of similarities between stimuli, while factor analysis requires subjects to rate those stimuli on some list of attributes (for which the factor analysis is performed).

In summary, MDS methods are applicable to a wide variety of research designs because distance measures can be obtained in any number of ways (for different examples, refer to the references provided at the beginning of this section).

(MDS) refers to a group of methods that is widely used especially in behavioral, econometric, and social sciences to analyze subjective evaluations of pairwise similarities of entities, such as commercial products in a market survey. The starting point of MDS is a matrix consisting of the pairwise dissimilarities of the entities. In this thesis only distances between pattern vectors in a Euclidean space will be considered, but in MDS the dissimilarities need not be distances in the mathematically strict sense. In fact, MDS is perhaps most often used for creating a space where the entities can be represented as vectors, based on some evaluation of the dissimilarities of the entities.

The goal in this thesis is not merely to create a space which would represent the relations of the data faithfully, but also to reduce the dimensionality of the data set to a sufficiently small value to allow visual inspection of the set. The MDS methods can be used to fulfill this goal, as well.

There exists a multitude of variants of MDS with slightly different cost functions and optimization algorithms. The first MDS for metric data was developed in the 1930s (historical treatments and introductions to MDS have been provided by, for example, Kruskal and Wish, 1978; de Leeuw and Heiser, 1982; Wish and Carroll, 1982; Young, 1985), and later generalized for analyzing nonmetric data and even the common structure in several dissimilarity matrices corresponding to, for instance, evaluations made by different individuals.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

The algorithms designed for analyzing a single dissimilarity matrix, and which can thus be used for reducing the dimensionality of a data set, can be broadly divided into two basic types, metric and nonmetric MDS.

In the original *metric MDS* (Torgerson, 1952; cf. Young and Householder, 1938) the distances between the data items have been given, and a configuration of points that would give rise to the distances is sought. Often a linear projection onto a subspace obtained with PCA is used. The key idea of the method, to approximate the original set of distances with distances corresponding to a configuration of points in a Euclidean space can, however, also be used for constructing a

nonlinear projection method. If each item \mathbf{x}_k is represented with a lower-dimensional, say, two-

dimensional data vector \mathbf{x}'_k , then the goal of the projection is to optimize the representations so that the distances between the items in the two-dimensional space will be as close to the original

distances as possible. If the distance between \mathbf{x}_k and \mathbf{x}_l is denoted by $d(k,l)$ and the distance

between \mathbf{x}'_k and \mathbf{x}'_l in the two-dimensional space by $d'(k,l)$, the metric MDS tries to approximate $d(k,l)$ by $d'(k,l)$. If a square-error cost is used, the objective function to be minimized can be written as

$$E_M = \sum_{k \neq l} [d(k,l) - d'(k,l)]^2. \quad (2)$$

A perfect reproduction of the Euclidean distances may not always be the best possible goal, especially if the components of the data vectors are expressed on an ordinal scale. Then only the *rank order* of the distances between the vectors is meaningful, not the exact values. The projection should try to match the rank order of the distances in the two-dimensional output space to the rank order in the original space. The best possible rank ordering for a given configuration of points can be guaranteed by introducing a monotonically increasing function f that acts on the original distances, and always maps the distances to such values that best preserve the rank order. *Nonmetric MDS* uses such a function whereby the normalized cost function becomes

$$E_N = \frac{1}{\sum_{k \neq l} [d'(k,l)]^2} \sum_{k \neq l} [f(d(k,l)) - d'(k,l)]^2. \quad (3)$$

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

$$x_k^f$$

For any given configuration of the projected points x_k^f , f is always chosen to minimize Equation 3.

Although the nonmetric MDS was motivated by the need of treating ordinal-scale data, it can also be used of course if the inputs are presented as pattern vectors in a Euclidean space. The projection then only tries to preserve the order of the distances between the data vectors, not their absolute values. A demonstration of nonmetric MDS, applied in a dimension reduction task, is given in Figure 3.

