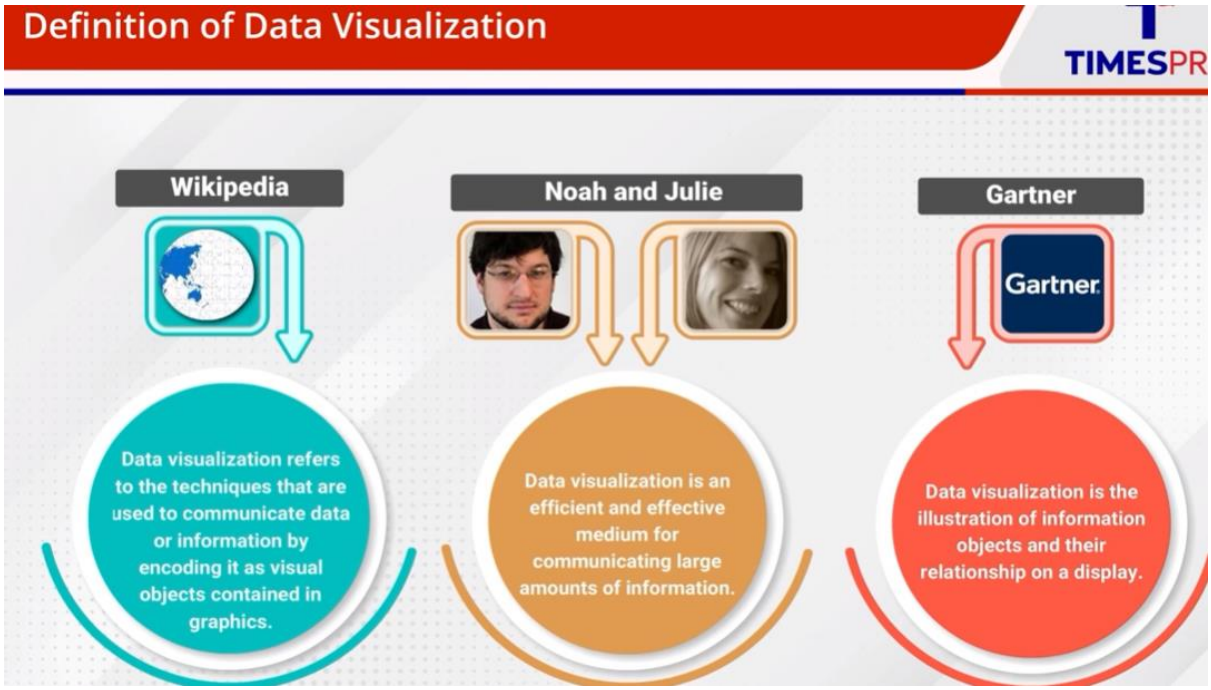
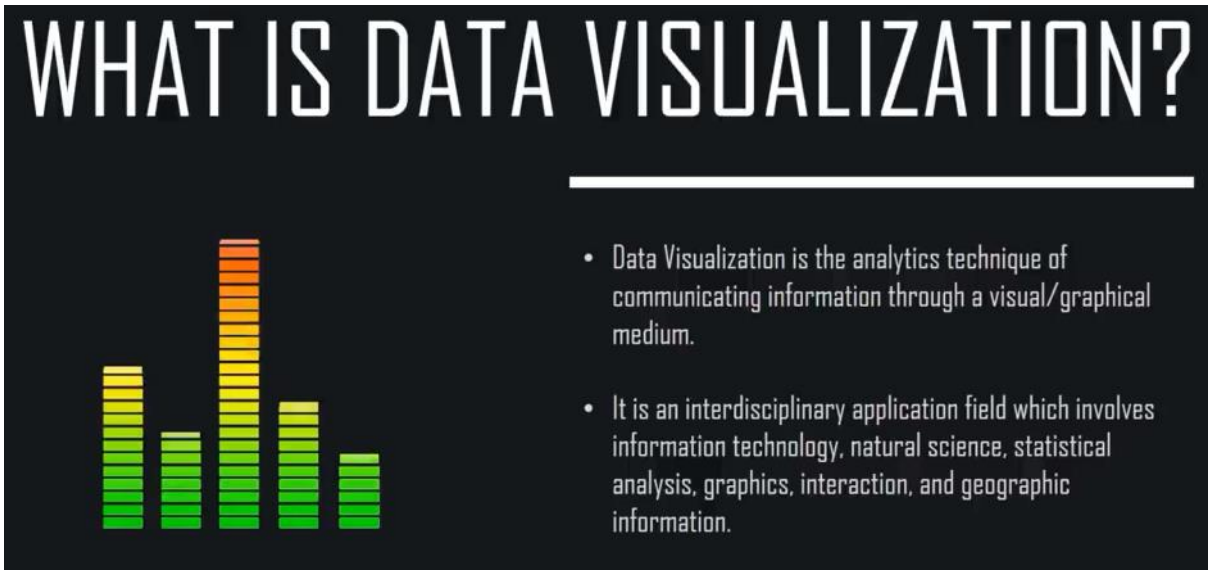
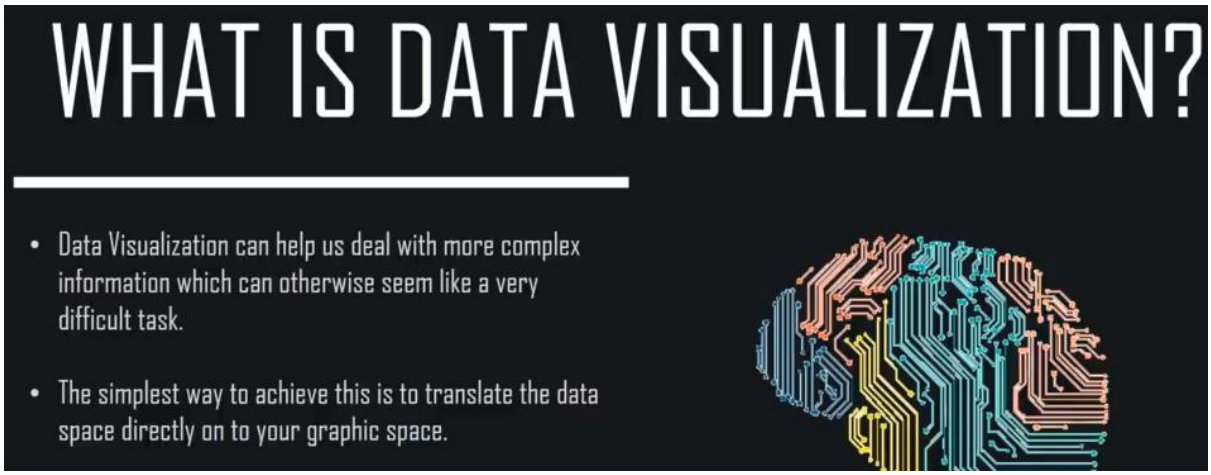
DMV UNIT 5


Definition of Data Visualization
TIMESPR
Wikipedia
Noah and Julie
Gartner
Gartner
Data visualization refers to the techniques that are used to communicate data or information by encoding it as visual objects contained in graphics.
Data visualization is an efficient and effective medium for communicating large amounts of information.
Data visualization is the illustration of information objects and their relationship on a display.


WHAT IS DATA VISUALIZATION?
• Data Visualization is the analytics technique of communicating information through a visual/graphical medium.
• It is an interdisciplinary application field which involves information technology, natural science, statistical analysis, graphics, interaction, and geographic information.

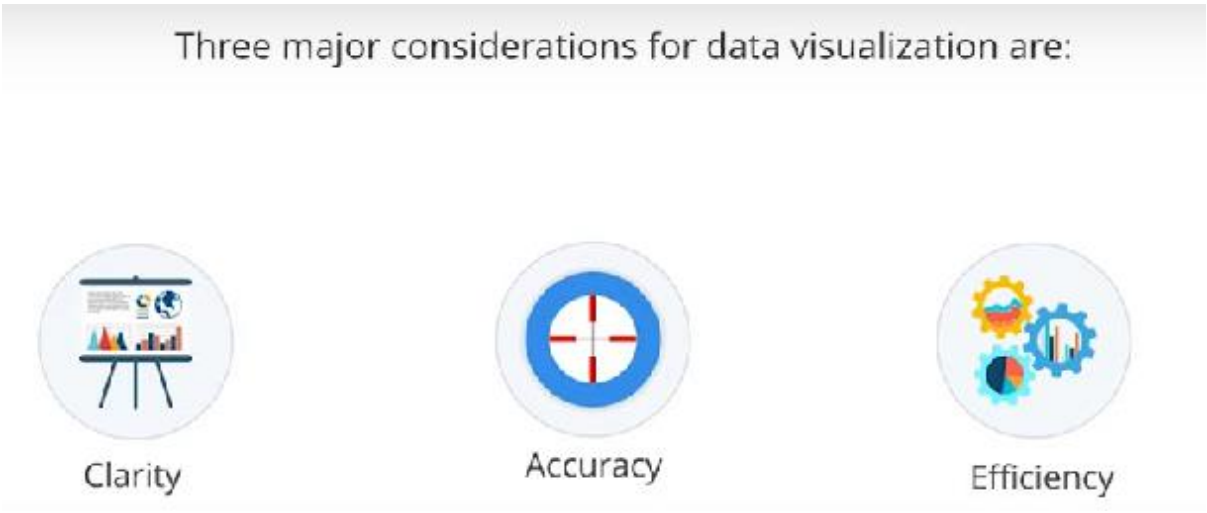

WHAT IS DATA VISUALIZATION?
• Data Visualization can help us deal with more complex information which can otherwise seem like a very difficult task.
• The simplest way to achieve this is to translate the data space directly on to your graphic space.

Why is Data Visualization Important?

- 1. Data Visualization Discovers the Trends in Data
- 2. Data Visualization Provides a Perspective on the Data
- 3. Data Visualization Puts the Data into the Correct Context
- 4. Data Visualization Saves Time
- 5. Data Visualization Tells a Data Story

the benefits of good data visualization:

1. **Simplifies Complexity**: Distills complex data into easy-to-understand visuals.
2. **Accelerates Comprehension**: Enables quick understanding by processing visuals faster than text.
3. **Informs Decisions**: Supports more informed and effective decision-making.
4. **Reveals Patterns**: Identifies trends and patterns not obvious in raw data.
5. **Enhances Communication**: Facilitates clearer data-driven communication to stakeholders.
6. **Engages Audience**: Captures and maintains attention better than text alone.
7. **Improves Retention**: Increases long-term memory retention of information.
8. **Provides Quick Reference**: Serves as an efficient reference for important data.
9. **Broadens Accessibility**: Makes data comprehensible to a wider audience.
10. **Aids Cross-Disciplinary Understanding**: Bridges gaps between different fields through a common visual language.
11. **Saves Time**: Reduces analysis time, enhancing workflow efficiency.
12. **Automates Reporting**: Allows for automated generation of visual reports.
13. **Facilitates Exploration**: Enables dynamic and interactive data exploration.
14. **Detects Errors**: Highlights data inaccuracies and inconsistencies.
15. **Persuades and Motivates**: Influences opinions and drives actions through compelling visuals.



the nature of data visualization, described concisely:

1. **Clarity**: Presents data clearly and avoids ambiguity.
2. **Accuracy**: Reflects data truthfully without distortion.
3. **Simplicity**: Uses minimalism to convey information without overcomplication.
4. **Relevance**: Aligns with the audience's needs and highlights key data aspects.
5. **Aesthetics**: Ensures visual appeal with consistent design elements.
6. **Interactivity**: Engages users with features like zooming and filtering.
7. **Functionality**: Serves practical purposes and ensures ease of use.
8. **Scalability**: Adapts to varying data sizes and devices.
9. **Accessibility**: Considers all users, ensuring readability and inclusivity.
10. **Narrative**: Guides the viewer through a data-driven story with logical flow.
11. **Comparability**: Uses consistent scales and units for easy comparison.
12. **Transparency**: Clearly indicates data sources and methodologies.
13. **Timeliness**: Uses up-to-date data and offers real-time updates when applicable.
14. **Multidimensionality**: Visualizes multiple data dimensions and layers.
15. **Customization**: Offers options for user preferences and flexible formats.

Depending on the number of variables used for plotting the visualization and the type of variables, there could be different types of charts which we could use to understand the relationship. Based on the count of variables, we could have

- *Univariate* plot(involves only one variable)
- *Bivariate* plot(more than one variable in required)

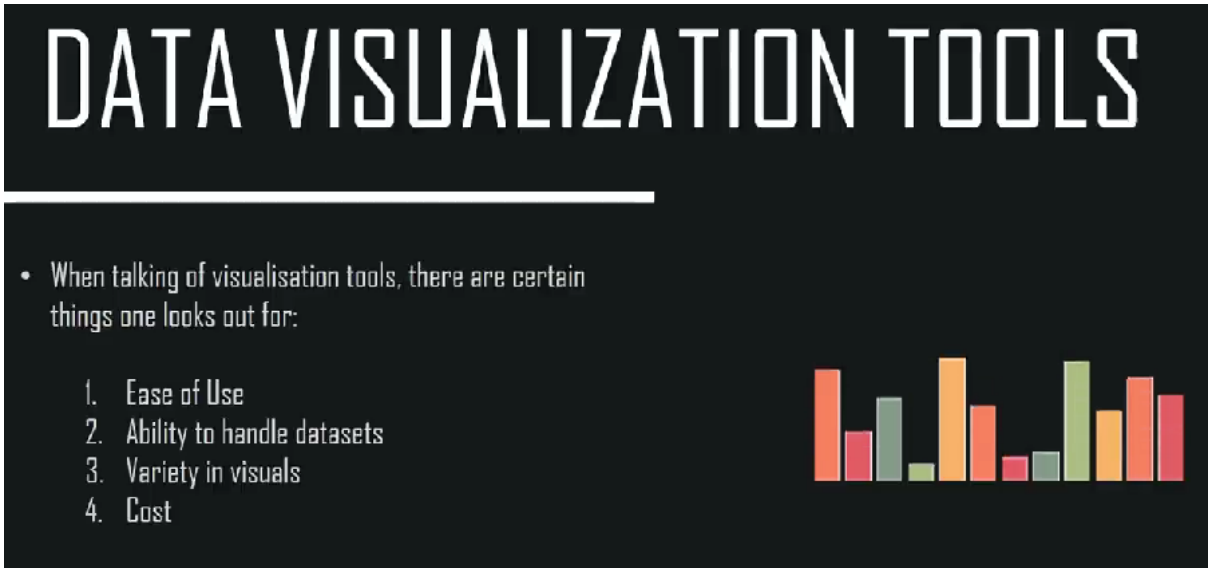
A *Univariate* plot could be for a continuous variable to understand the spread and distribution of the variable while for a discrete variable it could tell us the count Similarly, a *Bivariate* plot for continuous variable could display essential statistic like correlation, for a continuous versus discrete variable could lead us to very important conclusions like understanding data distribution across different levels of a categorical variable. A *bivariate* plot between two discrete variables could also be developed.

Advantages of Data Visualization:

- **Enhanced Comparison:** Visualizing performances of two elements or scenarios streamlines analysis, saving time compared to traditional data examination.
- **Improved Methodology:** Representing data graphically offers a superior understanding of situations, exemplified by tools like Google Trends illustrating industry trends in graphical forms.
- **Efficient Data Sharing:** Visual data presentation facilitates effective communication, making information more digestible and engaging compared to sharing raw data.
- **Sales Analysis:** Data visualization aids sales professionals in comprehending product sales trends, identifying influencing factors through tools like heat maps, and understanding customer types, geography impacts, and repeat customer behaviors.
- **Identifying Event Relations:** Discovering correlations between events helps businesses understand external factors affecting their performance, such as online sales surges during festive seasons.
- **Exploring Opportunities and Trends:** Data visualization empowers business leaders to uncover patterns and opportunities within vast datasets, enabling a deeper understanding of customer behaviors and insights into emerging business trends.

