Exploratory Data Analysis (EDA) on the Titanic Dataset

1.Introduction and Objectives

This report details the Exploratory Data Analysis (EDA) performed on the Titanic training dataset (‘train.csv’) using the Python libraries Pandas, Matplotlib, and Seaborn.

The primary objective was to extract meaningful **insights and patterns** regarding passenger survival, identifying the key variables that influenced whether a passenger lived or perished.

2. Statistical Data Summary

- **Overall Survival Rate:** The mean of the Survived column is **0.384**. This means only **38.4%** of the passengers in the dataset survived the disaster.

- **Age:** The **median age (28.0)** and the **mean age (29.7)** were very close, suggesting a relatively symmetrical age distribution.

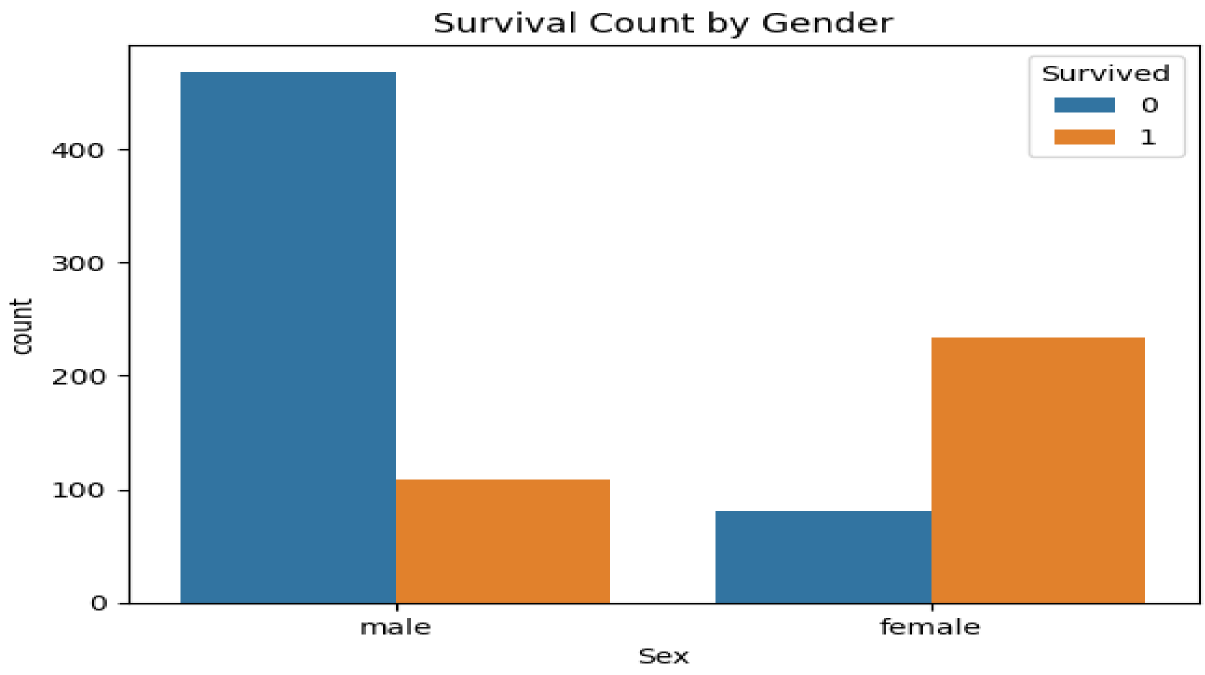
- **Fare:** The **mean fare ($$$32.20)** is significantly higher than the **median ($$$14.45)**. This indicates that the Fare data is highly **right-skewed** due to a small number of extremely expensive tickets (up to $$$512.33).

- **Missing Data:** Missing values in the Age column (177 entries) were imputed using the median, and the few missing values in Embarked were filled with the mode. Identifier columns (Name, Ticket) and the highly missing Cabin column were dropped.

3. Key Findings from Visual Analysis

The visual analysis of the relationship between features and the target variable (Survived) revealed clear patterns:

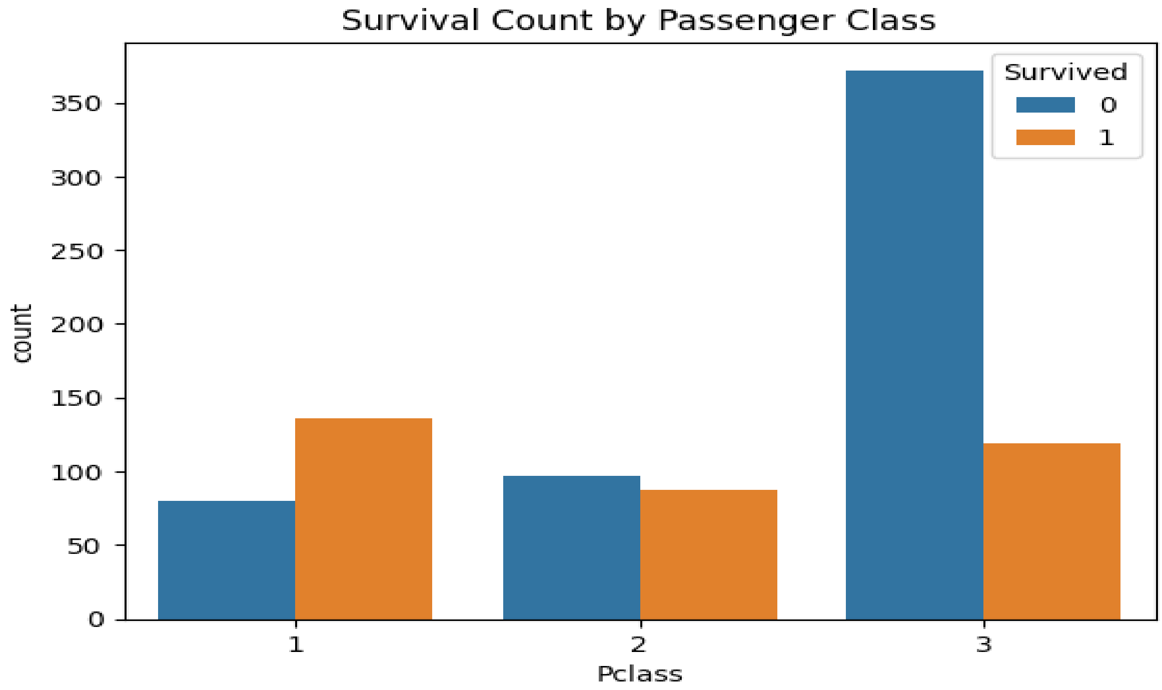
Finding A: Gender was the Dominant Factor