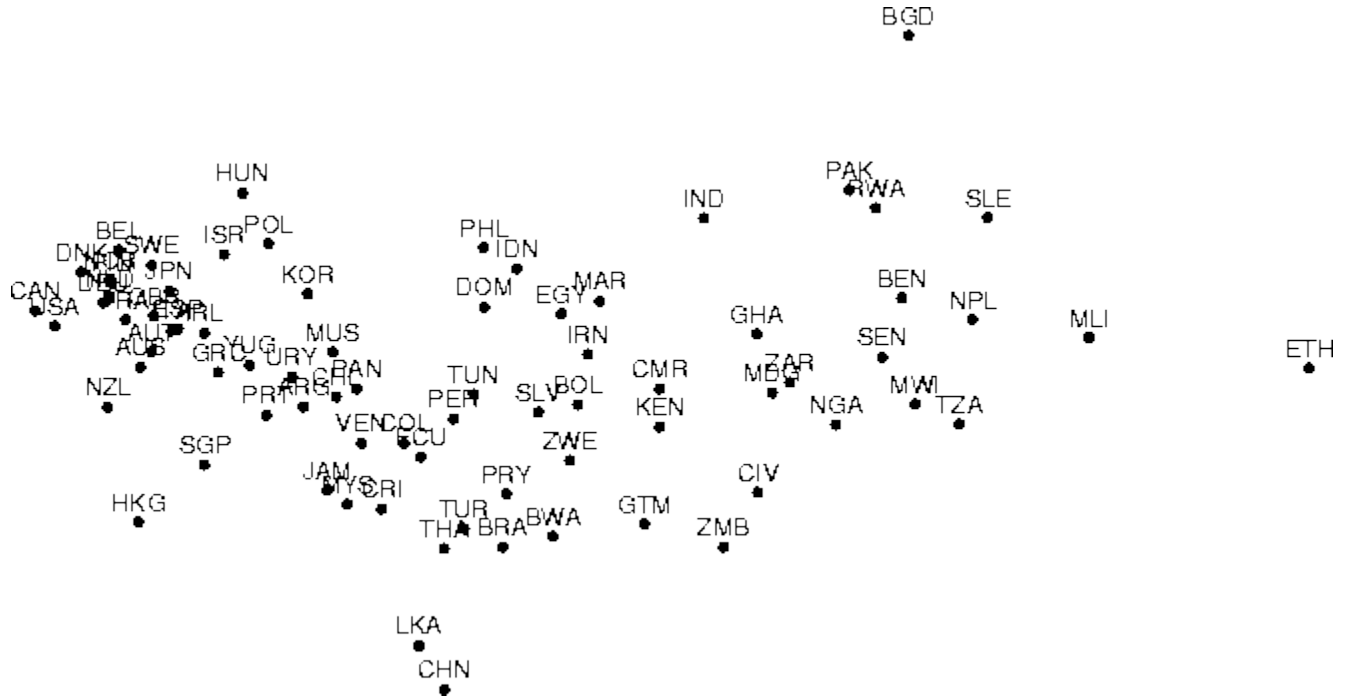


Figure: A nonlinear projection constructed using nonmetric MDS. The data set is the same as in Figure 2. Missing data values were treated by the following simple method, which has been demonstrated to produce good results at least in the pattern recognition context [Dixon, 1979]. When computing the distance between a pair of data items, only the (squared) differences between component values that are available are computed. The rest of the differences are then set to the average of the computed differences.

4.5 Sammon's Mapping

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

Sammon's mapping is closely related to the metric MDS version described above. It, too, tries to optimize a cost function that describes how well the pairwise distances in a data set are preserved. The cost function of Sammon's mapping is (omitting a constant normalizing factor)

$$E_S = \sum_{k \neq l} \frac{[d(k, l) - d'(k, l)]^2}{d(k, l)} \quad (4)$$

The only difference between Sammon's mapping and the (nonlinear) metric MDS (Eq. 2) is that the errors in distance preservation are normalized with the distance in the original space. Because of the normalization the preservation of small distances will be emphasized.

