EXPLORATORY DATA ANALYSIS

(EDA)

WHAT IS EDA?

- Exploratory Data Analysis (EDA) is like exploring a new place. You look around, observe things, and try to understand what’s going on.
- Similarly, in EDA data science, you look at a dataset, check out the different parts, and try to figure out what’s happening in the data.
- It involves using statistics and visual tools to understand and summarize data, helping data scientists and data analysts inspect the dataset from various angles without making assumptions about its contents.

- EDA is a significant step to take before diving into statistical modeling or machine learning, to ensure the data is really what it is claimed to be and that there are no obvious errors. It should be part of data science projects in every organization.

Why Exploratory Data Analysis is Important?

- Here are some of the key reasons why EDA is a critical step in the data analysis process:

1. **Understanding Data Structures**
2. **Identifying Patterns and Relationships**
3. **Detecting Anomalies and Outliers**
4. **Testing Assumptions**:
5. **Informing Feature Selection and Engineering**
6. **Optimizing Model Design**
7. **Facilitating Data Cleaning**
8. **Enhancing Communication**:

AIM OF THE EDA

- The goal of EDA is to open-mindedly explore data.
- John W. Tukey: *EDA is detective work… Unless detective finds the clues, judge or jury has nothing to consider.*
- Here, judge or jury is a confirmatory data analysis
- John W. Tukey: *Confirmatory data analysis goes further, assessing the strengths of the evidence.*
- With EDA, we can examine data and try to understand the meaning of variables. What are the abbreviations stand for.

Exploratory vs Confirmatory Data Analysis

EDA	CDA
• No hypothesis at first	• Start with hypothesis
• Generate hypothesis	• Test the null hypothesis
• Uses graphical methods (mostly)	• Uses statistical models

Here’s a Typical Process

- **Look at the Data**: Gather information about the data, such as the number of rows and columns, and the type of information each column contains. This includes understanding single variables and their distributions.
- **Clean the Data**: Fix issues like missing or incorrect values. Preprocessing is essential to ensure the data is ready for analysis and predictive modeling.
- **Make Summaries**: Summarize the data to get a general idea of its contents, such as average values, common values, or value distributions. Calculating quantiles and checking for skewness can provide insights into the data’s distribution.

- **Visualize the Data**: Use interactive charts and graphs to spot trends, patterns, or anomalies. Bar plots, scatter plots, and other visualizations help in understanding relationships between variables. Python libraries like pandas, NumPy, Matplotlib, Seaborn, and Plotly are commonly used for this purpose.
- **Ask Questions**: Formulate questions based on your observations, such as why certain data points differ or if there are relationships between different parts of the data.
- **Find Answers**: Dig deeper into the data to answer these questions, which may involve further analysis or creating models, including regression or linear regression models.

Types of EDA Techniques

Univariate Analysis: Univariate analysis examines individual variables to understand their distributions and summary statistics.

- This includes calculating measures such as mean, median, mode, and standard deviation, and visualizing the data using histograms, bar charts, box plots, and violin plots..

Bivariate Analysis: Bivariate analysis explores the relationship between two variables.

- It uncovers patterns through techniques like scatter plots, pair plots, and heatmaps. This helps to identify potential associations or dependencies between variables.

