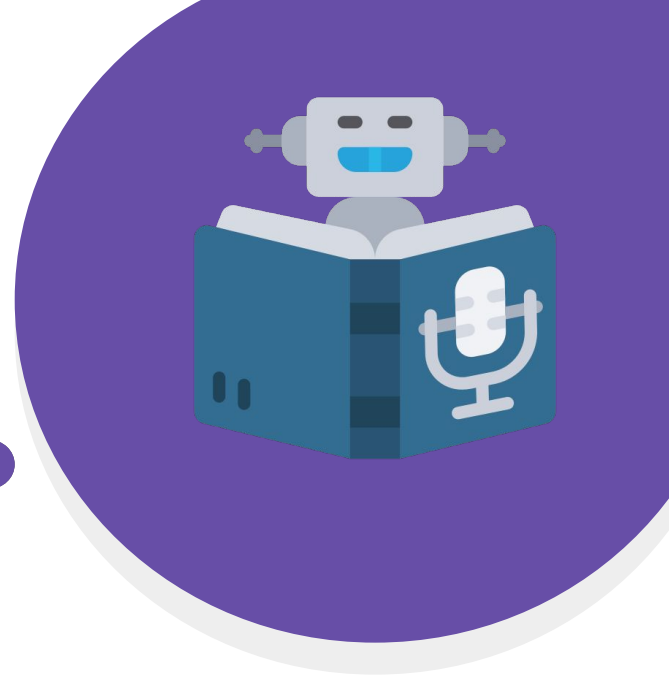


Exploratory Data Analysis

Nisal Mihiranga





Facilitator

Nisal Mihiranga

Areas of Interest & Expertise:

AI, Technology, Science, Teaching, Consulting, Mentoring

Experience:

Head of AI and Data Science,
Architect at
Zone24x7 pvt Ltd
Corporate Trainer

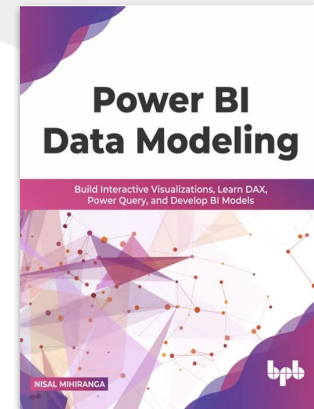
12 Years of Industry exposure
to Data Engineering, Data
Science and Business
Intelligence

Credentials:

M.Sc in Data Science

B.Sc in Information
Technology

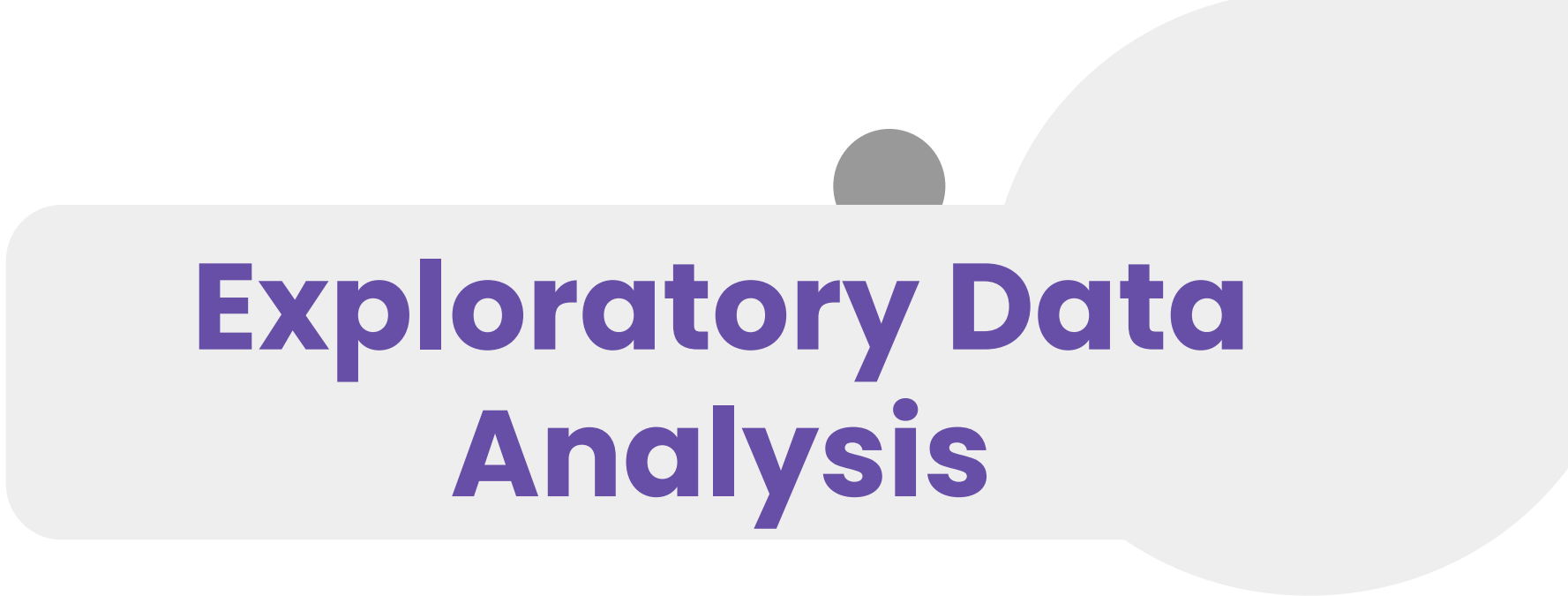
Microsoft Certified Trainer



Week	Module
Week 1	Python for Machine Learning
Week 2	Introduction to Machine Learning
Week 3	Data Transformation and Analysis
Week 4	Regression Analysis
Week 5	Classification, KNN, DT, SVM, Ensemble Systems
Week 6	Clustering Algorithms
Week 7	Neural Networks
Week 8	MLOPS, Machine Learning in Cloud

Learning Outcomes

- Understand the fundamental concepts of Exploratory Data Analysis
- Perform descriptive statistical analyses
- Visualize data effectively using various plotting techniques
- Handle missing data and outliers appropriately
- Apply EDA techniques using Python
- Recognize and implement best practices in EDA



Exploratory Data Analysis

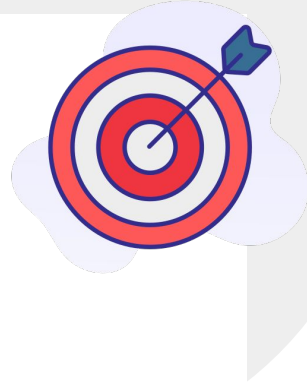
Exploratory Data Analysis

- **Exploratory Data Analysis** (EDA) is an approach to analyzing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- EDA was first introduced by **John W. Tukey** in the 1970s.
- It focuses on discovering patterns, spotting anomalies, testing hypotheses, and checking assumptions through graphical representations and summary statistics.

