📜 TASK -5 : Exploratory Data Analysis (EDA) Report

Author: *PARAPATLA SAI KUMAR*

Date: *29 September 2025*

Dataset: Titanic Dataset (titanic.csv)

Tools Used: Python, Pandas, Matplotlib, Seaborn,

📌 Objective

To explore the Titanic dataset using statistical and visual techniques to uncover patterns, trends, and anomalies that may influence passenger survival.

📂 Dataset Overview :

The Titanic dataset contains information about passengers aboard the RMS Titanic. Key features include:

- **Survived**: Target variable (0 = No, 1 = Yes)
- **Pclass**: Passenger class (1st, 2nd, 3rd)
- **Sex**: Gender
- **Age**: Age in years
- **SibSp**: Number of siblings/spouses aboard
- **Parch**: Number of parents/children aboard
- **Fare**: Ticket fare
- **Embarked**: Port of embarkation
- **Cabin**: Cabin number (many missing values)

🧹 Data Cleaning :

Python code :

```
df['Age'].fillna(df['Age'].median(), inplace=True)

df.drop(columns=['Cabin'], inplace=True)
```

- Age imputed with median.

- Cabin dropped due to excessive missing values.

🔍 **Initial Exploration :**

Data Summary

Python code :

df.info()

df.describe()

df.isnull().sum()

- **Missing Values**: Age (~20%), Cabin (~77%), Embarked (2 entries)
- **Data Types**: Mix of categorical and numerical
- **Target Distribution**: 38% survived, 62% did not

📊 **Univariate Analysis :**

Age Distribution

Python code :

sns.histplot(df['Age'].dropna(), kde=True)

- Most passengers are aged between 20–40.
- KDE curve shows a slight right skew.

Fare Boxplot :

Python code :

sns.boxplot(x='Fare', data=df)

- Median fare is around $15.
- Significant outliers above $200, likely first-class passengers.

🔗 **Bivariate Analysis :**

Survival by Gender :

Python code :

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

- Females had a much higher survival rate than males.

Survival by Class :

Python code :

```
sns.barplot(x='Pclass', y='Survived', data=df)
```

- First-class passengers had the highest survival rate.
- Third-class had the lowest.

Age vs Fare (Survival) :

Python code :

```
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
```

- High-fare passengers tended to survive more.
- Younger passengers show mixed survival outcomes.

🔥 **Correlation Analysis :**

Heatmap:

Python code :

```
sns.heatmap(df.corr(), annot=True)
```

- **Survived** correlates positively with **Fare** and negatively with **Pclass**.
- **Age** has weak correlation with survival.

Pairplot :

Python code :

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

- Clear separation in survival based on class and fare.

📌 **Summary of Findings :**

- **Gender**: Females had a significantly higher survival rate.
- **Class**: First-class passengers were more likely to survive.
- **Fare**: Higher fare correlated with survival.

- **Age**: No strong correlation, but children had slightly better survival.
- **Missing Data**: Cabin feature dropped; Age imputed.

✅ Conclusion

The EDA reveals that **gender**, **class**, and **fare** are strong indicators of survival. These insights can guide feature selection for predictive modeling and highlight social dynamics aboard the Titanic.