Titanic Dataset – Exploratory Data Analysis

Objective

To perform Exploratory Data Analysis (EDA) on the Titanic dataset using Python libraries to identify key patterns, trends, and re

Dataset Used

- Source: Kaggle Titanic Challenge
- File: train.csv
- Rows: 891
- Columns: 12
- Target Variable: Survived (0 = No, 1 = Yes)

Tools Used

Python (Jupyter Notebook)
Pandas, NumPy
Seaborn, Matplotlib

Data Overview

- Used .info() to check data types and nulls
- Used .describe() for statistical summary
- Checked missing values: Age and Cabin had missing data

Univariate Analysis

- Survival Count: More passengers died than survived
- Gender Distribution: More males than females
- Class Distribution: Most passengers were in 3rd class
- Age: Right-skewed; most were young adults
- Fare: Right-skewed; most paid low fares

Bivariate Analysis

- Gender vs Survival: Females had much higher survival rate than males
- Class vs Survival: 1st class had highest survival, 3rd class the lowest
- Age vs Survival: Younger passengers had higher survival rate

Multivariate/Correlation

- Created a heatmap to visualize correlations
- Key findings:
  - Pclass negatively correlated with Survived
  - Fare positively correlated with Survived
  - Sex (after encoding) showed strong survival relevance

Key Insights

- Females were prioritized during evacuation → high survival
- First Class passengers had better outcomes than lower classes
- Fare paid was a proxy for social class → correlated with survival
- Younger passengers had slightly higher survival rate

Summary

This analysis demonstrates the value of EDA in:

- Identifying key predictors (Sex, Pclass, Fare)
- Spotting missing values and distributions
- Understanding how different factors influenced survival

Next Steps

- Use these insights for feature engineering
- Build classification models
- Address missing values in Age, Cabin more rigorously

Attachments

- Titanic_EDA_Task5.ipynb
- Titanic_EDA_Task5.pdf (this document)