Business Anaytics

M. Tech QROR – 2nd yr (2024)

DATA VISUALIZATION

(Visual Analytics)

Prasun Das

SQC & OR Unit, ISI, Kolkata

prasun@isical.ac.in

Outline

- Data Visualization – WHY?
- Types and Basis of Selection
- Techniques
- Application Tools vs. Techniques: Snapshot

“A picture is worth a thousand words”

Copyright@ 2024 ISI SQCOR Kolkata

Why Visualization ?

To look at the patterns of data followed by understanding and analysing of it, Visualization of Data is extremely important both as a part of Data Pre-processing and post-analysis checking.

Data aggregation, summarization and visualization are some of the main pillars supporting Descriptive Analytics, the core components of a data science project.

In today’s age of AI, Data Visualization has been a powerful tool and has been widely adopted by organizations owing to its effectiveness in abstracting out the right information, understanding and interpreting the results with more clarity.

> ***“The greatest value of a picture is when it forces us to notice what we never expected to see.”***
>
> *— John Tukey*

Copyright@ 2024 ISI SQCOR Kolkata

Types of Visualization

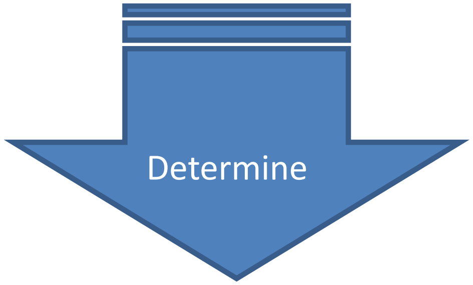
"Effective data visualization is both an art as well as a science"

Copyright@ 2024 ISI SQCOR Kolkata

How to select the right Technique ???

Four basic presentation types:

- Comparison
- Composition
- Distribution
- Relationship

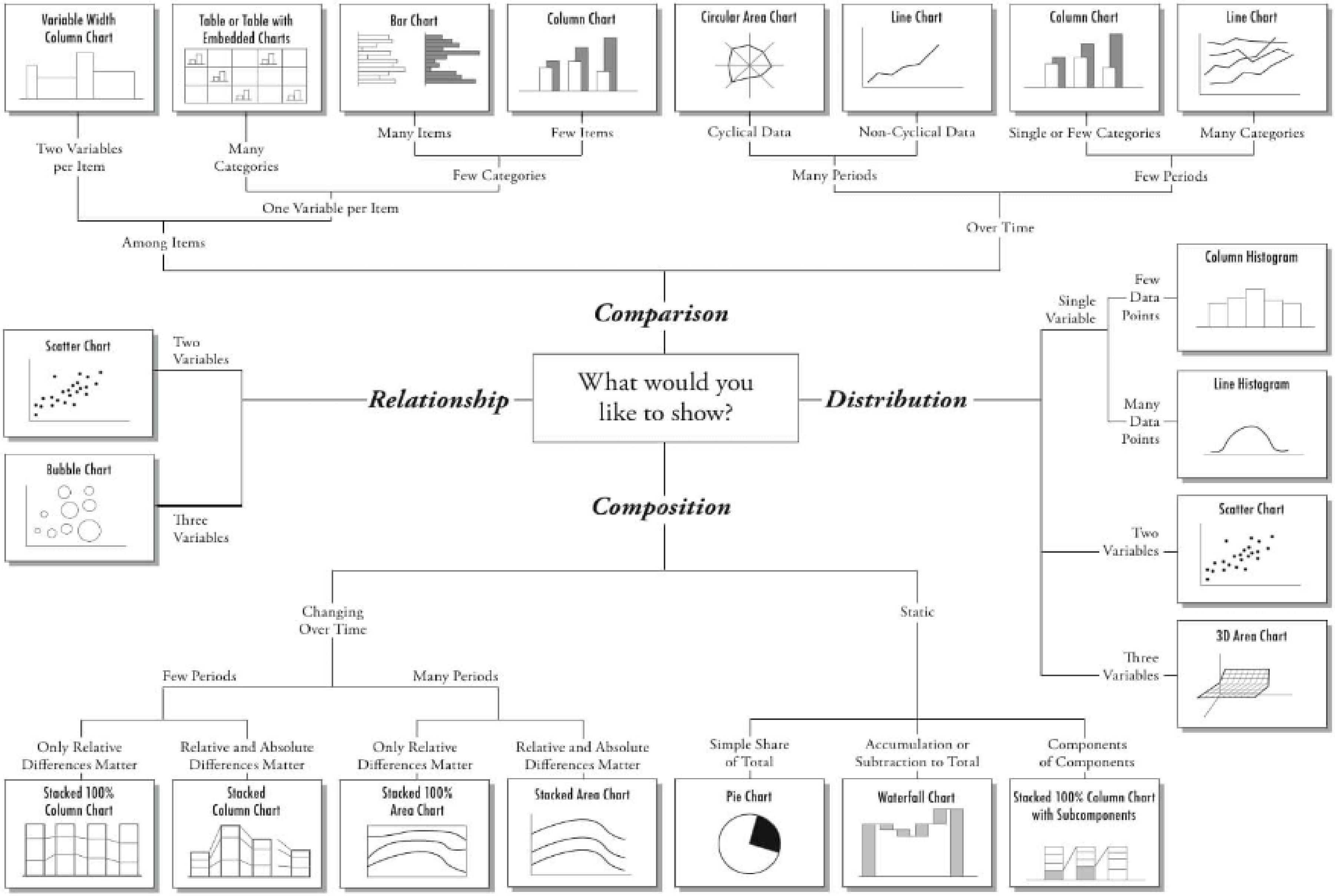


Determine

- How many **variables** need to show in single chart??
- How many **data points** needed to be displayed for each variable??
- Will values be displayed over a **period of time** or **among items or groups**??

Copyright@ 2024 ISI SQCOR Kolkata

Chart Suggestions—A Thought-Starter



List of Few Visualization Techniques

Sl. No.	Name		Sl.No.	Name
1.	Bar Chart / Column Chart		15.	Pareto Chart
2.	Line Chart / Area Chart		16.	Stem-and-Leaf Plot
3.	Sunburst Chart		17.	Box Plot
4.	Pie Chart / Doughnut Chart		18.	Parallel Coordinates Plot
5.	Tree Map		19.	Funnel Chart
6.	Bubble Chart		20.	Surface Chart
7.	Radar Plot		21.	Geographical Mapping
8.	Waterfall Chart		22.	Chernoff Faces
9.	Stock Chart		23.	Stick Figures
10.	Scatter Plot/ Matrix Plot		24.	Heat Map
11.	Correlogram		25.	Mosaic Plot
12.	Marginal Plot		26.	Sieve Plot
13.	Histogram		27.	Dendogram
14.	Dot Plot		28.	Silhouette Plot

Copyright@ 2024 ISI SQCOR Kolkata

Data Visualization Tool: MS Excel

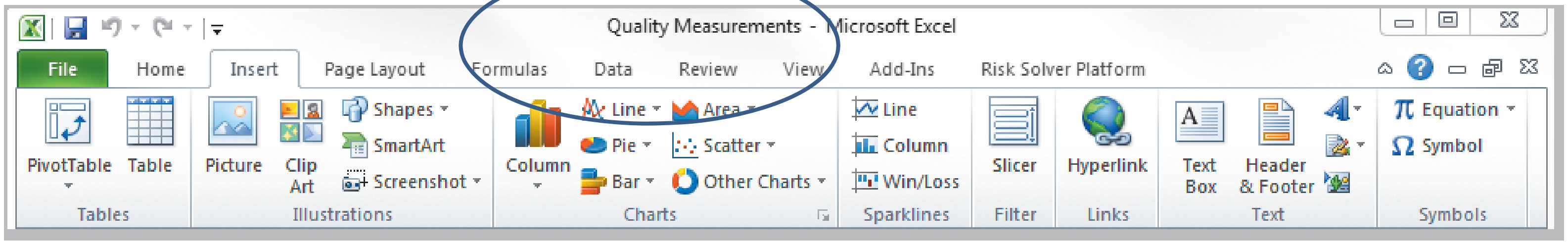
Creating Charts in Microsoft Excel

- Microsoft Excel supports statistical analysis in two ways:

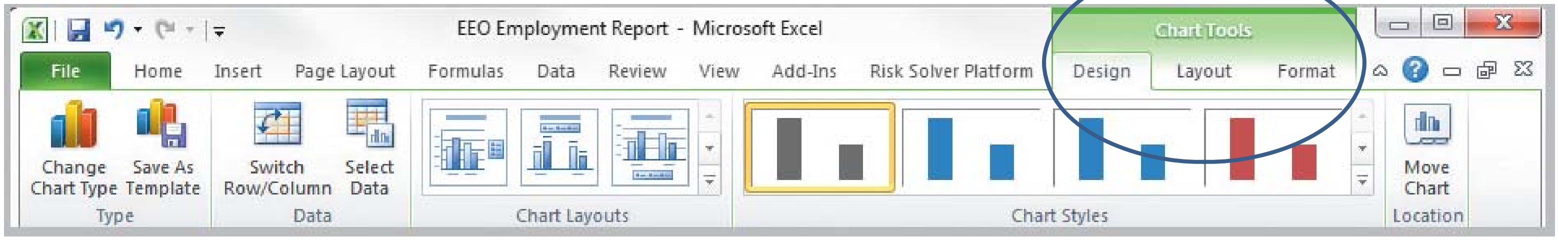
  1. Statistical functions

  2. *Analysis Toolpak* add-on

- Select the *insert* tab.
- Highlight the data.
- Click on chart type, then subtype.



- Use *chart tools* to customize.

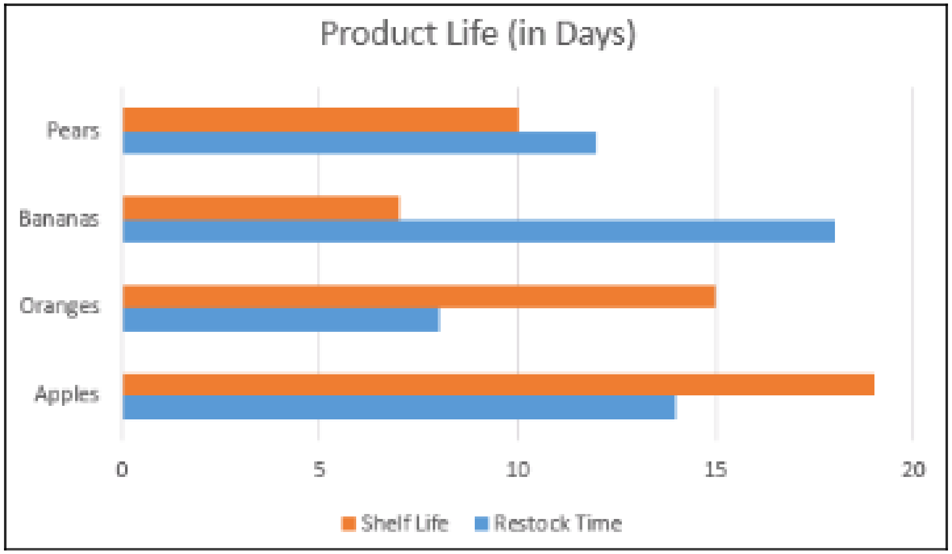


Bar Chart

Horizontal bars with the axis values for the bars displayed on the bottom of the graph.

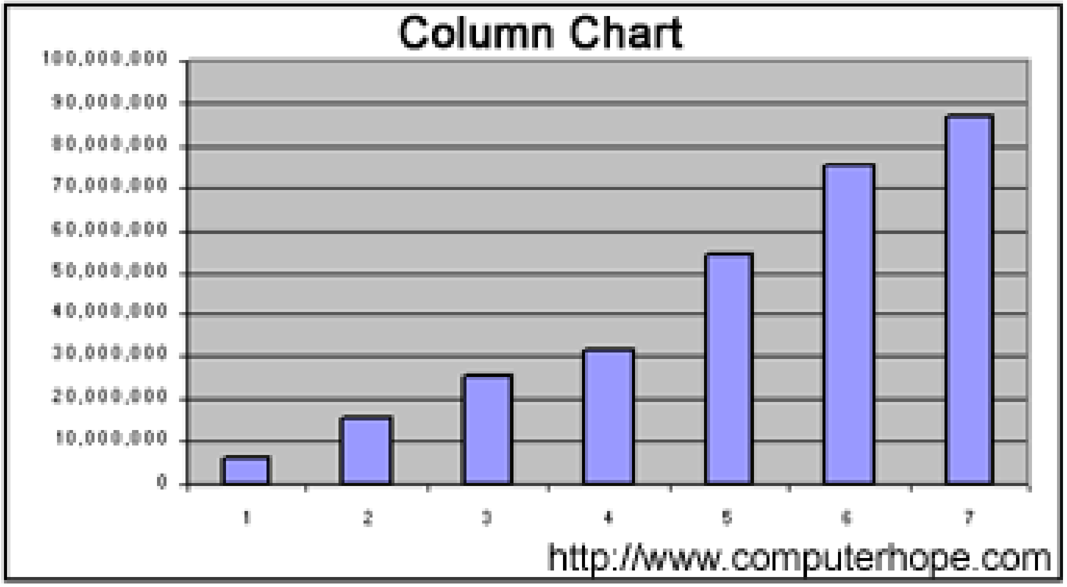
Use of Bar chart:

- To compare values across categories.
- The category text is long and difficult to display in a column chart.
- To show duration in a graph.



Column Chart

Column charts display vertical bars with the values of axis being displayed on the left side of the chart. The following column chart displays the no. of visitors that Computer Hope has received between the years of 2000 and 2006. As can be seen in this example, you can immediately see a gentle increase of users without reading any data.



Copyright@ 2024 ISI SQCOR Kolkata

Stack Column Chart

This chart allows part-to-whole comparisons over time, or across categories. In a stacked column chart, data series are stacked one on top of the other in vertical columns. Stacked column charts can show changes over time because it's easy to compare total column lengths.

Pros

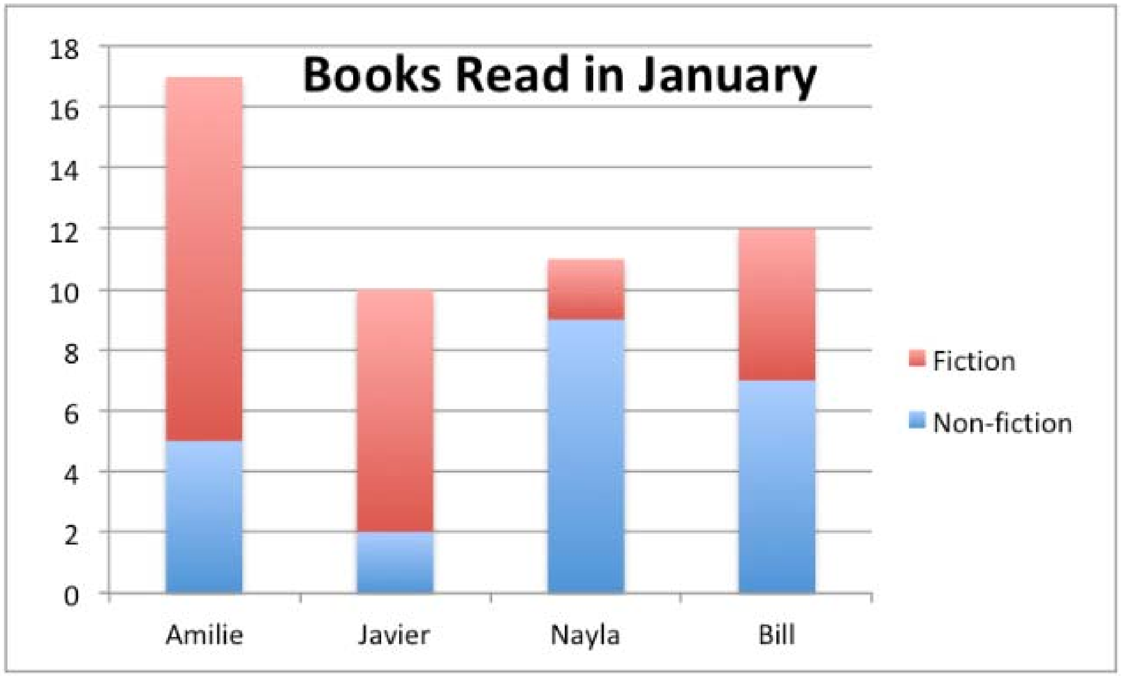
- Multiple categories and data series in compact space
- Can show change over time

Cons

- Difficult to compare all but the first series
- Become visually complex as categories or series are added

Copyright@ 2024 ISI SQCOR Kolkata

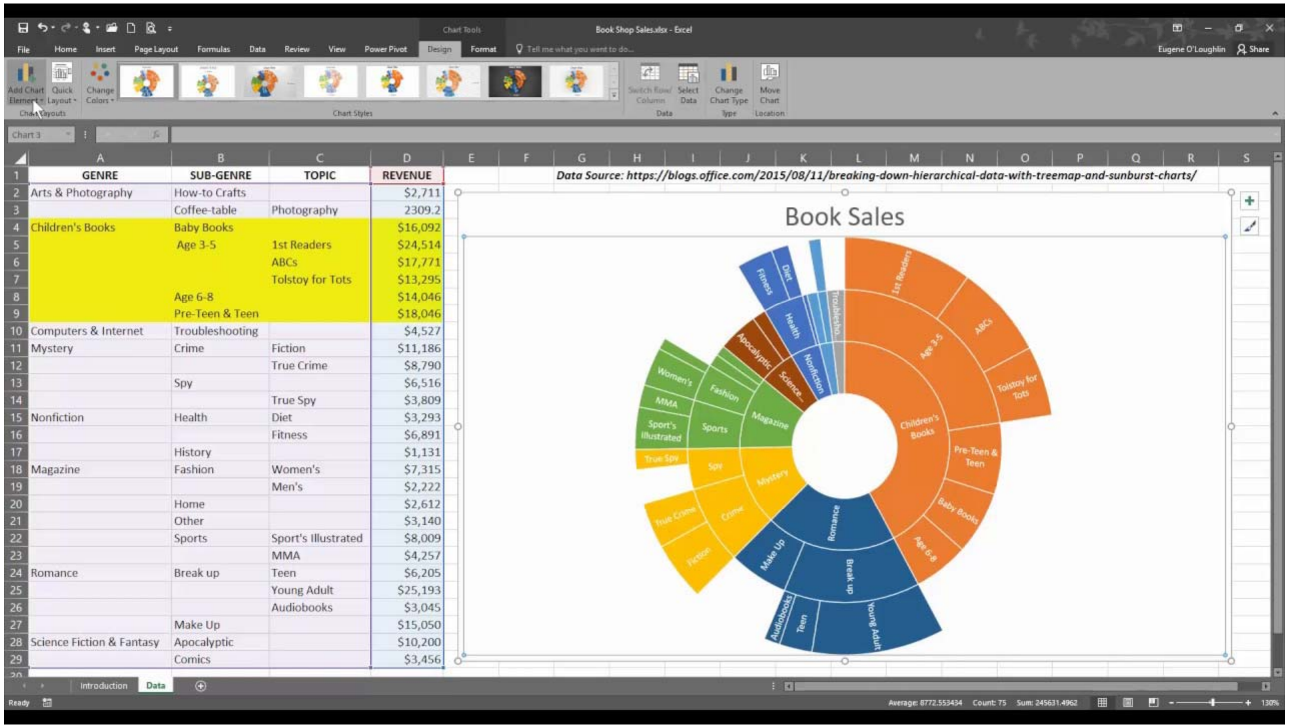
Stack Column Chart



Copyright@ 2024 ISI SQCOR Kolkata

Sunburst Chart

A sunburst chart, also known as Radial Treemap / Ring Chart / multi-level Pie Chart, is often implemented as a visual aid for hierarchical data structures. A sunburst chart highlights hierarchy through the use of concentric rings. Every ring is a level of the hierarchy.