A demonstration of Sammon's mapping is presented in Figure 4.

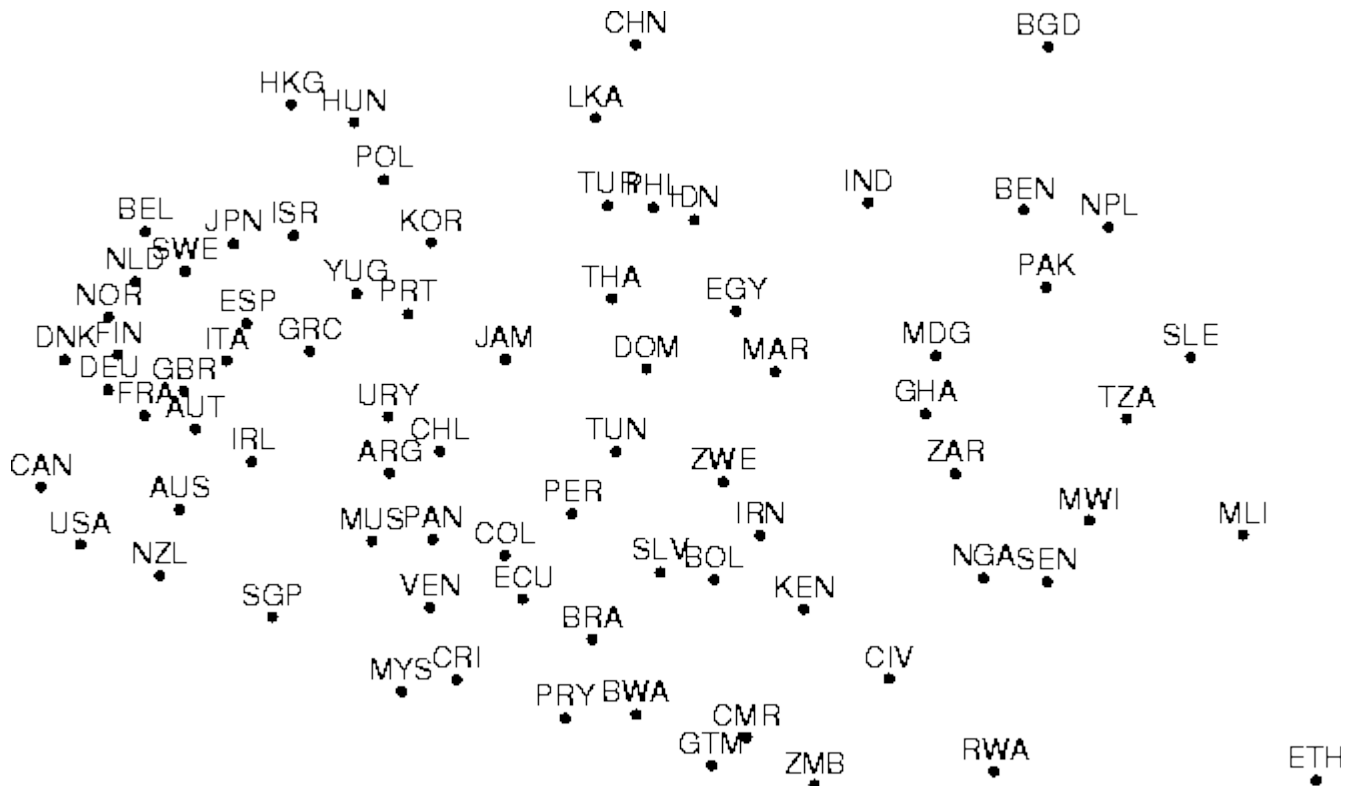


Figure: Sammon's mapping of the data set which has been projected using PCA in Figure 2 and nonmetric MDS in Figure 3. Missing data values were treated in the same manner as in forming the nonmetric MDS.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