Disadvantages of Data Visualization:

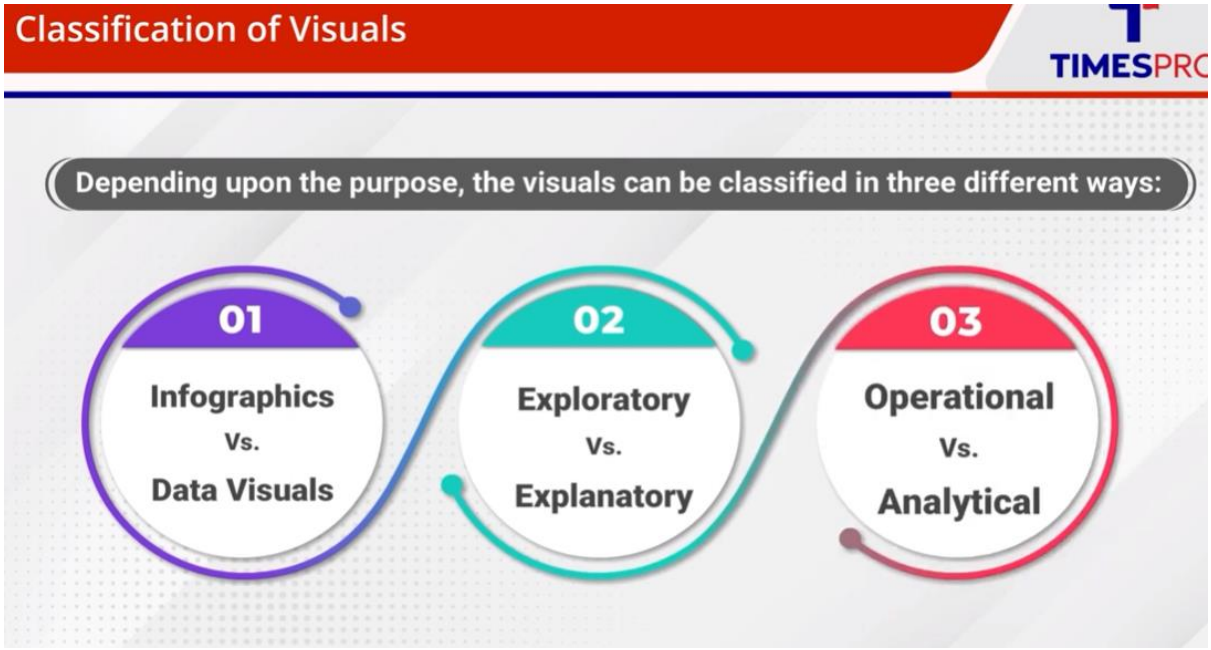
- **Can be time-consuming:** Creating visualizations can be a time-consuming process, especially when dealing with large and complex datasets.
- **Can be misleading:** While data visualization can help identify patterns and relationships in data, it can also be misleading if not done correctly. Visualizations can create the impression of patterns or trends that may not exist, leading to incorrect conclusions and poor decision-making.
- **Can be difficult to interpret:** Some types of visualizations, such as those that involve 3D or interactive elements, can be difficult to interpret and understand.
- **May not be suitable for all types of data:** Certain types of data, such as text or audio data, may not lend themselves well to visualization. In these cases, alternative methods of analysis may be more appropriate.
- **May not be accessible to all users:** Some users may have visual impairments or other disabilities that make it difficult or impossible for them to interpret visualizations. In these cases, alternative methods of presenting data may be necessary to ensure accessibility.



Tools for Visualization of Data

The following are the 10 best Data Visualization Tools

1. Tableau
2. Looker
3. Zoho Analytics
4. Sisense
5. IBM Cognos Analytics
6. Qlik Sense
7. Domo
8. Microsoft Power BI
9. Klipfolio
10. SAP Analytics Cloud



1. Infographics

Infographics are designed to present information quickly and clearly through a combination of visuals and text. They are often used for storytelling, marketing, or educational purposes.

- **Purpose**: Simplify complex information, engage audience, and tell a story.
- **Examples**:
  - **Static Infographics**: A visual representation combining images, charts, and minimal text to explain a concept or tell a story.
  - **Process Infographics**: Show steps in a process or workflow.

2. Data Visuals

Data visuals refer to any visual representation of data intended to communicate information clearly and effectively.

- **Purpose**: Represent data to make it easier to understand and interpret.
- **Examples**:
  - **Bar Charts**: Compare different categories.
  - **Line Charts**: Show trends over time.
  - **Pie Charts**: Display parts of a whole.

3. **Exploratory Visuals**

Exploratory visualizations are used for data exploration, allowing users to interact with the data to uncover patterns, relationships, and insights.

- **Purpose**: Facilitate data exploration and discovery.
- **Examples**:
  - **Interactive Dashboards**: Allow users to filter, drill down, and interact with data.
  - **Scatter Plot Matrices**: Compare multiple variables simultaneously.
  - **Heat Maps**: Visualize data density or intensity.

4. **Explanatory Visuals**

Explanatory visualizations are designed to convey specific insights or findings from the data. They focus on clear communication of results and are often used in presentations and reports.

- **Purpose**: Communicate insights and findings clearly.
- **Examples**:
  - **Annotated Charts**: Charts with notes and highlights explaining key points.
  - **Storytelling Dashboards**: Structured to guide the viewer through a narrative.
  - **Case Study Visuals**: Specific examples illustrating broader trends.

5. **Operational Visuals**

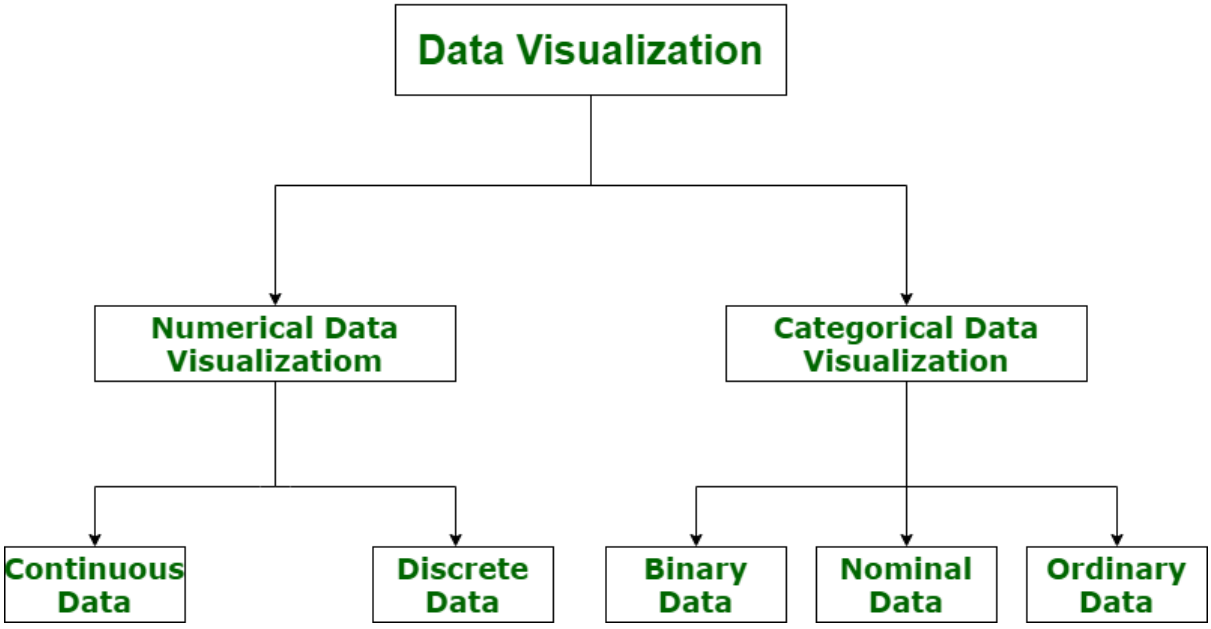
Operational visualizations are used for real-time monitoring and management of operations. They display live data to help in decision-making and performance tracking.

- **Purpose**: Monitor and manage ongoing operations.
- **Examples**:
  - **Real-Time Dashboards**: Display live data feeds, often used in operations centers.
  - **Gantt Charts**: Track project timelines and task progress.
  - **KPIs Dashboards**: Track key performance indicators in real-time.

6. **Analytical Visuals**

Analytical visualizations are used to support detailed analysis and deep dives into data. They help analysts to understand complex datasets, identify trends, and make data-driven decisions.

- **Purpose**: Facilitate in-depth data analysis.
- **Examples**:
  - **Histograms**: Show the distribution of data.
  - **Box Plots**: Display statistical summaries and identify outliers.
  - **Regression Analysis Charts**: Explore relationships between variables.



What is Data Visualization?

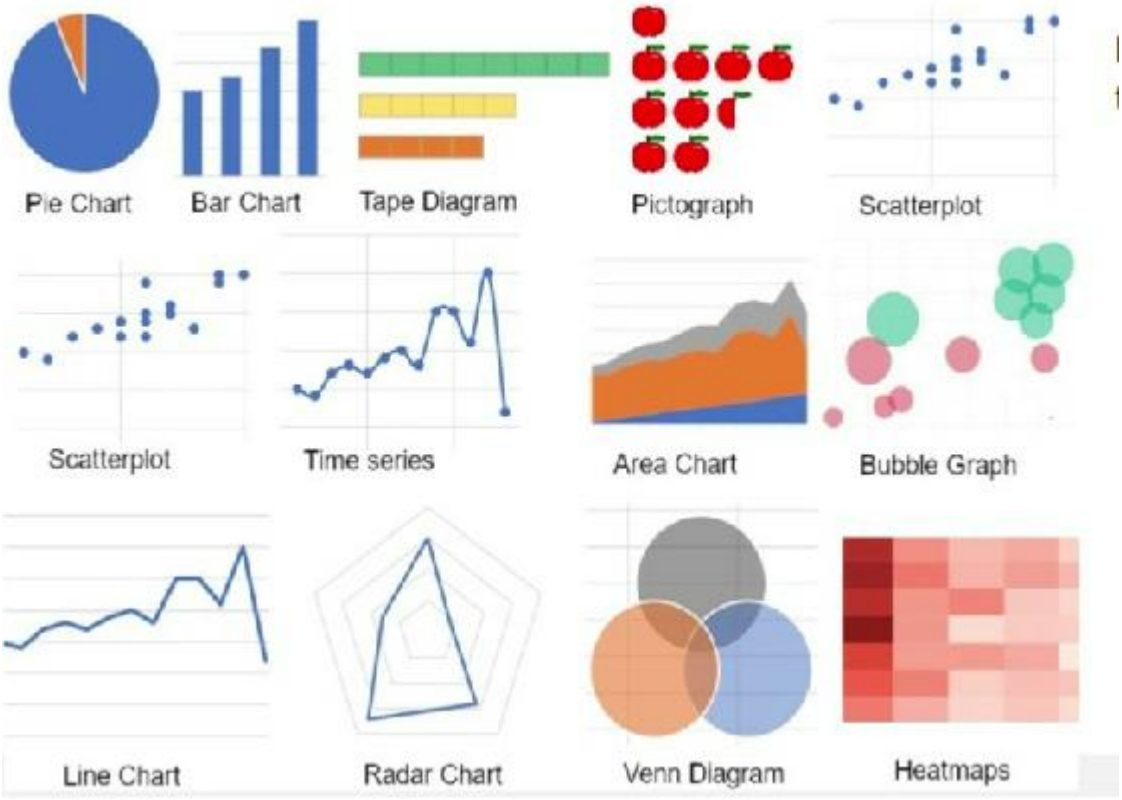
Data visualization translates complex data sets into visual formats that are easier for the human brain to comprehend. This can include a variety of visual tools such as:

- **Charts**: Bar charts, line charts, pie charts, etc.
- **Graphs**: Scatter plots, histograms, etc.
- **Maps**: Geographic maps, heat maps, etc.
- **Dashboards**: Interactive platforms that combine multiple visualizations.

Types of Data Visualization Techniques

Various types of visualizations cater to diverse data sets and analytical goals.

1. **Bar Charts:** Ideal for comparing categorical data or displaying frequencies, bar charts offer a clear visual representation of values.
2. **Line Charts:** Perfect for illustrating trends over time, line charts connect data points to reveal patterns and fluctuations.
3. **Pie Charts:** Efficient for displaying parts of a whole, pie charts offer a simple way to understand proportions and percentages.
4. **Scatter Plots:** Showcase relationships between two variables, identifying patterns and outliers through scattered data points.
5. **Histograms:** Depict the distribution of a continuous variable, providing insights into the underlying data patterns.
6. **Heatmaps:** Visualize complex data sets through color-coding, emphasizing variations and correlations in a matrix.
7. **Box Plots:** Unveil statistical summaries such as median, quartiles, and outliers, aiding in data distribution analysis.
8. **Area Charts:** Similar to line charts but with the area under the line filled, these charts accentuate cumulative data patterns.
9. **Bubble Charts:** Enhance scatter plots by introducing a third dimension through varying bubble sizes, revealing additional insights.
10. **Treemaps:** Efficiently represent hierarchical data structures, breaking down categories into nested rectangles.
11. **Violin Plots**: Violin plots combine aspects of box plots and kernel density plots, providing a detailed representation of the distribution of data.
12. **Word Clouds**: Word clouds are visual representations of text data where words are sized based on their frequency.
13. **3D Surface Plots**: 3D surface plots visualize three-dimensional data, illustrating how a response variable changes in relation to two predictor variables.
14. **Network Graphs**: Network graphs represent relationships between entities using nodes and edges. They are useful for visualizing connections in complex systems, such as social networks, transportation networks, or organizational structures.
15. **Sankey Diagrams**: Sankey diagrams visualize flow and quantity relationships between multiple entities. Often used in process engineering or energy flow analysis.

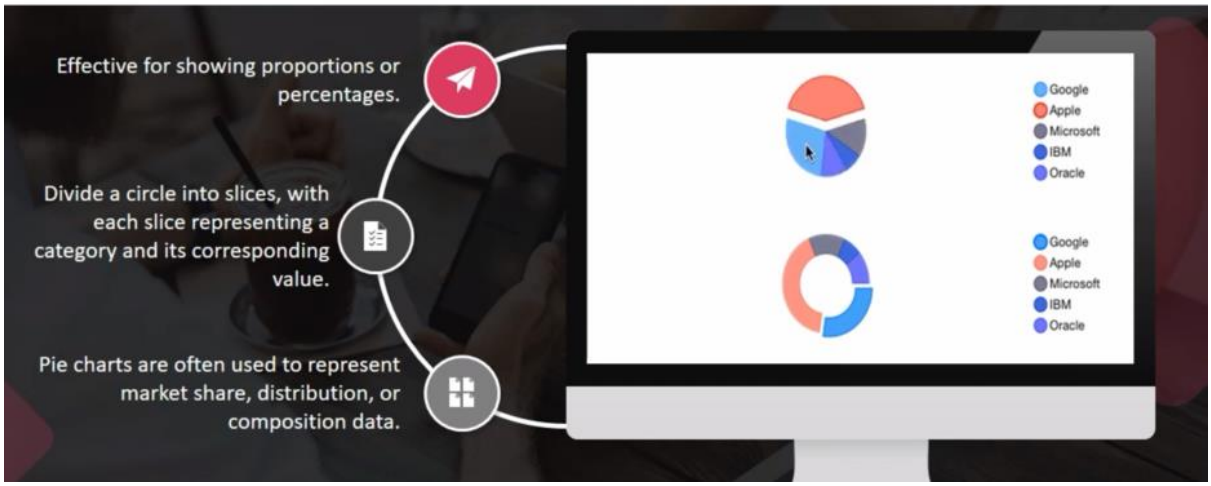


BAR CHARTS & COLUMN CHARTS

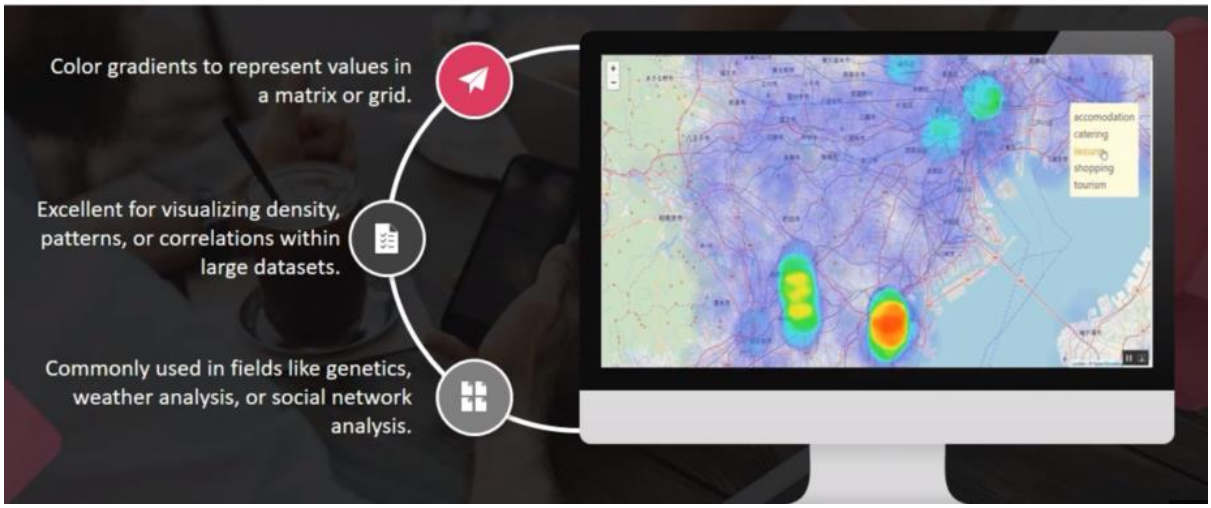
LINE PLOTS AND TIME SERIES

SCATTER PLOTS

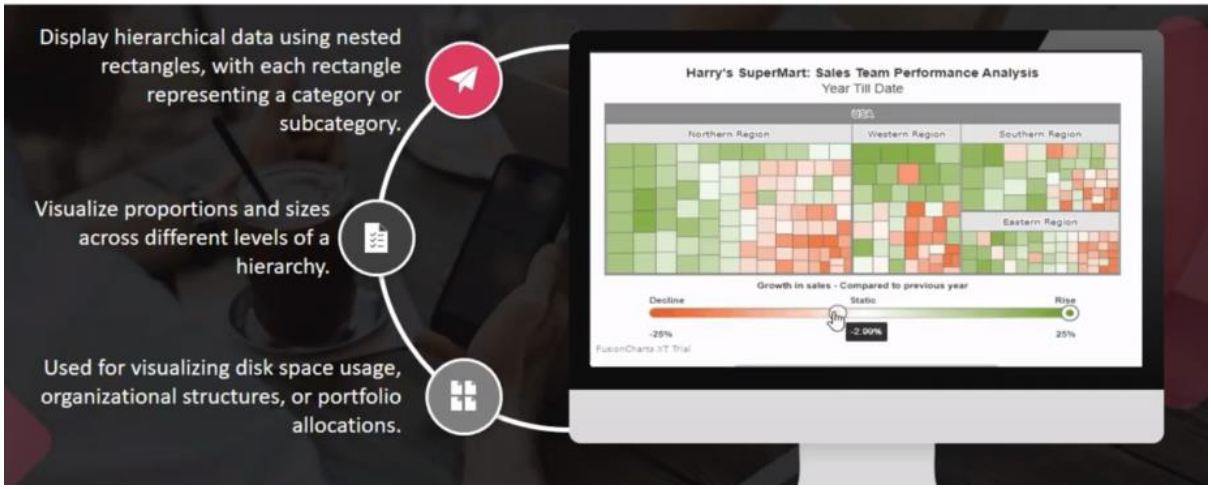
PIE CHARTS AND DONUT CHARTS



HEATMAPS

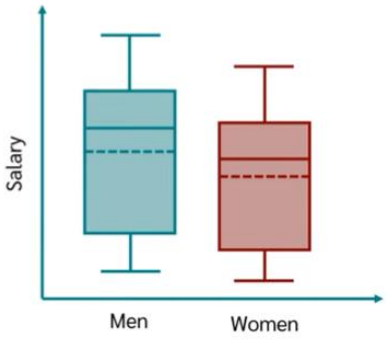


TREE MAPS

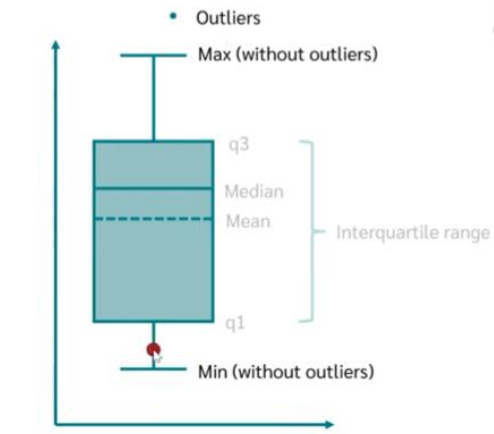


Boxplots

- Box plots are used in statistics to graphically display various parameters at a glance.
- In a boxplot the median, the interquartile range and the outliers can be read.
- The data must have metric scale level.
- A boxplot is often created to compare and contrast two or more groups.



