STACK ROUTE AN NIIT VENTURE

Data Visualization Using Matplotlib and Seaborn

EMPOWERING High Performance Technology Teams

OVERVIEW

Overview

- Total Sessions (6 hours)
- Focus majorly on Data Visualization using Matplotlib and Seaborn

Agenda

- Introduction to visualization
- Types of visualization
- Python libraries for visualization
- Introduction to Matplotlib
- Introduction to Seaborn
- Nature of Visualization

Data Visualization

- The graphical representation of data and information using visual components like charts, graphs, maps and info graphics is known as data visualization.
- It is the process of converting unprocessed data into understandable, interpretable and communicable visual formats.

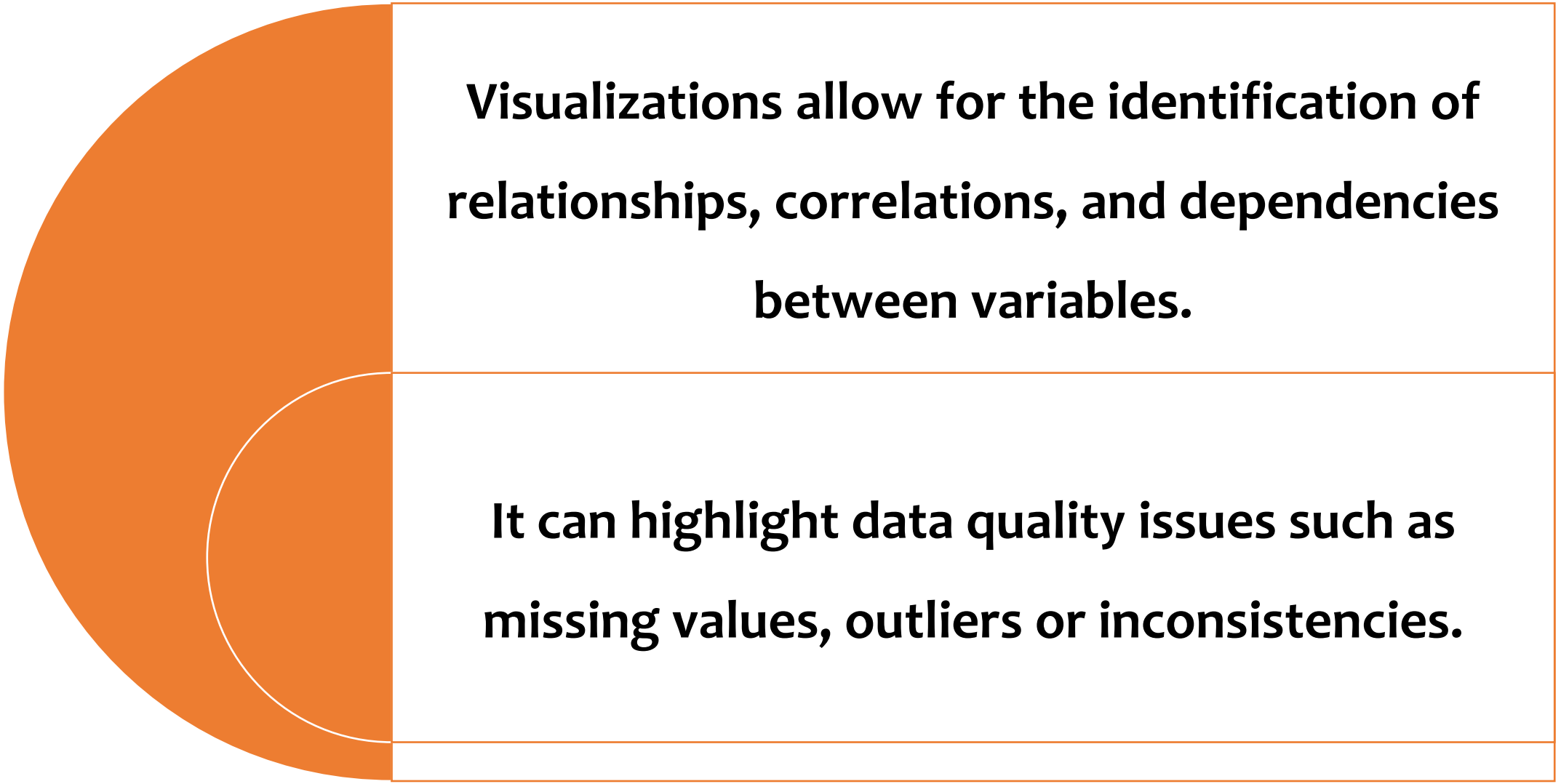
Consideration of Data Visualization

- It is considered to identify patterns, trends, relationships, and insights by visualizing data, which allows them to make more informed decisions and comprehend complex data sets.
- Data visualization is essential for data analysis, narrative development and the clear, engaging communication of information.

