# Exploratory Data Analysis (EDA) on Titanic Dataset

## 1. Introduction

Exploratory Data Analysis (EDA) is a fundamental step in data science used to understand data structure, quality, and hidden patterns. In this project, EDA was performed on the Titanic dataset to analyze factors that influenced passenger survival during the disaster.

## 2. Objective of the Project

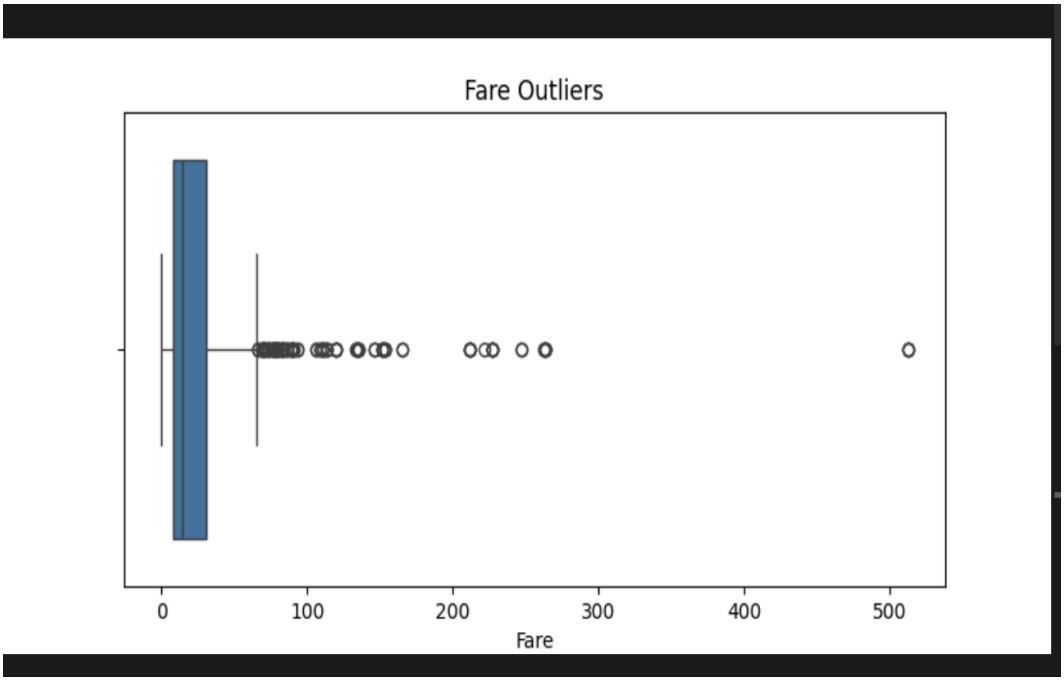
The objectives include understanding the dataset, identifying missing values, analyzing survival trends, detecting outliers, and discovering relationships between variables using visual and statistical methods.

