Exploratory Data Analysis on Titanic Dataset

Objective:

To analyze the Titanic dataset using Exploratory Data Analysis (EDA) techniques and extract insights through visual and statistical exploration.

Tools Used:

- Python
- Pandas
- Seaborn
- Matplotlib
- Jupyter Notebook

1. Dataset Overview:

The Titanic dataset includes details of passengers such as age, gender, ticket fare, class, and survival status. The goal is to identify trends, patterns, and relationships among the variables.

2. Data Summary:

- Total records: 891
- Key features: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
- Missing Values:
  - Age: ~19.8% missing
  - Cabin: ~77% missing
  - Embarked: 2 missing values

3. Univariate Analysis:

- **Survival Count:** More people died than survived.
- **Gender Distribution:** Males > Females
- **Age Distribution:** Right-skewed with a peak between 20–30 years

4. Bivariate Analysis:

- **Survival by Gender:** Females had higher survival rates than males
- **Survival by Class:** 1st class passengers had a better survival rate
- **Boxplot of Age vs Class:** Younger passengers are distributed across all classes, but 1st class passengers tend to be slightly older

5. Correlation Analysis:

- Fare and Pclass show negative correlation
- Survival is positively correlated with Fare and negatively with Pclass

6. Visualizations Used:

- Countplots (Survived, Gender, Class)
- Histogram (Age)
- Boxplot (Age vs Pclass)
- Heatmap (Correlation matrix)
- Pairplot (Survived, Age, Fare, Pclass)

---

7. Key Insights:

- Females and passengers in 1st class had better survival chances
- Age and Fare are skewed; consider transformation/imputation
- Cabin column has too many missing values and may not be reliable
- Pclass and Fare are important features in survival prediction

8. Conclusion:

This EDA helped uncover important patterns in survival across gender and class. The insights can be used for feature selection in future ML models.

9. Next Steps (Optional for Future Work):

- Impute missing age using median by Pclass/Sex
- Drop or engineer Cabin feature
- Train a classification model on cleaned dataset