Advantages of Data Visualization

- Visualizing data allows for a better understanding of complex information by presenting it in a visual format.
- Data visualization enables faster and more efficient data analysis.
- It is an effective way to communicate data-driven insights to stakeholders.

Advantages of Data Visualization



Disadvantages of Data Visualization

- Data visualizations can be manipulated or designed in a way that distorts the true nature of the data.
- Visual cues such as scaling, labeling, or color choices can unintentionally or intentionally misrepresent the information – leading to biased interpretations or incorrect conclusions.
- Overly cluttered or complex visualizations may hinder comprehension and lead to confusion or cognitive overload.

Python Visualization Libraries

Matplotlib is a widely used plotting library that provides a flexible and comprehensive set of tools for creating various types of visualizations, including line plots, scatter plots, bar plots, histograms, etc.

Seaborn is a higher-level library built on top of Matplotlib. It offers a simplified interface and provides additional statistical visualization capabilities, such as box plots, violin plots, heat maps, and categorical plots.

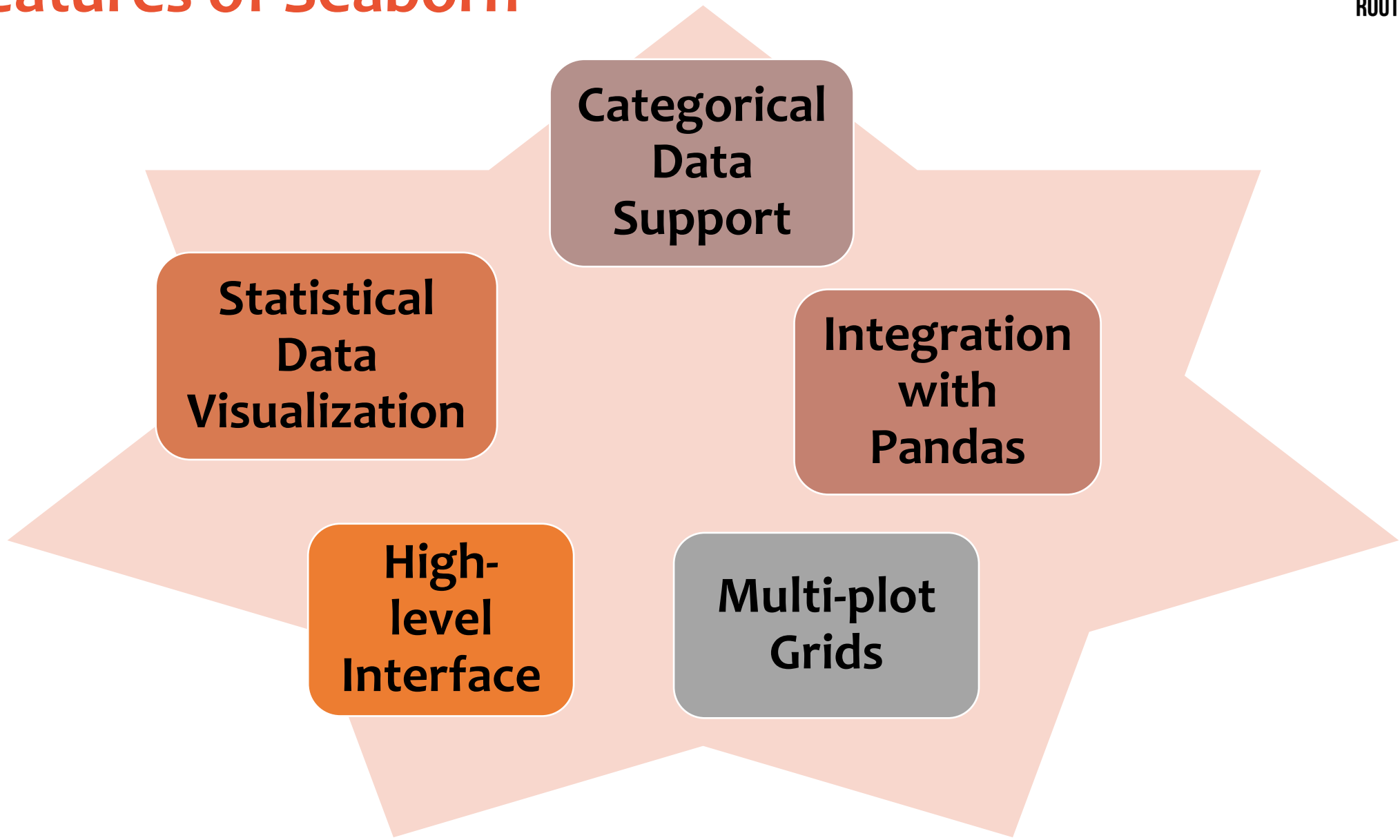
Benefits and Features of Matplotlib

- Matplotlib has a user-friendly interface, making it easy for beginners to get started with creating basic plots quickly.
- Its syntax is simple and intuitive, allowing users to generate plots with just a few lines of code.
- Matplotlib is designed to generate high-quality plots.