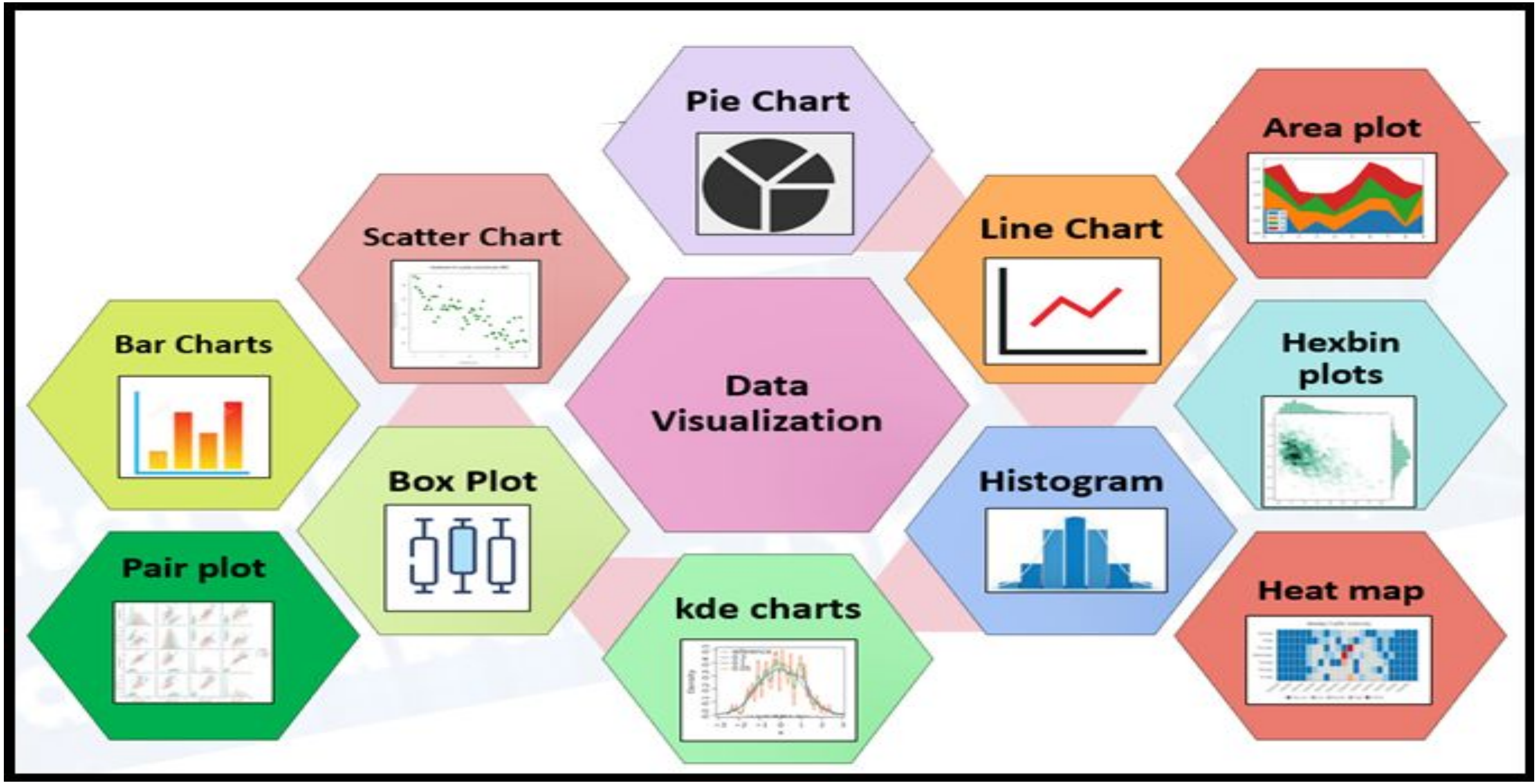
Multivariate Analysis: Multivariate analysis involves examining more than two variables simultaneously to understand their relationships and combined effects.

- Techniques such as contour plots, and principal component analysis (PCA) are commonly used in multivariate EDA.

Visualization Techniques: EDA relies heavily on visualization methods to depict data distributions, trends, and associations. Various charts and graphs, such as bar charts, line charts, scatter plots, and heatmaps, are used to make data easier to understand and interpret.

Outlier Detection: EDA involves identifying outliers within the data—anomalies that deviate significantly from the rest of the data. Tools such as box plots, z-score analysis, and scatter plots help in detecting and analyzing outliers.

- **Statistical Tests:** EDA often includes performing statistical tests to validate hypotheses or discern significant differences between groups. Tests such as t-tests, chi-square tests, and ANOVA add depth to the analysis process by providing a statistical basis for the observed patterns.



Tools for Performing Exploratory Data Analysis

Exploratory Data Analysis (EDA) can be effectively performed using a variety of tools and software, each offering unique features suitable for handling different types of data and analysis requirements.

1. Python Libraries

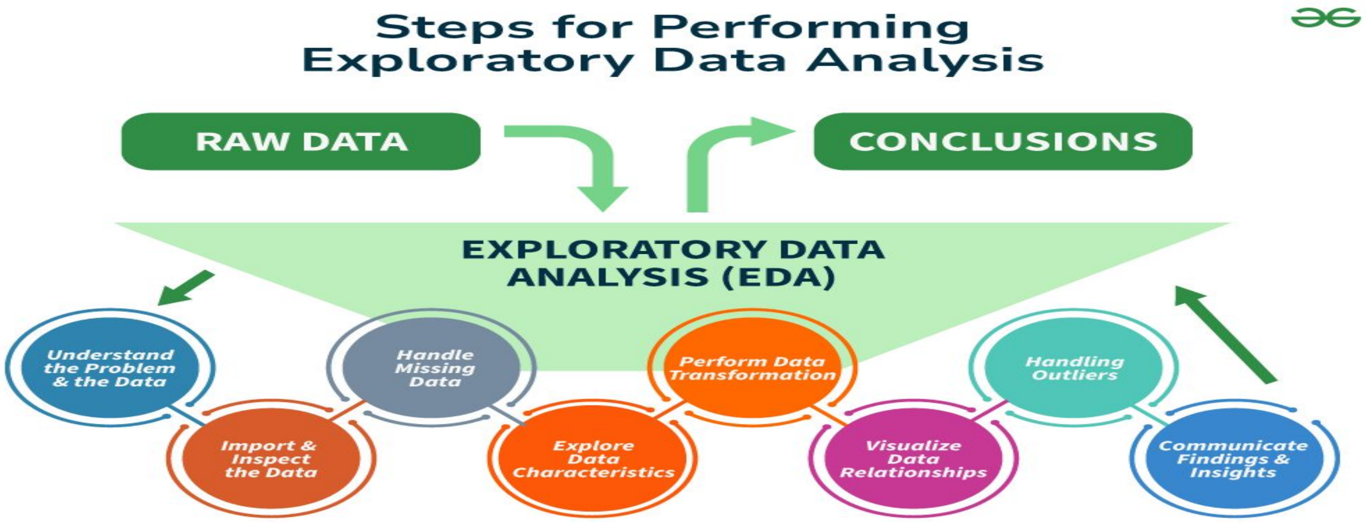
- **Pandas**: Provides extensive functions for data manipulation and analysis, including data structure handling and time series functionality.
- **Matplotlib**: A plotting library for creating static, interactive, and animated visualizations in Python.
- **Seaborn**: Built on top of Matplotlib, it provides a high-level interface for drawing attractive and informative statistical graphics.
- **Plotly**: An interactive graphing library for making interactive plots and offers more sophisticated visualization capabilities.