The box indicates the range in which the middle 50% of all data lies.



Thus, the lower end of the box is the 1st quartile and the upper end is the 3rd quartile.

Between q1 and q3, the interquartile range is thus

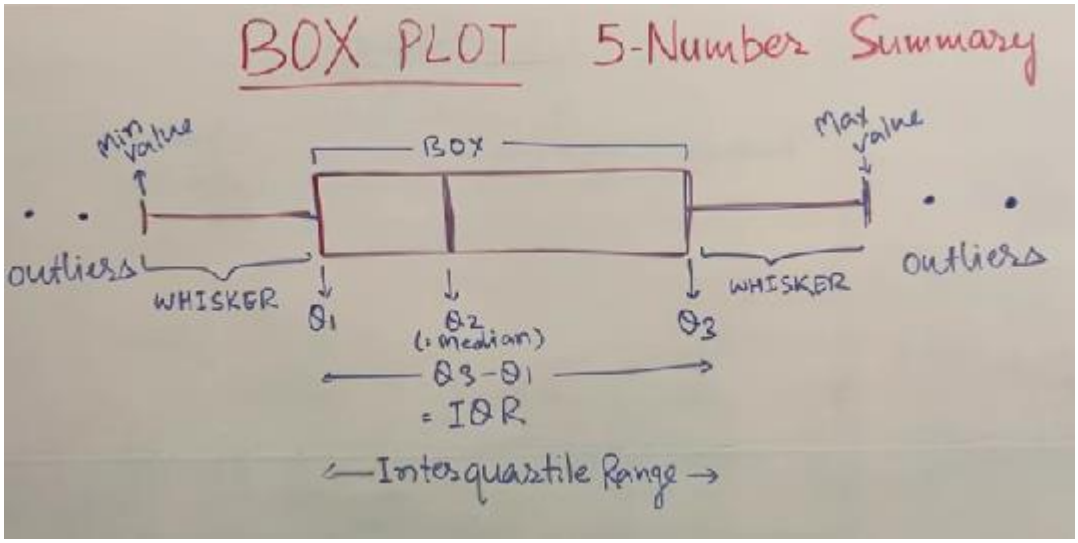
In the boxplot, the solid line indicates the median and the dashed line indicates the mean.

The T-shaped whiskers go to the last point, which is still within 1.5 times the interquartile range.

Points that are further away are considered outliers.

Suppose ek range given hai
the age between 21 to 50 years we have to choose from a given set then here the below q1 range will represent the 25 % and above q3 the other 25% and the box will represent the 50%

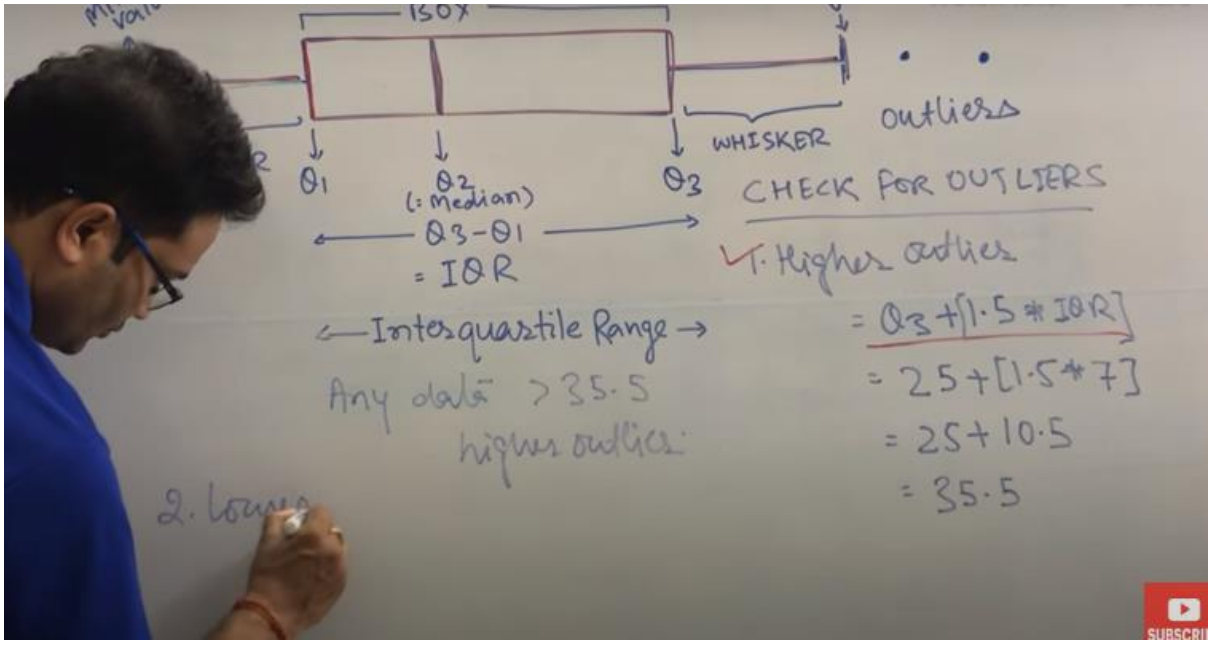
Supposed
below q1 is below 21 years and above q3 is above 50

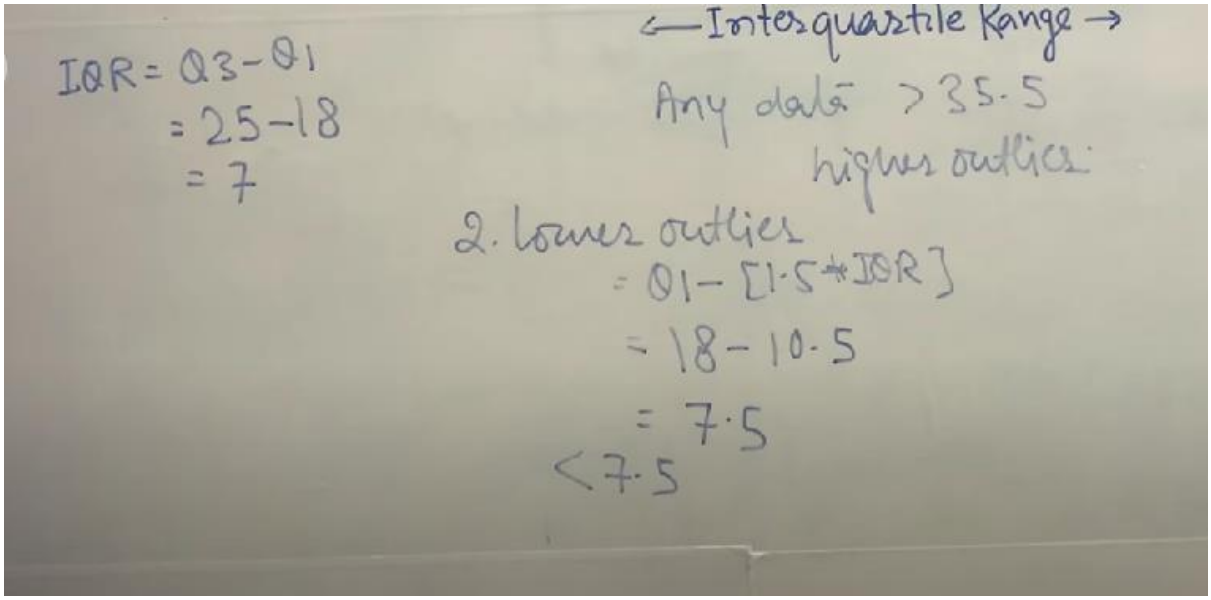


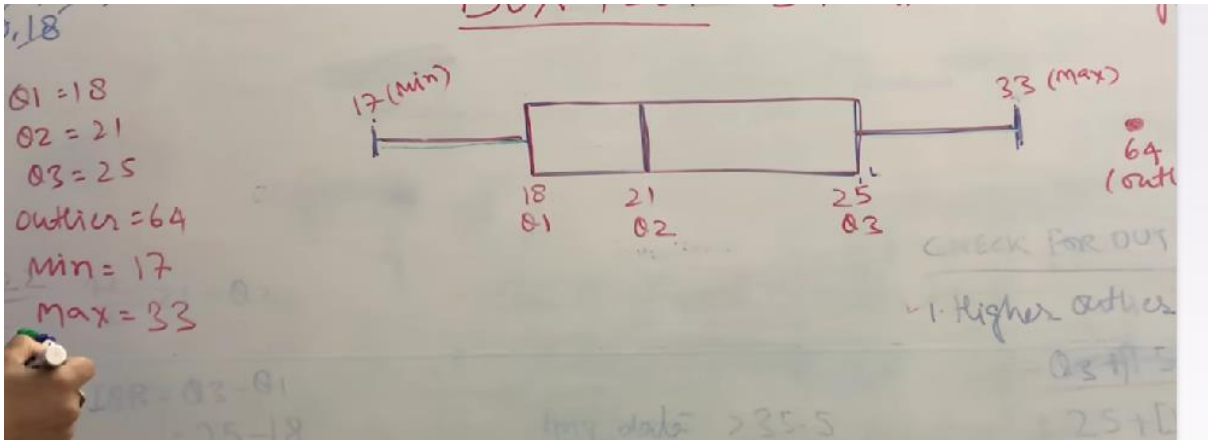
22, 28, 17, 19, 33, 64,
23, 17, 20, 18

17
17
18 — Q1
19
20
22 — $\frac{20+22}{2} = \frac{42}{2} = 21 = Q2$
23
25 — Q3
33
64

outl





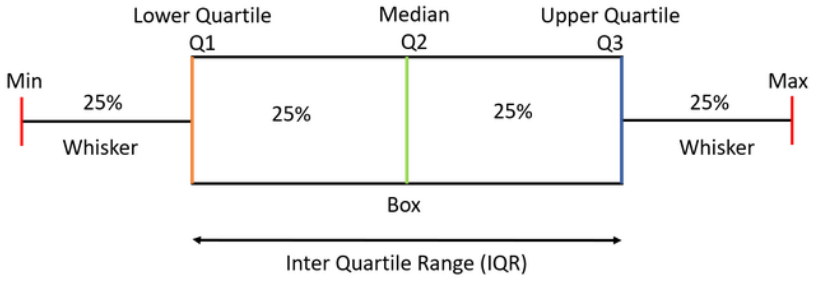


Box plot is also known as a whisker plot, box-and-whisker plot, or simply a box-and whisker diagram. Box plot is a graphical representation of the distribution of a dataset. It displays key summary statistics such as the median, quartiles, and potential outliers in a concise and visual manner. By using Box plot you can provide a summary of the distribution, identify potential and compare different datasets in a compact and visual manner.

Elements of Box Plot

A box plot gives a five-number summary of a set of data which is-

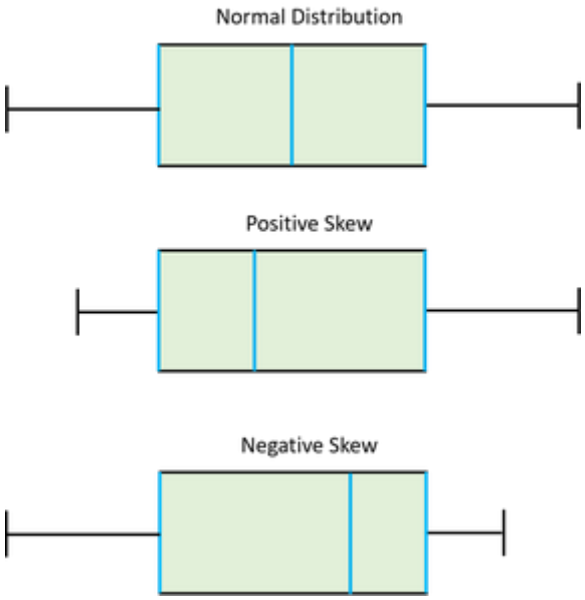
- **Minimum** – It is the minimum value in the dataset excluding the outliers.
- **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
- **Median (Q2)** – It is the mid-point of the dataset. Half of the values lie below it and half above.
- **Third Quartile (Q3)** – 75% of the data lies below the Third (Upper) Quartile.
- **Maximum** – It is the maximum value in the dataset excluding the outliers.



Use-Cases of Box Plot

- Box plots provide a visual summary of the data with which we can quickly identify the average value of the data, how dispersed the data is, whether the data is skewed or not (skewness).
- The Median gives you the average value of the data.
- Box Plots shows Skewness of the data-

- **a)** If the Median is at the **center** of the Box and the **whiskers** are almost the
- **same on both the ends** then the data is **Normally Distributed.**
- **b)** If the Median lies **closer to the First Quartile** and if the **whisker at the lower**
- **end is shorter** (as in the above example) then it has a **Positive Skew (Right Skew).**
- **c)** If the Median lies **closer to the Third Quartile** and if the **whisker at the**
- **upper end is shorter** than it has a **Negative Skew (Left Skew).**

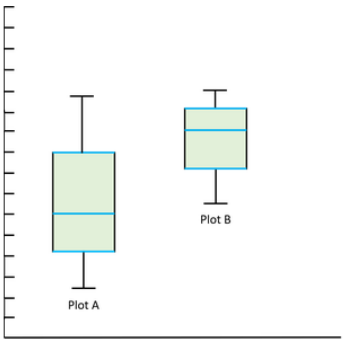


- The dispersion or spread of data can be visualized by the minimum and maximum values which are found at the end of the whiskers.
- The Box plot gives us the idea of about the Outliers which are the points which are numerically distant from the rest of the data.

How to compare box plots?

As we have discussed at the beginning of the article that box plots make comparing characteristics of data between categories very easy. Let us have a look at how we can compare different box plots and derive statistical conclusions from them.