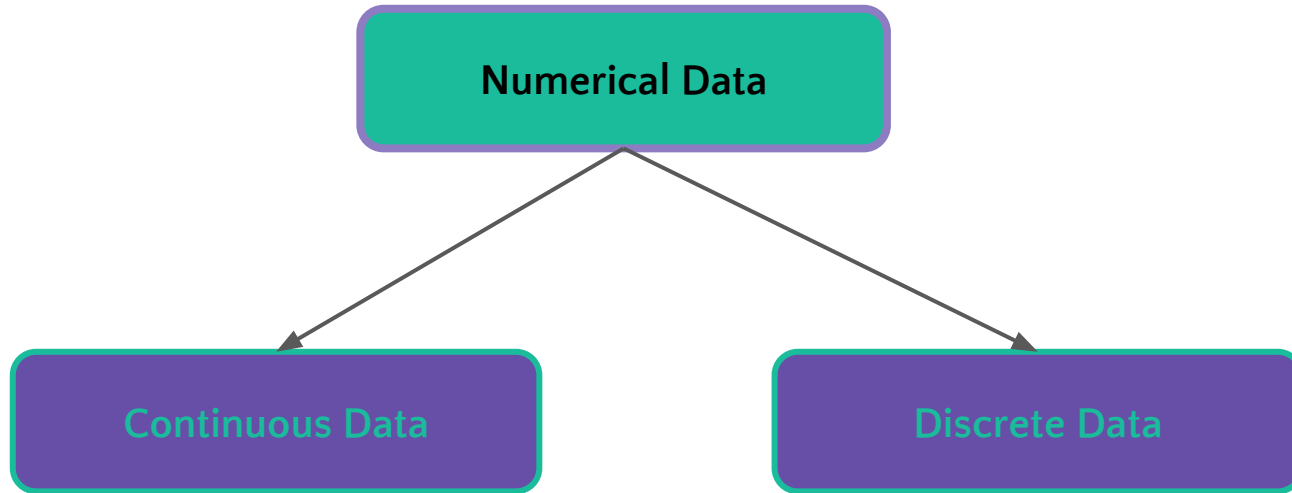


Goal of a Successful EDA

- 1. Gain Insights into Data:**
 - a. Understand the underlying structure and variables in the dataset.
 - b. Identify relationships among variables.
- 2. Detect Anomalies and Outliers:**
 - a. Find unusual observations that may affect the analysis.
- 3. Assess Data Quality:**
 - a. Identify missing values, errors, or inconsistencies.
- 4. Formulate Hypotheses:**
 - a. Develop initial ideas about the data that can be tested with further analysis.
- 5. Guide Modeling Decisions:**
 - a. Inform the selection of appropriate statistical models or algorithms.
- 6. Communicate Findings:**
 - a. Use visualizations to effectively convey data insights to stakeholders.



Data Types: Numerical



- Continuous variables can take on any value within a given range.
- They are measurable quantities and can be infinitely divided into smaller parts.



Height: A person's height can be 170.5 cm, 182.2 cm, etc.

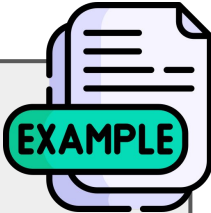
Weight: Measured in kilograms or pounds, like 65.8 kg or 145.2 lbs.

Temperature: Such as 36.6°C, 98.6°F.

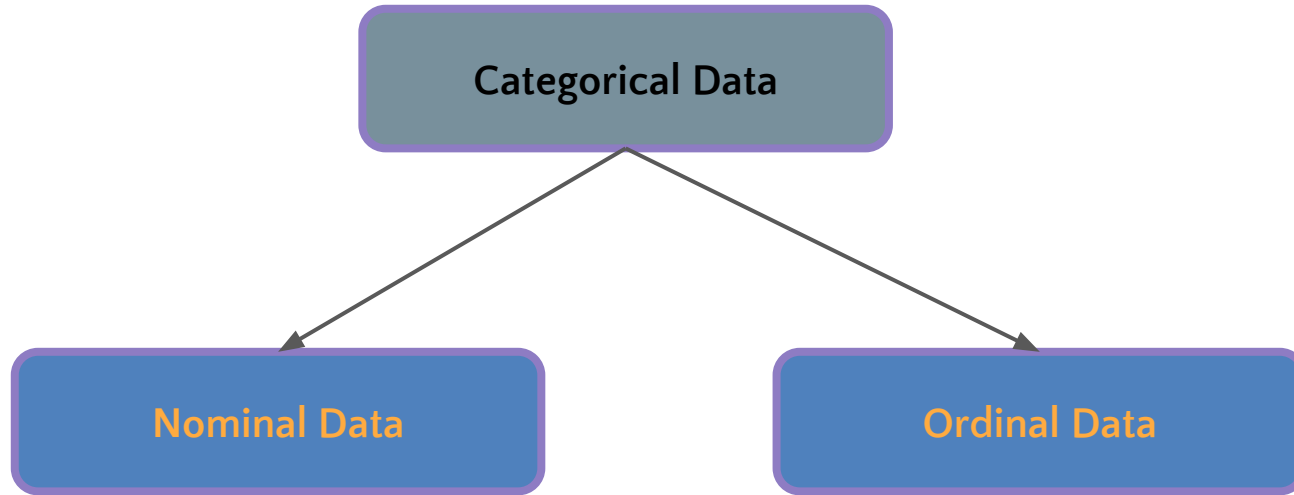
Time: Time taken to run a race, e.g., 9.58 seconds.