2. R Packages

- ggplot2: Part of the tidyverse, it's a powerful tool for making complex plots from data in a data frame.
- dplyr: A grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges.
- tidyr: Helps to tidy your data. Tidying your data means storing it in a consistent form that matches the semantics of the dataset with the way it is stored.

Steps for Performing Exploratory Data Analysis

- Performing Exploratory Data Analysis (EDA) involves a series of steps designed to help you understand the data you’re working with, uncover underlying patterns, identify anomalies, test hypotheses, and ensure the data is clean and suitable for further analysis.
-


Steps for Performing Exploratory Data Analysis
RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
Understand the Problem & the Data
Import & Inspect the Data
Handle Missing Data
Explore Data Characteristics
Perform Data Transformation
Visualize Data Relationships
Handling Outliers
Communicate Findings & Insights

- **Step 1: Understand the Problem and the Data**
- The first step in any information evaluation project is to sincerely apprehend the trouble you are trying to resolve and the statistics you have at your disposal. This entails asking questions consisting of:
- What is the commercial enterprise goal or research question you are trying to address?
- What are the variables inside the information, and what do they mean?
- What are the data sorts (numerical, categorical, textual content, etc.) ?
- Is there any known information on first-class troubles or obstacles?
- Are there any relevant area-unique issues or constraints?

Step 2: Import and Inspect the Data

- Once you have clean expertise of the problem and the information, the following step is to import the data into your evaluation environment (e.g., Python, R, or a spreadsheet program). During this step, looking into the statistics is critical to gain initial know-how of its structure, variable kinds, and capability issues.

Step 3: Handle Missing Data

- Missing records is a joint project in many datasets, and it can significantly impact the quality and reliability of your evaluation. During the EDA method, it’s critical to pick out and deal with lacking information as it should be, as ignoring or mishandling lacking data can result in biased or misleading outcomes.

Step 4: Explore Data Characteristics

- After addressing the facts that are lacking, the next step within the EDA technique is to explore the traits of your statistics. This entails examining your variables' distribution, crucial tendency, and variability and identifying any ability outliers or anomalies. Understanding the characteristics of your information is critical in deciding on appropriate analytical techniques, figuring out capability information first-rate troubles, and gaining insights that may tell subsequent evaluation and modeling decisions.

Step 5: Perform Data Transformation

- Data transformation is a critical step within the EDA process because it enables you to prepare your statistics for similar evaluation and modeling.

Step 6: Visualize Data Relationships

Step 7: Handling Outliers

Step 8: Communicate Findings and Insights

Classification of EDA*

- Exploratory data analysis is generally cross-classified in two ways.
  - First, each method is either non-graphical or graphical. And
  - second, each method is either univariate or multivariate (usually just bivariate).
- Non-graphical methods generally involve calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way.
- Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships. Usually our multivariate EDA will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables.
- *It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.*

Data Types and Measurement Scales

- Variables may be one of several types, and have a defined set of

valid values.

- Two main classes of variables are:

Continuous Variables: (Quantitative, numeric).

Continuous data can be rounded or \binned to create categorical data.

Categorical Variables: (Discrete, qualitative).

Some categorical variables (e.g. counts) are sometimes treated as continuous.

Categorical Data

- Unordered categorical data (nominal)

  2 possible values (binary or dichotomous)

  Examples: gender, alive/dead, yes/no.

  Greater than 2 possible values - No order to categories

  Examples: marital status, religion, country of birth, race.

- Ordered categorical data (ordinal)

  Ratings or preferences

  Cancer stage

  Quality of life scales,

  National Cancer Institute's NCI Common Toxicity Criteria (severity grades 1-5)

  Number of copies of a recessive gene (0, 1 or 2)

EDA : Summarizing Data With Tables and Plots

Examine the entire data set using basic techniques before starting a formal statistical analysis.

- Familiarizing yourself with the data.
- Find possible errors and anomalies.
- Examine the distribution of values for each variable.

Summarizing Variables

- Categorical variables

  Frequency tables - how many observations in each category?

  Relative frequency table - percent in each category.

  Bar chart and other plots.

