# Exploratory Data Analysis (EDA) Report

1. **Introduction:**

   The purpose of this Exploratory Data Analysis (EDA) is to understand the structure, quality, and patterns within the **Titanic Dataset**. The dataset contains passenger information such as demographic details, travel class, fares, and survival status. Using Python (Pandas, Matplotlib, Seaborn), various statistical summaries and visualizations were generated to uncover relationships, identify anomalies, and prepare the data for further insights or modeling.

**2. Dataset Overview**

- **Dataset Name:** Titanic Dataset
- **Total Rows:** ~500
- **Total Columns:** 12
- **Key Columns:** passenger_id, survived, p_class, name, sex, age, sib_sp, par_ch, ticket, fare, cabin, embarked

**3. Initial Data Inspection**

The dataset was first explored using:

- df.info() to check structure, datatypes, and null values
- Summary statistics with describe() for numerical columns
- value_counts() for categorical variables

This provided a broad overview of distributions, unique values, and data anomalies.

## Result:

```
# Check Overall Structure
df.info()
```

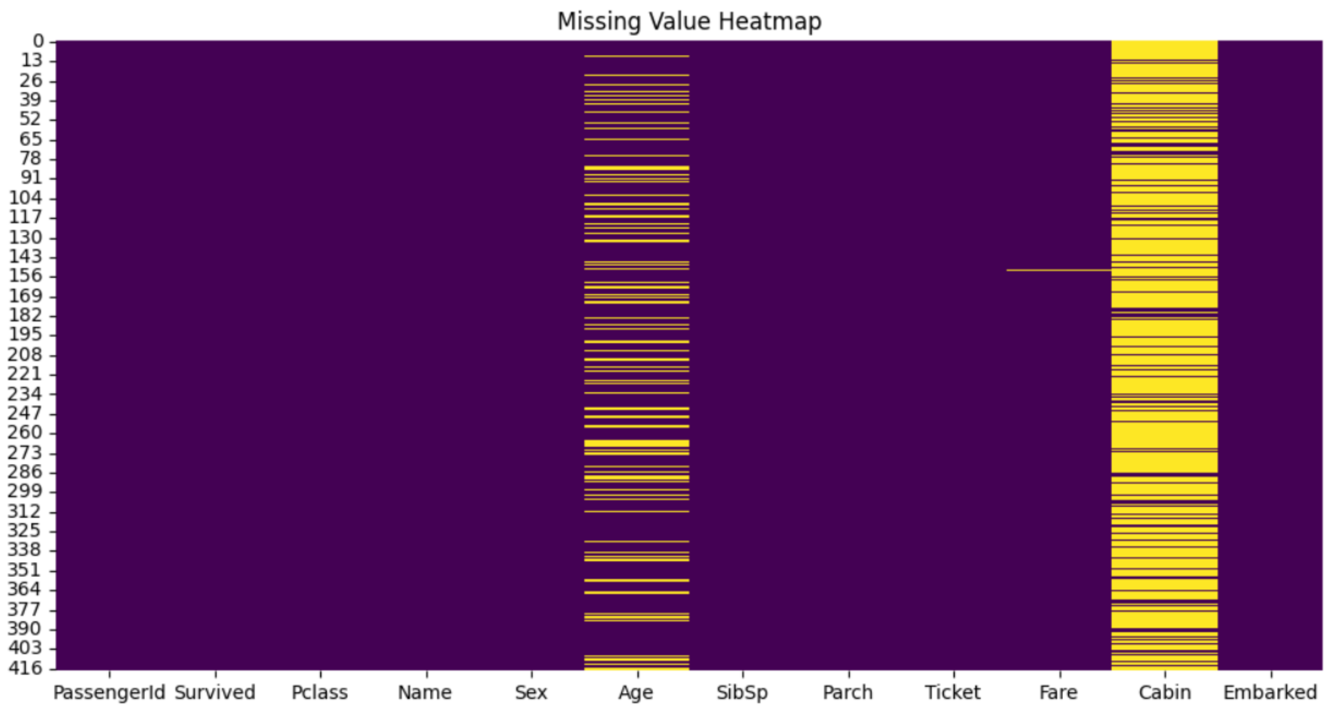
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Survived     418 non-null    int64
 2   Pclass       418 non-null    int64
 3   Name         418 non-null    object
 4   Sex          418 non-null    object
 5   Age          332 non-null    float64
 6   SibSp        418 non-null    int64
 7   Parch        418 non-null    int64
 8   Ticket       418 non-null    object
 9   Fare         417 non-null    float64
 10  Cabin        91 non-null     object
 11  Embarked     418 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 39.3+ KB
```

# 4. Missing Value Analysis

```
# Count Missing Values in Each Column
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

A heatmap was plotted to visually detect missing values.