Let us take the below two plots as an example: –

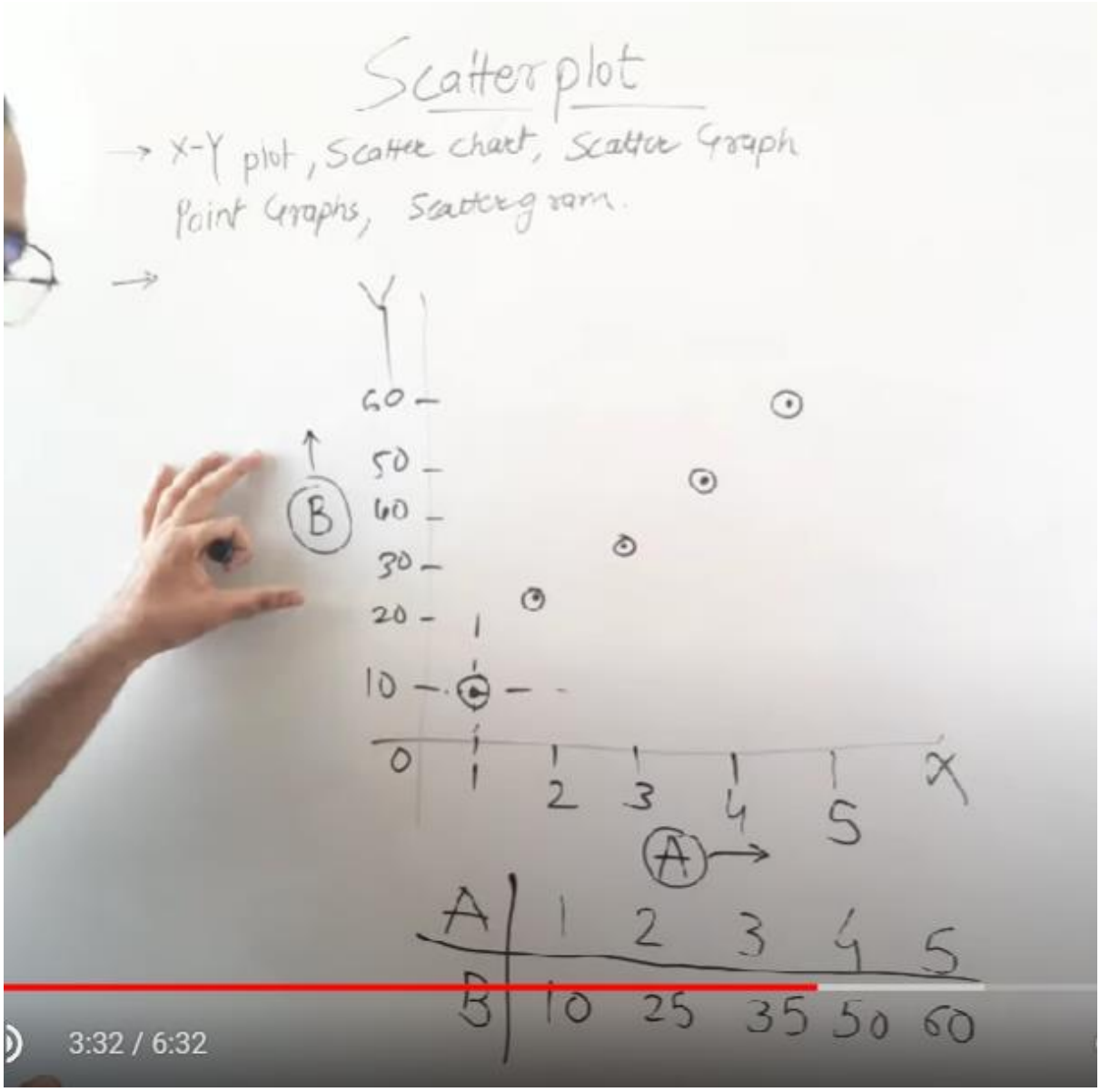


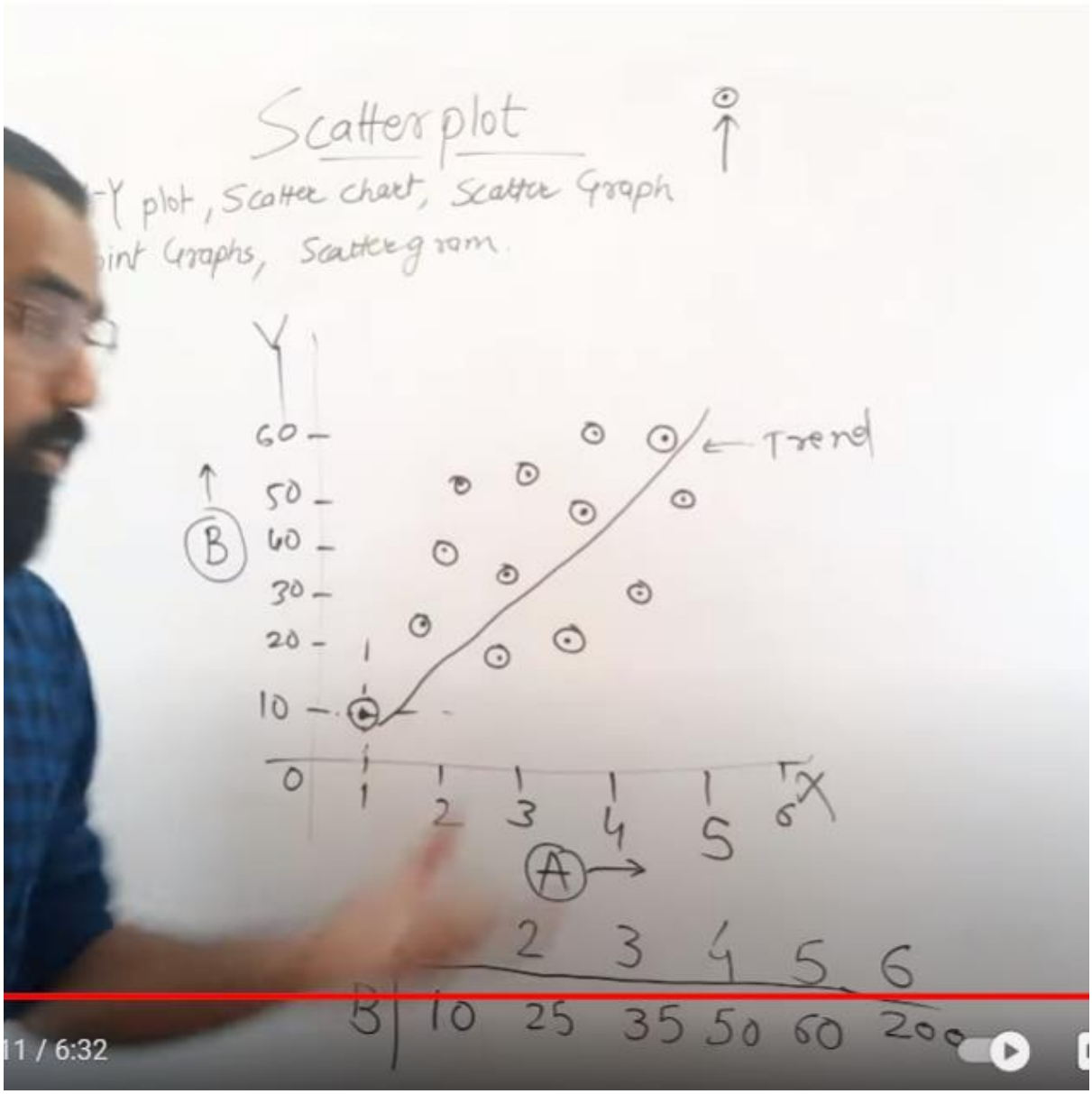
- **Compare the Medians** — If the median line of a box plot lies outside the box of the other box plot with which it is being compared, then we can say that there is likely to be a difference between the two groups. Here the Median line of the plot B lies outside the box of Plot A.
- **Compare the Dispersion or Spread of data** — The Inter Quartile range (length of the box) gives us an idea about how dispersed the data is. Here Plot A has a longer length than Plot B which means that the dispersion of data is more in plot A as compared to plot B. The length of whiskers also gives an idea of the overall spread of data. The extreme values (minimum &maximum) give the range of data distribution. Larger the range more scattered the data. Here Plot A has a larger range than Plot B.
- **Comparing Outliers** — The outliers give the idea of unusual data values which are distant from the rest of the data. More number of Outliers means the prediction will be more uncertain. We can be more confident while predicting the values for a plot which has less or no outliers.
- **Compare Skewness** — Skewness gives us the direction and the magnitude of the lack of symmetry. We have discussed above how to identify skewness. Here Plot A is Positive or Right Skewed and Plot B is Negative or Left Skewed.

Box plots are a useful tool due to several key benefits:

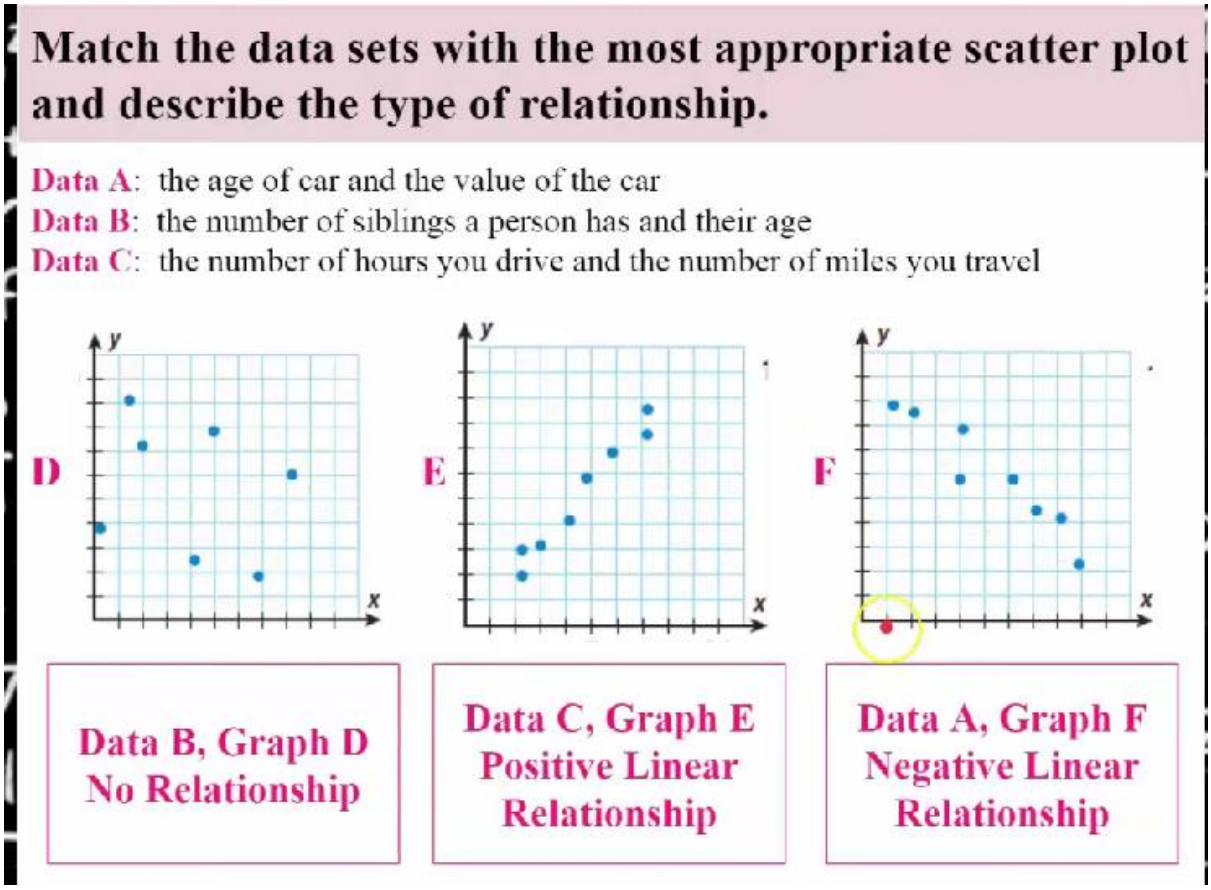
1. **Summarizes Data Distribution:** Provides a clear summary of the distribution, including median, quartiles, and potential outliers at a glance.
2. **Identifies Outliers:** Clearly marks outliers beyond the whiskers, helping to identify data points that may skew interpretation.
3. **Compares Data Sets:** Facilitates quick visual comparison between multiple datasets or different groups within a single dataset.
4. **Shows Skewness and Symmetry:** Indicates whether the data is symmetrically distributed or skewed to one side.
5. **Handles Large Datasets:** Effective for visualizing distributions of large datasets or datasets with many variables.
6. **No Assumptions About Distribution:** Does not assume a particular distribution of the data, making it versatile for various types of data.
7. **Intuitive Interpretation:** Easy to interpret for both technical and non-technical audiences, aiding in communication of findings.
8. **Complements Other Visualizations:** Can be used alongside histograms or density plots to provide a more comprehensive view of the data.
9. **Useful in Exploratory Data Analysis:** Valuable in exploratory data analysis (EDA) to understand the spread and central tendency of the data quickly.
10. **Supports Decision Making:** Helps in making data-driven decisions by providing insights into data variability and central tendency effectively.

SCATTER PLOTS





trend line or line of best fit



Scatter plot is one of the most important data visualization techniques and it is considered one of the Seven Basic Tools of Quality. A scatter plot is used to plot the relationship between two variables, on a two-dimensional graph that is known as Cartesian Plane on mathematical grounds.

It is generally used to plot the relationship between one independent variable and one dependent variable, where an independent variable is plotted on the x-axis and a dependent variable is plotted on the y-axis so that you can visualize the effect of the independent variable on the dependent variable. These plots are known as Scatter Plot Graph or Scatter Diagram.

Applications of Scatter Plot

As already mentioned, a scatter plot is a very useful data visualization technique. A few applications of Scatter Plots are listed below.

- **Correlation Analysis**: Scatter plot is useful in the investigation of the correlation between two different variables. It can be used to find out whether two variables have a positive correlation, negative correlation or no correlation.
- **Outlier Detection**: Outliers are data points, which are different from the rest of the data set. A Scatter Plot is used to bring out these outliers on the surface.
- **Cluster Identification**: In some cases, scatter plots can help identify clusters or groups within the data.

How to Construct a Scatter Plot?

To construct a scatter plot, we have to follow the given steps.
Step 1: Identify the independent and dependent variables
Step 2: Plot the independent variable on x-axis
Step 3: Plot the dependent variable on y-axis
Step 4: Extract the meaningful relationship between the given variables.

Types of Scatter Plot

On the basis of correlation of two variables, Scatter Plot can be classified into following types.

- Scatter Plot For Positive Correlation
- Scatter Plot For Negative Correlation
- Scatter Plot For Null Correlation

Scatter Plot For Positive Correlation

In this type of scatter-plot, value on y-axis increases on moving left to right. In more technical terms, if one variable is directly proportional to another, then, the scatter plot will show positive correlation. Positive correlation can be further classified into Perfect Positive, High Positive and Low Positive.

Scatter Plot For Negative Correlation

In this type of scatter-plot, value on the y-axis decreases on moving left to right. In other words, the value of one variable is decreasing with respect to the other. Positive correlation can be further classified into Perfect Negative, High Negative and Low Negative.

Scatter Plot For Null Correlation

In this type of scatter-plot, values are scattered all over the graph. Generally this kind of graph represents that there is no relationship between the two variables plotted on the Scatter Plot.

Solved Examples on Scatter Plot

Example 1: Draw a scatter plot for the given data that shows the number of IPL matches played and runs scored in each instance.

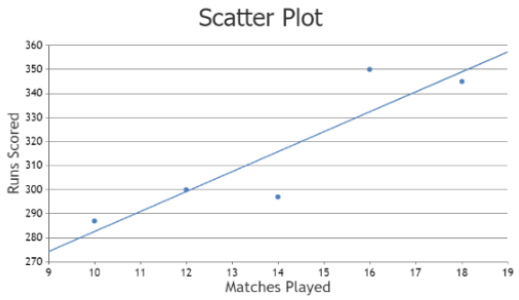
Matches Played	10	12	14	16	18
Runs Scored	287	300	297	350	345

Solution:

X-axis: Number of Matches Played

Y-axis: Number of Runs Scored

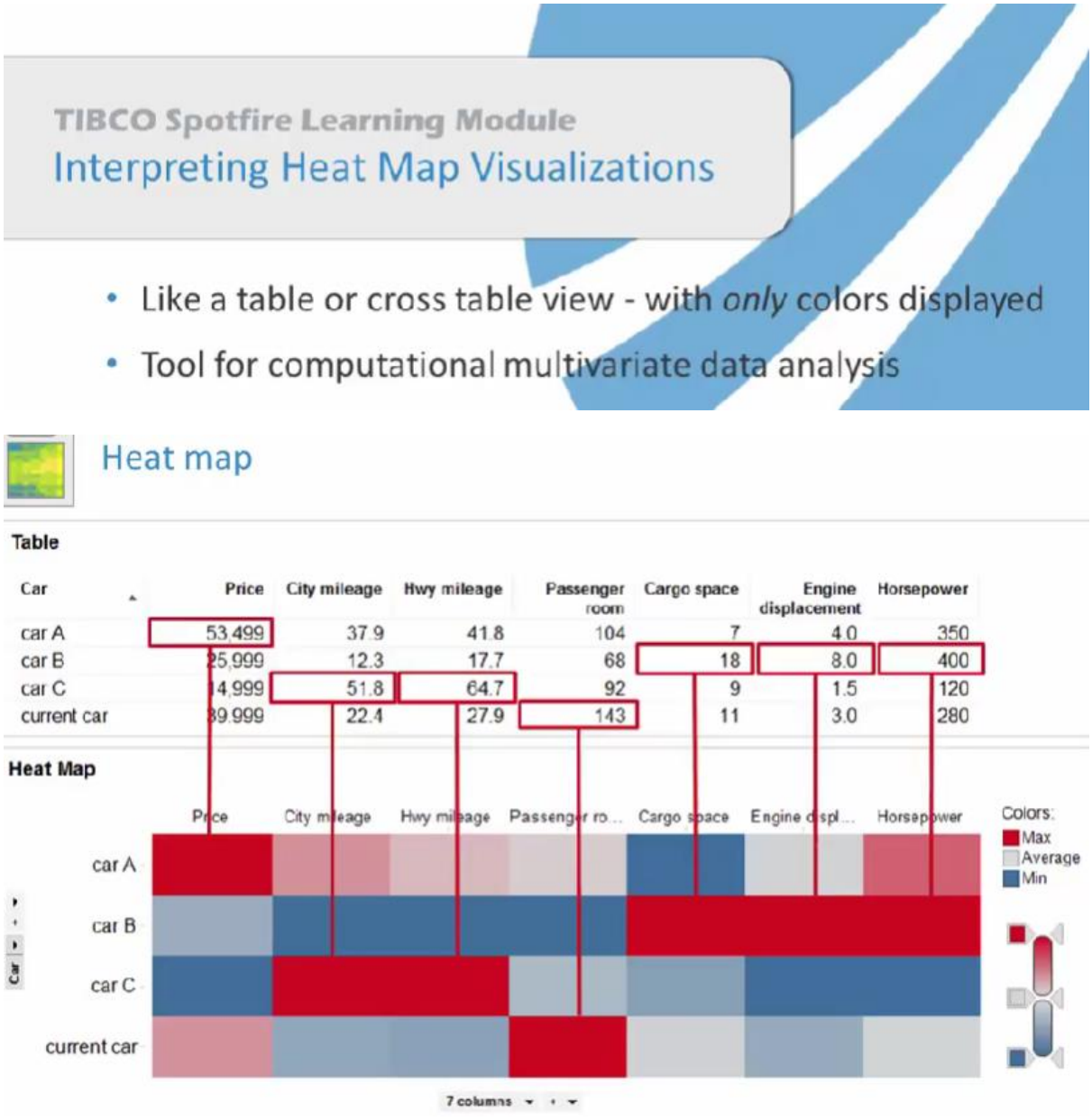
Graph:

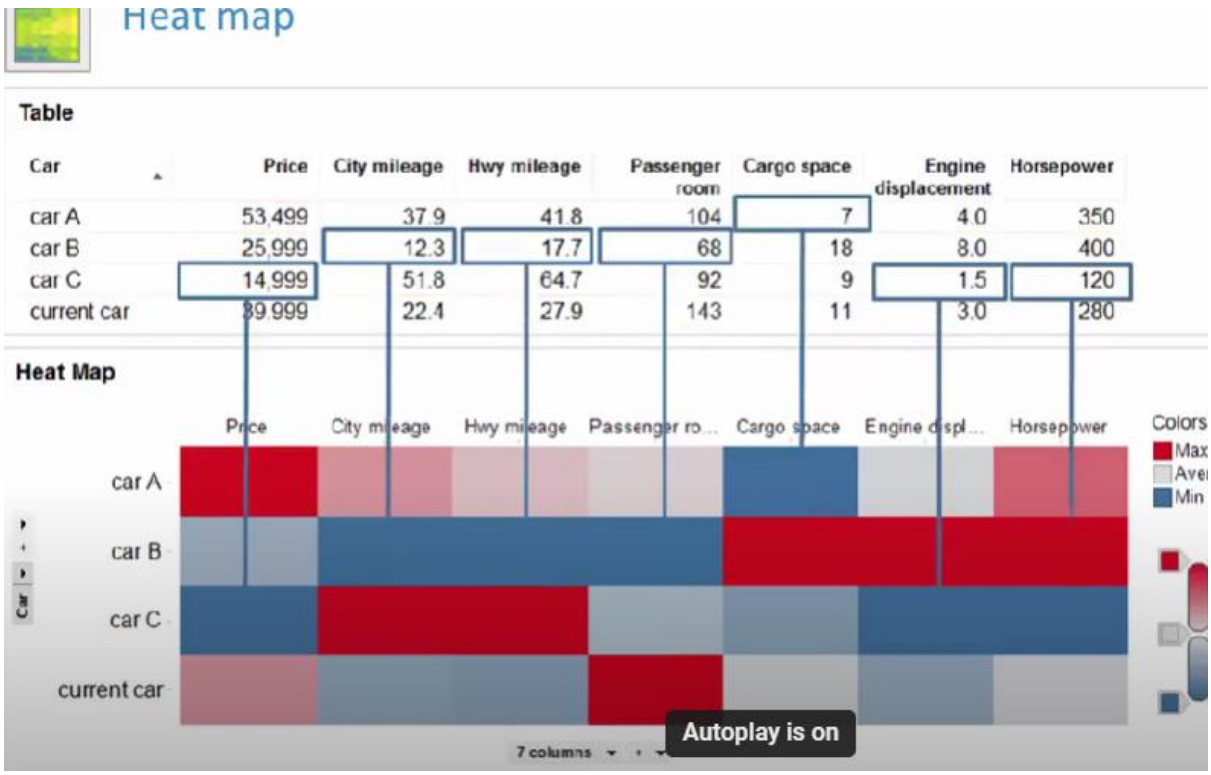


Scatter plots offer several benefits in visualizing data relationships:

1. **Visualizes Relationships:** Clearly displays the relationship between two variables, showing patterns, trends, and correlations.
2. **Identifies Clusters:** Highlights clusters or groups within the data, indicating potential subpopulations or patterns.
3. **Detects Outliers:** Easily identifies outliers or unusual data points that may affect analysis or modeling.
4. **Assesses Correlation:** Provides a quick visual assessment of the strength and direction of correlation between variables.
5. **Shows Trends:** Helps in identifying trends such as linear, quadratic, or exponential relationships between variables.
6. **Facilitates Pattern Recognition:** Allows for pattern recognition and understanding of how changes in one variable affect another.
7. **Supports Predictive Modeling:** Useful in predictive modeling by visually assessing potential predictive power and relationships.
8. **Enhances Interpretation:** Provides a more intuitive understanding of data compared to numerical summaries alone.
9. **Supports Hypothesis Testing:** Aids in hypothesis testing by visually examining relationships suggested by the data.
10. **Useful in Exploratory Data Analysis:** Essential in exploratory data analysis (EDA) to uncover insights and formulate hypotheses for further investigation.

HEAT MAPS

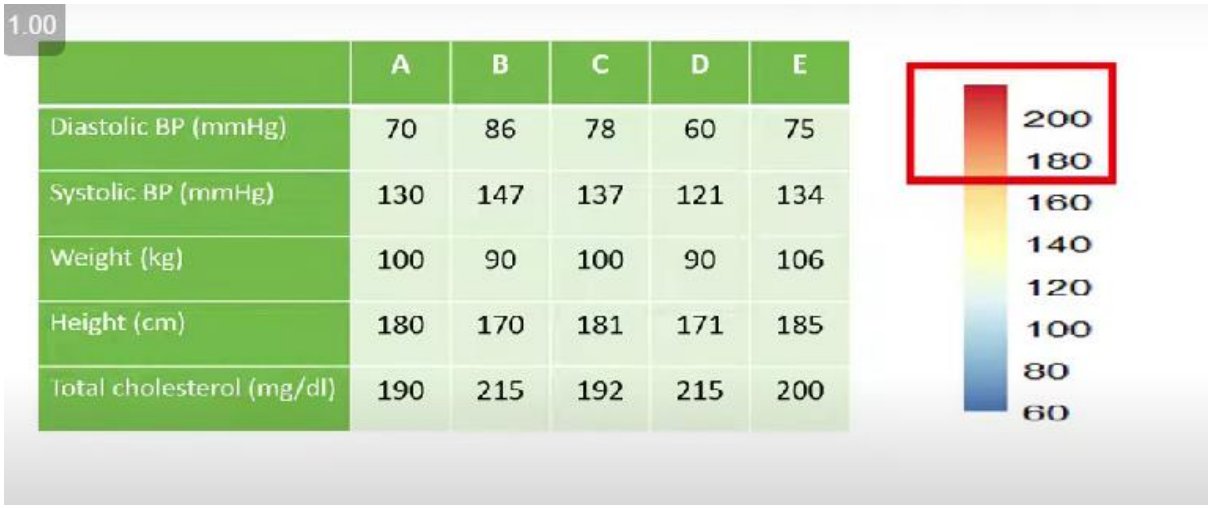

TIBCO Spotfire Learning Module
Interpreting Heat Map Visualizations
• Like a table or cross table view - with *only* colors displayed
• Tool for computational multivariate data analysis
Heat map
Table
Car
Price
City mileage
Hwy mileage
Passenger room
Cargo space
Engine displacement
Horsepower
car A 53,499 37.9 41.8 104 7 4.0 350
car B 25,999 12.3 17.7 68 18 8.0 400
car C 14,999 51.8 64.7 92 9 1.5 120
current car 39,999 22.4 27.9 143 11 3.0 280
Heat Map
Price
City mileage
Hwy mileage
Passenger ro...
Cargo space
Engine displ...
Horsepower
car A
car B
car C
current car
Car
Colors:
Max
Average
Min
7 columns



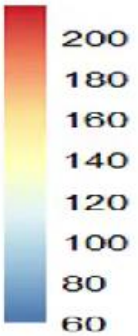
Cluster heatmaps

A heatmap is a graphical illustration of data where values are represented by colors.

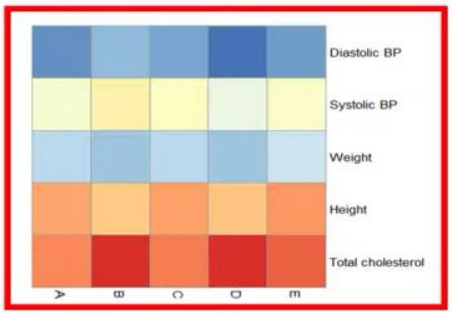
A cluster heatmap is a heatmap where the rows and columns of a data matrix have been ordered according to the output from clustering.

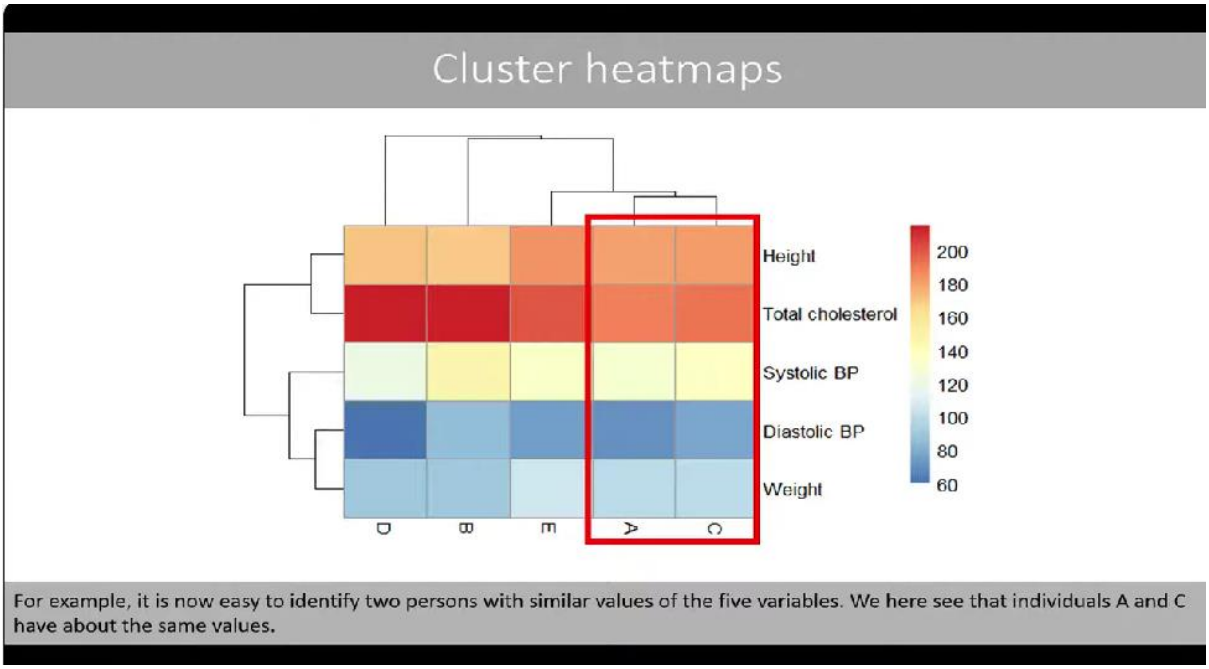


	A	B	C	D	E
Diastolic BP (mmHg)	70	86	78	60	75
Systolic BP (mmHg)	130	147	137	121	134
Weight (kg)	100	90	100	90	106
Height (cm)	180	170	181	171	185
Total cholesterol (mg/dl)	190	215	192	215	200



	A	B	C	D	E
Diastolic BP (mmHg)	70	86	78	60	75
Systolic BP (mmHg)	130	147	137	121	134
Weight (kg)	100	90	100	90	106
Height (cm)	180	170	181	171	185
Total cholesterol (mg/dl)	190	215	192	215	200





Heatmap is defined as a graphical representation of data using colors to visualize the value of the matrix. In this, to represent more common values or higher activities brighter colors basically reddish colors are

used and to represent less common or activity values, darker colors are preferred. Heatmap is also defined by the name of the shading matrix.

What is a heat map (heatmap)?

A heat map is a two-dimensional representation of data in which various values are represented by colors. A simple heat map provides an immediate visual summary of information across two axes, allowing users to quickly grasp the most important or relevant data points. More elaborate heat maps allow the viewer to understand complex data sets.

A heat map is a way to represent data points in a data set in a visual manner. All heat maps share one thing in common -- they use different colors or different shades of the same color to represent different values and to communicate the relationships that may exist between the variables plotted on the x-axis and y-axis. Usually, a darker color or shade represents a higher or greater quantity of the value being represented in the heat map.

EXAMPLE

For instance, a heat map showing the rain distribution (range of values) of a city grouped by month may use varying shades of red, yellow and blue. The months may be mapped on the y axis and the rain ranges on the x axis. The lightest color (i.e., blue) would represent the lower rainfall. In contrast, yellow and red would represent increasing rainfall values, with red indicating the highest values.

Other types of website heat maps include the following:

- **Click maps.** These maps show where users clicked on the website's pages, allowing webmasters to understand how people use the various pages and the elements on each page, and if there are navigational issues.
- **Scroll maps.** Scroll maps show site visitors' scrolling behavior and enable webmasters to determine the ideal length and content type for various webpages.

- **Mouse-tracking heat maps.** These maps show mouse hover patterns, which can reveal areas of visitor friction and enable webmasters to take action to optimize the website.

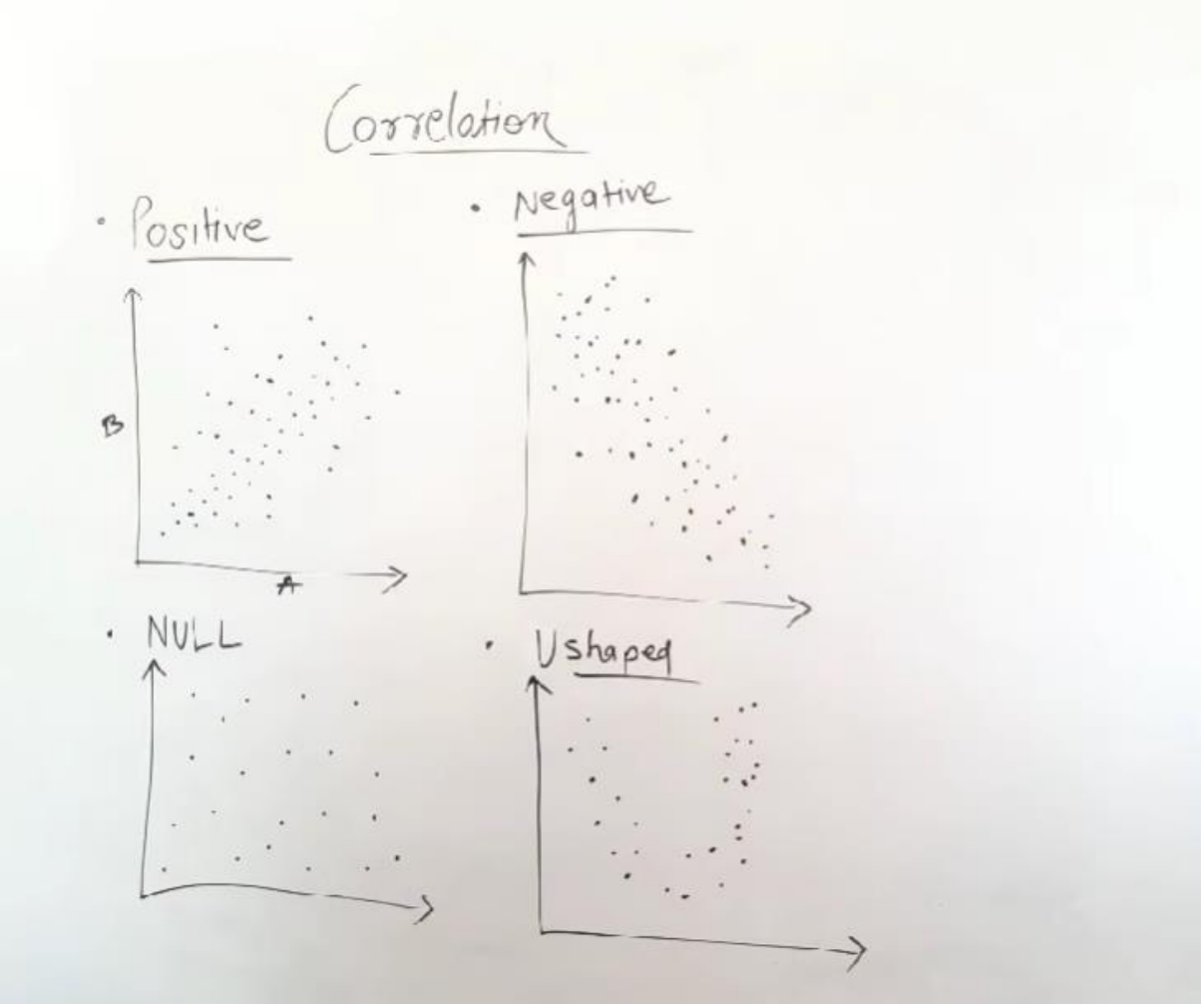
common applications of heat maps include the following:

- **Retail.** To analyze foot traffic in retail stores by hour, day, week or month and improve resource allocation for high-traffic locations
- **Manufacturing.** To monitor and optimize the performance of production bays in terms of defects, production count, downtime or other factors.
- **Population studies.** To understand various parameters of a population, such as per-capita-income or employment rate.