Sammon's mapping is a non-linear mapping that maps a set of input points on a plane trying to preserve the relative distance between the input points approximately. It can be used to visualize a SOM by mapping the values of codebook vectors on a plane. Furthermore, the topological relations can be drawn using lines between neighboring neurons to enhance the net-like look. Sammon's mapping can be applied directly to data sets, but is computationally very intensive. The SOM quantizes the input data to a small number of codebook vectors, so the burden of computation is not so heavy.

4.6 Interaction Techniques

Interaction technique, user interface technique or input technique is a combination of hardware and software elements that provides a way for computer users to accomplish a single task. For example, one can go back to the previously visited page on a Web browser by either clicking a button, pressing a key, performing a mouse gesture or uttering a speech command. It is a widely used term in human-computer interaction. In particular, the term "new interaction technique" is frequently used to introduce a novel user interface design idea.

From the computer's perspective, an interaction technique involves:

- One or several input devices that capture user input,
- One or several output devices that display user feedback,
- A piece of software that:
 - interprets user input into commands the computer can understand,
 - produces user feedback based on user input and the system's state.

Consider for example the process of deleting a file using a contextual menu. This assumes the existence of a mouse (input device), a screen (output device), and a piece of code that paints a menu and updates its selection (user feedback) and sends a command to the file system when the user clicks on the "delete" item (interpretation). User feedback can be further used to confirm that the command has been invoked.

The user's view

From the user's perspective, an interaction technique is a way to perform a single computing task and can be informally expressed with user instructions or usage scenarios. For example, "to delete a file, right-click on the file you want to delete, then click on the delete item".

The designer's view

From the user interface designer's perspective, an interaction technique is a well-defined solution to a specific user interface design problem. Interaction techniques as conceptual ideas can be

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

refined, extended, modified and combined. For example, contextual menus are a solution to the problem of rapidly selecting commands. Pie menus are a radial variant of contextual menus. Marking menus combine pie menus with gesture recognition.

An interaction task is "the unit of an entry of information by the user", such as entering a piece of text, issuing a command, or specifying a 2D position. A similar concept is that of domain object, which is a piece of application data that can be manipulated by the user.

Interaction techniques are the glue between physical I/O devices and interaction tasks or domain objects. Different types of interaction techniques can be used to map a specific device to a specific domain object. For example, different gesture alphabets exist for pen-based text input.

In general, the less compatible the device is with the domain object, the more complex the interaction technique. For example, using a mouse to specify a 2D point involves a trivial interaction technique, whereas using a mouse to rotate a 3D object requires more creativity to design the technique and more lines of code to implement it.

A current trend is to avoid complex interaction techniques by matching physical devices with the task as close as possible, such as exemplified by the field of tangible computing. But this is not always a feasible solution. Furthermore, device/task incompatibilities are unavoidable in computer accessibility, where a single switch can be used to control the whole computer environment.

Interaction techniques essentially involve data entry and manipulation, and thus place greater emphasis on input than output. Output is merely used to convey affordances and provide user feedback. The use of the term *input technique* further reinforces the central role of input. Conversely, techniques that mainly involve data exploration and thus place greater emphasis on output are called visualization techniques. They are studied in the field of information visualization.

4.7 Histograms

The purpose of a histogram is to graphically summarize the distribution of a univariate data set.

The histogram graphically shows the following:

1. center (i.e., the location) of the data;
2. spread (i.e., the scale) of the data;
3. skewness of the data;
4. presence of outliers; and
5. presence of multiple modes in the data.