Advantages of Seaborn

- Seaborn specializes in statistical visualizations, offering a wide range of plots for exploring and analyzing relationships in data.
- It including scatter plots, bar plots, box plots, violin plots and heat maps.
- Seaborn seamlessly integrates with Pandas, a popular data manipulation library, allowing users to easily visualize data stored in DataFrames.

Key Features of Seaborn



Nature of Visualization

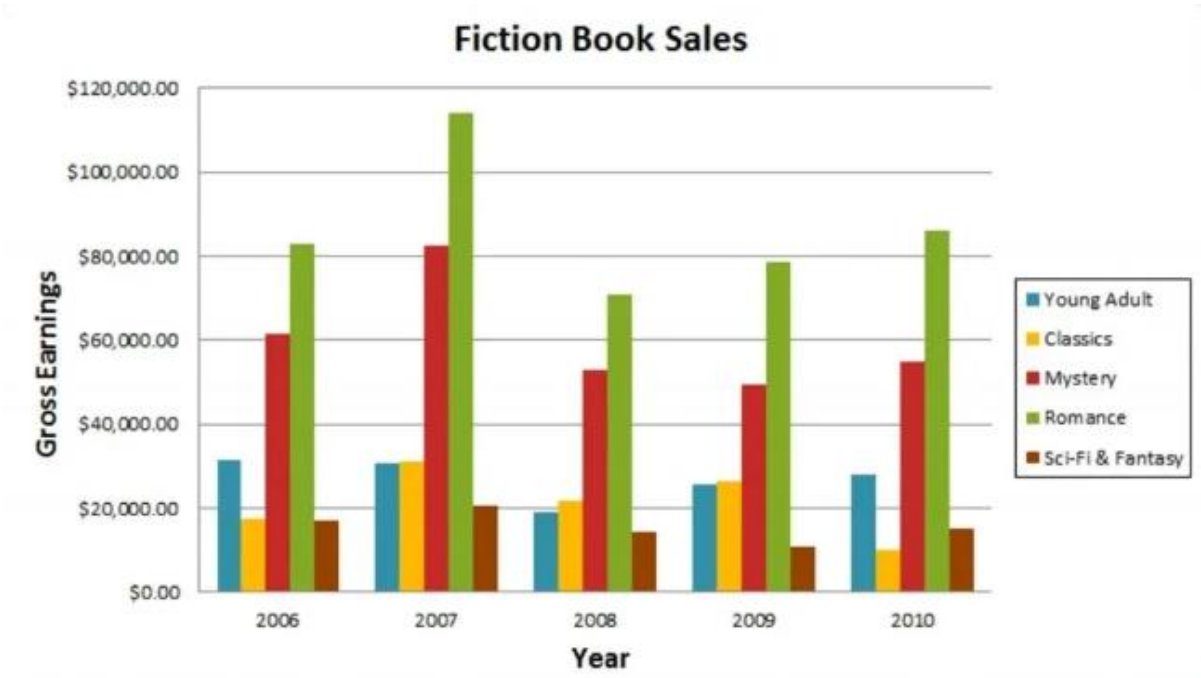
- The choice of visualization depends on several factors including the number and types of variables involved.
- Different types of charts are available to help us better understand the relationships within our data, catering to specific scenarios and variable types.
- Based on the count of variables:
  - Univariate plot (one variable)
  - Multivariate plot (more than one variable)

Bar Plot

- A bar plot, also known as a bar chart, is a type of plot that uses rectangular bars to represent categorical data.
- It shows the relationship between a categorical variable on the x-axis and a numerical variable on the y-axis.
- It is useful for comparing and visualizing data across different categories or groups.