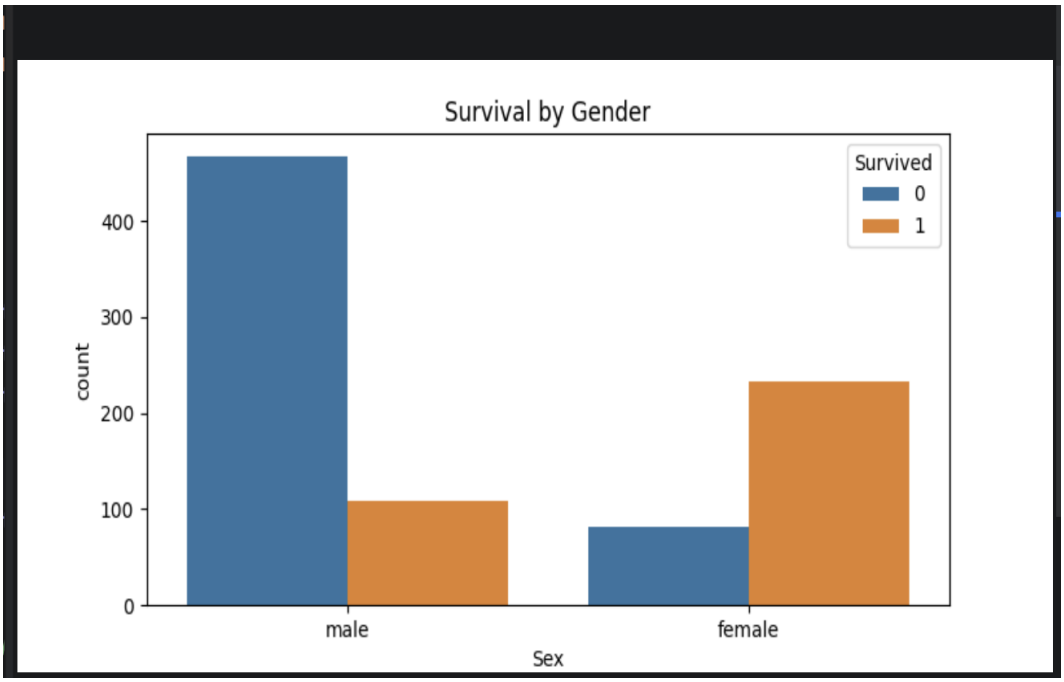
## 3. Dataset Overview

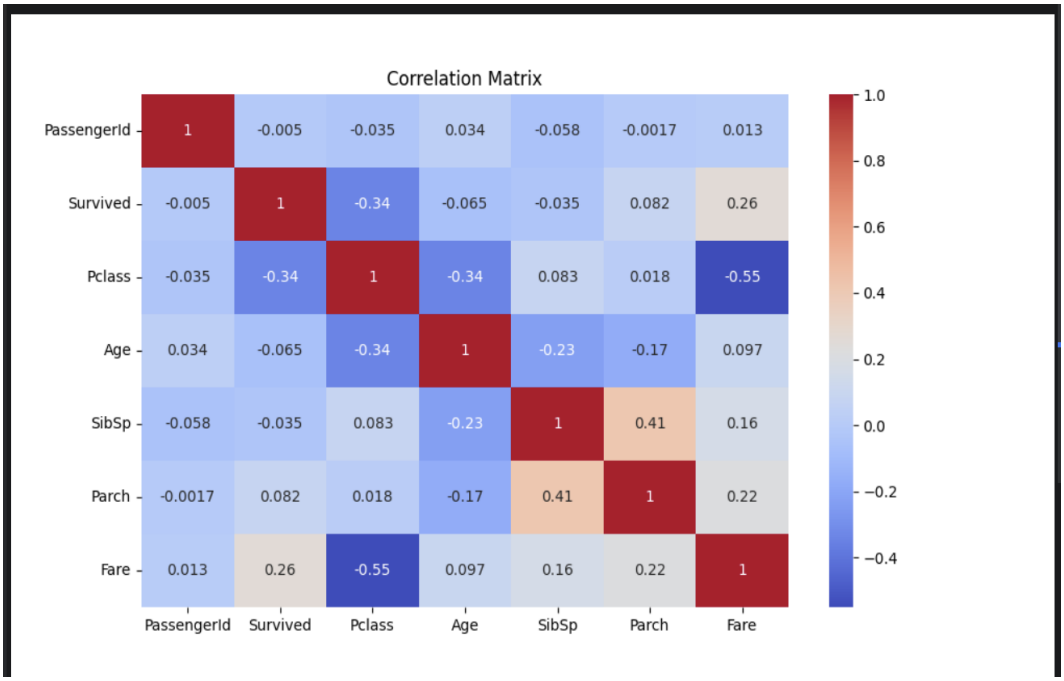
The dataset contains 891 records and 12 attributes including passenger class, gender, age, fare, family size, and survival status. It includes both numerical and categorical data.