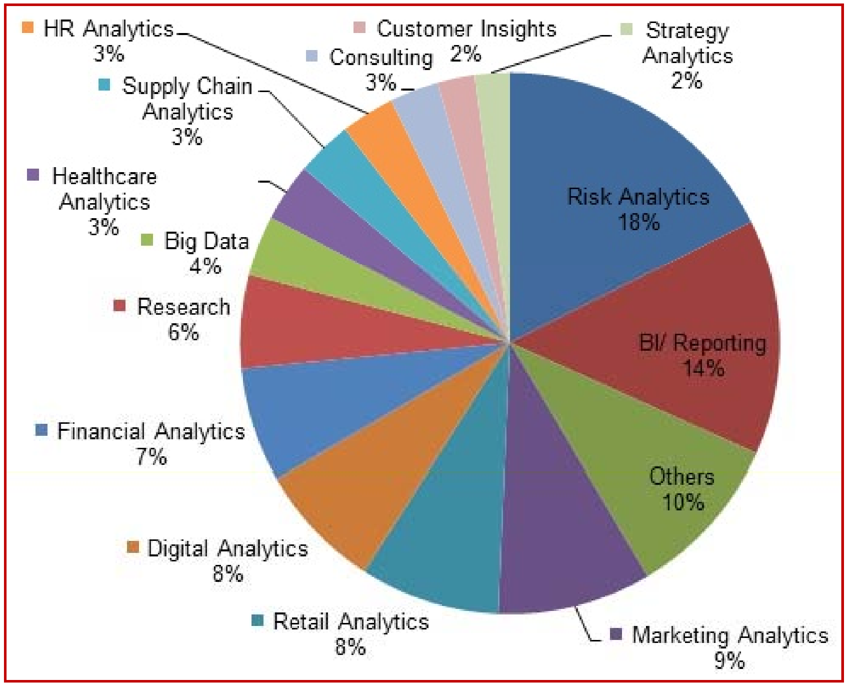
	A	B	C	D
1	GENRE	SUB-GENRE	TOPIC	REVENUE
2	Arts & Photography	How-to Crafts		$2,711
3		Coffee-table	Photography	2309.2
4	Children's Books	Baby Books		$16,092
5		Age 3-5	1st Readers	$24,514
6			ABCs	$17,771
7			Tolstoy for Tots	$13,295
8		Age 6-8		$14,046
9		Pre-Teen & Teen		$18,046
10	Computers & Internet	Troubleshooting		$4,527
11	Mystery	Crime	Fiction	$11,186
12			True Crime	$8,790
13		Spy		$6,516
14			True Spy	$3,809
15	Nonfiction	Health	Diet	$3,293
16			Fitness	$6,891
17		History		$1,131
18	Magazine	Fashion	Women's	$7,315
19			Men's	$2,222
20		Home		$2,612
21		Other		$3,140
22		Sports	Sport's Illustrated	$8,009
23			MMA	$4,257
24	Romance	Break up	Teen	$6,205
25			Young Adult	$25,193
26			Audiobooks	$3,045
27		Make Up		$15,050
28	Science Fiction & Fantasy	Apocalyptic		$10,200
29		Comics		$3,456

Data Source: https://blogs.office.com/2015/08/11/breaking-down-hierarchical-data-with-treemap-and-sunburst-charts/

Copyright@ 2024 ISI SQCOR Kolkata

Pie Chart

A Pie Chart is a type of circular graph where the pieces of the graph are proportional to the fraction of the whole in each category. In other words, each slice of the pie is relative to the size/portions of that category in the group as a whole where the entire “pie” represents 100 percent.



Job opening in Analytics (2012)

Doughnut Chart

A doughnut chart is a chart whose visualization function is similar to pie charts. The categories represented in this chart are parts, and together they express the whole data in the chart. We can only use the data in rows or columns in creating a doughnut chart in Excel.

Donut charts are used **to show the proportions of categorical data, with the size of each piece representing the proportion of each category**. Each of the donut arcs has the same width, but a different length.

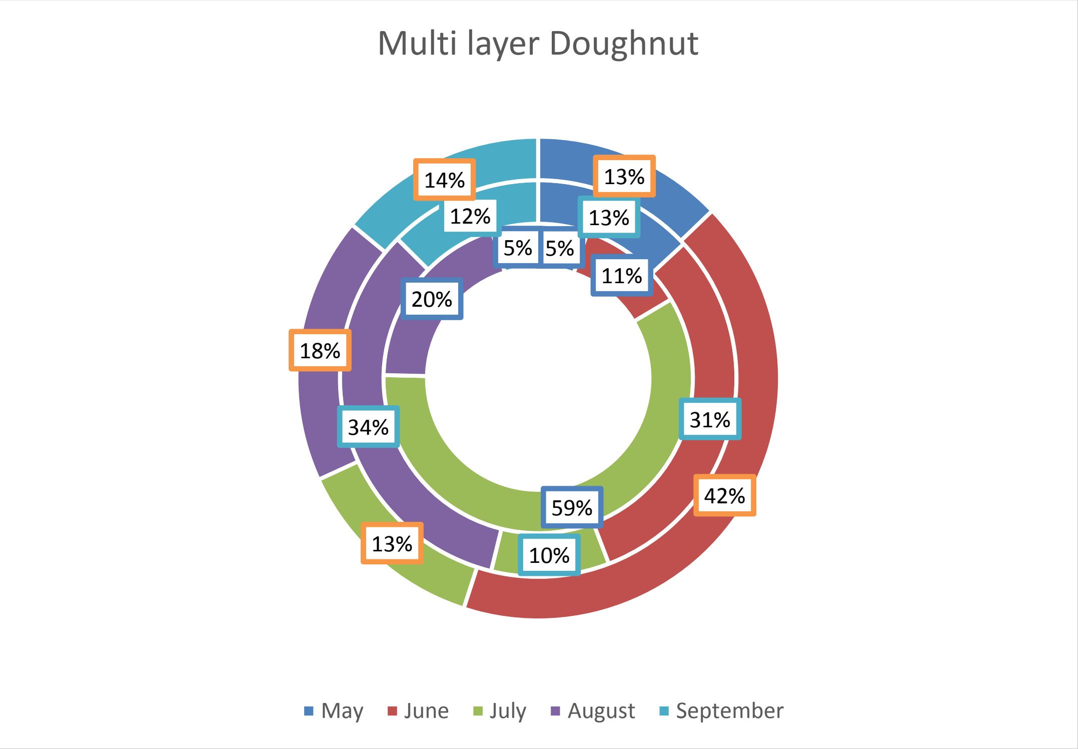
Donut charts can make it easier for users to compare individual dimensions or categories to the larger whole, as compared to pie chart.

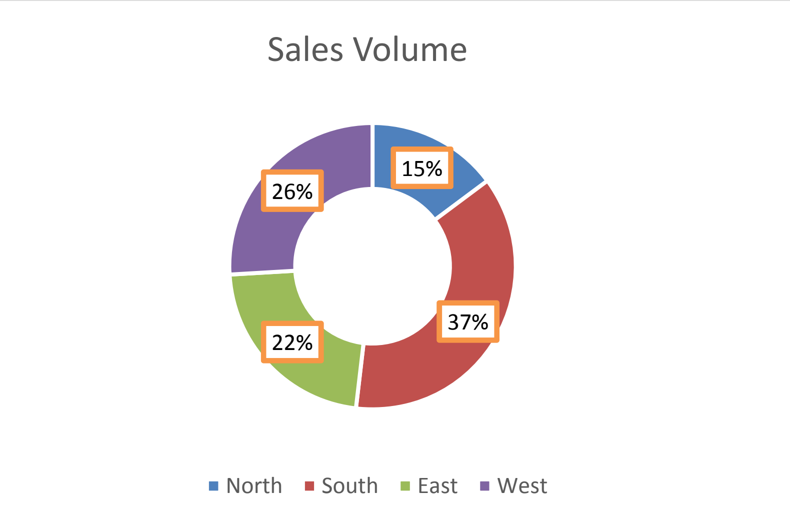
Doughnut Chart

Sales Report

Month	Retail	Online	Web Booking
May	24490	8206	32506
June	53191	19430	107237
July	278230	6018	33400
August	93484	21013	45374
September	22874	7863	35609
Total	**472269**	**62530**	**254126**

Zone	Sales Volume
North	20000
South	50000
East	30000
West	35000

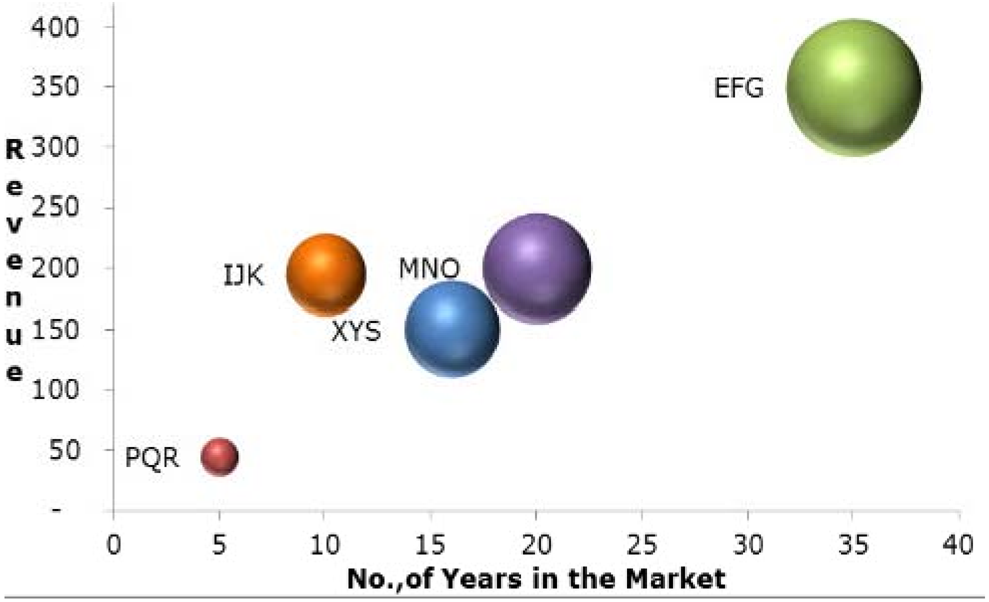




Bubble Chart

Bubble chart is used to represent three sets of data in a graphical way. Out of those three data triplet, it shows two axes of the chart in a series of X-Y coordinates and a third set shows the data points. With the help of a bubble chart, we can show the relationship between different datasets.

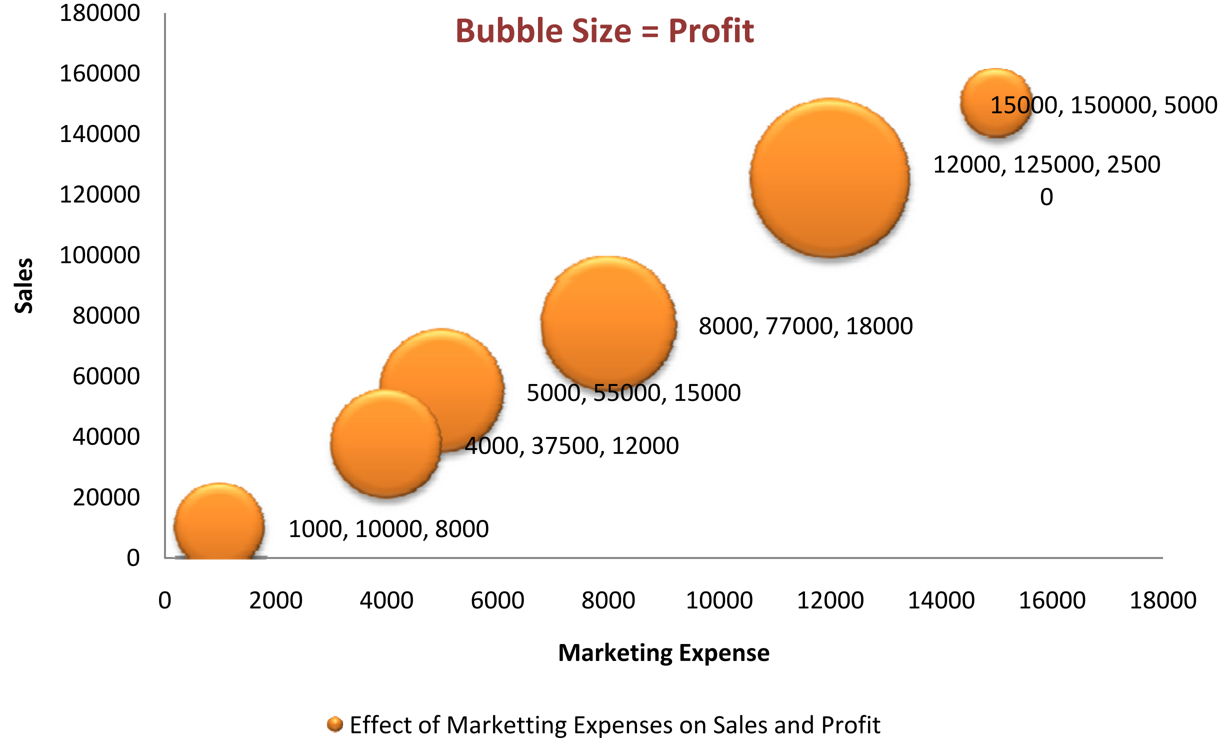
The following graph describe three basic information (**No. of years in Market, Revenue and Market Value**) of five companies in X-Y plane. First two information are represented in X and Y axes respectively whereas, **market value** is represented by the bubble size.



Copyright@ 2024 ISI SQCOR Kolkata

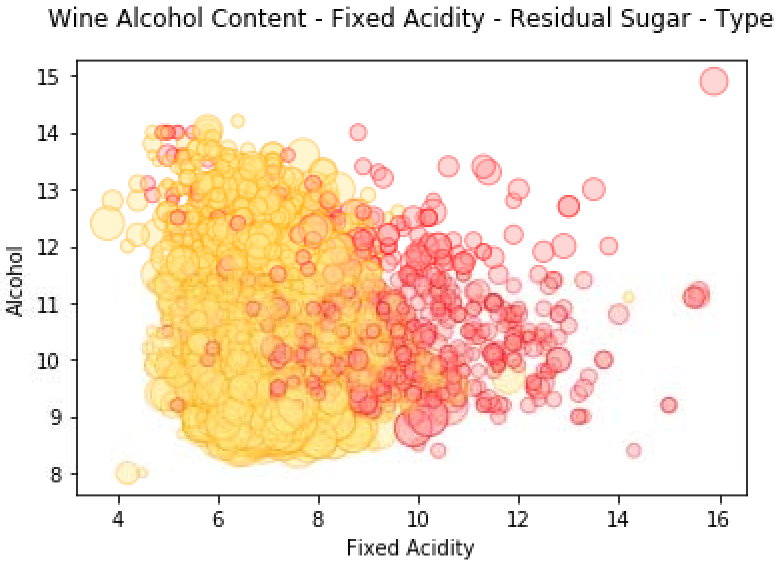
Bubble Chart

Product	Marketing Expenses	Sales	Profit
A	15000	150000	5000
B	12000	125000	25000
C	1000	10000	8000
D	5000	55000	15000
E	8000	77000	18000
F	4000	37500	12000

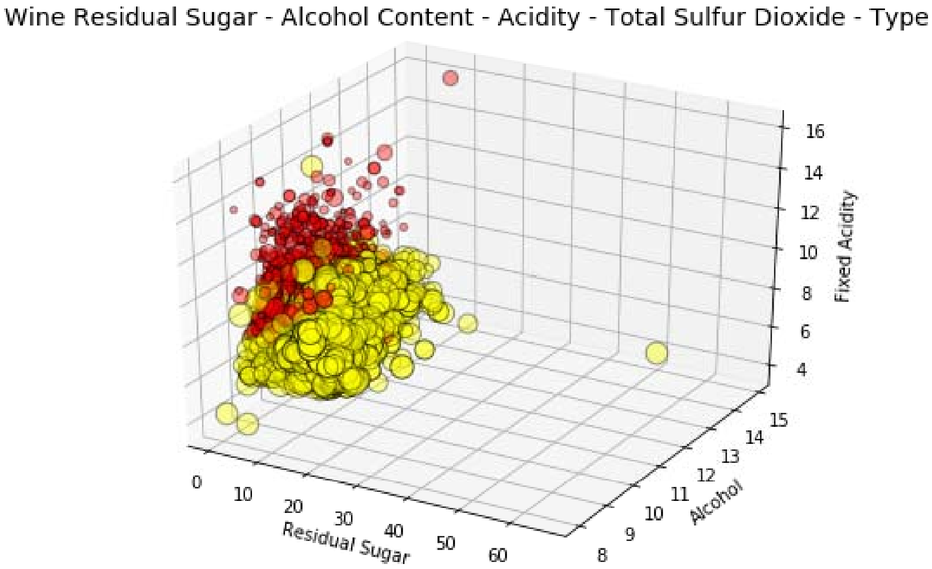


Copyright@ 2024 ISI SQCOR Kolkata

Bubble Chart – 4D & 5D



Bubble Size = Residual Sugar



Bubble Size = Total Sulphur Dioxide

WINE QUALITY DATA SET:
https://archive.ics.uci.edu/dataset/186/wine+quality

Copyright@ 2024 ISI SQCOR Kolkata

Bubble Chart

Advantages:

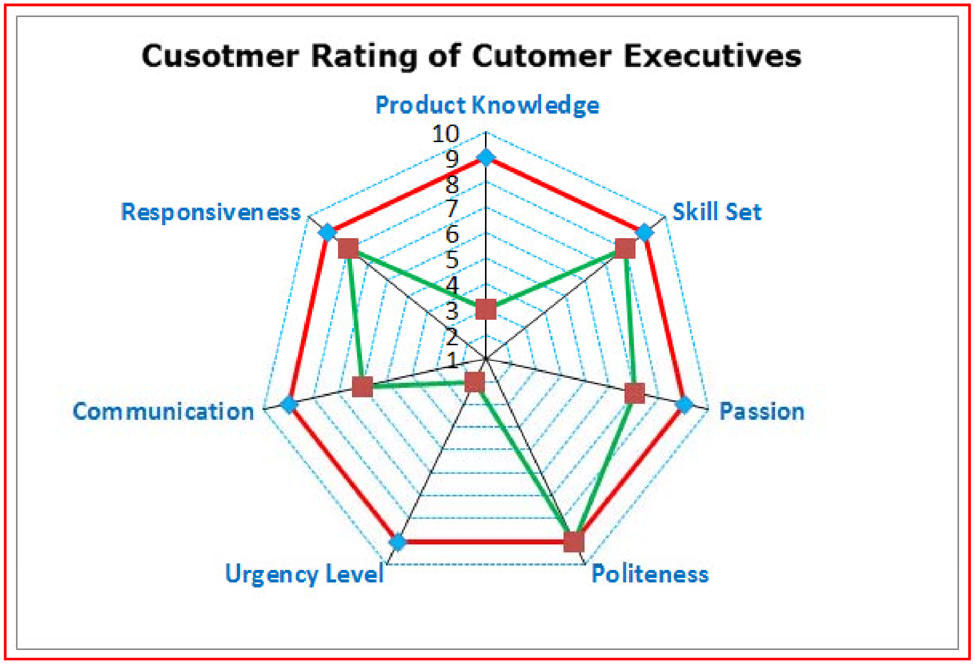
- The bubble chart is a better chart when it is applied for a 3-dimensional data set, even better for 4D & 5D.
- Eye-catching bubble size will attract the investigator.
- Visually appearance is better than the table format.

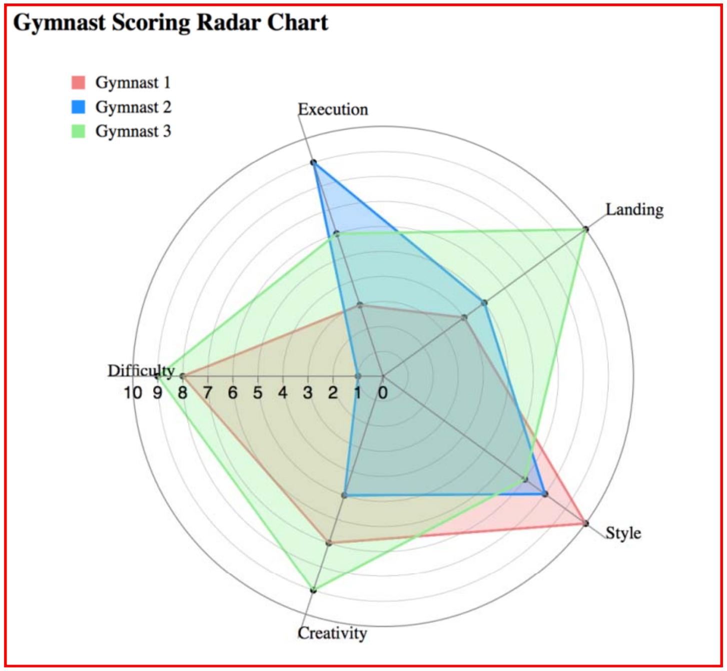
Disadvantages:

- May be difficult for a first time user to understand very quickly.
- Sometimes it gets confused with the bubble size.

