Titanic Dataset - Exploratory Data Analysis (EDA) Report

---

1. Introduction

The purpose of this Exploratory Data Analysis (EDA) is to understand the Titanic dataset's structure, detect patterns, find anomalies, and extract useful insights.
We used **Python**, along with **Pandas**, **Matplotlib**, and **Seaborn** for data exploration and visualization.

---

2. Dataset Overview

- **Dataset:** Titanic - Machine Learning from Disaster
- **Source:** Kaggle Titanic Dataset
- **Rows:** 891
- **Columns:** 12
- **Target Variable:** Survived (0 = No, 1 = Yes)

---

3. Data Information

- Used .info(), .describe(), .isnull() .sum() to inspect data.

Key observations:

- Columns Age, Cabin, and Embarked have missing values.
- Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked are important for analysis.
- Name, Ticket, Cabin are textual and may need preprocessing if used later.

---

4. Univariate Analysis (Single Column)

4.1 Age Distribution

python

Copyedit

df['Age'].hist(bins=30)

Observation:

- Most passengers are between 20-40 years old.
- Age distribution is right-skewed.

- Some missing values.

4.2 Survival Count

python

CopyEdit

sns.countplot(x='Survived', data=df)

Observation:

- More passengers did not survive than those who survived.

4.3 Passenger Class Distribution

python

CopyEdit

sns.countplot(x='Pclass', data=df)

Observation:

- 3rd Class passengers are the majority.

---

5. Bivariate Analysis (Two Columns)

5.1 Survival by Gender

python

CopyEdit

sns.countplot(x='Survived', hue='Sex', data=df)

Observation:

- Females had a higher survival rate compared to males.

5.2 Survival by Passenger Class

python

CopyEdit

sns.countplot(x='Survived', hue='Pclass', data=df)

Observation:

- 1st Class passengers had higher survival rates.

5.3 Boxplot of Age vs Pclass

python

CopyEdit

sns.boxplot(x='Pclass', y='Age', data=df)

Observation:

- 1st class passengers were older on average than 3rd class passengers.

---

6. Multivariate Analysis (More than Two Columns)

6.1 Correlation Heatmap

python

CopyEdit

numeric_df = df.select_dtypes(include=['int64', 'float64'])

sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')

Observation:

- Fare and Pclass are negatively correlated.
- Survival is moderately correlated with Fare and Pclass.

6.2 Pairplot

python

CopyEdit

sns.pairplot(df[['Survived', 'Age', 'Fare', 'Pclass']])

Observation:

- Higher fare passengers had better survival rates.
- Pclass and Fare are visibly related.

---

7. Key Findings and Summary

- Females and 1st class passengers had a much higher chance of survival.

- Passengers paying higher fares (1st class) had better survival rates.
- Age has a slight impact; young children and young adults survived better.
- There are missing values in Age, Cabin, and Embarked to be handled.
- Correlation between features is mostly low except between Pclass and Fare.