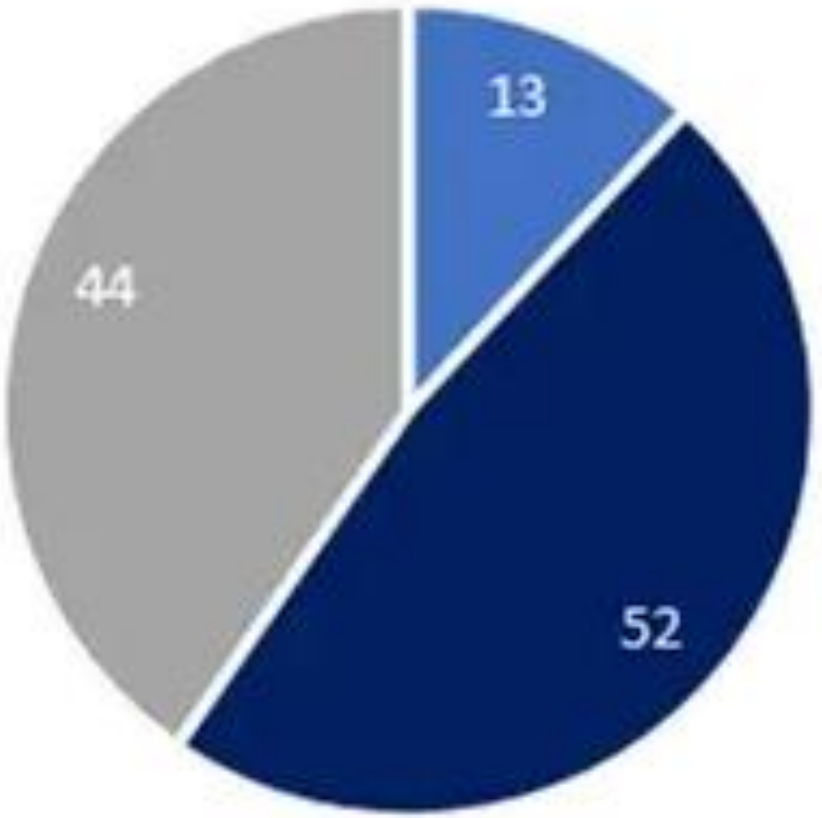
Grouped Bar Plot

- A grouped bar plot is a variation of a bar plot that allows for the comparison of multiple categorical variables within each category or group.
- Grouped bar plots are effective in comparing values across multiple groups and sub-groups, highlighting variations and patterns in the data.



Pie Chart

- Pie chart is a circular statistical plot used to visualize the proportionality and relativity of the data.
- Pie chart is one of the best visualizations for depicting a portion to whole relationship in the data.
- The area of the pie chart is the total percentage of the whole data.

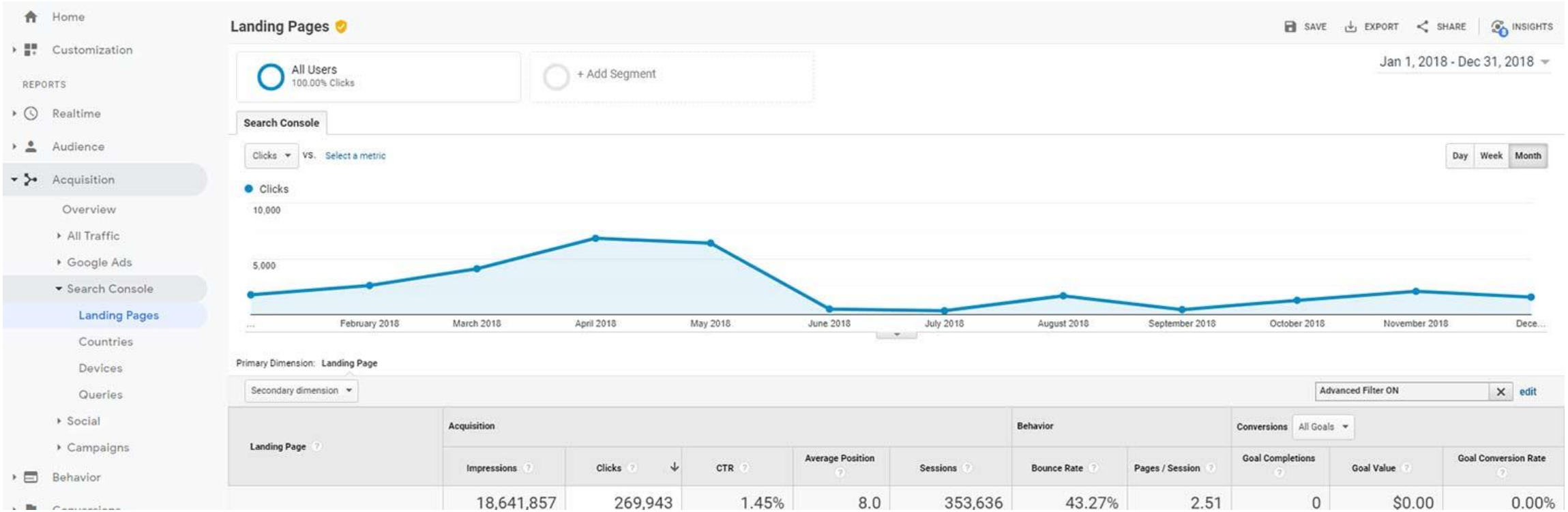


Donut Chart

- A donut chart is a pie chart with part of the centre removed.
- Pie charts are sometimes criticized for drawing readers' attention to how the slices' areas relate to one another and to the entire chart.
- This makes it challenging to distinguish between slices, especially when comparing multiple Pie Charts at once.