## 5. Data Cleaning Steps

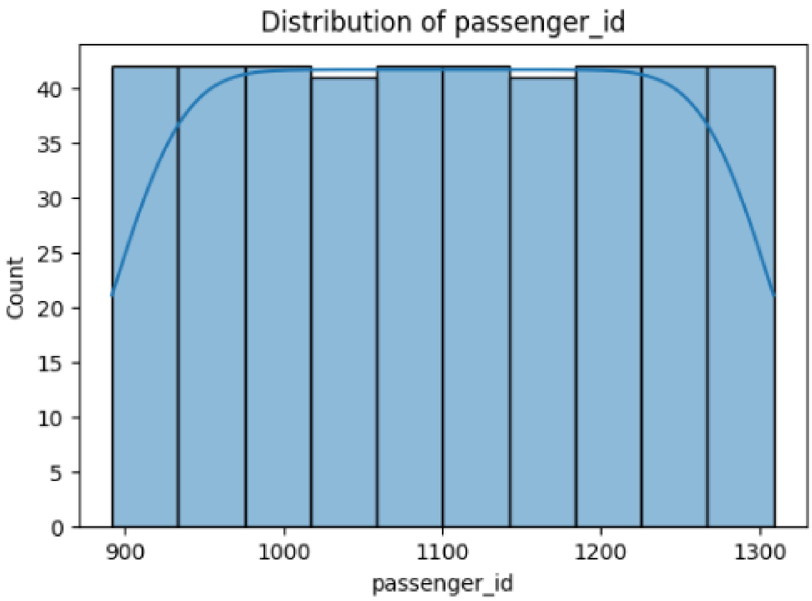
- Renamed and standardized column names
- Converted **Age** from float to int
- Filled missing values in **Age** and **Fare**
- Dropped **Cabin** column completely
- Checked for duplicate or inconsistent values
- Ensured clean categorical labels (Sex, Embarked)

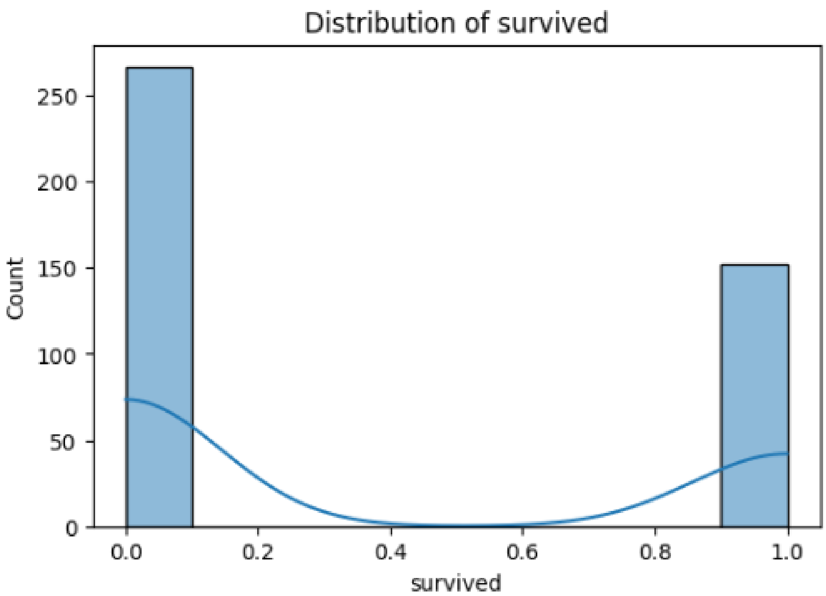
These steps improved dataset reliability for analysis.