- Observation**:** The visual clearly shows that most female passengers survived, while most male passengers did not survive.

- Insight**:** Gender was the single most powerful factor determining survival, strongly suggesting adherence to the **"women and children first"** protocol.

Finding B: Socio-Economic Class was a Critical Factor



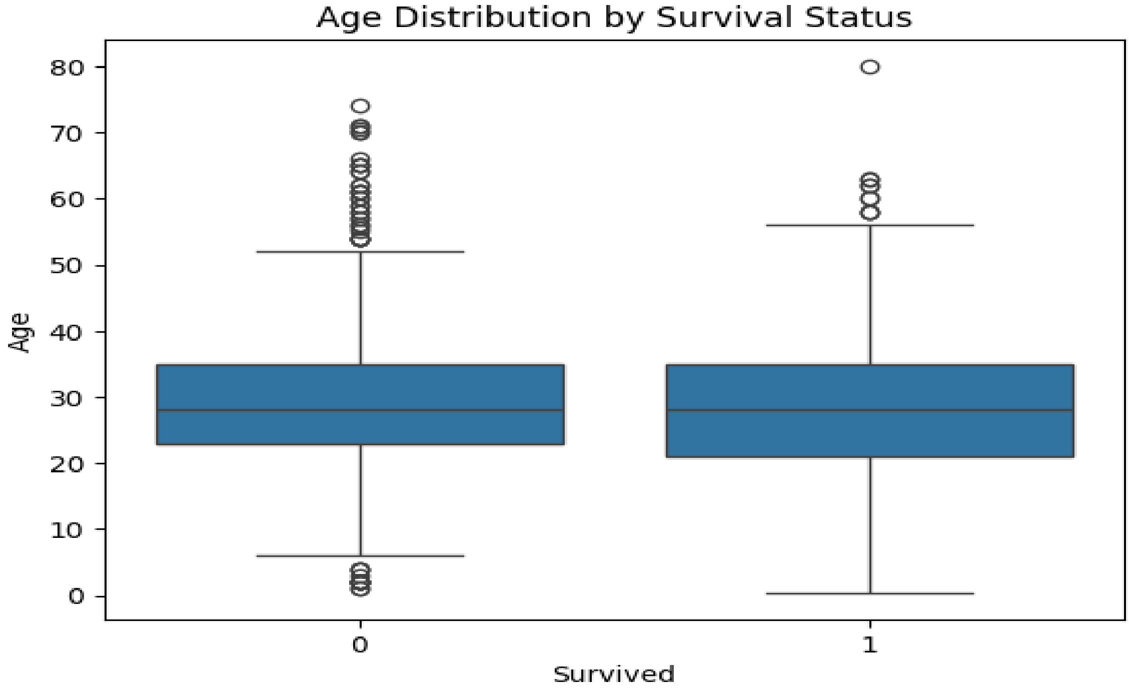
- Observation**:** There is a strong visual gradient in survival probability across the three classes.

  -First Class (Pclass=1): More passengers survived than perished.

  -Third Class (Pclass=3): The largest number of fatalities and had the lowest survival rate.

- Insight**: Wealth and social standing** (indicated by Pclass) provided a significant advantage in securing a lifeboat or accessing higher deck areas.

Finding C: Age was a Weak Predictor



- Observation**:** The median age and the central 50% distribution of ages are **nearly identical** for both the non-survivor (0) and survivor (1) groups.

- Insight**:** The overall age of a passenger, when viewed linearly, was not a strong determining factor in survival, although age-specific policies (like saving infants) may be present in the extremes of the distribution.

4. Correlation Analysis

The correlation matrix measures the linear relationship between numerical features and survival.

Feature	Correlation to Survived	Strength	Interpretation
Pclass	-0.338	Strong	As class number increases (goes from 1st to 3rd), the chance of survival decreases.
Fare	**+0.257**	Moderate	Higher fare is associated with a higher chance of survival.
Age	**-0.065**	Very Weak	Age has almost no linear relationship with survival probability.
Parch	**+0.082**	Very Weak	Having parents/children onboard has a negligible positive impact.

5. Conclusion

The Exploratory Data Analysis revealed that survival during the sinking of the Titanic was highly non- random and dependent on social factors.

The two dominant factors for survival were:

1. **Gender** (with females having a dramatically higher survival rate).
2. **Socio-Economic Class** (with First Class passengers having the highest survival rate).

Based on the evidence, the most likely profile for a survivor was a **female passenger traveling in First or Second Class**.