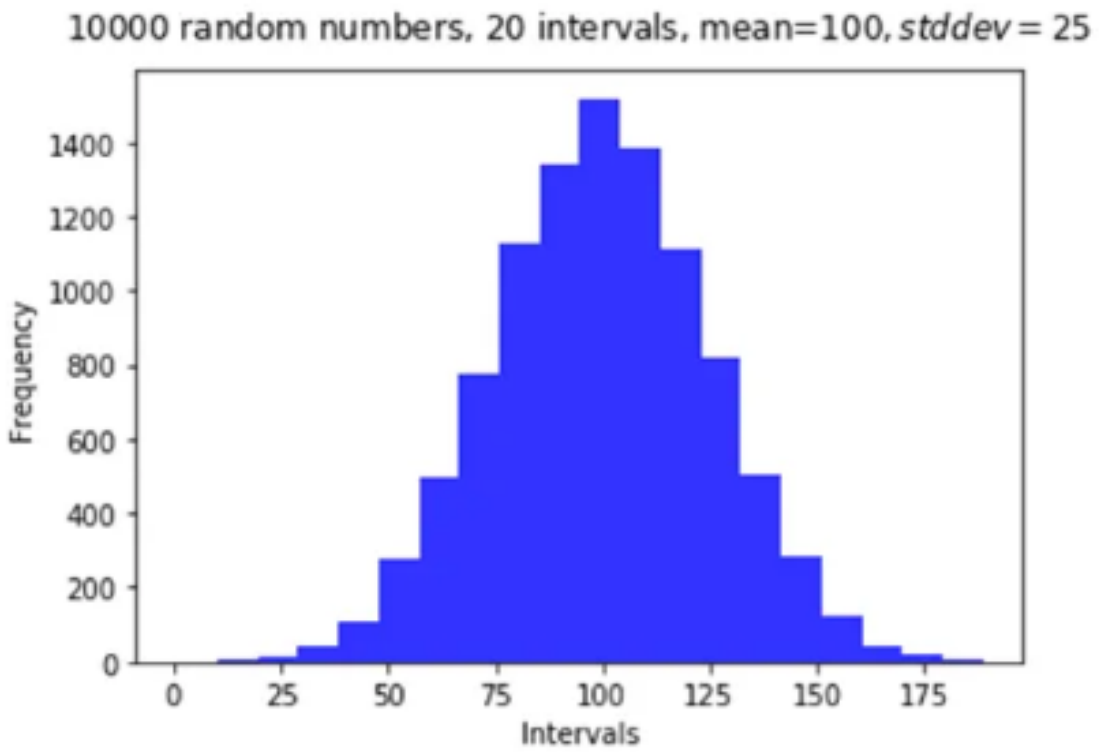
Line Chart

- Line charts are also used to depict the relationship between two variables through points but jointed together in a sequential manner.
- The trend of the relationship can be visualised by the line plot.



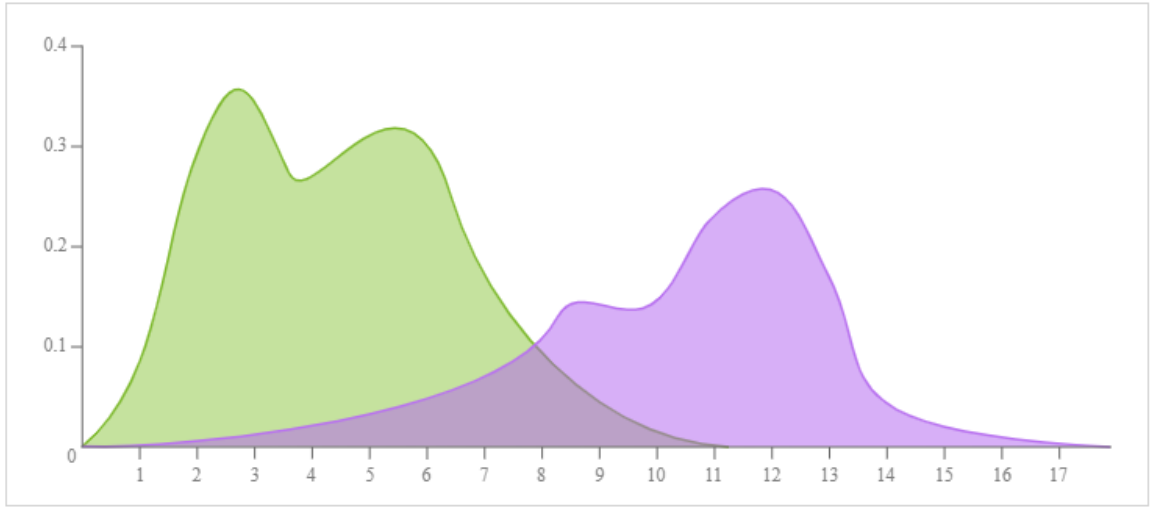
Histogram

- A histogram is a graphical representation of the distribution of a continuous variable.
- Histograms are commonly used to visualize the shape, central tendency, and spread of data.
- It allows to identify the presence of peaks, gaps, or outliers in the distribution.