Distance: The length between two points, e.g., 5.23 kilometers.

- Discrete variables can only take specific, separate values.
- They are countable and often represent counts or integers.

- 
- Number of Children: 0, 1, 2, 3, etc.
 - Number of Cars Owned: 1 car, 2 cars.
 - Dice Roll Outcome: Possible results are 1, 2, 3, 4, 5, or 6.
 - Exam Scores (if given in whole numbers): 85, 90, 95.

Data Types: Categorical



- Nominal variables categorize data without any intrinsic ordering.
- Categories are mutually exclusive and have no logical order.



- **Gender:** Male, Female, Non-binary.
- **Blood Type:** A, B, AB, O.
- **Marital Status:** Single, Married, Divorced, Widowed.
- **Eye Color:** Blue, Brown, Green, Hazel.

- Ordinal variables represent categories with a meaningful order or ranking.
- The intervals between categories are not necessarily equal or known.

- **Education Level:**
 - High School < Bachelor's Degree < Master's Degree < Doctorate.
- **Customer Satisfaction Rating:**
 - Very Unsatisfied < Unsatisfied < Neutral < Satisfied < Very Satisfied.
- **Socioeconomic Status:**
 - Low < Middle < High.

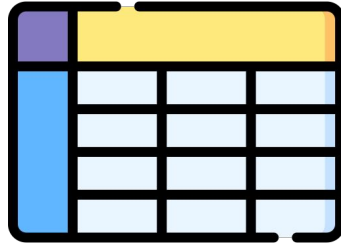


- Data that is organized into a predefined schema, often in tabular format. Easy to search and analyze and Uses rows (records) and columns (fields)

DataFrames in Pandas: The Titanic dataset as a table with rows and columns.



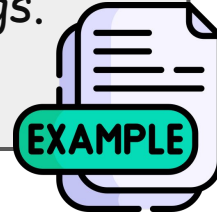
- Data that is organized into a predefined schema, often in tabular format. Easy to search and analyze and Uses rows (records) and columns (fields)



A **DataFrame** is a two-dimensional, tabular data structure commonly used in data analysis, especially in Python with the pandas library. It is similar to a table in a database or an Excel spreadsheet. Each row represents an observation or entry, and each column represents a feature, variable, or field associated with that entry..

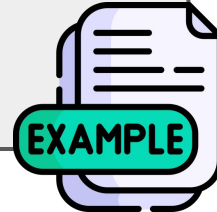
- Data without a predefined format or structure which are more difficult to process and analyze. Requires specialized tools and techniques.

- **Text Files:** Emails, social media posts.
- **Images and Videos:** Photographs, recordings.



- Data that does not conform to a rigid structure but has some organizational properties. Contains tags or markers to separate elements.

- **XML and JSON Files:** Web data, configuration files.



Python Libraries for EDA



Tabular Data



Collection of
mathematical
functions



Visualizations

Rows (Observations):

- Individual data points collected during the study.
- Important to ensure observations are independent when applying certain statistical methods.
 - Each row represents a single data point or record.
 - Example: Each passenger in the Titanic dataset is an observation.

Columns (Variables):

- Each column represents a feature or attribute of the data.
- Example: 'Age', 'Fare', 'Sex' are variables.

