SNS COLLEGE OF TECHNOLOGY

Coimbatore - 35

23BAE734 – Visual Analytic and Storytelling

Unit II – Tools and Techniques

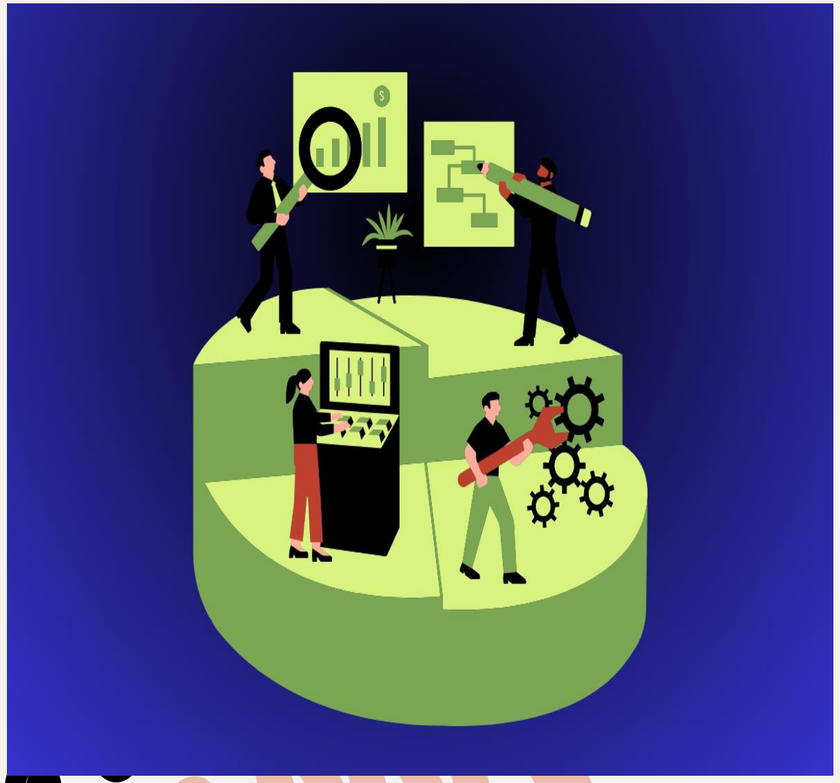
Presented by

Ms.S.D.Shamini

Design Thinker

Guess the Topic!!!