Radar Chart

Radar chart is also known as the Spider chart / Web chart / Polar chart. It is used to demonstrate data in a 2D-chart for three or more than two data series, the axes start on the same point in radar chart. This chart is used to do the comparison between more than two variables/characteristics/individuals. The relative positions, on an average, represent distances from the target set. The angles formed by the arms of radar (polygon) are usually uninformative.



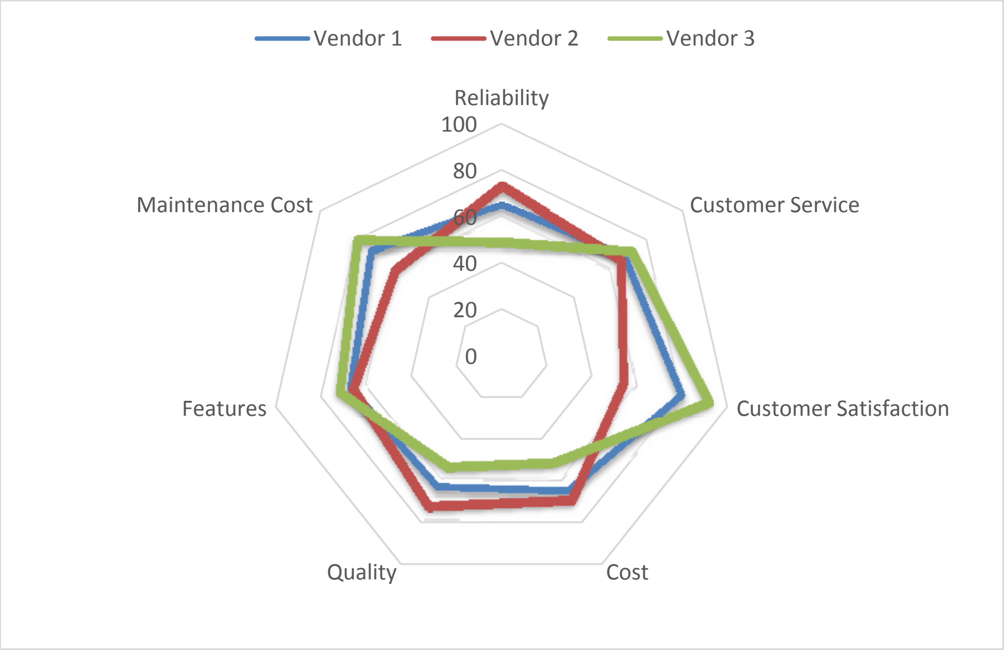


Copyright@ 2024 ISI SQCOR Kolkata

Radar Chart

Telephony Service Review Score (Rating:1-100)

	Vendor 1	Vendor 2	Vendor 3
Reliability	65	73	49
Customer Service	68	66	72
Customer Satisfaction	80	54	92
Cost	66	70	52
Quality	64	73	54
Features	67	66	72
Maintenance Cost	72	59	80

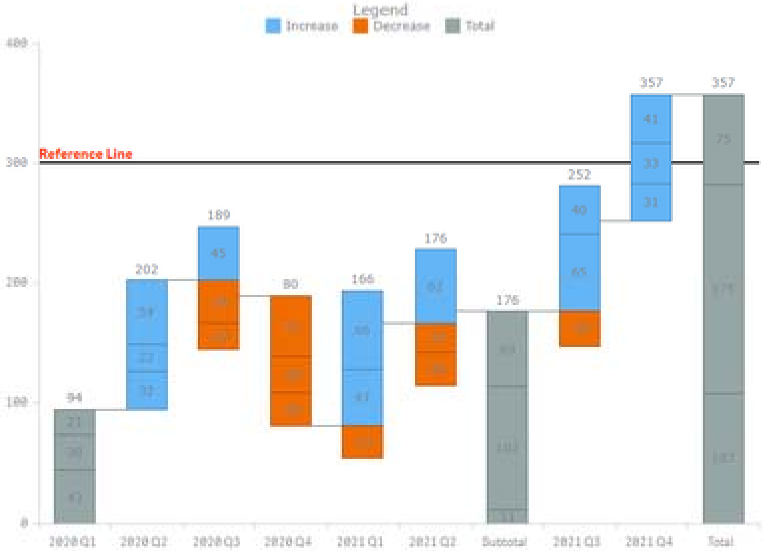


Copyright@ 2024 ISI SQCOR Kolkata

Waterfall Chart

A Waterfall chart **helps in understanding the cumulative effect of sequentially introduced positive or negative values**. A typical Waterfall chart is used to show how an initial value is increased and decreased by a series of intermediate values, leading to a final value.

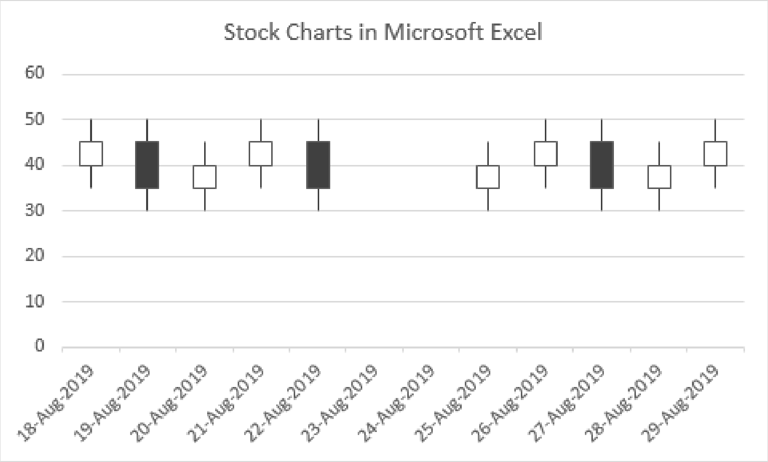
A Waterfall chart is actually a special type of column chart. It is normally used to demonstrate how the starting position either increases or decreases through a series of changes. The first and the last columns in a typical waterfall chart represent total values. The intermediate columns appear to float, and show positive or negative change from one period to another, ending up in the final total value.

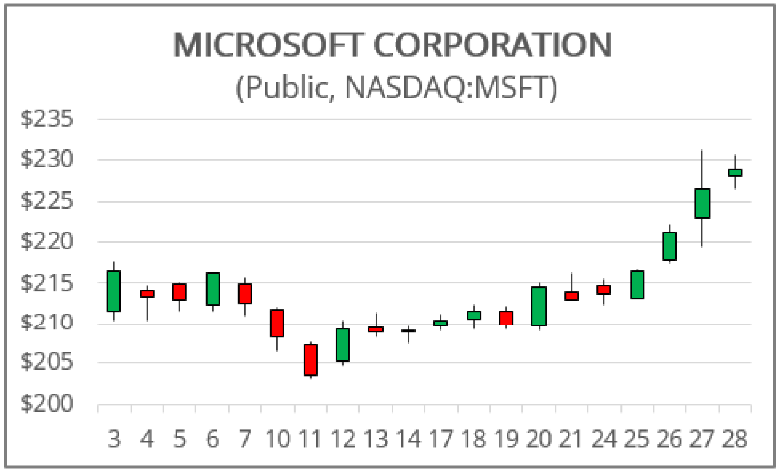


(Candlestick) Stock Chart

Stock charts, as the name indicates are mostly useful to show fluctuations in stock prices. However, these charts are useful to show fluctuations in other data also, such as daily rainfall or annual temperatures. If you use a Stock chart to display the fluctuation of stock prices, you can also incorporate the trading volume.

For Stock charts, the data needs to be in a specific order. For example, to create a simple open-high-low-close Stock chart, arrange your data with opening price, high price, low price, and closing price entered as column headings, in that order.

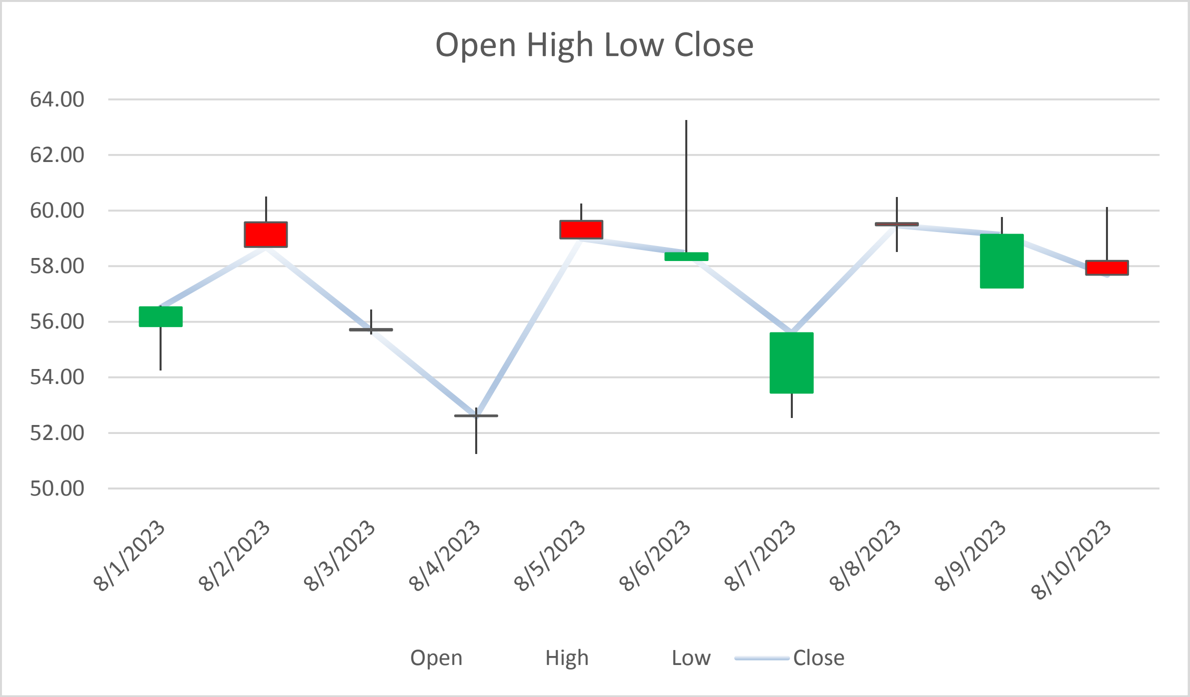




Stock Chart

Stock Details

Date	Volume	Open	High	Low	Close
8/1/2023	423454	55.84	56.58	54.25	56.52
8/2/2023	534535	59.58	60.50	59.12	58.69
8/3/2023	464255	55.74	56.44	55.55	55.69
8/4/2023	462123	52.63	52.90	51.25	52.60
8/5/2023	724552	59.63	60.25	60.00	59.00
8/6/2023	452426	58.22	63.25	61.25	58.47
8/7/2023	623562	53.45	55.25	52.55	55.59
8/8/2023	245621	59.55	60.48	58.50	59.47
8/9/2023	631531	57.23	59.77	57.80	59.13
8/10/2023	222455	58.19	60.13	5	[illegible]



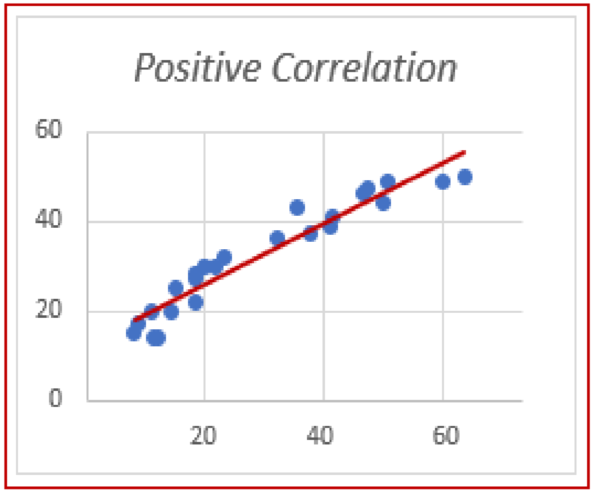
Copyright@ 2024 ISI SQCOR Kolkata

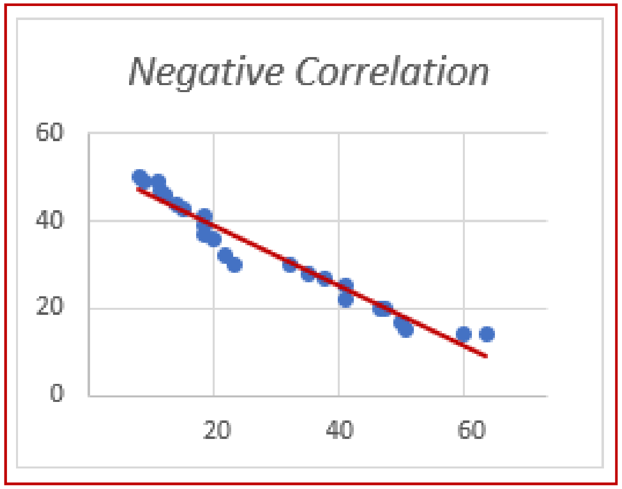
Scatter Plot

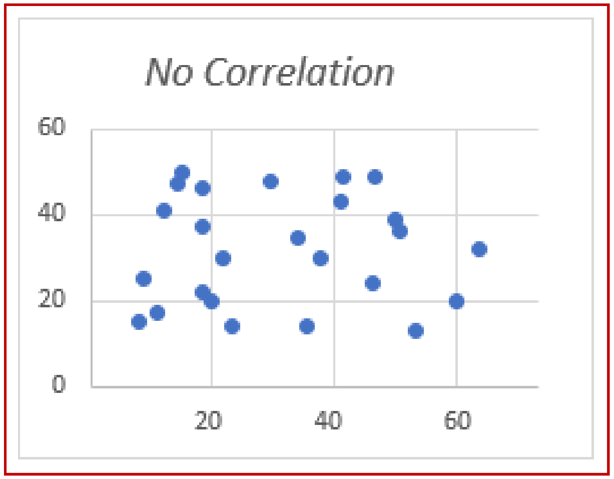
A **scatter plot** (also called *XY graph*, or *scatter diagram*) is a two-dimensional chart that shows the ***nature of relationship*** between two variables based on their numerical values in pairs.

Typically, the independent variable is on the x-axis, and the dependent variable on the y-axis. The chart displays paired values of two variables w.r.t (x,y) coordinates.

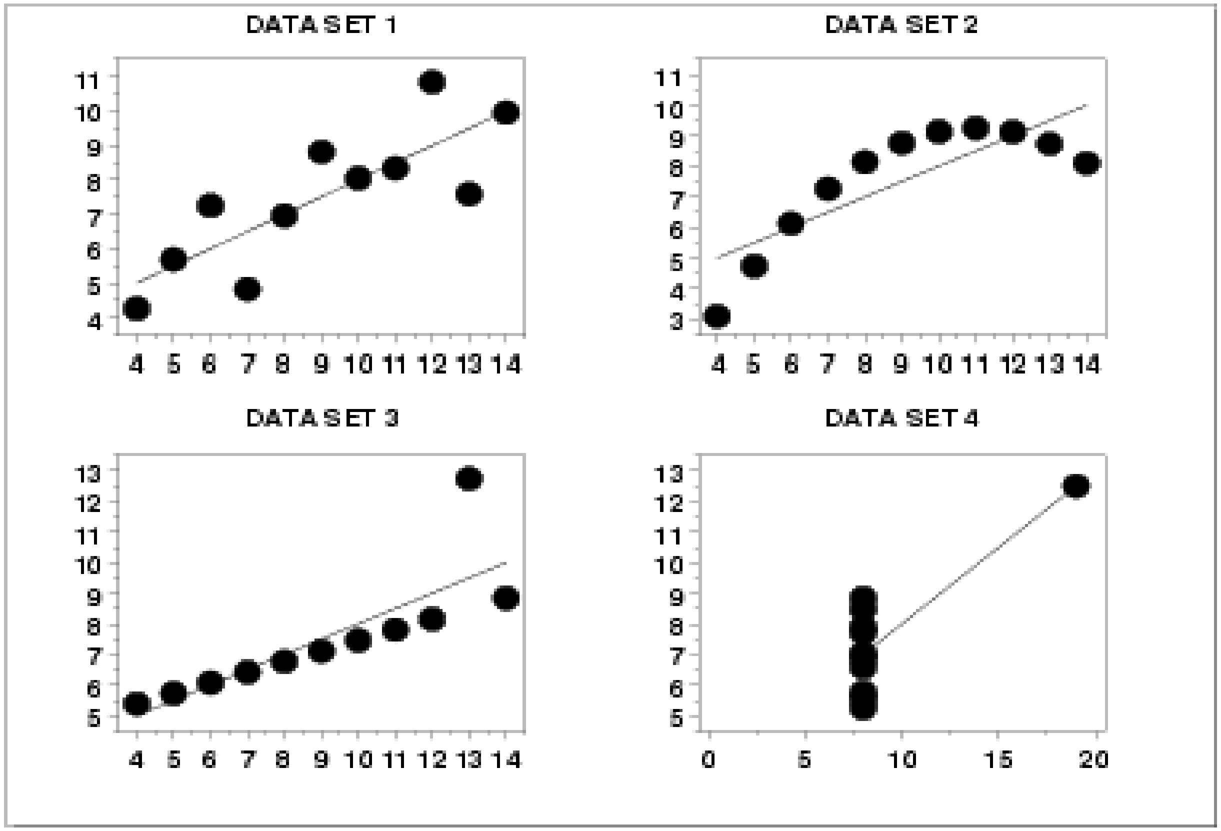
The main purpose of a scatter plot is to show how strong the **linear relationship (correlation)** is between the two variables.







SCATTER PLOT



(X_i, Y_i), $i = 1 \ldots 11$

Average of $X = 9.0$

Average of $Y = 7.5$

Intercept $(c) = 3$

Slope $(m) = 0.5$

Correlation $(r) = 0.816$

ANSCOMBE, FRANK J. (1973). "GRAPHS IN STATISTICAL ANALYSIS“, AMERICAN STATISTICIAN, VOL. 27, FEBRUARY 1973.

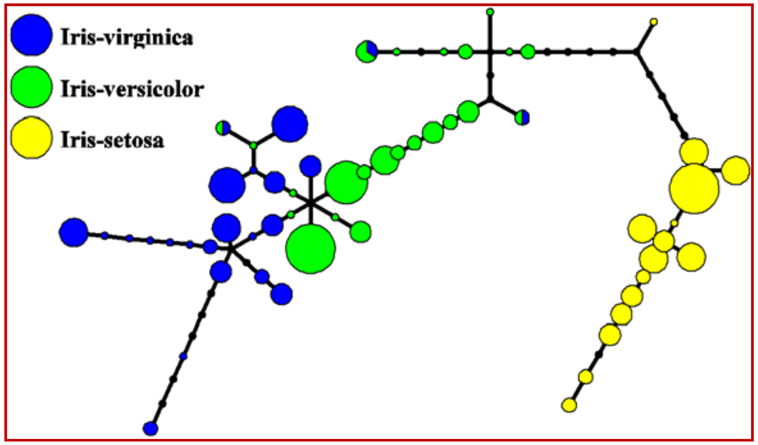
Copyright@ 2024 ISI SQCOR Kolkata

Scatter plot Matrix: 4D+

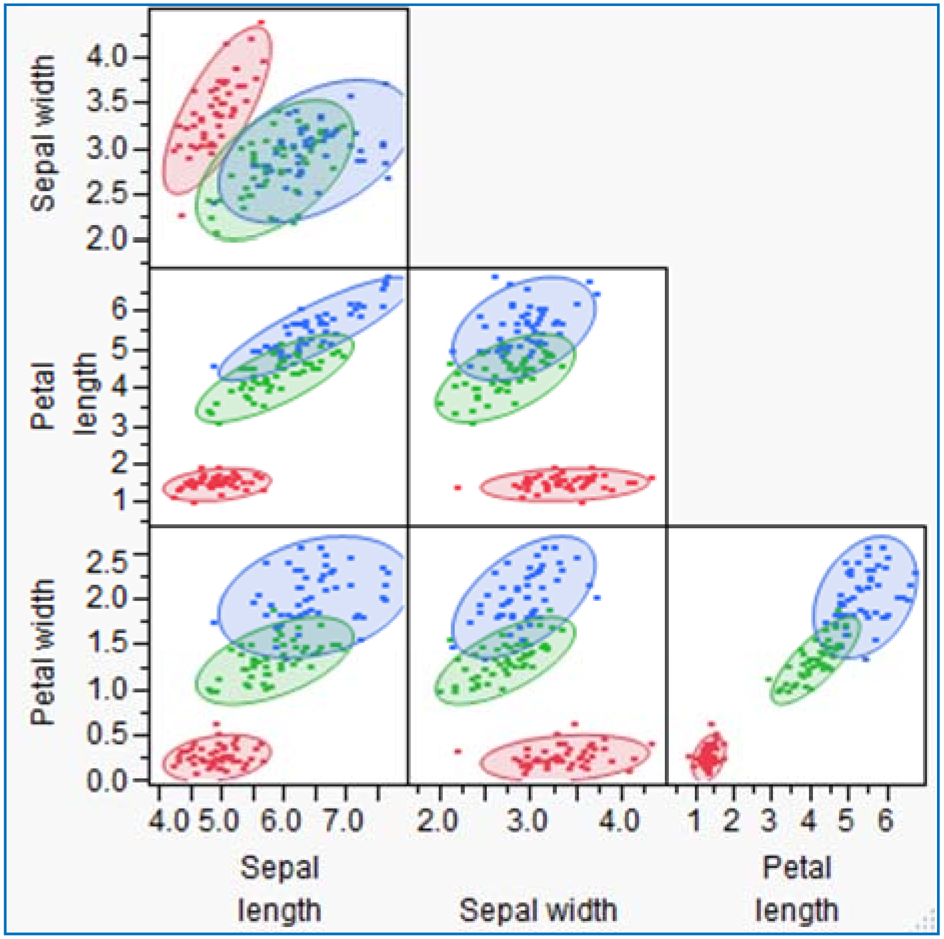
Association of multiple (>2) characteristics of an individual.

