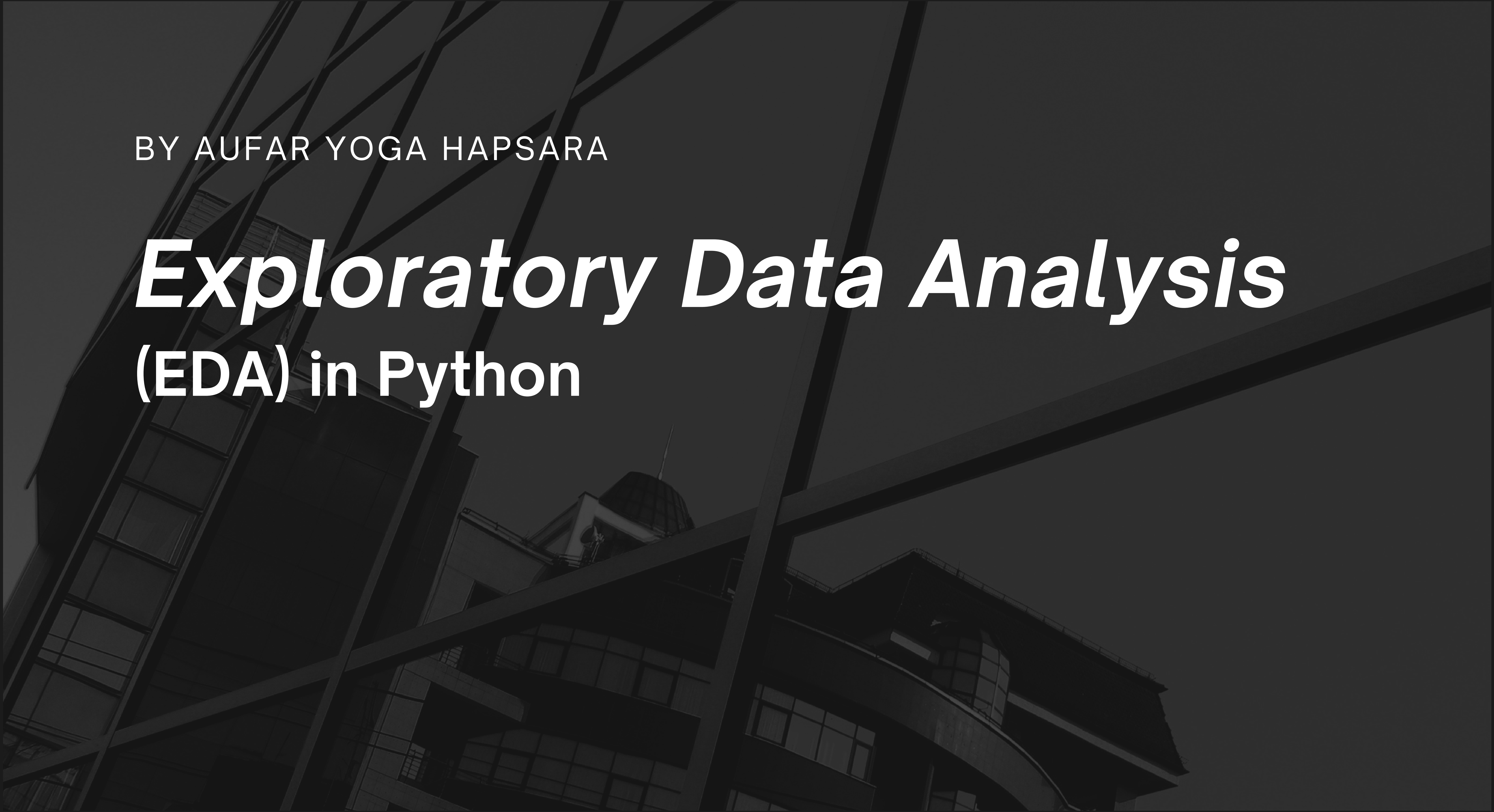


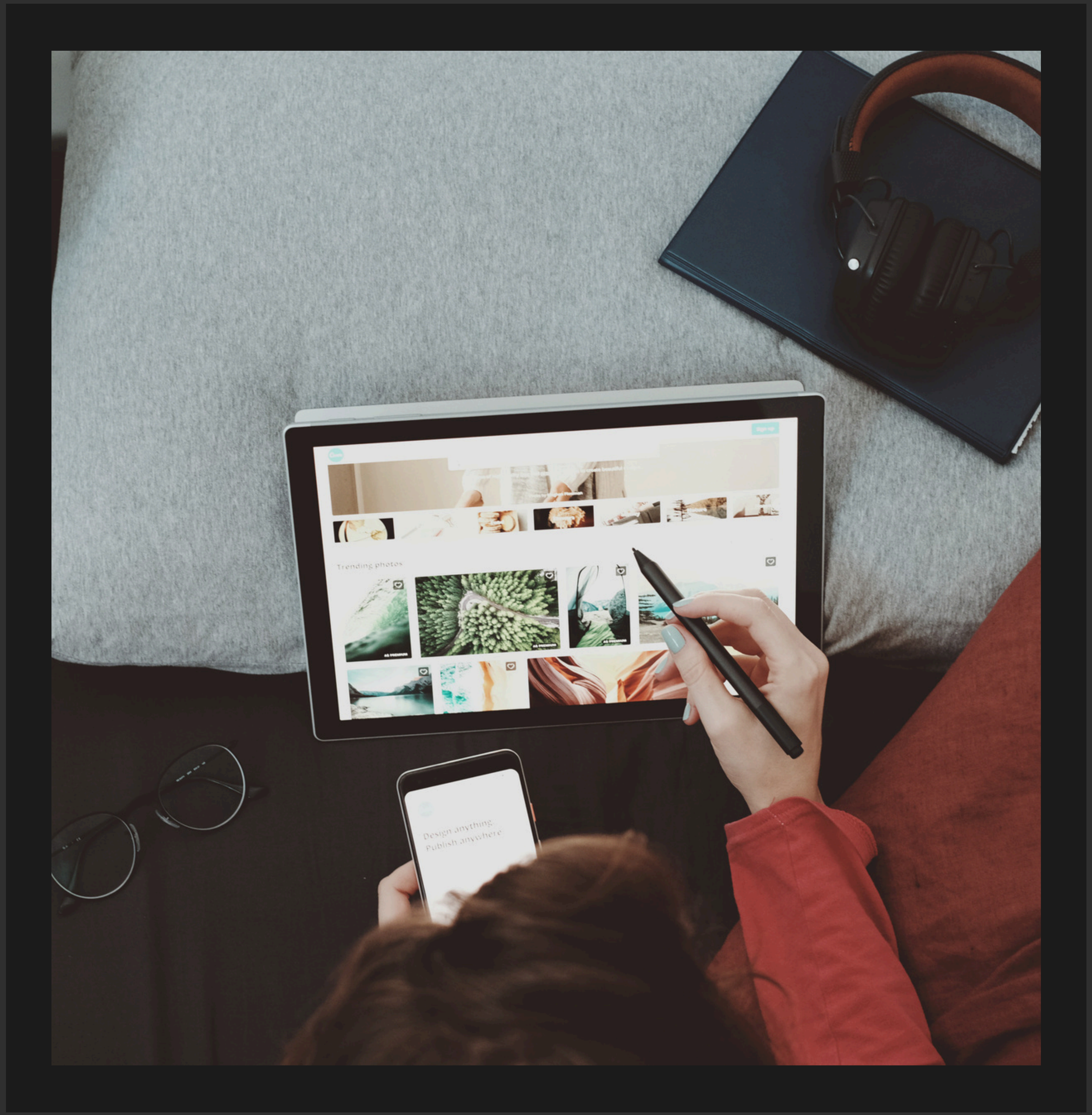
BY AUFAR YOGA HAPSARA

Exploratory Data Analysis

(EDA) in Python

What is Exploratory Data Analysis (EDA)?

is the process of examining and visualizing datasets to summarize their main characteristics, identify patterns, detect anomalies, test hypotheses, and check assumptions all before applying formal modeling techniques.



Purpose of EDA:

- To understand the structure and quality of the data
- To identify missing values, duplicates, or outliers
- To explore relationships between variables
- To gain insights that help in feature selection and model preparation

Several steps to perform in EDA

1. Input and inspect the data
2. Dropping irrelevant column
3. Handle missing value
4. Handling duplicated data
5. Handling outliers
6. Analyse data
7. etc.

Libraries that are commonly used

NUMPY

is used for scientific computing using high-performance arrays

PANDAS

is used for data manipulation and analysis

MATPLOTLIB

is used for creating static, animated, and interactive visualizations

SEABORN

is a data visualization library built on top of Matplotlib

Inspect the Data

1

DATA.SHAPE

is used to check the number of rows (observations) and columns (features) in a dataset.

2

DATA.HEAD()

is used to display the first few rows of a dataset (by default, the first 5 rows)

3

DATA.INFO()

It is useful for understanding the structure and completeness of the data.

4

DATA.DESCRIBE()

generates descriptive statistics of the numerical columns in a dataset. It includes metrics such as count, mean, standard deviation, minimum, maximum, and quartiles.

Data Celaning

1. **DROPPING IRRELEVANT COLUMN**

   In certain cases, there are columns/features that are rarely used, so they can be removed

2. **HANDLE MISSING DATA**

   refers to the process of dealing with data entries that are incomplete or contain null/NaN values.

3. **HANDILING DUPLICATED DATA**

   involves detecting and removing or resolving rows that appear more than once in a dataset.

4. **HANDLING OUTLIERS**

   detecting and addressing unusually high or low values that can distort statistical analyses and machine learning models.