## 6. Univariate Analysis

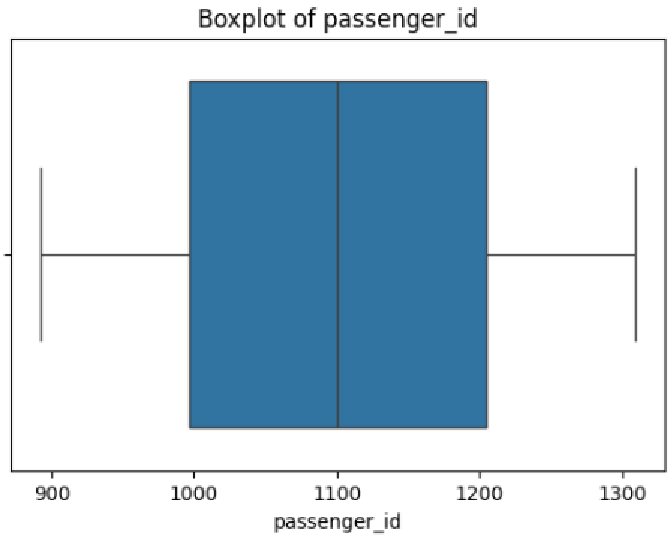
Visualizations used:

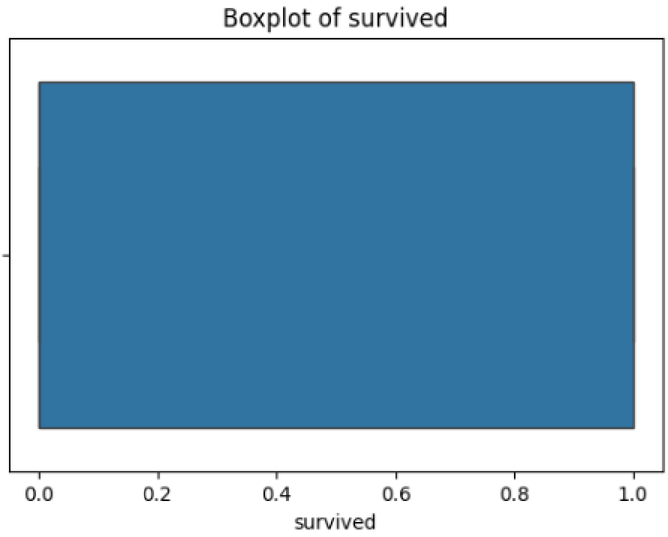
- **Histplot:** To examine numerical distributions (Age, Fare).
  - Age distribution showed a right-skew (younger passengers more common)
  - Fare distribution had extreme values.




- **Boxplot:** Identified clear outliers in Fare and some variation in Age.