Individual: IRIS Flower

Characteristics: Sepal width, Petal width, Sepal length *and* Petal length



Metro Map

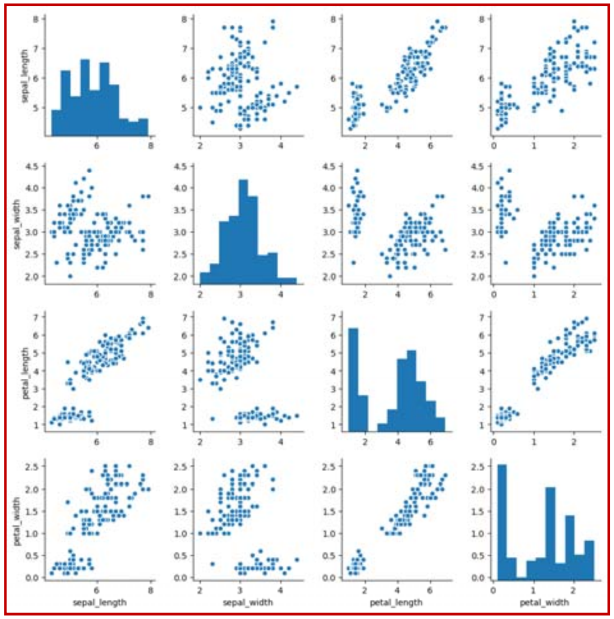


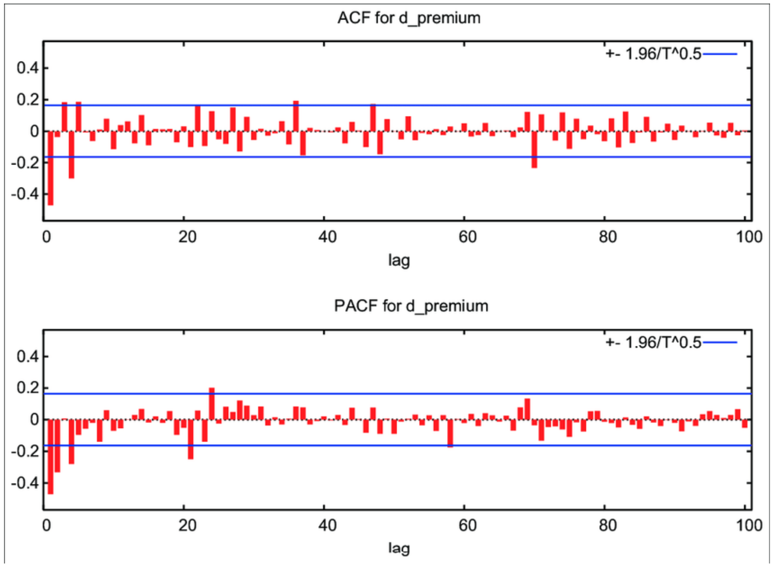
Copyright@ 2024 ISI SQCOR Kolkata

Correlogram (2 concepts)

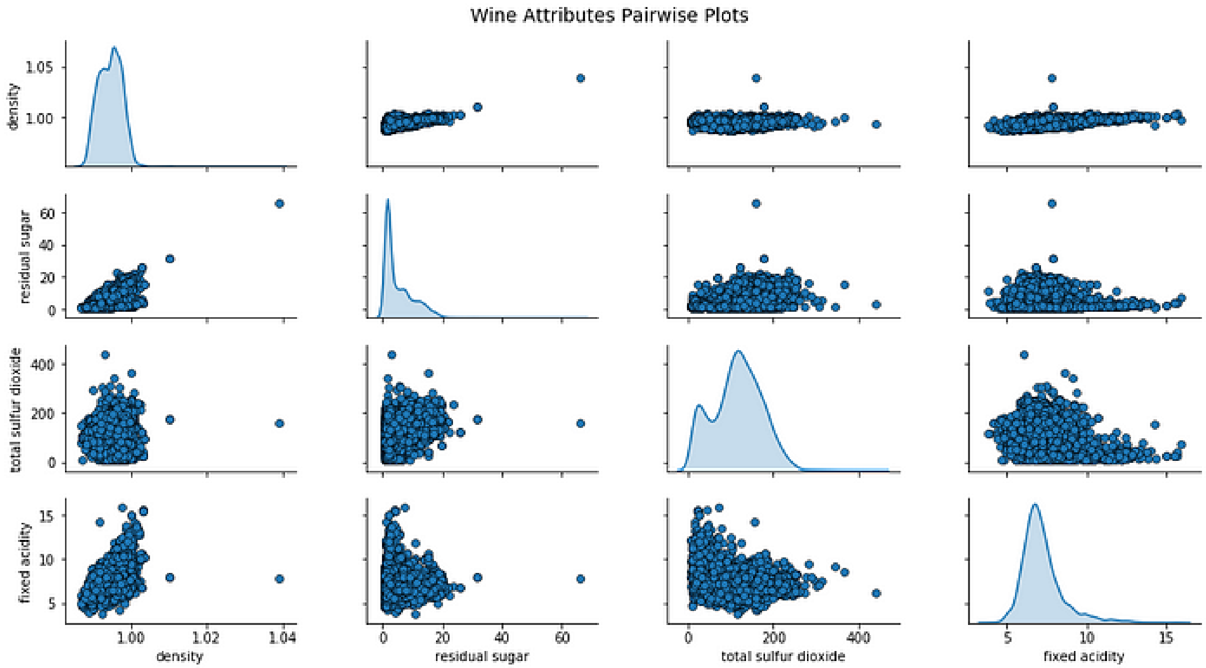
The correlogram represents the correlations for all pairs of variables. **Positive correlations are displayed in blue and negative correlations in red**. The intensity of the color is proportional to the correlation coefficient so the stronger the correlation (i.e., the closer to -1 or 1), the darker the boxes.

The correlogram is a commonly used tool for **checking randomness in a data set**. If random, autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.





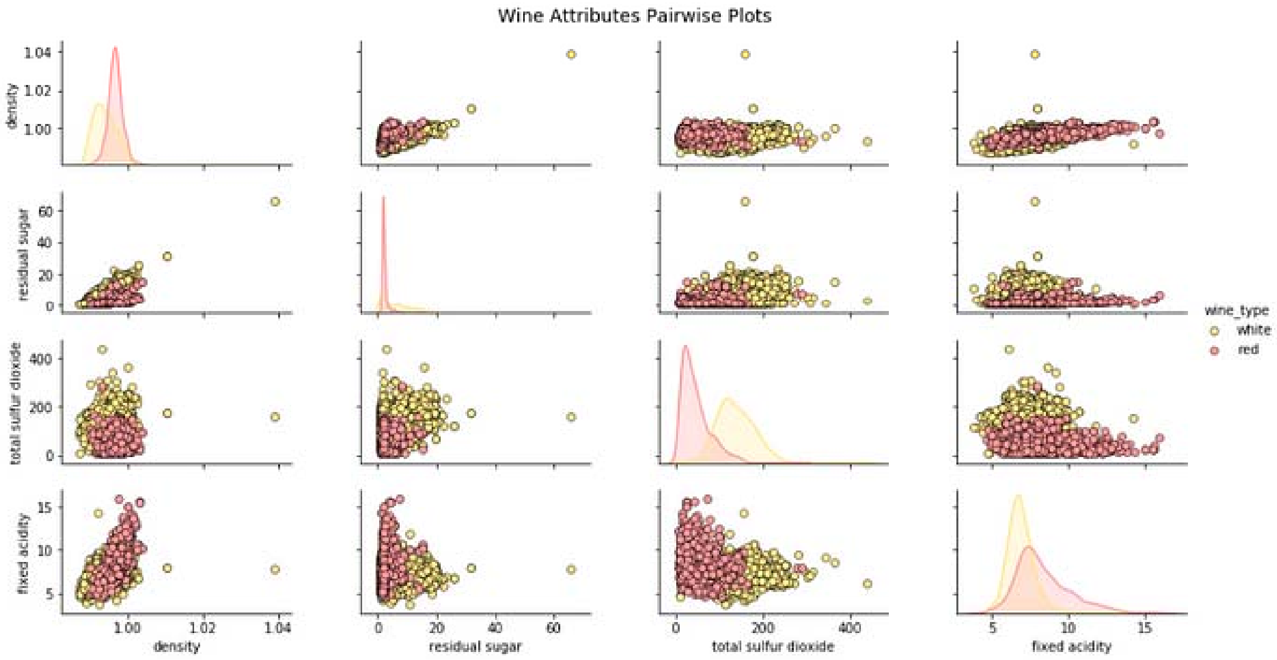
Correlogram – 2D



WINE QUALITY DATA SET:
https://archive.ics.uci.edu/dataset/186/wine+quality

Copyright@ 2024 ISI SQCOR Kolkata

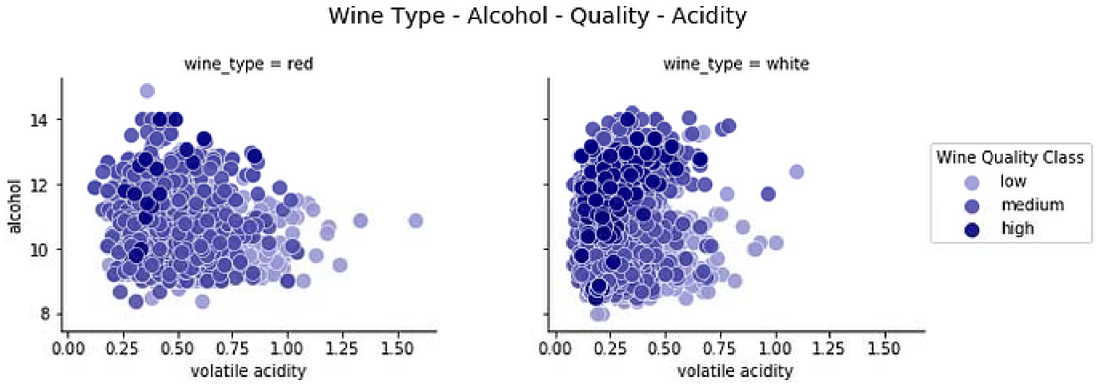
Correlogram – 3D

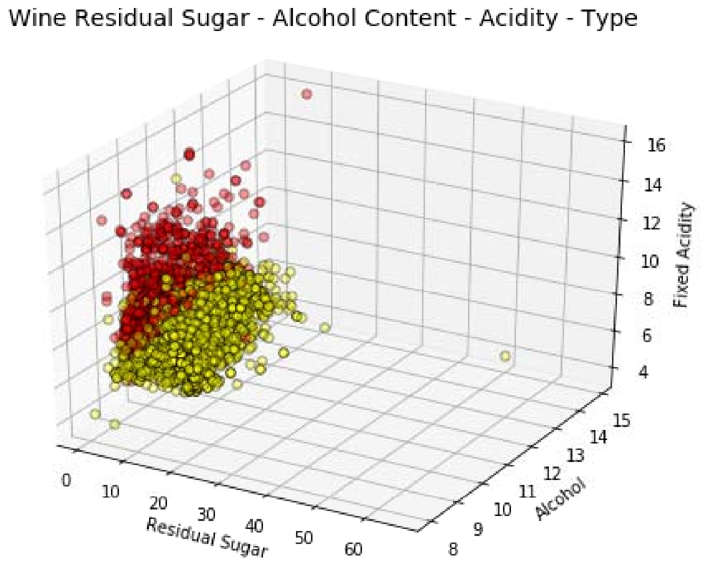


WINE QUALITY DATA SET:
https://archive.ics.uci.edu/dataset/186/wine+quality

Copyright@ 2024 ISI SQCOR Kolkata

Correlogram – 4D & 5D

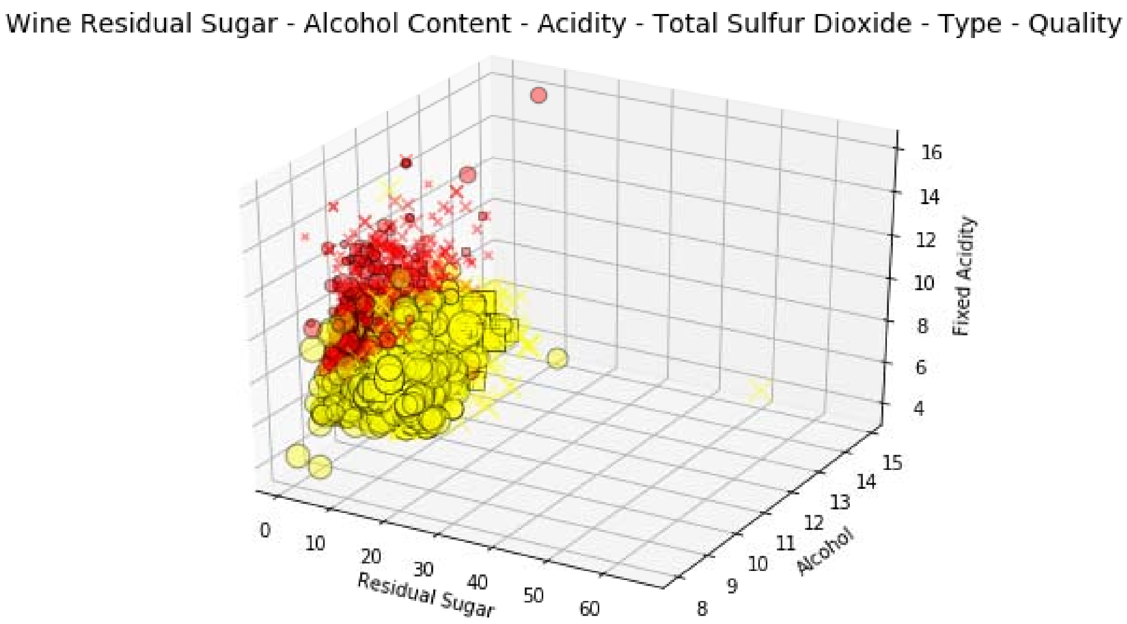




WINE QUALITY DATA SET:
https://archive.ics.uci.edu/dataset/186/wine+quality

Copyright@ 2024 ISI SQCOR Kolkata

Correlogram – 6D



- wine **quality_label** depicted by ***shapes - ****high* (the squared pixel), *medium* (the X marks) and *low* (the circles) quality wines.
- **wine_type** is represented by ***hue****.*
- **fixed acidity** is represented by the ***depth****.*
- **total sulfur dioxide** is represented by ***size****.*

Copyright@ 2024 ISI SQCOR Kolkata

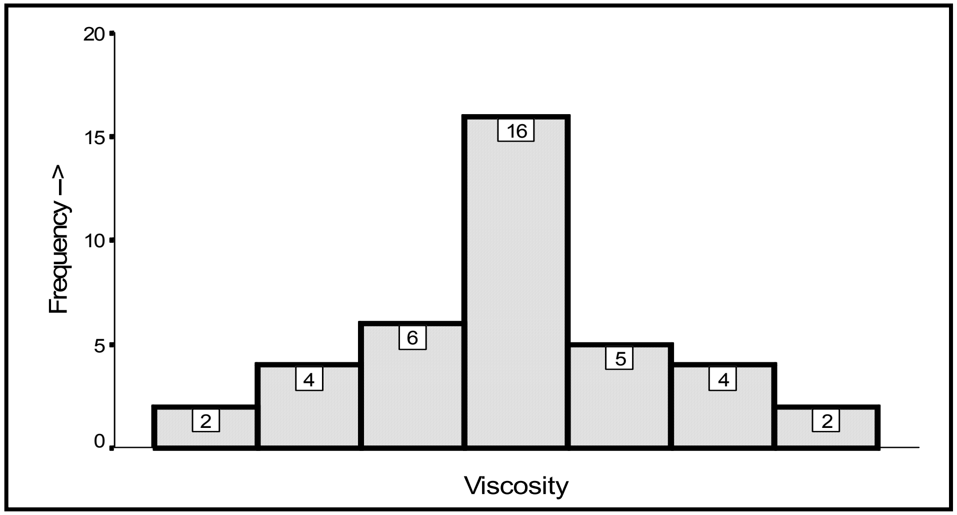
Frequency Distribution & Histogram

In order to study the extent and pattern of variation of Viscosity, following data were gathered.

60, 55, 60, 60, 53, 60, 55, 60, 52, 67, 57, 60, 63, 55, 58, 55, 60, 57, 60, 58, 59, 57, 65, 62,

60, 65, 62, 64, 58, 62, 60, 65, 62, 66, 60, 60, 60, 60, 60

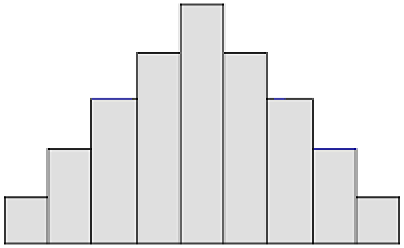
Class Interval	Tally Mark	Frequency
51.5 – 53.8	//	2
53.8 – 56.1	////	4
56.1 – 58.4	卌 /	6
58.4 – 60.7	卌 卌 卌 /	16
60.7 – 63.0	卌	5
63.0 – 65.3	////	4
65.3 – 67.6	//	2
Total	**--**	**39**



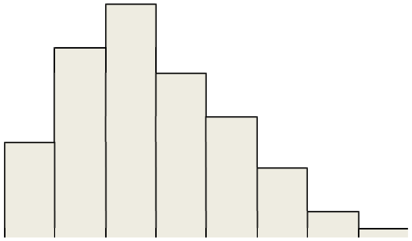
A Histogram is an accurate representation of the distribution of numerical data. It gives an idea of the probability distribution of a continuous variable.