Handling Missing Data and Outliers

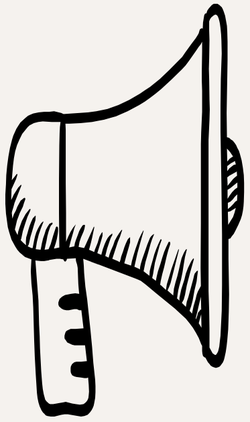


Recap

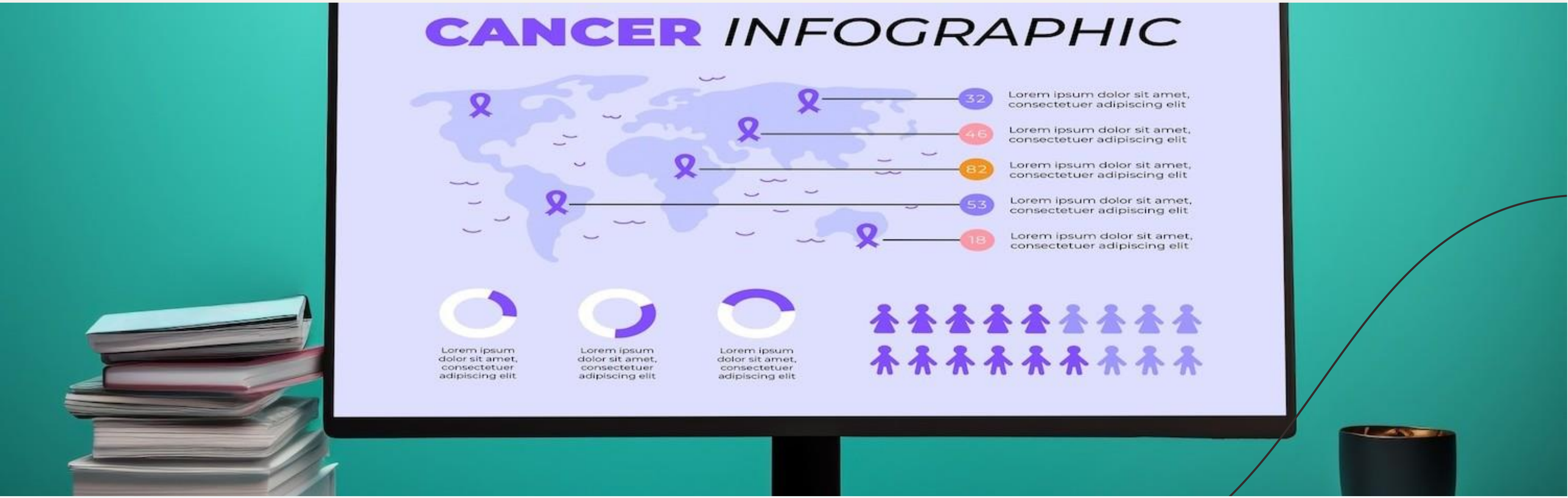
- Reason to Study Data Cleaning
- Data Cleaning
- Identifying and Handling Missing Data
- Detecting and Removing Outliers
- Handling Inconsistent and Duplicate Data
- Data Transformation and Normalization
- Feature Engineering and Selection
- Best Practices for Effective Data Preprocessing

Discussion about....

- Missing Data
- Types of Missing Data
- Strategies for Handling Missing Data
- Outliers
- Handling outliers

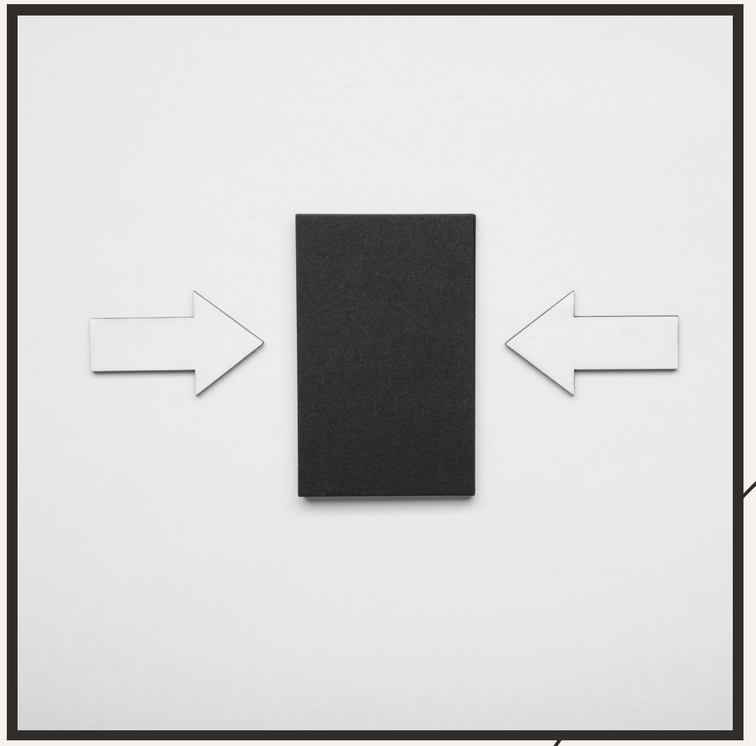


Missing data can significantly impact the quality of data analysis. Understanding the **causes** and **implications** of missing values is crucial for effective management. This presentation explores various strategies to handle missing data and **outliers** to ensure robust and reliable analysis.