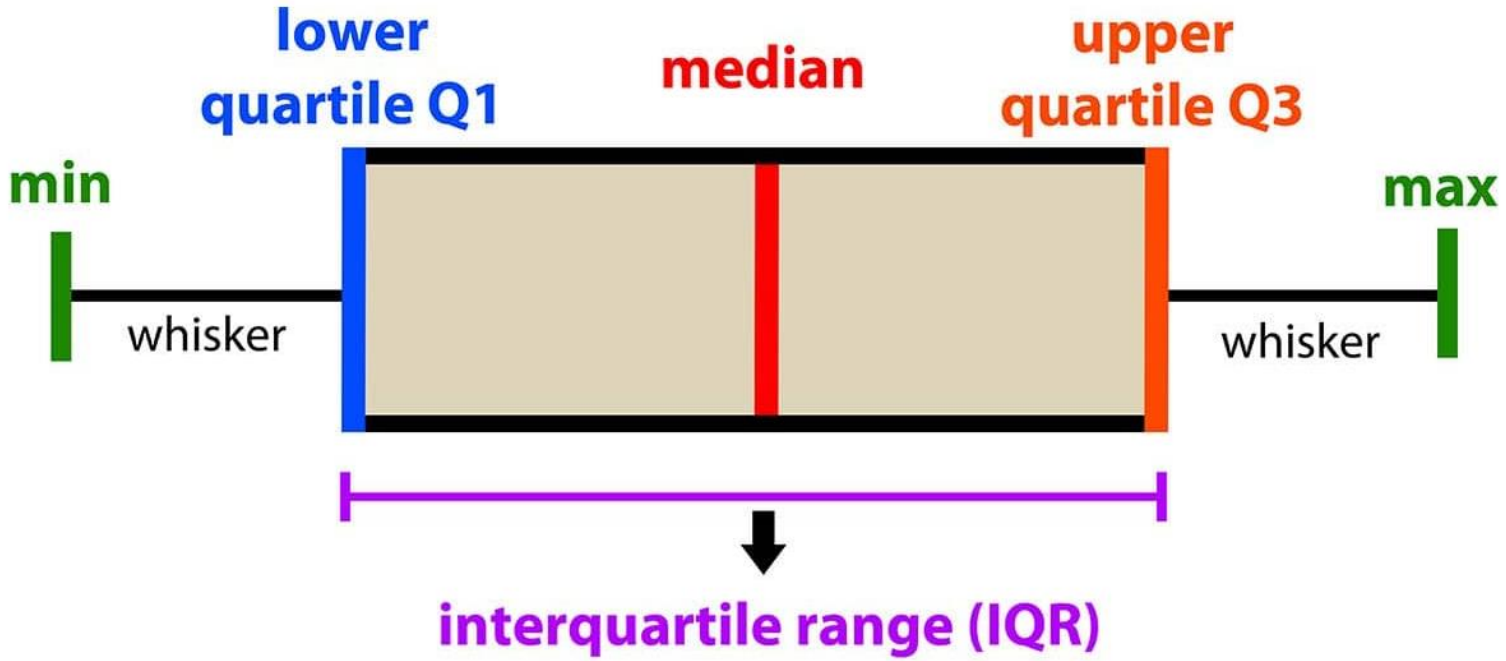
Density Plot

- A density plot, also known as a kernel density plot, is a data visualization technique that provides an estimate of the underlying probability density function of a continuous variable.
- Density plots are particularly useful for understanding the distribution of data when dealing with large datasets or when the data does not follow a specific parametric distribution.