Importance of Heatmaps

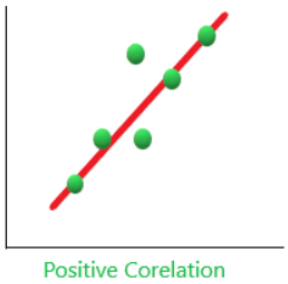
- **Visual Clarity**: Heatmaps present complex data in a visually accessible format, making it easier to grasp patterns and trends at a glance.
- **Data Exploration**: They facilitate in-depth exploration of data, allowing users to uncover hidden insights and anomalies that may not be apparent in raw data tables.
- **User-Centric Design**: In UX design, heatmaps enable designers to understand how users interact with websites and apps, leading to user-centric design improvements.
- **Data-Driven Decision Making**: Heatmaps provide actionable insights, empowering businesses to make data-driven decisions for website optimization, product development, and marketing strategies.
- **Efficient Problem-Solving**: They help identify issues and areas of improvement quickly, whether it's optimizing conversion paths, fixing UI/UX problems, or refining content placement.
- **Performance Evaluation**: Heatmaps aid in evaluating the effectiveness of marketing campaigns, email engagement, and ad placements, guiding strategies for better results.
- **Enhanced User Experience**: By analyzing user interactions through heatmaps, websites and apps can be fine-tuned to offer a more engaging and user-friendly experience.
- **Effective Communication**: Heatmaps simplify the communication of complex data findings to stakeholders, making it easier to convey insights and recommendations.
- • **Identifies Patterns:** Easily identifies patterns, clusters, and trends in large datasets that may not be apparent in raw data.
- • **Highlights Relationships:** Clearly displays relationships and correlations between variables through color gradients.
- • **Facilitates Comparison:** Allows for easy comparison of multiple categories or variables simultaneously.

CORRELATION PLOT

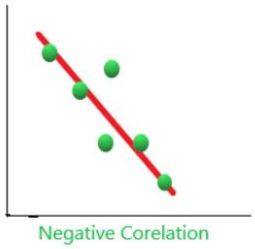


Correlation means an association, It is a measure of the extent to which two variables are related.

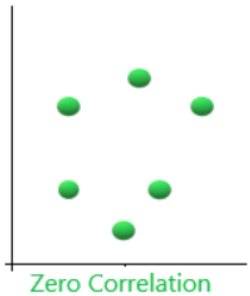
1. Positive Correlation: When two variables increase together and decrease together. They are positively correlated. '1' is a perfect positive correlation. For example – demand and profit are positively correlated the more the demand for the product, the more profit hence positive correlation.



2. Negative Correlation: When one variable increases and the other variable decreases together and vice-versa. They are negatively correlated. For example, If the distance between magnet increases their attraction decreases, and vice-versa. Hence, a negative correlation. '-1' is no correlation



3. Zero Correlation(No Correlation): When two variables don't seem to be linked at all. '0' is a perfect negative correlation. For Example, the amount of tea you take and level of intelligence.



What is a Correlation Coefficient? A coefficient of correlation is a value between `-1` and `+1` that denotes both the *strength* and *directionality* of a relationship between two variables.

- The closer the value is to 1 (or -1), the stronger a relationship.
- The closer a number is to 0, the weaker the relationship.

Calculating a Correlation Coefficient

Suppose we have the following values, with scores for $N = 5$ people on the variables X and Y:

X	Y
1	6
2	4
3	5
4	3
5	2

Correlation – Stating H_0 and H_1

First, we'll state our null (H_0) and alternative (H_1) hypotheses:

$H_0: \rho_{XY} = 0$ ← Population correlation is zero; there is *no* relationship between X and Y in the population. ρ = *'rho'*

$H_1: \rho_{XY} \neq 0$ ← Population correlation is *not* zero; there *is* a relationship between X and Y in the population – could be positive or negative, i.e., 2-tailed test.

Calculating a Correlation Coefficient

In terms of calculations, first we'll begin with calcu
the mean of X and the mean of Y.

1. Find mean of X and Y:

Mean of X = (1 + 2 + 3 + 4 + 5)/5

Mean of X = 15/5 = 3

Mean of Y = (6 + 4 + 5 + 3 + 2)/5

Mean of Y = 20/5 = 4

X	Y
1	6
2	4
3	5
4	3
5	2

Calculating a Correlation Coefficient

2. Next, we need to calculate deviation scores for each variable. Deviation scores subtract the mean from each value. They are called *deviation* scores because they indicate how far each value *deviates* (or departs from) the mean.

Calculating a Correlation: Deviation Scores

		Deviation score	Deviation score
X	Y	X – mean of X	Y – mean of Y
1	6	1 – 3 = –2	6 – 4 = 2
2	4	2 – 3 = –1	4 – 4 = 0
3	5	3 – 3 = 0	5 – 4 = 1
4	3	4 – 3 = 1	3 – 4 = –1
5	2	5 – 3 = 2	2 – 4 = –2

Mean X = 3 **Mean Y = 4**

Calculating a Correlation: Deviation Scores

		Deviation score	Deviation score
X	Y	X – mean of X	Y – mean of Y
1	6	1 – 3 = –2	6 – 4 = 2
2	4	2 – 3 = –1	4 – 4 = 0
3	5	3 – 3 = 0	5 – 4 = 1
4	3	4 – 3 = 1	3 – 4 = –1
5	2	5 – 3 = 2	2 – 4 = –2

Mean X = 3 **Mean Y = 4**

The deviaiton scores for a variable should always add to zero (by definiton).

Calculating a Correlation Coefficient

3. After finding deviation scores, we need to square each of these values.

(We square each of the deviation scores because it gets rid of the negative numbers.)

Calculating a Correlation

		Square each deviation score of X	Square each deviation score of Y
X	Y	$(X - \text{mean of } X)^2$	$(Y - \text{mean of } Y)^2$
1	6	$-2^2 = 4$	$2^2 = 4$
2	4	$-1^2 = 1$	$0^2 = 0$
3	5	$0^2 = 0$	$1^2 = 1$
4	3	$1^2 = 1$	$-1^2 = 1$
5	2	$2^2 = 4$	$-2^2 = 4$

Calculating a Correlation

4. Next, we need to add up these squared values. Technically speaking, we call this the sum of the squared deviation scores (or SS).

		Deviation score	Deviation score
X	Y	$SS_X = \Sigma(X - \text{mean of } X)^2$	$SS_Y = \Sigma(Y - \text{mean of } Y)^2$
1	6	$-2^2 = \mathbf{4}$	$2^2 = \mathbf{4}$
2	4	$-1^2 = \mathbf{1}$	$0^2 = \mathbf{0}$
3	5	$0^2 = \mathbf{0}$	$1^2 = \mathbf{1}$
4	3	$1^2 = \mathbf{1}$	$-1^2 = \mathbf{1}$
5	2	$2^2 = \mathbf{4}$	$-2^2 = \mathbf{4}$

$SS_X = 4 + 1 + 0 + 1 + 4$
$\mathbf{SS_X = 10}$

$SS_Y = 4 + 0 + 1 + 1 + 4$
$\mathbf{SS_Y = 10}$

equal humesha ho zaruri nahi

Calculating a Correlation Coefficient

		Deviation score	Deviation score	Sum of Products: SP (Cross Products)
X	Y	X – mean of X	Y – mean of Y	(X – mean X) (Y – mean Y)
1	6	1 – 3 = –2	6 – 4 = 2	(–2)(2) = **–4**
2	4	2 – 3 = –1	4 – 4 = 0	(–1)(0) = **0**
3	5	3 – 3 = 0	5 – 4 = 1	(0)(1) = **0**
4	3	4 – 3 = 1	3 – 4 = –1	(1)(–1) = **–1**
5	2	5 – 3 = 2	2 – 4 = –2	(2)(–2) = **–4**

Goal: Find SP.
SP = Sum of Cross Products

SP = –4 + 0 + 0 + (–1) + (–4)
SP = –9

Calculating a Correlation Coefficient

Now we have all of the values to calculate Pearson's *r*:

$$r = \frac{SP}{(\sqrt{SS_X})(\sqrt{SS_Y})}$$

Calculating a Correlation Coefficient

$SS_X = 10$, $SS_Y = 10$, $SP = -9$

$$r = \frac{-9}{(\sqrt{10})(\sqrt{10})} = \frac{-9}{(\sqrt{100})} = \frac{-9}{10} = -.9$$

$$r = -.9$$

Therefore, the value of Pearson's *r* is –.9.

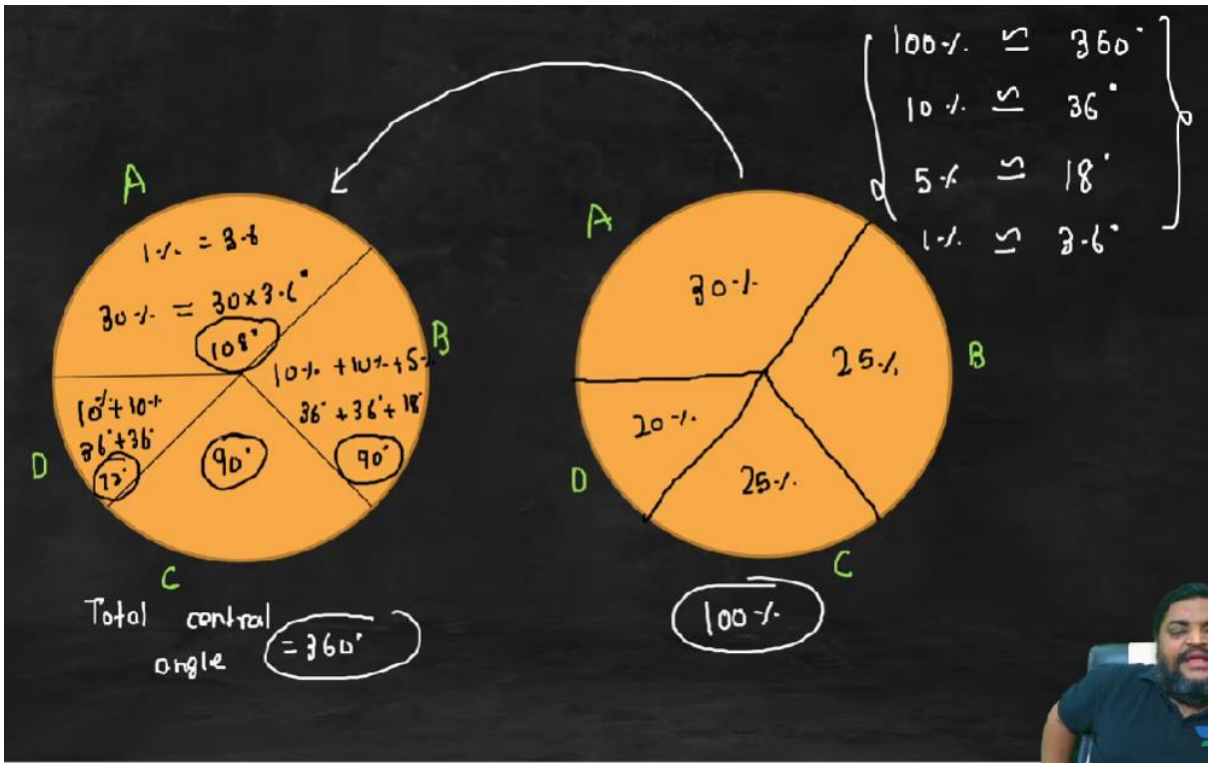
the benefits of correlation plots :

1. **Visualizes Relationships:** Shows the strength and direction of relationships between pairs of variables in a dataset.
2. **Identifies Patterns:** Quickly identifies which variables are positively, negatively, or not correlated with each other.
3. **Highlights Strong Relationships:** Easily identifies strong correlations that may indicate important connections in the data.
4. **Simplifies Multivariate Analysis:** Provides a compact representation of correlations across multiple variables, simplifying complex datasets.

5. **Detects Multicollinearity:** Helps in detecting multicollinearity, where independent variables are highly correlated, which can affect regression models.
6. **Guides Feature Selection:** Aids in feature selection by identifying variables that are most strongly correlated with the target variable or with each other.
7. **Supports Hypothesis Testing:** Facilitates hypothesis testing by visually examining relationships suggested by the data.
8. **Quantifies Relationships:** Quantifies the strength of correlations using correlation coefficients, typically ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation).
9. **Enhances Data Exploration:** Useful in exploratory data analysis (EDA) to understand data structure and relationships before further analysis or modeling.
10. **Facilitates Decision Making:** Provides insights into which variables may be most influential or related in the context of decision making or predictive modeling.

PIE CHARTS

A Pie Chart is a pictorial representation of the data. It uses a circle to represent the data and is called a Circle Graph. In a Pie Chart, we present the data by dividing the whole circle into smaller slices or sectors, and each slice or sector represents specific data.



There are [different ways of data representation](). A pie chart is one of the types of charts in which the data is represented in a circular shape. The pie chart circle is further divided into multiple sectors/slices; those sectors show the different parts of the data from the whole.

Pie charts, also known as circle graphs or pie diagrams, are very useful in representing and interpreting data. The data can be compared easily with the help of a pie chart

Types of Pie Chart

There are various variation or types of pie chart, some of the common types include:

- **3D Pie Chart:** A 3D pie chart adds depth to the traditional two-dimensional pie chart by rendering it in three dimensions.
- **Doughnut Chart:** A doughnut chart is similar to a pie chart but with a hole in the center.
- **Exploded Pie Chart:** In an exploded pie chart, one or more slices are separated from the rest of the pie to emphasize their importance or to make them stand out.
- **Nested Pie Chart:** Also known as a multi-level pie chart or hierarchical pie chart, this type of chart consists of multiple rings of pie charts, with each ring representing a different level of data hierarchy.
- **Ring Chart:** A ring chart is similar to a doughnut chart but consists of multiple rings instead of just one. Each ring represents a different category of data, with the size of each segment within the ring corresponding to its proportion of the whole.

Pie Chart Formula

The total value or percentage of the pie is 100% always. Here it contains different sectors and segments in which each sector or segment of the chart corresponds to a certain portion of the net or total percentage (or data). **The total or sum of all the data can be summed up to 360 degrees.**

- Converting the data into degrees on a pie chart. The formula for a pie chart can be summed up as:

$$(\text{Given Data} / \text{Total Value of Data}) \times 360^\circ$$

- Calculating the percentage of each sector from degrees in a pie chart.

To work out with degrees in a pie chart, we need to follow the following steps:

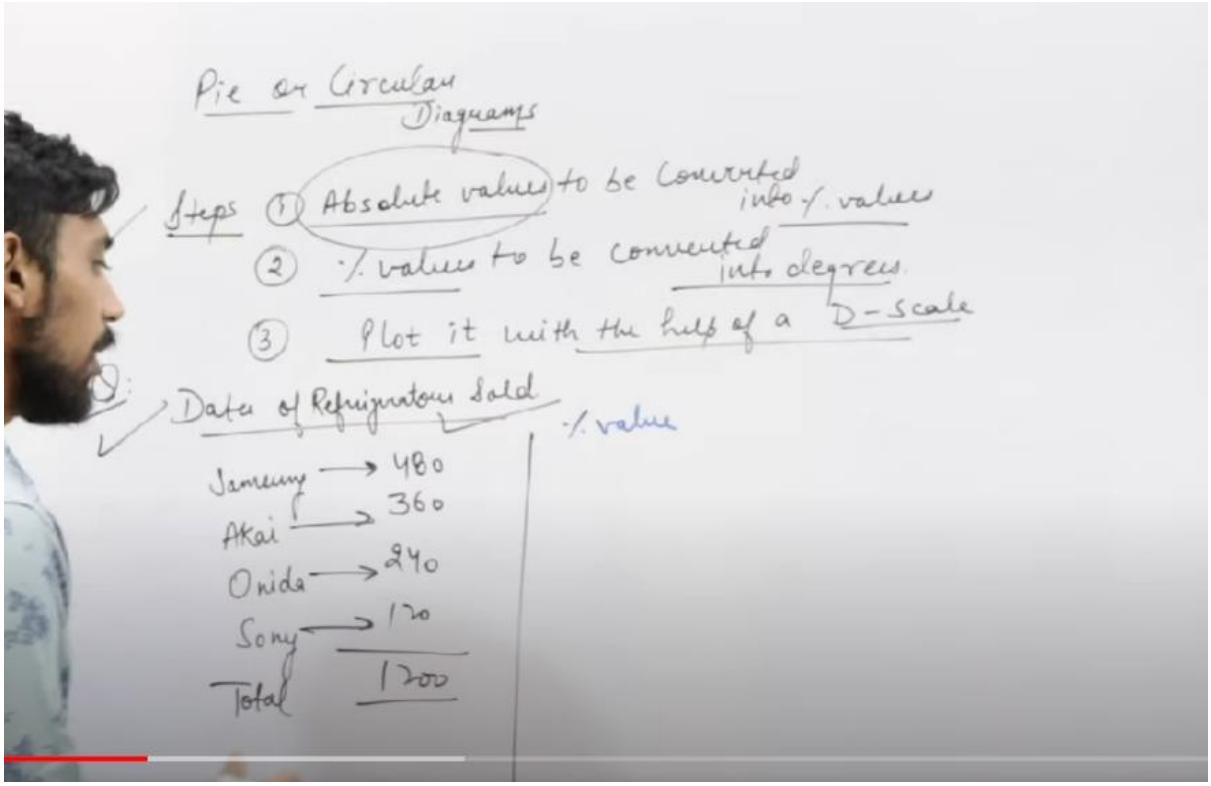
- First, we need to measure every slice of the chart.
- Then we need to divide it by 360°.
- Finally, multiply the obtained result by 100.