Classroom Activity

Duration: *15 mins*



1. Import the titanic Dataset
2. Describes the dimensions of the dataset.
3. Find the examples for continuous, discrete, nominal and ordinal data



Descriptive Statistics

Descriptive Statistics

Descriptive statistics summarize and organize characteristics of a dataset. They provide simple summaries about the sample and the measures, offering a way to understand and interpret data effectively.

- The mean is the **average** of all data points.
- Calculated by summing all values and dividing by the number of observations.

Formula:

$$\text{Mean}(\mu) = \frac{\sum_{i=1}^n x_i}{n}$$

- x_i = each value in the dataset
- n = number of observations

Ex: Mean Age of Passengers

- The median is the **middle value** when data points are arranged in ascending or descending order.
- If the number of observations is even, the median is the average of the two middle numbers.

How to calculate:

1. Arrange data in order.
2. Identify the middle value.

Ex: Median fare

- The mode is the value that appears most frequently in a dataset.
- A dataset can have more than one mode (bimodal, multimodal).

Ex: Mode of Embarked Port:

Determine which embarkation port ('C', 'Q', 'S') occurs most frequently.

- **Mean:** Sensitive to outliers; best for symmetric distributions without extreme values.
- **Median:** Not affected by outliers; useful for skewed distributions.
- **Mode:** Best for categorical data to identify the most common category.

Measure of Dispersion

Measures of dispersion describe the spread or variability in a dataset.

- Range
- Variance
- Standard Deviation
- Interquartile Range

20 mins Break



- The range is the difference between the maximum and minimum values.
- Simplest measure of dispersion.

Formula:

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

Ex: Range of Age

Subtract the youngest passenger's age from the oldest passenger's age.

- Variance measures the average squared deviation from the mean.
- Indicates how data points are spread out around the mean.

Formula:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1}$$

- σ^2 = variance
- x_i = each value
- μ = mean
- n = number of observations

Ex: Variance of fare: Calculate how much fare prices vary from the mean fare.

Standard Deviation

- The standard deviation is the square root of the variance.
- Represents the average distance from the mean.

Formula:

$$\sigma = \sqrt{\sigma^2}$$

Ex: SD of Age: Understand how much passenger ages vary around the mean age.

Interquartile Range

- The IQR measures the middle 50% of data.
- Calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1).
-

Formula:

$$\text{IQR} = Q3 - Q1$$

Ex: IQR of fare to determine the spread of the middle half of fare prices.

Summary of Dispersion Measures

Range: Gives a quick sense of the spread but is affected by outliers.

Variance and Standard Deviation: Provide information on overall variability.

IQR: Useful for identifying outliers and understanding the spread of the central portion of data.

Understanding the shape of a dataset's distribution helps in selecting appropriate statistical models and in interpreting the data correctly.

Skewness:

- Skewness measures the asymmetry of the distribution.
- A skewness of zero indicates a symmetrical distribution.

Types of Skewness

Positive Skew (Right Skew):

- Tail on the right side is longer.
- $\text{Mean} > \text{Median} > \text{Mode}$.