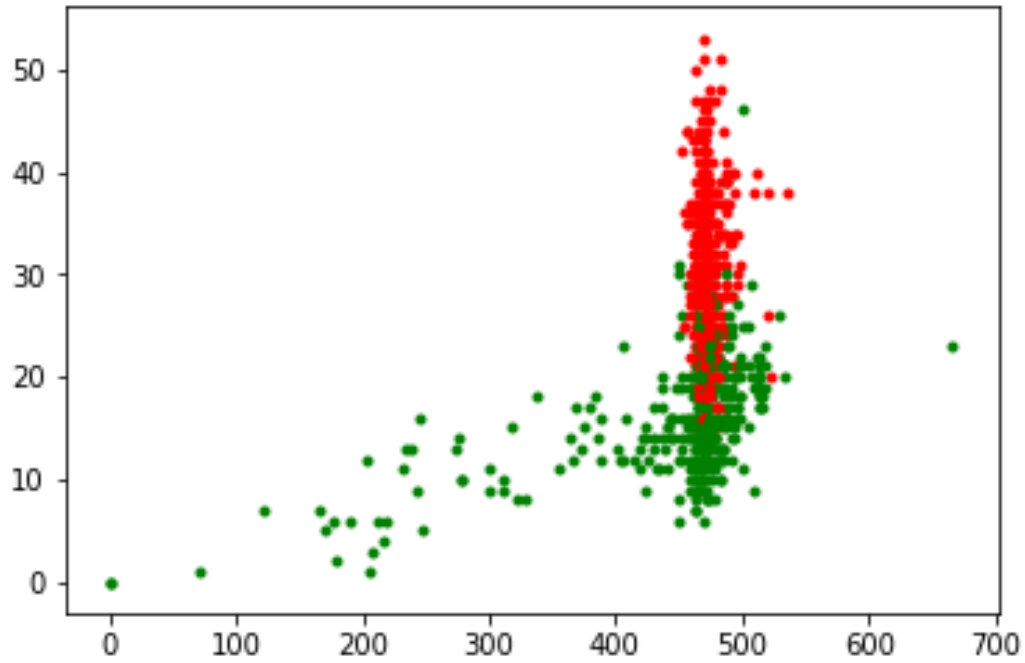
Box and Whiskers Plot

- Box and whiskers plot is beneficial for visualizing the distribution of numerical data.
- Displays a box that represents the interquartile range (IQR), with a line inside representing the median.



Scatter Plot

- A scatter plot shows the relationship between two continuous variables.
- Data points are represented as individual dots on a two-dimensional plane, with one variable on the x-axis and the other on the y-axis.
- Scatter plots are useful for recognizing patterns, trends, and relationships between variables visually.