- Continuous variables

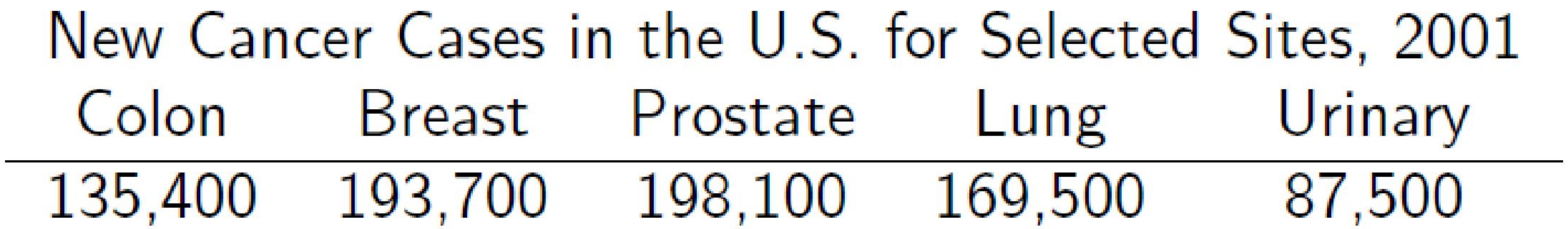
  Bin the observations (create categories .e.g., (0-10), (11-20), etc.) then, treat as ordered categorical.

  Plots specific to Continuous variables.

The goal for both categorical and continuous data is data reduction while preserving/extracting key information about the process under investigation.

Categorical Data Summaries

Tables



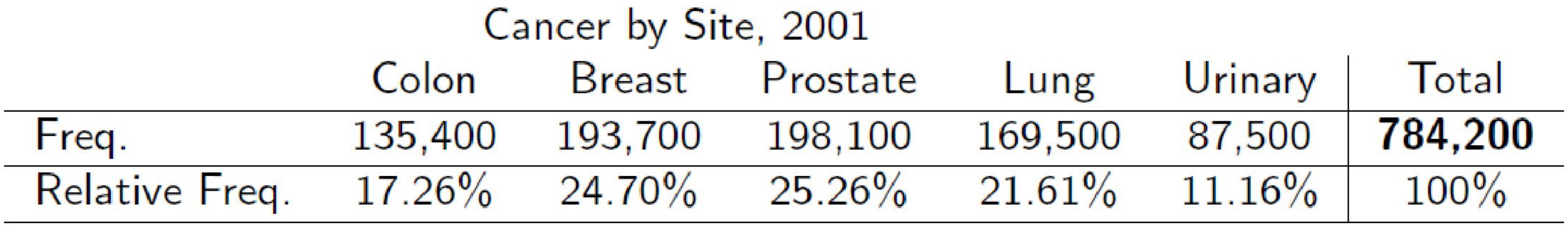
New Cancer Cases in the U.S. for Selected Sites, 2001

Colon	Breast	Prostate	Lung	Urinary
135,400	193,700	198,100	169,500	87,500

Cancer site is a variable taking 5 values

- categorical or continuous?
- ordered or unordered?

Frequency Table



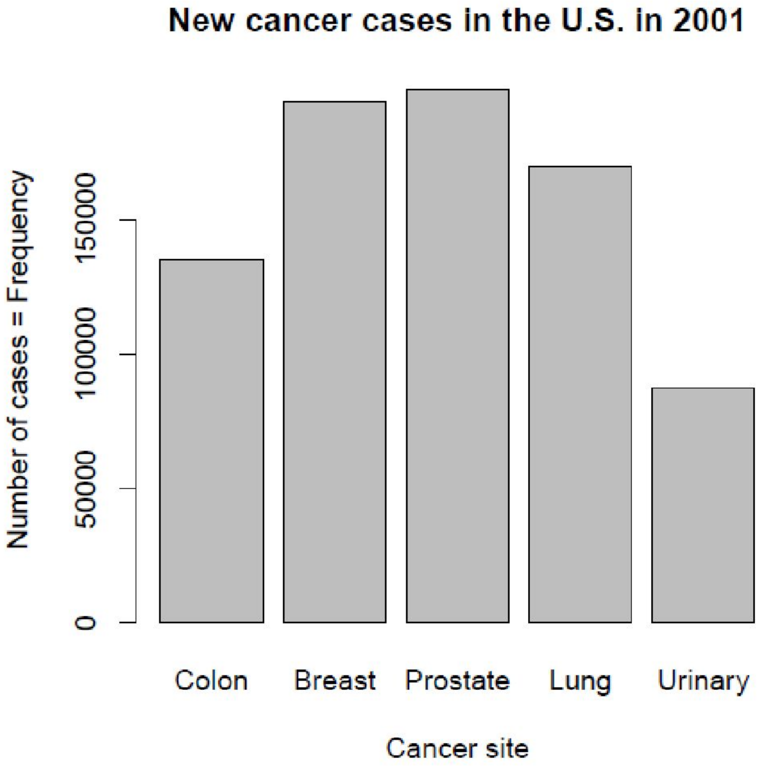
Cancer by Site, 2001

	Colon	Breast	Prostate	Lung	Urinary	Total
Freq.	135,400	193,700	198,100	169,500	87,500	**784,200**
Relative Freq.	17.26%	24.70%	25.26%	21.61%	11.16%	100%

