Intelligent Data Analysis - Lab4 Report

Exploratory Data Analysis (EDA) of Titanic Dataset

Student: Bekibaeva Aigerim 230121001 MATDAIS23
Professor: Mr. Mekuria

Lab Objective

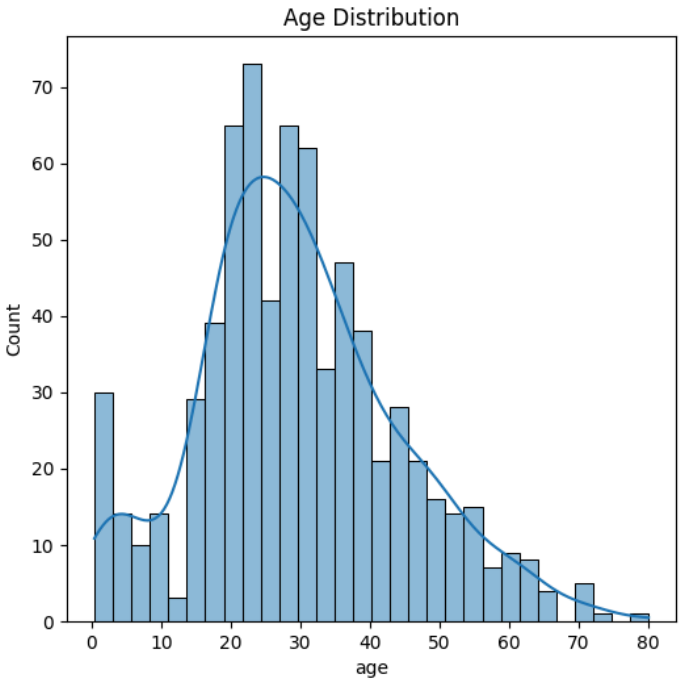
This report presents an exploratory data analysis (EDA) on the Titanic dataset to uncover patterns, relationships, and potential predictors of passenger survival. The dataset includes demographic and ticketing information for passengers on board of the Titanic, with a binary target variable '`survived`'.

---

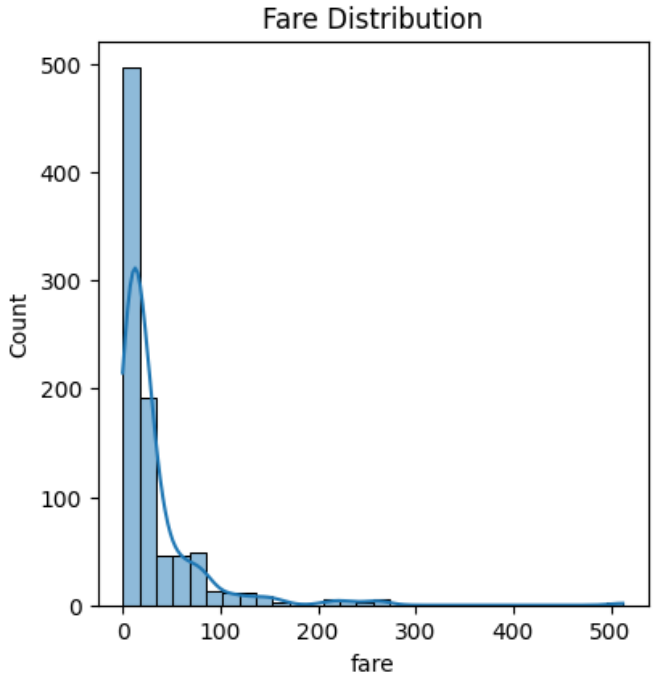
1. Univariate Analysis

We first explored the distribution of individual features to understand their ranges and frequencies:

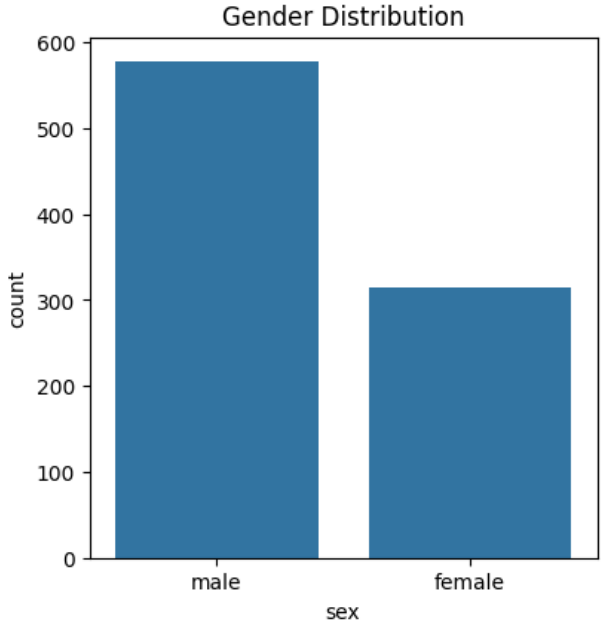
- **Age Distribution:** Most passengers were aged between 20 and 40 years. Some age data is missing (~20%), which suggests imputation will be necessary during preprocessing.



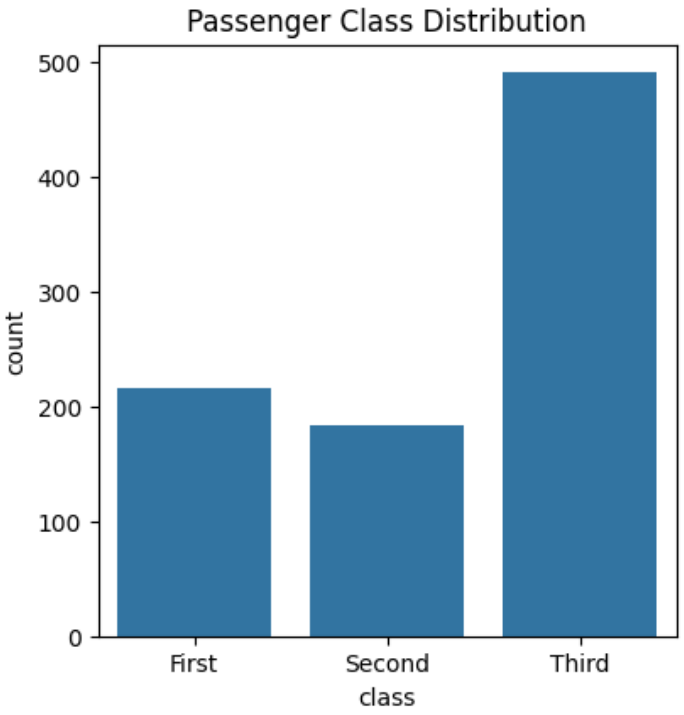
- **Fare Distribution:** The fare variable is right-skewed, with most passengers paying less than 50 units. A few high-paying passengers paid above 250, indicating outliers or premium ticket classes.



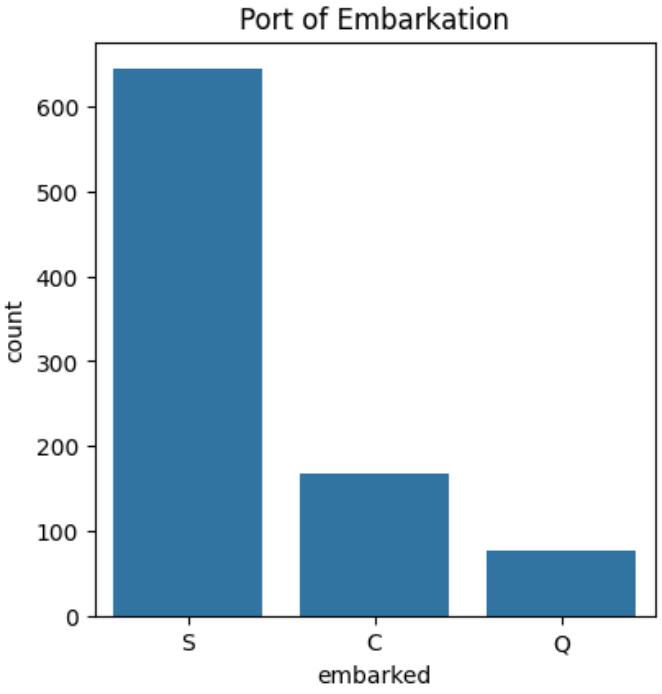
- **Sex Distribution:** About 65% of the passengers were male and 35% were female.



- **Class Distribution:** The majority of passengers traveled in third class, followed by first and second classes. This suggests socio-economic diversity.



