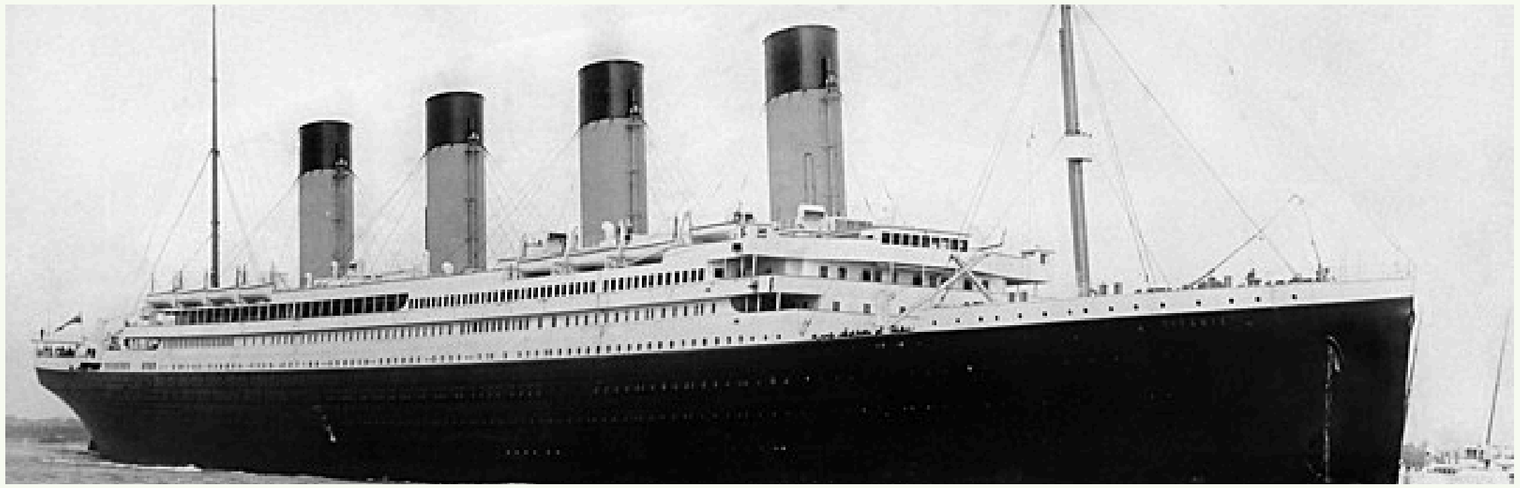


TITANIC

ANALISIS DATA EKSPLORASI (EDA) MENGGUNAKAN PYTHON

HALO,

I AM ALDI PUTRA PRATAMA

I will do an analysis of the Titanic dataset.

PROBLEMS

- Many Titanic passenger records have unknown survival status.
- The dataset contains duplicate entries and missing values.
- We need to identify patterns to understand who was more likely to survive.

OBJECTIVE

- Clean the dataset.
- Analyze the data distribution.
- Discover which factors most affected passengers' chances of survival.

COMPLETION STEPS

- ## IMPORT DATA & DATA PREVIEW

  Importing datasets into the environment using the pandas library.