The pie chart formula is given below:

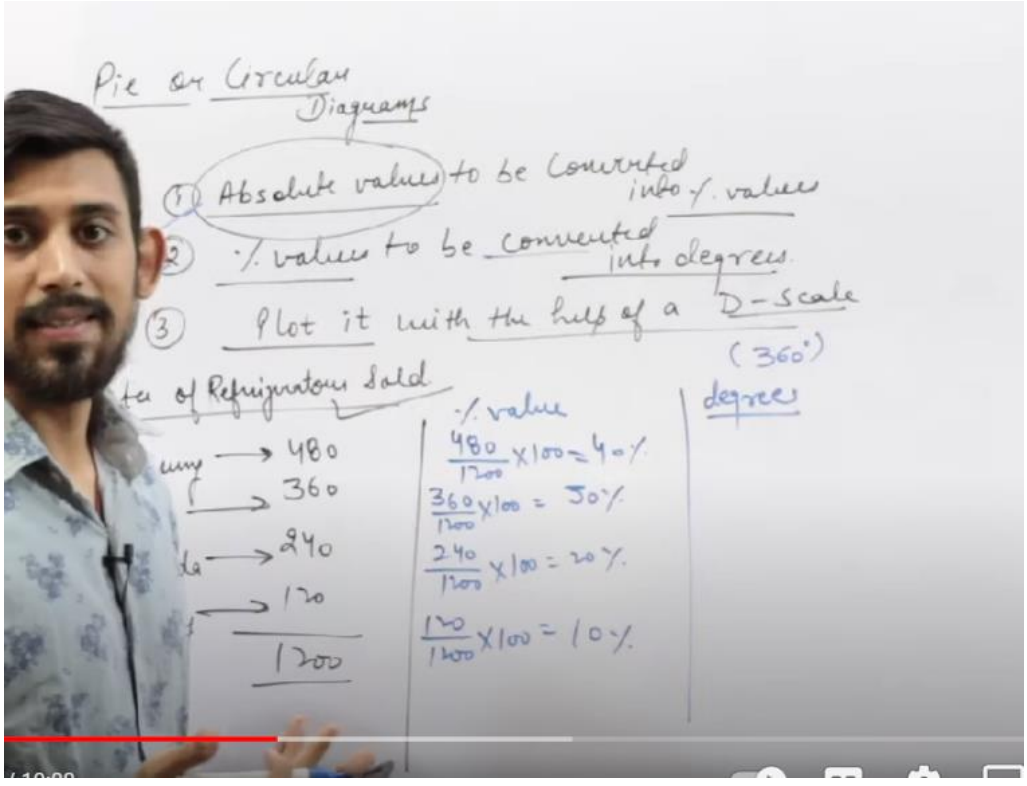
$$(\text{Frequency})/(\text{Total Frequency}) \times 100$$

Calculating Number of Sectors on a Pie Chart

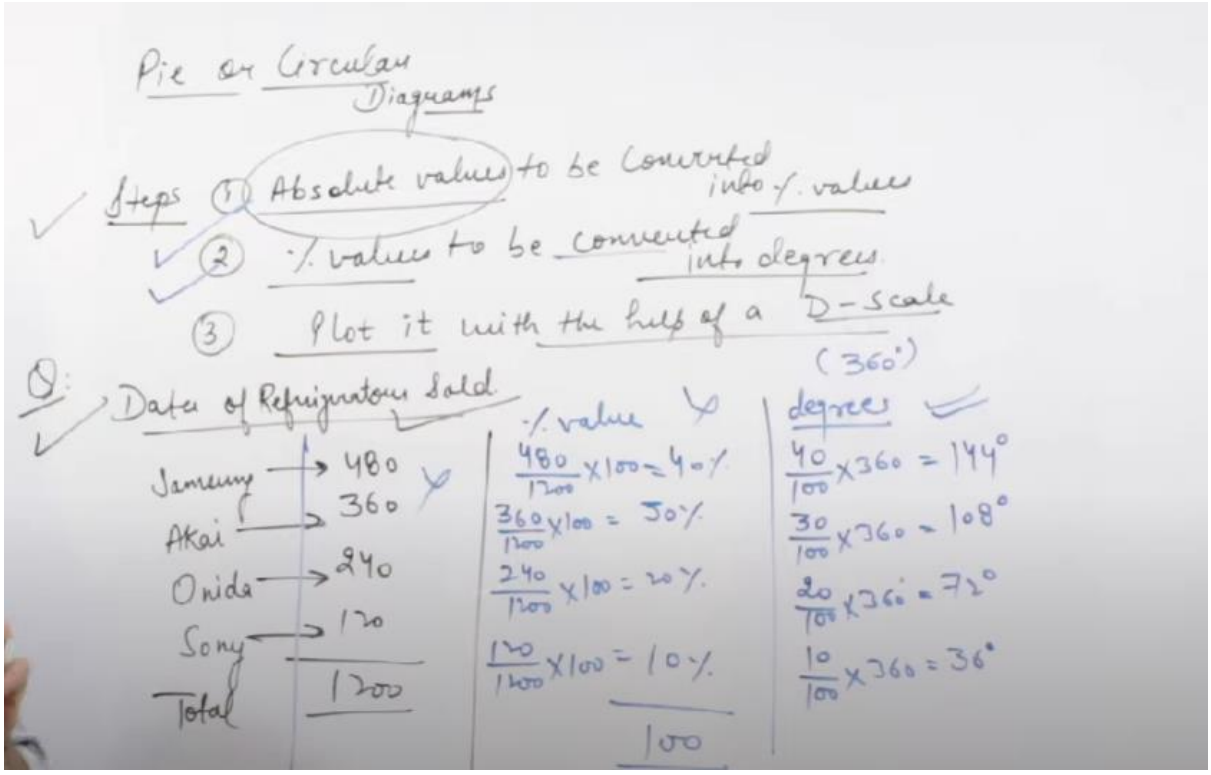
To calculate the total number of slices or sectors on a pie chart, we need to multiply the sector's percentage by the total value of the data and finally divide the result by 100.

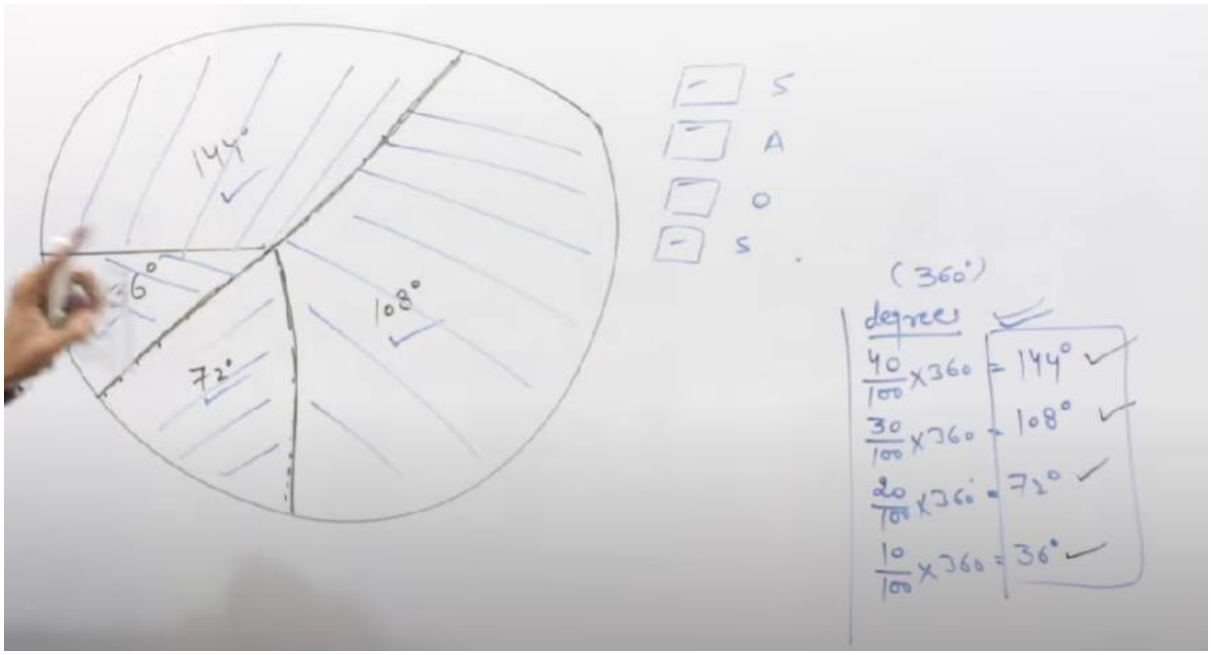


Now calculate % value



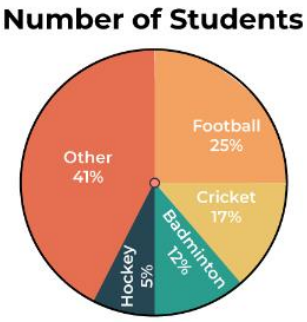
Now calculate degree





Pie Chart Examples

Let's take a look at an example for a better understanding of pie charts. In a class of 200 students, a survey was done to collect each student's favorite sports. The pie chart of the data is given below:



Since the pie chart is provided and the total number of students is given, we can easily take the original data out for each sport.

- Cricket = 17/100 × 200 = 34 students
- Football = 25/100 × 200 = 50 students
- Badminton = 12/100 × 200 = 24 students
- Hockey = 5/100 × 200 = 10 students
- Other = 41/100 × 200 = 82 students

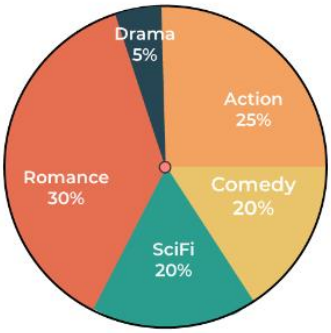
The original data for the pie chart shown above is given below:

Sport	Number of Students
Cricket	34
Football	50
Badminton	24
Hockey	10
Other	82

How to Read Pie Chart

In order to read a pie chart, the first thing to notice is the data presented in the pie chart. If the data is given in percentage, it should be converted accordingly in order to analyze and interpret the data. Let's take a look at an example in order to learn how to interpret pie charts.

Example: In a survey done among 300 people, it was observed which type of genre each person prefers. The pie chart of the same is mentioned below. Analyze and interpret the pie chart accordingly to find the original data.



Solution:

While observing the pie chart, it came to notice that the data is present in percentage. Let's convert the data to obtain the original value.

- *Number of people who like comedy = 20/100 × 300 = 60 people.*
- *Number of people who like action = 25/100 × 300 = 75 people.*
- *Number of people who like romance = 30/100 × 300 = 90 people.*
- *Number of people who like drama = 5/100 × 300 = 15 people.*
- *Number of people who like sci-fi = 20/100 × 300 = 60 people.*

Aspect	Pie Chart	Bar Graph
Representation	Circular display of data	rectangular display of data
Purpose	Shows parts of a whole	Compares discrete categories
Data presentation	Depicts percentages or proportions	Shows exact values or quantities.
Number of variables	Typically one variable	Can represent multiple variables.
Visualization	Easily shows relative proportions	Effective for comparing quantities.
Comparison	Might be difficult to compare precise values	Allows for easy comparison between categories.
Data complexity	Works well with simple datasets	Suitable for complex datasets
Interpretation	Provides a holistic view	Allows for detailed analysis.
Space efficiency	Not efficient with large datasets	Efficient for displaying large datasets

Pie Chart Advantages

Pie Chart is very useful for finding and representing data. Various advantages of the pie chart are,

- Pie chart is easily understood and comprehended.
- Visual representation of data in a pie chart is done as a fractional part of a whole.
- Pie chart provides an effective mode of communication to all types of audiences.
- Pie chart provides a better comparison of data for the audience.

Pie Chart Disadvantages

There are some disadvantages also of using pie charts and some of them are added below,

- In the case of too much data, this presentation becomes less effective using a pie chart.
- For multiple data sets, we need a series to compare them.
- For analyzing and Assimilating the data in a pie chart, it is difficult for readers to comprehend.

Uses of Pie Chart

Whenever a fraction or fractions are represented as a part of the whole, pie charts are used. Pie charts are used to compare the data and to analyze which data is bigger or smaller. Hence, while dealing with discrete data, pie charts are preferred. Let’s take a look at the uses of the pie chart:

- Pie charts are used to compare the profit and loss in businesses.
- In schools, the grades can be easily compared using a pie chart.
- The relative sizes of data can be compared using a pie chart.
- The marketing and sales data can be compared using a pie chart.

popular Python libraries used for data visualization:

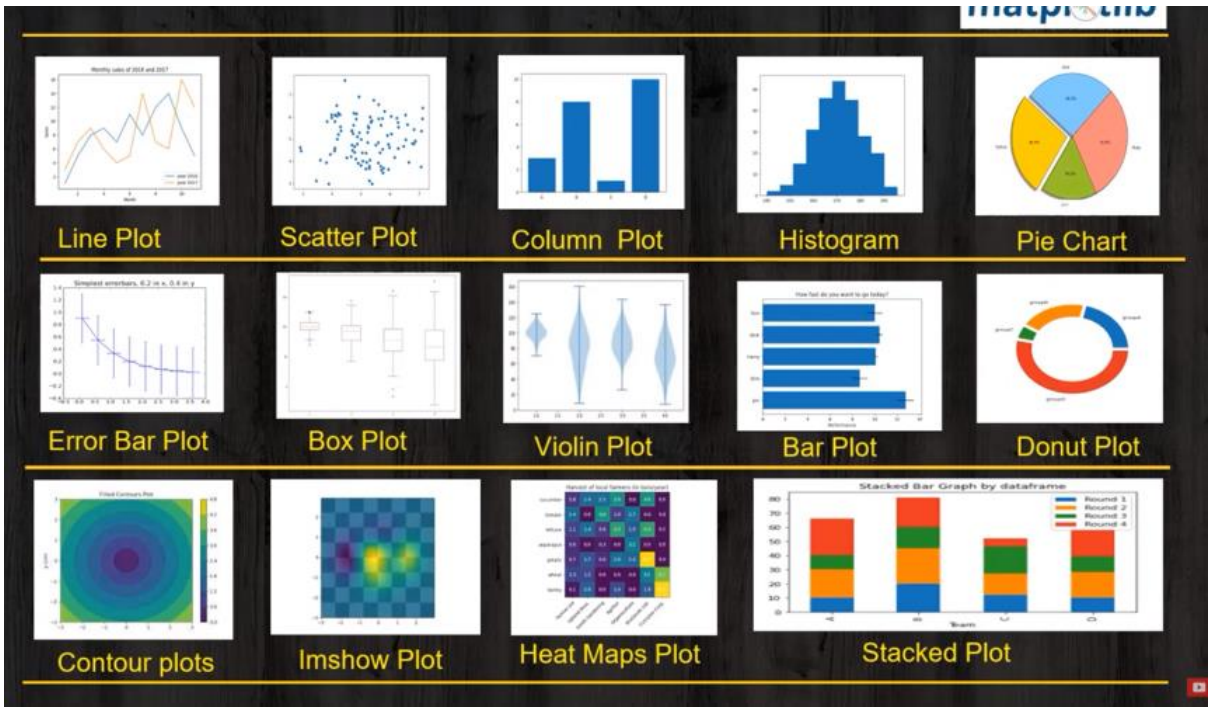
1. **Matplotlib**
    - **Purpose:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
    - **Functionality:** It provides a wide range of plotting options, from basic charts (like line plots, bar charts, scatter plots) to complex visualizations (like heatmaps, contour plots, 3D plots).
    - **Key Features:** Customizable plots, support for LaTeX rendering, integration with Jupyter Notebooks, and export to various file formats.
2. **Seaborn**
    - **Purpose:** Seaborn is built on top of Matplotlib and provides a high-level interface for drawing attractive statistical graphics.
    - **Functionality:** It simplifies the creation of complex visualizations such as heatmaps, violin plots, pair plots, and categorical plots with minimal code.
    - **Key Features:** Automatic estimation and plotting of complex statistical relationships, color palettes for aesthetic appeal, and integration with Pandas DataFrames.
3. **Plotly**
    - **Purpose:** Plotly is a library for creating interactive visualizations and web-based dashboards.
    - **Functionality:** It supports a wide range of chart types (scatter plots, line plots, bar charts, histograms) with interactive features like zooming, panning, tooltips, and hover effects.
    - **Key Features:** Creation of interactive plots directly in Jupyter Notebooks, integration with Dash for building web applications, and support for exporting plots as HTML or static images.
4. **Bokeh**
    - **Purpose:** Bokeh is another library for creating interactive visualizations and plots that are suitable for web deployment.
    - **Functionality:** It emphasizes interactivity, allowing users to create plots with linked brushing, hovering, and widgets for dynamic user interaction.
    - **Key Features:** Support for streaming data, large datasets, and interactive tools like sliders, buttons, and dropdowns for exploring data interactively.
5. **Altair**
    - **Purpose:** Altair is a declarative statistical visualization library based on Vega and Vega-Lite.
    - **Functionality:** It simplifies the creation of complex visualizations by using a concise grammar of graphics syntax.
    - **Key Features:** Automatic handling of data transformation, support for interactive visualizations in Jupyter Notebooks, and easy customization through a JSON-like specification.
6. **Pandas Plotting (built-in)**
    - **Purpose:** Pandas, a popular data manipulation library, includes basic plotting functionality for quick exploratory data analysis.
    - **Functionality:** It provides a simple interface for creating basic plots (like line plots, bar charts, histograms) directly from Pandas DataFrames or Series.
    - **Key Features:** Convenient for rapid visualization during data preprocessing and exploratory data analysis without needing to import additional libraries.