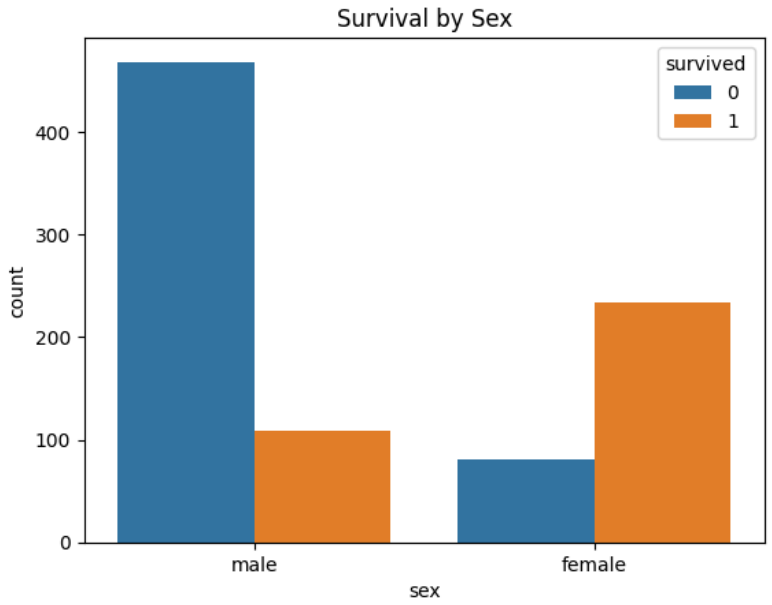
- **Embarked Port Distribution:** Most passengers boarded at **Southampton**, followed by Cherbourg and Queenstown.



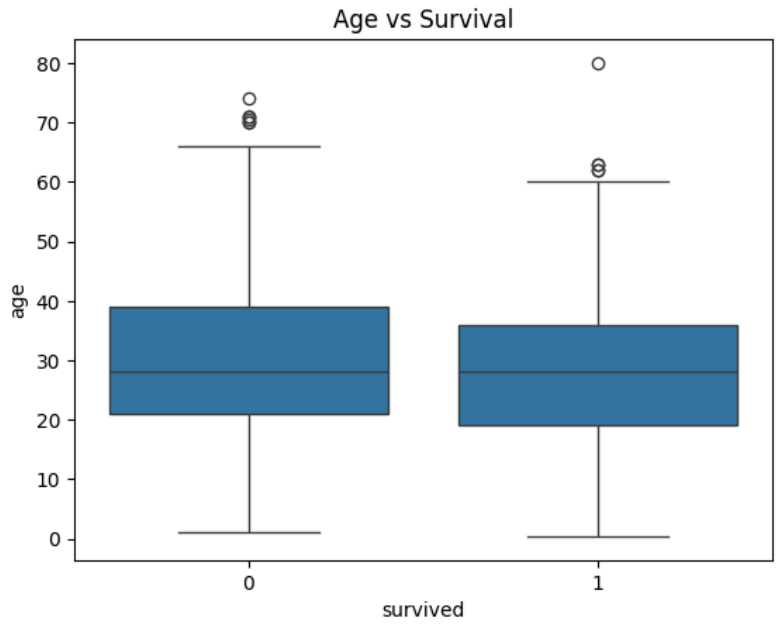
2. Bivariate Analysis

We examined relationships between features and the survival outcome:

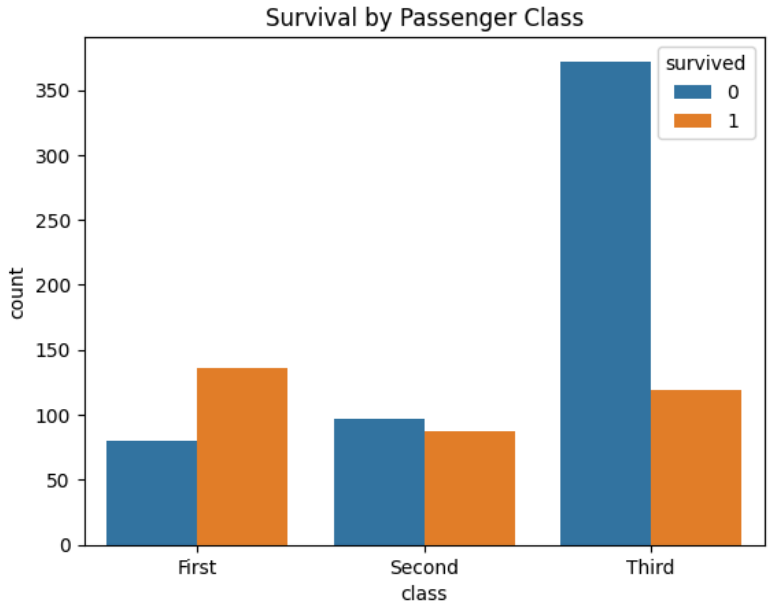
- **Survival by Sex:** A clear survival advantage for females is observed. Most survivors were women, while most non-survivors were men.