- ## INITIAL DATA INSPECTION

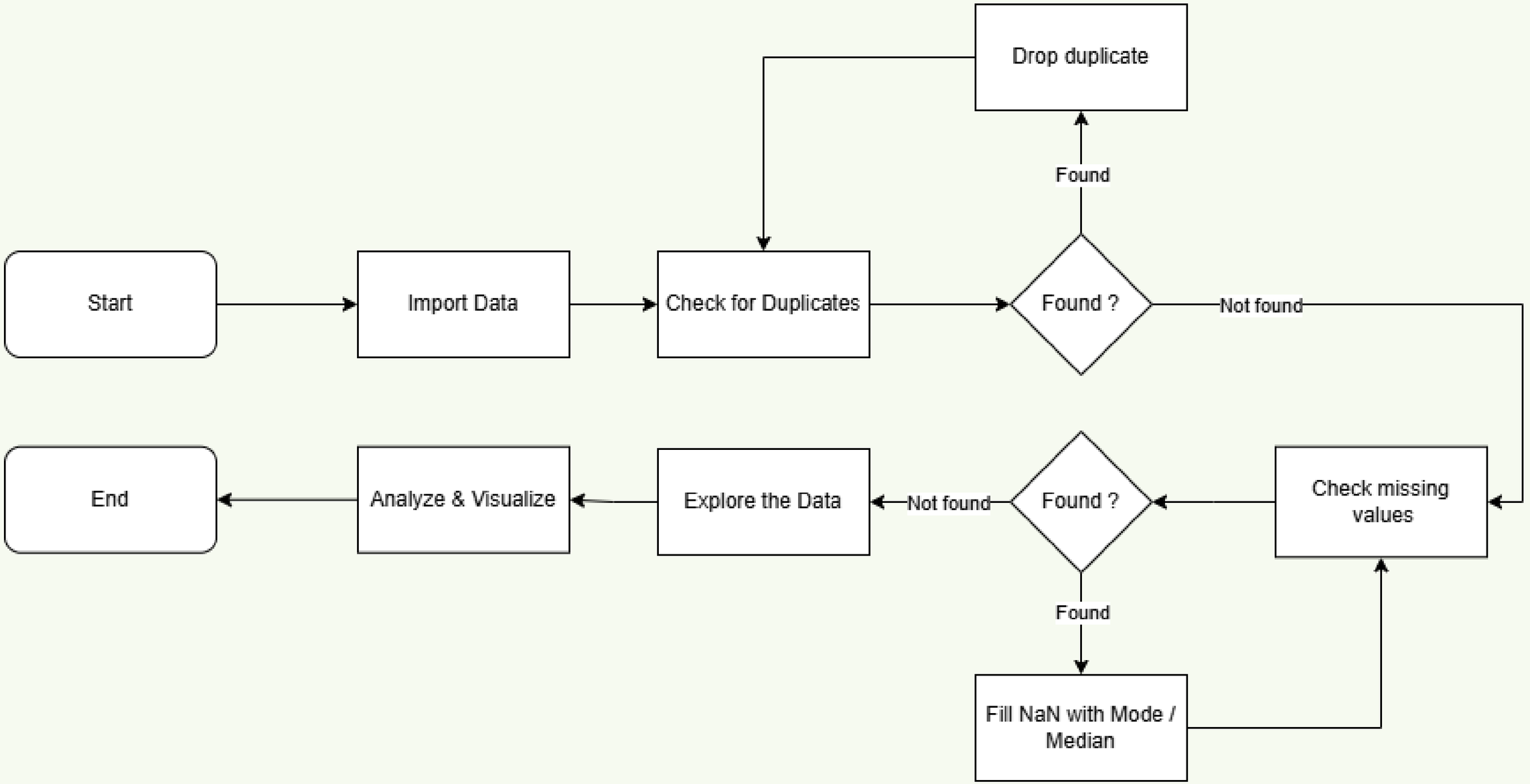
  - To understand the structure of the data
  - To identify the types of columns (numeric, categorical, or text)
  - To get a quick view of the typical values in each column

COMPLETION STEPS

- ## DUPLICATE DETECTION AND HANDLING
  - Counting duplicate data
  - Displays and sorts duplicate rows by frequency.
  - Duplicates are removed to ensure accurate analysis.

- ## MISSING VALUE DETECTION AND HANDLING
  - Check for missing values in each column.
  - Displays the number and percentage of missing data.
  - Object columns are filled with the mode (the most frequently occurring value).
  - Numeric columns are filled with the median (to avoid bias from outliers).

FLOWCHART


Drop duplicate
Found
Start
Import Data
Check for Duplicates
Found ?
Not found
End
Analyze & Visualize
Explore the Data
Not found
Found ?
Check missing values
Found
Fill NaN with Mode / Median

CONCLUSSION

- The Titanic dataset initially contained duplicate records and missing values, which were successfully cleaned.
- Exploratory Data Analysis (EDA) revealed that survival was influenced by gender and age group.
- Data cleaning is a crucial step before performing further modeling or predictions.