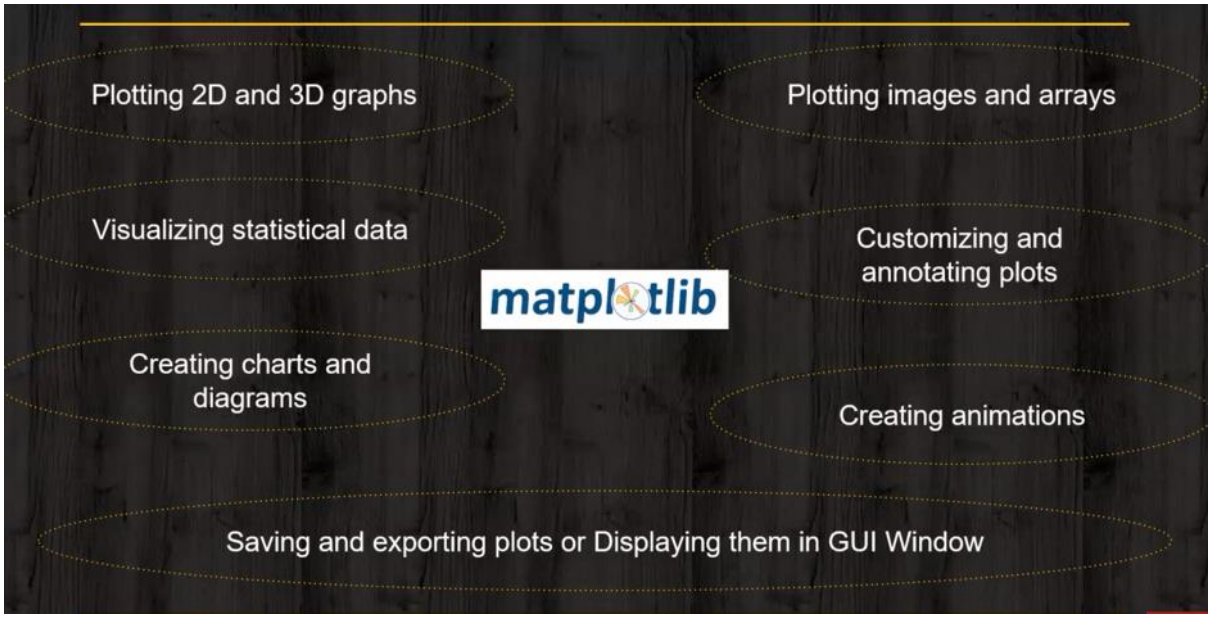
data visualization using python with matplotlib and seaborn

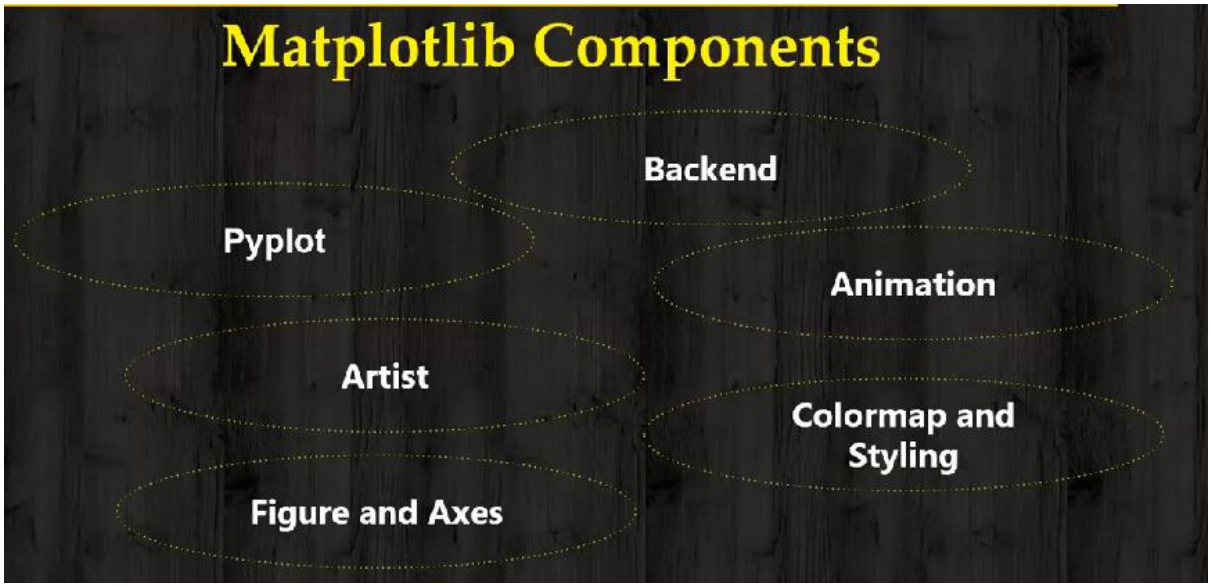
MATPLOTLIB

It is an amazing visualization library in Python for 2D plots of arrays, It is a multi-platform data visualization library built on *NumPy* arrays and designed to work with the broader *SciPy* stack. It was introduced by John Hunter in the year 2002. Let's try to understand some of the benefits and features of *matplotlib*

- It's fast, efficient as it is based on *numpy* and also easier to build
- Has undergone a lot of improvements from the open source community since inception and hence a better library having advanced features as well
- Well maintained visualization output with high quality graphics draws a lot of users to it
- Basic as well as advanced charts could be very easily built
- From the users/developers point of view, since it has a large community support, resolving issues and debugging becomes much easier



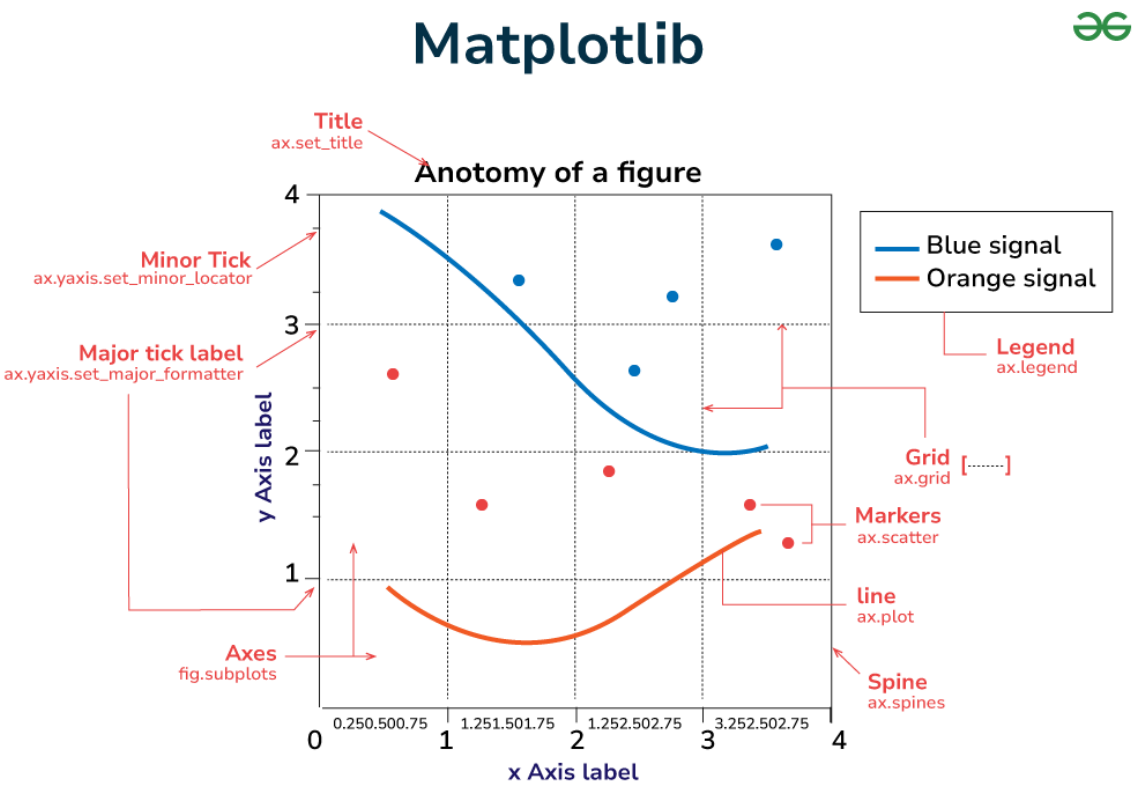




Key Features of Matplotlib:

1. **Versatility**: Matplotlib can generate a wide range of plots, including line plots, scatter plots, bar plots, histograms, pie charts, and more.
2. **Customization**: It offers extensive customization options to control every aspect of the plot, such as line styles, colors, markers, labels, and annotations.
3. **Integration with NumPy**: Matplotlib integrates seamlessly with NumPy, making it easy to plot data arrays directly.
4. **Publication Quality**: Matplotlib produces high-quality plots suitable for publication with fine-grained control over the plot aesthetics.

5. **Extensible**: Matplotlib is highly extensible, with a large ecosystem of add-on toolkits and extensions like Seaborn, Pandas plotting functions, and Basemap for geographical plotting.
6. **Cross-Platform**: It is platform-independent and can run on various operating systems, including Windows, macOS, and Linux.
7. **Interactive Plots**: Matplotlib supports interactive plotting through the use of widgets and event handling, enabling users to explore data dynamically.



- **Figure:** Top-level container for the plot, akin to a blank canvas.
- **Axes:** Rectangular areas within the figure where data is plotted, providing the coordinate system.
- **Axis:** Represents the x-axis and y-axis, defining limits, ticks, labels, and scales.
- **Marker:** Symbols like circles or squares used in plots to mark individual data points.
- **Lines:** Connect data points in plots, showing relationships or trends.
- **Title:** Text element providing a descriptive title for the plot.

- **Axis Labels:** Labels for x-axis and y-axis, identifying plotted data.
- **Ticks:** Marks along axes indicating specific data points or intervals.
- **Tick Labels:** Text labels displaying values corresponding to tick marks.
- **Legend:** Key explaining symbols or colors in the plot for different data series.
- **Grid Lines:** Horizontal and vertical lines aiding in data visualization.
- **Spines:** Borders around the plot area, customizable to alter plot appearance.

Matplotlib is a suitable choice for various data visualization tasks, including exploratory data analysis, scientific plotting, and creating publication-quality plots. It excels in scenarios where users require fine-grained control over plot customization and need to create complex or specialized visualizations.

SEABORN

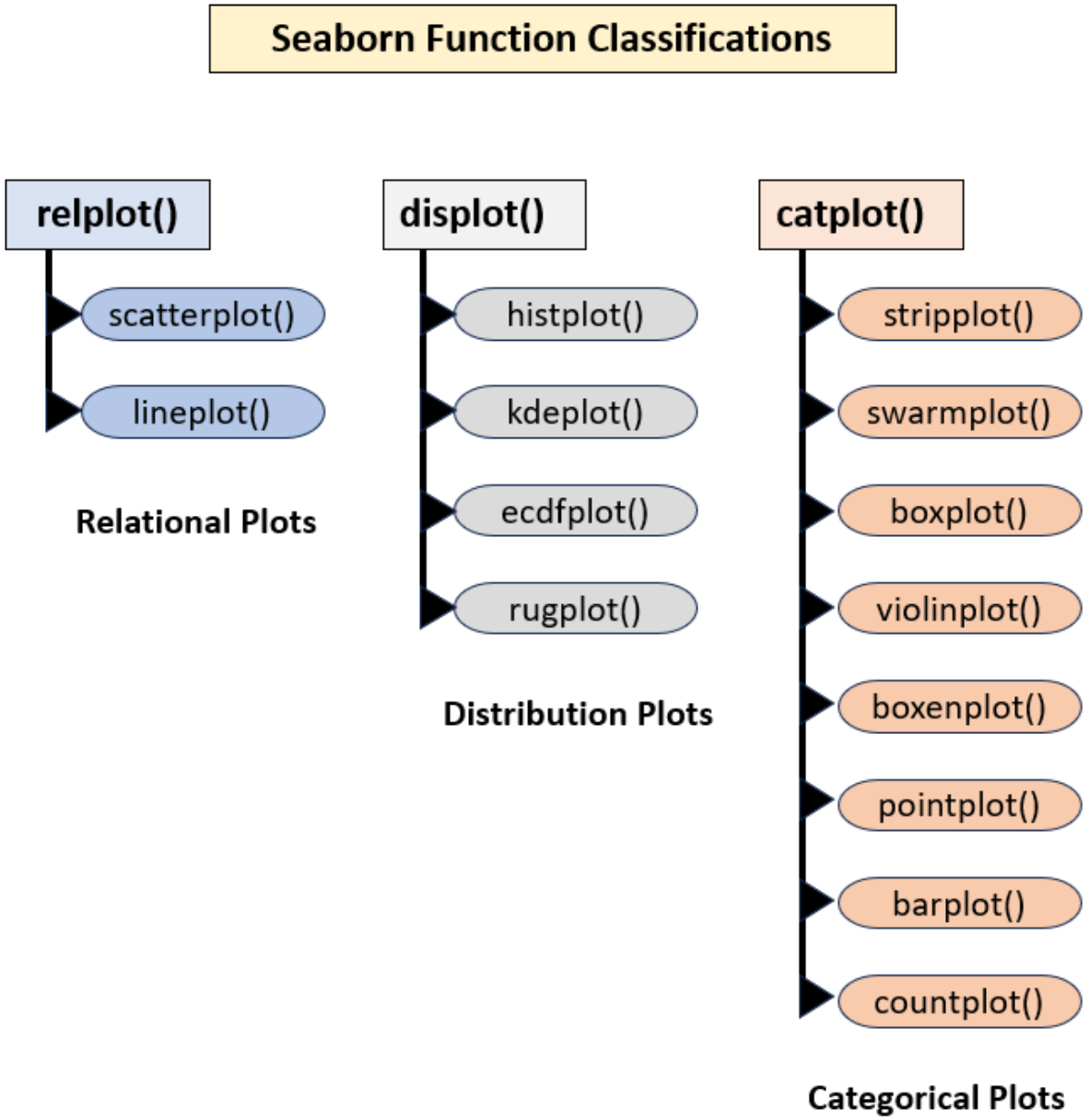
Conceptualized and built originally at the Stanford University, this library sits on top of *matplotlib*. In a sense, it has some flavors of *matplotlib* while from the visualization point, it is much better than *matplotlib* and has added features as well. Below are its advantages

- Built-in themes aid better visualization
- Statistical functions aiding better data insights
- Better aesthetics and built-in plots
- Helpful documentation with effective examples

Different categories of plot in Seaborn

Plots are basically used for visualizing the relationship between variables. Those variables can be either completely numerical or a category like a group, class, or division. Seaborn divides the plot into the below categories –

- **Relational plots:** This plot is used to understand the relation between two variables.
- **Categorical plots:** This plot deals with categorical variables and how they can be visualized.
- **Distribution plots:** This plot is used for examining univariate and bivariate distributions
- **Regression plots:** The regression plots in Seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses.
- **Matrix plots:** A matrix plot is an array of scatterplots.
- **Multi-plot grids:** It is a useful approach to draw multiple instances of the same plot on different subsets of the dataset.


Seaborn Function Classifications
relplot()
scatterplot()
lineplot()
Relational Plots
displot()
histplot()
kdeplot()
ecdfplot()
rugplot()
Distribution Plots
catplot()
stripplot()
swarmplot()
boxplot()
violinplot()
boxenplot()
pointplot()
barplot()
countplot()
Categorical Plots