These features provide strong indications of the proper distributional model for the data.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

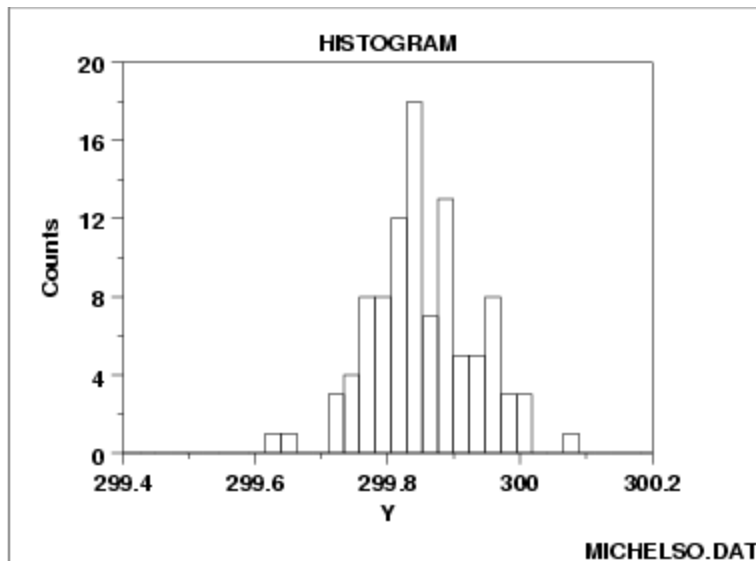
UNIT IV

Subject

Code : SIT1303

The [probability plot](#) or a [goodness-of-fit](#) test can be used to verify the distributional model.

The [examples](#) section shows the appearance of a number of common features revealed by histograms.



The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). Then for each bin, the number of points from the data set that fall into each bin is counted. That is

- Vertical axis: Frequency (i.e., counts for each bin)
- Horizontal axis: Response variable

The cumulative histogram is a variation of the histogram in which the vertical axis gives not just the counts for a single bin, but rather gives the counts for that bin plus all bins for smaller values of the response variable.

Both the histogram and cumulative histogram have an additional variant whereby the counts are replaced by the normalized counts. The names for these variants are the relative histogram and the relative cumulative histogram.

There are two common ways to normalize the counts.

1. The normalized count is the count in a class divided by the total number of observations. In this case the relative counts are normalized to sum to one (or 100 if a percentage scale is used). This is the intuitive case where the height of the histogram

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
COURSE MATERIAL

Subject Name : DATA ANALYTICS

UNIT IV

Subject

Code : SIT1303

bar represents the proportion of the data in each class.

2. The normalized count is the count in the class divided by the number of observations times the class width. For this normalization, the area (or integral) under the histogram is equal to one. From a probabilistic point of view, this normalization results in a relative histogram that is most akin to the probability density function and a relative cumulative histogram that is most akin to the cumulative distribution function. If you want to overlay a probability density or cumulative distribution function on top of the histogram, use this normalization. Although this normalization is less intuitive (relative frequencies greater than 1 are quite permissible), it is the appropriate normalization if you are using the histogram to model a probability density function.

4.8 Spectral Analysis

One of the most widely used methods for analyzing time series in geophysics, oceanography, atmospheric science, astronomy...

Spectral analysis is one of several statistical techniques necessary for characterizing and analyzing sequenced data. Sequenced data are observations that have been taken in one, two, or three dimensional space, and/or time. Examples might be observations of population density along a road, or of rainfall over an area, or of daily births at a hospital. One important limitation is that the observations be equally spaced in order that the analysis proceed efficiently. Spectral analysis refers to the decomposition of a sequence into oscillations of different lengths or scales. By this process, the observations in what is called the data domain are converted into the spectral domain. The reasons for doing this are that: (a) some forms of manipulation are easier in the spectral domain; and (b) the revealed scales are necessary statistical descriptors of the data and may suggest important factors that affect or produce such data. The following will provide brief descriptions of: (a) Fourier analysis and its use in manipulating data that are assumed to be periodic; (b) relevant statistics; and (c) one approach to spectral analysis of non-periodic data including an example.