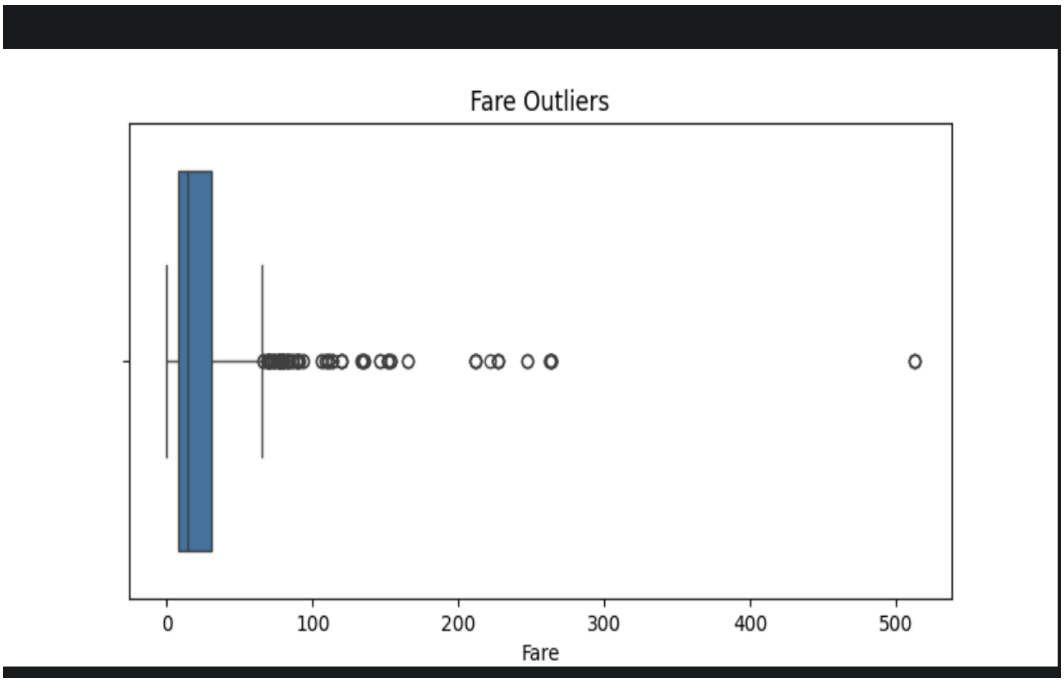
## 4. Methodology

Steps followed: data loading, missing value analysis, statistical summary, data imputation, visual exploration using plots, correlation analysis, and outlier detection using boxplots and Z-scores.


Fare Outliers
0
100
200
300
400
500
Fare


Survival by Gender
Survived
0
1
400
300
200
100
0
count
male
female
Sex


Correlation Matrix
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1 -0.005 -0.035 0.034 -0.058 -0.0017 0.013
-0.005 1 -0.34 -0.065 -0.035 0.082 0.26
-0.035 -0.34 1 -0.34 0.083 0.018 -0.55
0.034 -0.065 -0.34 1 -0.23 -0.17 0.097
-0.058 -0.035 0.083 -0.23 1 0.41 0.16
-0.0017 0.082 0.018 -0.17 0.41 1 0.22
0.013 0.26 -0.55 0.097 0.16 0.22 1
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Fare Outliers
0
100
200
300
400
500
Fare

```
Data Shape: (891, 12)

Data Types:
 PassengerId         int64
Survived             int64
Pclass               int64
Name                object
Sex                 object
Age                float64
SibSp                int64
Parch                int64
Ticket              object
Fare               float64
Cabin               object
Embarked            object
dtype: object

Missing Values:
 PassengerId        0
Survived            0
Pclass              0
Name                0
Sex                 0
Age               177
SibSp               0
Parch               0
Ticket              0
Fare                0
Cabin             687
Embarked
```

```
 PassengerId      0
Survived          0
Pclass            0
Name              0
Sex               0
Age             177
SibSp             0
Parch             0
Ticket            0
Fare              0
Cabin           687
Embarked          2
dtype: int64

Summary Stats:
        PassengerId    Survived      Pclass  ...        Fare    Cabin  Embarked
count    891.000000  891.000000  891.000000  ...  891.000000      204       889
unique          NaN         NaN         NaN  ...         NaN      147         3
top             NaN         NaN         NaN  ...         NaN  B96 B98         S
freq            NaN         NaN         NaN  ...         NaN        4       644
mean     446.000000    0.383838    2.308642  ...   32.204208      NaN       NaN
std      257.353842    0.486592    0.836071  ...   49.693429      NaN       NaN
min        1.000000    0.000000    1.000000  ...    0.000000      NaN       NaN
25%      223.500000    0.000000    2.000000  ...    7.910400      NaN       NaN
50%      446.000000    0.000000    3.000000  ...   14.454200      NaN       NaN
75%      668.500000    1.000000    3.000000  ...   31.000000      NaN       NaN
max      891.000000    1.000000    3.000000  ...  512.329200      NaN       NaN

[11 rows x 12 columns]
```

```
Survived           0          1
Pclass
1          37.037037  62.962963
2          52.717391  47.282609
3          75.763747  24.236253
Found 20 fare outliers

Process finished with exit code 0
```

## 5. Key Findings

• Females had a much higher survival rate than males.
• First-class passengers survived more compared to second and third class.
• Higher fare showed a positive correlation with survival.
• Several fare outliers were identified but retained as valid data.

## 6. Applications

This analysis can be applied to risk modeling, emergency planning, business decision-making, and as a foundation for predictive machine learning models.

## 7. Conclusion

The Titanic EDA project demonstrates the importance of data exploration before modeling. Through visualization and statistical analysis, meaningful insights were derived, making the dataset ready for further predictive analysis.