Copyright@ 2024 ISI SQCOR Kolkata

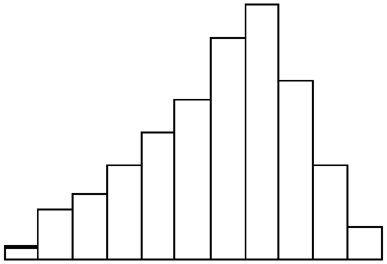
Shapes of Histogram



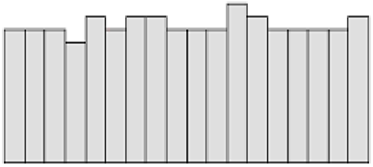
Symmetric Distribution



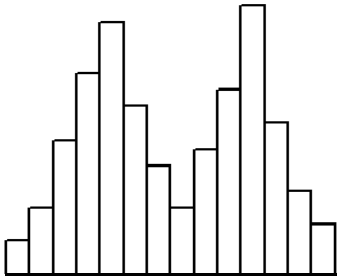
Skewed Distribution



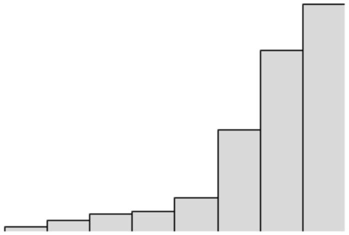
Skewed Distribution



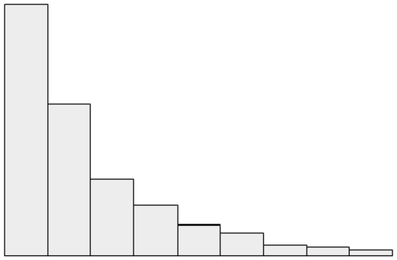
Uniform Distribution



Bimodal Distribution



J-shaped

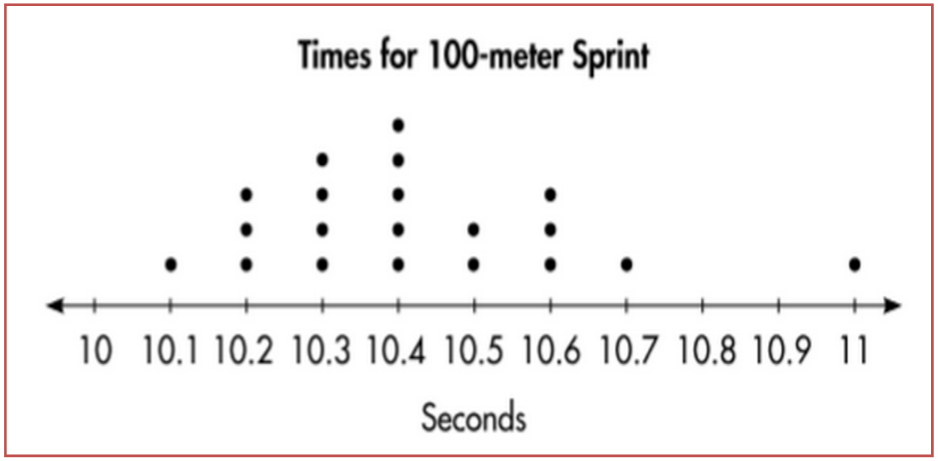


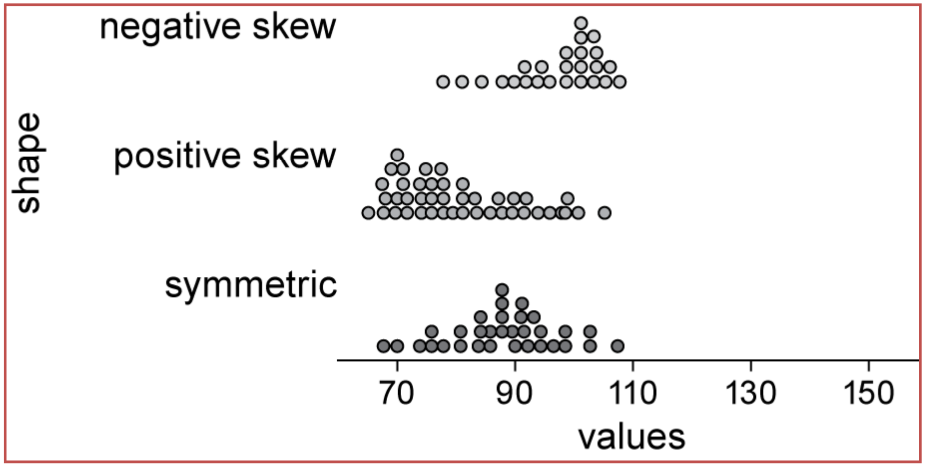
Reverse J-shaped

Copyright@ 2024 ISI SQCOR Kolkata

DOT PLOTS

A statistical chart consisting of data points **plotted** on a fairly simple scale, typically using filled in circles.





Copyright@ 2024 ISI SQCOR Kolkata

Pareto Chart

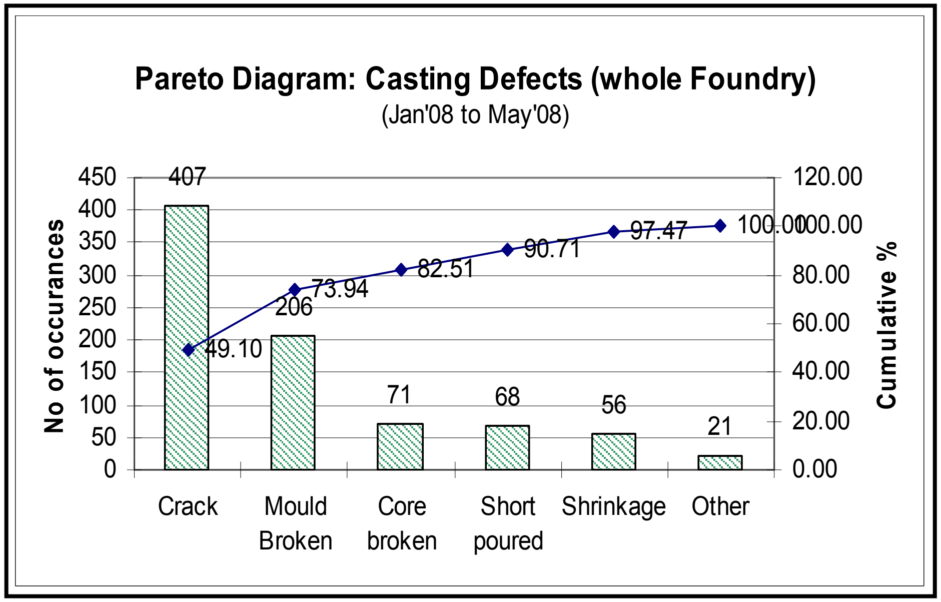
It is a ranked comparison of factors related to a quality problem, from most frequent, down to the least frequent.

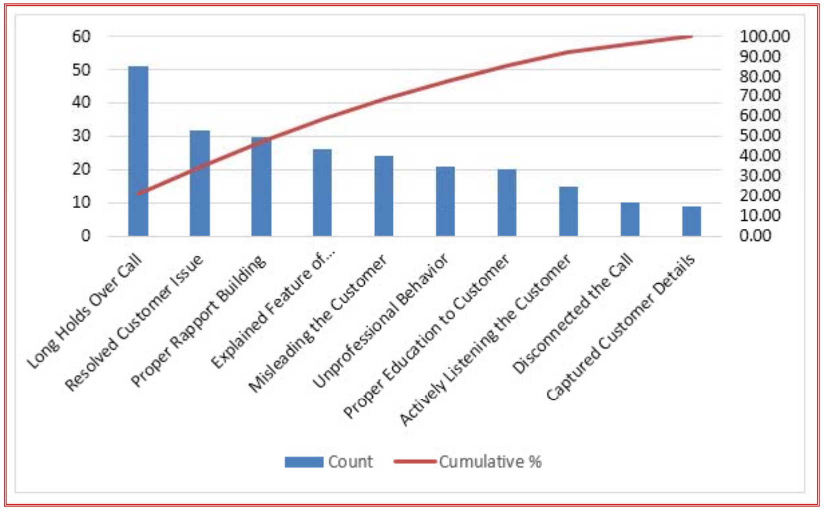
Utility: to identify and focus on the “vital few” factors

Vilfredo Pareto observed that a relative few people held the majority of the wealth. Dr. Joseph Juran has coined the terms *vital few* and *useful many* to refer to those few contributions which account for the bulk of the effect and to those many others which account for a smaller proportion of the effect.

The essential basic administer behind the Pareto principle is that in relatively every case, 80% of the aggregate issues brought about are caused by 20% of the issue causes.

Pareto Chart: Examples

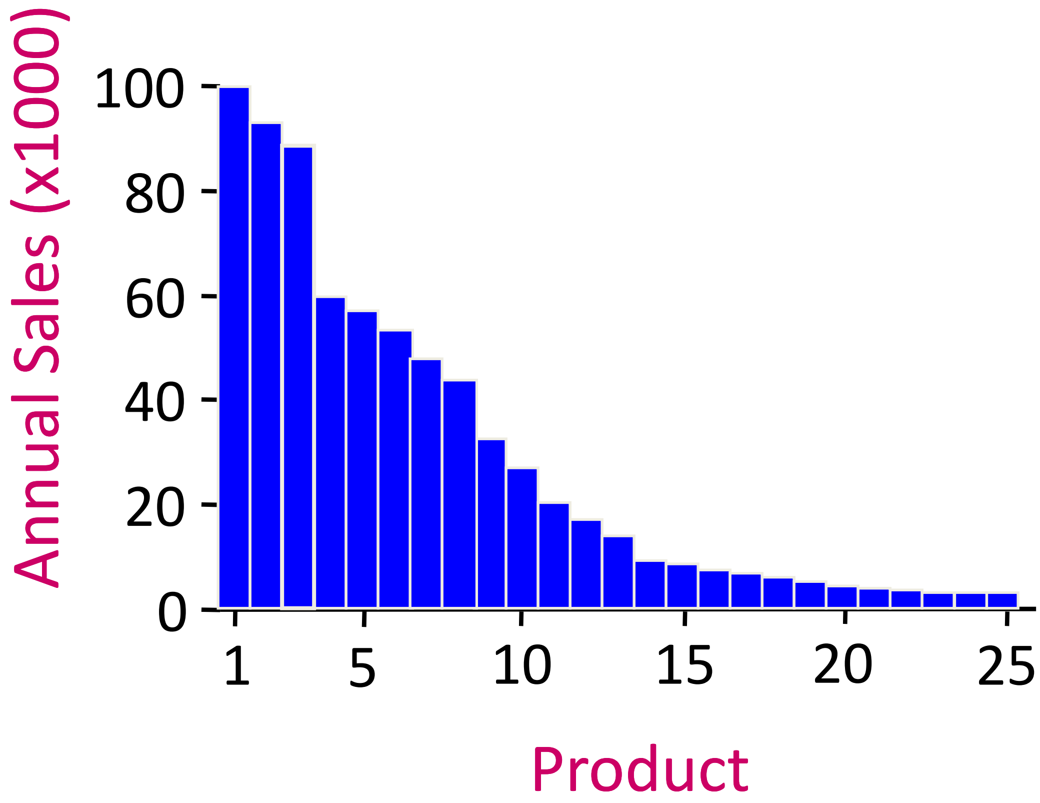




Copyright@ 2024 ISI SQCOR Kolkata

Pareto in Inventory Management: Classifying Items as ABC

25 products sorted by Annual Sales Volume

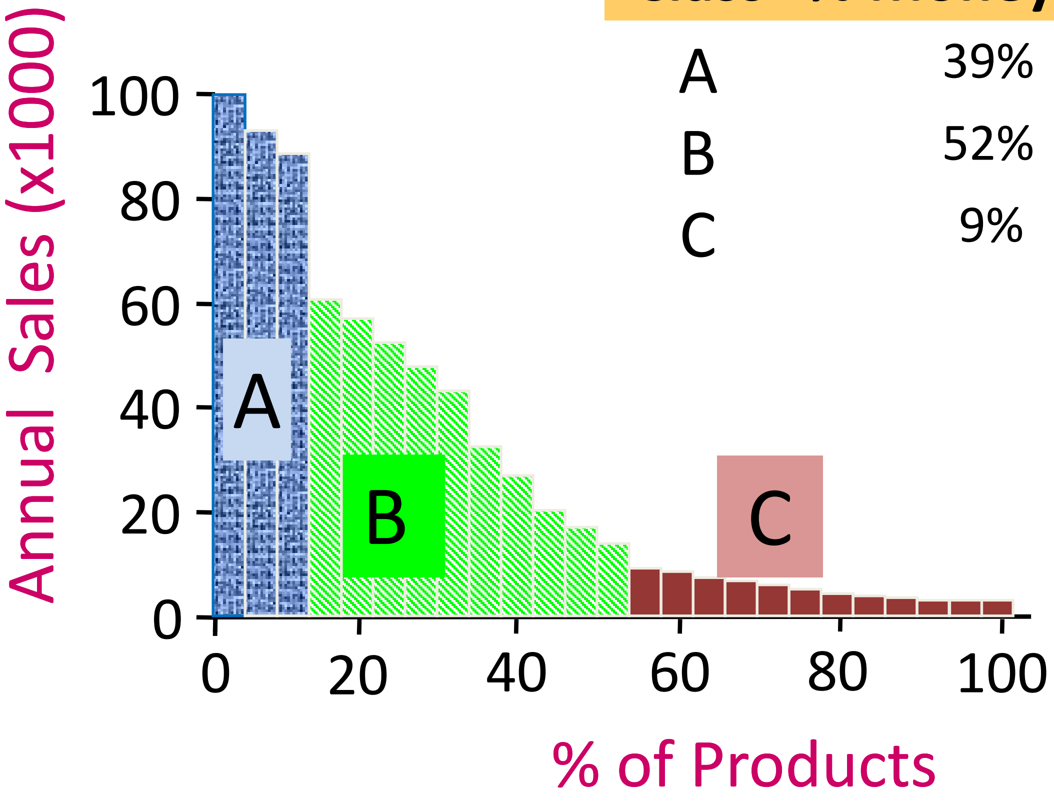


Product	Sales	%
1	100	14
2	92	13
3	88	12
4	60	8
5	58	8
6	53	7
7	49	7
8	41	6
9	32	4
10	26	4
11	21	3
12	18	2
13	16	2
14-25	66	9
Total	720	

Copyright@ 2024 ISI SQCOR Kolkata

Pareto in Inventory Management: Classifying Items as ABC

Class	% Money Val	% Items
A	39%	12% (3/25)
B	52%	40% (10/25)
C	9%	48% (12/25)



Copyright@ 2024 ISI SQCOR Kolkata

Stem-and-Leaf Plot

Example:

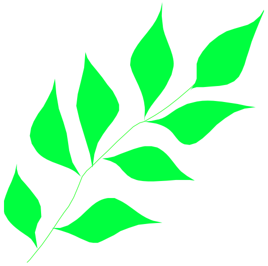
The manager of an Auto Repair company would like to have a better understanding of the cost of part used in the engine tune-ups performed in the shop. The company examines 50 customer invoices for tune-ups. The costs of Parts, are shown below in INR (ín hundreds):

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

5	2 7
6	2 2 2 2 5 6 7 8 8 8 9 9 9
7	1 1 2 2 3 4 4 5 5 5 6 7 8 9 9 9
8	0 0 2 3 5 8 9
9	1 3 7 7 7 8 9
10	1 4 5 5 9

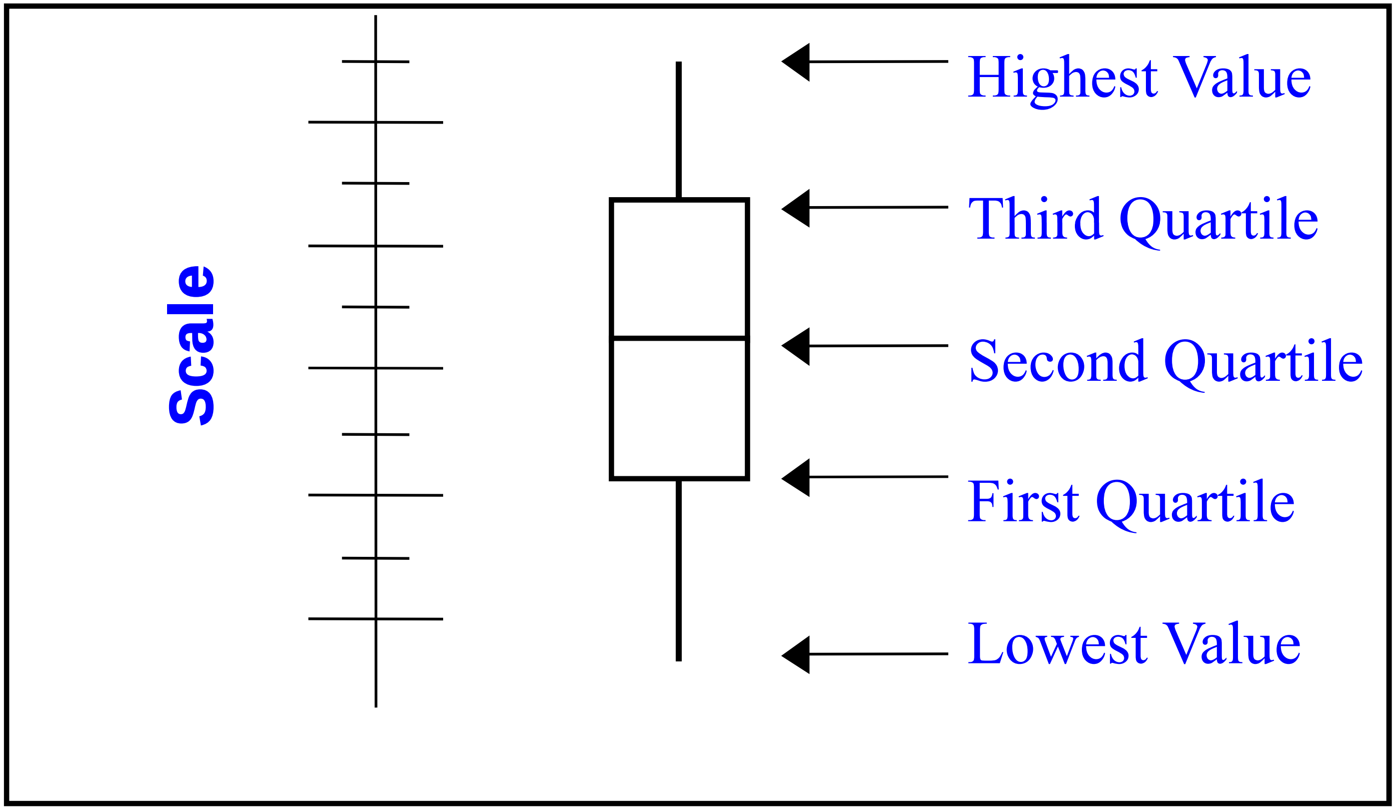
a leaf

a stem



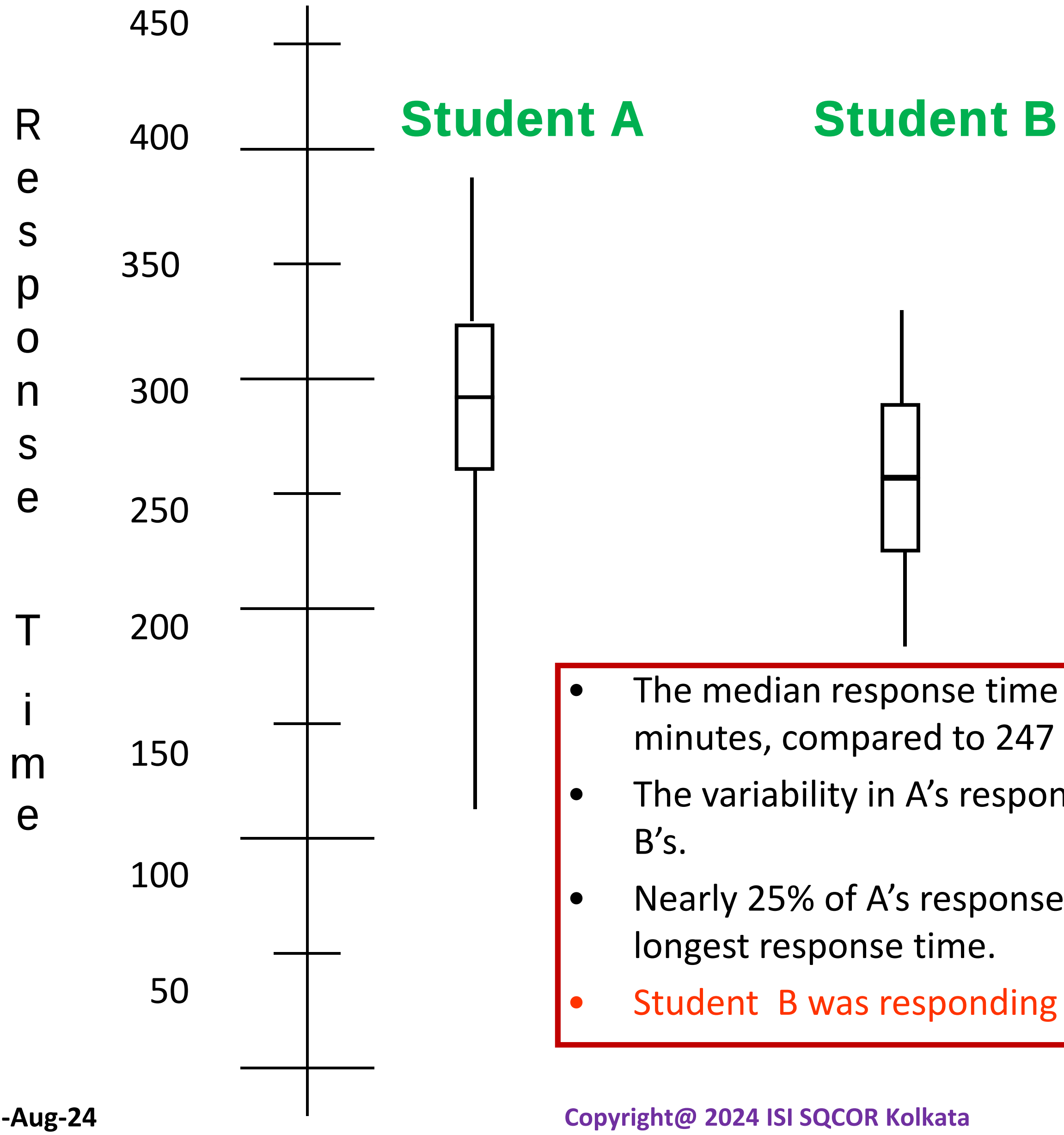
Copyright@ 2024 ISI SQCOR Kolkata

BOX PLOT: Structure (5 – point)