Handle missing data

DROPPING DATA

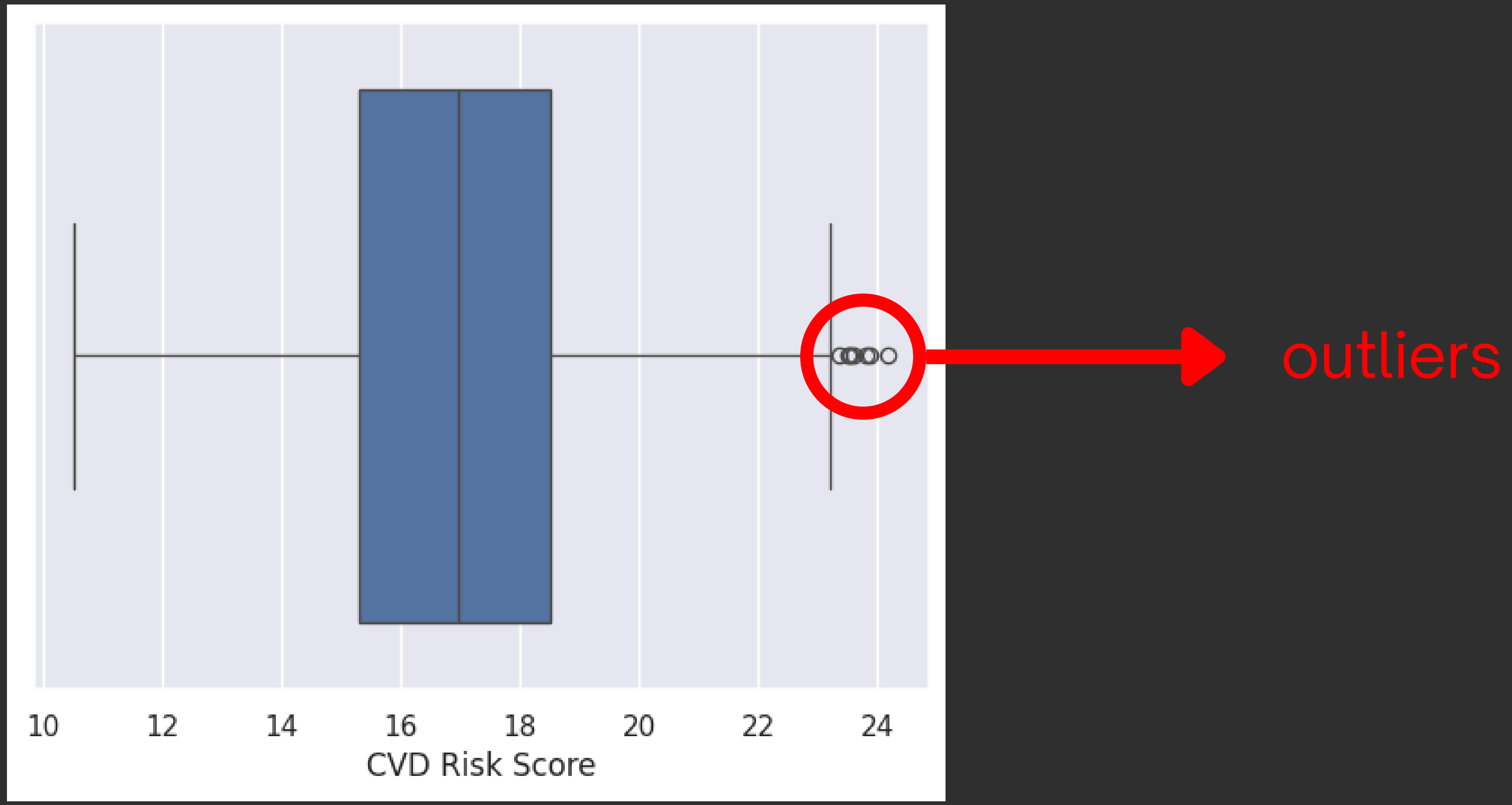
is a method of handling missing values by removing rows or columns that contain null or missing entries.

DATA IMPUTATION

is the process of filling in missing values with estimated or substitute values, such as the mean, median, mode, or using advanced techniques like regression or KNN.

Handling outliers

Outliers are data points that deviate significantly from the majority of other values in a dataset. They can be much larger or much smaller compared to other values. Outliers can have a significant impact on statistical analysis results and can affect the performance of machine learning models in making predictions.



Data analysis that is simple and commonly used

- **UNIVARIATE ANALYSIS**

  it deals with analyzing data within a single column or variable and is mostly used to describe data.

- **BIVARIATE ANALYSIS**

  involves analyzing data with two variables or columns. This is usually a way to explore the relationships between these variables and how they influence each other, if at all.

- **MULTIVARIATE ANALYSE**

  Data analysis that involves more than two variables.

Thank you for your attention

"This is only a part of the many processes that form the basis of conducting EDA, so there are still many other types of data analysis."