- Frequency Table: Categories with counts
- Relative Frequency Table: Percentage in each category

Graphing a Frequency Table - Bar Chart:

Plot the number of observations in each category:

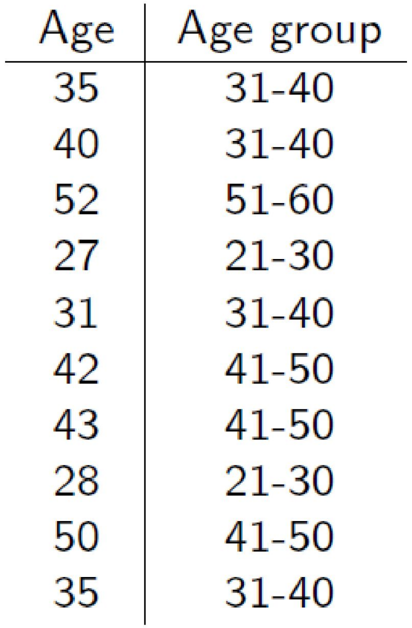


Continuous Data - Tables

Example: Ages of 10 adult leukemia patients:

35; 40; 52; 27; 31; 42; 43; 28; 50; 35

One option is to group these ages into decades and create a categorical age variable:

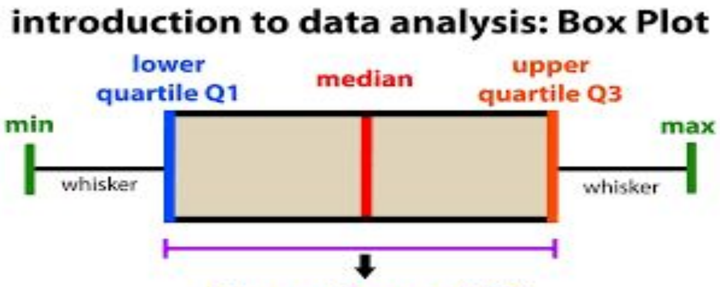


Age	Age group
35	31-40
40	31-40
52	51-60
27	21-30
31	31-40
42	41-50
43	41-50
28	21-30
50	41-50
35	31-40

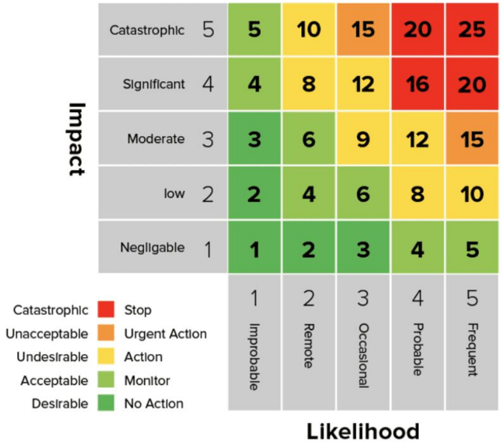
There are several types of data visualization diagrams that can be used to represent different aspects of a dataset. Here are some common ones:

- **Bar Chart**:
  - **Purpose**: Compare different categories or groups.
  - **Example**: Showing the sales of different products.
- **Line Chart**:
  - **Purpose**: Track changes over time.
  - **Example**: Visualizing stock prices over a period.
- **Pie Chart**:
  - **Purpose**: Show proportions or percentages of a whole.
  - **Example**: Representing the market share of companies.
- **Histogram**:
  - **Purpose**: Show the distribution of a single variable.
  - **Example**: Displaying the frequency of test scores in a class.

- **Scatter Plot**:
  - **Purpose**: Show the relationship between two continuous variables.
  - **Example**: Comparing height and weight of individuals.
- **Box Plot (Box-and-Whisker Plot)**:
  - **Purpose**: Summarize the distribution of a dataset.
  - **Example**: Displaying the spread and outliers of exam scores.


introduction to data analysis: Box Plot
lower quartile Q1
median
upper quartile Q3
min
max
whisker
whisker
interquartile range (IQR)

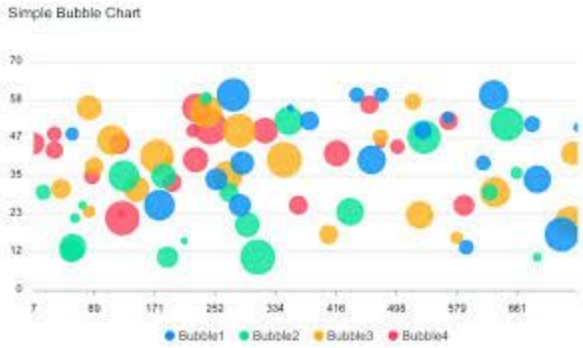
- **Heatmap**:
  - **Purpose**: Show data density or intensity.
  - **Example**: Correlation matrix of multiple variable