Copyright@ 2024 ISI SQCOR Kolkata

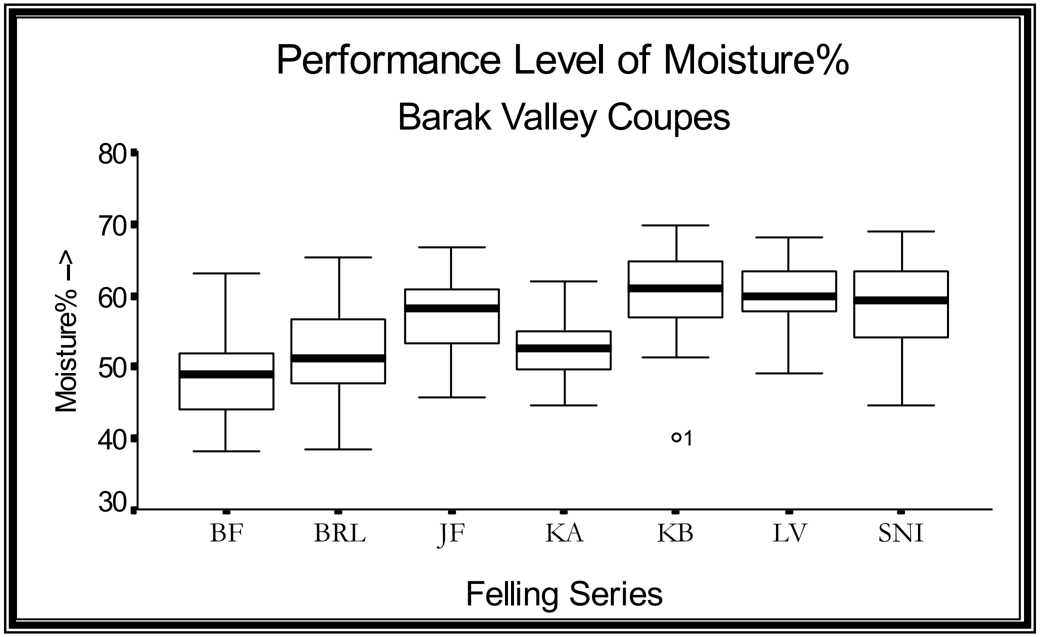
BOX PLOTS - Comparison

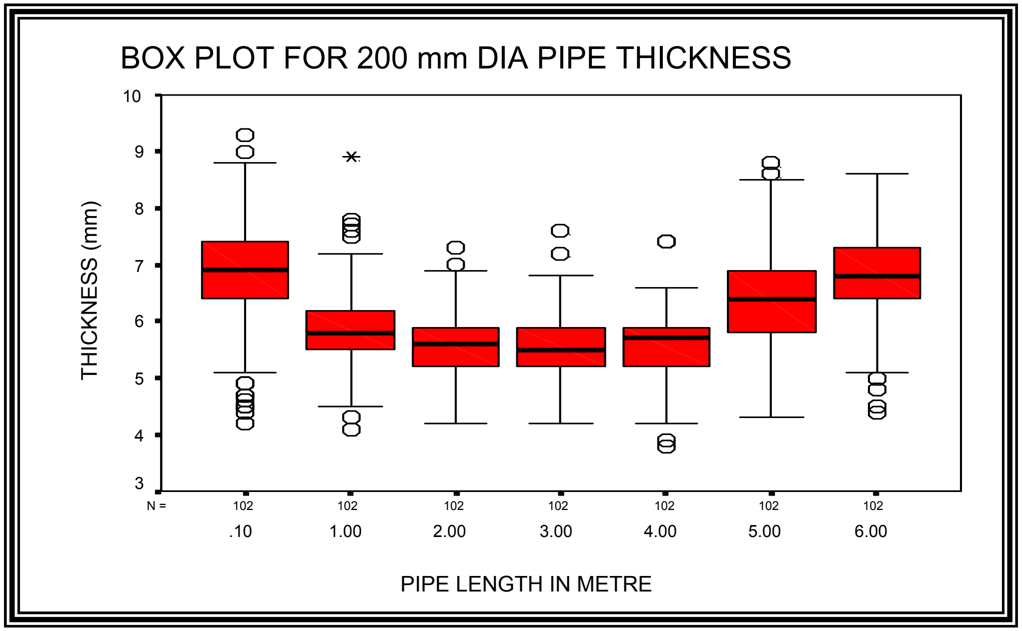


- The median response time of B was 190 minutes, compared to 247 minutes for A
- The variability in A's response time was greater than B's.
- Nearly 25% of A's response times were longer than B's longest response time.
- Student B was responding faster.

Copyright@ 2024 ISI SQCOR Kolkata

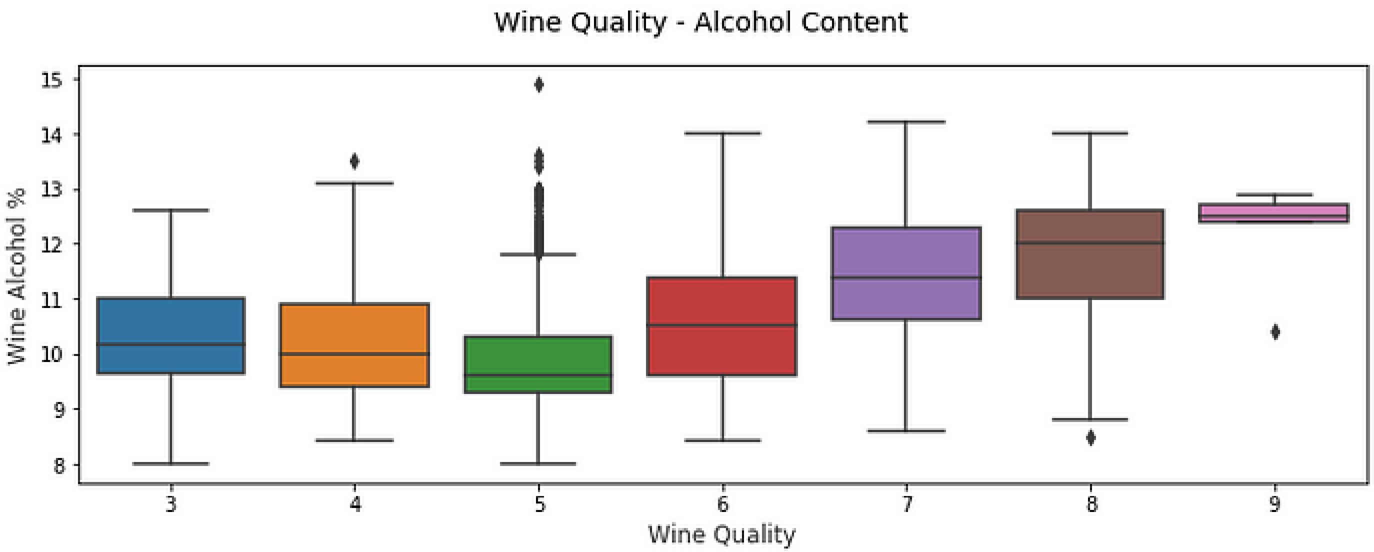
Some more examples





Copyright@ 2024 ISI SQCOR Kolkata

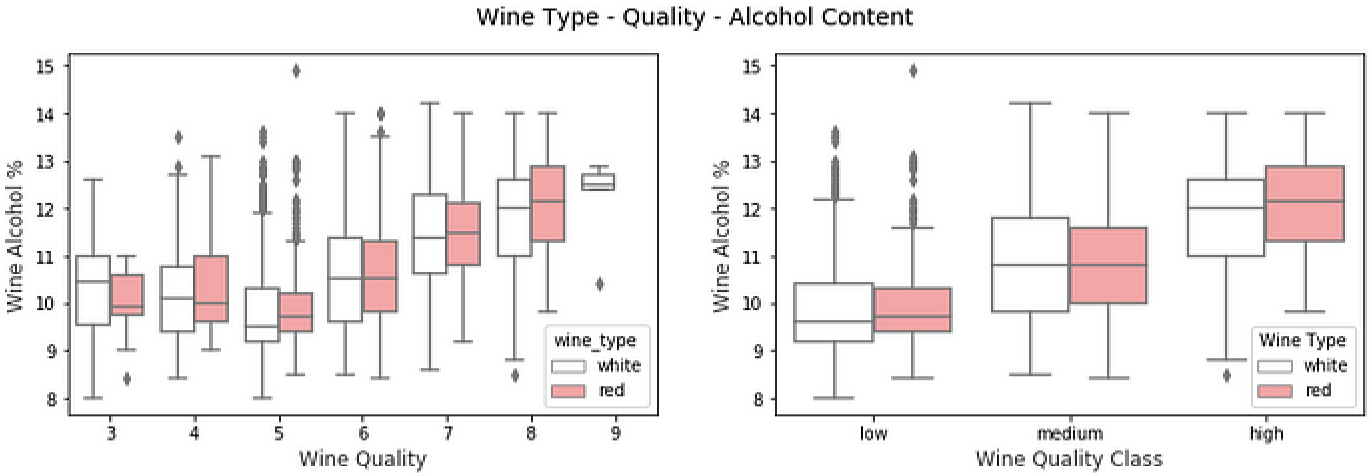
Multiple BOX PLOT – 2D



WINE QUALITY DATA SET:
https://archive.ics.uci.edu/dataset/186/wine+quality

Copyright@ 2024 ISI SQCOR Kolkata

Multiple BOX PLOT – 3D

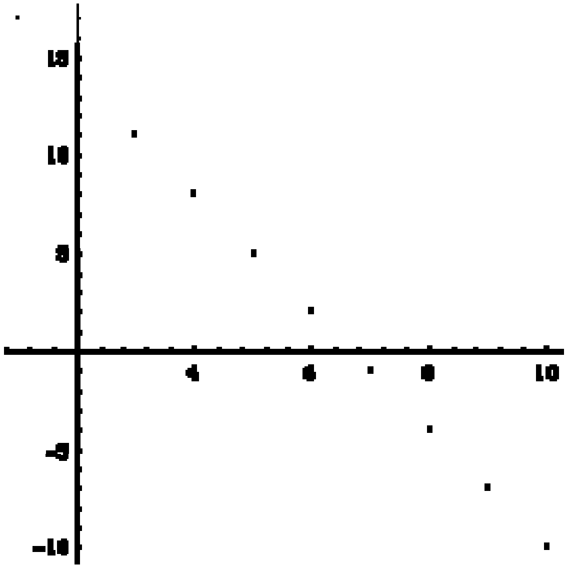


WINE QUALITY DATA SET:
https://archive.ics.uci.edu/dataset/186/wine+quality

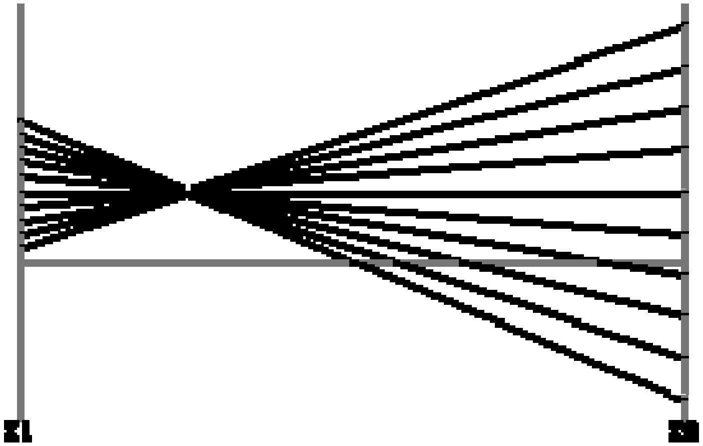
Copyright@ 2024 ISI SQCOR Kolkata

Parallel Coordinates: 4D+

- Encode variables along a horizontal row
- Vertical line specifies values



Dataset in a Cartesian coordinates



Same dataset in parallel coordinates

Invented by
Alfred Inselberg
while at IBM, 1985

Copyright@ 2024 ISI SQCOR Kolkata

Example: Iris Data

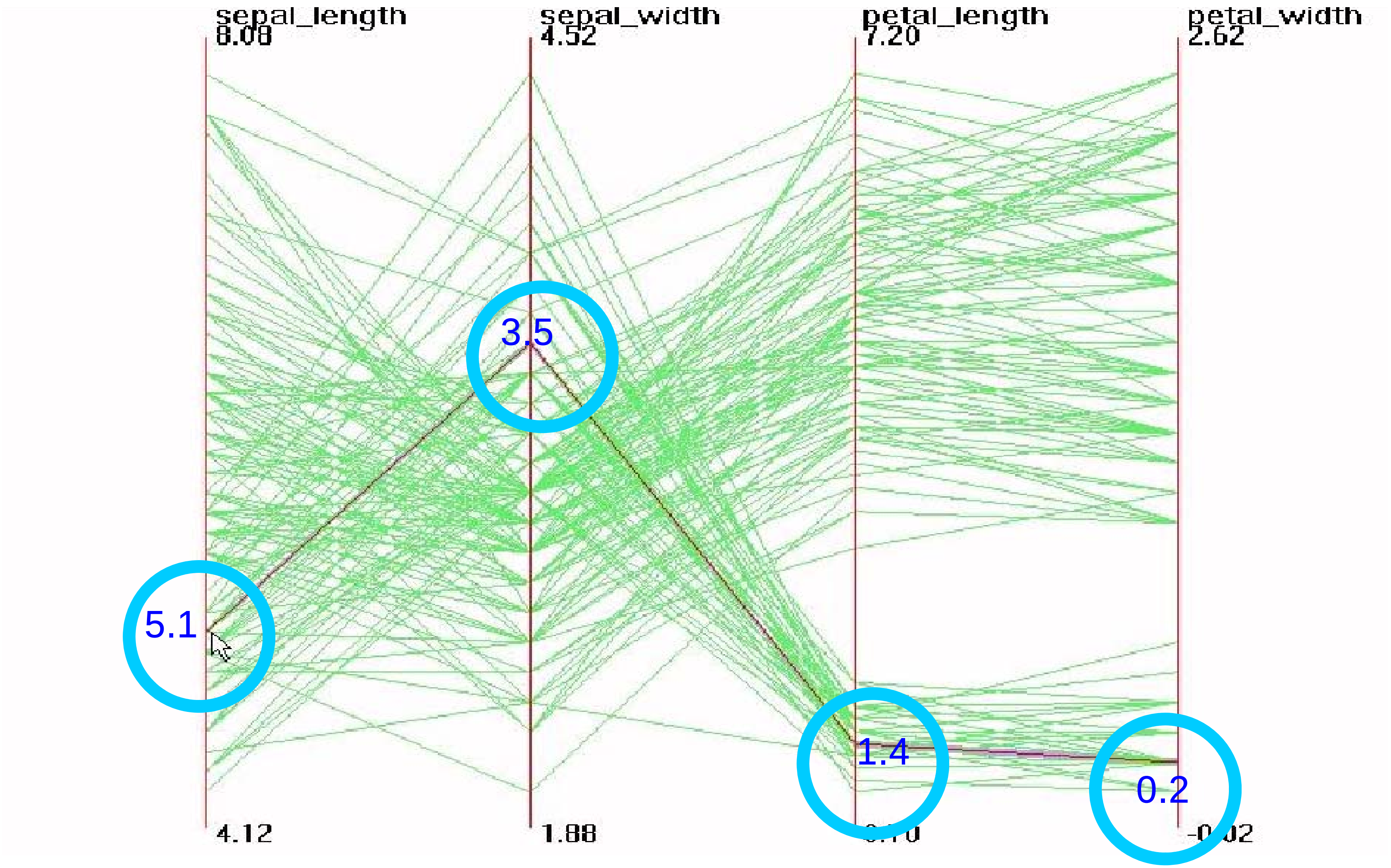
Iris setosa

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
…	…	…	…
5.9	3	5.1	1.8

Iris versicolor

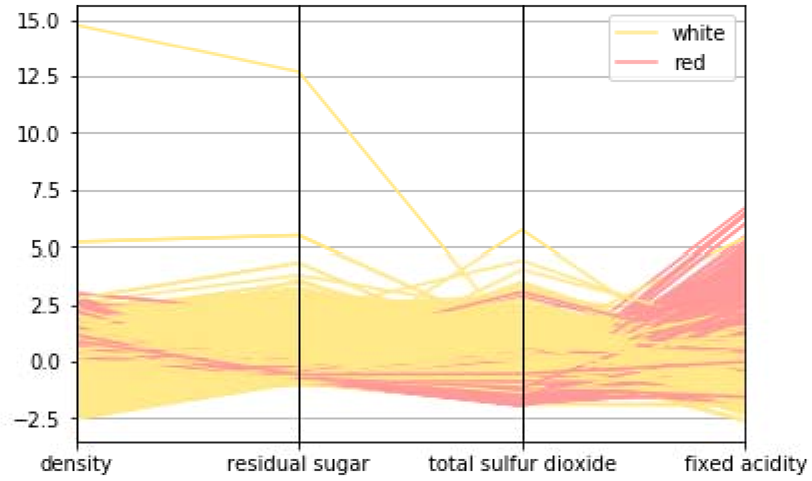
Iris virginica

Parallel Visualization: Iris Data



Copyright@ 2024 ISI SQCOR Kolkata

Parallel Visualization: Wine Data



WINE QUALITY DATA SET:
https://archive.ics.uci.edu/dataset/186/wine+quality

Copyright@ 2024 ISI SQCOR Kolkata

Parallel Visualization: Summary

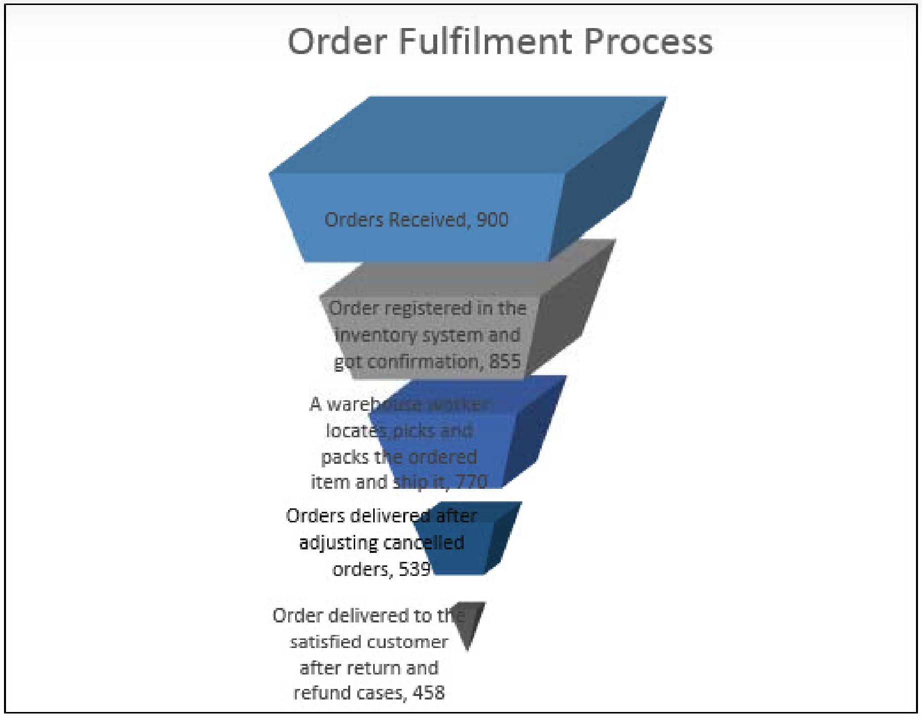
- Each data point is a line
- Similar points correspond to similar lines
- Lines crossing over correspond to negatively correlated attributes
- Interactive exploration and clustering