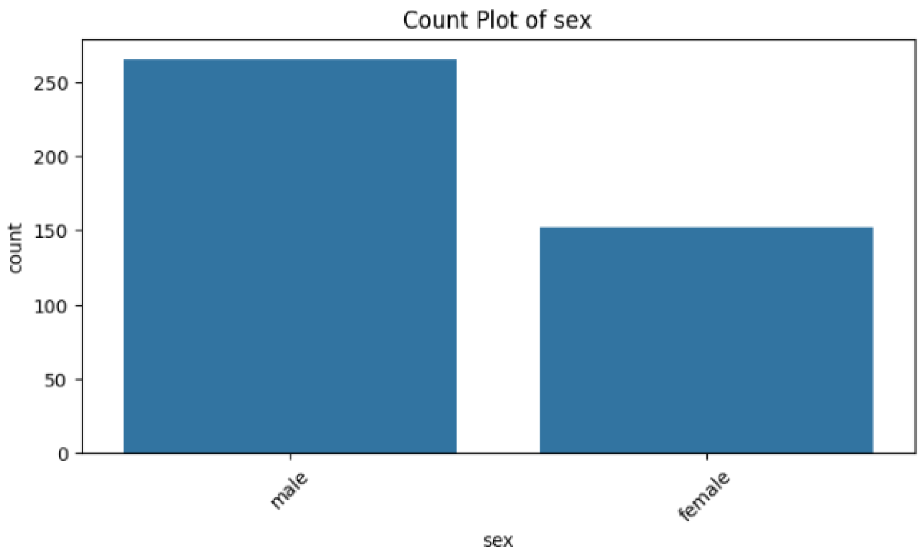


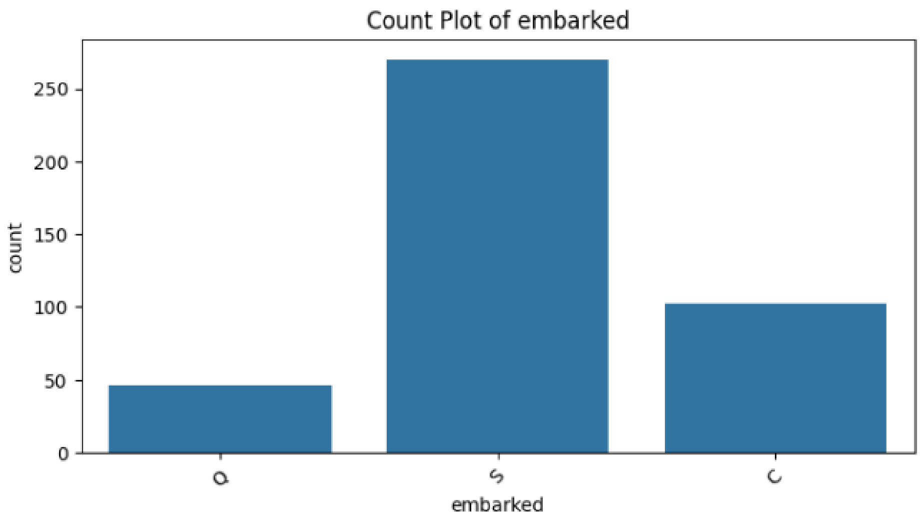

## Countplot:

- More males than females
- Most passengers belonged to **p_class 3**
- Embarked distribution showed highest boarding at Southampton

```
Skipping 'name' (too many unique values: 418)
```




```
Skipping 'ticket' (too many unique values: 363)
```