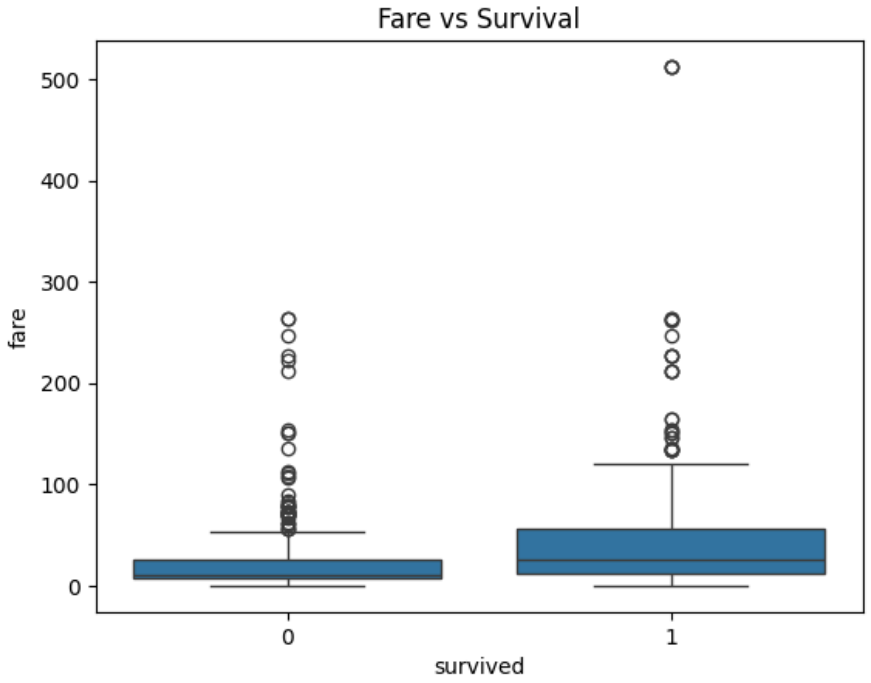
- **Age vs. Survival:** Younger passengers, particularly children, had higher survival rates. Older passengers (especially males) had a lower chance of survival.



- **Class vs. Survival:** First-class passengers had the highest survival rates. In contrast, third-class passengers had the lowest.