- **Problems**: order of axes, limit to ~20 dimensions

Copyright@ 2024 ISI SQCOR Kolkata

Funnel Chart

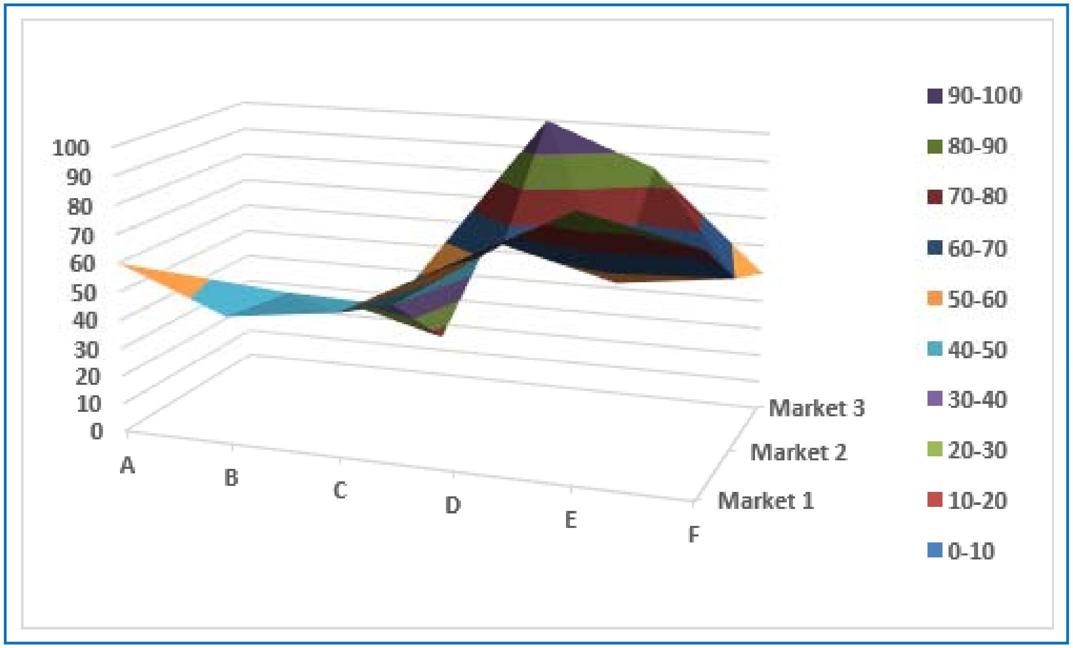
Funnel chart is similar to its name associated with it, it is used to represent data status in every stage as defined and as the values of the concerned event go on decreasing, thus making the shape of a funnel for the chart and so the name. (available in MS Excel 2019).



Copyright@ 2024 ISI SQCOR Kolkata

Surface Chart: 3D

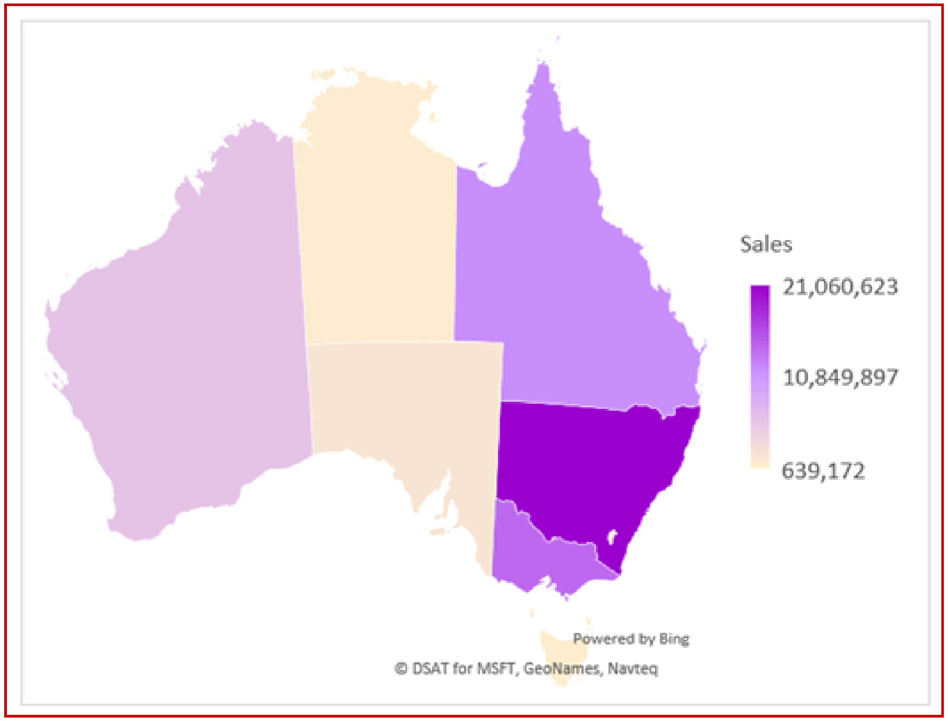
Surface Chart is a 3-dimensional chart. One can see the mesh kind of surface which helps us to find the optimum combination between two kinds of data points. A typical surface chart consists of three variable data points, let's call them as "X, Y, and Z". From these available three variables, we can categorize them into two sets i.e. Independent and Dependent variable. Two variables will be independent variables and one being the dependent variable.



Copyright@ 2024 ISI SQCOR Kolkata

Geographic Mapping

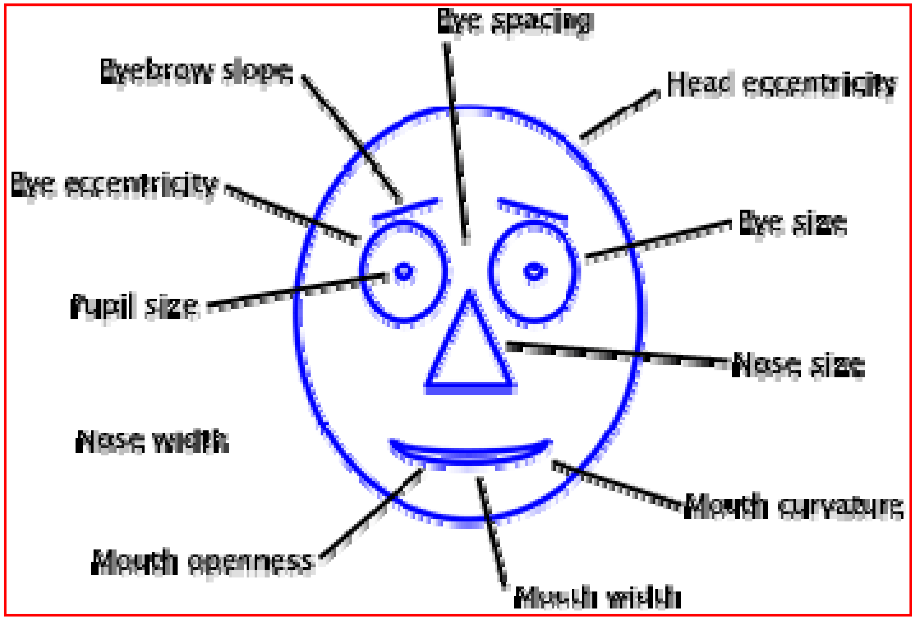
A map chart is visual representation of statistics within a map. This is mainly used where representation of data in a map make more sense than representing in normal form of the chart. There are several tools like Tableu, Microsoft Power BI has this functionalities. Latest version of excel has this functionality by this is very top level, latitude and longitude level customization is not possible.

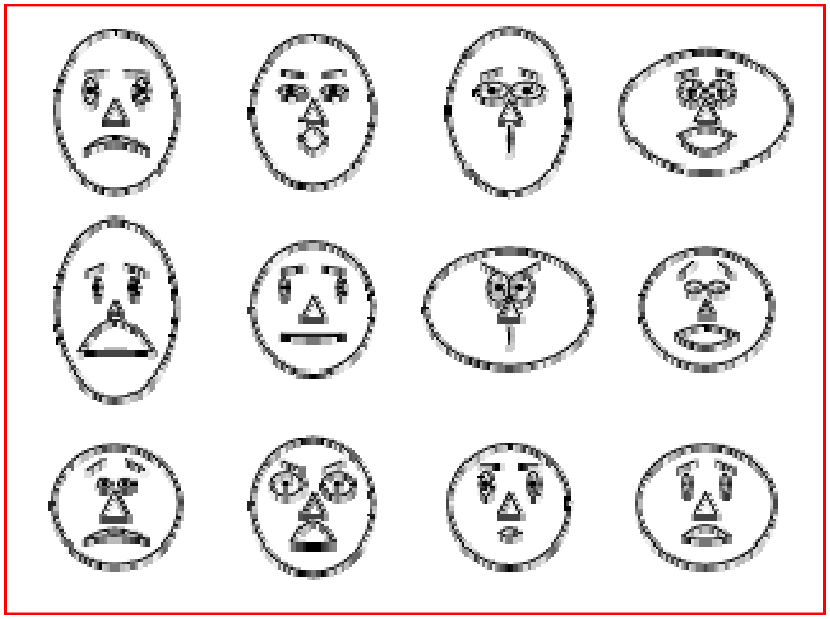


Copyright@ 2024 ISI SQCOR Kolkata

Chernoff Faces: 4D+

These faces display multivariate data in the shape of a human face. The individual parts, such as eyes, ears, mouth and nose represent values of the variables by their shape, size, placement and orientation. **(Herman Chernoff, 1973)**





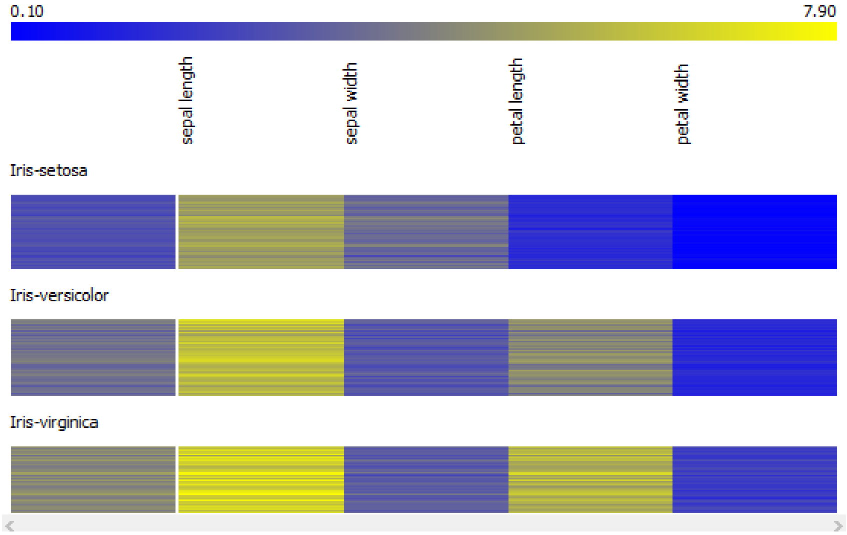
Copyright@ 2024 ISI SQCOR Kolkata

Heat Map

A heat map is a two-dimensional representation of data where the individual values contained in a matrix are represented by colors.

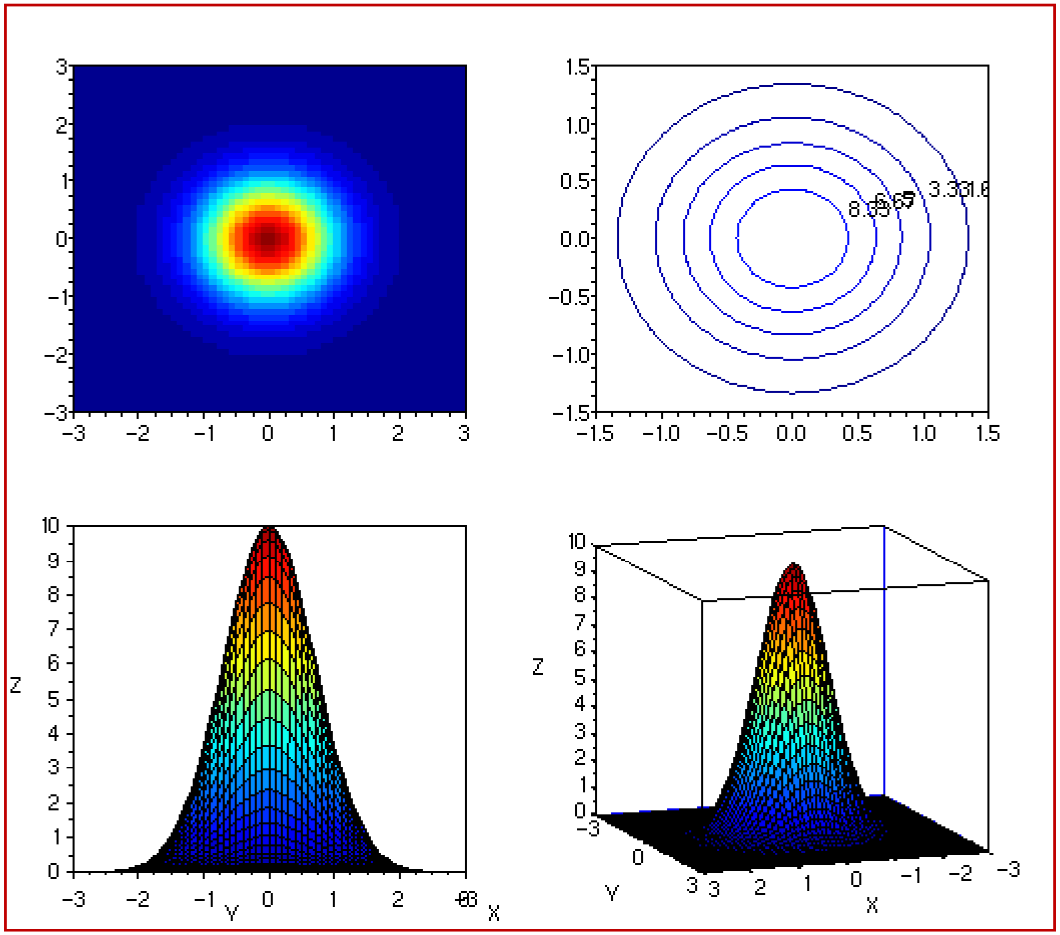
A simple **heat map** provides an immediate visual summary of information.

Now heat maps are the most-used tool for representing complex statistical data.



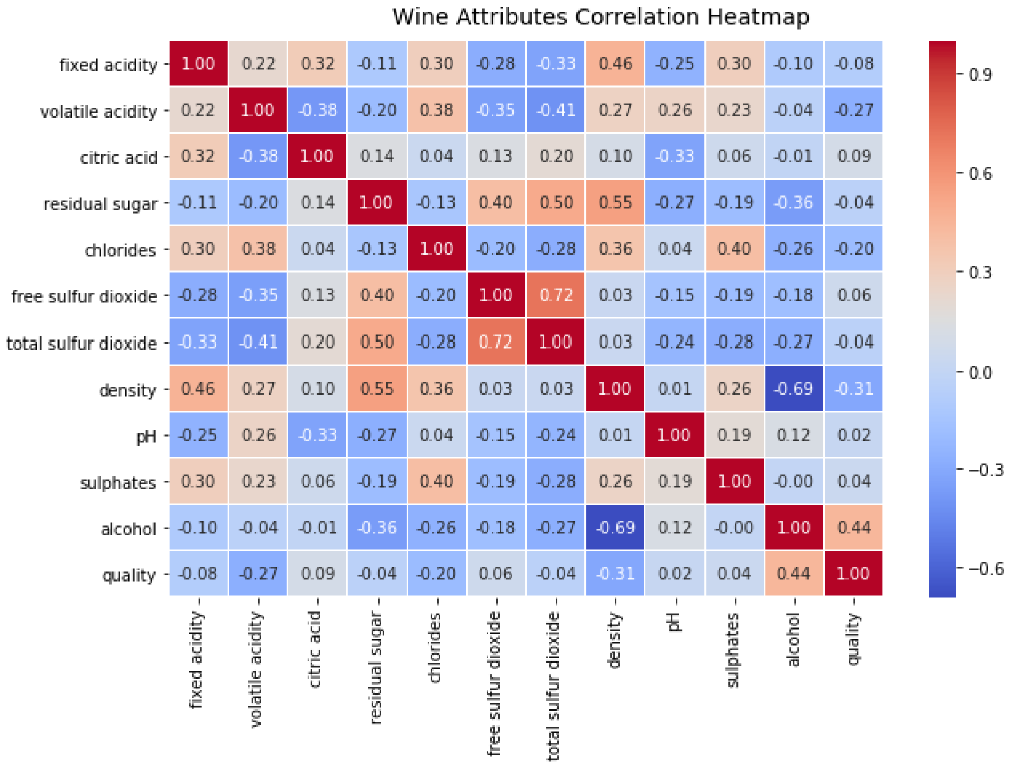
Copyright@ 2024 ISI SQCOR Kolkata

Relationships: heat map, surface plot, contour lines of the same data



Copyright@ 2024 ISI SQCOR Kolkata

Correlation Heat Map



Wine Attributes Correlation Heatmap

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	0.22	0.32	-0.11	0.30	-0.28	-0.33	0.46	-0.25	0.30	-0.10	-0.08
volatile acidity	0.22	1.00	-0.38	-0.20	0.38	-0.35	-0.41	0.27	0.26	0.23	-0.04	-0.27
citric acid	0.32	-0.38	1.00	0.14	0.04	0.13	0.20	0.10	-0.33	0.06	-0.01	0.09
residual sugar	-0.11	-0.20	0.14	1.00	-0.13	0.40	0.50	0.55	-0.27	-0.19	-0.36	-0.04
chlorides	0.30	0.38	0.04	-0.13	1.00	-0.20	-0.28	0.36	0.04	0.40	-0.26	-0.20
free sulfur dioxide	-0.28	-0.35	0.13	0.40	-0.20	1.00	0.72	0.03	-0.15	-0.19	-0.18	0.06
total sulfur dioxide	-0.33	-0.41	0.20	0.50	-0.28	0.72	1.00	0.03	-0.24	-0.28	-0.27	-0.04
density	0.46	0.27	0.10	0.55	0.36	0.03	0.03	1.00	0.01	0.26	-0.69	-0.31
pH	-0.25	0.26	-0.33	-0.27	0.04	-0.15	-0.24	0.01	1.00	0.19	0.12	0.02
sulphates	0.30	0.23	0.06	-0.19	0.40	-0.19	-0.28	0.26	0.19	1.00	-0.00	0.04
alcohol	-0.10	-0.04	-0.01	-0.36	-0.26	-0.18	-0.27	-0.69	0.12	-0.00	1.00	0.44
quality	-0.08	-0.27	0.09	-0.04	-0.20	0.06	-0.04	-0.31	0.02	0.04	0.44	1.00

WINE QUALITY DATA SET:
https://archive.ics.uci.edu/dataset/186/wine+quality

Copyright@ 2024 ISI SQCOR Kolkata

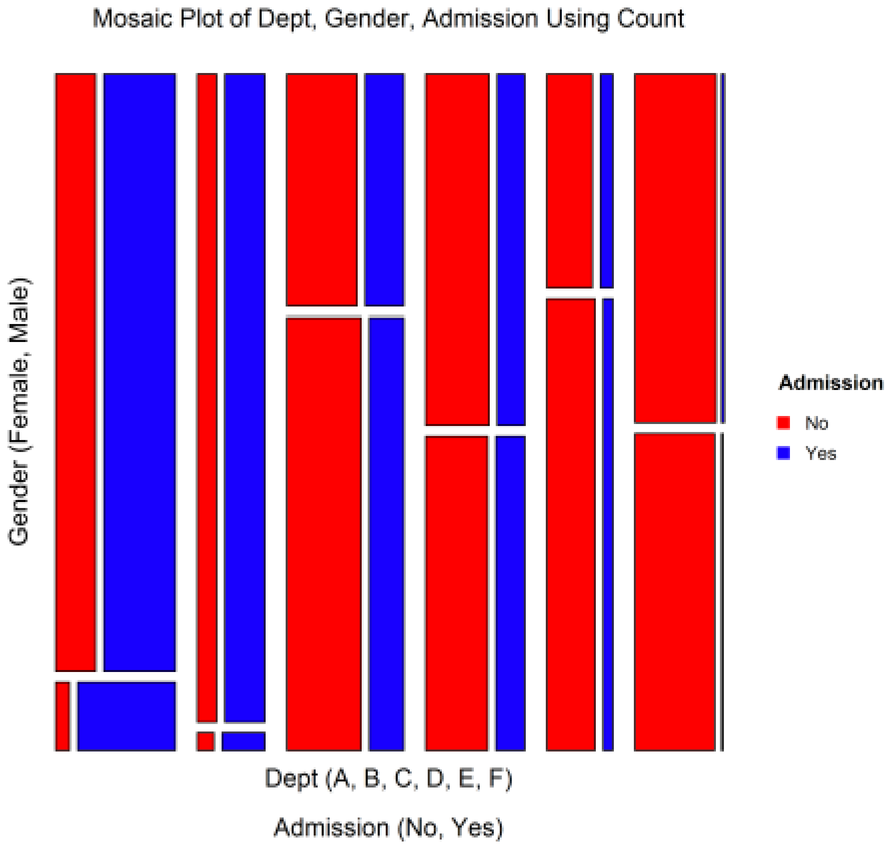
Mosaic Plot:2D+

A **mosaic plot** is a graphical method for visualizing data from two or more qualitative variables. It is the multidimensional extension of spineplots, which graphically display the same information for only one variable. It gives an overview of the data and makes it possible to recognize relationships between different variables. Independence can be shown when the boxes across categories all have the same areas. (Hartigan and Kleiner, 1981; Friendly, 1994)

Mosaic Plot

DATA TABLE

Admission	Gender	Dept	Count
No	Female	A	19
No	Female	B	8
No	Female	C	391
No	Female	D	244
No	Female	E	299
No	Female	F	317
No	Male	A	313
No	Male	B	207
No	Male	C	205
No	Male	D	279
No	Male	E	138
No	Male	F	351
Yes	Female	A	89
Yes	Female	B	17
Yes	Female	C	202
Yes	Female	D	131
Yes	Female	E	94
Yes	Female	F	24
Yes	Male	A	512
Yes	Male	B	353
Yes	Male	C	120
Yes	Male	D	138
Yes	Male	E	53
Yes	Male	F	22



Hint: Contingency 2x2 table

Mosaic Plot

By construction, the percent admitted within each gender-by-department combination is the width of the corresponding box.