## 7. Multivariate Analysis

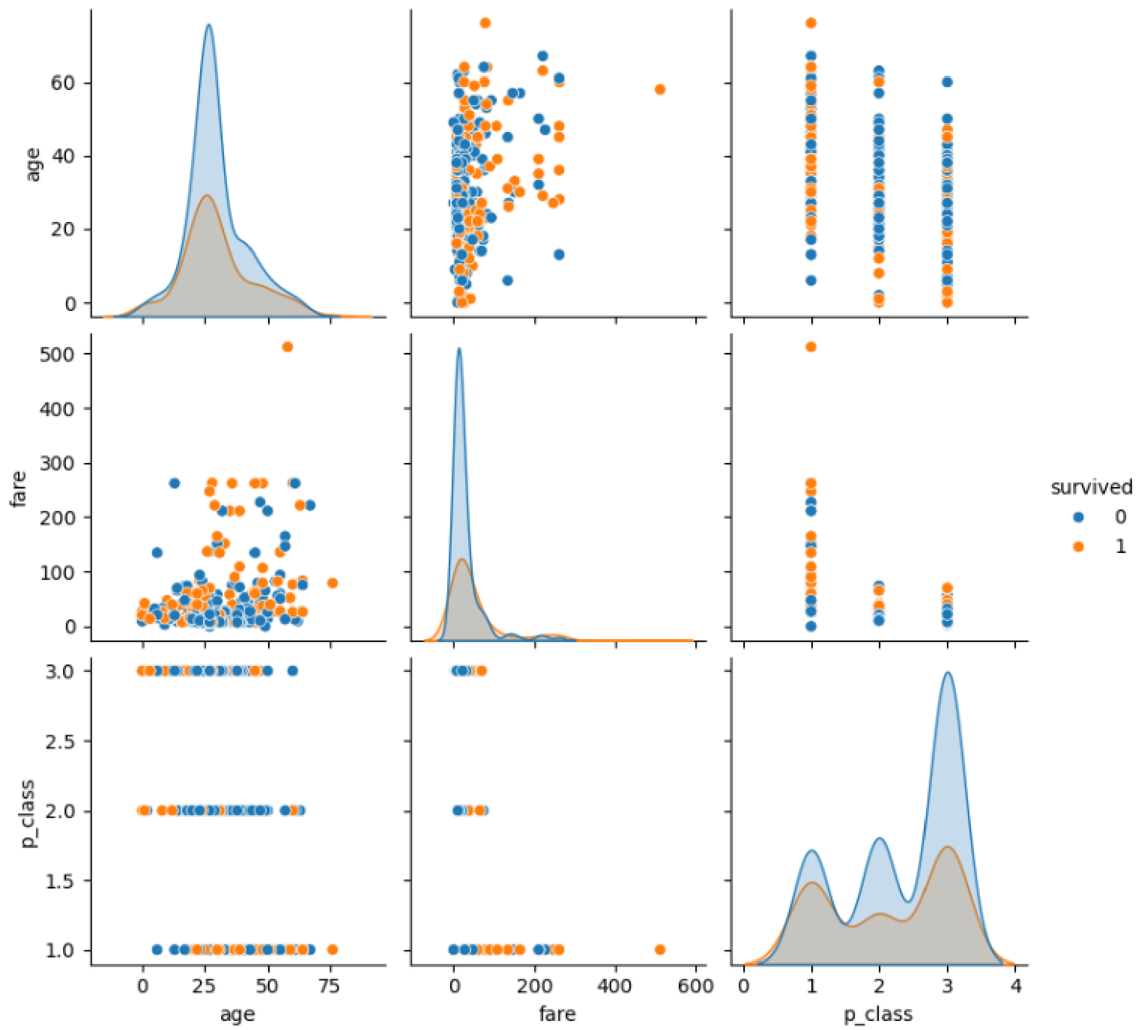
### Pairplot

A pairplot was used to observe interactions between:

- age
- fare
- p_class
- survived

Clear differences appeared:

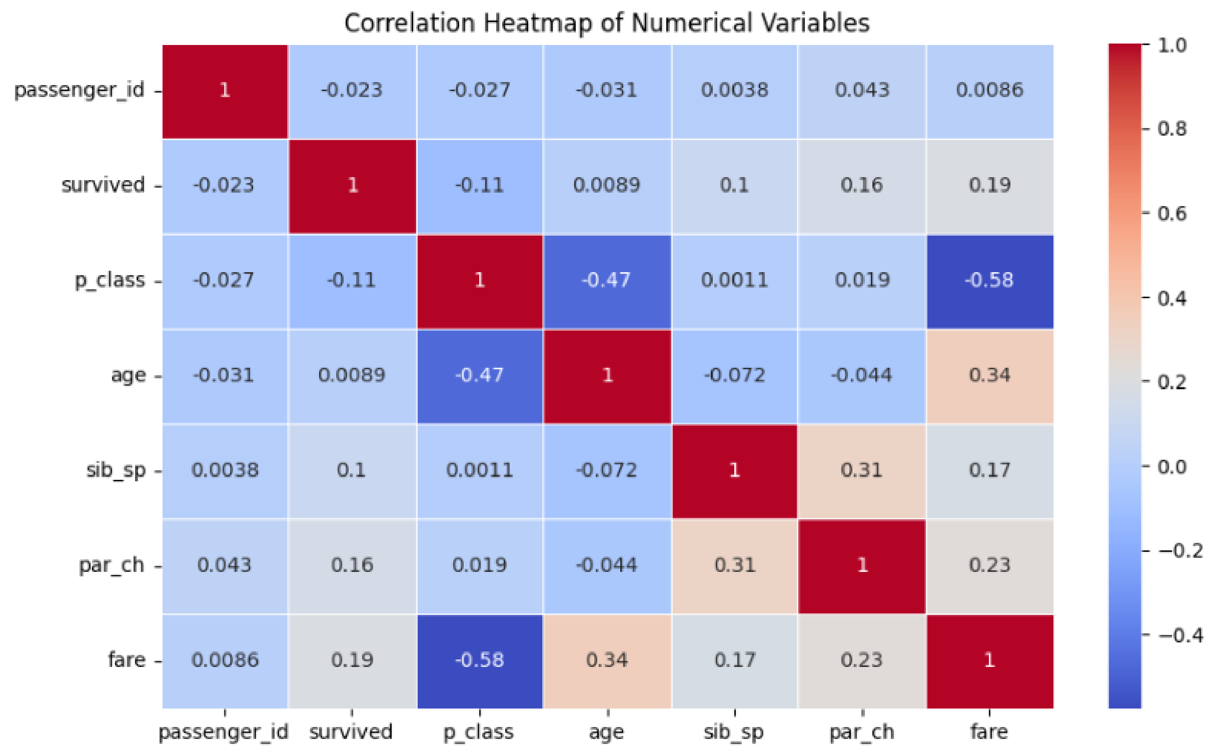
- Higher survival tendency among **females**
- Passengers in **p_class 1** tended to be older and paid higher fares
- Younger passengers mostly in p_class 3




