Exploratory Data Analysis (EDA) Report - Train Dataset

Introduction

This report explores the Train(Titanic) dataset, which contains information about the passengers Titanic. The dataset includes various numerical attributes such as Fare, Age etc. By analyzing these features, we aim to better understand the relationships between different components. The analysis includes data cleaning, handling missing values, detecting outliers, and performing statistical visualizations to uncover trends and patterns.

Data Cleaning

Loading the Dataset

- We started by loading the dataset into a Pandas DataFrame.
- The first few rows were examined to understand its structure.
- We checked the data types of each column to determine which variables were numerical and categorical.

Handling Missing Values

- The dataset was thoroughly checked for missing values.
- There are many missing values in the Age and fare column which are filled with mean and median to check how the distribution changes with different imputation techniques.

Removing Duplicate Records

- We looked for duplicate records to ensure data integrity.
- No duplicate entries were found, confirming the dataset was clean in this regard.

Detecting and Treating Outliers

- Box plots were used to identify outliers in key numerical features like age, fare.
- The Interquartile Range (IQR) method was applied to detect extreme values.

Exploratory Data Analysis (EDA)

Univariate Analysis

- Summary Statistics: We calculated measures such as mean, median, variance, and skewness to understand the central tendency and spread of numerical features.
- Frequency Distribution: The target variable Passenger class was analysed to determine the ration of passengers travelling in different class, Embarked column was analysed to determine how many passengers embarked from which city.
- Histogram: The columns of Fare and Age was visualized to see the distribution.
- Box Plot: Used to detect potential outliers in the age and fare feature.

Bivariate Analysis

- Correlation Matrix: A heatmap was used to identify relationships between numerical variables and highlight strong correlations.
- Scatter Plot: A scatter plot of age vs. fare showed how these two properties are related.
- Box Plot: Distribution of age was analysed based on sex of passenger and whether he/she survived or not.
- Violin Plot: Distribution of age was analysed based on sex of passenger and whether he/she survived or not.

Multivariate Analysis

- Pair Plot: Multiple variables were analysed simultaneously to spot trends and interactions.
- Heatmap: The correlations between all numerical variables were visualized to identify patterns.
- Bar Chart: The mean values of different features were displayed to get an overall sense of the dataset's composition.