For example, the percentage of females that were admitted to department A (shown by the width of blue box at the lower left) is much larger than that of the males (shown by the width of the long blue box directly above the female box).

If you consider each department in turn by scanning from left to right across the plot, the width of the blue box on the bottom appears to be quite similar to the box directly above it. This indicates that in most departments the percent of females admitted is about the same as that of males admitted.

Copyright@ 2024 ISI SQCOR Kolkata

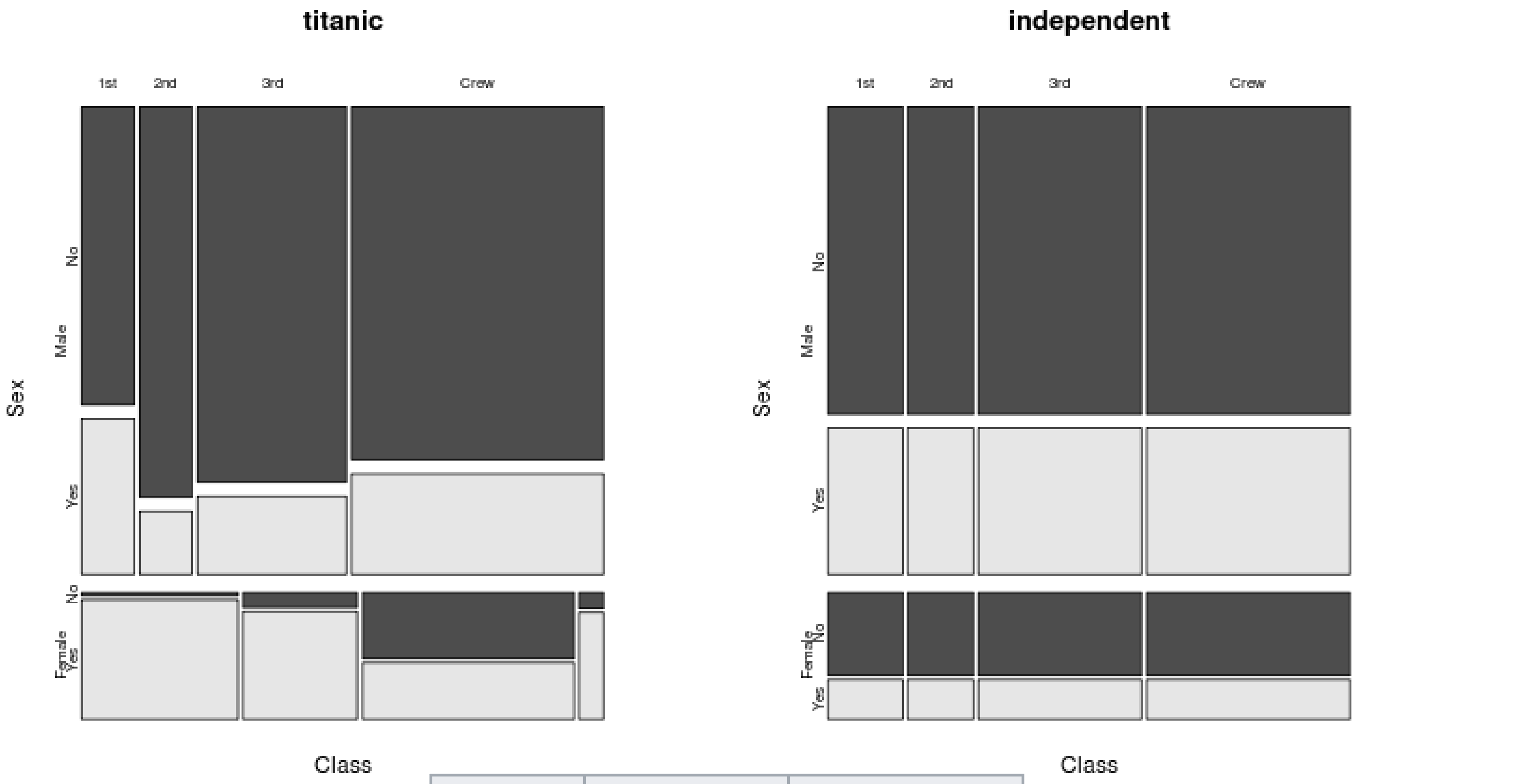
Mosaic Plot

Data from the passengers on the *Titanic:* 2201 observations and 3 variables.

Gender	Survived	1st Class	2nd Class	3rd Class	Crew
Male	No	118	154	422	670
	Yes	62	25	88	192
Female	No	4	13	106	3
	Yes	141	93	90	20

Order	Variable	Axis
1.	Gender	Vertical
2.	Class	Horizontal
3.	Survived	Vertical

Mosaic Plot



Order	Variable	Axis
1.	Gender	Vertical
2.	Class	Horizontal
3.	Survived	Vertical

Copyright@ 2024 ISI SQCOR Kolkata

Mosaic Plot

The categorical variables are first put in order. Then, each variable is assigned to an axis. In the table to the right, sequence and classification is presented for this data set. Another ordering will result in a different mosaic plot, i.e., the order of the variables is significant as for all multivariate plots.

At the left edge of the first variable we first plot "Gender," meaning that we divide the data vertically in two blocks: the bottom blocks corresponds to females, while the upper (much larger) one to males. One immediately sees that roughly a quarter of the passengers were female and the remaining three quarters male.

One then applies the second variable "Class" to the top edge. The four vertical columns therefore mark the four values of that variable (1st, 2nd, 3rd, and crew). These columns are of variable thickness, because column width indicates the relative proportion of the corresponding value on the population. Crew plainly represents the largest male group, whereas third-class passengers are the largest female group. The number of female crew members is also seen to have been marginal.

The last variable ("Survived") is finally applied, this time along the left edge with the result highlighted by shade: dark grey rectangles represent people that did not survive the disaster, light grey ones people that did. Women in the first class are immediately seen to have had the highest survival probability. The survival probability for females is seen to have been higher than that for men (marginalised over all classes). Similarly, a marginalization over gender identifies first-class passengers as most probable to survive. Overall, about 1/3 of all people survived (proportion of light gray areas).

Copyright@ 2024 ISI SQCOR Kolkata

Mosaic Plot

Properties:

- The displayed variables are categorical or in ordinal scales.
- The plot is of at least two variables. There is no upper limit, but too many variables may be confusing in graphic form.
- The number of observations is not limited, but not read in the image.
- The surfaces of the rectangular fields that are available for a combination of features are proportional to the number of observations that have this combination of features.
- Unlike, for example, the Box Plot or QQ plot, it is not possible for the mosaic plot to plot a confidence interval. The significance of different frequencies of the various characteristic values can therefore not be observed visually.

Copyright@ 2024 ISI SQCOR Kolkata

Dendogram: Tree Clustering

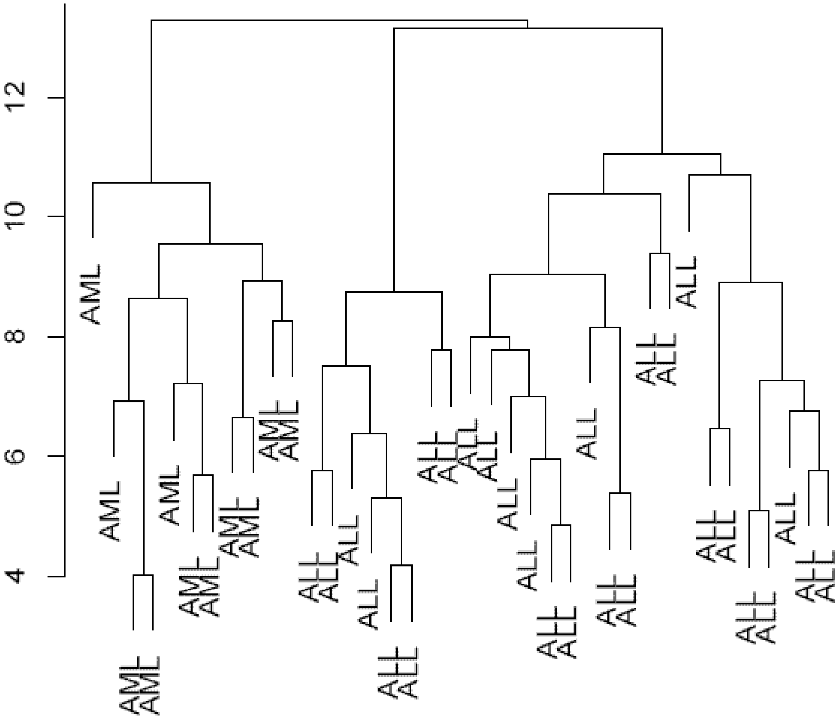
Cluster: set of objects that are similar to each other and separated from the other objects.

Method (Algorithm) of Clustering: K-means, PAM, SOM, Hierarchical

Distance Between Clusters: **MEASURE??**

Hierarchical Clustering: Similarity of objects represented in a tree structure (**Dendrogram**)

Application: Document clustering, Microarray data/genes and many…

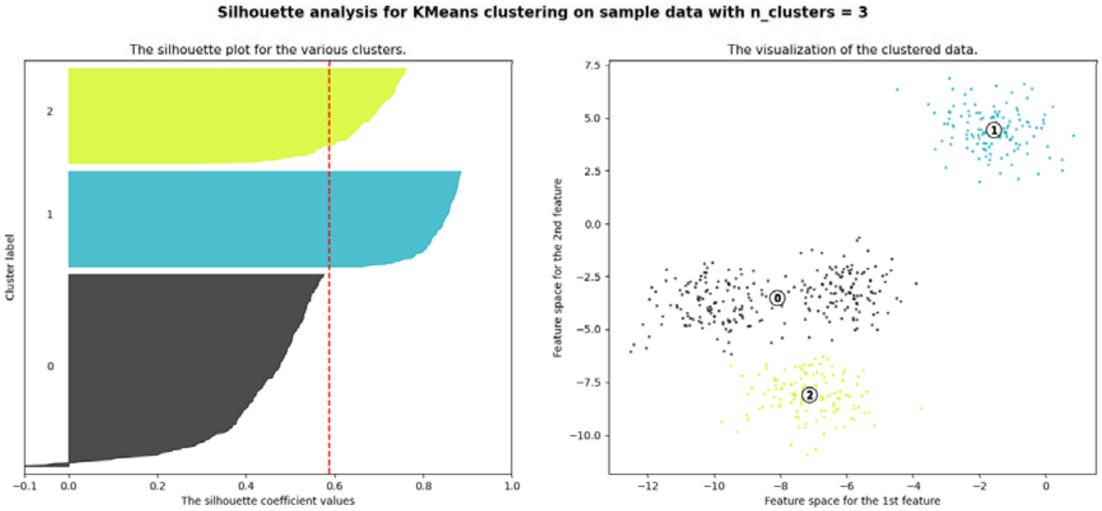


Golub data: different types of leukemia. Clustering based on the 150 genes with highest variance across all samples.

Copyright@ 2024 ISI SQCOR Kolkata

Silhouette Plot (clustering)

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure (Silhouette Coefficient) has a range of [-1, 1].



Silhouette coefficient:

- ~ +1: the sample data point is very close to its own cluster and far away from the neighboring clusters.
- ~ 0: the sample data point is on or very close to the decision boundary between two neighboring clusters
- ~ -1: those samples might have been assigned to the wrong cluster.

Thickness of the silhouette plot: represents the cluster size.

Copyright@ 2024 ISI SQCOR Kolkata

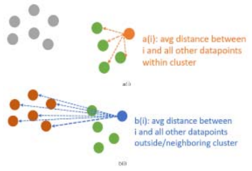
Silhouette Coefficient (clustering)

The **Silhouette coefficient** is a metric that measures how well each data point fits into its assigned cluster. It combines information about both the **cohesion** (how close a data point is to other points in its own cluster) and the **separation** (how far a data point is from points in other clusters) of the data point.

A higher silhouette score indicates that the data points are well-clustered, with clear separation between clusters and tight cohesion within each cluster. Conversely, a lower silhouette score suggests that the clustering may be less accurate, with overlapping clusters or points that are not well-assigned to their respective clusters.

Calculation:

1. *Average distance to all other data points within the same cluster (cohesion).*
2. *Average distance to all data points in the nearest neighboring cluster (separation).*
3. *Silhouette coefficient* = *(separation — cohesion) / max(separation, cohesion)*


a(i): avg distance between
i and all other datapoints
within cluster
a(i)
b(i): avg distance between
i and all other datapoints
outside/neighboring cluster
b(i)

Calculate the average silhouette coefficient across all data points to obtain the overall silhouette score for the clustering result (*higher the better*).

The silhouette coefficient provides a quantitative measure to evaluate the quality of clustering results. By considering both the cohesion and separation of data points, it offers insights into the effectiveness of the clustering algorithm and the distinctness of the clusters.

Copyright@ 2024 ISI SQCOR Kolkata

Data Visualization

Important Links

Easy Excel :
https://www.excel-easy.com/examples/pareto-chart.html

Excel Modeling :
https://www.wallstreetmojo.com/category/financial-modeling/excel-modeling/

Statistics : How to
https://www.statisticshowto.datasciencecentral.com/

Smart Sheet:
https://www.smartsheet.com/stacked-bar-chart-graph

http://jcsites.juniata.edu/faculty/rhodes/iv/multivarviz.html

https://docs.google.com/spreadsheets/d/1PR5StHxg2jlMCb4IUilGSEwhylXn-3q3EJucSaVolCU/edit#gid=0
https://www.tatvic.com/blog/7-visualizations-learn-r/

Visualization Tool: Dashboard

In the BIG data age with the availability of lots of data, some data visualization tools produce excellent meaningful output and require no technical knowhow indeed.

tableau

This is most popular visualization tool, produces excellent graphs. Graph making process is very simple, just drag and drop.

Power BI is another popular visualization tool, it also produces excellent interactive graphs. Graph making process is simple, only drag and drop.

Qlik

This is another popular visualization tool, it also produces excellent graphs. Graph making process is similar to Tableau, drag and drop.

orange
DATA MINING
FRUITFUL&FUN

Orange is another popular visualization tool, it also produces excellent interactive graphs. Graph making process is simple, only drag and drop.

Visualization Techniques vs. Available Software

		Applicable Dashboard/Software							
Sl. No.	Visualization Techniques	Tableau	Qlikview	Microsoft Power BI	Amazon QuickSight	Google Data Studio	Jaspersoft	Pentaho BI	Domo
1	Frequency Distribution	✓	✓	✓	✓	✓	✓	✓	✓
2	Frequency Polygon	✓	✓	✓	✓	✓	✓	x	x
3	Ogive / NPP	✓	✓	✓	✓	✓	✓	x	x
4	Histogram	✓	✓	✓	✓	✓	✓	✓	✓
5	Box Plot	✓	✓	✓	✓	✓	✓	✓	✓
6	Dot Plot	✓	✓	✓	✓	✓	✓	✓	✓
7	Stem-and-Leaf Plot	✓	x	✓	✓	✓	x	x	x
8	Bar Chart	✓	✓	✓	✓	✓	✓	✓	✓
9	Column Chart	✓	✓	✓	✓	✓	✓	✓	✓
10	Stack Column Chart	✓	✓	✓	✓	✓	✓	✓	✓
11	Line Chart	✓	✓	✓	✓	✓	✓	✓	✓
12	Time Series Plot	✓	✓	✓	✓	✓	✓	✓	✓
13	Parallel Coordinates	✓	✓	✓	✓	✓	x	x	x
14	Pie Chart	✓	✓	✓	✓	✓	✓	✓	✓
15	Area Chart	✓	✓	✓	✓	✓	✓	✓	✓
16	Scatter plot	✓	✓	✓	✓	✓	✓	✓	x
17	Scatter plot Matrix	✓	✓	✓	✓	✓	✓	✓	x
18	Bubble Chart	✓	✓	✓	✓	✓	✓	✓	✓
19	Radar Plot	✓	✓	✓	✓	✓	✓	✓	x
20	Pareto Chart	✓	✓	✓	✓	✓	✓	✓	✓
21	Funnel Chart	✓	✓	✓	✓	✓	✓	✓	✓
22	Surface Chart	✓	✓	✓	✓	✓	✓	✓	✓
23	Geographical Mapping	✓	✓	✓	✓	✓	✓	✓	✓
24	Tree Diagram/Dendogram	✓	✓	✓	✓	✓	✓	✓	✓
25	Chernoff Faces	✓	x	x	x	x	x	x	x
26	Stick Figures	✓	✓	✓	x	x	x	x	x
27	Heat Map	✓	✓	✓	✓	✓	✓	✓	✓

Data Visualization: Challenges

Choosing the Right Visualization: selecting the right one requires an understanding of the data and the message that needs to be conveyed.

Data Quality: requires high-quality data. Inaccurate, incomplete, or inconsistent data can lead to misleading or incorrect visualizations.

Data Overload: handling large and complex datasets.

Over-Emphasis on Aesthetics: While aesthetics are important, overemphasizing the visual appeal of the visualization at the expense of accuracy and effectiveness can be problematic.

Audience Understanding: ensuring that the target audience can interpret and understand the visualizations. Visualizations should be designed with the audience in mind and should be clear and concise.

Technical Expertise: Data analysts and data scientists need to be familiar with programming languages, visualization tools, and statistical concepts to create effective visualizations.

Copyright@ 2024 ISI SQCOR Kolkata

References:

Andrews (1972), Plots of High-Dimensional Data

Chernoff (1973), The use of faces to represent points in k-dimensional space graphically

E.R. Tufte (1983), The Visual Display of Quantitative Information

Becker, R. A. and Cleveland, W. S. (1996a). The design and control of Trellis display

Becker, R. A. and Cleveland, W. S. (1996b). Trellis Graphics User's Manual

Fisherkeller, M. A., Friedman, J. H. and Tukey, J. W. (1988). PRIM-9: An interactive multidimensional data display and analysis system