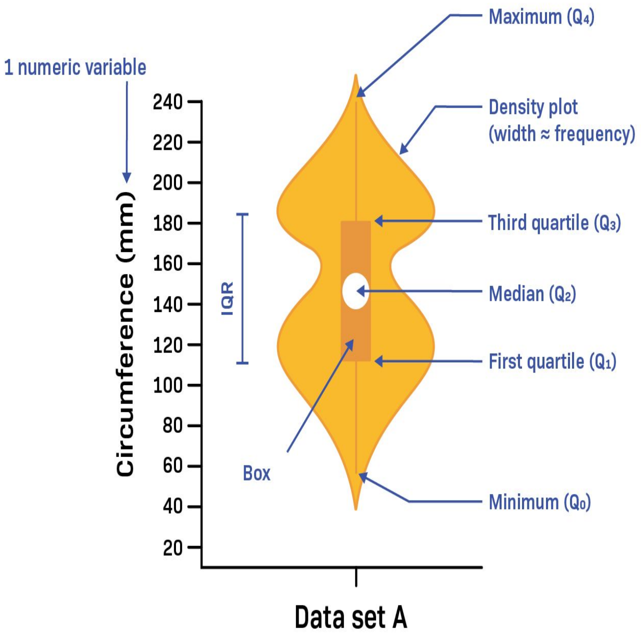
Impact					
Catastrophic 5	5	10	15	20	25
Significant 4	4	8	12	16	20
Moderate 3	3	6	9	12	15
low 2	2	4	6	8	10
Negligable 1	1	2	3	4	5
Likelihood	1 Improbable	2 Remote	3 Occasional	4 Probable	5 Frequent

Catastrophic	🟥	Stop
Unacceptable	🟧	Urgent Action
Undesirable	🟨	Action
Acceptable	🟩	Monitor
Desirable	🟩	No Action

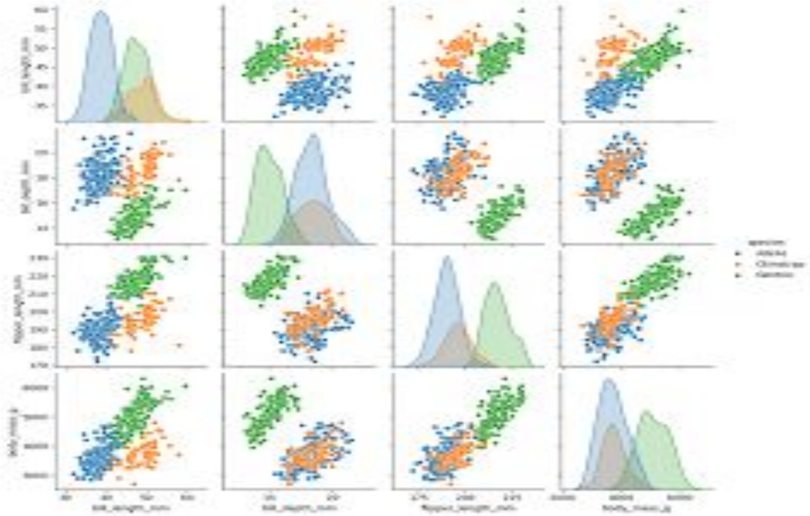
- **Bubble Chart**:
  - **Purpose**: Show relationships between three variables using circles.
  - **Example**: GDP, life expectancy, and population of countries.



- **Violin Plot**:
  - **Purpose**: Show the distribution of the data across different categories.
  - **Example**: Visualizing the distribution of exam scores across different classes.

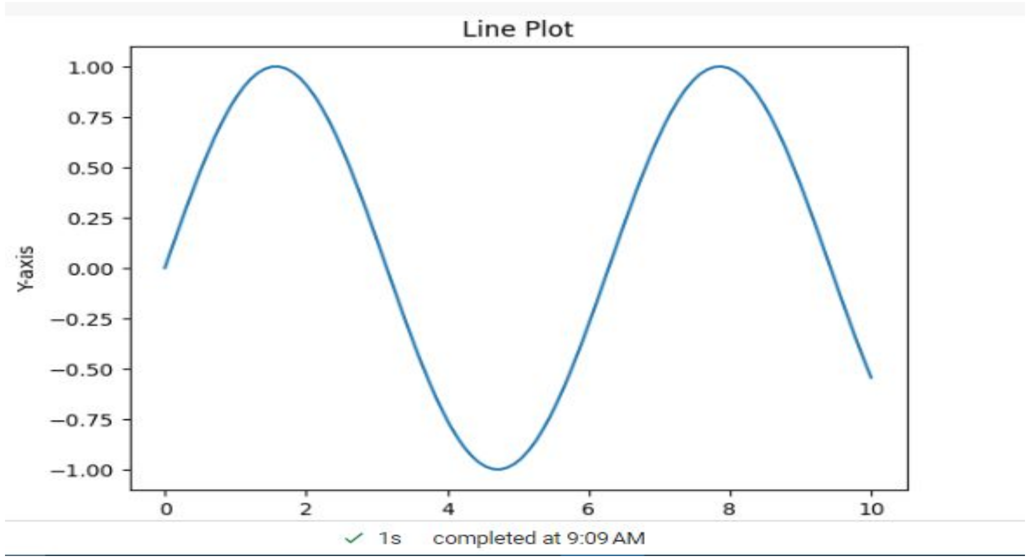


- **Pair Plot**:
  - **Purpose**: Display the pairwise relationship between variables in a dataset.
  - **Example**: Exploring the relationships in a dataset with multiple variables, like the Iris dataset.

