Titanic Dataset - EDA Report

1. Introduction

This report presents an exploratory data analysis (EDA) of the Titanic dataset. The goal is to understand the data and uncover patterns that explain passenger survival.

2. Dataset Overview

The dataset includes variables such as Survival, Pclass (ticket class), Sex, Age, Fare, and more. Initial inspection reveals missing values, particularly in the 'Age' column.

3. Data Quality and Missing Values

Missing values are primarily present in the Age feature. These need to be addressed before building predictive models.

4. Univariate Analysis

- Age: Age distribution shows a concentration of younger passengers.
- Survival: A majority of the passengers did not survive.
- Gender: Female passengers had a notably higher survival rate than males.

5. Bivariate Analysis

- Survival vs Gender: Females had a significantly better chance of survival.
- Survival vs Pclass: Passengers in higher classes had higher survival rates.
- Age vs Pclass: Younger passengers were more common in higher classes.

6. Multivariate Analysis

Titanic Dataset - EDA Report

Pairplots were used to explore relationships among features like Age, Fare, and Pclass. Some separation is evident between survivors and non-survivors across multiple features.

7. Correlation Insights

Correlation heatmap reveals:

- A moderate positive correlation between Fare and Survival.
- A negative correlation between Pclass and Fare.
- Other variables show weaker correlations with survival.

8. Key Observations

- Female passengers had a higher survival rate.
- Younger and higher-class passengers were more likely to survive.
- Fare could be an indirect indicator of survival likelihood.
- Missing Age values should be handled carefully for accurate modeling.

9. Conclusion

The EDA has provided valuable insights into the Titanic dataset. Understanding these relationships is crucial for effective feature engineering and model development.