Types of Missing Data

Missing data can be classified into three types: MCAR (Missing Completely At Random), MAR (Missing At Random), and MNAR (Missing Not At Random). Identifying the type of missing data is essential for selecting the appropriate imputation method and ensuring accurate results.

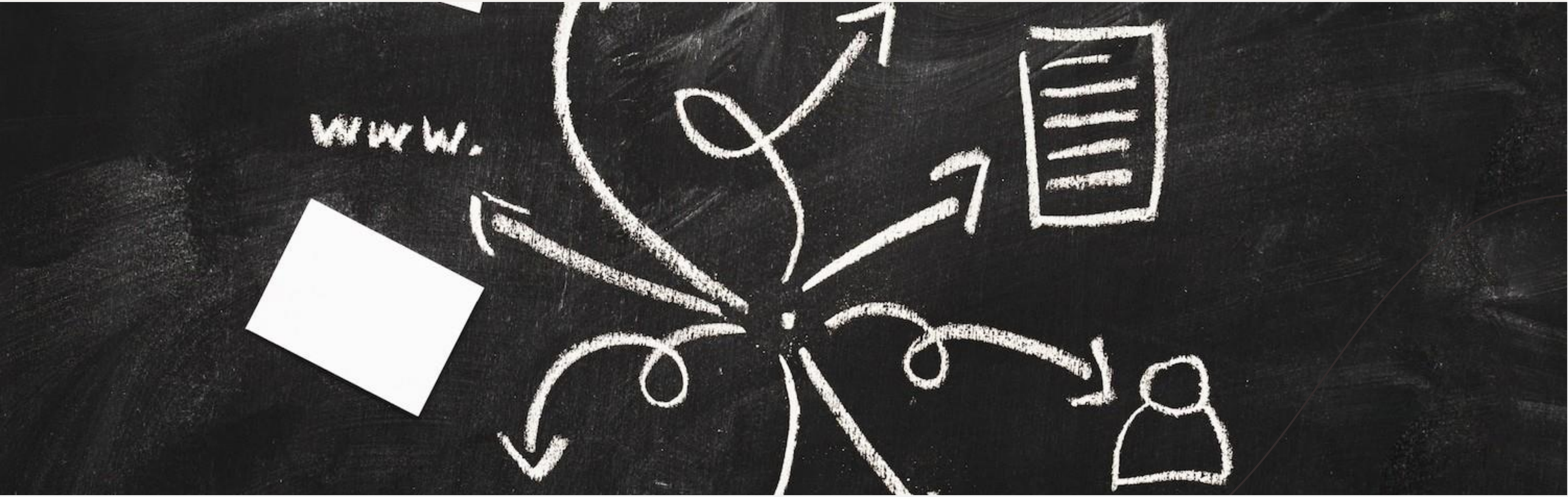


Strategies for Handling Missing Data

Effective strategies for managing missing data include **deletion methods**, **mean/mode imputation**, and **multiple imputation**. Each method has its advantages and disadvantages, and the choice should depend on the data characteristics and the analysis goals.

Outliers can skew results and lead to misleading conclusions. Techniques for identifying outliers include **statistical tests**, **visualizations** like box plots, and **Z-scores**. Recognizing outliers is the first step in deciding how to manage them effectively.

Once identified, outliers can be managed through methods such as **capping**, **transformation**, or even **removal**. The choice of method should be guided by the data context and the potential impact on the analysis outcomes.



Knowledge Check

Which of the following techniques is commonly used to handle missing data in a dataset before visualization?

A. Remove all missing data
B. Replace missing values with the mean, median, or mode
C. Ignore missing data and proceed
D. Delete the entire dataset

B) Replace missing values with the mean, median or mode

Summary

- Missing Data
- Types of Missing Data
- Strategies for Handling Missing Data
- Outliers
- Handling outliers

References

- https://www.linkedin.com/advice/3/how-can-you-identify-outliers-missing-values-your-dj0nc

REACH US

snsinstitutions

snsinstitutions

snsinstitutions

snsinstitutions

snsinstitutions

Thanks!