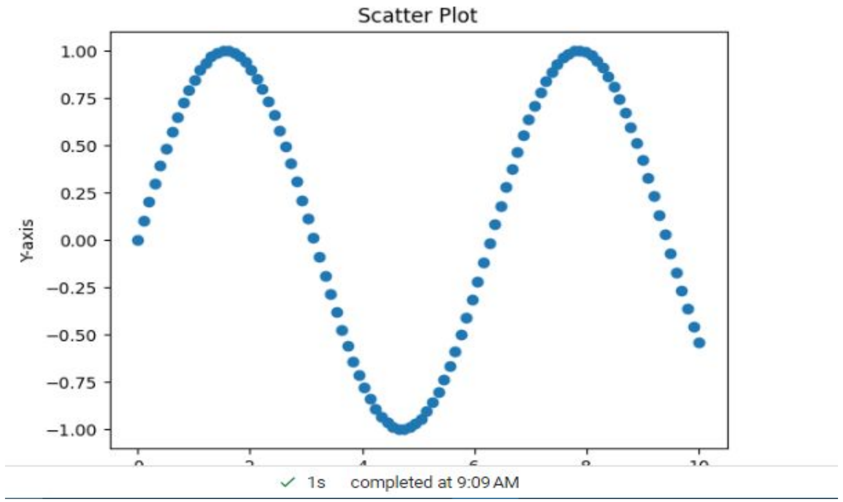


# prompt: data visualization techniques with pyhon code

```
 import matplotlib.pyplot as plt
•import numpy as np

 # Sample data
•x = np.linspace(0, 10, 100)
•y = np.sin(x)

 # Line plot
•plt.plot(x, y)
•plt.xlabel("X-axis")
•plt.ylabel("Y-axis")
•plt.title("Line Plot")
•plt.show()

 # Scatter plot
•plt.scatter(x, y)
•plt.xlabel("X-axis")
•plt.ylabel("Y-axis")
•plt.title("Scatter Plot")
•plt.show()
```



