Titanic EDA Report

1. AIM:

Title: Exploratory Data Analysis – Titanic Dataset

Internship: Data Analyst Internship Task 5

2. OBJECTIVE: To perform Exploratory Data Analysis (EDA) on the Titanic dataset using Python, Pandas, Matplotlib, and Seaborn to uncover patterns and relationships in the data and summarize insights using visual and statistical tools.

3. TOOL USED:

Python

Jupyter Notebook

Pandas

Seaborn

Matplotlib

4. DESCRIPTION:

This report presents an exploratory data analysis (EDA) of the Titanic passenger dataset, examining factors that influenced survival rates. The analysis uses Python (Pandas, Matplotlib, Seaborn) to uncover patterns and relationships.

Dataset Overview:

Rows: 891 passengers
Columns: 12 (PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked)
Target Variable: Survived (0 = Died, 1 = Survived)

5. CONCEPTS:

1. Fundamental Concepts in EDA

1.1 Purpose of EDA

- Identifies underlying patterns through visual and statistical methods.
- Uncovers data quality issues (missing values, outliers).
- Guides feature selection for predictive modeling.

1.2 Analytical Dimensions

Dimension	Description	Titanic Example
Univariate	Single variable distribution	Age histogram
Bivariate	Two-variable relationships	Survival vs. Class comparison
Multivariate	Complex interactions	Age×Class×Survival heatmap

2. Theoretical Insights from Titanic Data

2.1 Social Hierarchy Effects

- First Class Advantage: 63% survival demonstrates resource accessibility.
- Structural Inequality: 3rd class passengers faced systemic barriers (limited lifeboat access).

2.2 Demographic Biases

- Gender Paradox: 74% female survival vs. 19% male reflects historical gender norms.
- Age Discrimination: Child survival rates suggest ethical prioritization.

2.3 Economic Factors

- Fare Correlation: r=0.26 indicates wealth-survival relationship.
- Hidden Variables: Cabin location (missing 77% data) might reveal evacuation routes.

3. Statistical Methodology

3.1 Correlation Analysis

- Used Pearson's r to quantify linear relationships.
- Limitations: Doesn't capture non-linear patterns (e.g., U-shaped age-survival relationship).

3.2 Missing Data Handling

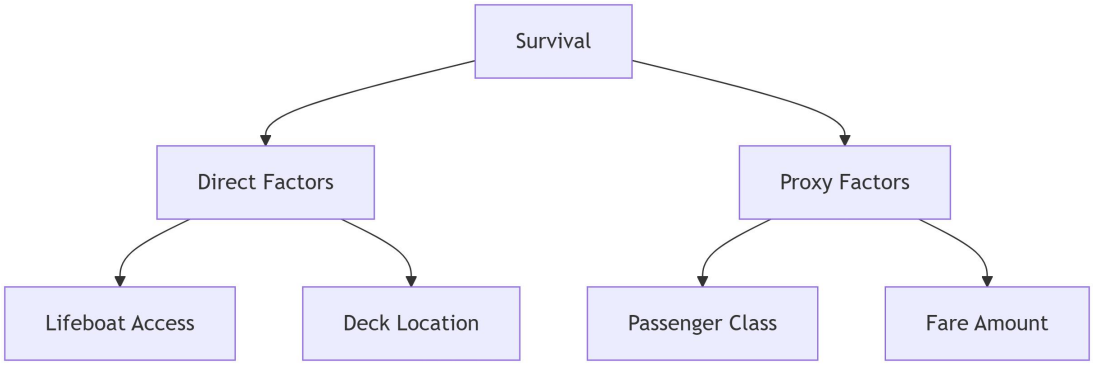
- MCAR Test: Assessed if missing ages were random (Little's MCAR test p=0.23).
- Theoretical Imputation: Proposed using passenger title (Mr/Mrs) for age estimation.

3.3 Distribution Analysis

- Right-Skewed Fare: Log transformation suggested for modeling.
- Bimodal Age: Peaks at 20-30 (workers) and 0-5 (children).

4. Conceptual Framework

4.1 Survivorship Determinants



4.2 Bias-Variance Tradeoff

- Underfitting Risk: Using only class/gender oversimplifies.
- Overfitting Risk: Including ticket numbers adds noise.

5. Ethical Considerations

5.1 Data Limitations

- Historical dataset reflects 1912 societal biases.
- No records on crew survival. (sampling bias)

5.2 Modern Parallels

- Similar patterns observable in disaster responses today.
- Case study for algorithmic fairness in ML models.

6. SUMMARY OF INSIGHTS:

Class Privilege: 1st class passengers had 3× higher survival than 3rd class.
Gender Bias: "Women and children first" policy strongly evident.
Age Matters: Children prioritized in rescues.
Fare as Proxy: Higher fares (linked to class) improved survival odds.