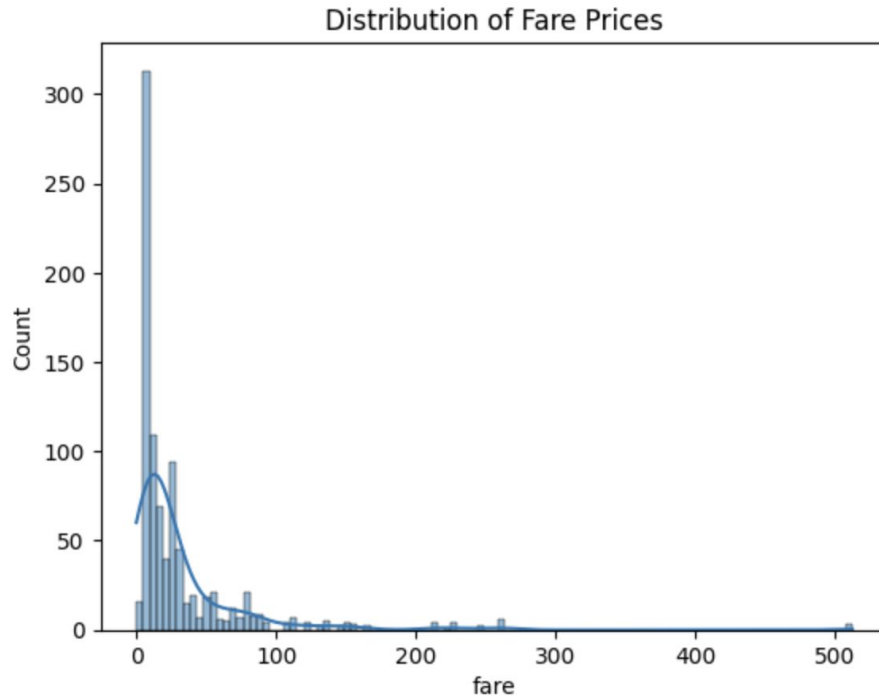
Negative Skew (Left Skew):

- Tail on the left side is longer.
- $\text{Mean} < \text{Median} < \text{Mode}$.

Ex: Fare Distribution: Often positively skewed due to high fare outliers.

Visualize the Distribution

Histogram: Visualize the distribution and identify skewness.



Classroom Activity

Duration: *20 mins*

Objective: Practice calculating descriptive statistics using the Titanic dataset.

1. Calculate the central tendency
 - Mean Age, Median fare, Mode of embark port
2. Calculate Measures of Dispersion:
 - Range of age, Standard deviation of fare, IQR of Age
3. Analyze the distribution:
 - Skewness of fare
4. Visualize data:
 - Histogram of Age, Box Plot of Fare by Pclass



Interpretation and Insights

1. Mean vs. Median:

- a. When They Differ:
 - i. A significant difference between mean and median indicates a skewed distribution.
- b. Example:
 - i. If the mean fare is higher than the median fare, the distribution is right-skewed due to high fare outliers.

2. High Standard Deviation:

Indicates that data points are spread out over a wider range.

Example:

- a. A high standard deviation in fare prices suggests significant variability in ticket costs.

2. Skewness and Data Transformation:

- a. Addressing Skewness:
 - i. Apply transformations (e.g., log transformation) to normalize data.
- b. Example:
- c. Applying a log transformation to fare prices may reduce skewness.

Considerations: Outliers

- Outliers can significantly impact the mean.
- Consider using the median in skewed distributions.

Considerations: Missing Data

- Missing values can distort statistical measures.
- Handle missing data appropriately before calculating statistics.

Example:

Before calculating the mean age, ensure that missing 'Age' values are addressed.

Classroom Activity

Duration: *10 mins*



Compare the survival rate by Age

1. Analyze whether age influenced survival chances.
2. Compare the mean ages to see if younger or older passengers had higher survival rates.

Classroom Activity

Duration: *10 mins*



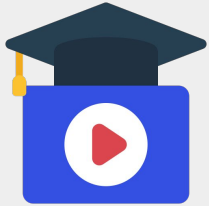
Analyze the fare by Embarked Port

1. Determine if fare prices varied by port of embarkation.
2. Compare the mean ages to see if younger or older passengers had higher survival rates.

Learning Resources



- [Statistics for Data Science by James D. Miller](#)
- [Practical Statistics for Data Scientists by Peter Bruce and Andrew Bruce](#)



- Khan Academy - Descriptive Statistics
- [Coursera - Basic Statistics](#)



- [Pandas Documentation](#)
- Numpy Documentation



Thank You



Nisal Mihiranga
Head of AI & DS at **Zone24x7**
Consultant, Trainer
M: +94 71 726 3044