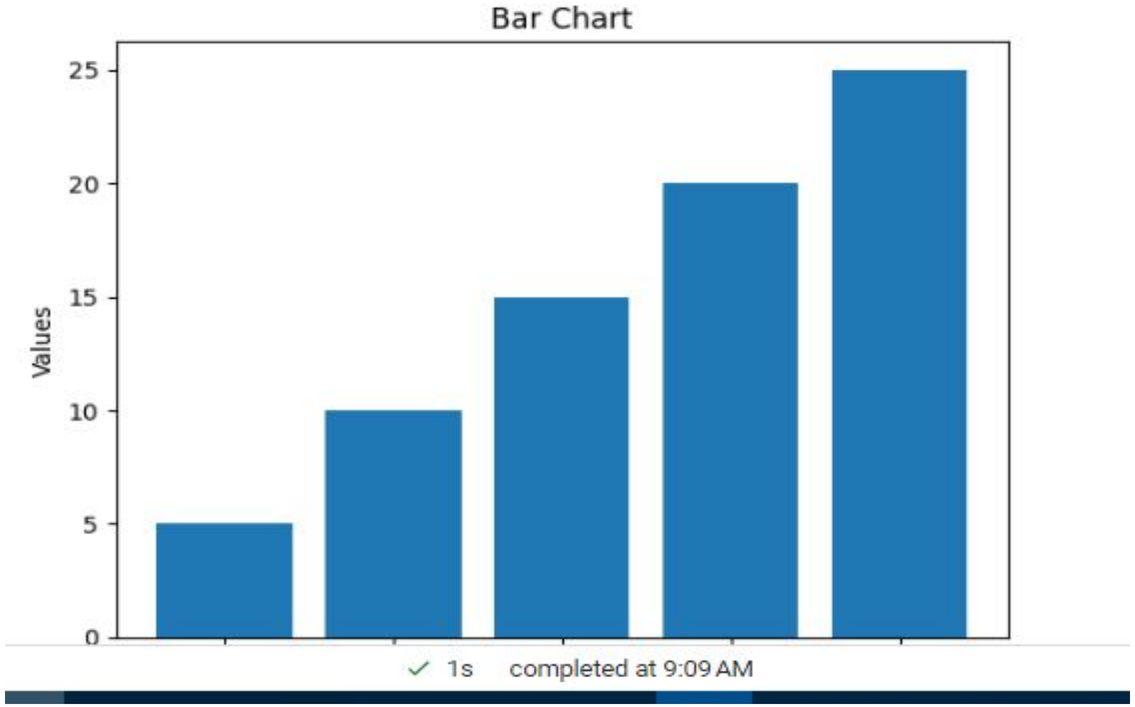
✓ 1s completed at 9:09 AM



✓ 1s completed at 9:09 AM

# Bar chart

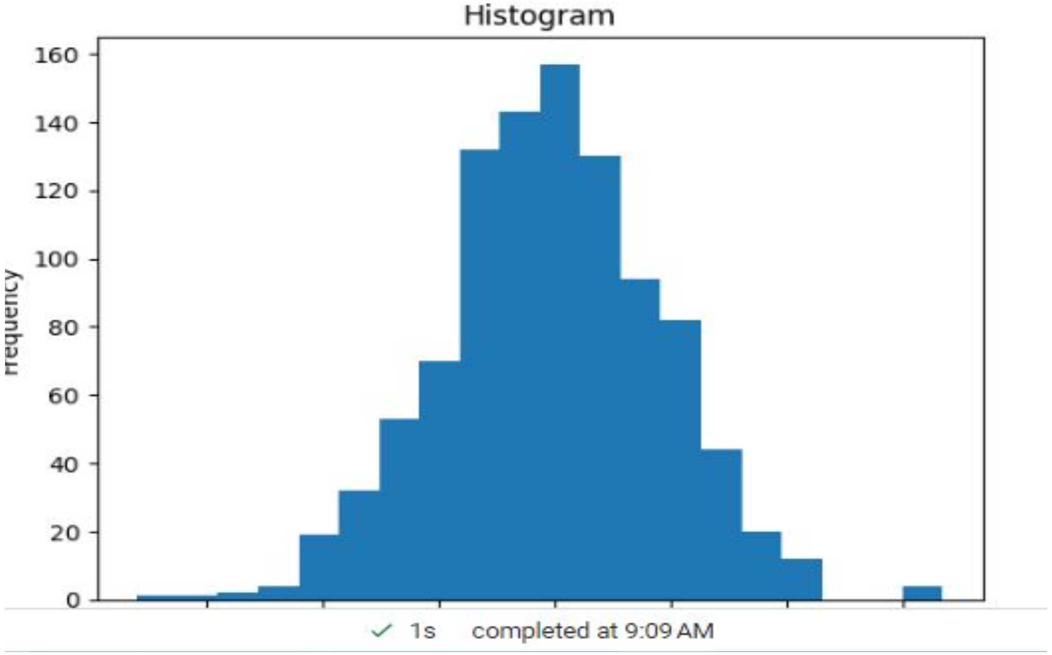
- data = [5, 10, 15, 20, 25]
- labels = ["A", "B", "C", "D", "E"]
- plt.bar(labels, data)
- plt.xlabel("Categories")
- plt.ylabel("Values")
- plt.title("Bar Chart")
- plt.show()



✓ 1s completed at 9:09 AM

# Histogram

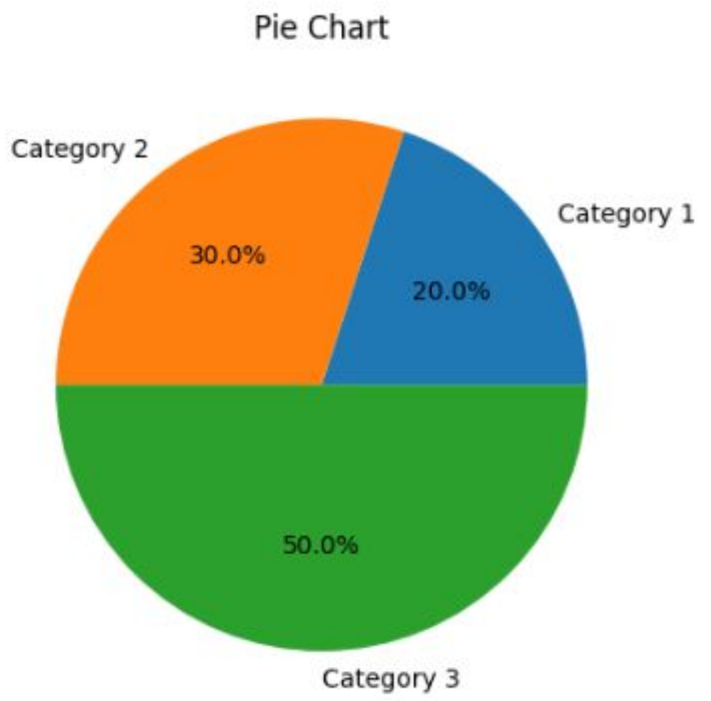
- data = np.random.randn(1000)
- plt.hist(data, bins=20)
- plt.xlabel("Values")
- plt.ylabel("Frequency")
- plt.title("Histogram")
- plt.show()



✓ 1s completed at 9:09 AM

# Pie chart

- data = [20, 30, 50]
- labels = ["Category 1", "Category 2", "Category 3"]
- plt.pie(data, labels=labels, autopct="%1.1f%%")
- plt.title("Pie Chart")
- plt.show()


Pie Chart
Category 2
Category 1
Category 3