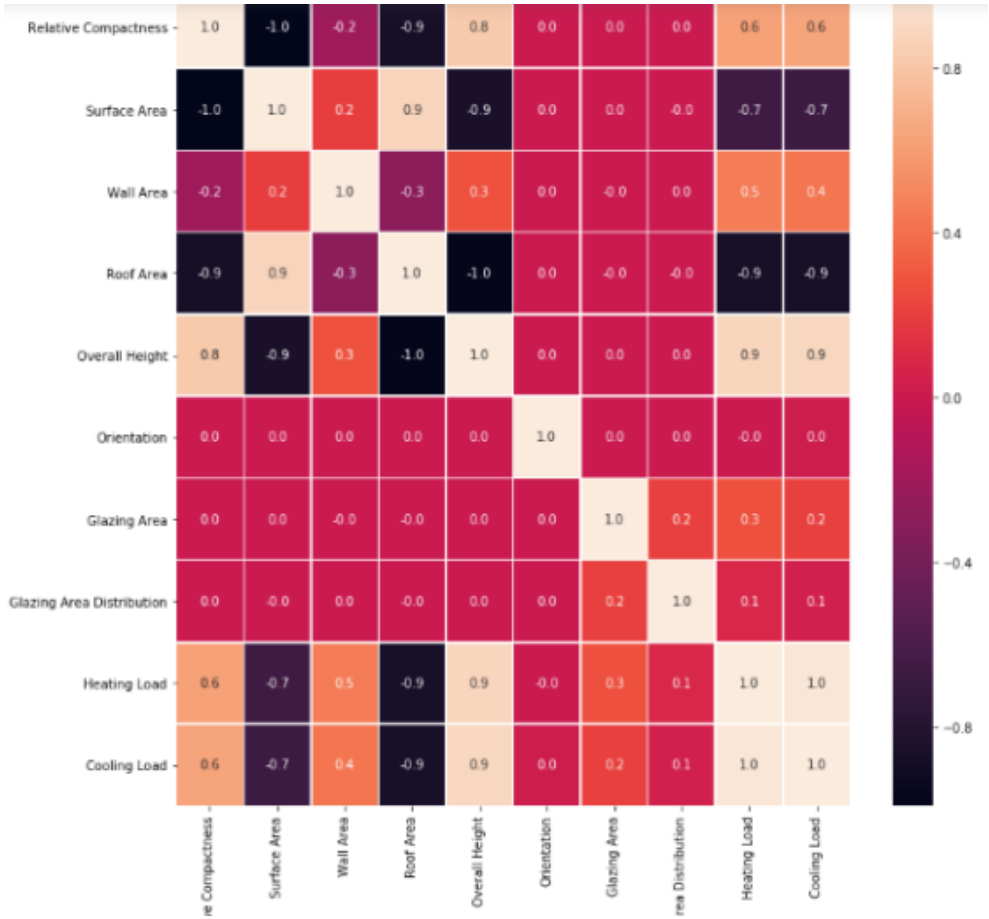


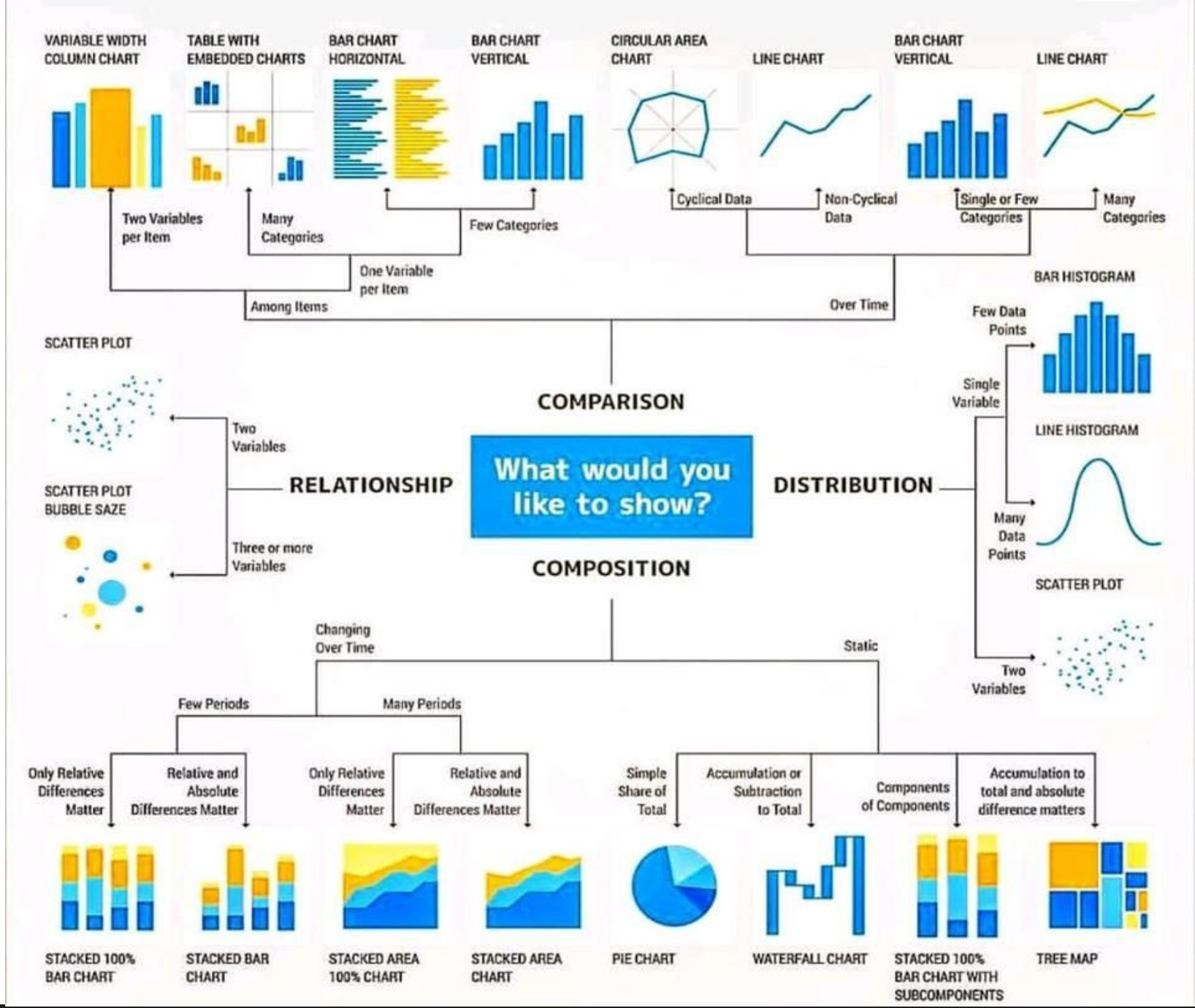
Pairwise Plots

- Pairwise plots, also called scatterplot matrices, are used to visualize the pairwise relationships between multiple variables in a dataset.
- Pairwise plots are advantageous in identifying correlations, clusters, or trends between variables.

Heatmap

- A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors.
- It uses a color gradient to visualize the magnitude or intensity of the values in the matrix.
- Each cell in the matrix is assigned a color based on its value, allowing patterns, trends, and relationships to be easily identified.





DEMO

SUMMARY

- Data visualization is essential for data analysis, narrative development and the clear, engaging communication of information
- It can highlight data quality issues such as missing values, outliers or inconsistencies
- Matplotlib and Seaborn are the major Python libraries for visualization
- When to use which visual tool is the key aspect to be considered before venturing into building the relevant visualizations, which in turn depends on several factors including the count and nature of variables involved