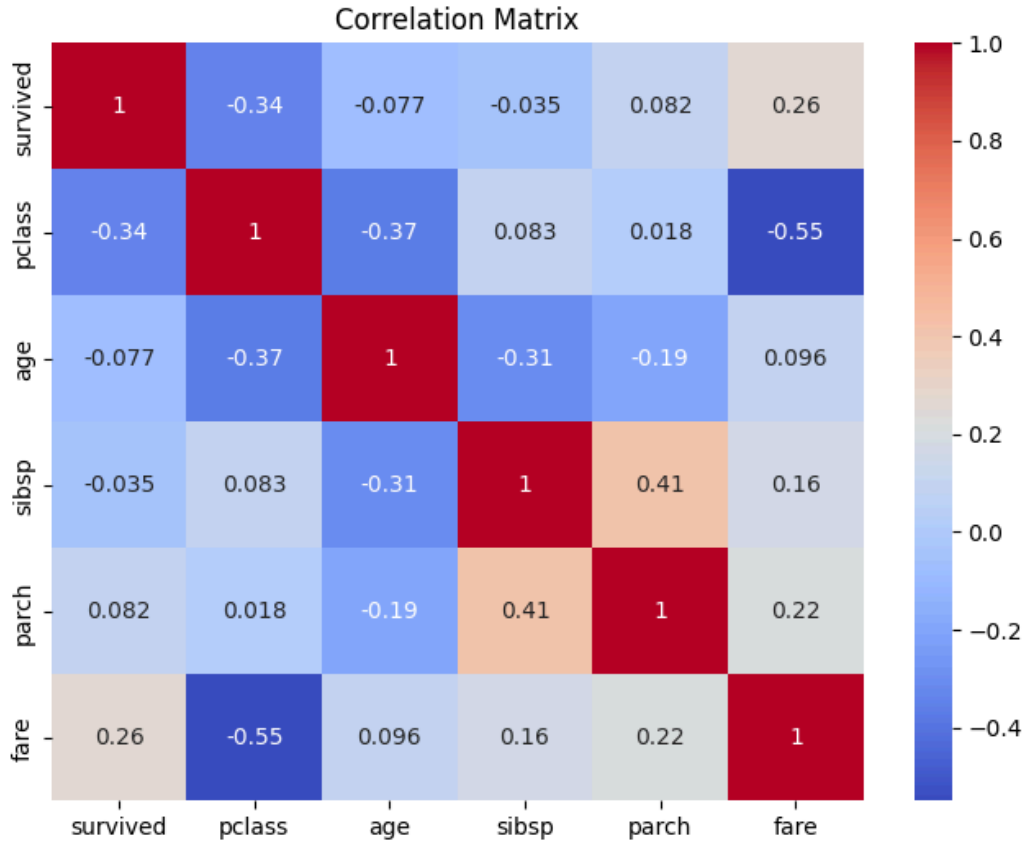


- **Fare vs. Survival:** Survivors generally paid higher fares, likely due to being in higher classes with better access to lifeboats.



3. Correlation Analysis

A correlation heatmap of numerical variables showed:



- **Fare and Pclass** were moderately correlated with survival (`Fare ~ 0.26`, `Pclass ~ -0.34`).
- **Age** showed a weak negative correlation with survival (`~ -0.08`), indicating older passengers were less likely to survive.
- **SibSp** and **Parch** (family-related features) showed weak but interesting correlations and may contribute to survival models.

This analysis helps identify potentially useful predictors and areas for feature engineering.

---

4. Target Leakage Check

```
Potential Leakage or Strong Predictors:
 fare      0.257307
parch     0.081629
sibsp    -0.035322
age      -0.077221
pclass   -0.338481
Name: survived, dtype: float64
```

Variables like `sex`, `fare`, and `class` showed strong relationships with the target variable. While these are legitimate predictors, derived columns such as `who`, `adult_male`, or `deck` could potentially leak post-event knowledge and should be reviewed carefully before modeling.

---

5. Draft Data Dictionary

```
Data Dictionary (Draft):
```

	Column	Data Type	Missing Values	Unique Values
0	survived	int64	0	2
1	pclass	int64	0	3
2	sex	object	0	2
3	age	float64	177	88
4	sibsp	int64	0	7
5	parch	int64	0	7
6	fare	float64	0	248
7	embarked	object	2	3
8	class	category	0	3
9	who	object	0	3
10	adult_male	bool	0	2
11	deck	category	688	7
12	embark_town	object	2	3
13	alive	object	0	2
14	alone	bool	0	2

This dictionary provides context for understanding each variable and preparing them for modeling.

---

10 Key Insights from EDA

1. **Young adults (20–40 years)** formed the largest age group aboard the Titanic.

2. **Male passengers** were the majority (~65%) but had a **much lower survival rate** than females.

3. **Females had a significantly higher survival rate**, likely due to rescue protocols favoring women and children.

4. **First-class passengers** had the highest survival rate, emphasizing the impact of social class on survival.

5. **Fare paid correlates with survival** – higher fare often meant better accommodations and access to lifeboats.

6. **Children (age < 15)** showed better survival outcomes than older passengers.

7. **Most passengers boarded at Southampton**, but survival rates varied by port.

8. **Age** showed a mild negative correlation with survival – older passengers were less likely to survive.

9. **Class and fare** are both strong indicators and potentially valuable predictors in modeling survival.

10. **Features like 'who', 'deck', and 'adult_male'** may represent **target leakage** if derived after the event or inferred from the target.

Conclusion

The EDA reveals strong socio-demographic influences on survival, especially **gender, class, and fare**. These variables, along with carefully cleaned age and embarked data, should be central in predictive modeling. Care should be taken to avoid using post-event derived variables that could leak information during training.

This analysis lays a solid foundation for the next stage: data preprocessing and model building.