## 8. Correlation Analysis

A correlation heatmap revealed:

- **survived** has a positive correlation with **fare** and **p_class (negative)**
  → Meaning higher-class passengers were more likely to survive
- Strong negative correlation between **p_class** and **fare**
- Age had weak correlation with survival




## 9. Groupby Analysis

Category-level patterns were observed using:

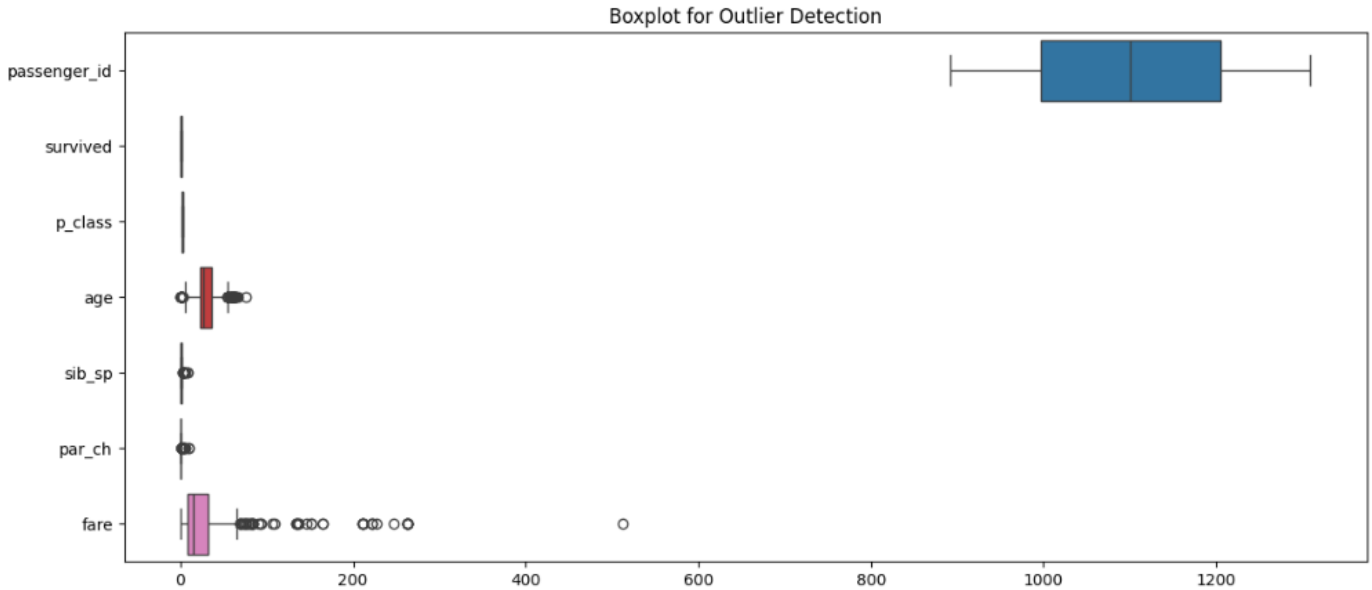
- df.groupby('sex')['survived'].mean()
  → Females had a significantly higher survival rate
- df.groupby('p_class')['fare'].mean()
  → Fare increases sharply with higher class
- df.groupby('Embarked')['survived'].mean()
  → Passengers from Cherbourg had the highest survival rate

Groupby analysis clearly highlighted survival patterns across categories.

## 10. Outlier and Skewness Analysis

Outliers were analyzed using:

- **Boxplots**
- **IQR method**
  → Fare column had multiple extreme outliers
  → Age had a more normal distribution with fewer outliers




## 11. Key Insights

- Females and first-class passengers had much higher survival rates.
- Fare displays high variation and strong correlation with passenger class.
- Age had missing values but after filling, distribution looked reasonable.
- Cabin column had excessive missingness and was appropriately dropped.
- Embarked location influenced survival—especially for passengers from Cherbourg.
- Outliers exist primarily in the Fare feature and may require transformation in modeling.

## 12. Conclusion

The EDA effectively cleaned the Titanic dataset, revealed missing value patterns, and highlighted important relationships between variables. The analysis shows clear survival patterns linked to demographic and travel-related factors. These findings provide a solid foundation for predictive modeling or deeper exploratory insights.